KARNATAKA STATE OPEN UNIVERSITY

MBA PROGRAMME

II YEAR- III SEMESTER

SPECIALIZATION: IT

BUSINESS INTELLIGENCE AND ANALYTICS

COURSE: MBA 117G/SC G3.2 BLOCKS: 1-4

DEPARTMENT OF STUDIES AND RESEARCH IN MANAGEMENT

KARNATAKA STATE OPEN UNIVERSITY

MUKTHAGANGOTHRI, MYSURU- 570 006.

DEPARTMENT OF STUDIES AND RESEARCH IN MANAGEMENT

M.B.A III SEMESTER

COURSE - MBSC - 3.2 G

BUSINESS INTELLIGENCE AND ANALYTICS

BLOCK - 1: INTRODUCTION AND BASICS

UNIT - 1
UNDERSTANDING BUSINESS INTELLIGENCE 01-16

UNIT - 2
FITTING BI WITH OTHER TECHNOLOGY DISCIPLINES 17-29

UNIT - 3
MEETING THE BI CHALLENGE 30-44

UNIT - 4
DATA WAREHOUSING 45-62

BLOCK-2: BUSINESS INTELLIGENCE USER MODELS

UNIT - 5
MATHEMATICAL MODELS AND DATA MINING 63-90

UNIT - 6
ONLINE ANALYTICAL PROCESSING 91-114

UNIT - 7
AGILE DEVELOPMENT 115-137

UNIT - 8
ADVANCED/EMERGING BI TECHNOLOGIES 138-155

BLOCK-3: THE BI LIFECYCLE

UNIT - 9
THE BI BIG PICTURE 156-171

UNIT - 10
HUMAN FACTORS IN BI IMPLIMENTATIONS 172-186

UNIT - 11
TAKING A CLOSER LOOK AT BI STRATEGIES 187-207

UNIT - 12
BUILDING A SOLID BI ARCHITECTURE AND ROADMAP 208-230

BLOCK-4:TABLEAU FOR VISUAL DATA ANALYTICS

UNIT - 13
INTRODUCTION TO TABLEAU 231-245

UNIT - 14
CONNECTING TO YOUR DATA 246-277

UNIT - 15
DATA VISUALISATION 278-325

UNIT - 16
CALCULATIONS WITH TABLEAU 326-354

CREDIT PAGE

Programme Name : MBA **Year/Semester : 2nd Year, 3rd Semester** **Block No: 1 to 4**

Course Name : Business Intelligence and Analytics **Credit : 04** **Units No: 1 to 16**

Course Design Expert Committee

Prof. Vidya Shankar Vice-Chancellor Karnataka State Open University Mukthagangothri, Mysore-06	**Chairman**	**Prof. Kamble Ashok** Dean (Academic) Karnataka State Open University Mukthagangothri, Mysore-06	**Member**

Course Designer/Course Co-ordinator

Dr. Rajeshwari H. Assistant Professor, DOS & R in Management KSOU, Mukthagangothri, Mysore-06	**BOS Chairman & Member**	**Dr. P. Savitha** Assistant Professor, DOS & R in Management KSOU, Mukthagangothri, Mysore-06	**Department Chairman & Member Convener**

Editorial Committee

Dr. Rajeshwari H. Assistant Professor DOS & R in Management KSOU, Mukthagangothri, Mysore-06	**BOS Chairman & Member**	**Prof. D. S. Guru** Dept. of Computer Science UoM, Mysuru	**External Subject Expert & Member**
Dr. Rajeshwari H. Assistant Professor DOS & R in Management KSOU, Mukthagangothri, Mysore-06	**Internal Subject Expert & Member**	**Dr. P. Savitha** Assistant Professor, DOS & R in Management KSOU, Mukthagangothri, Mysore-06	**Department Chairman & Member Convener**

Course Writers

Dr. Ashok Rao
Former Head, Network Project,
CEDT, IISc, Bangalore

Course Editor

Dr. Rashmi B S **Unit 01-04**
Assistant Professor **Unit 05-08**
DoS in Information Technology
KSOU, Mysore.

Dr. Kouser **Unit 09-12**
Assistant Professor **Unit 13-16**
DoS in Computer Science
Govt. First Grade College,
Gundulpet

Copy Right

Developed by the Department of Studies and Research in Management, KSOU, under the guidance of Dean (Academic), KSOU, Mysuru
Karnataka State Open University, January-2021

Printed and Published on behalf of Karnataka State Open University. Mysuru-570006 by **Registrar (Administration)-2021**

Karnataka State Open University

Mukthagangothri, Mysore – 570 006.

MBA in Information Technology

Business Intelligence and Analytics

Block 1: Introduction and Basics

Unit No.	Title	Page numbers
Unit 1	Understanding Business Intelligence	5- 16
Unit 2	Fitting BI with Other Technology Disciplines	17 - 30
Unit 3	Meeting the BI Challenge	31 - 45
Unit 4	Data warehousing	46 - 61

BLOCK INTRODUCTION

This module lay the BI groundwork and will keep get you covered for the study of Business Intelligence. You'll also get to know BI's family tree, where it all began, and what related technologies you should get to know. Not much bits-and-bytes talk is necessary because, as you'll see in this module, business intelligence is about business first, technology second.

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

Unit 1- Understanding Business Intelligence: Introduction, Limited Resources and Limitless Decisions, Business Intelligence Defined, The BI Value Proposition, A Brief History of BI, BI's Split Personality: Business and Technology, So, Are You BI Curious?

Unit 2- Fitting BI with Other Technology Disciplines: Best Friends for Life: BI and Data Warehousing, ERP and BI: Taking the Enterprise to Warp Speed, Customer's Always Right, BI-BUY! E-Commerce Takes BI Online, The Finance Function and BI

Unit 3- Meeting the BI Challenge: What's Your Problem? - What can go wrong, The BI Spectrum — Where Do You Want It? First Glance at Best (and Worst) Practices

Unit 4- Data warehousing: Definition of data warehouse**,** Data warehouse architecture**,** Evolution of information systems cubes and multidimensional analysis

UNIT -1: UNDERSTANDING BUSINESS INTELLIGENCE

Structure

1.0 Objectives

1.1 Introduction

1.2 Limited Resources, Limitless Decisions

1.3 Business Intelligence Defined

1.4 The BI Value Proposition

1.5 A Brief History of BI

1.6 BI's Split Personality: Business and Technology

1.7 Benefits of Business Intelligence

1.8 Check your progress

1.9 Summary

1.10 Keywords

1.11 Self Assessment Questions

1.12 References

1.0 OBJECTIVES

After studying this unit, you will be able to:

- ✓ Analyze the basic concepts
- ✓ Explain the business intelligence value proposition
- ✓ Seeing where BI came from and where it's going
- ✓ Previewing what works (and what doesn't)

1.1 INTRODUCTION

From the CEO down to the lowest levels of any organization, every minute of the day someone is making a decision that has an impact on the company's performance. Sometimes a decision is at a very high strategic level that affects the fate of the entire organization, and other times a decision might be narrowly defined and tactical, affecting a single person or department for a

very short window of time. When taken together, these decisions make up a significant portion of the "day in the life" at any given organization, be it a company, governmental agency, or nonprofit organization. In spite of the dramatic advances in technology and tools that aid in the decision-making process, however, far too many people still make decisions the old-fashioned way: by blending a gumbo of tidbits of current information, best recollections of the past, advice from others, and a whole lot of "gut instinct," and then assessing which path is likely to give the best possible outcome for the decision at hand. Decisions drive organizations. Making a good decision at a critical moment may lead to a more efficient operation, a more profitable enterprise, or perhaps a more satisfied customer.

So it only makes sense that the companies that make better decisions are more successful in the long run. That's where business intelligence comes in. Business intelligence is defined in various ways (our chosen definition is in the next section). For the moment, though, think of BI as using data about yesterday and today to make better decisions about tomorrow. Whether it's selecting the right criteria to judge success, locating and transforming the appropriate data to draw conclusions, or arranging information in a manner that best shines a light on the way forward, business intelligence makes companies smarter. It allows managers to see things more clearly, and permits them a glimpse of how things will likely be in the future

1.2 LIMITED RESOURCES, LIMITLESS DECISIONS

All organizations whether business, government, charitable, or otherwise, have limited resources for performing their missions. Companies are forced to make do with what they have — all the time. You can't put a Nobel laureate in every position, and you can't pour unlimited dollars into an endless quest to make all your factories and offices more efficient. The most precious resource is time. The marketplace is in constant motion, and companies must not only move correctly, they must move quickly. Otherwise competitors will fill any available vacuum in the market, resources will get used up, and your organization will inexorably wither away.

Business intelligence's entire reason for being is as an ally at those inflection points throughout the life of a business where a decision is required. Business intelligence is a flexible resource that can work at various organizational levels and various times — these, for example:

- A sales manager is deliberating over which prospects the account executives should focus on in the final-quarter profitability push
- An automotive firm's research-and-development team is deciding which features to include in next year's sedan
- The fraud department is deciding on changes to customer loyalty programs that will root out fraud without sacrificing customer satisfaction

The decisions can be strategic or tactical, grand or humble. But they represent two roads diverging in a yellow wood: Considered in the aggregate, the roads taken and those not taken represent the separation between successful and unsuccessful companies. Better decisions, with the help of business intelligence, can make all the difference.

1.3 BUSINESS INTELLIGENCE DEFINED

So what is business intelligence, anyway? In essence, BI is any activity, tool, or process used to obtain the best information to support the process of making decisions. Whether you're calling the Psychic Hotline, using an army of consultants, or have banks of computers churning your data; if it helps you get a better handle on your company's current situation, and provides insight into what to do in the future, it's BI. But by popular demand we'll narrow the definition just a tad. For our purposes, BI revolves around putting computing power (highly specialized software in concert with other more common technology assets) to work, to help make the best choices for your organization. But before digging into specifics, you should understand some context about how BI is defined, and who's defining it. The more you learn about BI, the more likely you are to encounter a wide swath of definitions for the term.

BI is technology and tools to support decision-making. Business intelligence is essentially timely, accurate, high-value, and actionable business insights, and the work processes and technologies used to obtain them. Today's common definitions of the essential BI components are markedly different from the definitions bandied about in the 1990s. What remains constant, though, is that BI's purpose has always been to produce timely, accurate, high-value, and actionable information

1.3.1 BI'S BIG FOUR

So what do we mean when we talk about insights that are accurate, valuable, timely, and actionable? As you dig into BI's main characteristics, you'll see why each is so important to the process. In fact, if the knowledge gained from BI fails to meet any of the four criteria, the process has failed

Accurate answers

When decisions are taken in your organization they are inevitably informed with conclusions drawn by a range of experts using important pieces of information about the enterprise's current state. For BI to be of any value in the decision making process, it must correctly reflect the objective reality of the organization, and adhere to rigid standards of correctness. As such, the first hallmark of insights produced from BI processes is their accuracy. As with any technology-related tool or process, the GIGO rule is in full effect with BI — that's Garbage In, Garbage Out. GIGO says that if the BI insights are not accurate, the decisions made are less likely to be the correct ones for your enterprise. Imagine a sample BI report that shows one of the company's sales territories lagging woefully behind the others. When folded into the decision-making process, that piece of knowledge might well lead executives to adjust the sales process (or perhaps the personnel). But if the picture is wrong — say the offices and departments were incorrectly aligned to the various territories, so sales dollars weren't correctly allocated — then the conclusions (and the resulting actions taken) not only fail to help the company, they might actually make things worse.

Getting it right is important from a political perspective as well. For BI to have an impact, company stakeholders (those key employees whose business domains affect, and are affected by, BI) must trust it. Nothing's more frustrating in the world of business intelligence than a development team toiling for months to produce a report that an executive looks at and, within 30 seconds, dismisses it by saying, "Those numbers aren't correct."

But such things are common. After all, BI insights are often surprising, counterintuitive, and even sometimes threatening to groups within an organization. The sales manager who is shown numbers that indicate her team is lagging behind will be motivated to find ways to challenge the validity of the report. Any errors, no matter how small, will call into question the veracity of the conclusions drawn from the data. BI must represent the absolute closest thing to the truth that's possible, not only to produce results, but to protect its reputation among the skeptics! Without

accuracy, insights that are the product of BI are worse than worthless. They can be harmful to the company. And once that happens, nobody will ever trust BI again

Valuable insights

Not all insights are created equal. Imagine, for example, that after a multimillion-dollar BI-driven probe of sales-history data, a grocery store chain finds that customers who bought peanut butter were also likely to buy jelly. BI insights like this are certainly accurate, but they are of limited value to the decision makers (who probably know that most supermarkets place those two items close together already). Part of what distinguishes BI is that its goal is not only to produce correct information, but to produce information that has a material impact on the organization — either in the form of significantly reduced costs, improved operations, enhanced sales, or some other positive factor. Further, high-value insights usually aren't easily deduced — even if data-driven analysis weren't readily available. Every company has smart people working for it who can connect the obvious dots. BI insights aren't always obvious, but their impact can be huge.

On-time information

Have you ever had a heated discussion with someone and thought of the perfect retort to their witless argument exactly five minutes after you walk away from them? You never think of your best comeback until you've left a person's apartment or office and are walking down the stairs in defeat. The lesson is simple: What makes people effective in a debate is that they can not only deliver sound information, they can do it at the precise time it's needed.

In business, information delays can make just as big a difference — and they can come in many forms:

- Sometimes it's a technology problem where the hardware or software can't compute fast enough to deliver information to users.
- Sometimes the problems relate strictly to workflow and logistics; the data isn't fed into the systems often enough.
- Logistics problems can pop up from time to time — for instance, what if a report has to be translated into a different language?

Every step in the process takes time, whether it involves microchips or humans. In the aggregate, those time intervals must be small enough to make the output of a BI process still relevant, useful, and valuable to a decision maker. Timeliness is as important a quality in your business insight as any other. The best decision support processes involve up to the minute information and analysis made available to decision makers in plenty of time to consider all the courses of action. Stock traders at hedge funds use massive spreadsheets full of constantly updated data.

The data streams in and is manipulated in a series of processes that makes it usable to the trader. He or she buys and sells stocks and bonds using the results of those calculations, making money for the firm and its clients. If the trader's applications were slower in producing translated data, they would miss opportunities to execute the most profitable trades.

Actionable conclusions

Accurate is one thing, actionable is another. Imagine if the conclusions reached at the end of the BI cycle were that the company would be better off if a competitor would go out of business, or if one of its factories were 10 years old instead of 30 years old. Those ideas might be accurate — and it's no stretch to believe that if either scenario came to pass, it would be valuable to the company. But what, exactly, are the bosses supposed to do about them? You can't wish a competing company out of business. You can't snap your fingers and de-age a factory. These are exaggerated examples but one of the biggest weaknesses of decision support tools is that they build conclusions that are not actionable.

To be actionable, there has to be a feasible course that takes advantage of the situation. It has to be possible to move from conclusion to action. Ideally, the BI team at your company would produce a report that would guide future actions. The executives would conclude that a price should be lowered, or perhaps that two items should be sold as a package. These are simple actions that can be taken — supported by BI — to improve the position of the company. In BI-speak, that means insights must be actionable

1.4 THE BI VALUE PROPOSITION

BI links information with action inside an organization. But because of the confusion over defining BI, it's not always clear where the value of a BI solution lies. What exactly do businesses get from a BI implementation? If you're thinking about BI, you're naturally wondering "What's in it for me?" The answer is that when companies utilize BI, the BI value comes from promoting good decision-making habits. Encompassing BI is a rational approach to a continuous improvement loop:

1. Gathering data
2. Making decisions and taking action based on that data
3. Measuring the results according to predetermined metrics for success
4. Feeding the lessons from one decision into the next

By using a continuous cycle of evidence-based actions, organizations adopt a rational approach to their decision-making process — and BI can support that cycle. Figure 1-1 shows how this continuous loop can work. Through business intelligence concepts and tools, companies glean meaningful insights from their operational data.

If the insights fit the four criteria of BI (timely, accurate, high-value, and actionable) the company can apply them to its regular decision-making process. Those decisions, now informed with BI insights, lead to actions — and, if all goes well, improved operational results and so the cycle begins anew; the first round of results becomes part of the historical data record, and the related BI insights are refined even further. The process of using data to make better decisions can involve just about any piece of an organization.

If there are lessons to be learned from operational data, be it customer behavior, financial information, or another category, BI can play a part. By using BI practices to transform raw data into meaningful conclusions, a team makes better decisions. The actions taken as a result of those decisions produce a new round of results — which can be fed back into the system as new empirical evidence to draw the next round of conclusions. BI can improve any decision by supplying it with timely, accurate, valuable, and actionable insights.

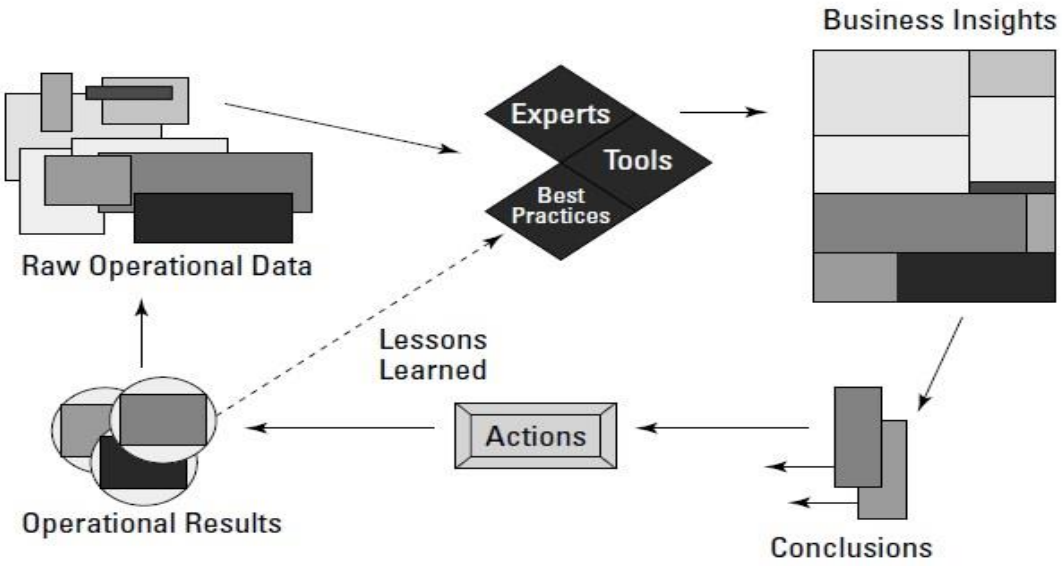


Fig.1.1 New data and results constantly feed the decision results cycle.

1.5 A BRIEF HISTORY OF BI

Business intelligence is an approach to solving business problems. It's a framework for managing tactical and strategic operations performance. BI is only possible because of advances in a number of adjunct technologies, such as computing power, data storage, computational analytics, reporting, and even networking technologies. But its origins are definitely more humble. In this section we'll take a look at how BI evolved to where it is today.

Data collection from stone tablets to databases

From the beginning of history, organizations have always had a need to collect and store data. Several thousand years ago, there were armies and imperial bureaucracies, working out ways to collect taxes, feed people, wage wars, and so on. The first recorded use of written language was data storage: Sumerian stone tablets that tracked shipments of wheat through the local granary

Record-keeping really came into its own as better forms of paper were invented. It allowed for more information to be stored and accessed in a smaller space. From silicon in stone to silicon in microchips, that challenge continues to this day: storing more and more information in smaller

and smaller space. The modern organization makes use of computer power for its data storage.

The growth of computing power and data storage

The first computers were tabulating machines, designed and built to perform one-off calculations. But scientists and inventors developed information storage capability almost neck and neck with the growth of computing power. After the 1940s, both technologies exploded. Mass storage began to take form when the properties of magnetic tape were used to store analog patterns of information. That turned to disk drives, a decades-old technology that is still in use today in a form that would be recognizable to its inventors, but on a scale that would blow their minds. To manage the growing mountains of stored data, programmers developed Database Management Systems (DBMSs) of growing power and complexity. Relational database technology came about as a response to the increasing information-storage demands. This was a revolutionary way to maintain data that dramatically sped up transaction time by splitting data elements into their component pieces and storing those pieces separately.

Transactional systems

As computing systems became more powerful and ubiquitous, businesses began taking advantage of them to manage their daily transactions.

Point-of-sale (POS) systems are the classic example of a transactional system. A POS system has one main purpose: to allow sales reps to quickly enter sales transactions, collect payment, and issue a receipt to the customer for that purchase. Handily enough, if the POS is some kind of computer it can be connected with accounting systems that gather and organize sales information for later use.

The emergence of decision support

In the late 1980s, companies began to recognize the potential value that the data represented. In response, they became motivated to build systems to extract the knowledge buried in their files. And so BI was born. Business intelligence came to encompass the wide range of technologies, protocols, and practices that is required to produce valuable business insights. What BI actually means to one company may be different from what it means to another because every company

represents a different situation, with different installed technology, and different needs. That's why business intelligence doesn't fit into a perfect definition you may have read on a vendor's website. BI means timely, accurate, high-value, and actionable insights and whatever it takes to produce those insights

1.6 BI'S SPLIT PERSONALITY: BUSINESS AND TECHNOLOGY

BI is built on the massive computing power available to today's enterprises. But it isn't just about bits and bytes. Business intelligence requires a company culture dedicated to the principles and practices that make high-quality, usable insights possible. Simply installing software and flipping a switch won't get a company to the promised land. The commitment to BI has to come from both the business and technology sides of a business:

- Business managers must engender a rational, measurement-based approach to setting strategy and running operations.
- IT must be prepared to support the BI culture to the extent that business managers are prepared to push it into all levels

BI: The people perspective

Business intelligence is about giving people new tools and perspectives; it's designed to let decision makers ponder what-if questions. That only works if those decision makers are not only able to use the BI tools but are also prepared to ask the right questions. That's where BI truly straddles the world between business and technology — it's both an art and a science. There is no set formula for determining the "right" reports and analytics for a particular company. No book explains every single possibility to consider in your analysis cycle. What's required is putting the right kind of people in positions where BI is to play a role. Or the BI attitude must be spread by the company's leadership. BI is about a commitment to a rational approach to making decisions — and that approach must be supported at all levels of the organization, by IT executives and business executives.

1.7 BENEFITS OF BUSINESS INTELLIGENCE

Would any organization benefit from a business intelligence solution? There is no automatic answer to that question, but nearly every company can see improvement from adding some rigor to the decision-making processes. The following list of questions you might ask about organization could indicate whether a BI approach makes sense:

- Can you view sales data in more than one view simultaneously? For example, if you wanted to see quarterly sales data by sales manager, product line, and customer type, how long would it take to produce the report?
- Is there data locked in transactional systems about your customers that you'd like to see but can't because the system just isn't designed to view the data the way you want?
- When your company makes strategic decisions, are you relying on hard data before you proceed or is it coin-toss time? Do you base your actions around evidence of the past and verifiable conclusions about the future? Do you consider statistical correlations between causes and effects?
- You know what items your customers buy the most, but do you know what items your customers buy in pairs?
- Do you know what your company does best? How do you know it? Is it a gut feeling or do you have metrics to back up your conclusions?

1.8 CHECK YOUR PROGRESS

1. Name any three BI tools available in the market
2. What is the purpose of BI?
3. What are the key advantages of using BI systems?
4. What do you understand by Business Intelligence?
5. What are the primary objectives of Business Intelligence?

Answers to Check your progress

1. i. Oracle Business Intelligence Enterprise Edition (OBIEE)
   ii. IBM Cognos Analytics
   iii. Micro Strategy
2. BI provides quick and simple methods to visualize company metrics, generate reports, and analyze data.
3. It helps boost productivity and makes it possible to create a business report with just a single click.
   ii. It also helps increase the visibility of the data analysis and possibly identify those areas that demand attention.
   iii. As per the organization's goals, the BI system sets the accountability.
4. The term Business Intelligence refers to a collective meaning, including technologies, tools, applications, practices for the data collection, and providing those data to the users, especially to help in running the business or a part of it.
5. i. Business Intelligence is leverage to make the following enterprise-level decisions.
   ii. Business Intelligence helps in identifying the wrong tracks and approaches of a business.
   iii. Business Intelligence can cluster the data for analysis and then compile them to monitor corrective actions.

1.9 SUMMARY

Just as the eyes are the windows to the soul, business intelligence is a window to the dynamics of a business. It reveals the performance, operational efficiencies, and untapped opportunities. Business intelligence (BI) is a set of technologies and processes that allow people at all levels of an organization to access and analyze data. Without people to interpret the information and act on it, business intelligence achieves nothing. For this reason, business intelligence is less about technology than about culture, creativity, and whether people view data as a critical asset. Technology enables business intelligence and analytics, but sometimes, too great a focus on technology can sabotage business intelligence initiatives. It is the people who will most make your BI efforts a wild success or an utter failure

1.10 KEYWORDS

- Business Intelligence (BI) - BI is using data about yesterday and today to make better decisions about tomorrow.
- Decision Support Systems - systems that would support the decision-making process
- DBMS - A database management system (DBMS) is a computerized system that enables users to create and maintain a database.
- Point of Sale(POS) - is the time and place where a retail transaction is completed

1.11 SELF ASSESSMENT QUESTIONS

1. Write a note on the history of BI
2. Explain BI's four main characteristics
3. What are the steps to implement company BI analytics from the ground up?
4. Explain the purpose of BI
5. What are the primary responsibilities of a BI developer?

1.12 REFERENCES

1. Swain Scheps - Business Intelligence For Dummies-For Dummies (2008), Wiley Publishing, Inc
2. Carlo Vercellis - Business Intelligence_ Data Mining and Optimization for Decision Making (2009), Wiley Publishing Inc

UNIT -2: FITTING BI WITH OTHER TECHNOLOGY DISCIPLINES

Structure

2.0 Objectives

2.1 Best Friends for Life: BI and Data Warehousing

2.2 ERP and BI: Taking the Enterprise to Warp Speed

2.3 Customer's Always Right

2.4 BI-BUY! E-Commerce Takes BI Online

2.5 The Finance Function and BI

2.6 Check your progress

2.7 Summary

2.8 Keywords

2.9 Self Assessment Questions

2.10 References

2.0 OBJECTIVES

After studying this unit you will be able to:

- ✓ Explain BI and data warehousing
- ✓ Connect BI to the enterprise with ERP
- ✓ Analyse customer data with CRM systems
- ✓ Design BI to plan for the future

2.1 BEST FRIENDS FOR LIFE: BI AND DATA WAREHOUSING

The collision of data-warehousing technologies with BI practices was a Eureka! moment for companies:

- Executives needed better access to the company's day-to-day data so they could evaluate conditions more accurately and make better decisions.

- The IT department was developing protocols and systems to bring widely dispersed and variable databases under one roof in order to run companywide statistical analysis and basic reporting.

BI and data warehousing are inextricably linked. The product of the two technology areas is more beneficial to companies than the sum of their parts. While each discipline is important in its own right, together they enable businesses to go beyond organizing operational data. BI and data warehousing make a transcendent combination — a powerful competitive weapon that can actually guide the business in ways previously considered impossible

2.1.1. THE DATA WAREHOUSE: NO FORKLIFT REQUIRED

The whole purpose of a BI implementation is to turn operational data into meaningful knowledge. That means BI must be connected with an organization's data to be effective. With data spilling out the doors and windows of any enterprise, the challenge is to put all the necessary data in one place, in one common format. Data warehouses are the perfect architecture to meet that challenge head on. A data warehouse is a single logical (but not necessarily

physical) repository for a company's transactional or operational data. The data warehouse itself does not create data; it's not a transactional system. Every byte of data inside the data warehouse has its origins elsewhere in the company. So what data are we talking about then? Most enterprises produce data in a variety of different departments or domains; there might be transactional sales information coming directly in from a point-of-sale system (POS), customer data from a Customer Relationship Management System (CRM), and an endless variety of operational systems that help the run. The data dispersed throughout all these different applications is likely saved in a variety of formats, on a range of hardware — say, a dedicated storage network, a mainframe, a database server on the Web, or even on various desktops. It could be anywhere

Data warehouses are different from standard transaction-based data management systems. A data warehouse aggregates information about a single subject area — and management then uses that resource in one of two ways:

- to create focused reports on one aspect of the enterprise
- to query in order to gain insights on that subject

Both activities are read-only. That makes sense because typically no data is deleted from a data warehouse. Transactional systems, on the other hand, add, delete, and update the data they store.

A data warehouse is a collection of data from different systems, focusing on one subject area. But because the data originates from a multitude of sources, it's going to be in different formats. Part of a data warehouse implementation involves manipulation — or transformation — of the data prior to storage so that it resides in a single common format.

2.1.2. DATA WAREHOUSES RESOLVE DIFFERENCES

Related data might be stored on completely different applications, in different storage media. Data might be missing or completely corrupted. Warehousing data can be an enormous task — you have to do three things to all that data from disparate sources:

- Put the information in a single format.

- Check for systemic data errors.
- Translate the data into useful units of knowledge.

In addition, your company may have organizational and geographical boundaries that separate information and prevent it from being used in concert with other key insights. So data warehousing technology must not only aggregate data of all flavors, it must also work with software and protocols that transform that data into common formats so information from one data source can be logically related to information from other data sources. Figure 2-1 shows a simple system where three different systems that collect similar data feed information into a data warehouse, which offers a single view of reality

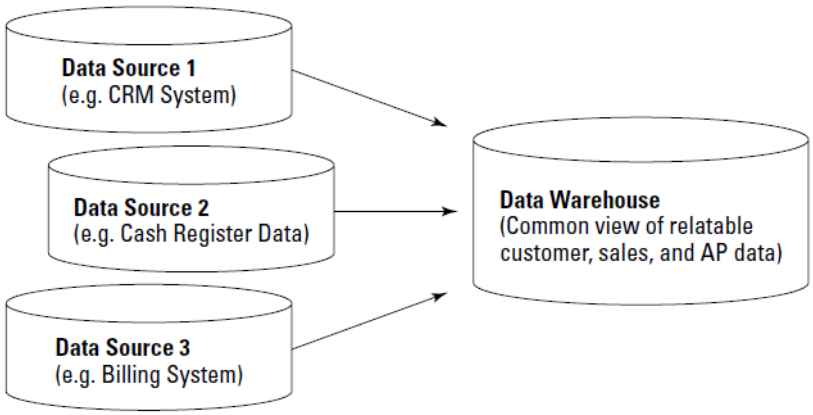


Fig. 2.1 A common data warehousing architecture, greatly simplified.

2.2 ERP AND BI: TAKING THE ENTERPRISE TO WARP SPEED

From the advent of business computers, up through the late 1980s and early 1990s, the systems that drove and supported the enterprise were designed — and run — mostly independently of each other. Even systems that would naturally fit together — say, finance and accounting or supply chain and inventory — were built and operated as separate entities. In the early 1990s, companies began to see computing power as a way to integrate vast and diverse transactional systems. As a result, Enterprise Resource Planning (ERP) systems were born.

2.2.1 From mainframe to client/server

The computing power that originated most legacy software was centreed on mainframe computers — gargantuan machines that existed in big rooms, or even took up entire floors of buildings. Their cousins, the minicomputers, had a similar role as centralized points for all of a company's processing. In those days, the IT staff consisted of as many electrical engineers as it did computer programmers — because the mainframes and minis were at least as much electromechanical machines as they were computers — and sometimes reprogramming meant rewiring. By the late 1980s, microcomputers (which we know now as personal computers) were getting small and powerful enough to place on workers' desktops, bringing about the rise of the client/server model. That led to pulling tasks away from the legacy systems and pushing them to workers' desktops. It also meant great advances in networking protocols and practices, linking people and data together for the first time. This had some fantastic advantages in terms of individual productivity, flexibility, and scalability. The term legacy application most commonly refers to older, mainframe oriented data-processing software. IT managers who cut their teeth on the newer client/server architecture came to view older mainframe applications as dinosaurs, an inherited burden that had to be upgraded or replaced ASAP. But today, legacy has become a catchall pejorative that refers to any last generation technology.

ERP came about as companies saw the need to integrate the core business computing systems to fit with their new client/server architecture. The battle cry arose: mainframe computing, with its magnetic tapes, punch cards, and outrageous electric bills, was dead. Why do data processing in one central place, they asked, when the work can be done on computers distributed throughout the business?

2.2.2 **The great migration**

The advantages of running ERP systems were clear. Workers could produce and consume data as never before. The old centralized data processing applications could now become more interactive, and customized to fit the company's needs. And yet the client/server architecture still provided centralized storage of information, which meant ERP applications could run and have everyone in a company looking at the same data. SAP was first on the scene, but others quickly followed. PeopleSoft, Baan, Oracle, Lawson, and JD Edwards were all pioneers in client/server

ERP.

2.2.3 **ERP leads to the foundations of BI**

The weaknesses of hard-coding reports was apparent — as were the problems associated with trying to use data from live transactional and operational systems in queries and reports. These challenges led ERP companies to begin incorporating some basic data-warehousing approaches into the new unified suites of applications — in conjunction with some advanced reporting capabilities.

The foundation for BI was laid. Forward-thinking ERP vendors started adding powerful reporting and analytic packages to their application suites to add further value for their customers. SAP was a trendsetter, introducing SAP BW in 1997. BW stands for Business (Information) Warehouse, a set of applications that offered customers advanced reporting and trend-spotting capabilities. And because BW was melded with the rest of the SAP application suite, clients could run the powerful BI tools against any set of data in the entire system — from manufacturing to sales. Business Warehouse was a sales success for SAP; other ERP vendors followed suit.

2.3 CUSTOMER'S ALWAYS RIGHT

CRM stands for Customer Relationship Management. It refers to software that handles all aspects of an organization's interactions with its customers. That may include the entire sales cycle, from winning new customers, to servicing and tracking existing customers, to providing post-sales services.

2.3.1 CRM joins ERP

CRM applications can touch lots of pieces of a business. Of course, the sales force itself relies heavily on robust CRM applications that can track leads, perform customer analysis, handle transactions and so forth. But beyond the sales force, CRM's tentacles reach into product management, inventory and procurement, accounting and finance, and others. Imagine using all that data about customer relationships to plan the use of enterprise resources. Some ERP vendors did. So, in the late 1990s, they began including CRM applications in their enterprise suites.

PeopleSoft made its CRM play through acquisition when they acquired a CRM company called Vantive in 1999. Vantive was a pure-play CRM vendor; that's the only class of software they sold. Other ERP vendors, such as Oracle, built their application in-house

Core CRM

Early CRM was always transactional in its approach, rather than analytical. The idea was to use technology to automate and facilitate the sales cycle as much as possible. So early incarnations of CRM would include features that tracked leads, scheduled sales operations, and recorded purchase history (along with other operational functions). But as CRM evolved, companies began to put greater demands on it to do more. Rather than just keep track of yesterday, CRM customers wanted the software to participate in the process — and help predict what customers were going to do. Companies began to see the potential for expanding CRM's role as they looked at all their contact points with customers. Call centres were foremost among these contact points; there hundreds of customer-service representatives would work phone banks while sitting in front of the custom-built applications they used to perform order entry or trouble-ticket functions on behalf of customers.

Customer decisions

In the late 1990s, data-crunching capabilities were on the rise in other parts of the company, it was only natural that CRM systems would become more involved in decision-support processes. E-commerce was also exploding in the late 1990s; the competitive fever required companies to wring every dollar they could from their online markets. That's when companies such as e.piphany came into being by merging traditional core CRM functions with BI-like analytical features and reporting capabilities.

Campaign management and more

Marketing was also transformed. Campaign management companies such as e.piphany guided their clients' marketing practices to make them more customer-centric. Yes, of course that's a silly tautological buzzword — customer-centric marketing. Products from this new generation of CRM companies allowed analysis and integration of customer data in a way that companies had never done before. That gave companies a new capability to create precision-guided campaigns,

to tweak their sales cycles to fit perfectly with the kinds of customers they attracted. Companies could suddenly measure the effectiveness of their sales force in ways never before imagined, and it spawned entire new marketing practices like customer loyalty management, churn management and customer reacquisition processes to help bring straying former customers back into the fold.

2.4 BI-BUY! E-COMMERCE TAKES BI ONLINE

Like the early CRM software, the original e-commerce applications were simple. Their one-dimensional functionality was good for facilitating fairly rudimentary business transactions over the Internet — at that time, almost entirely business-to-consumer sales. But in the early 1990s, the general public was just getting used to the idea of the Internet's existence; buying goods online represented a great leap forward. When the Internet first began popping up in homes, CRM software displayed little more than the same content you might find in sales brochures. There was little if any interaction with the company sites; it was strictly billboard ware.

E-commerce's early days

A few visionary companies started to change all that by actually selling products and interacting with customers on their Web sites. One of the first was Cyberian Outpost, who sold computers and peripherals off their site.

As retailers stumbled through e-commerce's growing pains, they had several key tasks to accomplish: They had to develop the software itself (catalogs, shopping carts, credit card processing) plus develop data-capture and reporting systems to match the transaction processing. As was the case with CRM and ERP systems, these functions had to be built from scratch, hard-coded by a mix of Web and traditional developers. At first the back-end analysis systems added little intelligence to the e-commerce process itself, it was little more than ex post facto reporting and analysis

E-commerce gets smart

Amazon founder Jeffrey Bezos and his apparent fascination with analytics were largely responsible for pushing BI into the e-commerce realm — and driving his company to billion-

dollar heights in the process. Early on, Bezos operated a company that lived, ate, and breathed data. Managers pulled metrics and reporting data from every conceivable piece of the Amazon operation. It started with a motivation to make the fulfillment and inventory systems as lean and efficient as possible — something Bezos knew could only be achieved by using hard data to drive management decisions. Then Bezos applied the same analysis-heavy mindset to the storefront too, where he tracked every move customers made — and had the Web application respond to those moves. The results of Amazon's BI culture manifest themselves today in the company's marketplace hegemony as well as in the user features you encounter when you shop online. When you log in, you see the last products you looked at, along with products judged to fit your tastes (according to your browsing and buying history). When you add books to your shopping cart, the system performs real-time analysis to recommend other books to you. BI actually helps shape customer behavior in real time.

Real-time business intelligence

Amazon's ability to influence customers is only possible because it collects a mountain of customer data. As you shop, sophisticated analytics are running in the background — comparing your habits and online activities with those of millions of customers who came before you. This customer-facing system can then react to you in real time (the industry calls it shopping time) and present you with options that make you more likely to spend more, return again and again, and be happy with the experience. That instantaneous reaction — BI capabilities shaping a Web site's behavior on the fly — represents a level of complexity and utility that isn't commonly seen in the other technology disciplines. ERP, CRM, and planning systems are most useful for looking at past data and doing one-time analysis to guide decision-making. But e-commerce brought BI into the present tense.

2.5 THE FINANCE FUNCTION AND BI

One more area of software functionality has been touched by business intelligence: financial reporting and analysis. The finance departments of companies of all shapes and sizes go through the process of assembling budgets, corporate planning initiatives, and performance forecasts. BI can help out there in unprecedented ways. The budgeting and planning process for organizations has always been intensely manual — low-level analysts and staff members would crunch

numbers and create individual spreadsheets that then had to be merged and summarized before becoming part of the next level. Team budgets would roll up into departmental budgets, from there up into divisional budgets, and eventually up into the overall corporate budget. This process left very little room for analysis. Any changes to the planning process would have to be cascaded up and down throughout the company to really understand its full effects — a process that's simply not feasible in most companies. As with ERP, CRM, and e-commerce, business intelligence found fertile soil in the global finance functions. CFO's were desperate to move beyond the pencil-and-paper processes and Excel spreadsheets that had dominated the area for so long. Business intelligence technology allows planners to perform what-if analyses, run budgets through predictive and profitability analyses, and create scorecards and dashboards to aid in corporate performance- management practices. BI not only speeds up these processes, it also gives the finance department far more confidence in the numbers themselves.

2.6 CHECK YOUR PROGRESS

1. Define CRMS
2. Who is CRMS designed for?
3. What is a Data warehouse?
4. What is an enterprise resource planning (ERP)
5. Write the difference between mainframe computing and client/server computing
6. What is Y2K bug?

Answers to Check your progress

1. 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. A CRM system helps companies stay connected to customers, streamline processes, and improve profitability.
2. A CRM system gives everyone — from sales, customer service, business development, recruiting, marketing, or any other line of business — a better way to manage the external interactions and relationships that drive success. A CRM tool lets you store customer and prospect contact information, identify sales opportunities, record service issues, and

manage marketing campaigns, all in one central location — and make information about every customer interaction available to anyone at your company who might need it.

3. A **data warehouse** is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics.
4. ERP is an acronym that stands for **enterprise resource planning** (ERP). It's a business process management software that manages and integrates a company's financials, supply chain, operations, commerce, reporting, manufacturing, and human resource activities.
5. In a **client/server** network, a number of network clients or workstations request resources or services from the network. One or more network servers manage and provide these resources or services. The clients are computers that depend on the server for data and software. In a **mainframe system**, all the processing is done by a single, very powerful computer. Individual terminals are used to access the mainframe computer but don't run any applications themselves.
6. The legacy systems (old data-processing software that was still running on corporate mainframes and minicomputers) weren't designed to understand that the year after 1999 was 2000, not 1900. That's because these systems were only capable of tracking years in terms of two digits, so 1971 was represented as 71, 1999 as 99 — and when 99 turned to 00

2.7 SUMMARY

In this unit, we looked at the most prominent technologies commonly associated with BI. Because these other classes of software are common in so many companies, we understood how business intelligence is related to them. The associations — some casual, some arrangements of convenience — are of interest and importance to anyone considering a BI implementation. These disciplines exist outside the immediate realm of BI — but in each case, business intelligence concepts and approaches have had a dramatic effect and helped the underlying technology round itself out into full form. The relationship between BI and each of these technologies is a two-way street. Each technology area has benefited from the BI process, and BI has grown and in response to the widespread adoption and evolution of these technologies — especially data warehousing, Customer Relations

2.8 KEYWORDS

- a Customer Relationship Management System (CRM) - Customer relationship management (CRM) is a technology for managing all your company's relationships and interactions with customers and potential customers
- Data warehouse – A **data warehouse** is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics.
- ERP -- is an acronym that stands for **enterprise resource planning** (ERP). It's a business process management software that manages and integrates a company's financials, supply chain, operations, commerce, reporting, manufacturing, and human resource activities.
- Mainframe computers – are high-performance computers with large amounts of memory and processors
- Client/server computing - Client–server model is a distributed application structure that partitions tasks or workloads between the providers of a resource or service, called servers, and service requesters, called clients.

2.9 SELF ASSESSMENT QUESTIONS

1. Explain how data warehousing technologies influences BI practices?
2. Explain how ERP systems influence business decisions
3. Write a note on CRMS
4. Explain the role of BI in E-Commerce.
5. Explain how financial reporting and analysis is influenced by BI?

2.10 REFERENCES

1. Swain Scheps - Business Intelligence For Dummies-For Dummies (2008), Wiley Publishing, Inc
2. Carlo Vercellis - Business Intelligence_ Data Mining and Optimization for Decision Making (2009), Wiley Publishing Inc

UNIT -3: MEETING THE BI CHALLENGE

Structure

3.0 Objectives

3.1 Challenges of BI

3.2 The BI Spectrum

3.3 First Glance at Best (and Worst) Practices

3.4 Check your progress

3.5 Summary

3.6 Keywords

3.7 Self Assessment Questions

3.8 References

3.0 OBJECTIVES

After studying this unit, you will be able to:

- ✓ Identifying your BI needs
- ✓ Looking at the BI continuums
- ✓ Use best and worst practices
- ✓ Avoid vendor hype

3.1 CHALLENGES OF BI

A typical business intelligence solution has many moving parts — including an array of software and hardware that must work together in concert. With BI's heavy reliance on IT, it's no surprise that many companies squander the bulk of their planning process on technology. This tech focus is a misconception that will get you into trouble

Hardware and software selection is not actually the most difficult part of a BI implementation. A company that focuses on choosing server vendors, designing architectures, and the like is missing the real problem: identifying accurately what the business actually needs

The real challenge of BI is coming up with a workable answer to the question "What's our problem?" For business intelligence solutions to go to work for you delivering important business insights, your organization must determine the purpose of the implementation. You do need to ask some preliminary questions to identify the kinds of business questions you want BI to answer.

Some of these questions are obvious, some require in-depth research, some might bring a range of responses depending on whom you ask within your company, and some may have no single answer — just an educated guess will have to do. Here's a sampling of those questions:

- What data is currently available to be measured and analyzed?
- What measurements can we monitor that indicate success in one or more areas of the business?

- When do we need the answers? Do we need them all at once? How will we take action if certain insights demand it?

- How prepared is the company culture to effect change once we get our answers?

And there are more of these — many more — to answer before you start thinking about technology.

What can go wrong

If you move ahead with a BI solution but fail to properly or completely identify the problem you want it to solve, you don't get much benefit from BI. In fact, a BI implementation can be a disruptive force to your organization in many ways. Consider the following areas of impact.

Cost

If you had an unlimited budget you could try everything and keep only what works. But BI can be an expensive proposition — and it doesn't stop with new software licenses or consultant bills. There's always a need for extra hardware and infrastructure. Then you have to factor in maintenance costs, training costs, and all sorts of hidden expenses (such as a temporary decline in worker productivity as they adapt to a new paradigm, training classes for users and project team members, opportunity costs, and so on).

Time

Working through a fruitless BI implementation means you aren't doing something else that could be useful to the company. Waiting on a BI implementation to complete can cause delays in tackling strategic problems. The resources dedicated to the task could be repurposed to other tasks with immediate payoffs like getting a product ready for market or improving your internal processes

Credibility

Not all high-level executives are convinced that BI is more than just an empty buzzword. If one of the doubters sees your project fail, you'll just confirm their pre-conceived notions. Worst of

all, it makes it that much harder to gain support the next time around. A good BI implementation depends on people at all levels making a commitment to the process. Skepticism is a self-fulfilling prophesy, so you should take steps to ensure you get it right the first time.

Bad advice

The worst consequence of all is that you may think you're getting what you need from your BI solution, only to find out the analysis has produced recommendations that lead in the direction precisely opposite to success. Most BI wounds are self-inflicted. When BI projects go wrong, it's not because software breaks or consultant teams lie. It's because companies either don't do adequate preparation or don't think through their goals and capabilities well enough (or both).

3.2 THE BI SPECTRUM

If you consider the Big Four characteristics of business insights produced by BI projects (timely, accurate, high-value, and actionable) you'll note that they are all abstract notions. Notice that nowhere in there does the definition include insights that save you money, or help long-range planning. The four characteristics of BI insights are intentionally left a safe distance from your company's particular success criteria

That's because BI projects come in all shapes and sizes. The CFO might be directing a massive BI project to look at the global finance functions and how they affect the entire enterprise. Or a regional sales manager could be the "customer" of your company's business intelligence system, and have a much narrower outlook. That person is not interested in enterprise-level financial insights — and isn't particularly concerned with building knowledge about the entire company. For the regional sales chief, a successful BI implementation offers insights on the sales process alone. Any insights beyond that aren't actionable at the regional-sales level. What makes BI insights good is not the breadth of their application in a company, it's whether they're answering the questions they were designed to answer — and producing information that is timely, accurate, high-value, and actionable

Good BI insights can look different depending on the shape and scope of your implementation. It's for you to decide what scope makes sense. Companies are different, as are their internal processes and organizational structures. Part of the challenge of implementing a sensible

business intelligence solution is managing the scope so it applies exactly where it's needed. Sometimes that means finding the place where BI insights can have the biggest impact; sometimes it's about delivering a usable solution on time. Whatever it is, determining your project's scope — how far the implementation will reach, what areas it will touch — is paramount to its success. In the next sections, we talk about some of the dimensions you should consider when you're specifying the scope for your BI project.

Enterprise versus departmental BI

For most companies, the scale of the BI implementation is predetermined by the team that originated the idea, the size of the project's budget, and the level of buy-in from the company's leadership:

- For smaller implementations, the focus rests on a single department.
- Larger projects are an entire enterprise solution that affects the entire company

As a best practice, your company's first BI implementation should be at the departmental level. Keeping it narrowly focused affords you the chance to learn and make mistakes, to see what does and doesn't work. Then, when you're ready for the big time, you can widen the scope for the second generation of the BI project. If you must implement an enterprise-wide BI project, at least make sure you keep the scope focused on a single function or geographic region. Trying to do it all right off the bat (multifunctional, enterprise-scale) is a challenge that has stymied even the best BI gurus.

Characteristics of enterprise BI

Enterprise BI projects are, naturally, broad in scope. They affect multiple functional areas of a business; typically they involve taking a unified view of the entire company (or of an entire self-contained business unit). Often heavy on analytics and forecasting, these projects produce insights that affect long term decisions. Enterprise-wide business intelligence is operated and sponsored by "C-level" people whose titles start with Chief — primarily CEOs and CFOs. That makes sense because they're the only ones who have the juice to force all business units to cooperate. Why wouldn't they cooperate? Read on. Some CEOs run Darwinian businesses where

their departments or business units compete against each other for survival. In this environment, bonuses and stock options are doled out only to the most successful pieces of the business. That's a problem because BI requires cooperation. BI projects involve — first and foremost — sharing potentially sensitive data between teams. It also requires buy-in of technology and business resources from all sides. While the surface may appear calm before work begins, internecine conflicts tend to bubble to the surface during a BI implementation. In big projects — say, installing enterprise-level BI systems — sooner or later everyone has to put their cards on the table. At that point, usually you find that someone is bluffing.

Characteristics of departmental BI

Unlike enterprise-level business intelligence projects, departmental BI exists at the operational level; where the rubber meets the road in an organization. Projects at this level are built to produce insights that fine-tune daily processes and improve short-term decisions. With this type of project, you're looking for insights in one narrowly defined piece of your business, such as marketing, sales, or finance.

Departmental itself is an abstract term that in essence means "anything that's not enterprise-wide." There are all levels of departmental BI, from single-team to multi-department. The word department itself could describe a traditional business team, or business functions that share a common geographic region or mission. There is no universally accepted definition, so when you discuss implementing a "departmental" BI solution with others, make sure you all agree on what the term means.

Mind the (operational) gap

Departmental and enterprise BI solutions can be a challenge mainly because traditionally there's been no smooth continuum of solutions that can handle the whole range — small-scale departmental BI, medium-sized business unit BI, and the largest enterprise-wide BI. Vendors normally produced software that covered only one specific level of an organization — perhaps even a single specialty, such as accounting, travel services, human resources, or e-commerce

This operational gap — which separates BI into different scales of activity in one company — is one of the main challenges the industry faces today. How do you build a bridge between BI user

groups at different levels of the company? The essence of the problem is that when strategic insights occur at the top of the company, there's not always an easy method of translating those changes into the company's day-to-day operations.

That task is so important, it's earned its own buzzword: operationalizing business intelligence. Enterprise BI tools may produce valuable strategic insights or reveal a long-term destination for the company. But that doesn't mean the way forward is always clear for individual workers or departments. To operationalize BI is to create an action plan of achievable day-to-day improvements that moves the company toward its strategic goals. If your aim is to change the way workers do specific departmental tasks, but the scope of your BI is enterprise-wide, you may be asking for trouble. You have to find a way to translate those strategic insights into a tactical format. A good BI solution will account for this operational gap and take steps to accommodate the inevitable "translation" problems between BI insights and their target audience

Strategic versus tactical business intelligence

In addition to the operational scale of BI that differentiates between departmental and enterprise-wide solutions, there's also a range of differences in the scope of the decisions that BI insights indicate. Strategic BI deals with the big picture and the long view; tactical BI handles today's immediate decision making process and other details of getting the job done, day in and day out

Strategic decisions

This kind of BI involves gaining business insights for long-term corporate decisions, whose arc covers the broad direction of the company or business unit as a whole. Imagine a question such as, "What is the optimal product mix for the New England sales region in Q1 2009?" The insights that can be gained from questions like that may not affect the day-to-day operation of the company: the receptionist will still answer the phone, the sales team will continue to make calls on customers, the factory will continue to run the same way. At some point, however, such questions may prompt a strategic shift that leads to a long-term change in how the company operates.

Tactical and operational decisions

These decisions guide how a company operates on a day-to-day basis; they're usually peculiar to a given department or business sub-unit — for example, a decision on the size of the discount that certain customer affinities might receive, or setting the price a business pays for a certain commodity. These are choices that affect day-to-day operations, and as such may change from day to day. It's difficult to fully automate these decisions; reporting and other computational tools can play a support role, but at some point in the chain, carbon-based neurons have to wrestle with them

If transactions are like traffic, then operational decisions are like high-volume stoplights at your company. They're the yes/no gates that might approve or decline credit to a customer, kick off a coupon-awarding routine (or fraud alert), or perhaps initiate a late charge in a billing system. The BI processes that touch these decisions — tactical BI — involve insights that might affect a business on a day-to-day or week-to-week basis. The insights might lead to immediate changes or adjustments to previous decisions. Imagine a BI application that could give you the answer to a specific question such as, "Who are the bottom three sales representatives in the Pacific sales region, year-to-date?" The answer to that might prompt an adjustment in pricing, product mix, or approach to the sales process. The decisions taken from tactical BI usually don't involve the entire business unit; they just touch a department or single functional area. There are two scales to think about: Enterprise versus Departmental BI, and Strategic versus Tactical BI. While they often go together, there is no automatic correlation between enterprise/strategic and departmental/tactical. So be careful that the solution you design answers the most important questions for your organization

Power versus usability in BI tools

This is really a question of your users. Are you likely to have a large population of limited-ability users? Or is your BI implementation going to be operated exclusively by a small group of experts? BI used to be the exclusive domain of highly experienced users — IT experts, sent to expensive training classes, would then operate on behalf of the business community to draw information from the BI application. This was the model that BI vendors counted upon, and they focused all their attention on making more powerful analytic and forecasting tools. But as BI has

spread throughout the organization, non-expert business users — such as CEOs — need to be able to harness the power of the applications so they can tap into the insights. As a result, more vendors are at least paying lip service to usability, creating self-service applications. Nevertheless there is a tension between tools that focus on ease of use and those with greater power, complexity, and flexibility. Suppose you anticipate your users to ask questions like, "What are the profit margins in the first quarter of this year for the 10 least profitable products of the last five years?" Fair enough — a good SQL query writer could build that query to work against a standard database, but not all BI tools can handle a question of that complexity

A good way to gauge your needs along the two primary BI continuums (enterprise/strategic and operational/tactical) is to anticipate what kinds of questions your users will want to ask — and then compare the complexity of those questions with the likely ability of those same users to operate the BI tools. Even the easiest BI applications are complex. Sure, the vendors have made it easy to drag and drop dimensions and metrics into tables, but many users will be stumped by such a simple move. Some folks are going to be so set in their ways; they won't be willing to run even the most basic of reports. Better to sniff out that problem ahead of time rather than after you're deep into your implementation. If your BI system is unusable to the people you're designing it for, it puts the entire project in jeopardy

Reporting versus predictive analytics

BI gives companies the ability to peer into the past, slicing and dicing historical data in revealing ways. But applications are available that dig through yesterday's information to form predictions about what the future will be like. Coupled with data warehouses, BI tools give users access to snapshots of information from the organization's operational and transaction-based systems. The capability to do complex drill-downs into this historical record is where the system provides its unique value for reporting purposes. On the other hand, some BI vendors focus on predictive-analysis tools. This software uses advanced statistical techniques to create tomorrow's forecasts based on yesterday's data. Not a crystal ball, but the next best thing.

3.3 First Glance at Best (and Worst) Practices

Why BI is as much an art as a science

The technology is mature enough that best practices are available — and (as we mentioned) BI architecture and software are only moderately complicated. So why do so many implementations fail?

The answer is that there is an art form to juggling all the competing priorities and coming up with the right answer. Beyond finding the right place on the continuum, there are political considerations, budget considerations, and a host of other concerns. There are process questions that must be answered that nobody ever thinks about: How do you deal with historical data? How do you handle different levels of user expertise? How do you validate requirements? How do you prioritize them? When you're about to miss a deadline, how can you tide over the key players who were promised a solution? How do you ensure that something else doesn't go wrong because of your BI project?

It would be nice to have a BI project manager always handy — someone who knows the answers to these questions.

Avoiding all-too-common BI traps

Statistics support that notion that more BI implementations fail than succeed. That's not meant to scare you off; it's a reality check. In some cases it just wasn't meant to be. But more often than not it was one of the well-known quicksand traps below that brought the expedition to an untimely end.

Thinking technology alone can cover it

Sheer wishful thinking. Even if you're prepared to spend millions on software licenses, it's impossible to do a plug-and-play BI implementation without getting buy-in from the data consumers — and, to some extent, the data producers. They may be silent and invisible, but you can't spell Business Intelligence without the letters T-E-A-M. Even modest BI rollouts require collaboration among many disciplines. "If you build it, they will come," is not a valid mantra.

Thinking people alone can handle it

Not to depend entirely on the talent of your company's IT team, you can always go for some components off the shelf (so to speak) instead of taking the total do-it-yourself route. The vendors may be full of themselves, but there is a kernel of truth in their hype. Business intelligence solutions aren't vastly more complicated than other software rollouts — but they are complex enough that building the entire package from scratch usually isn't worth your while.

Loving your data just the way it is

Data is hard to love if it's useless. One common BI failure is not paying enough attention to the quality of the data that's feeding the system. Bringing the data under one roof in the data warehouse is not (by itself) enough to ensure success. It has to be examined for a host of common problems — from format incompatibility to missing data — and then processed and integrated to make it useful. Remember that data warehouses are typically one-way streets. Transforming the data once it gets to the warehouse doesn't put your original data at risk. You shouldn't be afraid to twist, tweak, edit, or delete the information that's flowing into the data warehouse.

Confusing causality and coincidence

Although it's true that analyses and reports are only as good as the data in the data warehouse, it's also true that the untrained shouldn't draw conclusions. It's easy to misread reports, and see things that aren't actually there. Statistics and metrics can be made to tell a story that's more fiction than non-fiction.

One more continuum: hope versus hype

That BI vendors often oversell their products shouldn't come as a surprise to you; in fact, it makes BI just like every other technology. The different acronyms make it difficult to keep track of what's what in the industry — and that, combined with other factors, means the vendors may try to talk you in circles.

No matter which part of their pitch you listen to, there's an awful lot of noise. And to add more confusion to the mix, the industry can change month by month as companies acquire each other

and go out of business. Just keep that in mind at your next product evaluation. Many of the total BI packages are really just hodgepodge collections of products built by different companies and acquired one-at-a-time over the years by the company that's pitching its wares to you right now. That doesn't mean the packages won't work — but as always, caveat emptor.

Remember, BI vendors are there to make you think your business will transform itself overnight if only you would buy their products. Just remember that for every survivor in the BI marketplace, there is a company that fell by the wayside — and they said exactly the same thing to their clients right up until the paychecks started bouncing. In fact, it's almost guaranteed the vendor you're talking to right this moment will claim to have a perfect solution for you, no matter where you are in the BI life cycle. If you're in the midst of a failed implementation, they'll tell you their product is the missing piece of the puzzle that will stabilize the system and make it start spitting out amazing organizational insights. If you know very little about BI and are just starting down that road, the vendor will take the role of the wise doctor — and the prescription they give you will just happen to include a healthy dose of their products and services. As with medical doctors, when in doubt, get a second opinion, and a third, and a fourth.

3.4 CHECK YOUR PROGRESS

1. Define KPI.
2. Write the difference between an enterprise BI and a department BI
3. What is predictive analytics?
4. What is the difference between the tactical decision and operational decision
5. What is a Strategic decision?

Answers to Check your progress

1. KPI stands for **key performance indicator**, a quantifiable measure of performance over time for a specific objective. KPIs provide targets for teams to shoot for, milestones to gauge progress, and insights that help people across the organization make better decisions.
2. For smaller implementations, the focus rests on a single department. Larger projects are an entire enterprise solution that affects the entire company.

3. Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning that analyze current and historical facts to make predictions about future or otherwise unknown events.
4. Tactical decisions are **decisions and plans that concern the more detailed implementation of the directors' general strategy and** Operational decisions are **specific business decisions made every day within every business**
5. Strategic decisions are the **decisions that are concerned with whole environment in which the firm operates**

3.5 SUMMARY

Just about everybody agrees that having timely, accurate, high-value, and actionable insights available before making critical business decisions would be extremely helpful. marketing executive want a report that clearly shows the optimum mix of products to send to the marketplace as a package? V.P. of Sales like reliable figures on which territories and accounts are the most profitable? But wanting something doesn't make it a reality. We saw in this unit that While good business intelligence is within virtually any company's grasp, many obstacles can get in the way. There are technology landmines, project hurdles, and even political challenges. We understood as you prepare for a BI implementation, you need to be ready to answer some tough questions about how your company operates. You have to make some decisions about what exactly you want out of your BI solution. And you should brush up on your diplomatic skills to lay the groundwork for unifying your team behind a common goal. In this unit we discussed identifying all the things that can go wrong with BI. If you know in advance which problems lie in wait for you, they're much easier to solve (or avoid altogether). If you charge ahead without considering what can go wrong, you'll join the ranks of companies whose BI implementations either never got off the ground or foundered once they were launched.

3.6 KEYWORDS

- KPI- KPI stands for **key performance indicator**, a quantifiable measure of performance over time for a specific objective.
- Enterprise BI – deployment of BI throughout a large corporation.

- Departmental BI – BI restricted to a department in an organization
- Analytics - Analytics is the systematic computational analysis of data or statistics
- Strategic decision - about evaluating the pros and cons of a situation and developing a step-wise approach to realize your goals
- Operational decision - specific business decisions made every day within every business

3.7 SELF ASSESSMENT QUESTIONS

1. Explain how BI implementation can be a disruptive force to an organization
2. Compare and contrast the Enterprise versus departmental BI
3. Explain Strategic and tactical business intelligence.
4. What are BI Best Practices?

3.8 REFERENCES

1. Swain Scheps - Business Intelligence For Dummies-For Dummies (2008), Wiley Publishing, Inc
2. Carlo Vercellis - Business Intelligence_ Data Mining and Optimization for Decision Making (2009), Wiley Publishing Inc.

UNIT -4: DATA WAREHOUSING

Structure

4.0 Objectives

4.1 Definition of data warehouse

4.2 Data warehouse architecture

4.3 Evolution of information systems cubes and multidimensional analysis

4.4 Check your progress

4.5 Summary

4.6 Keywords

4.7 Self Assessment Questions

4.8 References

4.0 Objectives

After studying this unit, you will be able to

- ✓ Describe the features of data warehouses and data marts
- ✓ Examine the architecture of a data warehouse
- ✓ Perform on-line analytical processing operations and analyses

4.1 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-centre activities, customer assistance, execution of marketing campaigns;
- web-based systems, that gather sales transactions on e-commerce websites, 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 therefore 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 referred to 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.

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 A solution is then to direct complex queries for OLAP analyses to the data warehouse, physically separated from the 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.

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 warehouse 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.

4.1.1 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 framework. 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.

4.1.2 DATA QUALITY

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. However, 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.

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.

4.2 DATA WAREHOUSE ARCHITECTURE

The reference architecture of a data warehouse, shown in Figure 4.1, includes the following major functional components.

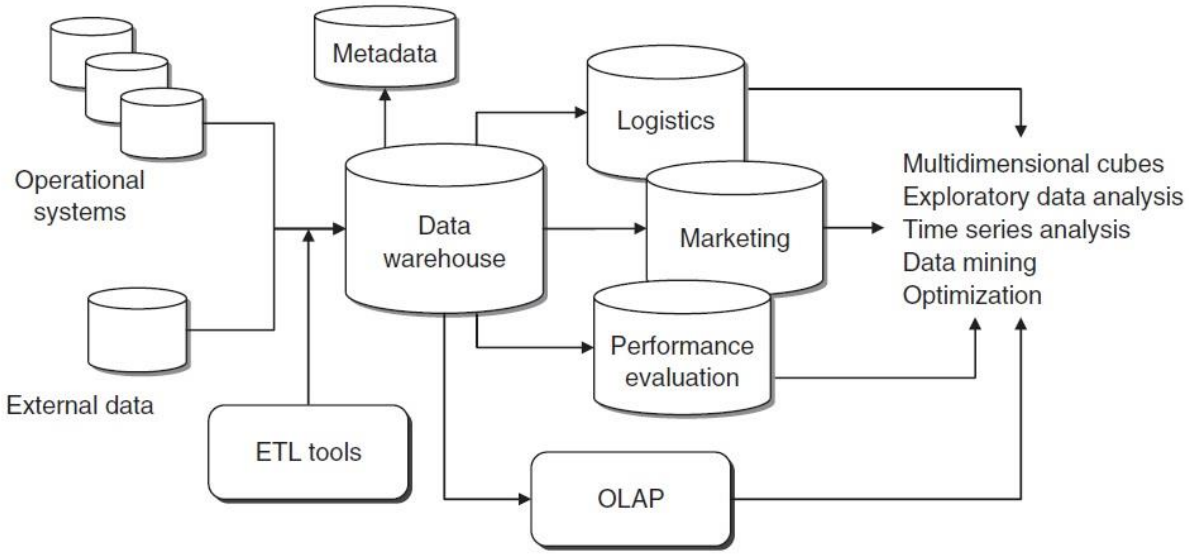


Fig.4.1. Architecture and functions of a data warehouse

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

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 therefore 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.

4.2.1 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 into the 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, which in 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 similarity. 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.

4.2.2 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 field 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.

4.3 EVOLUTION OF INFORMATION SYSTEMS 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 most cases 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 OLAP analyses.

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.

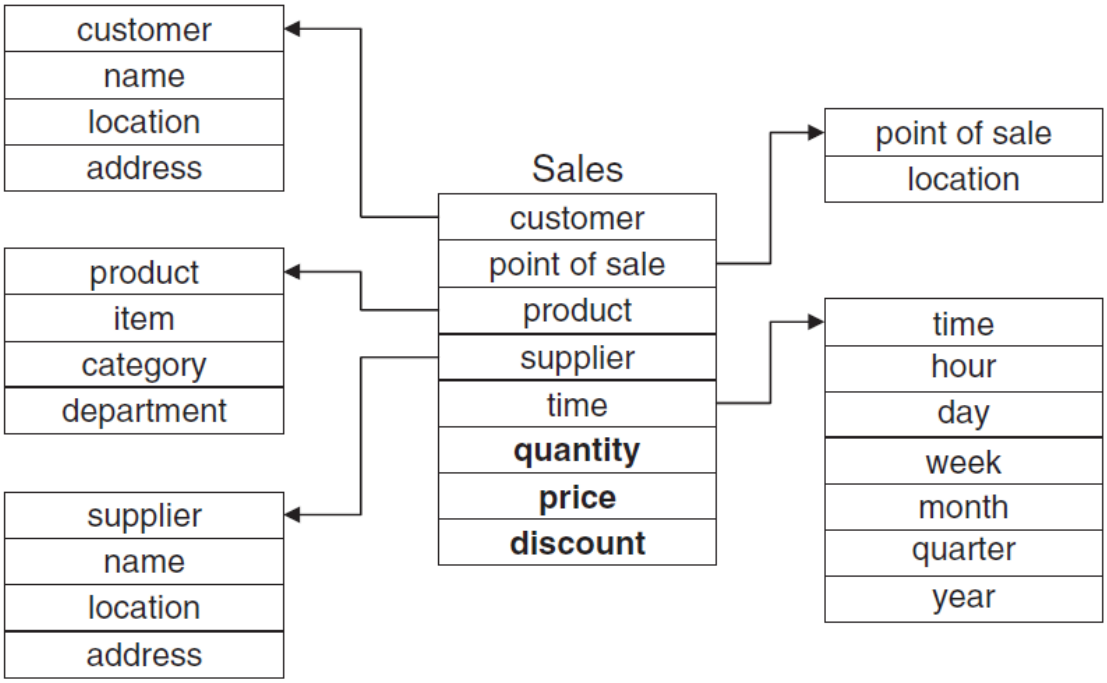


Fig. 4.2 Example of a star schema

Figure 4.2 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, 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. 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 4.3, is termed a galaxy schema.

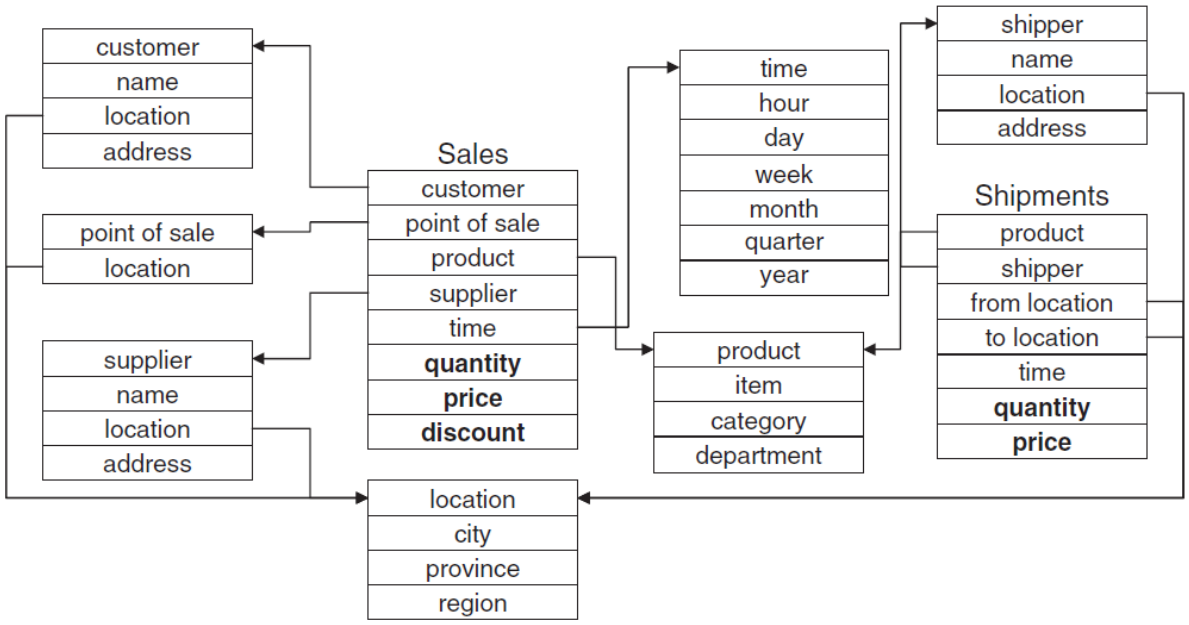


Fig 4.3 Example of a galaxy schema

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

4.3.1 HIERARCHIES OF CONCEPTS AND OLAP OPERATIONS

In many instances, OLAP analyses are based on hierarchies of concepts to consolidate the data and to create logical views along the dimensions of a data warehouse. A concept hierarchy defines a set of maps from a lower level of concepts to a higher level.

For example, the {location} dimension may originate a totally ordered hierarchy, developing along the {address, municipality, province, country} relationship. The temporal dimension, on the other hand, originates a partially ordered hierarchy.

Hierarchies of concepts are also used to perform several visualization operations dealing with data cubes in a data warehouse.

Roll-up. A roll-up operation, also termed drill-up, consists of an aggregation of data in the cube, which can be obtained alternatively in the following two ways.

- Proceeding upwards to a higher level along a single dimension defined over a concepts hierarchy. For example, for the {location} dimension it is possible to move upwards from the {city} level to the {province} level and to consolidate the measures of interest through a group-by conditioned sum over all records whereby the city belongs to the same province.

- Reducing by one dimension. For example, the removal of the {time} dimension leads to consolidated measures through the sum over all time periods existing in the data cube.

Roll-down

A roll-down operation, also referred to as drill-down, is the opposite operation to roll-up. It allows navigation through a data cube from aggregated and consolidated information to more detailed information. The effect is to reverse the result achieved through a roll-up operation. A drill-down operation can therefore be carried out in two ways.

- Shifting down to a lower level along a single dimension hierarchy. For example, in the case of the {location} dimension, it is possible to shift from the {province} level to the {city} level and to disaggregate the measures of interest over all records whereby the city belongs to the same province.

- Adding one dimension. For example, the introduction of the {time} dimension leads to disaggregate the measures of interest over all time periods existing in a data cube.

Slice and dice

Through the slice operation the value of an attribute is selected and fixed along one dimension. The dice operation obtains a cube in a subspace by selecting several dimensions simultaneously.

Pivot.

The pivot operation, also referred to as rotation, produces a rotation of the axes, swapping some dimensions to obtain a different view of a data cube.

4.4 CHECK YOUR PROGRESS

1. Which are the three main categories of data feeding into a data warehouse?
2. Define OLTP and OLAP
3. Define a data mart
4. What is Metadata?
5. What is a roll-up operation?
6. What is roll-down operation
7. What is Slice and dice.?
8. What is The pivot operation?

Answers to Check your progress

1. Internal, external, personal
2. **Online analytical processing** (OLAP) is a technology that organizes large business databases and supports complex analysis. OLTP or **Online Transaction Processing** is a type of data processing that consists of executing a number of transactions occurring concurrently—online banking, shopping, order entry, or sending text messages, for example
3. 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 busi ness intelligence analyses and executing decision support applications specific to the function itself. false
4. Data on data
5. A roll-up operation, also termed drill-up, consists of an aggregation of data in the cube
6. A roll-down operation, also referred to as drill-down, is the opposite operation to roll-up. It allows navigation through a data cube from aggregated and consolidated information to more detailed information.

7. Through the slice operation the value of an attribute is selected and fixed along one dimension. The dice operation obtains a cube in a subspace by selecting several dimensions simultaneously.
8. The pivot operation, also referred to as rotation, produces a rotation of the axes, swapping some dimensions to obtain a different view of a data cube.

4.5 SUMMARY

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 leant the features of data warehouses and data marts, understood the factors that led to their conception, and highlighting the major differences between them and operational systems, and discussed the requirements concerning data quality. Then we examined the architecture of a data warehouse, pointing out the role of ETL tools and metadata. The last part of the unit was on on-line analytical processing operations and analyses that can be performed by using multidimensional cubes and hierarchies of concepts.

4.6 KEYWORDS

- Internal data - The data stored in the operational systems
- External data – Data collected from external entities such as GIS
- Personal data – Worksheets or local databases of a user
- OLTP – Online Transaction Processing
- OLAP - Online analytical processing
- ETL – Extract Transform and Load

4.7 SELF ASSESSMENT QUESTIONS

1. Write reasons for implementing a data warehouse separately from the databases supporting OLTP applications in an enterprise.

2. Write Differences between OLTP and OLAP systems
3. Explain major factors that may affect data quality.
4. Explain the architecture and functions of a data warehouse
5. Explain different design approaches to a data warehouse design

4.8 REFERENCES

1. Swain Scheps - Business Intelligence For Dummies-For Dummies (2008), Wiley Publishing, Inc
2. Carlo Vercellis - Business Intelligence_ Data Mining and Optimization for Decision Making (2009), Wiley Publishing Inc

BLOCK 2 INTRODUCTION

In this block we will focus on the main characteristics shared by different mathematical models embedded into business intelligence systems. We will also develop a taxonomy of the most common classes of models, identifying for each of them the prevailing application domain and we 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. Unit 6 we understand OLAP basic operations such as drill down, drill up and drill through. In unit 7 we look at agile development. 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. agile for BI is more widely accepted, and advocating it as a best practice. In unit 8, we will describe guided analysis and data visualization.

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

Unit 5- Mathematical models for decision making and Data mining**:**

Structure of mathematical models, Development of a model, Classes of models, Definition of data mining, Representation of input data, Data mining process, Analysis methodologies

Unit 6- OLAP: Online Analytical Processing:

OLAP in Context**,** OLAP Application Functionality**,** Multidimensional Analysis**,** OLAP Architecture, What OLAP Can Really Do**,** Drill team: Working with Multidimensional Data, OLAP versus OLTP, Looking at Different OLAP Styles and Architecture

Unit 7- Agile Development:

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- : Advanced / Emerging BI Technologies:

Catching a Glimpse of Visualization**,** Steering the Way with Guided Analysis**,** Data Mining: Hype or Reality?**,** Other Trends in BI**.**

UNIT -5: MATHEMATICAL MODELS FOR DECISION MAKING AND DATA MINING

Structure

5.0 Objectives

5.1 Structure of mathematical models

5.2 Development of a model

5.3 Classes of models

5.4 Definition of data mining

5.5 Representation of input data

5.6 Data mining process

5.7 Analysis methodologies

5.8 Check your progress

5.9 Summary

5.10 Keywords

5.11 Self Assessment Questions

5.12 References

5.0 OBJECTIVES

After studying this unit, you will be able to

- ✓ Explain Fundamental features shared by mathematical models
- ✓ Identify Phases of mathematical model development
- ✓ Describe Different classes of models
- ✓ Examine Basics of data mining

5.1 STRUCTURE OF MATHEMATICAL MODELS

Mathematical models have been developed and used in many application domains, ranging from physics to architecture, from engineering to economics.

The models adopted in the various contexts differ substantially in terms of their mathematical structure. However, it is possible to identify a few fundamental features shared by most models.

Generally speaking, a model is a selective abstraction of a real system. In other words, a model is designed to analyse and understand from an abstract point of view the operating behaviour of a real system, regarding which it only includes those elements deemed relevant for the purpose of the investigation carried out. Figure 5.1 expresses in graphical terms the definition of a model.

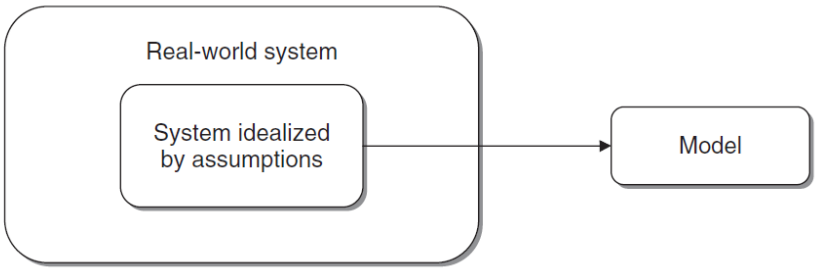


Fig. 5.1 *A model is a selective abstraction of reality*

Scientific and technological development has turned to mathematical models of various types for the abstract representation of real systems. As an example, consider *what-if analyses* that can be easily performed using a simple spreadsheet to find an answer to questions such as: given a

model for calculating the budget of a company, how are cash flows affected by a change in the payment terms, such as 90 days vs. 60 days, of invoices issued in favour of the main customers? According to their characteristics, models can be divided into *iconic*, *analogical* and *symbolic*.

Iconic. An *iconic* model is a material representation of a real system, whose behaviour is imitated for the purpose of the analysis. A miniaturized model of a new city neighbourhood is an example of iconic model.

Analogical. An *analogical* model is also a material representation, although it imitates the real behaviour by analogy rather than by replication. A wind tunnel built to investigate the aerodynamic properties of a motor vehicle is an example of an analogical model intended to represent the actual progression of a vehicle on the road.

Symbolic. A *symbolic* model, such as a mathematical model, is an abstract representation of a real system. It is intended to describe the behaviour of the system through a series of symbolic variables, numerical parameters and mathematical relationships.

Business intelligence systems are exclusively based on symbolic models. A further relevant distinction concerns the probabilistic nature of models, which can be either *stochastic* or *deterministic*.

Stochastic. In a *stochastic* model some input information represents random events and is therefore characterized by a probability distribution, which in turn can be assigned or unknown. Predictive models as well as waiting line models briefly mentioned below in this unit, are examples of stochastic models.

Deterministic. A model is called *deterministic* when all input data are supposed to be known a priori and with certainty. Since this assumption is rarely fulfilled in real systems, one resorts to deterministic models when the problem at hand is sufficiently complex and any stochastic elements are of limited relevance. Notice, however, that even for deterministic models the hypothesis of knowing the data with certainty may be relaxed. Sensitivity and scenario analyses, as well as what-if analysis, allow one to assess the robustness of optimal decisions to variations in the input parameters. A further distinction concerns the temporal dimension in a mathematical

model, which can be either *static* or *dynamic*.

Static. *Static* models consider a given system and the related decision-making process within one single temporal stage. For instance, the optimization model described in this unit determines an optimal plan for the distribution of goods in a specific time frame.

Dynamic. *Dynamic* models consider a given system through several temporal stages, corresponding to a sequence of decisions. In many instances the temporal dimension is subdivided into discrete intervals of a previously fixed span: minutes, hours, days, weeks, months and years are examples of discrete subdivisions of the time axis. *Discrete-time* dynamic models, which largely prevail in business intelligence applications, observe the status of a system only at the beginning or at the end of discrete intervals. *Continuous-time* dynamic models consider a continuous sequence of periods on the time axis.

5.2 DEVELOPMENT OF A MODEL

It is possible to break down the development of a mathematical model for decision making into four primary phases, shown in Figure 5.2. The figure also includes a *feedback* mechanism which takes into account the possibility of changes and revisions of the model.

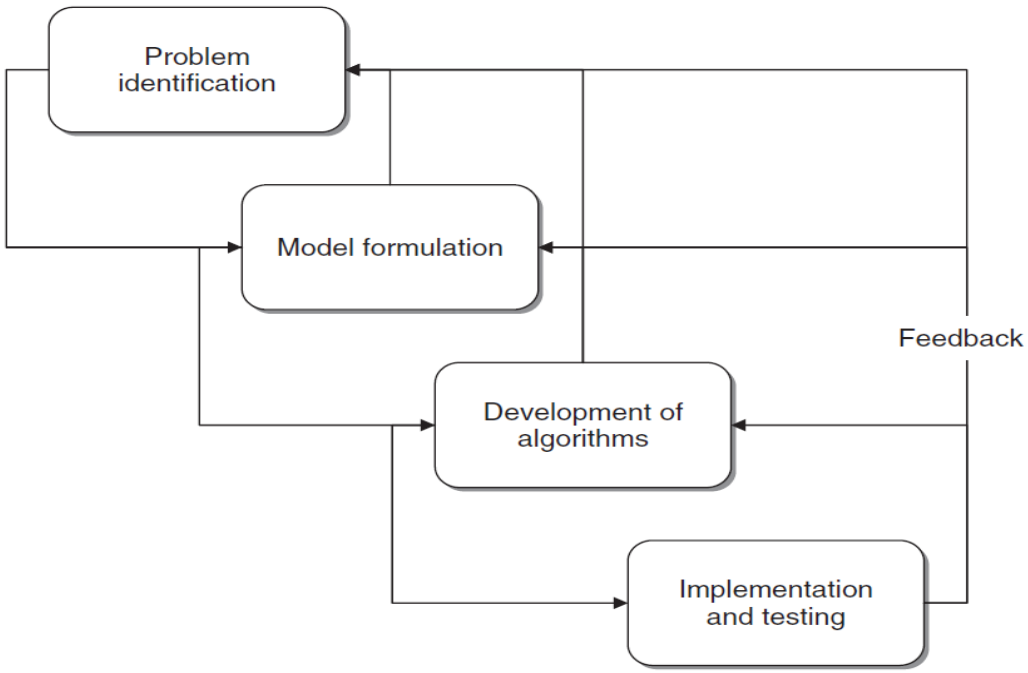


Fig. 5.2 *Phases in the development of mathematical models for decision making*

The figure also includes a *feedback* mechanism which takes into account the possibility of changes and revisions of the model.

Problem identification

First of all, the problem at hand must be correctly identified. The observed critical symptoms must be analysed and interpreted in order to formulate hypotheses for investigation. For example, too high a stock level, corresponding to an excessive stock turnover rate, may possibly represent a symptom for a company manufacturing consumable goods. It is therefore necessary to understand what caused the problem, based on the opinion of the production managers. In this case, an ineffective production plan may be the cause of the stock accumulation.

Model formulation

Once the problem to be analysed has been properly identified, effort should be directed toward defining an appropriate mathematical model to represent the system. A number of factors affect and influence the choice of model, such as the *time horizon*, the *decision variables*, the *evaluation criteria*, the *numerical parameters* and the *mathematical relationships*.

Time horizon. Usually a model includes a temporal dimension. For example, to formulate a tactical production plan over the medium term it is necessary to specify the production rate for each week in a year, whereas to derive an operational schedule it is required to assign the tasks to each production line for each day of the week. As we can see, the time span considered in a model, as well as the length of the base intervals, may vary depending on the specific problem considered.

Evaluation criteria. Appropriate measurable performance indicators should be defined in order to establish a criterion for the evaluation and comparison of the alternative decisions. These indicators may assume various forms in each different application, and may include the following factors:

- monetary costs and payoffs;
- effectiveness and level of service;

- quality of products and services;
- flexibility of the operating conditions;
- reliability in achieving the objectives.

Decision variables. Symbolic variables representing alternative decisions should then be defined. For example, if a problem consists of the formulation of a tactical production plan over the medium term, decision variables should express production volumes for each product, for each process and for each period of the planning horizon.

Numerical parameters. It is also necessary to accurately identify and estimate all numerical parameters required by the model. In the production planning example, the available capacity should be known in advance for each process, as well as the capacity absorption coefficients for each combination of products and processes.

Mathematical relationships. The final step in the formulation of a model is the identification of mathematical relationships among the decision variables, the numerical parameters and the performance indicators defined during the previous phases. Sometimes these relationships may be exclusively deterministic, while in other instances it is necessary to introduce probabilistic relationships. In this phase, the trade-off between the accuracy of the representation achieved

through the model and its solution complexity should be carefully considered. It may turn out more helpful at a practical level to adopt a model that sacrifices some marginal aspects of reality in the representation of the system but allows an efficient solution and greater flexibility in view of possible future developments.

Development of algorithms

Once a mathematical model has been defined, one will naturally wish to proceed with its solution to assess decisions and to select the best alternative. In other words, a solution algorithm should be identified and a software tool that incorporates the solution method should be developed or acquired. An analyst in charge of model formulation should possess a thorough knowledge of current solution methods and their characteristics.

Implementation and test

When a model is fully developed, then it is finally implemented, tested and utilized in the application domain. It is also necessary that the correctness of the data and the numerical parameters entered in the model be preliminarily assessed. These data usually come from a data warehouse or a data mart previously set up. Once the first numerical results have been obtained using the solution procedure devised, the model must be validated by submitting its conclusions to the opinion of decision makers and other experts in the application domain. A number of factors should be taken into account at this stage:

- the plausibility and likelihood of the conclusions achieved;
- the consistency of the results at extreme values of the numerical parameters;
- the stability of the results when minor changes in the input parameters are introduced.

5.3 CLASSES OF MODELS

There are several classes of mathematical models for decision making, which in turn can be solved by a number of alternative solution techniques. Each model class is better suited to represent certain types of decision-making processes. In this section we will cover the main categories of mathematical models for decision making, including:

- predictive models;
- pattern recognition and learning models;
- optimization models;
- project management models;
- risk analysis models;
- waiting line models.

Predictive models

A significant proportion of the models used in business intelligence systems, such as optimization models, require input data concerned with future events. For example, the results of random events determine the future demand for a product or service, the development of new scenarios of technological innovation and the level of prices and costs. As a consequence,

predictive models play a primary role in business intelligence systems, since they are logically placed upstream with respect to other mathematical models and, more generally, to the whole decision-making process. Predictions allow input information to be fed into different decision-making processes, arising in strategy, research and development, administration and control, marketing, production and logistics. Basically, all departmental functions of an enterprise make some use of predictive information to develop decision making, even though they pursue different objectives.

Predictive models can be subdivided into two main categories. The purpose of *explanatory* models is to functionally identify a possible relationship between a dependent variable and a set of independent attributes. *Regression* models, belong to this category as well as *classification* models The purpose of *time series* models is to functionally identify any temporal pattern expressed by a time series of observations referred to the same numerical variable.

Pattern recognition and machine learning models

In a broad sense, the purpose of pattern recognition and learning theory is to understand the mechanisms that regulate the development of intelligence, understood as the ability to extract knowledge from past experience in order to apply it in the future. Mathematical models for learning can be used to develop efficient algorithms that can perform such task. This has led to intelligent machines capable of learning from past observations and deriving new rules for the future, just like the human mind is able to do with great effectiveness due to the sophisticated mechanisms developed and fine-tuned in the course of evolution.

Besides an intrinsic theoretical interest, mathematical methods for learning are applied in several domains, such as recognition of images, sounds and texts; biogenetic and medical diagnosis; relational marketing, for segmenting and profiling customers; manufacturing process control; identification of anomalies and fraud detection.

Mathematical models for learning have two primary objectives. The purpose of *interpretation* models is to identify regular patterns in the data and to express them through easily understandable rules and criteria. *Prediction* models help to forecast the value that a given random variable will assume in the future, based on the values of some variables associated with the entities of a database, as for explanatory models. Based on the existence or not of a *target* attribute, the learning process may be *supervised* or *unsupervised*. In the first case, the target

attribute expresses for each record either the membership class or a measurable quantity. *Classification* and *regression* models belong to this category. In the second case, no target attribute exists and consequently the purpose of the analysis is to identify regularities, similarities and differences in the data. It is also possible to derive *association rules.* Alternatively one can determine groups of records, called *clusters*, characterized by similarity within each cluster and by dissimilarity among the elements of distinct clusters.

Optimization models

Many decision-making processes faced by companies or complex organizations can be cast according to the following framework: given the problem at hand, the decision maker defines a set of *feasible* decisions and establishes a criterion for the evaluation and comparison of alternative choices, such as monetary costs or payoffs. At this point, the decision maker must identify the *optimal* decision according to the evaluation criterion defined, that is, the choice corresponding to the minimum cost or to the maximum payoff. The conceptual paradigm outlined determines a wide and popular class of mathematical models for decision making, represented by *optimization* models.

In general, optimization models arise naturally in decision-making processes where a set of limited resources must be allocated in the most effective way to different entities. These resources may be personnel, production processes, raw materials, components or financial factors. Among the main application domains requiring an optimal allocation of the resources we find:

- logistics and production planning;
- financial planning;
- work shift planning;
- marketing campaign planning;
- price determination.

Mathematical optimization models represent a fairly substantial class of optimization problems that are derived when the objective of the decision making process is a function of the decision variables, and the criteria describing feasible decisions can be expressed by a set of mathematical equalities and inequalities in the decision variables. Mathematical optimization models offer ample opportunities for application due to the high flexibility of their formulations and the

availability of efficient solution methods. In light of the structure of the objective function and of the constraints, optimization models may assume different forms:

- linear optimization;
- integer optimization;
- convex optimization;
- network optimization;
- multiple-objective optimization.

Project management models

A *project* is a complex set of interrelated activities carried out in pursuit of a specific goal, which may represent an industrial plant, a building, an information system, a new product or a new organizational structure, depending on the different application domains. The execution of the project requires a planning and control process for the interdependent activities as well as the human, technical and financial resources necessary to achieve the final goal. *Project management* methods are based on the contributions of various disciplines, such as business organization, behavioural psychology and operations research.

Mathematical models for decision making play an important role in project management methods. In particular, network models are used to represent the component activities of a project and the precedence relationships among them. These models allow the overall project execution time to be determined, assuming a deterministic knowledge of the duration of each activity.

Stochastic models, on the other hand, usually referred to as *project evaluation and review techniques* (*PERT*), are used to derive the execution times when stochastic assumptions are made regarding the duration of the activities, represented by random variables. Finally, different classes of optimization models allow the analysis to be extended to the complex problem of optimally allocating a set of limited resources among the project activities in view of execution costs and times.

Risk analysis models

Some decision problems can be described according to the following conceptual paradigm: the decision maker is required to choose among a number of available alternatives, having uncertain

information regarding the effects that these options may have in the future. For example, assume that senior management wishes to evaluate different alternatives in order to increase the company's production capacity. On the one hand, the company may build a new plant providing a high operating efficiency and requiring a high investment cost. On the other hand, it may expand an existing plant with a lower investment but with higher operating costs. Finally, it may subcontract to external third parties part of its production: in this case, the investment cost is minimized but the operating costs are the highest among the available alternatives.

Clearly, in this situation the effects of the different options are strongly influenced by future stochastic events. In particular, a high level of future demand makes the construction of a new plant advantageous, while low demand levels tend to favour the subcontracting option. At an intermediate level of demand the expansion of an existing plant may be convenient.

However, the decision maker is forced to make a choice *before* knowing with absolute certainty the level of future demand. At best, she may obtain some stochastic information regarding the likelihood of occurrence of future events by carrying out some market research. In situations of this type, the methodological support offered by risk analysis models, primarily based on Bayesian and utility theories, may prove quite helpful. Indeed, this class of models is used successfully in several application domains, such as technology investment, design of new products, research and development, and financial and real estate investment.

Waiting line models

The purpose of *waiting line* theory is to investigate congestion phenomena occurring when the demand for and provision of a service are stochastic in nature. If the arrival times of the customers and the duration of the service are not known beforehand in a deterministic way, conflicts may arise between customers in the use of limited shared resources. As a consequence, some customers are forced to wait in a line.

Schematically, a waiting line system is made up of three main components: a *source* generating a stochastic process in which entities, also referred to as customers, arrive at a given location to obtain a service; a set of *resources* providing the service; a *waiting area* able to receive the entities whose requests cannot immediately be satisfied.

Waiting line models allow the performance of a system to be evaluated once its structure has been defined, and therefore are mostly useful within the

system design phase. Indeed, in order to determine the appropriate values for the parameters that characterize a new system, relevant economic factors are considered, which depend on the service level that the system should guarantee when operating in optimal conditions. More precisely, a model takes into account the cost of meeting the requests, which increases as the service level increases, and the cost generated by customer waiting times, which decreases as the level of service provided decreases. Since customers hope for shorter waiting times while the provider is interested in holding down the cost of service provision, the structure of the system is defined in such a way as to obtain an optimal trade-off between service costs and waiting line costs. In other words, the optimal service level, which in turn determines the ideal structure of the system, can be found at the minimum point of the curve expressing the sum of the two types of cost.

The main components of a waiting line system are the *population*, the *arrivals process*, the *service process*, the *number of stations*, and the *waiting line rules*. The population, which can be finite or infinite, represents the source from which potential customers are drawn and to which they return once the requested service has been received. The arrivals process describes how customers arrive at the system entry point. In general, this is a stochastic process described by the probability distribution of the inter-arrival times, that is, the time intervals between the arrival of two consecutive customers. The service process describes how the providers meet the requests of the customers waiting in line. This in turn is a stochastic process, defined by the probability distribution of the service time, that is, the amount of time that customers spend with the resources providing the service. The number of existing stations and the number of providers assigned to each station are additional relevant parameters of the waiting line system. The waiting rules describe the order in which customers are extracted from the line to be admitted to the service. A primary role is finally played by priority schemes in which a level of priority is assigned to each customer. The customer with the highest priority is then served before all the other customers waiting in line.

5.4 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.

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. The models and patterns identified in this way may take on different forms, such as linear equations, sets of rules in *if–then–else* form, clusters, charts and trees.

5.4.1 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.

5.4.2 DATA MINING, CLASSICAL STATISTICS AND OLAP

Data mining projects differ in many respects from both classical statistics and OLAP analyses 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.

5.4.2 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 detection.

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 behaviour 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 represent 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.

5.5 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 Rs 270 and Rs 360 respectively, it makes sense to claim that the difference between the amounts spent by the two customers is equal to Rs 90 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.

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

To represent a generic dataset *D*, we will denote by *m* the number of observations, or rows, in the two-dimensional table containing the data and by *n* the number of attributes, or columns. Furthermore, we will denote by

$$\mathbf{X} = [x_{ij}], \quad i \in \mathcal{M} = \{1, 2, \ldots, m\}, \quad j \in \mathcal{N} = \{1, 2, \ldots, n\},$$

the matrix of dimensions $m \times n$ that corresponds to the entries in the dataset *D*. We will write

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})$$

$$\mathbf{a}_j = (x_{1j}, x_{2j}, \ldots, x_{mj})$$

for the *n*-dimensional row vector associated with the *i*th record of the dataset and the *m*-dimensional column vector representing the *j* th attribute in *D*, respectively.

5.6 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.

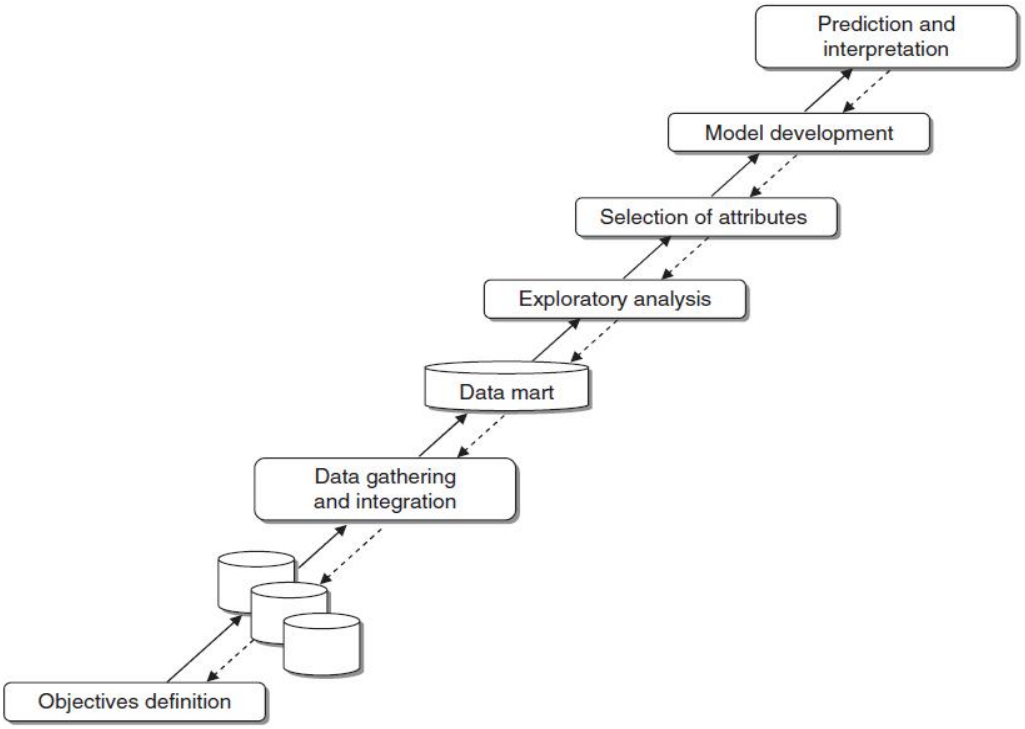


Fig. 5.1 *Data mining process*

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 geomarketing variables, or with lists of names of potential customers, termed *prospects*, not yet existing in the company information system.

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 favourable 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 behaviour 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.

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 values are also highlighted.

Attribute Selection.

In the subsequent phase, the relevance of the different attributes are 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 standpoint – 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.

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. Moreover, 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 interaction 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 supervise the access to the information sources;
- an expert in the mathematical theory of learning and statistics, for exploratory data analysis and for the generation of predictive models.

5.7 ANALYSIS METHODOLOGIES

Data mining activities can be subdivided into a few major categories, based on 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 or expresses a measurable quantity. As an example of the supervised perspective, 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 behaviour, 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. 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 target attribute, 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 classification and 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 catalogues. 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 predefined 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 sometimes 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.

5.8 CHECK YOUR PROGRESS

1. Define a stochastic model
2. What is a risk analysis model
3. Define a waiting line model
4. Define data mining.
5. What is the difference between OLAP, classical statistics and data mining.
6. What is supervised learning?
7. What is unsupervised learning?

Answers to Check your progress

1. A *symbolic* model, such as a mathematical model, is an abstract representation of a real system. It is intended to describe the behaviour of the system through a series of symbolic variables, numerical parameters and mathematical relationships.

2. risk analysis models is primarily based on Bayesian and utility theories

3. awaiting line system is made up of three main components:
   a *source* generating a stochastic process in which entities, also referred to as customers, arrive at a given location to obtain a service; a set of *resources* providing the service; a *waiting area* able to receive the entities whose requests cannot immediately be satisfied.
4. 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.

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

5. In a supervised (or *direct*) learning analysis, a target attribute either represents the class to which each record belongs or expresses a measurable quantity,

6. 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

5.9 SUMMARY

In the unit we have emphasized the critical role played by mathematical models in the development of business intelligence environments and decision support systems aimed at providing *active* support for knowledge workers. here we focused on the main characteristics shared by different mathematical models embedded into business intelligence systems. We also developed a taxonomy of the most common classes of models, identifying for each of them the prevailing application domain.

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 community, 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 described and characterized data mining activities with respect to investigation purposes and analysis methodologies. The relevant properties of input data is also be discussed. Finally, we described the data mining process and its articulation in distinct phases.

5.10 KEYWORDS

- **Iconic. -** An *iconic* model is a material representation of a real system
- **Analogical -** An *analogical* model is also a material representation
- **Symbolic -** A *symbolic* model, such as a mathematical model, is an abstract

representation of a real system.

- **Stochastic -** In a *stochastic* model some input information represents random events and is therefore characterized by a probability distribution
- **Deterministic.** A model is called *deterministic* when all input data are supposed to be known a priori and with certainty
- **Static.** *Static* models consider a given system and the related decision-making process within one single temporal stage.
- **Dynamic.** *Dynamic* models consider a given system through several temporal stages, corresponding to a sequence of decisions.
- **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.
- **Categorical -** Categorical attributes assume a finite number of distinct values
- **Numerical -** Numerical attributes assume a finite or infinite number of values and lend themselves to subtraction or division operations

5.11 SELF ASSESSMENT QUESTIONS

1. Explain classification of mathematical models for decision making.
2. Explain phases in the development of mathematical models for decision making.
3. Write a brief note on different classes of mathematical models.
4. Write a brief note on application of data mining.
5. Explain data mining process.

5.12 REFERENCES

1. Swain Scheps - Business Intelligence For Dummies-For Dummies (2008), Wiley Publishing, Inc
2. Carlo Vercellis - Business Intelligence_ Data Mining and Optimization for Decision Making (2009), Wiley Publishing Inc
3. Howson, Cindi - Successful Business Intelligence-McGraw-Hill (2014)

UNIT -6: OLAP: ONLINE ANALYTICAL PROCESSING

Structure

6.0 Objectives

6.1 OLAP in Context

6.2 OLAP Application Functionality

6.3 Multidimensional Analysis

6.4 OLAP Architecture

6.5 Benefits of OLAP

6.6 Drill team: Working with Multidimensional Data

6.7 OLAP versus OLTP

6.8 Looking at Different OLAP Styles and Architecture

6.9 Check your progress

6.10 Summary

6.11 Keywords

6.12 Self Assessment Questions

6.13 References

6.0 OBJECTIVES

After studying this unit, you will be able to

- ✓ Explain OLAP basics
- ✓ Drill into data
- ✓ Look at OLAP on different platforms
- ✓ Analyze the benefits of a hybrid approach

6.1 OLAP IN CONTEXT

Up until now, our focus has been on the technology foundations of a BI solution. First, you identify the important data in your company; that is, the transactional information that resides on ERP, CRM and other operational systems. Next you gather the data together into a single place, and in a single format; that's the job of the ETL processes, the data warehouse and related components.

Finally, you provide access to that mountain of data in the form of querying and reporting tools. Producing unified reports against companywide operational data is likely to increase the managers' visibility into how all the pieces interact, and make it more likely that valuable insights might come to light.

But years ago some clever programmers and database engineers realized that they could get more from company transactional data by conceptualizing it differently than was traditionally called for. The concept was multi-dimensional data rather than relational data — manifested as online analytical processing, or OLAP (pronounced "OH-lap").

Instead of just aggregating and summarizing your data, OLAP tools give BI systems the ability to look at it in a truly new way. Add the increased computing horsepower and innovative software tools available at the time, and a whole new paradigm was born.

6.2 OLAP APPLICATION FUNCTIONALITY

An OLAP application is software designed to allow users to navigate, retrieve, and present business data. Rather than taking data from a relational system, writing complex queries to

retrieve it, then manually inserting it into a report to analyze, OLAP tools cut out the middle steps by actually storing the data in a *report-ready* format.

Traditional data processing was like plumbing: Your query would get flushed down into the system. You'd wait a while, and hope that what came out of the pipes on the other end was what you wanted. That's not the case with OLAP. While it's true you have to make choices about the kind of data you want to get a look at initially, it's possible to twist and cut the results immediately. There's no need to work through confusing query logic or write long SQL strings. With a little drag-and-drop action, you're looking at the data in a whole new way. the *online* and *analytical* aspects of OLAP add up to a key difference. Consider: Rather than a set-in concrete report, OLAP allows for fully interactive sessions between users and software. If it used to take you a week to build the report of your dreams to hand in to your boss, you're in luck: OLAP can take you through the process in a matter or minutes and leave enough time to do the analysis yourself.

6.3 MULTIDIMENSIONAL ANALYSIS

An OLAP data-and-reporting environment is different from a traditional database environment — mainly because of the way data is conceptualized and stored and we aren't talking about putting it on magnetic tapes instead of floppy disks. In an OLAP system, users work with data in dimensional form rather than relational form. OLAP's bread and butter is multidimensional data. a dimension is nothing more than a way to categorize data.

The *A* in OLAP is not just any old kind of analysis — it's specifically *quantitative* — based on good old number crunching — rather than qualitative.

OLAP software is designed to work with numeric data. That's why most of the examples you'll see are accounting, finance, or some other calculation heavy subject.

6.3.1 LONELY NUMBERS

At their core, financial records and sales data are really just numbers — prices, costs, margins, quantities ordered, and time periods. A list of numbers means nothing unless you know something about what those numbers represent. What makes numbers meaningful (and available for analysis) is the details: their *descriptions* and *qualifiers*.

If you see a number in a vast table, it will mean little to you without the context provided by the *title* of the table (or chart or graph) and the names and elements of the axes.

6.3.2 One-dimensional data

Consider a table whose title is “Sales Data” with the vertical axis labeled “Region” — and you can begin to draw some business conclusions right away, as in the Table 6-1.

Table 6.1 A One-dimensional Table

	Annual sales data
Region	**Amount**
Northeast	Rs. 45,091
Southeast	Rs. 73,792
Central	Rs 88,122
West	Rs.63,054
TOTAL:	Rs. 270,059

Although the data is (of course) displayed in two dimensions (horizontal and vertical), in OLAP terms this is considered one dimensional data. In other words, we’re looking at sales data from one particular perspective (or dimension): in this case, by region.

Since companies usually record lots of information about each individual sales transaction, you could probably look at the same batch of transactions in a different way. Table 6-2 shows the same underlying sales data in a different dimension: by product type. The total sales amount is identical with the total in the sales table by region because the exact same transactions are used to calculate both. We’re just looking at it broken out by product types rather than by regions.

Table 6.2 The Same Data from a Slightly Different Direction

	Annual Sales Data
Product	**Amount**
Gizmo	Rs. 88,697
Widget	Rs 181,362
TOTAL:	Rs 270,059

Another important factor in OLAP analysis is time, so let's take a look at what the sales figures look like through the time dimension. Table 6-3 shows the same underlying transactions broken out by calendar quarters:

Table 6.3. Yet another view of the sales data

	Annual sales data
Quarter	Amount
Q1	Rs. 55,837
Q2	Rs. 87,659
Q3	Rs. 23,598
Q4	Rs 102,915
Total	Rs. 270,059

6.3.3 Setting the table

By running reports that show a breakdown of annual sales figures by region, product, and quarter, we know what the total sales figure is — and we can recognize which of those dimensions shows sales strength and weakness. But where we go from there is a different story. One option is to combine two of the dimensions into a table like Table 6.4

Annual sales data

Product			Totals
Quarter	Gizmo	Widget	
Q1	Rs. 23,199	Rs 32,688	Rs 55,887
Q2	Rs 24,798	Rs 62,861	Rs 87,659
Q3	Rs 14,555	Rs 9,043	Rs 23,598
Q4	Rs 26,145	Rs 76,770	Rs 102,915
Totals	Rs 88,697	Rs 181,362	Rs 270,059

With two-dimensional data, you can begin to think of dimensions as a kind of *coordinate system*. With quarters listed down the vertical axis and product type listed across the horizontal axis,

each unique pair of values of these two dimensions correspond to a single point of data. For example, we know that Quarter 2's Widget sales were Rs 24,798.

We also have regional data as well, so how can we incorporate that information here? One way would be to create a series of two-dimensional tables. We could map regional sales figures by quarter, and we could map regional sales figures by product type. But that doesn't necessarily get us where we ultimately want to go.

6.3.4 Seeing in 3-D

An analyst or a manager would find it especially helpful to be able to see all the dimensions — regional, quarterly, and by product type — on the same table at the same time.

But since spreadsheets and tables render all information in two spatial dimensions, we have to rig our table to handle the extra complexity. In Table 5-5, we've drilled deeper into the sales data. The individual data points from our original three one-dimensional tables are now the subtotals on the edges of this lone three-dimensional table. Annual sales data Region

	Northeast	**Southeast**	**Central**	**West**	**Total**
GIZMOS					
Q1	Rs 1,543	Rs 14,098	Rs 1,991	Rs 5,567	Rs 23,199
Q2	Rs 6,811	Rs 2,822	Rs 13,300	Rs 1,865	Rs 24,798
Q3	Rs 5,190	Rs 5,050	Rs 2,106	Rs 2,209	Rs 14,555
Q4	Rs 2,347	Rs 8,005	Rs 5,900	Rs 9,893	Rs 26,145
WIDGETS					
Q1	Rs 3,555	Rs 5,520	Rs 6,828	Rs 16,785	Rs 32,688
Q2	Rs 9,158	Rs 15,999	Rs 18,096	Rs 19,608	Rs 62,861
Q3	Rs 2,486	Rs 1,297	Rs 4,247	Rs 1,013	Rs 9,043
Q4	Rs 14,001	Rs 21,001	Rs 35,654	Rs 6,114	Rs 76,770
TOTALS	Rs 45,091	Rs 73,792	Rs 88,122	Rs 63,054	Rs 270,059

With this table, you begin to see the value of multidimensional analysis. The totals and subtotals are the same, but with all three dimensions represented together, you begin to see how these factors interact with each other in an operational sense.

As an example of some analysis you might do, notice how the Northeast and Central regional sales figures are disproportionately lower in the first quarter for Gizmos, relative to those of the other regions and the other products. Taking note of that anomaly might spur a call to the regional managers, who would then explain to you that Gizmos don't work well in the snow, making them hard to sell in Q1 because it's wintertime.

6.3.5 Beyond the third dimension

There's (theoretically) no limit to the number of dimensions you can use to describe your data. It all depends on what information your operational and transactional systems capture — and how finely detailed you want the picture to be. For example, your company's CRM system might also have data on specific sales reps. The accounting system could probably break down the time dimension further into months, weeks, and days.

There are practical limits on the software you use. Storing and manipulating multidimensional data is resource-intensive; it takes a lot of number crunching. So you should make sure the OLAP application you work with matches with your data model.

Even though *multi* means more than one, *multidimensional* in the OLAP world typically refers to data that can be described in three or more dimensions (as in our earlier example where you have sales data by time, by product, and by region). OLAP applications can certainly handle two-dimensional data, but one of the main reasons you'd use it is to handle information you can't easily generate with a table.

The output or results of an interactive session with OLAP data is often just a one- or two-dimensional data view. OLAP gives the analyst the freedom to view information from different vantage points and in different ways. After cleaving and rearranging the multidimensional data, the analyst will find or calculate the essential data and put it into tabular form.

6.4 OLAP ARCHITECTURE

OLAP systems are fundamentally different from other forms of data conceptualizations because it handles data in the same way people do when they're creating reports.

These systems are designed to work in concert with the other tools in the business intelligence architecture. The OLAP system typically comprises two distinct categories of software:

- The *OLAP cube* houses the multidimensional data
- The *access tools* allow users to build and massage information into forms appropriate for analysis.

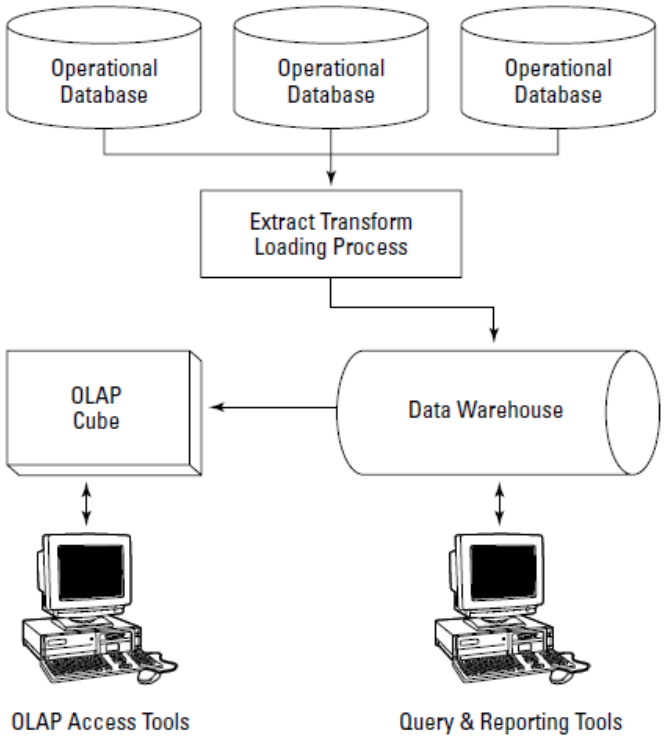


Fig.6.1: OLAP tools in the BI architecture

6.4.1 The OLAP Cube

The table is the most common representation of numeric data because it's a highly useful and easily recognizable structure. It also lends itself to the two dimensional world of paper and computer screens. But tables are, well, flat. tabular list or a row-and-column matrix that uses the

vertical and horizontal to represent the key characteristics. So when it comes time to add more complexity — in the form of extra data dimensions — tables quickly lose their utility.

6.4.2 It's a data structure

To store multidimensional data, we conceptualize it as a *cube*, where each of the three axes represents a different dimension of the same information. A cube version of the previous example would look like Figure 6-2.

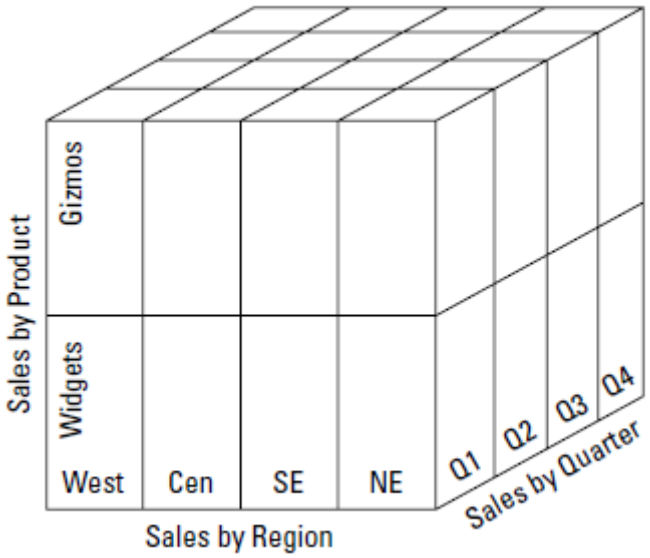


Fig. 6.2 The cube full of sales data

The cube has 32 cells of data, just like the three-dimensional table we drew. Another way to think about it is that we're taking two 2D reports and stacking them on top of each other — in this case Gizmo sales by region and by quarter, atop Widget sales by region and quarter.

That's the concept of a cube — but how does it work in real life?

In an OLAP environment, a cube is a specialized data store designed specifically to handle multidimensional summary data. But rather than being held in relational tables (which are built to process transactions rapidly), cube data is stored in cells; its structure is like a 3D spreadsheet.

In OLAP-speak, a *dimension* refers only to a characteristic of the data, not a direction in space. So when you hear talk about a cube with 25 dimensions, don't give yourself a headache trying to

picture the thing. In reality, it's just a list of sales figures that can be jockeyed in many different ways. Likewise, *cube* is a convenient proxy term for *any* multidimensional data structure. After

all, if you add a dimension to a flat two-dimensional table, you've made a cube. And since we don't have words for any shape beyond three-dimensions, *cube* is the closest we can get.

Some vendors use the term *cube* as a general term to describe all multidimensional data in an OLAP environment; others say their software employs many cubes at once, one for each general data subject. For example, one vendor may describe an OLAP system where the tools access several different sales cubes, inventory cubes, and so forth. But another vendor may simply say that all tools access *the* cube. There's no difference beyond semantics.

6.4.2 OLAP access tools

The OLAP access tools are the client environments that allow users to twist and turn the cube's data and ultimately produce meaningful business intelligence.

The goal of the OLAP access tool is to present large quantities of information to the user in a way that lets them produce business insights without losing sight of the larger context of what they're doing. That means the tools need to have the following characteristics:

- **Easy:** The coin of the realm with OLAP end-user software is its ease of use. That means not only its ability to quickly and intuitively navigate what can be extremely complicated data arrangements, but also drag and- drop controls, and simple administration of the data views and files.
- **Beautiful:** Spotting anomalies can be a matter of turning a river of numbers into just the right kind of visualization. The best tools will also have lots of charting, graphing, and reporting options.
- **Smart:** Users need to be able to tune the OLAP access tool to the right kinds of tasks. That means intelligent searching functions and robust analysis capabilities to help the user identify trends or create forecasts. In the early days of OLAP, vendors packaged their cubes and access tools into a single pseudo-client-server package; there were two distinct applications that had to be deployed together. That approach is still common today, but plenty of companies

specialize in one application or the other, building their software to work with that of other companies.

6.5 BENEFITS OF OLAP

You can take specific actions with data in an OLAP system that you can't do in other environments:

- **Consolidation:** This is another word for rolling up data into the next higher level of abstraction. For example, sales offices can be rolled up into districts, and districts rolled up into regions or complex expressions that connect interrelated data.
- **Drilling into the data:** With OLAP applications, you aren't just looking at a static report. Nor are you forced to thumb through 1000 different reports covering different areas of the business. OLAP lets you work through the data in a natural, intuitive way. See a data point on which you need more information? Click it — the software will re-orient the view around that piece of information. Consider our three-dimensional sales report showing subtotals by quarter, product type, and region. Clicking on one of the sales figures might take you to a view of the individual transactions. Clicking on a product category might open up detailed sales information on that category. OLAP drilling lets you go straight to the information you need.
- **Computation:** The formal word for number crunching. Because OLAP's focus is typically vast amounts of numeric data, the applications have built-in mathematical functions to help users turn more raw data into *less* raw data. For example, if your data warehouse contains sales and production-cost numbers, you can include the derived data in your final report that shows profit margins.
- **Pivoting**: If you've worked with a pivot table in Excel, you know the value of being able to view your data or report from different perspectives.

OLAP has its own peculiar lexicon that you need to become familiar with if you do any work with multidimensional analysis and reporting. Different vendors or consultants complicate matters by using their own proprietary glossaries that may not correspond to everyone else's. Nevertheless, here are some common terms for the OLAP concepts you'll encounter most frequently:

- **Attribute:** A descriptive detail, or a way to subdivide or categorize dimensions. In a way, attributes are members of dimensions. For example, if you are storing sales in a multidimensional environment, your dimensions might be time, product, and location. Each of the three dimensions has attributes that describe it; product's attributes might be "product name", "product type", "product family" and "product ID#". Attributes often have meaningful hierarchies too; you could create your multidimensional table so that certain product families combine to make a single product type. The time dimension is made up of "month", "quarter", and "year" which are all related to one another in ways familiar to everyone.
- **Cell:** A single data point — like Rs5439 — identifiable by a coordinate system. A table is made up of axes and a lot of cells.
- **Measure**: Think of the title of your report as the long version of your measure — the general description of what's going on in the table. The measure usually corresponds to a description of what is represented by the cells (data points). In our example, the data points were dollar figures representing daily sales by store; the measure would be simply "sales." You sometimes see the word "measure" used to mean much the same as "fact." Technically there are some subtle distinctions between a measure and a fact, but to get a handle on the basics of OLAP, it's usually okay to consider the two terms synonymous.
- **Member:** One discrete element within a dimension. In the preceding example, we listed out stores by number along the horizontal axis of the table; any individual store (like Store #49) as a member of the "location"
  dimension.

6.5.1 Remember the Big Four BI criteria

Think for a moment about the essential characteristics of business intelligence. Regardless of how you get them, what applications are used, and what acronyms are involved, you need insights that are timely, accurate, high-value, and actionable. Here's how those criteria look from an OLAP perspective:

- **Timely:** Manipulating data in a modern OLAP system is a much faster way of producing relevant business data and presenting it in an intuitive, intelligent way than querying and assembling data from a relational database.
- **Accurate:** Multidimensional data doesn't just provide mere accuracy; you can trace it. You can dig into any cell instantly and find out what its constituent data elements are. In an OLAP report that shows quarterly sales data (for example), you can drill down and see monthly, weekly and daily information as needed to make your point.
- **High-value:** Even though an OLAP system deals in multidimensional data, it's still the same information that's housed in the transactional systems. Using OLAP, you can perform advanced data manipulation and analysis on just about any data stored in your enterprise. That means you can move quickly through data that doesn't matter to get to the data that does matter. And when high-value data is easier to get, good valuable insights are likely to appear.
- **Actionable:** Your OLAP system is especially good at aiding in analysis, allowing precise trends to be plotted and activities to be monitored. That means analysts can recommend immediate action to take advantage of a situation — or prevent a problem from growing worse.

6.6 DRILL TEAM: WORKING WITH MULTIDIMENSIONAL DATA

Users get more utility from reporting-and-analysis applications when they can explore the data on the fly. A typical *dumb* report is just a list of numbers, a table, or a series of tables. OLAP links all the data together at various levels in the system, then gives users access to it.

Because OLAP data dimensions are usually arranged in terms of hierarchies, it's important that the users be allowed to navigate up or down the different levels as their jobs require. That lets 'em get granular and settle on the right level of detail for performing a mandated analysis.

OLAP terms this maneuverability *drill capabilities*. Viewing a data point's constituent parts is called *drilling down*. This allows you to see business insights in greater detail. If you're examining how a data point aggregates into broader figures and calculations, you're *drilling up*. And viewing related data is called *drilling through* — a metaphor of moving sideways instead of vertically.

6.6.1 Gaining insight through drill-down analysis

Most users start their analysis with figures that are broad in scope, and from there they move into finer levels of detail as they see anomalous data or interesting trends. This is the *drill-down* process, an essential tool in the OLAP user environment.

Specifically, drilling down is the action of moving through the data hierarchy to the next (more specific) level of detail. As an example, a user might take an initial look at the OLAP cube and see quarterly production data, as shown in Table 6-6.

Table 6-6 **A Drill-Down Example**

2007 Production Data

Quarter ➪	Q1	Q2	Q3	Q4
Factory A	577	529	499	442
Factory B	301	306	299	280
Factory C	753	731	819	648

The user might notice that each of the three factories were down in the fourth quarter of production — and double click the Q4 cell to open a new table that looks like Table 6-7.

Table 6-7 **The Drill-Down Destination**

2007 Production Data

Month	Oct	Nov	Dec
Factory A	122	116	104
Factory B	93	107	80
Factory C	220	227	201

Okay, here's the user's action translated into OLAP-speak: Looking at the "2007 Production Data", the user double-clicked one of the dimension members (Q4) and drilled down by one level, opening a new table that displays the next level "down" in the hierarchy. This table shows more granularity — finer details of the fourth quarter. The user is still looking at factory production data — but can now look at each factory's production for the three specific months that make up the fourth quarter.

Okay, this is a simplistic example, but it gives you an idea of how useful an OLAP access tool can be. Instead of having to rewrite a report, the user can simply drill into the data to reach the desired level of detail.

6.6.2 Going in the other direction: drill-up analysis

As with drill-down analysis, the user can also take advantage of multidimensional data's hierarchical nature by looking at larger groupings of data — also with a single click. This is called *drilling up* in a report. Drilling up effectively telescopes the more detailed levels of the data hierarchy into the next level up, and consolidates (*rolls up*) their data totals. If the user in the previous example wanted more of a bird's-eye view of the data, he or she could *drill up* from the quarterly table and view (say) the *annual* production total (that is, the quarterly totals combined). The table would provide a context across other years of output numbers, as in Table 6-8.

Table 6-8 **A Drill-Up Example**

Production Data				
Year ➪	2007	2008	2009	2010
Factory A	2047	1998	2166	2141
Factory B	1186	1094	1101	1136
Factory C	2951	2795	2788	2993

Again, the rest of the table remains the same, as it was in the drill-down analysis. We're still looking at production data over time, but now we're seeing yearly data with less granularity (detail). You'll be comforted to know the totals for 2007 by each factory are the same as before — and why not? It's exactly the same underlying production data.

6.6.3 Getting to the source: drill-through

Lying behind any item of multidimensional data are the many — sometimes millions — of rows of source information. The OLAP cube aggregates this data and transforms it into the multidimensional form you see in your OLAP front-end tools. But what if you need to go back and look at the original source data? *drill-through* capabilities

Drill-through allows analysts to move between the OLAP table view and the source data. For example, suppose an auto dealership has an OLAP cube in place to provide instant analysis and reporting — but in the process of producing the annual sales report to the ownership group, the lot manager notices that the margins on a certain car model are far lower than they are for other models.

In this case, drilling down into more granular levels of the data hierarchy won't help. The information that's really needed is the individual transactions themselves — and that data is stored on the relational databases of the operational systems. OLAP access tools that can drill through to those databases (on the same level of detail) give the user quick and seamless access to the source transactions that make up the aggregations shown in the OLAP cube.

So why not just keep all your multidimensional data in a relational database? Well, there are some really good reasons not to — and they have to do with the difference between OLAP and OLTP, detailed in the next section.

6.7 OLAP versus OLTP

The reason behind building an OLAP database is because it excels where regular relational databases don't do such a good job. It might be useful to remember what's involved in an OLTP system. For day-to-day operations such as running the accounting department or call centre, almost every company relies on transactional databases. Transactional systems are built to perform data actions in a few fundamental ways:

- **Rapidly:** Transactional systems must allow their users to read, write, and delete data quickly. For example, picture a point-of-sale (POS) system of a large retailer where the back-end database must allow rapid, simultaneous processing of cash-register-like transactions. Every time a customer buys a stick of gum, the cashier scans the gum's bar code, accesses the database, and retrieves the price and product information. Then the system adds a record of the final purchase — all fast enough to get the customer out the door with minimum wait time.
- **In vast numbers:** In addition to working rapidly, transactional systems must be able to address billions of rows of data. Imagine an inventory system for a multinational enterprise with hundreds of warehouses all over the planet. Every addition, deletion, and change to a warehouse's contents must be recorded in a database.

- **In real time:** Transactional systems operate on a more or less continuous basis, reacting to user actions immediately and processing transactions on demand. Transactional systems that support a company's basic operations are called *Online Transaction Processing* systems or *OLTP*. Even though ERP software provides some unity in the back-office applications and data, most big companies must run a variety of OLTP systems to support day-to-day operations.

6.8 LOOKING AT DIFFERENT OLAP STYLES AND ARCHITECTURE

A multidimensional approach to data can be useful in different kinds of architectures.

6.8.1 MOLAP: multidimensional OLAP

MOLAP — *Multidimensional Online Analytical Processing* — is the architecture based on cubes. It's a version of OLAP that's built for speed: MOLAP stores data in logical structures uniquely constructed to accelerate retrieval; these are separate from the relational database system (RDBMS). In spite of its unappealing acronym, this is the traditional cube architecture described so far. The M for *multidimensional* simply indicates that in a MOLAP environment, a cube structure sits as a layer between the OLAP access tools and the relational database.

6.8.1.1 Old days of MOLAP

OLAP can trace its origins to the 1970s, when the relational database model was starting to gain a foothold in the programming community. An IBM engineer named Ken Tomlinson developed a high-level language called APL (you might expect it to stand for something technical and important, but APL was really just A Programming Language.) APL was radical because it included built-in functions to handle multidimensional data.

APL was as much mathematics as computer science. The old saying "It's Greek to me" applied (literally and figuratively) to APL; many of its programmatic operators were letters from the Greek alphabet, making it difficult to code on a standard keyboard and ASCII-based system.

APL did, however, spark some interest in the multidimensional approach. Although relational databases became far more prevalent, the ancestor of OLAP continued to evolve in the shadows through products such as TM1, an early spreadsheet application. There were also analysis and reporting companies such as Comshare and Information Resources (IRI) that produced databases with many multidimensional characteristics.

6.8.1.2 MOLAP in modern times

MOLAP-based systems began to hit the mainstream in 1992 with the release of the Arbor Software (later Hyperion) product Essbase, which IBM later integrated with its DB/2 relational-database product. Then Oracle bought IRI's product with the intention of turning it into a packaged layer of their traditional relational product. About that time, the term OLAP came into being, often attributed to Dr. E.F. Codd, best known as the father of the *relational* database.

Other OLAP products came on the market in the 1990s — Cognos Powerplay, SAP BW, and early versions of Microsoft Analysis Services — each of which still holds an important position in the market today.

If there is a downside to a MOLAP environment (versus other analysis-and reporting architectures), it is that MOLAP necessarily involves an extra layer — the cube — and the architecture is already complex. It requires specific expertise beyond what the average database administrator can offer.

6.8.1.3 ROLAP: relational OLAP through "normal" databases

RDBMSs didn't get along well with OLAP at first; they were designed for transaction processing and efficient storage of data, not for analysis and reporting. Nevertheless (to the shock and consternation of many OLAP traditionalists), a breed of OLAP that goes completely cube less — ROLAP (Relational OLAP) — emerged, and is still alive and well.

Instead of proprietary multidimensional databases, ROLAP *emulates* a cube layer, inserting a semantic layer between the database and the end-user tool that mimics the data cube's actions. The OLAP access tools access the semantic layer as if they were speaking to OLAP cube.

The downside here, of course, is that relational databases weren't traditionally structured to deal with data in a dimensional format. When the information takes on more dimensions and increased hierarchical complexity, performance lags. A decade ago, there simply was not enough computing horsepower to handle such a load — but today's relational databases are more capable.

On the other hand, using ROLAP allows companies to use the relational database they've already installed (say, Oracle or DB2). What's more, they don't have to hire experts in building and maintaining multidimensional cubes and integrating all that stuff with the relational system.

6.8.1.4 HOLAP:

There are strong arguments on both sides of the ROLAP-versus-MOLAP debate. And the acronym-slinging doesn't end there. HOLAP (Hybrid OLAP) is an attempt to combine the best of both worlds. In the early 1990s, it was necessary to make a firm commitment to one technology or the other, but these days' vendors have taken serious steps to integrate the most useful functions of both worlds. Products like Microsoft SQL Server, with its integrated Analysis Services package, mean that organizations don't have to choose either ROLAP or MOLAP.

It's a lot like the hybrid automobiles on the market today. After all the questions about whether to buy an electric car, the manufacturers ended the debate by building a hybrid. When the driver needs power, the good old gas-guzzling, exhaust-spitting, V-8 internal combustion engine roars to life and boosts the car onto the freeway. But when the car is cruising, the gasoline-powered motor shuts off and gives way to the more efficient electric motor.

HOLAP systems perform the same kind of behind-the-curtains tricks as hybrid automobiles, switching modalities back and forth outside the view of the user. The cube structure is in place to handle large numbers of dimensions spanning many levels of hierarchy. They offer rapid performance and fast refresh times for workers performing analysis and creating complex reports. Meanwhile, the hybrid systems can rely upon the space-saving ROLAP architecture to store larger volumes of raw data, funnelling only the necessary summary information to the cube.

And when the user needs drill-through capabilities to dig into the source transactions, the OLAP access tools can work directly with the relational system without a hitch.

6.9 CHECK YOUR PROGRESS

1. Define OLAP
2. What is OLAP cube?
3. What is drill down operation?
4. What is drill up operation?
5. What is drill through operation?
6. What is MOLAP?
7. What is ROLAP?

Answers to Check your progress

1. **Online analytical processing** (OLAP) is a technology that organizes large business databases and supports complex analysis.
2. It's a data structure to store 3d data.
3. Viewing a data point's constituent parts is called *drilling down*. This allows you to see business insights in greater detail.
4. If you're examining how a data point aggregates into broader figures and calculations, you're *drilling up*.
5. viewing related data is called *drilling through* — a metaphor of moving sideways instead of vertically.
6. *Multidimensional Online Analytical Processing* — is the architecture based on cubes. It's a version of OLAP that's built for speed: MOLAP stores data in logical structures uniquely constructed to accelerate retrieval
7. It is a breed of OLAP that goes completely cubeless

6.10 SUMMARY

Online analytical processing, or **OLAP** is an approach to answer multi-dimensional analytical (MDA) queries swiftly in computing. OLAP is part of the broader category of business intelligence, which also encompasses relational databases, report writing and data mining. Typical applications of OLAP include business reporting for sales, marketing, management reporting, business process management (BPM) budgeting and forecasting, financial reporting and similar areas, with new applications emerging, such as agriculture.

The term *OLAP* was created as a slight modification of the traditional database term online transaction processing (OLTP).

OLAP tools enable users to analyze multidimensional data interactively from multiple perspectives. OLAP consists of three basic analytical operations: consolidation (roll-up), drill-down, and slicing and dicing. Consolidation involves the aggregation of data that can be accumulated and computed in one or more dimensions. For example, all sales offices are rolled up to the sales department or sales division to anticipate sales trends. By contrast, the drill-down is a technique that allows users to navigate through the details. For instance, users can view the sales by individual products that make up a region's sales. Slicing and dicing is a feature whereby users can take out (slicing) a specific set of data of the OLAP cube and view (dicing) the slices from different viewpoints. These viewpoints are sometimes called dimensions (such as looking at the same sales by salesperson, or by date, or by customer, or by product, or by region, etc.).

Databases configured for OLAP use a multidimensional data model, allowing for complex analytical and ad hoc queries with a rapid execution time. They borrow aspects of navigational databases, hierarchical databases and relational databases.

OLAP is typically contrasted to OLTP (online transaction processing), which is generally characterized by much less complex queries, in a larger volume, to process transactions rather than for the purpose of business intelligence or reporting. Whereas OLAP systems are mostly optimized for read, OLTP has to process all kinds of queries (read, insert, update and delete).

6.11 KEYWORDS

- OLAP - online analytical processing
- OLAP Cube - It's a data structure to store multidimensional data
- OLTP - Online Transaction Processing.
- drilling down - is the action of moving through the data hierarchy to the next (more specific) level of detail
- MOLAP: multidimensional OLAP
- ROLAP: relational OLAP
- HOLAP (Hybrid OLAP)

6.12 SELF ASSESSMENT QUESTIONS

1. With the help of examples, explain multidimensional data.
2. Explain OLAP architecture.

UNIT -7: AGILE DEVELOPMENT

Structure

7.0 Objectives

7.1 Waterfall Development Process

7.2 Agile Development Techniques

7.3 Basic Concepts of Scrum

7.4 Agile Culture at Netflix

7.5 Medtronic: Agile for the Right Projects

7.6 Sharper BI at 1-800 CONTACTS

7.7 Best Practices for Successful Business Intelligence

7.8 Check your progress

7.9 Summary

7.10 Keywords

7.11 Self Assessment Questions

7.12 References

7.0 OBJECTIVES

After studying this unit, you should be able to

- ✓ Explain waterfall development process
- ✓ Examine 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.

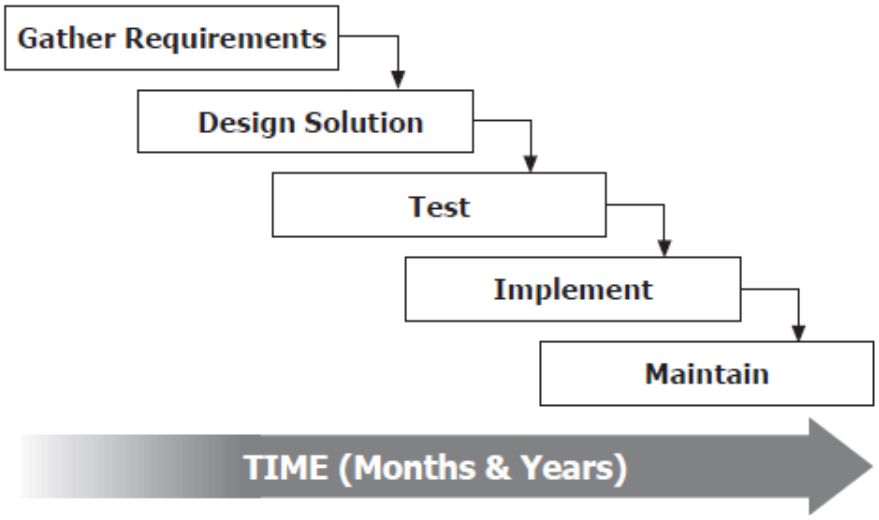


Fig. 7.1 Waterfall project methodology

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.

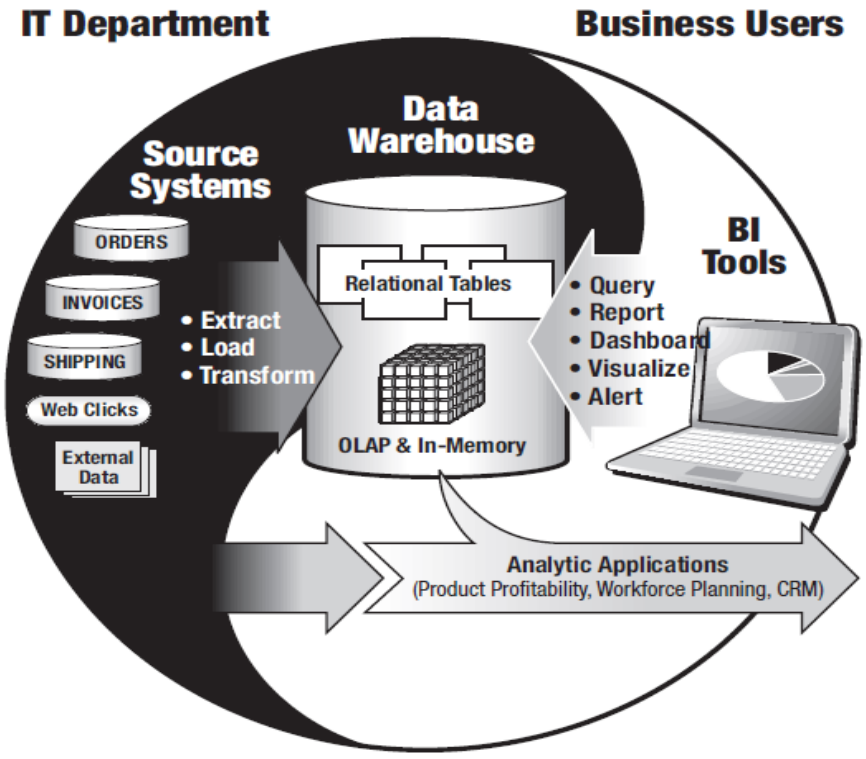


Fig. 7.2 Major components in the business intelligence life cycle

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.

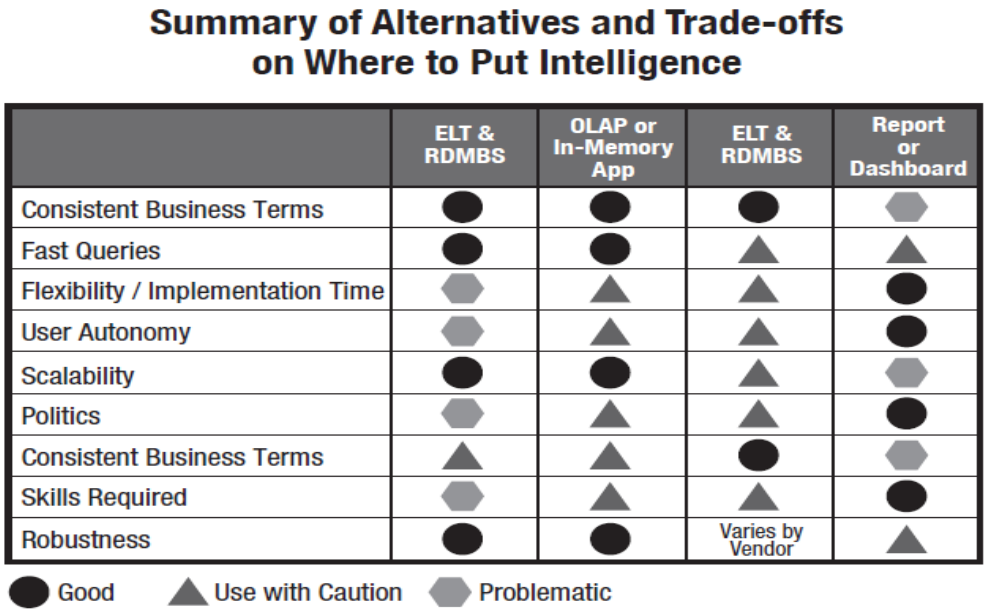


Figure 7-3 Alternatives and trade-offs in where to put the intelligence

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.

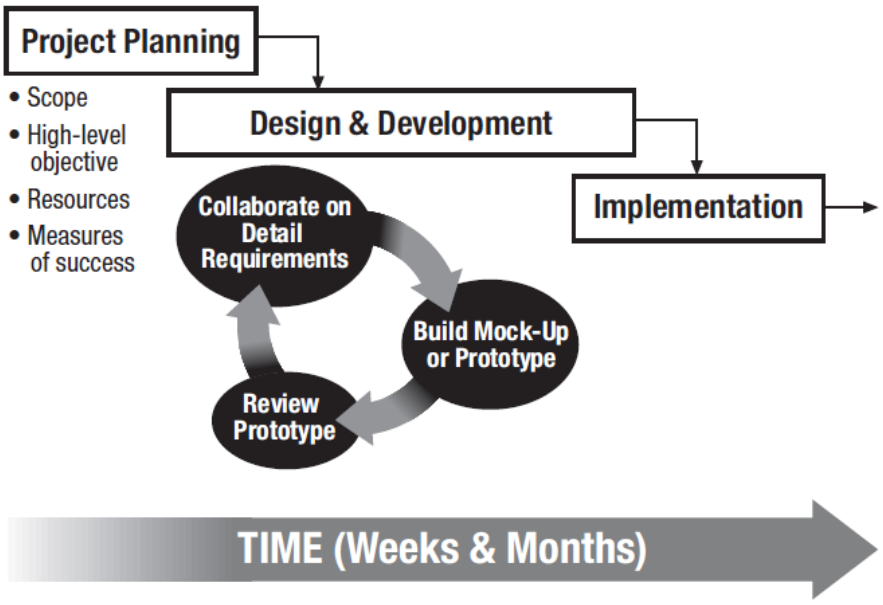


Figure 7-4 Iterative approach to delivering BI capabilities

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.

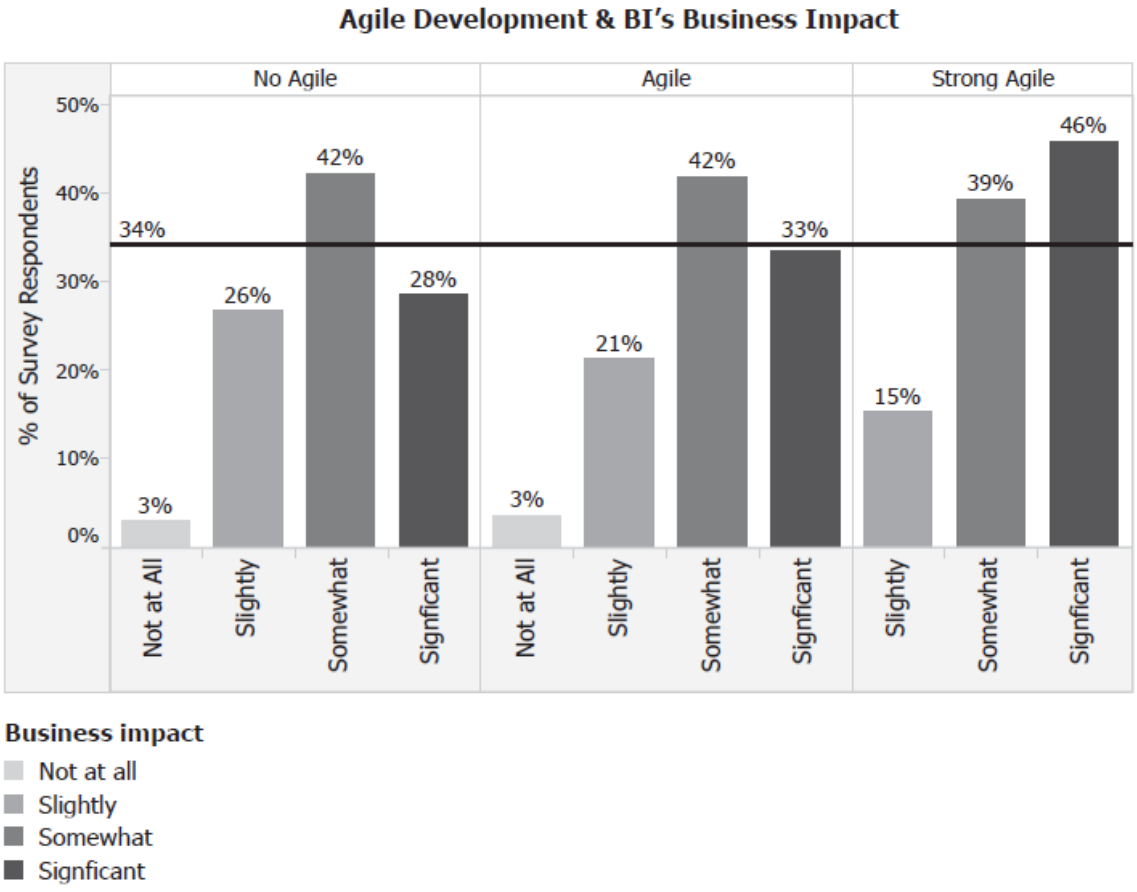


Fig 7.5 Use of agile development relates to greater business impact.

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.

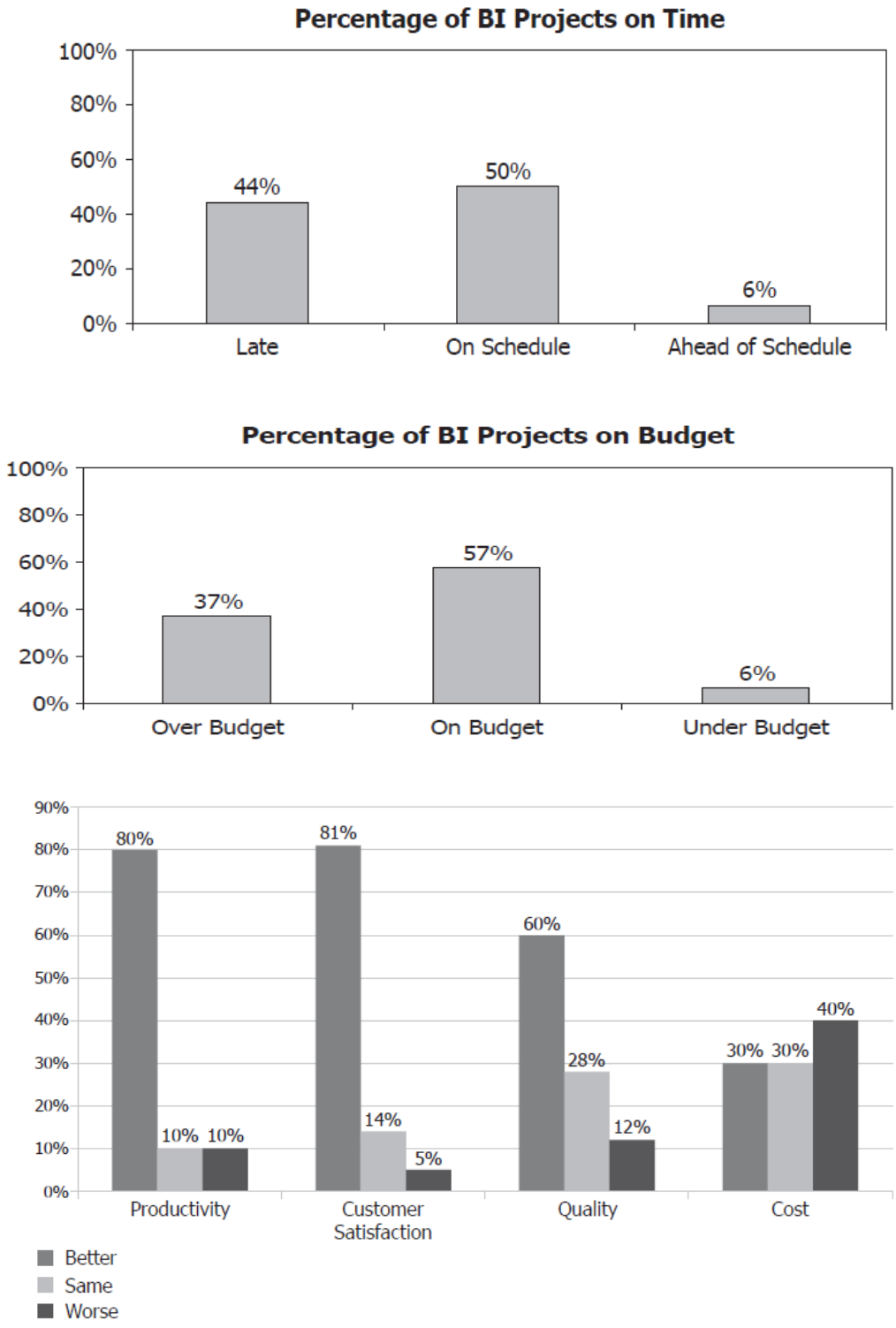


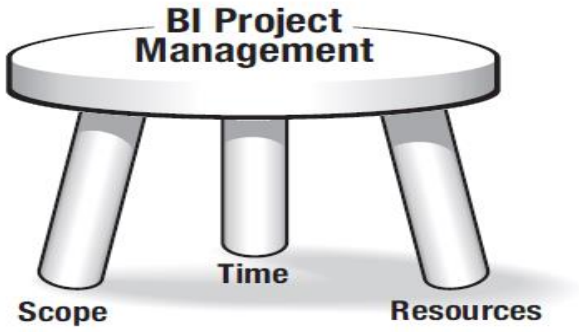
Fig 7.7 Ralph Hughes and TDWI: Agile's impact on BI project key performance indicators

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 centre 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 SELF ASSESSMENT QUESTIONS

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. Swain Scheps - Business Intelligence For Dummies-For Dummies (2008), Wiley Publishing, Inc
2. Carlo Vercellis - Business Intelligence_ Data Mining and Optimization for Decision Making (2009), Wiley Publishing Inc
3. Howson, Cindi - Successful Business Intelligence-McGraw-Hill (2014)

UNIT -8: ADVANCED / EMERGING BI TECHNOLOGIES

Structure

8.0 Objectives

8.1 Catching a Glimpse of Visualization

8.2 Steering the Way with Guided Analysis

8.3 Data Mining

8.4 Other Trends in BI

8.5 Check your progress

8.6 Summary

8.7 Keywords

8.8 Self Assessment Questions

8.9 References

8.0 OBJECTIVES

After studying this unit, you will be able to

- ✓ Make choices with guided analysis
- ✓ Explain data mining
- ✓ See visualization clearly

8.1 CATCHING A GLIMPSE OF VISUALIZATION

The mission of any business intelligence program is to get information to people when and where they need it. But as a secondary task, a BI system also has to make that information usable once it reaches its destination. One of the ways BI software can make information more usable is through visualization techniques.

Visualization means presenting numbers, statistics, metrics, and other facts in a graphical format that makes them easier to understand and interpret. Representing poll results as a pie chart is a simple example of visualization.

8.1.1 BASIC VISUALIZATION

As data-warehousing and querying software grew more powerful and widespread, so did the need for ever-more-complex ways to present the output data. The result was stand-alone reporting software — either separate from or packaged with basic query tools — that could help the user arrange, transform, and present data to audiences in a variety of formats. Reporting software made the information as easy to understand as possible.

As BI was applied throughout the organization, the insights it provided grew in strategic value. Presenting data in a compelling format was no longer a luxury; in fact it became a top priority. Companies turned to tools that could transform their numbers into charts, graphs, and other accessible and understandable representations. The lesson was clear: as important as standard

row-and-column reporting is, presenting data graphically can make the communication of complex data more efficient and more powerful.

8.1.2 WORTH A THOUSAND WORDS

Data has to be understood as having an impact on business. And representing data with charts, graphs, and other images is a powerful way to communicate insights to team members, managers, partners and customers. A concept or trend that may not be dramatic, or even clear, in tabular form, often comes alive in the right graphical format.

But this bit of everyday magic is easy to dismiss as insignificant. By creating a bar graph, all we've done is represent numbers from the table as proportionally sized bars on a graph. But that simple change is powerful: Instead of seeing numerals sitting in a table or on a page — thinking to ourselves, "Self, that number sure is bigger than that other number," and then analyzing the results of our assigning relative sizes to each data point — we can actually *see* those relationships. The bar graph allows us to skip a cognitive obstacle between us and the *meaning* of the numbers.

It's true that, you might well have spotted the sales pattern just by focusing on the 12 numbers in the table. But imagine a table with a *thousand* points of data, or a million. In those cases visualization isn't just a bonus or a shortcut; it's a necessary step to performing meaningful analysis and obtaining business insights.

Just as Excel grew beyond mere grid-style representation to include its well known charts and graphs toolset, BI reporting tools have grown to include basic visualization techniques, similar to those you find in spreadsheets.

8.1.3 OFF THE CHARTS

The charting tool is the core of a visualization tool set. At the most basic level, that means static representations of data points like pie charts where the size of a given "slice" of a disc shows its

relative share of the total amount. BI reporting tools available today include visualization packages — but most of those are still fairly simple. Analysts who need to translate their visually barren reports into compelling stories do have some special tools available that can help them with that job. Just as reporting tools ride atop the rest of the BI stack, visualization tools plug into reporting engines — and can translate data into cool pictures that convey the message about the data much more immediately than the data itself can.

Turning large complex data sets into meaningful images is the domain of advanced visualization tools. Instead of simple charts and graphs, graphics packages allow users to render data into complicated geometric shapes and vector graphics, all in vibrant colors. The goal is to make that information easy to interpret; instead of poring over tables to find profitability hot spots in a company's product line, a visualization tool can create an image that will bring the full profitability picture to life, and put it into context with other business factors.

8.1.4 VISUALIZING TOMORROW

Vendors have to grapple with an unavoidable challenge inherent in visualization techniques: A graphical representation of data must be compelling enough to look at, informative and truthful in its portrayal of the data — without giving the user a severe case of visual overload.

This balancing act is no mean task. As visualization software expands to include charts with multiple layers, drill-through capabilities, and navigation links, they risk becoming just as challenging to comprehend as the report they're attempting to simplify! BI administrators and visualization tool users must be vigilant that users and information consumers aren't getting buried in too much detail.

Nevertheless, graphical representation has given BI a jolt of life by making business insights compelling and convincing. In addition, users in many different jobs are used to graphics-based interfaces — and high-end visualization is the next logical step. Dashboards have become permanent tools for BI family; vendors such as Micro strategy are taking advantage of dashboard space as a place to represent data with visualization. The newest visualization tools offer some fairly jazzy features, especially when you think of BI and processing truckloads of numerical data:

- **Aesthetic appeal:** Vendors have realized that rendering data in a visual format is only useful if the audience is willing to read, view, and digest the information — and to do that, they have to *look* at it first. Making a control on a dashboard beautiful, rather than just giving it bare-bones functionality, helps attract the user's attention in the same way a memo or position paper can do if it's engaging, reader-friendly, and suited to its audience. Most knowledge workers with dashboards on their desktops aren't pilots, engineers, or mathematicians; accuracy by itself isn't enough to make a graphical representation of data useful.
- **Interactivity:** The original dashboard model relied on graphical controls to be read just like any other static report; the data was translated into a chart or graph that could then be interpreted by the reader. The next generation, however, takes advantage of greater computing power and speedier data transformation — and turns a static report into more of a dialogue. This goes beyond simply clicking a dashboard control to see a second, deeper-level control. Imagine putting the mouse pointer over one word on the report and having all the other graphical controls transform or pivot in reaction — can do. The newest controls also allow for rapid toggling and tabbing.
- **Customizable tools:** Vendors can't anticipate everything, so they build programmatic hooks into their tools to allow developers on your team to dip into the toolbox and make the dashboards and controls just right to fit with your system. This approach also allows developers to import and use a wide range of third-party tools.

8.2.1 REALLY COOL, NEXT-GENERATION VISUALIZATION

Quality visualization goes beyond just more slippery sliders and tastier-looking pie charts. BI vendors are trying to incorporate enough visual tools to allow design professionals to turn data into meaningful presentation material. These days most tools present complex, three-dimensional geometric renderings of data, overlaid with traditional visualizations such as bar graphs or pie charts. For example, a tool might render the data points of a simple two dimensional table into elevation points on a smooth 3-D terrain; on top the terrain would be a bar that corresponded to each elevation point and could show another dimension of the data.

Advanced visualization tools are only worthwhile when there is an expert there to create the graphics. That means a BI manager needs analysts who can not only build queries and reports, but also pilot the tools that create the advanced representations. Vendors make every attempt to make the tools easy to use, but to take full advantage of the latest tools that create scientific grade charts, advanced geometric representations, and other next-generation visualizations, you'd better have somebody on the team who can both

- Understand what's going on with the underlying BI process.
- Use the visualization tools that present the information most effectively to a specific audience.

On top of that, visualization requires good data-management practices. Companies typically need visualization tools when they're dealing with massive data sets that resist interpretation by other methods. If you've reached a point where visualization tools make sense, then your BI environment as a whole must be able to move and manipulate gargantuan volumes of information. If it can't do that, it won't be able to support the kind of visualization tools you need.

8.1.5 SPATIAL VISUALIZATION

One of the hottest trends (and latest buzzwords) in business intelligence is presenting data by way of *spatial visualization*. This approach takes advantage of today's mapping technologies to weave business information into maps and other geospatial representations. What you get is an immediate impression of (say) where business processes are taking place and how they compare with each other, as in Figure 8-1. Of course, using space as a data dimension is nothing new. (After all, where else *do* you visualize something but in space?) But its burgeoning integration with business intelligence is a symbiosis that allows companies to represent information about customers, vendors, shipping points, or any other entities that reside in the real world, and whose locations are an important business consideration.

These tools have grown with the advent of ever-more-accurate GPS technology that can create data with a coordinate system in much the same way that transactional data might be stamped with date, marking its "location" in the time dimension.

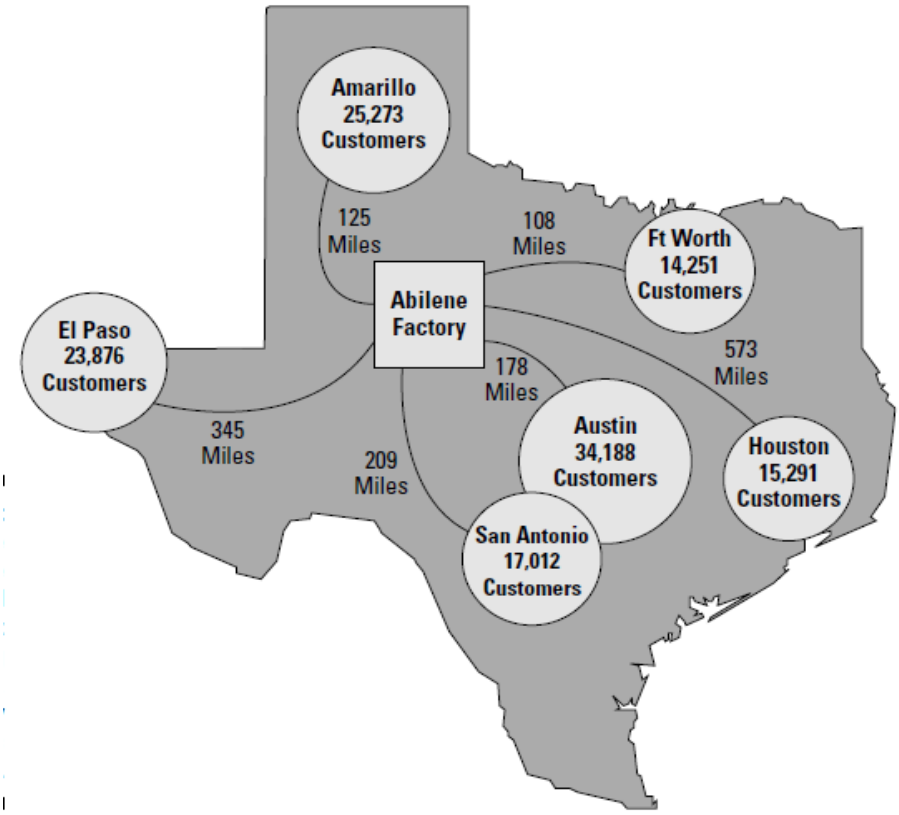


Fig. 8.1 Simple example of spatial visualization: mapping data points by geographic coordinates.

8.2 STEERING THE WAY WITH GUIDED ANALYSIS

Part of the challenge of business intelligence is that powerful tools create so many possibilities. It can all be overwhelming to an analyst. Capabilities keep proliferating — usually way too many for the limited range of problems that need to be solved immediately. Wouldn't it be nice if someone could just show you what to *do* with all these buttons and options and toolbars?

That's where *guided analysis* can help. It's an application that sits atop other BI tools and literally directs the user to take certain follow-on steps that depend on initial results. It's like having a BI tutor sitting next to you saying, "Okay, now you'll want to rerun that same query, except expand it to include the Northwest region." Having trouble visualizing what guided analysis is like? Think of a piece of well-designed tax software. The goal is to complete an overall tax return, but the software helps break up the task into manageable chunks. The application guides the user through a series of steps that build to an ultimate solution.

8.2.1 DANCING THE BI TWO-STEP

The typical BI process is still very much a two-step sequence:

1. The user poses a query of some kind to the system, or perhaps signs up to receive a certain array of information on a regular basis. Then, when the system delivers that information to the user, the second step begins: analysis.
2. The user takes the information set and performs whatever manipulations and data gymnastics might be needed to draw a useful conclusion. For many complex business tasks, this involves performing a series of queries, followed by intermediate steps of data manipulation, and finally producing the necessary answer. Guided analysis is a process that helps users get from point A to point B more efficiently and effectively.

8.2.2 OLD IDEA, NEW MOVES

The more BI spreads to the masses of a company, the more important it is that the end users have some form of guidance as they manipulate data to make decisions. This is particularly true when a department or team makes regular decisions where the variables are well understood, even if those decision processes are long and complicated. Guided analysis is an element of BI where developers configure decision-support tools to offer users direct help during their regular processes.

For example, a bank employee processing a mortgage application might have to enter hundreds of pieces of information about the customer and the loan. That process likely has dozens of branches that are taken depending on what data gets entered. On top of that, the employee might need to access BI systems to view market reports, credit profiles, and other information on the fly.

Guided analysis wraps this complex process into a single “skin”, prompting the user from one screen to the next, delivering contextual information as needed. The basic idea of guided analysis is nothing new; software vendors have always sought to combine computing tasks into sequences to create applications with broader reach. And since BI systems are built to support the decision making process, you’d think guided-analysis techniques would be a natural fit with BI-driven

tasks. But because BI tools were built to be strictly functional, users were expected to design their own workflow to solve their specific problems.

Designers rarely considered wrapping tools into a consistent package that could link insight-gathering processes together for users. With so much data being stored and analyzed in so many different corners of the today's business world, it only makes sense that vendors embrace the guided analysis paradigm. Users in any discipline can combine transactional tasks with analysis tasks — two actions that were traditionally separate. Teams can build playbooks and end-to-end processes around BI tools, making an already-strategic function that much more powerful.

8.2.3 GUIDING LIGHTS

Guided-analysis systems do more than simply add a few help icons to a reporting tool. BI tools with guided-analysis features might have (for example) process wizards that walk users from step to step, or set up event alerts that monitor the system — or the user's activity — and only offer help when it's needed.

When certain trigger conditions are met, the user is alerted and the guided analysis tools step in to direct the user on how to proceed. The trigger may cause any of the following to happen:

- The user is alerted to a problem and given advice on how to resolve it.
- The system automatically runs processes or queries that the user might need to perform the task that's been identified.
- The system leads the user down a known best-practices path, presenting screens and tasks that have been pre-scripted.

Using complex, multistep processes, the guided-analysis engine serves several functions:

- Helps keep users on track and ensures that they've supplied complete information
- Keeps track of the progress of branch tasks that might need to be revisited later

Gives contextual help about progress made toward the solution Result: users don't just work in a vacuum of numbers and spreadsheets; they work in the context of broader business goals. Tight process management combined with established best practices to help with users' workflow and collaboration make guided analysis a useful manifestation of business intelligence capabilities.

8.3 DATA MINING

The amount of information maintained by companies has reached levels that are truly astonishing. Most of the corporate packrat habit can be traced to the rapid advances in storage technologies. Data that used to require an entire storage room full of documents can now be digitally rendered and put on a few square millimeters of magnetic tape or a hard drive platter. There's also a just-in-case attitude that many corporations take: Hang on to your data because in a more litigious society, you never know when you'll need it.

But it's probable that *hope* drives some of the trend in data storage. There are people who dedicate their lives to finding hidden trends in stock prices and economic indicators in the hope that they can transform that information into profit. Top-level executives take the same attitude: All that company data must be harbouring secrets and trends that, if harnessed, could help make the business more successful than ever before.

That's what data mining is all about — examining oceans of past business data to find useful insights about the past that can act as a guide to the future. The twin trends of increased information storage and the steady advance of processing power means that dream may not be out of reach.

8.3.1 DIGGING THROUGH DATA MINING'S PAST

The concept of data mining has been around since the 1950s when the first computers were moving out of entire floors of buildings and into single rooms. As scientists' eyes were being opened to the possibilities of machines that could resolve math problems, they also dreamt about far more complex problems, and even of machines that could find their own problems to solve.
The term *artificial intelligence* was coined in 1956 at Dartmouth. Computer scientists originally saw AI as a way to tackle evolving problems by including feedback loops in code. The idea is simple: When an AI application attempts to solve a problem, it "learns" from its incorrect guesses by noting what variables change — and is programmed to look for why they change.

AI gave birth to so-called *expert systems* (a trend that peaked in the late 1980s) — computer programs that accept inputs and apply to them a set of *heuristics* (a fancy term that computer scientists use to refer to formulas and rules) to produce a result. Programmers typically fed the system a vast amount of past data and worked in a randomness that the system would use as a model for its predictions of future results.

In theory, you turn the power of the computer loose on the data and wait for a solution to emerge. In practice though, it was hard to implement. If they weren't set up correctly, or if they weren't fed the correct data, expert systems turned out to be not so smart. But the idea was sound, and as other information technologies around them improved, the expert systems of the 1980s evolved into ever-more powerful pattern-matching software of the 1990's and the twenty-first century. Data mining came into its own.

8.3.2 DIGGING FOR DATA GOLD

Like "business intelligence," *data mining* is a catchall phrase; it refers to any computational technique that attempts to transform reams of raw data into meaningful business insights. Data mining software often includes advanced pattern-matching algorithms and high-end statistical analyses — anything to help the user draw useful links between the past and the future. No matter what it's called, these are the questions that get answered:

What *really* happened in the past? This is where the mountain of data points comes in. We're not just talking about showing you reports of yesterday's sales. Data mining tools use advanced mathematical models to find patterns and themes in historical data that wouldn't otherwise be evident.

Why did it happen? Knowing *what* happened is valuable — but it's even better if you know what the root causes were. Data mining compares many different variables and looks for subtle correlations over long periods of time.

What is likely to happen in the future? If you can apply the patterns of the past to the current conditions of the company, it might be possible to predict the outcome of certain business activities.

Those three questions lie at the heart of all doctrine, whether it's business, medicine, warfare, or any other discipline. If you can recognize that you're currently following a specific chain of

events that has happened before and led to a certain outcome, it gives you the opportunity to act to improve that outcome this time around.

It's worth saying again in the context of data mining: Don't confuse causality with coincidence. Just because two things happen at roughly the same time doesn't mean that one event caused the other.

8.3.3 DATA MINING TODAY

Data mining has been successfully applied to business problems, especially over the last decade. Some industries (for example, banking and insurance) use it to attach likely outcomes to certain behavioural patterns — which helps determine major business variables such as financial risk.

As data mining techniques have become more refined, it's become a mainstream tool for non-financial businesses as well. More software vendors now view data mining as an essential BI component, and are starting to include it in their core DBMS (database management system) products. For example, Microsoft installed data mining tools in the latest version of SQL Server 2005, allowing you to work all kinds of cool statistical wizardry upon your data. If you know how to take advantage of things like multiple regression and nonparametric analysis, data mining is for you.

If you're using data mining, make sure you don't become the sorcerer's apprentice and unleash magic you can't control. Advanced data-mining and statistical tools are like weapons: Only those well trained in their use and application should be allowed near them. The problem is simple: It's both hard to create meaningful results and (unfortunately) easy to turn perfectly good data into total highfalutin garbage through the miracle of statistics. In spite of the challenges, data mining should be on every BI project manager's roadmap as a critical part of the overall toolkit. Used wisely, it can add tremendous value to the business.

8.4 OTHER TRENDS IN BI

The business intelligence universe is in constant flux. Read an article in one of the many magazines or Web sites covering the field, and you'll likely learn about another hot topic on the minds of BI professionals. Whether it's how to organize a team, the latest in applications and

architectures, or manage ongoing BI processes, keeping up with trends can be a full-time job. Below are a few examples of other broad directions in business intelligence.

8.4.1 BI FOR ONE AND ALL

The consistent trend in business intelligence is to offer the ability to retrieve and identify useful insights "down" the corporate food chain from the great white sharks to plankton. Whether it's high-level executives, senior managers, mid-level managers, or individual contributors, if an employee makes better decisions, it's a good thing for the organization.

8.4.2 UNSTRUCTURED DATA

Oceans of corporate data reside in non-standard formats. Imagine the wealth of information stored in documents, Web pages, or even video. And up until now, that information has been untamed — difficult (if not impossible) to search, sort, and report on.

That's why there's a veritable gold rush to develop tools that reach into a business's nooks and crannies to ferret out hard-to-reach information. This growing element of the BI space is cantered on new search technologies that allow a user to find — and make use of — information that doesn't fit into a known pattern.

The problem with unstructured data that exists in documents — or other formats — is that there's no context to explain how that data is arranged and what it means. As a result, the tools have to be extremely complex to be able to interpret what they're looking at.

For example, Business Objects has a technology called *Data Feed as a Universe* that accepts information from just about anywhere — including Excel files, Web-service feeds, RSS feeds, and any other source for external data. The idea is that users can combine information from varied sources with data warehouse data to create a richer environment for analysis and presentation.

8.5 CHECK YOUR PROGRESS

1. What is visualization?
2. What is guided analysis

3. What is spatial visualization
4. Define data mining
5. What is unsupervised learning?

Answers to Check your progress

1. *Visualization* means presenting numbers, statistics, metrics, and other facts in a graphical format that makes them easier to understand and interpret.

2. Spatial visualization ability or visual-spatial ability is the ability to mentally manipulate 2-dimensional and 3-dimensional figures. It is typically measured with simple cognitive tests and is predictive of user performance with some kinds of user interfaces.

3. interactive, goal-oriented BI that answers a specific set of questions in a structured way, prescribing certain actions based on the user's input and analysis

4. it refers to any computational technique that attempts to transform reams of raw data into meaningful business insights

8.6 SUMMARY

From its humble beginnings, business intelligence continues to evolve today as vendors offer more powerful and innovative solutions. With all those changes, the purpose of BI systems remains the same — timely, accurate, high-value, and actionable insights. The new capabilities all serve those simple goals in one way or another.

Some of the advances happening today are taking current capabilities and making them better, such as more powerful graphic representations of data. But BI innovation goes beyond simply adding bells and whistles to existing applications. The evolution of BI is happening along a broad front with the technology growing more powerful, more meaningful, and capable of putting business insights into the hands of more people.

Transforming data from raw facts into meaningful observations and rules about business operations is not so different from scientific research; every year sees incremental advances, not just in the conclusions reached, but in the research methodologies, tools, and technologies that the researchers rely on. The pace of change in BI is steady; occasional minor breakthroughs in how environments are built and used become next year's best practices and top-selling tools.

Of course, in a BI environment, *innovative*, *imaginative*, and *powerful* are all relative terms. Vendors are always rolling out new tools, but there's no guarantee that the newest thing can help you with your business problems; that's for you to evaluate, and for your team to make a reality.

Some of the areas where research is expanding rapidly include the following general categories:

- **Visualization:** using advanced graphics to make critical business insights clearer and more meaningful
- **Guided analysis:** interactive, goal-oriented BI that answers a specific set of questions in a structured way, prescribing certain actions based on the user's input and analysis
- **Data mining:** finding *the* needle you really need . . . in a huge stack of needles
- **Unstructured data tools:** turning irregular data into business insights with search and indexing capabilities for hard-to-quantify information formats

Implementations tend to lag BI innovations. That's a nice way of saying that companies rarely want to be the first on their block to try an untested technology. A new tool that blows away the current paradigm will take time to find its way into immediate widespread use — and it takes even longer to make everyone's "best practices" list.

Of course, a tool or feature that is so common in today's BI environments started as little more than a gleam in the collective eye of a team of software engineers in the design shop of a particularly innovative vendor. When it gained popularity in the marketplace other vendors would have copied that success by building their own version of that feature. Before long, it would become a must-have for any BI system. That process is still going on today as vendors continually seek to gain market share by building a better mouse trap; it's a continuing cycle of development in the BI world.

8.7 KEYWORDS

- *Visualization*- means presenting numbers, statistics, metrics, and other facts in a graphical format
- Spatial visualization - Spatial visualization is the study of two- and three-dimensional objects and the practice of mental manipulation of objects

- Unstructured data - is information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy,
- Structured data - is the data which conforms to a data model, has a well define structure, follows a consistent order and can be easily accessed and used by a person or a computer program. Structured data is usually stored in well-defined schemas such as Databases.
- 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.

8.8 SELF ASSESSMENT QUESTIONS

1. Explain the importance of data mining in BI.
2. With the help of an example, explain guided analysis.
3. Write some of the important features of visualization tools in BI.

8.9 REFERENCES

1. Swain Scheps - Business Intelligence For Dummies-For Dummies (2008), Wiley Publishing, Inc
2. Carlo Vercellis - Business Intelligence_ Data Mining and Optimization for Decision Making (2009), Wiley Publishing Inc
3. Howson, Cindi - Successful Business Intelligence-McGraw-Hill (2014)

BLOCK INTRODUCTION

In this block we talk about just what kind of BI tool you want to build, some of the key choices you have in front of you, and some tips on making good decisions in the early stages. . They are easily as important as the technology factors. Unit 11 drills farther into business intelligence strategy. Before you start making specific plans for your solution, however, take a quick look at certain realities that will become very important during the implementation. In unit 12, you put the knowledge gained in previous units into action and begin to build the project roadmap. As with any map that helps you on a journey, your BI roadmap is just a general guide to help you get from point A to point B. you have to find the best route between them.

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

Unit 9- The BI Big Picture:

So Many Methodologies - So Little Time, Customizing BI for Your Needs, Implementing BI: Get 'er Done

Unit 10- Human Factors in BI Implementations:

Star Techie: Skills Profile of a Core BI Team, Overruling Objections from the Court of User Opinion, Major in Competence

Unit 11- Taking a Closer Look at BI Strategy:

The Big Picture, Your Current BI Capabilities, Exploring "Should-Be" BI Alternatives, Deciding "Could-Be" Alternatives, Making your choice, Deciding on your strategy, Getting the necessary buy-in

Unit 12- : Building a Solid BI Architecture and Roadmap:

What a Roadmap Is (and Isn't), Centralized Versus Decentralized Architecture, BI Architecture Alternatives, Developing a Phased, Incremental BI Roadmap.

UNIT -9: THE BI BIG PICTURE

Structure

9.0 Objectives

9.1 Methodology of BI

9.2 Customizing BI for company Needs

9.3 Implementing BI

9.4 Check your progress

9.5 Summary

9.6 Keywords

9.7 Self Assessment Questions

9.8 References

9.0 OBJECTIVES

After studying this unit, you will be able to

- ✓ Choose the best BI software vendor for you
- ✓ Judge which methodology is correct for your situation
- ✓ Analyse all the options
- ✓ Take that first critical step

9.1 METHODOLOGY OF BI

When you take on the challenge of selecting a BI methodology, you're evaluating not only what you can afford to do with the technology budget; you're also thinking about what your company can actually do within the allotted timeframe, and why you're doing BI in the first place.

9.1.1 STARTING AT THE BEGINNING

What's your BI implementation going to accomplish? Have you settled on a precise scope and timeline for the project yet? If you haven't, then you should consider doing so before looking any farther down the road. Once the strategy for the project's been laid down, consider some road mapping exercises where you match the project goals with some concrete steps on how to achieve them. A roadmap is like a Gantt chart on a project plan. the results and ending of one phase become the inputs and starting whistle for the next phase. Working through the critical stages of the project gives you early perspective on the challenges of timing and coordinating activities, availability of interconnected resources, and the kinds of decisions that have to be made — now and in the future.

Every methodology you see will probably look easy on paper. And not just easy to do, but easy on the eyes as well, what with all the richly coloured diagrams of technology stacks and process flows where circles and boxes get magically connected by big, thick arrows. It's unlikely that a single, out-of-the box methodology pitched by one vendor or consultant will suit all your needs. There will inevitably be unexpected twists and turns in the implementation, and your organization will have nooks and crannies that won't be covered by a canned solution.

That's actually okay. It's to be expected that no one methodology will be exactly what you need. In the end, the best BI implementations are hodgepodges of ideas, best practices, and even

vendors. In the end, your technology solution will likely be a collage of software and processes that attach perfectly to each piece of your problem.

9.1.2. THE EXCEPTION TO THE RULE: MICRO-BI

It's possible to get away with a single-vendor, single-methodology BI solution if the scope of your project is narrow enough, and the needs are perfectly defined. A good example is when your BI problem concerns a single area of the business, like Shipping or Human Resources. Single-facet BI challenges can often be met with an off-the-shelf solution from an established vendor. In the case of HR, you'll find capable tools like the one from Knowledge Workers Inc. that can be installed and used right out of the box. Knowledge Workers is a specialty vendor that's been around for a long time providing a tightly defined solution with a well-established methodology to match. Companies shouldn't feel obligated to make their solution any more complicated than it needs to be.

For situations like this, where you are using a single vendor to solve a problem of limited scope, it's advisable to not seek out ways to custom fit the methodology to your company's contours. In fact, in these situations your best bet is to follow the vendor's step-by-step instructions as closely as possible. If your BI installation involves creating a solution for a single department and a minimal collection of data, it's possible to follow the recipe exactly. But sometimes things get more complicated than you originally planned — and you have to consider customizing or expanding your BI solution. Here are a few telltale signs that an off-the-shelf solution won't cut it:

- It goes beyond the one department you had in mind.
- It requires anything beyond static reports or light OLAP.
- It uses resources shared by other systems.

At this point, it's good to know when to throw the single methodology out the window and light out on your own path.

There is a difference between a BI methodology and a canned project plan. There's a temptation to confuse the two, but they encompass divergent ideas. A project plan only discusses steps and resources, while a methodology encompasses motives, vendors, and a strategic outlook. You

might be able to get by with just a project plan on a small-scale project. But for true BI solutions, you have to consider full methodologies if you want the job done right.

9.2 CUSTOMIZING BI FOR YOUR NEEDS

Differs from the one at the company down the street is a mystery that you need to unravel one step at a time. You should begin by taking a look at what your company's doing already to develop business insights. Then you can move on to more formal reviews of existing resources, software and hardware components, and needs. And if you do a good job, you might
just find yourself on the yellow brick road.

9.2.1 YOUR NOT-SO-CLEAN SLATE

The ideal situation for any BI project team would be to build a solution from the ground up. The landscape would consist only of operational data sources, with no tradition of information aggregation, few established reporting standards, and no analytics to speak of. With no pre-existing conditions to distort your design, each system component and process could be built to meet the business needs to a tee.

In the real world, there's no such thing as a clean slate. Every company has developed some form of decision-support apparatus, whether it's a Ouija board in the snack room or a full-fledged technology-based system. And like it or not, you'll have to consider what's there today before you can think about tomorrow.

Even if the words business intelligence has never crossed the minds of any of the managers or executives, it's a sure bet that reports get created and routed to an established schedule and with agreed-upon standards. There are processes for applying lessons learned and operational information to decision-making. Whatever it may be, someone at every company is attempting to find, use, and distribute business insights, and you must account for it in your process.

- **What is it?** Make an inventory of existing methods being used to deliver business insights — whether it's spreadsheets on desktops, old mainframe applications, or departmental data marts.

- **How does it overlap with your planned BI scope?** Concern yourself with how your planned system will either use or replace existing resources, but don't stop there. Think ahead, too — about what might happen when you make future upgrades to your BI implementation.
- **How effective is it?** This is the most challenging part: assessing whether anything in your current process is really worth keeping. Some elements of the existing system probably are worth keeping. Just because a technology is old doesn't automatically mean it's bad. Your job is to search for kernels of good, reusable business intelligence, if they exist. There may be reports or processes that are perfectly fine the way they exist today. If that's the case, be careful before you mess with them. Taking a step backward won't win you many friends in the user community, and it will take up precious time.

9.2.2 INITIAL ACTIVITIES

The early phases of the BI project consist of a set of evaluations on the organization's needs and an assessment of the company's current BI readiness, both in terms of technology and culture. With this information in hand, you're ready to start developing more concrete plans that will drive your project through to its conclusion.

Assessing your current BI state Does your company have a good solution in place that supports decision making? If so, what's it made of and how effective is it? You'll find parts of it (or maybe the whole thing) to be good, bad, or just plain ugly.

If you understand the general information needs that are driving the BI solution, then assessing the current systems' effectiveness is really just a process of comparing what's supposed to be delivered with what actually is being delivered. In most cases, the company's information needs are not being met. (Otherwise, why would you be installing a BI system in the first place, right?)

To mitigate the problems, you need to understand their sources. it's best to start at the natural flow of data and work your way toward the user After you identify the operational data sources, you need to evaluate their readiness to be integrated in a BI solution. Is the data entered accurately to begin with? Is it structured in such a way that makes it accessible down the line?

If your organization is merging data in a central repository, you need to assess how well that process is working. Does the quality of the end data meet the standards you need for your reporting and analysis systems? You should judge whether the current data-warehousing environment — if it exists — can handle the tasks you have in store for it. Finally, there's the reporting environment. It's safe to say that just about every company has an existing reporting environment. During the assessment phase, you need to evaluate whether that system can create the kinds of reports that will be commonplace in the BI solution — and whether the system can distribute the information in a way that meets the new standards.

Developing a sound BI strategy

It's best to develop your BI strategy in parallel with — or directly after — your assessment of the current BI state. Once you've figured out what your company can do now, it's time to really begin locking in the specific reports and functions that you expect the BI solution to handle. This is the time to begin talking about who will be using the reports and other applications — why they'll be using them, and what they hope to get from them. You also have to make judgments about who will be setting the rules, where the administrative responsibilities fall, and how you expect to keep the data secure.

There's a tendency to go straight from the assessment of current BI capabilities into evaluating software vendors, specific technology approaches, and products. But be cautious. Selecting a vendor inevitably narrows your solution options. One common pitfall in BI projects is a tendency to lock the project mindset in to one solution or another too early in the overall process. It's tempting to do because it gives the team a tangible starting point and some specific direction, but it also invests you in certain applications and protocols that may not (in the final analysis) be best for your organization.

Developing your BI architecture

At this point, you've got an idea of your ultimate destination, you've got a map to tell you where to go, and now it's time to plan the vehicle that will get you there. How well you assemble this plan will have a grave effect on the success or failure of the project. Piecing together the components of the solution is where you begin to see your BI vision start to take on a real form. When you're confident that you understand who needs what information and in what form, you can begin thinking about the kinds of systems that will meet those requirements.

The end result of this stage is a document (or set of documents) spelling out the specific requirements of the project — from an overall business perspective and in terms of a project-specific technology. Whether you label that document "Architecture" or "Systems Requirements" doesn't really matter. The point is what's in it:

- An itemized list of the components and major sub-components of your system
- Detailed functional requirements for system components
- Information about the data that flows through the system A good architecture document usually includes information on the following areas:

Source data: You should inventory the general domains of information that will be handled by the BI system, such as finance, human resources, or sales data. There's also the matter of assessing the current state of that information, as well as the databases and storage systems that house it.

Extraction, Transformation, and Loading (ETL): With the sources of data identified, a huge part of the architecture document will be devoted to how that data is moved to the central repository and made available for queries and reports. The architecture document will also cover data cleansing, and lay out thresholds for minimum data quality.

Data warehouse: The architecture document will include decisions about the final makeup of the dimensions and metrics tables, the metadata, the normalization mix in each table, and the business rules.

User tools: This will be a description of the functions available to the end users and administrators in the form of queries, reports, and analysis tools. Include descriptions of how precisely users will be able to manipulate the information they find, whether it's (say) drill-through capabilities or more advanced analysis.

Other sections of the architecture document will cover logistical concerns such as governance, administration, security. It's imperative to know where and how decisions get made during the ongoing use of the BI tool. Architecture documents are typically long on narrative and short on convoluted diagrams. But there should be some basic boxes-and-arrows artwork included to help readers picture the end results.

In the end, the architecture documentation is not just about the components and logistics of the system, but it will tell you how everything actually fits together. In complex systems such as business intelligence application suites, the individual pieces are also complicated; how those pieces should interconnect and interact isn't always intuitive.

Could-be versus should-be alternatives

Like any complex system, a business intelligence implementation boils down to making some key choices. Your team will have to make calls on some tough decisions throughout the design phase. your best bet is to look at more than just the first alternative that surfaces. These multiple choices are the could-be alternatives. What makes it harder is the fact that no single prescription can meet all of your enterprise's BI needs.

There certainly may be a solution that's preferable to the others, but as long as you have examined several could-be choices, you can just make your best call and move forward. A perfect example of many could-be solutions is in selecting your overarching BI business and technical architecture. If you want to do an enterprise-scale implementation, you have at least two ways to go about it:

- You could make your system highly centralized around a hub data warehouse.
- You could create a more distributed architecture, setting up departmental data marts.

Either alternative is a viable choice, but one or the other will make more sense for your company — depending on your available resources, the company culture, and how the business is organized. But it's up to you to weigh those factors and make the best possible call.

Seek outside expertise when you're confronted with a tough call that you worry might come back to haunt you down the road. Even better, you don't have to call a consultant or a vendor to get in touch with somebody who's been there before: There are multiple user communities open to BI professionals there you can get in touch with people who might have just the right piece of advice.

Selecting BI products and technologies With the strategy in place and the architecture plotted, it's finally time to begin speaking to vendors. Because you're fully prepared — with your assessments, documentation, and general knowledge about how you expect to integrate the

BI application into the company — you're in great shape to evaluate alternatives among applications.

Picking the right vendor

One approach to selecting the right product for your system is to focus first on finding a vendor the matches what you need, and then digging into their offerings to decide which pieces you need. If you want to evaluate vendors, you'd first assemble a list of companies that offer everything your BI project needs in general — and then narrow the list by examining criteria such as the following:

- Approach to license pricing: site (where use is limited to locations or servers) versus seat (where use is restricted to a certain number of people), and concurrent versus individual users (which concerns whether licenses are interchangeable between people).
- Availability of technical support and willingness to answer questions before you've issued a purchase order for a million bucks worth of licenses.
- Vendor stability and longevity.
- Product maturity and reputation in the marketplace.

Picking the right product

If you're simply going to go product by product and make evaluations, the questions get more granular, and focus on the specific capabilities of the tools each vendor offers:

- Suitable cost of ownership, including initial license costs as well as fees for ongoing support, training, and upgrades
- Compatibility with existing systems
- Response time and processing speeds
- Usability and ease of use that matches the profile of the user community in your organization
- Data-handling capabilities that fit with your variety of source-data platforms
- Customization capabilities so your BI developers can create tailor-made applications and reporting

9.3 IMPLEMENTING BI

We've taken a high-level look at all of the preparatory steps necessary in any business intelligence project. Your team has done assessments, made a detailed plan, and selected vendors and software platforms. Now it's time to start installing the system. For the professional who's comfortable with managing standard software implementation projects, the analysis phases are complete at this point

- now you're ready for the detailed design of the main components that will make up the system:
- the database design of the software infrastructure that will power the data warehouse the metadata repository
- The ETL process design that will wrangle the herds of data from the pastures and drive 'em home

When you get to this point, the theorizing is over. Now it's time to turn the general principles into specifics and start powering up pieces of the system one at a time until the BI solution is up and running.

9.3.1 ZEROING IN ON A TECHNICAL DESIGN

The assessments and high-level strategy and architecture documents will point the way to the technical design itself. This is the heart of the system, where there are no more abstractions, no more generalities. The technical design includes precise data definitions and user-interface designs. The central decisions surrounding the design of the data warehouse and data marts include the level of data granularity, working out how the fact tables and metrics will be constructed, the level of summarization of the data, which tables get normalized (and to what degree), and so forth.

This same technical design process continues for all pieces of the BI solution. That includes the user facing-tools — in most cases the querying and reporting applications plus any analytics software. Standard UI design processes apply here, just as they would with any application.

User interface design is a key factor in whether a BI solution is a success or not. That means it's a good idea to use tried and true methods for assessing usability — these, for example:

- Mock-ups and wireframes to ensure basic form-by-form or screen-by screen usability
- Cases and other UML tools to ensure that the system's navigation and activity flow makes sense in the context of the system's business purposes

UML is a great overall modelling tool for any kind of software system. For companies that buy packaged solutions, much of this work is already done; standard forms can be customized to suit specific needs. For companies building their own front-end applications, this is the moment when it's time to work out how information is presented to the user and which controls are in place.

In other words, the BI implementation team already knew what needed to be done; now it's actually working through the process of how to accomplish these tasks. This stage in the implementation will be dominated by the developers, data architects, database administrators, and other techies. But make sure the team isn't stuck on a virtual technology island, with no connection to the business "mainland". It's always important to maintain buy-in from every business organization affected by the technical design. This is where it becomes important to identify power users in the key business functions who can help the tech team stay grounded in the business strategy as they make technology side decisions.

9.3.2 PUTTING TOGETHER THE BI PROJECT PLAN

The project plan ties the many tasks of the technical design together, accounting for resources and dependencies. The project plan is at once the schedule for the BI implementation as well as a detailed inventory of the remaining steps, and a running assessment of the resources that are available to the BI team.

Standard tools are fine

There's nothing special about a BI project plan relative to other technology implementation project plans. Any software that offers the standard project planning, reporting, and display tools for tasks and resources will usually be fine. Microsoft Project is the most commonly used application for building and maintaining project plans, but there are certainly other capable offerings on the market.

Policing the process

As with any large-scale implementation, you need to maintain an adequate level of control over the quality of the technical development. It's best to build in such features as inspections, walk-through sequences, and a thorough quality-assurance and testing program to ferret out any bugs and defects.

Finishing the job

At the end of the line, when all the plans have been made and adhered to, when the development is complete, and the processes have all passed a white-glove inspection the project is finally done. Or is it? There's more to it than just building the solution and flipping the switch. Once the BI system is working the company has to be ready to actually use it — and benefit from the results.

As you get closer to the end of a long implementation, it's not uncommon for the troops to start grumbling. With the light at the end of the tunnel growing brighter, it's tempting to speed up, take shortcuts, or deviate from the plan. Without warning quality drops on the final phases of development. And even if you keep your team motivated and on target, sometimes the late-game pressure comes from the outside. Prospective users, managers, and others begin agitating to see the fruits of your labour. They'll grow more and more anxious to finally light the fuse and start taking advantage of the system before it's completely ready. They figure, just because the dashboard application hasn't gotten out of the quality-assurance testing phase doesn't mean we can't start querying the data warehouse, right? Wrong. There's nothing wrong with planning a phased rollout, but once the plan is set, stick to it. Hold the line at all costs. If you roll out the BI soufflé half baked, you risk it falling flat before anyone's been able to enjoy it. Part of the project plan should include time for training classes for the user community. The training strategy deserves just as much attention as other parts of the high-level plan, because it is one more area that can make or break the BI project. Here's a sample of the issues you have to wrestle to the ground:

- Will you have technical people train the user groups? Or will you hold train-the-trainer sessions and trust that the first round of learners can pass on the right information to their teams?
- How much material will you provide for the training? Will it be different for different user groups of varying skill levels?

- Will there be ongoing education as the system evolves?

These are not trivial questions. And they'll crop up again from time to time.

9.4 CHECK YOUR PROGRESS

1. What is A Gantt chart?
2. What is UML?
3. Give three examples of ETL tools.
4. Give three examples of BI tools.
5. What is supervised learning?
7. What is unsupervised learning?

Answers to Check your progress

1. A Gantt chart is a type of bar chart that illustrates a project schedule
2. The Unified Modeling Language is a general-purpose, developmental, modeling language in the field of software engineering that is intended to provide a standard way to visualize the design of a system
3. Informatica Power Centre, SAP Data Services, Talend Open Studio & Integration Suite

   Informatica Power Centre, SAP Data Services, Talend Open Studio & Integration Suite
4. Predictive modeling, data mining and contextual dashboards or KPIs
5. Supervised learning is a machine learning method in which models are trained using labeled data.
6. Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets.

9.5 SUMMARY

In this unit we understood careful first steps to take in a project. you are faced with a thousand decisions early on — including selecting a scope for your initial implementation, which vendor(s) to choose, and how to staff your team. But the consequences of rushing headlong into the project are too far-reaching and expensive to do anything but the best planning beforehand.

In this unit, we talked about just what kind of BI tool you want to build, some of the key choices you have in front of you, and some tips on making good decisions in the early stages.

9.6 KEYWORDS

- ETL - Extraction, Transformation, and Loading
- Data warehouse - In computing, a data warehouse, also known as an enterprise data warehouse, is a system used for reporting and data analysis and is considered a core component of business intelligence

9.7 SELF ASSESSMENT QUESTIONS

1. Write a note on how to Customize BI for Your Needs
2. Explain the early phases of the BI project
3. Explain Systems Requirements document
4. Explain how to pick the right BI vendor
5. Describe the steps involved in BI project plan in detail.

9.8 REFERENCES

1. Swain Scheps - Business Intelligence For Dummies-For Dummies (2008), Wiley Publishing, Inc
2. Carlo Vercellis - Business Intelligence_ Data Mining and Optimization for Decision Making (2009), Wiley Publishing Inc
3. Howson, Cindi - Successful Business Intelligence-McGraw-Hill (2014)

UNIT -10: HUMAN FACTORS IN BI IMPLEMENTATIONS

Structure

10.0 Objectives

10.1 Star Techie

10.2 Overruling Objections from the Court of User Opinion

10.3 Major in Competence

10.4 Check your progress

10.5 Summary

10.6 Keywords

10.7 Self Assessment Questions

10.8 References

10.0 OBJECTIVES

After studying this unit, you will be able to

- ✓ Find the right skills for your team
- ✓ Win over reluctant users
- ✓ Overcome resistance to change
- ✓ Run a BI competency centre

10.1 STAR TECHIE

Business intelligence projects are bridge-building exercises like no other. To guarantee success, the implementation team must develop linkages between the business goals and operations of the company and the elements of the IT infrastructure.

10.1.1 KEY PERFORMERS

Some members of the team will work on the BI project full time, while others will be *on loan* from other teams, or will supply expertise on an as-needed basis when the implementation touches their particular area.

- **Project manager (PM)**: This person will be the linchpin role for the entire implementation. The PM is tasked with oversight of the project, meaning they will establish the initial plan of attack, coordinate the necessary resources, and do whatever it takes to adhere to the schedule. A BI project manager should have the requisite skills inherent in good PMs; they should have a good grasp of both the business and technology side of things. They must understand how the BI project dovetails with the company's business objectives. The PM must be able to exercise all the key "soft" skills such as negotiation, mediation, and mentoring. It's often said that a good business intelligence PM has a sixth sense: able to *see* project dead-ends and pitfalls in time to avoid them, and able to *feel* when it's time to deviate from the plan to keep the project moving.
- **Business Analyst (BA)**: The BI business analyst shares the same basic qualities as counterparts working in other technology fields, with a few notable exceptions. All BAs are (in effect) translators who move between the technical staff and the business teams, enabling communication in both directions. Because they must create a common platform

for communication, BAs, like PMs, must have a good understanding of the underlying BI technology as well as a solid grasp of the business goals. BAs operating in a BI environment must have a particular handle on how data moves — not only within a company's operational systems but between those same systems — to succeed. Armed with that knowledge, their main objective is to help the user community get the reports and application functionality they need. For later phases of the project, they must also be able to grasp the complexity of analytical tools, especially when the requirements aren't as cut-and-dried as those that apply to simple transactions and reporting.

- **Data architects and designers (DAs):** These folks work at the lowest levels of the data itself — designing data models, database structure, and information flows through the various elements of a BI solution. They also make decisions about which methods to use for key processes. For instance, ETL (extract, transform, load — the essential processes of data warehousing that are responsible for moving data from source databases into the warehouse) can be performed in multiple combinations of pre-built tools and home-grown code; the DAs must recommend a combination that works best for the project. They must also work with the rest of the team on how and when to make new data available in the data warehouse. The developers may have designed the front-end tool based on (say) 30-minute data-refresh intervals, but the architect may determine that such a requirement is impossible to meet.

This is one position where the BI version of the role is quite different from the *normal* version. Unlike a *transactional* data architect, the *BI* data architect must understand dimensional modeling and be able to design a platform for multidimensional analysis. The data architect must be prepared to work closely with the business analysts to keep the data model in synch with the business model. As with many positions in the BI world, the DAs have to know just when to break certain rules, deviate from best practices, and mix in their own brands of solutions as they design the data model.

- **Data-quality analyst (DQA):** When a system relies on a data warehouse, a data-quality analyst in the project is a must. The DQA is there to assess the fitness of the data that courses through the operational and transactional systems for use in the data warehouse. The DQA has a strong hand in the ETL process, making the call on which cleansing routines must be used on data from each source before it's transferred into the warehouse.

The rules about how and when to implement the multidimensional data model (as opposed to a relational data model) are no longer as hard and fast as they used to be. Some experts will carry a copy of their favourite database-design textbook around and refer to it reverently. Most database designs follow one of two traditional directions:

- Relational databases are built for storing and accessing information.
- Multidimensional models are tailor-made for analysis.

But the lines are blurring; more systems aspire to do both these days.

10.1.2 YOUR OTHER TECHIES

The following folks don't have lead roles, but their support is important to the project nonetheless.

Front-line IT folks

This is the crew that actually wrestles the machines, programs, and data into submission:

- **BI infrastructure architect:** The person in this role oversees the technical foundation of the project, ensures that all the software parts work together, and ensures that the hardware is in place to handle the load.
- **Application lead developer:** This is a programmer who is responsible for assembling the front end tools. They are likely tasked with cobbling together the various querying, reporting, and analytics environments into a smooth, usable application.
- **Database administrator (DBA):** The database administrator's job is to take the logical model handed down by the data architect and turn it into an efficient physical model — selecting database hardware and software, and assembling the foundations of the data-handling software that will be accessed by other areas of the application.
- **Quality-assurance analyst:** In any technology environment, the testing phase is the first of many moments of truth. In a BI implementation, you're looking not just for classic script-testers and bug-finders, but also (in fact, especially) for people who will work to actively challenge the environment you've created. Once it's gone through the paces set by an experienced BI testing team, you can be confident that the system is rugged enough to handle any level of user and abuser.

Getting help from the IT-free

People in the organization who aren't necessarily technology professionals still have an important role to play in your BI implementation:

- **Users:** Yes, these people are effectively part of your team (even if they don't always want to be!). You need to identify and classify them in order to build the application to suit their needs. Remember how very different users can be across different business units and functions — and yet, if you can find the common aspects of what they need from the BI tool, your job is that much easier.
- **Subject-matter experts (SMEs):** It's important not just to get information and opinion from the end-users, but also to identify experts in the fields you're touching directly. For example, say you're building a BI tool that will serve the supply-chain team. Just because you speak with a few individual users doesn't necessarily mean you're getting the big picture. In fact, you can be led astray if you don't identify SMEs who can show you the 30,000-foot view of the company's supply-chain operations, the data that gets exchanged with the manufacturers and suppliers, what metrics are used to judge success inside the organization, and so on. Your team will inevitably develop a small army of SMEs to rely on again and again to resolve development dilemmas such as selecting from two competing business priorities. An SME can be anyone with full knowledge of that specific business function and how it fits into the overall business model. Ideally your SMEs will be people of considerable experience, but you don't necessarily want to aim for the most senior people; they're prone to mix opinion in with fact. Instead of a snapshot of reality from 30,000 feet, you could easily get an oil-on-canvas picture of how your SME would like it to be.

10.2 OVERRULING OBJECTIONS FROM THE COURT OF USER OPINION

Team leaders, mid-level managers, and people of all levels will seek to exact a toll of some kind from you in exchange for their cooperation in the BI implementation. After winning the fight to get the BI solution implemented, after duking it out over budget and resources, the final test awaits you: the user community.

Even if you do a superior job in every other respect, there will always be people who touch your implementation that will express mild doubt, scepticism, or downright antagonism.

The "other" half of the people battle is winning over the user community, as well as areas of the business that you must pass through.

10.2.1 CHANGES

People naturally fear change. For worker bees in a department that's about to get a shiny new BI application, it means learning a new skill. It opens them up to someone being "better" than they are — that is, losing their place in the pecking order. They might fall out of line for that promotion they were bucking for.

Just lay off, okay?

Worst of all, people will often fear for their jobs. So as you lead a BI implementation, you need to be able to reassure all those affected by the implementation that their jobs are absolutely, positively, *not* on the line. Except there's one little problem: Sometimes they're right. Their jobs may *be* on the line. Knowing that better decisions will result from the BI implementation is cold comfort for those who lose their jobs in the process; they won't be around to experience the fruits of those better decisions. It's an unfortunate side effect of working with *any* program that has the power to transform a company.

At the risk of sounding melodramatic, companies who don't adapt get left behind and eventually go out of business. Your company must be as efficient and smart as its competitors or *everyone* is out of a job. So (if you're looking for a little rationalization) think of BI as a way of saving jobs, not chopping them.

Cut through the buzz

Business intelligence is a powerful and far-reaching technology concept that has the power to transform organizations, both in action and in structure — in ways that can frighten the rank and file. BI solutions are often implemented right along with organizational restructuring initiatives. Frequently those involve putting processes under a microscope through programs such as *TQM, ISO 9000,* or *Six Sigma*— all of which are beyond the scope of this book but will produce millions of hits when run through your favourite search engine. Such efforts often go hand in hand with re-jiggering teams, moving employees around on the organizational chart, and (yes) downsizing.

Turn and face the strange

Being aware of the ramifications of a BI implementation is an important first step in combating the problems that can arise. Fear, uncertainty, and doubt will spawn resistance to the program through the ranks of the teams touched by the BI implementation. But there are some proven techniques and actions that can mitigate the drag-effect on your BI project:

- **Get sponsored:** Here's where top-down clout comes in handy. A BI implementation is far more likely to succeed if it carries the weight of a corner-office mandate with it. Anyone whose title starts with *Chief* and ends in *Officer* will do, because when they say something *will* happen, one of the main effects is that critics are immediately silenced. Your company's Chief Executive Officer is a case in point; you need that person's buy-in from the beginning. So show the CXO how the BI project will have an impact on things he or she cares about. Then keep that person in the loop — invitations to the kick-off meeting, the launch party, and as many points in-between as a busy exec is willing to attend. Run big decisions by the higher-ups you get on your side — both to get the benefit their expertise *and* to reinforce buy-in as they contribute what they know. Best of all, if you can generate enough enthusiasm for your project that your exec allies chat about it in the executive washrooms, you gain momentum
- **Employ champions:** In this case, *champions* are any managers or influential users who are willing not merely to set aside their antagonism toward the project, but to sing its praises. Long-time employees are especially good in this role because they not only hold the respect of their co-workers, but they also operate within networks of friends and colleagues that extend far and wide throughout the company. A few well-placed champions can generate more positive buzz about your BI implementation than a full-fledged marketing campaign.
- **Convert heretics:** Is there a particular user group or team whose participation is essential to the functioning of the entire program? Then focus your best sales effort on them — especially potential doubters — and bring them into the tent. If you involve sceptics early on in the process, ask their advice, and take their suggestions whenever possible, you're more likely to tamp down their negativity.
- **Accentuate the positive:** Simple positive reinforcement goes a long way for the implementation team, the extended technology resources, and the user community as

well. While you don't want to oversell BI capabilities, it's important to remind the relevant players of the end result, the magnitude of the anticipated improvement, and the level of value it will create within the company. It's what politicians sometimes refer to as *the vision thing:* where you give voice to a version of future events that people in the trenches sometimes lose sight of while dodging everyday bullets. When they're reminded of how great it's going to be (rather than being threatened or cajoled), they're more motivated to do their part to create that positive outcome for the company.

- **Do what works:** The wheel's been invented already, so there's no need to sit down at a drafting table with a picture of a manhole cover and a pencil. The biggest favour a BI project leader can do is simply to adhere to best practices. In almost every area of BI — from data management and integration to application design, training, and support — a standard has already been set. The lessons are all there: incremental rollouts, highest-value-first, and so on. Lean on that combined wisdom provided by the myriad experts who have gone before you down this path.
- **Archive knowledge:** BI implementations are not one-shot deals; they're designed to be a long-term transformative force inside companies. The systems will inevitably evolve over time as business priorities change, new technologies become available, and the people involved in the project rotate off or out of the company altogether. Preserving the lessons and institutional knowledge that accumulates over the life of a BI implementation is essential to keeping it efficient and relevant. Whether you create a competency centre, invest in knowledge management software, or find some other solution, responsible stewardship over the BI archives ensures the project's survival.

10.3 MAJOR IN COMPETENCE

Just building the BI solution can often be a multimillion-dollar enterprise that takes dozens of months to accomplish. But the enormity of that challenge is nothing compared to what it takes to actually *maintain* the BI system over several years. Setting and maintaining the ongoing business intelligence strategy is a role that often has no obvious home within an organization. And yet, it's a must for companies that want to protect their BI investment and get the most from that complicated system moving forward. A common trend in the BI world is to set up a permanent organization tasked with maintaining the company's BI effort.

Enter the BI Centre of Excellence (BICOE) — or (as it's also known) the Business Intelligence Competency Centre (BICC).

The two terms essentially mean the same thing. While there are no hard and fast statistics to prove it, *BICC* seems more commonly used so we use that here to talk about the organization in general — but if you have a choice, go with *BICOE* over BICC. The word *competence* has a connotation of bare minimum proficiency. Even though a BI Competency Centre will go well beyond mere proficiency, it sounds so *average*. But if you put *excellence* in the title, right away you have something everyone wants to identify with.

Find your centre

The purpose of the BICOE/BICC — okay, the BI *centre* by which ever name — is to act as a permanent body whose sole focus is to address every aspect of BI throughout the organization, from establishing standards and priorities, to driving the overall BI strategy.

BICCs don't issue orders or hand down mandates that must be followed. They make formal recommendations to the appropriate executive and management teams that actually govern the company. Even so, that advice — coming from a team of experts and representative users — typically carries a lot of weight.

Their evaluation of a vendor (for example) can make or break the relationship. If the BICC centre deems one project a success or failure, it can impact the future of the individuals involved. And the BICC will give a thumbs-up or thumbs-down to every major BI activity being considered — from new installations to upgrades to changes in the architecture. And it's all accomplished from one central organizational node.

If you think that a BI centre might be overkill in your organization, keep in mind that part of its purpose is *coordination*. BI activities often involve so many different players and teams that it's almost impossible to make a move without a central committee that gathers everyone into one place to hash through issues. And remember: There may be more than one BI effort going on at once. A BICC ensures that the multiple data warehouses and dozens of data-mart environments all follow the same standards and protocols.

Organization-based competency centres

In this model, the BICC acts as a cross-functional committee, filled with representatives from every relevant business unit and division that has a hand in the BI process. The committee members manage tasks like managing relationships, both internal and external, such as

establishing lines of communication with the legal department and bodies that govern the company as well as vendors and BI organizations. There are also sub-groups concerned with establishing common protocols and standard processes in the technology environment, as well as guiding principles for how projects are run.

Like competency, the word committee also has some negative connotations and/or bureaucratic baggage. But the BICC doesn't have to fall into the standard committee traps of doing too much or too little. As long as the mission and agenda of the BICC is clear, and all key players are required to participate, it has a good chance for success.

Competency centres — budgeting

A common mistake committed by companies is to create a BI centre without adjusting the mix between primary job responsibilities and committee responsibilities. If representatives from around the company are to take an active role in the BI centre, it must have people who can take the time to participate without jeopardizing their jobs. But at the same time, the BICC can't become a home to full-time committee members with no other responsibilities.

In the former case — the "PTA model" — the Competency Centre is made up of BI specialists who are essentially volunteers. Management doesn't officially make their work on the BICC a part of their performance plan. That leaves the BI centre devoid of a steady source of energy and influence as members drift in and out of the organization.

On the other hand, creating a fully budgeted BI centre — where members have no primary responsibility other than the work of the committee — creates another set of problems altogether. When members have no tasks in the day-to-day functioning of the BI system, they're likely to become divorced from the truth of what's working, what isn't, and where the company's BI initiative should go from there.

A BI centre that's just right

The middle ground is the safest place to be when it comes to forming a BI centre, somewhere between the two extremes mentioned earlier. Membership should be mostly voluntary, but those who serve on the committee should be compensated for doing so by their originating organization. That allows the members to stay in touch with the BI strategy without becoming so far removed from their primary jobs that they lose unique professional perspective on BI's direction in the company.

It’s also beneficial because there is no question of member loyalty. Everyone on the committee has a known dual allegiance: to their primary organization as well as to the company’s BI strategy. This balance provides a natural and positive tension that means actions won’t be rash or hasty, and BI conclusions will only be reached by active compromise where the company’s needs are placed ahead of all others.

Raising standards

The simple act of setting standards in a business intelligence environment can be daunting when you consider the number of source databases that the effort may have to draw from — and the sheer volume of reports that may be created. Sure, decisions can be made in the field, but that makes coordination and integration harder. Competency centres step in and apply best practices when it’s time to set standards throughout the company. Even something as mundane as column widths or report layout spacing isn’t out of reach for a BI Competency Centre, although such niggling attention to detail is rare. But when it comes to big-picture decisions — such as deciding on ETL timing and data-refresh rates for data warehouses throughout the company — standards are necessary to making the system run. The BI Competency Centre provides a mechanism for making such decisions.

10.4 CHECK YOUR PROGRESS

1. Name the full time roles in BI project.
2. List the front line IT roles in a BI project
3. Name the non IT people who play important role in BI implementation
4. What is the role of BI infrastructure architect?
5. What is the role of Application lead developer?.
6. What is supervised learning?
7. What is unsupervised learning?

Answers to Check your progress

1. Project Manger, Business analyst, data architect and designers, data quality analyst
2. BI infrastructure architect, Application lead developer, Database administrator, Quality-assurance analyst
3. Users, Subject matter experts

4. The person in this role oversees the technical foundation of the project, ensures that all the software parts work together, and ensures that the hardware is in place to handle the load.
5. This is a programmer who is responsible for assembling the front end tools.

10.5 SUMMARY

Every successful business intelligence implementation, no matter the size and scope, must address how the project is affected by human factors. You can assemble the best possible plans and purchase high-grade software and infrastructure components, but when it comes down to it, you'd better have the right people in place, or the whole thing could come crashing down around you.

You can't just have brainiest and geeks in your team. You need some *people skills.* That means you need internal salesmen, facilitators, negotiators, and diplomats. And in some cases, you need many of those skills packaged in one single person.

Companies are microcosmic societies, where egos, rivalries, and prejudgments are in constant motion — usually in the form of individuals and groups united for one cause or another. The business intelligence system will be a two-way street. It will have to draw on the resources of the community of experts, users, allies, and champions, and it will need to provide benefits, direct and indirect, to those same groups. This unit explains importance of people elements in your project. They are easily as important as the technology factors.

10.6 KEYWORDS

- BI manager - The role of the BI manager is crucial to successfully delivering BI Projects and managing stakeholder expectations
- business analyst - A business analyst is a person who analyzes and documents the market environment, processes, or systems of businesses
- Data architects and designers (DAs) - These people work at the lowest levels of the data itself — designing data models, database structure, and information flows through the various elements of a BI solution.
- Data-quality analyst (DQA): - The DQA is there to assess the fitness of the data that courses through the operational and transactional systems for use in the data warehouse

10.7 SELF ASSESSMENT QUESTIONS

1. Briefly Explain the role of core BI team members.
2. Explain the contribution of non-core team members of BI project
3. Explain How non IT people can contribute to the BI project.
4. Write some proven techniques and actions that can mitigate the drag-effect on your BI project
5. Write a note on importance BI centre.

10.8 REFERENCES

1. Swain Scheps - Business Intelligence For Dummies-For Dummies (2008), Wiley Publishing, Inc
2. Carlo Vercellis - Business Intelligence_ Data Mining and Optimization for Decision Making (2009), Wiley Publishing Inc
3. Howson, Cindi - Successful Business Intelligence-McGraw-Hill (2014)

UNIT -11: THE BI BIG PICTURE

Structure

11.0 Objectives

11.1 The Big Picture

11.2 Your Current BI Capabilities

11.3 Exploring “Should-Be” BI Alternatives

11.4 Deciding “Could-Be” Alternatives

11.5 Making your choice

11.6 Deciding on your strategy

11.7 Getting the necessary buy-in

11.8 Check your progress

11.9 Summary

11.10 Self Assessment Questions

11.11 References

11.0 OBJECTIVES

After studying this unit, you will be able to

- ✓ Analyse capabilities
- ✓ Look at potential "should be" states
- ✓ Dream about the "could be" possibilities
- ✓ Decide on your BI strategy

11.1 THE BIG PICTURE

The first part of your second look at BI strategy (as shown in Figure 11-1) includes a more detailed examination of the BI capabilities you already have (versus desired capabilities). The goal here is to collect information about your organization's current BI capabilities, assess its current BI needs, and chart a way forward to the next step: creating a roadmap, project plan, and requirements documentation. From there, you can begin building the project.

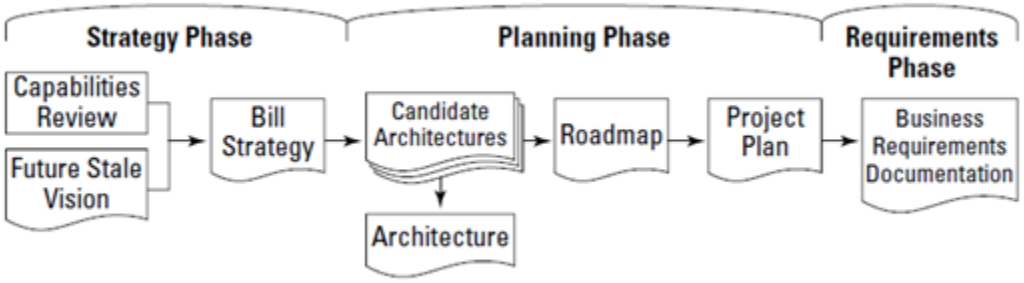


Fig.11.1 BI strategy: gathering ideas and information to shape implementation decisions

11.2 YOUR CURRENT BI CAPABILITIES

Traditional strategic planning begins with a (usually long) analysis-and planning phase: Leadership assesses a company's current situation, defines the finish line far over the horizon, and then goes through the long laborious process of developing the route to get from here to there. A BI implementation is no different: To know what's needed, you have to understand what's currently in place.

How you approach this assessment depends largely on the shape of the organization, and the scope of your planned offering:

- If the focus is going to be on a single business function (for example, customer-relationship management), you might look at how that function affects the entire organization: How does the sales department track customer data versus the service delivery group? How does the billing department use customer information, and which pieces of data do they use? You'll need to understand the ramifications of a focused BI solution everywhere it plays a role in the company.
- If your BI project is going to involve a comprehensive overhaul of a single business unit or department's processes, you'll start with that department or division, and then move outward to all the connection points throughout the organization.

Assessing your business infrastructure

Before you move into an assessment of the technology that's currently in place throughout the company that supports business intelligence capabilities, you'll want to understand the existing processes that govern the operation of your business. You're not yet at the point of gathering data on specific requirements. This is still preliminary discovery work to give you a good view of how a BI solution would fit. By stepping back to look at the big picture, you can work out the right questions to ask while you're identifying requirements. For now, it's about listening to individual contributors, middle management, and executives as they tell you how they go about normal work processes — as of today. Here's an overview of what to look for:

- **Business functions:** These are the broad categories of an organization's operational activities. When you make a map of them, what you get is a cousin of the organizational chart, which diagrams the divisions and teams of the company in a hierarchy of activity. For example, your company might subdivide the business into divisions based on the markets they serve, or perhaps base the divisions on different product lines. Within each of those divisions there might be a sales team, a sales support staff, a marketing arm, and a shared-services entity (further subdivided into accounting, finance, human resources, and so on). Chances are the company's functional breakdown has already been documented in some form; most people who've worked in an organization have a general understanding of how the company's large pieces fit together. For areas of the company that might be affected by the BI implementation, you'll want to create a pretty granular

functional map. Don't just list (for example) "Human Resources"; you'll want to do some detective work on how the HR department is assembled, and then list its major parts (such as the Benefits, Recruiting, Training, and Employee Relations teams).

- **Operational processes:** This category describes how things actually get done within each business function listed on the map you created earlier. This is a specific, step-by-step map of the major activities performed every day by each team. For example, in the payroll team under the accounting department, you'll build a list of the basic steps, roles, and interactions for the day-to-day functions like adding and deleting an employee from payroll, or changing employee withholding information. You're not looking to map out every single step the workers on a team take; there should be no entries in your notes that say, "Employee pulls out chair. Employee sits in chair. Employee sips coffee. Employee opens Report Wizz application by clicking the Run ReportWizz button. Employee waits 8 seconds." You want detail, but don't go overboard. You're trying to get down to the point where you understand at a high level only what data individuals use to perform their key everyday tasks, how those employees are using data to make their critical decisions, where they get their data from. You don't need to a play-by-play of every second of their day. The questions you'll ask will cover the following broad areas:
- **What are your current pain points?** In other words, what processes or applications don't serve your function well, and seem to be without an obvious or easy fix?
- **Are there any opportunities being missed because the data, reporting, or analysis infrastructure aren't working well?** What would be needed to take full advantage of those opportunities?
- **What are the positive and negative aspects of the tools and processes you work with every day to perform your job?** Does everyone in the department feel this way? Can that view be backed up with specific instances or evidence?

Business intelligence has to produce insights that are timely, accurate, high value, and actionable. Anything less than that and you're probably talking about a failed process or a tool that's not doing its job. That definition should never be far from you at any time during your evaluation process. It should inform your detective work and guide you as you gather information about quality of the business processes and their specific contributions to how the

business functions. In most cases, when employees deem a process or tool to be inadequate, they can identify the absence of at least one of the four key characteristics of BI.

As you discover more about how your company's current processes work, you also build an understanding of general policies and business rules. You're learning what people are doing, but along the way you also build a picture of why they're doing it that way. You figure out cause-and-effect relationships between the elements of the processes, along with business rules and team policies that affect business operations.

Mid-level managers are perfectly capable of expounding on the processes within their team and how their team interrelates with other teams. You'll want to understand how things are supposed to work as well as how things actually work. But ask a fundamental question about a business issue as a whole, and the answers start to vary widely. For example, you might ask a basic question like, "How do you calculate the profitability of a particular sales account?" or "Who are our main competitors?" and two different managers might have perspectives that are completely at odds with each other. Part of the task ahead of you, as you assess current capabilities, is to find "the truth" about how the business actually functions. That means finding agreement across divisions and teams about business rules and procedures, and sometimes even forcing agreement between opposing parties who have been doing things differently. It may actually come down to putting two managers into one room and letting them hash it out. That means you're going to interview lots of people, including representatives from the current end-user community, those performing analysis work, those making decisions based on the analysis, and just about everyone in-between. (It can be a long process. But you knew that.)

You may uncover some uncomfortable facts about business operations — say, a disagreement between managers over what constitutes the correct business process, or the correct data definition or rule. This is where your executive sponsor can go to work for you. Businesses are like families. When there's a dispute among siblings, it's up to a parent to resolve it. And when there's a dispute among managers at a company over how the business works, your task is to move up the organizational chart until you find the appropriate executive with enough power over all parties to get an agreement. Don't look at disputes about business rules as a bad thing necessarily; it's an opportunity for business intelligence to add value.

Now that your team has a clear picture of business policies and processes, as well as a solid understanding of how information is used within and between departments, you can move on to

the technology assessment, another important step on your way to developing a sound BI strategy. During this phase, you're attaching the processes described to you earlier by the managers and individual workers, and associating them with the technology tools currently in place.

Assessing the technology stack, top to bottom

The technology assessment proceeds similar to the non-technology assessment. At first you're involved in a simple discovery process where you gather vital statistics and basic information without pausing to perform a lot of analysis. Later you'll want to align what you found out with the project goals. Looking at the technology stack, here are the key points for each level:

- **Infrastructure:** This category starts with hardware; we're talking mostly about PCs, servers, and networking gear. But it goes beyond that and includes some low-level foundation software as well. The key variables to nail down are these:

  • What hardware platforms are currently in place in the departments in question — what is their life-cycle and are there any planned upgrades or changes to the company's approach to hardware?

  • What type of network are we using? Is the proper network in place to carry an adequate quantity of data between the appropriate business units? Are all departments connected in the way(s) they need to be?

  • Is enough server horsepower available to handle the kind of reporting, analysis, visualization, or advanced statistical tools that we're looking to install?

  • Are all potential end-users properly equipped to run the BI tools and applications? For example, if you're building a supply chain and inventory application — designed for your managers to access as they walk the warehouse floor — it is essential to know whether your managers carry personal mobile devices or tablet PCs that can run the application.

- **Security:** Because a BI initiative often involves moving large quantities of data (whether in raw form or as reports), you need to feel comfortable that the network and PCs are properly protected. That means understanding everything from data encryption on the network backbones to basic user management.

- **Information management:** Any software that has to do with the storage and manipulation of data is covered in this technology layer. For a BI project, this layer gets as much scrutiny during your current-state assessment as any other:

  • Is the company standardized on one database-management system (DBMS) or one single vendor? If not, what's each division using?

  • Are there any compatibility limitations with the DBMS used throughout the company?

  • Where does the relevant operational and transactional data reside right now?

  • How many different versions or views of each key data dimension are in use?

  For example, if you're planning to implement enterprise BI for the field sales team, it's important to know whether every entity in the company defines customer data in the same way, or whether you're dealing with islands of data, each with its own definition.

  • Is there a unified data-stewardship council that maintains the enterprise data model for companywide applications?

- **Application and user interface layer:** Applications can include middleware and other software that constitutes the foundation of the business technology environment, housing business logic, security, and communication functions. The user interface consists of any tools that stand between knowledge workers and the company's computing environment.

  • Is there an Enterprise Resource Planning system in place? For that matter, are there any other enterprise-wide tools that handle data that may be important to the BI implementation?

  • What software tools does the prospective user community currently use to perform its analysis tasks? How are queries built?

  • What is the current state of reporting in the organization? How are they produced and distributed, and when does that happen?

In the same way that you mapped out the business functions and operational processes, you're going to want to do an application inventory. This is essentially a list of all the major software being used by the teams that can potentially be affected by the BI implementation. A good application inventory should include

- a high-level view of the role each application plays in the delivery and manipulation of information
- an understanding of the users' subjective view of its effectiveness
- the relationship with the software vendor, sales and service points-of contact, and an overview of any contractual agreement (such as the license status)

Keep the good stuff

It's always possible that you're starting from a pure-green-fields situation, where nothing in the way of business intelligence is happening in your organization yet. But that's not likely. In most situations, companies have some existing elements that should be accounted for. You'll eventually have to evaluate whether it's best to use them in the BI implementation or simply leave them alone.

You're looking at both processes and software. And it's important you keep them separate in your evaluation. Sometimes the software tool being used is not adequate and you'll want to replace it. But that doesn't automatically mean the underlying process (or set of business rules) surrounding that BI function is no good. And the opposite may be true — good tools, bad rules.

The goal is to put an effective BI solution into place, but in all likelihood you've got a limited budget and limited time. So if you have software or processes in place that cover functions you've targeted for change during your business intelligence initiative, take advantage of it.

Improvement projects

As you assess the efficacy of a specific BI element, pay attention to the conversations you've had with the managers and workers in each department. Their opinion of what's working well and what isn't is your starting point — but it shouldn't be the final word. Evaluate their opinions in light of the overall goals of the project, keeping the future state (and your BI goals) firmly in mind.

In many cases, employees are so focused on their daily tasks that they miss the forest for the trees. In the case of a BI infrastructure assessment, it means they don't recognize that their current process could either be drastically improved with a new system (likely to a degree they can't envision), or that their current process will be inadequate for the future state of the team. In these cases, the user gives the tool, report, or process a thumbs-up, but the BI project team decides it must go anyway. To make sure that team stays on board with the implementation, be prepared to make your case to the folks who really like the old tools.

Beware the sacred cow. Some applications, protocols, or processes may be inextricably ingrained into the company's operational framework, not because it's the best-in-class, but because of politics. It happens all the time. Maybe the CEO's brother works for a software vendor whose product is doing a poor job supporting the call centre reps. Good luck upgrading your call centre in that situation! Make your recommendation, but do so knowing you may have to back down. Before you take on any sacred cow, be aware that there will be plenty of battles ahead, so pick the ones that count, and make sure the brouhaha is well worth it.

Hidden technology gems

At this stage, there's another task your team can do: Keep an eye open for functionality that already exists in the company's technology infrastructure, but which (for whatever reason) hasn't been utilized to its fullest. This can happen in several different ways. Sometimes an application that's been rolled out across the organization has a feature set that employees haven't been trained on. Rather than buying new software and installing it, you might be able to get by with developing a training class and getting your people to take full advantage of assets that are already in place.

In other cases, there is an application being used by one team or department that can be extended across the company. For example, if the company is looking at rolling out an advanced dashboard tool to all mid-level managers and above, it may be possible to take advantage of the fact that the vice presidents in the Finance organization have been using a dashboard for years. If the tool has been successful in its limited role, maybe it's time to make it the star of the show.

Think of the advantages of expanding an existing installation:

- The IT guys are already familiar with supporting it.
- You have a history with the vendor.
- You have some record or the software's performance.

Don't assume that an old or ugly process should be chucked out just because there's a replacement technology that's slicker and newer. Sometimes whizbangnew technology isn't the answer to every single problem in the world. (What a concept.) In some cases, the $100,000 application simply doesn't add enough value to a manual process to make it worthwhile.

Throw out the bad stuff

Of course, there's plenty of technology that can't be redeemed, no matter how hard you try. Since your organization is considering a BI initiative to begin with, in all likelihood there's a good deal of it that's either ineffective or doesn't exist at all. Figure 11-2 shows a typical relationship of processes to keep, remove, and add.

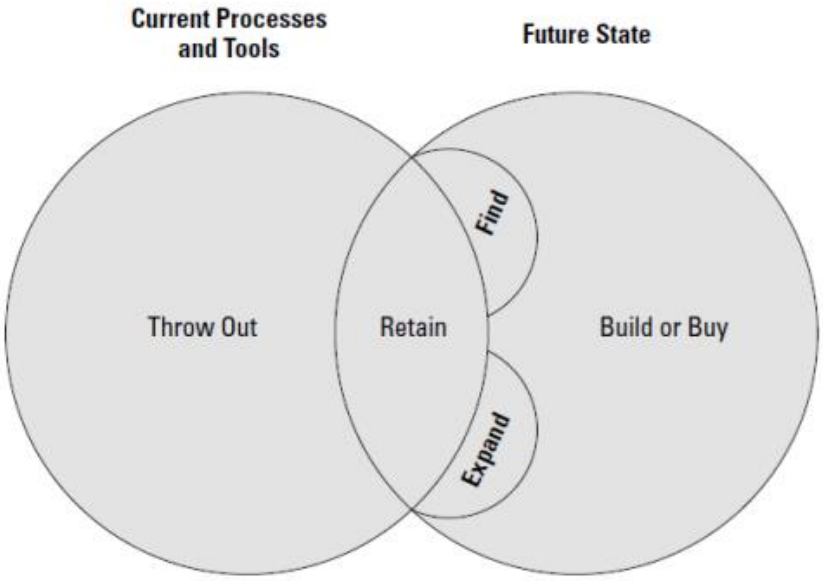


Fig 11-2 An infrastructure assessment's goal is to evaluate what's working and what isn't

The process of doing an infrastructure assessment is to evaluate how much of what you currently have is getting the job done. The future state of your BI tools and processes will be a combination of elements from the old days — whether used as-is, expanded from their original limited roles, or found to be underutilized — and (of course) lots of brand new elements.

11.3 EXPLORING "SHOULD-BE" BI ALTERNATIVES

Lots of folks you encounter during this process will have ideas about what it takes to fix BI-related processes. And at the same time, you'll be developing opinions of your own about where the company has gone wrong, and what tools are needed to make it right. These ideas are your should-be states; they're rough ideas about the direction BI should follow. They may involve changing processes, changing tools, altering corporate standards or policies, adjusting the way

departments interact, tweaking data definitions, and any of a hundred other ways complex processes can be taken apart and put back together.

Our goal is to find the best path between the way it is today, and the way it could be. The best path might be the shortest path, but often it isn't. Sometimes it's best to walk downstream until you find a bridge to cross; you may be tempted to wade right in, but you never know whether there are alligators lurking.

Utopian BI

After the assessment process is complete, an excellent exercise is to suspend reality for a moment. Remove the Earthly time and budget constraints and build the perfect imaginary team to go along with the highly cooperative user base.

So, what's it going to look like? Think of these areas and what your dream scenario entails:

- Changes to the technology infrastructure
  - • Hardware
  - • Networking
  - • Storage
  - • Personal computing devices
- Data management
  - • Transactional database systems
  - • Reporting systems
  - • Changes to data definition
- Changes to the business infrastructure
  - • Re-arrange organizational chart
  - • Adjust corporate policies
  - • New business processes

This is an important exercise because it gives you a baseline of sorts. Having a far shore to aim for is a good way to instill your team — and the rest of the stakeholders in the BI project — with a sense of common purpose. Plus, it helps to start thinking about what you want and why you want it. That's the first step to doing some real quantitative analysis on what needs to be included in the BI release, and what can be left out.

Don't skim over this step. There are some genuinely important lessons that come from this exercise that you can really take advantage of down the road. So don't just think about what you

want in an ideal world; think about why you want it, what the direct benefits would be, and what successes would be spawned by its implementation.

Coming back to reality: examining barriers to achieving your desired future state

You caught a glimpse of what your BI dream-state is, but now it's time to come back to Earth. Hopefully you'll descend filled with optimism and a sense of wonder at what might be. Truth to tell, most of us mere mortals aren't likely ever to reach that far shore. With the utopian BI vision planted firmly in your still dream-fogged head, slowly come out of your trance and begin thinking about what stands between you and that alternate BI universe. Set money and time aside for the moment and address other factors that play a role in technology implementations. What are the barriers that prevent you from getting to each item on the list? For starters, take a look at these thorny factors

- **Human:** The company doesn't have the right people in place to make the necessary changes.
- **Methodology:** The company doesn't foster an appropriate amount of cooperation, communication, or some other quality that's necessary to make it work.
- **Process:** The company doesn't follow sound or universal processes to achieve its strategic objectives.
- **Technology:** The company's technology environment is woefully inadequate to handle the load of such a system.
- **Political:** There are too many forces resistant to changes to think that such a system would even be possible.

After utopia, of course, this exercise can get a little depressing. But what makes it useful is that you'll start to identify potential obstacles to the BI initiative. You'll want to determine what the biggest barriers are, as well as which ones can be overcome, and what effort will need to be expended to do so.

From this stage of the analysis comes the diagnosis that can get you past the barriers — a game plan for dealing with negative forces while you try to build the system.

Deciding "Could-Be" Alternatives

In case you were wondering, the "correct" solution isn't likely to jump out at you. It's rarely the case that a single, obvious path presents itself (or even exists). For now, it's your job to winnow

the choices that don't fit for whatever reason, keeping those that do for later consideration.
Judging viability

In this stage of the process, you want to ask yourself if the alternatives swirling around your mind are even possible. Focus on the current state of the company; Here's a quick list of questions:

- Does the company have the technical capability to build, integrate, and support this approach?
- Does the user base have the appropriate skills to take full advantage of the tools that would be offered to them?
- Does sufficient budget exist to make this level of investment?
- Does this element or solution conform to existing corporate standards and policies?

If the answer to any of these questions is "no," you've identified an option that should be set aside. The hour is getting late, and this is the time to consider only those options that could actually get past all the usual hurdles and be implemented.

If every one of your alternatives produced a "no" answer, it's time to go back to the drawing board and create a smaller, less invasive set of BI options. Consider this to have been your feasibility study and the final verdict came down that your initial approach was infeasible.

Identifying risks and also how to mitigate those risks Slowly but surely, you're eliminating the impossible strategy elements from your list of candidates. What's left are the possible solutions (and pieces of solutions). That doesn't mean every choice remaining is practical or a version of the right choice, but you're getting warmer. The next step is to identify potential risks with the remaining choices. Before you go too far down the road with one of the alternatives, it's important to understand what risk exposure — and how much — you'd be subjecting your team and the company to.

Risks to BI projects come in all shapes and sizes. Some can be identified; others are latent, and won't show themselves until they've damaged your initiative.The first step to mitigating risk is to identify as many possible (realistic) risks as you can. Start with general categories and work your way to more precision as you go:

- **Data risks:** How stable is the data? How much clean up will be required?

- **Application risks:** Are we familiar with the tools? Are they reliable and easily integrated?
- **Organizational risks:** Are the users ready to accept this initiative? Will other teams cooperate with the implementation?
- **Financial:** Is this project fully funded? What could cause it to break the budget?

It's impossible to overcome every possible risk.

Gauging business value

Just as every strategy alternative has risks, each one also comes with rewards. The question is, how much of each? In the same way you identified the things that could go wrong, now it's time to talk about what could go right. Two principles help here:

- **Remember the Big Four.** The whole purpose of the BI initiative is to deliver a solution that provides business value in the form of insights that are timely, accurate, high-value, and actionable. But each will deliver its own unique combination of benefits and business value. As you did with risk, you have to create a matrix or list of business value attached to each potential strategy, along with some means of quantifying. And as with risk, you'll need to consider how likely the benefits are to be delivered, and how significant those benefits really are to the business.
- **Business value is not always tangible.** Improving a familiar metric — like wait time in a call centre, or profit margins — shows compelling value. But it's not the whole story; sometimes business value is indirect, hard to measure, or both. It's hard to quantify, but make sure you're considering more than just easily-measured benefits. For example, the BI initiative includes finally doing away with that clunky old application in the shipping department. Presumably there will be some tangible value produced in the form of improved departmental efficiency, but there's also the improved morale at the loading dock and the warehouse where the team is better managed, and doesn't have to wrestle with the old software.

That could have cascading benefits throughout the entire supply chain. The balanced-scorecard approach can be very helpful at this stage, as you attempt to measure the impact of the business value of certain strategy alternatives.

Aligning your alternatives with your organizational structure and culture

This process involves more than merely identifying costs, benefits, and risks that go along with each alternative. You have to consider the political challenges you might face as well. You must consider how any potential solution might align with the corporate culture. If it doesn't, even the utopian solution is destined to fail.

Any business intelligence strategy should work in concert with the power structure of the company. A decentralized organization, for example, lends itself to more entrepreneurial strategies. That's because decisions are made in the divisions rather than in a central control unit; the target users have more control over their experience, and are relied upon more to get the most value from the business insights they develop and find.

The power structure goes beyond centralized versus decentralized, however. The planning team must consider other aspects of the culture that may view certain solutions in an unfavorable light. For example, imagine a BI strategy that involved an aggressive fraud-detection system. It might be low-risk and provide tremendous business value in the form of reducing abuse. But if the corporate culture isn't accepting of intrusive technologies, installing the application might cause more trouble than it's ultimately worth.

If one choice seems to keep popping up as the right way to go, don't settle for less. Don't short-circuit the process just because there's an easy way to skip all those planning meetings you have on your calendar. This decision is important; it takes time for all the possible ramifications to bubble to the surface.

Let the process run its course so there's no doubt you've made the right choice when all is said and done.

11.4 MAKING YOUR CHOICE

Deciding on the BI strategy is the first step of the rest of your project. So take care in how you approach it.

Considering everything

An inevitable chicken-and-egg feel infuses this process: One decision affects another, which affects another, so it's hard to know where to start. You'll need to channel your school days — in particular, algebra class, when you had to solve simultaneous equations.

The way to solve simultaneous equations is simple: You find a way to hold one variable constant, and then adjust the other variable until it fits. In the BI world, that translates to making an assumption about what one or two important elements of your company's strategy will look like, and then seeing how that assumption affects the other pieces of the equation. Don't just think about the technology. This process requires a holistic approach to your business intelligence strategy and thinking about every possible variable — these, for example:

- Technology versus work processes
- Operational needs versus long-term goals
- Scope versus schedule
- Budget versus time
- Governance and control versus user flexibility
- Data integration versus performance
- Needs versus wants

Imagine you've identified the ideal reporting tool that suits the needs of your target users perfectly. But if you go with that tool, you likely cost yourself a month in implementation time because it doesn't integrate well with the existing applications you have in place. Optimizing one variable will cost you somewhere else, so the trick is finding that safe place where the high-priority needs are met, but the rest of the project doesn't pay an excessive price for it.

11.5 DECIDING ON YOUR STRATEGY

You have enough evidence to make a decision now. No really, you do. So just step off that cliff any time you're ready. What if there isn't a standout? If there's time, you can revisit the best available options, but more than likely, you need to make a choice and move on. At this point in the game, the reason you have two choices is because both are viable.

Several good techniques exist for picking two equal finalists. One way is to "game out" the next steps of the project, talking through likely scenarios that pre-suppose one of the two choices have been made. You've selected your path, so what happens next? At each fork in the hypothetical road, you'll talk through the likely impact on the core business processes, and assess the options available from that point forward. Most important as you walk through the imaginary architecture planning steps is to take note of any obvious roadblocks or problems that emerge by

taking one path versus the other. It involves making a lot of assumptions, but talking through a deployment path is a window into what the future just might hold.

Another way to choose between two seemingly even paths is to select a default choice and continue the debate. Pick one single fallback position that you'll take if no clear winner emerges. Then continue the discussion with a hardstop time limit. After time's up, if you haven't reached a different conclusion, move forward with the default selection and don't look back. If the differences between the two approaches are so minimal that you can't reach a definitive conclusion about which one's best after a full-throated debate, it's unlikely that some unrecognized disaster is lurking behind one of them. Flip a coin. With that decision made, it's time to get approval. If you don't, the project plan you're about to write won't be worth the paper it's not yet printed on.

11.6 GETTING THE NECESSARY BUY-IN

Just because your planning team thinks you've made the right call doesn't mean the rest of the company will. It's imperative that you get some outside perspective. A business intelligence initiative isn't just a technology installation; it precipitates real business changes in the way processes flow and decisions are made. Making such grand decisions in a vacuum is a sure path to failure. You'll need some level of buy-in from the teams that will be affected by the new systems. That means the folks who will be supporting the technology, contributing data, reaping the insights, and dealing with those systems at every point in between. And the best way to make that happen is to get everyone in the same place, and give them a single version of how it's going to go.

Conducting the stakeholders' summit meeting

One important option in making the solution choice is holding a stakeholders' summit. This meeting will introduce many of the key solution concepts and present some of the outstanding choices still to be made — to the people who will have to deal with their effects.

The breadth of the company should be represented at the summit, with the attendees being as high-level as possible. But this isn't a conference for bigwigs only; invite some mid-level managers and director-level folks who can provide solid expertise in their operational areas. Invite ambassadors from the user community as well; they'll be able to comment on what's

likely to happen where the BI rubber hits the road. Finally, make sure there are a few people from the technology team as well, who can speak to the constraints and impact on the rest of the IT infrastructure.

The goal of the summit is to increase the participants' understanding and sense of ownership in the initiative. You'll be getting a lot of vital input that will help you determine the strategy to use, but more important is the buy-in from critical personalities inside the company. Don't underestimate the importance of a good facilitator at a stakeholder summit. There will be a lot of strong, loud voices in the room, each thinking what they have to say is more important than what everybody else has to say. During the meeting (or series of meetings) you'll be presenting the main approach you're planning on taking, and how it affects each department, both good and bad. Listen to the stakeholders' concerns and do your best to address them. At the end of the stakeholders' summit, there should be a general agreement on the company's BI strategy.

Locking in your decision and moving ahead

With any luck, you'll reach a point where one strategy stands out above the rest. At that moment, it's time to lock the decision into place and move on to the next phase without looking back. Different camps might emerge within and outside the BI planning team, and inevitably one will be miffed when their first choice isn't your first choice. Take it all in stride. Criticism is part of the process. Stay positive, don't reenact the same debate over and over. When the course is set, quickly move to the next phase of the project.

11.7 CHECK YOUR PROGRESS

1. What are the key points to look for in the technology stack before starting a BI project?
2. List risks to a BI project
3. Apart from technology list other variables to consider in a BI project

Answers to Check your progress

1. Infrastructure, Security, Information management, Application and user interface layer:
2. Data risks, Application risks, Organizational risks, Financial risk

3. Technology versus work processes, Operational needs versus long-term goals, Operational needs versus long-term goals, Budget versus time, Governance and control versus user flexibility, Data integration versus performance, Needs versus wants

11.8 SUMMARY

This unit drills farther into business intelligence strategy. Before you start making specific plans for your solution, however, take a quick, hard look at certain realities that will become very important during the implementation. This unit helps with that step, too.

11.9 SELF ASSESSMENT QUESTIONS

1. Explain how to Assess business infrastructure?
2. Describe how to assess the technology stack, top to bottom
3. Explain about Exploring “Should-Be” and “Could-Be” BI Alternatives

11.10 REFERENCES

1. Swain Scheps - Business Intelligence For Dummies-For Dummies (2008), Wiley Publishing, Inc
2. Carlo Vercellis - Business Intelligence_ Data Mining and Optimization for Decision Making (2009), Wiley Publishing Inc
3. Howson, Cindi - Successful Business Intelligence-McGraw-Hill (2014)

UNIT -12: BUILDING A SOLID BI ARCHITECTURE AND ROADMAP

Structure

12.0 Objectives

12.1 Roadmap

12.2 Centralized Versus Decentralized Architecture

12.3 BI Architecture Alternatives

12.4 Developing a Phased Incremental BI Roadmap

12.5 Check your progress

12.6 Summary

12.7 Keywords

12.8 Self Assessment Questions

12.9 References

12.0 OBJECTIVES

After studying this unit, you will be able to

- ✓ Getting the most out of your roadmap
- ✓ Selecting the degree of centralization
- ✓ Getting a laser-focused plan
- ✓ Analyzing your solution's degree of coupling

12.1 ROADMAP

A business intelligence roadmap is one or more documents that lays out the formal objectives, high-level requirements, and specific activities of the project. It is, in effect, a founding charter that the project team will use as a guiding North star to set schedules, allocate resources, and even to market the project internally. Your roadmap should tell you what you're trying to accomplish, what tools you're going to use to get the job done, how you plan to do it, and (to some extent) the justification for the approach you're taking. Here are some examples of sections of what to include in your roadmap:

- Statement of overall business problem(s) and the specific scope of the solution.
- Business perspective of the solution — for example, what information needs your system will meet that it wasn't meeting before.
- Initial financial analysis, including ROI projections.
- Current condition of the organization's information infrastructure — including a discussion of where all the relevant data is being housed and what condition it's in.
- High-level review of hardware requirements, emphasizing any new platforms you may need to implement.
- Discussion of existing and new software that will be utilized for the BI solution.
- General make-up of project team and division of responsibilities.
- A section on risks, constraints, and assumptions that lays out where things can go wrong, as well as the known limits of the BI implementation.

What you won't see anywhere on your roadmap is a detailed work plan. A roadmap is not a comprehensive task-level project plan. It's a strategic-level document that outlines the major decisions about how the business intelligence solution is to be implemented. Unlike a project plan — which lays out every step — a set of major choices transforms the goals and requirements of the system.

Before you can finish your roadmap, you'll be forced to make some big decisions about the architecture of your BI solution. Those initial decisions will determine what kind of tools you'll need, and what kind of talent you'll need on your team to get the job done.

As you build the roadmap, you'll have to take a microscope to your needs and abilities, think through more than a few ideas and possibilities (to make sure you haven't "group-thought" yourself into one solution when another might be better for you), and make contact with people who can help you — both within your organization and without. Don't short-circuit the planning process by thinking that the roadmap document itself is the goal. While you want to make your roadmap as complete as it can be, don't get stuck in the mud if you don't know with absolute certainty all the answers and have to leave some stub headings that have no attached content beneath them for the time being. That's an inevitable part of the planning process. For example, if you work for the International Widget Corporation and you know there's a chance the unprofitable Micro-Widget division will be sold, you may need to account for that variable in your roadmap by inserting an assumption stating how different outcomes could lead to different planning decisions. Above all, don't bring the project planning process to a halt while waiting to nail down every possible loose end.

12.2 CENTRALIZED VERSUS DECENTRALIZED ARCHITECTURE

The roadmap's shape and form will hinge on the overall shape of the solution. And one of the biggest drivers of that is the question of whether to build out a centralized or decentralized system. The two architectures involve very different approaches to the implementation process. In this section we'll take a look at why the question is so important, and talk about ways to approach the answer. It's considered the biggest decision because it's the first. Everything you do in your BI implementation will be affected.

A couple question

Some BI specialists use the term *coupling* to describe the degree to which a BI system is centralized and consistently applied. A highly centralized system might make a common set of tools available to the entire company, share the most effective practices across the organization, and (usually) give a single entity or team the responsibility for making key BI decisions that affect everyone. In a system like that, one department's BI system is coupled to the central BI regime — both technologically and procedurally. It makes for easy management, but can stifle the effectiveness of a BI solution if the business situation calls for more versatility; everyone has to use one-size-fits-all tools.

On the other hand, a decentralized system allows some variations of policy and practice across different BI domains and functional areas. Departments and business units might use the same tools, but they aren't obligated to do so as they would be in a centralized system. This kind of system, where tools and practices are built up independently and *de-coupled* from a central decision- making power, is often appealing because it can put ideally suited BI tools in the everyone's hands. (Of course, it can also be a tad inefficient and hard to manage, what with everyone doing their own thing.) lets consider departmental and enterprise BI arrangements, but that's a matter of scope. A BI system's scope isn't the same as how *coupled* it is. It's possible to have a highly decentralized departmental implementation of BI.

For example, imagine an international conglomerate that wants to introduce BI for its sales team, but the sales team has sub-units that work on different products with wildly different sales dynamics. Perhaps they're in different countries and don't even speak the same language. In that case, a one-sizefits- all solution wouldn't work; you'd have to allow those sub-units to customize the BI solution as needed.

How to choose

So which one do you need? A centralized or decentralized architecture? Well, that depends on a handful of variables that can be summed up under two headings:

- **Organizational culture.** Most companies have one of two basic cultures:

  • **Autocratic:** Decisions are made from the top down, with little room for interpretation. Autocratic cultures lend themselves to centralized BI; the apparatus is already in place for centrally controlled strategy and application administration, and usually that's what their people are used to.

• **Entrepreneurial:** Business decisions are made throughout the company and innovation is encouraged. Where departments and teams have the authority to dictate the terms of how their supporting tools are built, decentralized BI systems are more likely to flourish.

- **Organizational structure.** In most cases, but not all, the organizational structure aligns with the company's culture. For example, if decisions are made at the top, the structure of power and communication radiate from a handful of top executives down to the rest of the organization in a rigid way, while entrepreneurial companies are likely to be matrix managed rather than hierarchical.

It's important to be aware of organizational factors because they will dictate how data moves between business units and teams, and that will ultimately guide your centralized versus decentralized decision. Since BI typically needs to bring lots of disparate data together, a company where teams are not in the habit of sharing data or working together for the greater good could make it difficult to install a centralized BI solution. You'll want to understand how the business units work together to perform common functions, as well as where they are geographically.

It's probable the answer lies somewhere in between centralized and decentralized BI architecture. That may seem contradictory, but the reality is that while large chunks of the organization may require tools and practices in common — and a rigid control system — pockets of decentralization may persist as well. For example, a company might install an enterprise-wide, centralized, homogeneous BI system, but allow one key team with a unique function — say, the corporate-strategy group or the sales-operations crew — to build their BI solution as they see fit.

Here the first key decision is setting policy to specify how coupled your BI environment should be; it's arguably as important as any decision you'll make throughout the life of your project. Answering it forces you to look closely at many aspects of your existing systems — and of your company as well. It's not a question you can expect to answer in one sitting. When you answer this how-coupled-is-it question, you set the direction for the entire BI implementation. The degree of centralization you specify will cascade to the other steps in the plan; when you've decided on a coupling scheme, you're ready to move on to other major choices — such as what

kind of tools you're going to use and how the data will be handled. And none of those questions can be tackled without answering the first question. Bringing disparate data under a common umbrella in the form of a *data warehouse* (a dedicated repository for historical operational data) is a common choice made during the roadmap phase — but it may not be right for every situation. Integrating data into one big centralized melting pot is far easier said than done — and will inevitably occupy an enormous chunk of your project resources. A data warehouse may well be the answer you're looking for, but it's certainly not the only solution out there. Often alternatives are available that work just as well and spare you some heartburn along the way.

12.3 BI ARCHITECTURE ALTERNATIVES

Once you've developed either a centralized or decentralized model, you can make the other general architecture choices. You'll need to consider a number of vital factors, such as these:

- How and where will the data be maintained?
- What will the integration schedule be?
- What tools will sit on which desktops — throughout the organization?

You won't lay out every single detail of all possible solutions until you get into detailed requirements and design, but you'll want to get a broad perspective on the solutions available in the marketplace. In the beginning, you have a lot of questions and few answers, so you'll have to start by working through a list of all possible alternatives. After that, you narrow it down to a few good candidates.

Starting an architecture evaluation

So what factors are important when you're looking into possible solutions? You'll want to stick with the basics and view every alternative through the prism of your main business needs. For BI implementations, your architecture choices will almost always start with three major categories of technology:

- Hardware
- Data management
- End-user tools

Of course, each of those main categories can be broken down into subcategories (even sub-sub-categories), but start with the big three. Sure, there's more to an enterprise-wide BI solution than just hardware, databases, and front-end tools. For example, you'll probably have to consider the network infrastructure that acts as the conduit between system components, as well as a vast collection of *middleware* (software that acts as the connective tissue between network components, data sources, and applications). BI can affect all of it.

As you examine each architecture alternative, be sure to address each of the main three components. For example, you may be starting your evaluation with the constraint that the hardware environment can't be changed. That could happen for any number of reasons — budgetary, political, or otherwise. In that case, you'll be forced to use what's available; the project team will have to find data-handling and tools software that can be placed on existing infrastructure without affecting performance any more than absolutely necessary.

In general, every architecture solution involves three general components that mirror the "big three" technology categories:

- **Hardware:** A discussion of what, if any, hardware changes are required (for example, will you have to increase server processing horsepower to handle the complexity of the transactions expected?).
- **Data management:** A small list of data-management options (such as how the data will be transported and transformed, and what the target database will need to be able to do.)
- **End-user tools:** Recommendations of one or more tools that meet the system's general business needs directly (such as an advanced statistical package, or a managed reporting solution to meet the needs of the prospective user groups.) Start with the end-user tools. If you have a rough idea of the requirements, then the tools are a good brass-tacks place to start. You can develop a list of available software that will meet your querying, reporting, and analysis needs, then work backward to expand the evaluation to identify compatible data handling technology that can support the tools. Then you'll look for what hardware is required to run it.

So many choices

Laying out all plausible alternatives — and giving each its day in court — is a step in the process skipped by far too many organizations. This phase doesn't have to last long, but it's an important

step because it o pens you (and the team) up to technologies and techniques that may not be in the standard recipe.

You want to work through a wide variety of choices while keeping your project's basic constraints and objectives in mind. But feel free, at this stage, to be a little more open with your ideas; allow your team some latitude during your planning sessions.

You'll want to look at solution elements like these:

- Operating systems
- Network protocols
- Server hardware
- Primary database vendor
- Data Warehouse and Extract,Transform, and Load (ETL) processes
- The kinds of front-end BI tools you absolutely must have
- The kinds of front-end BI tools that would be nice to have

So little time

At the beginning of the project, the slate is almost completely clean. The possibilities stretch out before you — the software tools, hardware platforms, protocols, and processes that will comprise your BI implementation. The skies are blue, the fields are green; whatever you envision can become a reality.

Get real.

While you certainly want to have a free-flowing discussion about the possibilities and endless alternatives, you'll need to narrow down your options to a few main candidates pretty quickly.

The company's installed technology and on-the-record future direction will be a primary factor in deciding architecture. Are there internal standards that limit your choice of tools both today and in the future? For example, does your organization rely on one main vendor for its database-management system? If so, does that mean the BI system is similarly limited? Keeping such basic constraints in mind should yield a general idea of what your options really are. The planning process is a notorious quicksand zone, where projects get bogged down as team leaders

agonize over the initial choices, knowing their importance. You're not going to make the right call every time. Sometimes identified "best practices" aren't "best" for your organization. But don't get paralyzed by the fear of making a misstep. You have no choice but to get moving, and surround yourself with vendors, consultants, and team members you trust. Keep a close eye on what they're doing, but put your faith in their ability to perform due diligence and meet the priorities you've laid out. As the old saying goes, *trust but verify*.

During the planning phase you look at the architecture as a whole, as well as the individual pieces. The categories of questions you'll need to ask at this point:

- Which solution components work well together? And which don't?
- What infrastructure is currently in place and does it have spare capacity?
- Does the company have existing relationships with some of the target vendors?

During this phase, you'll also start to get a handle on which pieces of the puzzle have higher priority than others. That's important information; you'll need to put it to use very soon.

The short list

The goal is to produce a short list of architecture alternatives that satisfy all of your bare minimum requirements, and hopefully supply some nice-to-have features as well. The vetting process is far from over; you'll want to turn over the short list of alternatives to some key stakeholders and analysts on your team to pick apart and find reasons to narrow down the list.

Each short list will include the querying, reporting, analysis, and other frontend tools that the end-users throughout the company will be using. There will also be the underlying database technology — not just software, but configuration options and architectural considerations as well. Finally, be sure to put any hardware requirements on the short list. For example, if the short list includes a solution that involves creating a new centralized data-warehouse environment, the entry on the short list should include

- A basic analysis of existing processing and storage capacity (relative to the minimum amount needed)
- An ideal hardware configuration for maximum performance

Taking a second look at your short list

You're going to have to get your hands dirty now; it's time to stop looking at solutions in a vacuum. You'll want to judge their capabilities and constraints in the context of your infrastructure. You may have already analysed the gap between your existing systems' capabilities and the business requirements for the BI system; now it's necessary to put the solutions in context. You can do that by examining your short list solutions and visualizing how they'll actually *work* when they're installed in your company's environment. You're looking to identify compatibility issues, integration problems, and other potential roadblocks that might arise when you start introducing new hardware and software to your existing technology environment. Eventually you're going to have to chop the architecture alternatives that don't play well with the other kids in the sandbox.

At this point it makes sense to begin in-depth discussions with candidate vendors and consultants so you can get detailed information on their product capabilities, plus a full menu of their support options to go along with their wares. Invite them in for a discussion of your situation and take careful note of how they would approach the challenges you face.

This is a great time to have software vendors do give demonstrations of their products. You can put an application through the paces and see how it holds up. If it's an end-user tool, invite a few key end-users to the demo, and ask for their input on how appealing, usable, and useful the software actually is.

It helps if the vendors have an idea of what you're trying to do, so you should be prepared to share a little information with them about your project. Project details will help the third parties tailor their pitch toward what you actually need. Check with your legal team and see whether there is a standard Non- Disclosure Agreement that you should use.

For gigantic implementations, you might have a vendor do an extensive Proof of Concept (POC) implementation as a way to test a product's ability to meet your specific needs. For a business intelligence solution, POCs are particularly useful to demonstrate whether different brands of software work together in your environment without building out the entire solution. POCs move beyond the theoretical realm of PowerPoint presentations, white papers, and even canned product demonstrations and reveal something of the true nature of the software.

Examining costs for each alternative

So far, cost has not been a factor in the conversation Bottom line: software licenses cost money. Servers and networking gear cost money, as does integration vendors' time. Evaluating possible solutions without considering your budget affords you a certain amount of freedom to isolate the best alternatives, and identify the most important components without constraint. But sooner or later the piper must be paid. Costs can sneak into IT projects in a lot of ways. Keep your eyes open for the following expenses:

- **Software licenses:** Don't be haphazard in your approach to buying licenses. Many vendors have complex and confusing schemes that can lead you to pay for more seats than you end up needing. A software partner with simple and flexible licensing can be worth its weight in gold.
- **Hardware-acquisition costs:** Buying new gear for your BI solution can be an expensive proposition, especially where it concerns high-performance servers and networking gear. Having a scalable solution is a good way to save money; you can start small and work your way up to the hardware capacity you need, but not until you actually need it.
- **Service and maintenance costs:** Many vendors make their money not from the initial purchase of their product, but from ongoing fees exacted from customers for supporting their products. Make sure you account for all outlays over the life of the products you're buying for your BI system, not just the big check at the beginning.
- **Training and support costs:** Complex software means the end-users of your business-intelligence system will need help before they get full use out of it. It's important that quality education be made available for the user community, and that costs money.

Always remember the *business* in *business intelligence*. Your organization is trying to make money, and that can only happen if you increase revenues or reduce expenses. In most cases, a BI implementation is already a big investment for a company — but don't confuse executive approval of your project with license to go crazy with the company cheque book. With limited resources, you'll want to stretch your budget resources as far as you can.

As you examine the costs of each solution alternative, keep in mind that *the most expensive solution isn't always the best.* It's a common trap that has snared many a project manager. Price is clearly an important component, but make sure that money don't enter into your *qualitative* evaluation of products to meet your needs.

Looking at technology risks

To this point, you've taken your short list through the paces. You've performed an analysis of each solution's viability as a way to meet your business needs. You've examined the associated cost of each element of your solution as well. Now it's time to look at risk. We're talking about specific risks that go hand-in-hand with technology — say, a *medium* meteorite that crashes into your data centre.

It's a scary word, but in essence, technology risk is nothing more than a *variable expense* that you don't see coming. It's impossible to predict with any certainty, but if you do your best to see it coming, it's possible to minimize its impact. Fortunately, there are some common guidelines you can look to if you want to understand risk in an IT environment better because it can have a huge impact on your BI rollout. Every large, complex IT implementation has common risks associated with it — these, for example:

- Software has unknown bugs that pupate and hatch at inopportune times.
- Software doesn't perform as promised by the vendor.
- Products don't work together as well as projected.

Included in your architecture and solution assessment should be a risk analysis of each choice on your short-list. You should include a list of the most likely things that could go wrong with each solution. It's always a good idea to quantify the likelihood of a risk scenario coming to pass, and include a projected damage toll — including how it would affect your BI initiative. Suppose, for example, you find that Application A has a large chance of causing minor performance issues with the system while the alternative, Application B, has a tiny chance of bringing the entire BI implementation to a halt. Depending on the values you assign for the likelihood of each outcome, you might actually decide Application B makes the most sense from a risk perspective.

When in doubt, go with proven solutions. And while you want to look at vendors with a track record of stability, you should avoid version 1.0 of just about any species of software. First-generation applications often have kinks that still need to be worked out, and they're a risky bet

to build your environment around unless you're getting major concessions from the vendor to protect and compensate your company for any problems that might arise.

Making your decision

It's a great feeling when a single solution emerges as a winner. Your project is practically laid out before you on a silver platter when one candidate solution is the only choice that fits your needs and your constraints. Unfortunately, that's a rarity. More often than not, a few solutions score very close to one another and you end up with a dead heat.

In spite of the temptation to flip a coin, you should take this opportunity to do a fresh analysis of your candidate solutions. That will usually lead to finding a deeper set of criteria from which to judge each candidate solution, with the aim of selecting a winner.

These three steps can help you break any ties:

1. **Verify your information**. Make sure all your existing information is correct. That means re-working pricing numbers, compatibility issues, and functional capabilities. Go over the features of each product step by step; make sure the analysis you're reading is fair and unbiased (not to mention up-to-date, since software features can come and go from release to release).
2. **Revisit your criteria.** After you've verified your research data, make sure you haven't missed anything in terms of judging criteria. Are you basing your judgment on the *complete* set of business drivers? Or are there some considerations you initially left out because they seemed irrelevant to a particular architecture choice?
1. **Get a new perspective.** It's a great idea to get a fresh set of eyes on each solution. Sometimes you'll find if you stare at something long enough you lose all objectivity and perspective. An outside resource, even someone who has no direct expertise in the kind of system your building, might have an angle that you hadn't considered on why one solution is better or worse than another.

12.4 DEVELOPING A PHASED INCREMENTAL BI ROADMAP

Like the project itself, the roadmap is something you build one iteration at a time. You take a first pass at the document with candidate solutions, then narrow those down to a few, and finally

a winning architecture emerges. At each step of the way, the roadmap changes, becoming more focused, and providing a deeper level of detail.

With the architecture selection made, and the solution coming into tighter focus, it's time to start working through how you're going to make your vision a reality. The goal is not to create a full project plan with step-by-step instructions; instead, the roadmap must include what you're going to deliver and when. The roadmap is where you lay out a strategy for building your business intelligence solution in a way that keeps momentum up, maintains support throughout the organization, doesn't use up resources too quickly, and tolerates occasional failures. it's a good thing if it delivers on its promised business value as well.

Deciding where to start

Instead of starting with the first single step, you'll want to define what the entire first *phase* is going to look like. It's almost always to your benefit to start with a limited solution that grows into a full BI implementation after several subsequent phases. Doing that ensures that any early failures are small and can be overcome quickly. If you spend several years on a comprehensive, enterprise-wide BI implementation, the audience anticipation grows with each passing month. When the big day comes and the spotlight is on you as it's time to hit the switch, you'll be in big trouble if the Christmas tree doesn't light up as planned.

So instead of shooting for the moon, you should look for objectives within your grasp for the early phases of the project. No IT book would be complete without the low-hanging fruit metaphor, so here it goes: Your Phase I implementation should pluck the lowest-hanging, ripest, best tasting fruit from the tree. That is, the initial goal should be to start building a solution at the intersection of the highest-value, least risky functions that are also the easiest to deliver:

- **Highest value:** If you have a system in place now that's already working, only at a level that requires eventual change, then skip that functional area for one where the user community is clamouring for anything to make their lives easier.
- **Least risky:** Don't roll out executive dashboards first, or any other function whose failure might lead to the bigwigs pulling the plug on the whole shebang. And we're not just talking about political risk; it's wise to avoid implementations that might interfere with systems that are functioning perfectly well.

- **Easy to deliver:** You should also avoid implementations that are highly technically complex as well. A simple solution establishes your team, lets you develop your internal processes, and build a tradition of success with the company.

Out of those three qualities, chances are you may only find two, but it can't hurt to be optimistic. It's always best to do improvements to existing systems, rather than brand new systems. If (for example) you already have a sales analytics module up and running, it probably makes sense to make the early phases of your project include an upgrade to advanced sales analytics before building the HR function from the ground up.

Keeping score

It's easy to decide to go after the low-hanging fruit, but what if the fruit one branch up is extra-tasty? And what about the fruit that just fell off the tree and doesn't even require a ladder? It's not always easy to decide where your priorities should be. A back-of-the napkin scorecard system might make sense early on — where you lay out the key variables listed in the previous section, along with a few that are peculiar to your situation. Work through the possible first steps and grade them out based on how they fall in each category. From that scorecard, you'll get the optimal combination of functions for Phase I. If you do a rough-and-ready scorecard of categories to evaluate, be sure you score each category the same way — even if that seems a little counterintuitive. Figure 12-2 shows a list of four possible Phase I initiatives for your BI system. On this scorecard, a higher score simply means "more advantageous to the company," so under Value a score of 4 means *more valuable* and a score of 1 means *less valuable*. For the Risk category a score of 4 means safer — that is, *less risky* (since less risk is more advantageous to the company, right?) and a score of 1 means the *most risky*. When all the individual scores are added, we find that the best Phase I solution is the upgrade to the reporting tool, whose score totaled 9. Sure, it's the least valuable — but because it's going to be easy and virtually risk-free to install, it makes the most sense.

The scorecard example in Figure 12-2 assumes all three criteria matter equally. You can always adjust the scorecard differently depending on what you're trying to accomplish, or if there are any special circumstances surrounding the choice you have to make. For example, if your company is especially risk averse, you can rig the scorecard to be more sensitive to the risk

category by multiplying each Risk score by 2 prior to totalling the sores for each choice. Just remember: A perfectly balanced and fair scorecard is not the goal here; it's just a tool to help illuminate your options.

Phase I Solution Choices Scorecard

Project	Value	Risk	Easy	Overall
Sales Dashboard	2	2	3	7
Finance OLAP Conversion	3	1	2	6
Reporting Upgrade	1	4	4	9
Productivity Analytics	4	2	1	7

Fig 12.2 A sample score card for determining sensible steps for your BI project

Deciding what comes next

The decisions you make about Phase I will determine how to proceed. If Phase I includes a Sales analytics implementation, you'll proceed like you would with any other IT project, beginning with an informational or discovery phase, followed by an architect phase where you design the solution, and on into a build and test phase.

Deciding what comes next, and next, and next . . .

Now you're on your way. You've got your roadmap in place with Phase I's deliverables. Rinse, lather, and repeat with Phase II. If you want, you can simply look at the next best item on your scorecard and pencil that in as the next priority on your roadmap. Or you might consider adjusting the scorecard and adding the scores again; priorities may change after you have the first success under your belt.

Planning for contingencies

you'll need to get familiar with the practice of contingency planning, and build some emergency scenarios and decision points into your roadmap. A contingency plan is little more than a carefully designed set of alternative branches in a project roadmap. Under certain conditions, the contingency plan gets activated. For the purposes of your roadmap, you'll need to flush out

specific project risks that could hamper development, delay the release, or put the entire initiative in jeopardy. The simplest contingency plans are those that reduce the scope of the project in case something goes wrong. Or if a part of the release fails, your contingency plan could be as simple as preparing a pre-assembled trouble-shooting team that swings into action.

You need to understand that like any large complex high-visibility project, a business intelligence implementation has dozens of inflection points where problems can appear and derail your initiative.

Some examples of project risks include these:

- Higher-than-expected project staff turnover
- Loss of project champion or sponsor
- Higher-than-expected — or unexpected — expenses that cause you to blow through your budget

Technology-specific risks such as these:

- Integration problems with existing software and hardware
- Over-promised (or under-delivered) software functionality

A good contingency-planning process identifies risks like these and creates alternate pathways into the roadmap, and later, the project plan itself. In some cases, best practices can show you the way to a safe harbour in the event of a storm. But for certain problems — including those unique to your team, your implementation, or your company — you'll need to be ready to toss out the book and improvise.

If you read the transcript of the radio chatter between the spacecraft and Mission Control during the failed Apollo XIII lunar-landing mission, you'd think they were dealing with a backed-up space toilet rather than the grave problems the astronauts actually faced. That's because NASA emphasized contingency planning in the early days of the space program, and does so even today. The transcript shows only the slightest hints of concern from the astronauts and Mission Control — even after discovering the crew's oxygen was leaking into space.

Your BI project is like a space mission — a large, complex system where pieces can fail unexpectedly. The better you can anticipate problems — and work out solutions and plans ahead of time to deal with them — the more likely your project won't be lost in space when something goes wrong.

Dealing with moving targets

It would be nice if you could freeze the world in place as you build out your BI system. That way you wouldn't have to worry that while you were off solving one problem, another problem on the other side of the company changed shape without you knowing about it. Unfortunately, BI implementations don't happen in a vacuum; there are all kinds of dependencies and vital connections with resources in various parts of your organization. Given the fact that planning and design takes time, it's always possible that things will change in an unexpected way.

It's not always easy to do, but as you build your roadmap, it's incumbent upon you and your team to account and plan for as many external variables as possible. Suppose, for example, your data-warehouse system is designed to use a state-of-the-art storage network that you've been told will be launched only a month before the data warehouse comes online. It's probably a good idea to have a contingency plan available in case the new storage system *isn't* available.

And it's not just major IT systems that can change, either. Tiny changes to the data schema that feeds your data warehouse (or perhaps an adjustment to the network addressing scheme) may happen without you being any the wiser. Seemingly insignificant updates can have a big impact on your project. Open a line of communication with other project managers in your company who are working on IT-related initiatives parallel to yours. Be aware of their roadmaps; work with them to coordinate the dates on your project plan with theirs. Reach out to technology-governance boards wherever possible to ensure you're up to date on scheduled system and process changes. And most of all (again), do some good contingency planning.

Leaving time for periodic "architectural tune-ups"

In spite of what you might have heard about the Pyramids in Egypt, not all architectures are designed to last forever. And the time to face your BI architecture's mortality is now, rather than after it becomes obsolete, useless, or just aggravating to the users and administrators. A BI system is a constantly-evolving organism; there will be regular upgrades and tweaks to functionality. The applications might be rolled out to a new team one month, and a software patch might be rolled out the next month. In such an environment, it's easy for changes to pile up without paying attention to their combined effect on system performance.

One way to avoid problems is to plan for occasional code freezes in your project plan. During these periods (a good standard is *one quarter out of every two years*), the system gets tuned from time to time, but that's about the extent of the changes. The code freeze gives you an opportunity to replace servers, tune your databases, upgrade your front-end user tools, and perform other tasks necessary to keep your system in shape. It's also a good time to make an honest evaluation of the state of the system.

As you examine the various elements of your system for how effectively they're performing their tasks, you'll want to ask yourself some key questions not just about raw quantitative performance issues, but also about softer, more qualitative issues — for example, how user-friendly the system is, whether it's time to upgrade the training program, and so forth.

It makes sense to have some "meta-metrics" (metrics that keep track of the metrics) for your BI architecture; they're essentially performance indicators for the system itself. Build some universal benchmark tasks that you can execute every so often to get a glimpse of the system's health and allow you to compare performance over time. Keep track of system uptime and throughput. And it's always important to have a standard user-satisfaction survey that gets distributed on a regular basis. This will provide you a good headsup when issues start to crop up.

It can't hurt to stay in touch with the state of the marketplace for the main cogs in your system — the data warehouse, the ETL software, the querying and reporting tools, and so on. The goal is not to create a killer case of buyer's remorse for you, but rather to keep an eye open for new pieces to your puzzle that might improve your performance, extend your existing functionality, or allow you to extend BI's reach to a new set of users within your organization.

12.5 CHECK YOUR PROGRESS

1. Define *middleware.*
2. What is a business intelligence roadmap?
3. What is coupling in BI?
4. List a few BI project risks.

Answers to Check your progress

1. software that acts as the connective tissue between network components, data sources, and applications.

2. A business intelligence roadmap is one or more documents that lays out the formal objectives, high-level requirements, and specific activities of the project.

3. Some BI specialists use the term *coupling* to describe the degree to which a BI system is centralized and consistently applied
4. Higher-than-expected project staff turnover
    - Loss of project champion or sponsor
    - Higher-than-expected — or unexpected — expenses that cause you to blow through your budget
    - Technology-specific risks such as these:
      • Integration problems with existing software and hardware
      • Over-promised (or under-delivered) software functionality

12.6 SUMMARY

Right up to this point, a lot of the focus has been on theoretical issues — defining tools and technologies, and assessing how one piece fits together with another piece. But now it's time to put that knowledge into action and begin to build the project roadmap. As with any map that helps you on a journey, your BI roadmap is just a general guide to help you get from point A to point B. objective of this unit is to let you find the best route between them.

Notice that we're looking for the *best* route between where you are now and where you want to be, not simply the fastest or the cheapest route. As is the case with any IT project, the "impossible triangle" (Figure 12-3) is in effect for BI implementations. The concept is simple: You may strive to create a project that's cheap, good, and fast, but you can't have them all. For example if you want an inexpensive system that works well, you'll have to sacrifice time. On the other hand, if you're looking to install something immediately on a limited budget, it's not going to be very good.

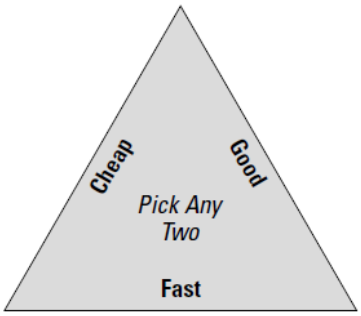


Fig 12.3 The well-known Impossible Triangle of IT projects.

12.7 KEYWORDS

- *coupling* - to describe the degree to which a BI system is centralized and consistently applied.
- *de-coupled* - **Is a** system, where tools and practices are built up independently
- Autocratic - Decisions are made from the top down, with little room for interpretation.
- Entrepreneurial - Business decisions are made throughout the company and innovation is encouraged

12.8 SELF ASSESSMENT QUESTIONS

1. Write in detail, what a typical BI project road map should include
2. Compare and contrast Centralized Versus Decentralized Architecture
3. How do you evaluate BI architecture alternatives?
4. List the expenses which should be considered while choosing a BI tool.
5. Explain How to develop a BI roadmap in a phased incremental manner.

12.9 REFERENCES

1. Swain Scheps - Business Intelligence For Dummies-For Dummies (2008), Wiley Publishing, Inc
2. Carlo Vercellis - Business Intelligence_ Data Mining and Optimization for Decision Making (2009), Wiley Publishing Inc
3. Howson, Cindi - Successful Business Intelligence-McGraw-Hill (2014)

BLOCK 4 INTRODUCTION

The purpose of this block 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

13.1 Overview of Visual Data Analytics Features

13.2 The Tableau Suite

13.3 Installing the Tableau Desktop

13.4 Tableau Desktop workspace

13.5 Check your progress

13.6 Summary

13.7 Keywords

13.8 Self Assessment Questions

13.9 References

13.0 OBJECTIVES

After studying this unit, you will be able to :

- ✓ Examine Importance of visual data analytics
- ✓ Identify Features of the Tableau software
- ✓ Install the Tableau desktop
- ✓ Create Tableau desktop workspace

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.

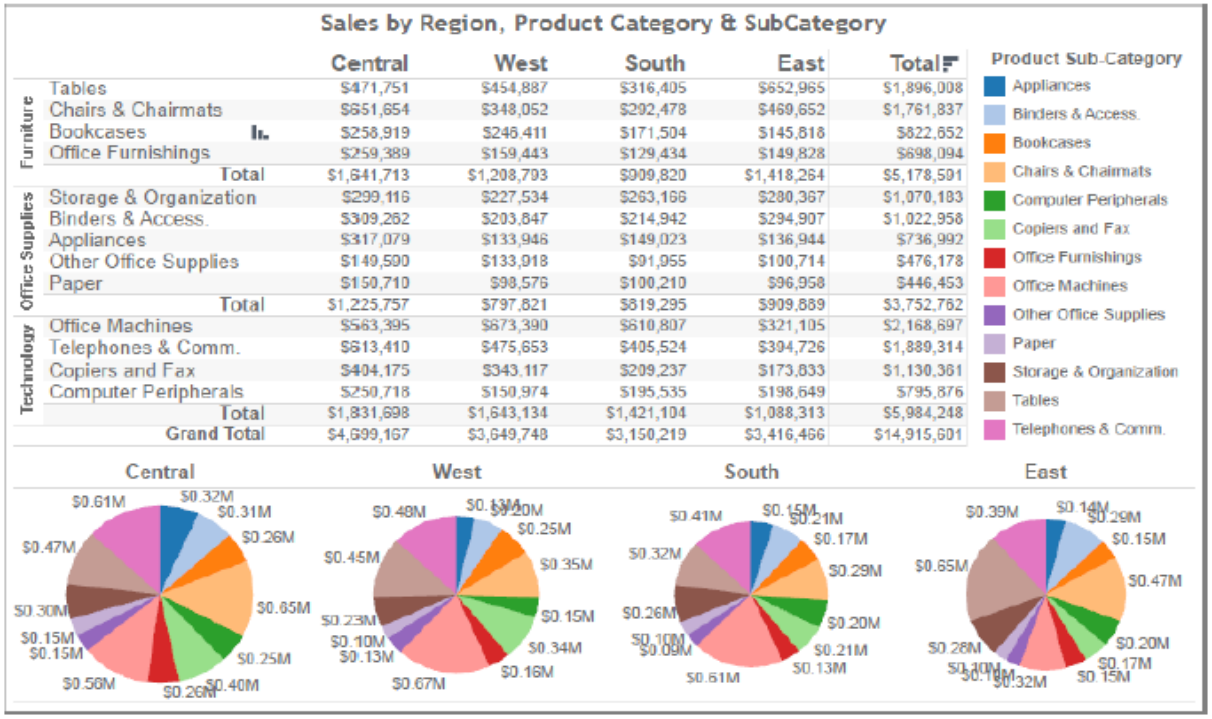


Fig.13.1 *Sales Mix Analysis using a crosstab and pie charts*

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.

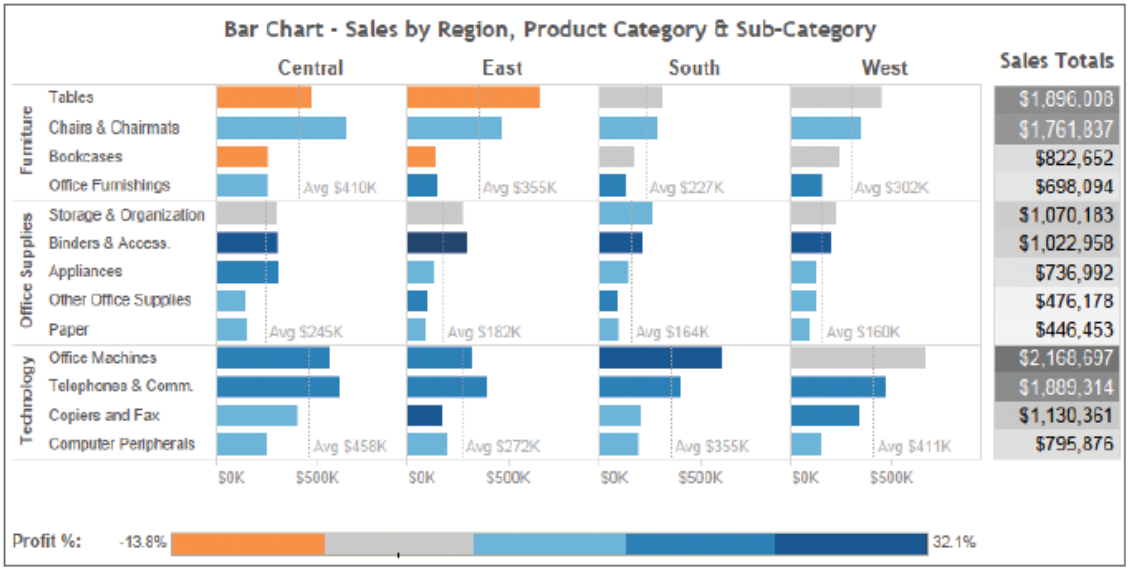


Fig.13.2 *Sales Mix Analysis using a bar chart and heat map*

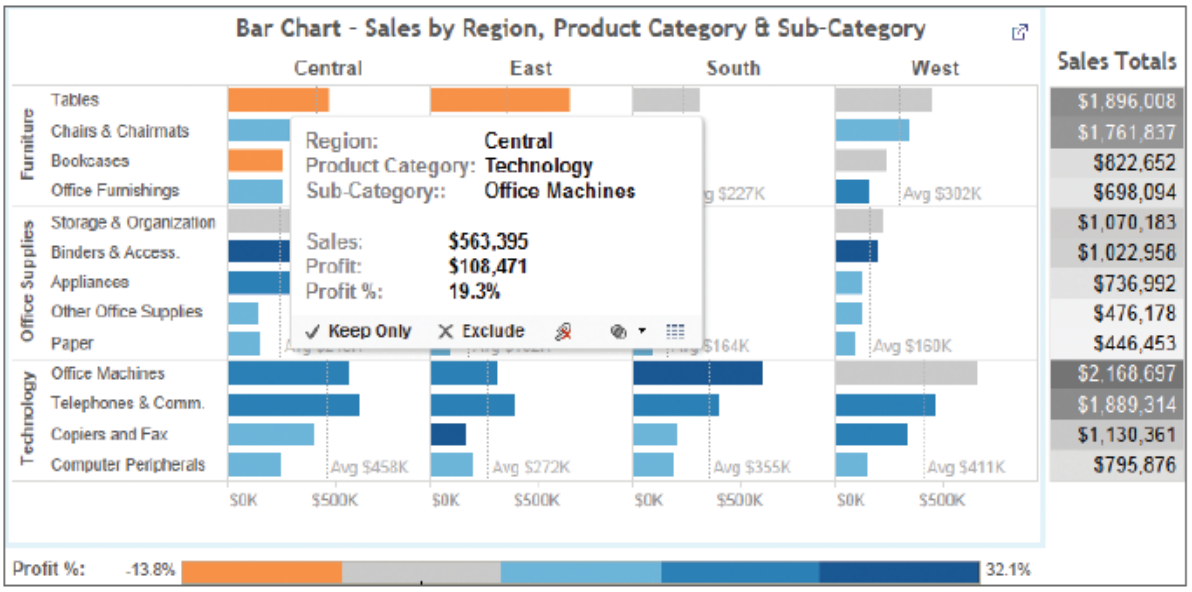


Fig. 13.3 *Adding labels and tooltips*

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.

Tableau Prep Tableau Prep is the latest addition to the Tableau product suite and is designed to help you prepare your data before you analyze it in Tableau Desktop. The visual interface allows you to quickly merge differently formatted datasets, clean the data, and unify the level of aggregation. Tableau Prep fits seamlessly into your analysis workflow.

Tableau Server Tableau Server is a platform for data analysis and is used by small family- run businesses and large Fortune 500 companies alike. It is intended for the organization- wide provision of data visualizations and dashboards that can be viewed in a browser and are frequently embedded into the organization's intranet.

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.

Tableau Reader Tableau Reader is a free desktop application that allows you to open and interact with Tableau workbook files that have been created in Tableau Desktop. However, it is not possible to make any changes to the visualizations in Tableau Reader.

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, I recommend the following additional specifications:

- 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

Figure 13.4 shows 15 numbered portions of the interface. The still- blank canvas (1), as it is called, includes the title Sheet 1 (2). On the left, you will find the Data pane (3). The tab next to it opens the Analytics pane

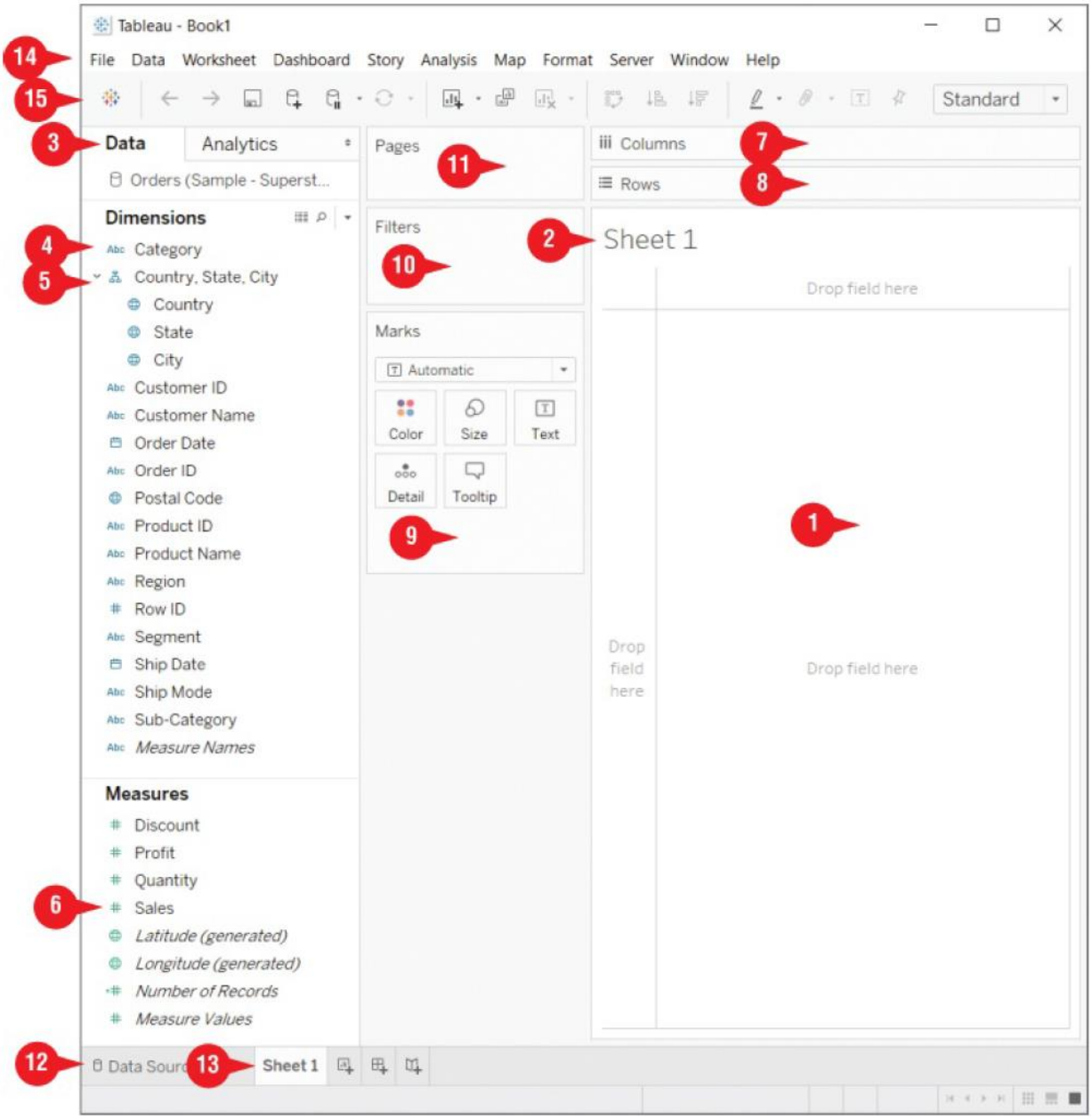


Fig. 13.4 The Tableau workspace with a yet to be filled canvas

In Tableau, most interactions are achieved by dragging and dropping items onto the canvas. This makes interacting with Tableau easy and intuitive. Both dimensions (4) (including hierarchical dimensions [5]) and measures (6) can be moved directly onto it. Alternatively, they can be placed on the Columns (7) and Rows (8) shelves, in order to add them to your visualization.

Fields from the Data pane can also be placed onto the Marks (9), Filters (10), and Pages (11) cards: for example, to change the color of marks or to only display marks for a filtered- out subset of the data.

The tabs bar at the bottom of the screen allows you to go back to the data source editor (12) and to toggle between your different worksheets (13), each containing a single visualization. With the three buttons to the right of the tabs, you can open additional worksheets, new dashboards, and stories, respectively. At the top of the screen, you can find the menu bar (14) and directly under that is the toolbar (15), with three important buttons:

The Tableau icon: This brings you back to the start screen, where, among other things, you can add additional data sources.

Undo: This allows you to go back a step so you can safely try out different ideas. You can go back as many steps as you like.

Redo: This allows you to restore any undone actions.

The Menu Bar

Even though most of the work in Tableau can be achieved by directly interacting with items using the mouse, there is also a menu bar at the top that lets you access additional features and settings. Let's take a closer look at some of the particularly useful entries:

File Menu The File menu contains the key functions Open, Save, and Save As.

The Print To PDF menu item allows you to export your worksheets and dashboards as PDF files. With the Repository Location option, you can look up and change the default location for Tableau files on your machine. With Export As Version, you can create workbooks for colleagues who might still be using an older version of Tableau Desktop.

Data Menu Here, the Insert function is especially interesting, as it presents a quick, ad hoc way to add a data table—for example, from a website. Simply select and copy the table in the original document, and click Insert in Tableau. This will add the data to your workbook as a new data source.

Worksheet Menu With Export, you can take your data out of Tableau by creating an image, a database file, or an Excel crosstab. Duplicate As Crosstab, on the other hand, opens a new worksheet in Tableau, showing a crosstab view of the data used in your visualization.

Dashboard Menu Dashboard actions that add interactivity to dashboards are set up and tweaked by clicking Actions.

Story Menu The Story menu entry lets you create a story from your worksheets and dashboards. In a story, content is arranged sequentially for presentation and enriched with annotations.

Analysis Menu With this menu, you can create and edit calculated fields. Here, you will also find options for tweaking table layouts as well as for showing grand totals, forecasts, and trend lines.

Map menu In the Map menu, you can choose between different background maps. The Offline option is particularly useful when you have no Internet connection and would like to access the built- in cartographic material.

Format Menu In this menu, you can set the font, alignment, shading, and other formatting options. In addition, you can set the overall workbook design and adjust the cell size.

Server Menu Use this menu for sharing your dashboard via Tableau Online,

Tableau Server, or Tableau Public. With the Create User Filter submenu, you can set audience- specific filters that grant specific users or user groups (which have been defined in Tableau Online or Tableau Server) access to selected subsets of the data.

Window Menu Use the Presentation Mode option to use the full screen for your dashboard.

Help Menu Via this menu, you have access to the Tableau online help, training videos, and sample workbooks. Use the Start Performance Recording option in the Settings And Performance submenu to analyze the processing time of your dashboard.

The Data Pane

The Data pane is divided into measures and dimensions. You control what visualizations you want to display by adding different combinations of measures and dimensions to the canvas.

Measures

Measures are numeric variables. By adding a measure to the view, you decide which values from your dataset to visualize. By default, Tableau automatically applies an aggregation function such as SUM or AVG (the arithmetic mean) to measures. That way, you can, for instance, show the sum or the average of a sales discount across different transactions. Measures typically (but not always) come with green symbols, which represent continuous variables.

Dimensions

Dimensions are descriptive, categorical variables. With dimensions, you can decide how to group the aggregated values of the used measures. For instance, the sum of sales revenue (a measure) could be broken down by country, product category, or both (i.e. two different dimensions). Typically, dimensions come with blue symbols in Tableau, which represent discrete variables.

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 datasources.

   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 SELF ASSESSMENT QUESTIONS

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

14.1 Data

14.2 Values

14.3 Knowing When to Use a Direct Connection or a Data Extract

14.4 Joining Database Tables with Tableau

14.5 Blending Different Datasources in a Single Worksheet

14.6 Data Quality Problems

14.7 Check Your Progress

14.8 Summary

14.9 Keywords

14.10 Self Assessment Questions

14.11 References

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 DATA

When you open Tableau you are taken to the home page where you can easily select from previous workbooks, sample workbooks, and saved data sources. You can also connect to new datasources 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 datasources 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

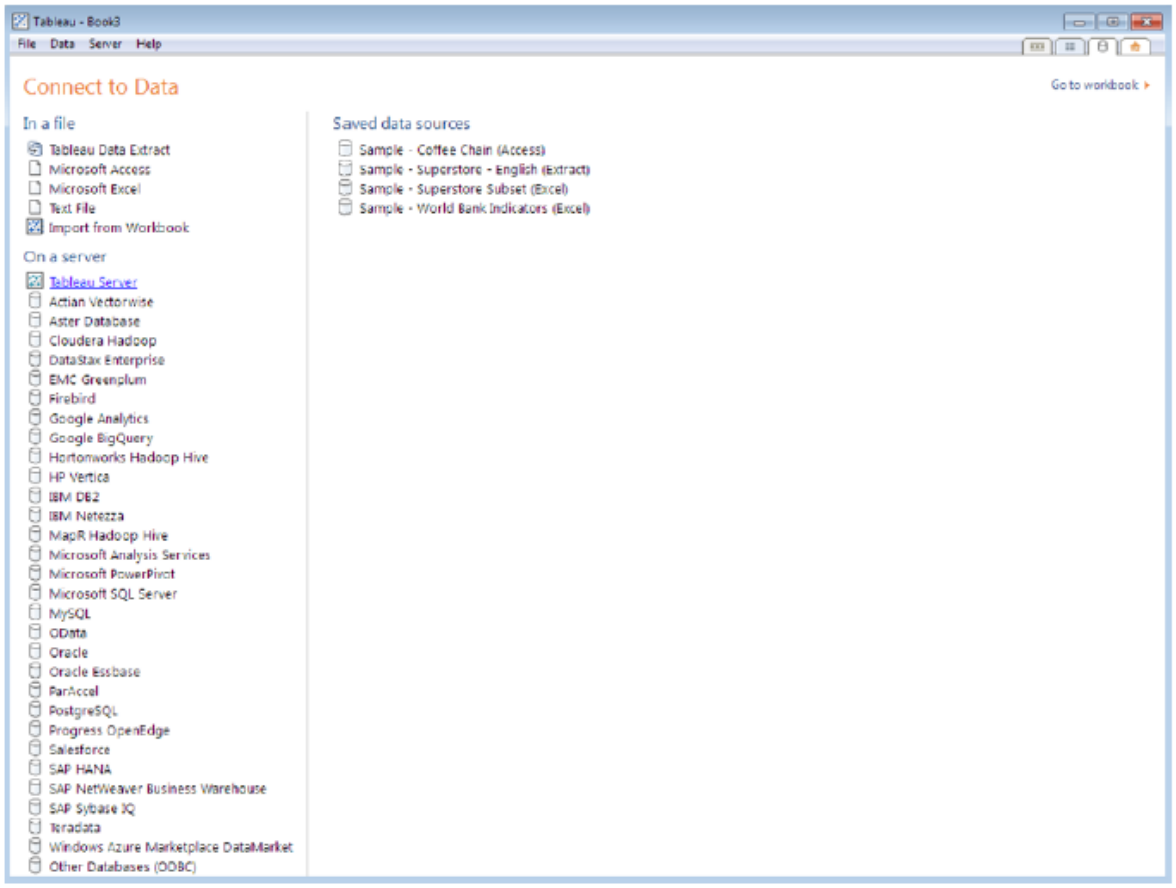


Fig 14.1 Connect to data screen

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 datasources 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.

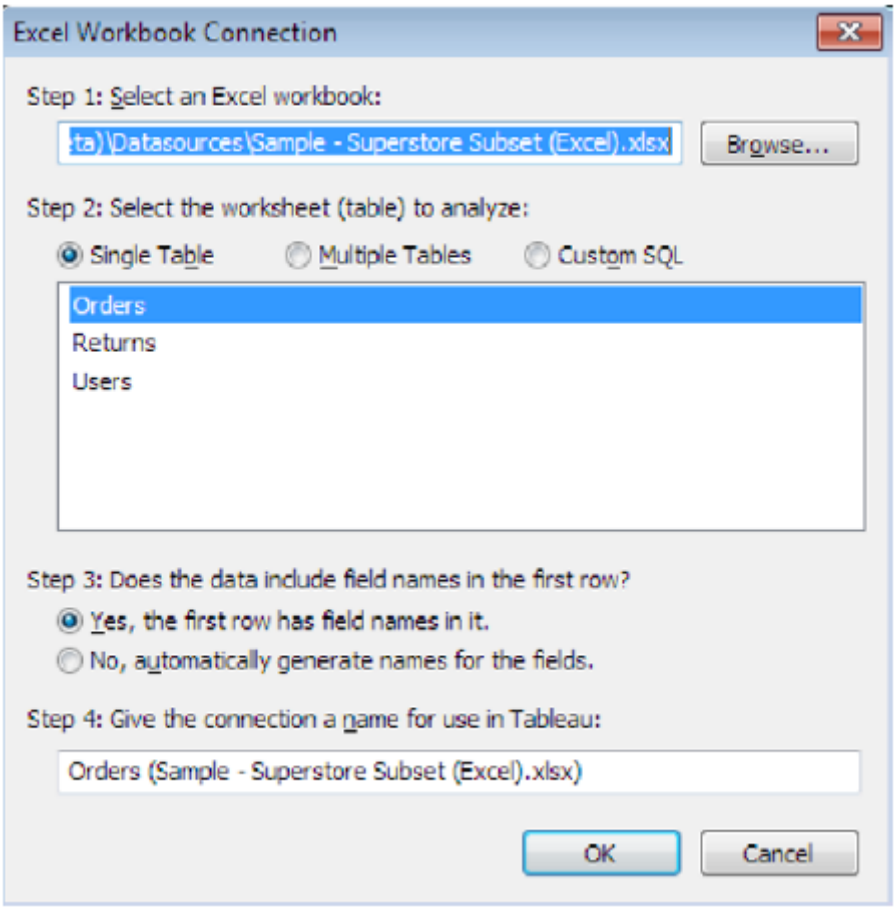


Fig. 14.2 The Connection Window

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.

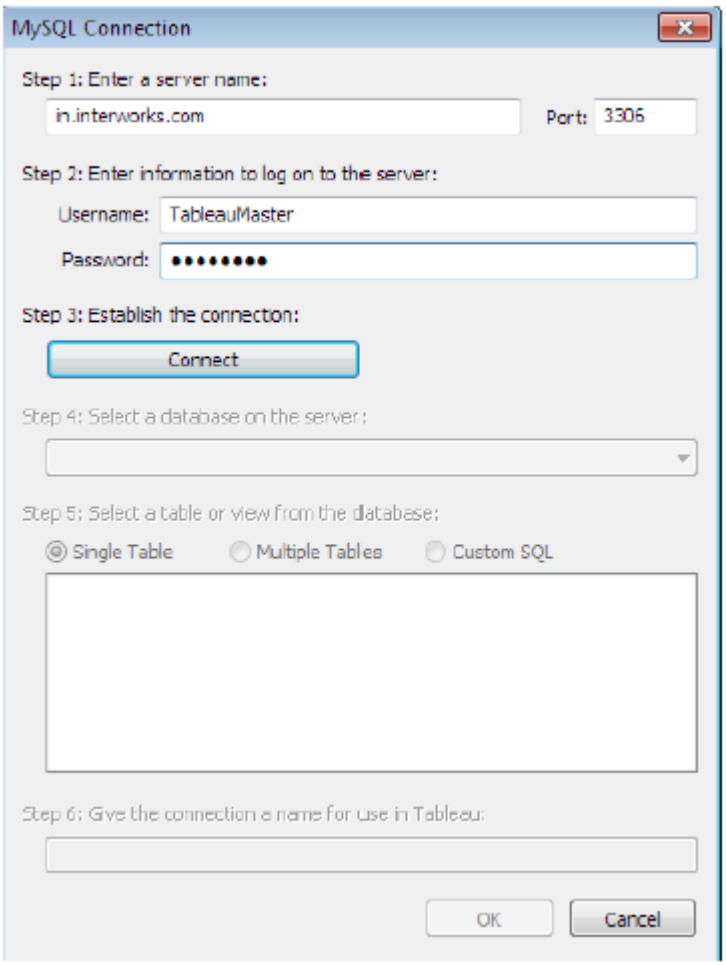


Fig. 14.3 Database connection window

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 datasources. 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 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 centre-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.

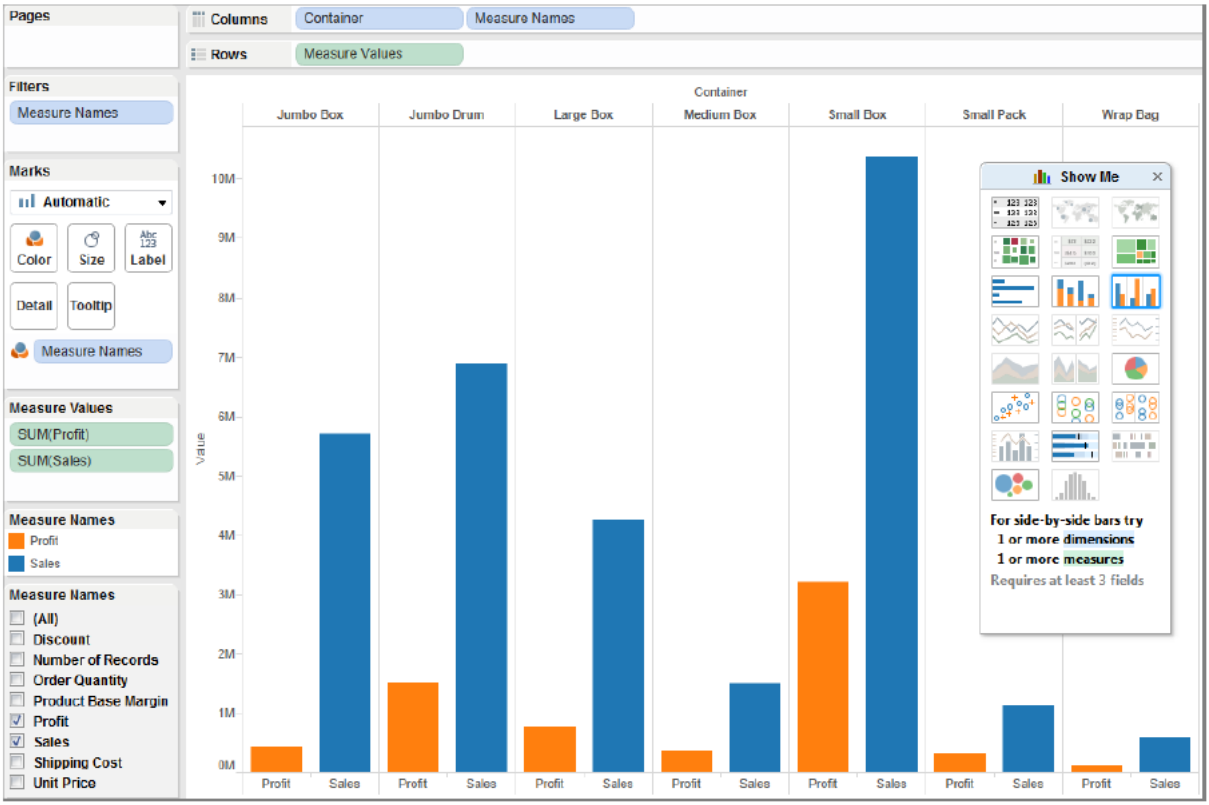


Fig 14.5 Measure values bar chart

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 centre 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.

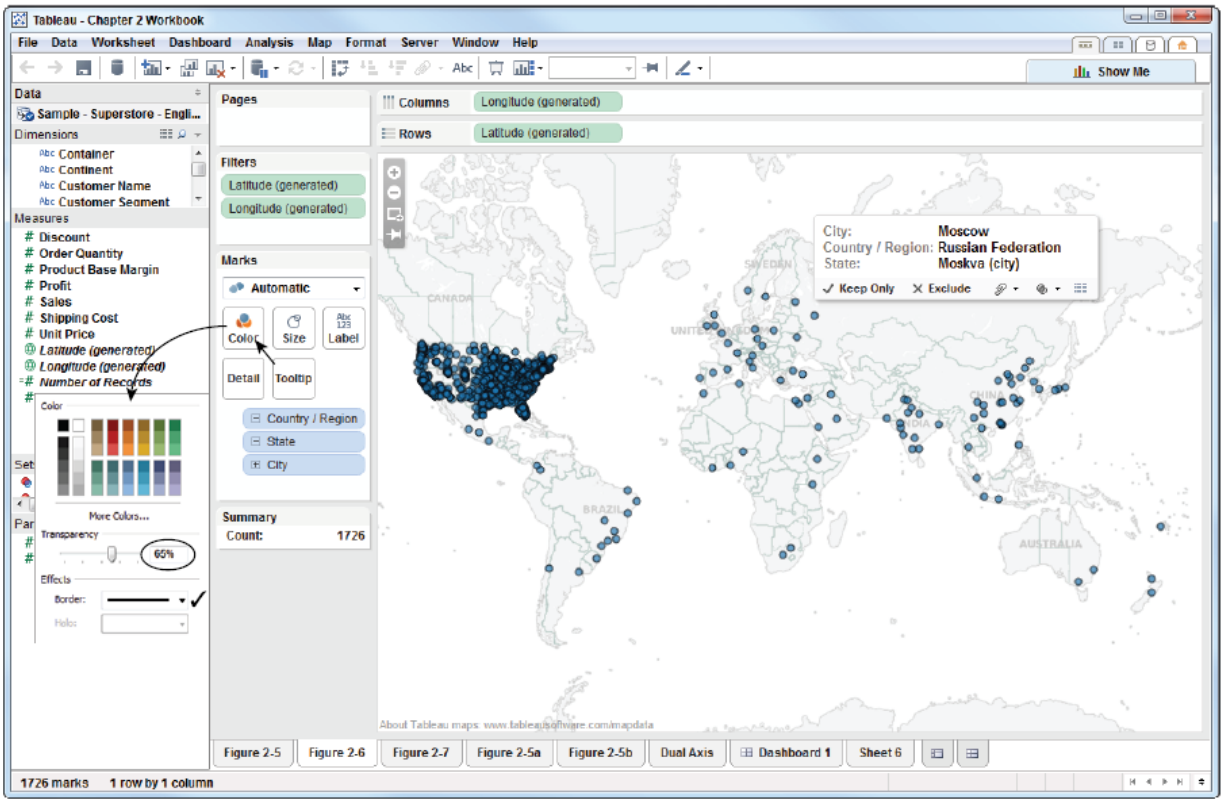


Fig 14.6 Latitude and longitude generated measures

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.

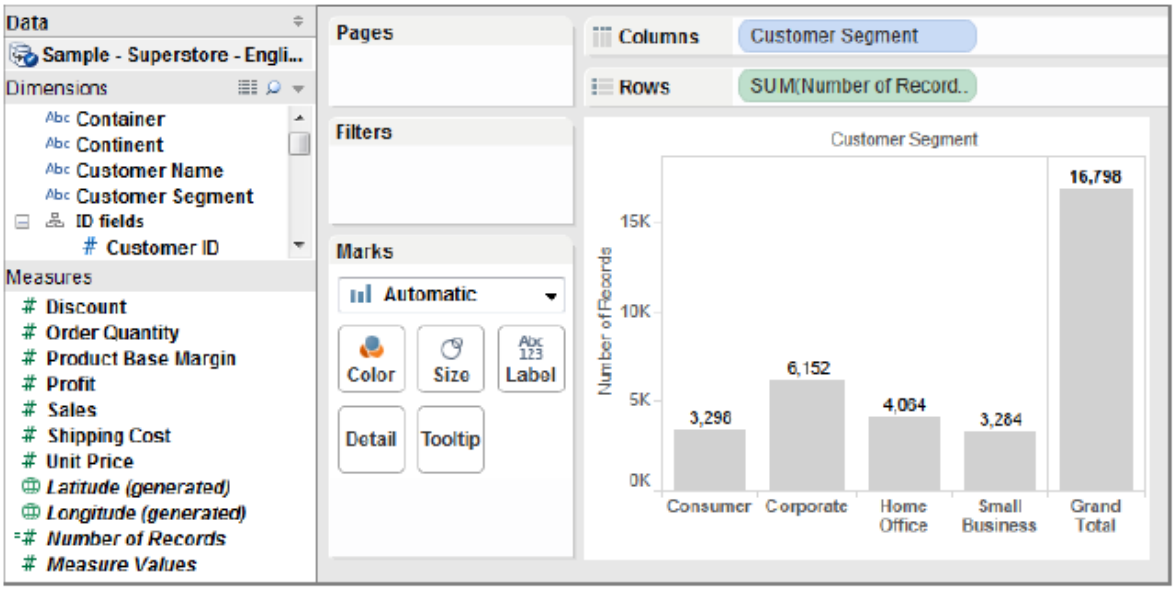


Fig 14-7 Number of records

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.

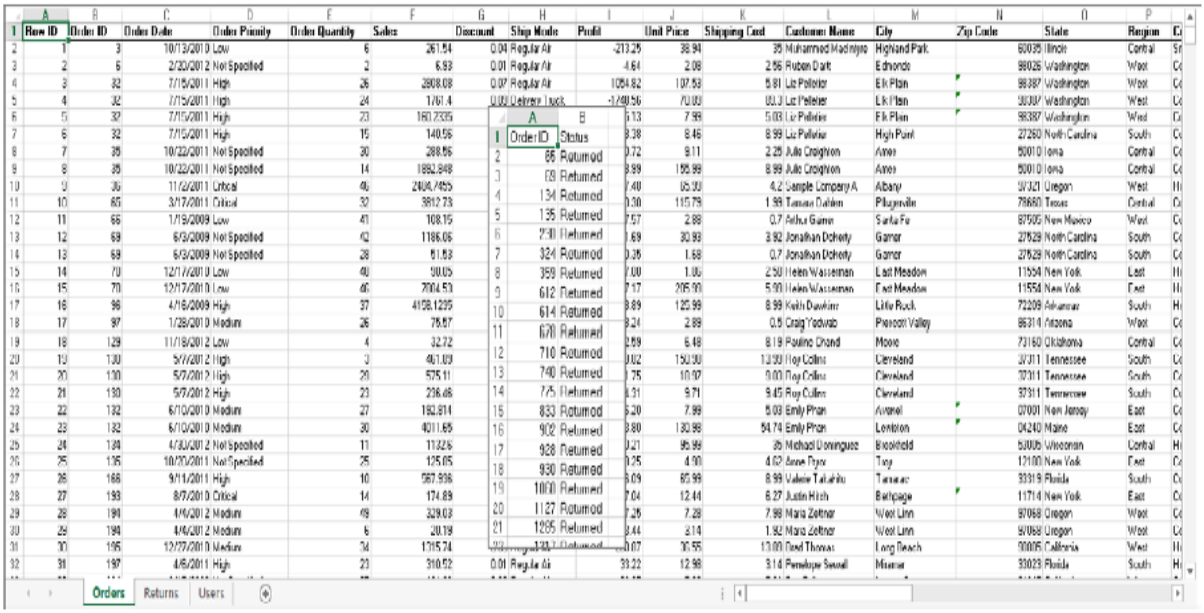


Fig 14-8 Superstore orders and return tables

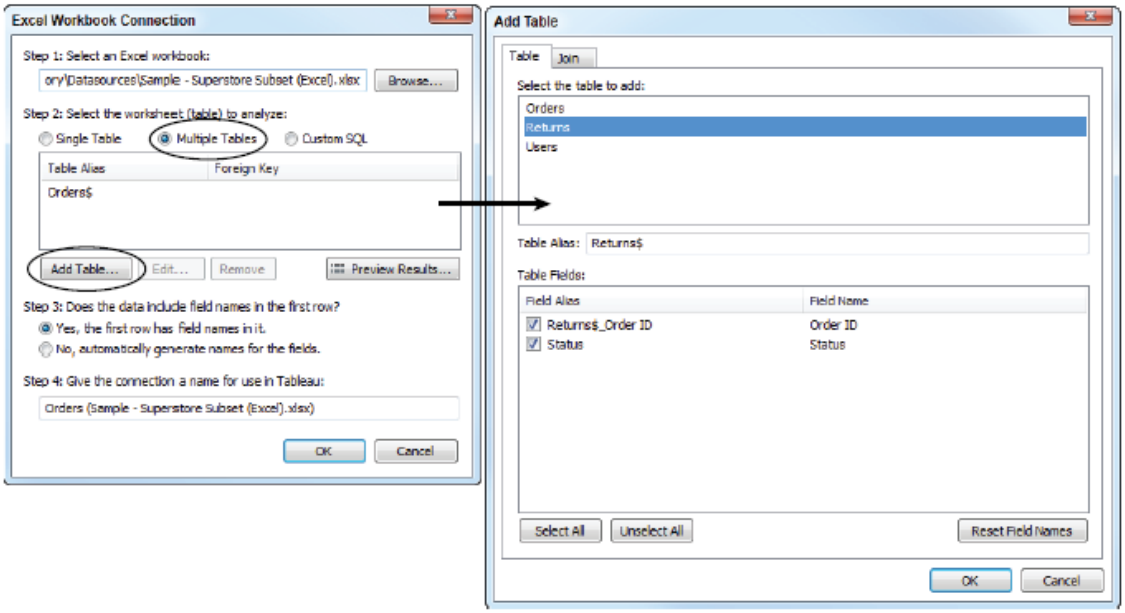


Fig 14-9. Joining multiple tables

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.

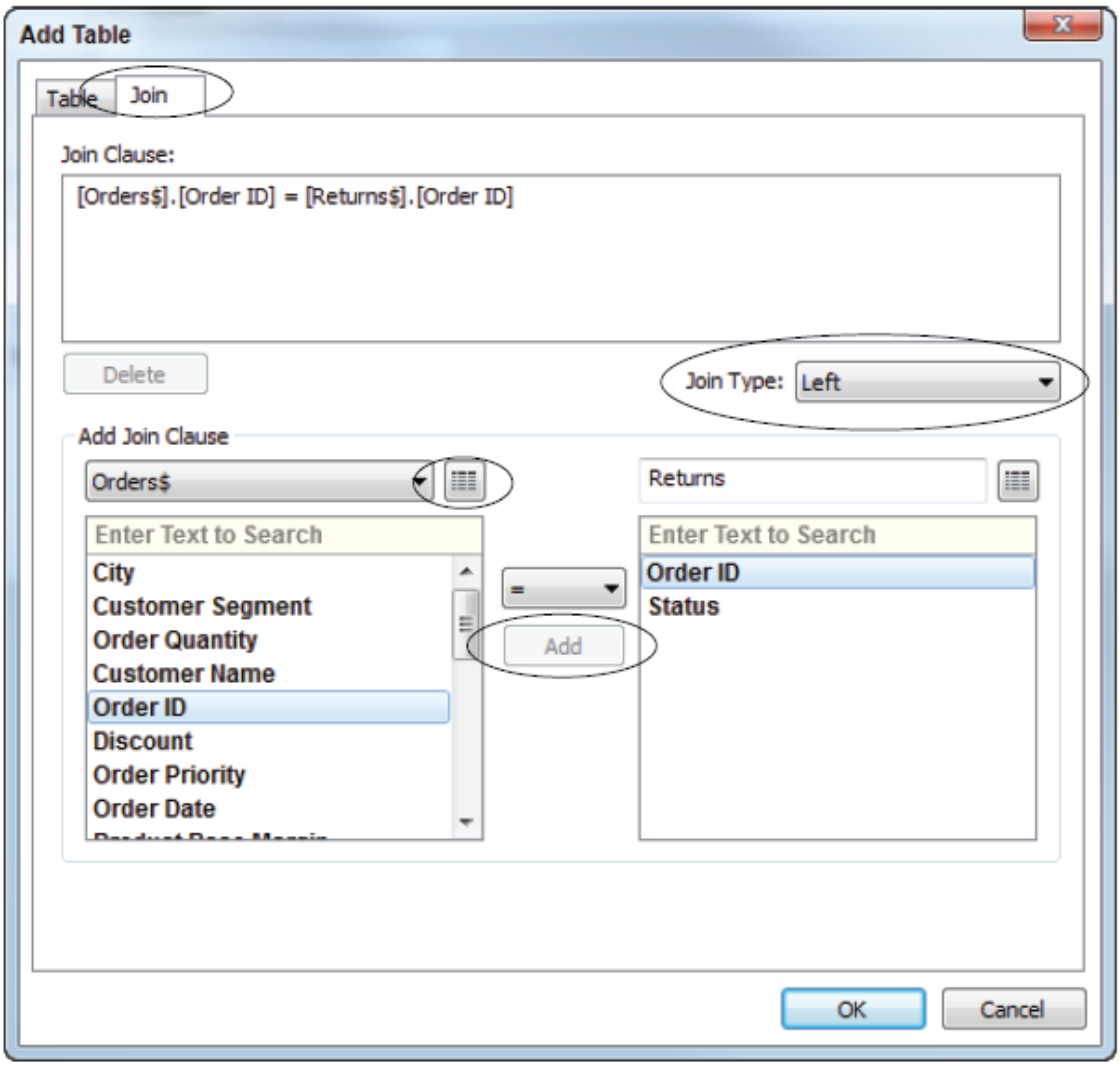


Fig. 14-10 Joining tables in Tableau

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.

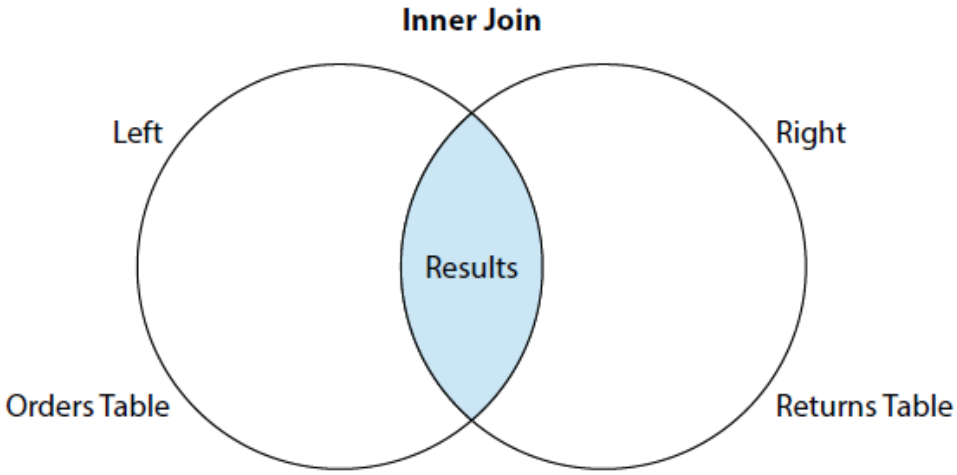


Fig.14-11 The inner join

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.

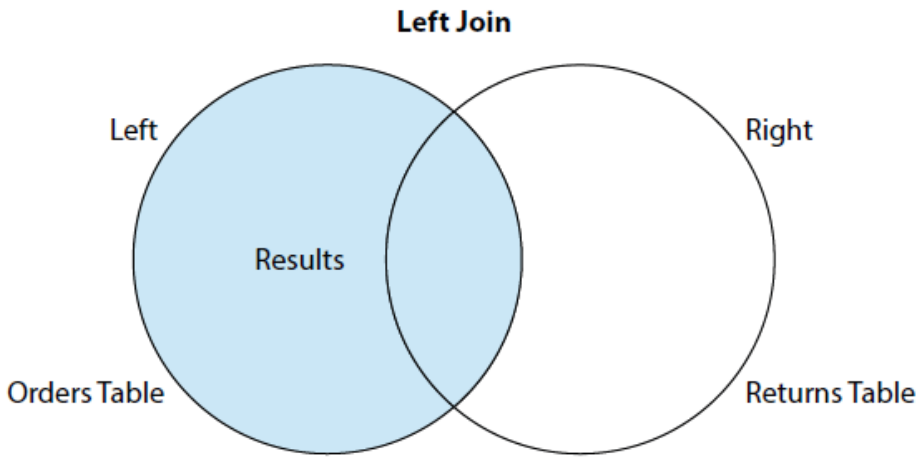


Fig 14-12. The left outer join

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

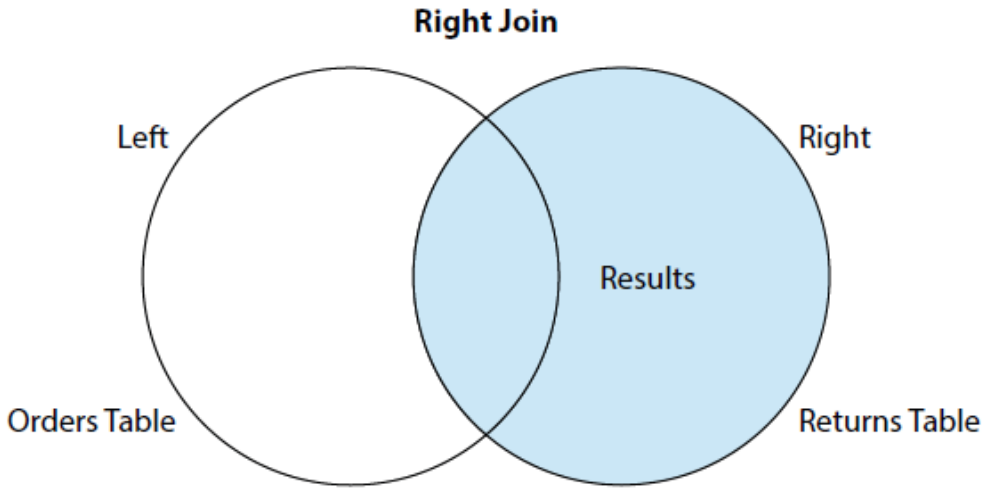


Fig 14-13 The Right outer join

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	Population
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.

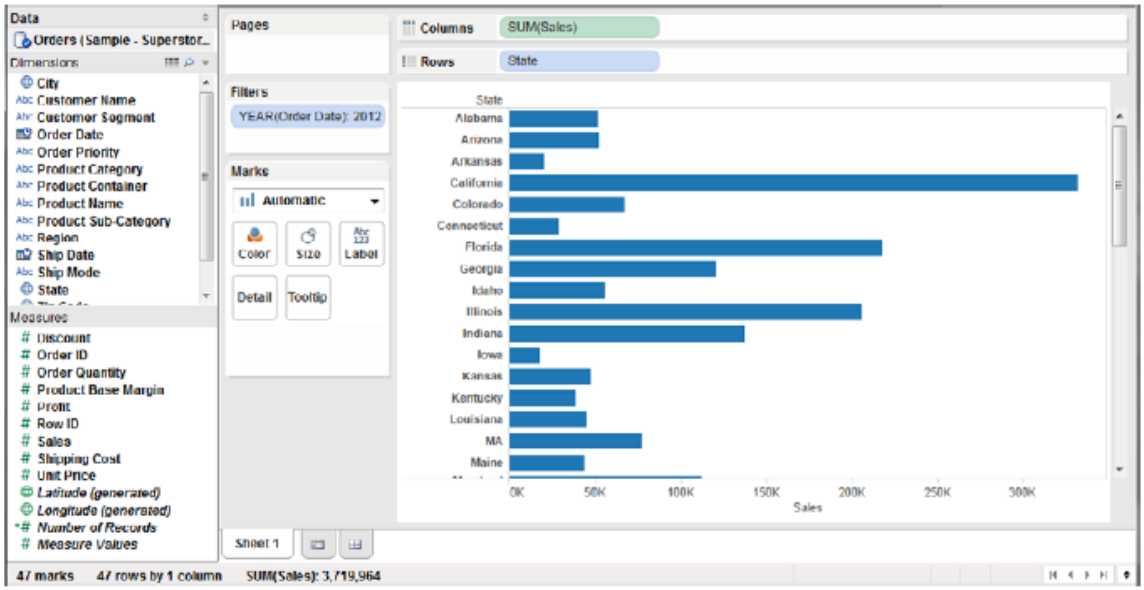


Fig. 14-18 Sales by state

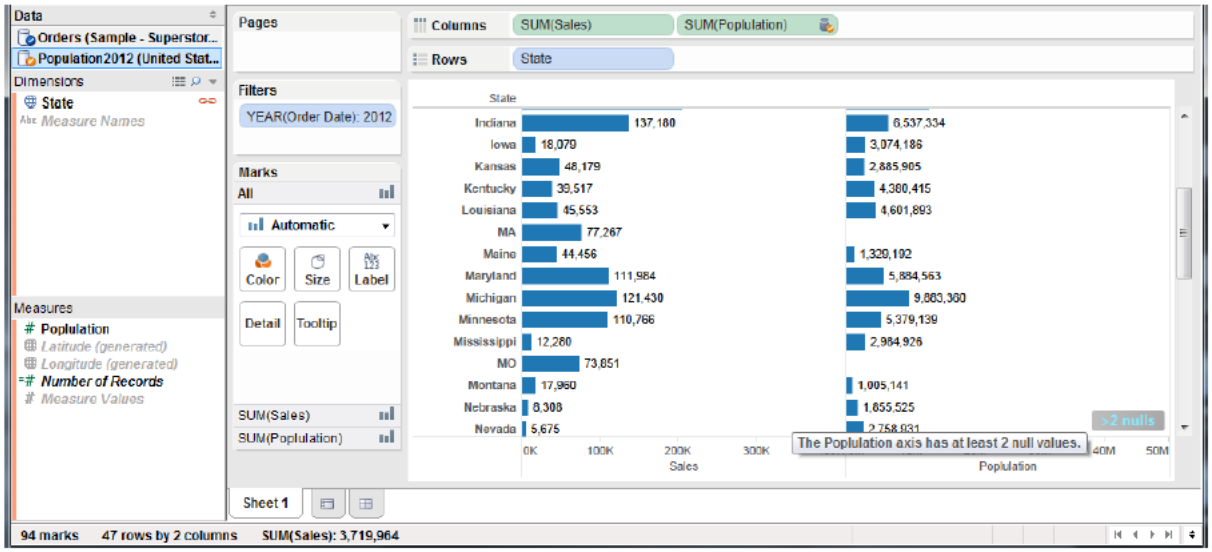


Fig 14-19 population data blended with superstore

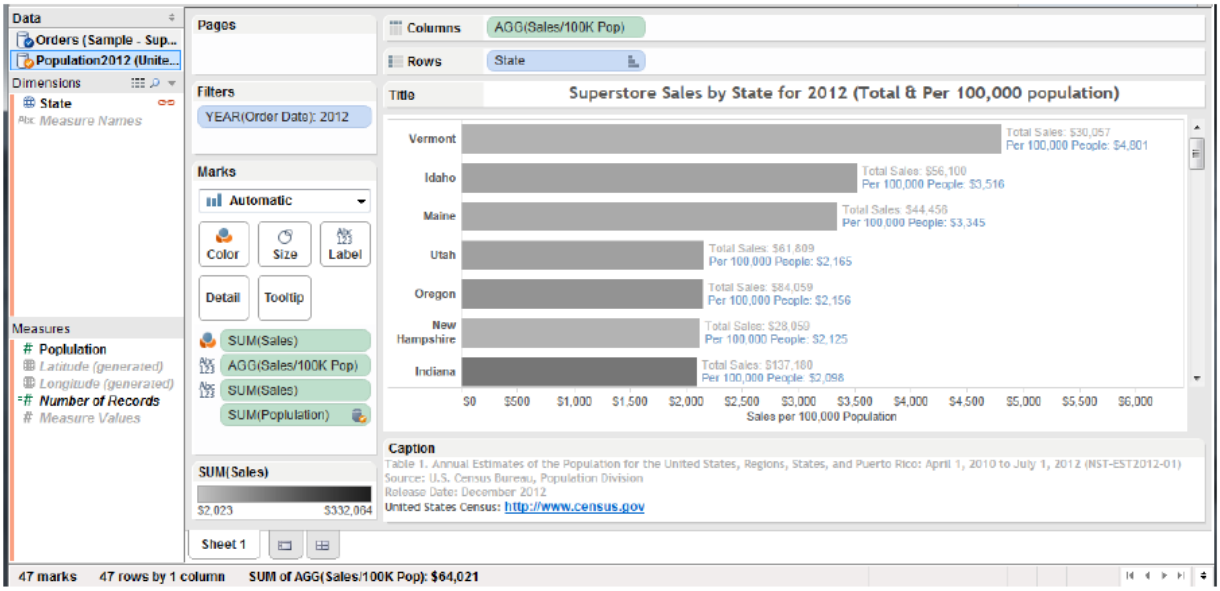
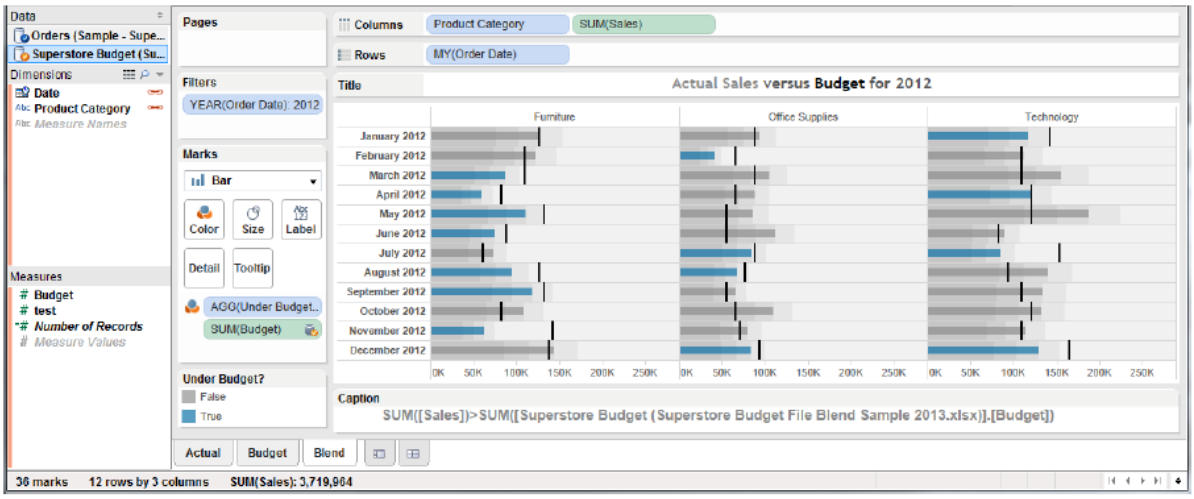


Fig 14-20 Bar chart using blended data

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*

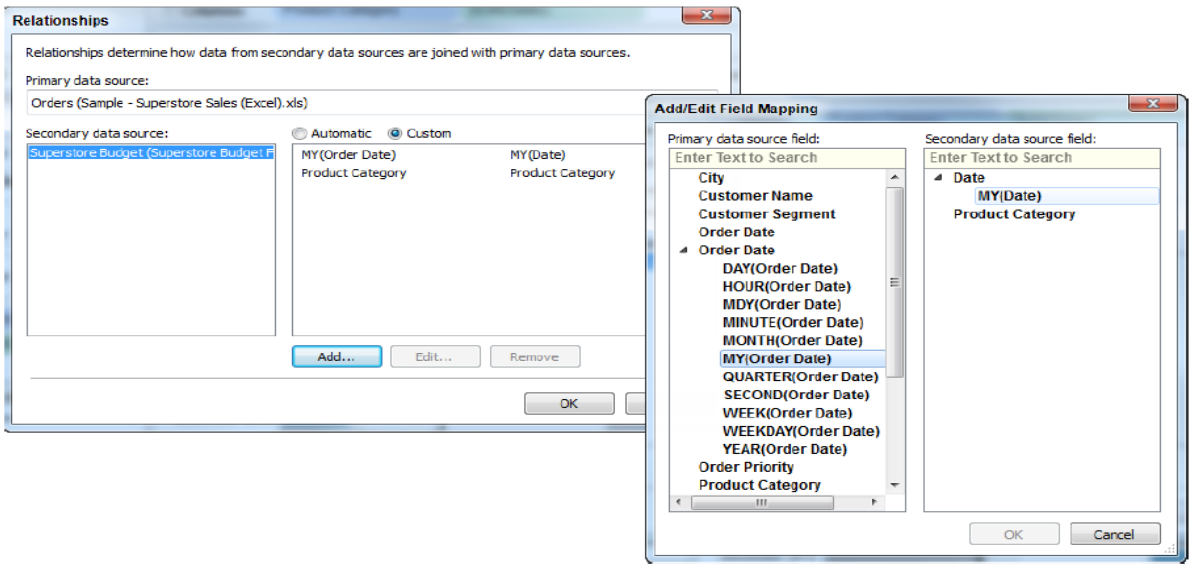


Fig14-22 *The edit relationships menu*

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 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

1. What is geocoding in Tableau?
2. What are the advantages of a data extract?
3. List different file types in Tableau.

4. 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 automati cally generate the longitude and latitude values for the centre 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 SELF ASSESSMENT QUESTIONS

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

15.1 Fast and Easy Analysis

15.2 *Show Me*

15.3 Trend Lines and Reference Lines

15.4 Sorting Data in Tableau

15.5 Enhancing Views with Filters Sets, Groups, and Hierarchies

15.6 Check your progress

15.7 Summary

15.8 Keywords

15.9 Self Assessment Questions

15.10 References

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

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.

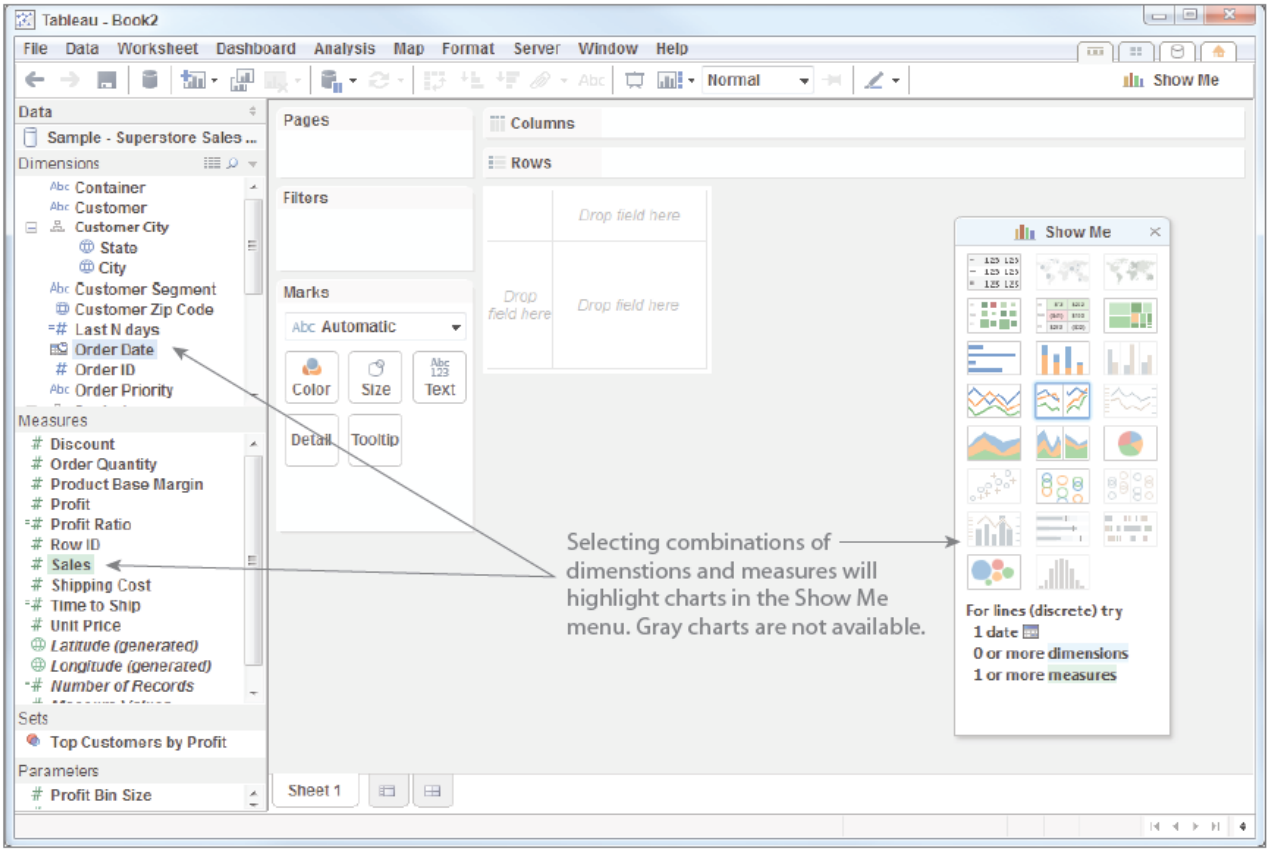


Fig 15-1. *Show Me* displays chart options

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.

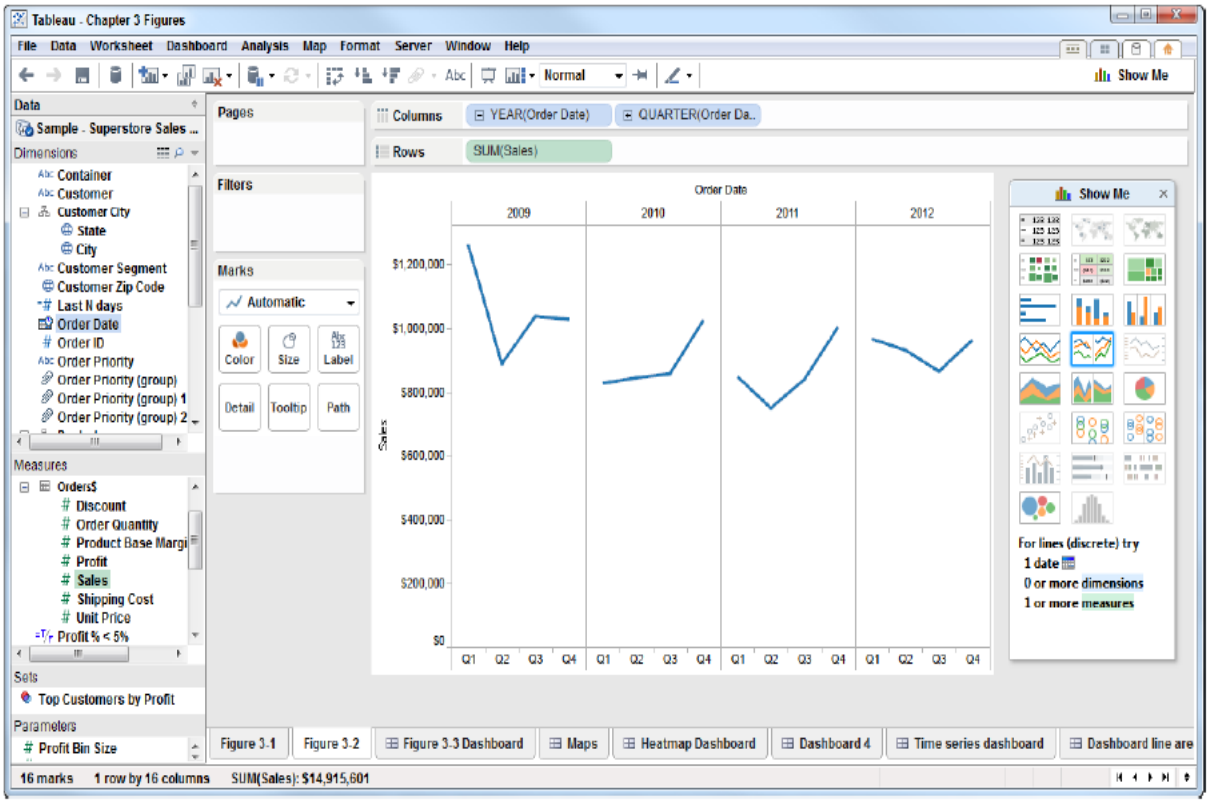


Fig 15-2. Discrete data time series chart

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.

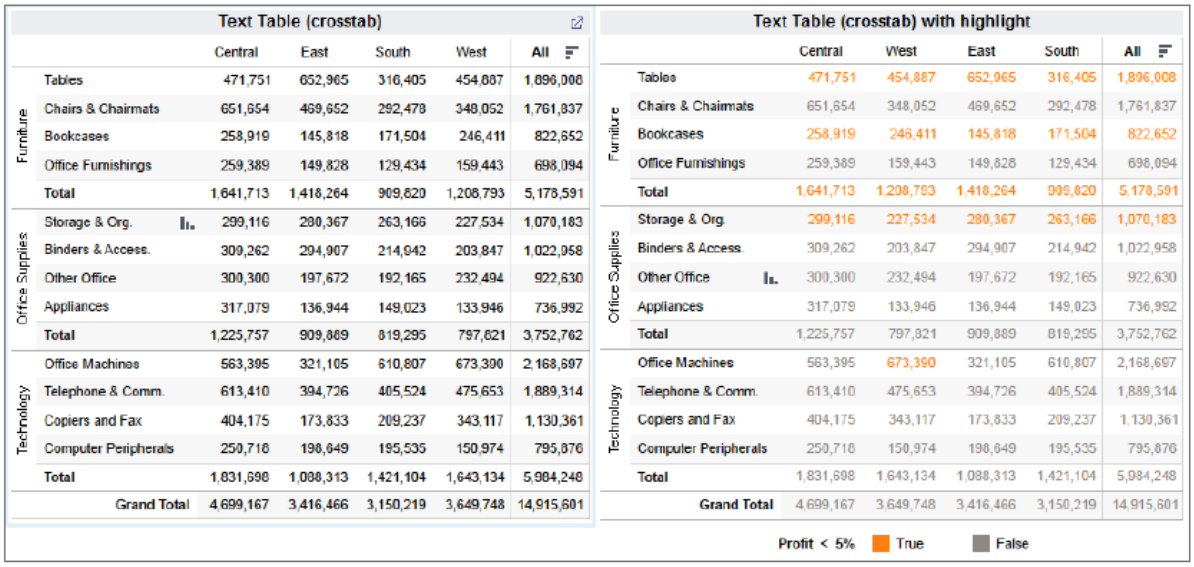
Chart Types Provided by the Show Me Butt on

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	250,716	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	250,716	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.

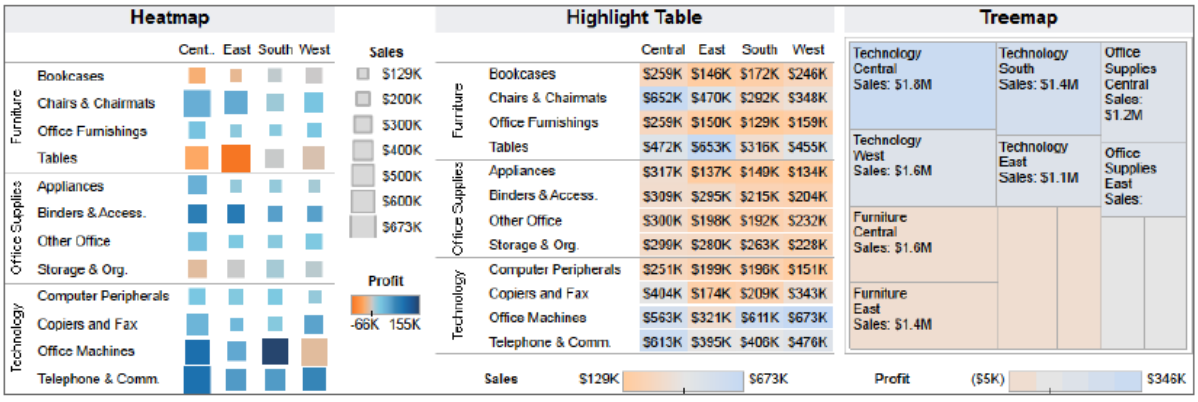


Fig 15-5. Heatmap, highlight table and treemap

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.

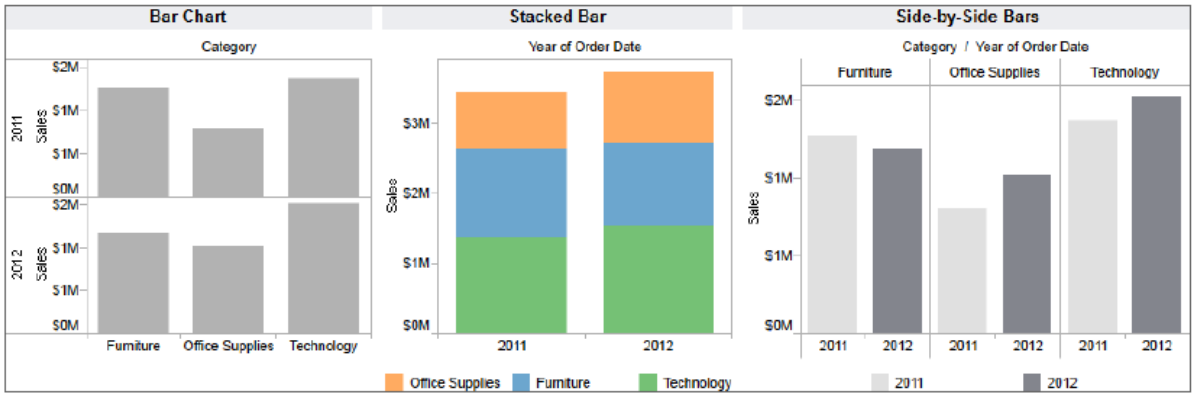


Fig 15-6. Bar chart, stacked bar chart and side by side bar chart

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*

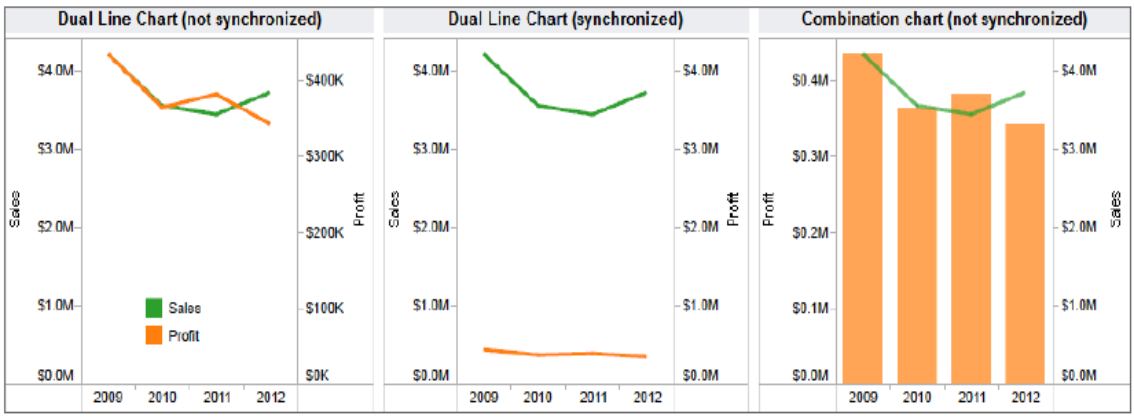


Fig 15-7 *Time series presented using continuous time*

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?

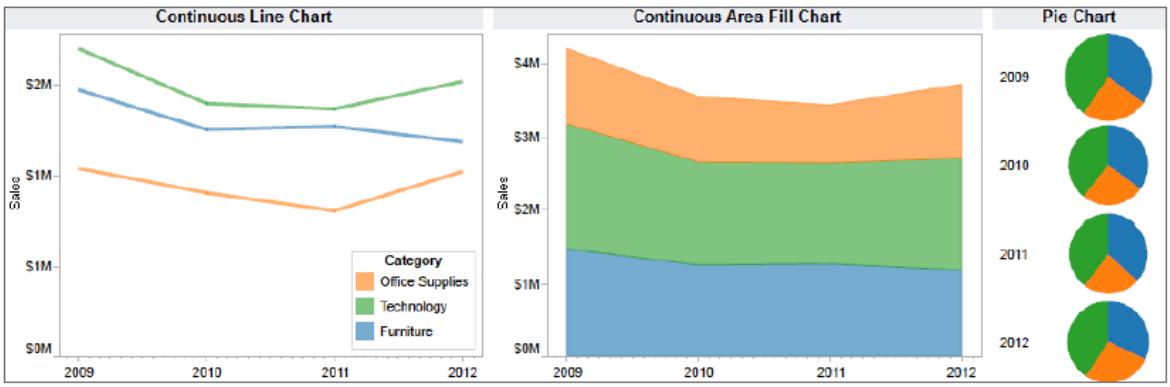


Fig 15-8 Continuous line chart, area fill chart, pie charts

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

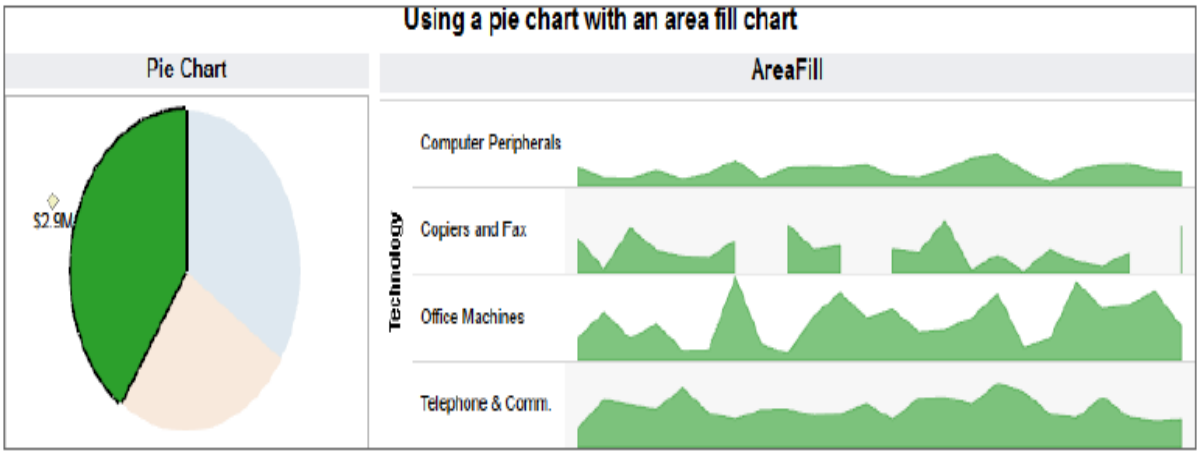


Fig 15-9. Pie chart and area fill chart

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.

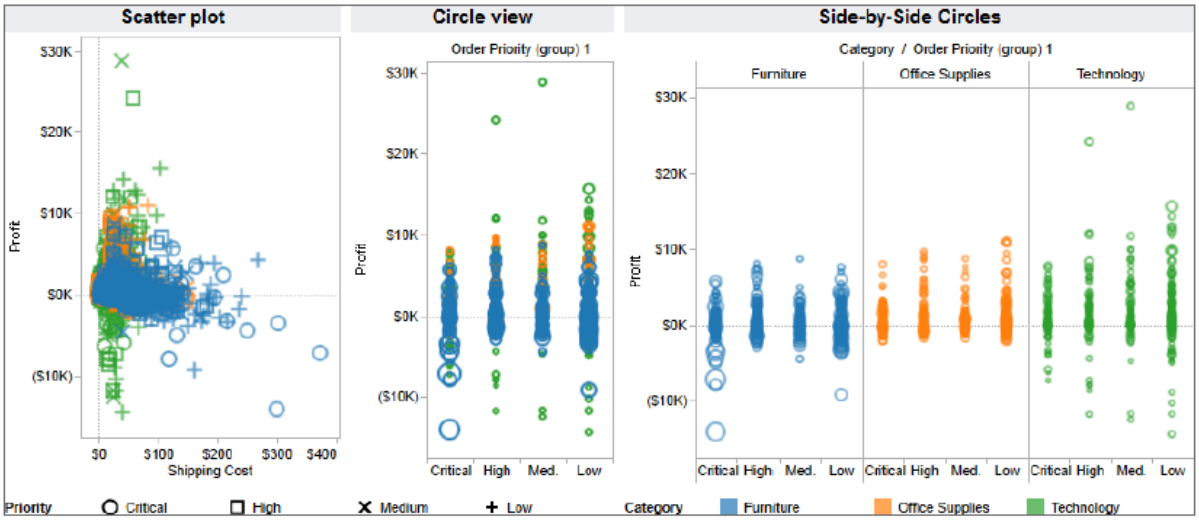


Fig 15-10 scatter plot, circle view, side-by-side circle view

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.

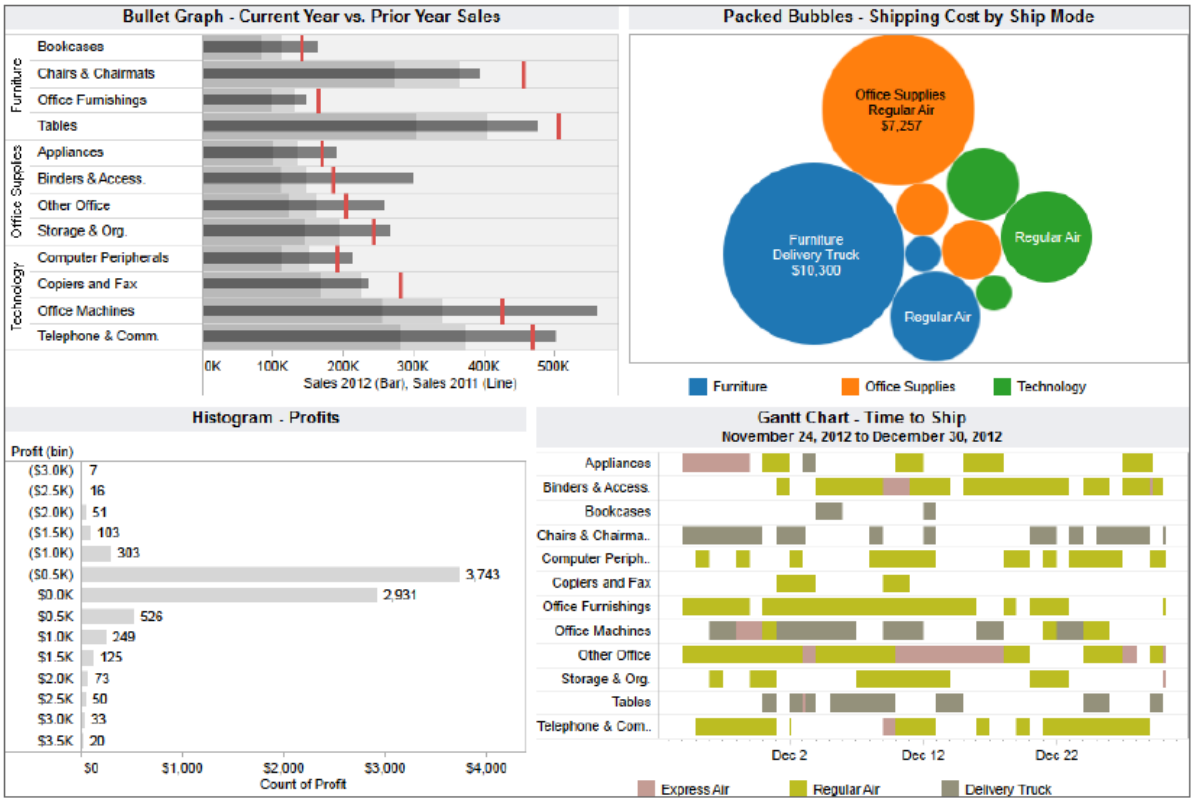


Fig 15-11. The last four Shoe Me Charts

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.

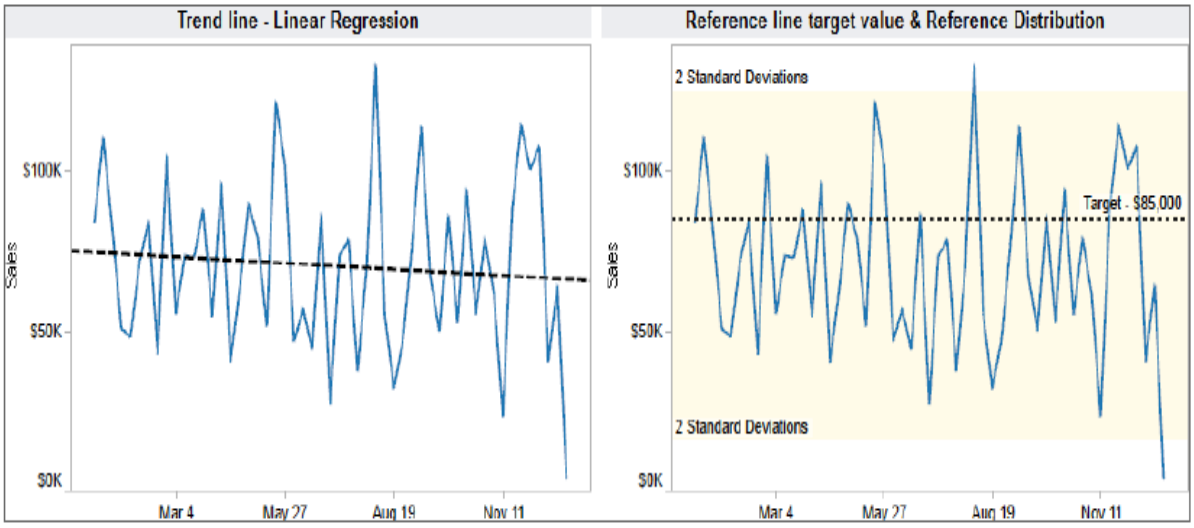


Fig. 15-12. Trend line and reference line

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.

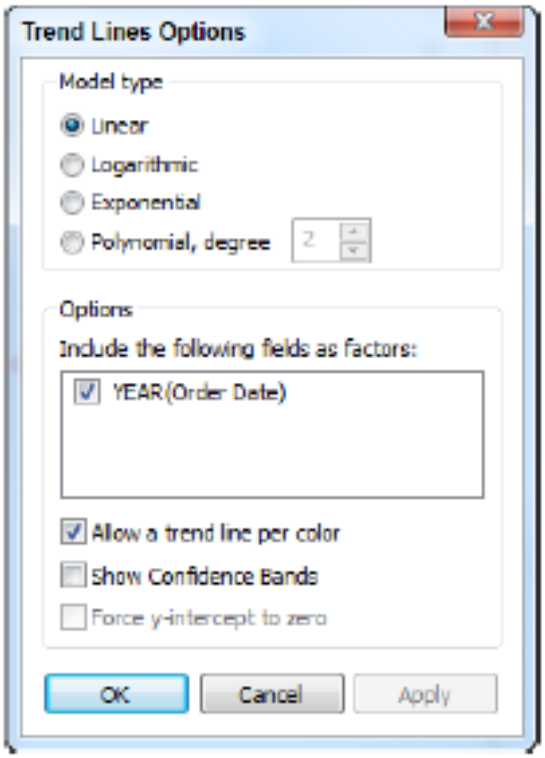


Fig 15-14. The trend line options menu

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.

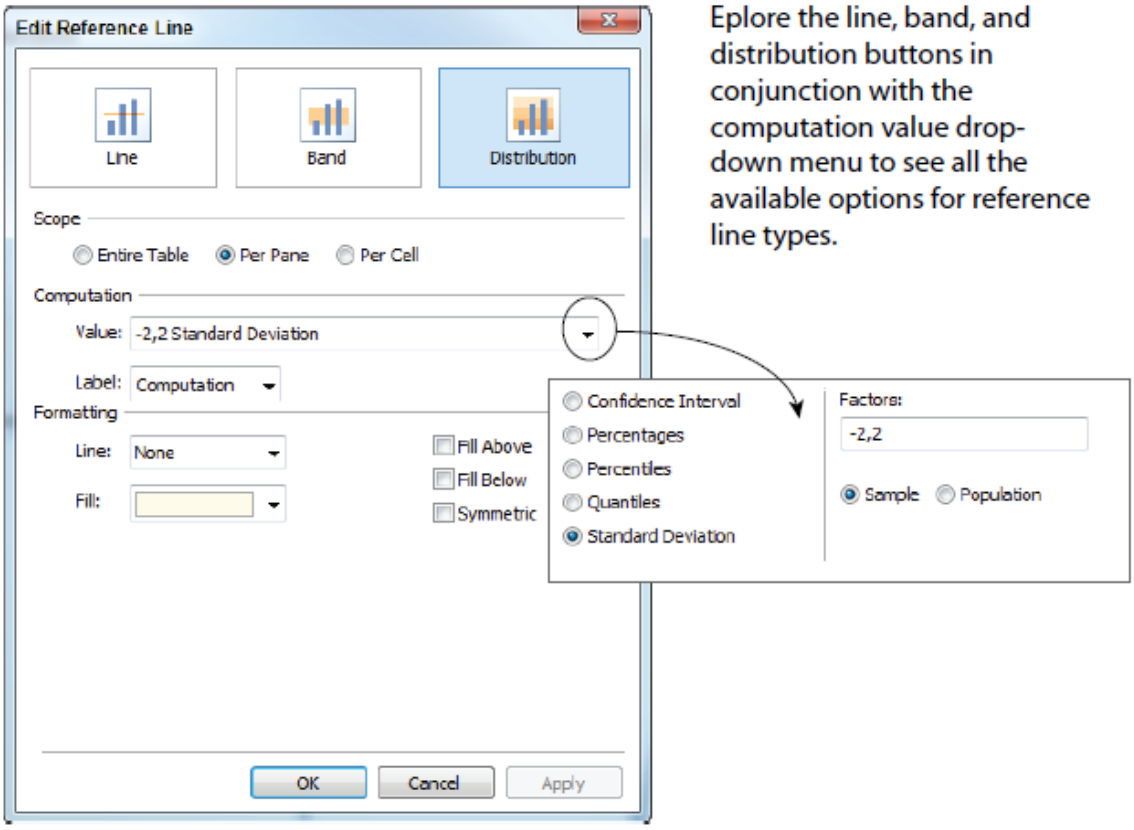


Fig 15-15 Reference line menu standard deviation reference lines

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.

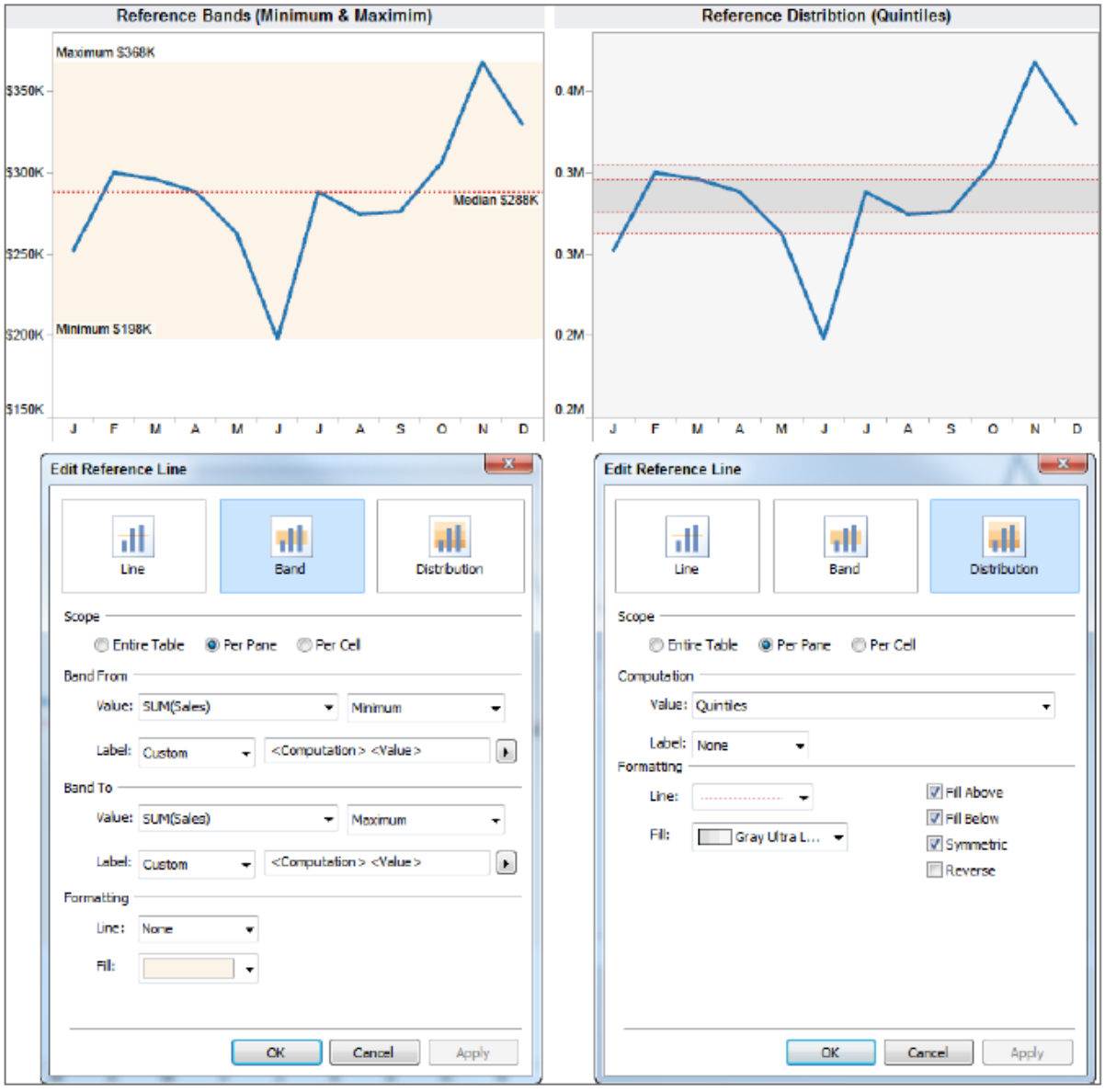


Fig 15-16 Reference bands and reference distributions

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.

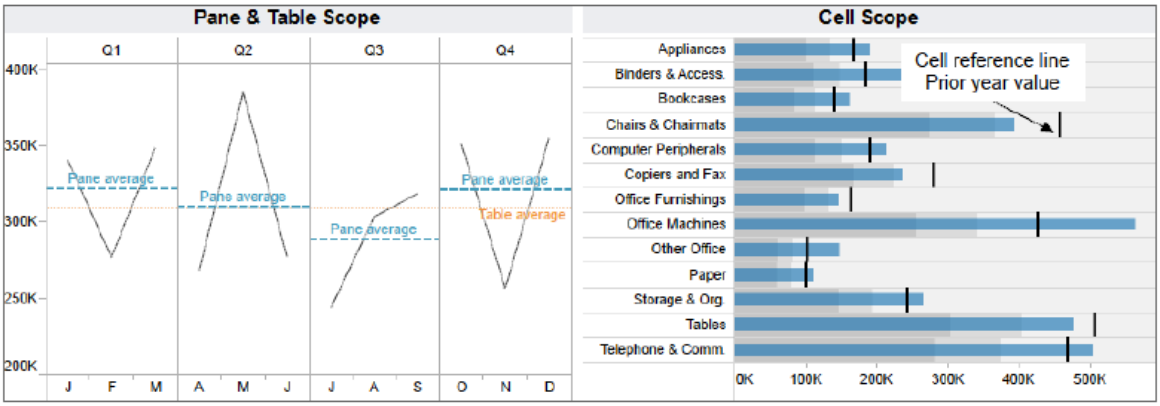


Fig. 15-17 Reference lines using entire table, pane and cell

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.

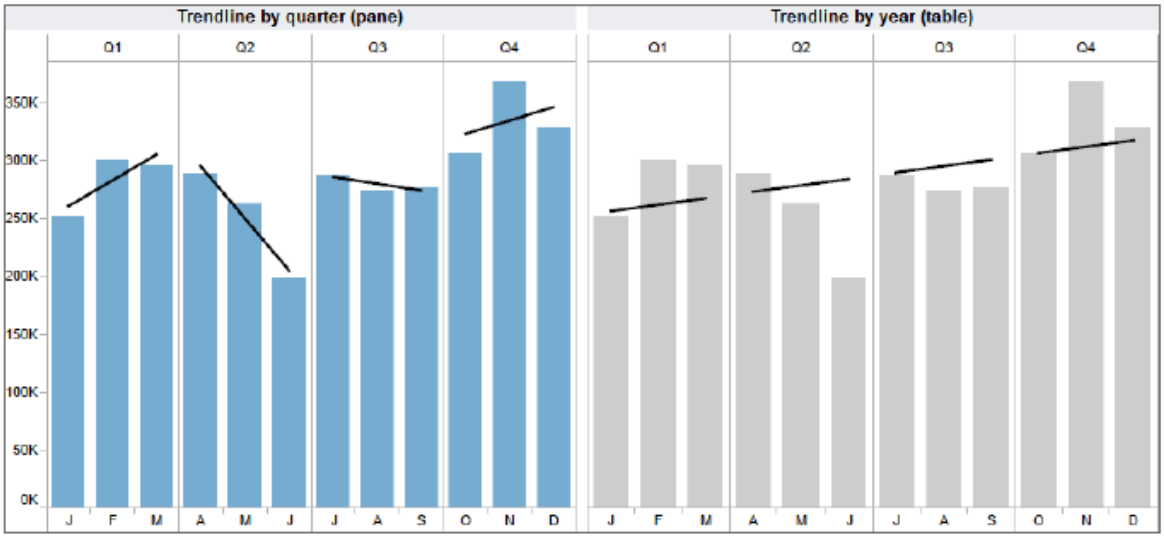


Fig 15-18 Trend lines using pane and table

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.

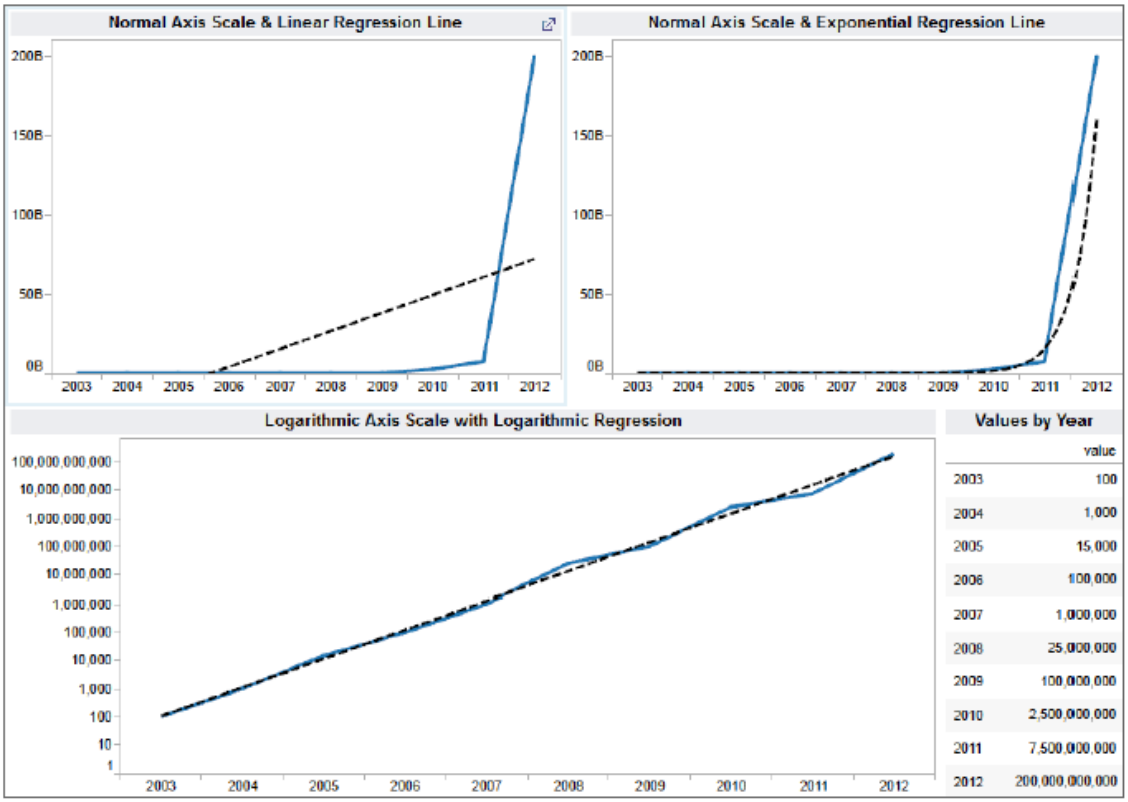


Fig 15-19 Rapidly increasing time series

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.

Man ual 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.

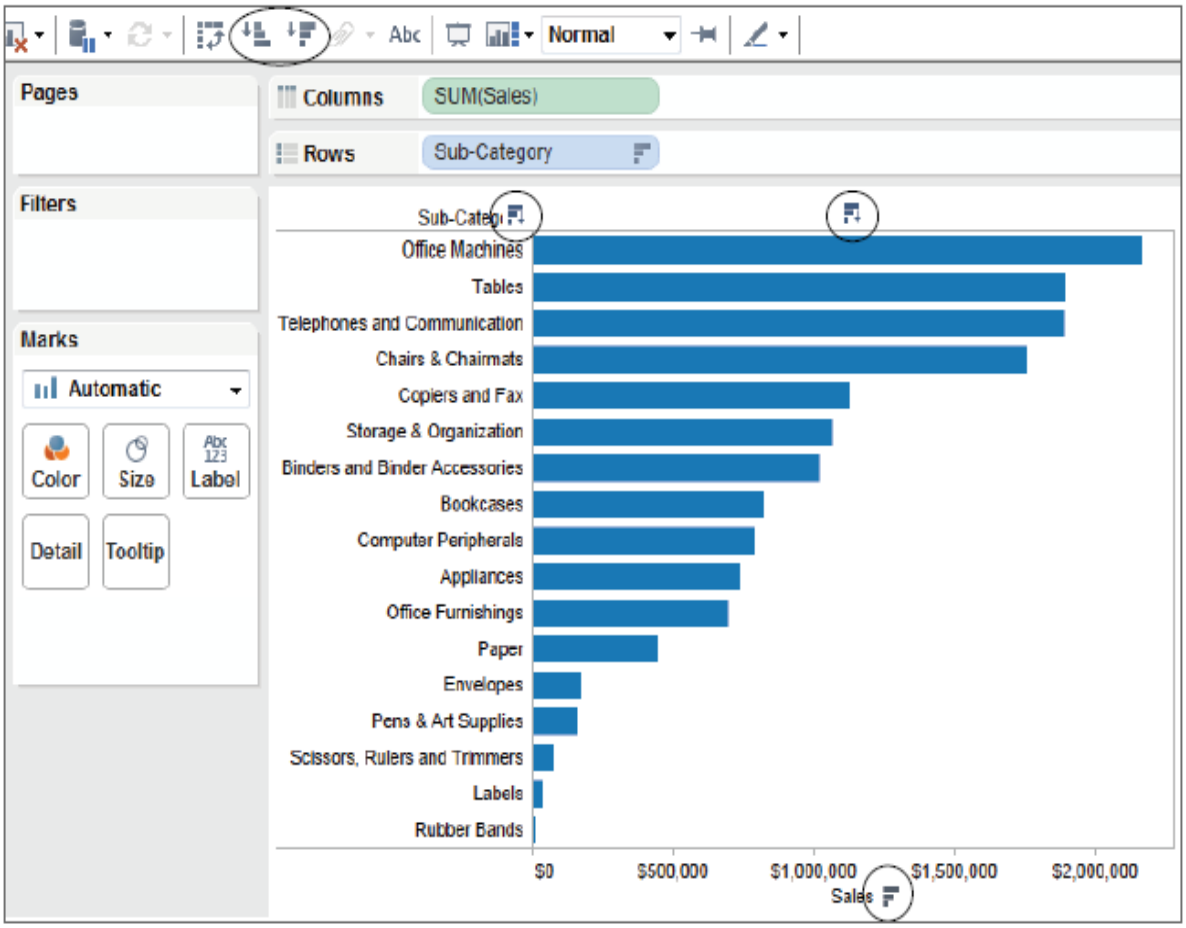


Fig. 15-20. Manual sorting applied from the toolbar icon

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.

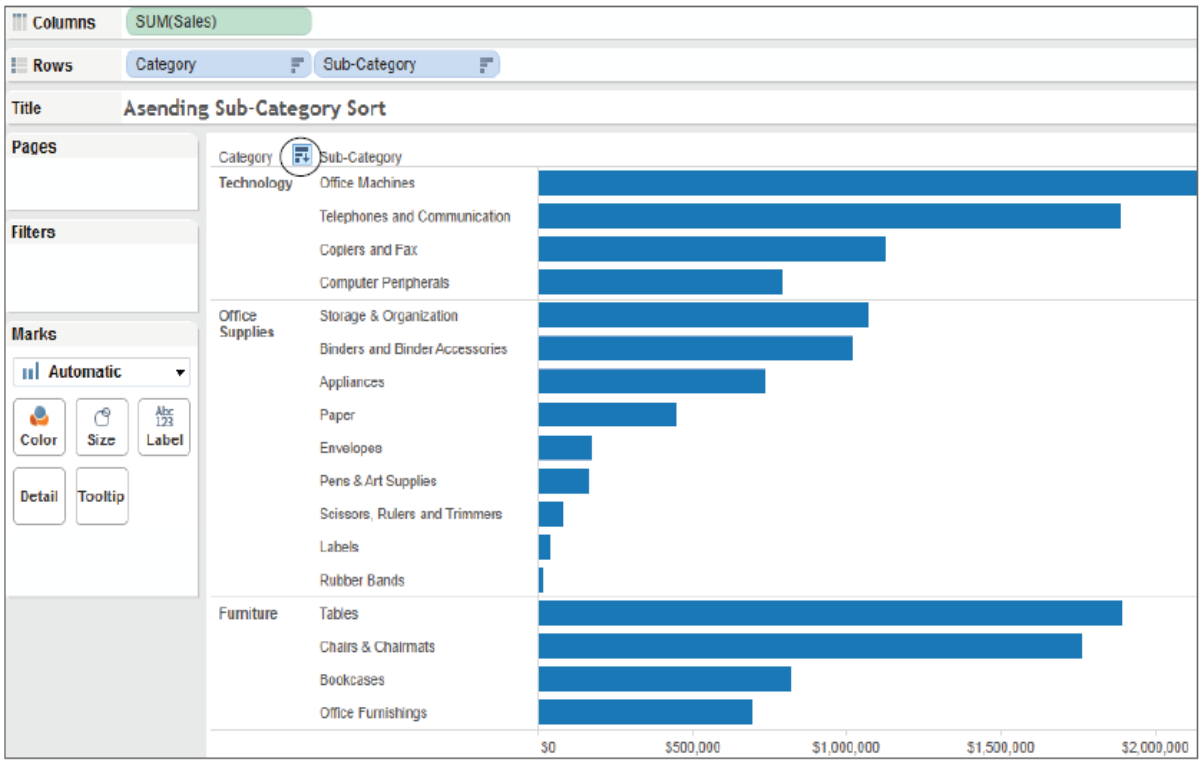


Fig. 15-21 Ascending sort by category

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.

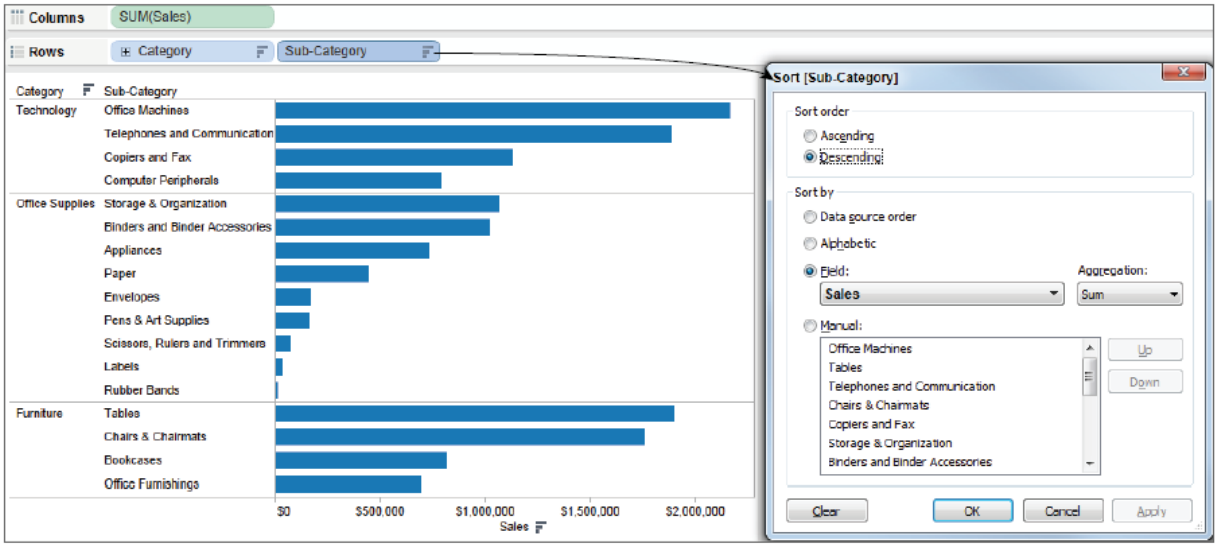


Fig. 15-22 Resorted computer peripherals and sort menu

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.

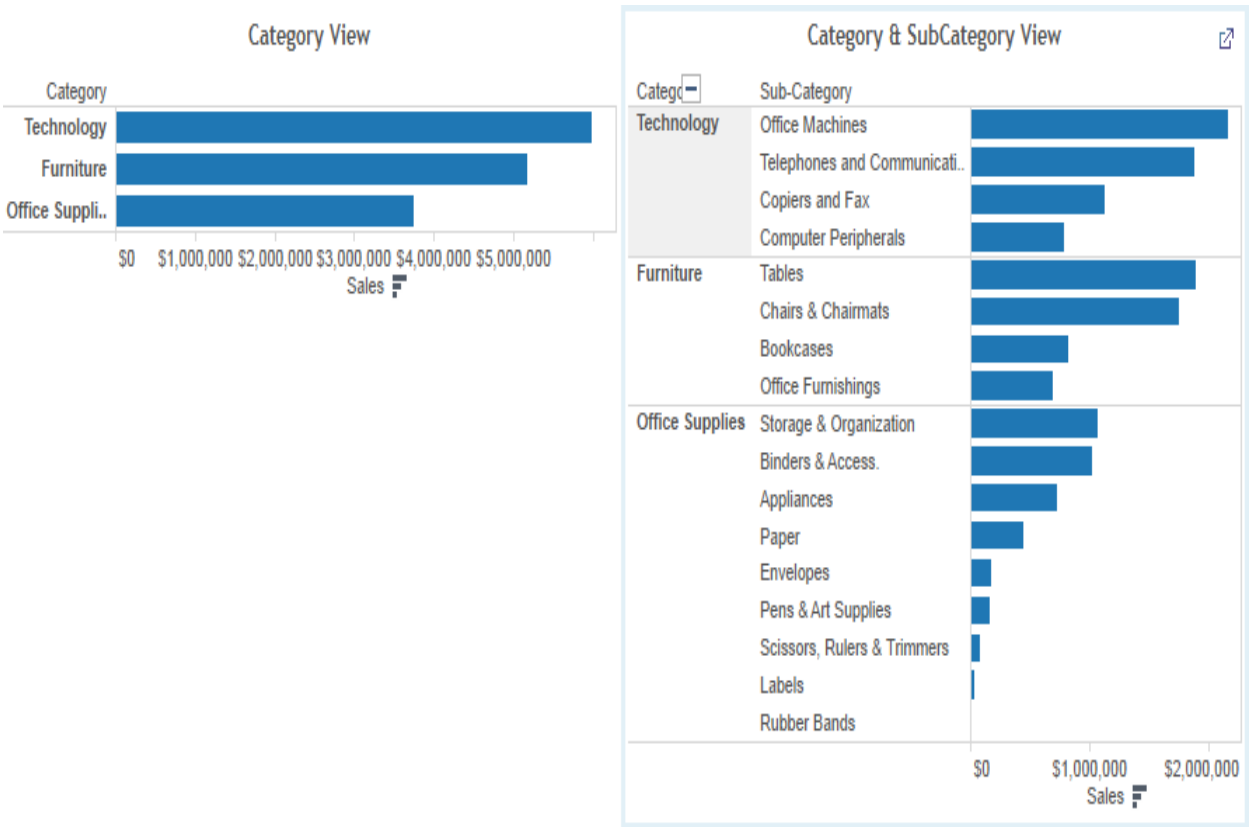


Fig 15-24. Hierarchy using category and subcategory

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.

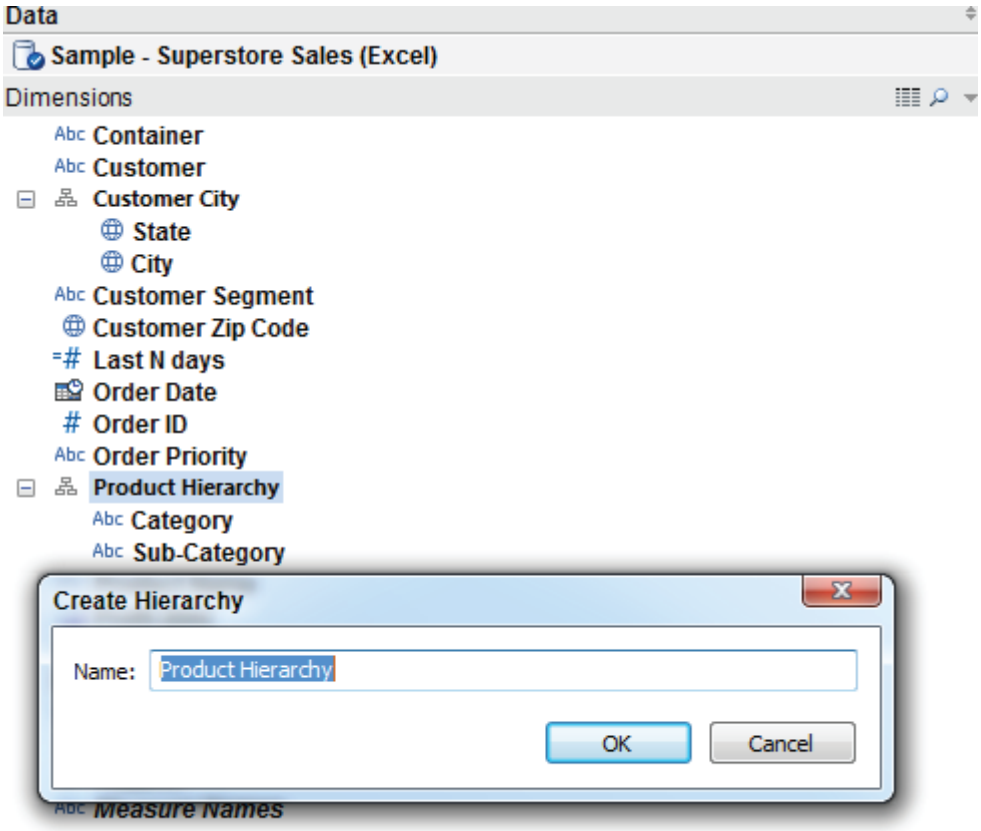


Fig 15-25. Making a custom hierarchy

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.

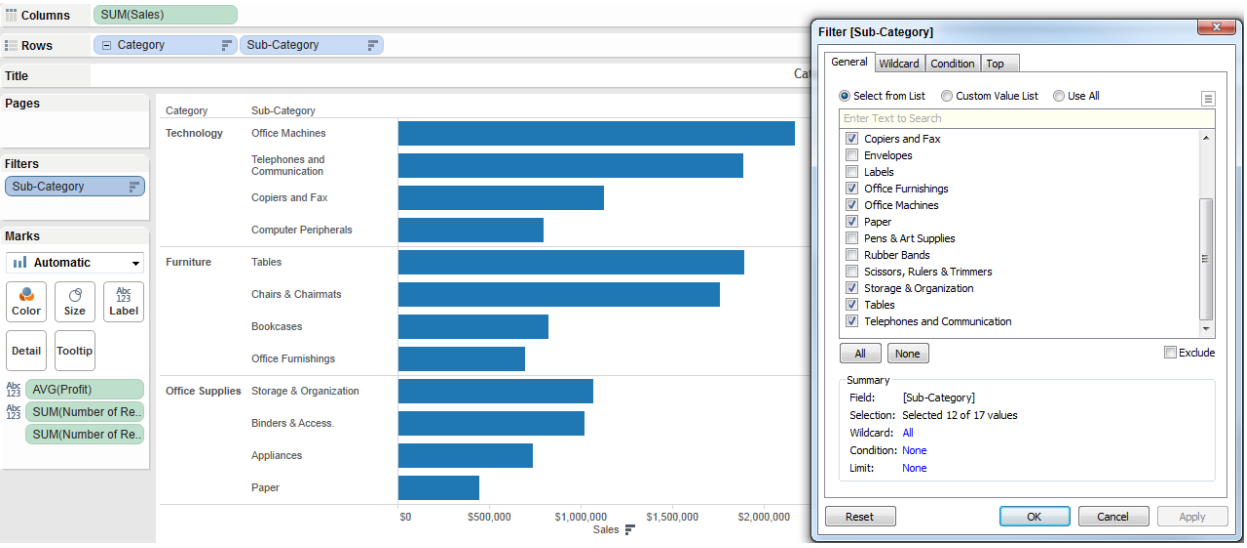


Fig.15-26. Applying a filter via the filter shelf

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.

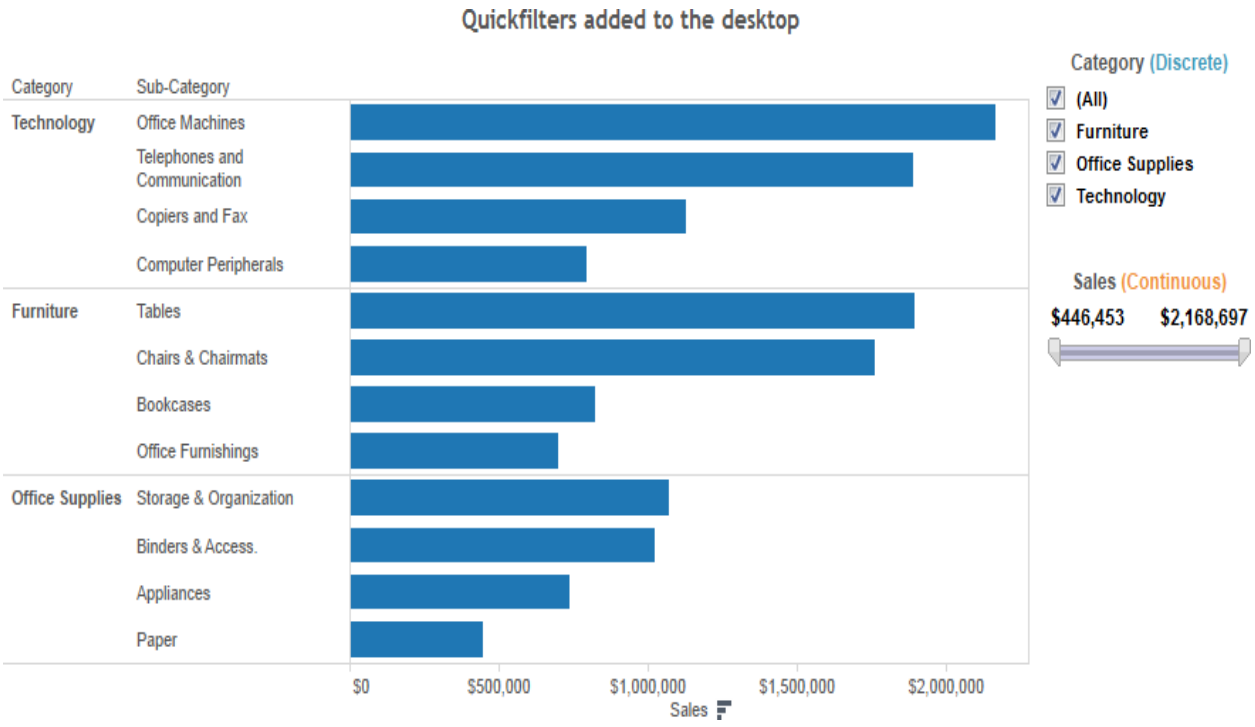


Fig.15-27. Adding a quick filters to the Desktop

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.

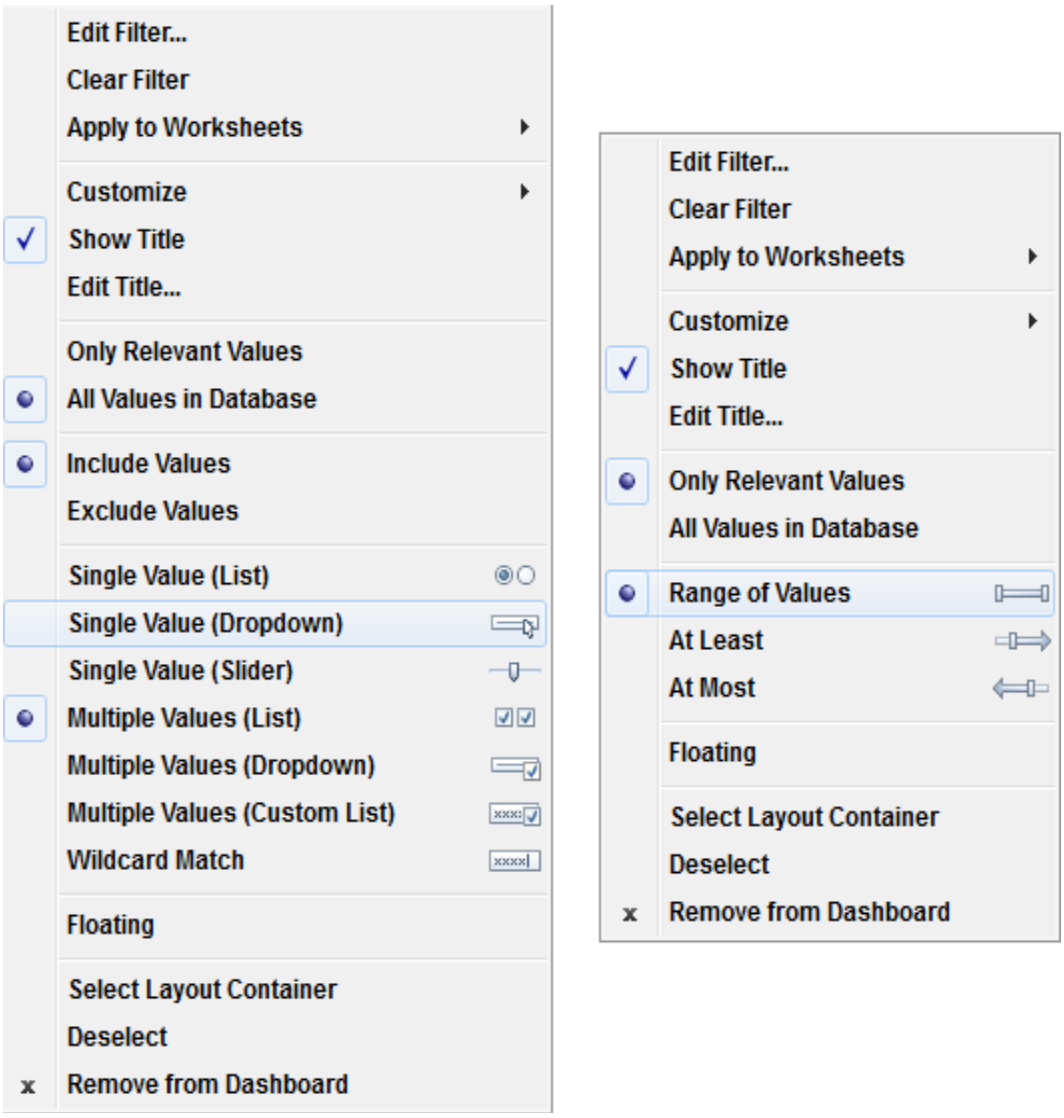


Fig. 15-28. Editing quick filter types

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 centreing 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).

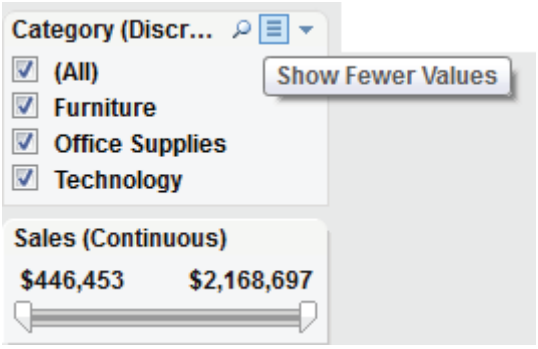


Fig 15-29 Including all or relevant values

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.

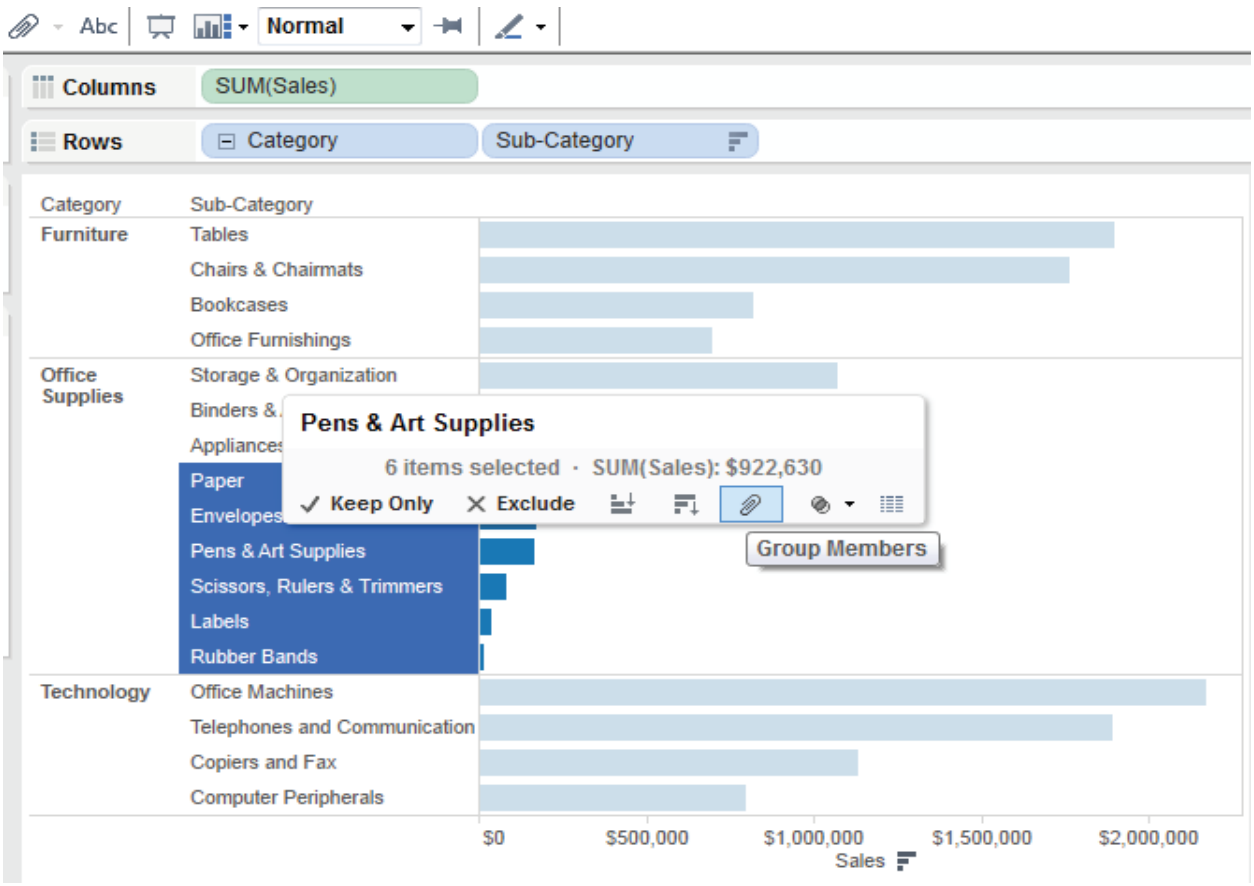


Fig 15-30 Grouping from headers

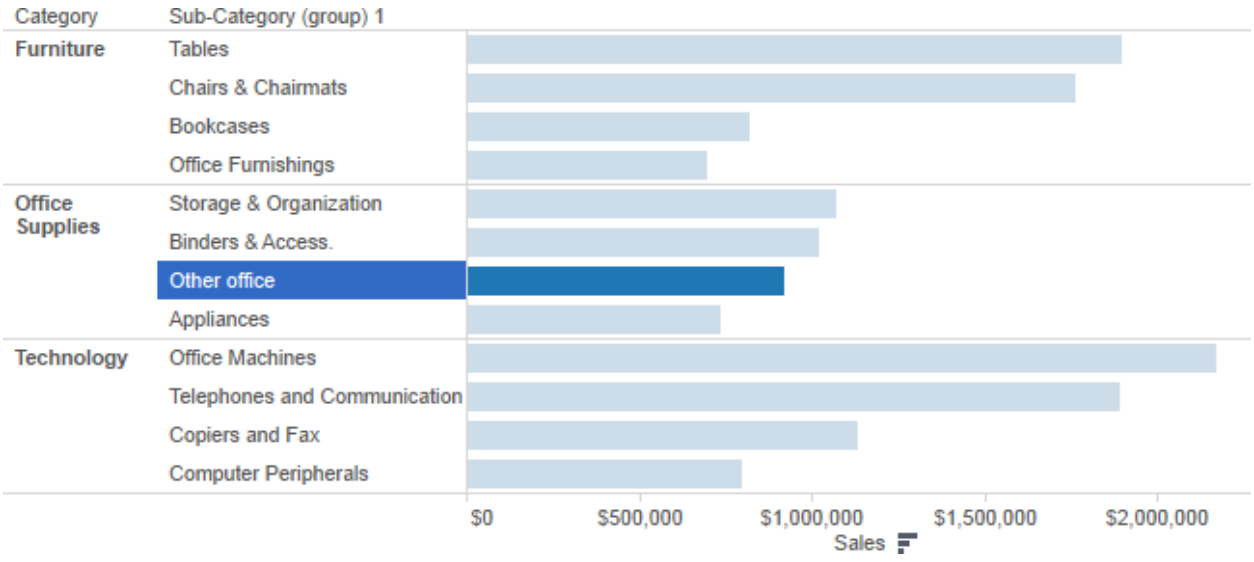
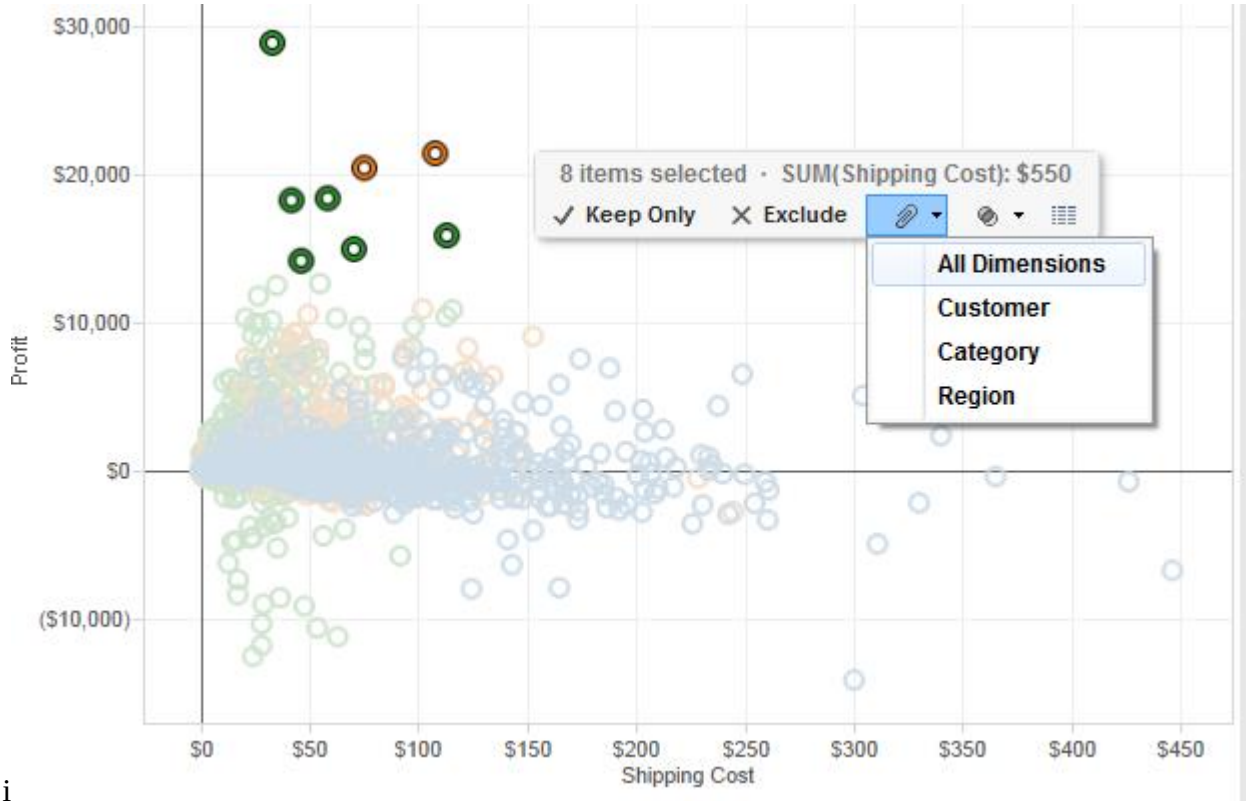


Fig 15-31 The ad hoc office group

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.



i

Fig 15-32 Grouping marks using all dimensions

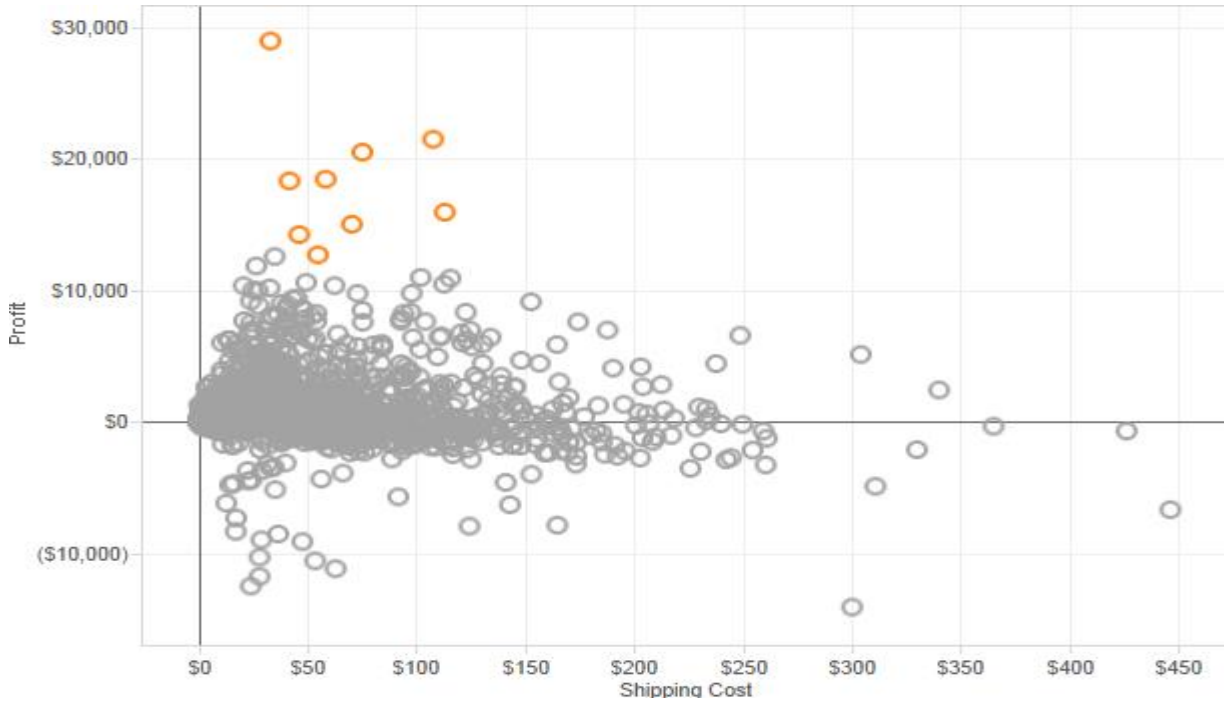


Fig. 15-33. Manually selecting a group

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.

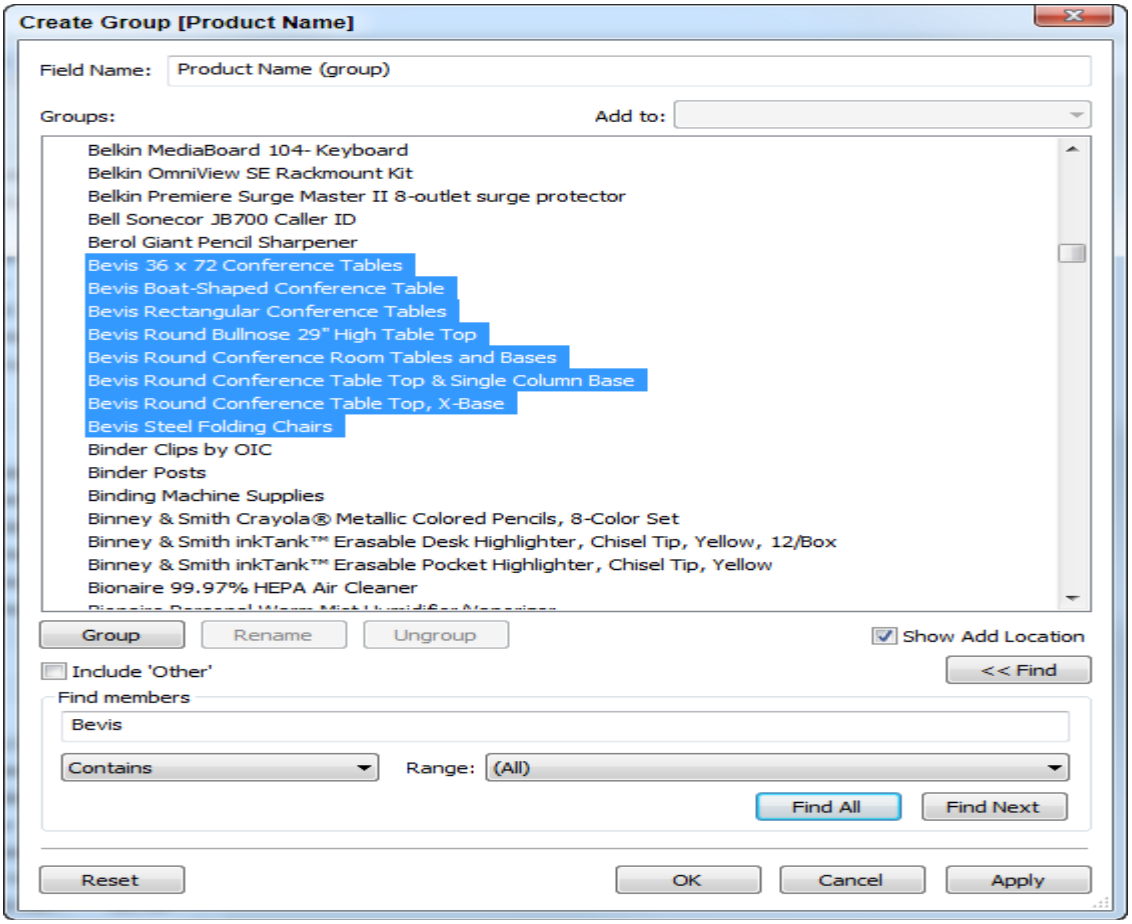


Fig. 15-34 Using string search to group

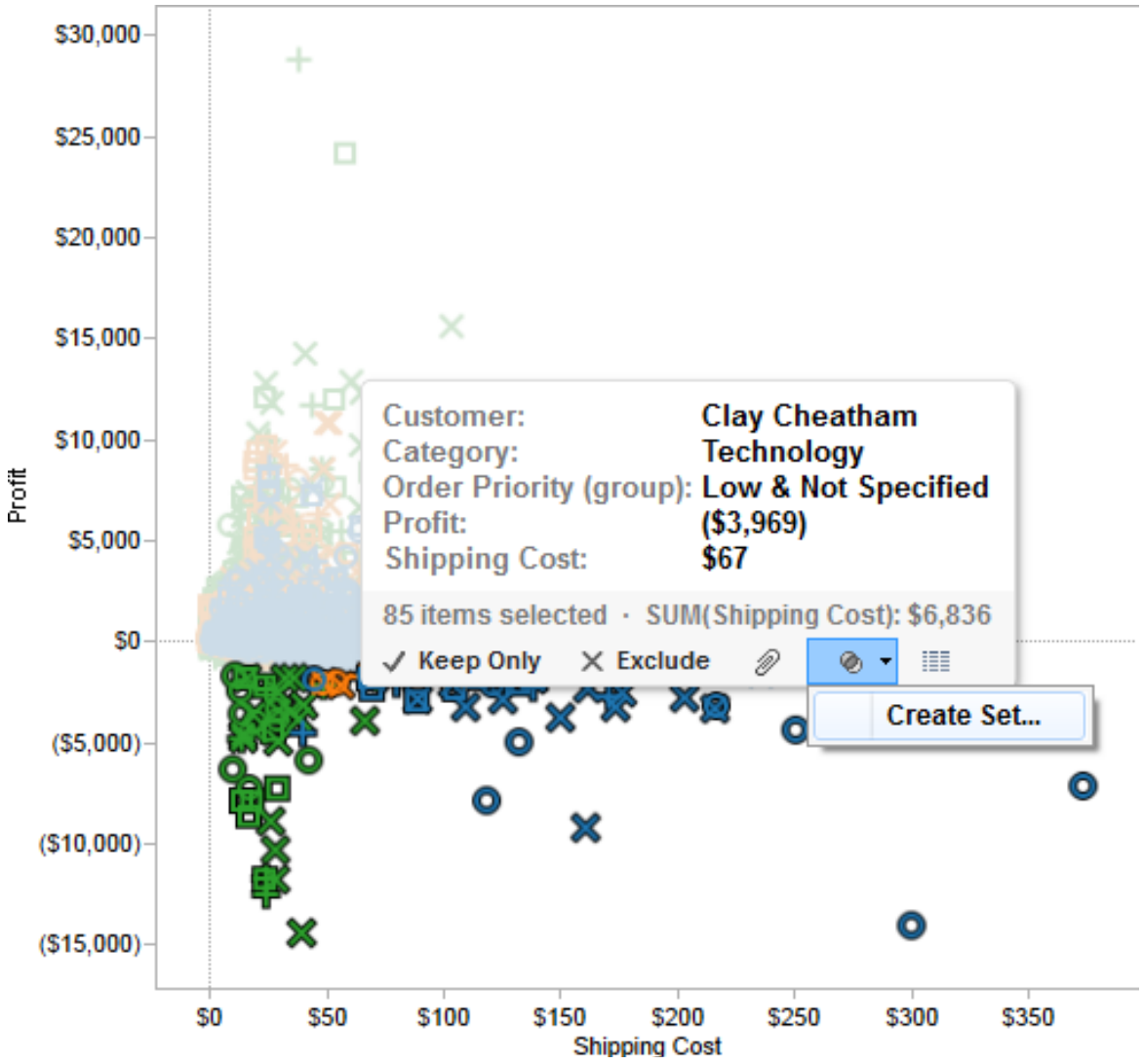


Fig 15-35 Selecting marks to create a set

Selecting the Create Set menu option exposes the dialog box in Figure 15-36.

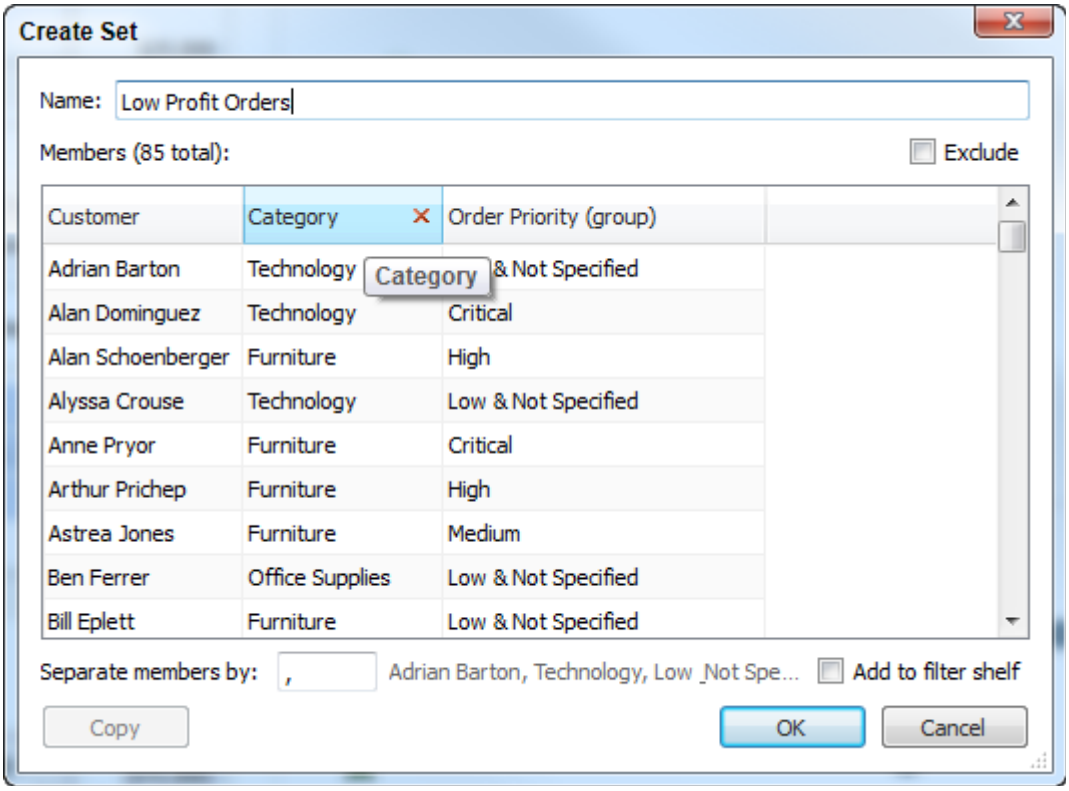


Fig. 15-36 Editing fields included in a set

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.

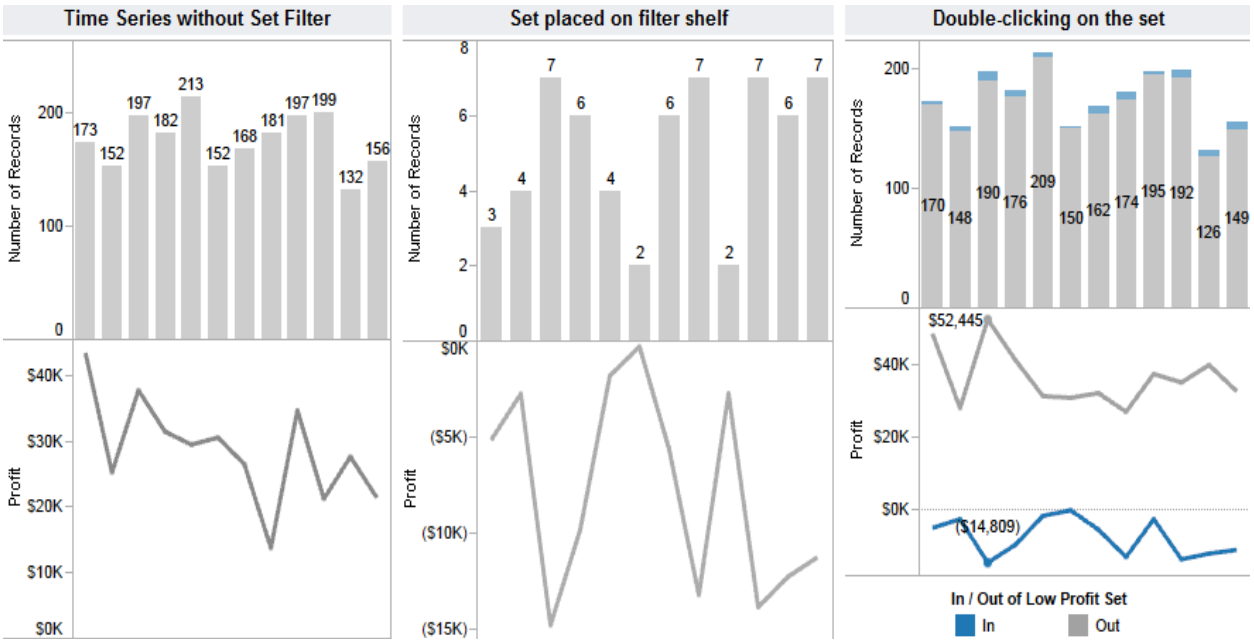


Fig. 15-37 Applying sets in different worksheets

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].

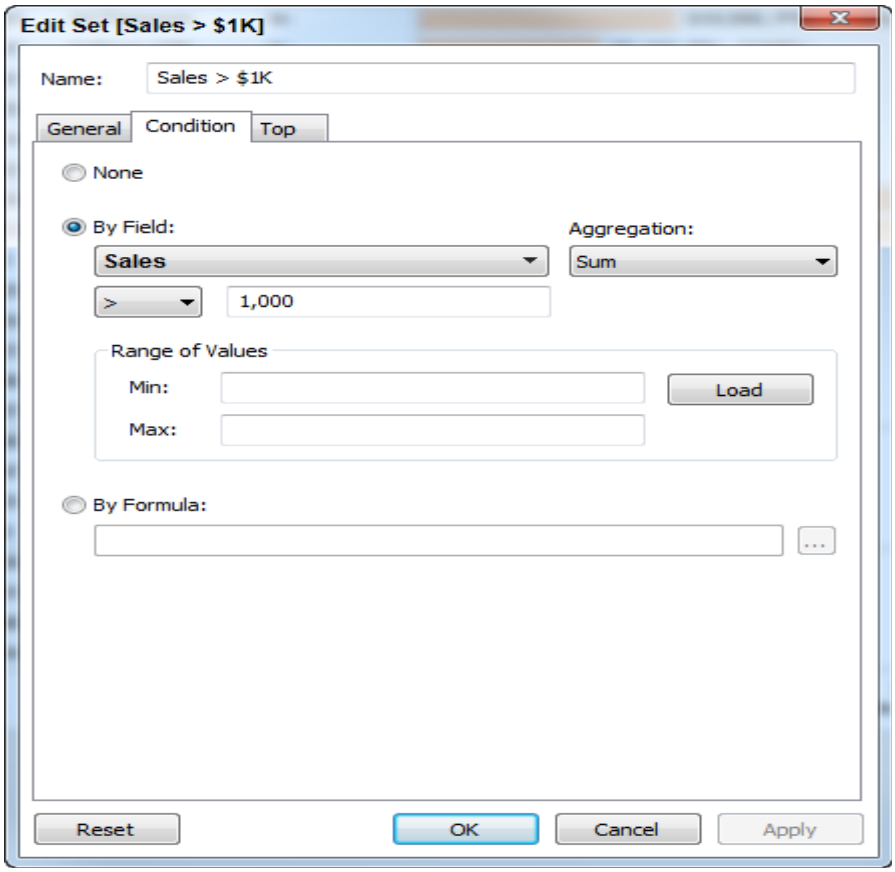


Fig .15-38 Making the sales set

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.

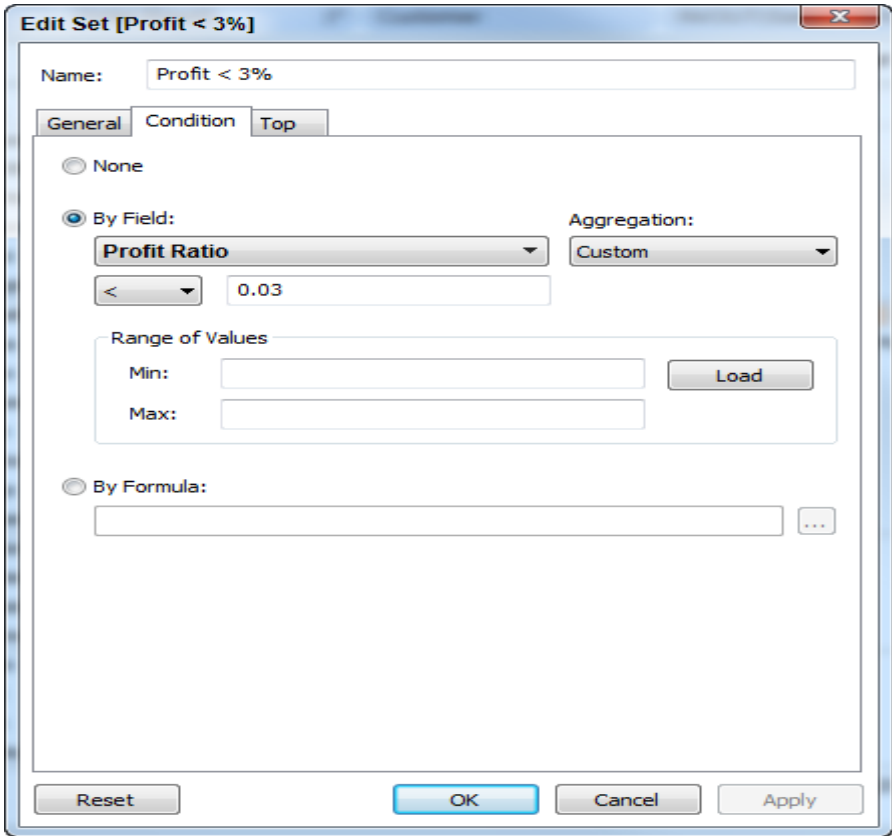


Fig 15-39 The low profit set condition defined

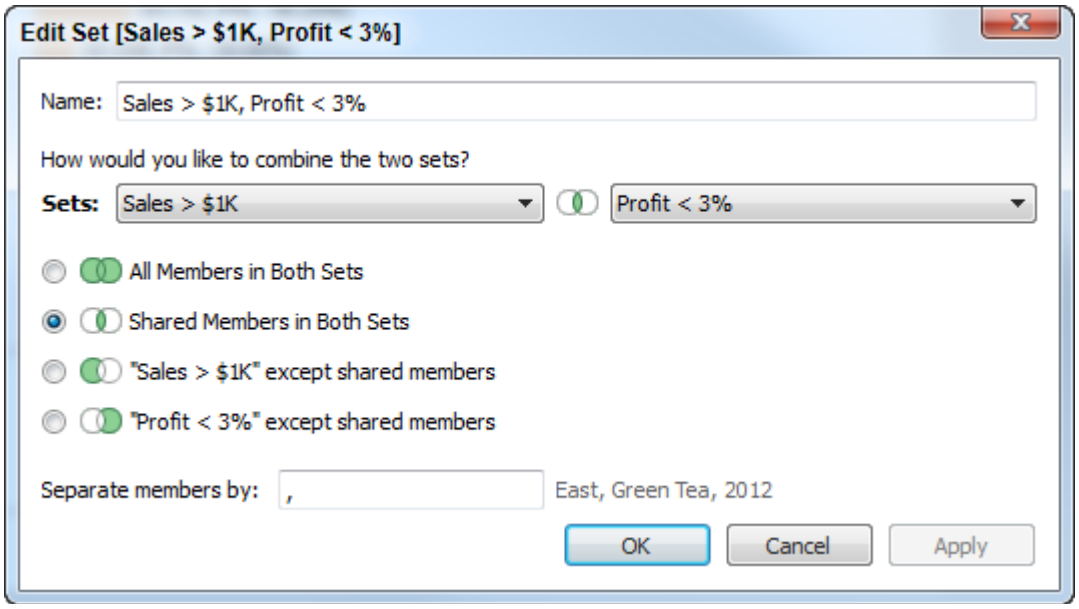


Fig. 15-40 Combination set dialogue box

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.

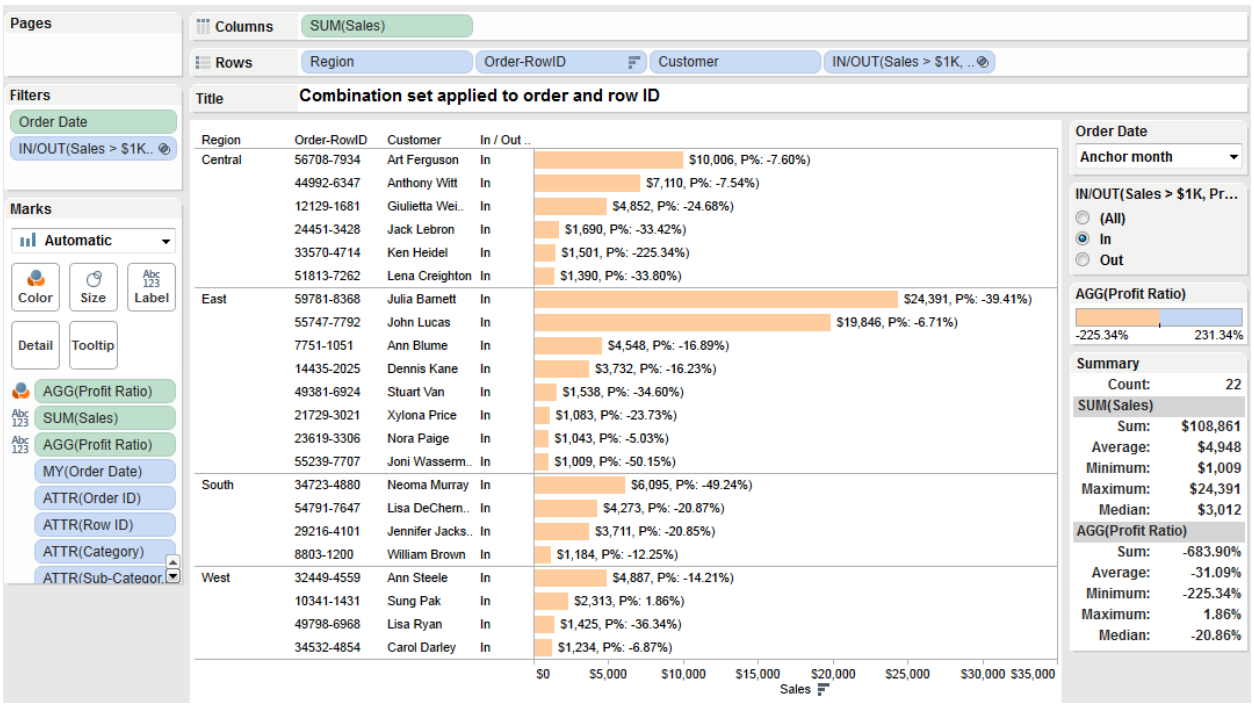


Fig 15-41 Combination filter applied to a bar chart

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.

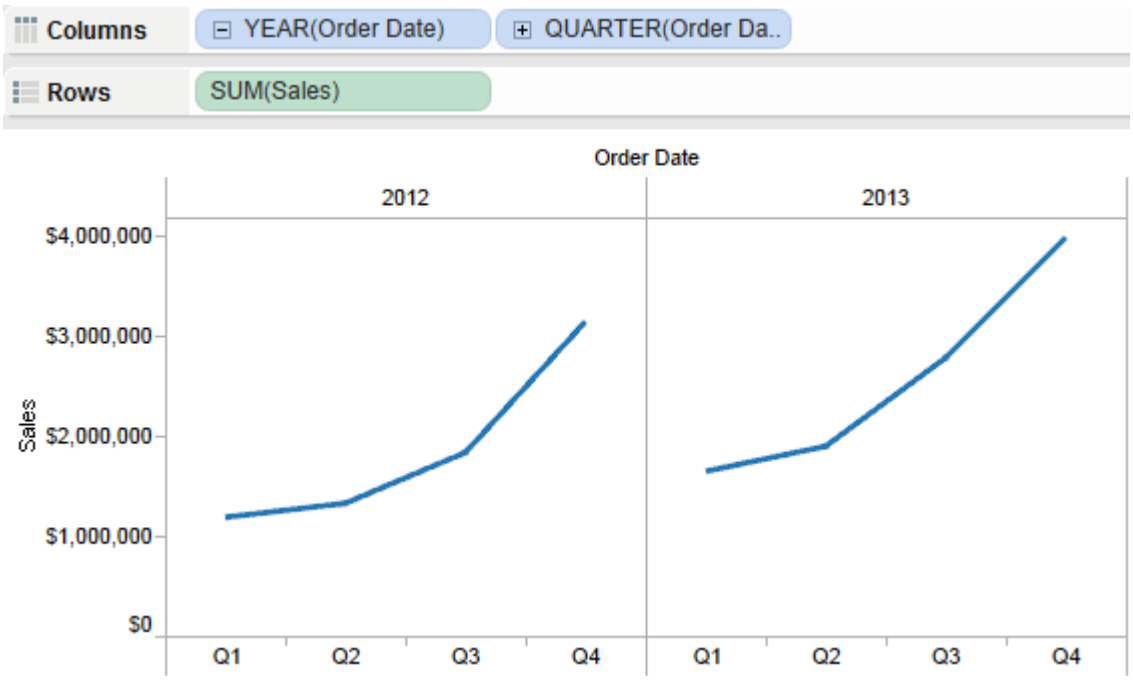


Fig. 15-42 Discrete time series

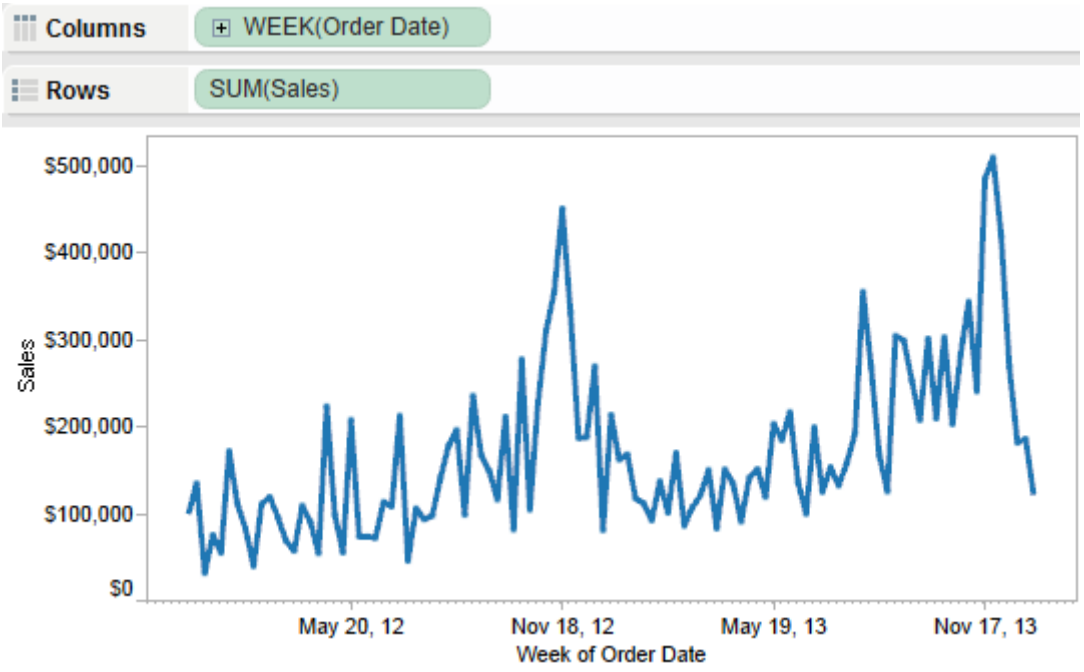


Fig. 15-43 Continuous time series

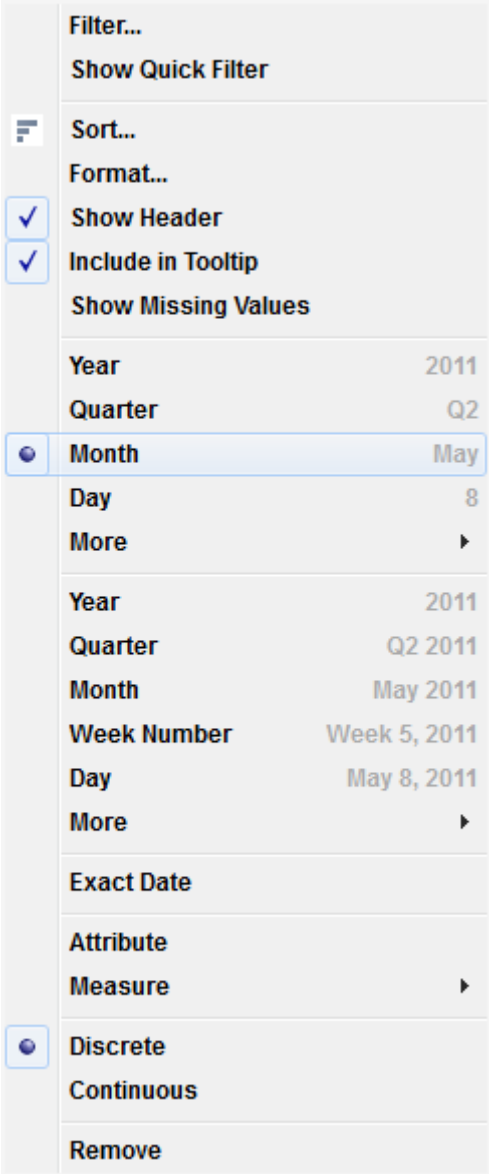


Fig 15-44 Changing the data level of detail

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.

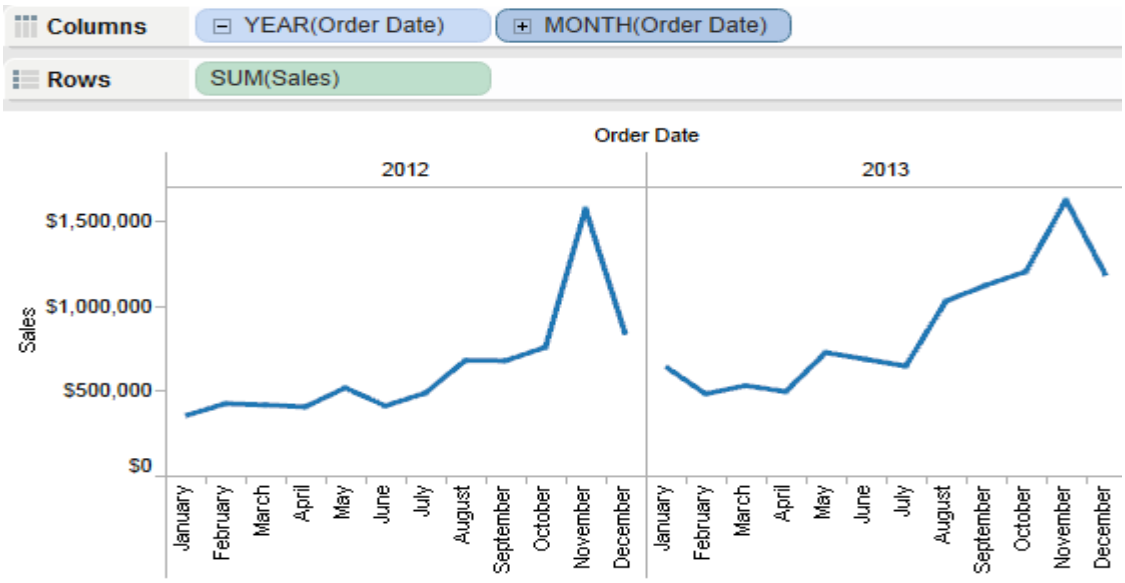


Fig 15-45 Time series displaying discrete year month

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 SELF ASSESSMENT QUESTIONS

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

16.1 Aggregation

16.2 Calculated Values and Table Calculations

16.3 Using the Calculation Dialog Box to Create

16.4 Building Formulas Using Table Calculations

16.5 Using Table Calculation Functions

16.6 Adding Flexibility to Calculations with Parameters

16.7 Using the Function Reference Appendix

16.8 Check your progress

16.9 Summary

16.10 Keywords

16.11 Self Assessment Questions

16.12 References

16.0 OBJECTIVES

After studying this unit, you will be able to:

- ✓ Explain when to use which type of aggregation
- ✓ Create and edit calculated fields
- ✓ Identify 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 *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.

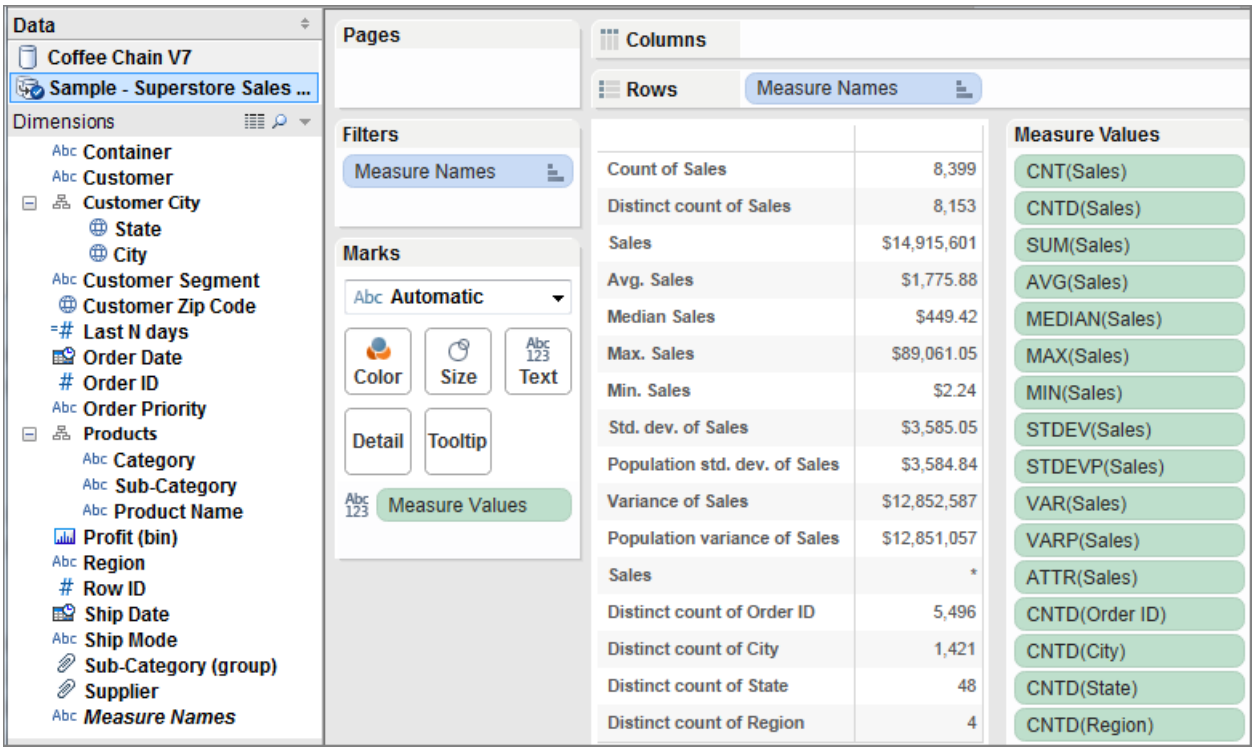


Fig. 16-1 Different aggregation of sales

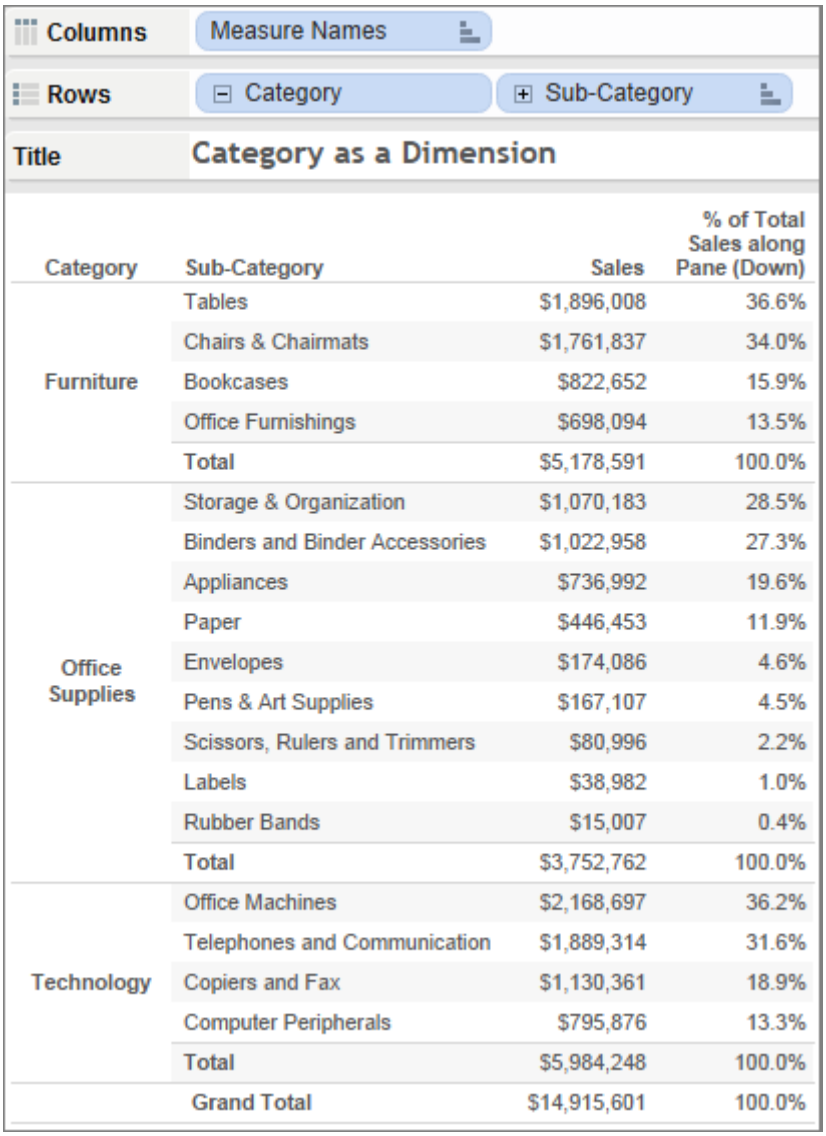
Notice that the bottom four rows are expressing Count Distinct values for different dimensions. By dragging each of those dimension fields into the crosstab using the right mouse button, the Count Distinct aggregation is expressed for each dimension. As you can see the data set includes over 5,000 different 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 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

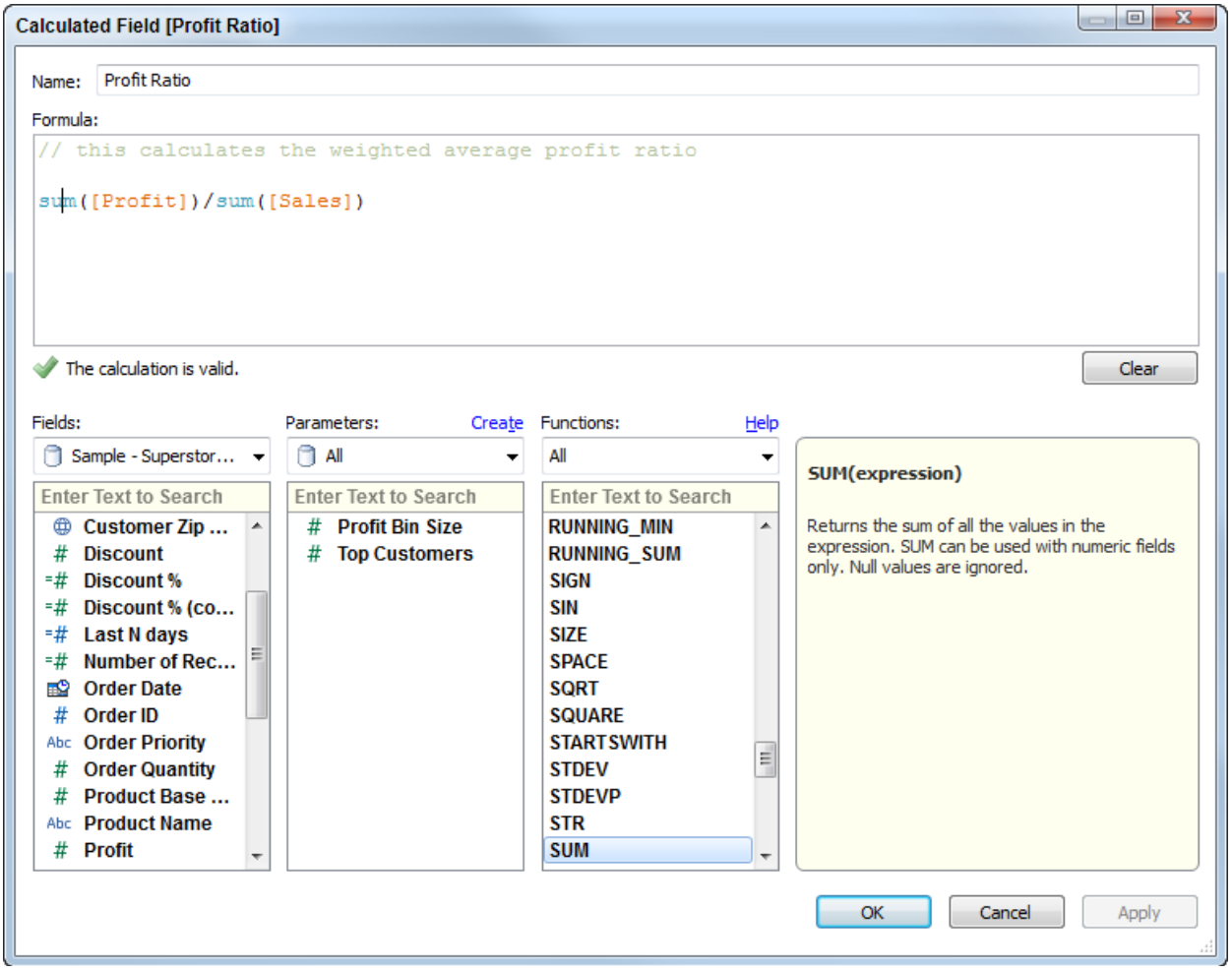


Fig 16-4 Formula dialog box or editing window

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

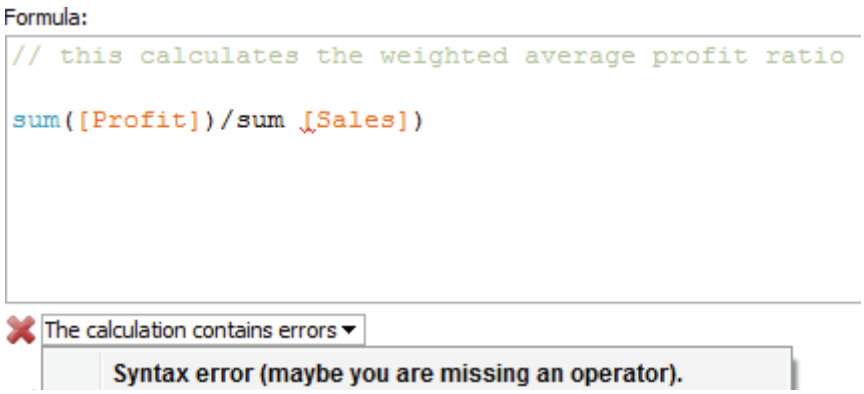


Fig 16-5 The real-time formula editor

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]).

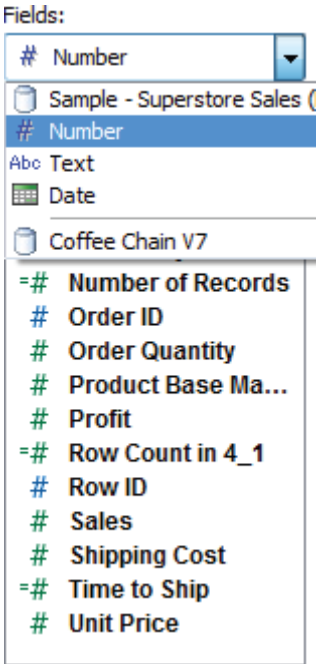


Fig 16.6 Filtering field selections for numbers

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.

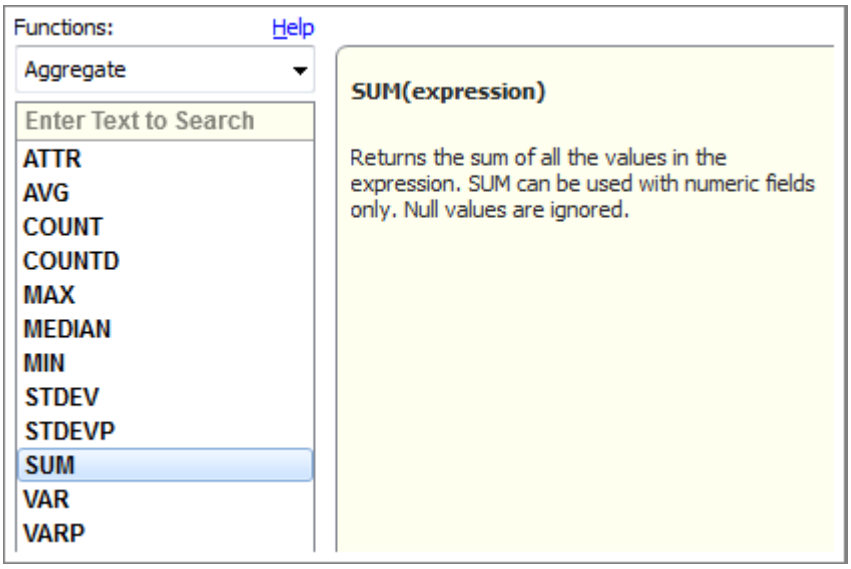


Fig. 16-7 Filter the function window

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.

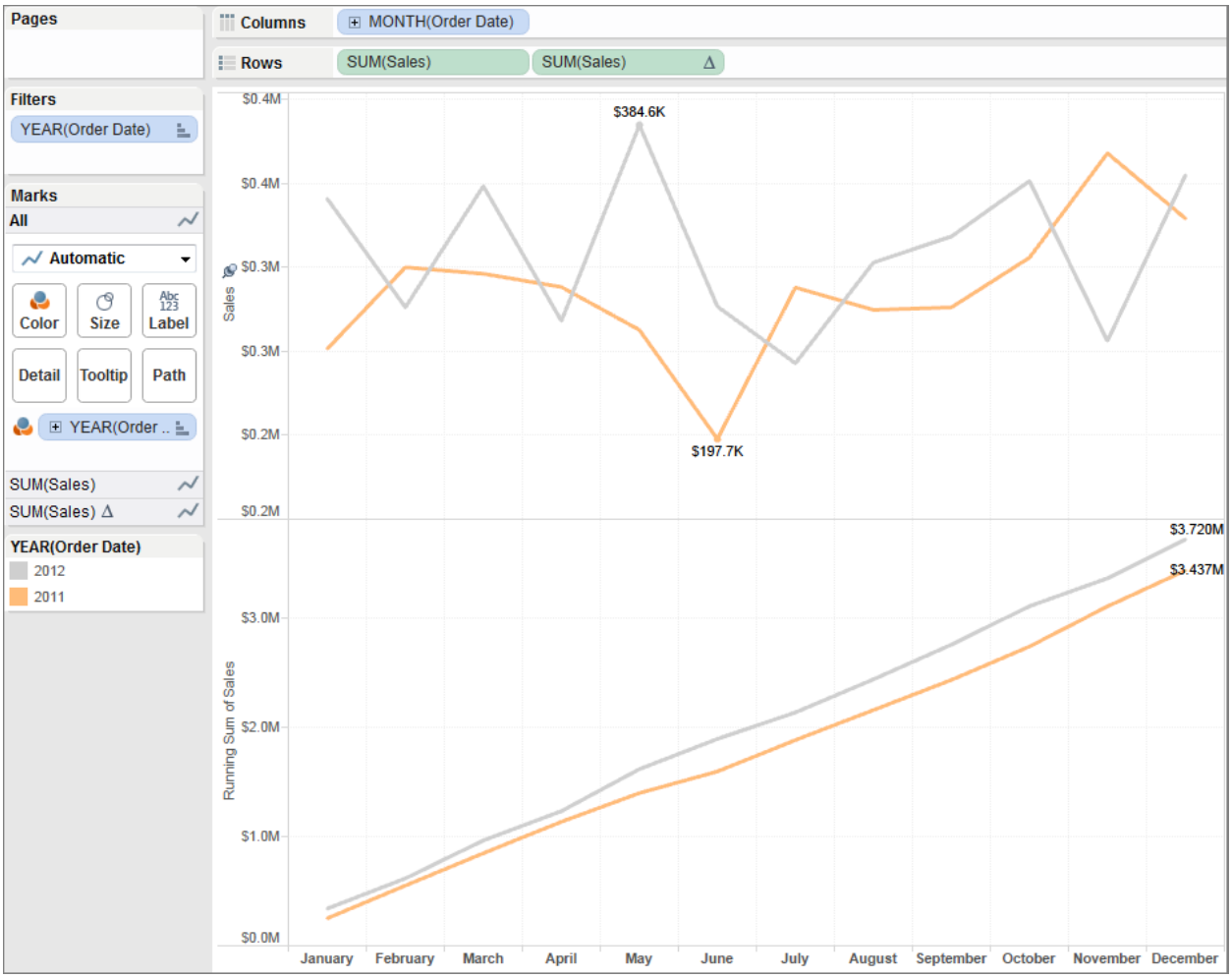


Fig 16-8 Time series using a running sum

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.

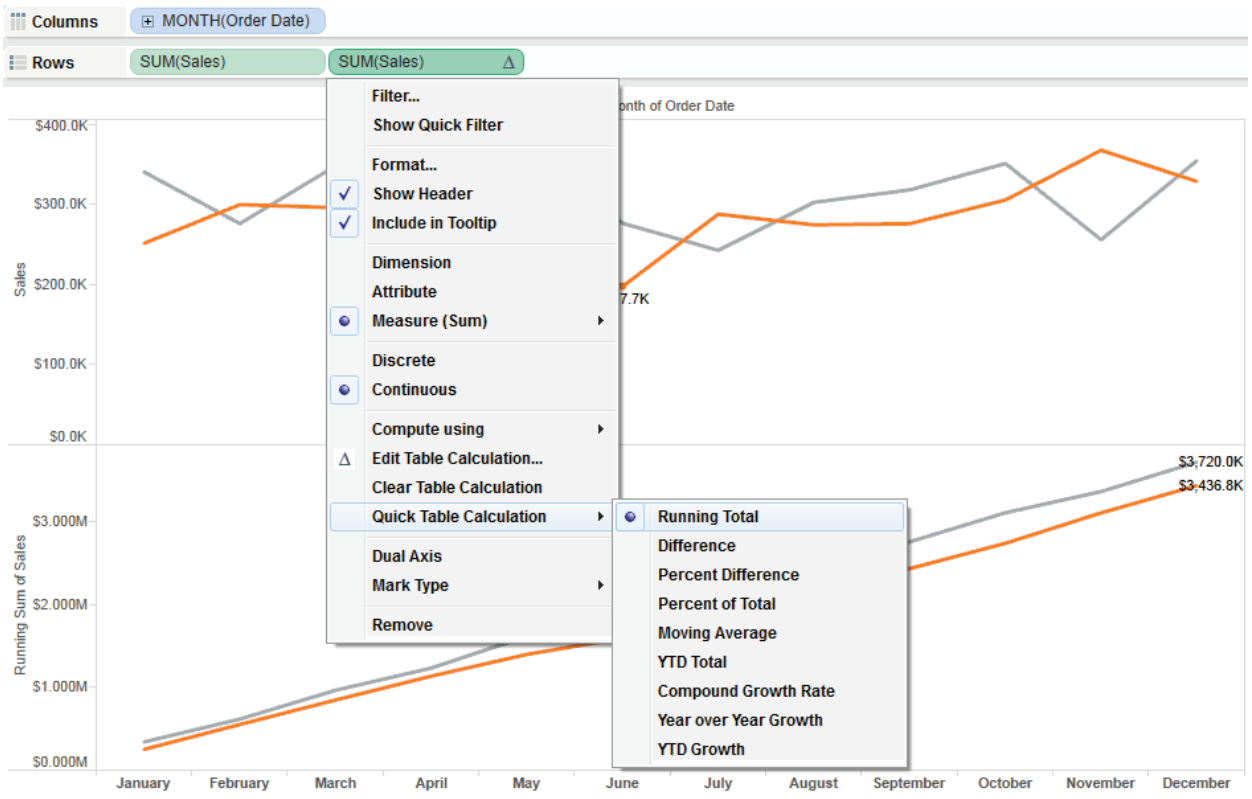


Fig 16-9 creating the quick table calculation

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

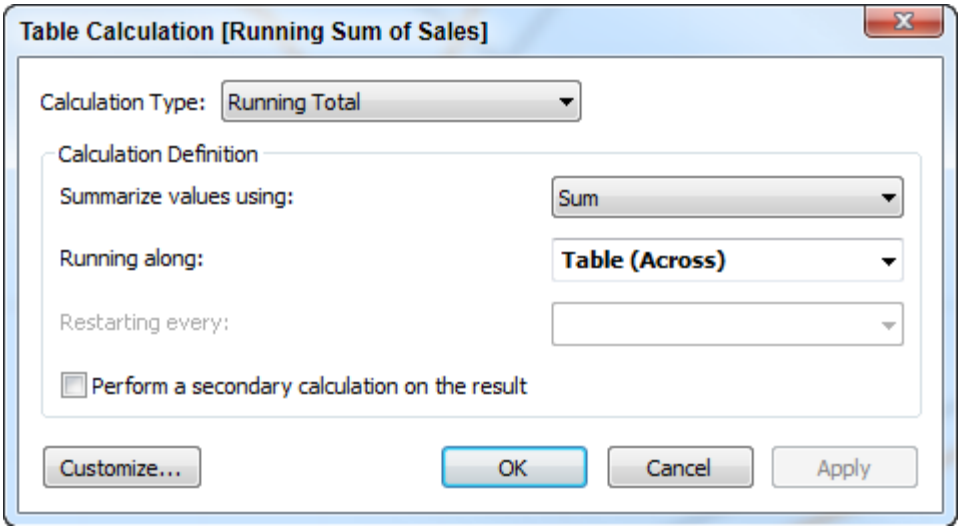


Fig 16-10 The table calculation editing menu

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.

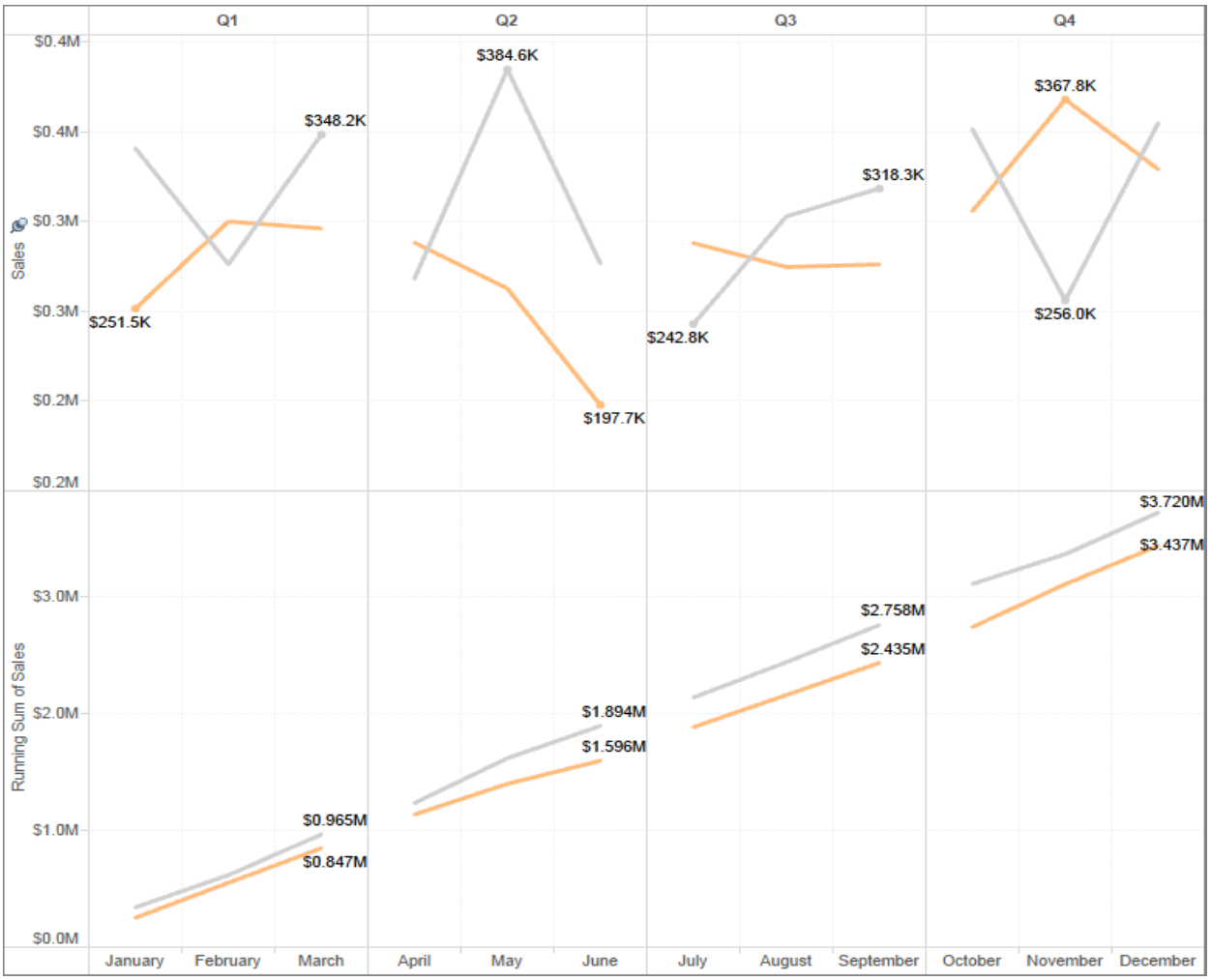


Fig. 16.11 using discrete quarter and month

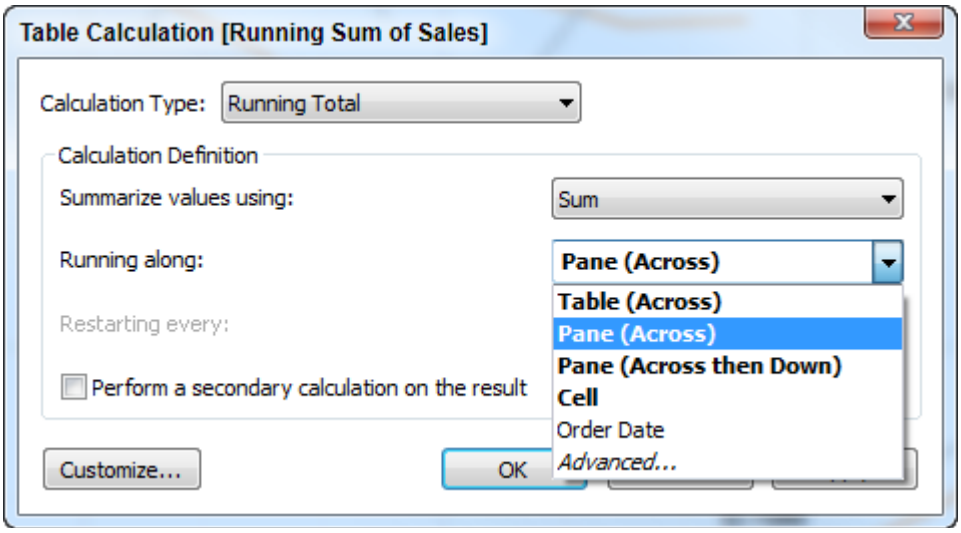


Fig 16-12 Changing table calculation scope

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.

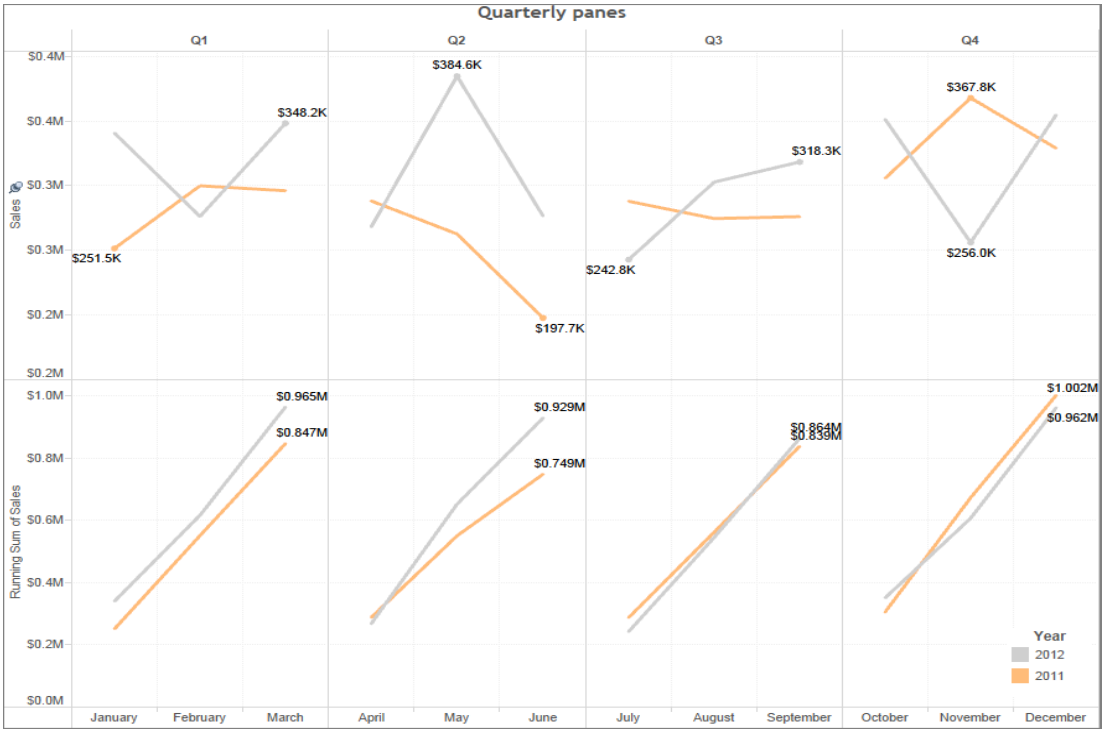


Fig 16-13 Running sum set to pane across

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.

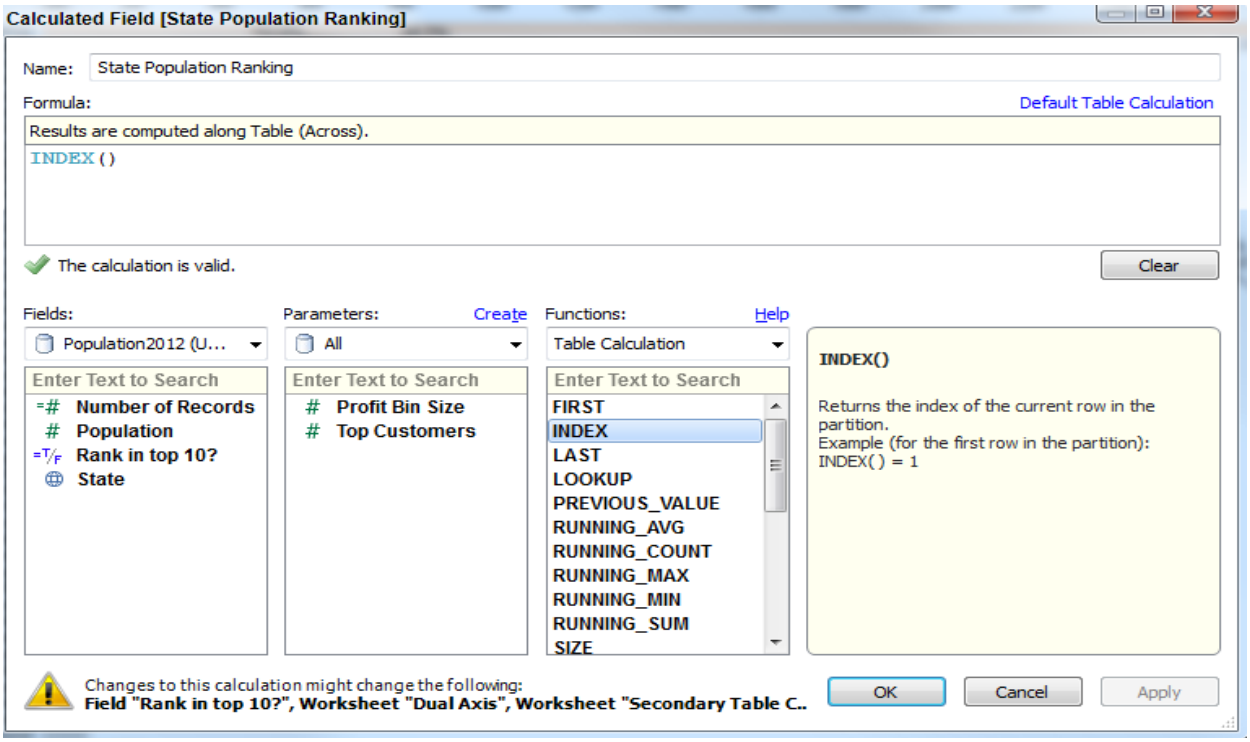


Fig 16-20. Creating a population rank

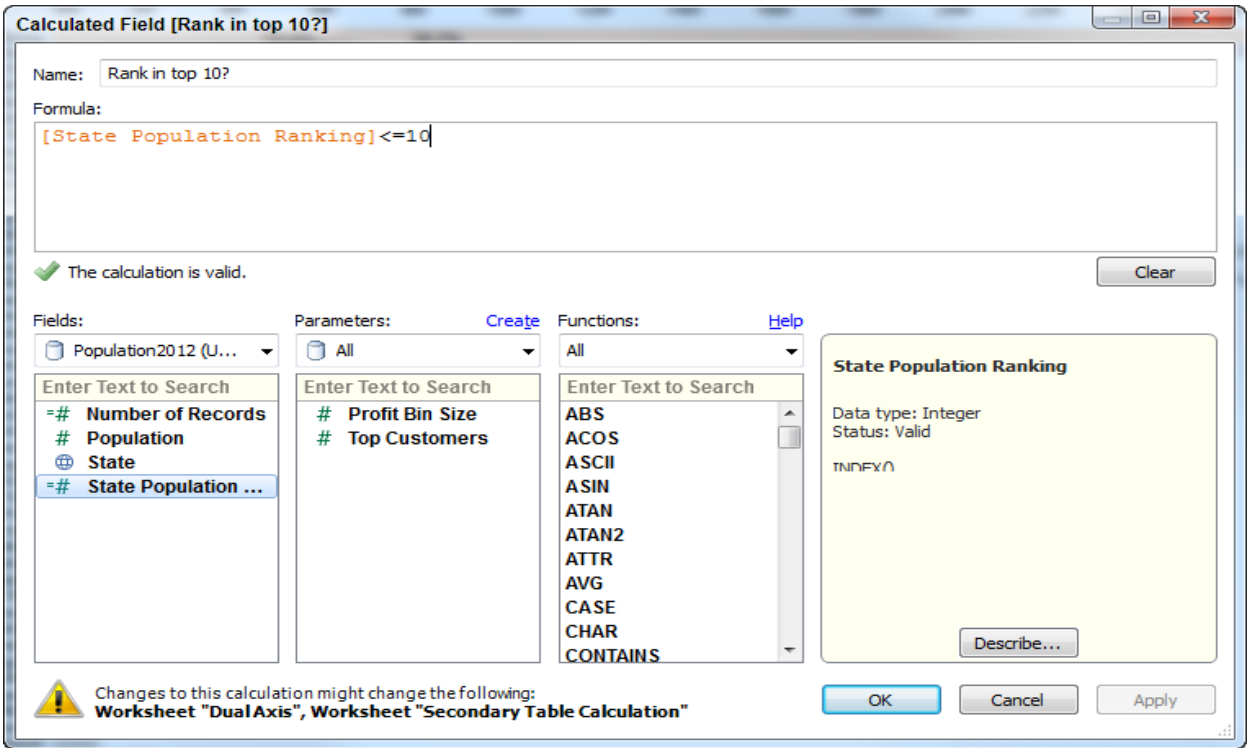


Fig 16-21 Creating a Boolean calculation

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.

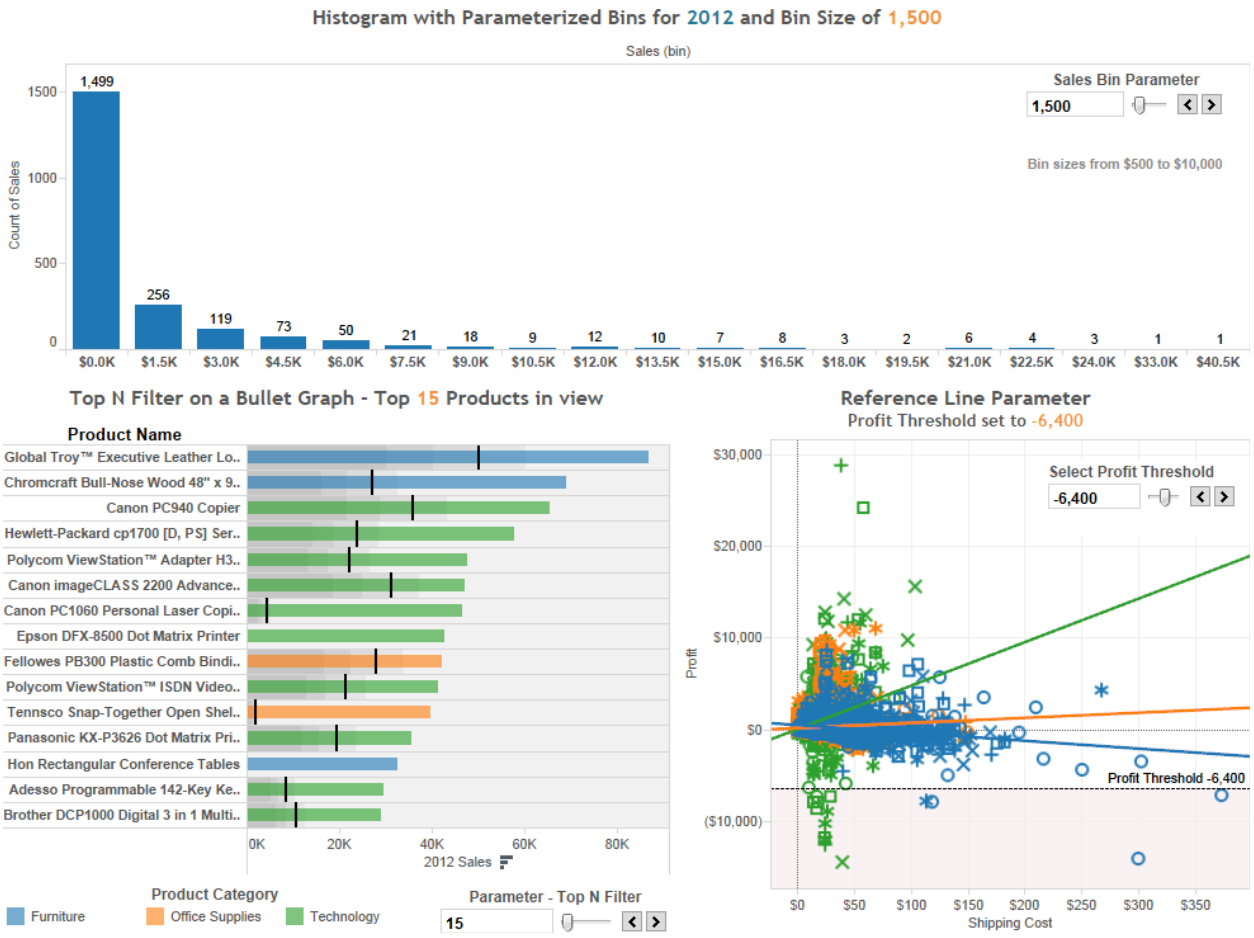


Fig. 16-22 Basic parameter controls

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.

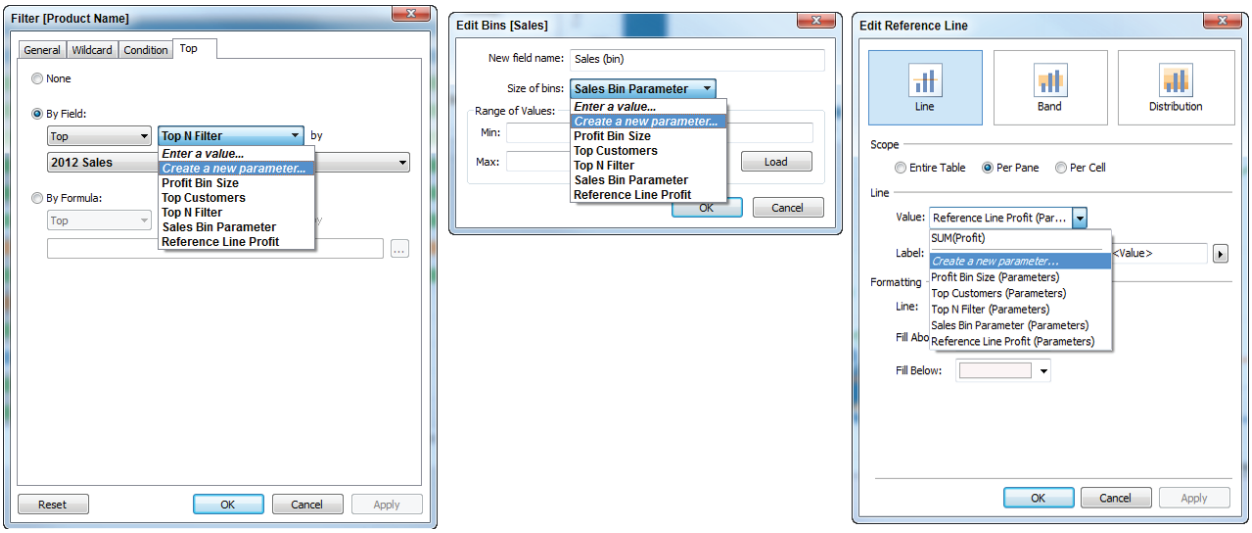


Fig. 16-23 Dialog windows for basic parameter creation

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.

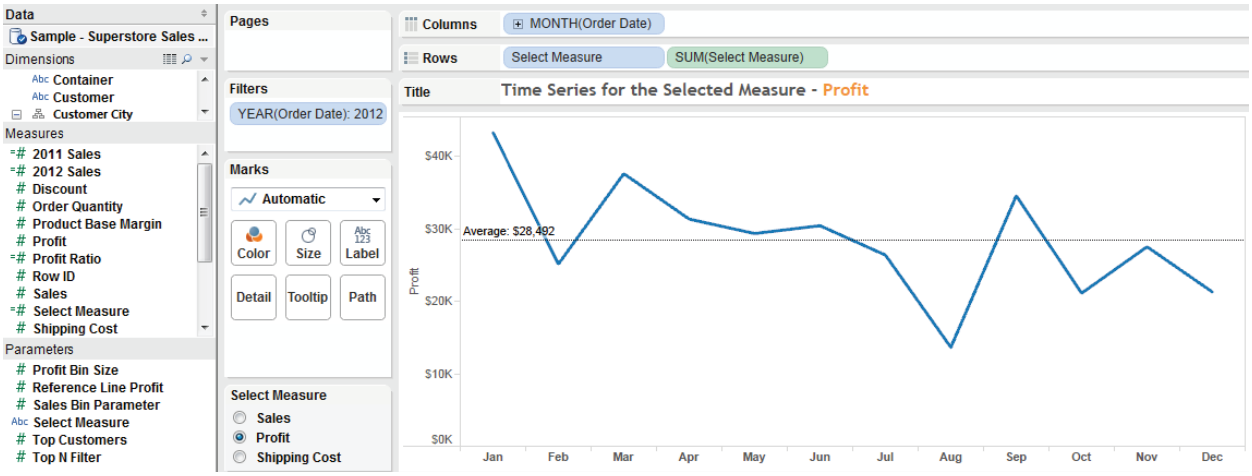


Fig. 16-24 Using a parameter to change the measure displayed in a view

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

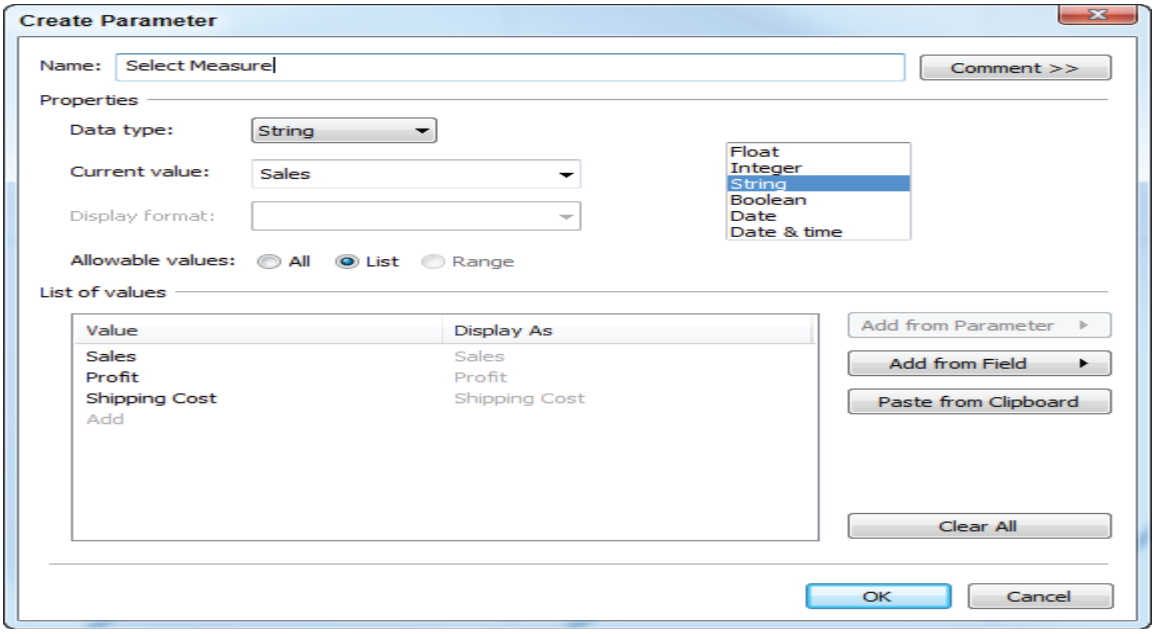


Fig 16-25 Defining a parameter control

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.

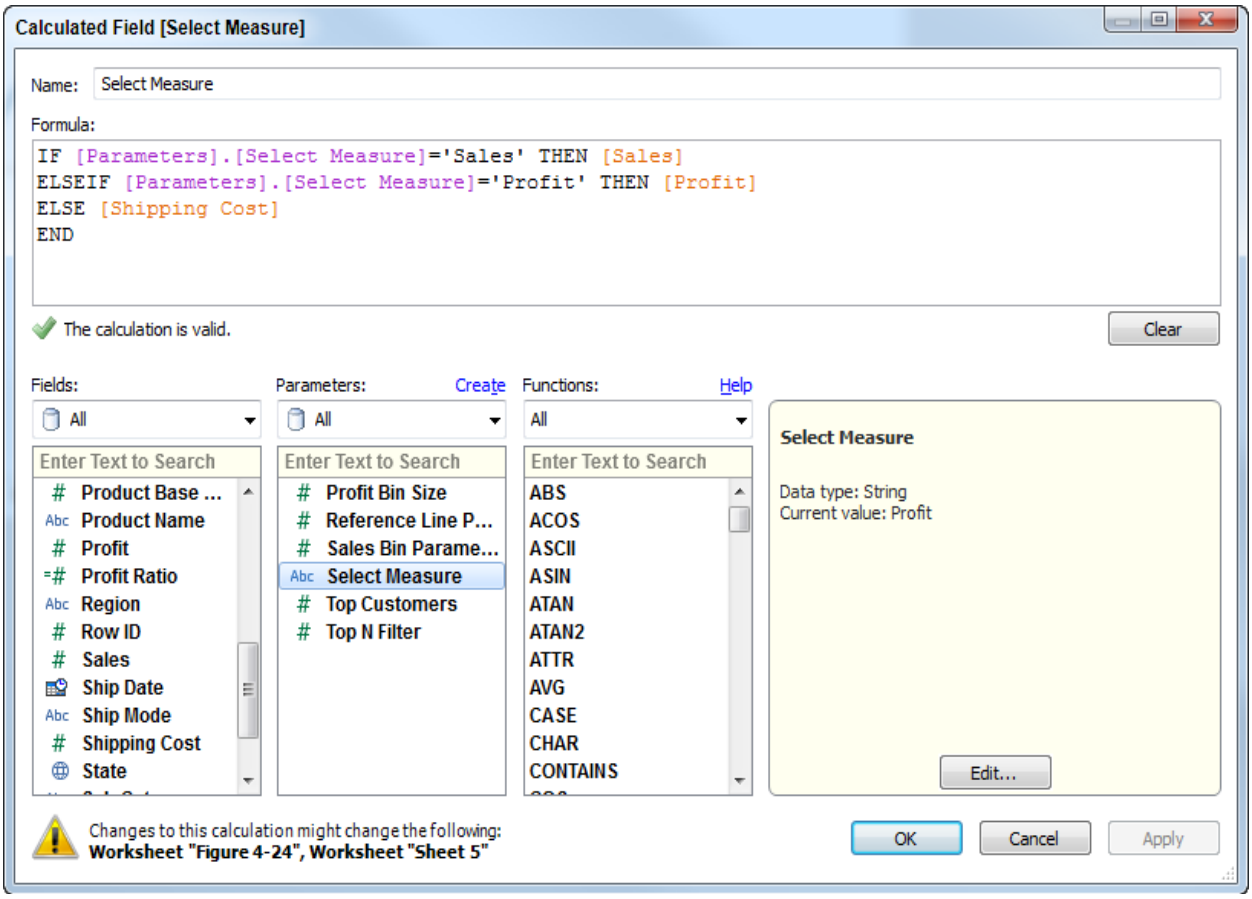


Fig. 16-26 using a parameter in a formula

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 SELF ASSESSMENT QUESTIONS

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)