Karnataka State Open University

Mukthagangothri, Mysore – 570 006.

Dept. of Studies and Research in Management

MBA IT Specialization

III Semester

Business Intelligence and Analytics

Block 1

PREFACE

Business intelligence (BI) can be described as “a set of techniques and tools for the acquisition and transformation of raw data into meaningful and useful information for business analysis purposes”. The term “data surfacing” is also more often associated with BI functionality. BI technologies are capable of handling large amounts of structured and sometimes unstructured data to help identify, develop and otherwise create new strategic business opportunities. The goal of BI is to allow for the easy interpretation of these large volumes of data. Identifying new opportunities and implementing an effective strategy based on insights can provide businesses with a competitive market advantage and long-term stability.

BI technologies provide historical, current and predictive views of business operations. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics.

The whole material is organized into four modules each with four units. Each unit lists the objectives of the study along with the relevant questions, illustrations and suggested reading to better understand the concepts.

Wish you happy reading!!!

Karnataka State Open University
Mukthagangothri, Mysore – 570 006.
Dept. of Studies and Research in Management

MBA. IT Specialization

III Semester

BUSINESS INTELLIGENCE AND ANALYTICS

BLOCK 1

BLOCK 1 INTRODUCTION

BI can be used to support a wide range of business decisions ranging from operational to strategic. Basic operating decisions include product positioning or pricing. Strategic business decisions include priorities, goals and directions at the broadest level. In all cases, BI is most effective when it combines data derived from the market in which a company operates (external data) with data from company sources internal to the business such as financial and operations data (internal data). When combined, external and internal data can provide a more complete picture which, in effect, creates an "intelligence" that cannot be derived by any singular set of data. Amongst myriad uses, BI tools empower organisations to gain insight into new markets, assess demand and suitability of products and services for different market segments and gauge the impact of marketing efforts

This block consists of 4 units and is organized as follows:

Unit 1- Applications in an Enterprise, Changing Business Environments and Evolving needs for Decision Support and Analytics, A Framework for Business Intelligence, Transaction Processing versus Analytic Processing, The Nature of Data, A Simple Taxonomy of Data

Unit 2- Levels of Analytics, A Brief Introduction to Big Data Analytics, An Overview of the Analytics Ecosystem

Unit 3- Data Models : Data Modeling and Data Models, The Importance of Data Models**,** Data Model Basic Building Blocks**,** Business Rules**,** The Evolution of Data Models**,** Degrees of Data Abstraction

Unit 4- Definition Of System, Representation of Decision-Making Process, Rationality and Problem Solving, The Decision-Making Process, Types Of Decisions, Definition of Decision Support System Development of a Decision Support System

UNIT 1 : AN OVERVIEW OF BUSINESS INTELLIGENCE, ANALYTICS, AND DATA SCIENCE

Structure

1.0 OBJECTIVES

After studying this unit, you will be able to:

- Discuss applications of Business Intelligence
- Give brief history of Business Intelligence
- Analyze architecture of Business Intelligence
- Distinguish Transaction Processing and Analytical Processing
- Discuss nature of data in Business Intelligence
- Give a simple taxonomy of data

1.1 INTRODUCTION

Business intelligence (BI) can be described as “a set of techniques and tools for the acquisition and transformation of raw data into meaningful and useful information for business analysis purposes”. The term “data surfacing” is also more often associated with BI functionality.

BI technologies are capable of handling large amounts of structured and sometimes unstructured data to help identify, develop and otherwise create new strategic business opportunities. The goal of BI is to allow for the easy interpretation of these large volumes of data. Identifying new opportunities and implementing an effective strategy based on insights can provide businesses with a competitive market advantage and long-term stability. BI technologies provide historical, current and predictive views of business operations. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics.

1.2 APPLICATIONS IN AN ENTERPRISE

BI can be used to support a wide range of business decisions ranging from operational to strategic. Basic operating decisions include product positioning or pricing. Strategic business decisions include priorities, goals and directions at the broadest level. In all cases, BI is most effective when it combines data derived from the market in which a company operates (external data) with data from company sources internal to the business such as financial and operations data (internal data). When combined, external and internal data can provide a more complete picture which, in effect, creates an "intelligence" that cannot be derived by any singular set of data. Amongst myriad uses, BI tools empower organisations to gain insight into new markets, assess demand and suitability of products and services for different market segments and gauge the impact of marketing efforts.

Business intelligence can be applied to the following business purposes, in order to drive business value.

1. Measurement – program that creates a hierarchy of performance metrics and benchmarking that informs business leaders about progress towards business goals (business process management).
2. Analytics – program that builds quantitative processes for a business to arrive at optimal decisions and to perform business knowledge discovery. Frequently involves: data mining, process mining, statistical analysis, predictive analytics, predictive modeling, business process modeling, data lineage, complex event processing and prescriptive analytics.

3. Reporting/enterprise reporting – program that builds infrastructure for strategic reporting to serve the strategic management of a business, not operational reporting. Frequently involves data visualization, executive information system and OLAP.
4. Collaboration/collaboration platform – program that gets different areas (both inside and outside the business) to work together through data sharing and electronic data interchange.
5. Knowledge management – program to make the company data-driven through strategies and practices to identify, create, represent, distribute, and enable adoption of insights and experiences that are true business knowledge. Knowledge management leads to learning management and regulatory compliance.

In addition to the above, business intelligence can provide a pro-active approach, such as alert functionality that immediately notifies the end-user if certain conditions are met. For example, if some business metric exceeds a pre-defined threshold, the metric will be highlighted in standard reports, and the business analyst may be alerted via e-mail or another monitoring service. This end-to-end process requires data governance, which should be handled by the expert.

1.3 CHANGING BUSINESS ENVIRONMENTS AND EVOLVING NEEDS FOR DECISION SUPPORT AND ANALYTICS

From traditional uses in payroll and bookkeeping functions, computerized systems have now penetrated complex managerial areas ranging from the design and management of automated factories to the application of analytical methods for the evaluation of proposed mergers and acquisitions. Nearly all executives know that information technology is vital to their business and extensively use information technologies.

Computer applications have moved from transaction processing and monitoring activities to problem analysis and solution applications, and much of the activity is done with cloud-based technologies, in many cases accessed through mobile devices. Analytics and BI tools such as data warehousing, data mining, online analytical processing (OLAP), dashboards, and the use of the cloud-based systems for decision support are the cornerstones of today's modern management. Managers must have high-speed, networked information systems (wire line or wireless) to assist them with their most important task: making decisions. In many cases, such decisions are routinely being automated, eliminating the need for any managerial intervention.

Besides the obvious growth in hardware, software, and network capacities, some developments have clearly contributed to facilitating growth of decision support and analytics in a number of ways, including the following:

- **Group communication and collaboration.** Many decisions are made today by groups whose members may be in different locations. Groups can collaborate and communicate readily by using collaboration tools as well as the ubiquitous smartphones. Collaboration is especially important along the supply chain, where partners— all the way from vendors to customers—must share information. Assembling a group of decision makers, especially experts, in one place can be costly. Information systems can improve the collaboration process of a group and enable its members to be at different locations (saving travel costs). More critically, such supply chain collaboration permits manufacturers to know about the changing patterns of demand in near real time and thus react to marketplace changes faster.
- **Improved data management.** Many decisions involve complex computations. Data for these can be stored in different databases anywhere in the organization and even possibly outside the organization. The data may include text, sound, graphics, and video, and these can be in different languages. Many times it is necessary to transmit data quickly from distant locations. Systems today can search, store, and transmit needed data quickly, economically, securely, and transparently.
- **Managing giant data warehouses and Big Data.** Large data warehouses (DWs), like the ones perated by Walmart, contain humongous amounts of data. Special methods, including parallel computing, Hadoop/Spark, and so on, are available to organize, search, and mine the data. The costs related to data storage and mining are declining rapidly. Technologies that fall under the broad category of Big Data have enabled massive data coming from a variety of sources and in many different forms, which allows a very different view into organizational performance that was not possible in the past.
- **Analytical support.** With more data and analysis technologies, more alternatives can be evaluated, forecasts can be improved, risk analysis can be performed quickly, and the

views of experts (some of whom may be in remote locations) can be collected quickly and at a reduced cost. Expertise can even be derived directly from analytical systems. With such tools, decision makers can perform complex simulations, check many possible scenarios, and assess diverse impacts quickly and economically. This, of course, is the focus of several chapters in the book.

- **Overcoming cognitive limits in processing and storing information.** According to Simon (1977), the human mind has only a limited ability to process and store information. People sometimes find it difficult to recall and use information in an error-free fashion due to their cognitive limits. The term cognitive limits indicates that an individual's problem-solving capability is limited when a wide range of diverse information and knowledge is required. Computerized systems enable people to overcome their cognitive limits by quickly accessing and processing vast amounts of stored information.
- **Knowledge management.** Organizations have gathered vast stores of information about their own operations, customers, internal procedures, employee interactions, and so forth, through the unstructured and structured communications taking place among the various stakeholders. Knowledge management systems have become sources of formal and informal support for decision making to managers, although sometimes they may not even be called KMS. Technologies such as text analytics and IBM Watson are making it possible to generate value from such knowledge stores.
- **Anywhere, anytime support.** Using wireless technology, managers can access information anytime and from any place, analyze and interpret it, and communicate with those involved. This perhaps is the biggest change that has occurred in the last few years. The speed at which information needs to be processed and converted into decisions has truly changed expectations for both consumers and businesses. These and other capabilities have been driving the use of computerized decision support since the late 1960s, but especially since the mid-1990s. The growth of mobile technologies, social media platforms, and analytical tools has enabled a different level of information systems (IS) support for managers. This growth in providing data-driven support for any decision extends to not just the managers but also to consumers. We will first study an overview

of technologies that have been broadly referred to as BI. From there we will broaden our horizons to introduce various types of analytics.

1.4 A FRAMEWORK FOR BUSINESS INTELLIGENCE

Definitions of BI

Business intelligence (BI) is an umbrella term that combines architectures, tools, databases, analytical tools, applications, and methodologies. It is, like DSS, a content-free expression, so it means different things to different people. Part of the confusion about BI lies in the flurry of acronyms and buzzwords that are associated with it (e.g., business performance management [BPM]). BI's major objective is to enable interactive access (sometimes in real time) to data, to enable manipulation of data, and to give business managers and analysts the ability to conduct appropriate analyses. By analyzing historical and current data, situations, and performances, decision makers get valuable insights that enable them to make more informed and better decisions. The process of BI is based on the transformation of data to information, then to decisions, and finally to actions.

A Brief History of BI

The term BI was coined by the Gartner Group in the mid-1990s. However, as the history in the previous section points out, the concept is much older; it has its roots in the MIS reporting systems of the 1970s. During that period, reporting systems were static, were two dimensional, and had no analytical capabilities. In the early 1980s, the concept of EISs emerged. This concept expanded the computerized support to top-level managers and executives. Some of the capabilities introduced were dynamic multidimensional (ad hoc or on-demand) reporting, forecasting and prediction, trend analysis, drill-down to details, status access, and critical success factors. These features appeared in dozens of commercial products until the mid-1990s. Then the same capabilities and some new ones appeared under the name BI.

Today, a good BI-based enterprise information system contains all the information executives need. So, the original concept of EIS was transformed into BI. By 2005, BI systems started to include artificial intelligence capabilities as well as powerful analytical capabilities. Figure 1.1 illustrates the various tools and techniques that may be included in a BI system.

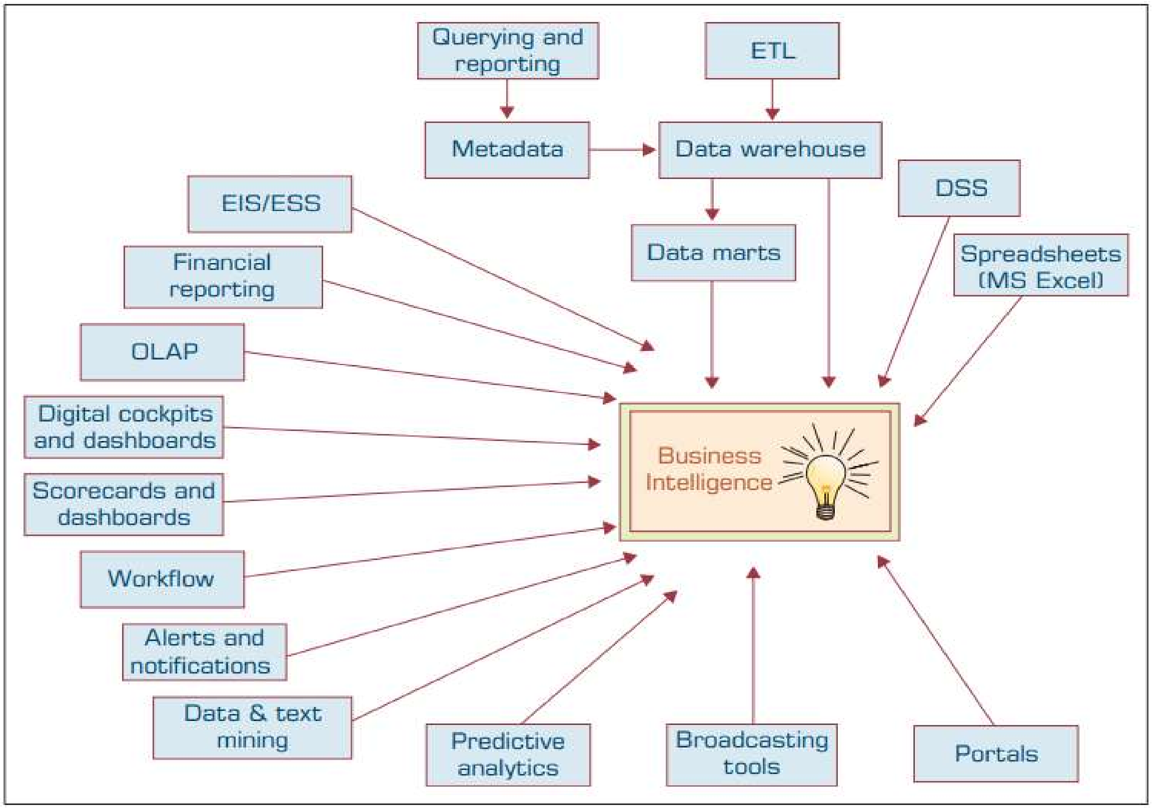


Figure 1.1 Evolution of Business Intelligence (BI)

The Architecture of BI

A BI system has four major components: a DW, with its source data; business analytics, a collection of tools for manipulating, mining, and analyzing the data in the DW; BPM for monitoring and analyzing performance; and a user interface (e.g., a dashboard). The relationship among these components is illustrated in Figure 1.2.

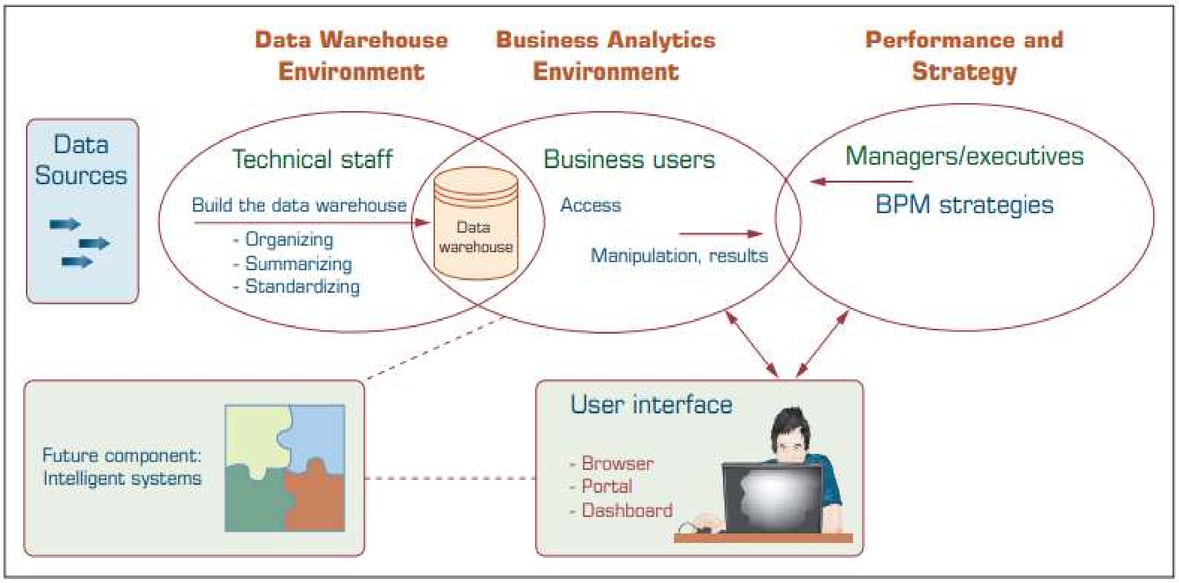


FIGURE 1.2 A High-Level Architecture of BI. (Source: Based on W. Eckerson, Smart Companies in the 21st Century: The Secrets of Creating Successful Business Intelligent Solutions. The Data Warehousing Institute, Seattle, WA, 2003, p. 32, Illustration 5.)

The Origins and Drivers of BI

Where did modern approaches to data warehousing and BI come from? What are their roots, and how do those roots affect the way organizations are managing these initiatives today? Today’s investments in information technology are under increased scrutiny in terms of their bottom-line impact and potential. The same is true of DW and the BI applications that make these initiatives possible. Organizations are being compelled to capture, understand, and harness their data to support decision making to improve business operations. Managers need the right information at the right time and in the right place. This is the mantra for modern approaches to BI.

1.5 TRANSACTION PROCESSING VERSUS ANALYTIC PROCESSING

To illustrate the major characteristics of BI, first we will show what BI is not—namely, transaction processing. We’re all familiar with the information systems that support our transactions, like ATM withdrawals, bank deposits, cash register scans at the grocery store, and so on. These transaction processing systems are constantly involved in handling updates to what we might call operational databases. For example, in an ATM withdrawal transaction, we need to reduce our bank balance accordingly; a bank deposit adds to an account; and a grocery store purchase is likely reflected in the store’s calculation of total sales for the day, and it should reflect an appropriate reduction in the store’s inventory for the items we bought, and so on. These online transaction processing (OLTP) systems handle a company’s routine ongoing business. In contrast, a DW is typically a distinct system that provides storage for data that will be used for analysis. The intent of that analysis is to give management the ability to scour data for information about the business, and it can be used to provide tactical or operational decision support, whereby, for example line personnel can make quicker and/or more informed decisions.

Most operational data in enterprise resources planning (ERP) systems—and in its complementary siblings like supply chain management (SCM) or CRM—are stored in an OLTP system, which is a type of computer processing where the computer responds immediately to user requests. Each request is considered to be a transaction, which is a computerized record of a discrete event, such as the receipt of inventory or a customer order. In other words, a transaction requires a set of two or more database updates that must be completed in an all-or-nothing fashion. The very design that makes an OLTP system efficient for transaction processing makes it inefficient for end-user ad hoc reports, queries, and analysis. In the 1980s, many business users referred to their mainframes as “black holes ” because all the information went into them, but none ever came back. All requests for reports had to be programmed by the IT staff, whereas only “prescanned” reports could be generated on a scheduled basis, and ad hoc real-time querying was virtually impossible.

1.6 THE NATURE OF DATA

Data is the main ingredient for any BI, data science, and business analytics initiative. In fact, it can be viewed as the raw material for what these popular decision technologies produce—information, insight, and knowledge. Without data none of these technologies could exist and be popularized—although, traditionally we have built analytics models using expert knowledge and experience coupled with very little or no data at all; however, those were the old days, and now data is of the essence. Once perceived as a big challenge to collect, store, and manage, data nowadays is widely considered among the most valuable assets of an organization, with the potential to create invaluable insight to better understand customers, competitors, and the business processes.

Data can be small or it can be very large. It can be structured (nicely organized for computers to process), or it can be unstructured (e.g., text that is created for humans and hence not readily understandable/consumable by computers). It can come in smaller batches continuously or it can pour in all at once as a large batch. These are some of the characteristics that define the inherent nature of today’s data, which we often call Big Data. Even though these characteristics of data make it more challenging to process and consume, it also makes it more valuable because it enriches the data beyond its conventional limits, allowing for the discovery of new and novel knowledge.

Traditional ways to manually collect data (either via surveys or via human-entered business transactions) mostly left their places to modern day data collection mechanisms that use Internet and/or sensor/RFID-based computerized networks. These automated data collection systems are not only enabling us to collect more volumes of data but also enhancing the data quality and integrity. Figure 1.3 illustrates a typical analytics continuum—data to analytics to actionable information.

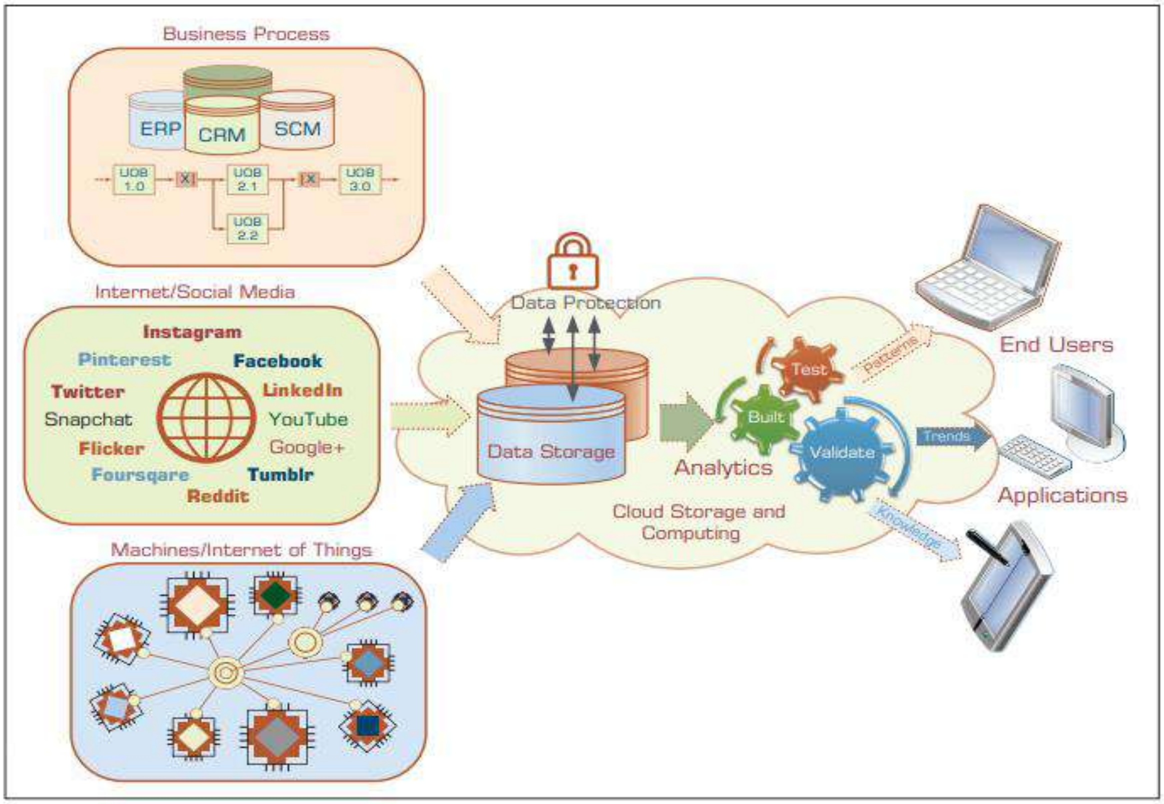


FIGURE 1.3 A Data to Knowledge Continuum.

Although its value proposition is undeniable, to live up its promise, the data has to comply with some basic usability and quality metrics. Not all data is useful for all tasks, obviously. That is, data has to match with (have the coverage of the specifics for) the task for which it is intended to be used. Even for a specific task, the relevant data on hand needs to comply with the quality and quantity requirements. Essentially, data has to be analytics ready. So what does it mean to make data analytics ready?

In addition to its relevancy to the problem at hand and the quality/quantity requirements, it also has to have a certain data structure in place with key fields/variables with properly normalized values. Furthermore, there must be an organization-wide agreed-on definition for common variables and subject matters (sometimes also called master data management), such as how you define a customer (what characteristics of customers are used to produce a holistic enough representation to analytics) and where in the business process the customer related information is captured, validated, stored, and updated.

Sometimes the representation of the data may depend on the type of analytics being employed. Predictive algorithms generally require a flat file with a target variable, so making data analytics ready for prediction means that data sets must be transformed into a flat-file format and made ready for ingestion into those predictive algorithms. It is also imperative to match the data to the needs and wants of a specific predictive algorithm and/or a software tool—for instance, neural network algorithms require all input variables to be numerically represented (even the nominal variables need to be converted into pseudo binary numeric variables) and decision tree algorithms do not require such numerical transformation, easily and natively handling a mix of nominal and numeric variables.

1.7 A Simple Taxonomy of Data

Data (datum in singular form) refers to a collection of facts usually obtained as the result of experiments, observations, transactions, or experiences. Data may consist of numbers, letters, words, images, voice recordings, and so on, as measurements of a set of variables (characteristics of the subject or event that we are interested in studying). Data are often viewed as the lowest level of abstraction from which information and then knowledge is derived. At the highest level of abstraction, one can classify data as structured and unstructured (or semistructured). Unstructured data/semistructured data is composed of any combination of textual, imagery, voice, and Web content. Structured data is what data mining algorithms use and can be classified as categorical or numeric. The categorical data can be subdivided into nominal or ordinal data, whereas numeric data can be subdivided into intervals or ratios. Figure 1.4 shows a simple data taxonomy.

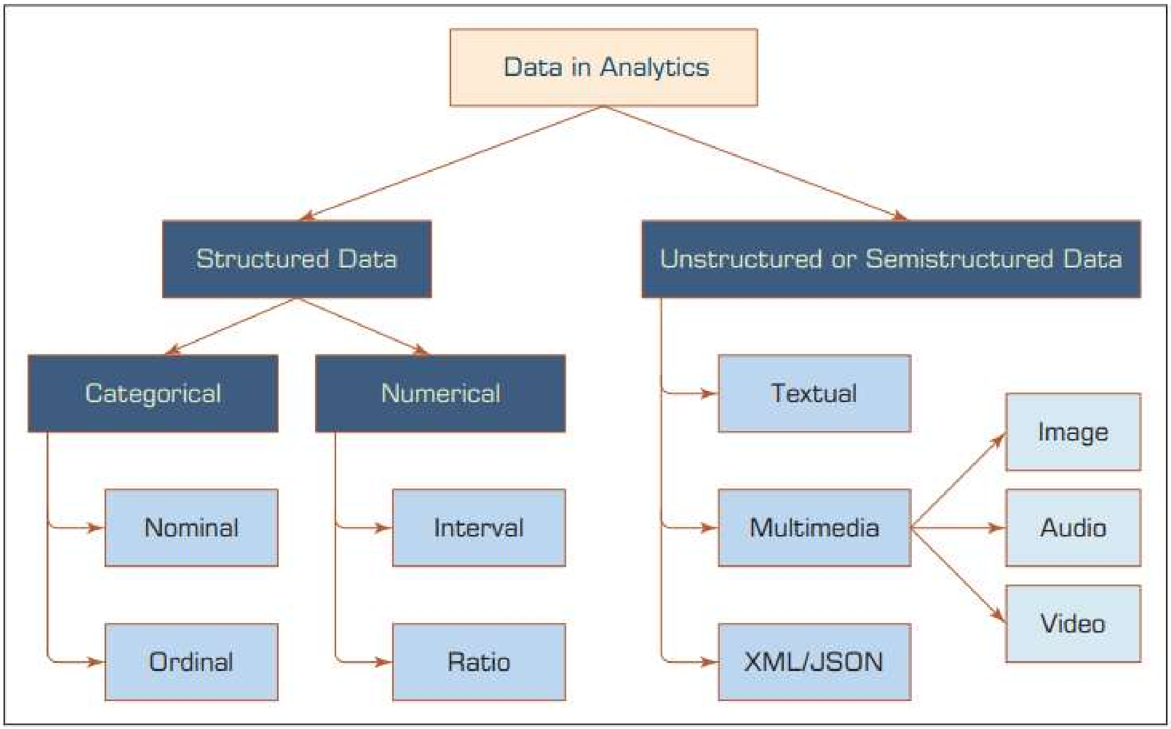


Figure 1.4 A Simple Taxonomy of Data.

- **Categorical data** represent the labels of multiple classes used to divide a variable into specific groups. Examples of categorical variables include race, sex, age group, and educational level. Although the latter two variables may also be considered in a numerical manner by using exact values for age and highest grade completed, it is often more informative to categorize such variables into a relatively small number of ordered classes. The categorical data may also be called discrete data, implying that it represents a finite number of values with no continuum between them. Even if the values used for the categorical (or discrete) variables are numeric, these numbers are nothing more than symbols and do not imply the possibility of calculating fractional values.
- **Nominal data** contain measurements of simple codes assigned to objects as labels, which are not measurements. For example, the variable marital status can be generally categorized as (1) single, (2) married, and (3) divorced. Nominal data can be represented with binomial values having two possible values (e.g., yes/no, true/ false, good/bad), or multinomial values having three or more possible values (e.g., brown/green/blue, white/black/Latino/Asian, single/married/divorced)

- **Ordinal data** contain codes assigned to objects or events as labels that also represent the rank order among them. For example, the variable credit score can be generally categorized as (1) low, (2) medium, or (3) high. Similar ordered relationships can be seen in variables such as age group (i.e., child, young, middle-aged, elderly) and educational level (i.e., high school, college, graduate school). Some predictive analytic algorithms, such as ordinal multiple logistic regression, take into account this additional rank-order information to build a better classification model.
- **Numeric data** represent the numeric values of specific variables. Examples of numerically valued variables include age, number of children, total household income (in U.S. dollars), travel distance (in miles), and temperature (in Fahrenheit degrees). Numeric values representing a variable can be integer (taking only whole numbers) or real (taking also the fractional number). The numeric data may also be called continuous data, implying that the variable contains continuous measures on a specific scale that allows insertion of interim values. Unlike a discrete variable, which represents finite, countable data, a continuous variable represents scalable measurements, and it is possible for the data to contain an infinite number of fractional values.
- **Interval data** are variables that can be measured on interval scales. A common example of interval scale measurement is temperature on the Celsius scale. In this particular scale, the unit of measurement is 1/100 of the difference between the melting temperature and the boiling temperature of water in atmospheric pressure; that is, there is not an absolute zero value.
- Ratio data include measurement variables commonly found in the physical sciences and engineering. Mass, length, time, plane angle, energy, and electric charge are examples of physical measures that are ratio scales. The scale type takes its name from the fact that measurement is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind. Informally, the distinguishing feature of a ratio scale is the possession of a nonarbitrary zero value. For example, the Kelvin temperature scale has a nonarbitrary zero point of absolute zero, which is equal to –273.15 degrees Celsius. This zero point is nonarbitrary because the particles that comprise matter at this temperature have zero kinetic energy.

1.8 CHECK YOUR PROGRESS

1. Define Business Intelligence.
2. List components of Business Intelligence.
3. _______is the main ingredient for any BI, data science, and business analytics initiative.
4. What is Unstructured data/semistructured data?
5. The categorical data is also called as ___________,

Answers to Check your progress

1 Business intelligence (BI) can be described as “a set of techniques and tools for the acquisition and transformation of raw data into meaningful and useful information for business analysis purposes”.

2 A BI system has four major components: a DW, with its source data; business analytics, a collection of tools for manipulating, mining, and analyzing the data in the DW.

3 Data

4 Unstructured data/semistructured data is composed of any combination of textual, imagery, voice, and Web content. Structured data is what data mining algorithms use and can be classified as categorical or numeric.

5 discrete data

1.9 SUMMARY

Data has become one of the most valuable assets of today’s organizations. Data is the main ingredient for any BI, data science, and business analytics initiative. Although its value proposition is undeniable, to live up its promise, the data has to comply with some basic usability and quality metrics. Data (datum in singular form) refers to a collection of facts usually obtained as the result of experiments, observations, transactions, or experiences. At the highest level of abstraction, data can be classified as structured and unstructured. Data in its original/raw form is not usually ready to be useful in analytics tasks.

This unit introduces about business intelligence and its varied applications. Some emerging technologies and their current applications are introduced and detailed. Evolution of BI is elaborated. A complete architecture of BI is detailed. The representation of data and its types used in business analytics are summarized.

1.10 KEYWORDS

- **Strategic business**: A strategic business unit, popularly known as SBU, is a fully-functional unit of a business that has its own vision and direction.
- **Benchmarking**: is defined as the process of measuring products, services, and processes against those of organizations known to be leaders in one or more aspects of their operations.
- **Analytics**: is the process of discovering, interpreting, and communicating significant patterns in data.
- **Knowledge management (KM)**:is the collection of methods relating to creating, sharing, using and managing the knowledge and information of an organization.

1.11 QUESTIONS FOR SELF STUDY

1 Discuss applications of Business Intelligence.

2 Give brief history of Business Intelligence

3 Explain the architecture of Business Intelligence

4 Distinguish Transaction Processing and Analytical Processing

5 Discuss nature of data in Business Intelligence

6 Discuss classification of data in business analytics.

1.12 REFERENCES

1. Sharda R, Delen D, Turban E. Business intelligence analytics and data science: A managerial perspective. Pearson, 2022.
2. Vercellis C. Business intelligence: data mining and optimization for decision making. John Wiley & Sons; 2011 Aug 10.
3. Bentley D. Business Intelligence and Analytics. Internet, link: https://www. pdfdrive. com/business-intelligence-and-analytics-e56416503. html. 2017.

UNIT 2: ANALYTICS OVERVIEW

Structure

2.1 INTRODUCTION

The word analytics has largely replaced the previous individual components of computerized decision support technologies that have been available under various labels in the past. Indeed, many practitioners and academics now use the word analytics in place of BI.

Although many authors and consultants have defined it slightly differently, one can view analytics as the process of developing actionable decisions or recommendations for actions based on insights generated from historical data. According to the Institute for Operations Research and Management Science (INFORMS), analytics represents the combination of computer technology, management science techniques, and statistics to solve real problems. Of course, many other organizations have proposed their own interpretations and motivations for analytics. For example, SAS Institute Inc. proposed eight levels of analytics that begin with standardized reports from a computer system. These reports essentially provide a sense of what is happening with an organization. Additional technologies have enabled us to create more customized reports that can be generated on an ad hoc basis.

The next extension of reporting takes us to OLAP-type queries that allow a user to dig deeper and determine specific sources of concern or opportunities. Technologies available today can also automatically issue alerts for a decision maker when performance warrants such alerts. At a consumer level we see such alerts for weather or other issues. But similar alerts can also be generated in specific settings when sales fall above or below a certain level within a certain time period or when the inventory for a specific product is running low. All of these applications are made possible through analysis and queries on data being collected by an organization. The next level of analysis might entail statistical analysis to better understand patterns. These can then be taken a step further to develop forecasts or models for predicting how customers might respond to a specific marketing campaign or ongoing service/product offerings. When an organization has a good view of what is happening and what is likely to happen, it can also employ other techniques to make the best decisions under the circumstances. These eight levels of analytics are described in more detail in a white paper by SAS (sas.com/news/sascom/analytics_levels.pdf).

2.2 LEVELS OF ANALYTICS

This idea of looking at all the data to understand what is happening, what will happen, and how to make the best of it has also been encapsulated by INFORMS in proposing three levels of analytics. These three levels are identified (informs.org/Community/Analytics) as descriptive, predictive, and prescriptive. Figure 2.1 presents a graphical view of these three levels of analytics. It suggests that these three are somewhat independent steps and one type of analytics applications leads to another. It also suggests that there is actually some overlap across these three types of analytics. In either case, the interconnected nature of different types of analytics applications is evident. We next introduce these three levels of analytics.

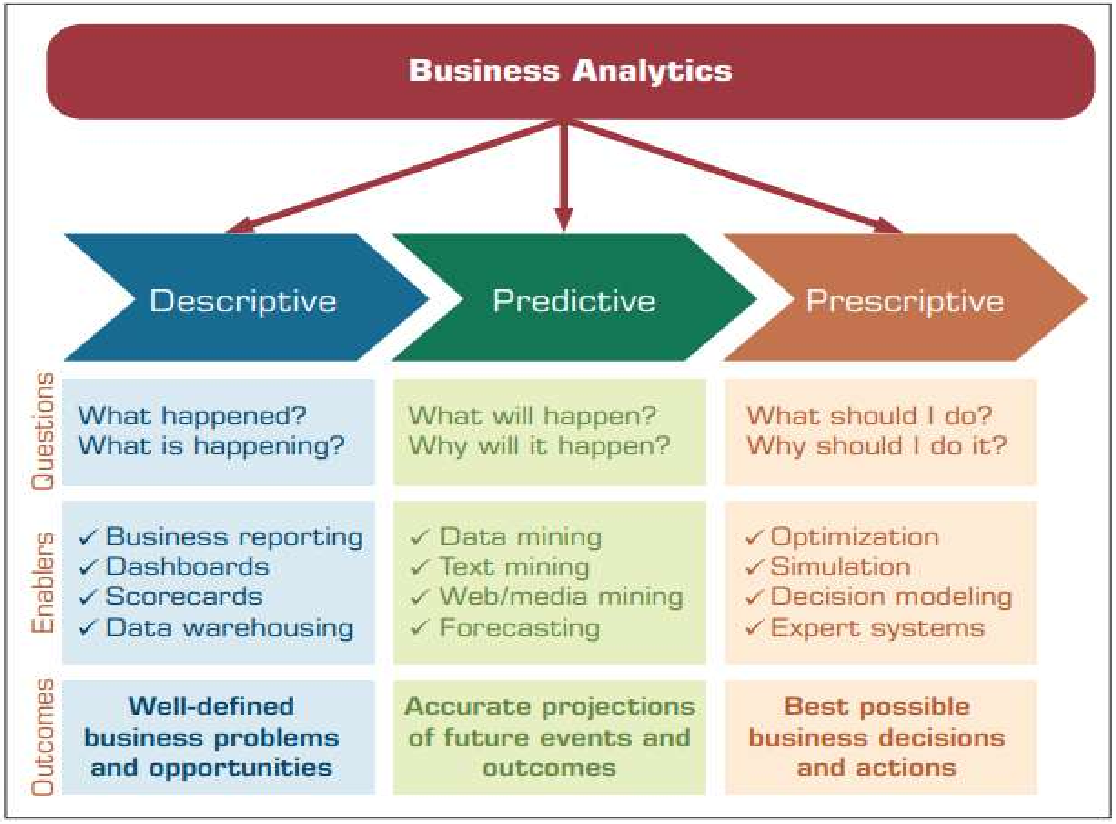


Figure 2.1 Three types of Analytics.

Descriptive Analytics

Descriptive (or reporting) analytics refers to knowing what is happening in the organization and understanding some underlying trends and causes of such occurrences. First, this involves the consolidation of data sources and availability of all relevant data in a form that enables appropriate reporting and analysis. Usually, the development of this data infrastructure is part of DWs. From this data infrastructure we can develop appropriate reports, queries, alerts, and trends using various reporting tools and techniques. A significant technology that has become a key player in this area is visualization. Using the latest visualization tools in the marketplace, we can now develop powerful insights in the operations of our organization.

Predictive Analytics

Predictive analytics aims to determine what is likely to happen in the future. This analysis is based on statistical techniques as well as other more recently developed techniques that fall under the general category of data mining. The goal of these techniques is to be able to predict if the customer is likely to switch to a competitor (“churn”), what the customer would likely buy next and how much, what promotions a customer would respond to, whether this customer is a creditworthy risk, and so forth. A number of techniques are used in developing predictive analytical applications, including various classification algorithms.

We can use classification techniques such as logistic regression, decision tree models, and neural networks to predict how well a motion picture will do at the box office. We can also use clustering algorithms for segmenting customers into different clusters to be able to target specific promotions to them. Finally, we can use association mining techniques to estimate relationships between different purchasing behaviors. That is, if a customer buys one product, what else is the customer likely to purchase? Such analysis can assist a retailer in recommending or promoting related products. For example, any product search on Amazon.com results in the retailer also suggesting other similar products that a customer may be interested in. The goal here is to provide a decision or a recommendation for a specific action. These recommendations can be in the form of a specific yes/no decision for a problem, a specific amount (say, price for a specific item or airfare to charge), or a complete set of production plans. The decisions may be presented to a decision maker in a report or may be used directly in an automated decision rules system (e.g., in airline pricing systems). Thus, these types of analytics can also be termed decision or normative analytics.

Prescriptive Analytics

The third category of analytics is termed prescriptive analytics. The goal of prescriptive analytics is to recognize what is going on as well as the likely forecast and make decisions to achieve the best performance possible. This group of techniques has historically been studied under the umbrella of OR or management sciences and are generally aimed at optimizing the performance of a system. The goal here is to provide a decision or a recommendation for a specific action. These recommendations can be in the form of a specific yes/no decision for a problem, a specific amount (say, price for a specific item or airfare to charge), or a complete set of production plans. The decisions may be presented to a decision maker in a report or may be used directly in an automated decision rules system (e.g., in airline pricing systems). Thus, these types of analytics can also be termed decision or normative analytics.

2.2 A BRIEF INTRODUCTION TO BIG DATA ANALYTICS

What Is Big Data?

Our brains work extremely quickly and efficiently and are versatile in processing large amounts of all kinds of data: images, text, sounds, smells, and video. We process all different forms of data relatively easily. Computers, on the other hand, are still finding it hard to keep up with the pace at which data is generated, let alone analyze it fast. This is why we have the problem of Big Data. So, what is Big Data? Simply put, Big Data is data that cannot be stored in a single storage unit. Big Data typically refers to data that comes in many different forms: structured, unstructured, in a stream, and so forth. Major sources of such data are clickstreams from Web sites, postings on social media sites such as Facebook, and data from traffic, sensors, or weather. A Web search engine like Google needs to search and index billions of Web pages to give you relevant search results in a fraction of a second. Although this is not done in real time, generating an index of all the Web pages on the Internet is not an easy task. Luckily for Google, it was able to solve this problem. Among other tools, it has employed Big Data analytical techniques.

There are two aspects to managing data on this scale: storing and processing. If we could purchase an extremely expensive storage solution to store all this at one place on one unit, making this unit fault tolerant would involve a major expense. An ingenious solution was proposed that involved storing this data in chunks on different machines connected by a network—putting a copy or two of this chunk in different locations on the network, both logically and physically. It was originally used at Google (then called the Google File System) and later developed and released as an Apache project as the Hadoop Distributed File System (HDFS).

However, storing this data is only half the problem. Data is worthless if it does not provide business value, and for it to provide business value, it has to be analyzed. How can such vast amounts of data be analyzed? Passing all computation to one powerful computer does not work; this scale would create a huge overhead on such a powerful computer. Another ingenious solution was proposed: Push computation to the data, instead of pushing data to a computing node. This was a new paradigm and gave rise to a whole new way of processing data. This is what we know today as the MapReduce programming paradigm, which made processing Big Data a reality. MapReduce was originally developed at Google, and a subsequent version was released by the Apache project called Hadoop MapReduce.

Today, when we talk about storing, processing, or analyzing Big Data, HDFS and MapReduce are involved at some level. Other relevant standards and software solutions have been proposed. Although the major toolkit is available as an open source, several companies have been launched to provide training or specialized analytical hardware or software services in this space. Some examples are HortonWorks, Cloudera, and Teradata Aster. Over the past few years, what was called Big Data changed more and more as Big Data applications appeared. The need to process data coming in at a rapid rate added velocity to the equation. An example of fast data processing is algorithmic trading. This uses electronic platforms based on algorithms for trading shares on the financial market, which operates in microseconds. The need to process different kinds of data added variety to the equation. Another example of a wide variety of data is sentiment analysis, which uses various forms of data from social media platforms and customer responses to gauge sentiments. Today, Big Data is associated with almost any kind of large data that has the characteristics of volume, velocity, and variety.

2.3 AN OVERVIEW OF THE ANALYTICS ECOSYSTEM

So you are excited about the potential of analytics and want to join this growing industry. Who are the current players, and what to do they do? Where might you fit in? The objective of this section is to identify various sectors of the analytics industry, provide a classification of different types of industry participants, and illustrate the types of opportunities that exist for analytics professionals.

Eleven different types of players are identified in an analytics ecosystem. An understanding of the ecosystem also gives the reader a broader view of how the various players come together. A secondary purpose of understanding the analytics ecosystem for the BI professional is also to be aware of organizations and new offerings and opportunities in sectors allied with analytics. The section concludes with some observations about the opportunities for professionals to move across these clusters.

Clearly, skill needs can vary between a strong mathematician to a programmer to a modeler to a communicator, and we believe this issue is resolved at a more micro/individual level rather than at a macro level of understanding the opportunity pool. We also take the widest definition of analytics to include all three types as defined by INFORMS—descriptive/reporting/visualization, predictive, and prescriptive as described earlier.

Figure 2.2 illustrates one view of the analytics ecosystem. The components of the ecosystem are represented by the petals of an analytics flower. Eleven key sectors or clusters in the analytics space are identified. The components of the analytics ecosystem are grouped into three categories represented by the inner petals, outer petals, and the seed (middle part) of the flower.

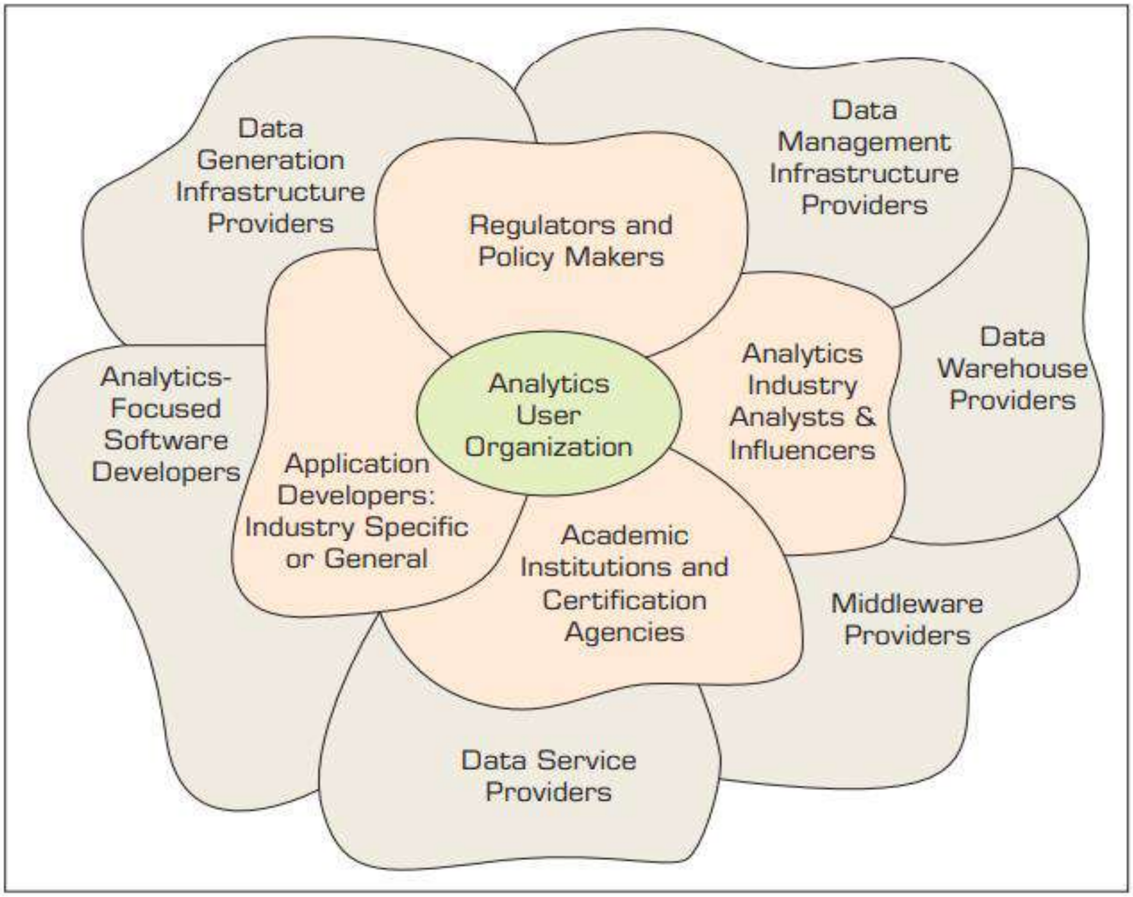

Data Generation Infrastructure Providers
Data Management Infrastructure Providers
Regulators and Policy Makers
Data Warehouse Providers
Analytics Industry Analysts & Influencers
Analytics User Organization
Analytics-Focused Software Developers
Application Developers: Industry Specific or General
Academic Institutions and Certification Agencies
Middleware Providers
Data Service Providers

Figure 2.2 Analytics Ecosystem.

The outer six petals can be broadly termed as the technology providers. Their primary revenue comes from providing technology, solutions, and training to analytics user organizations so they can employ these technologies in the most effective and efficient manner. The inner petals can be generally defined as the analytics accelerators. The accelerators work with both technology providers and users. Finally, the core of the ecosystem comprises the analytics user organizations. This is the most important component, as every analytics industry cluster is driven by the user organizations.

The metaphor of a flower is well-suited for the analytics ecosystem as multiple components overlap each other. Similar to a living organism like a flower, all these petals grow and wither together. We use the terms components, clusters, petals, and sectors interchangeably to describe the various players in the analytics space. We introduce each of the industry sectors next and give some examples of players in each sector. The list of company names included in any petal is not exhaustive. The representative list of companies in each cluster is just to illustrate that cluster’s unique offering to describe where analytics talent may be used or hired away.

Also, mention of a company’s name or its capability in one specific group does not imply that it is the only activity/offering of that organization. The main goal is to focus on the different analytic capabilities within each component of the analytics space. Many companies play in multiple sectors within the analytics industry and thus offer opportunities for movement within the field both horizontally and vertically.

Data Generation Infrastructure Providers

Perhaps the first place to begin identifying the clusters is by noting a new group of companies that enable generating and collection of data that may be used for developing analytical insights. Although this group could include all the traditional point-of-sale systems, inventory management systems, and technology providers for every step in a company’s supply/value chain and operations, we mainly consider new players where the primary focus has been on enabling an organization to develop new insights into its operations as opposed to running its core operations. Thus this group includes companies creating the infrastructure for collecting data from different sources.

One of the emerging components of such an infrastructure is the “sensor”. Sensors collect a massive amount of data at a faster rate and have been adopted by various sectors such as healthcare, sports, and energy. For example, health data collected by the sensors is generally used to track the health status of the users. Some of the major players manufacturing sensors to collect health information are AliveCor, Google, Shimmer, and Fitbit. Likewise, the sports industry is using sensors to collect data from the players and field to develop strategies and improve team play. Examples of the companies producing sports related sensors include Sports Sensors, Zepp, Shockbox, and others. Similarly, sensors are used for traffic management. These help in taking real-time actions to control traffic. Some of the providers are Advantech B+B SmartWorx, Garmin, and Sensys Network.

Sensors play a major role in the Internet of Things and are an essential part of smart objects. These make machine-to-machine communication possible. The leading players in the infrastructure of IoT are Intel, Microsoft, Google, IBM, Cisco, Smartbin, SIKO Products, Omega Engineering, Apple, and SAP. This cluster is probably the most technical group in the ecosystem. We will review an ecosystem for IoT in Chapter 8. Indeed, there is an ecosystem around virtually each of the clusters we identify here.

Data Management Infrastructure Providers

This group includes all of the major organizations that provide hardware and software targeting the basic foundation for all data management solutions. Obvious examples of these include all major hardware players that provide the infrastructure for database computing— IBM, Dell, HP, Oracle, and so on; storage solution providers like EMC (recently bought by Dell) and NetApp; companies providing indigenous hardware and software platforms such as IBM, Oracle, and Teradata; and data solution providers offering hardware and platform independent database management systems like the SQL Server family of Microsoft and specialized integrated software providers such as SAP fall under this group. This group also includes other organizations such as database appliance providers, service providers, integrators, developers, and so on, that support each of these companies’ ecosystems.

Several other companies are emerging as major players in a related space, thanks to the network infrastructure enabling cloud computing. Companies such as Amazon (Amazon Web Services), IBM (Bluemix), and Salesforce.com pioneered to offer full data storage and analytics solutions through the cloud, which now have been adopted by several companies listed earlier.

A recent crop of companies in the Big Data space are also part of this group. Companies such as Cloudera, Hortonworks, and many others do not necessarily offer their own hardware but provide infrastructure services and training to create the Big Data platform. This would include Hadoop clusters, MapReduce, NoSQL, Spark, Kafka, Flume, and other related technologies for analytics. Thus they could also be grouped under industry consultants or trainers enabling the basic infrastructure. Full ecosystems of consultants, software integrators, training providers, and other value-added providers have evolved around many of the large players in the data management infrastructure cluster. Some of the clusters listed below will identify these players because many of them are moving to analytics as the industry shifts its focus from efficient transaction processing to deriving analytical value from the data.

Data Warehouse Providers

Companies with a data warehousing focus provide technology and services aimed toward integrating data from multiple sources, thus enabling organizations to derive and deliver value from its data assets. Many companies in this space include their own hardware to provide efficient data storage, retrieval, and processing. Companies such as IBM, Oracle, and Teradata are major players in this arena.

Recent developments in this space include performing analytics on the data directly in memory. Another major growth sector has been data warehousing in the cloud. Examples of such companies include Snowflake and Redshift. Companies in this cluster clearly work with all the other sector players in providing DW solutions and services within their ecosystem and hence become the backbone of the analytics industry. It has been a major industry in its own right and, thus, a supplier and consumer of analytics talent.

Middleware Providers

Data warehousing began with a focus on bringing all the data stores into an enterprise wide platform. Making sense of this data has become an industry in itself. The general goal of the middleware industry is to provide easy-to-use tools for reporting or descriptive analytics, which forms a core part of BI or analytics employed at organizations. Examples of companies in this space include Micro strategy, Plum, and many others. A few of the major players that were independent middleware players have been acquired by companies in the first two groups.

For example, Hyperion became a part of Oracle, SAP acquired Business Objects, and IBM acquired Cognos. This sector has been largely synonymous with the BI providers offering dash boarding, reporting, and visualization services to the industry, building on top of the transaction processing data and the database and DW providers. Thus many companies have moved into this space over the years, including general analytics software vendors such as SAS or new visualization providers such as Tableau, or many niche application providers. A product directory at TDWI.org lists 201 vendors just in this category (http://www.tdwidirectory.com/category/business-intelligence-services) as of June 2016, so the sector has been robust. This is clearly also the sector attempting to move to a more data science segment of the industry.

Data Service Providers

Much of the data an organization uses for analytics is generated internally through its operations, but there are many external data sources that play a major role in any organization's decision making. Examples of such data sources include demographic data, weather data, data collected by third parties that could inform an organization's decision making, and so on. Several companies realized the opportunity to develop specialized data collection, aggregation, and distribution mechanisms. These companies typically focus on a specific industry sector and build on their existing relationships in that industry through their niche platforms and services for data collection.

For example, Nielsen provides data sources to their clients on customer retail purchasing behavior. Another example is Experian, which includes data on each household in the United States. Omniture has developed technology to collect Web clicks and share such data with their clients. Comscore is another major company in this space. Google compiles data for individual Web sites and makes a summary available through Google Analytics services. Other examples are Equifax, TransUnion, Acxiom, Merkle, Epsilon, and Avention. This can also include organizations such as ESRI.org, which provides location-oriented data to their customers.

There are hundreds of other companies that are developing niche platforms and services to collect, aggregate, and share such data with their clients. As noted earlier, many industry-specific data aggregators and distributors exist and are moving to offer their own analytics services. Thus this sector is also a growing user and potential supplier of analytics talent, especially with specific niche expertise.

Understand the nature of data as it relates to business intelligence (BI) and analytics

- Learn the methods used to make real world data analytics ready
- Describe statistical modeling and its relationship to business analytics
- Learn about descriptive and inferential statistics
- Define business reporting, and understand its historical evolution
- Understand the importance of data/ information visualization
- Learn different types of visualization techniques
- Appreciate the value that visual analytics brings to business analytics
- Know the capabilities and limitations of dashboards

In the age of Big Data and business analytics in which we are living, the importance of data is undeniable. The newly coined phrases like “data is the oil,” “data is the new bacon,” “data is the new currency,” and “data is the king” are further stressing the renewed importance of data. But what type of data are we talking about? Obviously, not just any data. The “garbage in garbage out—GIGO” concept/principle applies to today’s “Big Data” phenomenon more so than any data definition that we have had in the past. To live up to its promise, its value proposition, and its ability to turn into insight, data has to be carefully created/identified, collected, integrated, cleaned, transformed, and properly contextualized for use in accurate and timely decision making.

2.5 CHECK YOUR PROGRESS

1. Define Analytics.
2. Mention levels of analytics.
3. Distinguish Predictive Analytics and Prescriptive Analytics.
4. List the technologies used for analytics.

Answers to Check your progress

1. Analytics is the process of developing actionable decisions or recommendations for actions based on insights generated from historical data.
2. descriptive, predictive, and prescriptive
3. Predictive analytics is based on statistical techniques as well as other more recently developed techniques that fall under the general category of data mining. The goal of prescriptive analytics is to recognize what is going on as well as the likely forecast and make decisions to achieve the best performance possible.
4. Hadoop clusters, MapReduce, NoSQL, Spark, Kafka, Flume, are the technologies used for analytics.

2.6 SUMMARY

This unit introduces about analytics its varied applications. Different types of business analytics are detailed with varied applications. The significance of big data is analyzed with its distinguishing feature. A brief summary about analytics ecosystem is given.

2.7 KEYWORDS

- **Analytics:** is the process of discovering, interpreting, and communicating significant patterns in data.
- **Descriptive Analytics**: is the examination of data or content, usually manually performed, to answer the question "What happened?" (or What is happening?), characterized by traditional business intelligence (BI) and visualizations such as pie charts, bar charts, line graphs, tables, or generated narratives.
- **Predictive analytics**: is a branch of advanced analytics that makes predictions about future outcomes using historical data combined with statistical modeling, data mining techniques and machine learning.
- **Prescriptive Analytics** : is a form of advanced analytics which examines data or content to answer the question "What should be done?" or "What can we do to make _______ happen?", and is characterized by techniques such as graph analysis, simulation, complex

event processing, neural networks, recommendation engines, heuristics, and machine learning.

- **Big data:** is a combination of structured, semistructured and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modeling and other advanced analytics applications.

2.8 QUESTIONS FOR SELF STUDY

1. Give a brief account on analytics.
2. Describe different types of analytics.
3. What is Big Data? Write about its significance.
4. Give a brief overview of analytics ecosystem

2.9 REFERENCES

1. Sharda R, Delen D, Turban E. Business intelligence analytics and data science: A managerial perspective. Pearson, 2022.
2. Vercellis C. Business intelligence: data mining and optimization for decision making. John Wiley & Sons; 2011 Aug 10.
3. Bentley D. Business Intelligence and Analytics. Internet, link: https://www. pdfdrive. com/business-intelligence-and-analytics-e56416503. html. 2017.

UNIT 3: BUSINESS INTELLIGENCE ARCHITECTURE AND MODELS

Structure

3.0 OBJECTIVES

After studying this unit, you will be able to:

- ✓ Analyze significance of Business Intelligence.
- ✓ Discuss role of mathematical models.
- ✓ Differentiate data, information and knowledge.
- ✓ Develop a Business Intelligence System.
- ✓ Elucidate ethics and business intelligence.
- ✓ Discuss major building blocks of business intelligence system.
- ✓ Explain the architecture of business intelligence system.

3.1 INTRODUCTION

The advent of low-cost data storage technologies and the wide availability of Internet connections have made it easier for individuals and organizations to access large amounts of data. Such data are often heterogeneous in origin, content and representation, as they include commercial, financial and administrative transactions, web navigation paths, emails, texts and hypertexts, and the results of clinical tests, to name just a few examples. Their accessibility opens up promising scenarios and opportunities, and raises an enticing question: Is it possible to convert such data into information and knowledge that can then be used by decision makers to aid and improve the governance of enterprises and of public administration?

Business intelligence may be defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes. This unit will describe in general terms the problems entailed in business intelligence, highlighting the interconnections with other disciplines and identifying the primary components typical of a business intelligence environment.

3.2 EFFECTIVE AND TIMELY DECISIONS

In complex organizations, public or private, decisions are made on a continual basis. Such decisions may be more or less critical, have long or short-term effects and involve people and roles at various hierarchical levels. The ability of these knowledge workers to make decisions, both as individuals and as a community, is one of the primary factors that influence the performance and competitive strength of a given organization.

Most knowledge workers reach their decisions primarily using easy and intuitive methodologies, which take into account specific elements such as experience, knowledge of the application domain and the available information. This approach leads to a stagnant decision-making style which is inappropriate for the unstable conditions determined by frequent and rapid changes in the economic environment. Indeed, decision-making processes within today’s organizations are often too complex and dynamic to be effectively dealt with through an intuitive approach, and require instead a more rigorous attitude based on analytical methodologies and mathematical models.

The main purpose of business intelligence systems is to provide knowledge workers with tools and methodologies that allow them to make effective and timely decisions.

Effective decisions: The application of rigorous analytical methods allows decision makers to rely on information and knowledge which are more dependable. As a result, they are able to make better decisions and devise action plans that allow their objectives to be reached in a more effective way. Indeed, turning to formal analytical methods forces decision makers to explicitly describe both the criteria for evaluating alternative choices and the mechanisms regulating the problem under investigation. Furthermore, the ensuing in-depth examination and thought lead to a deeper awareness and comprehension of the underlying logic of the decision-making process.

Timely decisions: Enterprises operate in economic environments characterized by growing levels of competition and high dynamism. As a consequence, the ability to rapidly react to the actions of competitors and to new market conditions is a critical factor in the success or even the survival of a company.

Figure 3.1 illustrates the major benefits that a given organization may draw from the adoption of a business intelligence system. When facing problems, decision makers ask themselves a series of questions and develop the corresponding analysis. Hence, they examine and compare several options, selecting among them the best decision, given the conditions at hand.

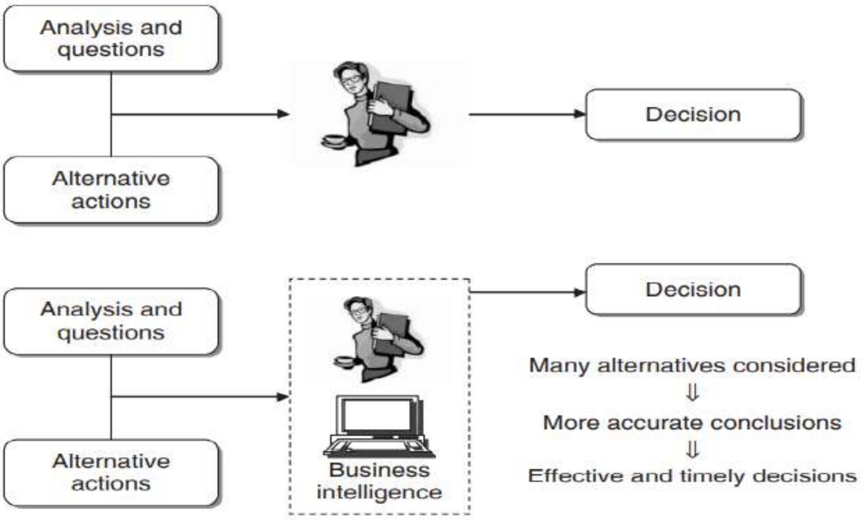


Figure 3.1 Benefits of a business intelligence system

If decision makers can rely on a business intelligence system facilitating their activity, we can expect that the overall quality of the decision-making process will be greatly improved. With the help of mathematical models and algorithms, it is actually possible to analyze a larger number of alternative actions, achieve more accurate conclusions and reach effective and timely decisions. We may therefore conclude that the major advantage deriving from the adoption of a business intelligence system is found in the increased effectiveness of the decision-making process.

3.3 DATA, INFORMATION AND KNOWLEDGE

As observed above, a vast amount of data has been accumulated within the information systems of public and private organizations. These data originate partly from internal transactions of an administrative, logistical and commercial nature and partly from external sources. However, even if they have been gathered and stored in a systematic and structured way, these data cannot be used directly for decision-making purposes. They need to be processed by means of appropriate extraction tools and analytical methods capable of transforming them into information and knowledge that can be subsequently used by decision makers. The difference between data, information and knowledge can be better understood through the following remarks.

Data: Generally, data represent a structured codification of single primary entities, as well as of transactions involving two or more primary entities.

For example, for a retailer data refer to primary entities such as customers, points of sale and items, while sales receipts represent the commercial transactions.

Information: Information is the outcome of extraction and processing activities carried out on data, and it appears meaningful for those who receive it in a specific domain.

For example, to the sales manager of a retail company, the proportion of sales receipts in the amount of over ¤100 per week, or the number of customers holding a loyalty card who have reduced by more than 50% the monthly amount spent in the last three months, represent meaningful pieces of information that can be extracted from raw stored data.

Knowledge: Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions. Therefore, we can think of knowledge as consisting of information put to work into a specific domain, enhanced by the experience and competence of decision makers in tackling and solving complex problems.

For a retail company, a sales analysis may detect that a group of customers, living in an area where a competitor has recently opened a new point of sale, have reduced their usual amount of business. The knowledge extracted in this way will eventually lead to actions aimed at solving the problem detected, for example by introducing a new free home delivery service for the customers residing in that specific area. We wish to point out that knowledge can be extracted from data both in a passive way, through the analysis criteria suggested by the decision makers, or through the active application of mathematical models, in the form of inductive learning or optimization.

Several public and private enterprises and organizations have developed in recent years formal and systematic mechanisms to gather, store and share their wealth of knowledge, which is now perceived as an invaluable intangible asset. The activity of providing support to knowledge workers through the integration of decision-making processes and enabling information technologies is usually referred to as knowledge management.

It is apparent that business intelligence and knowledge management share some degree of similarity in their objectives. The main purpose of both disciplines is to develop environments that can support knowledge workers in decision-making processes and complex problem-solving activities. To draw a boundary between the two approaches, we may observe that knowledge management methodologies primarily focus on the treatment of information that is usually unstructured, at times implicit, contained mostly in documents, conversations and past experience. Conversely, business intelligence systems are based on structured information, most often of a quantitative nature and usually organized in a database. However, this distinction is a somewhat fuzzy one: for example, the ability to analyze emails and web pages through text mining methods progressively induces business intelligence systems to deal with unstructured information.

3.4 THE ROLE OF MATHEMATICAL MODELS

A business intelligence system provides decision makers with information and knowledge extracted from data, through the application of mathematical models and algorithms. In some instances, this activity may reduce to calculations of totals and percentages, graphically represented by simple histograms, whereas more elaborate analyses require the development of advanced optimization and learning models.

In general terms, the adoption of a business intelligence system tends to promote a scientific and rational approach to the management of enterprises and complex organizations. Even the use of a spreadsheet to estimate the effects on the budget of fluctuations in interest rates, despite its simplicity, forces decision makers to generate a mental representation of the financial flows process.

Classical scientific disciplines, such as physics, have always resorted to mathematical models for the abstract representation of real systems. Other disciplines, such as operations research, have instead exploited the application of scientific methods and mathematical models to the study of artificial systems, for example public and private organizations.

The rational approach typical of a business intelligence analysis can be summarized schematically in the following main characteristics.

- First, the objectives of the analysis are identified and the performance indicators that will be used to evaluate alternative options are defined.
- Mathematical models are then developed by exploiting the relationships among system control variables, parameters and evaluation metrics.
- Finally, what-if analyses are carried out to evaluate the effects on the performance determined by variations in the control variables and changes in the parameters.

Although their primary objective is to enhance the effectiveness of the decision-making process, the adoption of mathematical models also affords other advantages, which can be appreciated particularly in the long term. First, the development of an abstract model forces decision makers to focus on the main features of the analyzed domain, thus inducing a deeper understanding of the phenomenon under investigation. Furthermore, the knowledge about the domain acquired when building a mathematical model can be more easily transferred in the long run to other individuals within the same organization, thus allowing a sharper preservation of knowledge in comparison to empirical decision-making processes. Finally, a mathematical model developed for a specific decision-making task is so general and flexible that in most cases it can be applied to other ensuing situations to solve problems of similar type.

3.5 BUSINESS INTELLIGENCE ARCHITECTURES

The architecture of a business intelligence system, depicted in Figure 3.2, includes three major components.

Data sources: In a first stage, it is necessary to gather and integrate the data stored in the various primary and secondary sources, which are heterogeneous in origin and type. The sources consist for the most part of data belonging to operational systems, but may also include unstructured documents, such as emails and data received from external providers.

Data warehouses and data marts: Using extraction and transformation tools known as *extract, transform*, *load* (ETL), the data originating from the different sources are stored in databases intended to support business intelligence analyses. These databases are usually referred to as *data warehouses* and *data marts*, and they will be the subject of Chapter 3.

Business intelligence methodologies: Data are finally extracted and used to feed mathematical models and analysis methodologies intended to support decision makers.

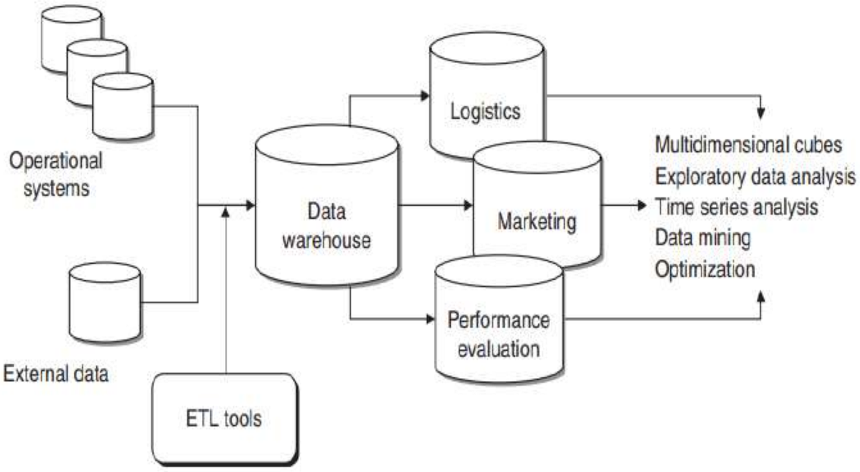


Figure 3.2 A typical Business Intelligence Architecture

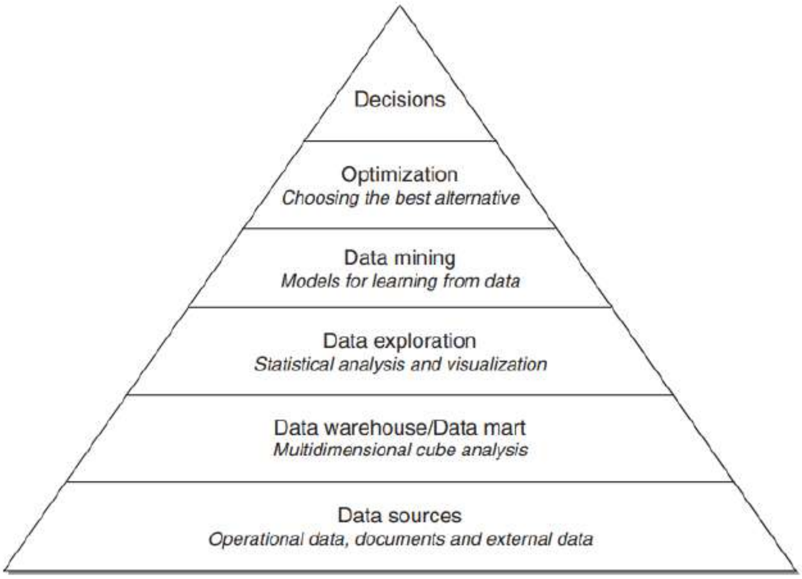

Decisions
Optimization
Choosing the best alternative
Data mining
Models for learning from data
Data exploration
Statistical analysis and visualization
Data warehouse/Data mart
Multidimensional cube analysis
Data sources
Operational data, documents and external data

Figure 3.3 The main components of a business intelligence system

The pyramid in Figure 3.3 shows the building blocks of a business intelligence system. So far, we have seen the components of the first two levels when discussing Figure 3.2. We now turn to the description of the upper tiers.

Data exploration: At the third level of the pyramid we find the tools for performing a *passive* business intelligence analysis, which consist of query and reporting systems, as well as statistical methods. These are referred to as passive methodologies because decision makers are requested to generate prior hypotheses or define data extraction criteria, and then use the analysis tools to find answers and confirm their original insight. For instance, consider the sales manager of a company who notices that revenues in a given geographic area have dropped for a specific group of customers. Hence, she might want to bear out her hypothesis by using extraction and visualization tools, and then apply a statistical test to verify that her conclusions are adequately supported by data. Statistical techniques for exploratory data analysis will be described in Chapters 6 and 7.

Data mining: The fourth level includes *active* business intelligence methodologies, whose purpose is the extraction of information and knowledge from data. These include mathematical models for pattern recognition, machine learning and data mining techniques, which will be dealt with in Part II of this book. Unlike the tools described at the previous level of the pyramid, the models of an active kind do not require decision makers to formulate any prior hypothesis to be later verified. Their purpose is instead to expand the decision makers' knowledge.

Optimization: By moving up one level in the pyramid we find optimization models that allow us to determine the best solution out of a set of alternative actions, which is usually fairly extensive and sometimes even infinite.

Decisions: Finally, the top of the pyramid corresponds to the choice and the actual adoption of a specific decision, and in some way represents the natural conclusion of the decision-making process. Even when business intelligence methodologies are available and successfully adopted, the choice of a decision pertains to the decision makers, who may also take advantage of informal and unstructured information available to adapt and modify the recommendations and the conclusions achieved through the use of mathematical models.

As we progress from the bottom to the top of the pyramid, business intelligence systems offer increasingly more advanced support tools of an active type. Even roles and competencies change. At the bottom, the required competencies are provided for the most part by the information systems specialists within the organization, usually referred to as database administrators. Analysts and experts in mathematical and statistical models are responsible for the intermediate phases. Finally, the activities of decision makers responsible for the application domain appear dominant at the top.

As described above, business intelligence systems address the needs of different types of complex organizations, including agencies of public administration and associations. However, if we restrict our attention to enterprises, business intelligence methodologies can be found mainly within three departments of a company, as depicted in Figure 3.4: marketing and sales; logistics and production; accounting and control.

3.6 CYCLE OF A BUSINESS INTELLIGENCE ANALYSIS

Each business intelligence analysis follows its own path according to the application domain, the personal attitude of the decision makers and the available analytical methodologies. However, it is possible to identify an ideal cyclical path characterizing the evolution of a typical business intelligence analysis, as shown in Figure 3.5, even though differences still exist based upon the peculiarity of each specific context.

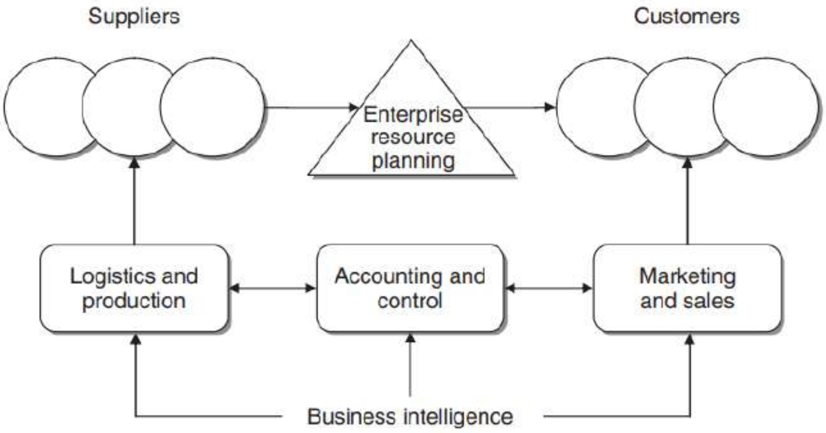


Figure 3.4 Departments of an enterprise concerned with business intelligence systems

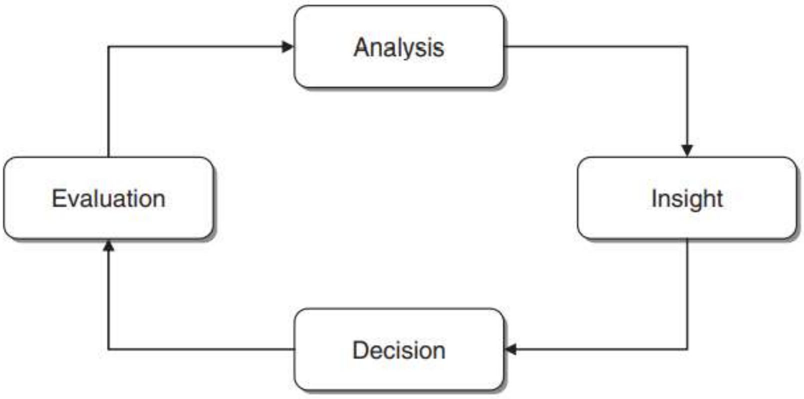


Figure 3.5 Cycle of a business intelligence analysis

Analysis: During the analysis phase, it is necessary to recognize and accurately spell out the problem at hand. Decision makers must then create a mental representation of the phenomenon being analyzed, by identifying the critical factors that are perceived as the most relevant. The availability of business intelligence methodologies may help already in this stage, by permitting decision makers to rapidly develop various paths of investigation. Thus, the first phase in the business intelligence cycle leads decision makers to ask several questions and to obtain quick responses in an interactive way.

Insight: The second phase allows decision makers to better and more deeply understand the problem at hand, often at a causal level. For instance, if the analysis carried out in the first phase shows that a large number of customers are discontinuing an insurance policy upon yearly expiration, in the second phase it will be necessary to identify the profile and characteristics shared by such customers. The information obtained through the analysis phase is then transformed into knowledge during the insight phase. On the one hand, the extraction of knowledge may occur due to the intuition of the decision makers and therefore be based on their experience and possibly on unstructured information available to them. On the other hand, inductive learning models may also prove very useful during this stage of analysis, particularly when applied to structured data.

Decision: During the third phase, knowledge obtained as a result of the insight phase is converted into decisions and subsequently into actions. The availability of business intelligence methodologies allows the analysis and insight phases to be executed more rapidly so that more effective and timely decisions can be made that better suit the strategic priorities of a given organization. This leads to an overall reduction in the execution time of the analysis– decision– action– revision cycle, and thus to a decision-making process of better quality.

Evaluation: Finally, the fourth phase of the business intelligence cycle involves performance measurement and evaluation. Extensive metrics should then be devised that are not exclusively limited to the financial aspects but also take into account the major performance indicators defined for the different company departments.

3.7 ENABLING FACTORS IN BUSINESS INTELLIGENCE PROJECTS

Some factors are more critical than others to the success of a business intelligence project: technologies, analytics and human resources.

Technologies: Hardware and software technologies are significant enabling factors that have facilitated the development of business intelligence systems within enterprises and complex organizations. On the one hand, the computing capabilities of microprocessors have increased on average by 100% every 18 months during the last two decades, and prices have fallen. This trend has enabled the use of advanced algorithms which are required to employ inductive learning methods and optimization models, keeping the processing times within a reasonable range. Moreover, it permits the adoption of state-of-the-art graphical visualization techniques, featuring real-time animations. A further relevant enabling factor derives from the exponential increase in the capacity of mass storage devices, again at decreasing costs, enabling any organization to store terabytes of data for business intelligence systems. And network connectivity, in the form of Extranets or Intranets, has played a primary role in the diffusion within organizations of information and knowledge extracted from business intelligence systems. Finally, the easy integration of hardware and software purchased by different suppliers, or developed internally by an organization, is a further relevant factor affecting the diffusion of data analysis tools.

Analytics: As stated above, mathematical models and analytical methodologies play a key role in information enhancement and knowledge extraction from the data available inside most organizations. The mere visualization of the data according to timely and flexible logical views, plays a relevant role in facilitating the decision-making process, but still represents a passive form of support. Therefore, it is necessary to apply more advanced models of inductive learning and optimization in order to achieve active forms of support for the decision-making process.

Human resources: The human assets of an organization are built up by the competencies of those who operate within its boundaries, whether as individuals or collectively. The overall knowledge possessed and shared by these individuals constitutes the organizational culture. The ability of knowledge workers to acquire information and then translate it into practical actions is one of the major assets of any organization, and has a major impact on the quality of the decision-making process. If a given enterprise has implemented an advanced business intelligence system, there still remains much scope to emphasize the personal skills of its knowledge workers, who are required to perform the analyses and to interpret the results, to work out creative solutions and to devise effective action plans. All the available analytical tools being equal, a company employing human resources endowed with a greater mental agility and willing to accept changes in the decision-making style will be at an advantage over its competitors.

3.8 DEVELOPMENT OF A BUSINESS INTELLIGENCE SYSTEM

The development of a business intelligence system can be assimilated to a project, with a specific final objective, expected development times and costs, and the usage and coordination of the resources needed to perform planned activities. Figure 3.6 shows the typical development cycle of a business intelligence architecture. Obviously, the specific path followed by each organization might differ from that outlined in the figure. For instance, if the basic information structures, including the data warehouse and the data marts, are already in place, the corresponding phases indicated in Figure 3.6 will not be required.

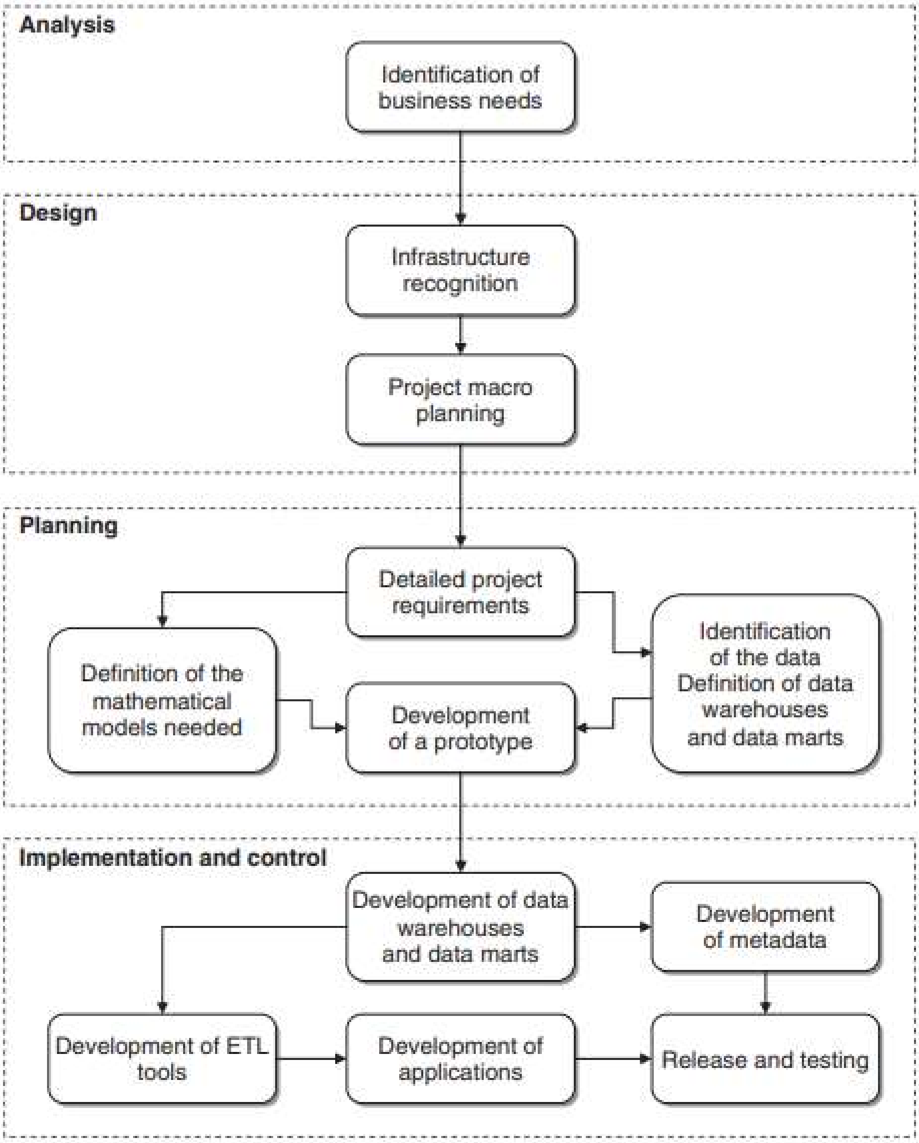

Analysis
Identification of business needs
Design
Infrastructure recognition
Project macro planning
Planning
Detailed project requirements
Definition of the mathematical models needed
Development of a prototype
Identification of the data Definition of data warehouses and data marts
Implementation and control
Development of data warehouses and data marts
Development of metadata
Development of ETL tools
Development of applications
Release and testing

Figure 3.6 Phases in the development of a business intelligence system

Analysis: During the first phase, the needs of the organization relative to the development of a business intelligence system should be carefully identified. This preliminary phase is generally conducted through a series of interviews of knowledge workers performing different roles and activities within the organization. It is necessary to clearly describe the general objectives and priorities of the project, as well as to set out the costs and benefits deriving from the development of the business intelligence system.

Design: The second phase includes two sub-phases and is aimed at deriving a provisional plan of the overall architecture, taking into account any development in the near future and the evolution of the system in the mid term. First, it is necessary to make an assessment of the existing information infrastructures. Moreover, the main decision-making processes that are to be supported by the business intelligence system should be examined, in order to adequately determine the information requirements. Later on, using classical project management methodologies, the project plan will be laid down, identifying development phases, priorities, expected execution times and costs, together with the required roles and resources.

Planning: The planning stage includes a sub-phase where the functions of the business intelligence system are defined and described in greater detail. Subsequently, existing data as well as other data that might be retrieved externally are assessed. This allows the information structures of the business intelligence architecture, which consist of a central data warehouse and possibly some satellite data marts, to be designed. Simultaneously with the recognition of the available data, the mathematical models to be adopted should be defined, ensuring the availability of the data required to feed each model and verifying that the efficiency of the algorithms to be utilized will be adequate for the magnitude of the resulting problems. Finally, it is appropriate to create a system prototype, at low cost and with limited capabilities, in order to uncover beforehand any discrepancy between actual needs and project specifications.

Implementation and control: The last phase consists of five main sub-phases. First, the data warehouse and each specific data mart are developed. These represent the information infrastructures that will feed the business intelligence system. In order to explain the meaning of the data contained in the data warehouse and the transformations applied in advance to the primary data, a *metadata* archive should be created. Moreover, ETL procedures are set out to extract and transform the data existing in the primary sources, loading them into the data warehouse and the data marts. The next step is aimed at developing the core business intelligence applications that allow the planned analyses to be carried out. Finally, the system is released for test and usage.

Figure 3.7 provides an overview of the main methodologies that may be included in a business intelligence system, most of which will be described in the following chapters. Some of them have a methodological nature and can be used across different application domains, while others can only be applied to specific tasks.

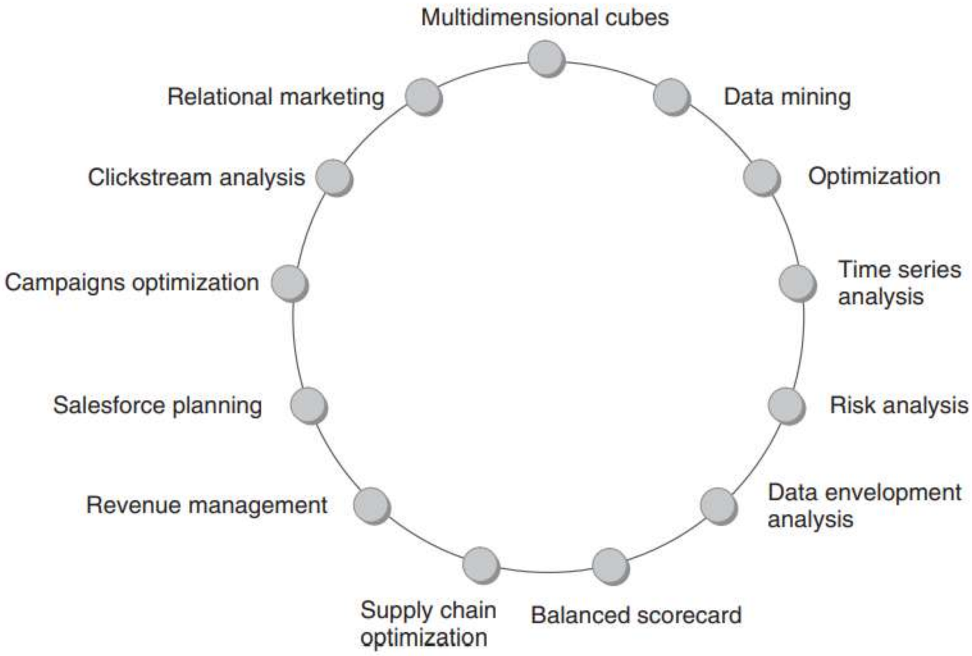


Figure 3.7 Portfolio of available methodologies in a business intelligence system

3.9 ETHICS AND BUSINESS INTELLIGENCE

The adoption of business intelligence methodologies, data mining methods and decision support systems raises some ethical problems that should not be over-looked. Indeed, the progress towards the information and Knowledge society opens up countless opportunities, but may also generate distortions and risks which should be prevented and avoided by using adequate control rules and mechanisms. Usage of data by public and private organizations that is improper and does not respect the individuals' right to privacy should not be tolerated. More generally, we must guard against the excessive growth of the political and economic power of enterprises allowing the transformation processes out- lined above to exclusively and unilaterally benefit such enterprises themselves, at the expense of consumers, workers and inhabitants of the Earth ecosystem.

However, even failing specific regulations that would prevent the abuse of data gathering and invasive investigations, it is essential that business intelligence analysts and decision makers abide by the ethical principle of respect for the personal rights of the individuals. The risk of overstepping the boundary between correct and intrusive use of information is particularly high within the relational marketing and web mining fields, described in Chapter 13. For example, even if disguised under apparently inoffensive names such as ‘data enrichment’, private information on individuals and households does circulate, but that does not mean that it is ethical for decision makers and enterprises to use it.

Respect for the right to privacy is not the only ethical issue concerning the use of business intelligence systems. There has been much discussion in recent years of the social responsibilities of enterprises, leading to the introduction of the new concept of *stakeholders*. This term refers to anyone with any interest in the activities of a given enterprise, such as investors, employees, labor unions and civil society as a whole. There is a diversity of opinion on whether a company should pursue the short-term maximization of profits, acting exclusively in the interest of shareholders, or should instead adopt an approach that takes into account the social consequences of its decisions.

As this is not the right place to discuss a problem of such magnitude, we will confine ourselves to pointing out that analyses based on business intelligence systems are affected by this issue and therefore run the risk of being used to maximize profits even when different considerations should prevail related to the social consequences of the decisions made, according to a logic that we believe should be rejected.

For example, is it right to develop an optimization model with the purpose of distributing costs on an international scale in order to circumvent the tax systems of certain countries? Is it legitimate to make a decision on the optimal position of the tank in a vehicle in order to minimize production costs, even if this may cause serious harm to the passengers in the event of a collision? As proven by these examples, analysts developing a mathematical model and those who make the decisions cannot remain neutral, but have the moral obligation to take an ethical stance.

3.10 CHECK YOUR PROGRESS

1. Define business intelligence
2. What is the main purpose of business intelligence and knowledge management?
3. Differentiate data and information.

4. Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions. (True/False)

5. Mathematical models are then developed by exploiting the relationships among ________, _____________ and ______________.

Answers to check your progress

1. Business intelligence may be defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes

2. The main purpose of both disciplines is to develop environments that can support knowledge workers in decision-making processes and complex problem-solving activities.

3. Data: Generally, data represent a structured codification of single primary entities, as well as of transactions involving two or more primary entities. Information: Information is the outcome of extraction and processing activities carried out on data, and it appears meaningful for those who receive it in a specific domain.

4. True

5. system control variables, parameters and evaluation metrics.

3.11 SUMMARY

This unit describes the problems entailed in business intelligence, highlighting the interconnections with other disciplines and identifying the primary components typical of a business intelligence environment. The role of mathematical models are discussed. Various strategies to develop a Business Intelligence System is highlighted.

3.112 KEYWORDS

- **Business Intelligence**: Business intelligence (BI) refers to the procedural and technical infrastructure that collects, stores, and analyzes the data produced by a company's activities
- **Business analytics (BA):** is a set of disciplines and technologies for solving business problems using data analysis, statistical models and other quantitative methods.

- **Data exploration**: Data exploration is the first step of data analysis used to explore and visualize data to uncover insights from the start or identify areas or patterns to dig into more.
- **Data mining:** Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes.
- **Data analysis** : is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.

3.13 QUESTIONS FOR SELF STUDY

1. Write the significance of Business Intelligence.
2. Write a note on effective and timely decisions.
3. Distinguish data, information and knowledge.
4. Explain the role of mathematical models.
5. Discuss major building blocks of business intelligence system.
6. Explain the architecture of business intelligence system.
7. Explain briefly cycle of a business intelligence analysis.
8. Mention ethics in adoption of business intelligence methodologies.

3.14 REFERENCES

1. Sharda R, Delen D, Turban E. Business intelligence analytics and data science: A managerial perspective. Pearson, 2022.
2. Vercellis C. Business intelligence: data mining and optimization for decision making. John Wiley & Sons; 2011 Aug 10.
3. Bentley D. Business Intelligence and Analytics. Internet, link: https://www. pdfdrive. com/business-intelligence-and-analytics-e56416503. html. 2017.

UNIT- 4: DECISION SUPPORT SYSTEMS

Structure

4.0 OBJECTIVES

After studying this unit, you will be able to:

- Illustrate Representation of Decision-Making process
- Elucidate Rationality and Problem Solving
- Account on representation of Decision-Making Process
- Analysis of Decision Support System
- Development of a Decision Support System

4.1 INTRODUCTION

A decision support system (DSS) is an interactive computer-based application that combines data and mathematical models to help decision makers solve complex problems faced in managing the public and private enterprises and organizations. As described in unit 3, the analysis tools provided by a business intelligence architecture can be regarded as DSSs capable of transforming data into information and knowledge helpful to decision makers. In this respect, DSSs are a basic component in the development of a business intelligence architecture.

In this unit, we will first discuss the structure of the decision-making process. Further on, the evolution of information systems will be briefly sketched. We will then define DSSs, outlining the major advantages and pointing out the critical success factors relative to their introduction. Finally, the development phases of a DSS project will be described, addressing the most relevant issues concerning its implementation.

4.2 DEFINITION OF SYSTEM

The term system is often used in everyday language: for instance, we refer to the solar system, the nervous system or the justice system. The entities that we intuitively denominate systems share a common characteristic, which we will adopt as an abstract definition of the notion of system: each of them is made up of a set of components that are in some way connected to each other so as to provide a single collective result and a common purpose.

Every system is characterized by boundaries that separate its internal components from the external environment. A system is said to be open if its boundaries can be crossed in both directions by flows of materials and information.

When such flows are lacking, the system is said to be closed. In general terms, any given system receives specific input flows, carries out an internal transformation process and generates observable output flows.

As can be imagined, this abstract definition of system can be used to describe a broad class of real-world phenomena. For example, the logistic structure of an enterprise is a system that receives as input a set of materials, services and information and returns as output a set of products, services and information. More generally, even an enterprise, taken as a whole or in part, may be represented in its turn as a system, provided the boundaries as well as input and output flows are clearly defined.

Figure 4.1 shows the structure that we will use as a reference to describe the concept of the system. A system receives a set of input flows and returns a set of output flows through a transformation process regulated by internal conditions and external conditions. The effectiveness and efficiency of a system are assessed using measurable performance indicators that can be classified into different categories. The figure shows the main types of metrics used to evaluate systems embedded within the enterprises and the public administration.

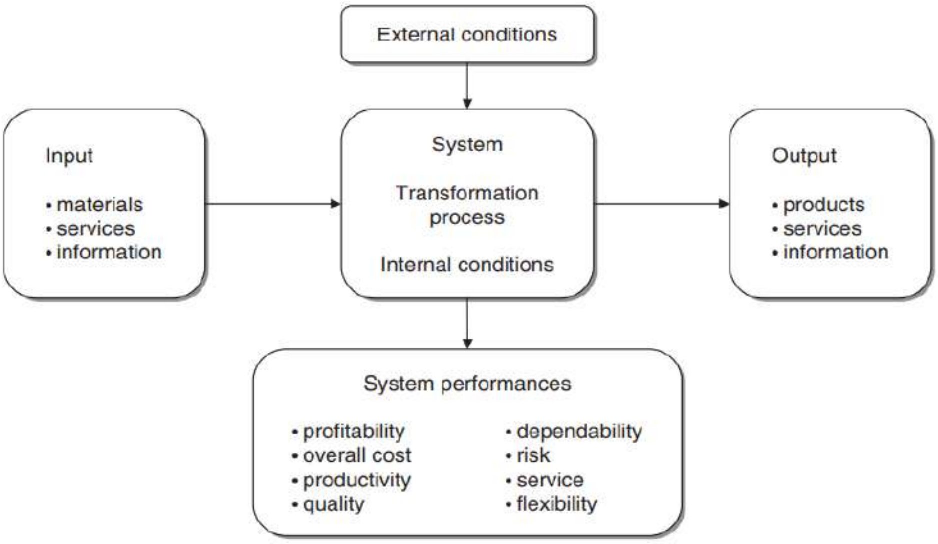


Figure 4.1 Abstract representation of a system

A system will often incorporate a feedback mechanism. Feedback occurs when a system component generates an output flow that is fed back into the system itself as an input flow, possibly as a result of a further transformation. Systems that are able to modify their own output flows based on feedback are called closed cycle systems. For example, the closed cycle system outlined in Figure 4.2 describes the development of a sequence of marketing campaigns. The sales results for each campaign are gathered and become available as feedback input so as to design subsequent marketing promotions.

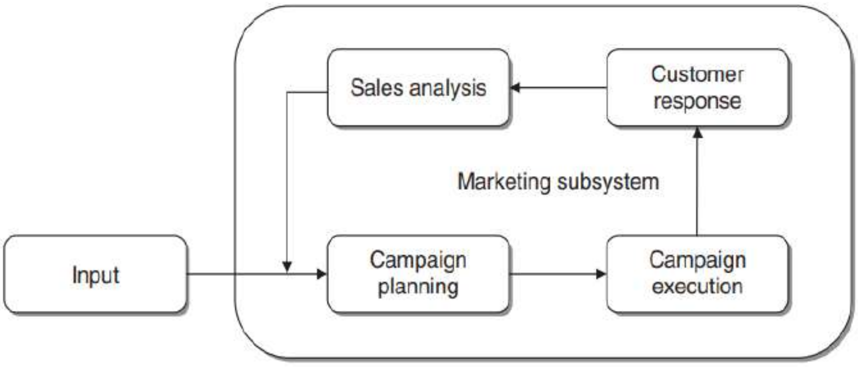


Figure 4.2 A closed cycle marketing system with feedback effects

In connection with a decision-making process, it is appropriate to categorize the evaluation metrics into two main classes: effectiveness and efficiency.

Effectiveness. Effectiveness measurements express the level of conformity of a given system to the objectives for which it was designed. The associated performance indicators are therefore linked to the system output flows, such as production volumes, weekly sales and yield per share.

Efficiency. Efficiency measurements highlight the relationship between input flows used by the system and the corresponding output flows. Efficiency measurements are therefore associated with the quality of the transformation process. For example, they might express the amount of resources needed to achieve a given sales volume.

Generally speaking, effectiveness metrics indicate whether the right action is being carried out or not, while efficiency metrics show whether the action is being carried out in the best possible way or not.

4.3 REPRESENTATION OF THE DECISION-MAKING PROCESS

In order to build effective DSSs, we first need to describe in general terms how a decision-making process is articulated. In particular, we wish to understand the steps that lead individuals to make decisions and the extent of the influence exerted on them by the subjective attitudes of the decision makers and the specific context within which decisions are taken.

4.3.1 RATIONALITY AND PROBLEM SOLVING

A decision is a choice from multiple alternatives, usually made with a fair degree of rationality. Each individual faces on a continual basis decisions that can be more or less important, both in their personal and professional life. In this section, we will focus on decisions made by knowledge workers in public and private enterprises and organizations. These decisions may concern the development of a strategic plan and imply therefore substantial investment choices, the definition of marketing initiatives and related sales predictions, and the design of a production plan that allows the available human and technological resources to be employed in an effective and efficient way.

Figure 4.3 outlines the structure of the problem-solving process. The alternatives represent the possible actions aimed at solving the given problem and helping to achieve the planned objective. In some instances, the number of alternatives being considered may be small. In the case of a credit agency that has to decide whether or not to grant a loan to an applicant, only two options exist, namely acceptance and rejection of the request. In other instances, the number of alternatives can be very large or even infinite. For example, the development of the annual logistic plan of a manufacturing company requires a choice to be made from an infinite number of alternative options.

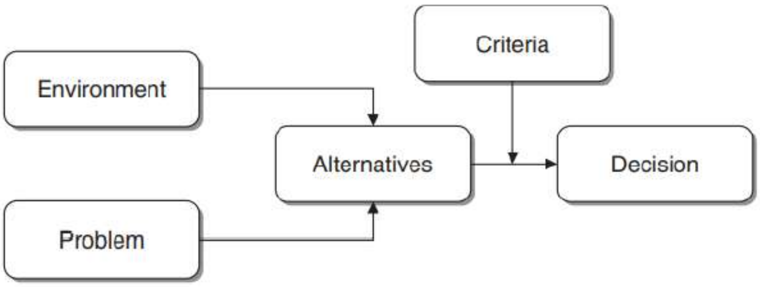


Figure 4.3 Logical flow of a problem-solving process

Criteria are the measurements of effectiveness of the various alternatives and correspond to the different kinds of system performance shown in Figure 4.1. A rational approach to decision making implies that the option fulfilling the best performance criteria is selected out of all possible alternatives. Besides economic criteria, which tend to prevail in the decision-making process within companies, it is however possible to identify other factors influencing a rational choice.

Economic. Economic factors are the most influential in decision-making processes, and are often aimed at the minimization of costs or the maximization of profits. For example, an annual logistic plan may be preferred over alternative plans if it achieves a reduction in total costs.

Technical. Options that are not technically feasible must be discarded. For instance, a production plan that exceeds the maximum capacity of a plant cannot be regarded as a feasible option.

Legal. Legal rationality implies that before adopting any choice the decision makers should verify whether it is compatible with the legislation in force within the application domain.

Ethical. Besides being compliant with the law, a decision should abide by the ethical principles and social rules of the community to which the system belongs.

Procedural. A decision may be considered ideal from an economic, legal and social standpoint, but it may be unworkable due to cultural limitations of the organization in terms of prevailing procedures and common practice.

Political. The decision maker must also assess the political consequences of a specific decision among individuals, departments and organizations.

The process of evaluating the alternatives may be divided into two main stages, shown in Figure 4.4: exclusion and evaluation. During the exclusion stage, compatibility rules and restrictions are applied to the alternative actions that were originally identified. Within this assessment process, some alternatives will be dropped from consideration, while the rest represent feasible options that will be promoted to evaluation. In the evaluation phase, feasible alternatives are compared to one another on the basis of the performance criteria, in order to identify the preferred decision as the best opportunity.

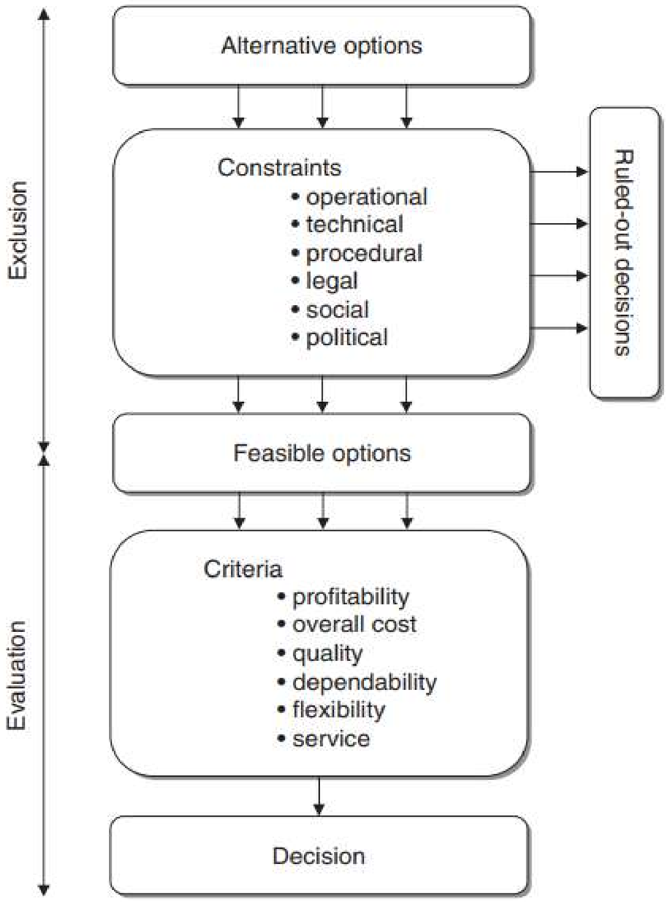


Figure 4.4 Logical structure of the decision-making process

4.3.2 THE DECISION-MAKING PROCESS

A compelling representation of the decision-making process was proposed in the early 1960s, and still remains today a major methodological reference. The model includes three phases, termed intelligence, design and choice. Figure 4.5 shows an extended version of the original scheme, which results from the inclusion of two additional phases, namely implementation and control.

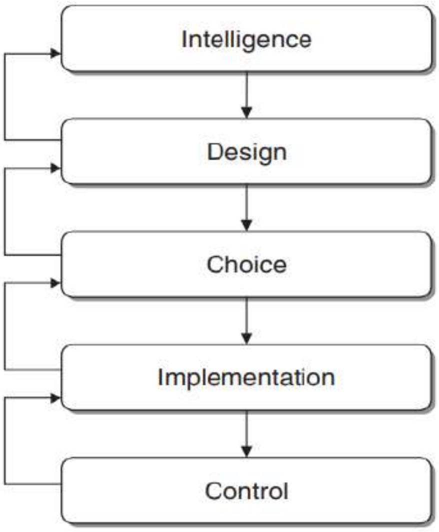


Figure 4.5 Phases of the decision-making process

Intelligence. In the intelligence phase, the task of the decision maker is to identify, circumscribe and explicitly define the problem that emerges in the system under study. The analysis of the context and all the available information may allow decision makers to quickly grasp the signals and symptoms pointing to a corrective action to improve the system performance. For example, during the execution of a project the intelligence phase may consist of a comparison between the current progress of the activities and the original development plan. In general, it is important not to confuse the problem with the symptoms.

Design. In the design phase actions aimed at solving the identified problem should be developed and planned. At this level, the experience and creativity of the decision makers play a critical role, as they are asked to devise viable solutions that ultimately allow the intended purpose to be achieved. Where the number of available actions is small, decision makers can make an explicit enumeration of the alternatives to identify the best solution. If, on the other hand, the number of alternatives is very large, or even unlimited, their identification occurs in an implicit way, usually through a description of the rules that feasible actions should satisfy. For example, these rules may directly translate into the constraints of an optimization model.

Choice. Once the alternative actions have been identified, it is necessary to evaluate them on the basis of the performance criteria deemed significant. Mathematical models and the corresponding solution methods usually play a valuable role during the choice phase. For example, optimization models and methods allow the best solution to be found in very complex situations involving count- less or even infinite feasible solutions.

The most relevant aspects characterizing a decision-making process can be briefly summarized as follows.

- Decisions are often devised by a group of individuals instead of a single decision maker.
- The number of alternative actions may be very high, and sometimes unlimited.
- The effects of a given decision usually appear later, not immediately.
- The decisions made within a public or private enterprise or organization are often interconnected and determine broad effects. Each decision has consequences for many individuals and several parts of the organization.
- During the decision-making process knowledge workers are asked to access data and information, and work on them based on a conceptual and analytical framework.
- Feedback plays an important role in providing information and knowledge for future decision-making processes within a given organization.
- In most instances, the decision-making process has multiple goals, with different performance indicators, that might also be in conflict with one another.
- Many decisions are made in a fuzzy context and entail risk factors. The level of propensity or aversion to risk varies significantly among different individuals.
- Experiments carried out in a real-world system, according to a trial-and- error scheme, are too costly and risky to be of practical use for decision making.
- The dynamics in which an enterprise operates, strongly affected by the pressure of a competitive environment, imply that knowledge workers need to address situations and make decisions quickly and in a timely fashion.

4.3.3 TYPES OF DECISIONS

Defining a taxonomy of decisions may prove useful during the design of a DSS, since it is likely that decision-making processes with similar characteristics may be supported by the same set of methodologies. Decisions can be classified in terms of two main dimensions, according to their nature and scope. Each dimension will be subdivided into three classes, giving a total of nine possible combinations, as shown in Figure 4.6.

According to their nature, decisions can be classified as structured, unstructured or semi-structured:

Structured decisions: A decision is structured if it is based on a well-defined and recurring decision-making procedure. In most cases structured decisions can be traced back to an algorithm, which may be more or less explicit for decision makers, and are therefore better suited for automation. More specifically, we have a structured decision if input flows, output flows and the transformations performed by the system can be clearly described in the three phases of intelligence, design and choice. In this case, we will also say that each component phase is structured in its turn. Actually, even decisions that appear fully structured require in most cases the direct intervention of decision makers to cope with unexpected events, caused for example by unusual values of some input flows.

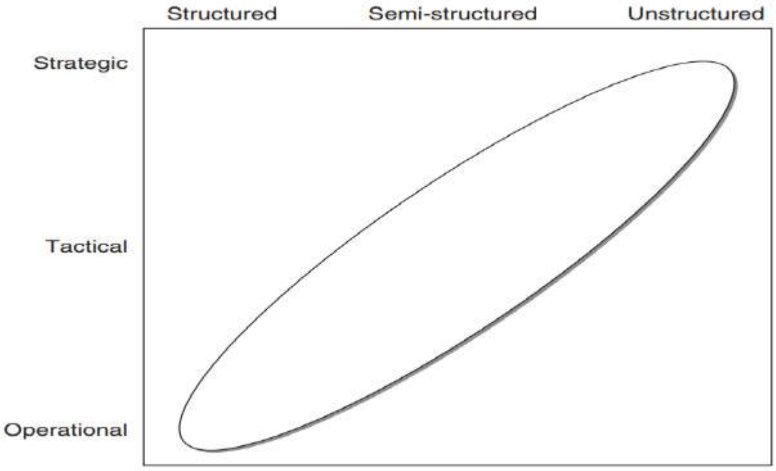


Figure 4.6 A taxonomy of decisions

Unstructured decisions: A decision is said to be unstructured if the three phases of intelligence, design and choice are also unstructured. This means that for each phase there is at least one element in the system (input flows, output flows and the transformation processes) that cannot be described in detail and reduced to a predefined sequence of steps. Such an event may occur when a decision-making process is faced for the first time or if it happens very seldom. In this type of decisions the role of knowledge workers is fundamental, and business intelligence systems may provide support to decision makers through timely and versatile access to information.

Semi-structured decisions: A decision is semi-structured when some phases are structured and others are not. Most decisions faced by knowledge workers in managing public or private enterprises or organizations are semi-structured. Hence, they can take advantage of DSSs and a business intelligence environment primarily in two ways. For the unstructured phases of the decision-making process, business intelligence tools may offer a passive type of support which translates into timely and versatile access to information. For the structured phases it is possible to provide an active form of support through mathematical models and algorithms that allow significant parts of the decision-making process to be automated.

Sometimes situations may arise where the nature of a decision cannot be easily identified unambiguously. When facing the same problem, such as establishing the sale price of a product, different decision makers operating in different organizations may come up with dissimilar choices. For example, a first decision maker may believe that the best sale price can be obtained by comparing cost and price – demand elasticity curves. As a consequence, such decision maker may consider the choice phase of the decision-making process as structured. By contrast, a second decision maker may believe that the elasticity curve does not reflect all the factors influencing the response of the market to price variations since some of these elements cannot be quantified. For this individual the choice phase turns out to be unstructured or at most semi-structured. Examples 4.1, 4.4 and 4.3 describe structured, semi-structured and unstructured decisions, respectively.

Example 4.1 – A structured decision

A paper mill produces for the company warehouse paper sheets in different standard sizes that are subsequently cut to size for customers. Specifically, customers submit orders in terms of type of paper, quantity and size. The sizes specified in the orders are usually smaller than standard sizes and must be cut out of these. The paper mill is therefore forced to consider how the sizes required to fulfill orders should best be combined and cut from standard sizes so as to minimize paper waste. This decision is common to many industries (paper, aluminum, wood, steel, glass, fabric) and can be very well supported by optimization models. However, even in connection with such structured decisions, particular circumstances and specific input values may require intervention by the decision maker to modify the plans obtained by means of optimization models. For example, the company may wish to favor a specific request of a customer considered strategic, introducing a fast-processing lane in the cutting plan, even if this may involve more wasted material during the cutting stage.

Example 4.2 – A semi-structured decision

The logistics manager of a manufacturing company needs to develop an annual plan. The logistic plan determines the allocation to each plant of the production volumes forecasted for the different market areas, the purchase of materials from each supplier with the related volumes and delivery times, the production lots for each manufacturing stage, the stock levels of sub-assemblies and end items, and the distribution of end items to the market areas. These decisions have a great economic and organizational impact that might greatly benefit from the adoption of a DSS based on large-scale optimization models. However, it is likely that in a real situation some elements are left to discretion of the decision makers, who may prefer a given logistic plan over another, even if it implies moderately higher costs compared to the optimal plan proposed by the model. For example, it might be appropriate to maintain unaltered the supply of parts purchased from a given supplier who is considered strategic for the future even though this supplier is less competitive than others, that are instead preferred by the optimization model in terms of minimum cost.

Example 4.3 – An unstructured decision

Consider an enterprise that is the target of a hostile takeover by a public offer made by a direct competitor. There are various possible defensive decisions and actions that are strongly dependent on the context in which the enterprise operates and the offer is made. It is difficult to envisage a systematic description of the decision process that might be later reproduced in other similar cases.

From the above examples it emerges that the nature of a decision process depends on many factors, including:

- the characteristics of the organization within which the system is placed;
- the subjective attitudes of the decision makers;
- the availability of appropriate problem-solving methodologies;
- the availability of effective decision support tools.

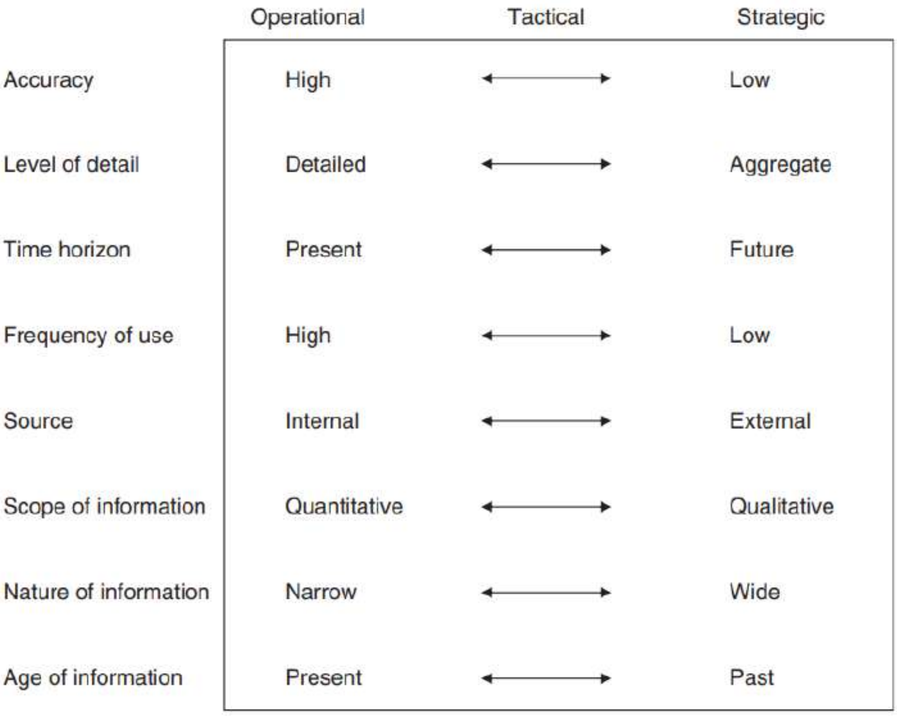
Depending on their scope, decisions can be classified as strategic, tactical and operational.

Strategic decisions. Decisions are strategic when they affect the entire organization or at least a substantial part of it for a long period of time. Strategic decisions strongly influence the general objectives and policies of an enterprise. As a consequence, strategic decisions are taken at a higher organizational level, usually by the company top management.

Tactical decisions. Tactical decisions affect only parts of an enterprise and are usually restricted to a single department. The time span is limited to a medium-term horizon, typically up to a year. Tactical decisions place them- selves within the context determined by strategic decisions. In a company hierarchy, tactical decisions are made by middle managers, such as the heads of the company departments.

Operational decisions. Operational decisions refer to specific activities carried out within an organization and have a modest impact on the future. Operational decisions are framed within the elements and conditions determined by strategic and tactical decisions. Therefore, they are usually made at a lower organizational level, by knowledge workers responsible for a single activity or task such as sub-department heads, workshop foremen, back-office heads. The characteristics of the information required in a decision-making process will change depending on the scope of the decisions to be supported, and consequently also the orientation of a DSS will vary accordingly. Figure 4.7 shows variations in the characteristics of the information as the scope of the decisions changes. The scheme can be used as an assessment tool when designing a DSS: once the scope of the decisions for which the system is intended has been set, the scheme can be used to establish whether the decision-making process is adequately supported by the right information.

Although nature and scope are not perfectly correlated, most real-world decisions fall within the ellipse shown in Figure 4.6: most strategic decisions are unstructured, while most operational decisions are structured and most tactical decisions are semi-structured. This empirical remark is useful when defining in advance the characteristics of a DSS to facilitate a decision-making process of specific nature and scope.



	Operational	Tactical	Strategic
Accuracy	High	↔	Low
Level of detail	Detailed	↔	Aggregate
Time horizon	Present	↔	Future
Frequency of use	High	↔	Low
Source	Internal	↔	External
Scope of information	Quantitative	↔	Qualitative
Nature of information	Narrow	↔	Wide
Age of information	Present	↔	Past

Figure 4.7 Characteristics of the information in terms of the scope of decisions

4.4 DEFINITION OF DECISION SUPPORT SYSTEM

Since the late 1980s, a decision support system has been defined as an interactive computer system helping decision makers to combine data and models to solve semi-structured and unstructured problems. This definition entails the three main elements of a DSS shown in Figure 4.8: a database, a repository of mathematical models and a module for handling the dialogue between the system and the users. It thus highlights the role of DSSs as the focal point of evolution trends in two distinct areas: on the one hand, data processing and information technologies; and on the other hand, the disciplines addressing the study of mathematical models and methods, such as operations research and statistics.

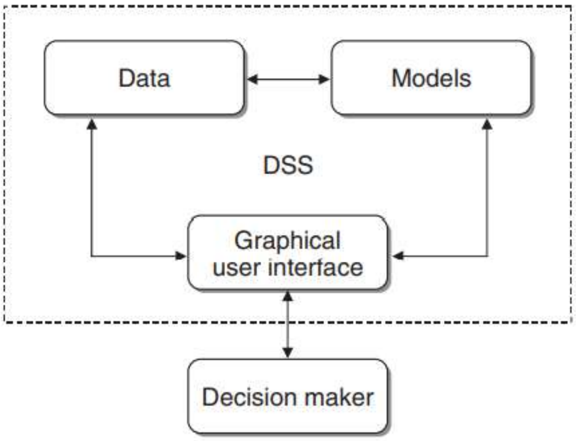


Figure 4.8 Structure of a decision support system

Indeed, despite significant improvements achieved in both areas, the actual implementation of applications that could be used by knowledge workers has been troublesome. Information technology in the early 1970s, as discussed in the previous section, mainly consisted of applications for accounting and administration, their capabilities restricted to the processing of large amounts of transactions and the production of summary reports. The partial failure of management information systems indicated that even the essential objective of attaining timely and versatile access to information was not within easy reach.

On the other hand, although mathematical models were made more flexible by resort to innovative user interaction techniques, they were still mostly based on regulatory methods, suitable for operational decisions characterized by unique objectives rather than tactical and strategic decisions, usually less structured. As we have observed, semi-structured and unstructured decision processes are heavily influenced by subjective opinions, so that knowledge workers have to directly analyze multiple performance indicators which it would be difficult to reduce to a single decision-making criterion.

It is worth highlighting the relevant features of a DSS in order to circumscribe the definition given above and to better understand its role.

Effectiveness. Decision support systems should help knowledge workers to reach more effective decisions. In this respect they are a fundamental component of business intelligence architectures. Note that this does not necessarily imply an increased efficiency in the decision-making process. In fact, the adoption of a DSS may entail a more accurate analysis and therefore require a greater time investment by decision makers. However, the greater effort required will usually result in better decisions.

Mathematical models. In order to achieve more effective decisions, a DSS makes use of mathematical models, borrowed from disciplines such as operations research and statistics, which are applied to the data contained in data warehouses and data marts. The use of analytical models to transform data into knowledge and provide active support is the main characteristic that sets apart a DSS from a simple information system.

Integration in the decision-making process. A DSS should provide help for different kinds of knowledge workers, within the same application domain, particularly in respect of semi-structured and unstructured decision processes, both of an individual and a collective nature. Further, a DSS is intended for decision-making processes that are strategic, tactical and operational in scope. Moreover, decision makers should have the opportunity to integrate in a DSS their preferences and competencies, adapting it to their needs rather than passively accepting what comes out of it. In this way, a DSS may progressively take the role of a key component in the problem-solving methodology adopted by decision makers, enabling a proactive and perceptive decision making style, instead of a reactive and by-exception attitude, in order to anticipate any rapidly evolving dynamic phenomenon.

Organizational role. In many situations the users of a DSS operate at different hierarchical levels within an enterprise, and a DSS tends to encourage communication between the various parts of an organization. By providing support for sequential and interdependent decision processes, a DSS can keep track of the analysis and the information that led to a specific decision.

Flexibility. A DSS must be flexible and adaptable in order to incorporate the changes required to reflect modifications in the environment or in the decision-making process. Moreover, it should be easy to use, with user-friendly and intuitive interaction methods and high-quality graphics for presenting the information extracted or generated. It is becoming increasingly common for DSSs to feature a web-browser interface to communicate with users.

The structure of the DSS shown in Figure 4.8 is extended in Figure 4.9 to include some new components.

Data management. The data management module includes a database designed to contain the data required by the decision-making processes to which the DSS is addressed.

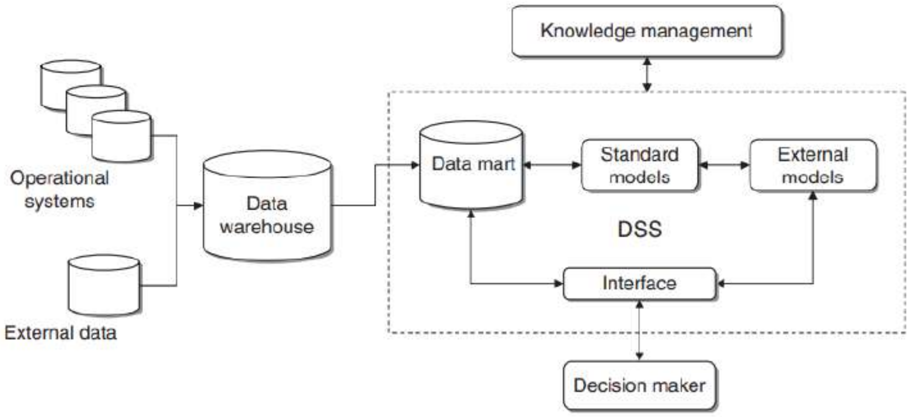


Figure 4.9 Extended structure of a decision support system

Model management. The model management module provides end users with a collection of mathematical models derived from operations research, statistics and financial analysis. These are usually relatively simple models that allow analytical investigations to be carried out that are very helpful during the decision-making process. To illustrate the role played by the model management module one can think of the analytical functions offered by current spreadsheets. These include simple optimization models, financial and actuarial analysis models and statistical functionalities. Moreover, the model management module helps the activities of knowledge workers by means of high-level languages for the development of ad hoc models. In certain applications such a module may be integrated with more complex models, referred to as external models in Figure 4.9, created to carry out specific analysis tasks. For example, a large-scale optimization model formulated to develop the annual logistic plan of a manufacturing company falls in this category.

Interactions. In most applications, knowledge workers use a DSS interactively to carry out their analyses. The module responsible for these interactions is expected to receive input data from users in the easiest and most intuitive way, usually through the graphic interface of a web browser, and then to return the extracted information and the knowledge generated by the system in an appropriate graphical form.

Knowledge management. The knowledge management module is also interconnected with the company knowledge management integrated system. It allows decision makers to draw on the various forms of collective knowledge, usually unstructured, that represents the corporate culture.

This section concludes with a summary of the major potential advantages deriving from the adoption of a DSS:

- an increase in the number of alternatives or options considered;
- an increase in the number of effective decisions devised;
- a greater awareness and a deeper understanding of the domain analyzed and the problems investigated;
- the possibility of executing scenario and what-if analyses by varying the hypotheses and parameters of the mathematical models;
- an improved ability to react promptly to unexpected events and unforeseen situations;
- a value-added exploitation of the available data;
- an improved communication and coordination among the individuals and the organizational departments;
- more effective development of teamwork;
- a greater reliability of the control mechanisms, due to the increased intelligibility of the decision process.

4.5 DEVELOPMENT OF A DECISION SUPPORT SYSTEM

In this section we will describe the development phases of a DSS. Unlike other software applications, such as information systems and office automation tools, DSSs are usually not available as standard programs. Multidimensional analysis environments have facilitated and standardized the access to passive business intelligence functions. However, in order to develop most DSSs a specific project is still required.

Figure 4.10 shows the major steps in the development of a DSS. The logical flow of the activities is shown by the solid arrows. The dotted arrows in the opposite direction indicate revisions of one or more phases that might become necessary during the development of the system, through a feedback mechanism. We describe now in detail how each phase is carried out.

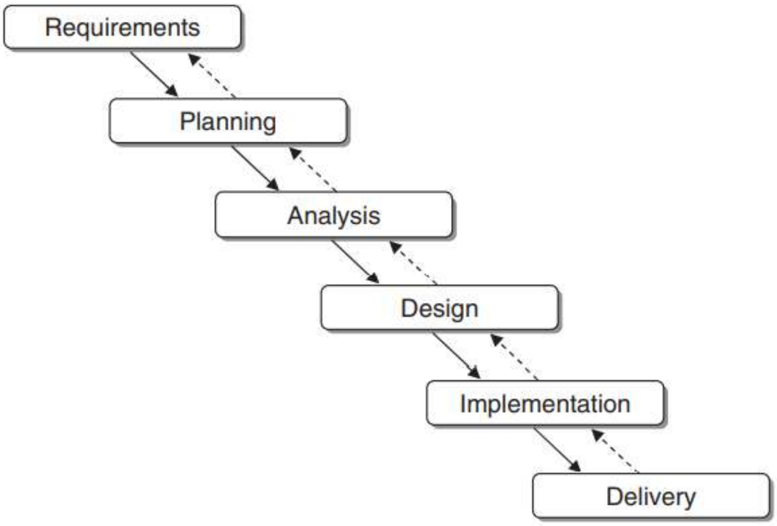


Figure 4.10 Phases in the development of a decision support system

Planning. The main purpose of the planning phase is to understand the needs and opportunities, sometimes characterized by weak signals, and to translate them into a project and later into a successful DSS. Planning usually involves a feasibility study to address the question: Why do we wish to develop a DSS? During the feasibility analysis, general and specific objectives of the system, recipients, possible benefits, execution times and costs are laid down. It is not easy to identify the benefits of a DSS.

Analysis. In the analysis phase, it is necessary to define in detail the functions of the DSS to be developed, by further developing and elaborating the preliminary conclusions achieved during the feasibility study. A response should therefore be given to the following question: What should the DSS accomplish, and who will use it, when and how? To provide an answer, it is necessary to analyze the decision processes to be supported, to try to thoroughly understand all interrelations existing between the problems addressed and the surrounding environment. The organizational implications determined by a DSS should be assessed. The analysis also involves mapping out the actual decision processes and imagining what the new processes will look like once the DSS is in place. Finally, it is necessary to explore the data in order to understand how much and what type of information already exists and what information can be retrieved from external sources

Design. During the design phase the main question is: How will the DSS work? The entire architecture of the system is therefore defined, through the identification of the hardware technology platforms, the network structure, the software tools to develop the applications and the specific database to be used. It is also necessary to define in detail the interactions with the users, by means of input masks, graphic visualizations on the screen and printed reports. In recent years the web browser has become an important interaction tool, and has certainly contributed to the harmonization of, and to the simplification of the problems related to, communication between knowledge workers and computers. A further aspect that should be clarified during the design phase is the make-or-buy choice – whether to subcontract the implementation of the DSS to third parties, in whole or in part.

Implementation. Once the specifications have been laid down, it is time for implementation, testing and the actual installation, when the DSS is rolled out and put to work. Any problems faced in this last phase can be traced back to project management methods. A further aspect of the implementation phase, which is often overlooked, relates to the overall impact on the organization determined by the new system. Such effects should be monitored using change management techniques, making sure that no one feels excluded from the organizational innovation process and rejects the DSS.

Sometimes a project may not come to a successful conclusion, may not succeed in fulfilling expectations, or may even turn out to be a complete failure. However, there are ways to reduce the risk of failure. The most significant of these is based on the use of rapid prototyping development where, instead of implementing the system as a whole, the approach is to identify a sequence of autonomous subsystems, of limited capabilities, and develop these subsystems step by step until the final stage is reached corresponding to the fully developed DSS.

4.6 CHECK YOUR PROGRESS

6. Define Decision Support System.
7. A decision is a choice from multiple alternatives, usually made with a fair degree of rationality.(True/False)
8. List phases of the decision-making process.
9. Decisions can be classified as _________, ___________ and _____________.
10. Differentiate effectiveness and efficiency.

Answers to Check your progress

1. A decision support system (DSS) is an interactive computer-based application that combines data and mathematical models to help decision makers solve complex problems faced in managing the public and private enterprises and organizations.

2. True
3. Phases of the decision-making process are intelligence, design, choice, implementation and control.
4. structured, unstructured or semi-structured:
5. Effectiveness measurements express the level of conformity of a given system to the objectives for which it was designed. Efficiency measurements highlight the relationship between input flows used by the system and the corresponding output flows.

4.7 SUMMARY

This unit describes structure of the problem-solving process. The different phases of the decision-making process are discussed. Various decision types are classified with suitable illustrations. The characteristics of a DSS to facilitate a decision-making process of specific nature and scope are listed. Analysis is made on the structure of decision support system. Also several phases in the development of a decision support system are detailed.

4.8 KEYWORDS

- Decision support system (DSS) : is a computerized program used to support determinations, judgments, and courses of action in an organization or a business.
- **Decision making**: is the process of making choices by identifying a decision, gathering information, and assessing alternative resolutions.
- **Problem Solving process** : consists of a sequence of sections that fit together depending on the type of problem to be solved.
- **Strategic decision-making**: is a process of understanding the interaction of decisions and their impact upon the organization to gain an advantage.
- **Tactical decisions**: are decisions and plans that concern the more detailed implementation of the directors' general strategy, usually with a medium-term impact on a company.
- **Operational decisions**: or Operating decisions are decisions made to manage day to day business.

4.9 QUESTIONS FOR SELF STUDY

1. Describe structure of the problem-solving process.
2. Explain different phases of the decision-making process.

3. How are decisions classified? Explain.
4. Mention characteristics of a DSS to facilitate a decision-making process of specific nature and scope.
5. Briefly explain structure of decision support system.
6. Discuss several phases in the development of a decision support system.

4.10 REFERENCES

1. Sharda R, Delen D, Turban E. Business intelligence analytics and data science: A managerial perspective. Pearson, 2022.
2. Vercellis C. Business intelligence: data mining and optimization for decision making. John Wiley & Sons; 2011 Aug 10.
3. Bentley D. Business Intelligence and Analytics. Internet, link: https://www. pdfdrive. com/business-intelligence-and-analytics-e56416503. html. 2017.

Karnataka State Open University

Mukthagangothri, Mysore – 570 006.

Dept. of Studies and Research in Management

MBA IT Specialization

III Semester

Business Intelligence and Analytics

Block 2

Karnataka State Open University

Mukthagangothri, Mysore – 570 006.

Dept. of Studies and Research in Management

MBA. IT Specialization

III Semester

BUSINESS INTELLIGENCE AND ANALYTICS

BLOCK 2

BLOCK 2 INTRODUCTION

BI can be used to support a wide range of business decisions ranging from operational to strategic. Basic operating decisions include product positioning or pricing. Strategic business decisions include priorities, goals and directions at the broadest level. In all cases, BI is most effective when it combines data derived from the market in which a company operates (external data) with data from company sources internal to the business such as financial and operations data (internal data). When combined, external and internal data can provide a more complete picture which, in effect, creates an “intelligence” that cannot be derived by any singular set of data. Amongst myriad uses, BI tools empower organisations to gain insight into new markets, assess demand and suitability of products and services for different market segments and gauge the impact of marketing efforts.

This block consists of 4 units and is organized as follows:

Unit 5- Definition of Data Mining Representation of Input Data, Data Mining process, Analysis of Methodologies

Unit 6- Definition of Data Warehouse, Data Warehouse Architecture

Unit 7- Waterfall Development Process, Agile Development Techniques, Basic Concepts of Scrum, Agile Culture at Netflix, Medtronic: Agile for the Right Projects, Sharper BI at 1-800 CONTACTS, Best Practices for Successful Business Intelligence

Unit 8- Text Analytics and Text Mining Overview, Sentiment Analysis, Web Usage Mining (Web Analytics), Social Analytics

UNIT- 5 : DATA MINING

5.0 OBJECTIVES

After studying this unit, you will be able to:

- Give an account on data mining.
- Describe data mining activities.
- Discuss applications of data mining.
- Analyze representation of data in data mining process.
- Explain the process of data mining.
- Elucidate different data mining tasks.

5.1 INTRODUCTION

The evolving technologies of information gathering and storage have made available huge amounts of data within most application domains, such as the business world, the scientific and medical com- munity, and public administration. The set of activities involved in the analysis of these large databases, usually with the purpose of extracting useful knowledge to support decision making, has been referred to in different ways, such as data mining, knowledge discovery, pattern recognition and machine learning. In particular, the term data mining indicates the process of exploration and analysis of a dataset, usually of large size, in order to find regular patterns, to extract relevant knowledge and to obtain meaningful recurring rules. Data mining plays an ever-growing role in both theoretical studies and applications. In this unit, we wish to describe and characterize data mining activities with respect to investigation purposes and analysis methodologies. The relevant properties of input data will also be discussed. Finally, we will describe the data mining process and its articulation in distinct phases.

5.2 DEFINITION OF DATA MINING

Data mining activities constitute an iterative process aimed at the analysis of large databases, with the purpose of extracting information and knowledge that may prove accurate and potentially useful for knowledge workers engaged in decision making and problem solving.

As described in Section 5.3, the analysis process is iterative in nature since there are distinct phases that might imply feedback and subsequent revisions. Usually such a process represents a cooperative activity between experts in the application domain and data analysts, who use mathematical models for inductive learning. Indeed, experience indicates that a data mining study requires frequent interventions by the analysts across the different investigation phases and therefore cannot easily be automated. It is also necessary that the knowledge extracted be accurate, in the sense that it must be confirmed by data and not lead to misleading conclusions.

The term data mining refers therefore to the overall process consisting of data gathering and analysis, development of inductive learning models and adoption of practical decisions and consequent actions based on the knowledge acquired. The term mathematical learning theory is reserved for the variety of mathematical models and methods that can be found at the core of each data mining analysis and that are used to generate new knowledge.

The data mining process is based on inductive learning methods, whose main purpose is to derive general rules starting from a set of available examples, consisting of past observations recorded in one or more databases. In other words, the purpose of a data mining analysis is to draw some conclusions starting from a sample of past observations and to generalize these conclusions with reference to the entire population, in such a way that they are as accurate as possible.

A further characteristic of data mining depends on the procedure for collecting past observations and inserting them into a database. Indeed, these records are usually stored for purposes that are not primarily driven by data mining analysis. For instance, information on purchases from a retail company, or on the usage of each telephone number stored by a mobile phone provider, will basically be recorded for administrative purposes, even if the data may be later used to perform some useful data mining analysis. The data gathering procedureis therefore largely independent and unaware of the data mining objectives, so that it substantially differs from data gathering activities carried out according to predetermined sampling schemes, typical of classical statistics. In this respect, data mining represent a secondary form of data analysis.

Data mining activities can be subdivided into two major investigation streams, according to the main purpose of the analysis: interpretation and prediction.

Interpretation:

The purpose of interpretation is to identify regular patterns in the data and to express them through rules and criteria that can be easily understood by experts in the application domain. The rules generated must be original and non-trivial in order to actually increase the level of knowledgeand understanding of the system of interest. For example, for a company in the retail industry it might be advantageous to cluster those customers who have taken out loyalty cards according to their purchasing profile. The segments generated in this way might prove useful in identifying new market niches and directing future marketing campaigns.

Prediction: The purpose of prediction is to anticipate the value that a random variable will assume in the future or to estimate the likelihood of future events. For example, a mobile phone provider may develop a data mining analysis to estimate for its customers the probability of churning in favor of some competitor. In a different context, a retail company might predict the sales of a given product during the subsequent weeks. Actually, most data mining techniques derive their predictions from the value of a set of variables associated with the entities in a database. For example, a data mining model may indicate that the likelihood of future churning for a customer depends on features such as age, duration of the contract and percentage of calls to subscribers of other phone providers.

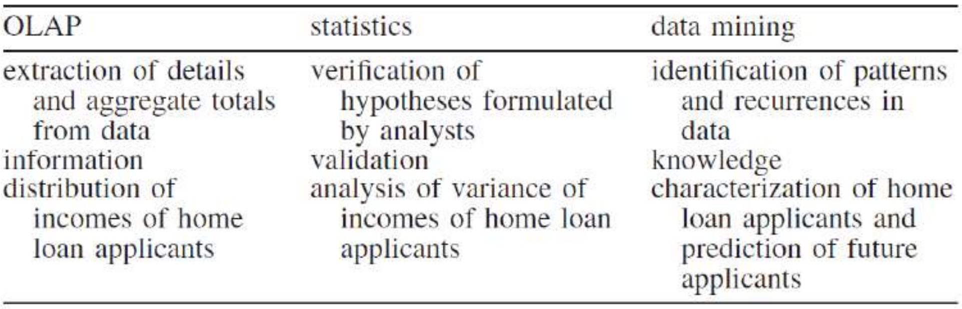
5.1.1 Models and methods for data mining

There are several learning methods that are available to perform the different data mining tasks. A number of techniques originated in the field of computer science, such as classification trees or association rules, and are referred to as machine learning or knowledge discovery in databases. In most cases an empirically based approach tends to prevail within this class of techniques. Other methods belong to multivariate statistics, such as regression or Bayesian classifiers, and are often parametric in nature but appear more theoretically grounded. More recent developments include mathematical methods for learning, such as statistical learning theory, which are based on solid theoretical foundations and place themselves at the crossroads of various disciplines, among which probability theory, optimization theory and statistics.

5.1.2 Data mining, classical statistics and OLAP

Data mining projects differ in many respects from both classical statistics and OLAP analyses. Such differences are shown in Table 5.1, with reference to an example.

Table 5.1 Differences between OLAP, statistics and data mining



OLAP	statistics	data mining
extraction of details and aggregate totals from data	verification of hypotheses formulated by analysts	identification of patterns and recurrences in data
information	validation	knowledge
distribution of incomes of home loan applicants	analysis of variance of incomes of home loan applicants	characterization of home loan applicants and prediction of future applicants

The main difference consists of the active orientation offered by inductive learning models, compared with the passive nature of statistical techniques and OLAP. Indeed, in statistical analyses decision makers formulate a hypothesis that then has to be confirmed on the basis of sample evidence. Similarly, in OLAP analyses knowledge workers express some intuition on which they base extraction, reporting and visualization criteria. Both methods – on one hand statistical validation techniques and on the other hand information tools to navigate through data cubes – only provide elements to confirm or disprove the hypotheses formulated by the decision maker, according to a top-down analysis flow. Conversely, learning models, which represent the core of data mining projects, are capable of playing an active role by generating predictions and interpretations which actually represent new knowledge available to the users. The analysis flow in the latter case has a bottom-up structure. In particular, when faced with large amounts of data, the use of models capable of playing an active role becomes a critical success factor, since it is hard for knowledge workers to formulate a priori meaningful and well-founded hypotheses.

5.1.3 Applications of data mining

Data mining methodologies can be applied to a variety of domains, from marketing and manufacturing process control to the study of risk factors in medical diagnosis, from the evaluation of the effectiveness of new drugs to fraud detects.

Relational marketing. Data mining applications in the field of relational marketing have significantly contributed to the increase in the popularity of these methodologies. Some relevant applications within relational marketing are:

- identification of customer segments that are most likely to respond to targeted marketing campaigns, such as cross-selling and up-selling;
- identification of target customer segments for retention campaigns;
- prediction of the rate of positive responses to marketing campaigns;
- interpretation and understanding of the buying behavior of the customers;
- analysis of the products jointly purchased by customers, known as market basket analysis

Fraud detection. Fraud detection is another relevant field of application of data mining. Fraud may affect different industries such as telephony, insurance (false claims) and banking (illegal use of credit cards and bank checks; illegal monetary transactions).

Risk evaluation. The purpose of risk analysis is to estimate the risk connected with future decisions, which often assume a dichotomous form. For example, using the past observations available, a bank may develop a predictive model to establish if it is appropriate to grant a monetary loan or a home loan, based on the characteristics of the applicant.

Text mining. Data mining can be applied to different kinds of texts, which rep- resent unstructured data, in order to classify articles, books, documents, emails and web pages. Examples are web search engines or the automatic classification of press releases for storing purposes. Other text mining applications include the generation of filters for email messages and newsgroups.

Image recognition. The treatment and classification of digital images, both static and dynamic, is an exciting subject both for its theoretical interest and the great number of applications it offers. It is useful to recognize written characters, compare and identify human faces, apply correction filters to photographic equipment and detect suspicious behaviors through surveillance video cameras.

Web mining. Web mining applications are intended for the analysis of so-called clickstreams – the sequences of pages visited and the choices made by a web surfer. They may prove useful for the analysis of e-commerce sites, in offering flexible and customized pages to surfers, in caching the most popular pages or in evaluating the effectiveness of an e-learning training course.

Medical diagnosis. Learning models are an invaluable tool within the medical field for the early detection of diseases using clinical test results. Image analysis for diagnostic purpose is another field of investigation that is currently burgeoning.

5.3 REPRESENTATION OF INPUT DATA

In most cases, the input to a data mining analysis takes the form of a two- dimensional table, called a dataset, irrespective of the actual logic and material representation adopted to store the information in files, databases, data warehouses and data marts used as data sources. The rows in the dataset correspond to the observations recorded in the past and are also called examples, cases, instances or records. The columns represent the information available for each observation and are termed attributes, variables, characteristics or features.

The attributes contained in a dataset can be categorized as categorical or numerical, depending on the type of values they take on.

Categorical:

Categorical attributes assume a finite number of distinct values, in most cases limited to less than a hundred, representing a qualitative property of an entity to which they refer. Examples of categorical attributes are the province of residence of an individual (which takes as values a series of names, which in turn may be represented by integers) or whether a customer has abandoned her service provider (expressed by the value 1) or remained loyal to it (expressed by the value 0). Arithmetic operations cannot be applied to categorical attributes even when the coding of their values is expressed by integer numbers.

Numerical:

Numerical attributes assume a finite or infinite number of values and lend themselves to subtraction or division operations. For example, the amount of outgoing phone calls during a month for a generic customer represents a numerical variable. Regarding two customers A and B making phone calls in a week for ¤27 and ¤36 respectively, it makes sense to claim that the difference between the amounts spent by the two customers is equal to ¤9 and that A has spent three fourths of the amount spent by B.

Sometimes a more refined taxonomy of attributes can prove useful.

Counts. Counts are categorical attributes in relation to which a specific property can be true or false. These attributes can therefore be represented using Boolean variables true, false or binary variables 0,1 . For example, a bank’s customers may or may not be holders of a credit card issued by the bank.

Nominal. Nominal attributes are categorical attributes without a natural ordering, such as the province of residence.

Ordinal. Ordinal attributes, such as education level, are categorical attributes that lend themselves to a natural ordering but for which it makes no sense to calculate differences or ratios between the values.

Discrete. Discrete attributes are numerical attributes that assume a finite number or a countable infinity of values.[1]

Continuous. Continuous attributes are numerical attributes that assume an uncountable infinity of values.

5.4 DATA MINING PROCESS

The definition of data mining given at the beginning of Section 5.1 refers to an iterative process, during which learning models and techniques play a key, though non-exhaustive role. Figure 5.1 shows the main phases of a generic data mining process.

[1] If a set A has the same cardinality as the set N of natural numbers, then we say that A is countable. In other words, a set is countable if there is a bisection between that set and N. There exist sets, such as the set R of real numbers, that are infinite and not countable, and are therefore called uncountable.

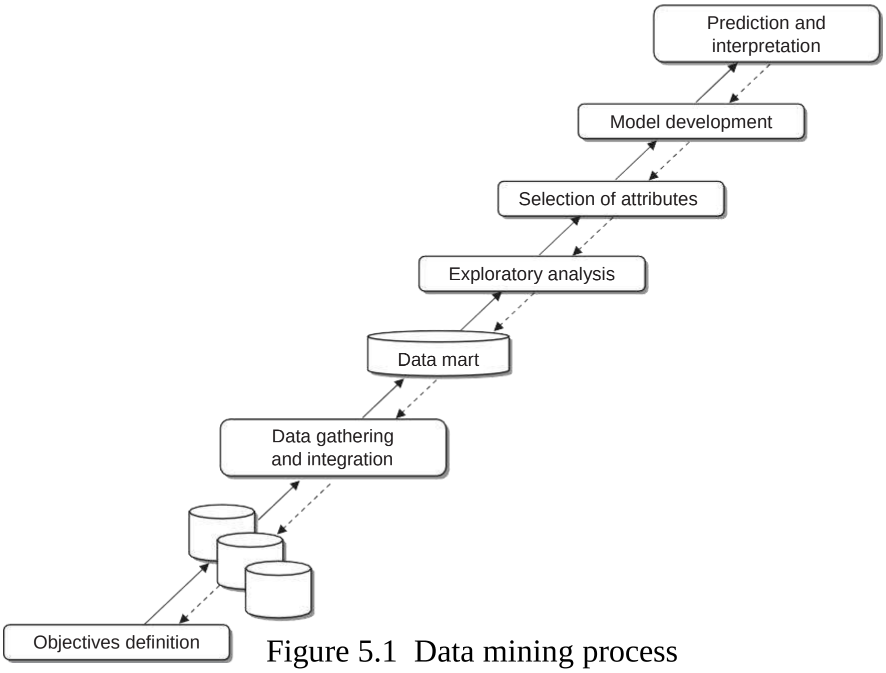


Figure 5.1 Data mining process

Definition of objectives. Data mining analyses are carried out in specific application domains and are intended to provide decision makers with useful knowledge. As a consequence, intuition and competence are required by the domain experts in order to formulate plausible and well-defined investigation objectives. If the problem at hand is not adequately identified and circumscribed one may run the risk of thwarting any future effort made during data mining activities. The definition of the goals will benefit from close cooperation between experts in the field of application and data mining analysts.

Data gathering and integration. Once the objectives of the investigation have been identified, the gathering of data begins. Data may come from different sources and therefore may require integration. Data sources may be internal, external or a combination of the two. The integration of distinct data sources may be suggested by the need to enrich the data with new descriptive dimensions, such as geo marketing variables, or with lists of names of potential customers, termed prospects, not yet existing in the company information sys- tem. In some instances, data sources are already structured in data warehouses and data marts for OLAP analyses and more generally for decision support activities. These are favorable situations where it is sufficient to select the attributes deemed relevant for the purpose of a data mining analysis. There is a risk, however, that, in order to limit memory uptake, the information stored in a data warehouse has been aggregated and consolidated to such an extent as to render useless any subsequent analysis. For example, if a company in the retail industry stores for each customer the total amount of every receipt, without keeping track of each individual purchased item, a future data mining analysis aimed at investigating the actual purchasing behavior may be compromised. In other situations, the original data have a heterogeneous format with no predefined structure. In this case, the process of data gathering and integration becomes more arduous and therefore more prone to errors. Regardless of the original structure, input datasets of data mining analyses almost always take the form of two-dimensional tables, as observed above. Unlike many standard sampling procedures of classical statistics, datasets for data mining represent samples extracted in accordance with an unknown distribution, with the analysts not being able to influence and affect the data gathering process. Chapter 6 will discuss data preparation issues in more detail.

Exploratory analysis. In the third phase of the data mining process, a preliminary analysis of the data is carried out with the purpose of getting acquainted with the available information and carrying out data cleansing. Usually, the data stored in a data warehouse are processed at loading time in such a way as to remove any syntactical inconsistencies. For example, dates of birth that fall outside admissible ranges and negative sales charges are detected and corrected. In the data mining process, data cleansing occurs at a semantic level. First of all, the distribution of the values for each attribute is studied, using histograms for categorical attributes and basic summary statistics for numerical variables. In this way, any abnormal values (outliers) and missing valuesare also highlighted. These are studied by experts in the application domain who may consider excluding the corresponding records from the investigation. Chapter 7 will discuss the techniques used to develop exploratory data analysis.

Attribute Selection. In the subsequent phase, the relevance of the different attributes is evaluated in relation to the goals of the analysis. Attributes that prove to be of little use are removed, in order to cleanse irrelevant information from the dataset. Furthermore, new attributes obtained from the original variables through appropriate transformations are included into the dataset. For example, in most cases it is helpful to introduce new attributes that reflect the trends inherent in the data through the calculation of ratios and differences between original variables. Exploratory analysis and attribute selection are critical and often challenging stages of the data mining process and may influence to a great extent the level of success of the subsequent stages.

Model development and validation. Once a high quality dataset has been assembled and possibly enriched with newly defined attributes, pattern recognition and predictive models can be developed. Usually the training of the models is carried out using a sample of records extracted from the original dataset. Then, the predictive accuracy of each model generated can be assessed using the rest of the data. More precisely, the available dataset is split into two subsets. The first constitutes the training set and is used to identify a specific learning model within the selected class of models. Usually the sample size of the training set is chosen to be relatively small, although significant from a statistical stand- point – say, a few thousands observations. The second subset is the test set and is used to assess the accuracy of the alternative models generated during the training phase, in order to identify the best model for actual future predictions. The most popular classes of learning models will be discussed in detail in the following chapters.

Prediction and interpretation. Upon conclusion of the data mining process, the model selected among those generated during the development phase should be implemented and used to achieve the goals that were originally identified. More-over, it should be incorporated into the procedures supporting decision-making processes so that knowledge workers may be able to use it to draw predictions and acquire a more in-depth knowledge of the phenomenon of interest.

The data mining process includes feedback cycles, represented by the dotted arrows in Figure 5.1, which may indicate a return to some previous phase, depending on the outcome of the subsequent phases.

Finally, we should emphasize the importance of the involvement and inter- action of several professional roles in order to achieve an effective data mining process:

- an expert in the application domain, expected to define the original objectives of the analysis, to provide appropriate understanding during the subsequent data mining activities and to contribute to the selection of the most effective and accurate model;
- an expert in the company information systems, expected to supervisethe access to the information sources;
- an expert in the mathematical theory of learning and statistics, forexploratory data analysis and for the generation of predictive models.

Figure 5.2 illustrates the competencies and the involvement in the different activities for each actor in the data mining process.

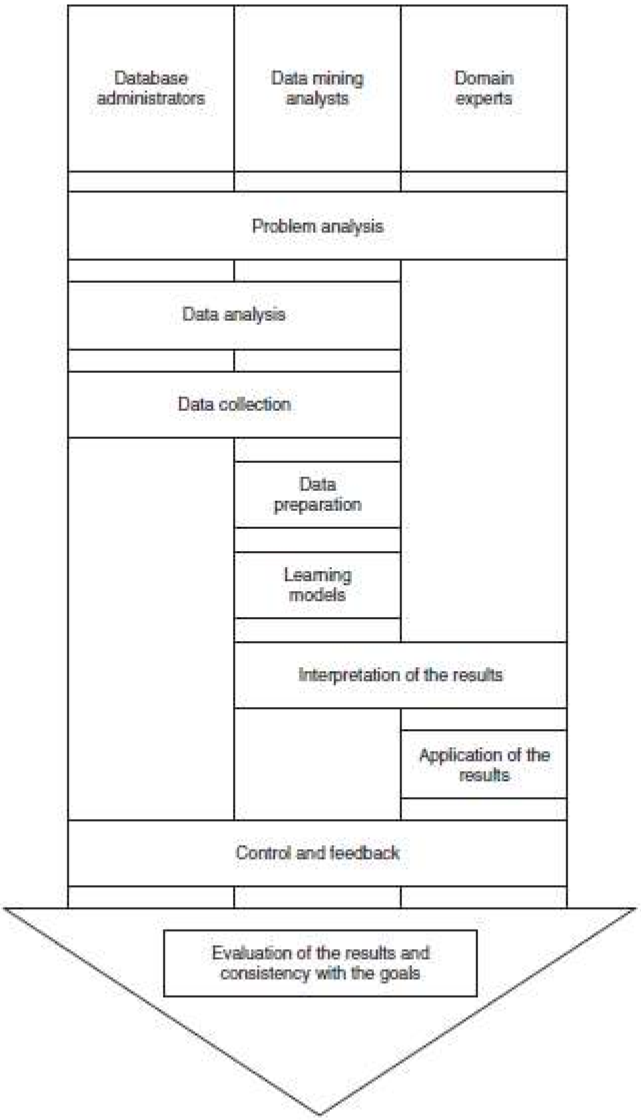


Figure 5.2 Actors and roles in a data mining process

5.5 ANALYSIS O F METHODOLOGIES

Data mining activities can be subdivided into a few major categories, basedon the tasks and the objectives of the analysis. Depending on the possible existence of a target variable, one can draw a first fundamental distinction between supervised and unsupervised learning processes.

Supervised learning. In a supervised (or direct) learning analysis, a target attribute either represents the class to which each record belongs. As an example, consider an investment management company wishing to predict the balance sheet of its customers based on their demographic characteristics and past investment transactions. Supervised learning processes are therefore oriented toward prediction and interpretation with respect to a target attribute.

Unsupervised learning. Unsupervised (or indirect) learning analyses are not guided by a target attribute. Therefore, data mining tasks in this case are aimed at discovering recurring patterns and affinities in the dataset. As an example, consider an investment management company wishing to identify clusters of customers who exhibit homogeneous investment behavior, based on data on past transactions. In most unsupervised learning analyses, one is interested in identifying clusters of records that are similar within each cluster and different from members of other clusters.

Taking the distinction even further, seven basic data mining tasks can be identified:

- characterization and discrimination;
- classification;
- regression;
- time series analysis;
- association rules;
- clustering;
- description and visualization.

The first four tasks correspond to supervised data mining analyses, since a specific target variable exists that must be explained based on the available attributes or throughout its evolution over time. The remaining three tasks represent unsupervised analyses whose purpose is the development of models capable of expressing the interrelationship among the available attributes.

Characterization and discrimination. Where a categorical target attribute exists, before starting to develop a classification model, it is often useful to carry out an exploratory analysis whose purpose is twofold. On the one hand, the aim is to achieve a characterization by comparing the distribution of the values of the attributes for the records belonging to the same class. On the other hand, the purpose is to detect a difference, through a comparison between the distribution of the values of the attributes for the records of a given class and the records of a different class, or between the records of a given class and all remaining records. This data mining task is primarily conducted by means of exploratory data analysis and therefore it is based on queries and counts that do not require the development of specific learning models. The information so acquired is usually presented to users in the form of histograms and other types of charts, as described in Chapter 7. The value of the information generated is, however, remarkable and may often direct the subsequent phase of attribute selection.

Classification. In a classification problem a set of observations is available, usually represented by the records of a dataset, whose target class is known. Observations may correspond, for instance, to mobile phone customers and the binary class may indicate whether a given customer is still active or has churned. Each observation is described by a given number of attributes whose value is known; in the previous example, the attributes may correspond to age, customer seniority and outgoing telephone traffic distinguished by destination. A classification algorithm can therefore use the available observations relative to the past in order to identify a model that can predict the target class of future observations whose attributes values are known. It is worth noting that the targetattribute, whose value is to be predicted, is categorical in classification problems and therefore takes on a finite and usually rather small number of values. In most applications the target is even represented by a binary variable. The categorical nature of the target determines the distinction between classificationand regression.

Regression. Unlike classification, which is intended for discrete targets, regression is used when the target variable takes on continuous values. Based on the available explanatory attributes, the goal is to predict the value of the target variable for each observation. If one wishes to predict the sales of a product based on the promotional campaigns mounted and the sale price, the target variable may take on a very high number of discrete values and can be treated as a continuous variable. A classification problem may be turned into a regression problem, and vice versa. To see this, a mobile phone company interested in the classification of customers based on their loyalty, may come up with a regression problem by predicting the probability of each customer remaining loyal.

Time series. Sometimes the target attribute evolves over time and is therefore associated with adjacent periods on the time axis. In this case, the sequence of values of the target variable is said to represent a time series. For instance, the weekly sales of a given product observed over 2 years represent a time series containing 104 observations. Models for time series analysis investigate data characterized by a temporal dynamics and are aimed at predicting the value of the target variable for one or more future periods.

Association rules. Association rules, also known as affinity groupings, are used to identify interesting and recurring associations between groups of records of a dataset. For example, it is possible to determine which products are purchased together in a single transaction and how frequently. Companies in the retail industry resort to association rules to design the arrangement of products on shelves or in catalogs. Groupings by related elements are also used to promote cross-selling or to devise and promote combinations of products and services.

Clustering. The term cluster refers to a homogeneous subgroup existing within a population. Clustering techniques are therefore aimed at segmenting a heterogeneous population into a given number of subgroups composed of observations that share similar characteristics; observations included in different clusters have distinctive features. Unlike classification, in clustering there are no pre- defined classes or reference examples indicating the target class, so that the objects are grouped together based on their mutual homogeneity. Sometimes, the identification of clusters represents a preliminary stage in the data mining process, within exploratory data analysis. It may allow homogeneous data to be processed with the most appropriate rules and techniques and the size of the original dataset to be reduced, since the subsequent data mining activities can be developed autonomously on each cluster identified.

Description and visualization. The purpose of a data mining process is some- times to provide a simple and concise representation of the information stored in a large dataset. Although, in contrast to clustering and association rules, descriptive analysis does not pursue any particular grouping or partition of the records in the dataset, an effective and concise description of information is very helpful, since it may suggest possible explanations of hidden patterns in the data and lead to a better understanding the phenomena to which the data refer. Notice that it is not always easy to obtain a meaningful visualization of the data. However, the effort of representation is justified by the remarkable conciseness of the information achieved through a well-designed chart.

5.6 CHECK YOUR PROGRESS

1. Define Data Mining.
2. Data mining activities can be subdivided into two major investigation streams, according to the main purpose of the analysis: __________ and ___________..
3. What is the purpose of interpretation and prediction?
4. Distinguish instances and features.

5. What is supervised and unsupervised learning?

Answers to Check your progress

1. Data mining activities constitute an iterative process aimed at the analysis of large databases, with the purpose of extracting information and knowledge that may prove accurate and potentially useful for knowledge workers engaged in decision making and problem solving.
2. interpretation and prediction
3. The purpose of interpretation is to identify regular patterns in the data and to express them through rules and criteria that can be easily understood by experts in the application domain. The purpose of prediction is to anticipate the value that a random variable will assume in the future or to estimate the likelihood of future events.
4. The rows in the dataset correspond to the observations recorded in the past and are also called examples, cases, instances or records. The columns represent the information available for each observation and are termed attributes, variables, characteristics or features.
5. In a supervised (or direct) learning analysis, a target attribute either represents the class to which each record belongs,
6. Unsupervised (or indirect) learning analyses are not guided by a target attribute.

5.7 SUMMARY

This unit highlights on data mining. The activities involved in data mining are elaborated. Various applications of data mining are detailed. The representation of data in data mining process are elucidated with suitable illustrations. Also, the process involved in data mining are described with neat diagram. The data mining tasks considering various cases are detailed.

5.8 KEYWORDS

- **Data Mining**: Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes.
- **Online analytical processing (OLAP)**: is a technology that organizes large business databases and supports complex analysis.

- **Data mining process**: Data Mining is a process to identify interesting patterns and knowledge from a large amount of data.
- **Classification:** is a data mining function that assigns items in a collection to target categories or classes.
- **Clustering**: is an unsupervised Machine Learning-based Algorithm that comprises a group of data points into clusters so that the objects belong to the same group.

5.9 QUESTIONS FOR SELF STUDY

1. Write a note on data mining.
2. Describe data mining activities.
3. Explain the applications of data mining.
4. Describe the representation of data in data mining process.
5. With a neat diagram explain the process of data mining.
6. Describe different data mining tasks.

5.10 REFERENCES

1. Sharda R, Delen D, Turban E. Business intelligence analytics and data science: A managerial perspective. Pearson, 2022.
2. Vercellis C. Business intelligence: data mining and optimization for decision making. John Wiley & Sons; 2011 Aug 10.
3. Bentley D. Business Intelligence and Analytics. Internet, link: https://www. pdfdrive. com/business-intelligence-and-analytics-e56416503. html. 2017.

UNIT-6 : DATA WAREHOUSING

Structure

6.0 OBJECTIVES

After studying this unit, you will be able to:

- Give an account on data warehouse.
- Identify three main categories of data feeding into a data warehouse.
- Explain different characteristics of data warehouse.
- Distinguish OLTP and OLAP.
- Mention the factors that may affect data quality.
- Describe the architecture of data warehouse.
- Significance of meta data.

6.1 INTRODUCTION

From the mid-1990s the need was felt for a logical and material separation between the databases feeding input data into decision support systems and business intelligence architectures on the one hand, and operational information systems on the other.

In this unit we will describe the features of data warehouses and data marts, illustrating the factors that led to their conception, and highlighting the major differences between them and operational systems, and discussing the requirements concerning data quality. Then we will examine the architecture of a data warehouse, pointing out the role of ETL tools and metadata. The last partof the unit will be devoted to on-line analytical processing operations and analyses that can be performed by using multidimensional cubes and hierarchies of concepts. We will focus our discussion on the goals and functions of a data warehouse, deliberately avoiding technical issues relating to their development.

6.2 DEFINITION OF DATA WAREHOUSE

As its name suggests, a data warehouse is the foremost repository for the data available for developing business intelligence architectures and decision support systems. The term data warehousing indicates the whole set of interrelated activities involved in designing, implementing and using a data warehouse.

It is possible to identify three main categories of data feeding into a data warehouse: internal data, external data and personal data:

Internal data. Internal data are stored for the most part in the databases, referred to as transactional systems or operational systems, that are the backbone of an enterprise information system. Internal data are gathered through transactional applications that routinely preside over the operations of a company, such as administration, accounting, production and logistics. This collection of transactional software applications is termed enterprise resource planning (ERP). The data stored in the operational systems usually deal with the main entities involved in a company processes, namely customers, products, sales, employees and suppliers. These data usually come from different components of the information system:

- back-office systems, that collect basic transactional records such as orders, invoices, inventories, production and logistics data;
- front-office systems, that contain data originating from call-center activities, customer assistance, execution of marketing campaigns;
- web-based systems, that gather sales transactions on e-commerce web- sites, visits to websites, data available on forms filled out by existing and prospective customers.

External data. There are several sources of external data that may be used to extend the wealth of information stored in the internal databases. For example, some agencies gather and make available data relative to sales, market share and future trend predictions for specific business industries, as well as economic and financial indicators. Other agencies provide data market surveys and consumer opinions collected through questionnaires.

A further significant source of external data is provided by geographic information systems (GIS), which represent a set of applications for acquiring, organizing, storing and presenting territorial data. These contain information relative to entities having a specific geographic position. Each entity is there- fore associated with latitude and longitude coordinates, along with some other attributes, usually originating from relational databases and actually depending on the application domain. Hence, these data allow to subject-specific analyses to be carried out on the data associated with geographic elements and the results to be graphically visualized.

Personal data. In most cases, decision makers performing a business intelligence analysis also rely on information and personal assessments stored inside worksheets or local databases located in their computers. The retrieval of such information and its integration with structured data from internal and external sources is one of the objectives of knowledge management systems.

Software applications that are at the heart of operational systems are referredto as on-line transaction processing (OLTP). On the other hand, the whole set of tools aimed at performing business intelligence analyses and supporting decision-making processes go by the name of on-line analytical processing (OLAP). We can therefore assume that the function of a data warehouse is to provide input data to OLAP applications.
There are several reasons for implementing a data warehouse separately from the databases supporting OLTP applications in an enterprise. Among them, we recall here the most relevant.

Integration. In many instances, decision support systems must access information originating from several data sources, distributed across different parts of an organization or deriving from external sources. A data warehouse integrating multiple and often heterogeneous sources is then required to promote and facilitate the access to information. Data integration may be achieved by means of different techniques – for example, by using uniform encoding methods, converting to standard measurement units and achieving a semantic homogeneity of information.

Quality. The data transferred from operational systems into the data warehouse are examined and corrected in order to obtain reliable and error-free information, as much as possible. Needless to say, this increases the practical value of business intelligence systems developed starting from the data contained in a data warehouse.

Efficiency. Queries aimed at extracting information for a business intelligence analysis may turn out to be burdensome in terms of computing resources and processing time. As a consequence, if a ‘killer’ query were directed to the transactional systems it would risk severely compromising the efficiency required by enterprise resource planning applications, with negative consequences onthe routine activities of a company. A better solution is then to direct complex queries for OLAP analyses to the data warehouse, physically separated fromthe operational systems.

Extendibility. The data stored in transactional systems stretch over a limited time span in the past. Indeed, due to limitations on memory capacity, data relative to past periods are regularly removed from OLTP systems and permanently archived in off-line mass-storage devices, such as DVDs or magnetic tapes. On the other hand, business intelligence systems and prediction models need to access all available past data to be able to grasp trends and detect recurrent patterns. This is possible due to the ability of data warehouses to retain historical information.

In light of the previous remarks, we can define a data warehouse as a collection of data supporting decision-making processes and business intelligence systems having the following characteristics:

Entity-oriented. The data contained in a data warehouse are primarily concerned with the main entities of interest for the analysis, such as products, customers, orders and sales. On the other hand, transactional systems are more oriented toward operational activities and are based on each single transaction recorded by enterprise resource planning applications. During a business intelligence analysis, orientation toward the entities allows the performance of a company to be more easily evaluated and any potential source of inefficiencies to be detected.

Integrated. The data originating from the different sources are integrated and homogenized as they are loaded into a data warehouse. For example, measurement units and encodings are harmonized and made consistent.

Time-variant. All data entered in a data warehouse are labeled with the time period to which they refer. We can fairly relate the data stored in a data ware- house to a sequence of nonvolatile snapshot pictures, taken at successive times and bearing the label of the reference period. As a consequence, the temporal dimension in any data warehouse is a critical element that plays a predominant role. In this way decision support applications may develop historical trend analysis.

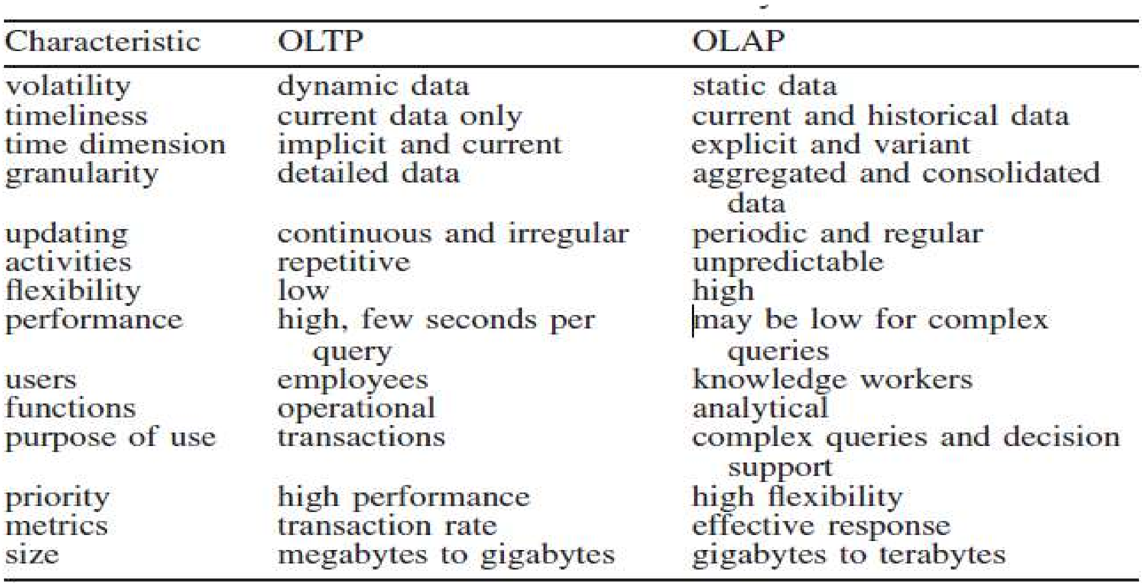
Persistent. Once they have been loaded into a data warehouse, data are usually not modified further and are held permanently. This feature makes it easier to organize read-only access by users and simplifies the updating process, avoiding concurrency which is of critical importance for operational systems.

Consolidated. Usually some data stored in a data warehouse are obtained as partial summaries of primary data belonging to the operational systems from which they originate. For example, a mobile phone company may store in a data warehouse the total cost of the calls placed by each customer in a week, subdivided by traffic routes and by type of service selected, instead of storing the individual calls recorded by the operational systems. The reason for such consolidation is twofold: on one hand, the reduction in the space required to store in the data warehouse the data accumulated over the years; on the other hand, consolidated information may be able to better meet the needs of business intelligence systems.

Denormalized. Unlike operational databases, the data stored in a data ware- house are not structured in normal form but can instead make provision for redundancies, to allow shorter response time to complex queries.

Granularity represents the highest level of detail expressed by the primary data contained in a data warehouse, also referred to as atomic data. Obviously, the granularity of a data warehouse cannot exceed that of the original data sources. In general, it is strictly lower due to consolidation aimed at reducing storage occupancy, as described above.

Table 6.1 Differences between OLTP and OLAP systems



Characteristic	OLTP	OLAP
volatility	dynamic data	static data
timeliness	current data only	current and historical data
time dimension	implicit and current	explicit and variant
granularity	detailed data	aggregated and consolidated data
updating	continuous and irregular	periodic and regular
activities	repetitive	unpredictable
flexibility	low	high
performance	high, few seconds per query	may be low for complex queries
users	employees	knowledge workers
functions	operational	analytical
purpose of use	transactions	complex queries and decision support
priority	high performance	high flexibility
metrics	transaction rate	effective response
size	megabytes to gigabytes	gigabytes to terabytes

The design philosophy behind data warehouses is quite different from that adopted for operational databases. Table 6.1 summarizes the main differences between OLTP and OLAP systems.

6.2.1 Data marts

Data marts are systems that gather all the data required by a specific company department, such as marketing or logistics, for the purpose of performing business intelligence analyses and executing decision support applications specific to the function itself. Therefore, a data mart can be considered as a functional or departmental data warehouse of a smaller size and a more specific type than the overall company data warehouse.

A data mart therefore contains a subset of the data stored in the company data warehouse, which are usually integrated with other data that the company department responsible for the data mart owns and deems of interest. For example, a marketing data mart will contain data extracted from the central data warehouse, such as information on customers and sales transactions, but also additional data pertaining to the marketing function, such as the results of marketing campaigns run in the past.

Data warehouses and data marts thus share the same technological frame- work. In order to implement business intelligence applications, some companies prefer to design and develop in an incremental way a series of integrated data marts rather than a central data warehouse, in order to reduce the implementation time and uncertainties connected with the project.

6.2.2 Data quality

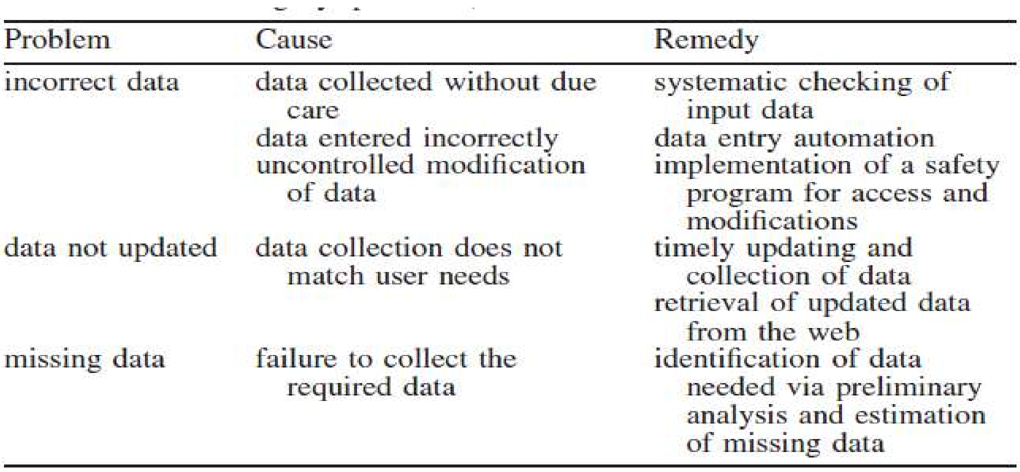
The need to verify, preserve and improve the quality of data is a constant concern of those responsible for the design and updating of a data warehouse. The main problems that might compromise the validity and integrity of the data are shown in Table 6.2. More generally, we can identify the following major factors that may affect data quality.

Accuracy. To be useful for subsequent analyses, data must be highly accurate. For instance, it is necessary to verify that names and encodings are correctly represented and values are within admissible ranges.

Completeness. In order to avoid compromising the accuracy of business intelligence analyses, data should not include a large number of missing values. How- ever, one should keep in mind that most learning and data mining techniques are capable of minimizing in a robust way the effects of partial incompleteness in the data.

Consistency. The form and content of the data must be consistent across the different data sources after the integration procedures, with respect to currency and measurement units.

Table 6.2 Data integrity: problems, causes and remedies



Problem	Cause	Remedy
incorrect data	data collected without due care data entered incorrectly uncontrolled modification of data	systematic checking of input data data entry automation implementation of a safety program for access and modifications
data not updated	data collection does not match user needs	timely updating and collection of data retrieval of updated data from the web
missing data	failure to collect the required data	identification of data needed via preliminary analysis and estimation of missing data

Timeliness. Data must be frequently updated, based on the objectives of the analysis. It is customary to arrange an update of the data warehouse regularly on a daily or at most weekly basis.

Non-redundancy. Data repetition and redundancy should be avoided in order to prevent waste of memory and possible inconsistencies. However, data can be replicated when the denormalization of a data warehouse may result in reduced response times to complex queries.

Relevance. Data must be relevant to the needs of the business intelligence system in order to add real value to the analyses that will be subsequently performed.

Interpretability. The meaning of the data should be well understood and correctly interpreted by the analysts, also based on the documentation available in the metadata describing a data warehouse, as illustrated in Section 6.2.2.

Accessibility. Data must be easily accessible by analysts and decision support applications.

6.3 DATA WAREHOUSE ARCHITECTURE

The reference architecture of a data warehouse, shown in Figure 6.1, includes the following major functional components. The data warehouse itself, together with additional data marts, that contains the data and the functions that allow the data to be accessed, visualized and perhaps modified.

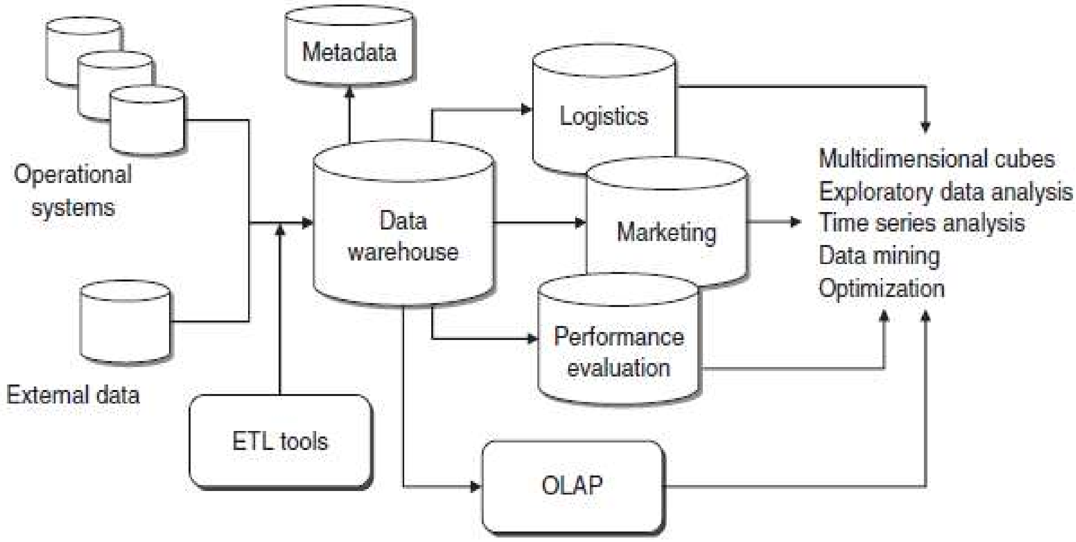


Figure 6.1 Architecture and functions of a data warehouse

- Data acquisition applications, also known as extract, transform and load(ETL) or back-end tools, which allow the data to be extracted, trans- formed and loaded into the data warehouse.

- Business intelligence and decision support applications, which represent the front-end and allow the knowledge workers to carry out the analyses and visualize the results.

- The three-level distinction applies to the architecture shown in Figure 3.1 even from a technological perspective.

- The level of the data sources and the related ETL tools that are usually installed on one or more servers.

- The level of the data warehouse and any data mart, possibly availableon one or more servers as well, and separated from those containing thedata sources. This second level also includes the metadata documentingthe origin and meaning of the records stored in the data warehouse.

- The level of the analyses that increase the value of the information contained in a data warehouse through query, reporting and possibly sophisticated decision support tools. The applications for business intelligence and decision support analysis are usually found on separate servers or directly on the client PC used by analysts and knowledge workers.

The same database management system platforms utilized to develop transactional systems are also adopted to implement data warehouses and data marts. Due to the response requirements raised by the complex queries addressed to a data warehouse, database management system platforms used for data ware- housing are subject to different structuring and parameterizations with respect to transactional systems.

A data warehouse may be implemented according to different design approaches: top-down, bottom-up and mixed.

Top-down. The top-down methodology is based on the overall design of the data warehouse, and is therefore more systematic. However, it implies longer development times and higher risks of not being completed within schedule since the whole data warehouse is actually being developed.

Bottom-up. The bottom-up method is based on the use of prototypes and there- fore system extensions are made according to a step-by-step scheme. This approach is usually quicker, provides more tangible results but lacks an overall vision of the entire system to be developed.

Mixed. The mixed methodology is based on the overall design of the data warehouse, but then proceeds with a prototyping approach, by sequentially implementing different parts of the entire system. This approach is highly practical and usually preferable, since it allows small and controlled steps to be taken while bearing in mind the whole picture.

The steps in the development of a data warehouse or a data mart can be summarized as follows.

- One or more processes within the organization to be represented in the data warehouse are identified, such as sales, logistics or accounting.
- The appropriate granularity to represent the selected processes is identified and the atomic level of the data is defined.
- The relevant measures to be expressed in the fact tables for multidimensional analysis are then chosen, as described in Section 6.3.
- Finally, the dimensions of the fact tables are determined.

6.3.1 ETL tools

ETL refers to the software tools that are devoted to performing in an automatic way three main functions: extraction, transformation and loading of data into the data warehouse.

Extraction. During the first phase, data are extracted from the available internal and external sources. A logical distinction can be made between the initial extraction, where the available data relative to all past periods are fed intothe empty data warehouse, and the subsequent incremental extractions that update the data warehouse using new data that become available over time. The selection of data to be imported is based upon the data warehouse design, whichin turn depends on the information needed by business intelligence analyses and decision support systems operating in a specific application domain.

Transformation. The goal of the cleaning and transformation phase is to improve the quality of the data extracted from the different sources, through the correction of inconsistencies, inaccuracies and missing values. Some of the major shortcomings that are removed during the data cleansing stage are:

- inconsistencies between values recorded in different attributes having the same meaning;
- data duplication;
- missing data;
- existence of inadmissible values.

During the cleaning phase, preset automatic rules are applied to correct most recurrent mistakes. In many instances, dictionaries with valid terms are used to substitute the supposedly incorrect terms, based upon the level of similar- ity. Moreover, during the transformation phase, additional data conversions occur in order to guarantee homogeneity and integration with respect to the different data sources. Furthermore, data aggregation and consolidation are performed in order to obtain the summaries that will reduce the response time required by subsequent queries and analyses for which the data warehouse is intended.

Loading. Finally, after being extracted and transformed, data are loaded into the tables of the data warehouse to make them available to analysts and decision support applications.

6.3.2 Metadata

In order to document the meaning of the data contained in a data warehouse, it is recommended to set up a specific information structure, known as metadata, i.e. data describing data. The metadata indicate for each attribute of a data warehouse the original source of the data, their meaning and the transformations to which they have been subjected. The documentation provided by metadata should be constantly kept up to date, in order to reflect any modification in the data warehouse structure. The documentation should be directly accessible to the data warehouse users, ideally through a web browser, according to the access rights pertaining to the roles of each analyst.

In particular, metadata should perform the following informative tasks:

- a documentation of the data warehouse structure: layout, logical views, dimensions, hierarchies, derived data, localization of any data mart;

- a documentation of the data genealogy, obtained by tagging the data sources from which data were extracted and by describing any trans- formation performed on the data themselves;

- a list keeping the usage statistics of the data warehouse, by indicating how many accesses to a or to a logical view have been performed;

- a documentation of the general meaning of the data warehouse with respect to the application domain, by providing the definition of the terms utilized, and fully describing data properties, data ownership and loading policies.

6.4 CUBES AND MULTIDIMENSIONAL ANALYSIS

The design of data warehouses and data marts is based on a multidimensional paradigm for data representation that provides at least two major advantages: on the functional side, it can guarantee fast response times even to complex queries, while on the logical side the dimensions naturally match the criteria followed by knowledge workers to perform their analyses.

The multidimensional representation is based on a star schema which contains two types of data tables: dimension tables and fact tables.

Dimension tables. In general, dimensions are associated with the entities around which the processes of an organization revolve. Dimension tables then correspond to primary entities contained in the data warehouse, and in mostcases they directly derive from master tables stored in OLTP systems, such as customers, products, sales, locations and time. Each dimension table is often internally structured according to hierarchical relationships. For example, the temporal dimension is usually based upon two major hierarchies: day, week, year and day, month, quarter, year . Similarly, the location dimension may be hierarchically organized as street, zip code, city, province, region, country, area . Products in their turn have hierarchical structures such as item, family, type in the manufacturing industry and item, category, department in the retail industry. In a way, dimensions predetermine the main paths along which OLAP analyses will presumably be developed.

Fact tables. Fact tables usually refer to transactions and contain two types of data:

links to dimension tables, that are required to properly reference the information contained in each fact table; numerical values of the attributes that characterize the corresponding transactions and that represent the actual target of the subsequent OLAPanalyses.

For example, a fact table may contain sales transactions and make reference to several dimension tables, such as customers, points of sale, products, suppliers, time. The corresponding measures of interest are attributes such as quantity of items sold, unit price and discount. In this example the fact table allows analysts to evaluate the trends of sales over time, either total, or referred to a single customer, or referred to a group of customers, that can be identified through any hierarchy induced by the dimension table associated with the customers. The analyst may also evaluate the trend over time of sales percentages relative to customers located in a specific region.

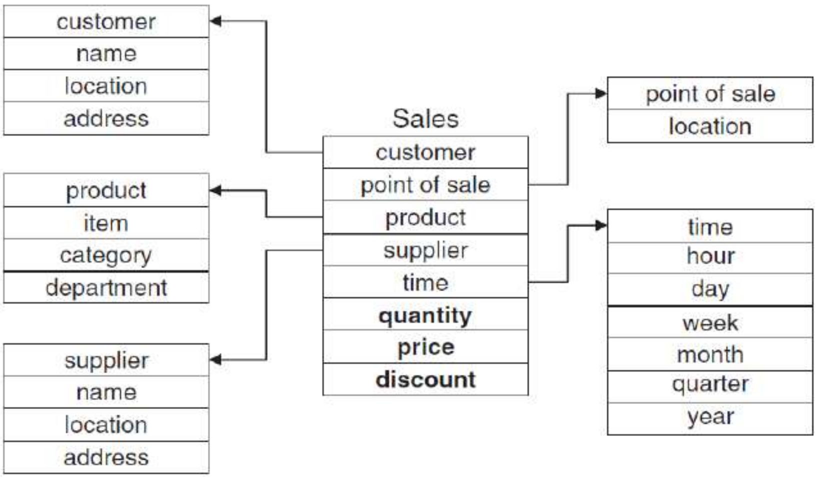


Figure 6.2 Example of a star schema

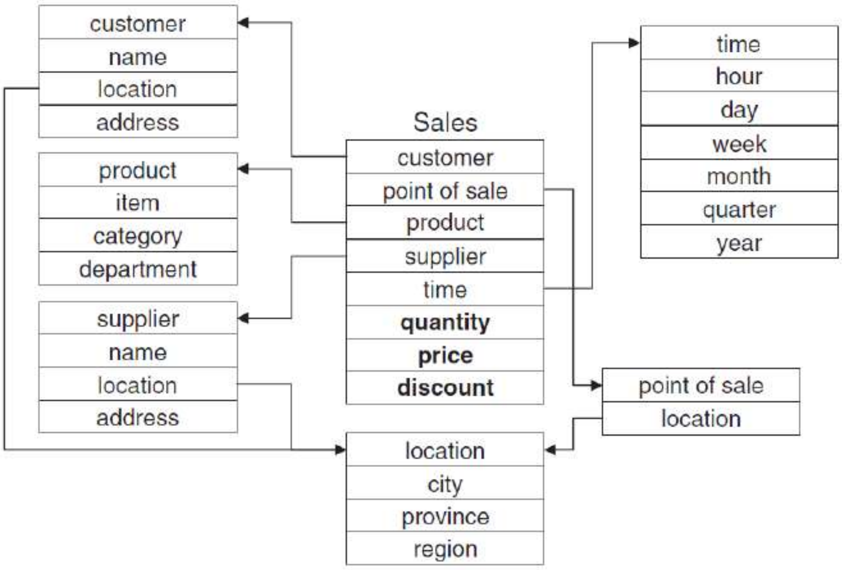


Figure 6.3 Example of a snowflake schema

Figure 6.2 shows the star schema associated with the fact table representing sales transactions. The fact table is placed in the middle of the schema and is linked to the dimension tables through appropriate references. The measures in the fact table appear in bold type.

Sometimes dimension tables are connected in their turn to other dimension tables, as shown in Figure 6.3, through a process of partial data standardization, in order to reduce memory use. In the given example the dimension table, referring to the location is in turn hierarchically connected with the dimension table containing geographical information. This brings about a snowflake schema.

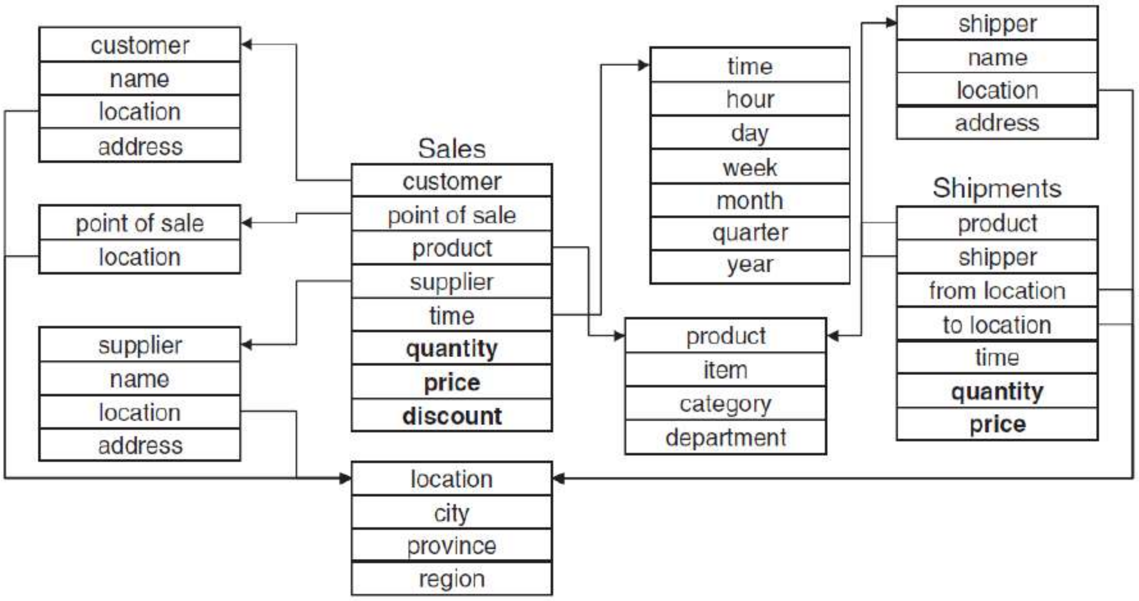


Figure 6.4 Example of a galaxy schema

A data warehouse includes several fact tables, interconnected with dimension tables, linked in their turn with other dimension tables. The latter type of schema, shown in Figure 6.4, is termed a galaxy schema.

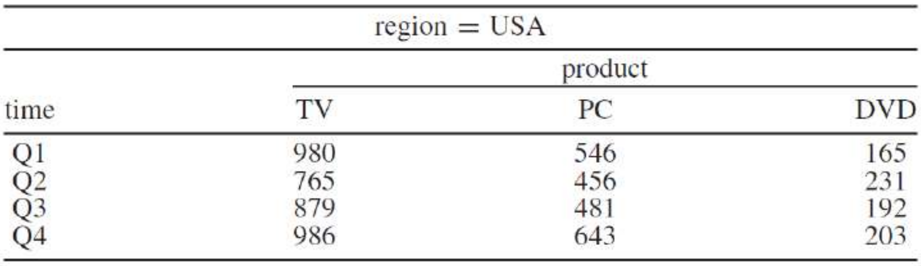
A fact table connected with n dimension tables may be represented by an n-dimensional data cube where each axis corresponds to a dimension. Multidimensional cubes are a natural extension of the popular two-dimensional spreadsheets, which can be interpreted as two-dimensional cubes. For instance, consider a sales fact table developed along the three dimensions of time, product, region. Suppose we select only two dimensions for the analysis, such as time, product, having preset the region attribute along the three values USA, Asia, Europa . In this way we obtain the three two-dimensional tables in which the rows correspond to quarters of a year and the columns to products (see Tables 6.3 –6.5).

The cube shown in Figure 6.5 is a three-dimensional illustration of the same sales fact table. Atomic data are represented by 36 cells that can be obtained by crossing all possible values along the three dimensions: time Q1, Q2, Q3, Q4 , region USA, Asia, Europa and product TV, PC, DVD. These atomic cells can be supplemented by 44 cells corresponding to the summary values obtained through consolidation along one or more dimensions, as shown by the cube in the figure.

Suppose that the sales fact table also contains a fourth dimension represented by the suppliers. The corresponding data cube constitutes a structure in four-dimensional space and therefore cannot be represented graphically. How- ever, we can obtain four logical views composed of three-dimensional cubes, called cuboids, inside the four-dimensional cube, by fixing the values of one dimension.

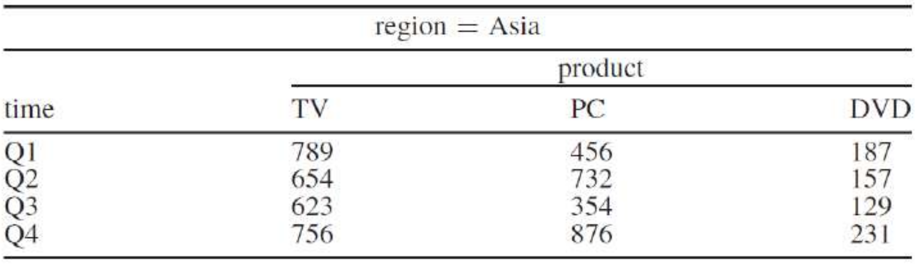
More generally, starting from a fact table linked to n dimension tables, it is possible to obtain a lattice of cuboids, each of them corresponding to a different level of consolidation along one or more dimensions. This type of aggregation is equivalent in structured query language (SQL) to a query sum derived from a group-by condition. Figure 3.6 illustrates the lattice composed by the cuboids obtained from the data cube defined along the four dimensions {time, product, region, supplier}.

Table 6.3 Two-dimensional view of sales data in the USA



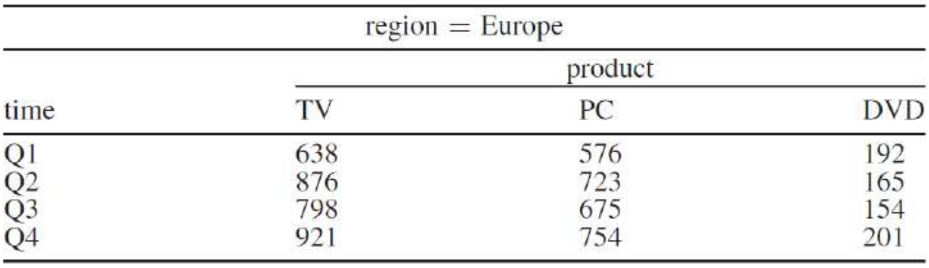
region = USA			
	product		
time	TV	PC	DVD
Q1	980	546	165
Q2	765	456	231
Q3	879	481	192
Q4	986	643	203

Table 6.4 Two-dimensional view of sales data in Asia



region = Asia			
	product		
time	TV	PC	DVD
Q1	789	456	187
Q2	654	732	157
Q3	623	354	129
Q4	756	876	231

Table 6.5 Two-dimensional view of sales data in Europe



region = Europe			
	product		
time	TV	PC	DVD
Q1	638	576	192
Q2	876	723	165
Q3	798	675	154
Q4	921	754	201

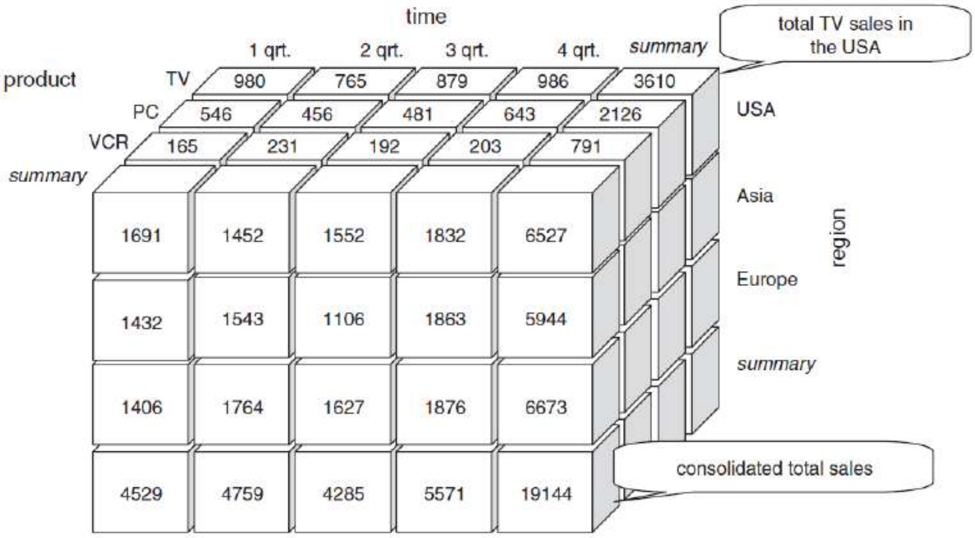


Figure 6.5 Example of a three-dimensional cube

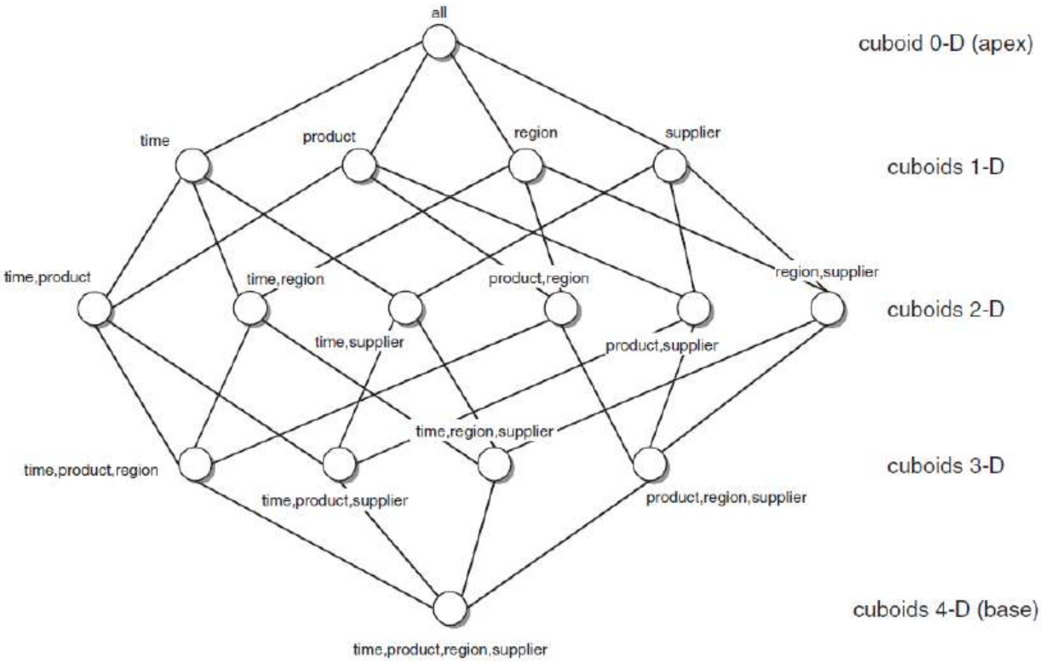


Figure 6.6 Lattice of cuboids derived from a four-dimensional cube

The cuboid associated with the atomic data, which therefore does not imply any type of consolidation, is called a base cuboid. At the other extreme, the apex cuboid is defined as the cuboid corresponding to the consolidation along all dimensions, therefore associated with the grand total of the measure of interest.

6.4 CHECK YOUR PROGRESS

1. Define Data warehouse and data warehousing.
2. Acronym for GIS.
3. Distinguish internal and external data.
4. What is data mart?
5. Fact tables usually refer to ____________

Answers to Check your progress

1. As its name suggests, a data warehouse is the foremost repository for the data available for developing business intelligence architectures and decision support systems. The term data warehousing indicates the whole set of interrelated activities involved in designing, implementing and using a data warehouse.
2. geographic information systems
3. Internal data are stored for the most part in the databases, referredto as transactional systems or operational systems, that are the backbone of anenterprise information system. There are several sources of external data that may be used to extend the wealth of information stored in the internal databases.
4. Data marts are systems that gather all the data required by a specific company department, such as marketing or logistics, for the purpose of performing business intelligence analyses and executing decision support applications specific to the function itself.
5. transactions

6.5 SUMMARY

This unit highlights on data warehouse. The three main categories of feeding the data into data warehouse are detailed. Different characteristics of data warehouse are discussed with suitable illustrations. Differentiation between OLTP and OLAP is analyzed. The factors that may affect the data quality are detailed in the context of data warehouse. Data integrity problems, its causes and remedies are highlighted. Architecture of data warehouse is illustrated and explained with suitable diagram. Metadata analysis has been performed.

6.6 KEYWORDS

- **Data warehouse**: is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics.
- **Data integrity**: Data integrity refers to the accuracy and consistency (validity) of data over its lifecycle.
- **OLTP and OLAP**: Online transaction processing (OLTP) captures, stores, and processes data from transactions in real time. Online analytical processing (OLAP) uses complex queries to analyze aggregated historical data from OLTP systems.
- **Data mart:** A data mart is a subset of a data warehouse focused on a particular line of business, department, or subject area.
- **Fact table** :A fact table is the central table in a star schema of a data warehouse. A fact table stores quantitative information for analysis and is often denormalized.

6.7 QUESTIONS FOR SELF STUDY

1. Write a note on data warehouse.
2. Describe three main categories of data feeding into a data warehouse.
3. Explain different characteristics of data warehouse.
4. Distinguish OLTP and OLAP.
5. Give an account on the factors that may affect data quality.
6. What are data integrity problems, causes and remedies?
7. Describe the architecture of data warehouse.
8. Write the significance of meta data.

6.8 REFERENCES

1. Sharda R, Delen D, Turban E. Business intelligence analytics and data science: A managerial perspective. Pearson, 2022.
2. Vercellis C. Business intelligence: data mining and optimization for decision making. John Wiley & Sons; 2011 Aug 10.
3. Bentley D. Business Intelligence and Analytics. Internet, link: https://www. pdfdrive. com/business-intelligence-and-analytics-e56416503. html. 2017.

UNIT -7: AGILE DEVELOPMENT

Structure

7.0 OBJECTIVES

After studying this unit, you will learn:

- ✓ Basics of waterfall development process
- ✓ Agile development techniques
- ✓ Basic concepts of scrum
- ✓ Best practices for BI

7.1 WATERFALL DEVELOPMENT PROCESS

Traditional systems development projects often follow a waterfall project approach: A set of tasks is completed, and then another set, until several months or years later, you have a working piece of software (see Figure 7-1). The waterfall approach is heavy on defining requirements precisely up front. The thinking goes that if you get your requirements right up front, then you save development costs later in the process. The waterfall approach is also preferred when a development project is outsourced and a systems provider must build a solution to a specification.

Such a project approach is reasonable for *portions* of a business intelligence solution and as long as the time frames are reasonable, but it is less effective for business-facing solutions when requirements are difficult to articulate and frequently change and processes are fluid. With business intelligence, the project is never-ending and the focus is not on finishing, but rather, on delivering a certain set of capabilities within a defined period. one of the ways in which business intelligence is used is to uncover opportunities. Requirements for discovery-style applications, then, are not precisely known. Instead of a fixed report or dashboard, the BI application has to facilitate exploration of a broad set of data.

How BI can be most relevant to front-line workers? the requirements-definition process is much more collaborative versus the traditional, somewhat rigid process of "define requirements precisely and build to the specification." These fundamental aspects of business intelligence make the waterfall approach to project management inappropriate to much of the BI initiative. some of the early failures of data warehouse projects can be attributed to the use of a waterfall approach in which the data warehouse team spent a year or more building out enterprise architecture, later delivering a system not at all useful to the business.

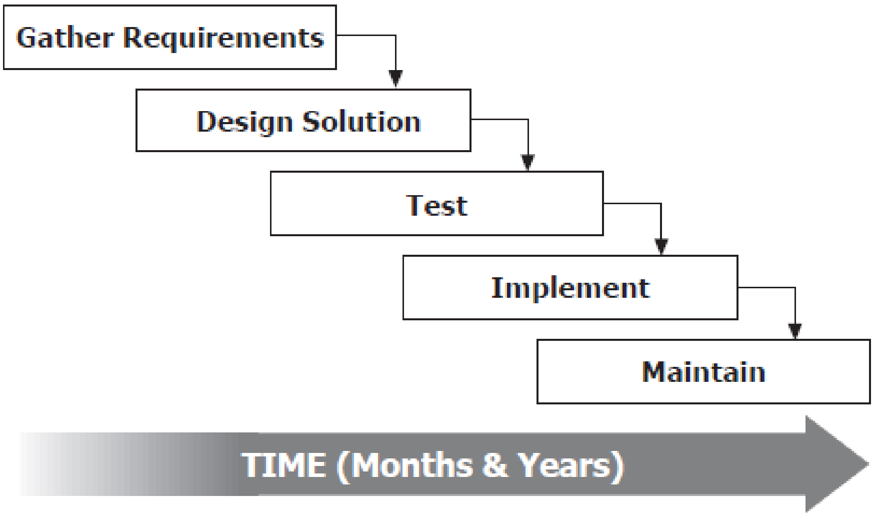


Fig. 7.1 Waterfall project methodology

Within the BI architecture (see Figure 7-2.), making changes to items on the far left (source systems and extract, transform, and load [ETL] processes) is often more costly to do, requires more time, has a greater risk, and may have less of an immediate value-add to the business. Items farther on the right (dashboards, reports, alerts) are less time-consuming to change and therefore more adaptable to changing business requirements.

Specific elements are listed in Table 7-1. For each portion of the BI architecture, you may want to adopt a periodic release schedule, but a schedule that balances the need for stability with responsiveness. Items on the far left may only change every few years; those in the middle, once a quarter; and items on the further right, on an as-needed basis (daily, weekly, or monthly). The frequency for change varies due to the cost of change, the degree of difficulty to change, the number of people and related components affected by the change, risk, and the corresponding business value provided by the change.

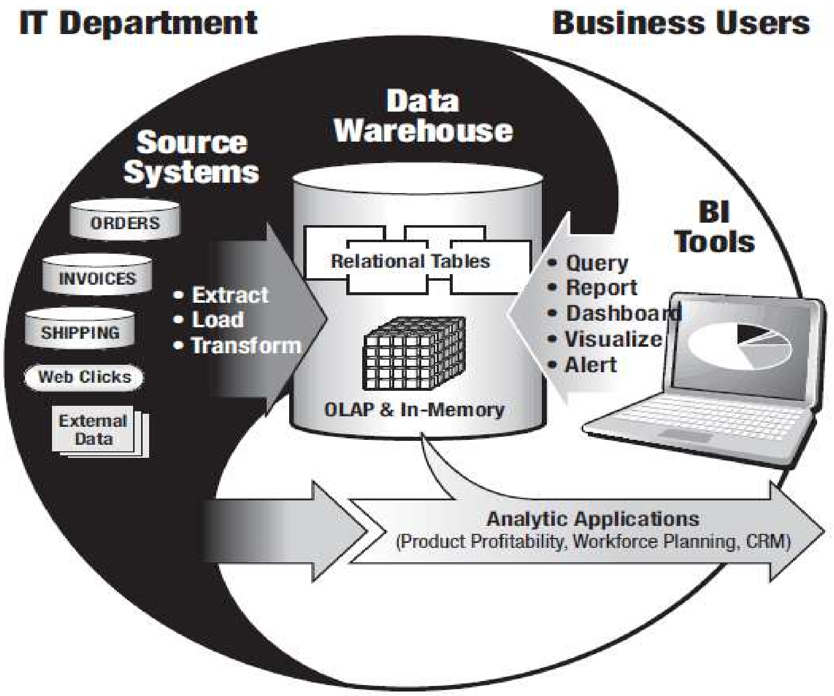
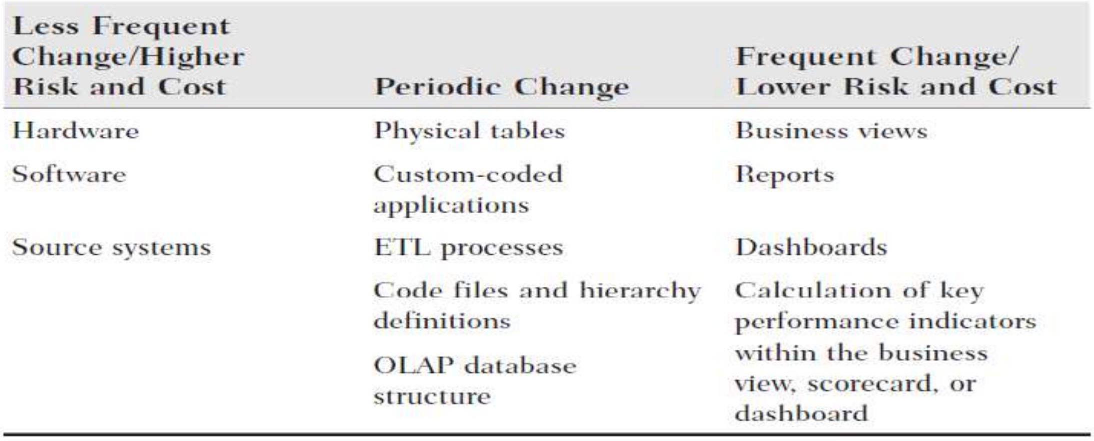


Fig. 7.2 Major components in the business intelligence life cycle

Table 7.1 Specific elements requiring change in the BI architecture

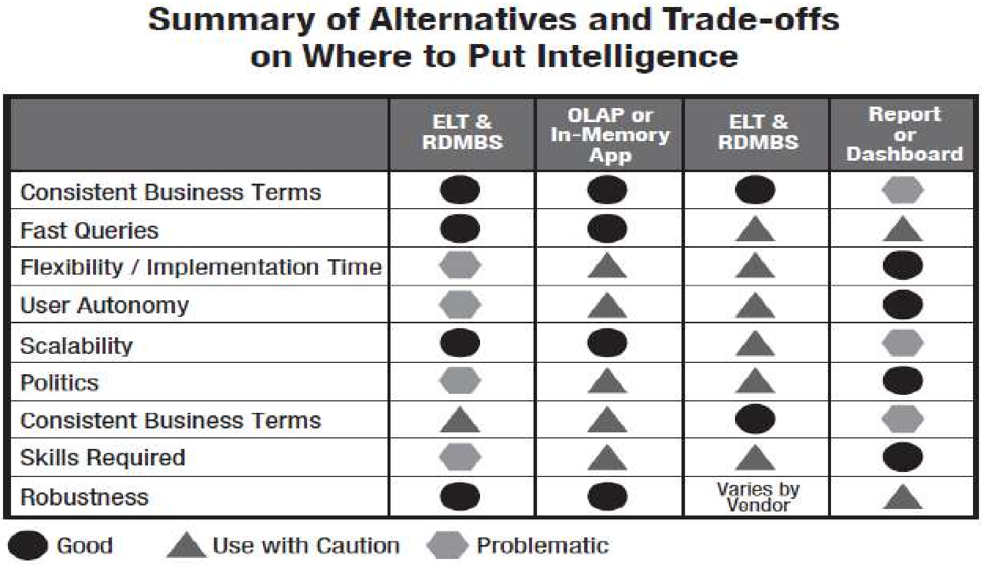


Less Frequent Change/Higher Risk and Cost	Periodic Change	Frequent Change/ Lower Risk and Cost
Hardware	Physical tables	Business views
Software	Custom-coded applications	Reports
Source systems	ETL processes	Dashboards
	Code files and hierarchy definitions	Calculation of key performance indicators within the business view, scorecard, or dashboard
	OLAP database structure	

As an example, getting various stakeholders and individual lines of business to agree on consistent business definitions is difficult and time-consuming. Important metrics such as “customer churn” or “product profitability” can be calculated in a myriad of ways. Once everyone agrees on a definition, however, implementing a consistent calculation of such business metrics within a business view or scorecard is something that can be implemented rapidly. If, however, the definition or calculation logic has been hard-coded into ETL processes or into physical tables in the data warehouse, then consolidating and changing these business rules can mean a major overhaul to multiple programs.

Sometimes developers will hard-code business definitions into individual reports or dashboards: Stakeholders can’t agree, so a report is the “easiest” and fastest place to define an element. This has some short-term value until there is a new business rule. Now those hundreds of instances of “customer churn” or “product profitability” have to be changed in hundreds of individual reports, as opposed to in one business view. Such business-facing capabilities demand flexibility. Other components, such as the hardware for the BI server or data warehouse, may only need to be changed when a company wants to update the infrastructure, add capacity, or exploit a new technology.

For every BI element, consider carefully where to place the capability and what promotes the most reusability and flexibility while balancing the trade-offs in risk, cost, and business benefit. Figure 7-3 provides a summary of trade-offs in cost, benefit, and flexibility of where to put the intelligence in various parts of the BI life cycle.



Summary of Alternatives and Trade-offs on Where to Put Intelligence

	ELT & RDMBS	OLAP or In-Memory App	ELT & RDMBS	Report or Dashboard
Consistent Business Terms	●	●	●	⬢
Fast Queries	●	●	▲	▲
Flexibility / Implementation Time	⬢	▲	▲	●
User Autonomy	⬢	▲	▲	●
Scalability	●	●	▲	⬢
Politics	⬢	▲	▲	●
Consistent Business Terms	▲	▲	●	⬢
Skills Required	⬢	▲	▲	●
Robustness	●	●	Varies by Vendor	▲

● Good ▲ Use with Caution ⬢ Problematic

Figure 7-3 Alternatives and trade-offs in where to put the intelligence

For example, if your requirement is to calculate customer churn, you may write the logic to do this in:

- The ETL or ELT script that then populates the data warehouse
- An Online Analytical Processing (OLAP) cube or in-memory application that an OLAP viewer, visual discovery tool, or dashboard may access
- The business view or business meta data layer of a BI tool
- As a calculation within an individual report or dashboard At one end of the spectrum in which IT is strongly involved in developing the solution, logic inside an ETL or ELT script provides the following benefits:
- Consistency of business terms across all applications and reports that would use this metric
- Fast performance, as queries that use the calculation would access data physically stored in the relational data warehouse or loaded into memory
- Good scalability, as large volumes of data and large numbers of users can reuse this
- Low cost to maintain after the initial implementation, but frequent changes can be expensive
- Robust modeling and calculation logic that can handle multiple data passes, if-then-else logic, and so on

However, building intelligence in the ETL script provides the following disadvantages:

- Less flexibility and a longer implementation time up front.
- No business user autonomy to change the way something is calculated.
- Political challenges to establish how to calculate the metric, requiring consensus from all business units and stakeholders. If marketing defines churn differently from finance, such differences in definitions need to be resolved before the ETL process can be written.
- Highly skilled ETL developers are required to understand distinct data sources, data integration tools, and programming, so there may be a bottleneck or additional cost. At the other end of the BI life cycle, an individual business power user may calculate customer churn inside a dashboard or report. This approach provides the following advantages:
- Strong flexibility and a fast implementation time.
- Strong business user autonomy to change the way something is calculated.

- Minimal to no political obstacles. Only the requirements of the individual business unit are considered in defining the calculation logic. The needs of the larger organization do not need to be considered.
- Business users can implement the design and only need limited training in a BI tool.
- When a business user implements intelligence inside a report or dashboard, it poses the following disadvantages:
- Inconsistent business terms when other report authors or dashboard developers want to use a similar metric that they may inadvertently or intentionally calculate differently.
- Variable performance, depending on if the back-end source is an in memory application or relational database. Query performance may suffer when there is complex SQL generated at query run time.
- Poor scalability when there are large volumes of data or large numbers of users accessing the calculation.
- Higher cost to maintain because, when there is a change, each individual report or dashboard needs to be modified.
- Less robust calculation logic than with other points in the BI life cycle, but capabilities vary widely.

7.2 AGILE DEVELOPMENT TECHNIQUES

The concept of agile software development emerged from an informal gathering of software engineers in 2001.The group published a manifesto, some of whose principles aptly apply to business intelligence. With agile development, BI developers do not work from a precise list of requirements, in stark contrast to the waterfall approach. Instead, they work from a broad requirement, with specific capabilities that are identified and narrowed down through a prototyping process. This prototyping process may involve sample screens mocked up within an Excel spreadsheet, or reports and dashboards built within a BI tool. When using packaged BI software, building a report or dashboard takes a matter of minutes and hours, not days and weeks of custom-coded solutions. Discarding a prototype after a collaborative session is more expeditious than asking the business users to list precisely their requirements, having someone build a solution to those requirements, and then discovering that the requirements have changed or that there was a misinterpretation.

A project plan for a BI solution using agile development techniques is illustrated in Figure 7-4. A specific task is iterated and recycled until the project team is satisfied with the capabilities, within a defined time frame, and in adherence to the resource constraints (time and people) agreed upon in the planning stage. Time frames are usually measured in weeks (as opposed to months and years in waterfall-style projects). In this way, there is not a concept of a project being late. Instead, requirements and deliverables are time boxed. So the question is not whether or not the project was late, but rather, were the requirements met and of an appropriate quality.

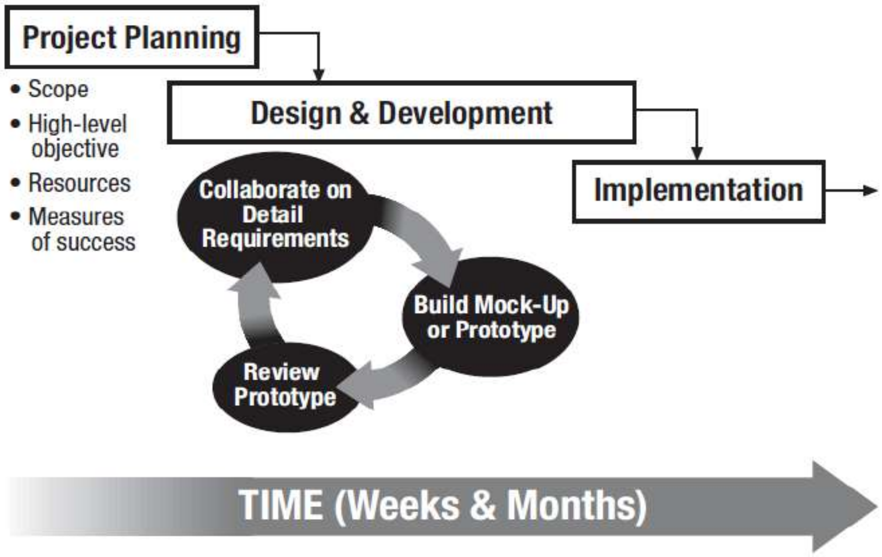


Figure 7-4 Iterative approach to delivering BI capabilities

A Subset of Principles from the Agile Manifesto

- Our highest priority is to satisfy the customer through early and continuous delivery of valuable software.
- Welcome changing requirements, even late in development. Agile processes harness change for the customer’s competitive advantage.
- Businesspeople and developers must work together daily throughout the project.
- The most efficient and effective method of conveying information to and within a development team is face-to-face conversation.
- The sponsors, developers, and users should be able to maintain a constant pace indefinitely.
- Continuous attention to technical excellence and good design enhances agility.
- Simplicity—the art of maximizing the amount of work not done— is essential.

- The best architectures, requirements, and designs emerge from self-organizing teams.

For this iterative process to be successful, the business users and the IT developers must work closely together in a collaborative fashion. Some BI project teams will establish "war rooms" to facilitate collaboration in which business users and IT developers routinely meet to review prototypes and hash out requirements. In addition to logistical issues such as co-location in war rooms, in order for such collaborative development to be successful, the business and IT must have a strong Partnership.

The State of Agile Software Development

According to the Successful BI survey, 15 percent of respondents strongly agree that they are using agile development techniques, and 44 percent are using them to some extent. A sizable minority (41 percent) are not using agile at all. The influence on business impact, though, is significant. As shown in Figure 7-5, those that strongly agree they use agile, 46 percent, report significant business impact, 12 percentage points higher than the industry average of 34 percent.

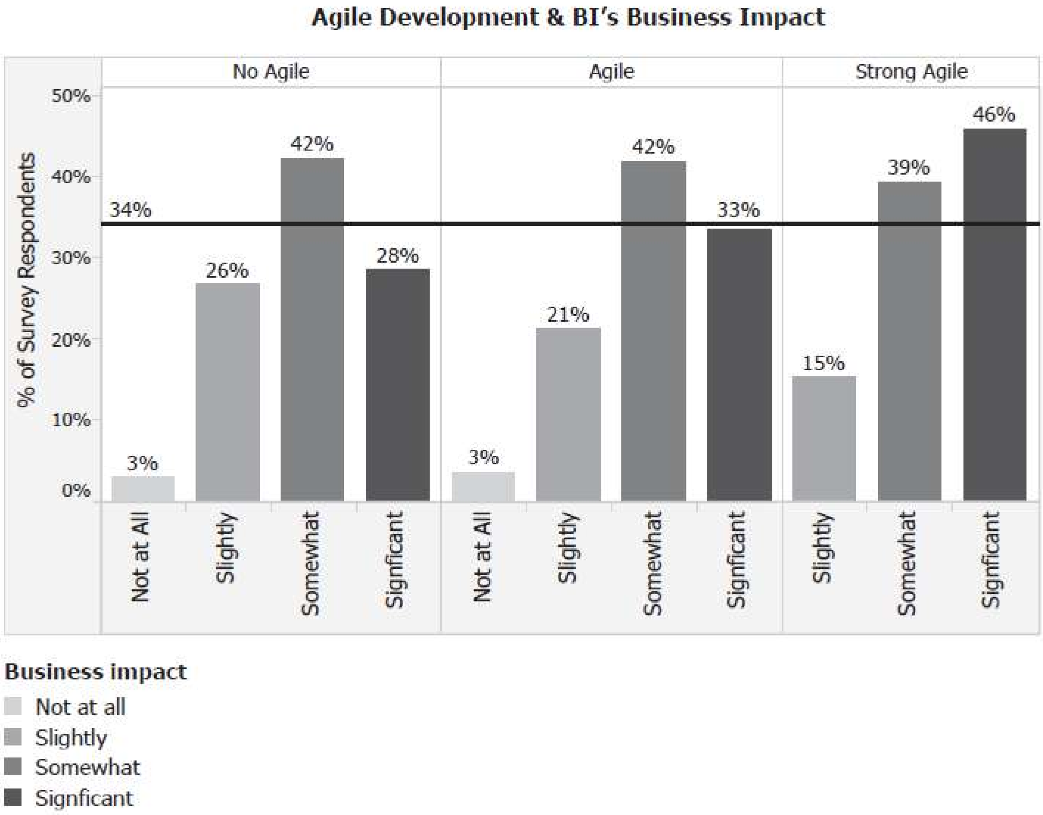


Fig 7.5 Use of agile development relates to greater business impact.

Industry literature suggests that some of the barriers to adoption of agile development are concerns about higher costs, loss of control, and inability for the business and IT to partner together. Scott Ambler, an author of several books on agile software development, conducted a broad survey in March 2007 (781 respondents) with an updated version in July 2010 (233 respondents). Some key findings in support of agile software development include the following:

- Small teams of one to ten people report the highest success rates (83 percent).
- Co-located agile projects are more successful on average than non co-located, which in turn, are more successful than projects involving off-shoring.
- Regardless of team size, agile showed higher success rates than traditional waterfall development.

A Recognized Need for Agile

With the frenetic pace of business, business intelligence needs to be able to adapt at an equally rapid pace to new requirements and changes. Agile development can help achieve flexibility and rapid delivery, but it requires the right culture, business–IT partnership, and an understanding of new development approaches. A number of Successful BI survey respondents wrote of the need for more agility in delivering BI solutions. A senior systems accountant voiced frustration at the disconnect: “IT is very reluctant to get involved with business requirements and manages projects in a very linear, waterfall approach, which turns quite basic data warehousing and BI requests into long, drawn-out process which fail to deliver what is needed as an end output. The business goes back to workarounds and Excel.”

A supply chain manager in manufacturing blames their lack of BI success on slow delivery times. “Lack of a lean IT deployment process; it takes too much time; is too costly, and is not prepared to anticipate future needs and developments.” Conversely, a systems developer who has been using agile development credits their BI success to this development approach. “A good relationship with business is essential, and we have a good experience of scrum with the business BI-manager as the product owner.”

7.3 Basic Concepts of Scrum

There are different approaches to agile development, but scrum seems to be the most widely used. Scrum.org publishes a guide on scrum development techniques and provides training and certifications. It uses self-organizing teams to develop capabilities within a specific time frame. Following are some of the key terms that anyone involved with a BI team using agile should be familiar with:

- **Product owner** A single person responsible for the completed product and for deciding what's in scope and what's out of scope, setting priorities, and managing the list of requirements or product backlog.
- **Scrum master** The team leader who ensures scrum theories and practices are being followed.
- **Sprints** A development time, usually a month, in which a set of product capabilities is delivered. A release cycle may be composed of multiple sprints.
- **Product backlog** A list of requirements or capabilities needed in the deliverable. These may be captured as user stories.
- **Co-location** IT developers and business users will be located in the same physical room to facilitate collaborative development.
- **Task board** A wall or chart that shows the progress of each story. It usually consists of the following columns: Story by Priority, Tasks Waiting, Tests Written, Under Development, Waiting Validation, and Ready to Demo. The last step, Ready to Demo, is when the development team confirms with the product owner that all requirements for that sprint have been met.
- **Swim lanes** Because the task board has been organized into columns that appear as swim lanes in a lap pool. Items can be reshuffled in priority and phase within the task board.

7.3.1 Basic Concepts of Kanban

Kanban is another agile development approach. In Japanese, Kanban means "signal card" and is an approach that Toyota uses in its production system to signal when a phase of work has completed decentralized manufacturing. Where scrum is time boxed, Kanban is focused on continuous development. Both approaches rely heavily on the concept of teams. Several of the Successful BI case study companies use a combination of kanban and scrum. Kanban includes four main principles:

- Assess current development processes
- Pursue incremental, evolutionary change
- Respect the current process, roles, responsibilities, and titles
- Leadership at all levels

With Kanban, the focus is on reducing work in progress and continuing to move outstanding requests through the development process.

7.3.2 How Well Are BI Projects Managed?

Agile development processes may require different and perhaps stronger project management skills than a waterfall approach. Collaborative design sessions that are characteristic of agile development can too easily slip into never-ending tweaks to the system. Without a detailed requirement document, it's harder for project personnel to declare a particular item is out of scope. According to the Successful BI survey results, having a well-managed BI program ranked sixth in importance for organizational factors, with 24 percent rating this as essential to a successful business intelligence deployment. It seems that data warehouse failures, wasted investments, and late projects were reported more often in the mid-1990s, when the concepts of data warehousing and business intelligence were still new.

Nonetheless, the stigma of project failures still seems to linger and is perhaps exaggerated. we hear that new vendors and consulting companies saying most BI projects fail, which the survey results clearly show is not true. Research by Professor Hugh Watson of the Terry College of Business at the University of Georgia in 2005 showed that only a slight majority of data warehouse projects then were on time and on budget. A sizable portion of data warehouse projects, percent on average, were late.

The degree to which data warehouse projects were over budget was also sizable at 37 percent.

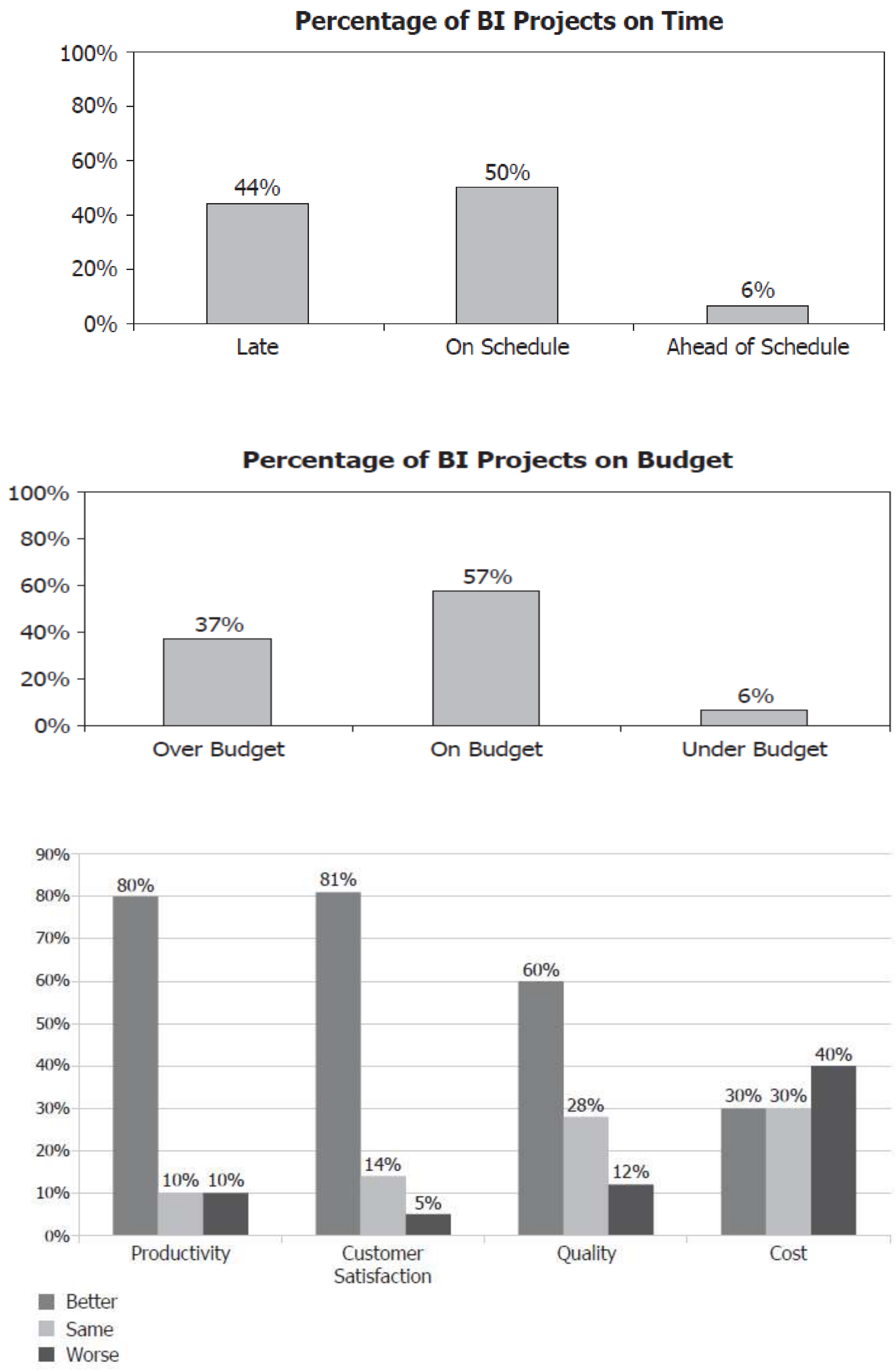


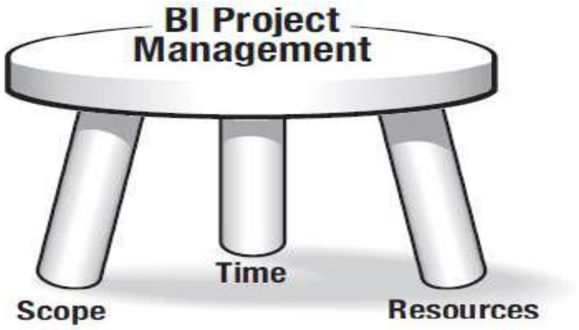
Fig 7.7 Ralph Hughes and TDWI: Agile's impact on BI project key performance indicators

More recent data shows improvement in productivity, customer satisfaction, and quality when agile development methodologies are used. As shown in Figure 7-7, a joint survey conducted by Ralph Hughes of Ceregenics and TDWI in 2012 (204 respondents) found that 80 percent had better productivity, 81 percent had better customer satisfaction, and 60 percent had better quality when using agile over traditional waterfall development. The only project performance indicator that did not have a major improvement was cost, for which 40 percent said the cost was worse.

There are three key variables in managing a BI project effectively:

- **Scope** For example, the subject areas and data accessible for analysis, the underlying infrastructure, the BI tool capabilities, and the quality
- **Resources** The amount of money and number of people you have available to invest in the project
- **Time** The deadline for delivering a set of capabilities

Like a three-legged stool, when any one of these variables changes, it affects the other variables.



So when the business asks for more data than originally agreed upon in the scope, either

- You need more *resources* or better productivity to deliver the changed scope on time. or
- The resources will stay fixed and the project *timeline* must be renegotiated.

Unfortunately, 44 percent of the Successful BI survey respondents said they do not have adequate time and funding to be successful.

Quality is part of the project scope, and this is an aspect that can sabotage the timeliness of any project, no matter how well planned. When the severity of data quality problems is not known, allowing appropriate time to handle such issues is guesswork. In an ideal world, data would be 100 percent accurate, software would be bug-free, and functionality would be as expected. That's not reality. One of the most challenging aspects to project management, then, is delivering a solution whose quality is good enough within the agreed-upon time constraints and available resources.

7.4 Agile Culture at Netflix

Agile is not just a development approach at Netflix; it is part of the company culture. The company actively recruits people who are willing to take risks, think out of the box, and work with a great deal of freedom as a team member. One of the major differences in waterfall development versus agile development is the idea of control and individual freedom. With waterfall development, a developer is assigned a task by a supervisor. It is much more suited to a hierarchical organization and culture. With agile development, the team will agree on who works on which tasks for maximum value and efficiency. Team members are free to make decisions and voice concerns or alternatives. In fact, normally a daily stand up is part of the agile development process. This type of work style requires the right people and culture.

To a certain extent, that Netflix is in the entertainment industry and is an innovator allows and requires agility, so they actively recruit top performers able to work in such an environment. Netflix CEO Reed Hastings says, "In procedural work, the best are two times better than the average. In creative/inventive work, the best are ten times better than the average, so there is a huge premium on creating effective teams of the best."

Across the industry, IT often has been criticized for moving too slowly, but conversely, what happens when the business moves too fast? For example, in 2011, Netflix announced changes to its subscription plans, initially trying to separate DVD and streaming customers. There was a big customer backlash that sent the share price plunging. Later in the fall, CEO Reed Hastings announced that a separate company, Qwikster, would handle DVD subscribers, and a month later, the company backtracked.

In explaining these changes, CEO Hastings said, "There is a difference between moving quickly—which Netflix has done very well for years—and moving too fast, which is what we did in this case."

IT has to keep pace with such changing business priorities. Andrew Dempsey, director of DVD BI and analytics, says the speed of the business can sometimes be a challenge in ensuring BI success.

Sometimes the business is too fast. The Netflix culture is faster than agile. There is a lot of freedom and responsibility so you need a higher level of communication. Changes in one system impact another, and they are done without really checking. For example, we'll get a new data feed in [the] morning, it will have something new in the afternoon, and it's impacted basic reporting. Whilst the rate of change of data does impact standard reporting, we also have the agility to react to it quickly and can thus stay in sync with all the changes going on around us.

The culture and right people have enabled Netflix to be agile, but so have rapid changes in technology. The use of public cloud and open source have been pivotal in allowing Netflix to launch streaming in new markets, such as to Europe in 2012. Ariel Tseitlin, director of cloud solutions, explains, "Every engineer who needed cloud resources was able to procure them at the click of a button. The elastic nature of the cloud makes capacity planning less crucial, and teams can simply add resources as needed."

While agile is part of Netflix, the company clarifies that they can adopt this approach because they "are in a creative-inventive market, not a safety-critical market like medicine or nuclear power." This is an important point of contrast for a company such as Medtronic.

7.5 Medtronic: Agile for the Right Projects

For Medtronic, one of the keys to success was using agile development techniques when and where it made sense rather than adopting the methodology in its entirety. Collaborative development is a fundamental concept of agile, and to this end, Medtronic had three full-time business analysts dedicated to the reporting aspect of the Global Complaint Handling (GCH) system. These business analysts teamed up with individuals in the business who knew the details on what had to go into FDA audit needs, weekly scorecards, or quarterly metrics for senior management.

"They worked side by side in flushing out the requirements," explains IT Director Sarah Nieters, who acted as the IT sponsor for GCH.17 Co-developing reports was new to Medtronic, and the team set up war rooms for individual businesses. The agile concept of a "task board" was used, with the status of various reports posted on the wall: Design, Complete, Written, Validated.

Another concept of agile is voicing alternatives to ensure maximum quality, value, and expediency. Sara Rottunda, business lead on the project, suggests there needs to be more of this mindset. "Don't just take the order. BI developers should push back and engage critical thinking. Tell us: Did you know that another business unit just asked for the same thing?"

Rottunda makes a valid point, but this is where company culture and adequate resources have to be in place before BI specialists or IT developers in general will challenge or probe business requirements. If a developer fears for his job or is perceived as being a second-guesser, such critical thinking and dialogue will rarely happen.

Similar to Netflix, changes in technology also played a role in allowing Medtronic to be more agile, but at the same time, use of agile development on the vendor's part presented its own set of challenges.

Medtronic was the fourth live customer in the United States on a new technology, SAP Hana, an in-memory appliance. Medtronic selected the technology for its performance, but also, because it could handle long text fields. In the past, Medtronic couldn't readily search or analyse comments because its relational data warehouse had a 60-character limit. Kiran Musunuru, the SAP HANA architect at Medtronic, recalls, "Bleeding edge technology had some challenges. We got a new vendor release every two weeks." Despite these challenges, Nieters says, "When you look at what we have now and the capability, it's a huge leap forward in capability. It's been worth the pain."

7.6 Sharper BI at 1-800 CONTACTS

1-800 Contacts implemented agile software development methodology early in its BI journey in 2005. Prior to this, users had to define their requirements in advance and formally submit them to the IT group. Now the BI team meets with various businesspeople on a weekly basis to plan the week's iterations. Dave Walker, the vice president of operations at 1-800 Contacts, describes the dynamics of agile development as one of the reasons for their success. "We are virtually one team. The IT people in the data warehouse team understand the call center so well, they could probably take some calls. There is partnership, high trust, and it's collaborative. It’s not ‘make a list, send it over.’ It’s very iterative. It takes lot of time and effort on both sides, but the end product is well worth it.”

The team still works within a high-level roadmap with yearly deliverables, and Jim Hill, director of data management, says these weekly planning sessions could not work without that roadmap. Disagreements about prioritizations and resource allocation are resolved by a finance director who reports to the executive sponsor.

In many respects, the BI technology itself allows for agile development because the business users themselves may be building the solution.

If users are building or customizing their own reports and dashboards, they most likely are not working from a documented list of requirements, but rather working from, at most, needs and thoughts jotted in an e-mail request. Chris Coon, a senior analyst at 1-800 CONTACTS, says the Microsoft Analysis Services OLAP cube allows for exploration. “Before the data warehouse and these cubes, we always had to go to the IT group who produced something static. It always took a long time. It didn’t facilitate a rapid response to change in sales volume or other business event.”

Now Coon estimates 80 percent of his requirements can be fulfilled by the OLAP database, allowing him to explore sales by new customers, by repeat customers, or by different products.

7.7 Best Practices for Successful Business Intelligence

Project managers should recognize that because of the ways in which business intelligence is used, solutions must be flexible and modifiable in response to changing business requirements. Given the lack of understanding of what is possible with BI and that users often don’t know what they want until they see it, agile development techniques are preferable to traditional waterfall development process for BI applications.

- Be prepared to change the business-facing parts of BI on a more rapid basis than the behind-the-scenes infrastructure.
- Use collaborative development and rapid prototyping.
- Repeat the project manager’s mantra: There is scope, resources, and time. When you change one aspect, expect it to affect the others.

- Understand how quality and the desire for perfection can sabotage a project's timeline. Manage expectations about quality early on, and agree upon acceptable quality levels.
- Recognize the role of culture and the right people in adopting agile development techniques.

7.8 Check your progress

1. Define waterfall model
2. Define agile model
3. Define scrum.
4. What is kanban.
5. Write best practices for successful BI

Answers to Check your progress

1. The waterfall model is a breakdown of project activities into linear sequential phases, where each phase depends on the deliverables of the previous one and corresponds to a specialization of tasks.
2. Agile modeling is a methodology for modeling and documenting software systems based on best practices. It is a collection of values and principles, that can be applied on an software development project
3. Within project management, scrum, sometimes written Scrum or SCRUM, is a framework for developing, delivering, and sustaining products in a complex environment, with an initial emphasis on software development, although it has been used in other fields including research, sales, marketing and advanced technologies.
4. Kanban is a lean method to manage and improve work across human systems. This approach aims to manage work by balancing demands with available capacity, and by improving the handling of system-level bottlenecks

5. 1. Be prepared to change the business-facing parts of BI on a more rapid basis than the behind-the-scenes infrastructure.

   2. Use collaborative development and rapid prototyping.

   3. Repeat the project manager's mantra: There is scope, resources, and time. When you change one aspect, expect it to affect the others. 4.Understand how quality and the

desire for perfection can sabotage a project's timeline. Manage expectations about quality early on, and agree upon acceptable quality levels.

5. Recognize the role of culture and the right people in adopting agile development techniques.

7.9 Summary

The role of agile development in BI success is one of those secrets that emerged only from a study of common themes in the successful BI case studies. In the beginning few companies were using agile software development, and even fewer were using it in BI. Today agile for BI is more widely accepted, and, advocating it as a best practice, The Data Warehousing Institute (TDWI) now focuses a number of conferences on agile. Despite broader awareness of agile development, awareness of it is not required for newly certified project management professionals. Instead, certification in agile development techniques are more often provided separately by organizations who offer consulting and education on agile.

7.10 Keywords

- Business Intelligence - Business intelligence comprises the strategies and technologies used by enterprises for the data analysis and management of business information
- Agile development - an iterative approach to project management and software development
- Water fall model **-** The waterfall model is a breakdown of project activities into linear sequential phases, where each phase depends on the deliverables of the previous one and corresponds to a specialization of tasks.
- **Scrum -** is a framework for developing, delivering, and sustaining products in a complex environment, with an initial emphasis on software development, although it has been used in other fields including research, sales, marketing and advanced technologies.
- **Data mining** - indicates the process of exploration and analysis of a dataset, usually of large size, in order to find regular patterns, to extract relevant knowledge and to obtain meaningful recurring rules.

7.11 Questions for self-study

1. Explain waterfall project methodology
2. Write the principles mentioned in agile manifesto
3. Explain agile development techniques.
4. Describe basic concepts of scrum.
5. Explain Basic Concepts of Kanban

7.12 References

1. Sharda R, Delen D, Turban E. Business intelligence analytics and data science: A managerial perspective. Pearson, 2022.

2. Vercellis C. Business intelligence: data mining and optimization for decision making. John Wiley & Sons; 2011 Aug 10.

3. Bentley D. Business Intelligence and Analytics. Internet, link: https://www. pdfdrive. com/business-intelligence-and-analytics-e56416503. html. 2017.

UNIT 8 : TEXT, WEB, AND SOCIAL MEDIA ANALYTICS

Structure

8.0 OBJECTIVES

After studying this unit, you will be able to:

- Distinguish text analytics and text mining.
- Give significance of Natural Language Processing
- Brief out on web analytics.
- Elucidate sentiment analysis.
- Explain off-site and on-site web analytics.
- Discuss social network types.
- Write a note on social media analytics.

8.1 INTRODUCTION

This unit provides a comprehensive overview of text analytics/mining and Web analytics/mining along with their popular application areas such as search engines, sentiment analysis, and social network/media analytics. As we have been witnessing in the recent years, the unstructured data generated over the Internet of Things (Web, sensor networks, radio-frequency identification [RFID]-enabled supply chain systems, surveillance networks, etc.) is increasing at an exponential pace, and there is no indication of it slowing down. This changing nature of data is forcing organizations to make Text and Web analytics a critical part of their business intelligence/analytics infrastructure.

8.2 TEXT ANALYTICS AND TEXT MINING OVERVIEW

The information age that we are living in is characterized by the rapid growth in the amount of data and information collected, stored, and made available in electronic format. A vast majority of business data are stored in text documents that are virtually unstructured. Because knowledge is power in today's business world, and knowledge is derived from data and information, businesses that effectively and efficiently tap into their text data sources will have the necessary knowledge to make better decisions, leading to a competitive advantage over those businesses that lag behind. This is where the need for text analytics and text mining fits into the big picture of today's businesses.

Even though the overarching goal for both text analytics and text mining is to turn unstructured textual data into actionable information through the application of natural language processing (NLP) and analytics, their definitions are somewhat different, at least to some experts in the field. According to them, text analytics is a broader concept that includes information retrieval (e.g., searching and identifying relevant documents for a given set of key terms), as well as information extraction, data mining, and Web mining, whereas text mining is primarily focused on discovering new and useful knowledge from the textual data sources.

Figure 8.1 illustrates the relationships between text analytics and text mining along with other related application areas. The bottom of Figure 8.1 lists the main disciplines (the foundation of the house) that play a critical role in the development of these increasingly more popular application areas. Based on this definition of text analytics and text mining, one could simply formulate the difference between the two as follows:

Text Analytics = Information Retrieval + Information Extraction + Data Mining + Web Mining

or simply

Text Analytics = Information Retrieval + Text Mining

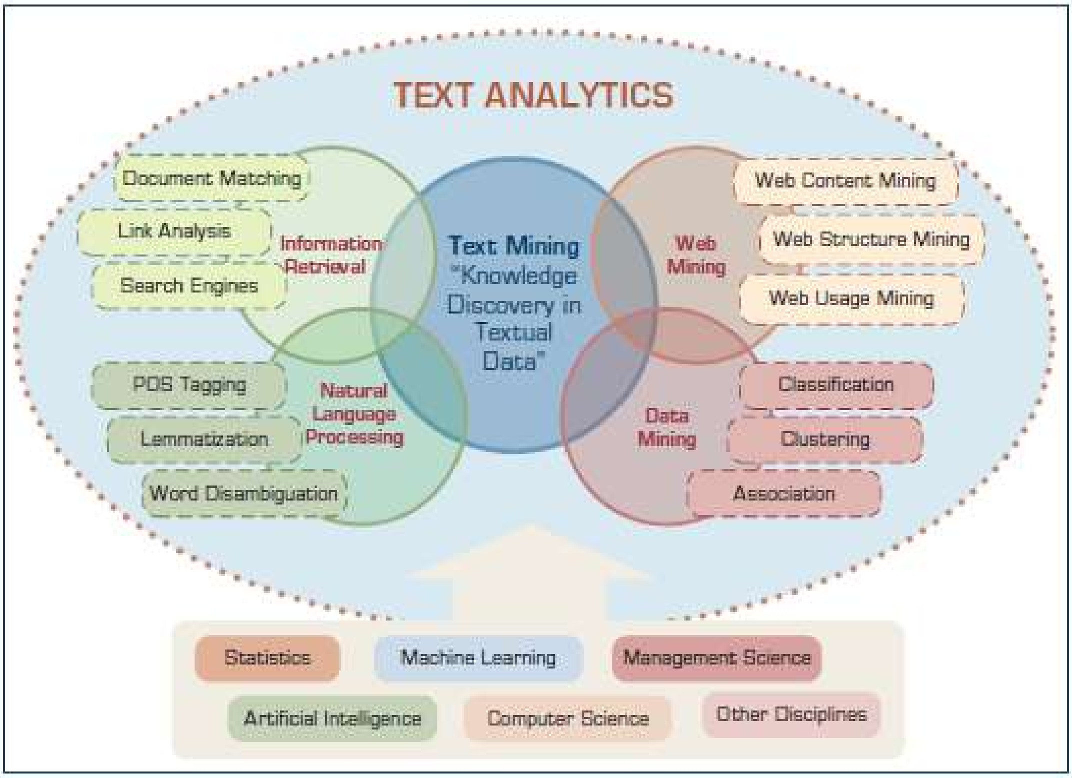


FIGURE 8.1 Text Analytics, Related Application Areas, and Enabling Disciplines.

Compared to text mining, text analytics is a relatively new term (e.g., consumer analytics, completive analytics, visual analytics, social analytics). Although the term text analytics is more commonly used in a business application context, text mining is frequently used in academic research areas.

Text mining (also known as text data mining or knowledge discovery in textual databases) is the semi-automated process of extracting patterns (useful information and knowledge) from large amounts of unstructured data sources. Remember that data mining is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases, where the data are organized in records structured by categorical, ordinal, or continuous variables.

Text mining is the same as data mining in that it has the same purpose and uses the same processes, but with text mining the input to the process is a collection of unstructured (or less structured) data files such as Word documents, PDF files, text excerpts, XML files, and so on. In essence, text mining can be thought of as a process (with two main steps) that starts with imposing structure on the text-based data sources followed by extracting relevant information and knowledge using data mining techniques and tools.

The benefits of text mining are obvious in the areas where very large amounts of textual data are being generated, such as law (court orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), and marketing (customer comments). For example, the free-form text-based interactions with customers in the form of complaints (or praises) and warranty claims can be used to objectively identify product and service characteristics that are deemed to be less than perfect and can be used as input to better product development and service allocations. Likewise, market outreach programs and focus groups generate large amounts of data. By not restricting product or service feedback to a codified form, customers can present, in their own words, what they think about a company's products and services. Another area where the automated processing of unstructured text has had a lot of impact is in electronic communications and e-mail.

The following list describes some commonly used text mining terms:

- **Unstructured data (versus structured data).** Structured data has a predetermined format. It is usually organized into records with simple data values (categorical, ordinal, and continuous variables) and stored in databases. In contrast, **unstructured data** does not have a predetermined format and is stored in the form of textual documents. In essence, the structured data is for the computers to process while the unstructured data is for humans to process and understand.
- **Corpus.** In linguistics, a **corpus** (plural corpora) is a large and structured set of texts (now usually stored and processed electronically) prepared for the purpose of conducting knowledge discovery.
- **Terms.** A term is a single word or multiword phrase extracted directly from the corpus of a specific domain by means of NLP methods.
- **Concepts.** Concepts are features generated from a collection of documents by means of manual, statistical, rule-based, or hybrid categorization methodology. Compared to terms, concepts are he result of higher level abstraction.
- **Stemming. Stemming** is the process of reducing inflected words to their stem (or base or root) form. For instance, stemmer, stemming, stemmed are all based on the root stem.
- **Stop words. Stop words** (or noise words) are words that are filtered out prior to or after processing of natural language data (i.e., text). Even though there is no

universally accepted list of stop words, most NLP tools use a list that includes articles (a, am, the, of, etc.), auxiliary verbs (is, are, was, were, etc.), and context-specific words that are deemed not to have differentiating value.

- **Synonyms and polysemes.** Synonyms are syntactically different words (i.e., spelled differently) with identical or at least similar meanings (e.g., movie, film, and motion picture). In contrast, **polysemes**, which are also called homonyms, are syntactically identical words (i.e., spelled exactly the same) with different meanings (e.g., bow can mean "to bend forward," "the front of the ship," "the weapon that shoots arrows," or "a kind of tied ribbon").
- **Tokenizing.** A token is a categorized block of text in a sentence. The block of text corresponding to the token is categorized according to the function it performs. This assignment of meaning to blocks of text is known as **tokenizing**. A token can look like anything; it just needs to be a useful part of the structured text.
- **Term dictionary.** A collection of terms specific to a narrow field that can be used to restrict the extracted terms within a corpus.
- **Word frequency.** The number of times a word is found in a specific document.
- **Part-of-speech tagging.** The process of marking up the words in a text as corresponding to a particular part of speech (such as nouns, verbs, adjectives, adverbs, etc.) based on a word's definition and the context in which it is used.
- **Morphology.** A branch of the field of linguistics and a part of NLP that studies the internal structure of words (patterns of word formation within a language or across languages).
- **Term-by-document matrix (occurrence matrix).** A common representation schema of the frequency-based relationship between the terms and documents in tabular format where terms are listed in columns, documents are listed in rows, and the frequency between the terms and documents is listed in cells as integer values.
- **Singular value decomposition (latent semantic indexing).** A dimensionality reduction method used to transform the term-by-document matrix to a manageable size by generating an intermediate representation of the frequencies using a matrix manipulation method similar to principal component analysis.

8.3 NATURAL LANGUAGE PROCESSING (NLP)

Some of the early text mining applications used a simplified representation called *bagof- words* when introducing structure to a collection of text-based documents to classify them into two or more predetermined classes or to cluster them into natural groupings. In the bag-of-words model, text, such as a sentence, paragraph, or complete document, is represented as a collection of words, disregarding the grammar or the order in which the words appear. The bag-of-words model is still used in some simple document classification tools.

For instance, in spam filtering an e-mail message can be modeled as an unordered collection of words (a bag-of-words) that is compared against two different predetermined bags. One bag is filled with words found in spam messages and the other is filled with words found in legitimate e-mails. Although some of the words are likely to be found in both bags, the "spam" bag will contain spam-related words such as *stock, buy etc.,* much more frequently than the legitimate bag, which will contain more words related to the user's friends or workplace. The level of match between a specific e-mail's bag-of-words and the two bags containing the descriptors determines the membership of the e-mail as either spam or legitimate.

Naturally, we (humans) do not use words without some order or structure. We use, words in sentences, which have semantic as well as syntactic structure. Thus, automated techniques (such as text mining) need to look for ways to go beyond the bag-of-words interpretation and incorporate more and more semantic structure into their operations. The current trend in text mining is toward including many of the advanced features that can be obtained using NLP.

Natural language processing (NLP) is an important component of text mining and is a subfield of artificial intelligence and computational linguistics. It studies the problem of "understanding" the natural human language, with the view of converting depictions of human language (such as textual documents) into more formal representations (in the form of numeric and symbolic data) that are easier for computer programs to manipulate. The goal of NLP is to move beyond syntax-driven text manipulation (which is often called "word counting") to a true understanding and processing of natural language that considers grammatical and semantic constraints as well as the context. The definition and scope of the word *understanding* is one of the major discussion topics in NLP. Considering that the natural human language is vague and that a true understanding of meaning requires extensive knowledge of a topic (beyond what is in the words, sentences, and paragraphs), will computers ever be able to understand natural language the same way and with the same accuracy that humans do? Probably not! NLP has come a long way from the days of simple word counting, but it has an even longer way to go to really understanding natural human language.

NLP has successfully been applied to a variety of domains for a wide range of tasks via computer programs to automatically process natural human language that previously could only be done by humans. Following are among the most popular of these tasks:

- **Question answering.** The task of automatically answering a question posed in natural language; that is, producing a human language answer when given a human language question. To find the answer to a question, the computer program may use either a restructured database or a collection of natural language documents (a text corpus such as the World Wide Web).
- **Automatic summarization.** The creation of a shortened version of a textual document by a computer program that contains the most important points of the original document.
- **Natural language generation.** Systems convert information from computer databases into readable human language.
- **Natural language understanding.** Systems convert samples of human language into more formal representations that are easier for computer programs to manipulate.
- **Machine translation.** The automatic translation of one human language to another.
- **Foreign language reading.** A computer program that assists a nonnative language speaker to read a foreign language with correct pronunciation and accents on different parts of the words.
- **Foreign language writing.** A computer program that assists a nonnative language user in writing in a foreign language.
- **Speech recognition.** Converts spoken words to machine-readable input. Given a sound clip of a person speaking, the system produces a text dictation.
- **Text-to-speech.** Also called *speech synthesis*, a computer program automatically converts normal language text into human speech.

- **Text proofing.** A computer program reads a proof copy of a text to detect and correct any errors.
- **Optical character recognition.** The automatic translation of images of handwritten, typewritten, or printed text (usually captured by a scanner) into machine editable textual documents.

The success and popularity of text mining depends greatly on advancements in NLP in both generation as well as understanding of human languages. NLP enables the extraction of features from unstructured text so that a wide variety of data mining techniques can be used to extract knowledge (novel and useful patterns and relationships) from it. In that sense, simply put, text mining is a combination of NLP and data mining.

8.3 SENTIMENT ANALYSIS

We humans are social beings. We are adept at utilizing a variety of means to communicate. We often consult financial discussion forums before making an investment decision; ask our friends for their opinions on a newly opened restaurant or a newly released movie; and conduct Internet searches and read consumer reviews and expert reports before making a big purchase like a house, a car, or an appliance. We rely on others' opinions to make better decisions, especially in an area where we don't have a lot of knowledge or experience.

Thanks to the growing availability and popularity of opinion-rich Internet resources such as social media outlets (e.g., Twitter, Facebook), online review sites, and personal blogs, it is now easier than ever to find opinions of others (thousands of them, as a matter of fact) on everything from the latest gadgets to political and public figures. Even though not everybody expresses opinions over the Internet—due mostly to the fast-growing number and capabilities of social communication channels—the numbers are increasing exponentially.

As a field of research, sentiment analysis is closely related to computational linguistics, NLP, and text mining. Sentiment analysis has many names. It's often referred to as opinion mining, subjectivity analysis, and appraisal extraction, with some connections to affective computing (computer recognition and expression of emotion). The sudden upsurge of interest and activity in the area of sentiment analysis (i.e., opinion mining), which deals with the automatic extraction of opinions, feelings, and subjectivity in text, is creating opportunities and threats for businesses and individuals alike. The ones who embrace and take advantage of it will greatly benefit from it. Every opinion put on the Internet by an individual or a company will be accredited to the originator (good or bad) and will be retrieved and mined by others (often automatically by computer programs).

Sentiment Analysis Applications

Compared to traditional sentiment analysis methods, which were survey based or focus group centered, costly, and time consuming (and therefore driven from a small sample of participants), the new face of text analytics–based sentiment analysis is a limit breaker. Current solutions automate very large-scale data collection, filtering, classification, and clustering methods via NLP and data mining technologies that handle both factual and subjective information. Sentiment analysis is perhaps the most popular application of text analytics, tapping into data sources like tweets, Facebook posts, online communities, discussion boards, Web logs, product reviews, call center logs and recordings, product rating sites, chat rooms, price comparison portals, search engine logs, and newsgroups.

8.5 WEB USAGE MINING (WEB ANALYTICS)

Web usage mining (also called Web analytics) is the extraction of useful information from data generated through Web page visits and transactions. Analysis of the information collected by Web servers can help us better understand user behavior. Analysis of this data is often called clickstream analysis. By using the data and text mining techniques, a company might be able to discern interesting patterns from the clickstreams. For example, it might learn that 60% of visitors who searched for “hotels in Maui” had searched earlier for “airfares to Maui.” Such information could be useful in determining where to place online advertisements. Clickstream analysis might also be useful for knowing when visitors access a site.

For example, if a company knew that 70% of software downloads from its Web site occurred between 7 and 11 p.m., it could plan for better customer support and network bandwidth during those hours. Figure 8.2 shows the process of extracting knowledge from clickstream data and how the generated knowledge is used to improve the process, improve the Web site, and most important, increase the customer value.

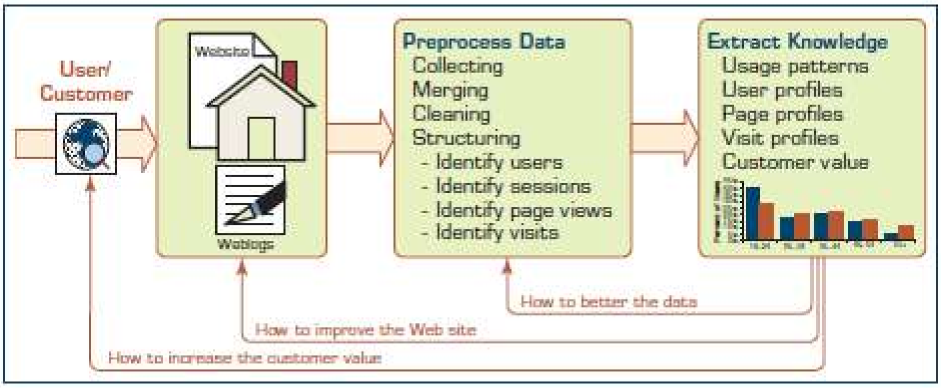


FIGURE 8.2 Extraction of Knowledge from Web Usage Data.

Web Analytics Technologies

There are numerous tools and technologies for Web analytics in the marketplace. Because of their power to measure, collect, and analyze Internet data to better understand and optimize Web usage, the popularity of Web analytics tools is increasing. Web analytics holds the promise of revolutionizing how business is done on the Web. Web analytics is not just a tool for measuring Web traffic; it can also be used as a tool for e-business and market research and to assess and improve the effectiveness of e-commerce Web sites. Web analytics applications can also help companies measure the results of traditional print or broadcast advertising campaigns. It can help estimate how traffic to a Web site changes after the launch of a new advertising campaign. Web analytics provides information about the number of visitors to a Web site and the number of page views. It helps gauge traffic and popularity trends, which can be used for market research.

There are two main categories of Web analytics: off-site and on-site. Off-site Web analytics refers to Web measurement and analysis about you and your products that takes place outside your Web site. It includes the measurement of a Web site's potential audience (prospect or opportunity), share of voice (visibility or word-of-mouth), and buzz (comments or opinions) that is happening on the Internet.

What is more mainstream has been on-site Web analytics. Historically, Web analytics has referred to on-site visitor measurement. However, in recent years this has blurred, mainly because vendors are producing tools that span both categories. On-site Web analytics measure visitors' behavior once they are on your Web site. This includes its drivers and conversions—for example, the degree to which different landing pages are associated with online purchases. On-site Web analytics measure the performance of your Web site in a commercial context. The data collected on the Web site is then compared against key.

For on-site Web analytics, there are two technical ways of collecting the data. The first and more traditional method is the server log file analysis, where the Web server records file requests made by browsers. The second method is page tagging, which uses JavaScript embedded in the site page code to make image requests to a third-party analytics-dedicated server whenever a page is rendered by a Web browser (or when a mouse click occurs). Both collect data that can be processed to produce Web traffic reports.

In addition to these two main streams, other data sources may also be added to augment Web site behavior data. These other sources may include e-mail, direct mail campaign data, sales and lead history, or social media–originated data.

8.5 SOCIAL ANALYTICS

Social analytics include mining the textual content created in social media (e.g., sentiment analysis, NLP) and analyzing socially established networks (e.g., influencer identification, profiling, prediction) for the purpose of gaining insight about existing and potential customers' current and future behaviors, and about the likes and dislikes toward a firm's products and services. Based on this definition and the current practices, social analytics can be classified into two different branches:

i. Social Network Analysis (SNA)
ii. Social Media Analytics (SMA)

i) Social Network Analysis

A social network is a social structure composed of individuals/people (or groups of individuals or organizations) linked to one another with some type of connections/relationships. The social network perspective provides a holistic approach to analyzing the structure and dynamics of social entities. The study of these structures uses SNA to identify local and global patterns, locate influential entities, and examine network dynamics. Social networks and the analysis of them is essentially an interdisciplinary field that emerged from social psychology, sociology, statistics, and graph theory.

A social network is a theoretical construct useful in the social sciences to study relationships between individuals, groups, organizations, or even entire societies (social units). The term is used to describe a social structure determined by such interactions. The ties through which any given social unit connects represent the convergence of the various social contacts of that unit. In general, social networks are self-organizing, emergent, and complex, such that a globally coherent pattern appears from the local interaction of the elements (individuals and groups of individuals) that make up the system.

Following are a few typical social network types that are relevant to business activities:

Communication Networks :

Communication studies are often considered a part of both the social sciences and the humanities, drawing heavily on fields such as sociology, psychology, anthropology, information science, biology, political science, and economics. Many communications concepts describe the transfer of information from one source to another and thus can be represented as a social network. Telecommunication companies are tapping into this rich information source to optimize their business practices and to improve customer relationships.

Community Networks:

Traditionally, community referred to a specific geographic location, and studies of community ties had to do with who talked, associated, traded, and attended social activities with whom. Today, however, there are extended "online" communities developed through social networking tools and telecommunications devices. Such tools and devices continuously generate large amounts of data, which can be used by companies to discover invaluable, actionable information.

Criminal Networks

In criminology and urban sociology, much attention has been paid to the social networks among criminal actors. For example, studying gang murders and other illegal activities as a series of exchanges between gangs can lead to better understanding and prevention of such criminal activities. Now that we live in a highly connected world (thanks to the Internet), much of the criminal networks' formations and their activities are being watched/pursued by security agencies using state-of-the-art Internet tools and tactics. Even though the Internet has changed the landscape for criminal networks and law enforcement agencies, the traditional social and philosophical theories still apply to a large extent.

Innovation Networks

Business studies on diffusion of ideas and innovations in a network environment focus on the spread and use of ideas among the members of the social network. The idea is to understand why some networks are more innovative, and why some communities are early adopters of ideas and innovations (i.e., examining the impact of social network structure on influencing the spread of an innovation and innovative behavior).

ii) Social Media Analytics

Social media refers to the enabling technologies of social interactions among people in which they create, share, and exchange information, ideas, and opinions in virtual communities and networks. Social media depends on mobile and other Web-based technologies to create highly interactive platforms for individuals and communities to share, co-create, discuss, and modify user-generated content. It introduces substantial changes to communication among organizations, communities, and individuals.

Social media analytics refers to the systematic and scientific ways to consume the vast amount of content created by Web-based social media outlets, tools, and techniques for the betterment of an organization's competitiveness. Social media analytics is rapidly becoming a new force in organizations around the world, allowing them to reach out to and understand consumers as never before. In many companies, it is becoming the tool for integrated marketing and communications strategies.

The exponential growth of social media outlets, from blogs, Facebook, and Twitter to LinkedIn and YouTube, and analytics tools that tap into these rich data sources offer organizations the chance to join a conversation with millions of customers around the globe every day.

8.6 CHECK YOUR PROGRESS

1. Acronym for RFID
2. What is the goal of NLP?
3. Distinguish text analytics and text mining.
4. What is structured and unstructured data?

5. What is social network?

Answers to Check your progress

1. Radio-Frequency Identification Device
2. The goal of NLP is to move beyond syntax-driven text manipulation (which is often called "word counting") to a true understanding and processing of natural language that considers grammatical and semantic constraints as well as the context.
3. Text analytics is a broader concept that includes information retrieval (e.g., searching and identifying relevant documents for a given set of key terms), as well as information extraction, data mining, and Web mining, whereas text mining is primarily focused on discovering new and useful knowledge from the textual data sources.
4. Structured data has a predetermined format. It is usually organized into records with simple data values (categorical, ordinal, and continuous variables) and stored in databases. In contrast, unstructured data does not have a predetermined format and is stored in the form of textual documents.
5. A social network is a theoretical construct useful in the social sciences to study relationships between individuals, groups, organizations, or even entire societies (social units).

8.7 SUMMARY

This unit highlights on text analytics and text mining. The application areas are detailed related with the mining process. One of the best usage of text mining is NLP. The significance of NLP is detailed with various needs. Web usage statistics has been analyzed and web analytics has been detailed. Various types of web analytics has been analyzed. Sentiment analysis has been studied in detail with various application. Social network types has been detailed considering various context.

8.8 KEYWORDS

- **Text Mining**: Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights

- **Text Analytics**: Text analytics is the process of extracting meaning out of text. For example, this can be analyzing text written by customers in a customer survey, with the focus on finding common themes and trends.
- **Social Media Analytics**: Social media analytics is the ability to gather and find meaning in data gathered from social channels to support business decisions — and measure the performance of actions based on those decisions through social media.
- **Web Analytics**: Web analytics is the process of analyzing the behavior of visitors to a website. This involves tracking, reviewing and reporting data to measure web activity, including the use of a website and its components, such as webpages, images and videos.
- **Social Analytics**: Social analytics is monitoring, analyzing, measuring and interpreting digital interactions and relationships of people, topics, ideas and content.

8.9 QUESTIONS FOR SELF STUDY

1. Write a note on text analytics and text mining.
2. What is NLP? Give its significance.
3. Write a short note on web analytics.
4. Write a short note on sentiment analysis.
5. Explain off-site and on-site web analytics.
6. Discuss social network types.
7. Write a note on social media analytics.

8.10 REFERENCES

1. Sharda R, Delen D, Turban E. Business intelligence analytics and data science: A managerial perspective. Pearson, 2022.
2. Vercellis C. Business intelligence: data mining and optimization for decision making. John Wiley & Sons; 2011 Aug 10.
3. Bentley D. Business Intelligence and Analytics. Internet, link: https://www. pdfdrive. com/business-intelligence-and-analytics-e56416503. html. 2017.

Karnataka State Open University

Mukthagangothri, Mysore – 570 006.

Dept. of Studies and Research in Management

MBA IT Specialization

III Semester

Business Intelligence and Analytics

Block 3

Karnataka State Open University

Mukthagangothri, Mysore – 570 006.
Dept. of Studies and Research in Management

MBA. IT Specialization

III Semester

Business Intelligence and Analytics

BLOCK 3

BLOCK 3 INTRODUCTION

BI can be used to support a wide range of business decisions ranging from operational to strategic. Basic operating decisions include product positioning or pricing. Strategic business decisions include priorities, goals and directions at the broadest level. In all cases, BI is most effective when it combines data derived from the market in which a company operates (external data) with data from company sources internal to the business such as financial and operations data (internal data). When combined, external and internal data can provide a more complete picture which, in effect, creates an "intelligence" that cannot be derived by any singular set of data. Amongst myriad uses, BI tools empower organisations to gain insight into new markets, assess demand and suitability of products and services for different market segments and gauge the impact of marketing efforts.

This block consists of 4 units and is organized as follows:

Unit 9: Categories of Business Reporting, Data Visualization, Different Types of Charts and Graphs, Specialized Charts and Graphs, The Emergence of Visual Analytics, Information Dashboards

Unit 10: Internet of Things, Cloud Computing and Business Analytics, Location-Based Analytics for Organizations, Issues of Legality, Privacy and Ethics

Unit 11: Relational Marketing, Motivation and Objectives.

Unit 12: History, Mobile Client Application, Web Applications vs. Device-specific, Applications for Mobile BI, Security Considerations for Mobile BI Apps, Real-Time Business Intelligence

Unit 9 : BUSINESS REPORTING

Structure

9.0 OBJECTIVES

After studying this unit, you will be able to:

- Analyze categories of Business Reporting
- Give significance of Data Visualization
- Discuss different types of Charts and Graphs
- Describe specialized Charts and Graphs
- Explore about the Emergence of Visual Analytics
- Signify Information Dashboards

9.1 INTRODUCTION

Decision makers are in need of information to make accurate and timely decisions. Information is essentially the contextualization of data. In addition to statistical means that were explained in the previous section, information (descriptive analytics) can also be obtained using online analytics processing [OLTP] systems. The information is usually provided to the decision makers in the form of a written report (digital or on paper), although it can also be provided orally. Simply put, a report is any communication artifact prepared with the specific intention of conveying information in a digestible form to whoever needs it, whenever and wherever they may need it. It is usually a document that contains information (usually driven from data) organized in a narrative, graphic, and/or tabular form, prepared periodically (recurring) or on an as-needed (ad hoc) basis, referring to specific time periods, events, occurrences, or subjects. Business reports can fulfill many different (but often related) functions. Here are a few of the most prevailing ones:

- To ensure that all departments are functioning properly
- To provide information
- To provide the results of an analysis
- To persuade others to act
- To create an organizational memory (as part of a knowledge management system)

9.2 CATEGORIES OF BUSINESS REPORTING

Business reporting (also called OLAP or BI) is an essential part of the larger drive toward improved, evidence-based, optimal managerial decision making. The foundation of these business reports is various sources of data coming from both inside and outside the organization (online transaction processing [OLTP] systems). Creation of these reports involves ETL (extract, transform, and load) procedures in coordination with a data warehouse and then using one or more reporting tools

Due to the rapid expansion of information technology coupled with the need for improved competitiveness in business, there has been an increase in the use of computing power to produce unified reports that join different views of the enterprise in one place. Usually, this reporting process involves querying structured data sources, most of which were created using different logical data models and data dictionaries, to produce a human-readable, easily digestible report. These types of business reports allow managers and coworkers to stay informed and involved, review options and alternatives, and make informed decisions. Figure 9.1 shows the continuous cycle of data acquisition, information generation, decision making, business process management. Perhaps the most critical task in this cyclical process is the reporting (i.e., information generation)— converting data from different sources into actionable information.

Key to any successful report are clarity, brevity, completeness, and correctness. The nature of the report and the level of importance of these success factors change significantly based on for whom the report is created. Most of the research in effective reporting is dedicated to internal reports that inform stakeholders and decision makers within the organization. There are also external reports between businesses and the government (e.g., for tax purposes or for regular filings to the Securities and Exchange Commission). Even though there are a wide variety of business reports, the ones that are often used for managerial purposes can be grouped into three major categories (Hill, 2016).

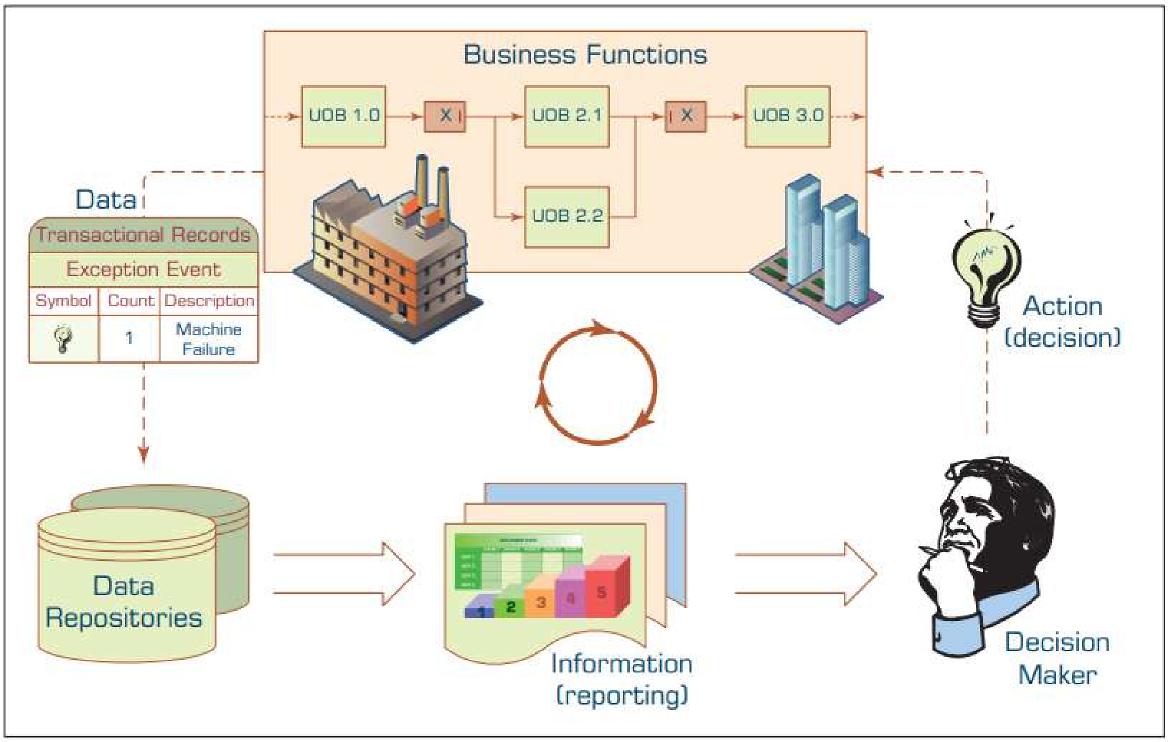


FIGURE 9.1 The Role of Information Reporting in Managerial Decision Making.

METRIC MANAGEMENT REPORTS

In many organizations, business performance is managed through outcome-oriented metrics. For external groups, these are service-level agreements. For internal management, they are key performance indicators (KPIs). Typically, there are enterprise-wide agreed targets to be tracked against over a period of time. They may be used as part of other management strategies such as Six Sigma or Total Quality Management.

DASHBOARD-TYPE REPORTS

A popular idea in business reporting in recent years has been to present a range of different performance indicators on one page, like a dashboard in a car. Typically, dashboard vendors would provide a set of predefined reports with static elements and fixed structure, but also allow for customization of the dashboard widgets, views, and set targets for various metrics. It’s common to have color-coded traffic lights defined for performance (red, orange, green) to draw management’s attention to particular areas. A more detailed description of dashboards can be found in later part of this chapter.

BALANCED SCORECARD–TYPE REPORTS

This is a method developed by Kaplan and Norton that attempts to present an integrated view of success in an organization. In addition to financial performance, balanced scorecard–type reports also include customer, business process, and learning and growth perspectives.

9.3 DATA VISUALIZATION

Data visualization (or more appropriately, information visualization) has been defined as “the use of visual representations to explore, make sense of, and communicate data” (Few, 2007). Although the name that is commonly used is data visualization, usually what is meant by this is information visualization. Because information is the aggregation, summarization, and contextualization of data (raw facts), what is portrayed in visualizations is the information and not the data. However, because the two terms data visualization and information visualization are used interchangeably and synonymously, in this chapter we will follow suit.

Data visualization is closely related to the fields of information graphics, information visualization, scientific visualization, and statistical graphics. Until recently, the major forms of data visualization available in both BI applications have included charts and graphs, as well as the other types of visual elements used to create scorecards and dashboards. To better understand the current and future trends in the field of data visualization, it helps to begin with some historical context.

9.4 DIFFERENT TYPES OF CHARTS AND GRAPHS

Often end users of business analytics systems are not sure what type of chart or graph to use for a specific purpose. Some charts or graphs are better at answering certain types of questions. Some look better than others. Some are simple; some are rather complex and crowded. What follows is a short description of the types of charts and/or graphs commonly found in most business analytics tools and what types of questions they are better at answering/analyzing.

Basic Charts and Graphs

What follows are the basic charts and graphs that are commonly used for information visualization.

LINE CHART

Line charts are perhaps the most frequently used graphical visuals for time series data. Line charts (or a line graphs) show the relationship between two variables; they are most often used to track changes or trends over time (having one of the variables set to time on the x-axis). Line charts sequentially connect individual data points to help infer changing trends over a period of time. Line charts are often used to show time-dependent changes in the values of some measure, such as changes on a specific stock price over a 5-year period or changes on the number of daily customer service calls over a month.

BAR CHART

Bar charts are among the most basic visuals used for data representation. Bar charts are effective when you have nominal data or numerical data that splits nicely into different categories so you can quickly see comparative results and trends within your data. Bar charts are often used to compare data across multiple categories such as percent of advertising spending by departments or by product categories. Bar charts can be vertically or horizontally oriented. They can also be stacked on top of each other to show multiple dimensions in a single chart.

PIE CHART

Pie charts are visually appealing, as the name implies, pie-looking charts. Because they are so visually attractive, they are often incorrectly used. Pie charts should only be used to illustrate relative proportions of a specific measure. For instance, they can be used to show the relative percentage of an advertising budget spent on different product lines, or they can show relative proportions of majors declared by college students in their sophomore year. If the number of categories to show is more than just a few (say more than four), one should seriously consider using a bar chart instead of a pie chart.

SCATTER PLOT

Scatter plots are often used to explore the relationship between two or three variables (in 2-D or 2-D visuals). Because they are visual exploration tools, having more than three variables, translating them into more than three dimensions is not easily achievable. Scatter plots are an effective way to explore the existence of trends, concentrations, and outliers. For instance, in a two-variable (two-axis) graph, a scatter plot can be used to illustrate the corelationship between age and weight of heart disease patients or it can illustrate the relationship between the number of customer care representatives and the number of open customer service claims. Often, a trend line is superimposed on a two-dimensional scatter plot to illustrate the nature of the relationship.

BUBBLE CHART

Bubble charts are often enhanced versions of scatter plots. Bubble charts, though, are not a new visualization type; instead, they should be viewed as a technique to enrich data illustrated in scatter plots (or even geographic maps). By varying the size and/or color of the circles, one can add additional data dimensions, offering more enriched meaning about the data. For instance, a bubble chart can be used to show a competitive view of college-level class attendance by major and by time of the day, or it can be used to show profit margin by product type and by geographic region.

9.5 SPECIALIZED CHARTS AND GRAPHS

The graphs and charts that we review in this section are either derived from the basic charts as special cases or they are relatively new and are specific to a problem type and/ or an application area.

HISTOGRAM

Graphically speaking, a histogram looks just like a bar chart. The difference between histograms and generic bar charts is the information that is portrayed. Histograms are used to show the frequency distribution of a variable or several variables. In a histogram, the x-axis is often used to show the categories or ranges, and the y-axis is used to show the measures/values/frequencies. Histograms show the distributional shape of the data. That way, one can visually examine if the data is normally or exponentially distributed. For instance, one can use a histogram to illustrate the exam performance of a class, where distribution of the grades as well as comparative analysis of individual results can be shown, or one can use a histogram to show age distribution of the customer base.

GANTT CHART

Gantt charts are a special case of horizontal bar charts that are used to portray project timelines, project tasks/activity durations, and overlap among the tasks/ activities. By showing start and end dates/times of tasks/activities and the overlapping relationships, Gantt charts provide an invaluable aid for management and control of projects. For instance, Gantt charts are often used to show project timelines, task overlaps, relative task completions (a partial bar illustrating the completion percentage inside a bar that shows the actual task duration), resources assigned to each task, milestones, and deliverables.

PERT CHART

PERT charts (also called network diagrams) are developed primarily to simplify the planning and scheduling of large and complex projects. They show precedence relationships among the project activities/tasks. A PERT chart is composed of nodes (represented as circles or rectangles) and edges (represented with directed arrows). Based on the selected PERT chart convention, either nodes or the edges may be used to represent the project activities/tasks (activity-on-node versus activity-on-arrow representation schema).

GEOGRAPHIC MAP

When the data set includes any kind of location data (e.g., physical addresses, postal codes, state names or abbreviations, country names, latitude/longitude, or some type of custom geographic encoding), it is better and more informative to see the data on a map. Maps usually are used in conjunction with other charts and graphs, as opposed to by themselves. For instance, one can use maps to show distribution of customer service requests by product type (depicted in pie charts) by geographic locations. Often a large variety of information (e.g., age distribution, income distribution, education, economic growth, or population changes) can be portrayed in a geographic map to help decide where to open a new restaurant or a new service station. These types of systems are often called geographic information systems (GIS).

BULLET

Bullet graphs are often used to show progress toward a goal. A bullet graph is essentially a variation of a bar chart. Often they are used in place of gauges, meters, and thermometers in a dashboard to more intuitively convey the meaning within a much smaller space. Bullet graphs compare a primary measure (e.g., year-to-date revenue) to one or more other measures (e.g., annual revenue target) and present this in the context of defined performance metrics (e.g., sales quotas). A bullet graph can intuitively illustrate how the primary measure is performing against overall goals (e.g., how close a sales representative is to achieving his/her annual quota).

HEAT MAP

Heat maps are great visuals to illustrate the comparison of continuous values across two categories using color. The goal is to help the user quickly see where the intersection of the categories is strongest and weakest in terms of numerical values of the measure being analyzed. For instance, one can use heat maps to show segmentation analysis of target markets where the measure (color gradient would be the purchase amount) and the dimensions would be age and income distribution.

HIGHLIGHT TABLE

Highlight tables are intended to take heat maps one step further. In addition to showing how data intersects by using color, highlight tables add a number on top to provide additional detail. That is, they are two-dimensional tables with cells populated with numerical values and gradients of colors. For instance, one can show sales representatives' performance by product type and by sales volume.

TREE MAP

Tree maps display hierarchical (tree-structured) data as a set of nested rectangles. Each branch of the tree is given a rectangle, which is then tiled with smaller rectangles representing sub branches. A leaf node's rectangle has an area proportional to a specified dimension on the data. Often the leaf nodes are colored to show a separate dimension of the data. When the color and size dimensions are correlated in some way with the tree structure, one can often easily see patterns that would be difficult to spot in other ways, such as if a certain color is particularly relevant. A second advantage of tree maps is that, by construction, they make efficient use of space. As a result, they can legibly display thousands of items on the screen simultaneously

Which Chart or Graph Should You Use?

Which chart or graph that we explained in the previous section is the best? The answer is rather easy: there is not one best chart or graph, because if there was we would not have these many chart and graph types. They all have somewhat different data representation "skills." Therefore, the right question should be, "Which chart or graph is the best for a given task?" The capabilities of the charts given in the previous section can help in selecting and using the right chart/graph for a specific task, but it still is not easy to sort out. Several different chart/graph types can be used for the same visualization task. One rule of thumb is to select and use the simplest one from the alternatives to make it easy for the intended audience to understand and digest.

Although there is not a widely accepted, all-encompassing chart selection algorithm or chart/graph taxonomy, Figure 9.2 presents a rather comprehensive and highly logical organization of chart/graph types in a taxonomy-like structure (the original version was published in Abela 2008). The taxonomic structure is organized around the questions of "What would you like to show in your chart or graph?"

That is, what the purpose of the chart or graph will be. At that level, the taxonomy divides the purpose into four different types—relationship, comparison, distribution, and composition—and further divides the branches into subcategories based on the number of variables involved and time dependency of the visualization.

Even though these charts and graphs cover a major part of what is commonly used in information visualization, they by no means cover it all. Nowadays, one can find many other specialized graphs and charts that serve a specific purpose. Furthermore, the current trend is to combine / hybridize and animate these charts for better-looking and more intuitive visualization of todays complex and volatile data sources.

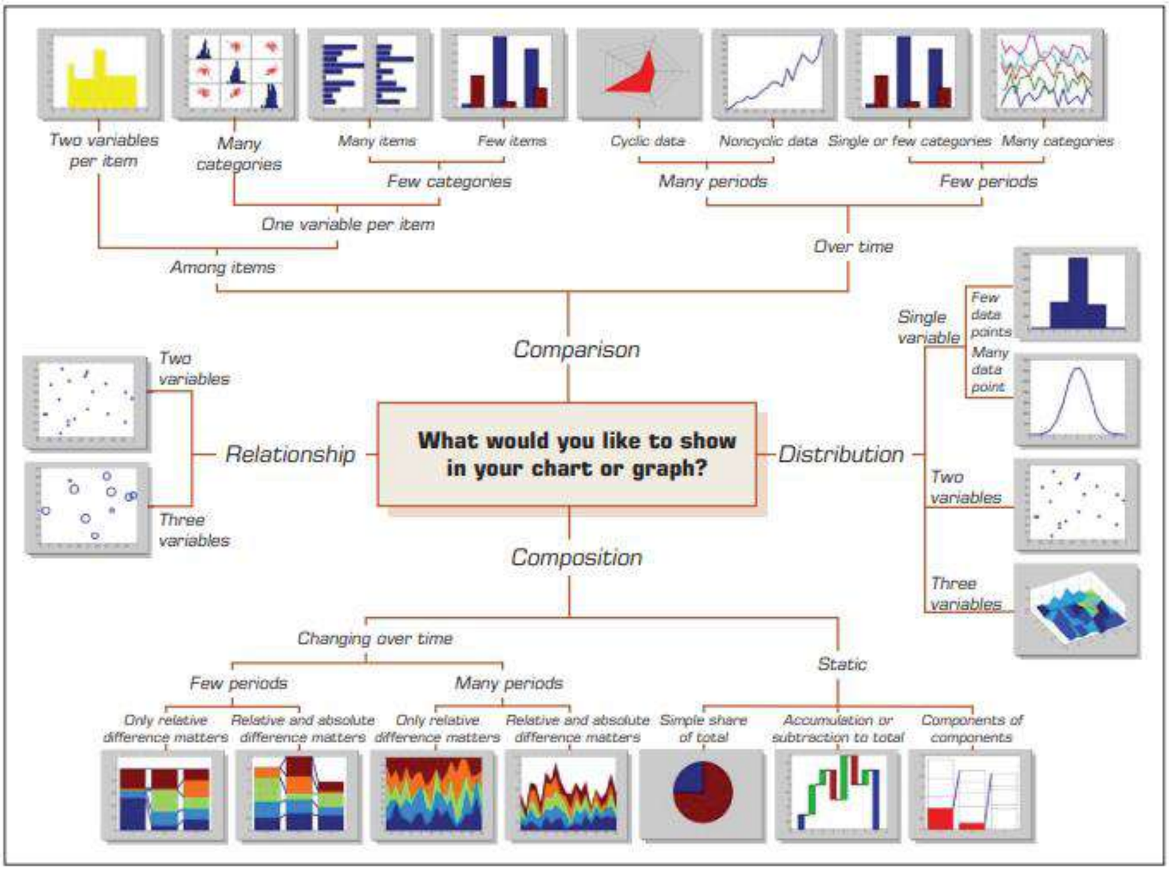


Figure 9.2 A Taxonomy of Charts and Graphs.

9.6 The Emergence of Visual Analytics

In BI and analytics, the key challenges for visualization have revolved around the intuitive representation of large, complex data sets with multiple dimensions and measures. For the most part, the typical charts, graphs, and other visual elements used in these applications usually involve two dimensions, sometimes three, and fairly small subsets of data sets. In contrast, the data in these systems reside in a data warehouse. At a minimum, these warehouses involve a range of dimensions (e.g., product, location, organizational structure, time), a range of measures, and millions of cells of data. In an effort to address these challenges, a number of researchers have developed a variety of new visualization techniques.

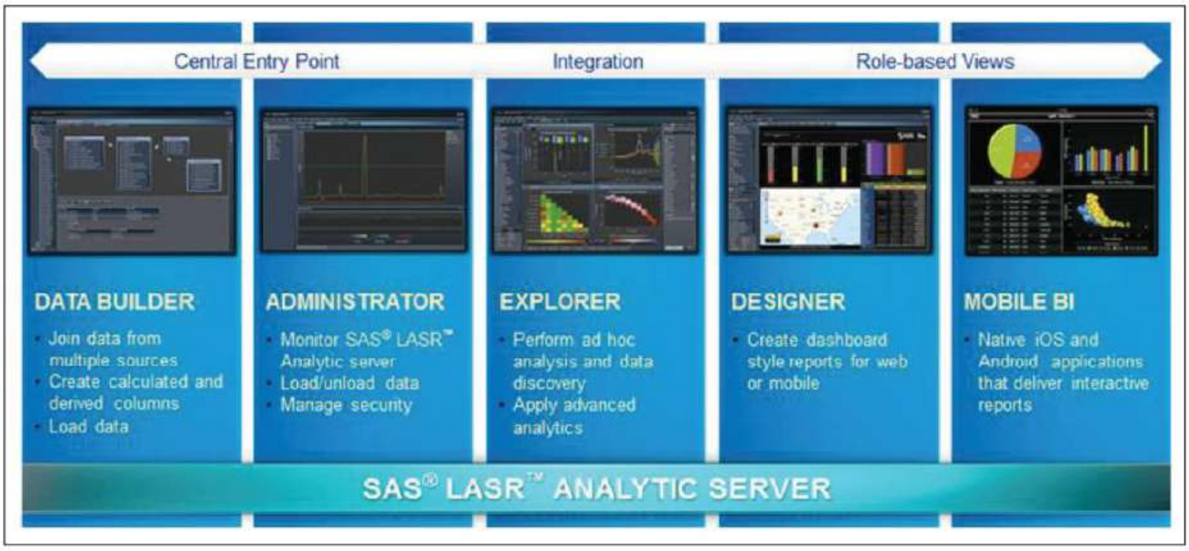
Visual Analytics

Visual analytics is a recently coined term that is often used loosely to mean nothing more than information visualization. What is meant by visual analytics is the combination of visualization and predictive analytics. Whereas information visualization is aimed at answering, “What happened?” and “What is happening?” and is closely associated with BI (routine reports, scorecards, and dashboards), visual analytics is aimed at answering, “Why is it happening?” “What is more likely to happen?” and is usually associated with business analytics (forecasting, segmentation, correlation analysis). Many of the information visualization vendors are adding the capabilities to call themselves visual analytics solution providers. One of the top, long-time analytics solution providers, SAS Institute, is approaching it from another direction. They are embedding their analytics capabilities into a high-performance data visualization environment that they call visual analytics.

Visual or not visual, automated or manual, online or paper based, business reporting is not much different than telling a story.

High-Powered Visual Analytics Environments

Due to the increasing demand for visual analytics coupled with fast-growing data volumes, there is an exponential movement toward investing in highly efficient visualization systems. With their latest move into visual analytics, the statistical software giant SAS Institute is now among those who are leading this wave. Their new product, SAS Visual Analytics, is a very high-performance computing, in-memory solution for exploring massive amounts of data in a very short time (almost instantaneously). It empowers users to spot patterns, identify opportunities for further analysis, and convey visual results via Web reports or a mobile platform such as tablets and smartphones. Figure 9.3 shows the high-level architecture of the SAS Visual Analytics platform. On one end of the architecture, there is a universal data builder and administrator capabilities, leading into explorer, report designer, and mobile BI modules, collectively providing an end-to-end visual analytics solution.



Central Entry Point

Integration

Role-based Views

DATA BUILDER

- Join data from multiple sources
- Create calculated and derived columns
- Load data

ADMINISTRATOR

- Monitor SAS® LASR™ Analytic server
- Load/unload data
- Manage security

EXPLORER

- Perform ad hoc analysis and data discovery
- Apply advanced analytics

DESIGNER

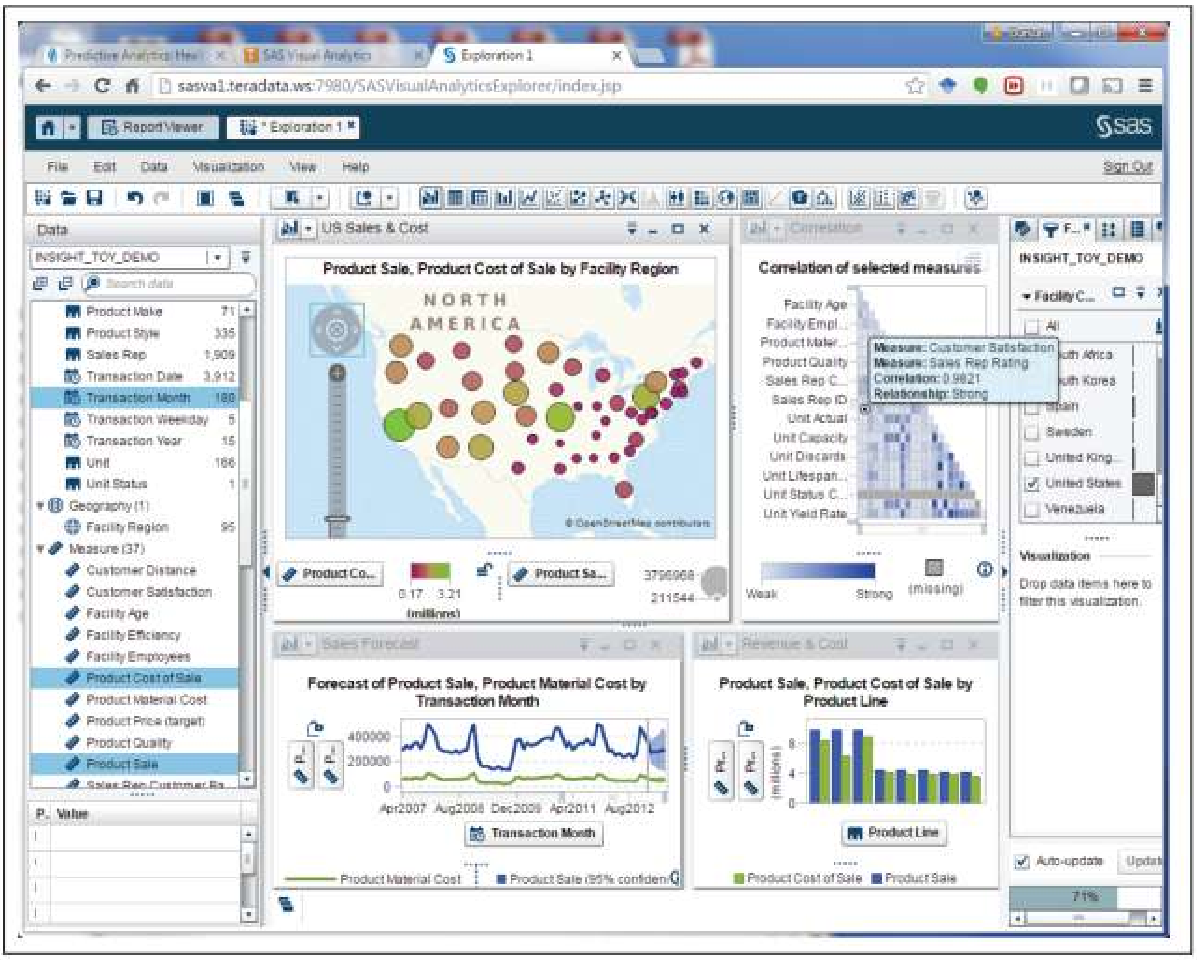
- Create dashboard style reports for web or mobile

MOBILE BI

- Native iOS and Android applications that deliver interactive reports

SAS® LASR™ ANALYTIC SERVER

FIGURE 9.3 An Overview of SAS Visual Analytics Architecture.



sasva1.teradata.ws:7980/SASVisualAnalyticsExplorer/index.jsp
Report Viewer
* Exploration 1
File
Edit
Data
Visualization
View
Help
Sign Out
Data
INSIGHT_TOY_DEMO
US Sales & Cost
Product Sale, Product Cost of Sale by Facility Region
NORTH AMERICA
Correlation
Correlation of selected measures
Measure: Customer Satisfaction
Measure: Sales Rep Rating
Correlation: 0.9821
Relationship: Strong
Sales Forecast
Forecast of Product Sale, Product Material Cost by Transaction Month
Revenue & Cost
Product Sale, Product Cost of Sale by Product Line
Transaction Month
Product Line
Product Material Cost
Product Sale (95% confiden
Product Cost of Sale
Product Sale
Visualization
Drop data items here to filter this visualization.
Auto-update

Figure 9.4 A Screenshot from SAS Visual Analytics. Source: SAS.com

Figure 9.4 shows a screenshot of an SAS Analytics platform where time series forecasting and confidence interval around the forecast are depicted.

9.7 Information Dashboards

Information dashboards are common components of most, if not all, BI or business analytics platforms, business performance management systems, and performance measurement software suites. Dashboards provide visual displays of important information that is consolidated and arranged on a single screen so that information can be digested at a single glance and easily drilled in and further explored. A typical dashboard is shown in Figure 9.5. This particular executive dashboard displays a variety of KPIs for a hypothetical software company called Sonatica (selling audio tools). This executive dashboard shows a high-level view of the different functional groups surrounding the products, starting from a general overview to the marketing efforts, sales, finance, and support departments. All of this is intended to give executive decision makers a quick and accurate idea of what is going on within the organization.

On the left side of the dashboard, we can see (in a time series fashion) the quarterly changes in revenues, expenses, and margins, as well as the comparison of those figures to previous years' monthly numbers. On the upper-right side we see two dials with color-coded regions showing the amount of monthly expenses for support services (dial on the left) and the amount of other expenses (dial on the right). As the color coding indicates, although the monthly support expenses are well within the normal ranges, the other expenses are in the red region, indicating excessive values. The geographic map on the bottom right shows the distribution of sales at the country level throughout the world. Behind these graphical icons there are variety of mathematical functions aggregating numerous data points to their highest level of meaningful figures. By clicking on these graphical icons, the consumer of this information can drill down to more granular levels of information and data. Dashboards are used in a wide variety of businesses for a wide variety of reasons

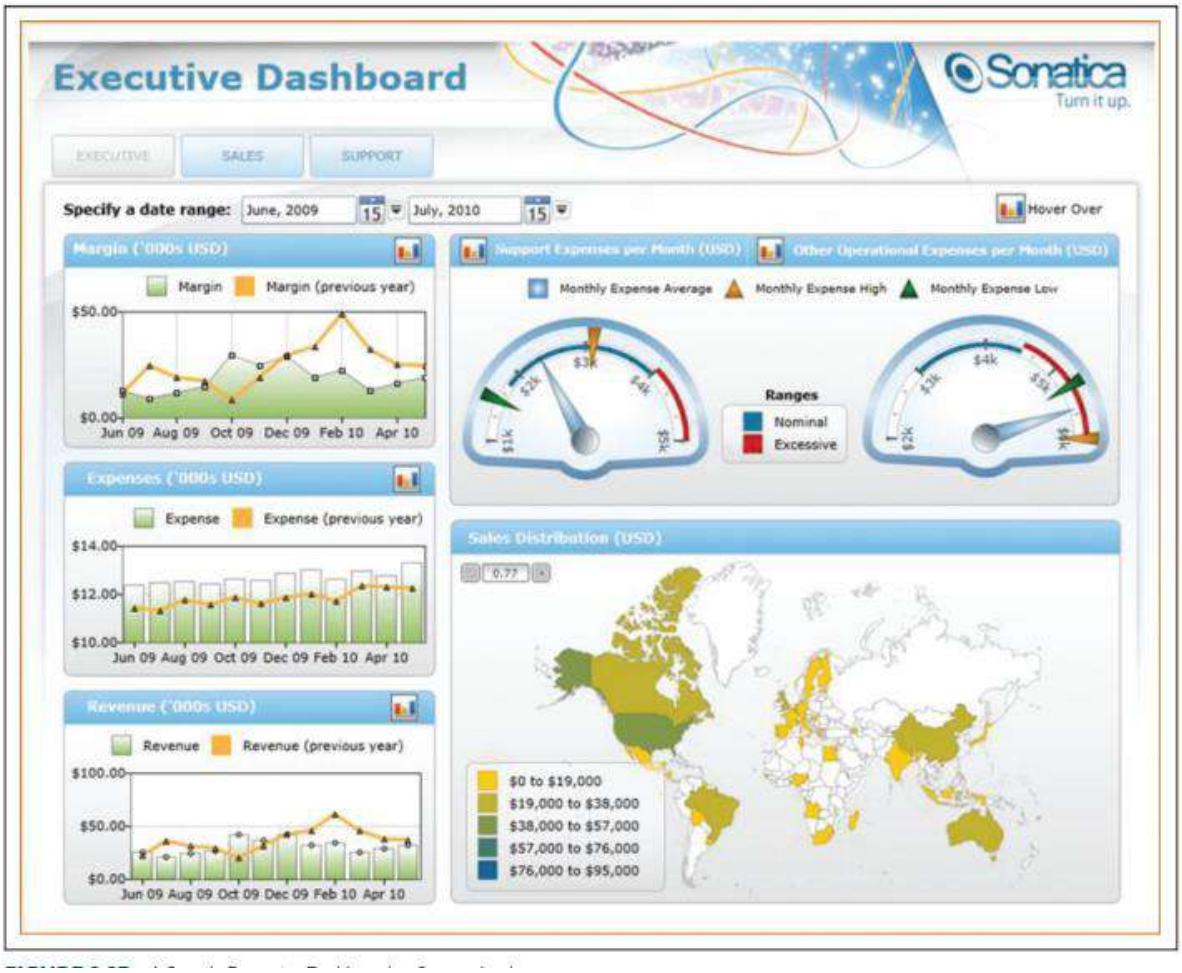


Figure 9.5 A Sample Executive Dashboard. Source: dundas.com

Dashboard Design

Dashboards are not a new concept. Their roots can be traced at least to the executive information system of the 1980s. Today, dashboards are ubiquitous. The Dashboard Spy Web site (dashboardspy.com/about) provides further evidence of their ubiquity. The site contains descriptions and screenshots of thousands of BI dashboards, scorecards, and BI interfaces used by businesses of all sizes and industries, nonprofits, and government agencies.

According to Eckerson (2006), a well-known expert on BI in general and dashboards in particular, the most distinctive feature of a dashboard is its three layers of information:

1. Monitoring: Graphical, abstracted data to monitor key performance metrics.
2. Analysis: Summarized dimensional data to analyze the root cause of problems.

3. Management: Detailed operational data that identify what actions to take to resolve a problem.

Because of these layers, dashboards pack a lot of information into a single screen. According to Few (2005), “The fundamental challenge of dashboard design is to display all the required information on a single screen, clearly and without distraction, in a manner that can be assimilated quickly.” To speed assimilation of the numbers, the numbers need to be placed in context. This can be done by comparing the numbers of interest to other baseline or target numbers, by indicating whether the numbers are good or bad, by denoting whether a trend is better or worse, and by using specialized display widgets or components to set the comparative and evaluative context. Some of the common comparisons that are typically made in BI systems include comparisons against past values, forecasted values, targeted values, benchmark or average values, multiple instances of the same measure, and the values of other measures (e.g., revenues versus costs).

Even with comparative measures, it is important to specifically point out whether a particular number is good or bad and whether it is trending in the right direction. Without these types of evaluative designations, it can be time consuming to determine the status of a particular number or result. Typically, either specialized visual objects (e.g., traffic lights, dials, and gauges) or visual attributes (e.g., color coding) are used to set the evaluative context.

What to Look for in a Dashboard ?

Although performance dashboards and other information visualization frameworks differ, they all do share some common design characteristics. First, they all fit within the larger BI and/or performance measurement system. This means that their underlying architecture is the BI or performance management architecture of the larger system. Second, all well-designed dashboard and other information visualizations possess the following characteristics (Novell, 2009):

- They use visual components (e.g., charts, performance bars, sparklines, gauges, meters, stoplights) to highlight, at a glance, the data and exceptions that require action.
- They are transparent to the user, meaning that they require minimal training and are extremely easy to use.

- They combine data from a variety of systems into a single, summarized, unified view of the business.
- They enable drill-down or drill-through to underlying data sources or reports, providing more detail about the underlying comparative and evaluative context.
- They present a dynamic, real-world view with timely data refreshes, enabling the end user to stay up to date with any recent changes in the business.
- They require little, if any, customized coding to implement, deploy and maintain.

9.8 CHECK YOUR PROGRESS

1. Define a report.
2. What is scatter plot?
3. Histograms are used to show the ____________of a variable or several variables.
4. Write the significance of dashboard.
5. Visual analytics is the combination of ______and ____________

Answers to Check your progress

1. A report is any communication artifact prepared with the specific intention of conveying information in a digestible form to whoever needs it, whenever and wherever they may need it. It is usually a document that contains information (usually driven from data) organized in a narrative, graphic, and/or tabular form, prepared periodically (recurring) or on an as-needed (ad hoc) basis, referring to specific time periods.
2. Scatter plots are often used to explore the relationship between two or three variables (in 2-D or 2-D visuals).
3. frequency distribution
4. Dashboards provide visual displays of important information that is consolidated and arranged on a single screen so that information can be digested at a single glance and easily drilled in and further explored.
5. Visualization and predictive analytics

9.9 SUMMARY

Data preprocessing is a tedious, time-demanding, yet crucial task in business analytics. A business report is a written document that contains information regarding business matters. The key to any successful business report is clarity, brevity, completeness, and correctness. Data visualization is the use of visual representations to explore, make sense of, and communicate data. Basic chart types include line, bar, and pie chart. Specialized charts are often derived from the basic charts as exceptional cases. Data visualization techniques and tools make the users of business analytics and BI systems better information consumers.

Visual analytics is the combination of visualization and predictive analytics. Increasing demand for visual analytics coupled with fast-growing data volumes led to exponential growth in highly efficient visualization systems investment. Dashboards provide visual displays of important information that is consolidated and arranged on a single screen so that information can be digested at a single glance and easily drilled in and further explored.

9.10 KEYWORDS

- **Business reporting:** or enterprise reporting refers to both "the public reporting of operating and financial data by a business enterprise," and "the regular provision of information to decision-makers within an organization to support them in their work." It is a fundamental part of the larger movement towards improved
- **Data preprocessing**: is a component of data preparation, describes any type of processing performed on raw data to prepare it for another data processing procedure. It has traditionally been an important preliminary step for the data mining process.
- **Data visualization:** is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
- **Dashboard**: is a visual display of all of your data. While it can be used in all kinds of different ways, its primary intention is to provide information at-a-glance, such as KPIs. A dashboard usually sits on its own page and receives information from a linked database.
- **Frequency distribution**: is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval. The interval size depends on the data being analyzed and the goals of the analyst.

9.11 QUESTIONS FOR SELF STUDY

1. Discuss categories of Business Reporting.
2. Write a note on Data Visualization.
3. Discuss different types of Charts and Graphs.
4. Explain various specialized Charts and Graphs in detail.
5. Elaborate on emergence of Visual Analytics.
6. Write a note on Information Dashboards.

9.12 REFERENCES

1. Sharda R, Delen D, Turban E. Business intelligence analytics and data science: A managerial perspective. Pearson, 2022.
2. Vercellis C. Business intelligence: data mining and optimization for decision making. John Wiley & Sons; 2011 Aug 10.
3. Bentley D. Business Intelligence and Analytics. Internet, link: https://www. pdfdrive. com/business-intelligence-and-analytics-e56416503. html. 2017.

UNIT 10 : FUTURE TRENDS, PRIVACY AND MANAGERIAL CONSIDERATIONS IN ANALYTICS

Structure

10.0 OBJECTIVES

After studying this unit, you will be able to:

- Discuss phenomenon of Internet of Things
- Elucidate Cloud Computing and Business Analytics
- Explain Location-Based Analytics for Organizations
- Give an account on issues of Legality, Privacy and Ethics

10.1 INTRODUCTION

This unit introduces several emerging technologies that are likely to have major impacts on the development and use of business intelligence (BI) applications. In a dynamic area such as analytics, the terms also evolve and overlap. As noted earlier, we can refer to these technologies as BI, analytics, data science, machine learning, artificial intelligence (AI), cognitive computing, Big Data, or by several other labels. We introduce and explain some emerging technologies and explore their current applications. We then discuss the organizational, personal, legal, ethical, and societal impacts of analytical support systems and issues that should be of importance to managers and professionals in analytics.

10.2 INTERNET OF THINGS

Internet of Things (IoT) is the phenomenon of connecting the physical world to the Internet, in contrast to the Internet of people that connects us humans to each other through technology. In IoT, physical devices are connected to sensors that collect data on the operation, location, and state of a device. This data is processed using various analytics techniques for monitoring the device remotely from a central office or for predicting any upcoming faults in the device. Perhaps the most common example of the IoT is the upcoming self-driving car. To drive on its own, a car needs to have enough sensors that automatically monitor the situation around it and take appropriate actions to adjust any setting necessary, including the car's speed, direction, and so on. Another common example of the IoT is a fitness tracker device that allows a user to keep track of physical activities such as walking, running, and sleep. Another example that illustrates the IoT phenomenon is a company called Smartbin. Smartbin has developed trash containers that include sensors to detect the fill levels. The trash collection company can automatically be informed to empty a trash container when the sensor detects it to be nearly full. Of course, the most common example people give in illustrating IoT is the idea of your refrigerator automatically ordering milk when it detects that the milk has run out! In all these examples, a human does not have to necessarily communicate with another human, or even with a machine in many cases. The machines can do the talking. That is why the term Internet of Things is used. There are many reasons IoT is growing exponentially:

1. **Hardware is smaller, affordable, and more powerful**: Costs of actuators and sensors have decreased significantly in the last 10 years, resulting in a much cheaper sensor overall. Cheap mobility: Costs of data processing, bandwidth, and mobile devices have gone down by 97% since the last decade.
2. **Availability of BI tools**: Now more and more companies are offering their BI tools both on premise and in the cloud at cheaper rates. Big Data and BI tools are widely available and are highly sophisticated.
3. **New and interesting use cases are emerging virtually every day**. We should also note that there is some disagreement about using the term Internet of Things. Some people also term this as the Web of Things. Others have argued to call it the Internet of Systems because in many ways it would be a combination of systems that would

communicate with one another. However, we will continue to refer to this phenomenon as the Internet of Things (IoT) in this section for the sake of consistency.

IoT Technology Infrastructure

From a bird's-eye view, IoT technology can be divided into four major blocks. Figure 10.1 illustrates these four blocks.

1. **Hardware**: It includes the physical devices, sensors, and actuators where data is produced and recorded. The device is the equipment that needs to be controlled, monitored, or tracked. IoT sensor devices could contain a processor or any computing device that parses incoming data.
2. **Connectivity:** There should be a base station or hub that collects data from the sensor-laden objects and sends that data to the cloud. Devices are connected to a network to communicate with each other or with other applications. These may be directly or indirectly connected to the Internet. A gateway enables devices that are not directly connected to Internet to reach the cloud platform.
3. **Software backend**: In this layer the data collected is managed. Software backend manages connected networks and devices and provides data integration. This may very well be in the cloud (again, see next section).
4. **Applications:** In this part of IoT, data is turned into meaningful information. Many of the applications may run on smartphones, tablets, and PCs and do something useful with the data. Other applications may run on the server and provide results or alerts through dashboards or messages to the stakeholders.

RFID Sensors

One of the earliest sensor technologies that has found a new life and is experiencing significant growth is radio-frequency identification (RFID). RFID is a generic technology that refers to the use of radio-frequency waves to identify objects. Fundamentally, RFID is one example of a family of automatic identification technologies, which also includes the ubiquitous barcodes and magnetic strips. Since the mid-1970s, the retail supply chain (and many other areas) has used barcodes as the primary form of automatic identification. The potential advantages of RFID have prompted many companies (led by large retailers such as Wal-Mart, Target, and Dillard's) to aggressively pursue this technology as a way to improve their supply chain and thus reduce costs and increase sales.

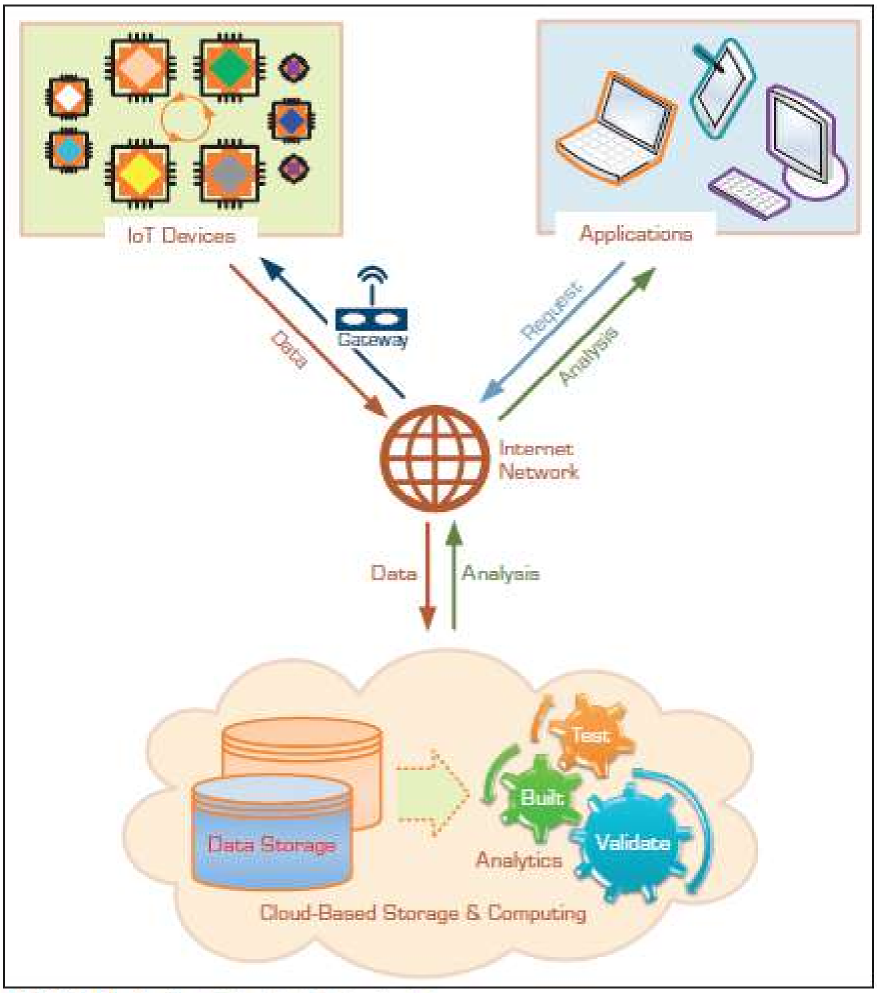


FIGURE 10.1 Building Blocks of IoT Technology Infrastructure.

How does RFID work? In its simplest form, an RFID system consists of a tag (attached to the product to be identified), an interrogator (i.e., reader), one or more antennae attached to the reader, and a computer (to control the reader and capture the data). At present, the retail supply chain has primarily been interested in using passive RFID tags. Passive tags receive energy from the electromagnetic field created by the interrogator (e.g., a reader) and backscatter information only when it is requested. The passive tag will remain energized only while it is within the interrogator's magnetic field.

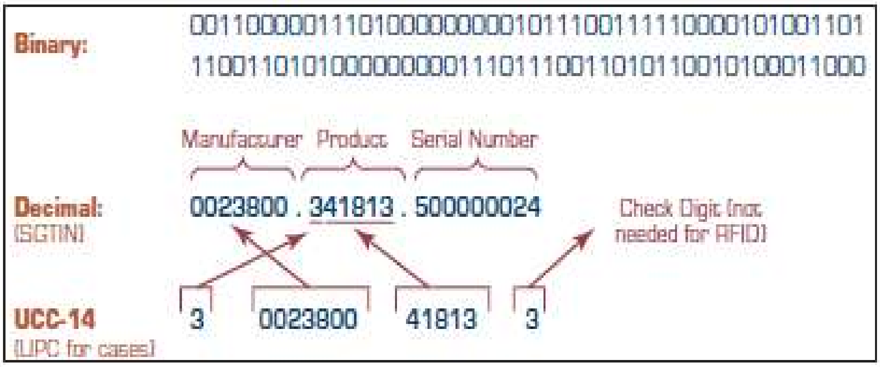


FIGURE 10.2 RFID Data Tag.

The most commonly used data representation for RFID technology is the Electronic Product Code (EPC), which is viewed by many in the industry as the next generation of the Universal Product Code (UPC), most often represented by a barcode. Like the UPC, the EPC consists of a series of numbers that identifies product types and manufacturers across the supply chain. The EPC code also includes an extra set of digits to uniquely identify items.

Currently, most RFID tags contain 96 bits of data in the form of serialized global trade identification numbers (SGTINs) for identifying cases or serialized shipping container codes for identifying pallets (although SGTINs can also be used to identify pallets). The complete guide to tag data standards can be found on EPCglobal's Web site (epcglobalinc .org). EPCglobal, Inc., is a subscriber-driven organization of industry leaders and organizations focused on creating global standards for the EPC to support the use of RFID.

As illustrated in Figure 10.2, tag data, in its purest form, is a series of binary digits. This set of binary digits can then be converted to the SGTIN decimal equivalent. As shown, an SGTIN is essentially a UPC (UCC-14, for shipping-container identification) with a serial number. The serial number is the most important difference between the 14-digit UPC used today and the SGTIN contained on an RFID tag. With UPCs, companies can identify the product family to which a case belongs (e.g., 8-pack Charmin tissue), but they cannot distinguish one case from another. With an SGTIN, each case is uniquely identified. This provides visibility at the case level, rather than the product-family level.

Fog Computing

One of the key issues in IoT is that the data produced by sensors is huge, and not all of it is useful. So how much should be uploaded to the cloud servers for analysis? A recent concept to address this question is the idea of fog computing. Fog extends the cloud to be closer to the things that produce and act on IoT data. These devices, called fog nodes, can be placed anywhere between the network connection. Any device with computing, storage, and network connectivity can be a fog node, for example, routers or switches. The following view illustrates this:

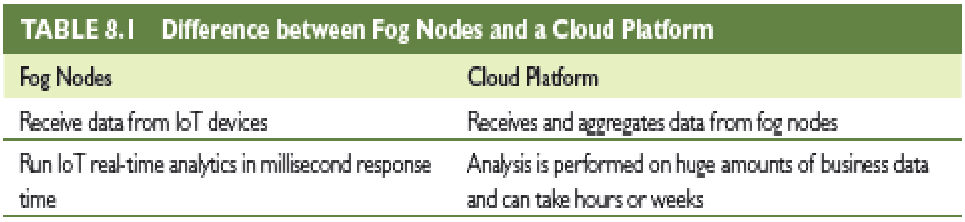


TABLE 8.1 Difference between Fog Nodes and a Cloud Platform

Fog Nodes	Cloud Platform
Receive data from IoT devices	Receives and aggregates data from fog nodes
Run IoT real-time analytics in millisecond response time	Analysis is performed on huge amounts of business data and can take hours or weeks

Data Center/Cloud ------->> Fog device ------->> Physical device/Sensors generating data

Analyzing data close to the devices minimizes latency. It also conserves bandwidth, as sending data to the cloud requires large bandwidth. Fog computing is crucial in situations when data need to be analyzed in less than a second, as in the case of a cascading system failure. Table 8.1 identifies two simple differences between the cloud and fog. Fog computing may also give better security, as fog nodes can be secured with the same security solution used in the other IT environments.

10.3 CLOUD COMPUTING AND BUSINESS ANALYTICS

Another emerging technology trend that business analytics users should be aware of is cloud computing. The National Institute of Standards and Technology (NIST) defines cloud computing as "a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, and services) that can be rapidly provisioned and released with minimal management effort or service-provider interaction".

Although we do not typically look at Web-based e-mail as an example of cloud computing, it can be considered a basic cloud application. Typically, the e-mail application stores the data (e-mail messages) and the software (e-mail programs that let us process and manage e-mails). The e-mail provider also supplies the hardware/software and all of the basic infrastructure. As long as the Internet is available, one can access the e-mail application from anywhere in the cloud. When the application is updated by the e-mail provider (e.g., when

Gmail updates its e-mail application), it becomes available to all customers without them having to download any new programs. Social networking Web sites like Facebook, Twitter, and LinkedIn, are also examples of cloud computing. Thus, any Web-based general application is in a way an example of a cloud application.

Another example of a general cloud application is Google Docs and Spreadsheets. This application allows a user to create text documents or spreadsheets that are stored on Google's servers and are available to the users anywhere they have access to the Internet. Again, no programs need to be installed as "the application is in the cloud." The storage space is also "in the cloud." A good general business example of cloud computing is Amazon.com's Web services.

Data as a Service (DaaS)

The concept of data as a service basically advocates the view that "where data lives"— the actual platform on which the data resides—doesn't matter. Data can reside in a local computer or in a server at a server farm inside a cloud-computing environment. With DaaS, any business process can access data wherever it resides. Data as a service began with the notion that data quality could happen in a centralized place, cleansing and enriching data and offering it to different systems, applications, or users, irrespective of where they were in the organization, computers, or on the network. This has now been replaced with master data management and customer data integration solutions, where the record of the customer (or product, or asset, etc.) may reside anywhere and is available as a service to any application that has the services allowing access to it. By applying a standard set of transformations to the various sources of data (for example, ensuring that gender fields containing different notation styles [e.g., M/F, Mr./Ms.] are all translated into male/female) and then enabling applications to access the data via open standards such as SQL, XQuery, and XML, service requestors can access the data regardless of vendor or system.

With DaaS, customers can move quickly thanks to the simplicity of the data access and the fact that they don't need extensive knowledge of the underlying data. If customers require a slightly different data structure or have location-specific requirements, the implementation is easy because the changes are minimal (agility). Second, providers can build the base with the data experts and outsource the analysis or presentation layers (which allows for very cost-effective user interfaces and makes change requests at the presentation layer much more feasible), and access to the data is controlled through the data services. It tends to improve data quality because there is a single point for updates.

Software as a Service (SaaS)

This model allows consumers to use applications and software that run on distant computers in the cloud infrastructure. Consumers need not worry about managing underlying cloud infrastructure and have to pay for the use of software only. All we need is a Web browser to connect to the cloud. Gartner estimates that SaaS revenue was around $32 billion in 2015 and is used in 77% of all organizations. Gmail, Picasa, and Flickr are examples of SaaS.

Platform as a Service (PaaS)

Using this model, companies can deploy their software and applications in the cloud so that their customers can use them. Companies don’t have to manage resources needed to manage their applications in cloud-like networks, servers, storage, or operating systems. This reduces the cost of maintaining underlying infrastructure for running their software and also saves time for setting up this infrastructure. Now, users can focus on their business rather than focusing on managing infrastructure for running their software. Examples of PaaS are Microsoft Azure, Amazon EC2, and Google App Engine.

Infrastructure as a Service (IaaS)

In this model, infrastructure resources like networks, storage, servers, and other computing resources are provided to client companies. Clients can run their application and have administrative rights to use these resources but do not manage underlying infrastructure. Clients have to pay for usage of infrastructure. A good example of that is Amazon .com’s Web services. Amazon.com has developed impressive technology infrastructure that includes data centers. Other companies can use Amazon.com’s cloud services on a pay-per-use-basis without having to make similar investments.

Essential Technologies for Cloud Computing

VIRTUALIZATION: Virtualization is the creation of a virtual version of something like an operating system or server. A simple example of virtualization is the logical division of a hard drive to create two separate hard drives in a computer. Virtualization can be in all three areas of computing:

- **Network virtualization:** It is the splitting of available bandwidth into channels, which disguises complexity of the network by dividing it into manageable parts. Then each bandwidth can be allocated to a particular server or device in real time.
- **Storage virtualization:** It is the pooling of physical storage from multiple network storage devices into a single storage device that can be managed from a central console.
- **Server virtualization:** It is the masking of physical servers from server users. Users don't have to manage the actual servers or understand complicated details of server resources. This difference in the level of virtualization directly relates to which cloud service one employs.

Cloud Deployment Models

Cloud services can be acquired in several ways, from building an entirely private infrastructure to sharing with others. The following three models are the most common:

- **Private cloud:** This can also be called internal cloud or corporate cloud. It is a more secure form of cloud service than public clouds like MS Azure and Google BigQuery. It is operated solely for a single organization having a mission critical workload and security concerns. It provides the same benefits as a public cloud-like service, scalability, changing computing resources on demand, and so on. Companies that have a private cloud have direct control over their data and applications. The disadvantage of having a private cloud is the cost of maintaining and managing the cloud because on-premise IT staff are responsible for managing it.
- **Public cloud:** In this model the subscriber uses the resources offered by service providers over the Internet. The cloud infrastructure is managed by the service provider. The main advantage of this public cloud model is saving time and money in setting up hardware and software required to run their business. Examples of public clouds are Microsoft Azure platform, Google App Engine, and Amazon AWS.
- **Hybrid cloud:** The hybrid cloud gives businesses great flexibility by moving workloads between private and public clouds. For example, a company can use hybrid cloud storage to store its sales and marketing data, and then use a public cloud platform like Amazon Redshift to run analytical queries to analyze its data. The main requirement is network connectivity and API (application program interface) compatibility between the private and public cloud.

10.4 LOCATION-BASED ANALYTICS FOR ORGANIZATIONS

Thus far, we have seen many examples of organizations employing analytical techniques to gain insights into their existing processes through informative reporting, predictive analytics, forecasting, and optimization techniques. In this section, we learn about a critical emerging trend—incorporation of location data in analytics. Figure 10.3 gives our classification of location-based analytic applications. We first review applications that make use of static location data that is usually called geospatial data. We then examine the explosive growth of applications that take advantage of all the location data being generated by today's devices.

Geospatial Analytics

A consolidated view of the overall performance of an organization is usually represented through the visualization tools that provide actionable information. The information may include current and forecasted values of various business factors and key performance indicators (KPIs). Looking at the KPIs as overall numbers via various graphs and charts can be overwhelming. There is a high risk of missing potential growth opportunities or not identifying the problematic areas. As an alternative to simply viewing reports, organizations employ visual maps that are geographically mapped and based on the traditional location data, usually grouped by postal codes. These map-based visualizations have been used by organizations to view the aggregated data and get more meaningful location- based insights. The traditional location-based analytic techniques using geocoding of organizational locations and consumers hamper the organizations in understanding "true location-based" impacts. Locations based on postal codes offer an aggregate view of a large geographic area. This poor granularity may not help pinpoint the growth opportunities within a region, as the location of target customers can change rapidly.

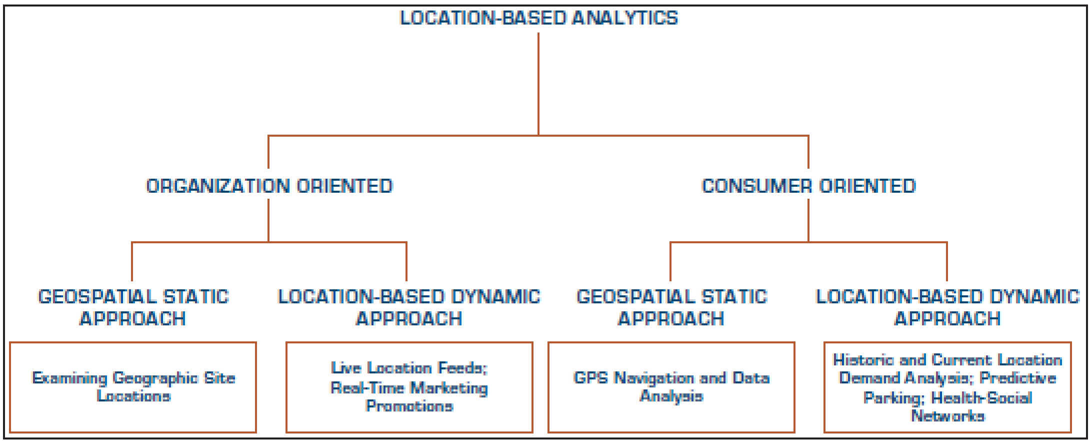

LOCATION-BASED ANALYTICS
ORGANIZATION ORIENTED
CONSUMER ORIENTED
GEOSPATIAL STATIC APPROACH
LOCATION-BASED DYNAMIC APPROACH
GEOSPATIAL STATIC APPROACH
LOCATION-BASED DYNAMIC APPROACH
Examining Geographic Site Locations
Live Location Feeds; Real-Time Marketing Promotions
GPS Navigation and Data Analysis
Historic and Current Location Demand Analysis; Predictive Parking; Health-Social Networks

Figure 10.3 Classification of Location-Based Analytics Applications.

Location-based data are now readily available from geographic information systems (GIS). These are used to capture, store, analyze, and manage data linked to a location using integrated sensor technologies, global positioning systems installed in smartphones, or through RFID deployments in the retail and healthcare industries. Organizations now create interactive maps that further drill down to details about any location, offering analysts the ability to investigate new trends and correlate location-specific factors across multiple KPIs. Analysts can now pinpoint trends and patterns in revenue, sales, and profitability across geographical areas.

By incorporating demographic details into locations, retailers can determine how sales vary by population level and proximity to other competitors; they can assess the demand and efficiency of supply-chain operations. Consumer product companies can identify the specific needs of customers and customer complaint locations and easily trace them back to the products. Sales reps can better target their prospects by analyzing their geography.

10.5 ISSUES OF LEGALITY, PRIVACY AND ETHICS

As data science, analytics, cognitive computing, and AI grow in reach and pervasiveness, everyone is affected by these applications. Just because something is doable through technology, does not make it appropriate, legal, or ethical. Data science professionals and managers have to be very aware of these concerns. Several important legal, privacy, and ethical issues are related to analytics. Here we provide only representative examples and sources. Popular media is usually quite keen to report on such breaches of legal and ethical behavior, so this is one section where you may find even more recent examples online.

Legal Issues

The introduction of analytics may compound a host of legal issues already relevant to computer systems. For example, questions concerning liability for the actions of advice provided by intelligent machines are beginning to be considered.

In addition to resolving disputes about the unexpected and possibly damaging results of some analytics, other complex issues may surface. For example, who is liable if an enterprise finds itself bankrupt as a result of using the advice of an analytic application? Will the enterprise itself be held responsible for not testing the system adequately before entrusting it with sensitive issues? Will auditing and accounting firms share the liability for failing to apply adequate auditing tests? Will the software developers of intelligent systems be jointly liable? As self-driving cars become more common, who is liable for any damage or injury when a car’s sensors, network, or the analytics fail to function as planned?

Privacy

Privacy means different things to different people. In general, privacy is the right to be left alone and the right to be free from unreasonable personal intrusions. Privacy has long been a legal, ethical, and social issue in many countries. The right to privacy is recognized today in every state of the United States and by the federal government, either by statute or by common law. The definition of privacy can be interpreted quite broadly. However, the following two rules have been followed fairly closely in past court decisions:

(1) The right of privacy is not absolute. Privacy must be balanced against the needs of society.

(2) The public’s right to know is superior to the individual’s right to privacy.

Collecting Information about Individuals

The complexity of collecting, sorting, filing, and accessing information manually from numerous government agencies was, in many cases, a built-in protection against the misuse of private information. It was simply too expensive, cumbersome, and complex to invade a person’s privacy. The Internet, in combination with large-scale databases, has created an entirely new dimension of accessing and using data. The inherent power in systems that can access vast amounts of data can be used for the good of society. For example, by matching records with the aid of a computer, it is possible to eliminate or reduce fraud, crime, government mismanagement, tax evasion, welfare cheating, family support filching, employment of illegal workers, and so on.

However, what price must the individual pay in terms of loss of privacy so that the government can better apprehend criminals? The same is true on the corporate level. Private information about employees may aid in better decision making, but the employees' privacy may be affected. Similar issues are related to information about customers. The implications for online privacy are significant

Mobile User Privacy

Many users are unaware of the private information being tracked through their smartphone usage. Many apps collect user data that track each phone as it moves from one cell tower to another, from GPS-enabled devices that transmit users' locations, and from phones transmitting information at Wi-Fi hotspots. Major app developers claim they are extremely careful and protective of users' privacy, but it is interesting to note how much information is available through the use of a single device.

Homeland Security and Individual Privacy

Using analytics technologies such as mining and interpreting the content of telephone calls, taking photos of people in certain places and identifying them, and using scanners to view your personal belongings are considered by many to be an invasion of privacy. However, many people recognize that analytic tools are an effective and efficient means to increase security, even though the privacy of many innocent people is compromised.

Recent Technology Issues in Privacy and Analytics

Most providers of Internet services such as Google, Facebook, Twitter, and others depend on monetizing their users' actions. They do so in many different ways, but all of these approaches in the end amount to understanding a user's profile or preferences on the basis of their usage. With the growth of Internet users in general and mobile device users in particular, many companies have been founded to employ advanced analytics to develop profiles of users on the basis of their device usage, movement, and the contacts of the users.

Who Owns Our Private Data?

With the recent growth of data from our use of technology and companies' ability to access and mine it, the privacy debate also leads to the obvious question of whose property any user's data is. Welch (2016) highlighted this issue in a Bloomberg Business week column. Take an example of a relatively new car. The car is equipped with many sensors starting with tire pressure sensors to GPS trackers that can keep track of where you have gone, how fast you were driving, when you changed lanes, and so on. The car may even know the passenger's weight added to the front seat. As Welch notes, a car connected to the Internet (most new cars are!) can be a privacy nightmare for the owner or a data "gold mine" for whoever can possess this data. A major battle is brewing between automobile manufacturers and technology providers such as Apple (CarPlay) and Google (Android Auto) on who owns this data and who can get access to this data. This is becoming more crucial because as cars become more self-driving, the driver/passenger in the car could be a highly targeted prospective customer for specific products and services whose profile is very well known to the organization who is able to create that profile.

For example, Google's Waze app collects user GPS data for over 50 million users to track traffic information and help users find the best route, but then displays pop-up ads on the users' screens. Yelp, Spotify, and other apps popularly used in the car have similar plans and applications. A similar battle is also brewing about users' health and biometric data. Because of security concerns, many users are moving to biometric log-in authentication using fingerprints, touch screens, iris scans, and so on. Because this information is highly unique to an individual, future profiling of a user may become even more precise. Thus the battle to own and relate this information to other data gathered is growing as well. Similarly, hospitals, medical professionals, labs, and insurance companies collect a lot of information about our medical history. Although in the United States there are strict laws in place (e.g., HIPAA) to protect a user's privacy, compilation of this information is unleashing major advances in health analytics. The privacy challenge, however, is still very real.

Bottom line, as a data analytics professional, be very aware of the legal and ethical issues involved in collecting information that may be privileged or protected. A general question to ask yourself is—would you like your own information to be included for the application you are contemplating?

Ethics in Decision Making and Support

The last question brings us to several ethical issues that are related to analytics. Representative ethical issues that could be of interest in analytics implementations include the following:

- Electronic surveillance
- Ethics in DSS design
- Software piracy
- Invasion of individuals' privacy
- Use of proprietary databases
- Use of intellectual property such as knowledge and expertise
- Exposure of employees to unsafe environments related to computers
- Computer accessibility for workers with disabilities
- Accuracy of data, information, and knowledge
- Protection of the rights of users
- Accessibility to information
- Use of corporate computers for non-work-related purposes
- How much decision making to delegate to computers

One story that made many users upset (although it was not illegal) some time back was Facebook's experiment to present different News Feeds to the users and monitor their emotional reactions as measured by replies, likes, sentiment analysis, and so on. Most companies, including technology companies, run user testing to identify the features most liked or disliked and fine-tune their product offerings. Because Facebook is so large, running this experiment without the users' informed consent was viewed as unethical. Indeed, Facebook acknowledged its error and instituted a more formal review through Internal Review Boards and other compliance mechanisms for future testing. Although they faced a lot of bad press initially, their timely response allowed them to recover quickly.

10.6 CHECK YOUR PROGRESS

1. Acronym for RFID
2. Define Internet of things.
3. What is fog computing?
4. Define Virtualization.

5. Acronym for KPIs.

Answers to Check your progress

1. Radio-Frequency Identification Device
2. Internet of Things (IoT) is the phenomenon of connecting the physical world to the Internet, in contrast to the Internet of people that connects us humans to each other through technology.
3. Fog extends the cloud to be closer to the things that produce and act on IoT data.
4. Virtualization is the creation of a virtual version of something like an operating system or server.
5. key performance indicators

10.7 SUMMARY

This unit introduces several emerging technologies that are likely to have major impacts on the development and use of business intelligence (BI) applications. Some emerging technologies and their current applications are introduced and detailed. Discussion is made on organizational, personal, legal, ethical, and societal impacts of analytical support systems and issues that should be of importance to managers and professionals in analytics

10.8 KEYWORDS

- **Cloud computing**: Cloud computing is a general term for anything that involves delivering hosted services over the internet
- **Sensors**: A sensor is a device that detects and responds to some type of input from the physical environment
- **Big data**: is a combination of structured, semi structured and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modeling and other advanced analytics applications.
- **Fog computing**: is a decentralized computing infrastructure in which data, compute, storage and applications are located somewhere between the data source and the cloud.
- **Google Docs**: is an online word processor that lets you create and format documents and work with other people. See our top five tips for Google Docs.

10.9 QUESTIONS FOR SELF STUDY

1. Write a note on Internet of Things(IoT).
2. Elucidate the reasons for IOT growing exponentially.
3. How does RFID work?
4. Explain cloud deployment models.

10.10 REFERENCES

1. Sharda R, Delen D, Turban E. Business intelligence analytics and data science: A managerial perspective. Pearson, 2022.
2. Vercellis C. Business intelligence: data mining and optimization for decision making. John Wiley & Sons; 2011 Aug 10.
3. Bentley D. Business Intelligence and Analytics. Internet, link: https://www. pdfdrive. com/business-intelligence-and-analytics-e56416503. html. 2017.

UNIT-11 : MARKETING MODELS

Structure

11.0 OBJECTIVES

After studying this unit, you will be able to:

- Importance of marketing models.
- Brief introduction on relational marketing.
- Analyze Relational marketing strategies
- Give an account on components of a relational marketing strategy
- Study the network of relationships involved in a relational marketing strategy

11.0 INTRODUCTION

Marketing decision processes are characterized by a high level of complexity due to the simultaneous presence of multiple objectives and countless alter- native actions resulting from the combination of the major choice options available to decision makers. Therefore, it should come as no surprise thata large number of mathematical models for marketing have been successfully developed and applied in recent decades.

The importance of mathematical models for marketing has been further strengthened by the availability of massive databases of sales transactions that provide accurate information on how customers make use of services or purchase products.

This unit will primarily focus on two prominent topics in the field of marketing intelligence. The first theme is particularly broad and concerns the application of predictive models to support relational marketing strategies, whose purpose is to customize and strengthen the relationship between a company and its customers.

11.2 RELATIONAL MARKETING

In order to fully understand the reasons why enterprises develop relational marketing initiatives, consider the following three examples: an insurance company that wishes to select the most promising market segment to target for a new type of policy; a mobile phone provider that wishes to identify those customers with the highest probability of churning, that is, of discontinuing their service and taking out a new contract with a competitor, in order to develop targeted retention initiatives; a bank issuing credit cards that needs to identify a group of customers to whom a new savings management service should be offered. These situations share some common features: a company owning a massive database which describes the purchasing behavior of its customers and the way they make use of services, wishes to extract from these data useful and accurate knowledge so as to develop targeted and effective marketing campaigns.

The aim of a relational marketing strategy is to initiate, strengthen, intensify and preserve over time the relationships between a company and its stakeholders, represented primarily by its customers, and involves the analysis, planning, execution and evaluation of the activities carried out to pursue these objectives.

Relational marketing became popular during the late 1990s as an approach to increasing customer satisfaction in order to achieve a sustainable competitive advantage. So far, most enterprises have taken at least the first steps in this direction, through a process of cultural change which directs greater attention toward customers, considering them as a formidable asset and one of the main sources of competitive advantage.

A relational marketing approach has been followed in a first stage by service companies in the financial and telecommunications industries, and has later influenced industries such as consumer goods, finally reaching also manufacturing companies, from automotive and commercial vehicles to agricultural equipments, traditionally more prone to a vision characterized by the centrality of products with respect to customers.

11.3 MOTIVATIONS AND OBJECTIVES

The reasons for the spread of relational marketing strategies are complex and interconnected. The increasing concentration of companies in large enterprises and the resulting growth in the number of customers have led to greater complexity in the markets.

Since the 1980s, the innovation – production – obsolescence cycle has progressively shortened, causing a growth in the number of customized options on the part of customers, and an acceleration of marketing activities by enterprises.

- The increased flow of information and the introduction of e-commerce have enabled global comparisons. Customers can use the Internet to compare features, prices and opinions on products and services offered by the various competitors.
- Customer loyalty has become more uncertain, primarily in the service industries, where often filling out an on-line form is all one has to do to change service provider.
- In many industries a progressive commoditization of products and ser- vices is taking place, since their quality is perceived by consumers as equivalent, so that differentiation is mainly due to levels of service.
- The systematic gathering of sales transactions, largely automated in most businesses, has made available large amounts of data that canbe transformed into knowledge and then into effective and targeted marketing actions.
- The number of competitors using advanced techniques for the analysis of marketing data has increased.

Relational marketing strategies revolve around the choices shown in Figure 11.1,which can be effectively summarized as formulating for each segment, ideally for each customer, the appropriate offer through the most suitable channel, at the right time and at the best price.

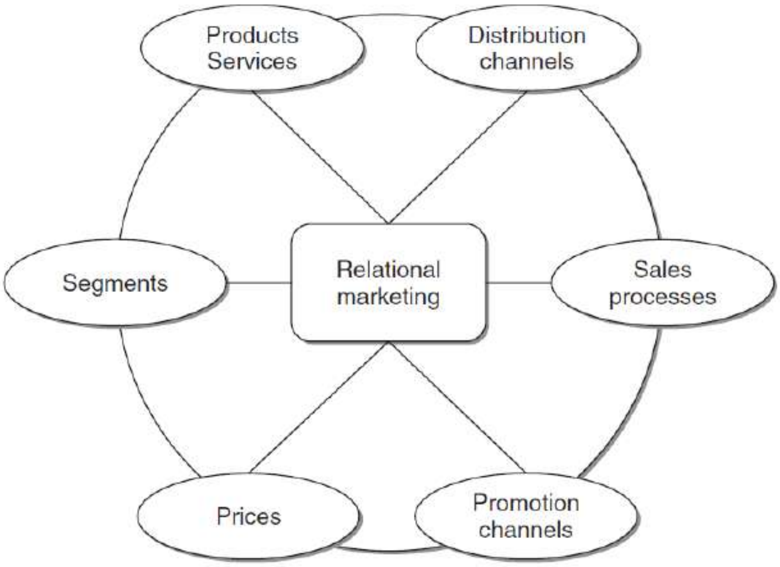


Figure 11.1 Decision-making options for a relational marketing strategy

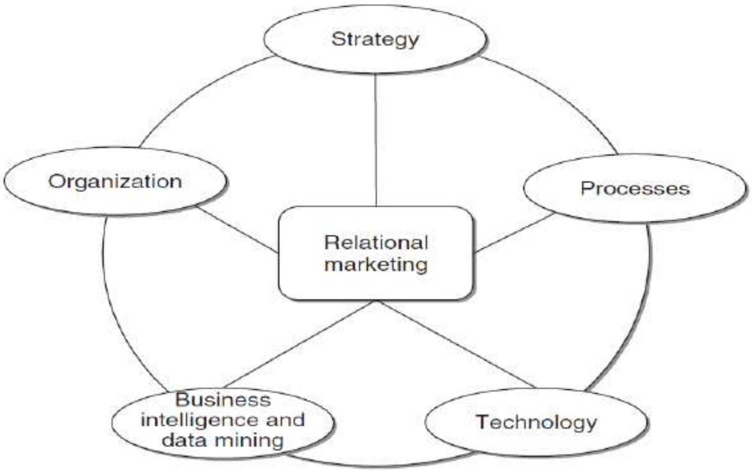


Figure 11.2 Components of a relational marketing strategy

The ability to effectively exploit the information gathered on customers' behavior represents today a powerful competitive weapon for an enterprise. A company capable of gathering, storing, analyzing and understanding the huge amount of data on its customers can base its marketing actions on the knowledge extracted and achieve sustainable competitive advantages. Enterprises may profitably adopt relational marketing strategies to transform occasional contacts with their customers into highly customized long-term relationships. In this way, it is possible to achieve increased customer satisfaction and at the same time increased profits for the company, attaining a win – win relationship.

To obtain the desired advantages, a company should turn to relational marketing strategies by following a correct and careful approach. In particular, it is advisable to stress the distinction between a relational marketing vision and the software tools usually referred to as customer relationship management (CRM). As shown in Figure 11.2, relational marketing is not merely a collection of software applications, but rather a coherent project where the various company departments are called upon to cooperate and integrate the managerial culture and human resources, with a high impact on the organizational structures. It is then necessary to create within a company a true data culture, with the awareness that customer-related information should be enhanced through the adoption of business intelligence and data mining analytical tools.

Based on the investigation of cases of excellence, it can be said that a successful relational marketing strategy can be achieved through the development of a company-wide vision that puts customers at the center of the whole organization. Of course, this goal cannot be attained by exclusively relying on innovative computer technologies, which at most can be considered a relevant enabling factor.

The overlap between relational marketing strategies and CRM software led to a misunderstanding with several negative consequences. On one hand, the notion that substantial investments in CRM software applications were in themselves sufficient to generate a relational marketing strategy represents a dangerous simplification, which caused many project failures. On the other hand, the high cost of software applications has led many to believe that a viable approach to relational marketing was only possible for large companies in the service industries. This is a deceitful misconception: as a matter offact, the essential components of relational marketing are a well-designed and correctly fed marketing data mart, a collection of business intelligence and data mining analytical tools, and, most of all, the cultural education of the decision makers. These tools will enable companies to carry out the required analyses and translate the knowledge acquired into targeted marketing actions.

The relationship system of an enterprise is not limited to the dyadic relation- ship with its customers, represented by individuals and companies that purchase the products and services offered, but also includes other actors, such as the employees, the suppliers and the sales network. For most relationships shown in Figure 11.3, a mutually beneficial exchange occurs between the different subjects involved. More generally, we can widen the boundaries of relational marketing systems to include the stakeholders of an enterprise. The relationship between an enterprise and its customers is sometimes mediated by the sales network, which in some instances can partially obstruct the visibility of the end customers.

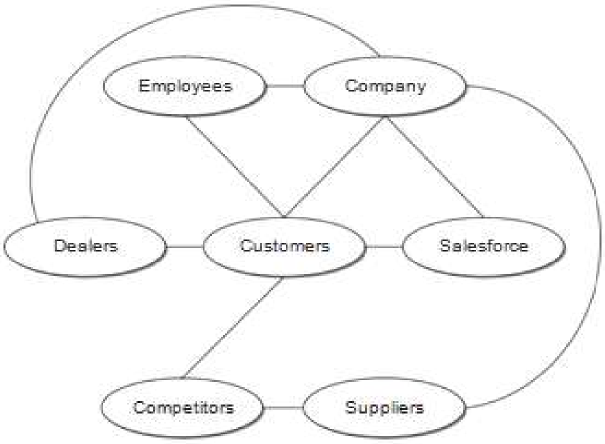


Figure 11.3 Network of relationships involved in a relational marketingstrategy

Let us take a look at a few examples to better understand the implications ofthis issue. The manufacturers of consumer goods, available at the points of saleof large and small retailers, do not have direct information on the consumers purchasing their products. The manufacturers of goods covered by guarantees, such as electrical appliances or motor vehicles, have access to personal information on purchasers, even if they rarely also have access to information onthe contacts of and promotional actions carried out by the network of dealers.

Likewise, a savings management company usually places shares in its investment funds through a network of intermediaries, such as banks or agents, and often knows only the personal data of the subscribers. A pharmaceutical enterprise producing prescription drugs usually ignores the identity of the patients that use its drugs and medicinal products, even though promotional activities toinfluence consumers are carried out in some countries where the law permits.

It is not always easy for a company to obtain information on its end customers from dealers in the sales network and even from their agents. Thesemay be reluctant to share the wealth of information for fear, rightly or wrongly,of compromising their role. In a relational marketing project specific initiatives should be devised to overcome these cultural and organizational barriers, usually through incentives and training courses.

The number of customers and their characteristics strongly influence the nature and intensity of the relationship with an enterprise, as shown in Figure 11.4. The relationships that might actually be established in a specific economic domain tend to lie on the diagonal shown in the figure. At one extreme, there are highly intense relationships existing between the company and a small number of customers of high individual value. Relationships of this type occur more frequently in business-to-business (B2B) activities, although they can also be found in other domains, such as private banking.

The high value of each customer justifies the use of dedicated resources, usually consisting of sales agents and key account managers, so as to maintain and strengthen these more intense relationships. In situations of this kind, careful organization and planning of the activities of sales agents is critical. Therefore, optimization models for sales- force automation (SFA), described in Section 11.2, can be useful in this context. At the opposite extreme of the diagonal are the relationships typical of consumer goods and business-to-consumer (B2C) activities, for which a high number of low-value customers get in contact with the company in an impersonal way, through websites, call centers and points of sale.

Data mining analyses for segmentation and profiling are particularly valuable especially in this con- text, characterized by a large number of fragmented contacts and transactions. Relational marketing strategies, which are based on the knowledge extracted through data mining models, enable companies develop a targeted customization and differentiation of their products and/or services, including companies more prone toward a mass-market approach

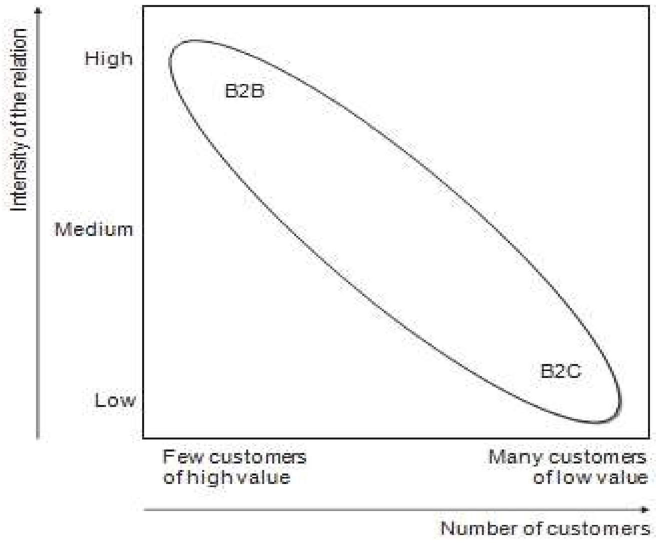

Intensity of the relation
High
Medium
Low
B2B
B2C
Few customers of high value
Many customers of low value
Number of customers

Figure 11.4 Intensity of customer relationships as a function of number of customers

Figure 11.5 contrasts the cost of sales actions and the corresponding revenues. Where transactions earn a low revenue per unit, it is necessary to implement low-cost actions, as in the case of mass-marketing activities. Moving down along the diagonal in the figure, more evolved and intense relationships with the customers can be found. The relationships at the end of the diagonal presuppose the action of a direct sales network and for the most part are typicalof B2B relational contexts.

Figure 11.6 shows the ideal path that a company should follow so as to be able to offer customized products and services at low cost and in a short time. On the one hand, companies operating in a mass market, well acquainted with fast delivery at low costs, must evolve in the direction of increased customization, by introducing more options and variants of products and services offered to the various market segments. Data mining analyses for relational marketing purposes are a powerful tool for identifying the segments to be targeted with customized products. On the other hand, the companies oriented toward make-to-order production must evolve in a direction that fosters reductions in both costs and delivery times, but without reducing the variety and the rangeof their products.

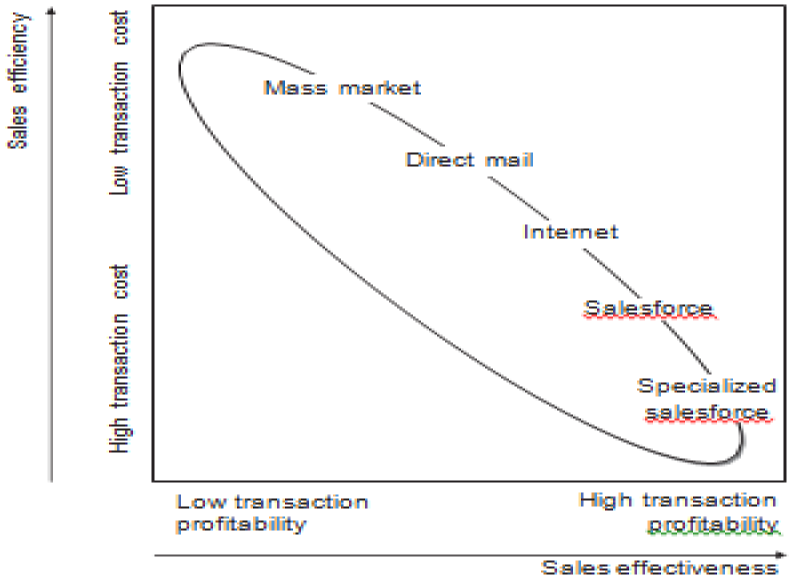


Figure 11.5 Efficiency of sales actions as a function of their effectiveness

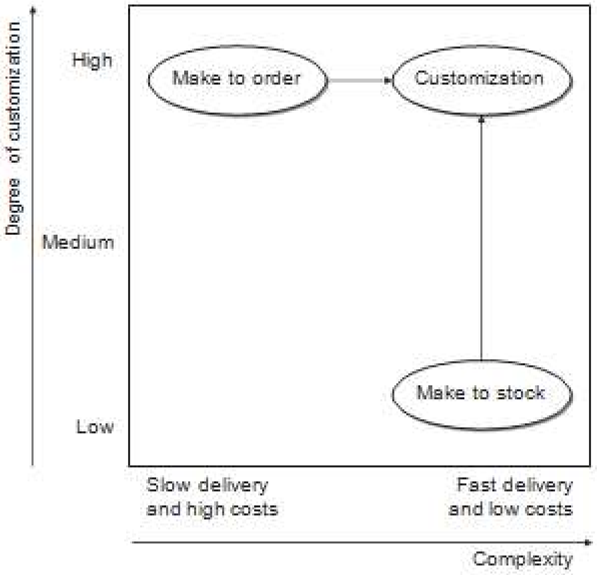


Figure 11.6 Level of customization as a function of complexity of products andservices

11.4 CHECK YOUR PROGRESS

1. What is the aim of relational marketing?
2. Give acronym for CRM
3. The relationship between an enterprise and its customers is sometimes mediated by the__________.

4. What is distribution channel?

Answers to Check your progress

1. The aim of a relational marketing strategy is to initiate, strengthen, intensify and preserve over time the relationships between a company and its stakeholders, represented primarily by its customers, and involves the analysis, planning, execution and evaluation of the activities carried out to pursue these objectives.
2. Customer Relationship Management (CRM).
3. sales network
4. A distribution channel represents a chain of businesses or intermediaries through which the final buyer purchases a good or service.

11.5 SUMMARY

A brief introduction to relational marketing is given along with the main streams of analysis that can be dealt with domain of application Also, indicating for each of them the classes of predictive models that are best suited to dealing with the problems are considered. The subjects discussed in this context can be partly extended to the relationship between citizens and the public administration.

11.6 KEYWORDS

- **Relationship marketing**: is a facet of customer relationship management (CRM) that focuses on customer loyalty and long-term customer engagement rather than shorter-term goals like customer acquisition and individual sales.
- **Customer relationship management** (CRM): is a technology for managing all your company's relationships and interactions with customers and potential customers. The goal is simple: Improve business relationships.
- **Organizational structure**: is a system that outlines how certain activities are directed in order to achieve the goals of an organization.
- **Stakeholder** : is a party that has an interest in a company and can either affect or be affected by the business.
- **Promotion Channel**: means a method or format for the placement of the Company Advertising, including, Internet advertising, outdoor advertising, television and/or radio advertising

11.7 QUESTIONS FOR SELF STUDY

1. Write the importance of marketing models.
2. Give a brief introduction on relational marketing.
3. Discuss Relational marketing strategies
4. Give an account on components of a relational marketing strategy using a diagram.
5. Narrate the network of relationships involved in a relational marketing strategy

11.8 REFERENCES

1. Sharda R, Delen D, Turban E. Business intelligence analytics and data science: A managerial perspective. Pearson, 2022.
2. Vercellis C. Business intelligence: data mining and optimization for decision making. John Wiley & Sons; 2011 Aug 10.
3. Bentley D. Business Intelligence and Analytics. Internet, link: https://www. pdfdrive. com/business-intelligence-and-analytics-e56416503. html. 2017.

UNIT 12: MOBILE BUSINESS INTELLIGENCE

Structure

11.0 OBJECTIVES

After studying this unit, you will be able to:

- Analyze Mobile Client Application
- Differentiate Web Applications vs. Device-specific Applications for Mobile BI
- Examine Security Considerations for Mobile BI Apps
- Analyze Real-Time Business Intelligence

12.1 INTRODUCTION

Mobile Business Intelligence (Mobile BI or Mobile Intelligence) is defined as "The capability that enables the mobile workforce to gain business insights through information analysis using applications optimized for mobile devices" Verkooij(2012) Business intelligence (BI) refers to computer-based techniques used in spotting, digging-out, and analyzing business data, such as sales revenue by products and/or departments or associated costs and incomes.

Although the concept of mobile computing has been prevalent for over a decade, Mobile BI has shown a momentum/growth only very recently. This change has been partly encouraged by a change from the 'wired world' to a wireless world with the advantage of smartphones which has led to a new era of mobile computing, especially in the field of BI.

According to the Aberdeen Group, a large number of companies are rapidly undertaking mobile BI owing to a large number of market pressures such as the need for higher efficiency in business processes, improvement in employee productivity (e.g., time spent looking for information), better and faster decision making, better customer service, and delivery of real-time bi-directional data access to make decisions anytime and anywhere. But despite the apparent advantages of mobile information delivery, mobile BI is still in the 'early adopter' phase. Some CFOs remain sceptical of the business benefits and with the perceived lack of specific business use cases and tangible ROI, mobile BI adoption is still behind the curve compared with other enterprise mobile applications.

12.2 HISTORY

Information Delivery to Mobile Devices

The predominant method for accessing BI information is using proprietary software or a Web browser on a personal computer to connect to BI applications. These BI applications request data from databases. Starting in the late 1990s, BI systems offered alternatives for receiving data, including email and mobile devices

Static Data Push

Initially, mobile devices such as pagers and mobile phones received pushed data using a short message service (SMS) or text messages. These applications were designed for specific mobile devices, contained minimal amounts of information, and provided no data interactivity. As a result, the early mobile BI applications were expensive to design and maintain while providing limited informational value, and garnered little interest.

Data Access Via a Mobile Browser

The mobile browser on a smartphone, a handheld computer integrated with a mobile phone, provided a means to read simple tables of data. The small screen space, immature mobile browsers, and slow data transmission could not provide a satisfactory BI experience. Accessibility and bandwidth may be perceived as issues when it comes to mobile technology, but BI solutions provide advanced functionality to predict and outperform such potential challenges.

While Web-based mobile BI solutions provide little to no control over the processing of data in a network, managed BI solutions for mobile devices only utilize the server for specific operations. In addition, local reports are compressed both during transmission and on the device, permitting greater flexibility for storage and receipt of these reports. Within a mobile environment, users capitalize on easy access to information because the mobile application operates within a single authoring environment that permits access to all BI content (respecting existing security) regardless of language or locale. Furthermore, the user will not need to build and maintain a separate mobile BI deployment. In addition, mobile BI requires much less bandwidth for functionality. Mobile BI promises a small report footprint on memory, encryption during transmission as well as on the device, and compressed data storage for offline viewing and use.

12.3 MOBILE CLIENT APPLICATION

In 2002, Research in Motion released the first BlackBerry smartphone optimized for wireless email use. Wireless e-mail proved to be the “killer app” that accelerated the popularity of the smartphone market. By the mid-2000s, Research in Motion’s BlackBerry had solidified its hold on the smartphone market with both corporate and governmental organizations. The BlackBerry smartphones eliminated the obstacles to mobile business intelligence. The BlackBerry offered a consistent treatment of data across its many models, provided a much larger screen for viewing data, and allowed user interactivity via the thumbwheel and keyboard. BI vendors re-entered the market with offerings spanning different mobile operating systems (BlackBerry, Windows, Symbian) and data access methods. The two most popular data access options were:

- to use the mobile browser to access data, similar to desktop computer, and
- to create a native application designed specifically for the mobile device. Research in Motion is continuing to lose market share to Apple and Android smartphones. In the first three months of 2011 Google’s Android OS gained 7 points of market share. During the same time period RIM’s market share collapsed and dropped almost 5 points.

Purpose-built Mobile BI Apps Apple quickly set the standard for mobile devices with the introduction of the iPhone. In the first three years, Apple sold over 33.75 million units. Similarly, in 2010, Apple sold over 1 million iPads in just under three months. Both devices feature an interactive touchscreen display that is the de facto standard on many mobile phones and tablet computers.

In 2008, Apple published the SDK for which developers can build applications that run natively on the iPhone and iPad instead of Safari-based applications. These native applications can give the user a robust, easier-to-read and easier-to-navigate experience.

Others were quick to join in the success of mobile devices and app downloads. The Google Play Store now has over 700,000 apps available for the mobile devices running the Android operating system.

More importantly, the advent of the mobile device has radically changed the way people use data on their mobile devices. This includes mobile BI. Business intelligence applications can be used to transform reports and data into mobile dashboards, and have them instantly delivered to any mobile device.

Google Inc.'s Android has overtaken Apple Inc.'s iOS in the wildly growing arena of app downloads. In the second quarter of 2011, 44% of all apps downloaded from app marketplaces across the web were for Android devices while 31% were for Apple devices, according to new data from ABI Research. The remaining apps were for various other mobile operating systems, including BlackBerry and Windows Phone 7.

Mobile BI applications have evolved from being a client application for viewing data to a purpose-built application designed to provide information and workflows necessary to quickly make business decisions and take action.

12.4 WEB APPLICATIONS VS. DEVICE-SPECIFIC APPLICATIONS FOR MOBILE BI

In early 2011, as the mobile BI software market started to mature and adoption started to grow at a significant pace in both small and large enterprises, most vendors adopted either a purpose-built, device-specific application strategy (e.g. iPhone or Android apps, downloaded from iTunes or the Google Play Store) or a web application strategy (browser-based, works on most devices without an application being installed on the device). This debate continues and there are benefits and drawbacks to both methods. One potential solution will be the wider adoption of HTML5 on mobile devices which will give web applications many of the characteristics of dedicated applications while still allowing them to work on many devices without an installed application.

Microsoft has announced their mobile BI strategy. Microsoft plans to support browser-based applications such as Reporting Services and PerformancePoint on iOS in the first half of 2012 and touch-based applications on iOS and Android by the second half of 2012. Despite popular perception that Microsoft only acknowledges its own existence, recent moves suggest the company is aware that it is not the only player in the technology ecosystem. Instead of attempting to squelch competition or suggesting new technology developments were ridiculous, the company has instead decided to make its technology accessible to a wider audience.

There are many mobile devices and platforms available today. The list is constantly growing and so is the platform support. There are hundreds of models available today, with multiple hardware and software combinations. The enterprise must select a device very carefully. The target devices will impact the mobile BI design itself because the design for a smartphone will be different than for a tablet. The screen size, processor, memory, etc. all vary. The mobile BI program must account for lack of device standardization from the providers by constantly testing devices for the mobile BI apps. Some best practices can always be followed. For example, a smartphone is a good candidate for operational mobile BI. However, for analytics and what-if analysis, tablets are the best option. Hence, the selection or availability of the device plays a big role in the implementation.

Demand Gartner analyst Ted Friedman believes that mobile delivery of BI is all about practical, tactical information needed to make immediate decisions – "The biggest value is in operational BI — information in the context of applications — not in pushing lots of data to somebody's phone."

Accessing the Internet through a mobile device such as a smartphone is also known as the mobile Internet or mobile Web. IDC expects the US mobile workforce to increase by 73% in 2011. Morgan Stanley reports the mobile Internet is ramping up faster than its predecessor, the desktop Internet, enabling companies to deliver knowledge to their mobile workforce to help them make more profitable decisions.

Michael Cooney from Gartner has identified bring-your-own-technology at work as becoming the norm, not the exception. By 2015 media tablet shipments will reach around 50% of laptop shipments and Windows 8 will likely be in third place behind Android and Apple. The net result is that Microsoft's share of the client platform, be it PC, tablet or smartphone, will likely be reduced to 60% and it could fall below 50%.

Business Benefits

In its latest Magic Quadrant for Business Intelligence Platforms, Gartner examines whether the platform enables users to "fully interact with BI content delivered to mobile devices." The phrase "fully interact" is the key. The ability to send alerts embedded in email or text messages, or links to static content in email messages hardly represents sophistication in mobile analytics. For users to benefit from mobile BI, they must be able to navigate dashboard and guided analytics comfortably—or as comfortably as the mobile device will allow, which is where devices with high-resolution screens and touch interfaces (like the iPhone and Android-based phones) have a clear edge over, say, earlier editions of BlackBerry. It is equally important to take a step back to define your purpose and adoption patterns. Which business users can benefit the most from mobile analytics—and what, exactly, is their requirement? You don't need mobile analytics to send a few alerts or summary reports to their handhelds—without interactivity, mobile BI is indistinguishable from merely informative email or text messages.

Applications

Similar to consumer applications, which have shown an ever increasing growth over the past few years, a constant demand for anytime, anywhere access to BI is leading to a number of custom mobile application development. Businesses have also started adopting mobile solutions for their workforce and are soon becoming key components of core business processes. In an Aberdeen survey conducted in May 2010, 23% of companies participating indicated that they now have a mobile BI app or dashboard in place, while another 31% indicated that they plan to implement some form of mobile BI in the next year.

Definitions

Mobile BI applications can be defined/segregated as follows:

- Mobile Browser Rendered App: Almost any mobile device enables Web-based, thin client, HTML-only BI applications. However, these apps are static and provide little data interactivity. Data is viewed just as it would be over a browser from a personal computer. Little additional effort is required to display data but mobile browsers can typically only support a small subset of the interactivity of a web browser.
- Customized App: A step up from this approach is to render each (or all) reports and dashboards in device-specific format. In other words, provide information specific to

the screen size, optimize usage of screen real estate, and enable device-specific navigation controls. Examples of these include thumb wheel or thumb button for BlackBerry, up/down/left/ right arrows for Palm, gestural manipulation for iPhone. This approach requires more effort than the previous but no additional software.

- Mobile Client App: The most advanced, the client app provides full interactivity with the BI content viewed on the device. In addition, this approach provides periodic caching of data which can be viewed and analyzed even offline.

Companies across all verticals, from retail to even non-profit organizations are realizing the value of purpose-specific mobile applications suited for their mobile workforce.

Development

Developing a native mobile BI app poses challenges, especially concerning data display rendering and user interactivity. Mobile BI App development has traditionally been a time-consuming and expensive effort requiring businesses to justify the investment for the mobile workforce. They do not only require texting and alerts, they need information customized for their line of work which they can interact with and analyze to gain deeper information.

Custom-coded Mobile BI Apps

Mobile BI applications are often custom-coded apps specific to the underlying mobile operating system. For example, the iPhone apps require coding in Objective-C while Android apps require coding in Java. In addition to the user functionality of the app, the app must be coded to work with the supporting server infrastructure required to serve data to the mobile BI app. While custom-coded apps offer near limitless options, the specialized software coding expertise and infrastructure can be expensive to develop, modify, and maintain.

Fixed-form Mobile BI Applications

Business data can be displayed in a mobile BI client (or web browser) that serves as a user interface to existing BI platforms or other data sources, eliminating the need for new master sources of data and specialized server infrastructure. This option offers fixed and configurable data visualizations such as charts, tables, trends, KPIs, and links, and can usually be deployed quickly using existing data sources. However, the data visualizations are not limitless and cannot always be extended to beyond what is available from the vendor.

Graphical Tool-developed Mobile BI Apps

Mobile BI apps can also be developed using the graphical, drag-and-drop development environments of BI platforms. The advantages including the following:

1. Apps can be developed without coding,
2. Apps can be easily modified and maintained using the BI platform change management tools,
3. Apps can use any range of data visualizations and not be limited to just a few,
4. Apps can incorporate specific business workflows, and
5. The BI platform provides the server infrastructure.

Using graphical BI development tools can allow faster mobile BI app development when a custom application is required.

12.5 SECURITY CONSIDERATIONS FOR MOBILE BI APPS

High adoption rates and reliance on mobile devices makes safe mobile computing a critical concern. The Mobile Business Intelligence Market Study discovered that security is the number one issue (63%) for organizations.

A comprehensive mobile security solution must provide security at these levels:

- Device
- Transmission
- Authorization, Authentication, and Network Security

Device Security

A senior analyst at the Burton Group research firm recommends that the best way to ensure data will not be tampered with is to not store it on the client device (mobile device). As such, there is no local copy to lose if the mobile device is stolen and the data can reside on servers within the data center with access permitted only over the network. Most smartphone manufacturers provide a complete set of security features including full-disk encryption, email encryption, as well as remote management which includes the ability to wipe contents if device is lost or stolen. Also, some devices have embedded third-party antivirus and firewall software such as RIM’s BlackBerry.

Transmission Security

Transmission security refers to measures that are designed to protect data from unauthorized interception, traffic analysis, and imitative deception. These measures include Secure Sockets Layer (SSL), iSeries Access for Windows, and virtual private network (VPN) connections. A secure data transmission should enable the identity of the sender and receiver to be verified by using a cryptographic shared key system as well as protect the data to be modified by a third party when it crosses the network. This can be done using AES or Triple DES with an encrypted SSL tunnel.

Authorization, Authentication, and Network Security

Authorization refers to the act of specifying access rights to control access of information to users. Authentication refers to the act of establishing or confirming the user as true or authentic. Network security refers to all the provisions and policies adopted by the network administrator to prevent and monitor unauthorized access, misuse, modification, or denial of the computer network and network-accessible resources. The mobility adds to unique security challenges. As data is trafficked beyond the enterprise firewall towards unknown territories, ensuring that it is handled safely is of paramount importance. Towards this, proper authentication of user connections, centralized access control (like LDAP Directory), encrypted data transfer mechanisms can be implemented.

Role of BI for Securing Mobile Apps

To ensure high security standards, BI software platforms must extend the authentication options and policy controls to the mobile platform. Business intelligence software platforms need to ensure a secure encrypted keychain for storage of credentials. Administrative control of password policies should allow creation of security profiles for each user and seamless integration with centralized security directories to reduce administration and maintenance of users

Products

A number of BI vendors and niche software vendors offer mobile BI solutions. Some notable examples include:

- CollabMobile
- Cognos

- Cherrywork
- Dimensional Insight
- InetSoft
- Infor
- Information Builders
- MicroStrategy
- QlikView
- Roambi
- SAP
- Tableau Software
- Sisense
- TARGIT Business Intelligence

12.6 REAL-TIME BUSINESS INTELLIGENCE

Real-time business intelligence (RTBI) is the process of delivering business intelligence (BI) or information about [business operations] as they occur. Real time means near to zero latency and access to information whenever it is required.

The speed of today's processing systems has moved classical data warehousing into the realm of real-time. The result is real-time business intelligence. Business transactions as they occur are fed to a real-time BI system that maintains the current state of the enterprise. The RTBI system not only supports the classic strategic functions of data warehousing for deriving information and knowledge from past enterprise activity, but it also provides real-time tactical support to drive enterprise actions that react immediately to events as they occur. As such, it replaces both the classic data warehouse and the enterprise application integration (EAI) functions. Such event-driven processing is a basic tenet of real-time business intelligence.

In this context, "real-time" means a range from milliseconds to a few seconds (5s) after the business event has occurred. While traditional BI presents historical data for manual analysis, RTBI compares current business events with historical patterns to detect problems or opportunities automatically. This automated analysis capability enables corrective actions to be initiated and/or business rules to be adjusted to optimize business processes.

RTBI is an approach in which up-to-a-minute data is analyzed, either directly from Operational sources or feeding business transactions into a real time data warehouse and Business Intelligence system. RTBI analyzes real time data.

Real-time business intelligence makes sense for some applications but not for others – a fact that organizations need to take into account as they consider investments in real-time BI tools. Key to deciding whether a real-time BI strategy would pay dividends is understanding the needs of the business and determining whether end users require immediate access to data for analytical purposes, or if something less than real time is fast enough

Architectures

Event-based

Real-time Business Intelligence systems are event driven, and may use Complex Event Processing, Event Stream Processing and Mashup (web application hybrid) techniques to enable events to be analysed without being first transformed and stored in a database. These in- memory techniques have the advantage that high rates of events can be monitored, and since data does not have to be written into databases data latency can be reduced to milliseconds.

Data Warehouse

An alternative approach to event driven architectures is to increase the refresh cycle of an existing data warehouse to update the data more frequently. These real-time data warehouse systems can achieve near real-time update of data, where the data latency typically is in the range from minutes to hours. The analysis of the data is still usually manual, so the total latency is significantly different from event driven architectural approaches.

Server-less Technology

The latest alternative innovation to “real-time” event driven and/or “real-time” data warehouse architectures is MSSO Technology (Multiple Source Simple Output) which removes the need for the data warehouse and intermediary servers altogether since it is able to access live data directly from the source (even from multiple, disparate sources). Because live data is accessed directly by server-less means, it provides the potential for zero-latency, real-time data in the truest sense.

Process-aware

This is sometimes considered a subset of Operational intelligence and is also identified with Business Activity Monitoring. It allows entire processes (transactions, steps) to be monitored, metrics (latency, completion/failed ratios, etc.) to be viewed, compared with warehoused historic data, and trended in real-time. Advanced implementations allow threshold detection, alerting and providing feedback to the process execution systems themselves, thereby 'closing the loop'.

Technologies that Support Real-time Analytics

Technologies that can be supported to enable real-time business intelligence are data visualization, data federation, enterprise information integration, enterprise application integration and service oriented architecture. Complex event processing tools can be used to analyze data streams in real time and either trigger automated actions or alert workers to patterns and trends.

Data Warehouse Appliance

Data warehouse appliance is a combination of hardware and software product which was designed exclusively for analytical processing. In data warehouse implementation, tasks that involve tuning, adding or editing structure around the data, data migration from other databases, reconciliation of data are done by DBA. Another task for DBA was to make the database to perform well for large sets of users. Whereas with data warehouse appliances, it is the vendor responsibility of the physical design and tuning the software as per hardware requirements. Data warehouse appliance package comes with its own operating system, storage, DBMS, software, and required hardware. If required data warehouse appliances can be easily integrated with other tools.

There are very limited vendors for providing Mobile business intelligence; MBI is integrated with existing BI architecture. MBI is a package that uses existing BI applications so people can use on their mobile phone and make informed decision in real time.

Application Areas

- Algorithmic trading
- Fraud detection
- Systems monitoring

- Application performance monitoring
- Customer Relationship Management
- Demand sensing
- Dynamic pricing and yield management
- Data validation
- Operational intelligence and risk management
- Payments & cash monitoring
- Data security monitoring
- Supply chain optimization
- RFID/sensor network data analysis
- Work streaming
- Call center optimization
- Enterprise Mashups and Mashup Dashboards

Transportation industry

Transportation industry can be benefited by using real-time analytics. For an example railroad network. Depending on the results provided by the real-time analytics, dispatcher can make a decision on what kind of train he can dispatch on the track depending on the train traffic and commodities shipped.

12.7 CHECK YOUR PROGRESS

1. Define Authorization
2. What is Authentication?
3. Write the role of Network Security.
4. Administrative control of password policies should allow creation of security profiles for each user and seamless integration with centralized security directories to reduce administration and maintenance of users (True/False)
5. What is Real-time Business Intelligence?

Answers to Check your progress

1. Authorization refers to the act of specifying access rights to control access of information to users.
2. Authentication refers to the act of establishing or confirming the user as true or authentic.

3. Network security refers to all the provisions and policies adopted by the network administrator to prevent and monitor unauthorized access, misuse, modification, or denial of the computer network and network-accessible resources.
4. True
5. Real-time business intelligence (RTBI) is the process of delivering business intelligence (BI) or information about [business operations] as they occur. Real time means near to zero latency and access to information whenever it is required.

12.8 SUMMARY

A brief introduction on Mobile BI Apps has been given in this unit. The evolution of Mobile BI Apps has been discussed. The security measures on mobile apps has been elaborated. Significance of Real-time Business Intelligence has been given. Also the architectures of Real-time Business Intelligence has been detailed considering different domains.

12.9 KEYWORDS

- **Business intelligence software**: is a type of application software designed to retrieve, analyze, transform and report data for business intelligence. The applications generally read data that has been previously stored, often - though not necessarily - in a data warehouse or data mart.
- **Demand sensing**: is a software solution that leverages advanced technologies such as artificial intelligence (AI), machine learning and real-time data capturing to close the gap between your demand planning and what is happening in your supply chain.
- **Real-time analytics** is the discipline that applies logic and mathematics to data to provide insights for making better decisions quickly. For some use cases, real time simply means the analytics is completed within a few seconds or minutes after the arrival of new data.
- **Event processing:** is computing that performs operations on events as they are reported in a system that observes or listens to the events from the environment. Common information processing operations include reading, creating, transforming, and processing events.
- **Operational intelligence** (OI) is an approach to data analysis that enables decisions and actions in business operations to be based on real-time data as it's generated or collected by companies.

12.10 QUESTIONS FOR SELF STUDY

1. Write the advantages of Mobile BI Apps.
2. How security is maintained in Mobile? Explain.
3. Write the significance of Real-time Business Intelligence.
4. Discuss segregation of Mobile BI applications
5. Explain the architectures of Real-time Business Intelligence.
6. Write a short note on Data Warehouse Appliance.

12.11 REFERENCES

1. Sharda R, Delen D, Turban E. Business intelligence analytics and data science: A managerial perspective. Pearson, 2022.
2. Vercellis C. Business intelligence: data mining and optimization for decision making. John Wiley & Sons; 2011 Aug 10.
3. Bentley D. Business Intelligence and Analytics. Internet, link: https://www. pdfdrive. com/business-intelligence-and-analytics-e5

Karnataka State Open University

Mukthagangothri, Mysore – 570 006.

Dept. of Studies and Research in Management

MBA IT Specialization

III Semester

Business Intelligence and Analytics

Block 4

Karnataka State Open University

Mukthagangothri, Mysore – 570 006.

Dept. of Studies and Research in Management

MBA. IT Specialization

III Semester

Business Intelligence and Analytics

BLOCK 4

BLOCK 4: INTRODUCTION

The purpose of unit 13 is to introduce you to the different products that make up the Tableau application suite, the Tableau user interface, and to how Tableau processes your data. This unit provides a first glimpse of the possibilities that Tableau gives you for creating data visualizations. unit 14 we understand how to connect data residing in different places. Your data is scattered over multiple databases, text files, spreadsheets, and public services. Connecting to a wide variety of data sources directly, Tableau makes it much easier to analyze data residing in different places. You can analyze spreadsheets, public data tools, analytic databases, Hadoop, and a large variety of general-purpose databases as well as data cubes. In unit 15 you can start building visualizations. In this unit you will learn about all of the chart types provided by the Show Me button. You will discover how to add trend lines, reference lines, and control the way your data is sorted and filtered. You'll see how creating ad hoc groups, sets, and hierarchies can produce information not available in the data source. Tableau's discrete and continuous data hierarchies will be explained, and how you can alter Tableau's default date hierarchies by creating your own custom dates. In unit 16, you will learn how to use calculated values and table calculations to derive facts and dimensions that don't exist in your source data. Tableau's Formula Editing window will be explained as well as the Quick Table Calculation menu, and how to modify Quick Table defaults to address your specific needs. This block consists of 4 units and is organized as follows:

Unit 13- Introduction to Tableau:

Overview of Visual Data Analytics, The Tableau Suite, Installing the Tableau Desktop, Tableau Desktop workspace

Unit 14- Connecting your Data:

How to Connect to Your Data, What Are Generated Values?, Knowing When to Use a Direct Connection or a Data Extract, Joining Database Tables with Tableau, Blending Different Data sources in a Single Worksheet, How to Deal with Data Quality Problems .

Unit 15- Data Visualization:

Fast and Easy Analysis via Show Me, How Show Me Works, Trend Lines and Reference Lines, Sorting Data in Tableau, Enhancing Views with Filters, Sets, Groups, and Hierarchies.

Unit 16- : Calculations with Tableau:

What is Aggregation?, What Are Calculated Values and Table Calculations?, Using the Calculation Dialog Box to Create, Building Formulas Using Table Calculations, Using Table Calculation Functions, Adding Flexibility to Calculations with Parameters, Using the Function Reference Appendix

UNIT -13: Introduction to Tableau

Structure

13.0 OBJECTIVES

After studying this unit, you will be able to :

- ✓ Importance of visual data analytics
- ✓ Features of the Tableau software
- ✓ How to install the Tableau desktop
- ✓ How to create Tableau desktop workspace

3.1 Introduction

Tableau Software started ten years ago as a desktop application, but as the tool has matured it has become popular in the enterprise and is being used in "Big Data" environments. The enterprise means any type of significant organization—a global business or non-profit, such as a large university, small college or hospitals, banks, retailers, or internet-based data companies that have accumulated massive data sets. Or, this might refer to a small business with only a few employees. A short list of the types of organizations using Tableau should include:

- Multi-national financial institutions
- Federal Government
- International police organizations

- The military
- Government intelligence organizations
- Media companies
- Financial institutions
- Hospitals
- Book publishers
- Internet-based business (with Big Data)
- Insurance companies
- Non-profit entities
- Manufacturing and Distribution companies
- Education (Universities, Colleges, Charter Schools, Public Schools)
- Law firms
- Consulting firms
- Retailers
- Consumer product companies
- Accounting firms
- Consulting firm

Any entity that needs to see and understand data is a candidate for using Tableau Software. Tableau does a good job listening to their customers and partners. They've improved the speed, security, and added more visualization types to Tableau's capabilities. Today, many large enterprises use Tableau because they find it increases user adoption rates. It also allows business users to create their own reports with relative ease—reducing the report backlog that accumulates within information technology departments. Smaller enterprises are using Tableau because it provides a low-cost way to turn data into useful information.

The seeds for Tableau were planted in the early 1970s when IBM invented Structured Query Language (SQL) and later in 1981 when the spreadsheet became the killer application of the personal computer. Data creation and analysis fundamentally changed for the better. Our ability to create, and store data increased exponentially.

The business information (BI) industry was created with this wave; each vendor providing a product "stack" based on some variant of SQL. The pioneering companies invented foundational technologies and developed sound methods for collecting and storing data. Recently, a new generation of NOSQL2 (Not Only SQL) databases are enabling web properties like Facebook to mine massive, multi-petabyte3 data streams.

Deploying these systems can take years. Data today resides in many different proprietary databases and may also need to be collected from external sources. The traditional leaders in the BI industry have created reporting tools that focus on rendering data from their proprietary products. Performing analysis and building reports with these tools requires technical expertise and time. The people with the technical chops to master them are product specialists that don't always know the best way to present the information.

The scale, velocity, and scope of data today demands reporting tools that deploy quickly. They must be suitable for non-technical users to master. They should connect to a wide variety of data sources. And, the tools need to guide us to use the best techniques known for rendering the data into information.

The Business Case for Visual Analysis Whether the entity seeks profits or engages in non-profit activities, all enterprises use data to monitor operations and perform analysis. Insights gleaned from the reports and analysis are then used to maintain efficiency, pursue opportunity, and prevent negative outcomes. Supporting this infrastructure (from the perspective of the information consumer) are three kinds of data.

Three Kinds of Data that Exist in Every Entity

Reports, analysis, and ad hoc discovery are used to express three basics kinds of data.

Known Data (type 1)

Encompassed in daily, weekly, and monthly reports that are used for monitoring activity, these reports provide the basic context used to inform discussion and frame questions. Type 1 reports aren't intended to answer questions. Their purpose is to provide visibility of operations.

Data You Know You Need to Know (type 2)

Once patterns and outliers emerge in type 1 data the question that naturally follows is: why is this happening? People need to understand the cause of the outliers so that action can be taken. Traditional reporting tools provide a good framework to answer this type of query as long as the question is anticipated in the design of the report.

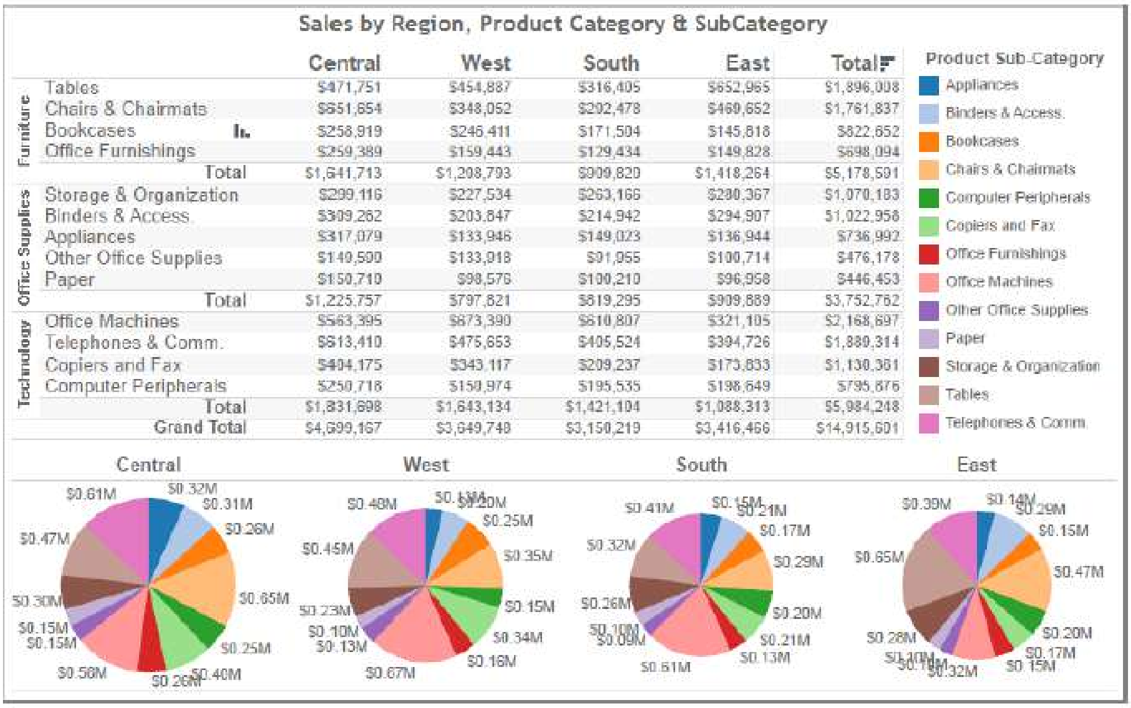
Data You Don't Know You Need to Know (type 3)

By interacting with data in real-time while using appropriate visual analytics, Tableau provides the possibility of seeing patterns and outliers that are not visible in type 1 and type 2 reports. The process of interacting with granular data yields different questions that can lead to new actionable insights. Software that enables quick-iterative analysis and reporting is becoming a necessary element of effective business information systems.

Distributing type 1 reports in a timely manner is important, but speed in the design and build stage of type 1 reports are also important when a new Type 1 report is created. To effectively enable Type 2 and 3 analyses the reporting tool must adapt quickly to ad hoc queries and present the data in intuitively understandable ways.

13.1 OVERVIEW OF VISUAL DATA ANALYTICS FEATURES

Rendering data accurately with appropriate visual analytics reduces the time required to achieve understanding. Review the following examples to see how visual analytics can reduce the time to insight. The goal of these reports is to provide sales analysis by region, product category, and product sub-category. Figure 13-1 presents data using a grid of numbers (crosstab) and pie charts. Crosstabs are useful for finding specific values. Pie Charts are intended to show one-to-many comparisons of dimensions. The pie charts compare sales by product sub-category.



Sales by Region, Product Category & SubCategory

		Central	West	South	East	Total
Furniture	Tables	$471,751	$454,887	$316,405	$652,965	$1,896,008
	Chairs & Chairmats	$551,654	$348,052	$292,478	$469,652	$1,761,837
	Bookcases	$258,919	$246,411	$171,504	$145,818	$822,652
	Office Furnishings	$259,389	$159,443	$129,434	$149,828	$698,094
	Total	$1,641,713	$1,208,793	$909,820	$1,418,264	$5,178,591
Office Supplies	Storage & Organization	$299,116	$227,534	$263,166	$280,367	$1,070,183
	Binders & Access.	$309,262	$203,847	$214,942	$294,907	$1,022,958
	Appliances	$317,079	$133,946	$149,023	$136,944	$736,992
	Other Office Supplies	$149,590	$133,918	$91,955	$100,714	$476,178
	Paper	$150,710	$98,576	$100,210	$96,958	$446,453
	Total	$1,225,757	$797,821	$819,295	$909,889	$3,752,762
Technology	Office Machines	$563,395	$673,390	$610,807	$321,105	$2,168,697
	Telephones & Comm.	$613,410	$475,653	$405,524	$394,726	$1,889,314
	Copiers and Fax	$404,175	$343,117	$209,237	$173,833	$1,130,361
	Computer Peripherals	$250,718	$159,974	$195,535	$198,649	$795,876
	Total	$1,831,698	$1,643,134	$1,421,104	$1,088,313	$5,984,248
	Grand Total	$4,699,167	$3,649,748	$3,150,219	$3,416,466	$14,915,601

Fig.13.1 *Sales Mix Analysis using a crosstab and pie charts*

Crosstabs are not the most effective way to make one-to-many comparisons or identify outliers. Pie charts are commonly used for comparisons but are one of the least effective ways to compare values across dimensions. It is difficult to make precise comparisons especially between slices, and even more so when there are many slices.

Figure 13-2 employs a bar chart and heat map to convey the same information. Bar charts provide a better means for comparing product sub-categories. The heat map on the right provides total sales for each category. The gray scale color range highlights the high and low selling product sub-categories. The color encoding in the bar chart provides additional information on profit ratio. Reference lines in the bar chart display the average sales for all product subcategories within each region.

Clearly the bar chart and heat map communicate the sales values more quickly while adding profit ratio information with the use of color. The reference lines within each region and product category provide average sales values. One could argue that the bar chart doesn’t communicate the details available in the crosstab, but in Figure 13-3 those details and more are provided via tooltips that pop out when you point your mouse at a mark. Appropriate visual analytics improve decision-making by making it easier to see summary trends and outliers without sacrificing desired details by making those details available on demand.

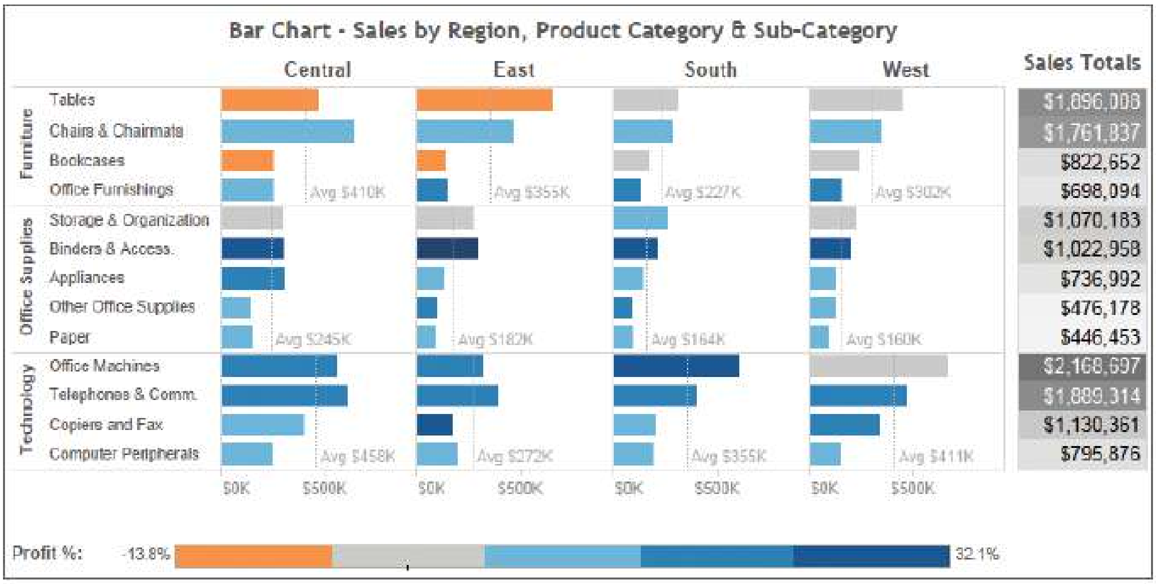


Fig.13.2 *Sales Mix Analysis using a bar chart and heat map*

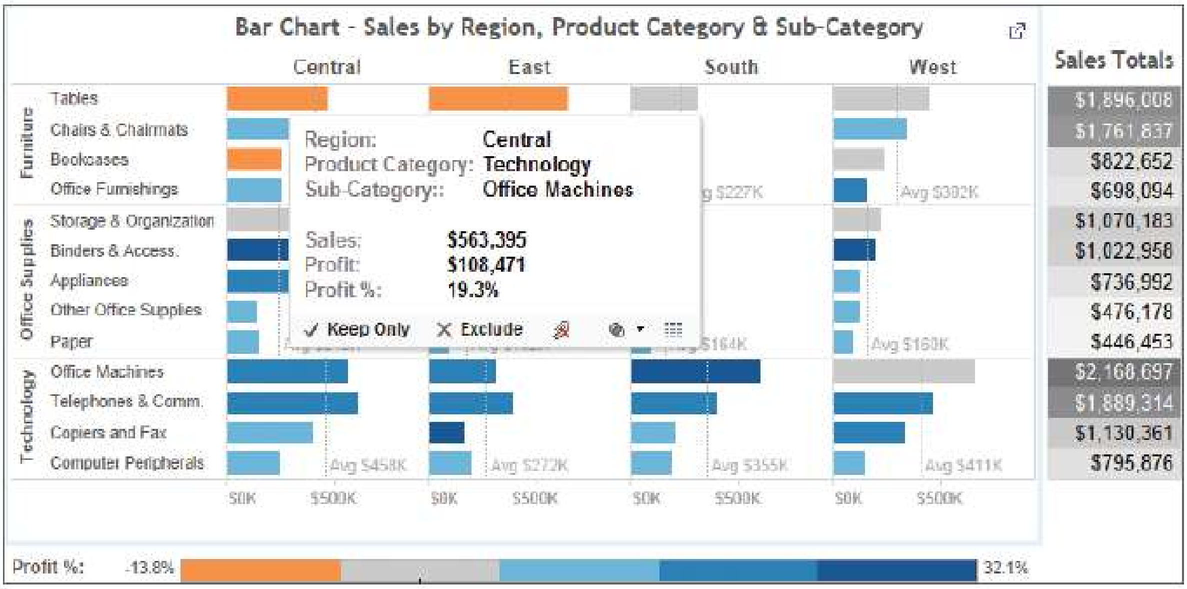


Fig. 13.3 *Adding labels and tooltips*

Turning Data into Information with Visual Analytics

Data that is overly summarized loses its ability to inform. When it’s too detailed, rapid interpretation of the data is compromised. Visual analytics bridges this gap by providing the right style of data visualization and detail for the situational need. The ideal analysis and reporting tool should possess the following attributes:

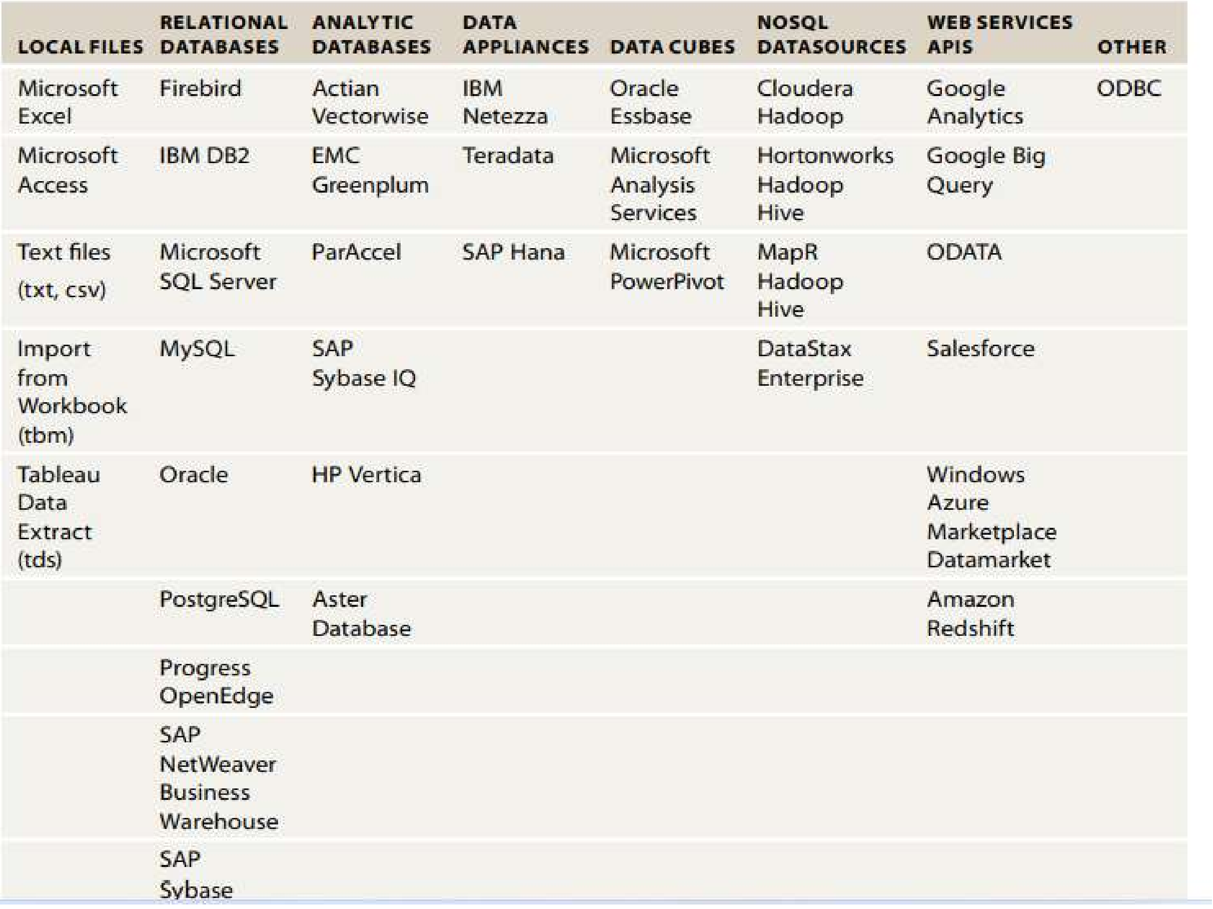
- Simplicity—Be easy for non-technical users to master.
- Connectivity—Seamlessly connect to a large variety of data sources.
- Visual Competence—Provide appropriate graphics by default.
- Sharing—Facilitate sharing of insight.
- Scale—Handle large data sets.

Traditional BI reporting solutions aren’t adapted for the variety of data sources available today. Analysis and reporting can’t occur in these tools until the architecture is created within the proprietary product stack. Tableau Software was designed to address these needs.

13.2 THE TABLEAU SUITE

Tableau Desktop Tableau Desktop is an application for Windows and Mac, appreciated by both analysts and business users. In Tableau Desktop, you can connect to flat files (such as Excel and CSV files) and save your workbooks to your local hard drive. To tap into an organization's IT infrastructure, you can also use Tableau Desktop to connect to a host of different database solutions, and you can share your workbooks via Tableau Server or the cloud-based Tableau Online. Table 1-1 displays the available connections arranged by the type of data source. Personal Edition only connects to local files.

Table 1-1 Data sources Accessible to Tableau Desktop



LOCAL FILES	RELATIONAL DATABASES	ANALYTIC DATABASES	DATA APPLIANCES	DATA CUBES	NOSQL DATASOURCES	WEB SERVICES APIS	OTHER
Microsoft Excel	Firebird	Actian Vectorwise	IBM Netezza	Oracle Essbase	Cloudera Hadoop	Google Analytics	ODBC
Microsoft Access	IBM DB2	EMC Greenplum	Teradata	Microsoft Analysis Services	Hortonworks Hadoop Hive	Google Big Query	
Text files (txt, csv)	Microsoft SQL Server	ParAccel	SAP Hana	Microsoft PowerPivot	MapR Hadoop Hive	ODATA	
Import from Workbook (tbm)	MySQL	SAP Sybase IQ			DataStax Enterprise	Salesforce	
Tableau Data Extract (tds)	Oracle	HP Vertica				Windows Azure Marketplace Datamarket	
	PostgreSQL	Aster Database				Amazon Redshift	
	Progress OpenEdge						
	SAP NetWeaver Business Warehouse						
	SAP Sybase						

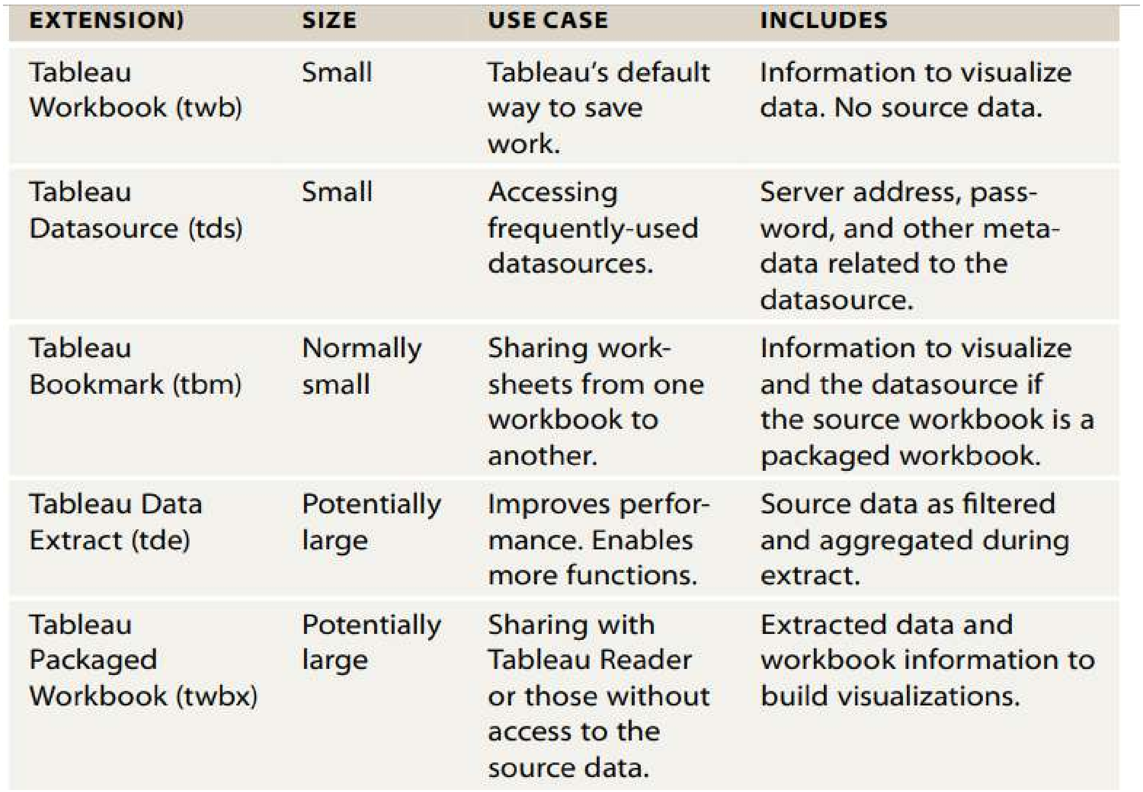
Tableau Reader

Tableau also permits you to share content with another desktop tool. Tableau Reader is a free version that allows users to consume Tableau Desktop reports without the need for a paid license. The only requirement is that the Tableau report be saved as a packaged workbook.

File Types

You can save and share data using a variety of different file types. The differences between each file type relates to the amount and type of information being stored in the file. Table 1-2 summarizes different Tableau file types.

Table 1-2 Tableau File Type



EXTENSION)	SIZE	USE CASE	INCLUDES
Tableau Workbook (twb)	Small	Tableau's default way to save work.	Information to visualize data. No source data.
Tableau Datasource (tds)	Small	Accessing frequently-used datasources.	Server address, password, and other metadata related to the datasource.
Tableau Bookmark (tbm)	Normally small	Sharing worksheets from one workbook to another.	Information to visualize and the datasource if the source workbook is a packaged workbook.
Tableau Data Extract (tde)	Potentially large	Improves performance. Enables more functions.	Source data as filtered and aggregated during extract.
Tableau Packaged Workbook (twbx)	Potentially large	Sharing with Tableau Reader or those without access to the source data.	Extracted data and workbook information to build visualizations.

When you save your work in desktop the default save method creates a workbook (twb) file. If you need to share your work with people that don't have a Tableau Desktop license or don't have access to the data source you can save your work as a packaged workbook (twbx) by using the Save As option when saving your file.

Tableau Data sources (tds) are useful when you frequently connect to a particular data source or you have edited the metadata associated with that data source in some way (renaming or grouping fields for example). Using saved data sources reduces the time required to connect to the data.

Tableau Bookmarks (tbm) allow you to share a single worksheet from your workbook with others. To create a bookmark file, access the main file menu window/bookmark/create bookmark option.

Tableau Data Extracts (tde) leverage Tableau’s proprietary data engine. When you create an extract your data is compressed. If your data source is from a file (Excel, Access, text) Data Extracts add formula functions that don’t exist in those sources—including count distinct and median. If you are publishing workbooks via Tableau Server, Data Extracts provide an effective way to separate the analytical load Tableau generates from your source database

Tableau Server

If you produce a large number of workbooks that have to be updated regularly or you have a large number of people consuming your work, Tableau Server will save you time. Server allows people to view and interact with your work via a web browser. Server will also automatically refresh data extracts that have been published to Tableau Server

Tableau Public

Tableau Public is a free hosted web service that can be used to publish Tableau Reports on the web. Commonly used content management systems like WordPress, Tumblr, and Typepad are supported. Tableau’s licensed desktop editions can also publish content to Tableau Public. Tableau also offers a free Public desktop edition for creating and publishing reports.'

Tableau Online Tableau Online is a Tableau-hosted solution for storing and deploying dashboards. It provides similar functionality to Tableau Server but is a cloud-based service. No purchase and maintenance of server hardware is necessary here.

Tableau Public Tableau Public is a hosting service for the publication of data visualizations to the web. It is used by newsrooms and bloggers but also by companies, research institutes, governmental bodies, and non governmental organizations that aim to get their data stories into the public eye. The interactive visualizations can be viewed in the browser directly on the Tableau Public platform, or they can be embedded into blogs and websites.

13.3 INSTALLING THE TABLEAU DESKTOP

Installing Tableau Desktop is a simple process and takes only a few minutes. Therefore, this will be a very brief section.

System Requirements for Tableau Desktop

Before installing Tableau Desktop, be sure your machine meets the necessary requirements for this application. Tableau Desktop is available for Windows and Mac. These are the official minimum requirements for a Windows installation:

- Microsoft Windows 7 or later (64 bit)
- Microsoft Server 2008 R2 or later
- Intel Pentium 4 or AMD Opteron processor or later
- 2 GB RAM
- At least 1.5 GB of free hard disk space
- These are the official minimum requirements for a Mac installation:
- iMac/MacBook 2009 or later
- OS X 10.10 or later
- At least 1.5 GB of free hard disk space

Should you wish to work with large datasets, the following additional specifications are recommended:

- Latest service pack or update for your operating system
- Intel Core i3/i5/i7/i9 or AMD FX processor or later
- At least 8 GB RAM
- Solid state drive (SSD) with at least 20 GB of free space
- Full HD resolution (1920 × 1080 pixels) or higher with 32 bit color depth

Downloading and Installing Tableau Desktop

If you don't already have Tableau Desktop installed on your machine, use this link to download the latest trial version:

https://www.tableau.com/products/desktop.

Make sure you are logged in to your machine as administrator and that you have the rights to install software on the machine. Run the installer as you normally would, given your operating system:

On a Windows Machine Open the setup (EXE) file, and accept any safety prompts from your OS.

On a Mac Open the image (DMG) file, and double-click the installation package (PKG) file to start the installation.

Follow the prompts during the setup process. Changes to the installation path or similar changes usually are not required.

Registering and Activating Tableau Desktop

Once the installation process is completed, open Tableau Desktop. A registration form will appear, which you can use to register and activate your Tableau Desktop installation using the product key.

If you do not have a product key for Tableau Desktop yet, you can test it for free for two full weeks. You will be able to use all the features of the software during this trial period.

13.4 TABLEAU DESKTOP WORKSPACE

Tableau Workspace

Clicking on the far left icon (with three squares) displays the Tableau Worksheet page and exposes the contents of the worksheet tab selected at the bottom of the screen. When you connect to a new data source this is also the default workspace view. Go to the home page and select the Sample-Superstore SalesExcel spread sheet. You just opened a connection to a saved data source and should have a blank worksheet open.

There are many ways you can open a workspace page; for example, if you display Tableau's icon on your desktop and you have a data source displayed on your desktop. Dragging any data source icon and dropping it on the Tableau icon opens Tableau's worksheet page for the selected data source. Keep in mind that you can open as many connections as you want in Tableau by going to the start page or data connection page and selecting a new connection.

Figure 1-7 is worksheet-connected to the Sample-Superstore Sales-Excel dataset used to create scatter plots.

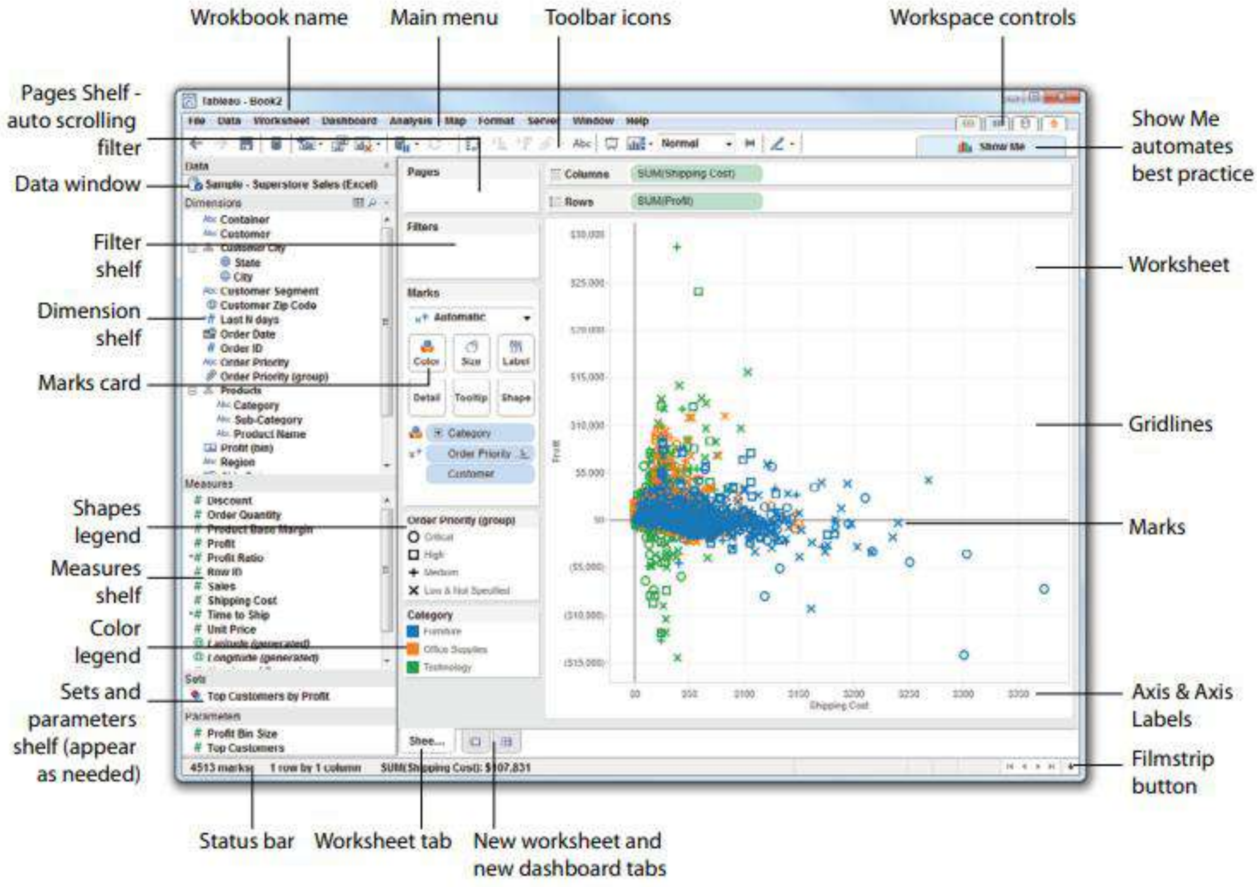
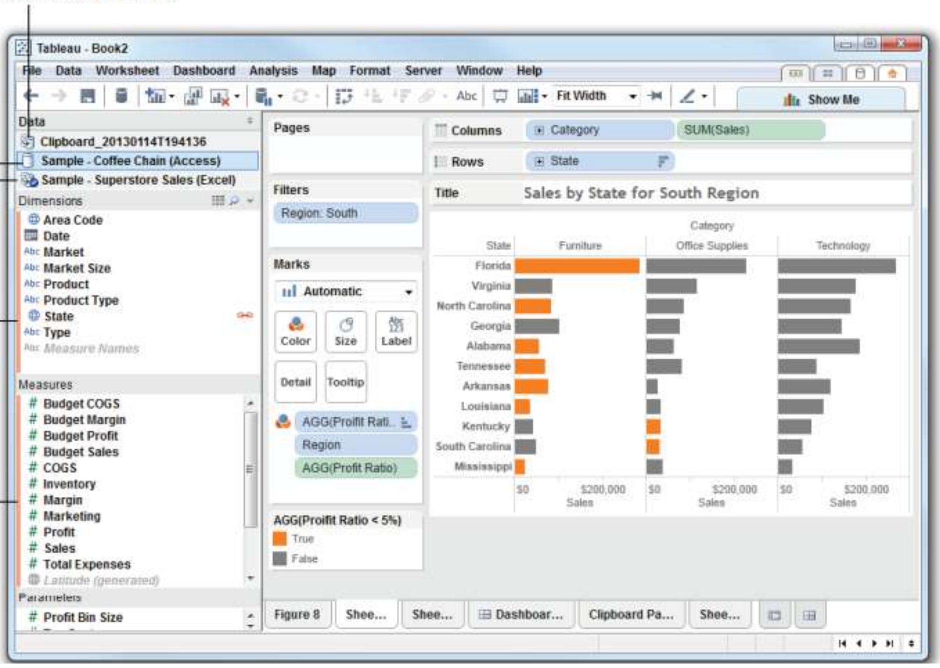


Figure 1-7 Worksheet page

The Data Window, Data Types, and Aggregation

When you connect Tableau to a data source it is expressed in the data window. You can connect to as many different data sources as you want in a single workbook. The small icons associated with data connections provide additional details about the nature of the connection. Figure 1-10 shows a workbook with three different data connections



Two cans with an arrow means the connection is an active data extract.

Blue highlight means the fields from this datasource are being displayed in the dimension and measures shelf.

The blue check mark indicates this data connection is active in this worksheet.

The **Dimensions** shelf contains filtering and sorting elements, text, dates, or geographic fields.

The **Dimensions** shelf contains numbers that you may want to use in formulas.

Figure 1-10 Data shelf

Data Types

Tableau expresses fields and assigns data types automatically. If the data type is assigned by the datasource, Tableau will use that data type. If the datasource doesn't specifically assign a data type, Tableau will assign one. Tableau supports the following data types:

- Text values
- Date values
- Date and time values
- Numerical values
- Geographic values (latitude and longitude used for maps)
- Boolean values (true/false conditions)

Look at Figure 1-10 and focus on the icons next to the fields in the dimension and measures shelves. These icons denote specific data types. Small globes are geographic features;

calendars are dates. A calendar with a clock is a date/time field. Numeric values have pound signs, and text fields are denoted by “abc” icons. Boolean fields have “T/F” icons. Explore Tableau’s manual for more examples.

Aggregation

It is often useful to look at numeric values using different aggregations. Tableau supports many different aggregation types including:

- Sum
- Average
- Median
- Count
- Count Distinct
- Minimum
- Maximum
- Standard Deviation
- Standard Deviation of a Population
- Variance
- Variance of a Population
- Attribute (ATTR)
- Dimension

Adding fields into your visualization results in default aggregations being displayed. Tableau allows you to change the default aggregation or just alter the aggregation level for a specific view. To change the default aggregation, right-click on that field inside the data shelf and change its default by selecting the menu option (default properties/aggregation). You can also change the aggregation of a field for a specific use in a worksheet. Figure 1-11 provides an example. By right-clicking on the SUM (Sales) pill and selecting the Measure (SUM) menu option, you can select any of the aggregations highlighted.

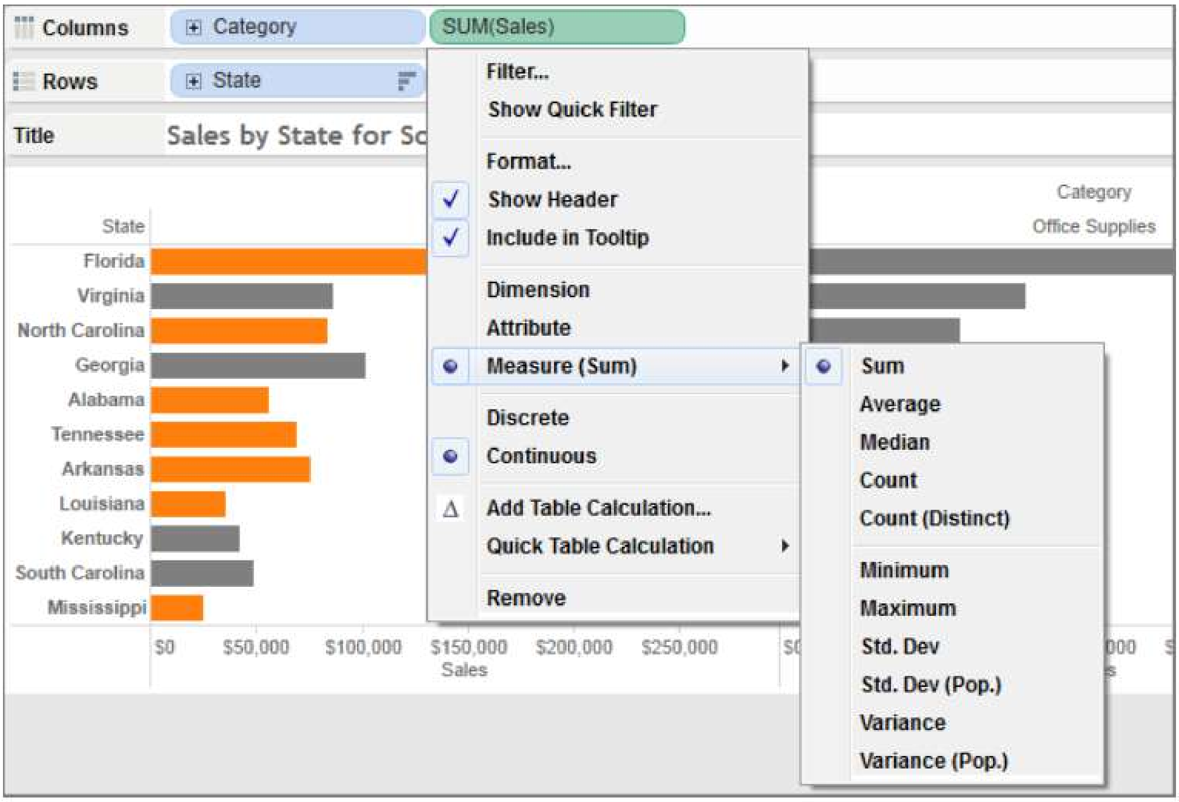


Figure 1-11 Changing aggregation

13.5 CHECK YOUR PROGRESS

1. List the Uses of Show Me options in Tableau.
2. List the Three Essential Tableau concepts
3. Write Three Kinds of Data that Exist in Every Entity.
4. What are the Attributes possessed by the ideal analysis and reporting tool?
5. List the data types supported by Tableau

Answers to Check your progress

1. Efficiency, Inspiration, Inspiration

2. Dimensions and measures Row level, aggregate level, and table level Continuous and discrete

3. Known Data, Data You Know You Need to Know, Data You Don’t Know You Need to Know

4. Simplicity—Be easy for non-technical users to master.

   Connectivity—Seamlessly connect to a large variety of data sources.

   Visual Competence—Provide appropriate graphics by default.

   Sharing—Facilitate sharing of insight.

   Scale—Handle large data sets

5. Text values
   Date values
   Date and time values
   Numerical values
   Geographic values (latitude and longitude used for maps)
   Boolean values (true/false conditions)

13.6 SUMMARY

The seeds for Tableau were planted in the early 1970s when IBM invented Structured Query Language (SQL) and later in 1981 when the spreadsheet became the killer application of the personal computer. Data creation and analysis fundamentally changed for the better. Our ability to create, and store data increased exponentially.

The business information (BI) industry was created with this wave; each vendor providing a product “stack” based on some variant of SQL. The pioneering companies invented foundational technologies and developed sound methods for collecting and storing data. Recently, a new generation of NOSQL2 (Not Only SQL) databases are enabling web properties like Facebook to mine massive, multi-petabyte data streams.

Deploying these systems can take years. Data today resides in many different proprietary databases and may also need to be collected from external sources. The traditional leaders in the BI industry have created reporting tools that focus on rendering data from their proprietary products. Performing analysis and building reports with these tools requires technical expertise and time. The people with the technical chops to master them are product specialists that don’t always know the best way to present the information.

The scale, velocity, and scope of data today demands reporting tools that deploy quickly. They must be suitable for non-technical users to master. They should connect to a wide variety of data sources. And, the tools need to guide us to use the best techniques known for rendering the data into information.

13.7 KEYWORDS

- Tableau Desktop **-** Tableau Desktop is where visualizations are created
- Tableau Server **-** Tableau Server provides a secure, web-based environment where end users can access visualizations created in Desktop either through a browser or via the Tableau Mobile app for Android and iPhone
- Tableau Reader **-** Reader is used for viewing
- The Tableau workspace consists of menus, a toolbar, the Data pane, cards and shelves, and one or more sheets. Sheets can be worksheets, dashboards, or stories

13.8 QUESTIONS FOR SELF-STUDY

1. Explain different aggregation types supports Tableau.
2. Explain how *Show Me* button is used in Tableau
3. Write a note on Visual Data Analytics.
4. Write a brief note importance of aggregate functions in Tableau.
5. Write about the Toolbar icons in Tableau

13.9 REFERENCES

1. Alexander Loth - Visual Analytics with Tableau-Wiley (2019)
2. Dan Murray - Tableau Your Data!_ Fast and Easy Visual Analysis with Tableau Software-Wiley (2013)
3. David Baldwin - Mastering Tableau-Packt Publishing (2017)

UNIT -14: CONNECTING YOUR DATA

Structure

14.0 OBJECTIVES

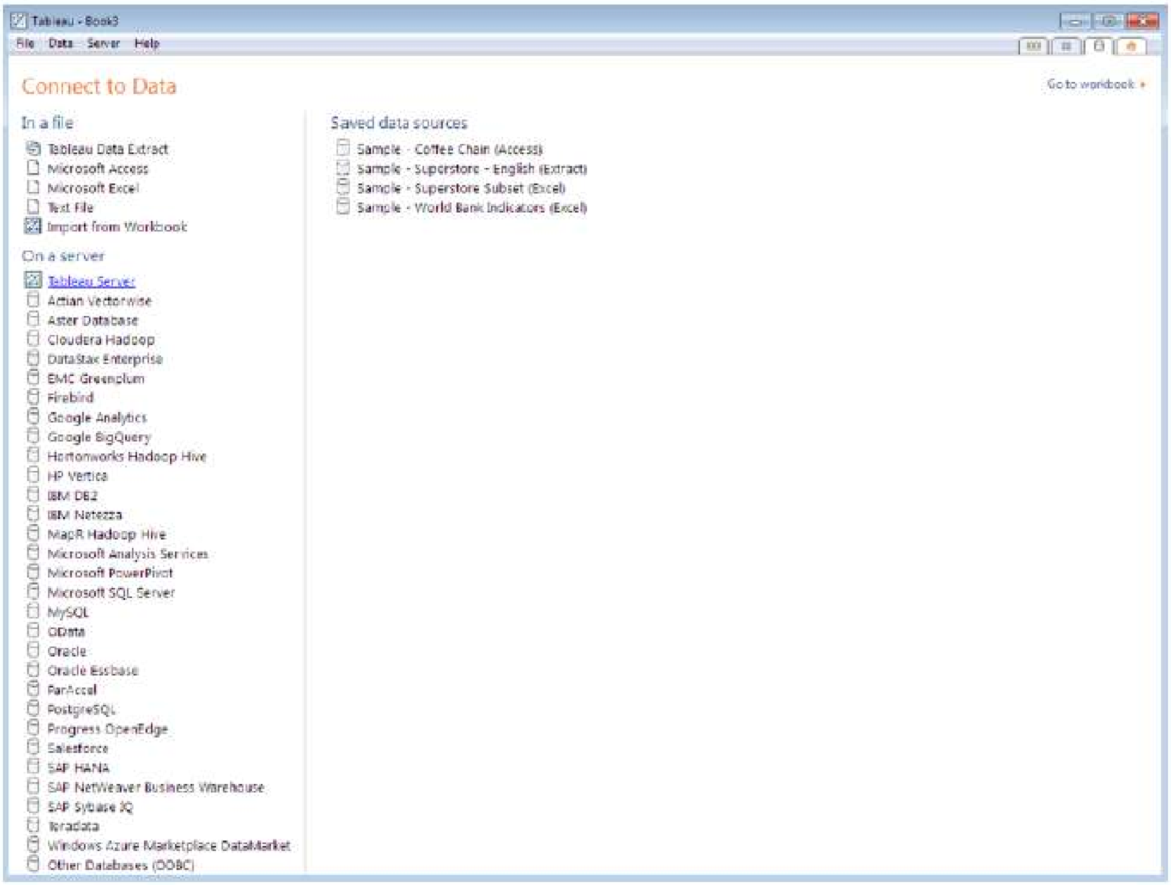
After studying this unit, you will be able to:

- Create connections to files and databases.
- Combine different data tables using joins and unions
- Deal with Data Quality Problems

14.1 HOW TO CONNECT TO YOUR DATA?

When you open Tableau you are taken to the home page where you can easily select from previous workbooks, sample workbooks, and saved datasources. You can also connect to new data sources by selecting Connect to Data. Figure 14-1 displays the screen. The option in a File is for connecting to locally stored data or file based data. Tableau Personal edition can only access Excel, Access, and text files (txt, csv). You can also import from data sources stored in other workbooks.

The options listed beneath On a Server' link to data stored in a database, data cube, or a cloud service. Although all of these databases have very different ways of storing and looking up data, the pop-up window is very user friendly and requires little or no understanding of the underlying technology. Most of these databases will require you to install a driver particular to each tool. Installation normally requires a few minutes and you can find all the connectors at: http://www.tableausoftware.com/support/drivers



Connect to Data

In a file

- Tableau Data Extract
- Microsoft Access
- Microsoft Excel
- Text File
- Import from Workbook

On a server

- Tableau Server
- Actian Vectorwise
- Aster Database
- Cloudera Hadoop
- DataStax Enterprise
- EMC Greenplum
- Firebird
- Google Analytics
- Google BigQuery
- Hortonworks Hadoop Hive
- HP Vertica
- IBM DB2
- IBM Netezza
- MapR Hadoop Hive
- Microsoft Analysis Services
- Microsoft PowerPivot
- Microsoft SQL Server
- MySQL
- OData
- Oracle
- Oracle Essbase
- ParAccel
- PostgreSQL
- Progress OpenEdge
- Salesforce
- SAP HANA
- SAP NetWeaver Business Warehouse
- SAP Sybase IQ
- Teradata
- Windows Azure Marketplace DataMarket
- Other Databases (ODBC)

Saved data sources

- Sample - Coffee Chain (Access)
- Sample - Superstore - English (Extract)
- Sample - Superstore Subset (Excel)
- Sample - World Bank Indicators (Excel)

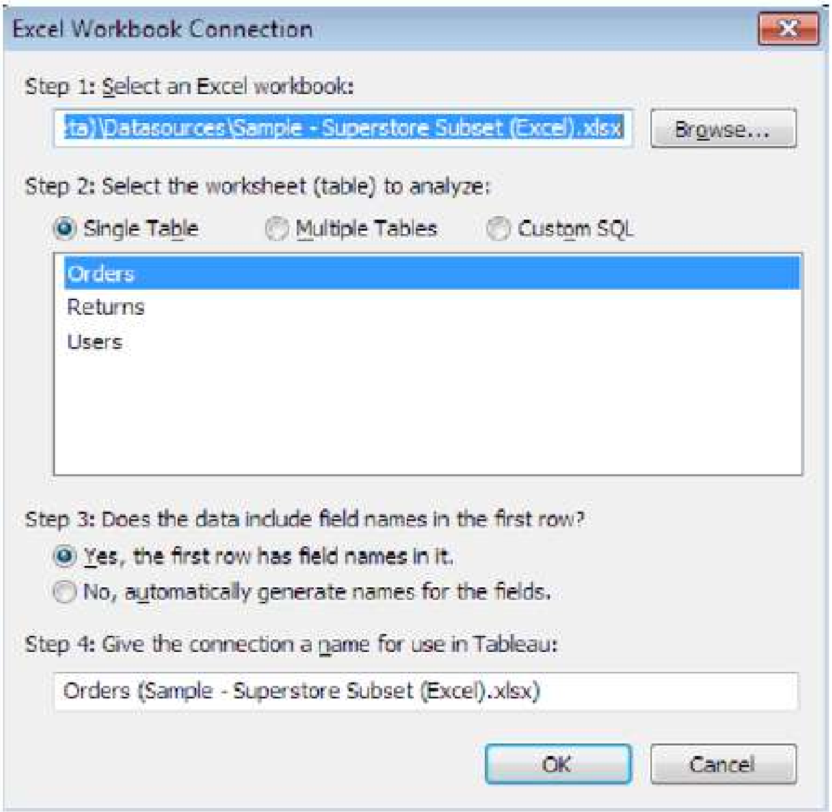
Fig 14.1 Connect to data screen

If your database isn't listed try the other database connector (ODBC) that utilizes the Open Database Connectivity standard. You will also see a list of saved data sources on the right. Saving data sources that you use frequently saves time. We will explain how to save a data source in the Tableau Data Source Files section later in this chapter.

Saved data source files (.tds) are found on your computer's hard disk in the data sources directory under the My Tableau repository. If you are logged into Tableau Server you may also see saved data sources on your server's repository.

Connecting to Desktop Sources

If you click on one of the desktop source options under the In a File list you will get a directory window to select the desired file. Once you have chosen your file you will be taken to the Connection Options window. There are small differences in the connection dialog depending on the data source you are connecting to but the menus are self-explanatory. Figure 14-2 shows the connection window with the Superstore sample spreadsheet being the file that is being accessed.



Excel Workbook Connection

Step 1: Select an Excel workbook:

ta)\Datasources\Sample - Superstore Subset (Excel).xlsx

Step 2: Select the worksheet (table) to analyze:

● Single Table ○ Multiple Tables ○ Custom SQL

Orders
Returns
Users

Step 3: Does the data include field names in the first row?

● Yes, the first row has field names in it.
○ No, automatically generate names for the fields.

Step 4: Give the connection a name for use in Tableau:

Orders (Sample - Superstore Subset (Excel).xlsx)

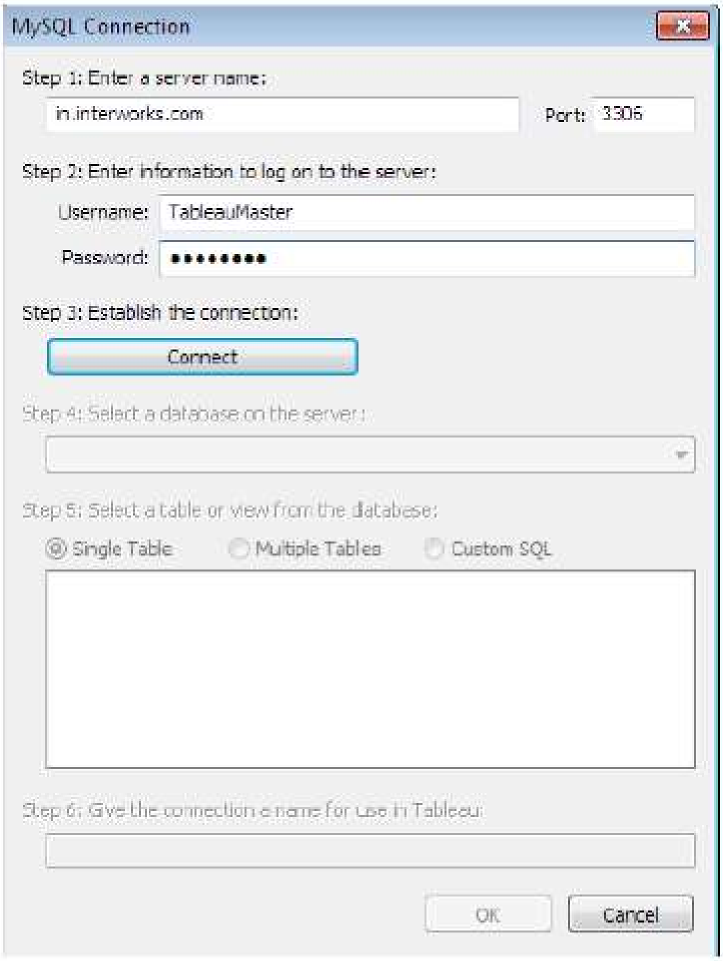
Fig. 14.2 The Connection Window

There are three tabs in the spreadsheet file. Tableau interprets these tabs the same way it views different tables in a database. The same is true of text files stored within the same folder. If the tabs contain related information, Tableau can join these just like it can join tables in a database. Joining options are the same regardless of the file or database type.

Once you have selected and customized your data connection, you will be taken to the second Data Connection window where you must decide whether or not to create an extract. There are advantages to extracting the data into Tableau's data engine, particularly when you are using Excel, Access, or text files as your data source. Clicking the OK button creates the connection and opens the workbook authoring environment.

Connecting to Database Sources

Databases have an additional level of security—requiring you to enter a server name and user credentials to access the data. The username and password you enter are assigned in the database, meaning the security credentials and the amount of access granted are controlled by the database—not Tableau. Figure 14–3 shows the connection window to a MySQL database.



MySQL Connection

Step 1: Enter a server name:

in.interworks.com Port: 3306

Step 2: Enter information to log on to the server:

Username: TableauMaster

Password: ••••••••

Step 3: Establish the connection:

Connect

Step 4: Select a database on the server:

Step 5: Select a table or view from the database:

◉ Single Table ○ Multiple Tables ○ Custom SQL

Step 6: Give the connection a name for use in Tableau:

OK Cancel

Fig. 14.3 Database connection window

The remaining steps in the connection window guide you through the process of selecting the database, database tables, and defining the joins between the tables in the data source. The final step is to decide whether you want to directly connect to the data or to extract data from the database into Tableau’s data engine. Following these steps completes the process of connecting to a database.

Connecting to Public Data sources

The increasing quantity and variety of data available via the Internet falls into three categories:

- Public domain data sets
- Commercial data services

- Cloud database platforms

For example, United State Census provides free data via the Internet. The World Bank provides a variety of data, and many other government public data repositories have sprouted all over the world. This data can be accessed by downloading files and then connecting Tableau to those files.

There are also a growing number of commercial data sources. At this time Tableau provides connectors to several, including:

- Google Analytics
- Google Big Query
- Amazon Redshift
- Salesforce
- Open Data Protocol (ODATA)
- Windows Azure Marketplace

The Google Analytics connector can be used to create customized click stream analysis of web pages. Google Big Query and Amazon Redshift connectors allow you to leverage the computing capacity of Google and Amazon. Both are designed to allow you to purchase petabyte- scale database processing capacity for a fee. There is also a connector for the popular cloud-based CRM tool—Salesforce.

Microsoft supplies data over the web via the Windows Azure Marketplace and ODATA. Tableau's own free cloud service—Tableau Public—allows you to create and share your workbooks and dashboards on the web.

Tableau Public is a great way to embed live/interactive dashboards on the web. Be careful not to publish proprietary data there as it is available to everyone without restriction.

14.2 WHAT ARE GENERATED VALUES?

Tableau has built-in fields that make difficult tasks easier. These are found on the left side of the screen at the bottom of the dimensions list and the bottom of the measures list. When you perform an operation (such as double clicking on a geographic field) these Tableau generated fields are automatically added to the design window. Generated values include:

- Measure Names and Measure Values
- Longitude and Latitude
- Number of Records

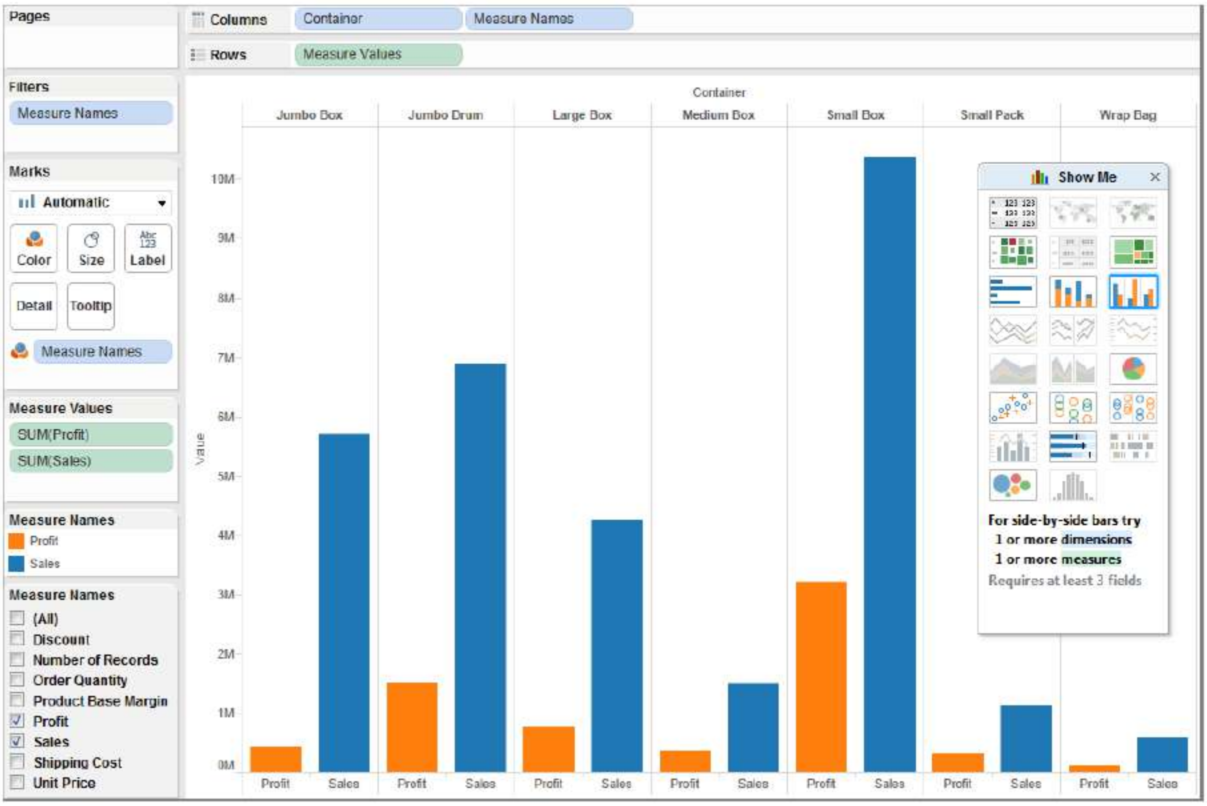
Measure Names, Measure Values, and Number of Records are always present. If your dimensions include standard geographic place names, Tableau will also automatically generate center-point geocodes.

Measure Names and Measure Values

Measure Names and Measure Values can be used to quickly express all the different measures in your dataset or to express multiple measures on a single axis.

In Figure 14–5 you can see that two measures are shown, SUM (Profit) and SUM (Sales). These are shown as separate columns in the same bar chart. The generated value, Measure Names, is used in the column shelf to separate the bars. Measure Name is also used on the marks card to distinguish color and on the filters shelf to limit the number of measures shown in the view. Measure

Value contains the data and this is shown as rows as you would expect from this type of bar chart. The side-by-side bar chart in Figure 14.5 was created by multi-selecting one dimension Container and two measures Sales and Profit. Using the Show Me button, the side-by-side bar chart was selected. Tableau automatically applied Measure Names to the column shelf and separated the two measures being plotted on the horizontal axis. The Measure Names Quick Filter was exposed by right-clicking on the Measure Names dimension on the Filter Shelf. Other measures can be added to the axis using the Quick Filter.



Columns: Container | Measure Names

Rows: Measure Values

Filters: Measure Names

Marks: Automatic

Color: Measure Names

Measure Values: SUM(Profit), SUM(Sales)

Measure Names

- 🟧 Profit
- 🟦 Sales

Measure Names

- ☐ (All)
- ☐ Discount
- ☐ Number of Records
- ☐ Order Quantity
- ☐ Product Base Margin
- ☑ Profit
- ☑ Sales
- ☐ Shipping Cost
- ☐ Unit Price

Show Me

For side-by-side bars try
1 or more dimensions
1 or more measures
Requires at least 3 fields

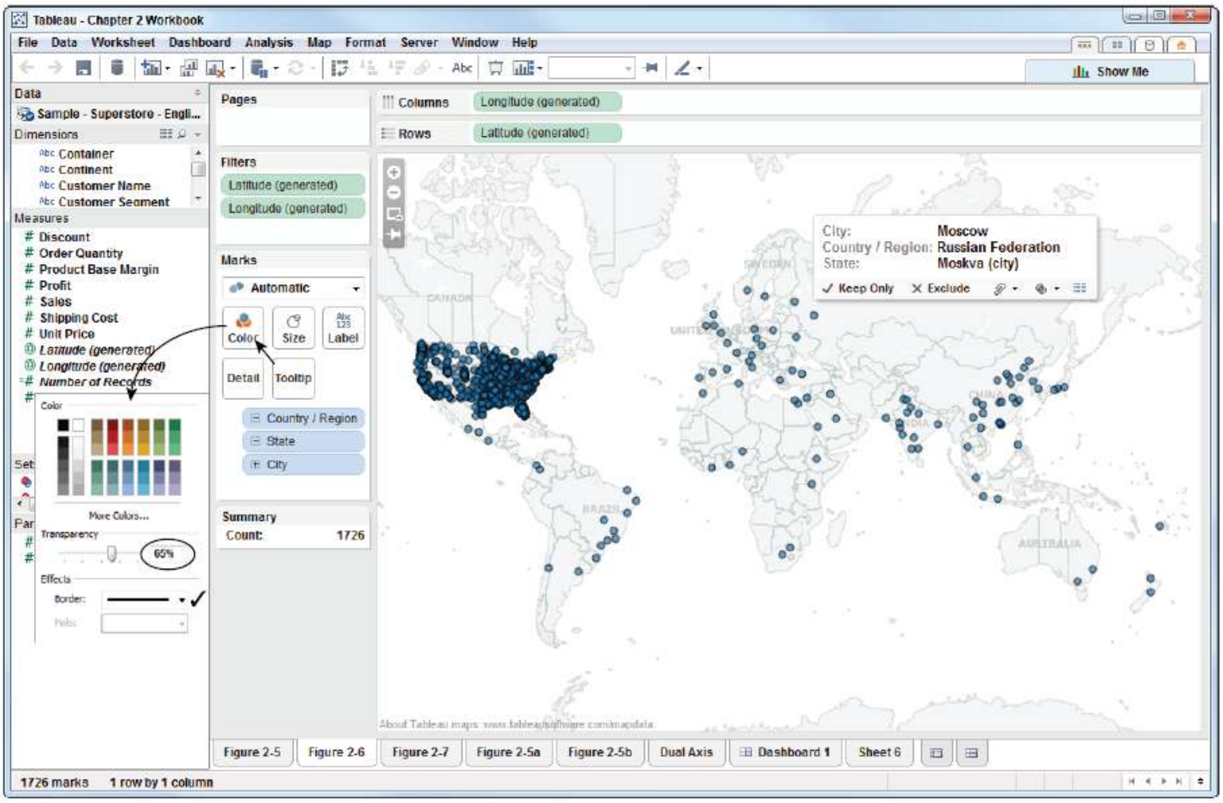
Fig 14.5 Measure values bar chart

The view could also be created by dragging Container to the column shelf, dragging the Sales Measure to the row shelf, then dragging Profit on to the left axis and dropping the measure when a light green ruler appears. The Measure Names and Measure Values pills will automatically appear when the second measure is placed on the vertical axis.

Tableau Geocoding

If your data includes standard geographic fields like country, state, province, city, or postal codes—denoted by a small globe icon—Tableau will automatically generate the longitude and latitude values for the center points of each geographic entity displayed in your visualization. If for some reason Tableau doesn’t recognize a geographic dimension, you can change the geographic role of the field by right-clicking on the field and selecting the appropriate geographic role. Figure 14–6 shows a map created using country, state, and city, then using Show Me to display the symbol map.

The Map Option menu seen on the left was exposed from the map menu, Map Option Selection. The marks in the map were styled from the Color button— changing the color transparency and adding a black border. Overlapping clusters of marks are easier to see. Hovering over any mark exposes the Tooltip that includes the geographic entities exposed in the marks card. The summary card was exposed in the view so that you can see that 1,726 marks are plotted. If Tableau failed to recognize any location, a small gray pill would appear in the lower right of the map. Clicking on that pill would expose a menu that would help you identify and correct the geocoding.



Columns: Longitude (generated)
Rows: Latitude (generated)

City: Moscow
Country / Region: Russian Federation
State: Moskva (city)

Transparency: 65%
Border: ✓

Summary
Count: 1726

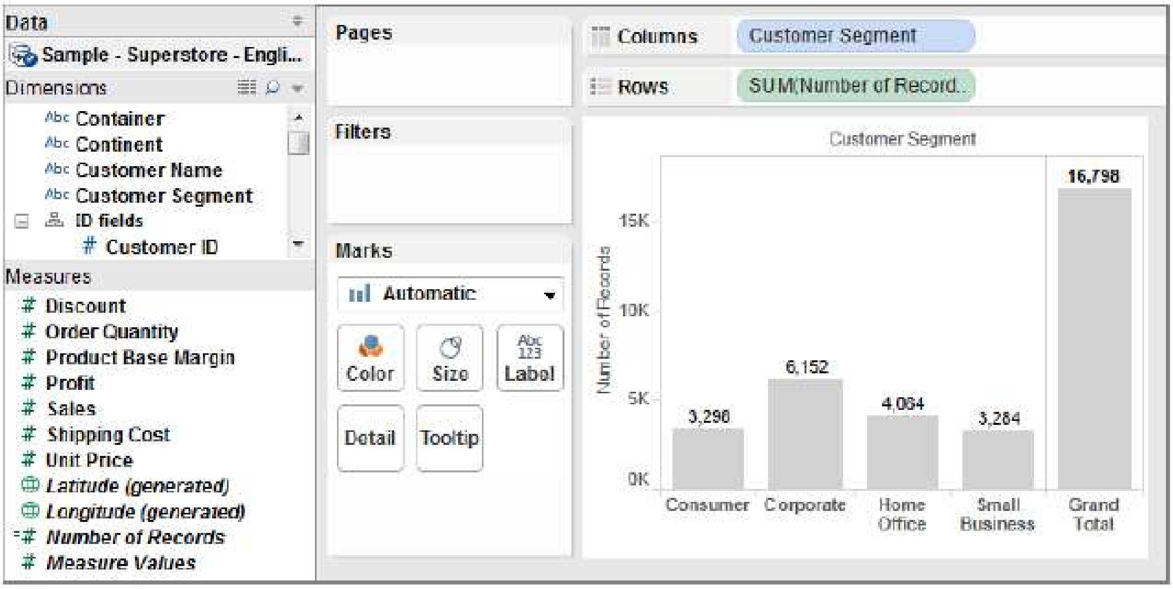
1726 marks 1 row by 1 column

Fig 14.6 Latitude and longitude generated measures

Number of Records

The final generated value provided is a calculated field near the bottom of the measures shelf called Number of Records. Any icon that includes an equals sign denotes a calculated field. The number of records calculation formula includes only the number one. This is how Tableau generates record counts.

The bar chart in Figure 14–7 displays the record count for each customer segment and grand total. Number of Records helps you understand the row count in your data set. It is particularly helpful when you begin to join other tables. Monitoring how the record count changes helps you understand data quality issues or design challenges that you may need to address.



Columns: Customer Segment

Rows: SUM(Number of Record..

Fig 14-7 Number of records

14.3 KNOWING WHEN TO USE A DIRECT CONNECTION OR A DATA EXTRACT

Direct connections allow you work with live data. When you extract data you import some or all of your data into Tableau’s data engine. This is true in Tableau Desktop and Server. Which connection method is the best to use? There is no straightforward answer. It is entirely dependent on your situation, requirements, and network resources.

The Flexibility of Direct Connections

Connecting to your data source with a direct connection means you are always visualizing the most up-to-date facts. If your database is being updated in real-time you only need to refresh the Tableau visualization via the F5 function key or by right-clicking on the data source in the data window and selecting the Refresh option.

If you connect to massive data, the visualization is very dense, or your data is in a high-performance enterprise-class database, you may get faster response time with a direct connection. Choosing a direct connection doesn’t preclude the possibility of extracting the data later. You can also swap from an extract to a live connection by right-clicking the datasource and un-checking the Use Extract option.

The Advantages of a Data Extract

Data extracts don't have the advantage real-time updating that a direct connection provides, but using Tableau's data engine provides a number of benefits:

- Performance improvement
- Additional functions
- Data portability

Performance Improvement

Perhaps your primary database is already heavily loaded with requests. Using Tableau's data engine enables you to split the load from your primary database server to the Tableau Server. Tableau's extract may be updated daily, weekly, or monthly during off-peak hours. Tableau's Server can also refresh extracts incrementally and in time intervals as low as fifteen minutes. In many cases, the small time consumed during the data extract update is more than offset by the performance gains.

There are several options available for creating an extract. First, you can aggregate the extract, which will roll up rows so that only the aggregation and fields used are updated for the visible dimensions and measures. Aggregating for Visible Dimensions when performing a data extract will reduce the amount of data that Tableau is importing. The appropriate level of fidelity is provided but the size of the extract file is significantly reduced—making the extract file more portable but also improving security.

Extracting incrementally also speeds refresh time because Tableau isn't updating the entire extract file. It is adding only new records. To do incremental extracts you must specify a field to use as the index; Tableau will only refresh the row if the index has changed, so you need to be aware that changes to a row of data which doesn't change the index field will be excluded from the update.

Another way to speed extracts is to apply filters when extracting the data. If the analysis doesn't require your entire dataset you can filter the extract to include only the records required. If you have a very large dataset you will rarely need to extract the entire contents of the database. For example, your database may include ten years of historical data but you may only require one year of history. Once you have created an extract you may append another file. This may be a great alternative to custom SQL if you are considering a table UNION. This technique might be useful if you need to combine monthly data that is stored in separate tables.

Additional Functions

If your data source is from a file (Excel, Access, text) doing an extract will add calculation functions (median and count distinct) that are not supported by the data source.

Data Portability

Extracts can be saved locally and used when the connection to your data source is not available. A direct connection doesn't work if you don't have access to your data source via a local network or the Internet. Perhaps you need to supply a dashboard to an executive that will be flying to a remote location. Providing that person with a data extract (.tde) file provides that person with a fully-functional, high-performance, data source. Data extract files are also compressed and are normally much smaller than the host system database tables.

In enterprise environments, data governance is an important consideration. If you distribute many data extract files to field staff, keep in mind that you should consider the security of those files. Appropriate safeguards should be in place (non-disclosure agreements) before you provide these files to traveling or remote staff. Consider restricting what the extract includes via filters and aggregating to visible dimensions.

Using Tableau's File Types Effectively

Tableau provides flexible options for the sharing of data and design metadata. This is accomplished through a variety of file types:

- Tableau Workbook (.twb)
- Tableau Packaged Workbook (.twbx)
- Tableau Datasource (.tds)

- Tableau Bookmark (.twb)
- Tableau Data Extract (.tde)

You should see many of these files in your My Tableau repository folder which is normally located in My Documents. Data extract (.tde) files were covered in the previous section. Next you will see how the other file types can be used.

Tableau Workbook Files

Tableau Workbook files (.twb) are the main file type created by Tableau to save your entire workbook. These are normally small files because the only data they contain is the metadata related to your connection and the pill placements for rendering the views and dashboards in the workbook. What is not saved is the underlying data from the data source.

To clarify: A .twb file does not contain any of the actual data from the database. It contains the definition of how you wish to display data. This means workbook files will normally be very small. But, if you want to share the workbook with other people you need to be certain that they have access rights to the database or that you also provide the data source with the workbook.

Tableau Data source Files

Changes made within your data window (the left side of your workbook) alter the metadata of your connection. Grouping, sets, aliased names, field-type changes, and any other modifications made in your workbook are part of the metadata. Can you share just the metadata with others? The answer is yes. This is done by creating a Tableau Data Source (.tds) file.

A Tableau data source file defines where the source data is, how to connect to it, what fieldnames have been changed, and other changes applied in the dimensions and measures shelves. Data source files can be saved locally or published to Tableau Server. This is particularly helpful if you work in a large enterprise. Perhaps you have a small number of database experts that understand your database schema well. They can create the connection, define table joins, group or rename fields, and then publish the data source file for less experienced staff to use as a starting point.

To create a data source file right-click on the filename in your data window, then select the Add To Saved data sources option. Data source files are placed in the My Tableau repository/data source folder. Additionally, files placed in that folder are automatically displayed as saved data connections on Tableau’s home and connection tabs. Alternatively, you can publish data source files to Tableau Server and share them with other staff. The best part about this option—changes made to the data source file are automatically propagated to other people using that connection.

Tableau Bookmark Files

What if you have a massive workbook (with many worksheets) and you want to share one worksheet only with a colleague? This is done by using a bookmark (.tbm) file. Bookmark files save the data and metadata related to a worksheet within your workbook—including the connection and calculated fields.

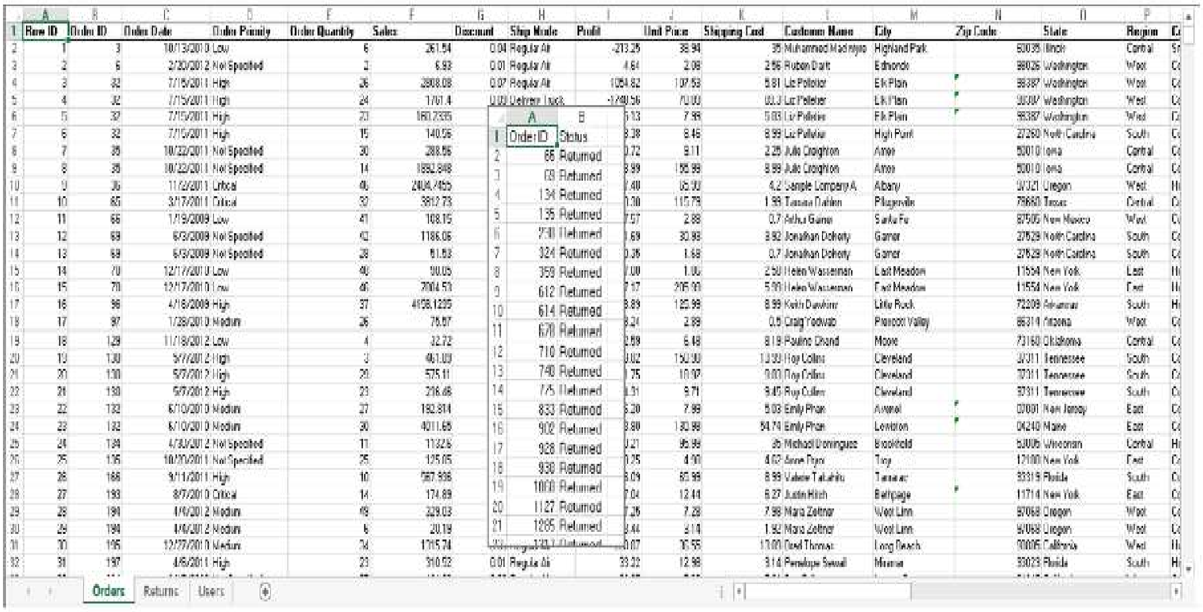
To create a bookmark file go to the Windows menu bar and look for the Bookmark menu option and select Create Bookmark. The bookmark will become visible when a new Tableau session is started. The file will appear in the Windows menu. Opening the bookmark file will initiate the connection and add it to the workbook. Tableau bookmark files are stored in your Tableau Repository in the Bookmarks folder.

14.4 JOINING DATABASE TABLES WITH TABLEAU

Most Tableau users aren’t database experts. This section introduces a fundamental database concept—joining tables. Seldom will your data source include every bit of information you need in a single table. Even if you normally connect to Excel it may be advantageous to use related data from more than one tab.

As long as the data resides in a single spreadsheet or database and each table includes unique identifiers that tie the tables or tabs together, you can perform joins of these tables within Tableau. These identifiers are called Key Records. Database joins can be complex, but the basic principle is to bring together related information in your view. In Tableau, you can define joins when you make your initial data connection or add them later. This example will use the Orders and Return tabs (tables) from the Superstore sample dataset. Figure 14–8 shows portions of both tables.

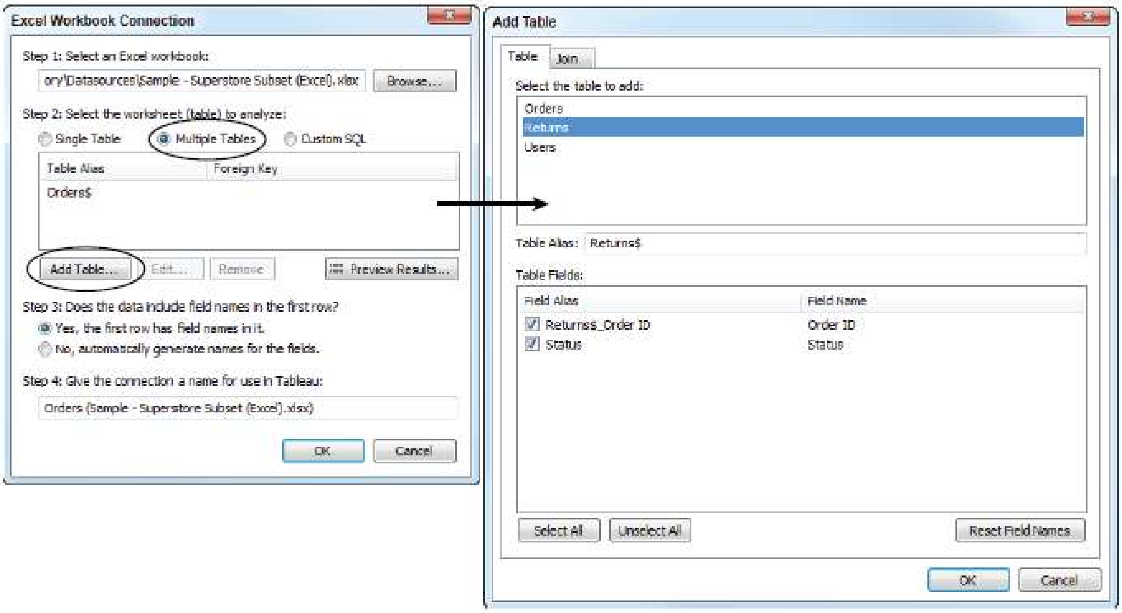
The Orders table includes billing information. The Returns tab includes the smaller returned order table. Start by connecting to the spreadsheet as you would if you were going to connect to one table. In the Connection Menu under Step 2, select Multiple Tables and click the Add Table button to expose the Add Table menu. Then select the Returns table as you see in Figure 14-9.



Row ID	Order ID	Order Date	Order Priority	Order Quantity	Sales	Discount	Ship Mode	Profit	Unit Price	Shipping Cost	Customer Name	City	Zip Code	State	Region
1	3	10/13/2010	Low	6	261.54	0.04	Regular Air	-213.25	38.94	35	Muhammed MacIntyre	Highland Park		60035	Illinois	Central
2	6	2/20/2012	Not Specified	2	6.93	0.01	Regular Air	4.64	2.08	2.56	Ruben Dartt	Edmonds	98026	Washington	West
3	32	7/15/2011	High	26	2808.08	0.07	Regular Air	1054.82	107.53	5.81	Liz Pelletier	Elk Plain	98387	Washington	West
4	32	7/15/2011	High	24	1761.4	0.09	Delivery Truck	-1748.56	70.89	89.3	Liz Pelletier	Elk Plain	98387	Washington	West
5	32	7/15/2011	High	23	160.2335				7.99	5.03	Liz Pelletier	Elk Plain	98387	Washington	West
6	32	7/15/2011	High	15	140.56				8.46	8.99	Liz Pelletier	High Point	27260	North Carolina	South
7	35	10/22/2011	Not Specified	30	288.56				9.11	2.25	Julie Creighton	Ames	50010	Iowa	Central
8	35	10/22/2011	Not Specified	14	1892.848				155.99	8.99	Julie Creighton	Ames	50010	Iowa	Central
9	36	11/3/2011	Critical	46	2484.7455				65.99	4.2	Sample Company A	Albany	97321	Oregon	West
10	65	3/17/2011	Critical	32	3812.73				115.79	1.99	Tamara Dahlen	Pflugerville	78660	Texas	Central
11	66	1/19/2009	Low	41	108.15				2.88	0.7	Arthur Gainer	Santa Fe	87505	New Mexico	West
12	69	6/3/2009	Not Specified	42	1186.06				30.98	3.92	Jonathan Doherty	Garner	27529	North Carolina	South
13	69	6/3/2009	Not Specified	28	51.53				1.68	0.7	Jonathan Doherty	Garner	27529	North Carolina	South
14	70	12/17/2010	Low	48	90.05				1.86	2.58	Helen Wasserman	East Meadow	11554	New York	East
15	70	12/17/2010	Low	46	7804.53				205.99	5.99	Helen Wasserman	East Meadow	11554	New York	East
16	96	4/16/2009	High	37	4158.1235				125.99	8.99	Keith Dawkins	Little Rock	72209	Arkansas	South
17	97	1/29/2010	Medium	26	75.57				2.89	0.5	Craig Yedwab	Prescott Valley	86314	Arizona	West
18	129	11/19/2012	Low	4	32.72				6.48	8.19	Pauline Chand	Moore	73160	Oklahoma	Central
19	130	5/7/2012	High	3	461.89				150.98	13.99	Roy Collins	Cleveland	37311	Tennessee	South
20	130	5/7/2012	High	29	575.11				10.97	9.03	Roy Collins	Cleveland	37311	Tennessee	South
21	130	5/7/2012	High	23	236.46				9.71	9.45	Roy Collins	Cleveland	37311	Tennessee	South
22	132	6/10/2010	Medium	27	192.814				7.99	5.03	Emily Phan	Avenel	07001	New Jersey	East
23	132	6/10/2010	Medium	30	4011.65				130.98	54.74	Emily Phan	Lewiston	04240	Maine	East
24	134	4/30/2012	Not Specified	11	1132.6				95.99	35	Michael Dominguez	Brookfield	53005	Wisconsin	Central
25	135	10/21/2011	Not Specified	25	125.85				4.98	4.62	Anne Pryor	Troy	12180	New York	East
26	166	9/11/2011	High	10	567.936				65.99	8.99	Valerie Takahito	Tamarac	33319	Florida	South
27	193	8/7/2010	Critical	14	174.89				12.44	6.27	Justin Hirsh	Bethpage	11714	New York	East
28	194	4/4/2012	Medium	49	329.03				7.28	7.98	Maria Zettner	West Linn	97068	Oregon	West
29	194	4/4/2012	Medium	6	20.19				3.14	1.92	Maria Zettner	West Linn	97068	Oregon	West
30	195	12/27/2010	Medium	34	1215.74				16.55	13.99	[illegible] Thomas	Long Beach	90805	California	West
31	197	4/6/2011	High	23	310.52	0.01	Regular Air	33.22	12.98	3.14	Penelope Sewall	Miramar	33023	Florida	South

Order ID	Status
65	Returned
69	Returned
134	Returned
135	Returned
230	Returned
324	Returned
359	Returned
612	Returned
614	Returned
678	Returned
710	Returned
740	Returned
775	Returned
833	Returned
902	Returned
928	Returned
930	Returned
1060	Returned
1127	Returned
1285	Returned

Fig 14-8 Superstore orders and return tables



Excel Workbook Connection

Step 1: Select an Excel workbook:

ory\Datasources\Sample - Superstore Subset (Excel).xlsx

Step 2: Select the worksheet (table) to analyze:

○ Single Table ● Multiple Tables ○ Custom SQL

Table Alias	Foreign Key
Orders$	

Step 3: Does the data include field names in the first row?

● Yes, the first row has field names in it.

○ No, automatically generate names for the fields.

Step 4: Give the connection a name for use in Tableau:

Orders (Sample - Superstore Subset (Excel).xlsx)

Add Table

Select the table to add:

- Orders
- Returns
- Users

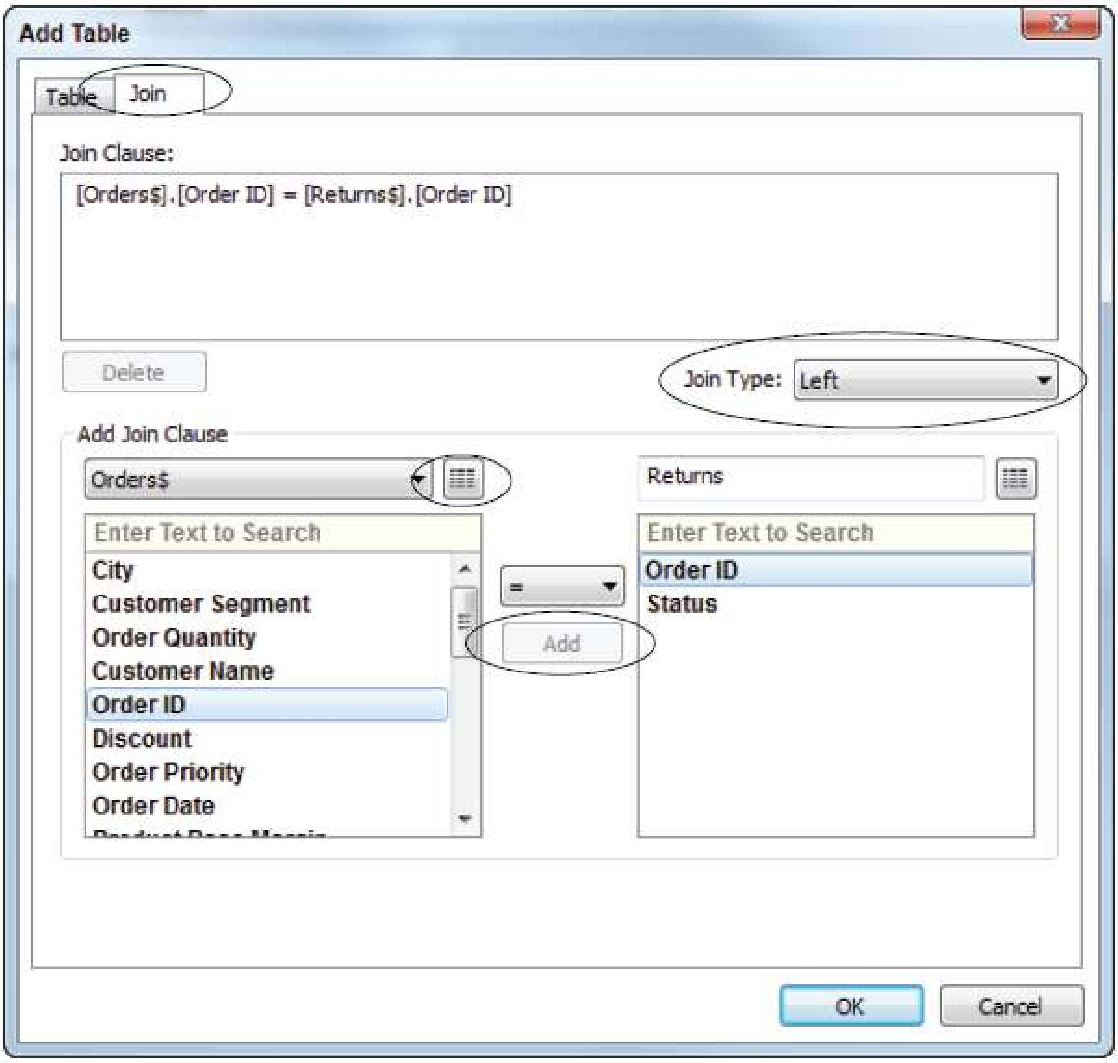
Table Alias: Returns$

Table Fields:

Field Alias	Field Name
☑ Returns_Order ID	Order ID
☑ Status	Status

Fig 14-9. Joining multiple tables

While in the Add Table menu ensure that the Returns table is highlighted and click the Join button. This will expose the menu in which you define the join type as you see in Figure 17–10. In the example, you see that the Left outer join type has been selected. If you preview the results you will see that the join will result in 9,426. Following these steps results in a left outer join between the Orders and Returns tables. Keep in mind that you can also join additional tables later just by pointing at the data source on your data shelf, right-clicking, and selecting the Edit Tables option. Using different join types can result in different record counts so it is important that you understand the different join types.



Add Table

Join Clause:

[Orders$].[Order ID] = [Returns$].[Order ID]

Join Type: Left

Add Join Clause

Orders$

Returns

City
Customer Segment
Order Quantity
Customer Name
Order ID
Discount
Order Priority
Order Date

Order ID
Status

Fig. 14-10 Joining tables in Tableau

The Default Inner Join

When you join two tables together Tableau will default to the inner join type. Figure 14–11 shows a Venn diagram that illustrates the inner join. Using an inner join returns only records that match in both the left and right tables. In the Superstore example this join type returns only ninety-eight records. It is a good practice when you join tables to know how many records there are in each table. If you're working with a spreadsheet you can look at each tab and note the total row counts in each. Remember to deduct the header from your row totals. Alternatively, as you are doing the join, utilize the preview buttons to check the row counts.

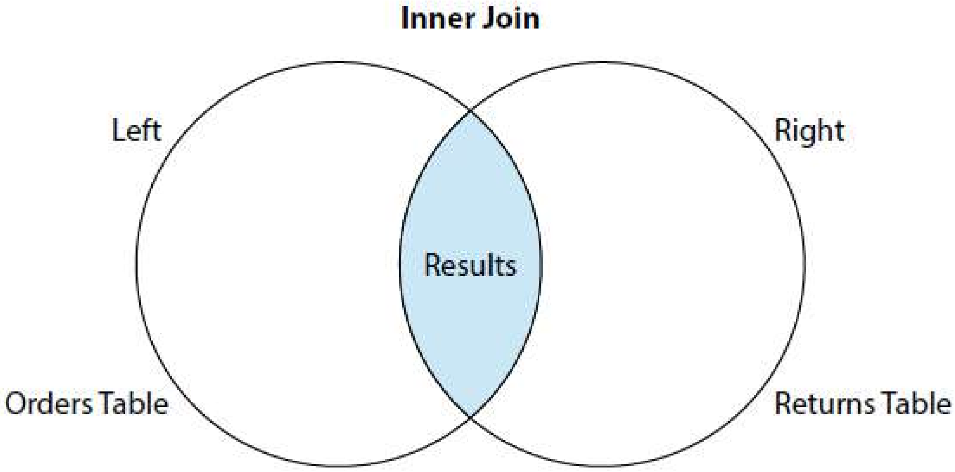


Fig.14-11 The inner join

When you complete the join you can also drag the record count field into the view to see how many total records are available. You can have more than one join clause to ensure that the correct results are returned. If you're a database expert this won't present any challenge. If you are like many Tableau users you are probably not accustomed to creating joins. If you run into problems, ask for help from a database expert.

The Left and Right Join Types

Tableau provides two other join types via point and click options in the Join menu. These join types give priority to either the left table or the right in the set returned. Pick the primary table first. In the previous example, the primary table is the Orders table so it is considered the left table. The new table added in the join is the Returns table on the right. Selecting left gives priority to the original table. Selecting right gives priority to the new table. But what does it mean? Figure 14–12 shows a Venn diagram of the left outer join type.

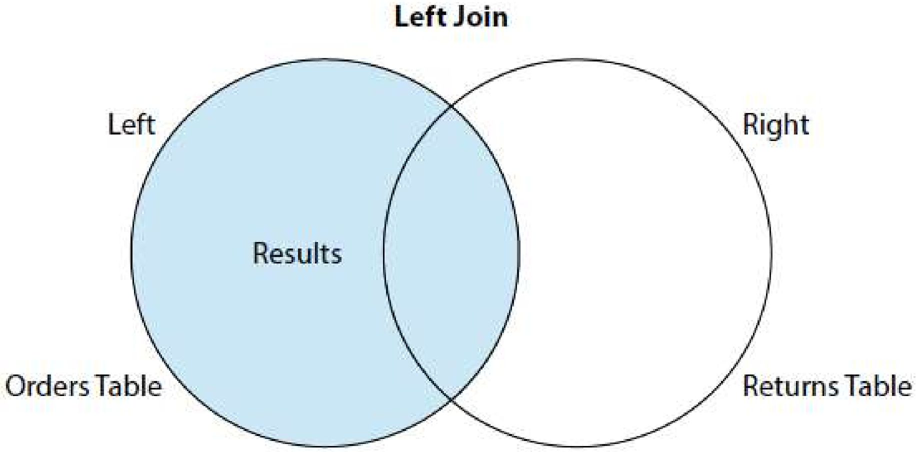


Fig 14-12. The left outer join

In the example, the left join returns every record in the orders table plus the matching records in the returns table. Earlier you saw that join generated over nine thousand records being returned. The right join gives priority to the right returns table as you see in Figure 14–13

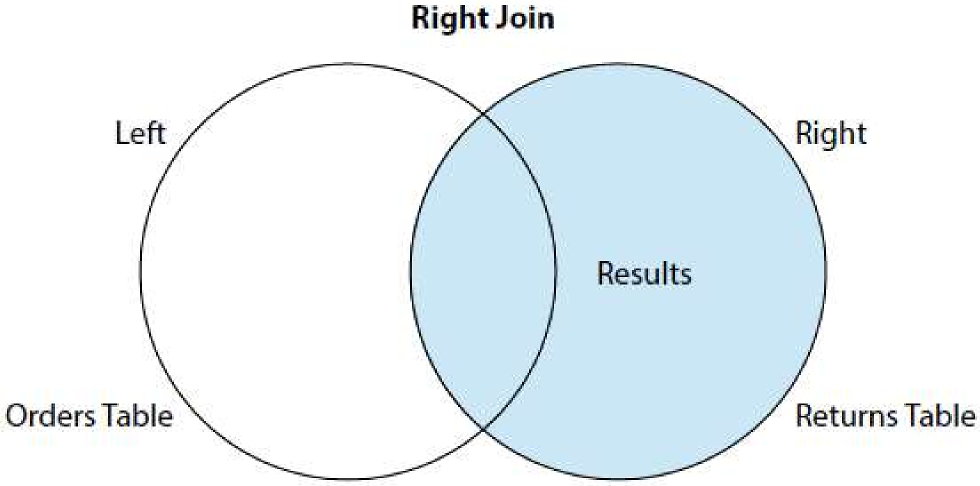


Fig 14-13 The Right outer join

Since there are fewer rows in the returns table the number of records will drop significantly and only include details from matching records in orders. If you preview results when using left and right joins you may see a lot of null fields in yellow. Or, if you check the record counts and place the key record that you use in the join on your row shelf, you will see the word null appear whenever a record exists in the primary table that is missing in the joined table. In Superstore, a right join would result in 1,673 records being returned, but only 98 of those records will be matched to the orders table. The remaining 1,573 records will return null. These are the order records in the order table that have no matching record in the returns table.

14.5 BLENDING DIFFERENT DATA SOURCES IN A SINGLE WORKSHEET

Wouldn’t it be wonderful if all the data you needed to create your analysis always resided in a single database? Many times this isn’t the situation. If you need to use data from more than one data source, Tableau provides a solution that does not require building a middle-layer data repository. As long as the disparate data sources have at least one common field, Tableau facilitates using both sources via data blending.

When to Use Blending vs Joins?

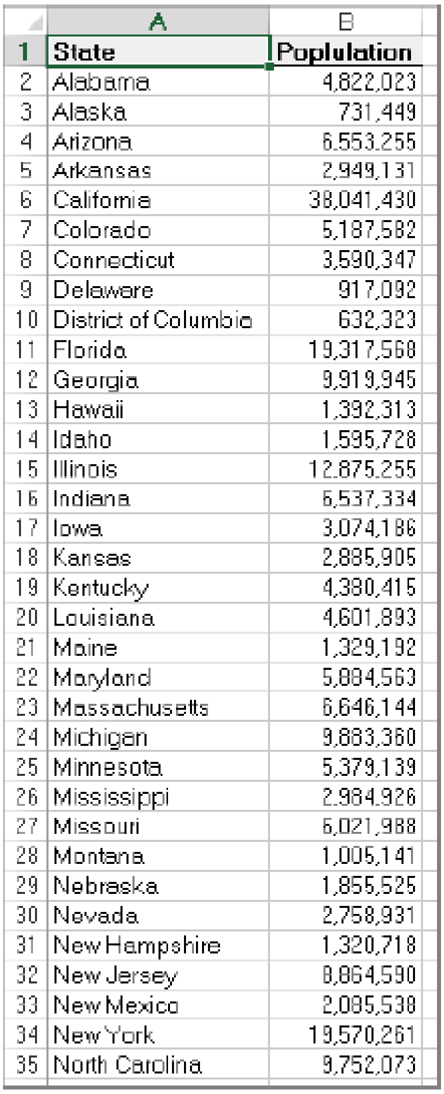
If your data does reside in a single source, it is always more desirable to use a join versus a data blend. In the last section you saw that Tableau provides plenty of flexibility for creating joins to your data source. Joins are normally the best option because joins are robust, persist everywhere in your workbook, and are more flexible than blending. However, if your data isn’t in one place, blending provides a viable way to quickly create a left-join-like connection between your primary and secondary data sources. Blends are more fragile than joins. They persist only on the worksheet page on which they are created. But blends offer a different kind of flexibility—the ability to alter the primary data source for each worksheet page.

How to Create a Data Blend?

Creating data blends requires a little planning. If you are going to bring data that doesn’t reside in your primary data source you have to think about what field(s) you may need in order to achieve the desired result. There are two ways you can create a blend—the automatic method or manually defining the blend.

Automatically-Defined Relationship

The automatic method works well if the field you are employing to create the blend has the same fieldname in both data sources, or if you alias the field names in Tableau so that they match. The Superstore data source contains geographic sales data. What if you wanted to know what the per capita sales for each state were for the year 2012? The Superstore data set doesn't include population data. But, the United States Census Bureau website has population data. The data in Figure 14–17 was downloaded from the web.



	A	B
1	**State**	**Poplulation**
2	Alabama	4,822,023
3	Alaska	731,449
4	Arizona	6,553,255
5	Arkansas	2,949,131
6	California	38,041,430
7	Colorado	5,187,582
8	Connecticut	3,590,347
9	Delaware	917,092
10	District of Columbia	632,323
11	Florida	19,317,568
12	Georgia	9,919,945
13	Hawaii	1,392,313
14	Idaho	1,595,728
15	Illinois	12,875,255
16	Indiana	6,537,334
17	Iowa	3,074,186
18	Kansas	2,885,905
19	Kentucky	4,380,415
20	Louisiana	4,601,893
21	Maine	1,329,192
22	Maryland	5,884,563
23	Massachusetts	6,646,144
24	Michigan	9,883,360
25	Minnesota	5,379,139
26	Mississippi	2,984,926
27	Missouri	6,021,988
28	Montana	1,005,141
29	Nebraska	1,855,525
30	Nevada	2,758,931
31	New Hampshire	1,320,718
32	New Jersey	8,864,590
33	New Mexico	2,085,538
34	New York	19,570,261
35	North Carolina	9,752,073

Fig 14-7 Population data

Just two columns of data are included in the table. It is important to note the field description for state. Once again—for automatic blending to work—the field name for the blend must be the same in Superstore and the census data file. If the fields are not the same you will need to edit the name in the spreadsheet or rename the fields in Tableau so that they match. To automatically blend the population data with the Superstore data build a view in Tableau that contains the state field. Figure 14–18 shows a view that will work.

Superstore is the primary data source. The bar chart is filtered for the desired year. The population data is from a completely different data source, but both data sources include the word State. Automatic blending can now be done by pointing at the population data spreadsheet and dragging it into the worksheet seen in Figure 14–18. Once that is done, the data from the population spreadsheet can be used in the workbook. The visualization in Figure 14–19 uses the blended population data to express sales per hundred thousand population by state.

Look at the data window in the upper left of Figure 14–19. The blue check next to the Superstore data source indicates that it is the primary data source. The orange check next to the population data denotes it is the secondary data source.

Since the secondary source is highlighted you see its dimension and measure fields below. The orange border on the left side of the dimension and measures shelves confirms that they come from the secondary data source and the orange link to the right of the State field indicates the field used for the blend. You can also see the State field in Figure 14–18 from the primary data source.

A warning—when you perform data blending you must ensure that all of the records you expected to blend actually came into the dataset. In Figure 14–19 that is clearly not the case. The states of Massachusetts (MA) and Missouri (MO) didn’t come over in the blend because the state names in the census data are not abbreviated. This can be fixed by right-clicking on the abbreviated state label for Missouri and Massachusetts and aliasing full spelling of each state name. After that is done, the population data from those states will be blended as well.

This is an important point with data blending. As the “designer” you must ensure the integrity of the data blend. The whole point in doing this exercise was to use the blended data to calculate per capita sales by state. Figure 14–20 displays the finished blend.

To save space, Figure 14–20 shows only the top seven states by per capita sales. The labels to the right of each bar show the sales per hundred thousand people. The color of each bar encodes the total sales of each state.

Manual Blending

What if your needs are more involved? A scenario that requires a more complicated blend would be the comparison of budget data from spreadsheet with actual data from a database. Assume that you have defined a budget by product category for each month in the year 2012, and that you want to create a visualization that will display the actual and the budgeted sales by month. Building this view will require a blend on the product category and the date field. The steps required are:

1. Connect to both data sources.
2. Use the edit relationship mean to define the blend.
3. Build the visualization.

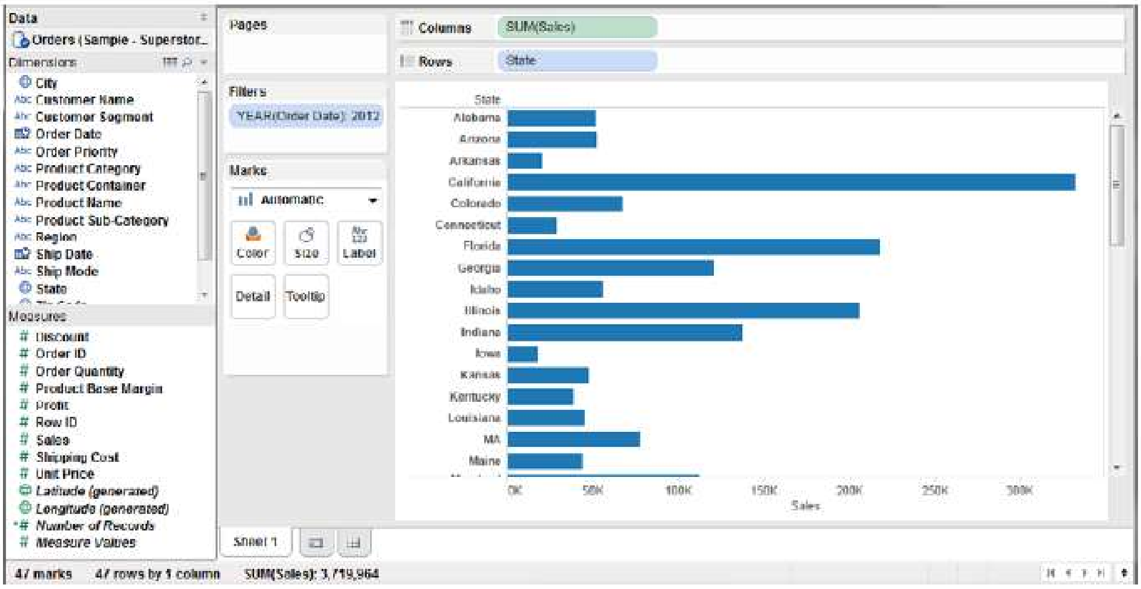


Fig. 14-18 Sales by state

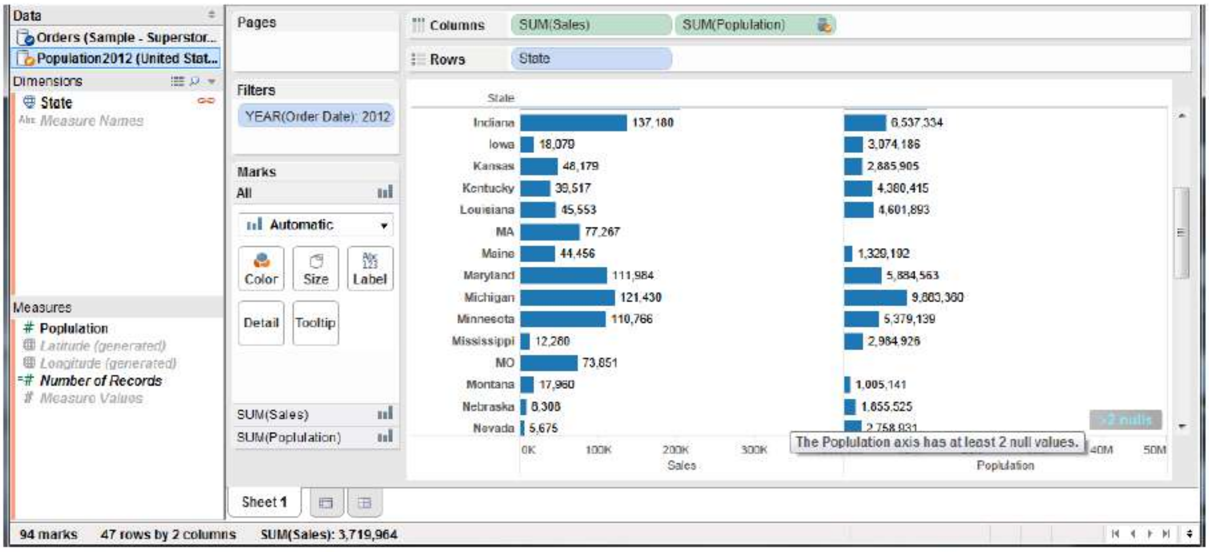


Fig 14-19 population data blended with superstore

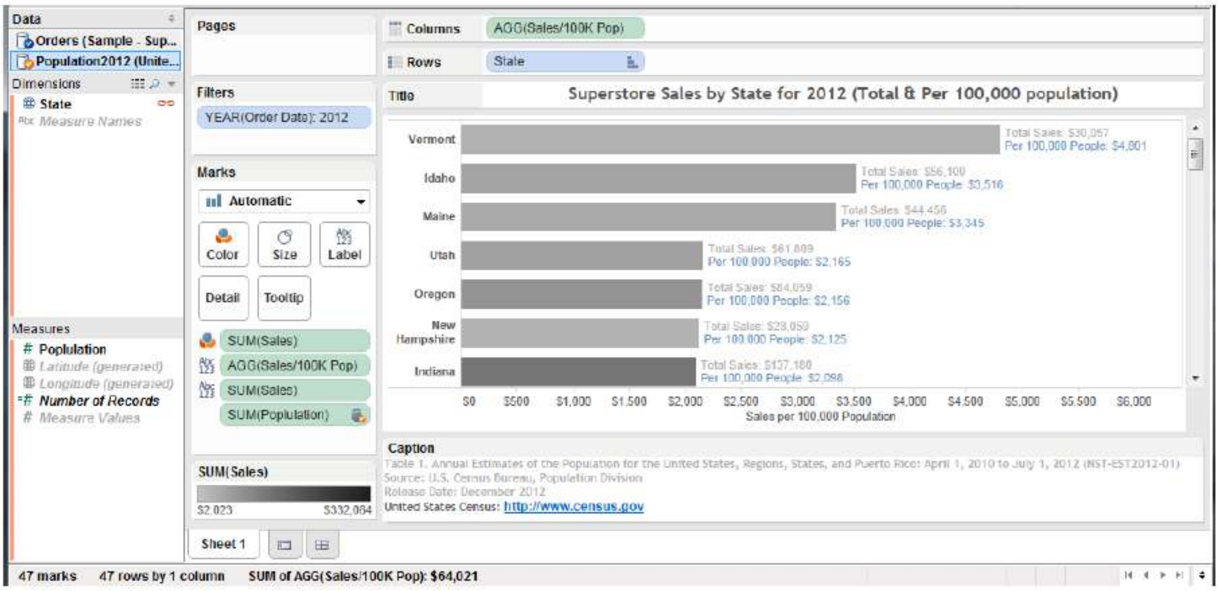
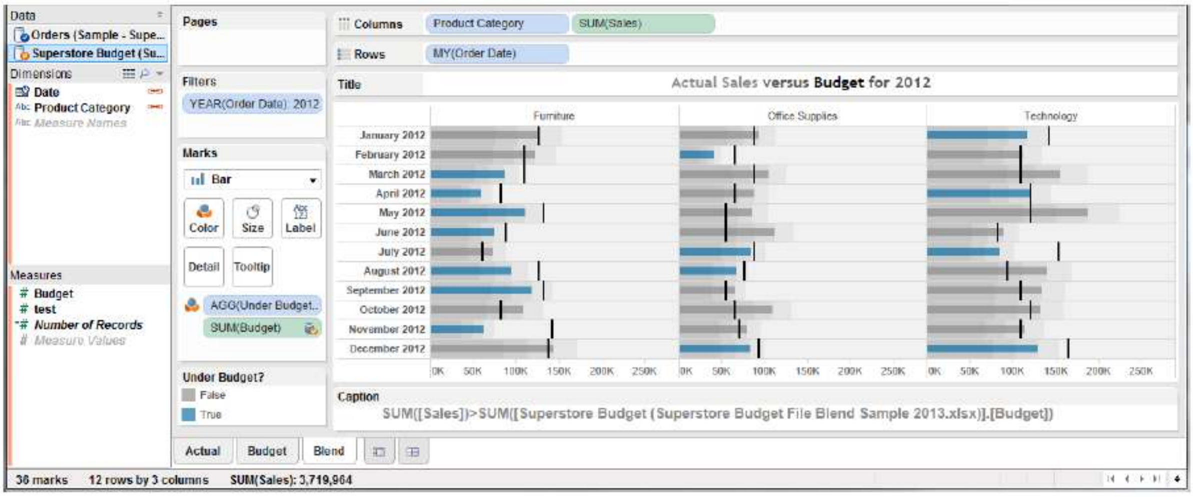

Data
Orders (Sample - Sup...
Population2012 (Unite...
Dimensions
State
Measure Names
Measures
Poplulation
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
YEAR(Order Date): 2012
Marks
Automatic
Color
Size
Label
Detail
Tooltip
SUM(Sales)
AGG(Sales/100K Pop)
SUM(Sales)
SUM(Poplulation)
SUM(Sales)
$2,023
$332,064
Columns
AGG(Sales/100K Pop)
Rows
State
Title
Superstore Sales by State for 2012 (Total & Per 100,000 population)
Vermont Total Sales: $30,057 Per 100,000 People: $4,801
Idaho Total Sales: $56,100 Per 100,000 People: $3,516
Maine Total Sales: $44,456 Per 100,000 People: $3,345
Utah Total Sales: $61,609 Per 100,000 People: $2,165
Oregon Total Sales: $84,059 Per 100,000 People: $2,156
New Hampshire Total Sales: $28,059 Per 100,000 People: $2,125
Indiana Total Sales: $137,180 Per 100,000 People: $2,098
$0 $500 $1,000 $1,500 $2,000 $2,500 $3,000 $3,500 $4,000 $4,500 $5,000 $5,500 $6,000
Sales per 100,000 Population
Caption
Table 1. Annual Estimates of the Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2012 (NST-EST2012-01)
Source: U.S. Census Bureau, Population Division
Release Date: December 2012
United States Census: http://www.census.gov
Sheet 1
47 marks
47 rows by 1 column
SUM of AGG(Sales/100K Pop): $64,021

Fig 14-20 Bar chart using blended data

After connecting to the Superstore dataset and the spreadsheet containing budgeted sales, it is possible to define the blend manually. The blending must include both the product category field and a date field. In this example, month and year are used. Figure 14–21 shows a bullet graph that uses the blended superstore data and budget data. As you can see in Figure 14–21, actual sales data from the primary data source (the orders table in Superstore) is displayed as blue or gray bars. Budgeted data from the secondary data source is plotted using vertical black reference lines for each cell. Notice the two orange links in the dimension shelf for the budget data source. Both fields are being used in the blend. How do you create a more multi-field blend? Figure 14–22 shows the Edit Relationships menu.



Figur e 14–21 *Bullet graph using blended data*

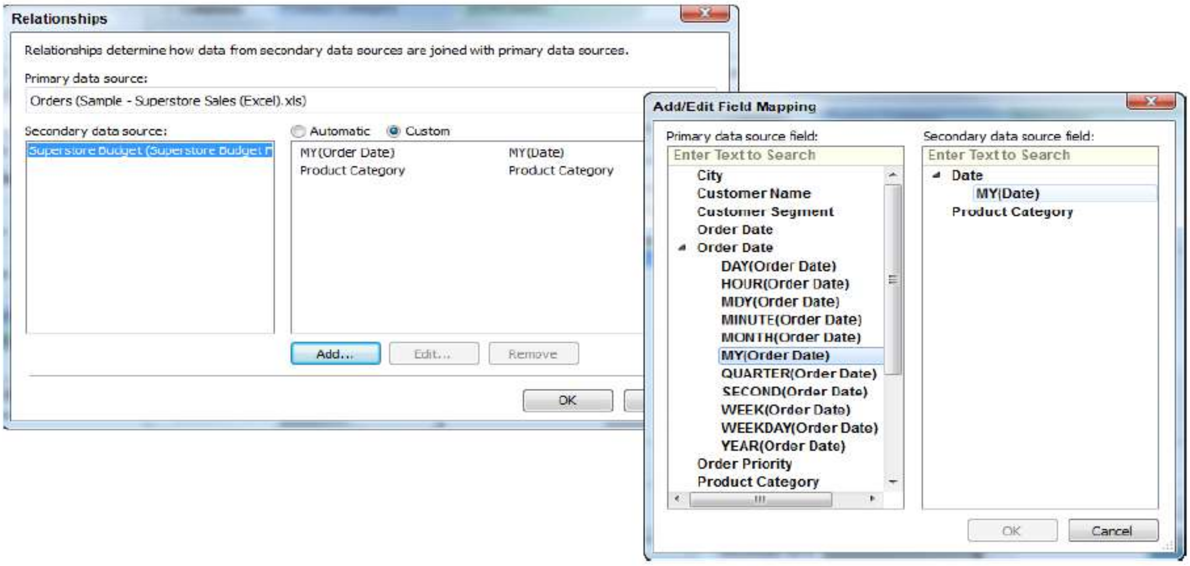


Fig14-22 *The edit relationships menu*

Select the Edit Relationships option from the data menu. This exposes the relationships window. By default, the automatic radio button option will be selected. Product category will appear automatically because that field name exists in both data sources. Since the view contains sales data by month and year for the year 2012, the custom option must be used to select the date field. Figure 14–21 displays the sales by month and year. Clicking the Add button exposes the add/edit field mapping window where the specific data aggregation can be selected from each data source. Clicking the OK button creates the second link.

Confirming that in the relationship window finalizes the links for both fields. Review the pill placements in Figure 14–21 to see where fields were placed to create the chart. The SUM (budget) pill (with the orange check mark on the marks card) was used to create the black reference line. The calculated field used to create the bar colors is displayed in the caption below the graph and is stored in the primary data source. Gray bars denote items above plan. The gray color gradient behind the sale bars comes from a reference distribution that uses color hue to show sales at 60%, 80%, 100%, and 120% of planned sales.

Fact ors that Affect the Speed of Your Data Connections

There are four areas that affect Tableau's speed:

- The Server hardware, which hosts the database
- The database, which hosts the data
- The network, over which the data is sent
- Your own computer's hardware, which has Tableau Desktop

Like any chain, the weakest link dictates overall performance.

Your Personal Computer

Tableau doesn't require high-end equipment to run. But, you will find that more internal memory, a new microprocessor, and a faster hard disk will all contribute to better performance, especially if you are accessing very large data sets. The video card and monitor resolution can contribute to the quality of how Tableau presents the visuals.

Random Access Memory (RAM)

Tableau 8 is a 32-bit application, which means the maximum memory that it can access is four gigabytes. Expanding system RAM beyond this level may not yield any benefit if you are running 32-bit Windows, but if you are running a 64-bit version of Windows you may see a performance boost if you have more RAM.

Processor

A faster processor will help Tableau's performance, but you only really get a chance to change the processor when you buy your computer. Buy the best you can justify and you should be fine.

Disk Access Times

Tableau is not normally a disk intensive program, but having a faster hard disk drive or a solid state drive (SSD) will help Tableau load faster. If you work with very wide and deep data sets that exceed your machine's internal memory capacity, it will slow down and will result in page-swapping to the hard disk drive. In this circumstance a fast hard drive will help performance.

Screen Size

The resolution of your screen will affect the level of detail that you're able to discern. The same visualization on a large, high-resolution screen may provide better insight into your data. If you have a very good monitor, you must consider that other people may be consuming your analysis with equipment that isn't as good. If they have a lower resolution video card, your visualization will not be the same on their computer.

Finally, consider the amount of work you are asking Tableau to do. While it is possible to plot millions of marks in a chart—ask yourself if all those marks add to understanding the data. If you run into performance issues, review the level of detail you're plotting. Using fewer marks in the view may improve the content's value and improve the rendering speed.

Your Server Hardware

The key consideration with regard to the specification of your server hardware is the volume and activity level you anticipate. Is your database currently deployed on the three-year-old production server with thousands of concurrent users? Does your server have other demanding applications running that may cause resource contention?

Tableau can run in the cloud and on servers that have other applications running, but as your deployment expands it is best to dedicate a server to Tableau. For massive deployments, Tableau core licenses can be divided across multiple servers.

Specifying server hardware is not a one-size-fits-all proposition. Tableau provides guidelines on their website, but each situation is unique and requires some detailed planning. In general, oversizing the hardware a little isn’t a bad idea. Tableau normally becomes very popular when it is deployed, so consider the potential for increasing demand and get professional assistance if you are unsure about the Server hardware you should purchase.

The Network

Like any other form of infrastructure (transport, power, water) data networking is a mundane but vital component for the efficient performance of any system. Networking is therefore the responsibility of specialists within your organization, and they can help you identify if there are choke points in your network that slow the performance of Tableau. For all but the very largest organizations, network capacity is seldom a bottleneck.

The Database

If you are using live connections to your data—as opposed to data extracts— the performance of your database is one of the most significant determining factors of speed. As more people in your organization use Tableau, it is important to monitor resource load on the Server, the network, and the database. Tuning your database is the responsibility of the database administrator. It is normally helpful if someone from the IT team is directly involved in the early phases of enterprise roll-outs, especially if it is expected that Tableau may create larger or different demands on the database.

If the database administrator understands the type, amount, and timing of the query loads that Tableau may generate—proper planning can ensure that system performance will not be degraded due to inadequately indexed database tables or an overloaded database server.

14.6 HOW TO DEAL WITH DATA QUALITY PROBLEMS

Why should you care about the cleanliness of your data? Inaccurate data can lead to bad decisions. Tableau is very good at visualizing data and making it understandable. If your data isn’t clean—when you connect Tableau to it, you will see the problem clearly. Fortunately, Tableau provides tools to help you deal with issues that don’t require intervention at the database-level to resolve (at least temporarily) unclean data problems. However, the best course of action when you find errors is to report them to the IT person responsible for data quality within the database you are using.

Quick Solutions in Tableau

There are several different ways you can correct data problems within Tableau that don't involve changing the source data.

Renaming

Renaming fields in Tableau is done by right-clicking on the field and renaming it. Field member names can be aliased. These changes do not alter the source database. Tableau "remembers" what you renamed without altering the source data.

Grouping

Let's assume that a company name has been entered as all of these: A&M, A & M, A and M, A+M. With Tableau you can Ctrl-Select each of these names and group them—and then create a name alias for the ad hoc grouping. So, all the versions of the name appear as one record in Tableau—A&M. This grouping and name alias will be saved as part of Tableau's metadata.

Aliases

Sometimes the name of something in the database is not a useful term for reporting purposes. For example, everybody on the team enters the customer type as P1, P2, G1, G2 where P2 denotes the size of the customer in annual revenue. For example, "Platinum level 2" could mean that the customer has an annual revenue of $1m to $5m. In Tableau, you can right-click on P2 and alias it with a more meaningful description.

Geographic Errors

Although Tableau has built-in mapping that works very well, there will be occasions when geographic locations are not recognized. Tableau will warn you by placing a small gray pill in the lower right area of your map. Clicking on that pill provides the ability to edit the offending locations or filter them out of view. This is also accessible from Tableau's map menu.

Null Values

When you see the word null appear in a view, that means Tableau can't match the record. You can filter out nulls, group them with non-null members of the set, or correct the join that is causing the null. There are many reasons why a null value could result. If you aren't sure how to correct the null, seek assistance from a qualified technical resource.

Correcting Your Source Data

Although it's quick and easy to address data quality issues directly in Tableau, It's important to bear in mind that the changes you have made in Tableau will only benefit those using the same Tableau file. There is no substitute for correcting the underlying data in the datasource. Report errors to the responsible staff quickly and provide them with your Tableau report. Expose the details so that the database is corrected.

14.7 CHECK YOUR PROGRESS

6. What is geocoding in Tableau?
7. What are the advantages of a data extract?
8. List different file types in Tableau.
9. List the factors that Affect the Speed of Your Data Connection in Tableau

Answers to Check your progress

1. If your data includes standard geographic fields like country, state, province, city, or postal codes—denoted by a small globe icon—Tableau will automatically generate the longitude and latitude values for the center points of each geographic entity displayed in your visualization

2. Data extracts don't have the advantage real-time updating that a direct connection provides, but using Tableau's data engine provides a number of benefits: Performance improvement, Additional functions, Data portability

3. Tableau Workbook (.twb), Tableau Packaged Workbook (.twbx),Tableau Data source (.tds), Tableau Bookmark (.twb) ,Tableau Data Extract (.tde)

4. There are four areas that affect Tableau's speed:
    - The Server hardware, which hosts the database

- The database, which hosts the data
- The network, over which the data is sent
- Your own computer's hardware, which has Tableau Desktop

14.8 Summary

It would be nice if all the data you needed to access resided in one place, but it Doesn't. Your data is scattered over multiple databases, text files, spreadsheets, and public services. Connecting to a wide variety of data sources directly, Tableau makes it much easier to analyze data residing in different places. Currently there are thirty-three different database connectors available with more being added every year. You can analyze spreadsheets, public data tools, analytic databases, Hadoop, and a large variety of general-purpose databases as well as data cubes.

14.9 Keywords

- **Null value -** empty
- **Aliases -** alternative
- **Table join -** combining tables
- **Left join –** return every row in the left table plus the matching ones in right table
- **Right join -** return every row in the right table plus the matching ones in left table

14.10 QUESTIONS FOR SELF-STUDY

1. Explain various ways of connecting to your data in Tableau.
2. Describe how to join database tables with Tableau.
3. Explain how to blend different data sources in a single worksheet.
4. What is the minimum hardware requirement to run Tableau?
5. Explain how to deal with data quality problems in Tableau.

14.11 References

1. Alexander Loth - Visual Analytics with Tableau-Wiley (2019)
2. Dan Murray - Tableau Your Data!_ Fast and Easy Visual Analysis with Tableau Software-Wiley (2013)
3. David Baldwin - Mastering Tableau-Packt Publishing (2017)

UNIT -15: Data Visualization

Structure

15.0 Objectives

After studying this unit, you will be able to:

- ✓ Choose between simple chart types, including bar charts, scatter plots, and line charts.
- ✓ Answer comprehensive questions with more-complex chart types including bullet graphs and waterfall charts.
- ✓ Add legends, filters, and hierarchies to your analysis.
- ✓ Follow the logic of how Tableau charts are assembled.

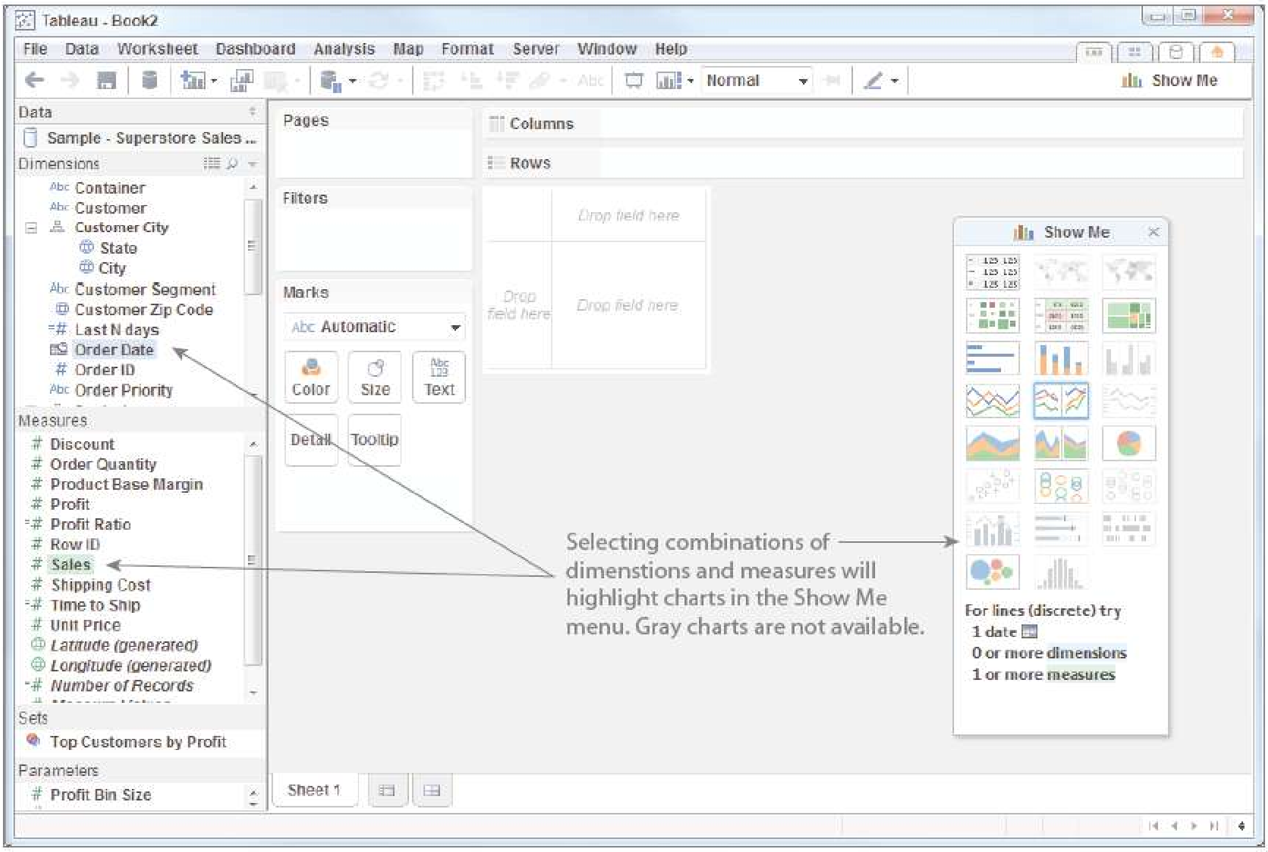
15.1 FAST AND EASY ANALYSIS VIA *SHOW ME*

Tableau's mission statement is to help you see and understand your data by enabling self-service visual analytics. The software is designed to facilitate analysis for non-technical information consumers. This is the concept behind Tableau's Show Me button. Consider Show Me to be your expert helper. Show Me tells you what chart to use and why. It will also help you create complicated visualizations faster and with less effort.

For example, advanced map visualizations are best started via Show Me because Tableau will properly place multiple dimensions and measures pills on the appropriate shelves with a single click. If you know what you want to see, Show Me will get you to your desired destination quickly.

15.2 HOW *SHOW ME* WORKS

Show Me looks at the combination of measures and dimensions you've selected and interprets what chart types display the data most effectively. Most of the examples in this section use the superstore sales Excel dataset. If you want to follow along, connect to that data source. Picking order date, sales, and then clicking Show Me will expose the options available for that combination that you see in Figure 15-1.

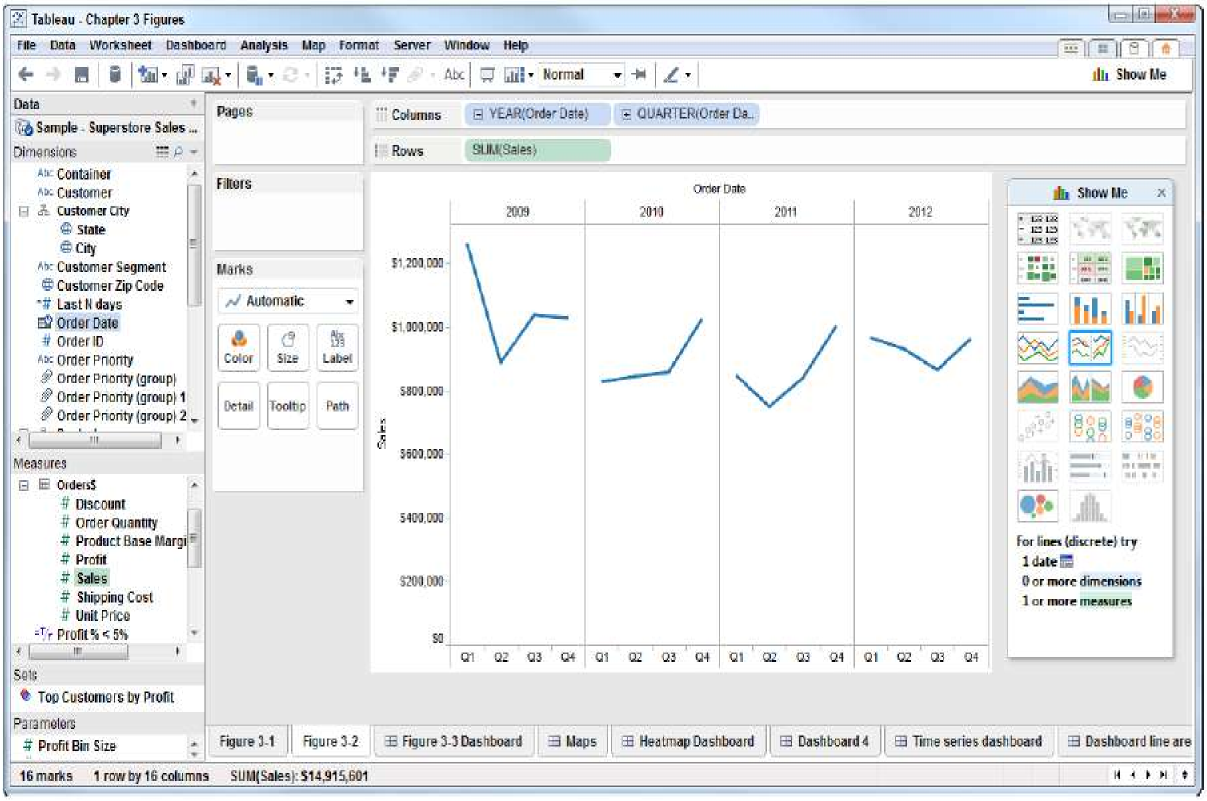


Selecting combinations of dimenstions and measures will highlight charts in the Show Me menu. Gray charts are not available.

For lines (discrete) try
1 date
0 or more dimensions
1 or more measures

Fig 15-1. *Show Me* displays chart options

Tableau recommends a line (discrete) time series chart in Show Me–denoted with a blue outline. At the bottom of the Show Me area you also see additional details regarding requirements needed for building any available chart. The time series chart requires one date, one measure, and zero or more dimensions. Selecting the highlighted chart causes the time series chart in Figure 15-2 to be displayed.



Columns: YEAR(Order Date) QUARTER(Order Da..

Rows: SUM(Sales)

For lines (discrete) try
1 date
0 or more dimensions
1 or more measures

16 marks 1 row by 16 columns SUM(Sales): $14,915,601

Fig 15-2. Discrete data time series chart

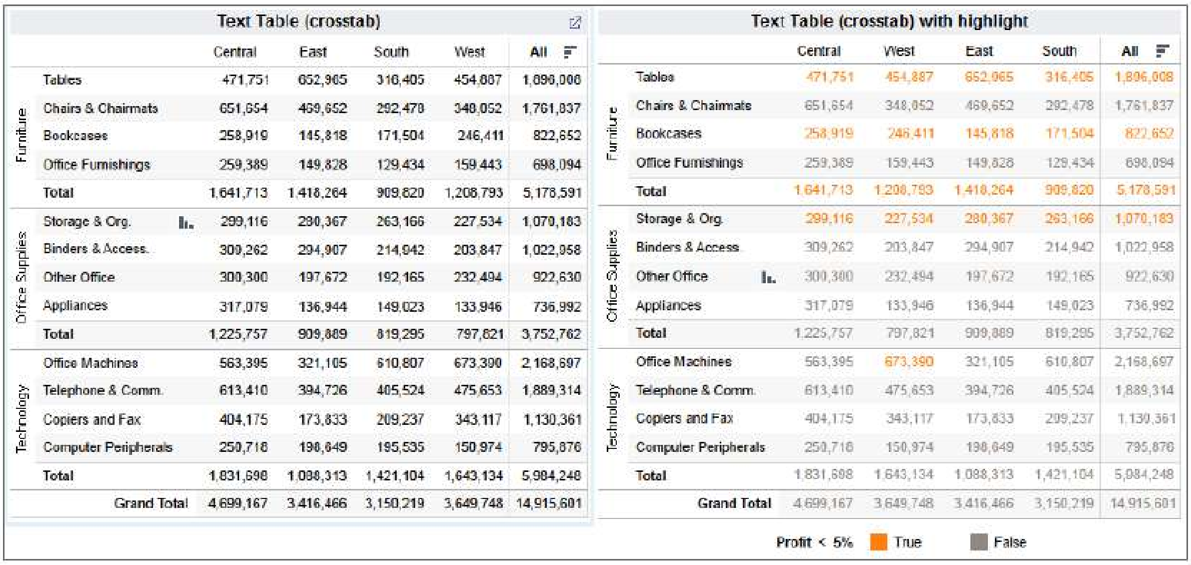
Pointing at other chart options in the Show Me menu changes the text at the bottom of the menu. This text provides guidance on the combination of data elements required for the chart being considered. Clicking on any of the highlighted Show Me icons alters the visualization in the worksheet.

Chart Types Provided by the Show Me Butt on

Show Me currently contains twenty-two chart types. Expect to see more charts added to in future releases. Advanced charts are normally variations on the basic pallet of charts you see in Show Me. Now take a look at each chart type provided by Show Me in more detail.

Text Tables (crosstabs)

Text tables look like grids of numbers in a spreadsheet. Crosstabs are useful for looking up values. Figure 15-3 shows a standard crosstab on the left. The text table on the right has been enhanced by adding a Boolean calculation to highlight items with less than five percent profit ratio. Individual cells (marks) that are greater than five percent profit ratio are gray.



Text Table (crosstab)

		Central	East	South	West	All
Furniture	Tables	471,751	652,965	316,405	454,887	1,896,008
	Chairs & Chairmats	651,654	469,652	292,478	348,052	1,761,837
	Bookcases	258,919	145,818	171,504	246,411	822,652
	Office Furnishings	259,389	149,828	129,434	159,443	698,094
	Total	1,641,713	1,418,264	909,820	1,208,793	5,178,591
Office Supplies	Storage & Org.	299,116	280,367	263,166	227,534	1,070,183
	Binders & Access.	309,262	294,907	214,942	203,847	1,022,958
	Other Office	300,300	197,672	192,165	232,494	922,630
	Appliances	317,079	136,944	149,023	133,946	736,992
	Total	1,225,757	909,889	819,295	797,821	3,752,762
Technology	Office Machines	563,395	321,105	610,807	673,390	2,168,697
	Telephone & Comm.	613,410	394,726	405,524	475,653	1,889,314
	Copiers and Fax	404,175	173,833	209,237	343,117	1,130,361
	Computer Peripherals	250,718	198,649	195,535	150,974	795,876
	Total	1,831,698	1,088,313	1,421,104	1,643,134	5,984,248
	Grand Total	4,699,167	3,416,466	3,150,219	3,649,748	14,915,601

Text Table (crosstab) with highlight

		Central	West	East	South	All
Furniture	Tables	471,751	454,887	652,965	316,405	1,896,008
	Chairs & Chairmats	651,654	348,052	469,652	292,478	1,761,837
	Bookcases	258,919	246,411	145,818	171,504	822,652
	Office Furnishings	259,389	159,443	149,828	129,434	698,094
	Total	1,641,713	1,208,793	1,418,264	909,820	5,178,591
Office Supplies	Storage & Org.	299,116	227,534	280,367	263,166	1,070,183
	Binders & Access.	309,262	203,847	294,907	214,942	1,022,958
	Other Office	300,300	232,494	197,672	192,165	922,630
	Appliances	317,079	133,946	136,944	149,023	736,992
	Total	1,225,757	797,821	909,889	819,295	3,752,762
Technology	Office Machines	563,395	673,390	321,105	610,807	2,168,697
	Telephone & Comm.	613,410	475,653	394,726	405,524	1,889,314
	Copiers and Fax	404,175	343,117	173,833	209,237	1,130,361
	Computer Peripherals	250,718	150,974	198,649	195,535	795,876
	Total	1,831,698	1,643,134	1,088,313	1,421,104	5,984,248
	Grand Total	4,699,167	3,649,748	3,416,466	3,150,219	14,915,601

Profit < 5% 🟧 True ■ False

Fig 15-3. Text tables (crosstabs)

Maps (Symbol and Filled)

Selecting a field with a small globe icon makes maps available in Show Me. Figure 15-4 shows examples of the two kinds of maps Show Me provides. Symbol maps are most effective for displaying very granular details, or if you need to show multiple members of a small dimension set. In Figure 15-4 Show Me used pie charts to display product category in the map on the left. In Filled maps it is a good idea to make the marks more transparent and add dark borders because marks tend to cluster around highly populated areas. Using the color button on the marks card to do this makes the individual marks

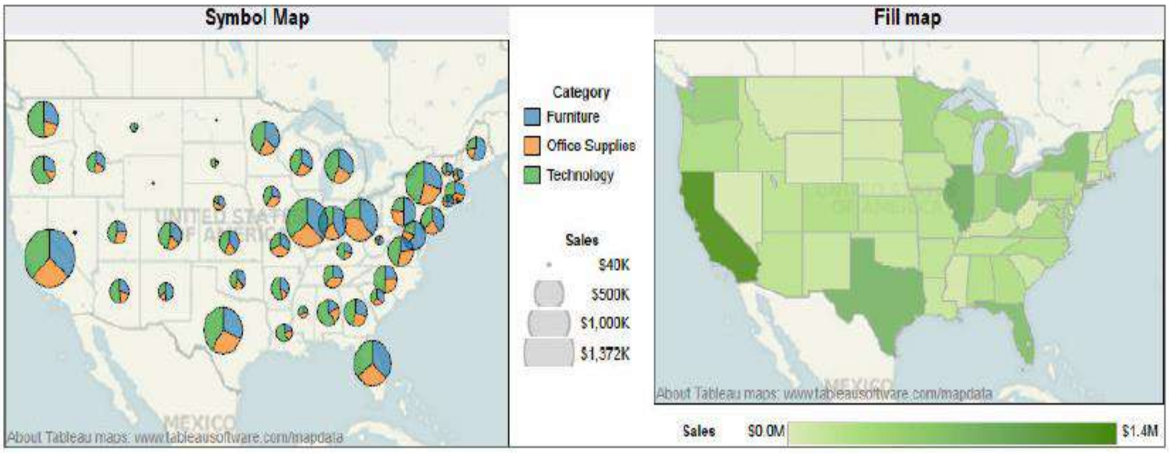
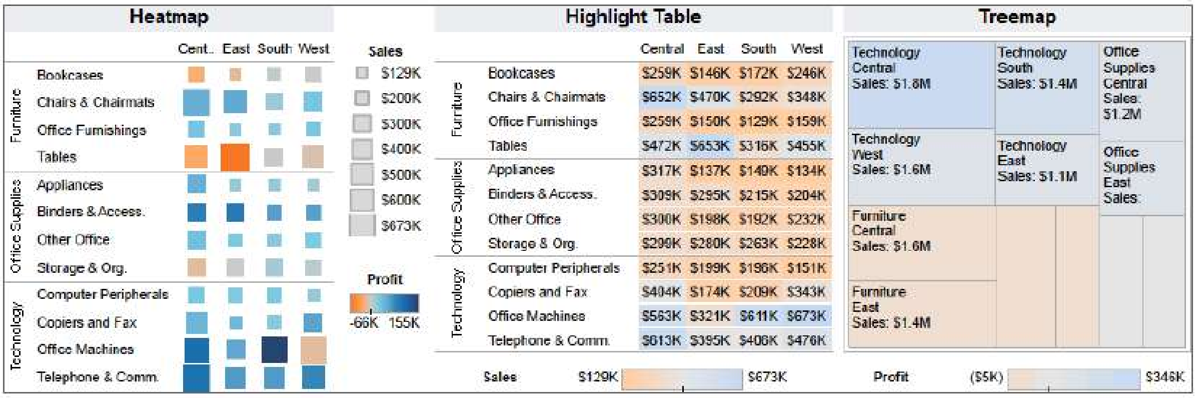


Fig.15-4. Symbol map and filled map

easier to see. The color and size legends in view are automatically provided by Tableau. Filled maps display a single measure using color within a geographic shape. If you restrict filled maps to smaller geographic areas (state, province) they effectively display more granular areas like county or postal code.

Heat Maps , Highlight Tables , Treemaps

Comparing granular combinations of dimensions and measures can be done effectively with each of these charts. Heat maps use color and size to compare up to two measures. Highlight tables can display one measure using a color gradient background to differentiate values. Treemaps effectively display larger dimension sets using color and size to display one or more dimensions and up to two measures. Figure 15-5 displays examples of each.



		Central	East	South	West
Furniture	Bookcases	$259K	$146K	$172K	$246K
	Chairs & Chairmats	$652K	$470K	$292K	$348K
	Office Furnishings	$259K	$150K	$129K	$159K
	Tables	$472K	$653K	$316K	$455K
Office Supplies	Appliances	$317K	$137K	$140K	$134K
	Binders & Access.	$309K	$295K	$215K	$204K
	Other Office	$300K	$198K	$192K	$232K
	Storage & Org.	$299K	$280K	$263K	$228K
Technology	Computer Peripherals	$251K	$199K	$196K	$151K
	Copiers and Fax	$404K	$174K	$209K	$343K
	Office Machines	$563K	$321K	$611K	$673K
	Telephone & Comm.	$613K	$395K	$406K	$476K

Fig 15-5. Heatmap, highlight table and treemap

These charts, and text tables, can also be used to replace quick filters on dashboards—providing more information in the same space that a multi-select filter would require.

Bar Chart , Stacked Bar, Side -by -Side Bars

These charts facilitate one-to-many comparisons. Figure 15-6 includes examples of each. Bar charts are the most effective way to compare values across dimensions— their linear nature making precise comparisons easy. Stacked bar charts should not be used when there are many different dimensions because they can be overwhelming if too many colors are plotted in each bar. Side-by-side bars provide another way to compare measures across and dimensions on a single axis.

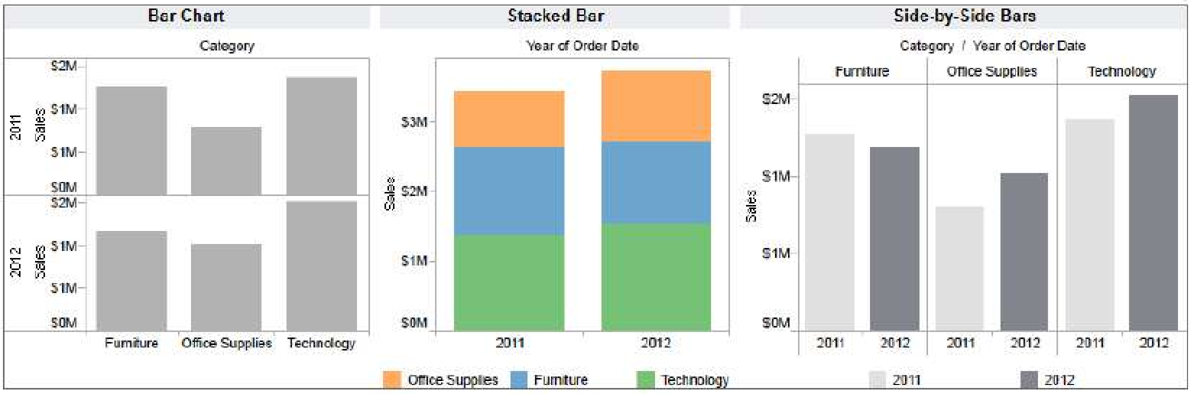


Fig 15-6. Bar chart, stacked bar chart and side by side bar chart

Line Charts For Time Series Analysis

Line charts are the most effective way to display time series data. One variable to consider when presenting time series is the treatment of time as a discrete (bucketed) entity or as a continuous (unbroken) series progression. Discrete line charts place breaks between time units (year, quarter, and month). Most people are familiar with time series charts that are presented in unbroken lines. Figure 15-2 presents a single measure (sales) using a discrete time series. Time is presented discretely by quarters within each year. Figure 15-7 provides three different time series charts that are plotting two measures with a continuous time axis. Figure 15-7 *Time series presented using continuous time*

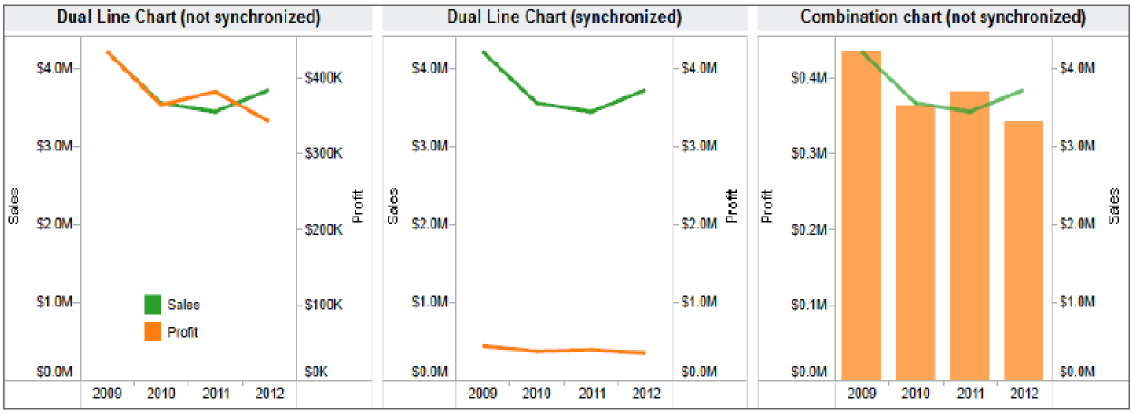


Fig 15-7 *Time series presented using continuous time*

The dual line chart presents two measures (sales and profit) using asynchronous axis ranges. Show Me assumes dual axis charts will be used to present values that are dissimilar and plots the marks using different axis ranges. The middle dual line chart, with synchronized axis, provides a better comparison of the relative values of sales and profit. The combination chart, using a bar for profit and line for sales, maintains asynchronous axis ranges, but the use of different mark types accentuates that there are different measures being plotted.

Area Fill Charts and Pie Charts

Figure 15-8 provides a comparison of lines, area fill, and pie charts. Compare the value of each kind of chart for displaying the information. All three charts are plotting the same data using Show Me to create the charts. Which one do you prefer?

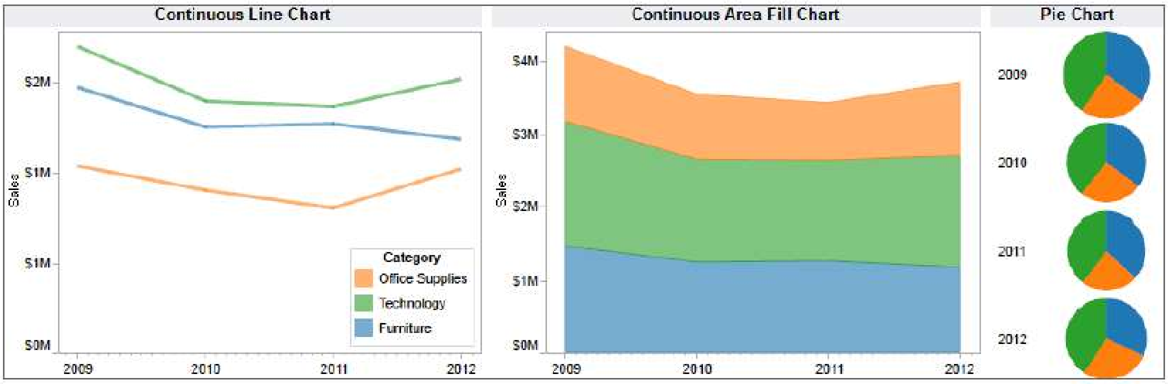


Fig 15-8 Continuous line chart, area fill chart, pie charts

The line chart facilitates accurate comparison of the relative sales by category. Since the area fill chart plots sales values as bands, it is easy to misinterpret the top band as being the largest value in the set. Area fill charts are best used for plotting a single dimension to avoid misunderstanding. Pie charts should be used for getting a general sense of magnitude and not for precise comparisons. A more effective use of a pie chart and area fill chart is provided in Figure 15-9

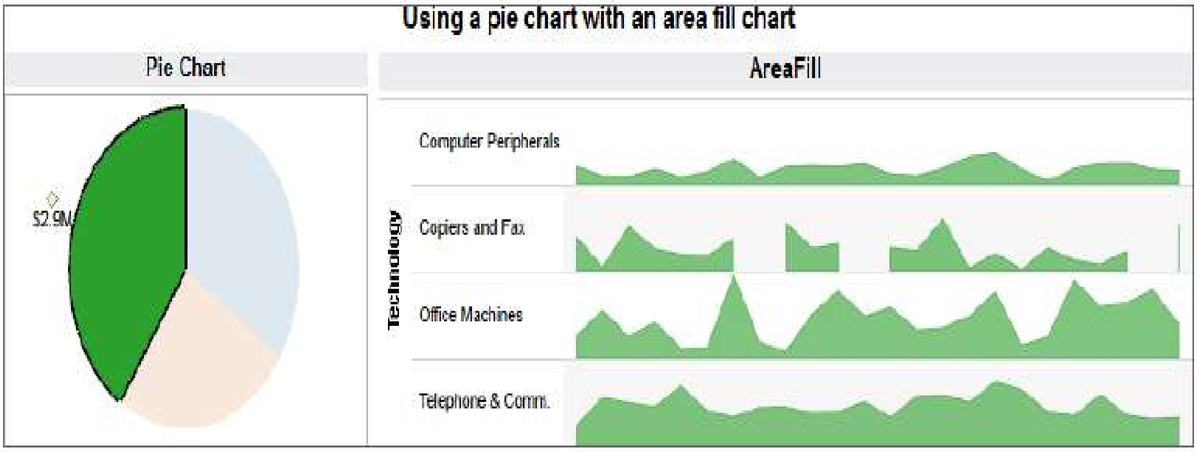


Fig 15-9. Pie chart and area fill chart

By limiting the area fill chart to one dimension on each axis and using a pie chart with only three slices—the combination of chart types presents the information effectively. The pie chart acts as a filter for the area fill chart in Figure 15-9. If you have limited space and are sure that your pie's slices won't be tiny, pie charts can be used effectively as filters.

Scatter Plot, Circle View , and Side -by -Side Circle Plots

Enabling analysis of granular data across multiple dimensions, Scatter plots, Circle views, and Side-by-Side circles can be used to identify outliers. Figure 15-10 provides example of each.

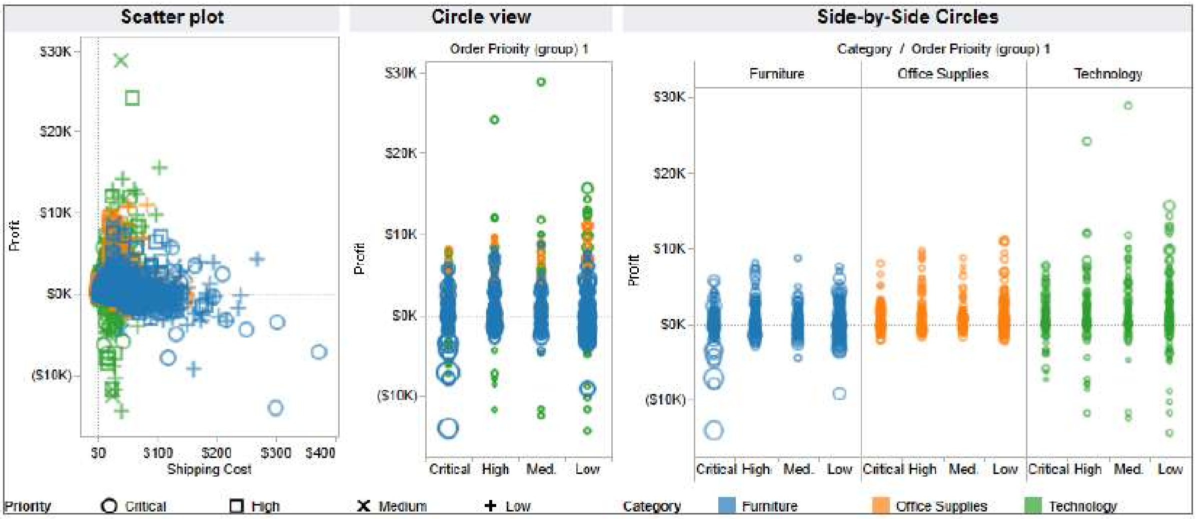


Fig 15-10 scatter plot, circle view, side-by-side circle view

All three charts in Figure 15-10 are plotting over four thousand marks in a very small space. The scatter plot uses two axes for comparing profit and shipping cost. Color and shape provide insight into two dimensions. Size isn't being used in the example but could be used for a third measure. The circle view uses one axis to plot a single measure. In both circle plots size is used to denote shipping cost amount. The side-by-side chart provides a more granular breakdown of the product categories.

Bullet Graph , Packed Bubble , Histogram , and Gantt Charts

The last four chart types provided by Show Me are completely different tools. Figure 15-11 shows them together but their uses are very different.

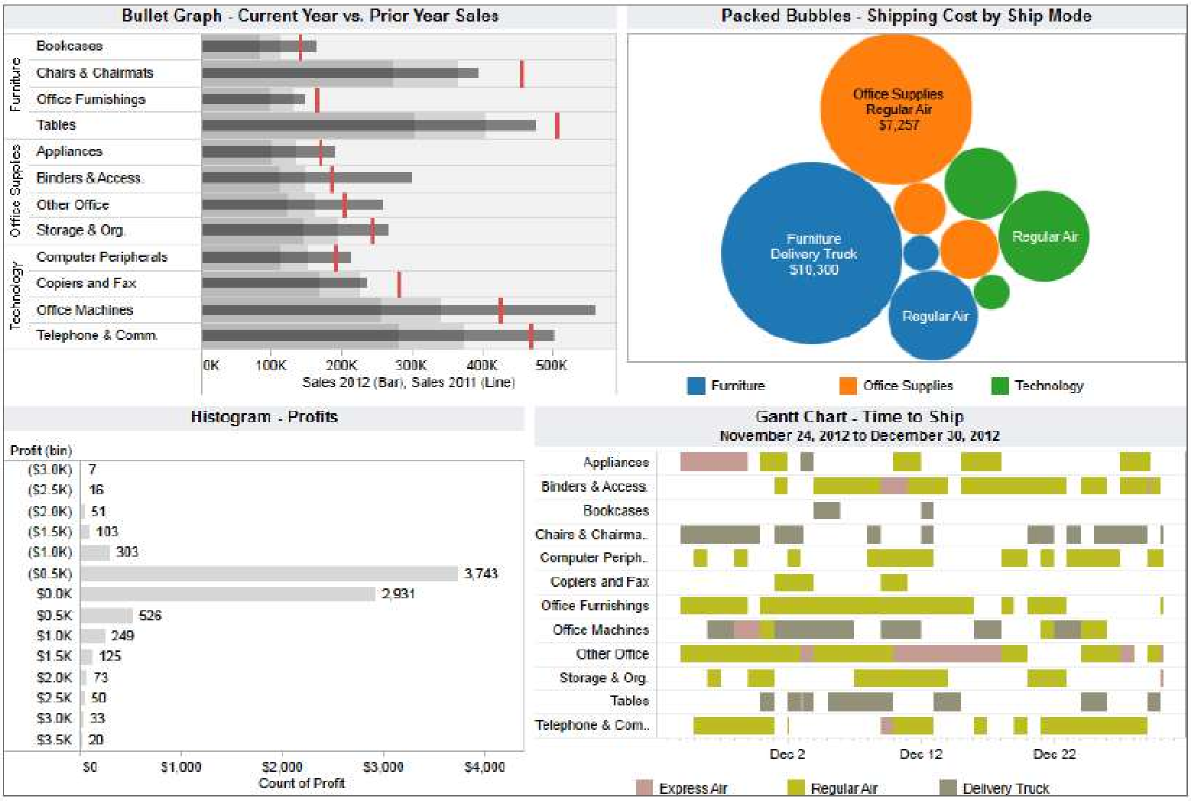


Fig 15-11. The last four Shoe Me Charts

Bullet graphs are bar charts that include a reference line and reference distribution for each cell in the plot. In the example, current year sales (bars) are compared to prior year sales (red reference lines). The color band behind the bar represent sixty and eight percent of the prior year sales. Bullet graphs pack a lot of information into a small space.

Bubble charts offer another way to present one-to-many comparisons by using size and color. They can be interesting to look at but do not allow for very precise comparisons between the different bubbles. For this reason limit the situations that don't require precise visual ranking of the bubbles. Histograms turn normally continuous measures into discretely-bucketed bins of numeric values. The example histogram breaks down profits into five hundred dollar increments. The bar's length shows the number of orders that fall within the band.

You've probably seen Gantt charts before being used in project planning. The length of each bar color in the example displays a time duration for an activity. These are particularly useful when you want to visualize the timing and duration of events. In the example, the length of the bar is the duration of time required to complete a shipment. The starting position of the bar is the date the order was received. Using Tableau View Structure to Create New Data when you are new to Tableau and don’t completely grasp how each shelf affects your chart’s appearance, Show Me will help you build charts without having to understand the mechanics. Show Me helps everyone achieve desirable results quickly, and it helps you gain an understanding of the mechanics of how each shelf and field type can change the appearance of your visualizations. Once you have a chart in view, you can use that chart structure to add additional information. Two common ways to do this are by adding trend lines or reference lines to your chart. The numbers used to derive trend lines and reference lines can come from the view in Tableau itself and don’t necessarily require that the data exist in your data source.

15.3 TREND LINES AND REFERENCE LINES

Visualizing granular data sometimes results in random-looking plots. Trend lines help you interpret the data by fitting a straight or curved line that best represents the pattern contained within detailed data plots. Reference lines provide visual comparisons to benchmark figures, constants, or calculated values that provide insight into marks that don’t conform to expected or desired values. Trend lines help you see patterns in data that are not apparent when looking at your chart of the source data by drawing a line that best fits the values in view. Reference lines allow you to compare the actual plot against targets or to create statistical analyses of the deviation contained in the plot; or the range of values based on fixed or calculated numbers. Trend lines help you see patterns that can provide predictive value. Reference lines alert you to outliers that may require attention or additional analysis. Figure 15-12 provides examples of a trend line and a reference line.

The chart on the left employs a linear regression line to plot the trend in volatile weekly sales figures The pattern of sales is volatile—making it difficult to see the overall pattern. The trend isn’t very pronounced, but the trend line helps you see that sales are trending down slightly. How reliable is the trend line plot? That question can be answered by pointing at the trend line and reviewing the statistical values displayed or by pointing at the trend line, right-clicking, and selecting Describe Trend Model. Figure 15-13 shows the more detailed description of the trend model statistics.

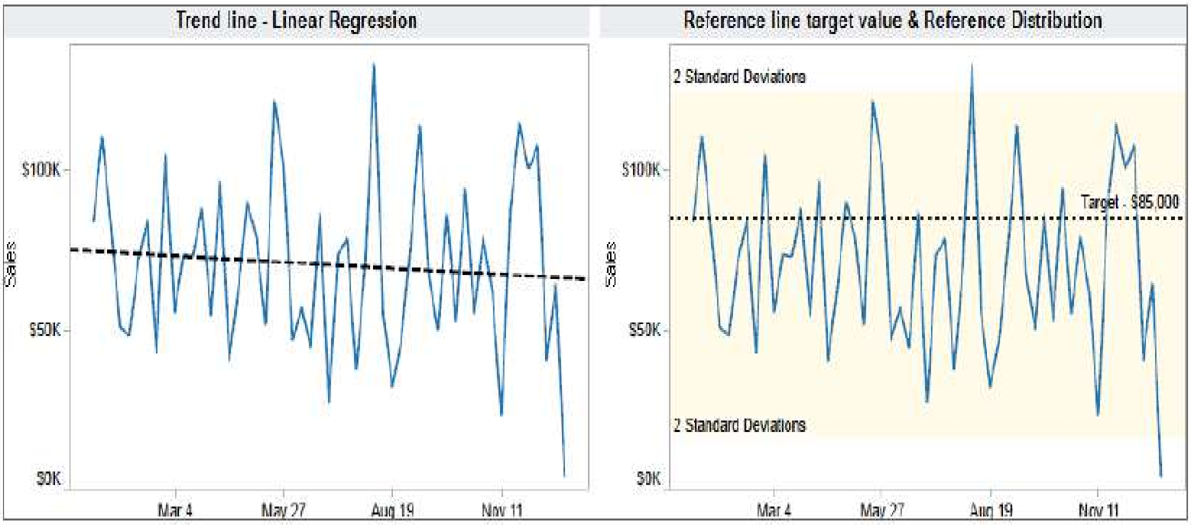
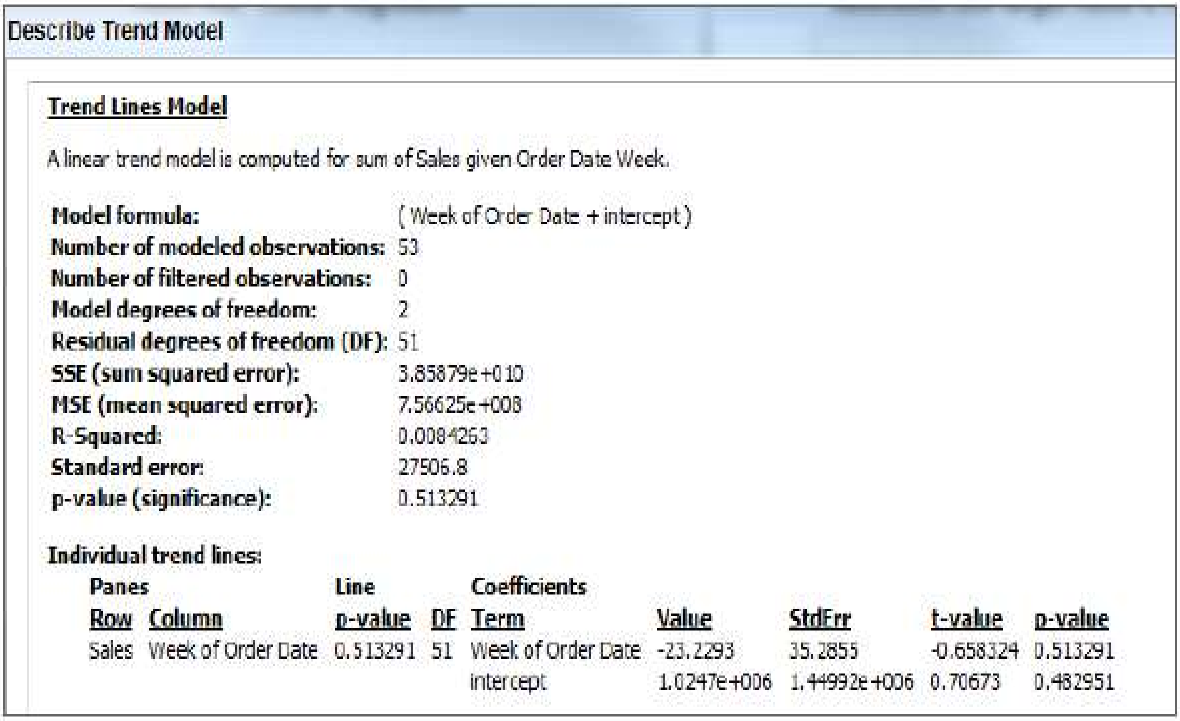


Fig. 15-12. Trend line and reference line



Describe Trend Model

Trend Lines Model

A linear trend model is computed for sum of Sales given Order Date Week.

```
Model formula:                      ( Week of Order Date + intercept )
Number of modeled observations:     53
Number of filtered observations:    0
Model degrees of freedom:           2
Residual degrees of freedom (DF):   51
SSE (sum squared error):            3.85879e+010
MSE (mean squared error):           7.56625e+008
R-Squared:                          0.0084263
Standard error:                     27506.8
p-value (significance):             0.513291

Individual trend lines:
     Panes                       Line            Coefficients
     Row    Column               p-value   DF    Term                 Value          StdErr         t-value     p-value
     Sales  Week of Order Date   0.513291  51    Week of Order Date   -23.2293       35.2855        -0.658324   0.513291
                                                 intercept            1.0247e+006    1.44992e+006   0.70673     0.482951
```

Fig 15-13. Describe the trend model

Describing the trend model exposes the statistical values that describe the trend line plot. If you are a statistician all the figures will mean something to you. If you aren't a statistics expert, focus on the P-Value and R-Squared figures. They help you evaluate the reliability and predictive value of the trend line plot. If the P-Value is greater than .05, then the trend line doesn't provide much predictive value. R-Squared provides an indicator of how well the line fits the individual marks. The linear regression trend line displayed on the left side of Figure 15-12 clearly doesn’t have much predictive value (P-Value is .513291, which implies a confident interval of less than 50%), nor does the line fit the marks particularly well. The R-Square value (.008) is very low indicating that the plot doesn’t fit the marks very precisely. Tableau does the best job it can fitting the line to the plot, but if the marks are randomly scattered, the R-Squared value will be low. The combination of low P-Value and R-Squared value means that the trend line on the left side of Figure 15-12 does not provide much predictive value.

The chart on the right in Figure 15-12 uses the same data as the chart on the left but this time a reference line has been applied to show the target value of $85,000. A reference distribution has also been calculated to show two standard deviations from the mean value of the plot. Assuming the data is normally distributed—marks outside of that range indicate abnormal variation that would warrant further investigation to determine the cause of the variance. You don’t need to become a statistics expert to use trend lines and reference lines. But, understanding the basics will certainly help you interpret what the plots indicate. A web search will provide more details regarding the mathematics if you are interested.

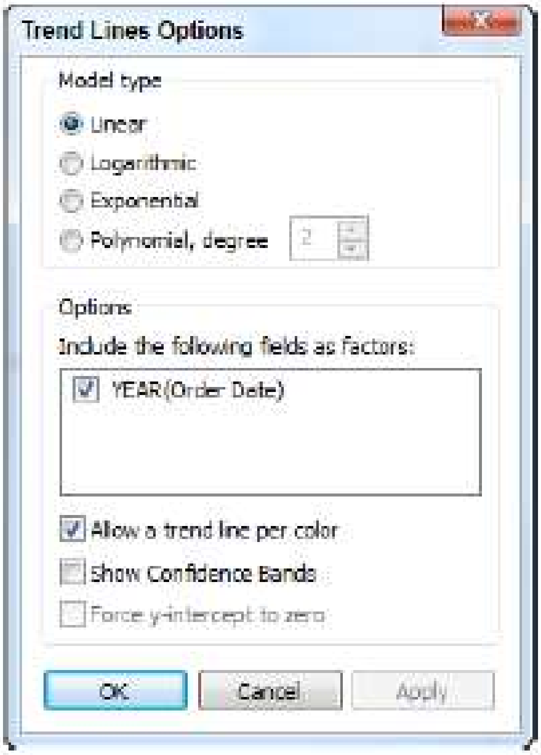
Adding Trendlines and Reference Lines to Your Charts

There are many options available for presenting trend lines and reference lines in Tableau. Take a look below at each in more detail.

Trend Lines

Add a trend line to your visualization by right-clicking on the white space in the worksheet and selecting the menu option Trend Lines/Show Trend Lines. This adds a linear regression line to the chart. More trend line options are available if you point at the trend line, right-click, and select Edit Trend Lines. This exposes the trend line menu in Figure 15-14.

The trend line menu provides options for changing the trend line type. If your chart uses color to express a dimension, you can choose to create separate trend lines for each colored line in the view—or not. Selecting Show Confidence Bands adds upper and lower bounding lines based on the variation of the data. If you’re applying trend lines in charts like scatter plots, you can also force the trend line to intercept the vertical y-axis at zero.



Trend Lines Options

Model type

● Linear

○ Logarithmic

○ Exponential

○ Polynomial, degree 2

Options

Include the following fields as factors:

☑ YEAR(Order Date)

☑ Allow a trend line per color

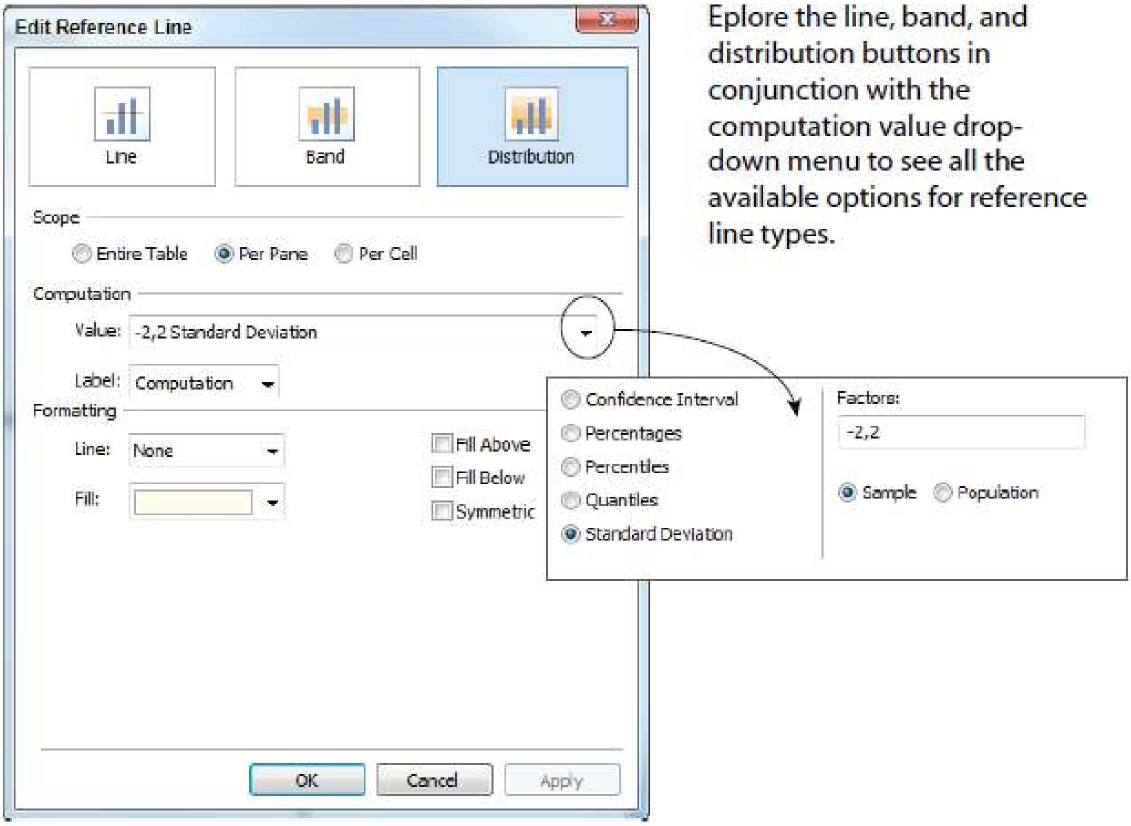
□ Show Confidence Bands

□ Force y-intercept to zero

Fig 15-14. The trend line options menu

Reference Lines

There are many different options for reference lines and you can apply more than one reference line to an axis. To add a reference line, right-click on the axis from which you want to apply the reference line. Be careful to point at the white space and not at a title or axis label. Figure 15-15 shows the reference line menu selections used to add the standard deviation reference distribution displayed in the time series plot on the right side of Figure 15-12.



Eplore the line, band, and distribution buttons in conjunction with the computation value drop-down menu to see all the available options for reference line types.

Edit Reference Line

Scope

○ Entire Table ● Per Pane ○ Per Cell

Computation

Value: -2,2 Standard Deviation

Label: Computation

Formatting

Line: None

Fill:

□ Fill Above
□ Fill Below
□ Symmetric

○ Confidence Interval
○ Percentages
○ Percentiles
○ Quantiles
● Standard Deviation

Factors: -2,2

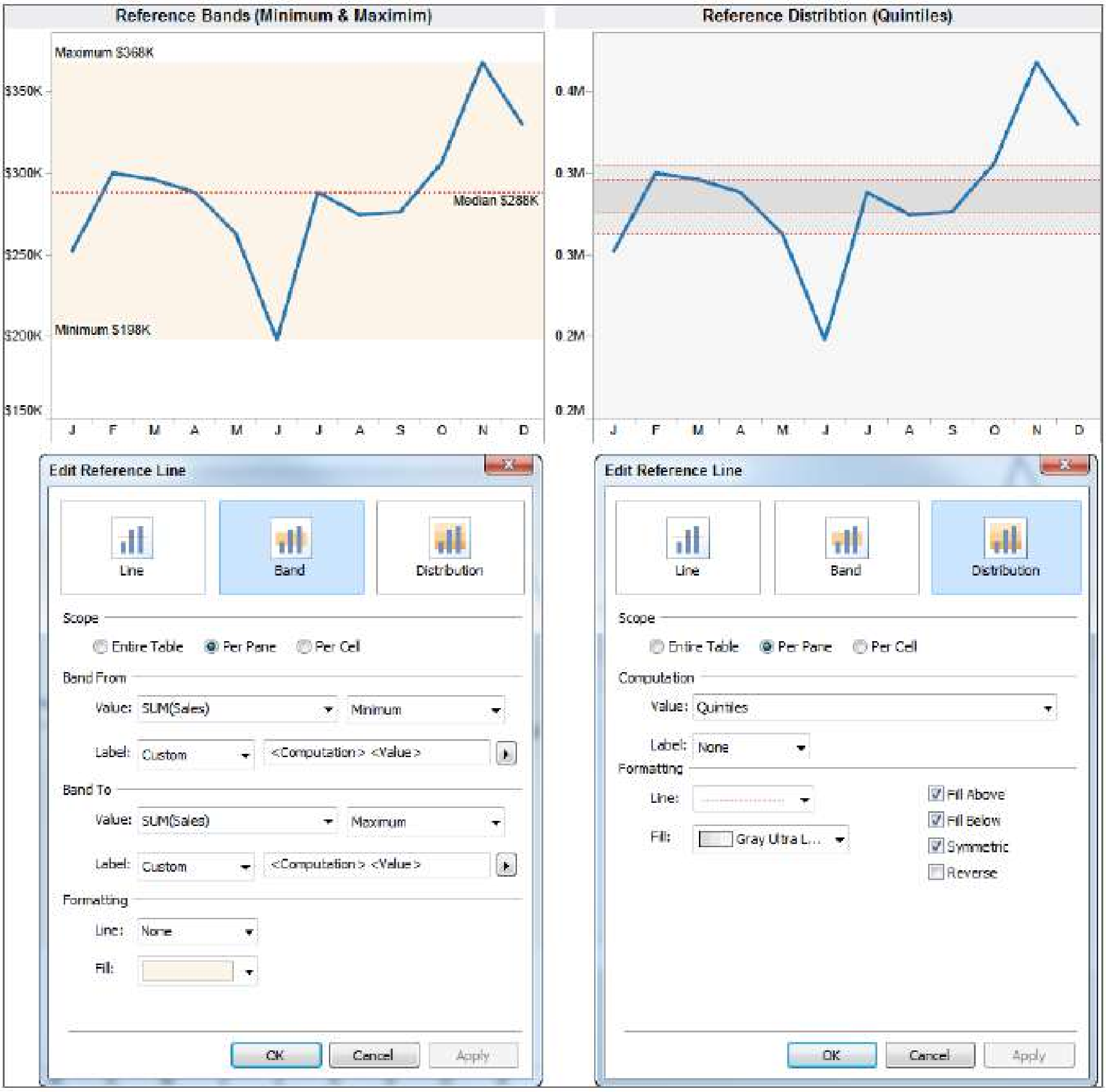
● Sample ○ Population

Fig 15-15 Reference line menu standard deviation reference lines

The same chart in Figure 15-12 includes a second reference line that displays a constant. This was added by selecting the reference "line" type to display a manually-entered constant value of $85,000. Two more reference line examples along with the related reference line menu selections can be seen in Figure 15-16.

The example on the left in Figure 15-16 combines a reference line displaying (median) with reference bands for maximum and minimum values. The chart on the right side of Figure 15-16 uses a reference distribution to plot quintile ranges. Note the use of the Symmetric Color option. Selecting this causes the color bands outside of the widest quintile lines to use the same color hue. If Symmetric Color wasn't selected, the band color would get darker from top to bottom. Alternatively, if the symmetric options were left unchecked and the reverse was selected, the color bands would get lighter from top to bottom.

Applying color fill above or below reference lines calls attention to specific areas of your visualization. Use trend lines and reference lines in moderation. They add insight to your visualizations but too many reference lines can lead to chart clutter and make it more difficult to understand.



Reference Bands (Minimum & Maximim)

Reference Distribtion (Quintiles)

Edit Reference Line

Line | Band | Distribution

Scope

○ Entire Table ● Per Pane ○ Per Cell

Band From

Value: SUM(Sales) Minimum

Label: Custom <Computation> <Value>

Band To

Value: SUM(Sales) Maximum

Label: Custom <Computation> <Value>

Formatting

Line: None

Fill:

OK Cancel Apply

Edit Reference Line

Line | Band | Distribution

Scope

○ Entire Table ● Per Pane ○ Per Cell

Computation

Value: Quintiles

Label: None

Formatting

Line:

Fill: Gray Ultra L...

☑ Fill Above

☑ Fill Below

☑ Symmetric

☐ Reverse

OK Cancel Apply

Fig 15-16 Reference bands and reference distributions

Why the Concept of Scope is Important

Understanding how the scope in trend line and reference line calculations determines the resulting appearance of the line is important not only for the deriving trend and reference lines, but for understanding how Calculated Values and Table Calculations work in Tableau. Figure 15-17 includes a time series chart on the left that contains two different reference lines and the bullet graph on the right contains a single reference line for each bar (cell) in the view.

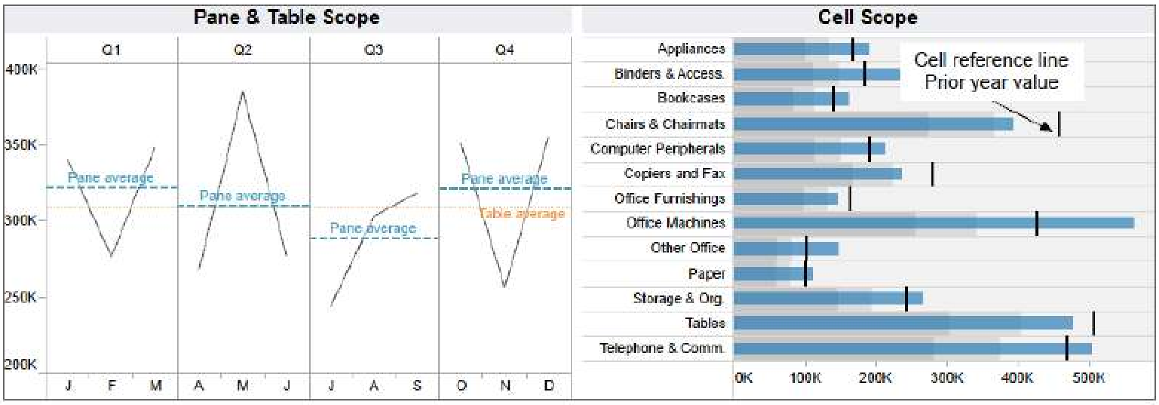


Fig. 15-17 Reference lines using entire table, pane and cell

The time series chart on the left employs discrete dates to create panes by quarter. Tableau outlines the panes using gray lines. The scope that the calculation Tableau uses to create the orange dotted reference line is the table. It shows the average value for the entire table. The scope of the blue dashed line is using the quarter panes to derive that reference line. By coincidence the table average and the pane average lines overlap in the second quarter. In all other quarters in the view, the pane average differs from the average for the entire year (table scope). The bullet graph on the right compares current year values (blue bars) with prior year values plotted using thick black reference lines. Those reference lines are applied using cell scope.

Changing the Scope of Trend Lines

Scope can also be used to change the appearance of trend lines. Figure 15-18 includes examples of trend lines that are applied by pane, and for the entire table.

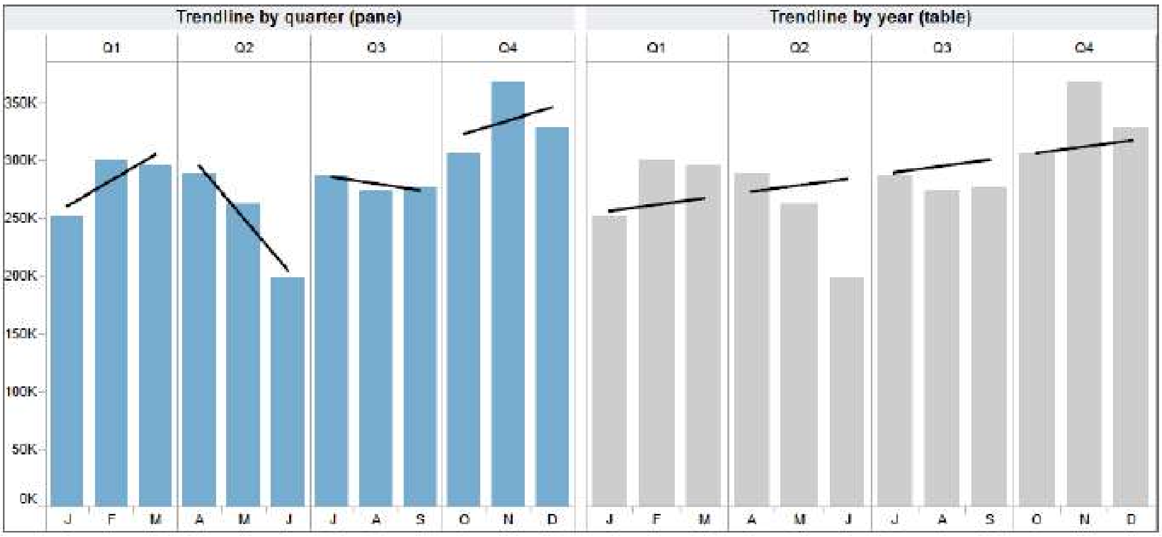
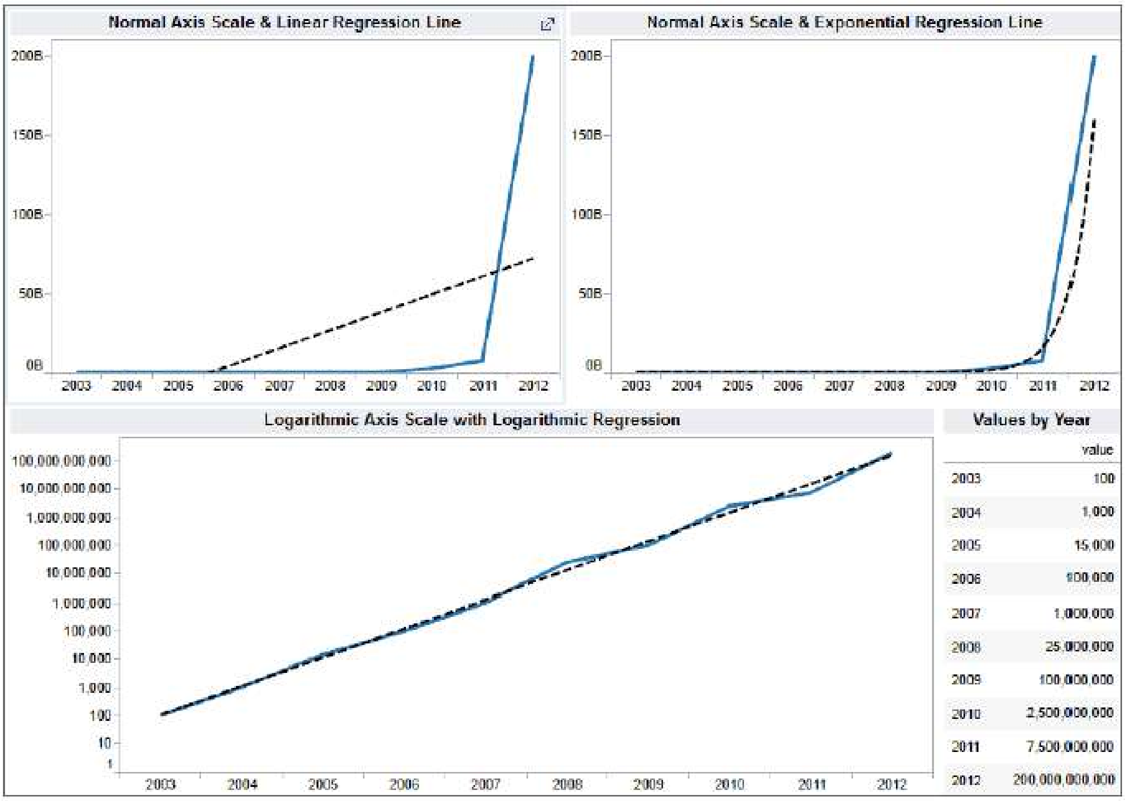


Fig 15-18 Trend lines using pane and table

Tableau provides four different kinds of trend lines (linear, logarithmic, exponential, and polynomial). Most people are accustomed to seeing linear (straight) regression lines in time series data. Polynomial regression provides a more curved line. Increasing the degrees of freedom will make the trend line follow the plot of the individual marks more closely. Logarithmic and exponential regression normally results in curved lines.

Different Trend Line and Axis Types

One reason for using trend lines is predictive analysis. To help you see a possible future condition. The choice of method for calculating trend lines requires some professional judgment and is dependent on the data. People associate the word “exponential” with rapid growth. A real-world example of this is provided by rapid advance of computing power over the past 40 years. Plotting numbers that change drastically and making those figures easy to interpret can be challenging. Figure 15-19 shows three different ways to plot a rapidly changing data set.



Values by Year

	value
2003	100
2004	1,000
2005	15,000
2006	100,000
2007	1,000,000
2008	25,000,000
2009	100,000,000
2010	2,500,000,000
2011	7,500,000,000
2012	200,000,000,000

Fig 15-19 Rapidly increasing time series

You can tell by looking at the top two time series plots that the values plotted are increasing very rapidly over a ten-year period. These charts use a linear axis scale. In the top left chart a linear trend line is also used to smooth the data. The top right chart uses an exponential regression line. It's obvious that the exponential trend line fits the data better. The bottom chart utilizes a logarithmic axis scale, which was altered by right-clicking in the white space of the axis and picking the logarithmic scale option. The trend line is also computed using logarithmic regression.

Tableau's logarithmic axis scale makes it easier to compare very different values in the same chart. The logarithmic regression line also makes it easier to see what next year's value might be. If you feel that logarithmic or exponential trend lines might benefit your analysis, you should arm yourself with the technical expertise to explain what the lines mean. As with all statistics, judgment should be applied. History may not repeat.

If you know a friendly statistician, ask them to explain the underlying theory and math. Alternatively, go to Kahn Academy's website https://www.khanacademy .org/math/probability/regression and watch the videos related to regression, statistics, and probability. Unless you understand the statistics supporting exponential and logarithmic smoothing, you should stick to what you feel comfortable explaining to your audience.

15.4 SORTING DATA IN TABLEAU

Tableau provides basic and advanced sorting methods that are easily accessed through icons or menus. Sorting isn’t limited to fields that are visible in the chart—any field in the data source can be used for sorting.

Manual Sorting via Icons

The most basic way to sort is via the icons that appear in the toolbar menu. The toolbar menu sort icons provide ascending and descending sorts. Figure 15-20 shows a bar chart in which a manual sort was applied from the toolbar icon.

Tableau also provides sorting icons near the headings and mark axis. If you Don’t see an icon, hover your mouse near the area and it will appear. Notice the icon that appears in the sub-category pill on the row shelf? The light gray descending sort icon that appears in that pill provides an indication that a sort has been applied on that sub-category field.

Clicking on the sort icon floating over the right-side of the sub-category heading provides ascending and descending sorts using the text of the product category headings. The sort icons that appear over and under the mark (bar) axis provide ascending and descending sorts based on the values displayed by the marks, and also add data source order sorting.

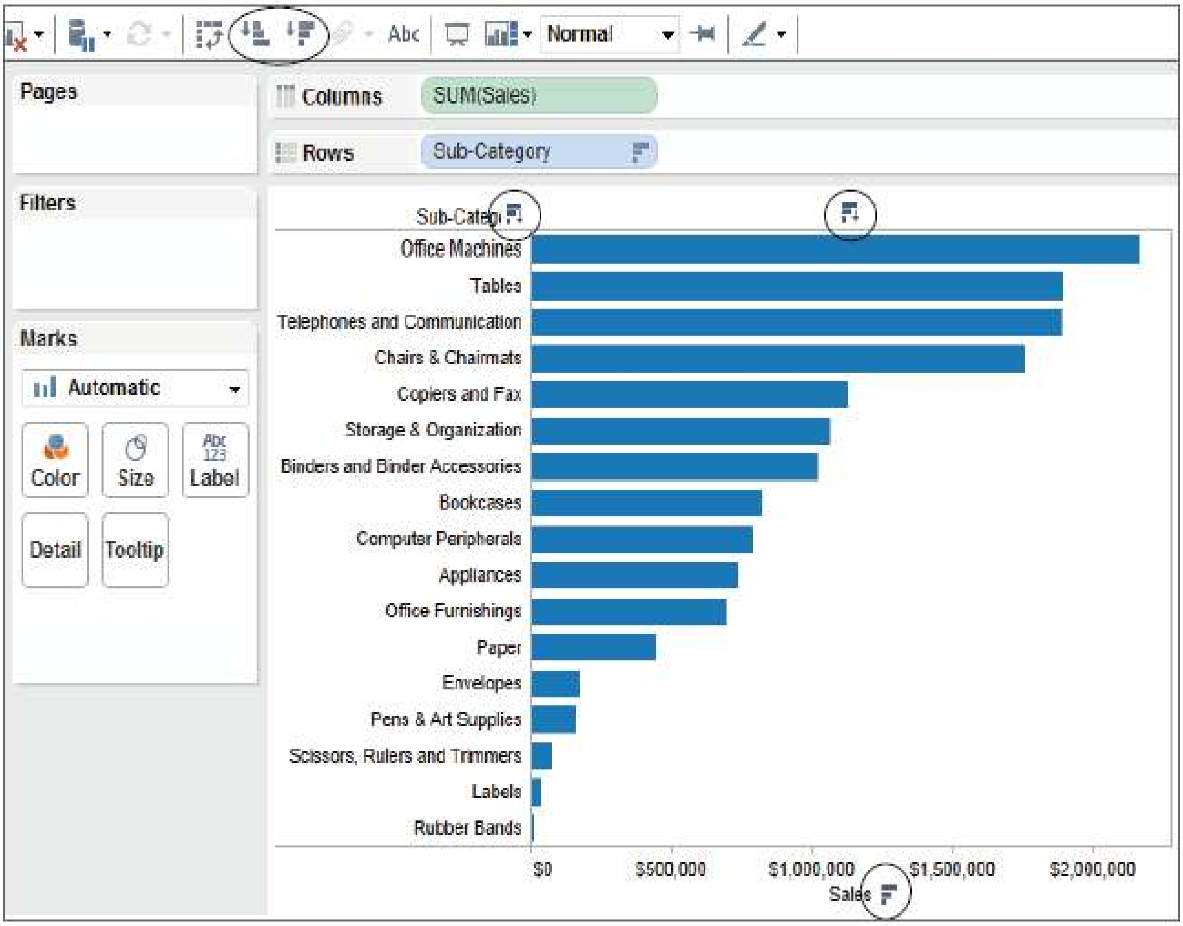
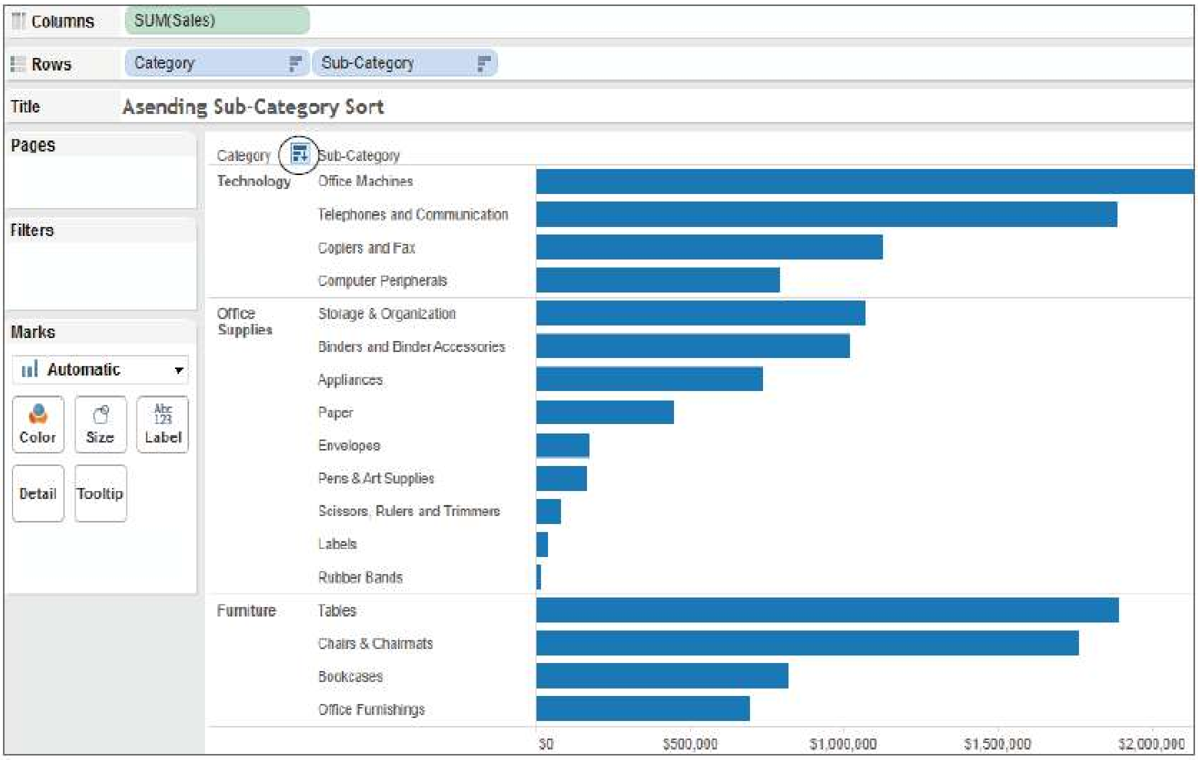


Fig. 15-20. Manual sorting applied from the toolbar icon

It doesn’t matter how many levels of hierarchy are added to the view, you can sort on each level. Figure 15-21 includes the category dimension and that pill has been sorted using an ascending sort.



Columns SUM(Sales)

Rows Category Sub-Category

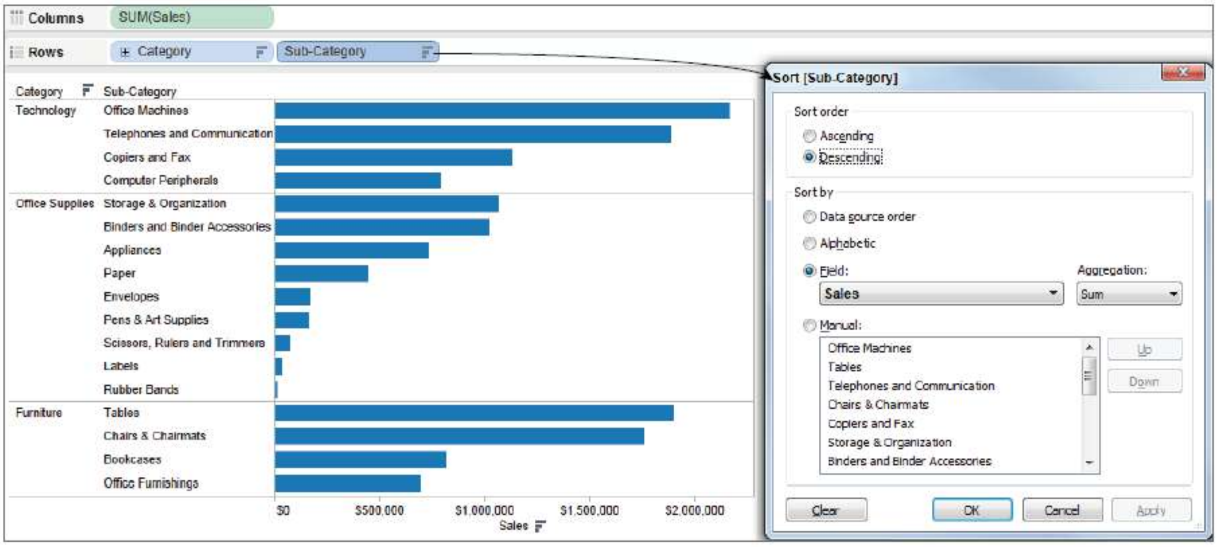
Title Asending Sub-Category Sort

Fig. 15-21 Ascending sort by category

In addition to sorting via the toolbar or the sort icons, you can point at and drag any one of the rows in the display and revise the sort to an arbitrary manual sort. For example, you could change the sort order by dragging computer peripherals to the top of the technology category and defining a new manual sort.

Calculated Sorts Using the Sort Men u

More advanced sorting can be accessed by pointing at a dimension pill, right clicking, and selecting the Sort option. Figure 15-22 shows the sort menu that displays when you right-click on a dimension pill--in this example the Sub- Category pill.



Sort [Sub-Category]

Sort order

○ Ascending

● Descending

Sort by

○ Data source order

○ Alphabetic

● Field: Sales — Aggregation: Sum

○ Manual:

Fig. 15-22 Resorted computer peripherals and sort menu

Tableau's sort menu allows you to more precisely define the default sort method and order. The sort by section includes a drop-down menu that currently displays the sales field using an aggregation of sum. However, it is possible to select any field in the data set and change the aggregation. For example, you could also apply ascending sort by average profit. Leaving the sort menu open and using the apply button at the bottom right side of the menu is useful. You can apply a variety of sort options and see the result. When you decide to keep the sort, click the OK button.

Sorting via Legends

Another useful sort feature is enabled within legends. Figure 15-23 shows two versions of the same bar chart. The left view orders the blue delivery truck dimension on the bottom. The chart on the right shows regular air at the bottom. Reordering the position of the colors displayed within the color legend causes the order of the colors appearing in the bars to change. Reposition the colors within the color legend by pointing at a color, holding down the left mouse button, and dragging the color to the desired position.

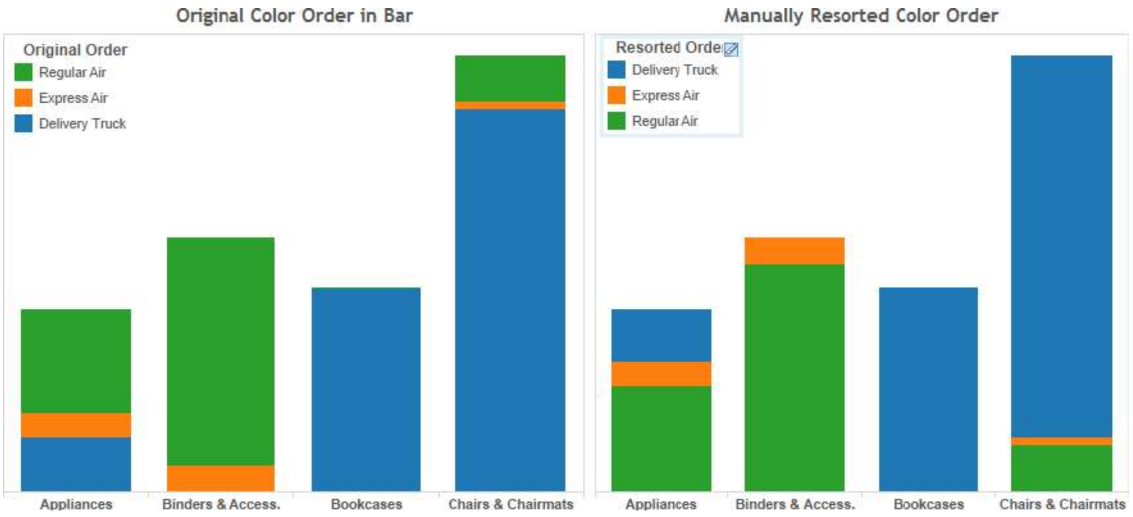


Fig. 15-23 Reordering the colors in chart

The ability to reorder colors in a stacked bar chart is important because precise comparisons are most easily made for the color that starts at the zero point on the axis. All of the other colors are not as easily compared because they don't start at the same value.

15.5 ENHANCING VIEWS WITH FILTERS, SETS, GROUPS, AND HIERARCHIES

Sorting isn't the only way to arrange data. Creating drill-down hierarchies is easy in Tableau. Perhaps your data includes a dimension set with too many members for convenient viewing. Grouping dimensions within a particular field is available. Interacting with your data may uncover measurement outliers that you would like to save and reuse in other visualizations. That capability is enabled via sets. Even groups of sets can be created on-the-fly

Making Hierarchies to Provide Drill -Down Capability

Hierarchies provide a way to start with a high-level overview of your data, and then drill down to lower levels of detail on demand. In Figure 15-21 you can see a two-level view of the data that included product category and then subcategory. That presentation may include more detail than you prefer to see. A hierarchy that combines category and subcategory can address both needs. Figure 15-24 uses a hierarchy to show category first and subcategory on demand.

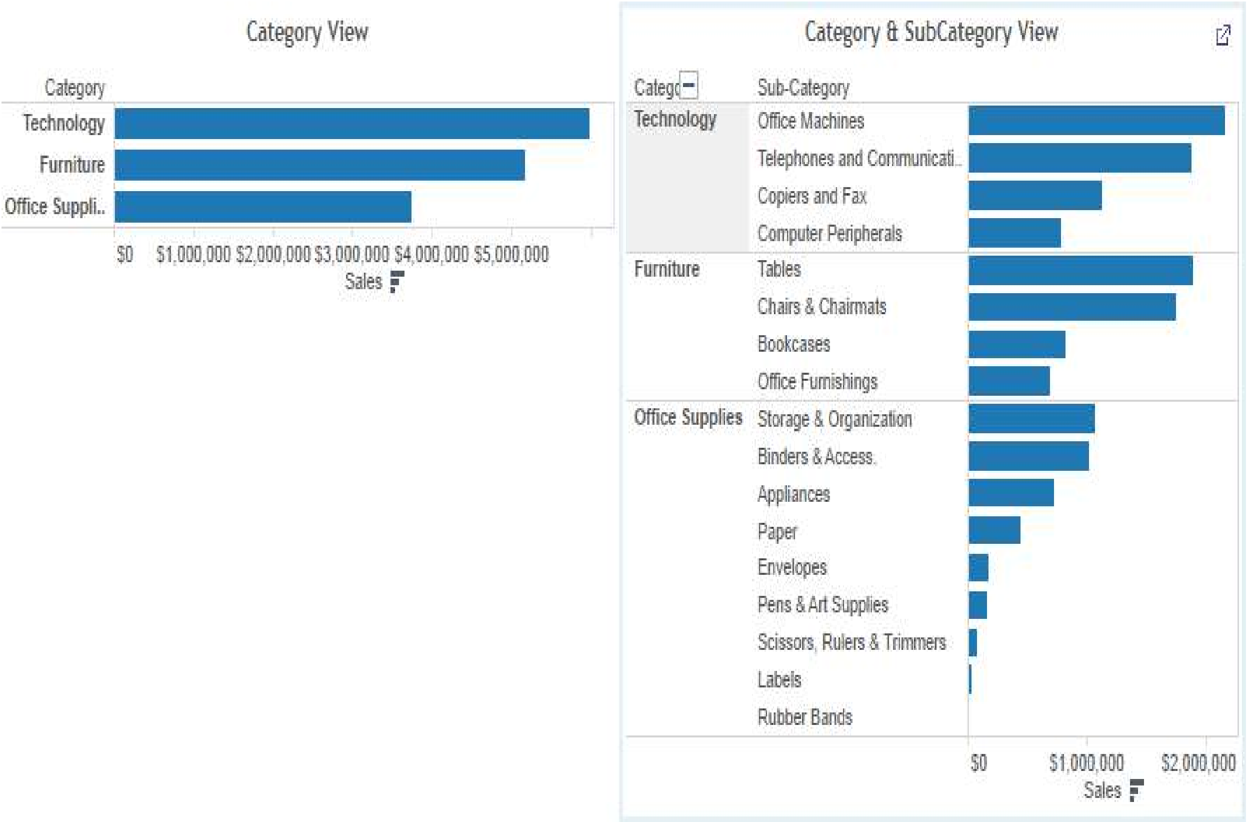
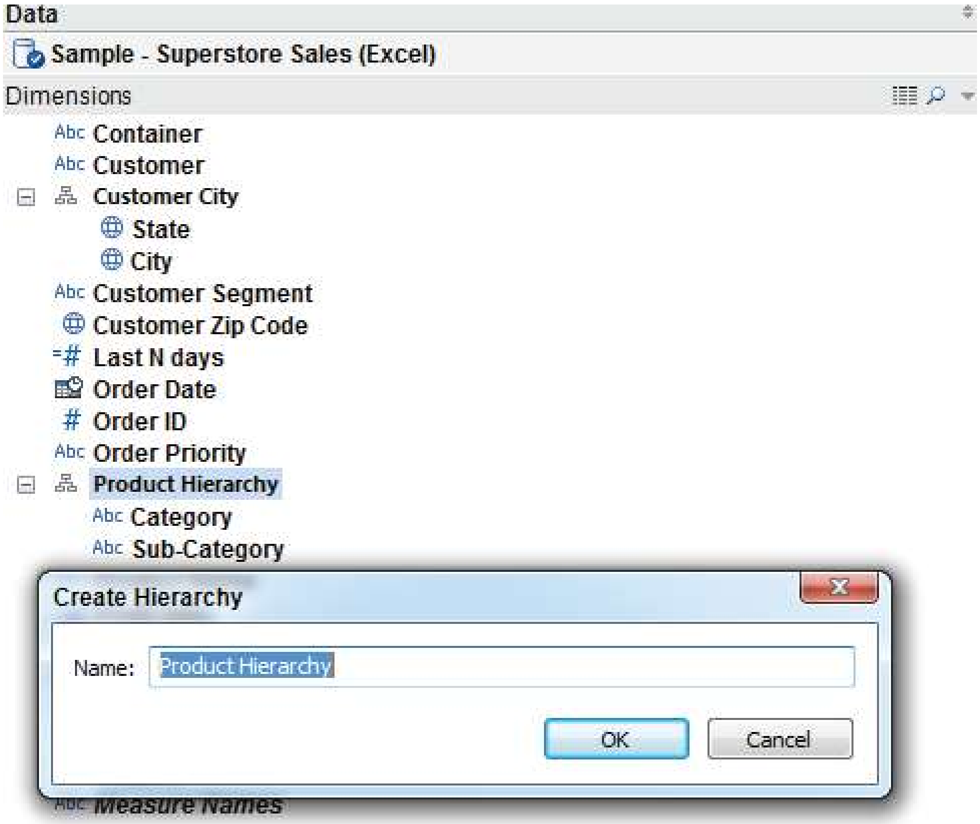


Fig 15-24. Hierarchy using category and subcategory

The bar chart on the left displays the summary product category. By pointing at the category heading a small plus sign will appear. Clicking on that causes the sub-category level of detail to be exposed. To collapse the hierarchy point at the category heading again and click on the minus sign. You can create as many levels in your hierarchy as you desire.

Hierarchies are created by pointing at a dimension field and dragging it on top of another field. The order of appearance is defined by dragging the field names contained within the hierarchy icon to the desired position. Figure 15-25 shows the

hierarchy icon with category and sub-category. You can change the hierarchy name by pointing at the text to the right of the hierarchy icon and typing **product hierarchy**. Other fields can be added to the hierarchy by positioning them in the order desired inside the hierarchy grouping on the dimension shelf.



Name: Product Hierarchy

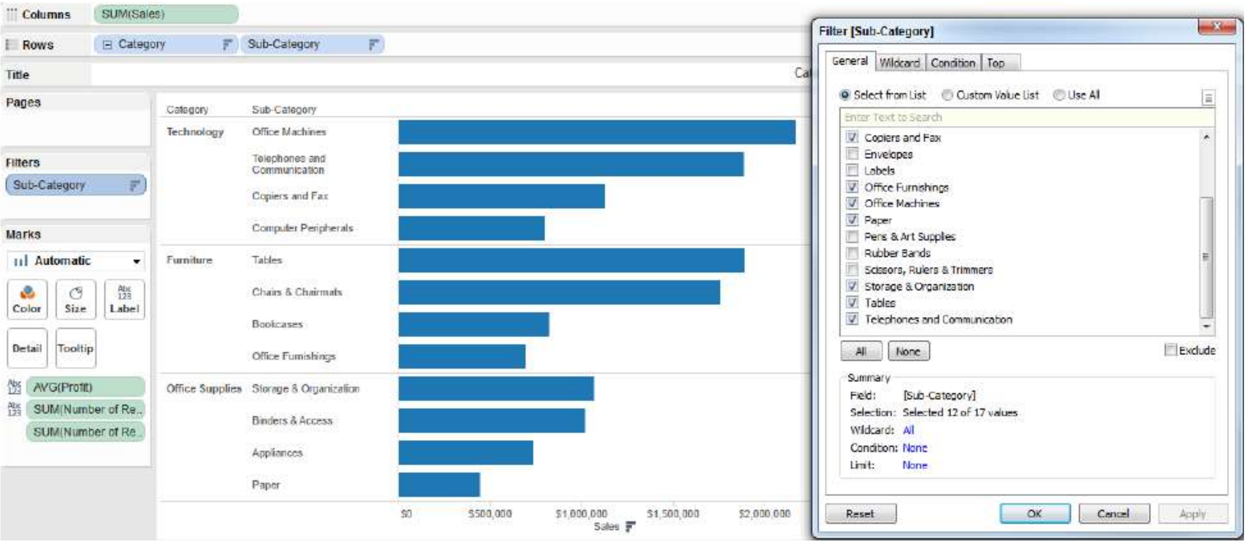
Fig 15-25. Making a custom hierarchy

Creating and Using Filters

There are a few different ways to add filtering to your visualization. Dragging any dimension or measure on to the filter shelf provides filtering that is accessible to the designer. Make that filter accessible to more people by turning it into a quick filter. This places it on the desktop where it is accessible to anyone—even those reading your report via Tableau Reader or Tableau Server. You can also create conditional filters that operate according to rules you define.

Creating a Filter with the Filter Shelf

In Figure 15-24 the category and subcategory view contains seventeen different rows of data. Suppose you want to hide five of those rows from view. Dragging the subcategory field from the dimension shelf and placing it in the filer shelf exposes the filter menu. Figure 15-26 shows the filtered data with the general tab of the filter menu. The subcategories that do not have check marks have been filtered out of view.



Filter [Sub-Category]

General | Wildcard | Condition | Top

● Select from List ○ Custom Value List ○ Use All

Enter Text to Search

☑ Copiers and Fax
☐ Envelopes
☐ Labels
☑ Office Furnishings
☑ Office Machines
☑ Paper
☐ Pens & Art Supplies
☐ Rubber Bands
☐ Scissors, Rulers & Trimmers
☑ Storage & Organization
☑ Tables
☑ Telephones and Communication

All None ☐ Exclude

Summary
Field: [Sub-Category]
Selection: Selected 12 of 17 values
Wildcard: All
Condition: None
Limit: None

Reset OK Cancel Apply

Fig.15-26. Applying a filter via the filter shelf

Notice that there are three other tabs on the filter menu. The Wildcard tab is typically used to search for text strings to filter. If you want to filter using another field that isn't in your view you can use the Condition tab to select any field in your datasource and filter using that field. The Top tab facilitates building top and bottom filtering or filtering requiring other formula conditions. If you use more than one of the filtering options tabs to define your filter, Tableau applies the conditions defined in each tab in the order the tabs appear from left to right. General conditions will be applied first, then wildcard, then condition, and the top tab conditions last. Below the general field list to the right of the None button is a check box for the Exclude option. If Exclude is checked, the items that include check marks are filtered out of view. Exclude filters can take a little longer to execute than Include filters, especially if your data set is very large.

Quick Filters

If you want to make the filter available for people that are viewing the report via Tableau Reader or Server you need to expose the filter control on the desktop. To create a quick filter, point at and right-click on any pill used on any shelf in your worksheet, then select the Show Quick Filter option. Figure 15-27 includes quick filters using the category and sales fields.

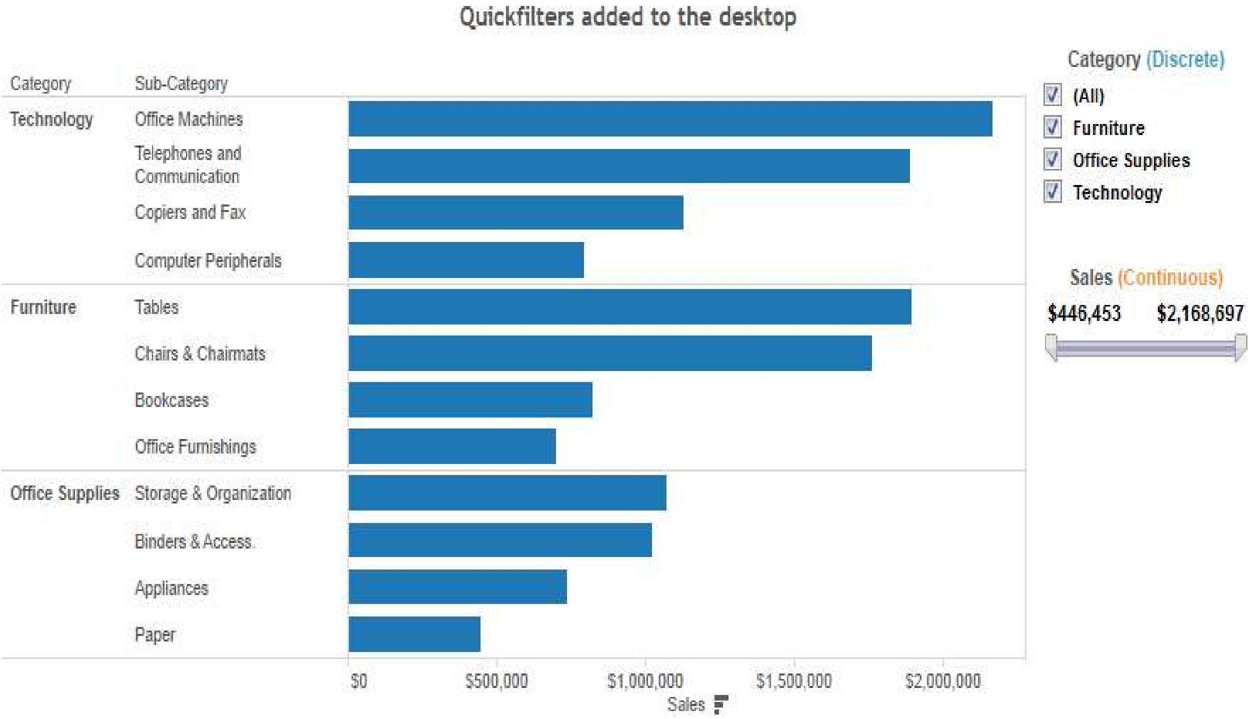


Fig.15-27. Adding a quick filters to the Desktop

The default quick filter styles are dependent on the type of field you apply within the quick filter control. In Figure 15-27 the discrete category field results in discrete filter options (furniture, office supplies, technology). Discrete filters are expressed using radio buttons or multi-select boxes. The second quick filter for sales (a continuous range of values) is expressed using slider-type filters. Editing the quick filter type can be done from inside the quick filter itself. Click on the title bar of the filter to expose the available options. Figure 15-28 shows examples of the menus that can be activated from the category and sales quick filter title bars.

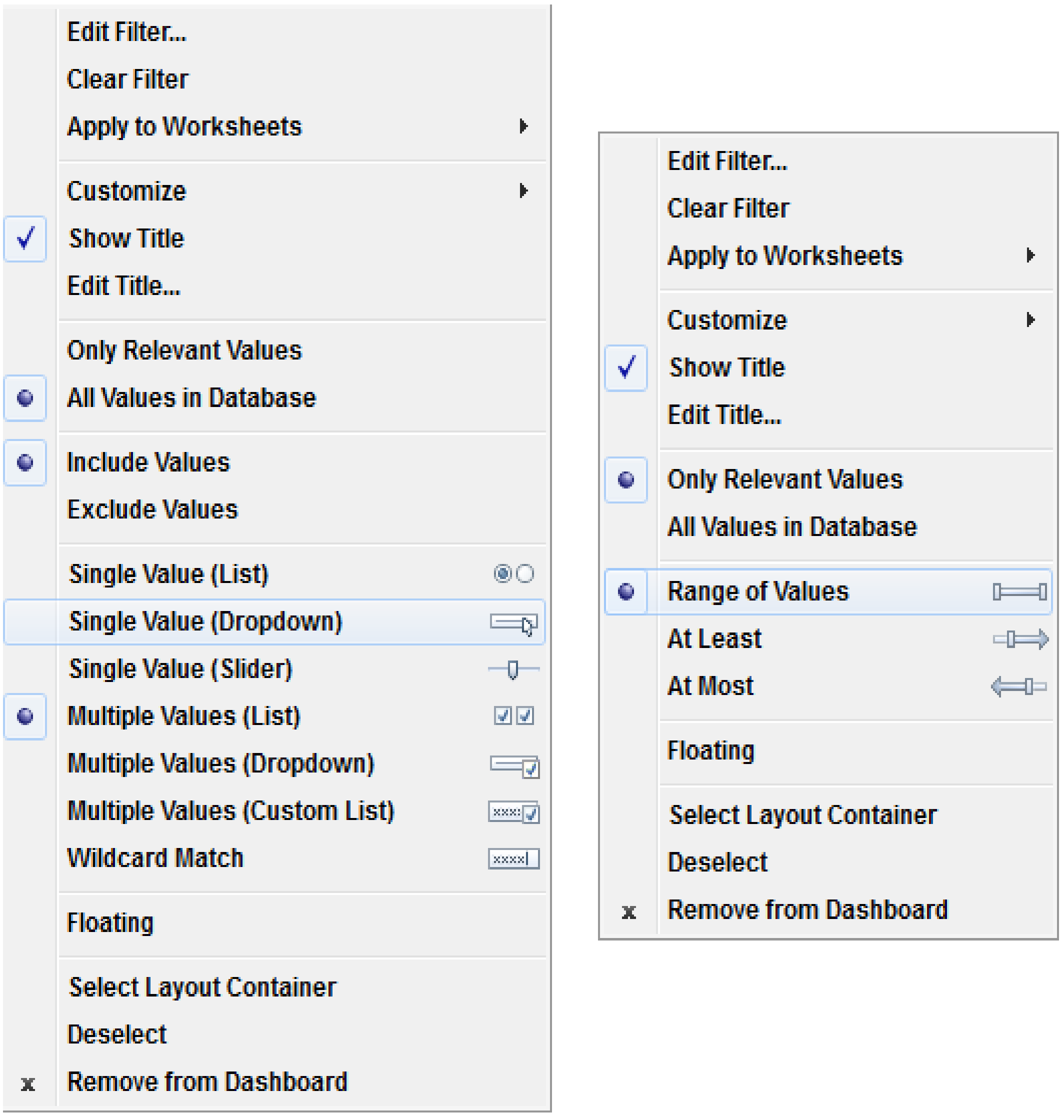


Fig. 15-28. Editing quick filter types

The menu on the left side of Figure 15-28 relates to discrete category filters. The right menu is for the continuous filters. In addition to controlling the filter style you can adjust many other attributes. You can edit the titles of each filter by including the words discrete and continuous and applying a different color to each word and centering the title. The quickfilter titles in Figure 15-27 have been modified in this way. These are the Quick Filter menus (both continuous and discrete)

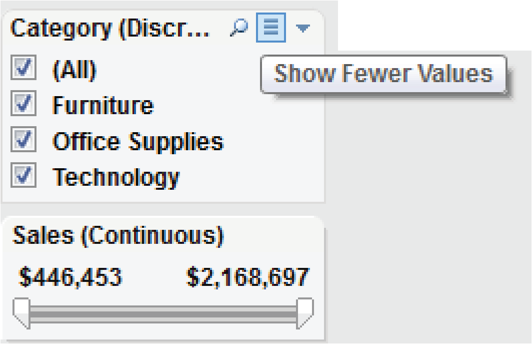
- Edit filter—Exposes the main filter menu.
- Clear filter—Removes the quick filter.
- Apply to worksheets—Apply the filter to all or selected worksheets.

- Customize—Turn on or off different filter controls.
- Show title—Turn off or on the quick filter title.
- Edit title—Modify the text in the quick filter title.
- Only relevant values—Turning this on reduces the set members displayed in the filter.
- Include values—Causes selected items in the filter to be included in the view.
- Exclude values—Causes selected items in the filter to be excluded from view.
- Hide card—Removes the quick filter from view but leaves it on the filter shelf.

These are the Quick Filter menu items that appear only if the quick filter is on a dashboard:

- Floating—If activated, allows the filter to float on top of other worksheet objects.
- Select layout container—Activates the layout container in the dashboard.
- Deselect—Removes the layout container selection in the dashboard.
- Remove from dashboard—Removes the quick filter from the dashboard.

The remaining sections of each filter type control the style of quick filter. There are seven styles of discrete and three styles of continuous quick filter types available. One other feature available directly from the quick filter is the ability to control the relevant values displayed directly from the desktop. Figure 15-29 displays a small control (three bars).


Category (Discr...

☑ (All)
☑ Furniture
☑ Office Supplies
☑ Technology

Sales (Continuous)

$446,453 $2,168,697

Fig 15-29 Including all or relevant values

This is important when you have several quick filters exposed in a view. For example, a hierarchy of quick filters might include a filter to select state, then city. Restricting the city filter to include only the relevant values means that if a particular state (Georgia) is selected in the first quickfilter, the city quickfilter would only display cities in the state of Georgia. If the city filter didn't apply only relevant values, the filter would contain every city in the United States.

Context Filters

One type of filter that many experienced Tableau users are unaware of is the context filter. Context filters do not only filter the data, they cause Tableau to create a temporary table that contains only the filtered data. For this reason they execute more slowly than a normal filter. Context filters are denoted by a gray colored pill. They can be useful if you want to work with a subset to achieve a particular result. Don't use a context filter if you plan to alter the filter frequently. Tableau provides robust filtering.

Grouping Dimensions

When you have a dimension that contains many members and your source data doesn't include a hierarchy structure, grouping can provide summarized views of the data. You can manually group items from headers or multi-select marks in a chart. Tableau also provides a menu option with fuzzy search that will help you group by searching strings in large lists of values. You can even group by selecting marks in a view. If you need to work with data that isn't structured the way you want it, grouping allows you to build that structure within Tableau.

Creating Groups Using Headers

Figure 15-30 includes a bar chart that compares product subcategories within each product category. The office supplies dimension has too many small members with very low sales values. Grouping the six smallest categories in office supplies into a single (ad hoc) category creates a grouping that is more comparable to the other subcategories.

There are three ways to group headings. The easiest way is to click on the paper clip icon in the Tooltips that appears when you multi-select the headers. The second way is to right click after selecting the headings and pick the Group option in the menu. One final option is available via the paper clip icon in the toolbar.

After creating the group, all six members will be combined into a single bar. The name that appears in the heading will be a concatenated list of the individual headings. To rename the combined list heading, right-click while pointing at the new group, choose edit alias, and type in a shorter name. The example group will be called (Other office). Figure 15-31 shows the new group and group name. Now each category includes four members—eliminating the tiny bars seen in Figure 15-30 that are difficult to see and compare. You can also create groups by selecting marks in the worksheet. This method is a great way to highlight items of interest when you are performing ad hoc analysis. In Figure 15-32 you see a cluster of marks that has been selected. These marks can be grouped using the paperclip icon inside the tooltip menu that appears when you point at any of the selected marks. You can select All Dimensions to create the group. The result is shown in Figure 15-33.

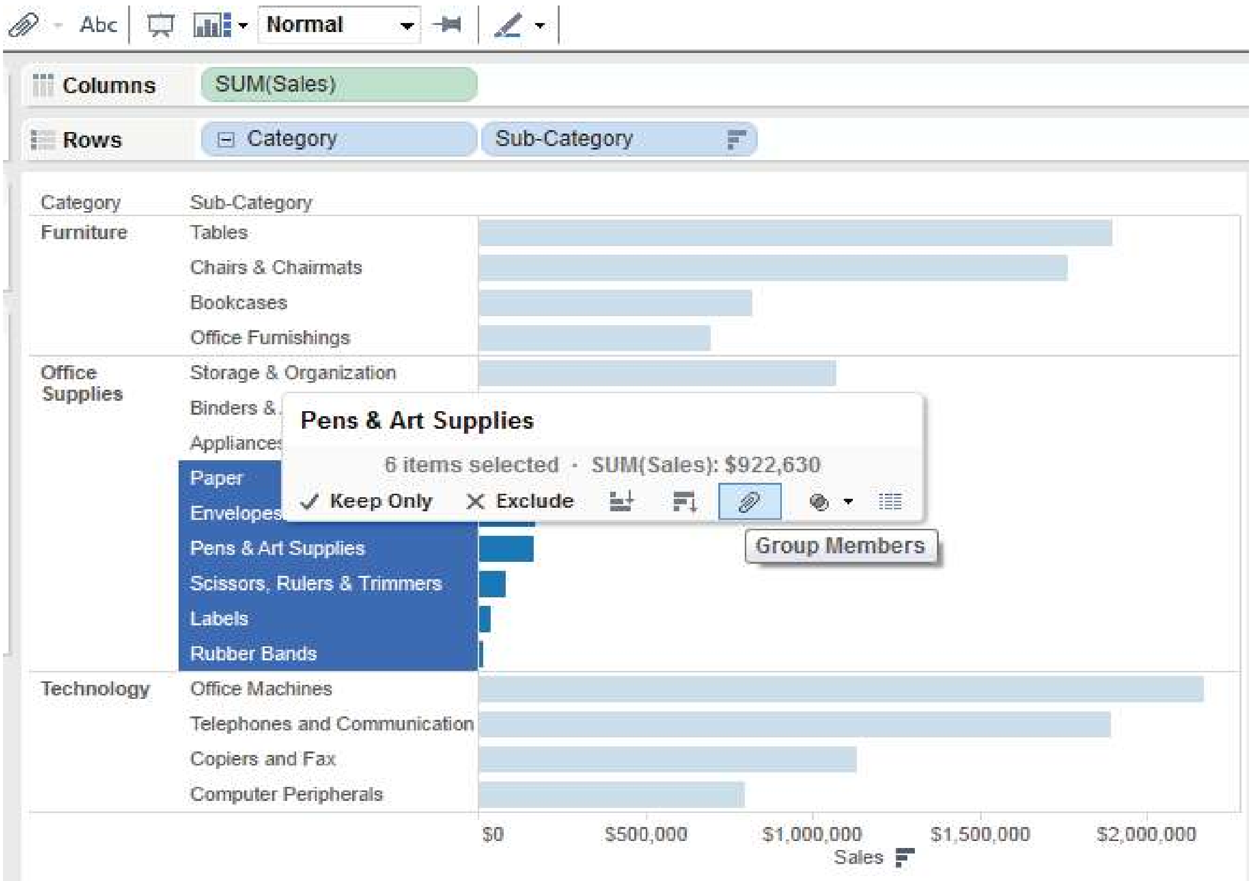


Fig 15-30 Grouping from headers

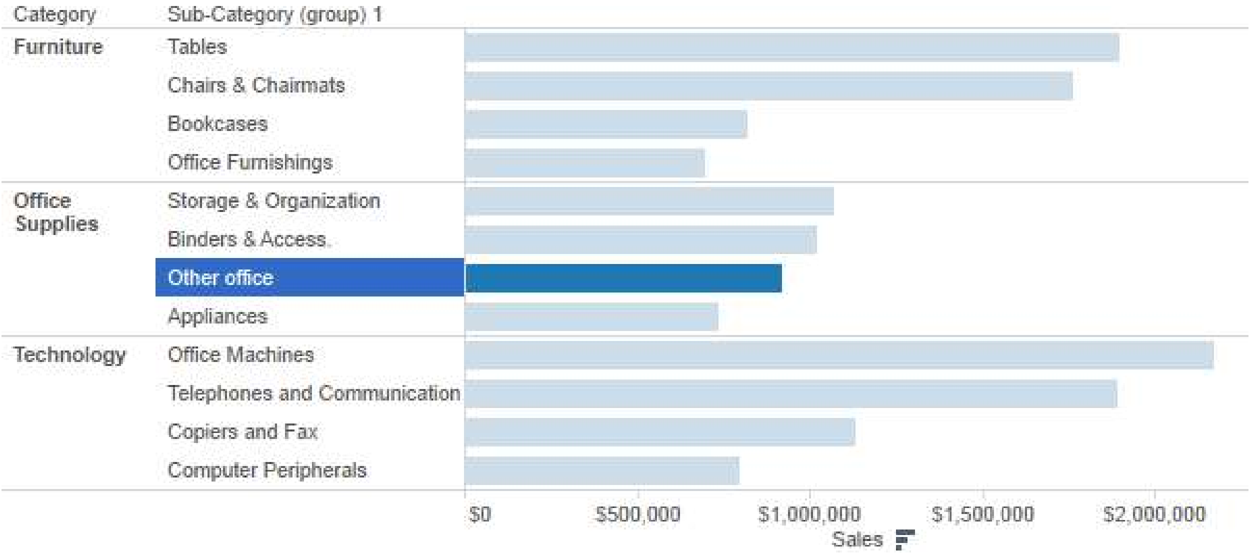


Fig 15-31 The ad hoc office group

Tableau's visual grouping causes the selected marks to be highlighted using a different color than the marks that are not included in the group. These methods work well if you have a small number of members to group or you can easily select the marks that you want to highlight.

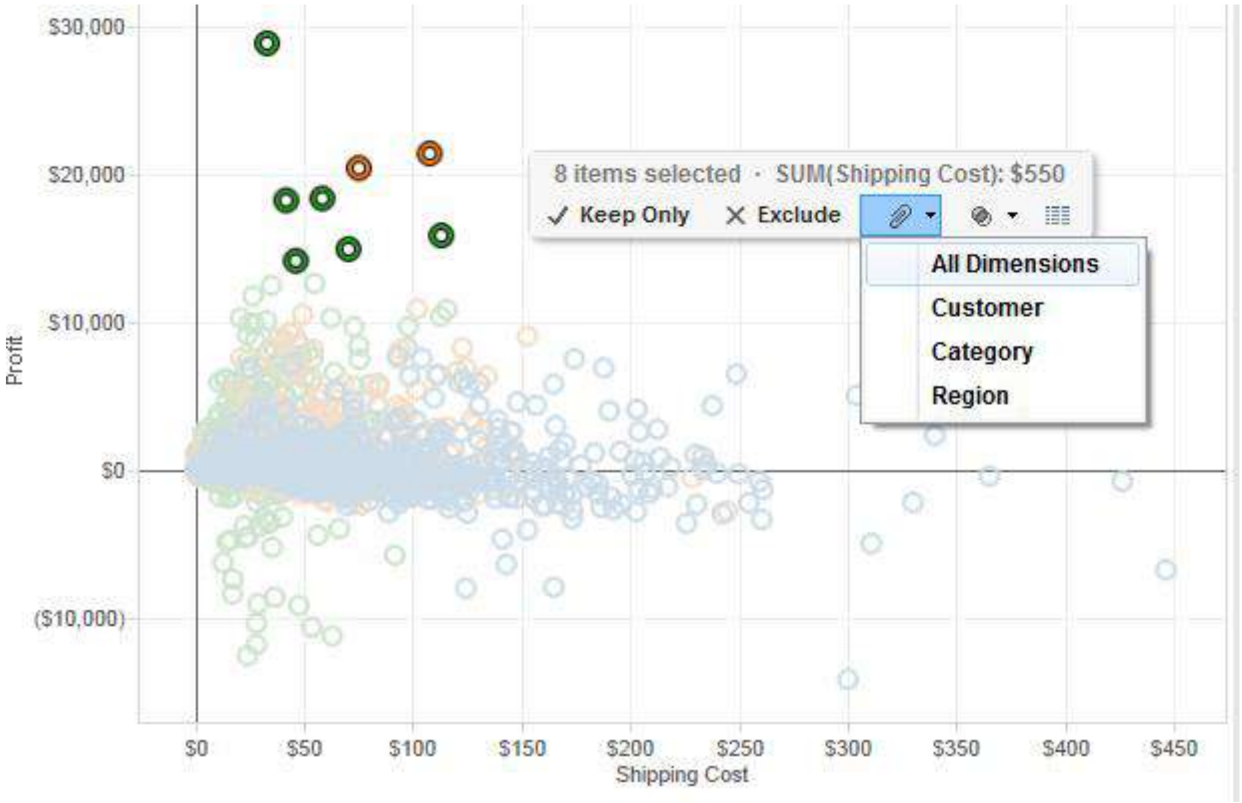


Fig 15-32 Grouping marks using all dimensions

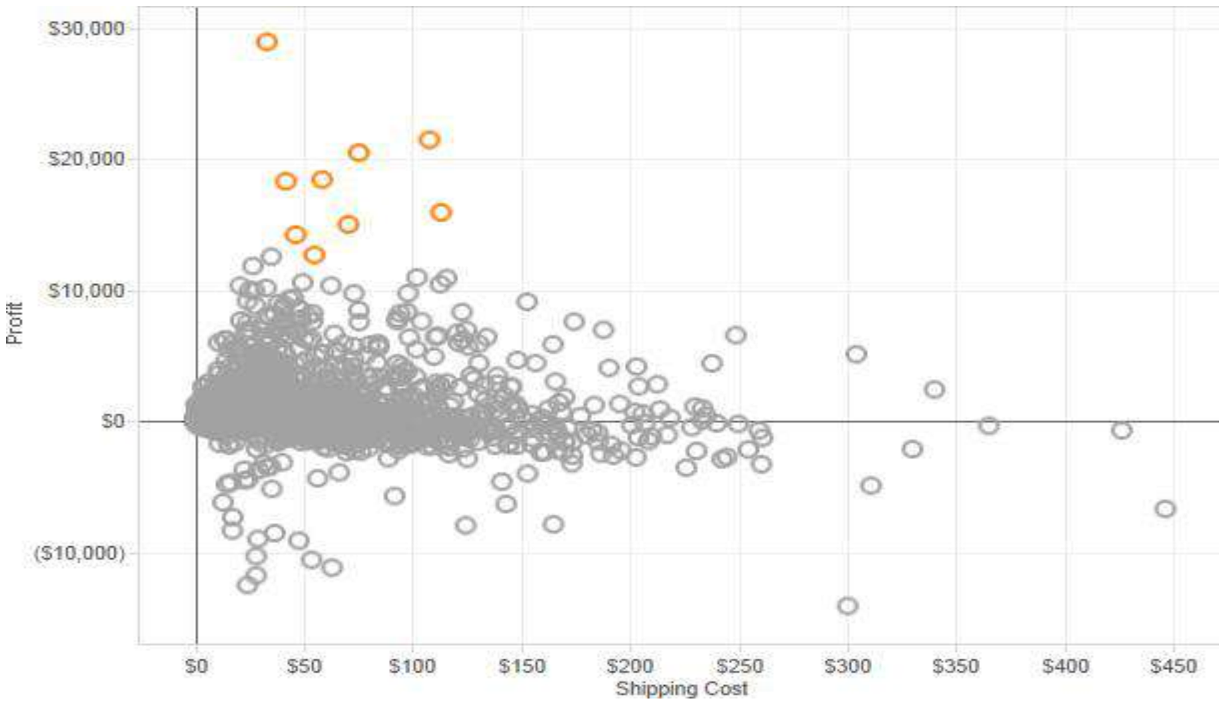


Fig. 15-33. Manually selecting a group

If you have a very large set of dimensions that you need to group, or the grouping must be created using portions of field names—these methods would be tedious. Tableau provides a more robust way to create groups using fuzzy search. Figure 15-34 shows another grouping menu that can be accessed by right-clicking on a specific dimension field within the dimension shelf.

You can also group products by vendor. Figure 15-34 shows a search for all products provided by the vendor Bevis. Using the Find Members search, Tableau executes a string search in all the product names that include that string. After checking to ensure that the group contains the correct information, clicking the Group button will create a new grouping of the products. You can also alias the group name within the menu. After completing all the vendor groups you require, selecting the Include Other check box will generate a group that contains all the other items in the dimension that haven't already been assigned to a vendor group.

Please note that any new group members that are added to your data source will not automatically appear in any group. You always have to add them manually the first time they appear in the data source.

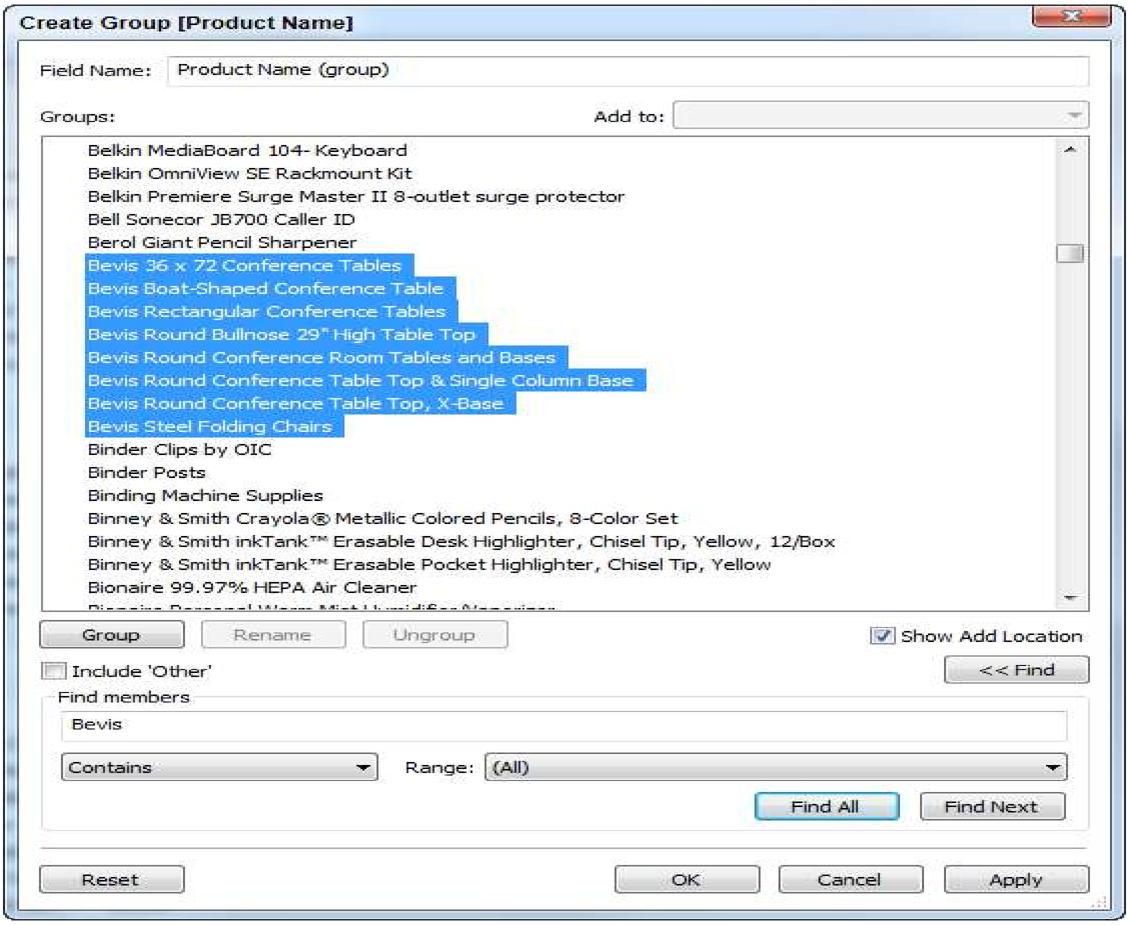
Using Sets to Filter for Specific Criteria

Think of sets as special kinds of filters that enable you to share findings made in one worksheet across other worksheets in your workbook. Or, perhaps you want to create an exception report that only displays records that meet specific criteria. Sets can be created several different ways:

- Multi-selecting marks
- Right-clicking on a field in the dimension shelf
- Combining sets on the set shelf

Saving Outliers by Multi-Selecting Marks

Creating a set by selecting marks in a view is fast and intuitive. Figure 15-35 shows a scatter plot that is comparing profit and shipping cost. If you want to create a set that includes low profit items, hold the left mouse button down and draw a box around the marks you want to save. This will automatically open the Tooltips.



Create Group [Product Name]

Field Name: Product Name (group)

Groups:

Belkin MediaBoard 104- Keyboard
Belkin OmniView SE Rackmount Kit
Belkin Premiere Surge Master II 8-outlet surge protector
Bell Sonecor JB700 Caller ID
Berol Giant Pencil Sharpener
Bevis 36 x 72 Conference Tables
Bevis Boat-Shaped Conference Table
Bevis Rectangular Conference Tables
Bevis Round Bullnose 29" High Table Top
Bevis Round Conference Room Tables and Bases
Bevis Round Conference Table Top & Single Column Base
Bevis Round Conference Table Top, X-Base
Bevis Steel Folding Chairs
Binder Clips by OIC
Binder Posts
Binding Machine Supplies
Binney & Smith Crayola® Metallic Colored Pencils, 8-Color Set
Binney & Smith inkTank™ Erasable Desk Highlighter, Chisel Tip, Yellow, 12/Box
Binney & Smith inkTank™ Erasable Pocket Highlighter, Chisel Tip, Yellow
Bionaire 99.97% HEPA Air Cleaner

☑ Show Add Location

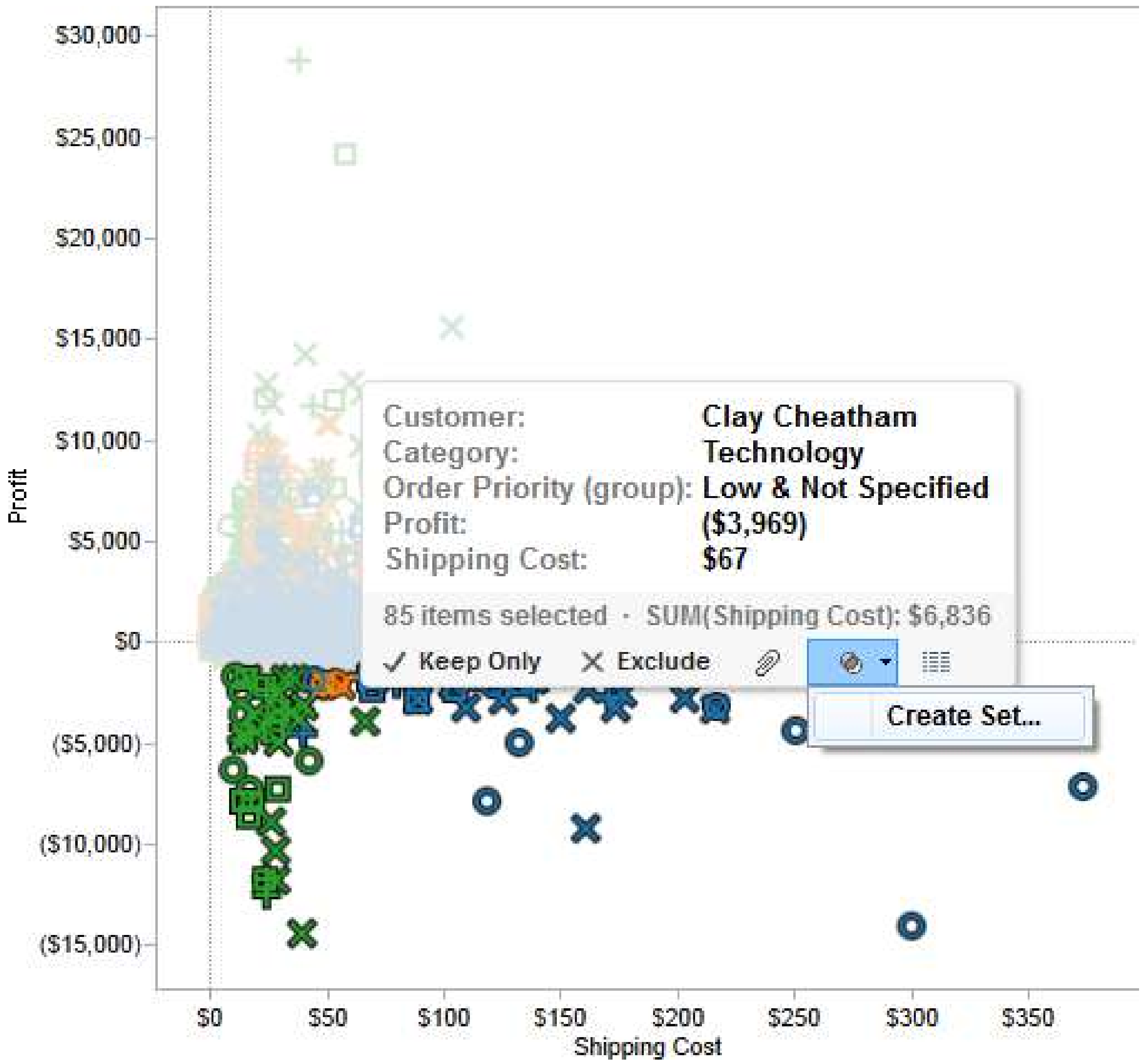
□ Include 'Other'

Find members: Bevis

Contains

Range: (All)

Fig. 15-34 Using string search to group

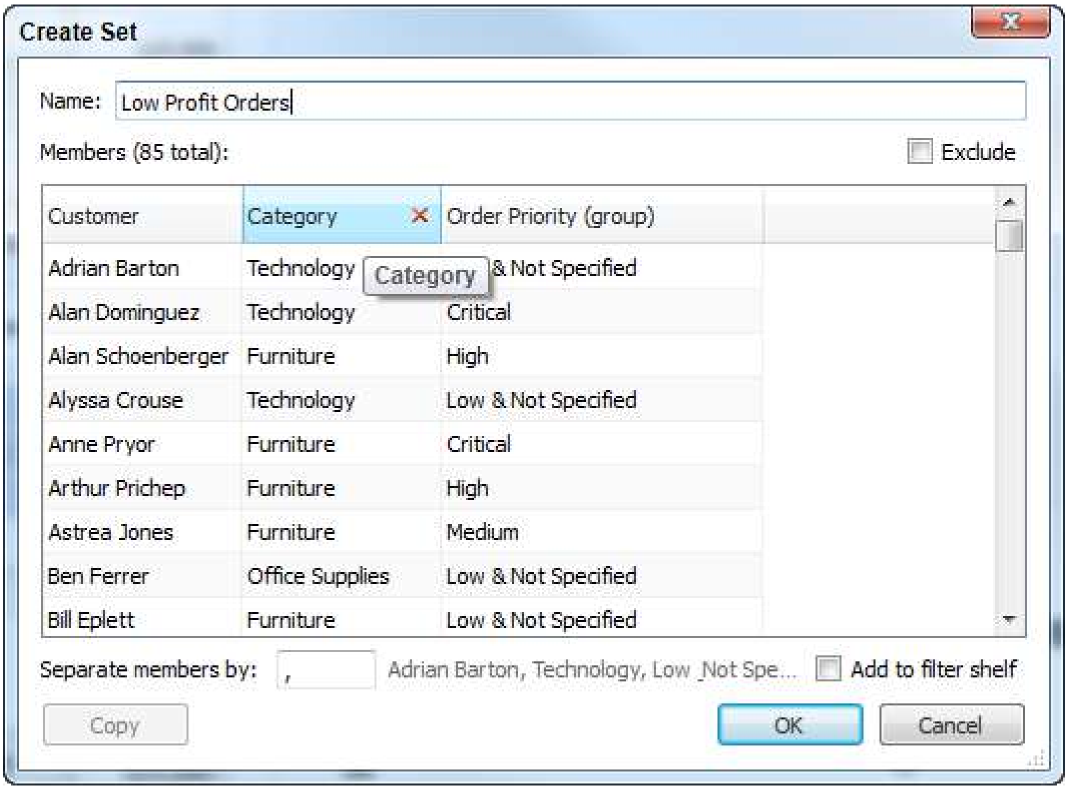


Customer: Clay Cheatham
Category: Technology
Order Priority (group): Low & Not Specified
Profit: ($3,969)
Shipping Cost: $67

85 items selected · SUM(Shipping Cost): $6,836

Fig 15-35 Selecting marks to create a set

Selecting the Create Set menu option exposes the dialog box in Figure 15-36.



Name: Low Profit Orders

Members (85 total):

□ Exclude

Customer	Category	Order Priority (group)
Adrian Barton	Technology	& Not Specified
Alan Dominguez	Technology	Critical
Alan Schoenberger	Furniture	High
Alyssa Crouse	Technology	Low & Not Specified
Anne Pryor	Furniture	Critical
Arthur Prichep	Furniture	High
Astrea Jones	Furniture	Medium
Ben Ferrer	Office Supplies	Low & Not Specified
Bill Eplett	Furniture	Low & Not Specified

Separate members by: , Adrian Barton, Technology, Low _Not Spe...

□ Add to filter shelf

Fig. 15-36 Editing fields included in a set

If you want to exclude a category from the set, hovering the mouse over the category header exposes a red (x) that if selected removes the category field from the set. Similarly, if you want to remove specific records, you could do that by pointing and clicking on the same control appearing in the row. For now, keep all dimensions and measures in this set. In addition, you can rename the set calling it **Low Profit Set**. Clicking the OK button adds a new shelf below the measures shelf that includes this set. You can also use the set in other worksheets within this workbook. Figure 15-37 shows different ways the set could be applied.

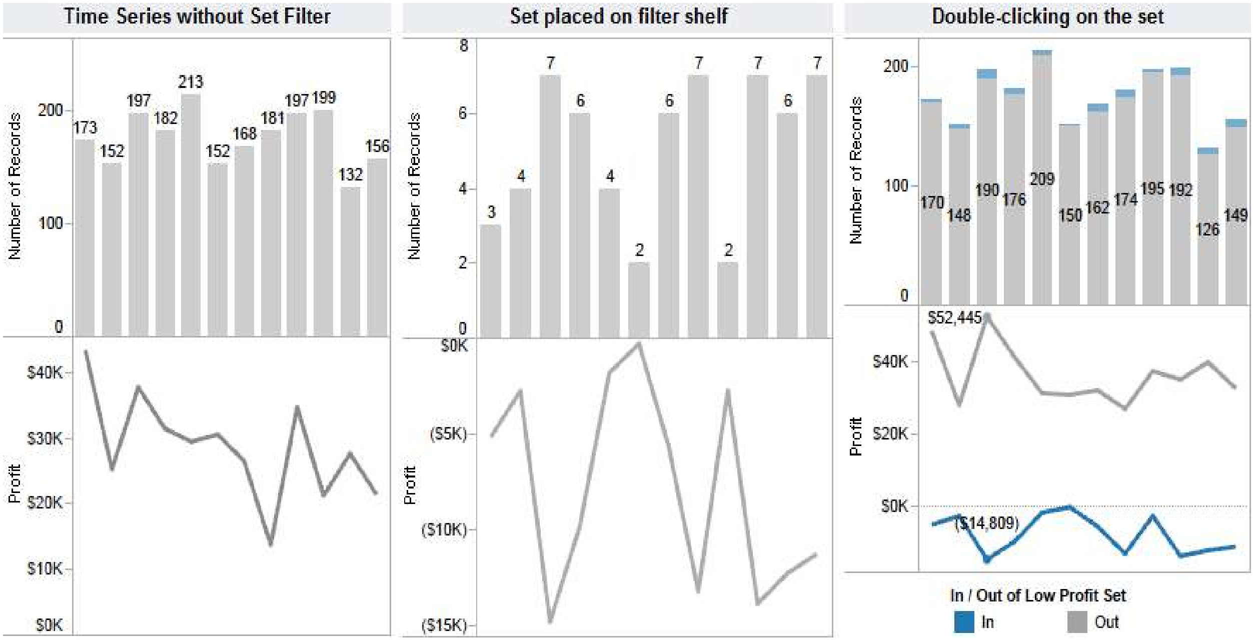


Fig. 15-37 Applying sets in different worksheets

The time series on the left displays record count and profit dollars for one year of sales. By dragging the low profit set to the filter shelf the view will change to reflect only the records included in the set. The middle view in Figure 15-37 shows the result. Notice the record count is much smaller and the profit pane has been filtered as well. Another way you could apply the set filter would be to double-click the low profit orders set on the set shelf. This option produces the visualization on the far right of Figure 15-37. The items that aren’t in the low profit set are gray and the low profit orders are blue.

Right-Clicking on a Field in the Dimension Shelf

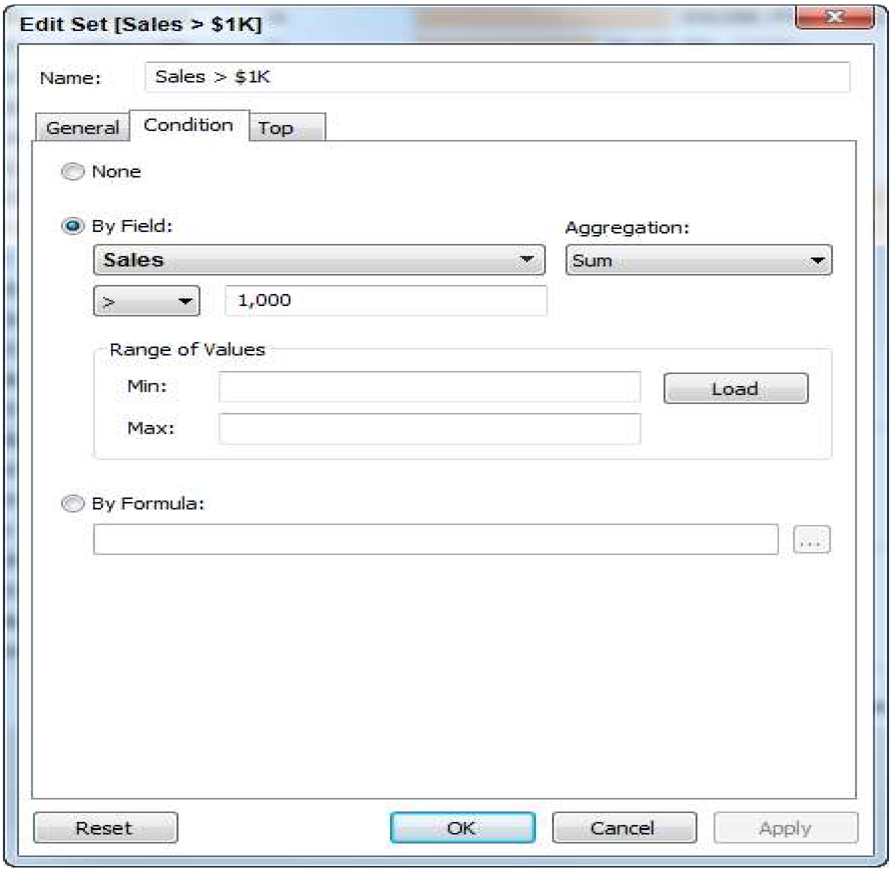
It is also possible to create a set by right-clicking on a field displayed in the dimension shelf and selecting the Create Set option. This will expose the dialog box in which you can apply filters manually or via calculations.

Combining Multiple Sets to Create a Combination Set

What if you want to create an exception report that only displays records that meet specific criteria? This can be achieved by joining two different sets in combination. You can see this in the following example and then use it to filter a chart. The desired combination set includes only order line detail for sales that are greater than one thousand dollars that have profit ratios of less than three percent. The steps required to create this combination set are:

- Create a concatenated field consisting of order id and row id.
- Make the set for sales greater than $1,000.
- Make the set for profit ratio less than three percent.
- Build a combination set consisting of the intersection of both sets.
- Display the result in a color-encoded bar chart.

Superstore includes information on each order down to each item included in the order. You want to display each order-row that is over one thousand dollars but less than three percent profit ratio. To enable this combination set, create a calculated field that uniquely combines order id and row id. Create a new field called Order-RowID by making a calculated field that concatenates the order id field and row id field. This can be done by using the following formula syntax: [Order ID]+”-”+[Row ID].



Edit Set [Sales > $1K]

Name: Sales > $1K

○ None

● By Field: Sales

Aggregation: Sum

> 1,000

Range of Values

Min:

Max:

○ By Formula:

Fig .15-38 Making the sales set

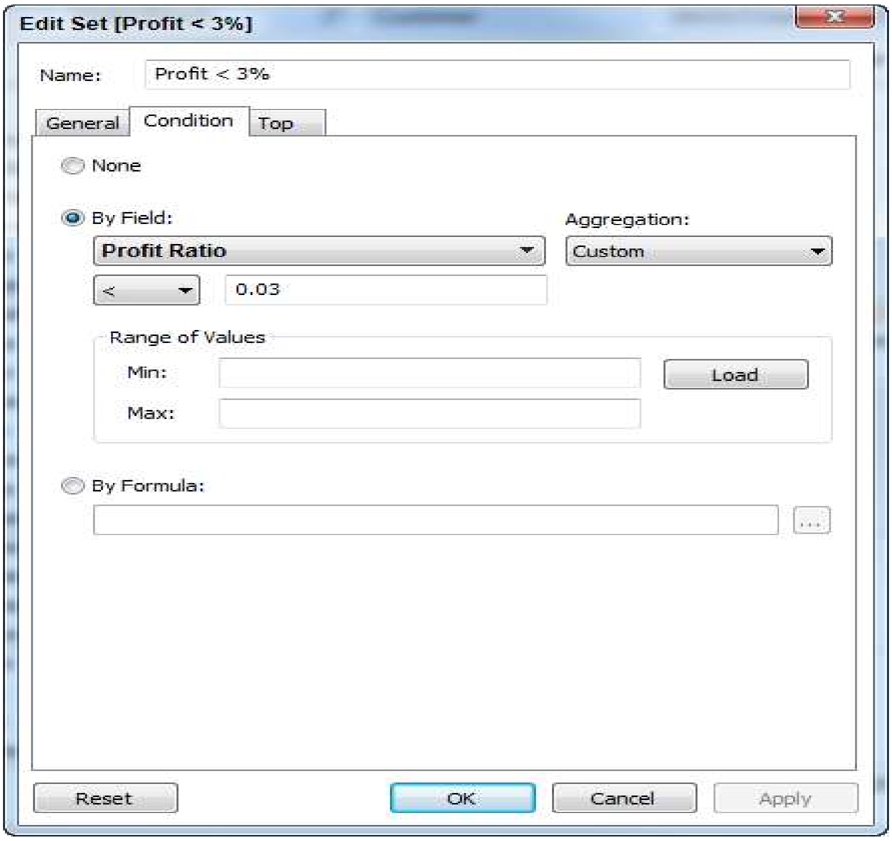
Make the Set for Sales Over One Thousand Dollars

Figure 15-38 shows how the set dialog box is exposed by right-clicking on the calculated field you just created for the combination of order and row id. On the general tab you will select all records. Using the condition tab you can choose the sales field for the sum of sales exceeding one-thousand dollars. Name the set (Sales > $1K) and click the OK button.

Building the Low Profit Set

Next you can create the set that will include only items with a profit ratio of less than three percent. Figure 15-39 shows the condition dialog box exposed after right-clicking on the Order-RowID field and selecting all records from the general tab, then defining the profit ratio limit.

After defining these sets you can now create a combination set. You do this by pointing at the set for sales over one thousand dollars, right-clicking, and selecting the Create Combined Set menu option. Figure 15-40 shows the dialog box that is displayed.



Edit Set [Profit < 3%]

Name: Profit < 3%

○ None

● By Field: Profit Ratio

Aggregation: Custom

< 0.03

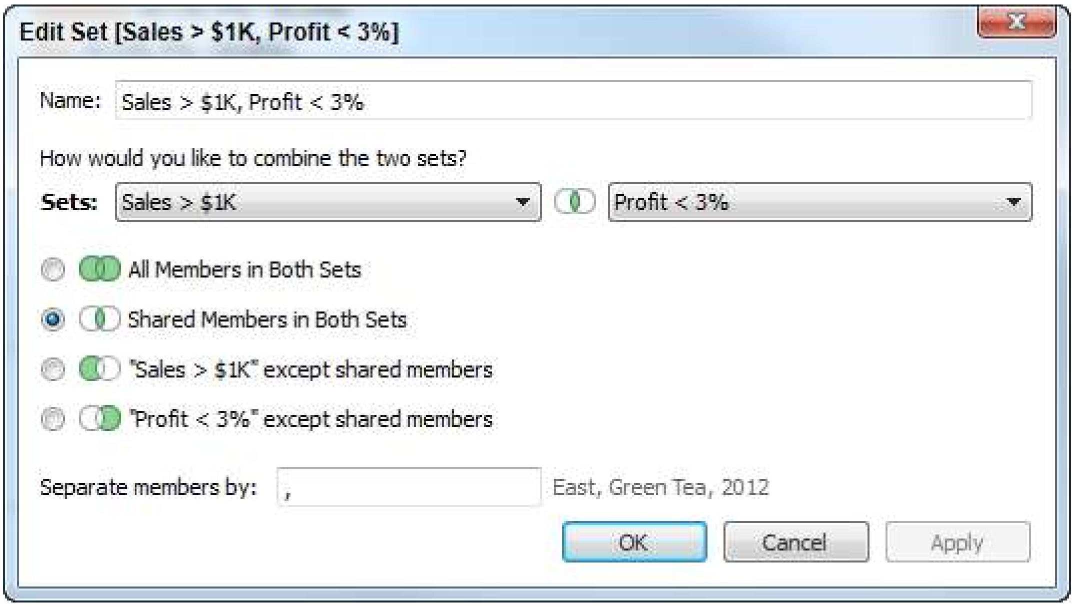
Range of Values

Min:

Max:

○ By Formula:

Fig 15-39 The low profit set condition defined



Edit Set [Sales > $1K, Profit < 3%]

Name: Sales > $1K, Profit < 3%

How would you like to combine the two sets?

Sets: Sales > $1K ▾ Profit < 3% ▾

○ All Members in Both Sets

● Shared Members in Both Sets

○ "Sales > $1K" except shared members

○ "Profit < 3%" except shared members

Separate members by: , East, Green Tea, 2012

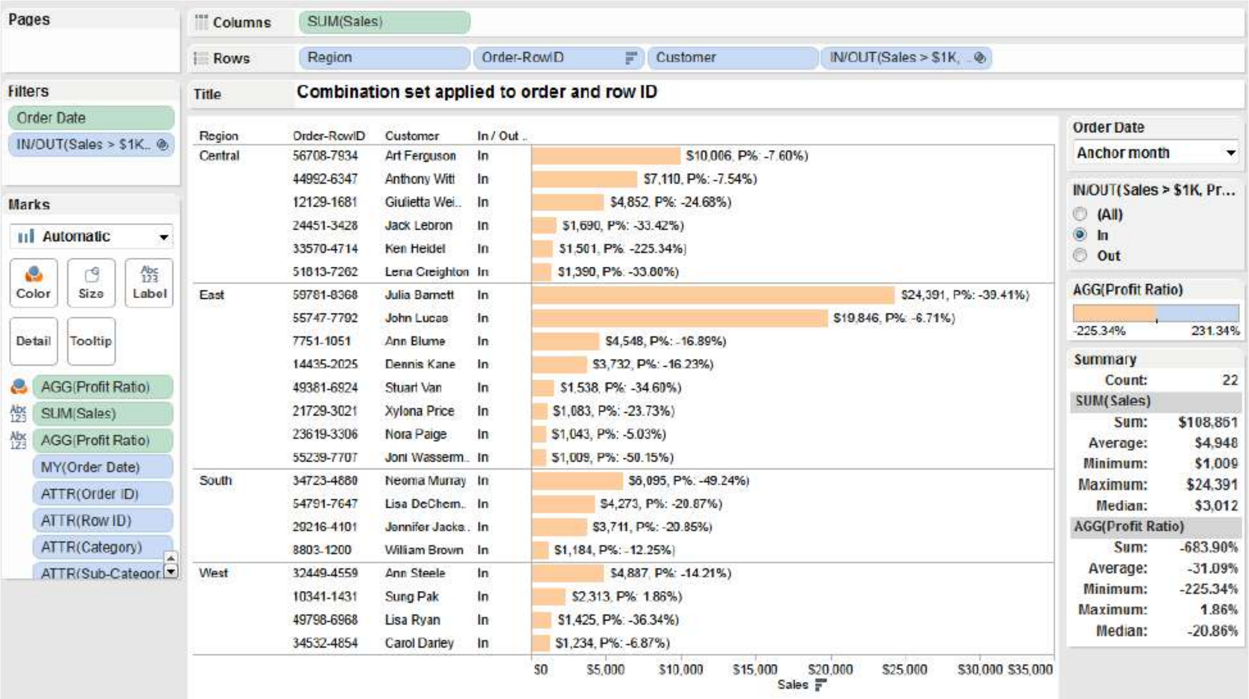
Fig. 15-40 Combination set dialogue box

The (Sales > $1K) set is already in the left set drop-down menu. The right drop-down menu was initially empty. Select the (Profit < 3%) set and the Radio button for the shared members option. Then click the OK button. This will generate another filter set that is the combination of the intersection of both sets. Figure 15-41 shows a bar chart that uses the combination set in the view.

Notice that the set option for displaying items in or out of the combination set has been selected. To make this chart easier to view, the color shelf has been edited to display items with profit ratios less than three percent using orange and over three percent using blue. Each bar is labeled with the sum of sales and profit ratio—providing visual confirmation that the data has been properly filtered by the combination set.

How Tableau Uses Date Fields?

Tableau recognizes dates that are contained in your source data and allows you to change the level of detail displayed via an auto-generated hierarchy. It is also possible to rearrange date levels by changing the order of date pills on the row or column shelves.



Columns: SUM(Sales)

Rows: Region | Order-RowID | Customer | IN/OUT(Sales > $1K, ..

Title: **Combination set applied to order and row ID**

Region	Order-RowID	Customer	In / Out ..	
Central	56708-7934	Art Ferguson	In	$10,006, P%: -7.60%)
	44992-6347	Anthony Witt	In	$7,110, P%: -7.54%)
	12129-1681	Giulietta Wei..	In	$4,852, P%: -24.68%)
	24451-3428	Jack Lebron	In	$1,690, P%: -33.42%)
	33570-4714	Ken Heidel	In	$1,501, P%: -225.34%)
	51813-7262	Lena Creighton	In	$1,390, P%: -33.80%)
East	59781-8368	Julia Barnett	In	$24,391, P%: -39.41%)
	55747-7792	John Lucas	In	$19,846, P%: -6.71%)
	7751-1051	Ann Blume	In	$4,548, P%: -16.89%)
	14435-2025	Dennis Kane	In	$3,732, P%: -16.23%)
	49381-6924	Stuart Van	In	$1,538, P%: -34.60%)
	21729-3021	Xylona Price	In	$1,083, P%: -23.73%)
	23619-3306	Nora Paige	In	$1,043, P%: -5.03%)
	55239-7707	Joni Wassem..	In	$1,009, P%: -50.15%)
South	34723-4880	Neoma Murray	In	$6,095, P%: -49.24%)
	54791-7647	Lisa DeChem..	In	$4,273, P%: -20.87%)
	29216-4101	Jennifer Jacks..	In	$3,711, P%: -20.85%)
	8803-1200	William Brown	In	$1,184, P%: -12.25%)
West	32449-4559	Ann Steele	In	$4,887, P%: -14.21%)
	10341-1431	Sung Pak	In	$2,313, P%: 1.86%)
	49798-6968	Lisa Ryan	In	$1,425, P%: -36.34%)
	34532-4854	Carol Darley	In	$1,234, P%: -6.87%)

Order Date: Anchor month

IN/OUT(Sales > $1K, Pr...
○ (All)
● In
○ Out

AGG(Profit Ratio)
-225.34% 231.34%

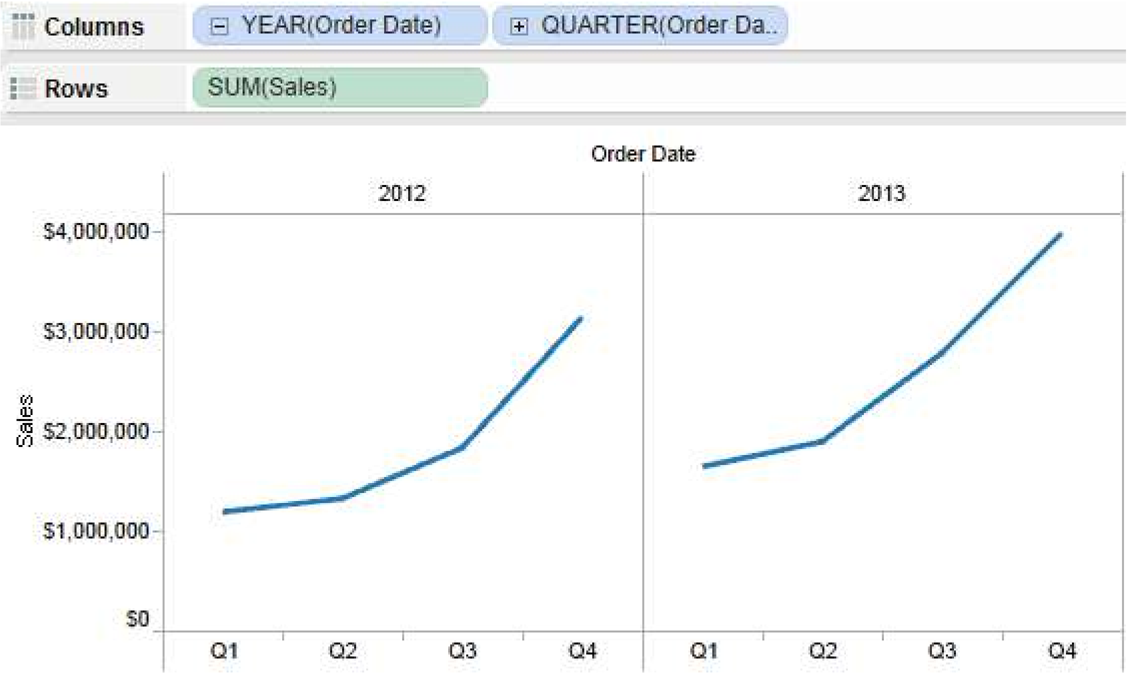
Summary	
Count:	22
SUM(Sales)	
Sum:	$108,861
Average:	$4,948
Minimum:	$1,009
Maximum:	$24,391
Median:	$3,012
AGG(Profit Ratio)	
Sum:	-683.90%
Average:	-31.09%
Minimum:	-225.34%
Maximum:	1.86%
Median:	-20.86%

Fig 15-41 Combination filter applied to a bar chart

Discrete and Continuous Time

You’ve probably noticed by now that some pills are green and others are blue. Similarly, icons can be in blue or green colors. Most beginners believe blue pills and icons denote dimensions while green pills are used to display measures.

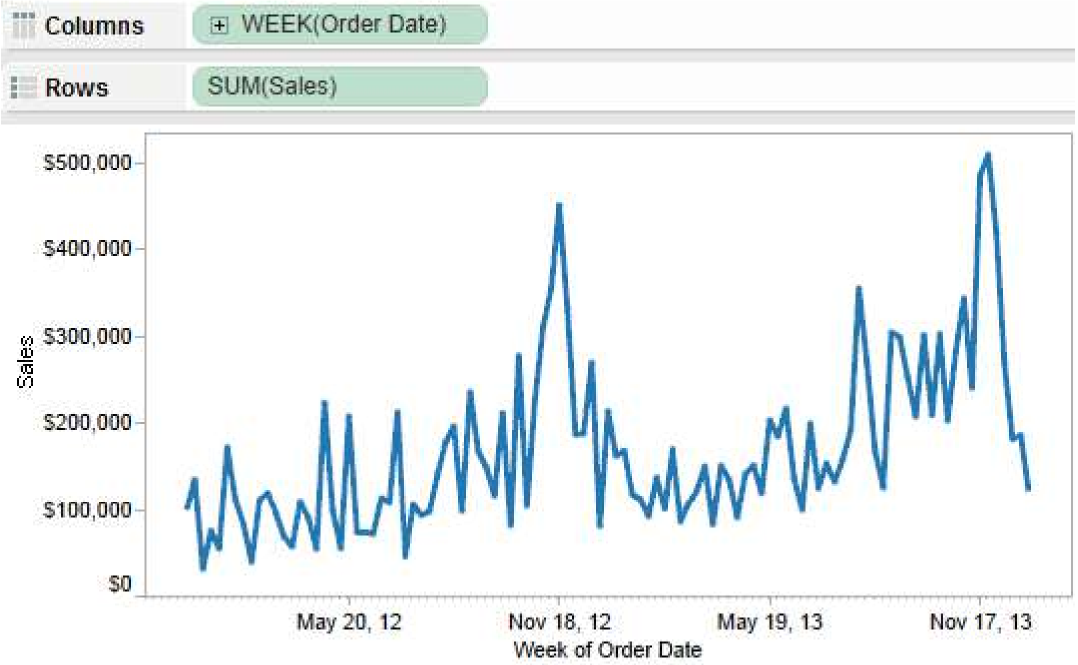
While this is frequently the case, the truth is more subtle. Blue pills/icons denote “discrete” fields. Green pills/icons denote “continuous” fields. Dates can be both discrete and continuous. Figure 15-42 shows Tableau’s default way of displaying time—as discrete time hierarchy. You can see that time has been discretely segmented in the time series chart by year. Clicking on the plus sign in the quarter pill would cause the date hierarchy to expand to include months, and panes for each quarter would be exposed. Continuous dates don’t discretely bucket time but will cause a drill down to a lower level of detail. Figure 15-43 shows a similar time series chart with continuous time being used and the level of detail being month. The green pill on the column shelf in Figure 15-43 indicates the level of detail being displayed. Notice that there are no panes in view. Time is continuously displayed as an unbroken line.



Columns: ⊟ YEAR(Order Date) ⊞ QUARTER(Order Da..

Rows: SUM(Sales)

Fig. 15-42 Discrete time series



Columns: ⊞ WEEK(Order Date)

Rows: SUM(Sales)

Fig. 15-43 Continuous time series

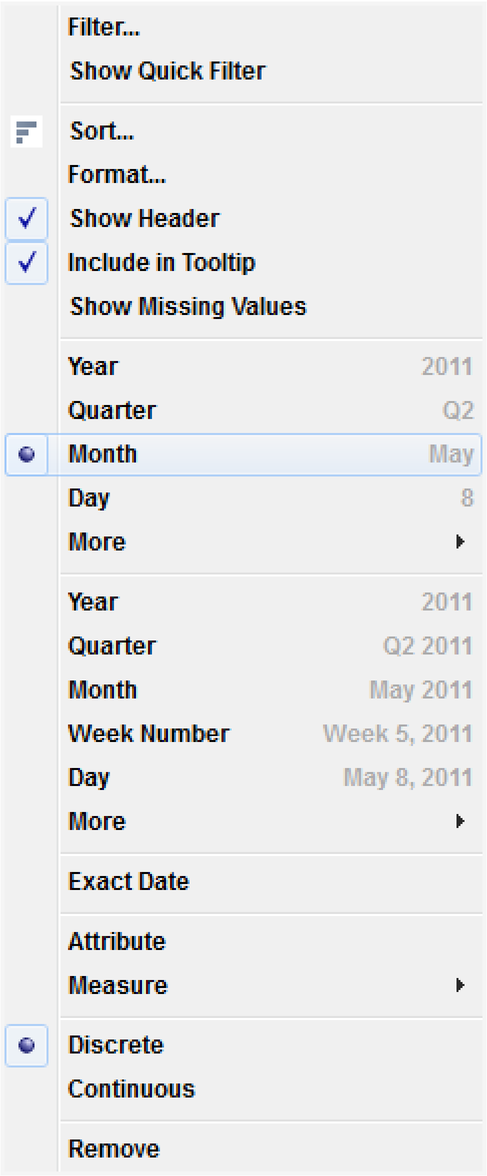


Fig 15-44 Changing the data level of detail

Tableau’s Date Hierarchy

Time can be expanded to more granular levels simply by clicking on the plus sign within the date pill. Experiment with this and note that you can rearrange time buckets just by changing the order of the pills by repositioning them. It’s also possible to change the level of detail displayed by right-clicking on the date pill. This exposes the menu in Figure 15-44. The menu includes two different date sections that start with year. The first group provides discrete date parts. The second group provides continuous date values. Figure 15-45 was created by changing the date displayed in Figure 15-42, altering the quarter pill to display month.

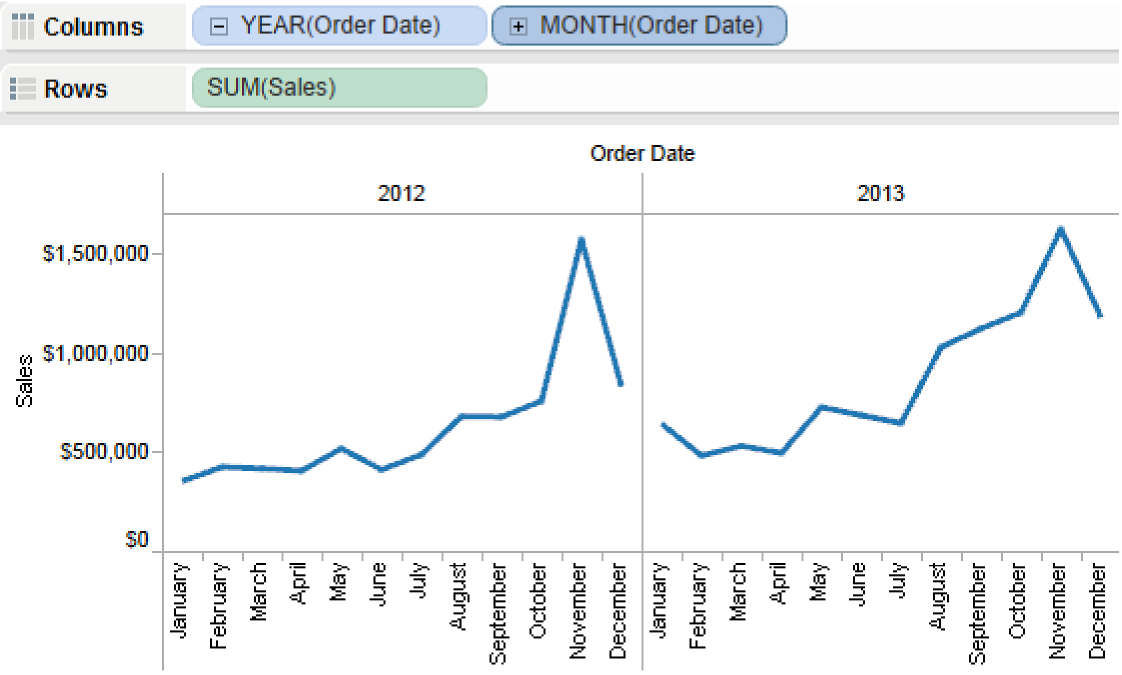


Fig 15-45 Time series displaying discrete year month

In Figure 15-44 note the menu option “more” appears twice. The first time it appears is within the discrete date section of the menu. The second time it is the continuous date section. Explore the menu option more in both the discrete and continuous time portions of the menu. The More menu options provide even more granular options for controlling how date and time are presented in your view.

15.6 CHECK YOUR PROGRESS

1. What are different time measures in Tableau
2. List at least three sorting feature embedded in Tableau
3. What is a context filter?

Answers to Check your progress

1. Continuous and discrete
2. Sort menu, sorting via legends, hierarchies to provide drill down capability
3. Context filters do not only filter the data, they cause Tableau to create a temporary table that contains only the filtered data.

15.7 SUMMARY

Now that you've learned how to connect Tableau to a variety of data sources you can start building visualizations. In this unit you learnt about all of the chart types provided by the Show Me button. You discovered how to add trend lines, reference lines, and control the way your data is sorted and filtered. You have seen how creating ad hoc groups, sets, and hierarchies can produce information not available in the datasource. Tableau's discrete and continuous data hierarchies explained, and how you can alter Tableau's default date hierarchies by creating your own custom dates.

15.8 KEYWORDS

- Trend lines - help you see patterns in data that are not apparent when looking at your chart of the source data by drawing a line that best fits the values in view **-** An *iconic* model is a material representation of a real system

- Reference lines - allow you to compare the actual plot against targets or to create statistical analyses of the deviation contained in the plot

- discrete time - signal is the one which is **not defined at intervals between two successive samples of a signal**

- A **continuous-time (CT)** - signal is a function, s (t), that is defined for all time t contained in some interval on the real line

15.9 QUESTIONS FOR SELF-STUDY

1. Explain how *Show me* button works?

2. Describe how to sort data in Tableau?

3. Explain quick filter menu in Tableau.

4. Describe how Tableau uses data fields?

15.10 REFERENCES

1. Alexander Loth - Visual Analytics with Tableau-Wiley (2019)
2. Dan Murray - Tableau Your Data!_ Fast and Easy Visual Analysis with Tableau Software-Wiley (2013)
3. David Baldwin - Mastering Tableau-Packt Publishing (2017)

UNIT -16: Calculations with Tableau

Structure

16.0 OBJECTIVES

After studying this unit, you will be able to:

- ✓ Understand when to use which type of aggregation
- ✓ Create and edit calculated fields
- ✓ Understand the order in which different work processes happen in Tableau and how that affects the different types of calculations.
- ✓ Use Table Calculations for in-depth analyses.

16.1 *WHAT IS AGGREGATION?*

Aggregation defines how values are expressed. Most Tableau functions are calculated at the database server with only the results being sent to Tableau. If you are familiar with SQL, you will find most of the functions in Tableau are an extension of SQL. Tableau uses the Sum aggregation by default. If the default aggregation isn't what you want, point at the pill of the measure you've placed into the view—right-click, and select a more appropriate aggregation. Supported aggregation types include:

- Sum
- Average
- Median
- Count
- Count Distinct
- Minimum
- Maximum
- Standard Deviation
- Standard Deviation of a Population
- Variance
- Variance of a Population

These are clearly defined in Tableau's online manual. Search the help menu to read more about each of them if you are unfamiliar with the type of aggregation each provides.

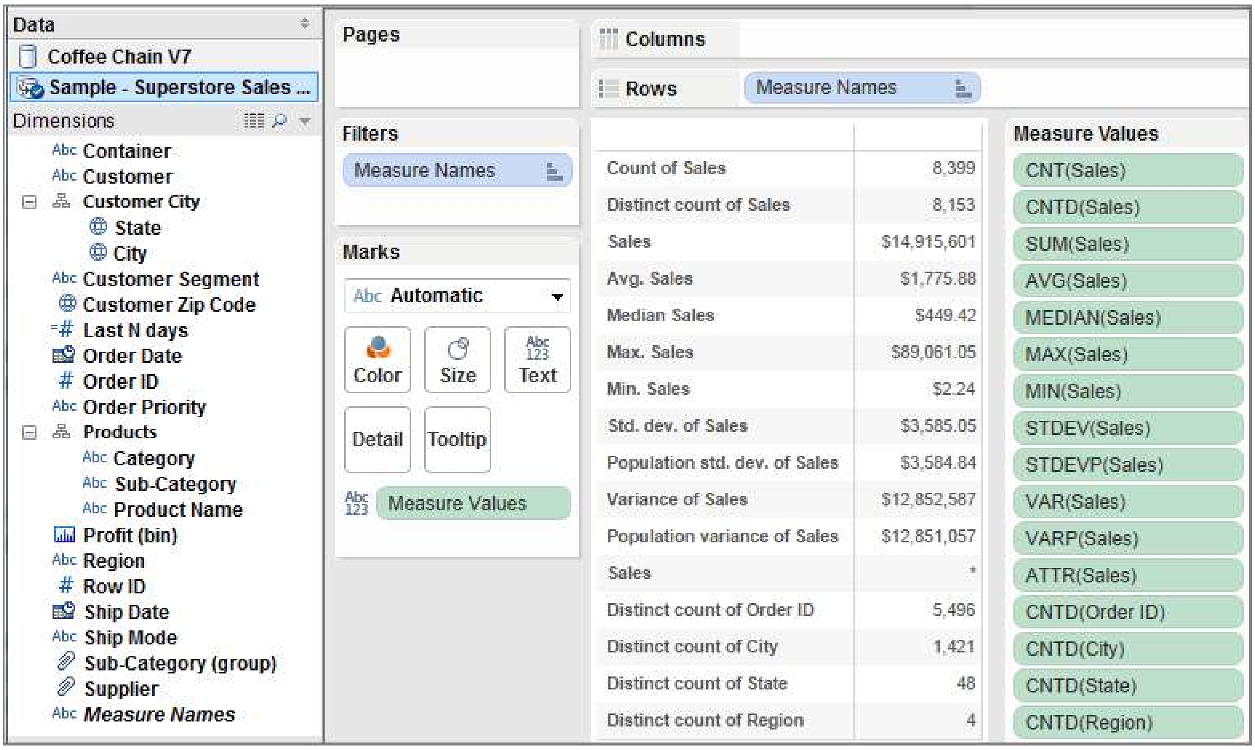
Count Distinct Versus Count

These functions count records in different ways. Consider a data set that includes 10,000 records with 20 different regions. Performing a Count Distinct on the Region field returns a value of 20. The purpose of Count Distinct is to count the unique instances of a particular item. A Count aggregation of 10,000 records will result with an answer of 10,000 because it counts all records.

Count Distinct is supported by relational database sources but is not supported by Excel, Access, or text files. You can add the ability to create Count Distinct aggregation when accessing those sources by performing a data extract. Tableau's extract files do support Count Distinct aggregation

Median

Similarly, Median is not supported by a direct connection from Tableau to Excel, Access, or text files. Performing a data extract will once again give you the ability to compute median values. Using the Superstore data set, Figure 16-1 shows a cross tab displaying all of the different aggregations available for the sales field in the data set.



Count of Sales	8,399
Distinct count of Sales	8,153
Sales	$14,915,601
Avg. Sales	$1,775.88
Median Sales	$449.42
Max. Sales	$89,061.05
Min. Sales	$2.24
Std. dev. of Sales	$3,585.05
Population std. dev. of Sales	$3,584.84
Variance of Sales	$12,852,587
Population variance of Sales	$12,851,057
Sales	*
Distinct count of Order ID	5,496
Distinct count of City	1,421
Distinct count of State	48
Distinct count of Region	4

Fig. 16-1 Different aggregation of sales

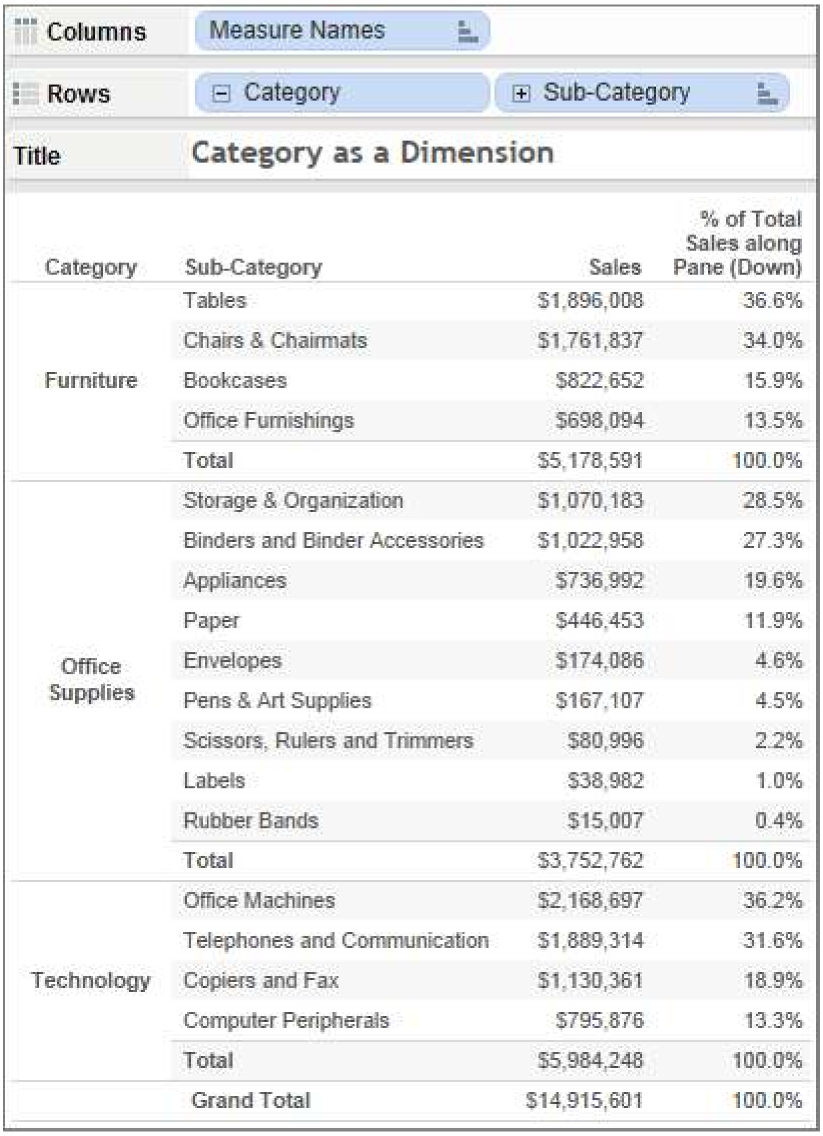
Notice that the bottom four rows are expressing Count Distinct values for different dimensions. By dragging each of those dimension fields into the crosstab using the right mouse button, the Count Distinct aggregation is expressed for each dimension. As you can see the data set includes over 5,000 different orders, over 1,400 cities, 48 states, and four regions.

Dimension versus Attribute

Aggregation behavior can be changed by altering the default method by which Tableau expresses dimensions. Figure 16-2 shows a cross tab containing sales by product category and sub-category. A table calculation is being used to display the percent of total sales that each row represents within each product category pane.

By default, Tableau partitions the result by the category dimension. Subtotals have been added by using the main menu option analysis/totals, then showing subtotal and column totals. You can see that in each category pane the amount of sales and percent of sales are totaled within each category pane. But, if the category dimension is changed to an attribute, the category dimension will become a label only and no longer cause the data to be partitioned. Figure 16-3 shows the same data set but with the category field changed to an attribute.

The view still shows the light gray boundary lines between each category, but because the category dimension has been changed to an attribute, it no longer partitions the view. The sales total reflects the total for the entire crosstab and the percent of total sales is now expressing the percentage of total sales, not the sales within each category. This may appear to be trivial, but as your skills advance and you begin to employ more advanced table calculations you will need to understand how attributes change Tableau’s behavior.



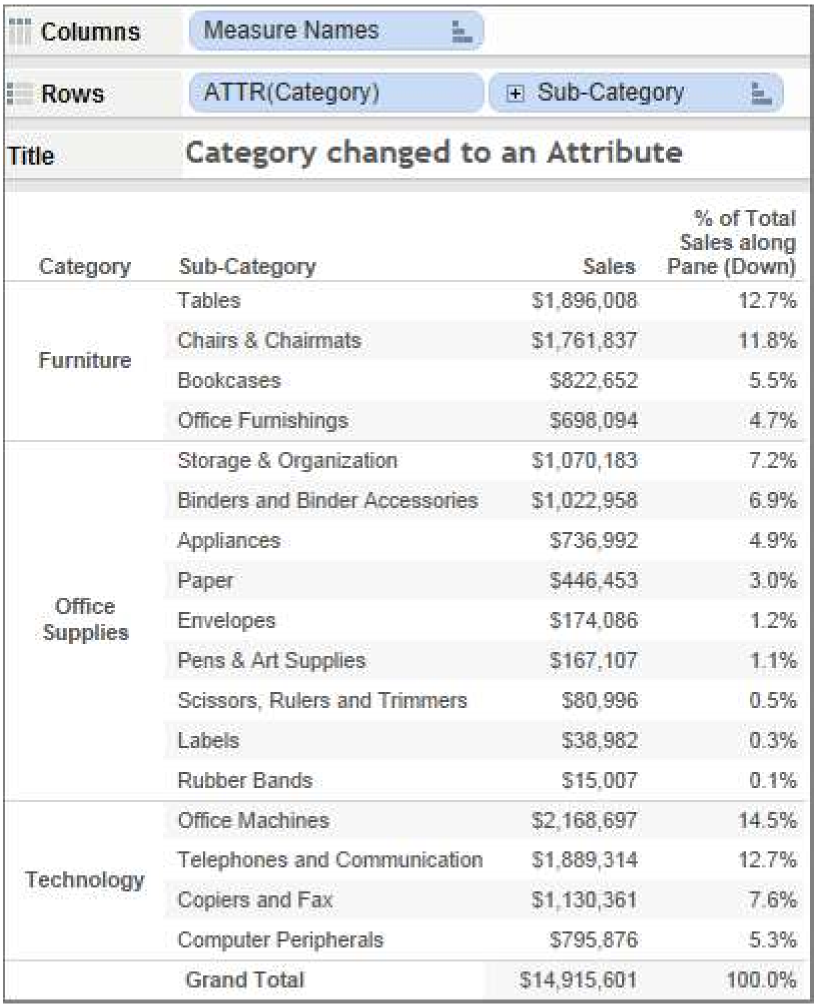
Columns: Measure Names

Rows: ⊟ Category ⊞ Sub-Category

Title: Category as a Dimension

Category	Sub-Category	Sales	% of Total Sales along Pane (Down)
Furniture	Tables	$1,896,008	36.6%
	Chairs & Chairmats	$1,761,837	34.0%
	Bookcases	$822,652	15.9%
	Office Furnishings	$698,094	13.5%
	Total	$5,178,591	100.0%
Office Supplies	Storage & Organization	$1,070,183	28.5%
	Binders and Binder Accessories	$1,022,958	27.3%
	Appliances	$736,992	19.6%
	Paper	$446,453	11.9%
	Envelopes	$174,086	4.6%
	Pens & Art Supplies	$167,107	4.5%
	Scissors, Rulers and Trimmers	$80,996	2.2%
	Labels	$38,982	1.0%
	Rubber Bands	$15,007	0.4%
	Total	$3,752,762	100.0%
Technology	Office Machines	$2,168,697	36.2%
	Telephones and Communication	$1,889,314	31.6%
	Copiers and Fax	$1,130,361	18.9%
	Computer Peripherals	$795,876	13.3%
	Total	$5,984,248	100.0%
	Grand Total	$14,915,601	100.0%

Fig 4.2 product category as a dimension



Columns Measure Names

Rows ATTR(Category) ⊞ Sub-Category

Title Category changed to an Attribute

Category	Sub-Category	Sales	% of Total Sales along Pane (Down)
Furniture	Tables	$1,896,008	12.7%
	Chairs & Chairmats	$1,761,837	11.8%
	Bookcases	$822,652	5.5%
	Office Furnishings	$698,094	4.7%
Office Supplies	Storage & Organization	$1,070,183	7.2%
	Binders and Binder Accessories	$1,022,958	6.9%
	Appliances	$736,992	4.9%
	Paper	$446,453	3.0%
	Envelopes	$174,086	1.2%
	Pens & Art Supplies	$167,107	1.1%
	Scissors, Rulers and Trimmers	$80,996	0.5%
	Labels	$38,982	0.3%
	Rubber Bands	$15,007	0.1%
Technology	Office Machines	$2,168,697	14.5%
	Telephones and Communication	$1,889,314	12.7%
	Copiers and Fax	$1,130,361	7.6%
	Computer Peripherals	$795,876	5.3%
	Grand Total	$14,915,601	100.0%

Fig. 16-3 product category as an attribute

16.2 WHAT ARE CALCULATED VALUES AND TABLE CALCULATIONS?

Calculated Values and Table Calculations allow you to add new data to your Tableau workbook, but the way you add the data, and where the calculations occur, is different for each method. Calculated Values are defined by entering a formula into Tableau's formula editing dialog box. For example, if you have gross margin dollars and sales dollars in your source data, you may want to add a new field called Gross Margin Percent by creating a calculated value. The formula to create the gross margin percent is: sum([gross margin dollars])/sum([sales dollars]).

The Sum aggregation function in front of each field name tells the source database what to return to Tableau. Calculated values are normally processed at the datasource. What this means is that the power of your database server is used to do the heavy number crunching, with the database returning only what is needed for Tableau to build the visualization. Table calculations are created in a different way—using your data visualization as the source for the formula.

Pre-defined Quick Table Calculations remove the need for you to create the formula manually, but these are always processed locally because they rely on the data presented in your view to derive the formula. Calculated values can also include table calculation functions. These are functions you use in calculated values that are processed locally just like Quick Table calculations.

How Do Calculated Values Work ?

Calculated Values can be used to generate numbers, dates, date-times, or strings. All calculated values require the following elements:

- Functions—including aggregate, number, string, date, type conversion, logical, user, and table calculation types.
- Fields—selected from the datasource.
- Operators—for math and comparison of values, dates, and text.
- Optional elements can be added within the formula dialog box including:
    - Parameters—for creating formula variables that are accessible to information consumers.
    - Comments—for documenting formula syntax and notes within the formula dialog box.

Start the formula dialog box via the main menu using the Analysis/Created Calculated Field option or by right-clicking on a field. The formula dialog is where you enter the functions, operators, and parameters to create the logic for your formula. Alternatively, right-clicking a field in the dimensions or measures shelves opens the formula dialog box as well, but also includes that field already entered in the formula editing area.

People experienced at writing SQL script or creating spreadsheet formulas normally have very little difficultly learning how to write formulas in Tableau. Those with very little experience writing formulas may need more help. Tableau provides assistance via a real-time formula editor and a help window in the formula editing window, as well as an online manual that is accessible from the editing window.

How Do Table Calculations Work ?

Table calculations are derived from the structure of the data included in your visualization, so table calculations are dependent on the source worksheet view contained in your workbook. That means these calculations are always derived locally using your personal computer's processor to return the result. Understanding exactly how Table Calculations work takes a little time because

Table Calculations can change as your visualization is altered. As with any new concept, after you create some Table Calculations you'll get comfortable with how they behave in different situations. Tableau's online manual has a large number of examples that you can view that provide a good basic introduction.

Creating a Table Calculation requires that you have a worksheet with a visualization. A good way to create them is to right click on a measure pill used in the view to expose the Quick Table Calculation menu. Quick Table Calculations are provided for:

- Running total
- Difference
- Percent difference
- Percent of total
- Moving average
- YTD total
- Compound growth rate
- Year over year growth
- YTD growth

Depending on the view of the data included in your worksheet some of these may be unavailable because your worksheet view doesn't support the calculation. Unavailable calculations will be visible in the menu but will appear grayed-out.

A Word on Calculations and Cubes

Tableau connects to relational databases, spreadsheets, columnar-analytic databases, data services, and data cubes (multi-dimensional datasources). Data cubes are different from regular database files because they pre-aggregate data and define hierarchies of dimensions in specific ways.

If you need to access pre-aggregated data that is stored in a multi-dimensional datasource, you can still perform calculations using Tableau formulas or create formulas using the standard query language of multi-dimensional databases, Multidimensional Expressions (MDX). The syntax is a bit more complex but MDX also provides the ability to create more complex formulas. If you desire to learn more about options for creating calculations when accessing Data Cubes, refer to Tableau Software's quick start guide Creating Calculated Fields-Cubes. Tableau's behavior when you connect it to a data cube is different because the cube controls aggregation. For example, date fields behave differently because the cube controls date aggregation in specific ways.

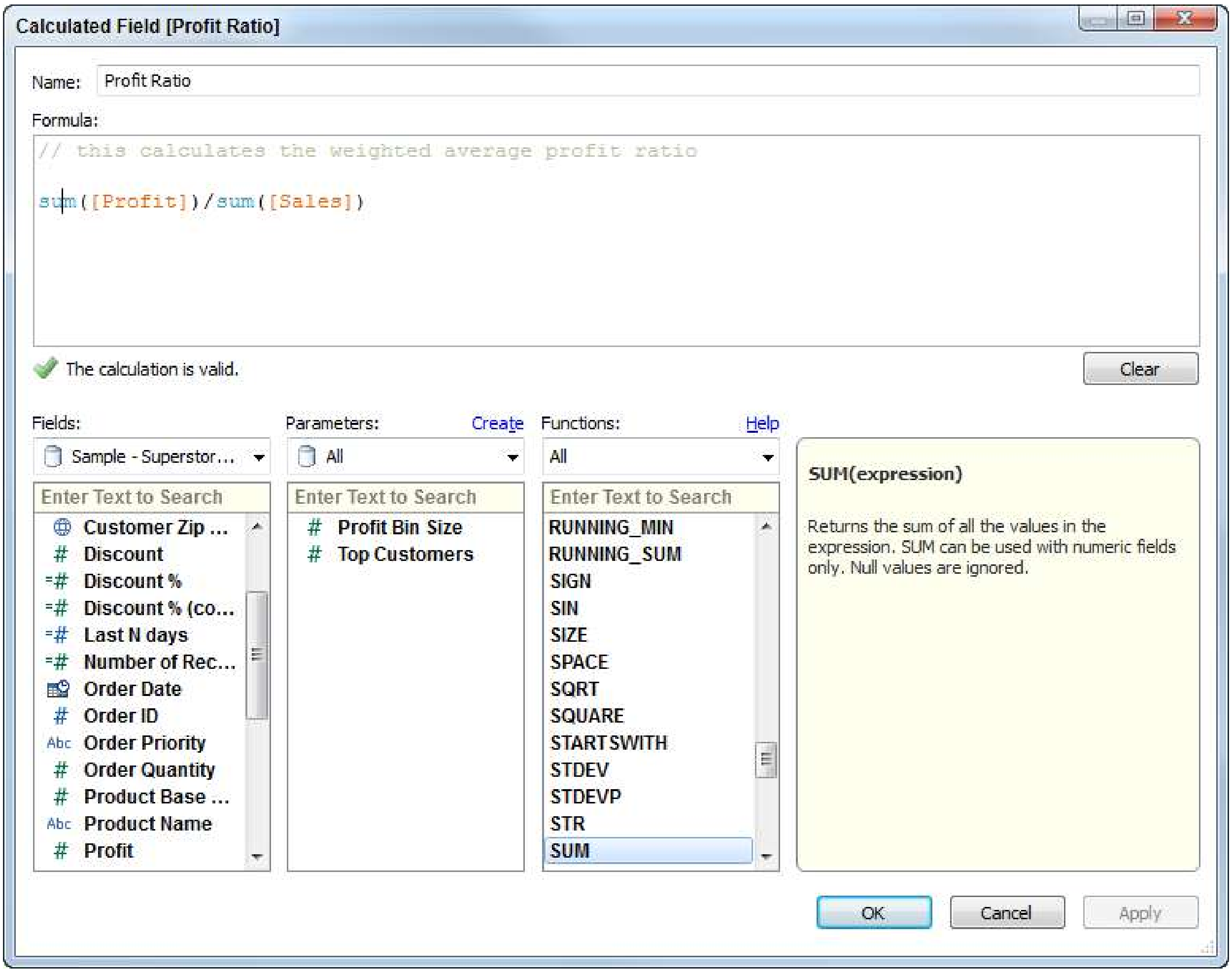
16.3 *USING THE CALCULATION DIALOG BOX TO CREATE*

Calculated Values require that you enter fields, functions, and operators. Tableau strives to make formula creation fast and easy, so it is possible to write formulas with minimal typing. Once you've connected to a datasource, you can create a calculated field from the main menu by selecting Analysis/Create Calculated Field. This example uses the Superstore spreadsheet. Figure 16-4 shows the Calculated Value editing window.

The figure shows a calculation for Profit Ratio that uses two fields from the Superstore file to derive the result. The Name field at the top of Figure 16-4 is where you type the name of your Calculated Value as you want it to appear in the data window of the worksheet. The Formula box is used to write the script for the formula.

You will also see that Tableau color-encodes different elements of formulas so that they are easy to separate visually. Fields are orange, Parameters are purple, and Functions are blue.

Notice the example in Figure 16-4 includes comments at the top, color-encoded in green. Comments are useful for documenting sections of complex formulas or for adding basic descriptive information to other analysts that may use your formula in their work. You can add comments anywhere in the formula window by typing two forward slashes (//) in front of the text



Calculated Field [Profit Ratio]

Name: Profit Ratio

Formula:

```
// this calculates the weighted average profit ratio

sum([Profit])/sum([Sales])
```

The calculation is valid.

Fields: Sample - Superstor...

Parameters: All

Functions: All

SUM(expression)

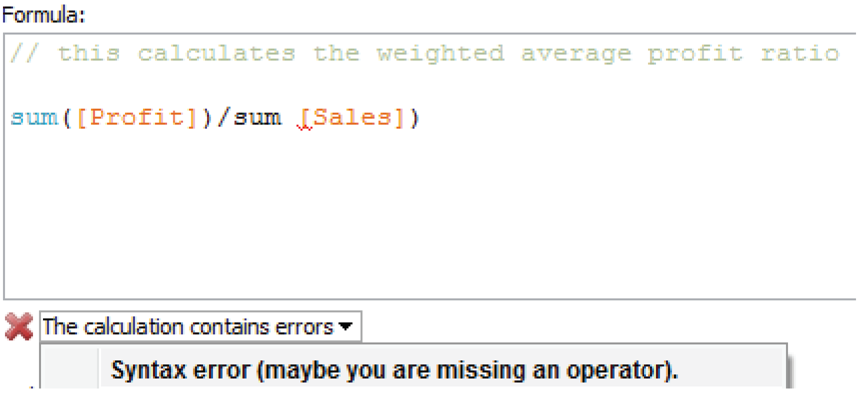
Returns the sum of all the values in the expression. SUM can be used with numeric fields only. Null values are ignored.

Fig 16-4 Formula dialog box or editing window

Below the formula window is a green check mark followed by the statement, The Calculation Is Valid. This is the formula editor that will help you correct syntax errors. If you get something wrong a red X will appear. In Figure 16-5, you see this in action.

For example, if the beginning parenthesis is omitted in front of the sales field, clicking on the error message—or in the formula near the crooked red line— will provide more information about the syntax error. Typing in the missing parenthesis will correct the problem. If you are new to writing formulas, or if you are creating a particularly complex formula, Tableau's editor will help you find and correct errors.

Referring to Figure 16-4 again you can see four panes on the bottom half of the window. These panes display the available fields, parameters, and functions. If you have a particular field or function selected, the yellow window at the far right provides a brief description of the field or the formula definition



Formula:

```
// this calculates the weighted average profit ratio

sum([Profit])/sum [Sales])
```

The calculation contains errors ▾

Syntax error (maybe you are missing an operator).

Fig 16-5 The real-time formula editor

Field Selection

Looking below the Fields title you will see a filter that allows you to select different data sources (if you have more than one being used in your worksheet), or filter for specific data types available (numbers, text, dates, etc.). Figure 16-6 shows this in action with the Number data types only being displayed below. If you have many fields in your source data, a high-level field selection filter may not prune the list enough. In Figure 16-4, notice the small boxes below each window that provide a fuzzy String search for a specific field name. Notice that the Parameter and Function windows also provide the same search capability. You can add fields to your Formula window by typing them manually, pasting them in from a text editor, or by double-clicking on the desired field from the Fields window. If you are new to writing formulas, use the double-click method. Tableau inserts the appropriate syntax automatically. For example, double-clicking on the Profit field in the Fields window will cause the following script to be entered: ([Profit]).

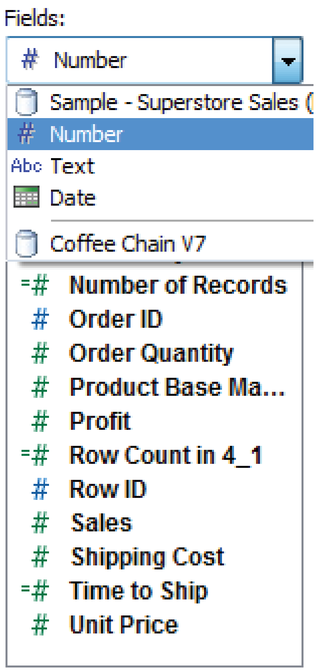
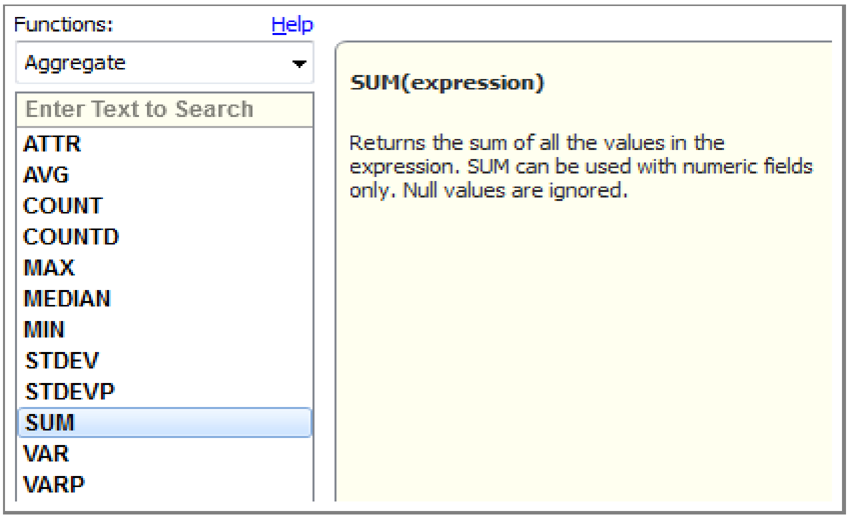


Fig 16.6 Filtering field selections for numbers

Function Selection

Functions can be added exactly the same way. The filter at the top of the Functions window lets you filter for a function category. To add functions without typing, place your cursor within the location of the formula window where you want the function to be placed and double-click on the desired function name in the function window below. Figure 16-7 shows the function window. When the Sum function is selected, the yellow help window displays a brief description of the function along with the function syntax. It you want a more detailed definition, selecting the help menu option will take you to Tableau Software's online manual.



SUM(expression)

Returns the sum of all the values in the expression. SUM can be used with numeric fields only. Null values are ignored.

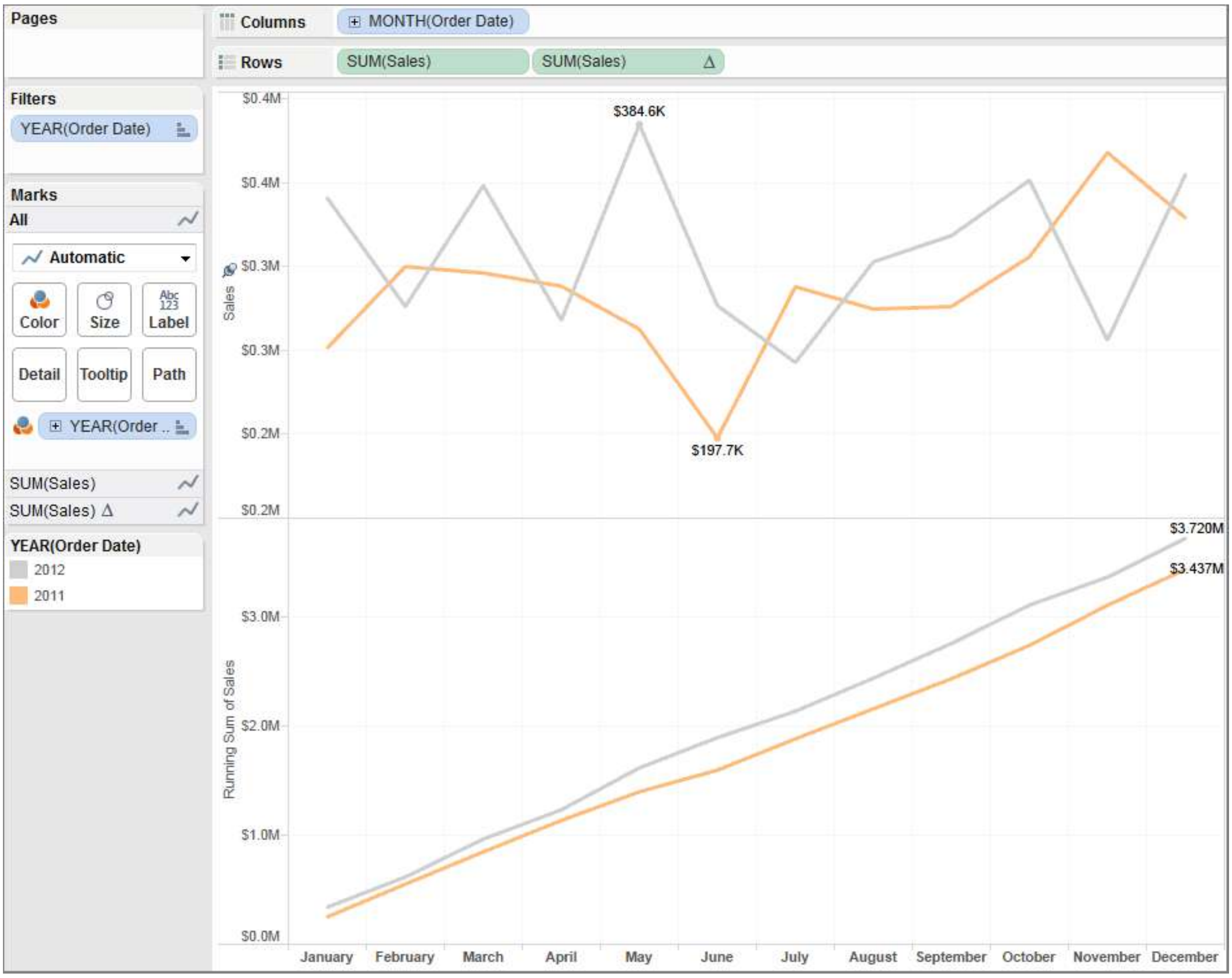
Fig. 16-7 Filter the function window

Parameter Selection

Parameters are optional elements that allow you to add variables in formulas. Figure 16-4 shows two parameters that are included with the Superstore sample file. When you complete editing the formula, don’t forget to click the OK button at the bottom because the new field isn’t created until you do that. If you get interrupted while writing a very long formula either keep your window open, or copy the script to a text editor and save it. When you resume work, you can paste that script back into the formula window and continue. Once you get comfortable with the formula editor and the available functions, you’ll find many ways to leverage Calculated Values.

16.4 *BUILDING FORMULAS USING TABLE CALCULATIONS*

In contrast to Calculated Values, Quick Table calculations use the data in your Visualization to create a formula. Before you can use Quick Table calculations you must first create a worksheet that includes Visualization. Using Superstore again, Figure 16-8 displays a time series of monthly sales on top. The bottom half employs a Quick Table calculation to derive the running total of sales as the year progresses.



Columns: MONTH(Order Date)

Rows: SUM(Sales) SUM(Sales) Δ

Filters: YEAR(Order Date)

Marks: All — Automatic

Color: YEAR(Order ..

SUM(Sales)

SUM(Sales) Δ

YEAR(Order Date)

- 2012
- 2011

Fig 16-8 Time series using a running sum

The steps required to build the charts in Figure 16-8 are:

- Add month to the column shelf.
- Add sales to the row shelf.
- Filter order date for the year(s) 2011 and 2012.
- Add order date to the color marks button.
- Turn on labeling for min/max values.

The data from the Sales Time-Series chart will serve as the datasource for a quick table calculation that will be used to create the chart in the bottom half of Figure 16-8. That chart displays the running sum of sales for each month within the displayed years. The steps required to add that portion of the view are:

1. Ctrl drag the sales pill on the row shelf to create a duplicate chart.

2. Right-click on the second sales pill.

3. Select Quick Table Calculation—Running Total.

4. Turn on field labels for the line ends and un-check Label Start of Line.

Figure 16-9 shows how right-clicking on the duplicate sales pill exposes the Quick Table Calculation menu.

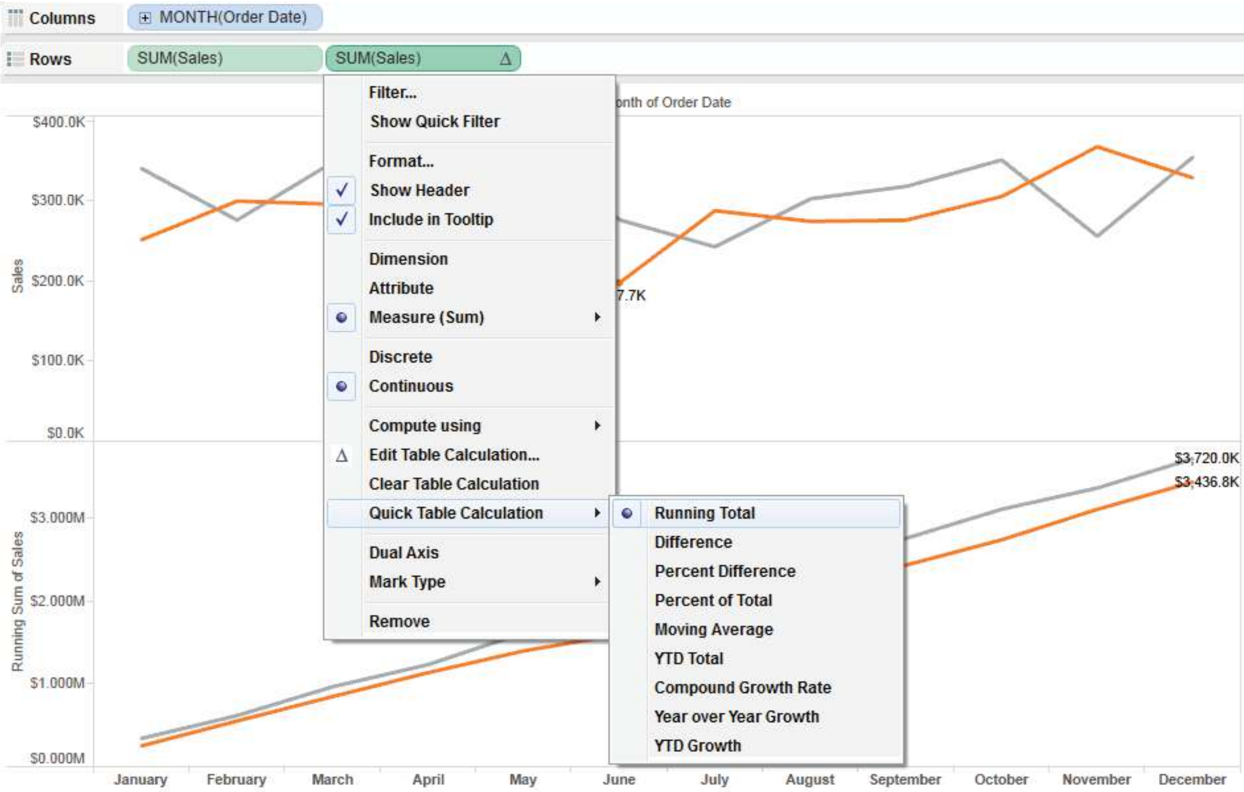


Fig 16-9 creating the quick table calculation

Selecting Running Total generates the table calculation that results in the Running Total Time Series chart. The label number format was also formatted to display the results in thousands in the top chart and millions in the lower chart. The total time required to build this chart was less than 60 seconds.

Editing Table Calculations to Suit Your Purpose

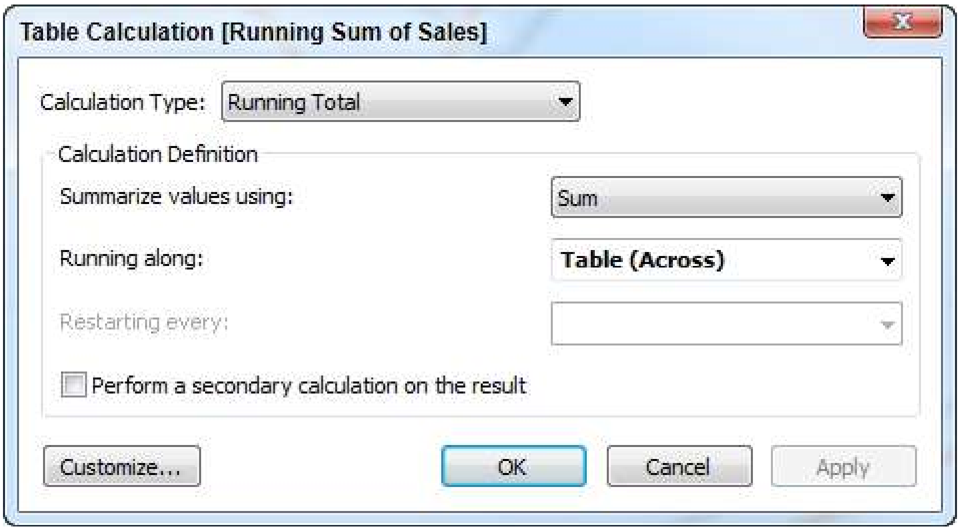


Table Calculation [Running Sum of Sales]

Calculation Type: Running Total

Calculation Definition

Summarize values using: Sum

Running along: Table (Across)

Restarting every:

□ Perform a secondary calculation on the result

Fig 16-10 The table calculation editing menu

You can also see in Figure 16-9 that there are many other Quick Table Calculation options available. There is also a menu option called Edit Table Calculation. In fact, the four rows in the menu below Continuous are all used to customize Table Calculations. Understanding how Table Calculations work takes a little time—playing with the options and looking at the results. Take a close look at the Edit Table Calculation menu option displayed in Figure 16-9.

Table Calculations require selections of the following options:

- Calculation type—as seen in Figure 16-10.
- Aggregation method—sum, average, median, (these will change depending on the content of your source).
- Running Along—defines the direction that the calculation travels (Table Across, Table Down, etc.).

The Restarting Every option is grayed-out in Figure 16-10 because there are no discrete time or other dimension panes dividing the Time Series. Modifying the Time Series to show time as discrete quarters and months creates quarterly partitions as seen in Figure 16-11.

The bottom Time Series showing the running sum of sales is still using Table Across to calculate the total. Right-clicking on the table calculation (denoted by a small triangle on the right side of the pill) and selecting the Edit Table Calculation Menu exposes the Running Along control. Figure 16-12 shows the Table Calculation editing menu for Running Along and includes more options. Adding the partition for quarter creates quarterly panes that can be used in the Table Calculation.

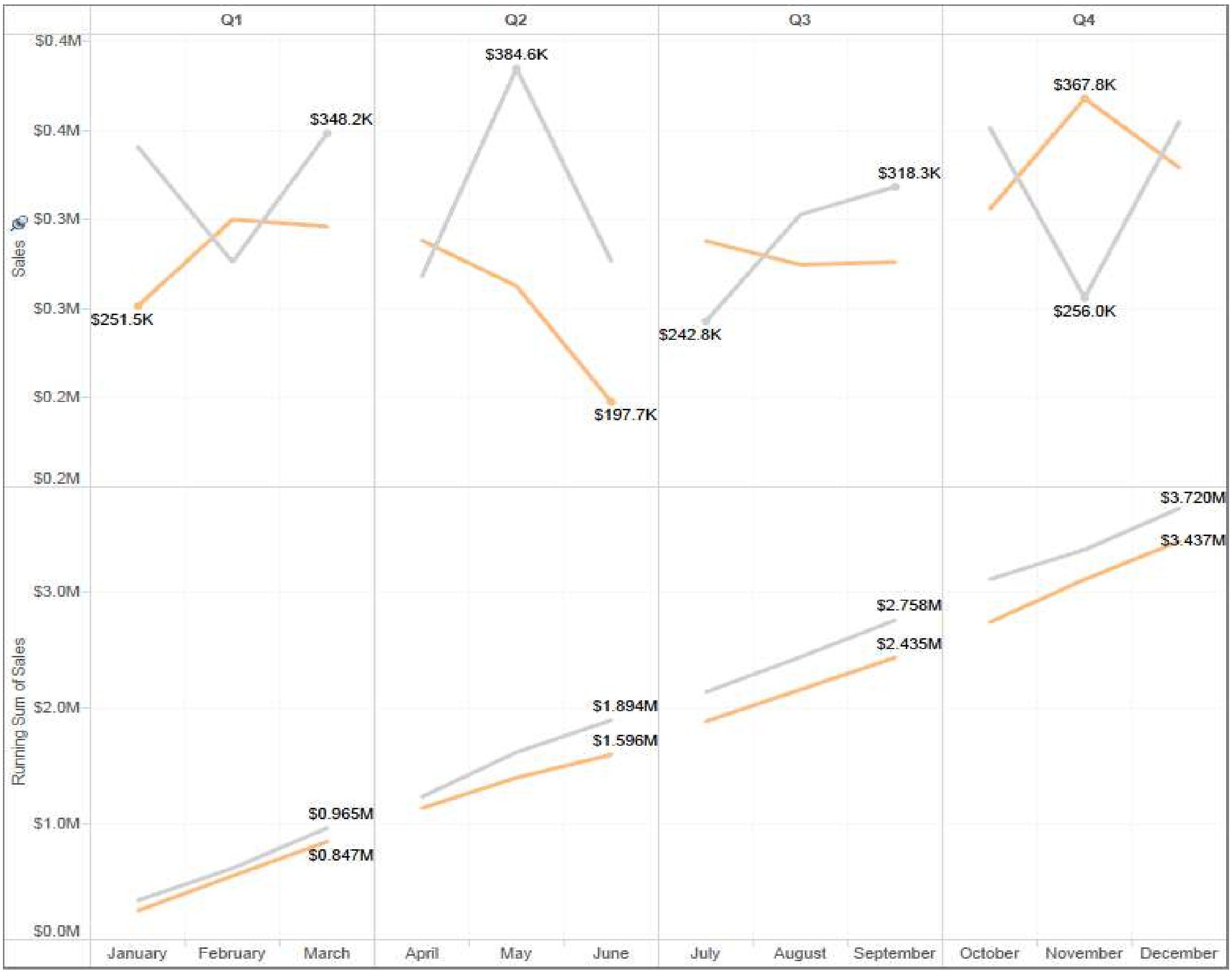


Fig. 16.11 using discrete quarter and month

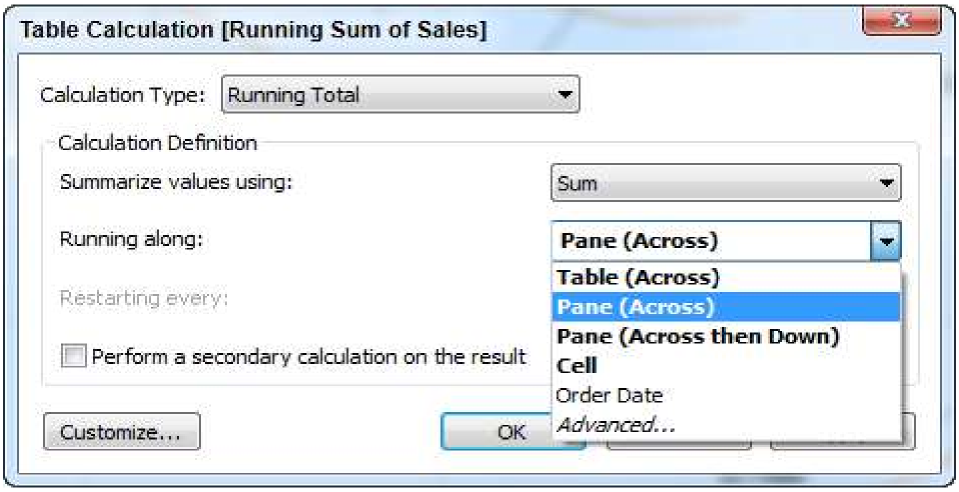


Table Calculation [Running Sum of Sales]

Calculation Type: Running Total

Calculation Definition

Summarize values using: Sum

Running along: Pane (Across)

Restarting every:

□ Perform a secondary calculation on the result

Fig 16-12 Changing table calculation scope

Changing the scope of the calculation to Pane Across causes the Running Sum calculation to reset every quarter (pane). Figure 16-13 reflects the revised scope in the lower pane. As you see, the running totals restart at the beginning of each quarter.

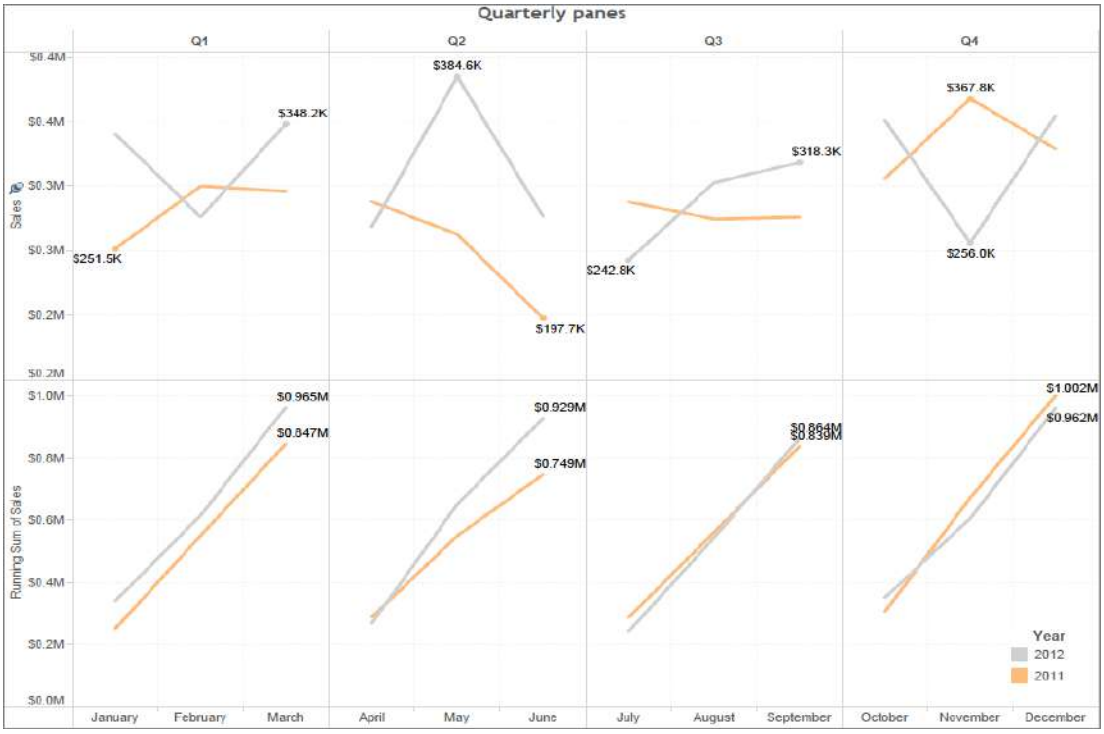


Fig 16-13 Running sum set to pane across

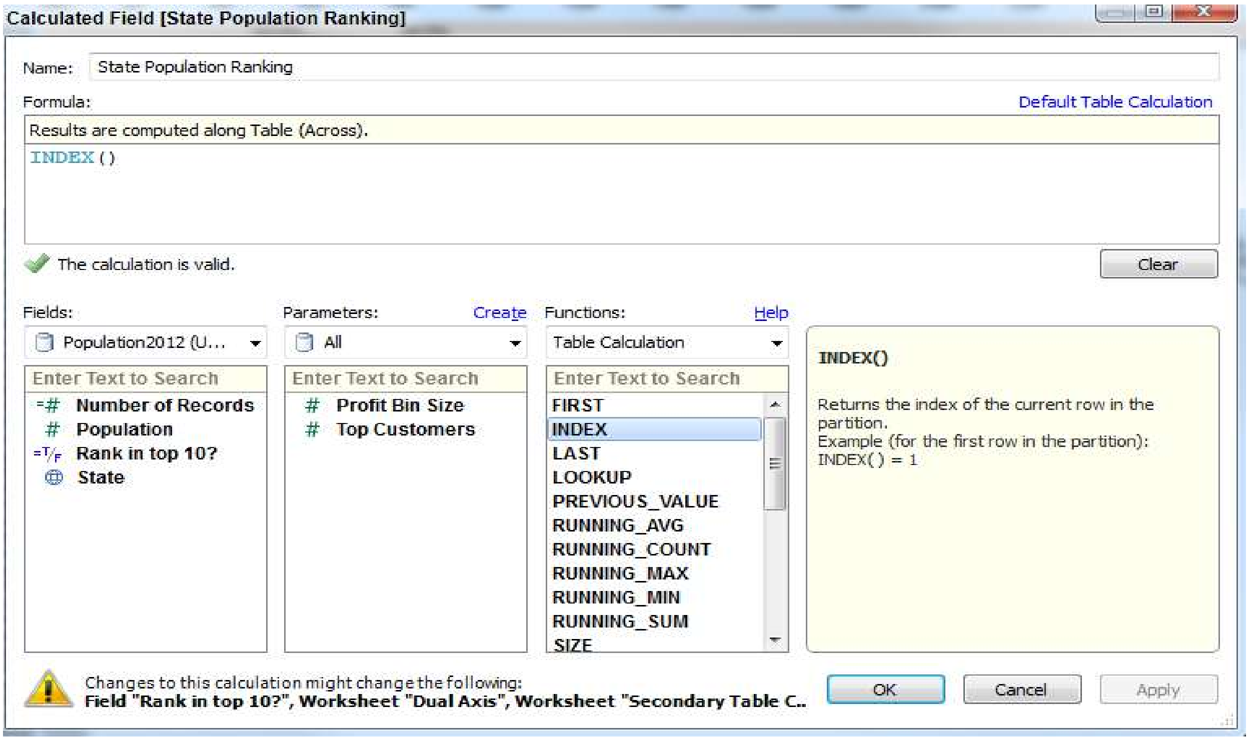
Understanding Table Calculation Behavior

Learning exactly how Table Calculations behave in different Visualizations takes a little time. The best way to learn is to build a crosstab report, then start playing with different options to see the results. Tableau's online manual provides many different examples. Figure 16-14 shows Percent of Total table calculations using all of the different standard Running Along scope options.

Notice that in this example the example for the Table scope returns exactly the same result as the Table Down Then Across scope. Also, the Cell scope is calculating the mark value of itself, resulting in 100 percent in every cell. Depending on the structure of your view it is not uncommon for different scope options to return the same values. In general, adding more dimensions to your view will increase the number of available options provided by Table Calculations. Experiment with different Visualization styles and Table Calculations. With practice you'll be able to anticipate how they behave in different situations.

16.5 *USING TABLE CALCULATION FUNCTIONS*

The Index function is a Table Calculation function that counts the position of a row or column in a set. A calculated value called State Population Ranking was created using this function. Figure 16-20 shows the Calculated Value using the Index function. Creating the Boolean Calculated Value compares the result of the Index to a top 10 ranking value. The resulting Calculated Value is placed on the color shelf for the bar chart to color encode the top 10 states a different color. Figure 16-21 shows the Boolean formula being created.



Calculated Field [State Population Ranking]

Name: State Population Ranking

Formula:

Results are computed along Table (Across).

```
INDEX()
```

The calculation is valid.

Fields: Population2012 (U...

Parameters: All

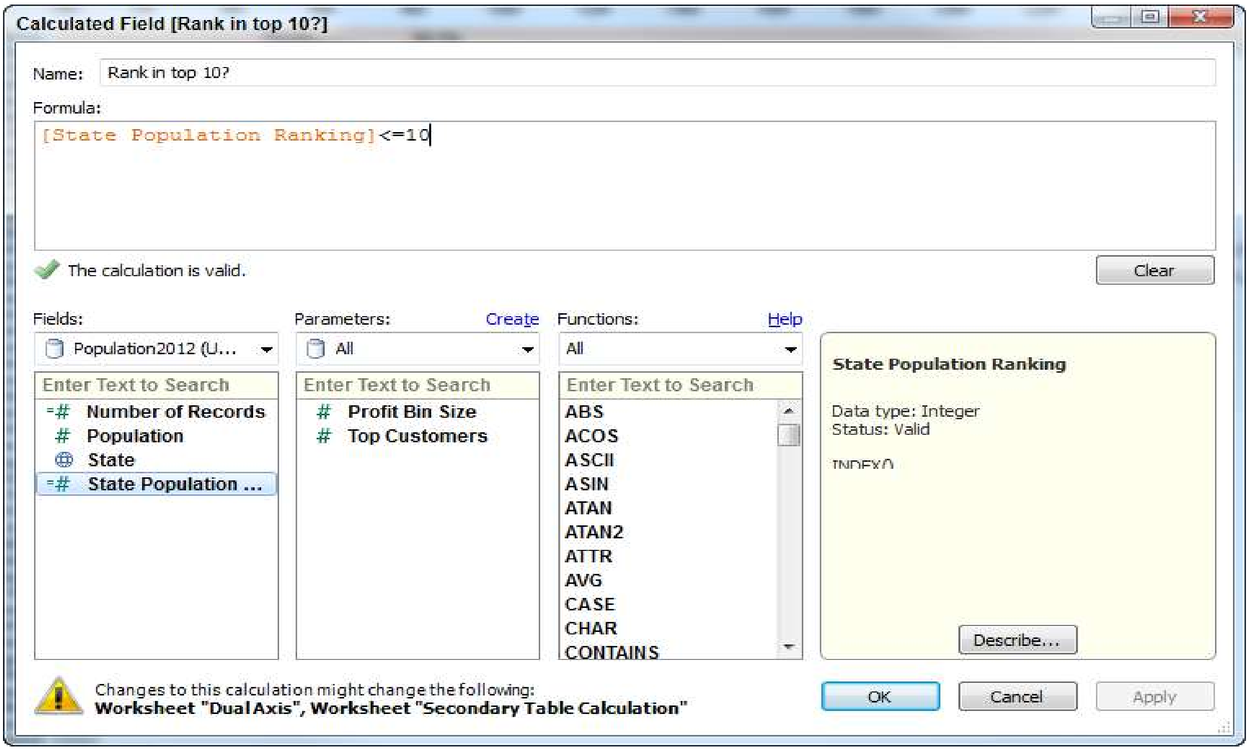
Functions: Table Calculation

INDEX()

Returns the index of the current row in the partition.
Example (for the first row in the partition):
INDEX( ) = 1

Changes to this calculation might change the following:
Field "Rank in top 10?", Worksheet "Dual Axis", Worksheet "Secondary Table C..

Fig 16-20. Creating a population rank



Calculated Field [Rank in top 10?]

Name: Rank in top 10?

Formula:

```
[State Population Ranking]<=10
```

The calculation is valid.

Fields: Population2012 (U...

- Number of Records
- Population
- State
- State Population ...

Parameters: All

- Profit Bin Size
- Top Customers

Functions: All

ABS, ACOS, ASCII, ASIN, ATAN, ATAN2, ATTR, AVG, CASE, CHAR, CONTAINS

State Population Ranking

Data type: Integer
Status: Valid

INDEX()

Changes to this calculation might change the following:
Worksheet "Dual Axis", Worksheet "Secondary Table Calculation"

Fig 16-21 Creating a Boolean calculation

The Boolean formula in Figure 16-21 compares the state population ranking with the number 10 to derive a true-false condition for the top 10 ranked states by population. The resulting Calculated Value is then added to the color button on the Marks card. The resulting color encoding is seen in Figure 16-18. Using Table Calculations in combination with Calculated Values that employ Table Calculation Functions helps you add more meaning and context to analysis. There really is no limit to the creative ways you can use Calculated Values and Table Calculations to enhance information.

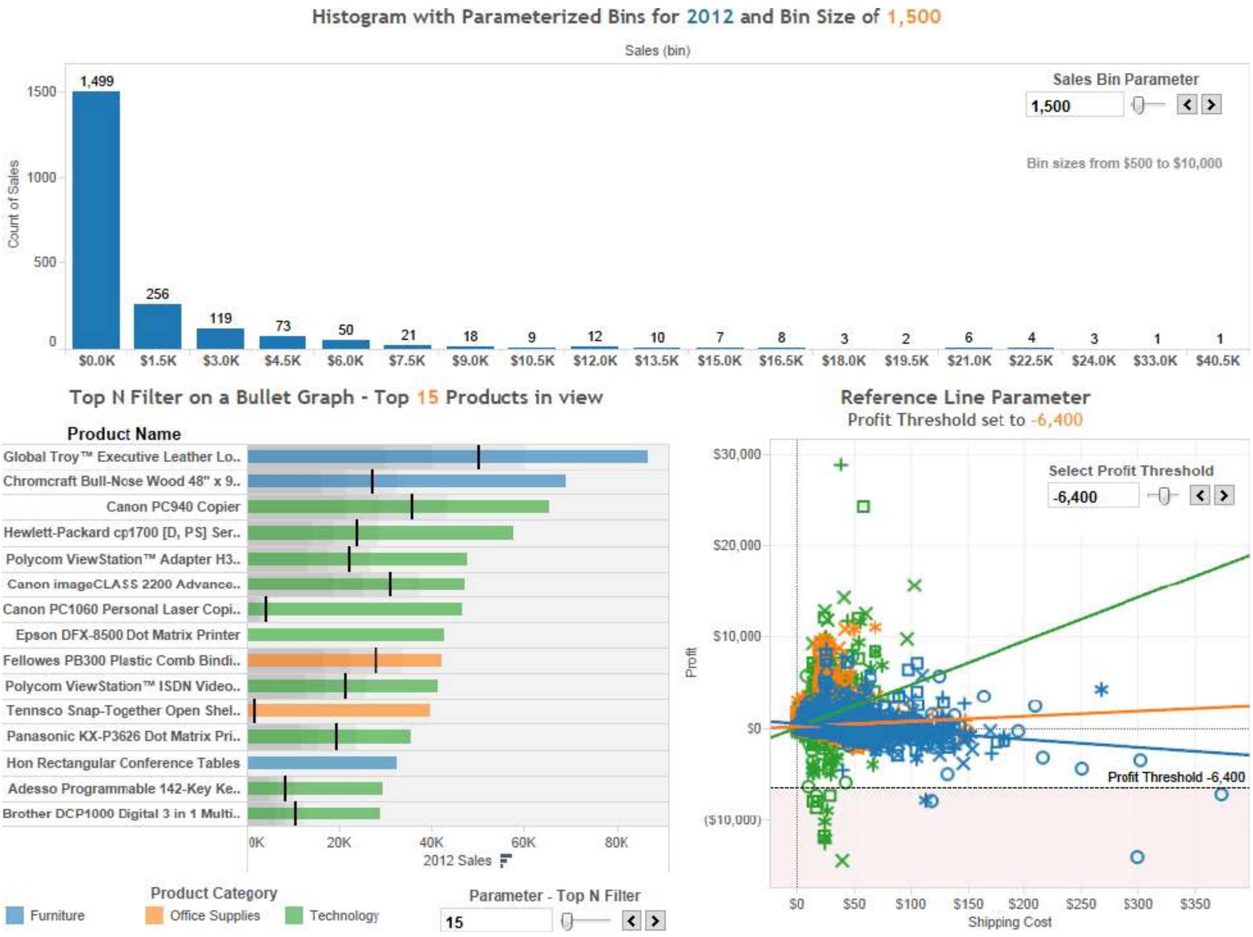
16.6 *ADDING FLEXIBILITY TO CALCULATIONS WITH PARAMETERS*

Parameters empower information consumers to change the content that appears in worksheets and dashboards. Basic parameter controls can be created using embedded options for a limited number of common use cases. Advanced parameters offer the ability to create parameters to address more unique use cases at the expense of a little more time developing the parameter control.

What are Basic Parameters ?

Basic parameters are variables that are provided in specific situations that reduce the number of steps required to create a parameter control. Basic Parameters are available to make flexible top or bottom filters for a specified number of items in a set. In histograms, a parameter can be added that allows users to specify the size of each bin. Reference lines include a parameter option that provides a way to make the reference line change based on a user-selectable parameter value. Figure 16-22 shows the three Basic Parameter controls in action.

The histogram on the top of Figure 16-22 displays order counts by the Size of Orders. The Sales Bin parameter allows the end user to change the size of each bin. The Parameter Size Range is from $500 to $10,000. The bullet graph in the lower left of Figure 16-22 compares sales (bars) to prior year sales (black reference lines) for every product name. The data set includes over 1,000 product names.



Histogram with Parameterized Bins for 2012 and Bin Size of 1,500

Sales Bin Parameter: 1,500

Bin sizes from $500 to $10,000

Top N Filter on a Bullet Graph - Top 15 Products in view

Parameter - Top N Filter: 15

Reference Line Parameter

Profit Threshold set to -6,400

Select Profit Threshold: -6,400

Fig. 16-22 Basic parameter controls

The parameter allows the user to change the number of products displayed through a flexible top down filter. You can see that currently the top 15 products are being displayed. The scatter plot in the lower right includes a reference line called Profit Threshold that allows the user to change the threshold value and change the position of the reference line and the corresponding shading below the line.

All of these are Basic Parameters that are selectable options for these uses. Parameterizing a histogram's bin size is accessed via a right-click on the bin field name that appears in the dimension shelf. The flexible filter in the bullet graph is accessed by right-clicking on the product name dimension and selecting the Top tab in the filter dialog. The reference line parameter is accessed when adding the reference line by clicking the Value drop down selector and picking the Create a Parameter Option. Figure 16-23 shows each of the menus. While Basic Parameters are very easy to create they are also currently limited to the specific use cases you see in Figure 16-23. Top or Bottom Filters, Bin Sizing, or Flexible Reference Lines; if you want to create more advanced parameters, these require a little more effort.

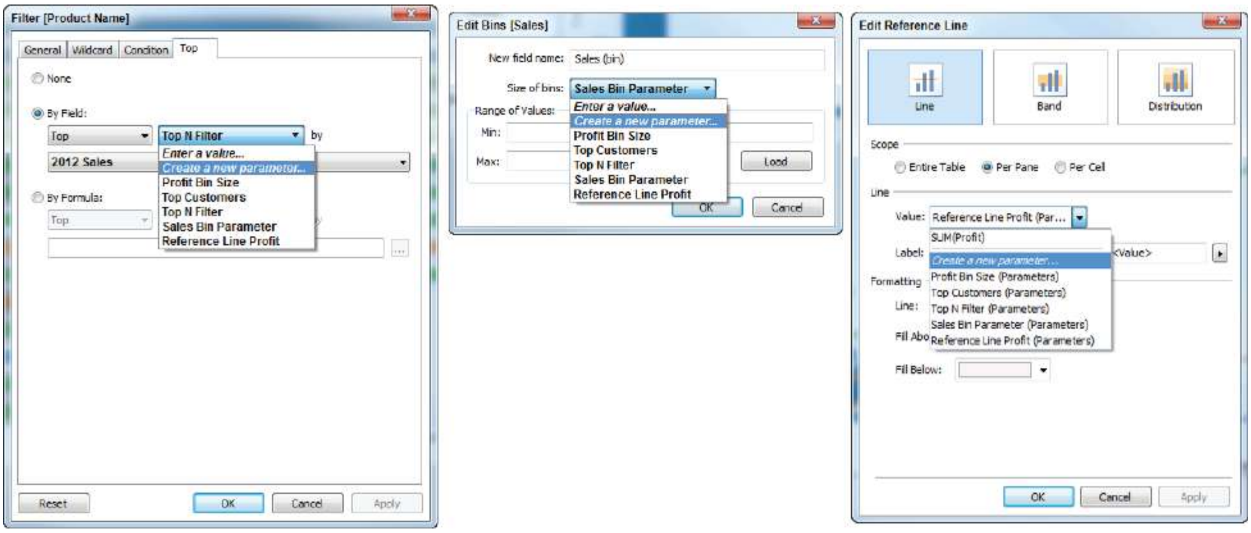


Fig. 16-23 Dialog windows for basic parameter creation

What Are Advanced Parameters ?

Advanced Parameters controls are limited only by your imagination. You can create multiple Parameter Controls. Parameter Controls can be chained together to create linked parameters. An entire book could be written on Parameter Controls because they provide programming-like functionality to Visualizations. Creating Advanced Parameter controls requires three or four steps:

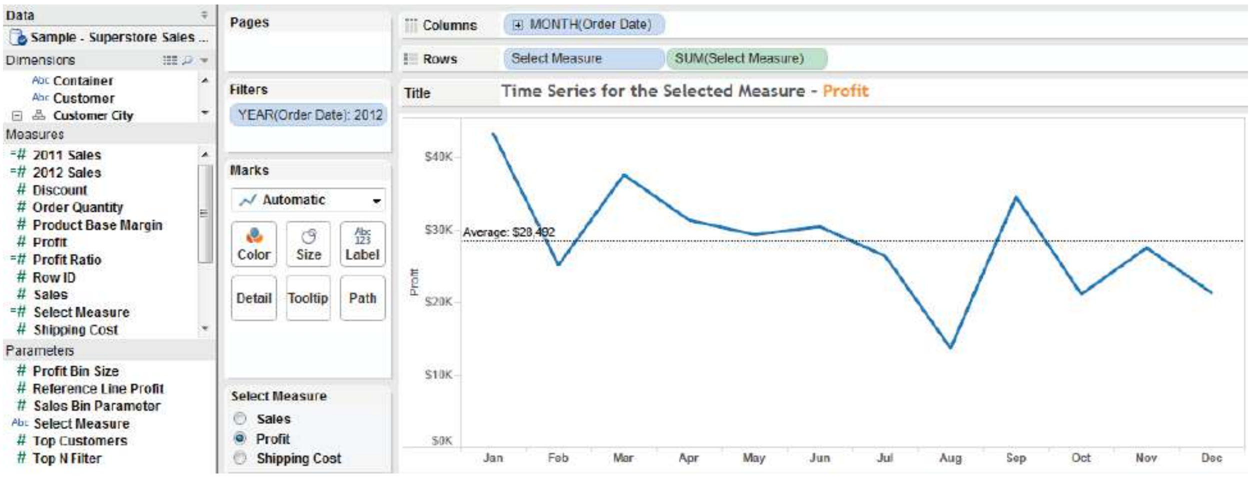
1. Create the parameter control.

2. Expose the parameter control on the desktop.

3. Use the parameter in a calculated value (optional).

4. Use the calculated value in the view.

If the parameter is being directly placed in the Visualization, it may be unnecessary to create a Calculated Value. The key point is that whatever the parameter is being used to change (typically a formula variable), that item must be used somehow in the Visualization in order for the Parameter Control to work. The most popular use cases for Advanced Parameter is that it permits users to change measures or dimensions being displayed in a view. The technique in either case is the same. Figure 16-24 shows a Time Series chart in which a parameter is being used to change the measure plotted.



Data: Sample - Superstore Sales ...

Dimensions: Container, Customer, Customer City

Measures: 2011 Sales, 2012 Sales, Discount, Order Quantity, Product Base Margin, Profit, Profit Ratio, Row ID, Sales, Select Measure, Shipping Cost

Parameters: Profit Bin Size, Reference Line Profit, Sales Bin Parameter, Select Measure, Top Customers, Top N Filter

Filters: YEAR(Order Date): 2012

Marks: Automatic

Columns: MONTH(Order Date)

Rows: Select Measure, SUM(Select Measure)

Title: Time Series for the Selected Measure - Profit

Select Measure
○ Sales
● Profit
○ Shipping Cost

Fig. 16-24 Using a parameter to change the measure displayed in a view

The Parameter Control appears below the Marks card in a radio-button style filter. It allows the user to select three different measures for the time series chart. Currently the view shows profit dollars. Notice that the title of the worksheet includes the parameter and the axis label also changes.

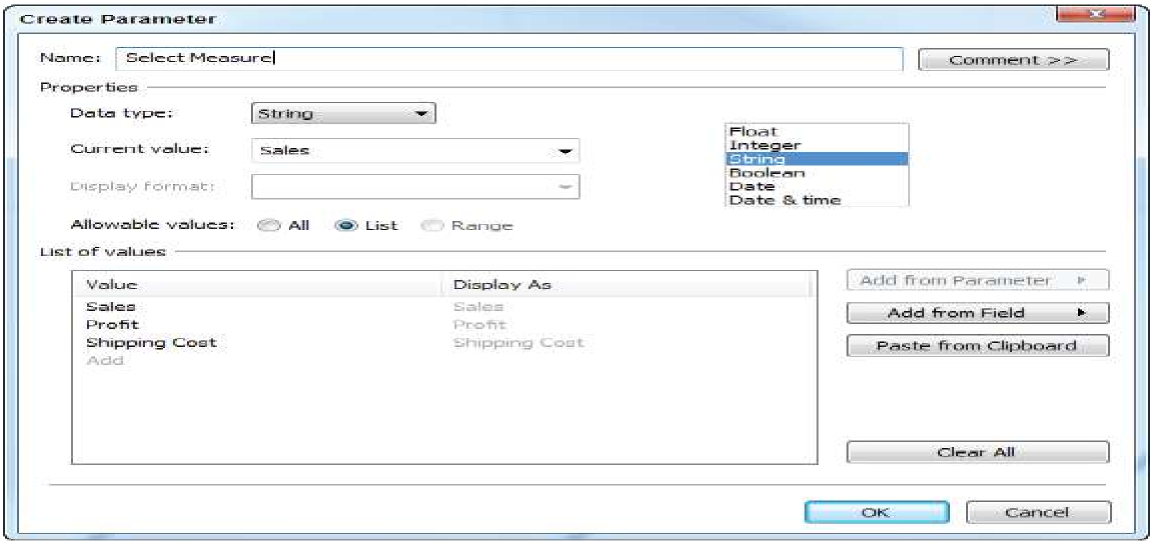
Adding a Parameter Description to the title bar is done by double-clicking on the title bar and selecting the parameter used in the view. To add the Parameter Name to an axis, drag the parameter from the Parameters shelf to the axis. Then edit the axis and erase the static title. This example also rotated the parameter label and removed the label heading. When a new selection is made from the Parameter Control, the Visualization will change along with the headings and reference line to reflect the selected value.

Creating the Parameter Control

This can be done directly in the Formula Editing window or by right-clicking on blank space in the Dimension, Measures, or Parameter shelf. Doing that exposes the dialog window that is used to define the parameter as you see in Figure 16-25. Enter the name of the Parameter as you want it to appear in the control that is placed on the desktop, and then define the data type. Parameters can be numbers (floating decimal point or integers), Strings, Boolean (true/false), and Date or Date and Time values.

The allowable values section is where you define the variables that will contain the Parameter. In Figure 16-24 there is a small list of Measure names defined. While it isn't always desirable, I suggest that for this type of parameter you exactly copy the field names of the Measures. This will make formula creation easier in the next step. However, if you find that the performance of your parameter is not good, use

numbers in a series (1,2,3…) as your value names in the parameter definition. It makes creating the formula in the next step a little more difficult; using numbers in the parameter definition will generally result in a more responsible parameter control. This is especially noticeable with larger data sets. Notice that there is a Display As option. This is used to create a name alias that will



Name: Select Measure

Properties

Data type: String

Float
Integer
String
Boolean
Date
Date & time

Current value: Sales

Display format:

Allowable values: ○ All ● List ○ Range

List of values

Value	Display As
Sales	Sales
Profit	Profit
Shipping Cost	Shipping Cost
Add	

Fig 16-25 Defining a parameter control

Notice that there is a Display As option. This is used to create a name alias that will appear instead of the actual field name. The options to the right of the List of Values section are not applicable to this example, but are useful for cases where you might be using values from another Parameter Control or adding members of a particularly large set. To complete the formula definition, click OK and the parameter will appear on the Parameter shelf.

Expose the Parameter in the Workspace

In order for users to access the Parameter Control it needs to be placed on the desktop. To do this, right-click on the Parameter name appearing in the Parameter shelf and select Show Parameter Control. If you access the parameter now, nothing will happen because you haven’t used the control yet in a formula or in any other way in the Visualization. This is because the parameter hasn’t been used in a formula yet or in any other way in the visualization. The next step is to use this parameter variable in a formula.

Create a Formula That Uses the Parameter Control

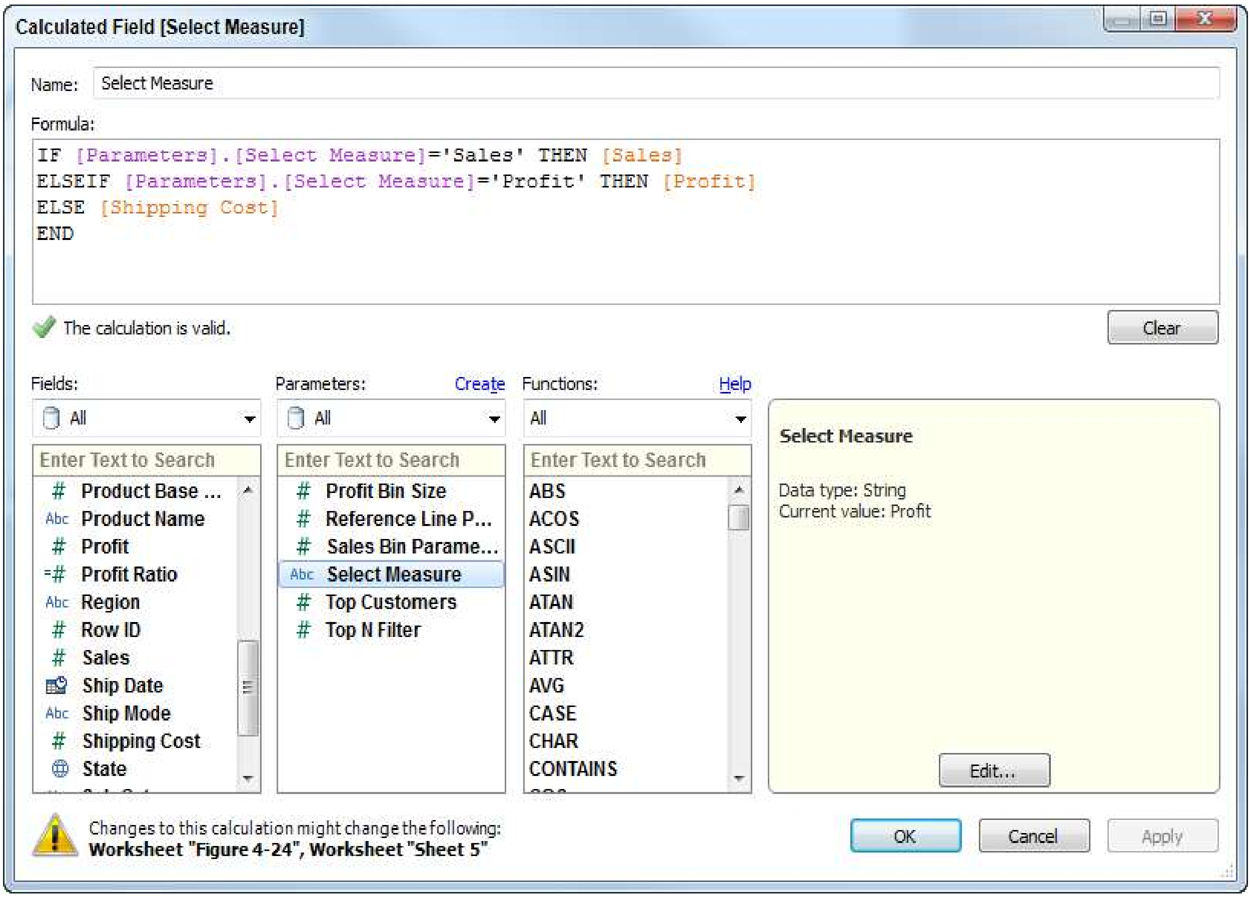
In Figure 16-24 the Parameter Control is used to change the Measure being plotted in the Time series. This requires a formula that will link the String values defined in the parameter to measure field names in the datasource. You can see the formula definition in Figure 16-26.

Now the parameter variable comes into play. The formula logic associates the selected parameter string with the related field name. This is why it is a good idea to define the Parameter String names to exactly match the field names you want to associate. It just makes writing the formula easier. But keep in mind that if performance degrades, using sequentially-ordered numeric values in the parameter definition will result in the best performance.

Clicking OK adds the Calculated Value to the Measure shelf with the name Select Measure. It’s also a good idea to give your parameter name the same name as the related calculations, especially if you have many parameters defined in the worksheet. This just makes it easier to retrace your work at a later date if you need to modify the Parameter Control to add or delete items.

Use the Calculated Value in the View

Dragging the Select Measures measure to the Row shelf will activate the Parameter Control. Each selection made in the parameter control will trigger changes in the Select Measure formula and will change the measure being displayed in the Time series.



Calculated Field [Select Measure]

Name: Select Measure

Formula:

```
IF [Parameters].[Select Measure]='Sales' THEN [Sales]
ELSEIF [Parameters].[Select Measure]='Profit' THEN [Profit]
ELSE [Shipping Cost]
END
```

The calculation is valid.

Select Measure

Data type: String
Current value: Profit

Changes to this calculation might change the following:
Worksheet "Figure 4-24", Worksheet "Sheet 5"

Fig. 16-26 using a parameter in a formula

Parameters can be used to create multi-purpose Visualizations. There are many different ways that Advanced Parameters can be used. The limit is your imagination. For more examples, go to Tableau Software's website and search for Parameters. You should find many different forum posts that relate to parameters and some training videos.

16.7 *USING THE FUNCTION REFERENCE APPENDIX*

Tableau provides good online documentation of Functions. The user forum on Tableau's website is also quite good. However, many novice users have asked for a more detailed reference for Tableau Functions that provide examples and explain the formula syntax in more detail. Functions are listed by function type, alphabetically. Each Function Reference entry provides a short description of the Function, typical use cases, and basic, intermediate, and advanced examples. Hopefully you'll find the Function Reference a useful addition to your tool set. As questions come in, the book's companion website will provide additional tips and tricks related to Functions, Parameters, dashboard building, and other topics that merit an ongoing discussion.

16.8 CHECK YOUR PROGRESS

1. What Are Calculated Values and Table Calculations?
2. Write the steps to create advanced parameter controls
3. List three reasons why Quick Table Calculations are provided?

Answers to Check your progress

1. Calculated Values are defined by entering a formula into Tableau's formula editing dialog box. Table calculations are created in a different way—using your data visualization as the source for the formula.

2. 1. Create the parameter control. 2. Expose the parameter control on the desktop. 3. Use the parameter in a calculated value (optional). 4. Use the calculated value in the view.

3. Running total

   Difference

   Percent difference

16.9 SUMMARY

Tableau provides two ways to enhance your data through the creation of new fields that don't exist in your datasource. Tableau also allows you to turn single-purpose dashboards and views into multi-purpose analysis environments though parameter controls. Parameters are formula variables that can be used to provide filter-like controls that allow users to change the measures and dimensions used in a dashboard or worksheet.

In this unit you learnt how to use calculated values and table calculations to derive facts and dimensions that don't exist in your source data. Tableau's Formula Editing window is explained as well as the Quick Table Calculation menu, and how to modify Quick Table defaults to address your specific needs. In the sections at the end of this unit on parameters, you have learnt parameter controls—basic and advanced—so that you can make views that address different needs using the same basic visual design. Tableau makes formula creation as easy as it can possibly be, but it helps to understand the concept of aggregation, and the functions and operators that are available to use before you start making formulas.

16.10 KEYWORDS

- Median **-** n statistics and probability theory, the median is the value separating the higher half from the lower half of a data sample

- Standard Deviation **- I**n statistics and probability theory, the median is the value separating the higher half from the lower half of a data sample

- Average: In ordinary language, an average is a single number taken as representative of a list of numbers, usually the sum of the numbers divided by how many numbers are in the list (the arithmetic mean).

- Count: o indicate or name by units or groups so as to find the total number of units involved : number

16.11 QUESTIONS FOR SELF-STUDY

1. List the Tableau supported aggregation types.

2. Explain how do calculated values and table calculations work in Tableau

3. Explain how to use the calculation dialogue box to create calculated values.

16.12 REFERENCES

1. Alexander Loth - Visual Analytics with Tableau-Wiley (2019)
2. Dan Murray - Tableau Your Data!_ Fast and Easy Visual Analysis with Tableau Software-Wiley (2013)
3. David Baldwin - Mastering Tableau-Packt Publishing (2017).