Transforming Data With Intelligence™

Previews of TDWI course books offer an opportunity to see the quality of our material and help you to select the courses that best fit your needs. The previews cannot be printed.

TDWI strives to provide course books that are content-rich and that serve as useful reference documents after a class has ended.

This preview shows selected pages that are representative of the entire course book; pages are not consecutive. The page numbers shown at the bottom of each page indicate their actual position in the course book. All table-of-contents pages are included to illustrate all of the topics covered by the course.

This page intentionally left blank.

tdwi

Transforming Data
With Intelligence™

TDWI Business Intelligence and Analytics Principles and Practices

Charting the Course to BI and Analytics Success

TDWI takes pride in the educational soundness and technical accuracy of all of our courses. Please send us your comments—we'd like to hear from you.

Address your feedback to info@tdwi.org

Publication Date: December 2017

TABLE OF CONTENTS

COURSE OBJECTIVES

You will learn:

- ✓ ***Meaningful and actionable definitions of business intelligence and analytics***
- ✓ ***Roadmap development techniques for your BI and analytics program and projects***
- ✓ ***A capability framework that links BI and analytics program initiatives to business goals***
- ✓ ***Best practices for BI and analytics services, including performance management, analytics, OLAP, reporting, visualization, and self-service***
- ✓ ***Types of data consumed by BI and analytics applications, including enterprise, external, and big data sources***
- ✓ ***Architecture, implementation, and operational practices for data integration services***
- ✓ ***Data management practices including data governance, data quality management, data profiling, and data cleansing***
- ✓ ***Technologies needed to support BI and analytics***

Module 1

Introduction to BI and Analytics

Definitions

Business Intelligence

“A set of concepts and methodologies to improve decision making in business through the use of facts and fact-based systems.”

HOWARD DRESNER

“the processes, technologies and tools needed to turn data into information, information into knowledge, and knowledge into plans that drive profitable business actions. Business intelligence encompasses data warehousing, business analytic tools, and content knowledge management.”

DAVID LOSHIN

“The use of information to improve business performance.”

CHRIS ADAMSON

“the ability of an organization to reason, plan, predict, solve problems, understand, innovate, and learn in ways that increase organizational knowledge, inform decision processes, enable effective actions, and help establish and achieve business goals.”

DAVE WELLS

Definitions

Business Intelligence

FACT-BASED DECISION MAKING

Howard Dresner, formerly with the Gartner Group, is credited with creating the term business intelligence in the early 1990s. Dresner first defined BI as “a set of concepts and methodologies to improve decision making in business through use of facts and fact-based systems.”

PROCESSES, TOOLS, AND TECHNOLOGIES

A decade later David Loshin defined BI as “the processes, technologies, and tools needed to turn data into information, information into knowledge, and knowledge into plans that drive profitable business actions.” (David Loshin, *Business Intelligence: The Savvy Manager’s Guide*, Addison-Wesley, 2003).

BUSINESS IMPACT

Chris Adamson defines BI as “the use of information to improve business performance.” His definition emphasizes impact, which does not have to be measured as profit. (Chris Adamson, 2014)
http://blog.chrisadamson.com/2014/09/business-intelligence-in-modern-era.html

BUSINESS AND INTELLIGENCE

Each of the definitions seen so far describe aspects of BI—decision making, tools and technology, people and information, and impact. All are correct but none describes the characteristics of an intelligent business. Dave Wells created and published a definition focused on business capabilities, answering the question

What does it mean to be an intelligent business?

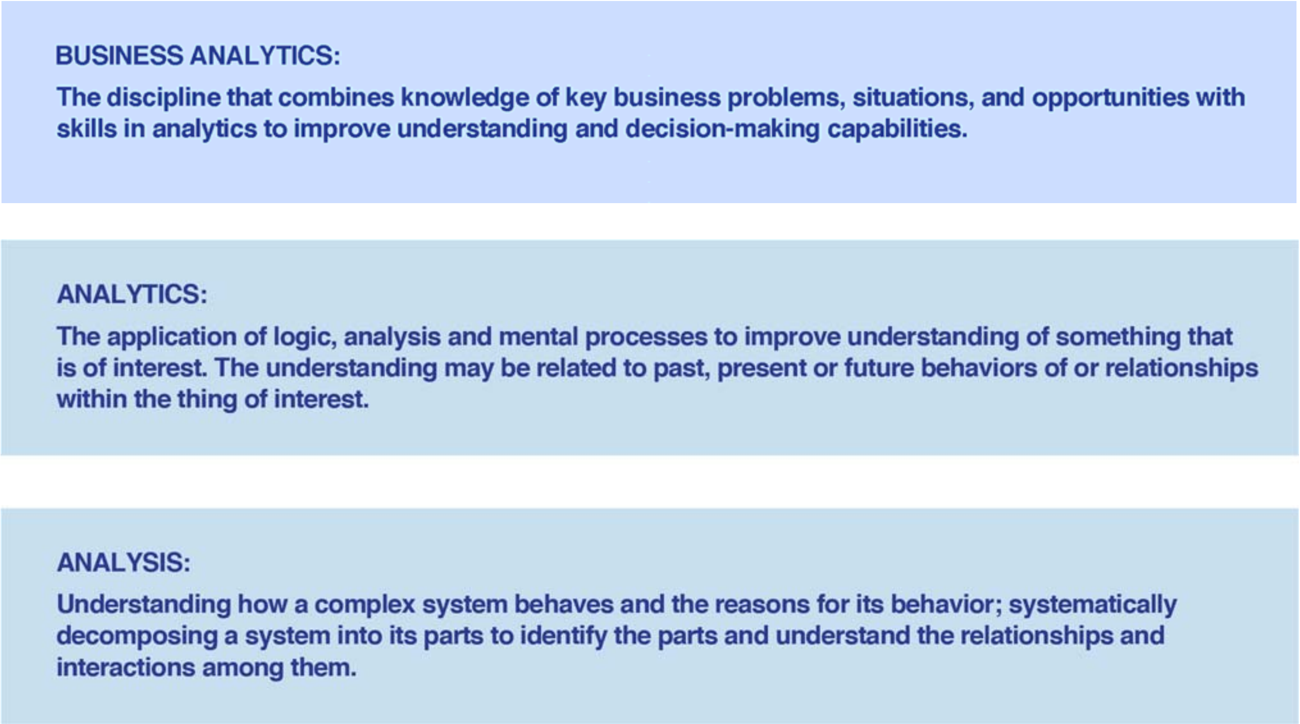
This definition describes business intelligence as “the ability of an organization to reason, plan, predict, solve problems, understand, innovate and learn....” The definition continues to describe the purpose of BI, answering the question

Why do we need business intelligence?

“... in ways that increase organizational knowledge, inform decision processes, enable effective actions, and help to establish and achieve business goals.”

Definitions

Business Analytics



BUSINESS ANALYTICS:

The discipline that combines knowledge of key business problems, situations, and opportunities with skills in analytics to improve understanding and decision-making capabilities.

ANALYTICS:

The application of logic, analysis and mental processes to improve understanding of something that is of interest. The understanding may be related to past, present or future behaviors of or relationships within the thing of interest.

ANALYSIS:

Understanding how a complex system behaves and the reasons for its behavior; systematically decomposing a system into its parts to identify the parts and understand the relationships and interactions among them.

Definitions

Business Analytics

BUSINESS ANALYTICS

For the purposes of this course, we'll define *business analytics* as the discipline that combines knowledge of key business problems, situations, and opportunities with skills in analytics to improve understanding and decision-making capabilities.

Business analytics expands organizational knowledge. The discovery processes of analytics extend the value of data well beyond the traditional uses of knowing *what*, *when*, and *how much*. Analytics reach—the power of exploration, experimentation, and discovery—is in answering the really hard questions: *why* and *what-if.*

ANALYTICS

Analytics is the application of logic, analysis, and mental processes to improve our understanding of something that is of interest to us. The understanding may be related to past, present, or future behaviors of, or relationships within, the thing of interest.

ANALYSIS

Analysis is a fundamental technique used to understand how a complex system behaves and what drives its behavior. The analysis technique is based on systematically decomposing a system to identify the key parts and to understand the relationships and interactions among them.

BI VS ANALYTICS

There are differing perspectives on the relationship between business intelligence and business analytics.

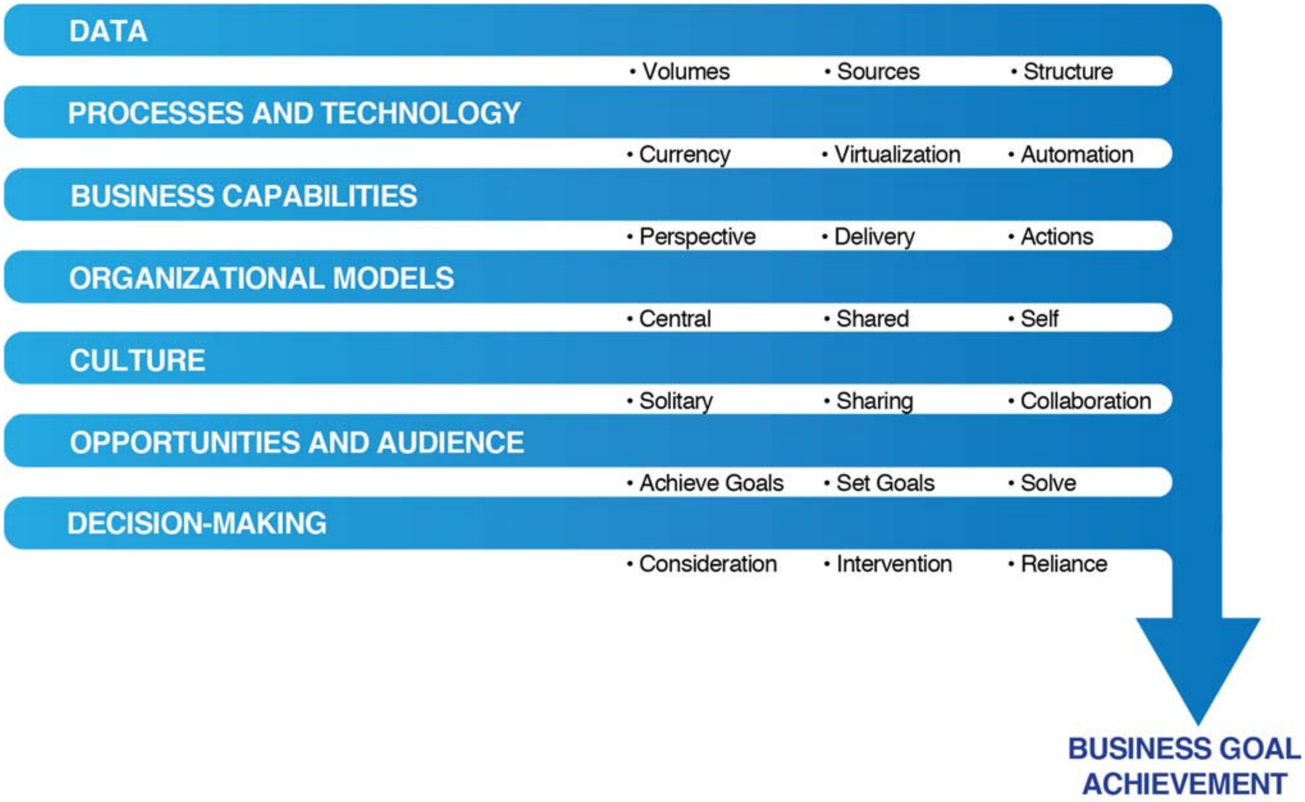
Historically, many organizations developed analytics capabilities separately from their BI programs. Typically, analytics skills grew within lines of business. As these organizations evolved, many chose to establish separate competency centers for analytics. Viewed from this perspective, business analytics and business intelligence are separate but overlapping disciplines.

Many organizations view business analytics as a form of business intelligence—another way to drive impact from data. Indeed, the definition of business analytics on this page is consistent with the various definitions for business intelligence on the prior pages.

Pragmatically speaking, it doesn't matter. For the purposes of this course, we often refer to "business intelligence and analytics" to encompass both without artificially drawing a line between them.

Definitions

Evolution of BI and Analytics



Definitions

Evolution of BI and Analytics

CHANGING LANDSCAPE

The graphic on the facing page depicts key elements of the definitions of BI and business analytics. In the past decade, there have been major changes in each of these areas.

DATA

Data volumes have increased significantly over time, from data warehouses measured in megabytes and gigabytes to ones that are measured in terabytes and petabytes. Exabyte-sized environments are in the foreseeable future.

Initially, data was obtained from internal operational systems or external sources and was highly structured in nature. As the ecosystem evolved, additional sources (e.g., machine-generated data, social media, and Internet applications) have been incorporated, encompassing other data structures that may be stable or in motion. These are described further in *Module 4: Data Integration*.

PROCESSES AND TECHNOLOGY

As demands have increased for more data, near real-time access, and additional functionality, supporting processes and technology have evolved.

Initial environments depended on manual coding of the data movement algorithms. ETL tools were then introduced. These have evolved into more comprehensive data integration and aggregation tools that employ other technologies, including data virtualization. Tools to automate major parts of the development process have also emerged. Tools and technology are described further in *Module 6: BI and Analytics Technology*.

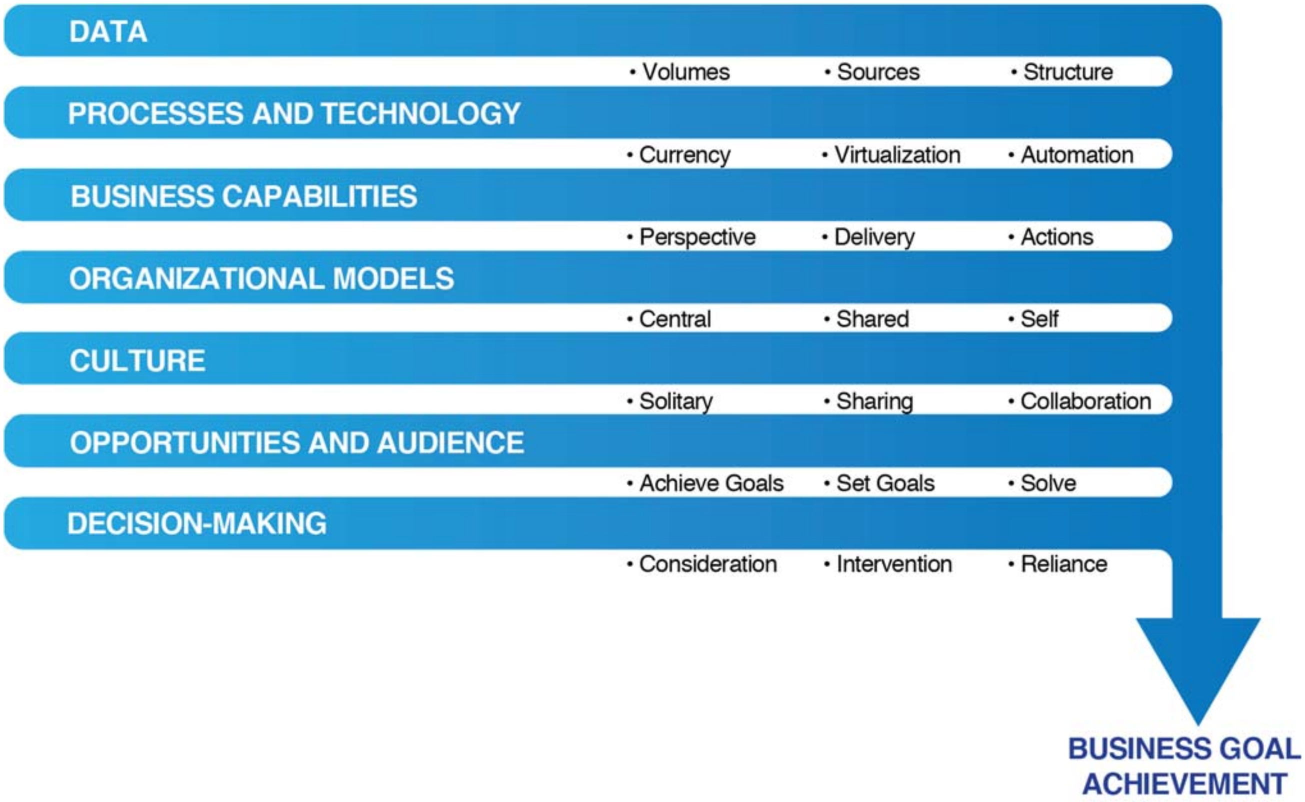
A word of caution is needed. There is no silver bullet. As you consider emerging technologies, be sure to carefully evaluate the offerings and your organization’s readiness to adopt them and to deal with potentially being on the leading edge.

BUSINESS CAPABILITIES

Early business intelligence solutions delivered information mostly in the form of queries and reports with some simple graphics. Today, the landscape has evolved to include dashboards and scorecards, analytic models and simulations, and data visualizations. These topics are discussed in *Module 2: Performance Management and Analytics, and Module 3: OLAP and Other Information Services.*

Definitions

Evolution of BI and Analytics



DATA
- Volumes
- Sources
- Structure

PROCESSES AND TECHNOLOGY
- Currency
- Virtualization
- Automation

BUSINESS CAPABILITIES
- Perspective
- Delivery
- Actions

ORGANIZATIONAL MODELS
- Central
- Shared
- Self

CULTURE
- Solitary
- Sharing
- Collaboration

OPPORTUNITIES AND AUDIENCE
- Achieve Goals
- Set Goals
- Solve

DECISION-MAKING
- Consideration
- Intervention
- Reliance

BUSINESS GOAL ACHIEVEMENT

Definitions

Evolution of BI and Analytics

ORGANIZATIONAL MODELS

Three common organizational models are applied within the business intelligence and analytics environment:

- *Central services:* In this model, which was initially dominant, standards, processes, the architecture, and the technology are prescribed, with a centralized team being responsible for development, deployment, and management of business intelligence and analytics solutions.
- *Self-service:* In this model, business units meet their own needs with support of business-oriented tools, architectures, frameworks, guidelines, examples, templates, etc.
- *Shared services:* In this model, the standardized architecture and processes are defined and a centralized team is maintained for shared work, with most project and process work occurring within project teams and distributed lines of business.

CULTURE

As BI and analytics capabilities have evolved, many organizations have developed a culture of information use and collaboration. Collaboration is the act of working jointly—two or more people combining their efforts toward achieving shared or intersecting goals. It includes data sharing, collective analysis, and coordinated decisions and actions. Analytics cultures use data to drive strategy, tactics, and operations.

OPPORTUNITIES AND AUDIENCE

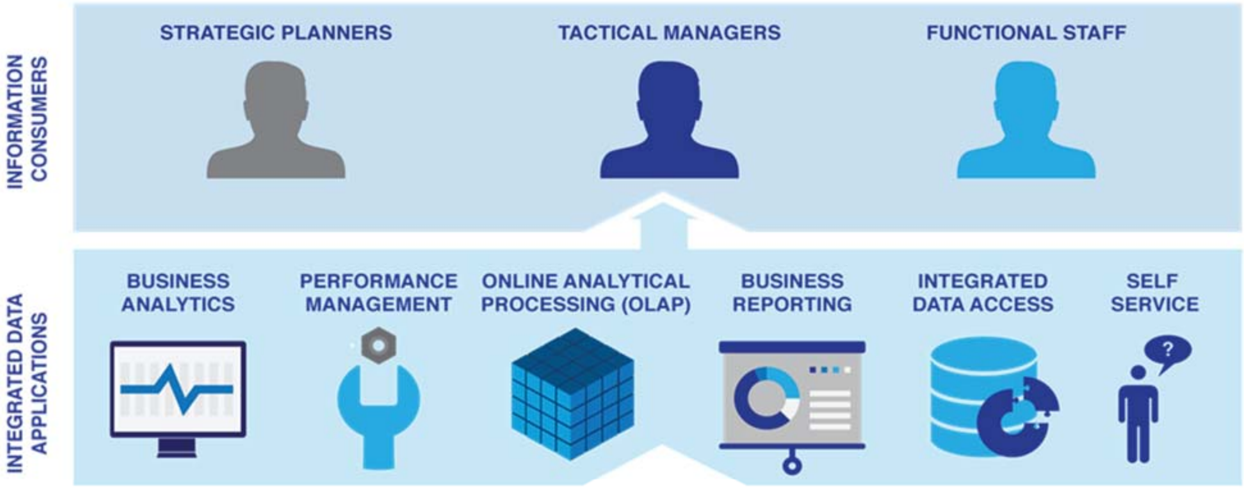
In the past, business intelligence solutions distributed information products to managers, executives, and analysts. Today, information use has expanded to encompass all levels of the organization—strategic, tactical, and operational.

DECISION MAKING

Historically, BI solutions focused on delivering data, relying on people to use it to achieve the goals. Today, solutions have evolved that can predict outcomes and even suggest a course of action.

Components

People and Applications



Components

People and Applications

PEOPLE

The ultimate consumers of business intelligence and analytics are the business executives, managers, and staff who use data and analysis to help them do the work of strategic planning, tactical management, and functional execution of business activities.

APPLICATIONS AND SERVICES

Applications are the systems and processes that access and process data and use it to deliver information to people.

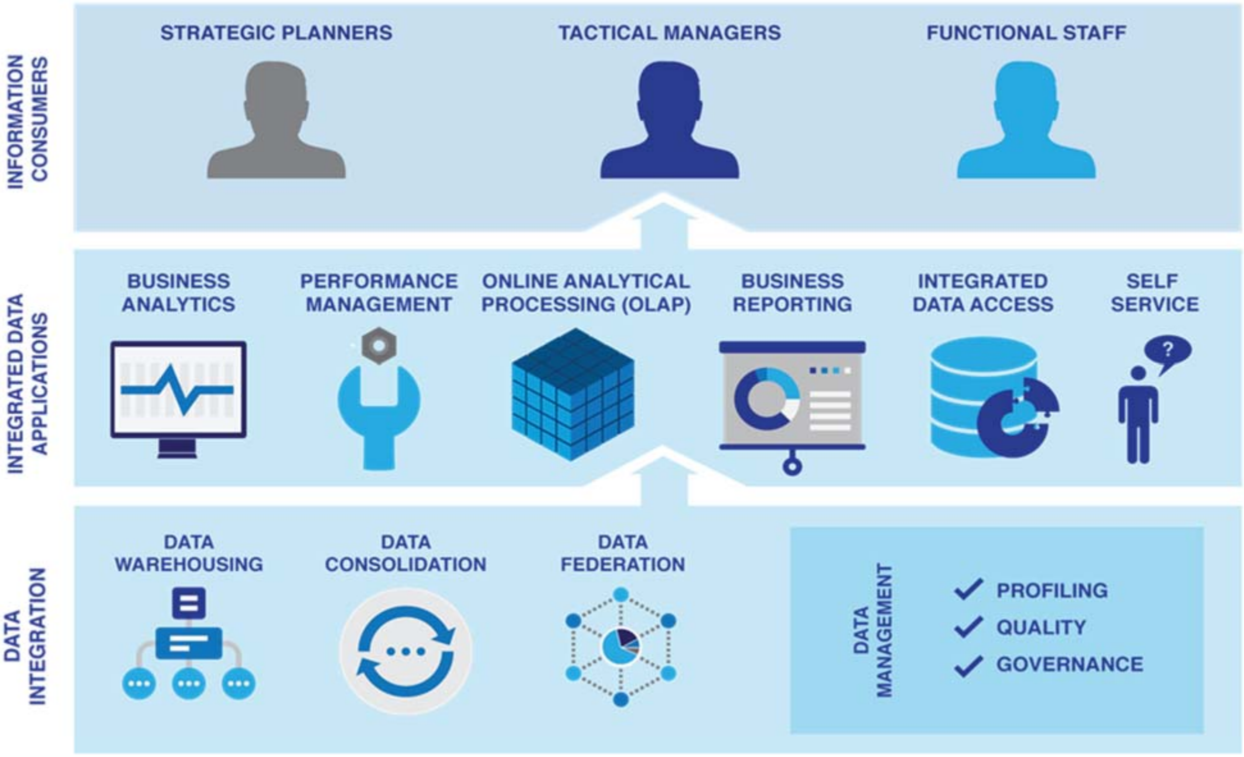
An application supports a business function or organization and provides one or more services.

For example, a human resources data mart is an application that provides OLAP services to the human resources department. A fraud detection model is an application that provides analytics services to the risk management group.

We'll discuss the service categories for BI and analytics applications in two parts. *Module 2: Business Metrics and Analytics Services* explores performance management and business analytics services. *Module 3: OLAP and Other Information Services* explores OLAP services, reporting and visualization services, data feeds, and self-service applications.

Components

Systems and Processes



Components

Systems and Processes

DATA INTEGRATION SYSTEMS

Integrated data is an essential and fundamental component of BI and analytics. Among the most vexing problems of information management are overlap, redundancy, and inconsistency across multiple sources of data.

Decision making becomes confused rather than informed when multiple data sources provide conflicting answers to questions. BI and analytics systems must reconcile and rationalize data disparity to deliver trustworthy information and enable confident decision processes.

Data integration systems include the data sources, data integration processes, and data stores that support BI and analytics applications.

Data integration systems are described in greater depth in *Module 4: Data Integration.*

DATA MANAGEMENT PROCESSES

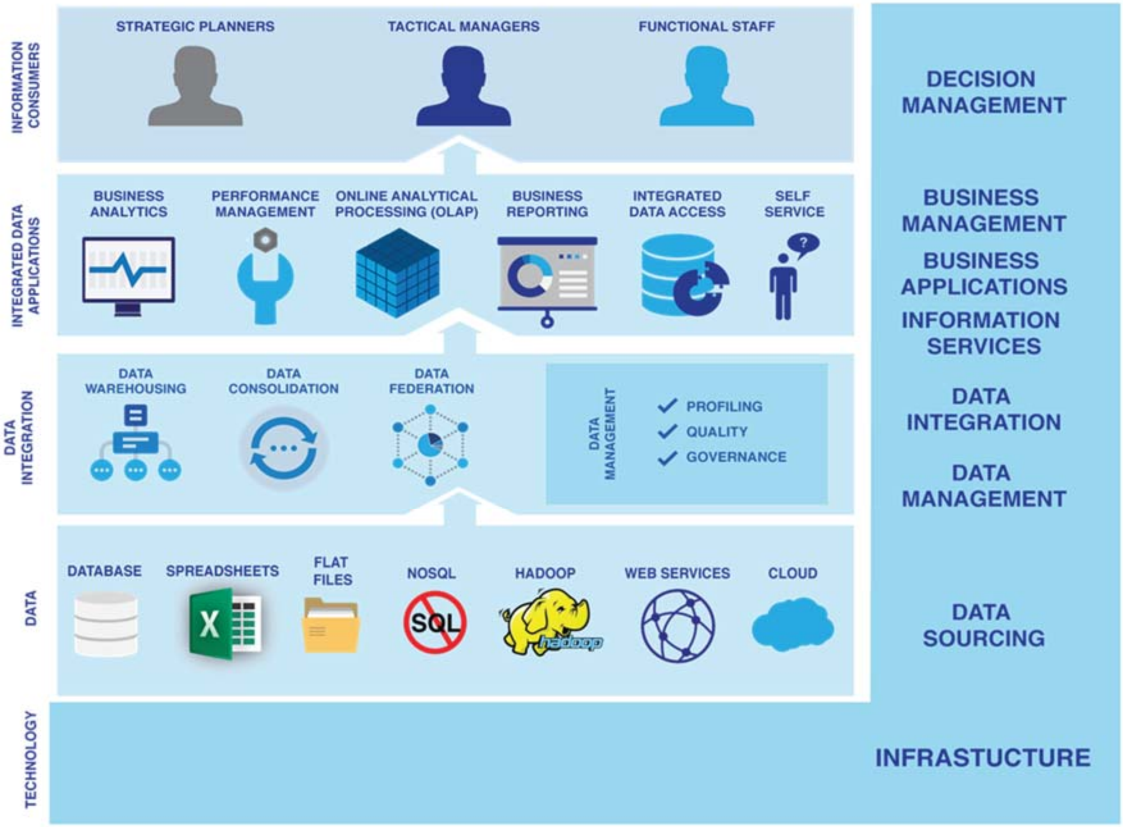
Integration alone is not sufficient to meet the data needs of BI and analytics. The data must be actively managed to ensure that it is fully understood, of high quality, secure, and used appropriately and compliant with corporate policy and externally imposed regulations.

Key data management processes include data governance, data quality management, and data profiling.

We'll discuss these processes further in *Module 5: Data Management.*

Components

Data and Technology



Components

Data and Technology

DATA

Data is the raw material from which information is created. There are a variety of sources ranging from enterprise transaction systems to big data sources. When identifying source data, consider a broad range of possibilities including:

- Databases, both internal to your organization and externally available through syndication and commercial services, are key data sources. Internal databases are likely to be the foundation for much of your information services. External databases enrich the data and offer opportunities to improve data quality.
- End-user data, including the many spreadsheets and occasional databases built to meet individual and departmental needs, may be more current and detailed than corresponding data in corporate databases. In some instances end-user data is the only available source of data needed for analysis.
- Flat files found throughout information systems are often overlooked. These are frequently the interface files between disparate applications and may have already accomplished some steps toward integration.
- Unstructured data such as text, documents, and images may enrich the data resource, particularly when data is needed for business analysis.
- A variety of data opportunities found on the Internet and embedded in email communications also offer data enrichment possibilities.

Module 4: Data Integration describes more about data types and sources.

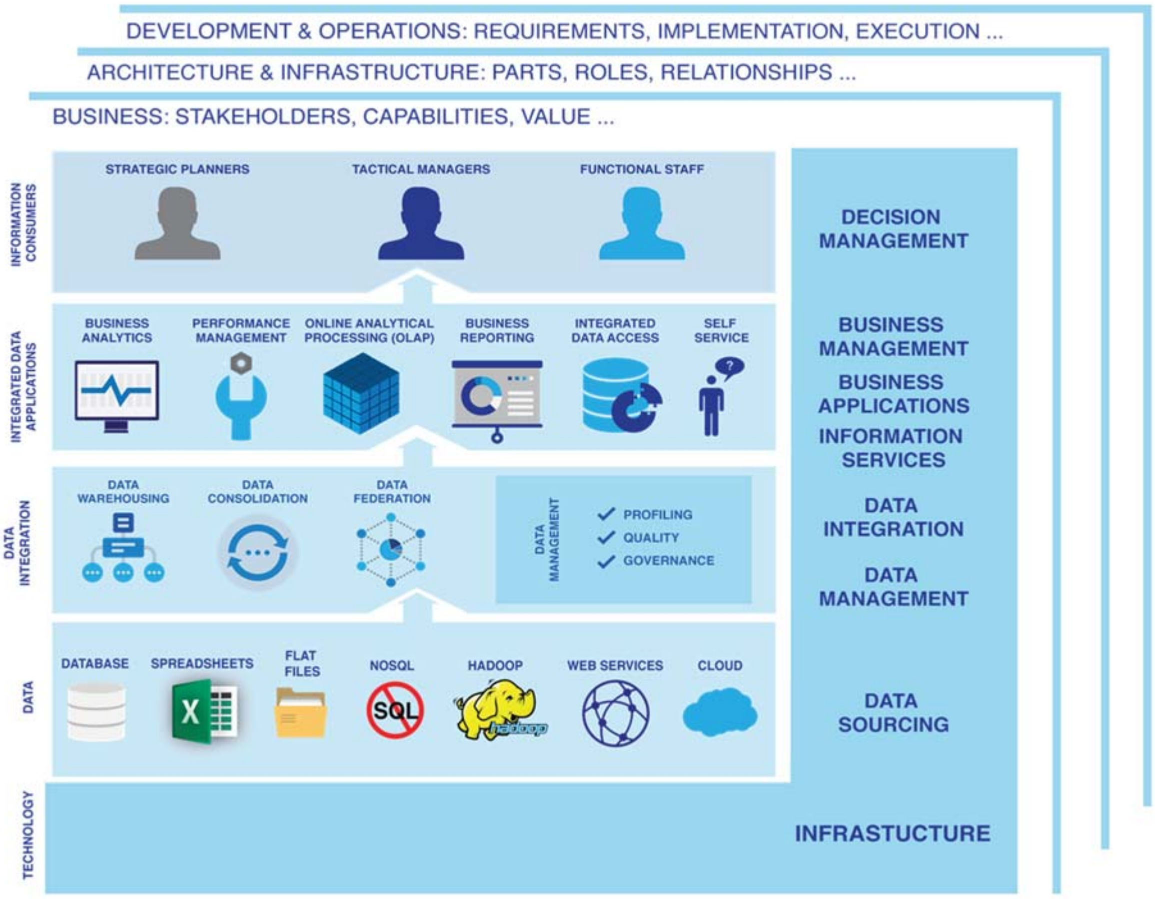
TECHNOLOGY

A variety of technology is needed to implement and operate BI and analytics systems, including support for infrastructure, data sourcing, data management, data integration, information services, business applications, business analytics, and decision management.

Module 6: BI and Analytics Technology describes the technology types, features, and functions at each level of the technology stack.

Perspectives

Points of View



Perspectives

Points of View

BUSINESS VIEW

The business view of BI and analytics is an enterprise perspective with focus on business value and sustainability. Program management, portfolio management, and stakeholder engagement are key considerations of business-driven BI and analytics.

Business perspective is particularly significant to the topics discussed in *Module 2: Business Metrics and Analytics Services,* and *Module 3; OLAP and Other Information Services.*

ARCHITECTURE & INFRASTRUCTURE VIEW

The architectural view of BI and analytics is an enterprise perspective that establishes the foundation for evolutionary and incremental development of BI and analytics systems that are cohesive, scalable, and adaptable to change. Architecture focuses on the components of BI and analytics, their roles, and the relationships among them.

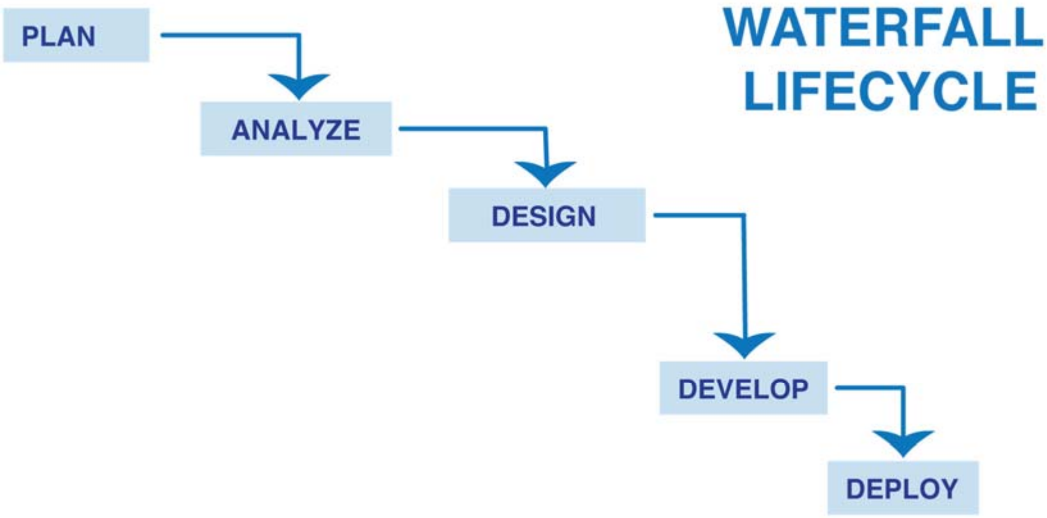
Architectural perspective has a role in every aspect of BI and analytics and implications for every module of this course. It has special significance in connecting the various parts of BI and analytics from enabling technology to information consumers.

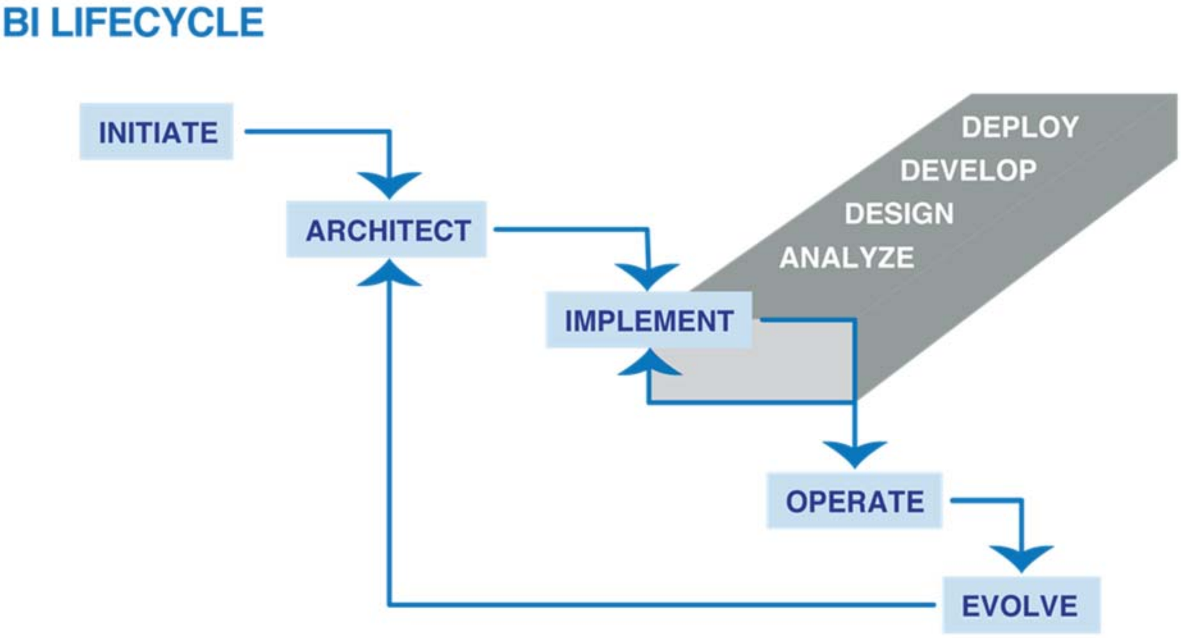
DEVELOPMENT & OPERATIONS VIEW

The development and operations view of BI and analytics is from the perspective of projects and processes. Development focuses on building new BI capabilities that fit neatly into the BI and analytics environment and systems that are already deployed. Operations is concerned with the day-to-day processes, both automated and manual, that must be carried out to make BI and analytics a living system that is naturally incorporated into business activities.

The BI and Analytics Roadmap

The BI and Analytics Lifecycle





The BI and Analytics Roadmap

The BI and Analytics Lifecycle

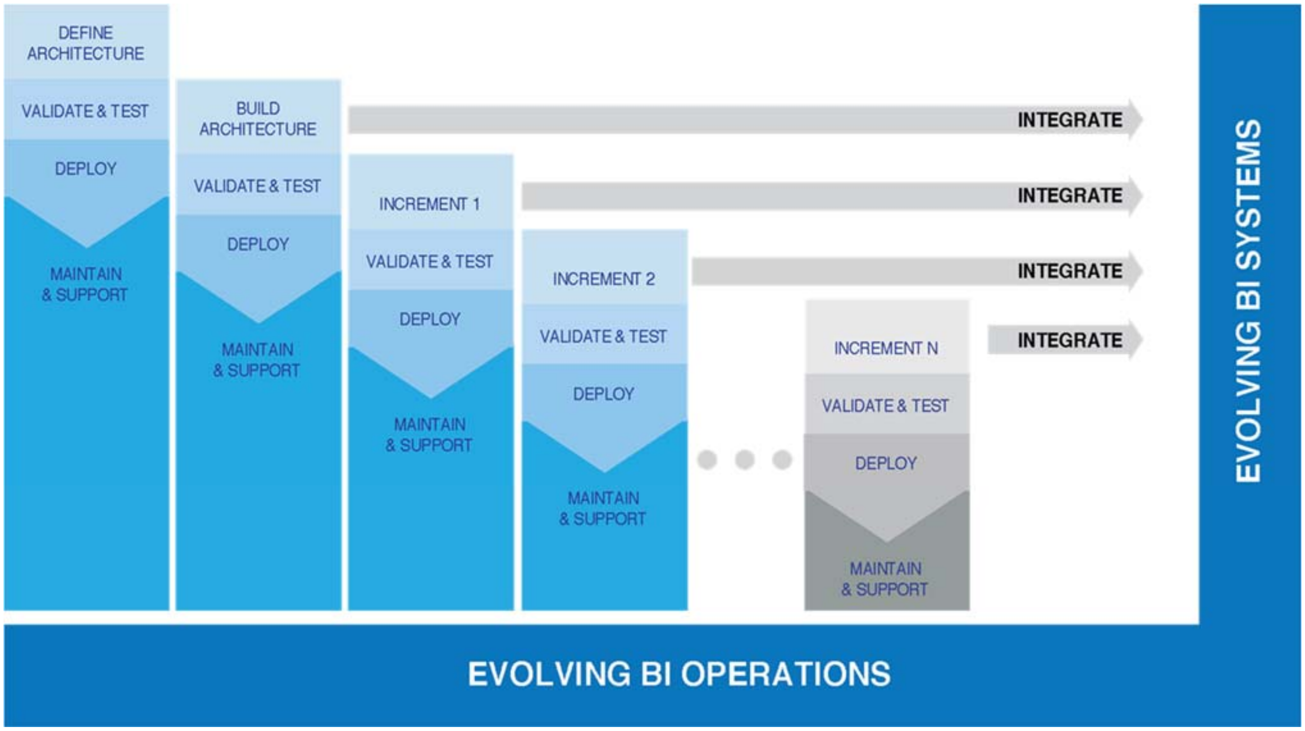
BI AND ANALYTICS PHASES

The BI and analytics lifecycle has some visual similarity to the waterfall, with a cascade effect through five phases. However, the similarities end with the visual likeness. Among the significant differences:

- The phases are distinctly different—Initiate, Architect, Implement, Operate, and Evolve.
- Where the waterfall describes a development process with a definite ending, the BI and analytics lifecycle describes a continuously evolving process with no end point.
- The BI and analytics lifecycle has feedback loops, but they are not among contiguous phases as often practiced with waterfall. The outer loop is a cycle of all phases except initiate. The inner loop is a continuous repetition of implementations.
- The implementation phase includes analysis, design, development, and deployment (the phases of waterfall) as steps or activities.

The BI and Analytics Roadmap

Evolving Capabilities



The BI and Analytics Roadmap

Evolving Capabilities

INCREMENTAL AND ITERATIVE

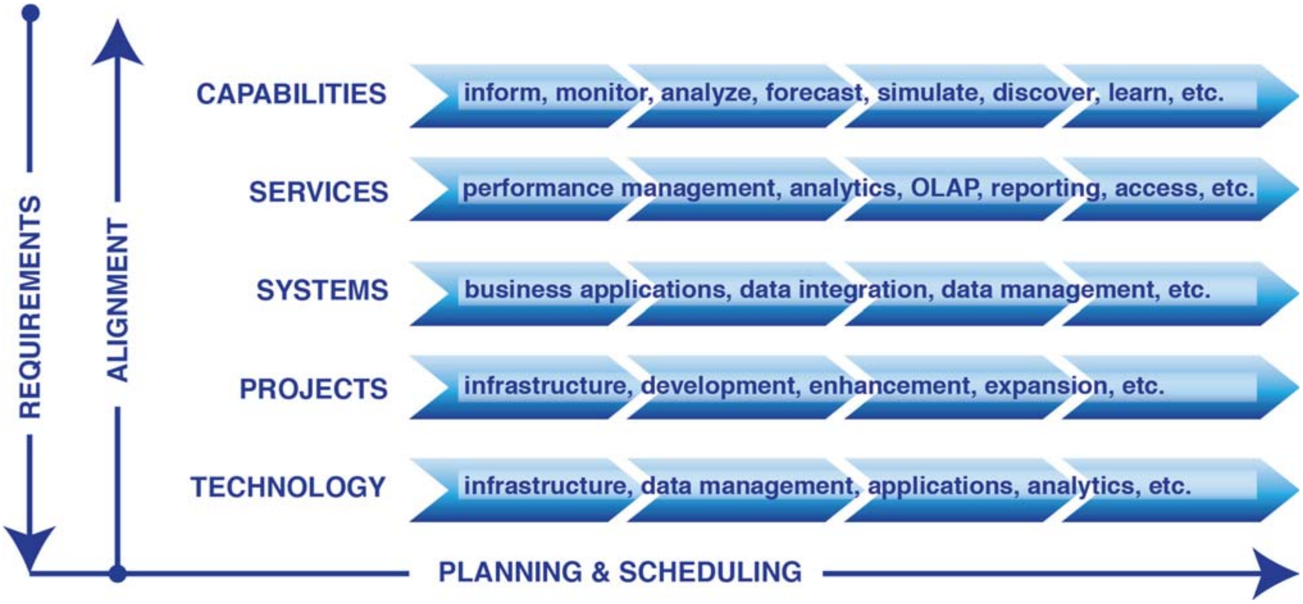
The work of building a BI and analytics system is not an isolated event. It is an ongoing process that is both incremental and iterative.

Incremental means increasing capability gradually, by regular additions. Incremental development focuses on the parts of a system—growing the range or scope by regular addition of new features or functions.

To iterate is to repeat—to perform the same activity or process over and over. An iterative process is one in which a phase or set of phases is repeated. In the BI and analytics lifecycle, iteration generally refers to the implementation phase which is repeated many times. Each cycle of iteration produces a relatively small part of the overall functionality of a BI system.

The BI and Analytics Roadmap

Parallel Paths


REQUIREMENTS
ALIGNMENT
CAPABILITIES
inform, monitor, analyze, forecast, simulate, discover, learn, etc.
SERVICES
performance management, analytics, OLAP, reporting, access, etc.
SYSTEMS
business applications, data integration, data management, etc.
PROJECTS
infrastructure, development, enhancement, expansion, etc.
TECHNOLOGY
infrastructure, data management, applications, analytics, etc.
PLANNING & SCHEDULING

The BI and Analytics Roadmap

Parallel Paths

PLANNED EVOLUTION

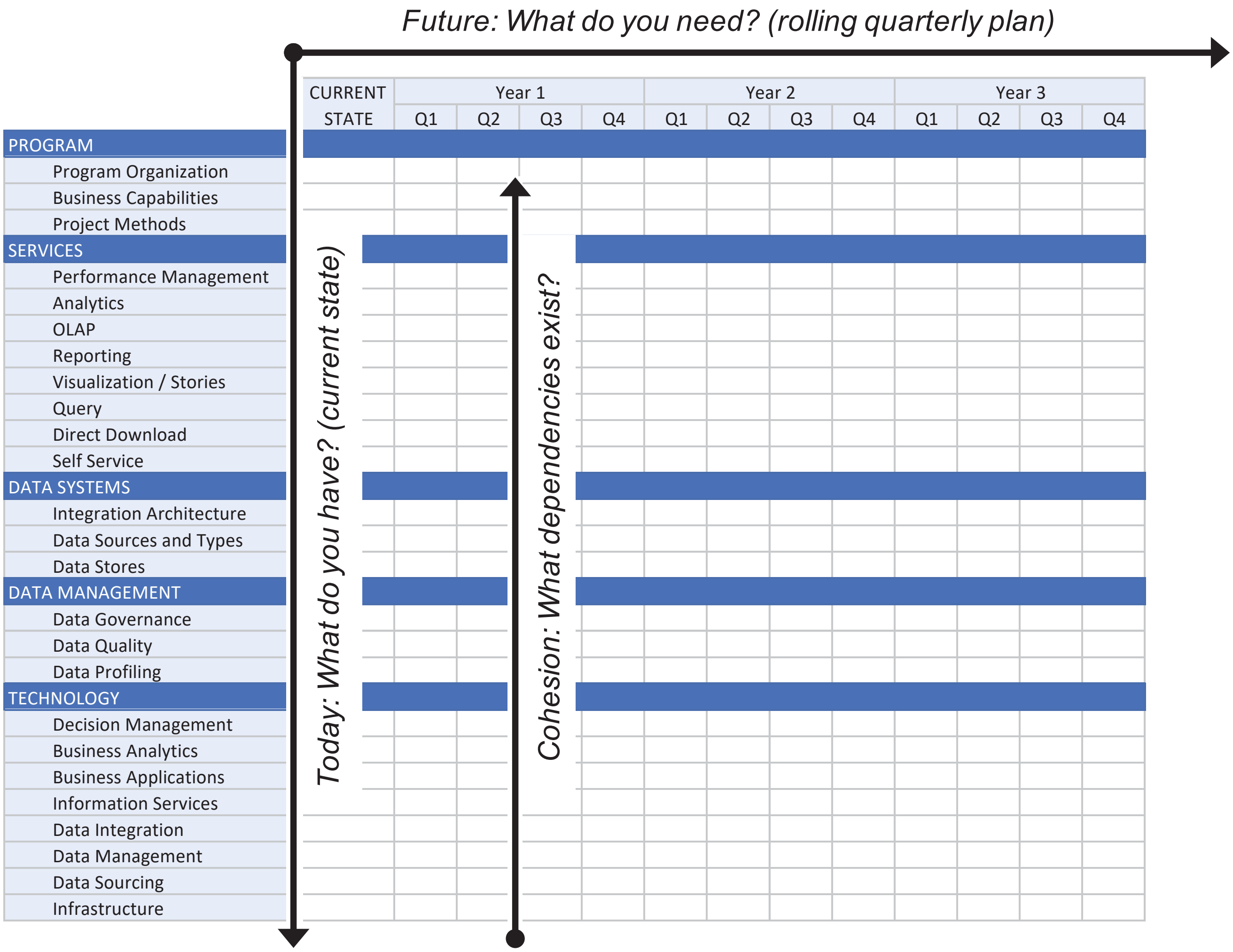
The BI and analytics roadmap is a plan for managed evolution of BI and analytics capabilities and systems. It is a timeline view of upcoming projects and activities needed to move from current state to future state as guided by portfolio management. The roadmap begins with planned capabilities and is a process of prioritization, dependency management, and planning that answers questions in four areas:

- Services—What services are needed to fill gaps or meet future business requirements? When are they needed? In what sequence?
- Systems—What systems need to be developed, enhanced, or modified to deliver the needed services? When and in what sequence?
- Projects—What projects must be planned and executed to produce the needed systems? What project dependencies exist? When and in what sequence should projects be scheduled?
- Technology—What technologies are needed to support and enable planned BI projects? When and in what sequence should technology be deployed? What projects must be planned and executed to deploy the technology?

The roadmap illustrates four parallel tracks—services, systems, projects, and technology—mapped to planned business capabilities and to a corresponding timeline.

The BI and Analytics Roadmap

Continuous Planning



Future: What do you need? (rolling quarterly plan)

Today: What do you have? (current state)

Cohesion: What dependencies exist?

	CURRENT	Year 1				Year 2				Year 3			
	STATE	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
PROGRAM													
Program Organization													
Business Capabilities													
Project Methods													
SERVICES													
Performance Management													
Analytics													
OLAP													
Reporting													
Visualization / Stories													
Query													
Direct Download													
Self Service													
DATA SYSTEMS													
Integration Architecture													
Data Sources and Types													
Data Stores													
DATA MANAGEMENT													
Data Governance													
Data Quality													
Data Profiling													
TECHNOLOGY													
Decision Management													
Business Analytics													
Business Applications													
Information Services													
Data Integration													
Data Management													
Data Sourcing													
Infrastructure													

The BI and Analytics Roadmap

Continuous Planning

BUILDING THE ROADMAP

The illustration on the facing page suggests how to structure your BI and analytics roadmap. The roadmap will be revisited at the end of each module of this course. For now, study its overall structure.

CURRENT STATE

The roadmap begins with a view of the current state. Planning for the future depends on first understanding what you have today. To begin the process of developing a roadmap, you may need to begin with an internal or external assessment of your business capabilities, organization, architecture and methodology, data management, and decision making—possibly using one of the maturity models described on the coming pages may be useful.

FUTURE STATE

The future view is a look ahead, typically at the next two to three years. The plan begins by looking at needed capabilities and plotting them on a timeline.

ROLLING PLAN

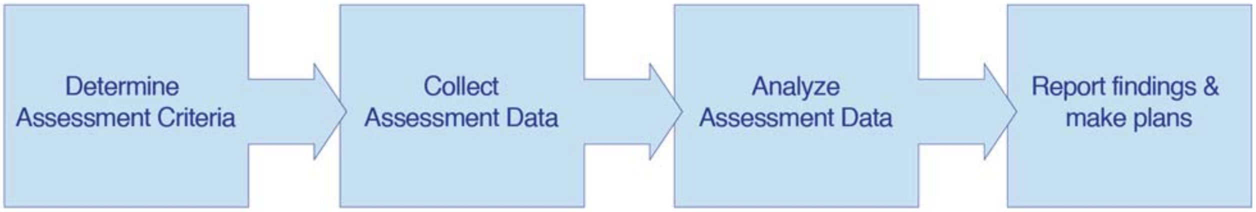
Planning is a continuous process of responding to change in business needs and technology growth. As the business intelligence and analytics ecosystem evolves and technical growth continues, the plan must be adapted and adjusted. A rolling quarterly plan works well for the roadmap. Plot your plan with the greatest confidence and detail for the immediate next quarter and with fewer specifics as you move further into the future. Then update the plan once each quarter by:

- Updating the current state inventory
- Adapting and adjusting to change
- Adding detail and specifics for upcoming quarters
- Extending the timeline by one quarter to sustain the 2-3 year look ahead

BI and Analytics Maturity

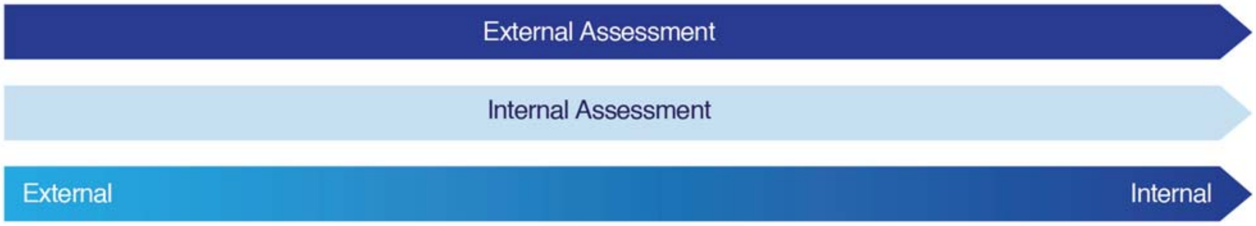
Readiness Assessment

ASSESSMENT PROCESS



Determine Assessment Criteria → Collect Assessment Data → Analyze Assessment Data → Report findings & make plans

ASSESSMENT OPTIONS

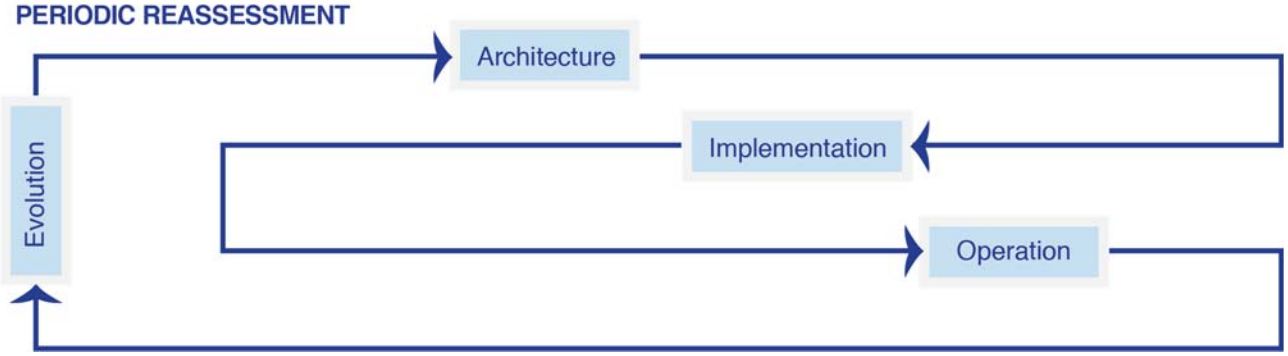


External Assessment

Internal Assessment

External → Internal

PERIODIC REASSESSMENT



Architecture

Implementation

Operation

Evolution

BI and Analytics Maturity

Readiness Assessment

ASSESSMENT OVERVIEW

Program assessment is a structured process for appraising and measuring characteristics of a business intelligence and analytics program. Program assessment examines characteristics such as value, quality, satisfaction, alignment with best practices, and probability of success.

Business intelligence and analytics programs are challenging endeavors. Recognizing and responding to the strengths, weaknesses, and risks in your program is essential to ensuring success and delivering real value. Changing business needs, user expectations, and technologies compound the challenge and drive the need for regular assessment.

ASSESSMENT TYPES

Assessments are equally valuable for organizations starting out and for ones with an ongoing program. For enterprises starting out, a readiness assessment should be conducted, with emphasis on the factors that need to be in place to maximize the likelihood of success. Companies with an ongoing program should conduct an interim assessment, which emphasizes an examination of what’s going well, what’s not, and what needs to be done to be poised to meet future needs.

EXTERNAL AND INTERNAL ASSESSMENTS

Two techniques—external assessment and self-assessment—are common for evaluating business intelligence and analytics readiness. These techniques may be used, individually or in combination, to meet particular assessment needs:

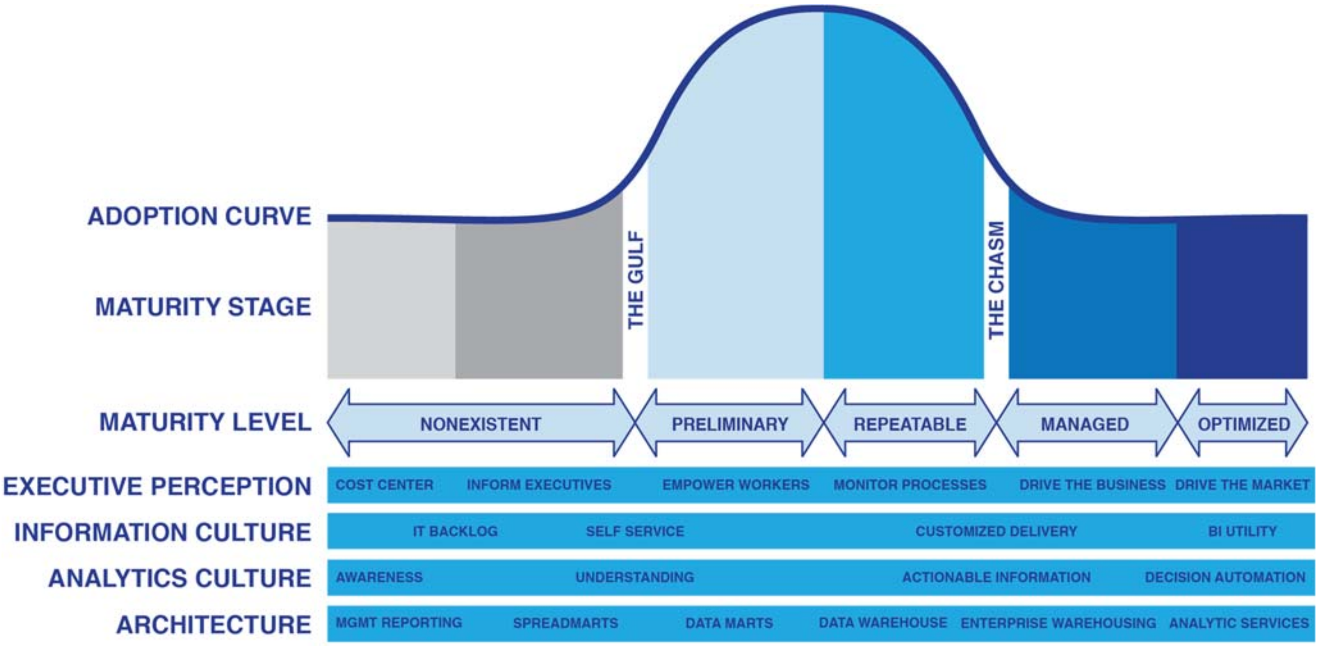
An external assessment is typically a formal report produced by an unbiased, experienced consultant. The report should describe both the conclusions and the processes and data used to reach them. Business, IT, and technology readiness should all be examined. The roadmap is the most important part of an external assessment.

An internal assessment measures an organization’s perceptions about its existing or planned business intelligence and analytics activities. It typically identifies strengths, weaknesses, and risk areas and is a good way to create awareness of any major issues.

Assessments are only useful when the results are used to adapt the program plan. Findings from assessments help to both plan for success and adapt the plan to organizational and individual realities.

BI and Analytics Maturity

Maturity Models



BI and Analytics Maturity

Maturity Models

BI MATURITY MODEL

The diagram on the facing page illustrates some aspects of TDWI's BI Maturity Model. The model illustrates evolution through the stages of BI from early (management reports and spreadsheets) to optimized (enterprise data warehousing and analytics services). This model provides useful context for BI assessment because it describes many BI success factors.

OTHER MATURITY MODELS

Maturity models exist for other aspects of the business intelligence and analytics ecosystem. TDWI's website (www.tdwi.org) provides self-assessment tools for evaluating maturity in several areas, including:

- Business Analytics
- Big Data
- Hadoop
- The Internet of Things (IoT)

Each of these assessment tools includes questionnaires related to the area being assessed. Upon completing the questionnaire, the tool provides scores that quantify your readiness. These help you focus your efforts to best move forward.

Other maturity models are available from various vendors.

Mistakes to Avoid

When Validating Direction

- ONLY FOCUSING ON MEETING CURRENT REQUIREMENTS
- ASSUMING THE ARCHITECTURE AND METHODOLOGY ARE SOUND
- IGNORING THE BUSINESS PERSPECTIVE
- IGNORING THE TECHNICAL PERSPECTIVE
- IGNORING OTHER INITIATIVES
- ONLY FOCUSING ON QUERY AND REPORTING NEEDS
- IGNORING THE INDUSTRY AND BUSINESS DIRECTION
  IGNORING DATA QUALITY
- IGNORING DATA GOVERNANCE
- VALIDATING YOUR BI/DW DIRECTION HAPHAZARDLY

Source: *Ten Mistakes to Avoid When Validating Your Business Intelligence or Data Warehousing Direction* by Jonathan Geiger. © TDWI

Mistakes to Avoid

When Validating Direction

TEN MISTAKES TO AVOID

Ten Mistakes to Avoid When Validating Your Business Intelligence or Data Warehousing Direction by Jonathan Geiger.

It is often said that "If it ain't broke, don't fix it," but how do you know if your program isn't broken? Even if it meets some business needs and provides business users with data to support analyses and decision making, there are always opportunities for improvement. Companies should periodically review where they've been with their program and more important, where they're heading. However, exploring your direction without a well-thought-out approach and ignoring critical areas may lead to inaccurate conclusions about the environment and suboptimal improvements. Mistakes to avoid are shown on the facing page.

Mistakes to Avoid

When Delivering Business-Driven BI

- NOT SOLVING A REAL BUSINESS PROBLEM
- HAVING “SOLUTION ENVY”
- NOT INCLUDING BUSINESS USERS WHEN DATA MODELING
- LACK OF ONGOING COMMUNICATION
- DESIGNING DASHBOARDS WITHOUT DATA
- NOT PROVIDING EDUCATION
- NOT PROVIDING DOCUMENTATION
- PUTTING AN EMPHASIS ON SIZZLE, NOT QUALITY
- IMPLEMENTING SELF-SERVICE WITHOUT A LIFE PRESERVER
- NOT MEASURING IMPACT

Mistakes to Avoid

When Delivering Business-Driven BI

TEN MISTAKES TO AVOID

Ten Mistakes to Avoid When Delivering Business-Driven BI by Laura Reeves.

Most organizations have a reporting mechanism in place. Too often, these reports are not enough to meet all business needs. There is growing pressure to move forward to an environment that supports more sophisticated analysis. In this vision, business users can use the environment themselves to access and manipulate the necessary data and drive the analytics. This is commonly called *self-service business intelligence* (BI). The most successful BI solutions are those whose design and subsequent use are driven by the business itself.

This is much easier said than done. It is easy to get caught up in a frenzy of activity in the attempt to build and deliver something. Too often, what is delivered is not well-received by the business community, or worse, met with disappointment or resistance. The most common mistakes for avoiding this are shown on the facing page.

Discussion

Introduction to BI and Analytics

LET’S TALK ABOUT IT!

- How do your definitions for business intelligence and analytics relate to the ones presented?
- What is the current state of your BI and analytics program? How do you know?
- Do you have a roadmap?
- What is the best way for your organization to move forward?

Discussion

Introduction to BI and Analytics

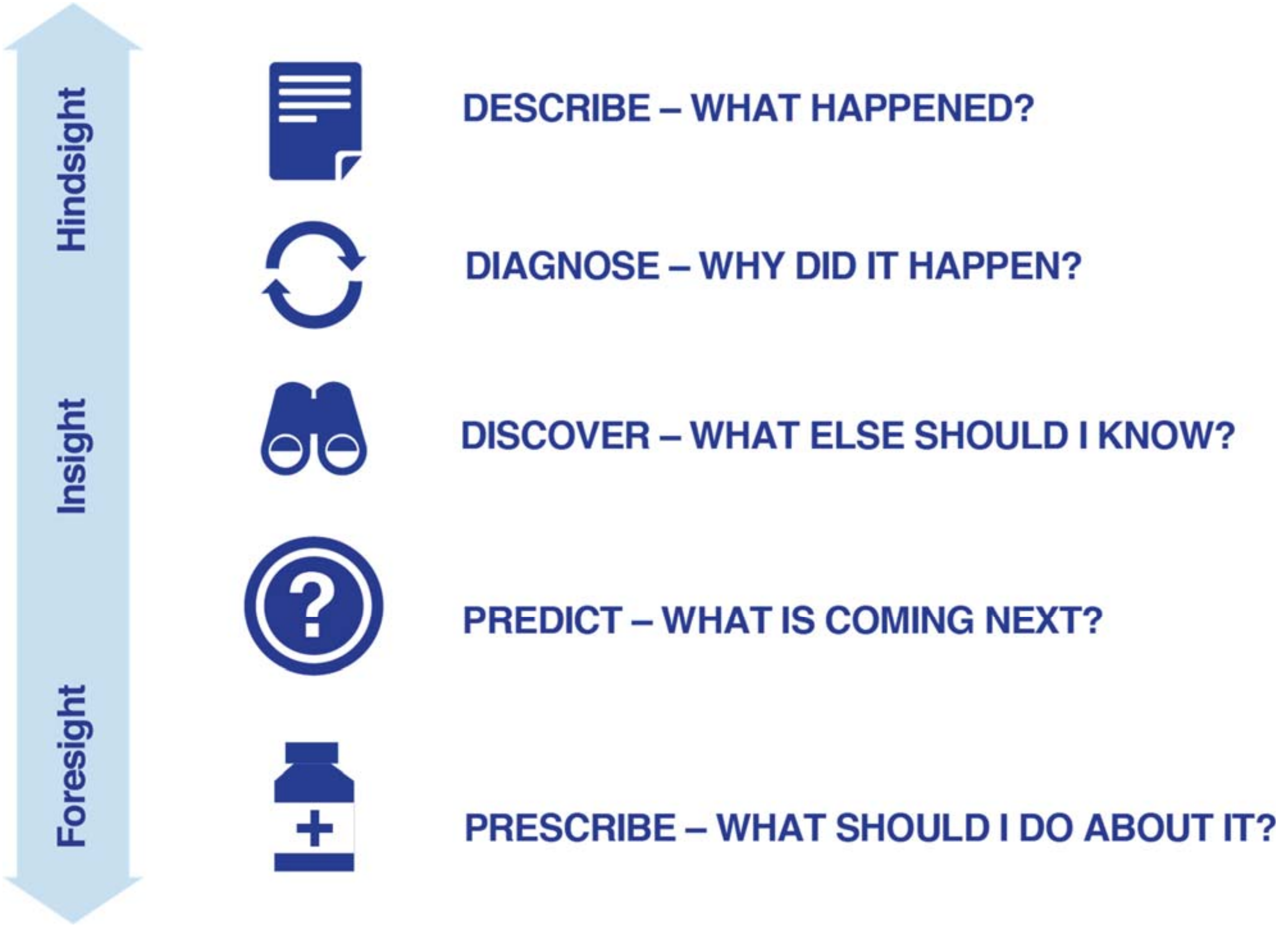
Notes:

Module 2

Business Metrics and Analytics

Business Capabilities

Purpose



Hindsight

Insight

Foresight

DESCRIBE – WHAT HAPPENED?

DIAGNOSE – WHY DID IT HAPPEN?

DISCOVER – WHAT ELSE SHOULD I KNOW?

PREDICT – WHAT IS COMING NEXT?

PRESCRIBE – WHAT SHOULD I DO ABOUT IT?

Business Capabilities

Purpose

PURPOSE

Business intelligence and analytics are not “one size fits all” endeavors. Five kinds of business capability can be delivered—each with a different purpose:

- *Descriptive business intelligence and analytics* provide information about past actions, events, and outcomes: What happened? How much? When did it happen?

- *Diagnostic business intelligence and analytics* also look at past events, but with focus on cause and effect: Why do things happen? Why do outcome measures go up or down? What are the influences on business outcomes? Why do things happen at specific points in time or on particular time cycles?

- *Discovery business intelligence and analytics* seek to learn that which is unknown. It can be thought of as learning analytics—seeking new and useful knowledge. Bridging from data to business, discovery analytics may pursue data insights, business insights, or both.

- *Predictive business intelligence and analytics* use data to look into the future, forecasting probabilities of future events and conditions—to answer questions about what is likely to happen. The forecasts are typically related to expected behaviors of individuals (customers, employees, machines, parts, etc.) and are used in planning, goal setting, and decision making.

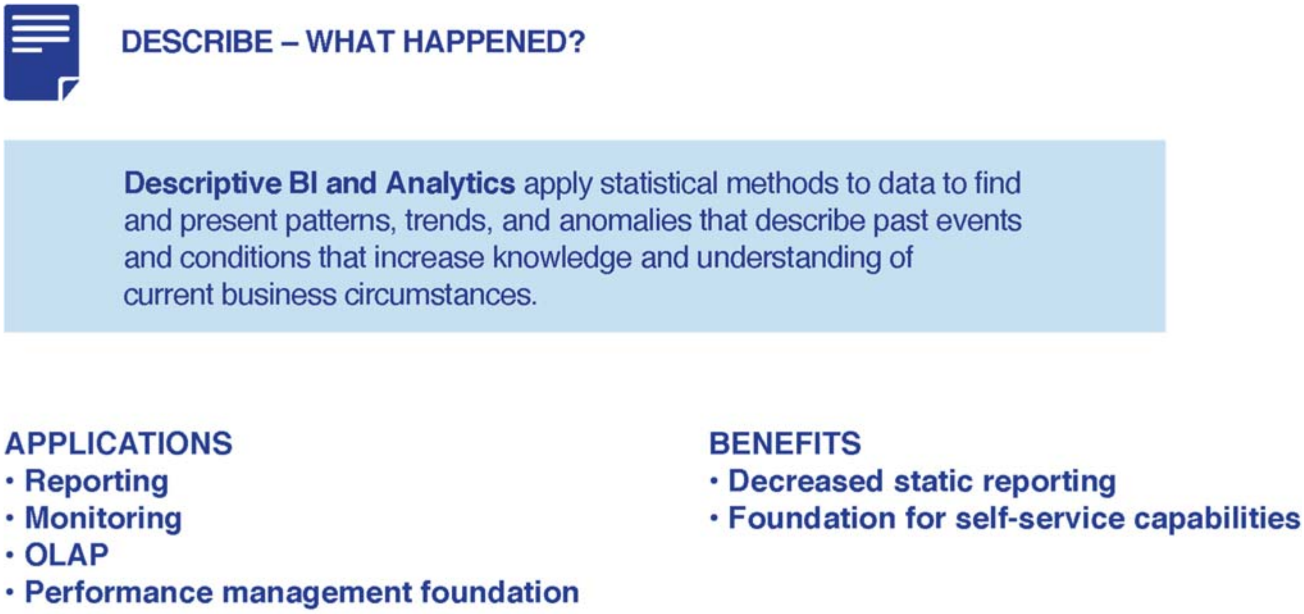
- *Prescriptive business intelligence and analytics* are used to recommend and/or automate decisions with the goal of optimizing future outcomes. Prescriptions as recommendations help decision makers to choose among alternative responses to a set of circumstances. When prescription is used to automate decision making, the analytics model determines a single best response.

CAPABILITY DEPENDENCIES

Though not a hard-and-fast rule, there are some dependencies among the capabilities. Diagnostic work is difficult without first performing descriptive work to understand the nature of the data and the events that it describes. Discovery work benefits from results of descriptive and diagnostic findings. Predictive work benefits from discovery, diagnosis, and description, and prescriptive capabilities are built on a foundation of predictive insights.

Business Capabilities

Descriptive BI and Analytics



DESCRIBE – WHAT HAPPENED?

Descriptive BI and Analytics apply statistical methods to data to find and present patterns, trends, and anomalies that describe past events and conditions that increase knowledge and understanding of current business circumstances.

APPLICATIONS

- **Reporting**
- **Monitoring**
- **OLAP**
- **Performance management foundation**

BENEFITS

- **Decreased static reporting**
- **Foundation for self-service capabilities**

Business Capabilities

Descriptive BI and Analytics

WHAT HAPPENED?

Descriptive business intelligence and analytics describe what has happened in the past. The purpose of descriptive analytics is to provide quantitative descriptions of systems, processes, activities, events, and other domains of interest over time and at various points in time. Quantitative descriptions provide a foundation for process monitoring, trend analysis, and performance management.

BUSINESS APPLICATIONS

The most basic technique of descriptive analytics is measurement—turning data into metrics that quantify things of business interest. Descriptive statistics, time series analysis, and basic visualization techniques may also be applied to find and present the meaning in the data.

Descriptive business intelligence and analytics provide standard and ad hoc query, reporting, and monitoring capabilities. These combine to provide a foundation for performance management.

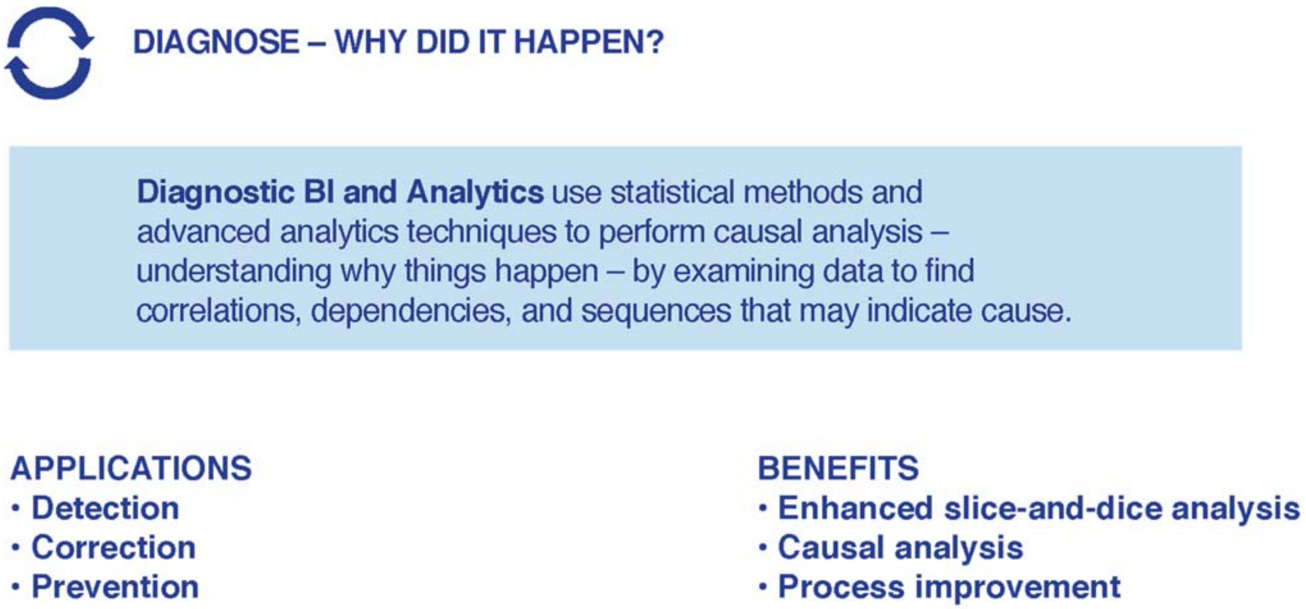
VALUE

Business and technical benefits from descriptive business intelligence and analytics include:

- Reduced need for static reports
- Foundation for self-service capabilities

Business Capabilities

Diagnostic BI and Analytics



DIAGNOSE – WHY DID IT HAPPEN?

Diagnostic BI and Analytics use statistical methods and advanced analytics techniques to perform causal analysis – understanding why things happen – by examining data to find correlations, dependencies, and sequences that may indicate cause.

APPLICATIONS

- **Detection**
- **Correction**
- **Prevention**

BENEFITS

- **Enhanced slice-and-dice analysis**
- **Causal analysis**
- **Process improvement**

Business Capabilities

Diagnostic BI and Analytics

WHY DID IT HAPPEN?

The purpose of diagnostic business intelligence and analytics is to detect unusual situations or abnormal conditions in the context of a business or operational process. This detection capability is combined with root cause analysis methods to gain insights into causal relationships driving the observed or predicted fault. It provides fault detection and root cause analysis capabilities to the different categories of people who need to respond to the situation and take corrective action.

Remember that the various types of analytics enable and support each other—this form of analytics depends on discovery analytics for base models and descriptive analytics for statistical properties of key processes. Diagnostic analytics detects and identifies current unusual conditions. However, this capability can be combined with predictive and prescriptive analytics to create predictive diagnostics, identifying potential problems before they actually occur. Prescriptive analytics may apply to recommend action for either detected or predicted problems.

BUSINESS APPLICATIONS

Diagnostic BI and analytics aim to understand why things happened. The capabilities that are delivered are detection—understanding the underlying causes, correction—dealing with symptoms to improve the situation, and prevention—dealing with the root causes to prevent recurrences.

These capabilities build on the basic OLAP and performance management foundation provided with descriptive BI and analytics

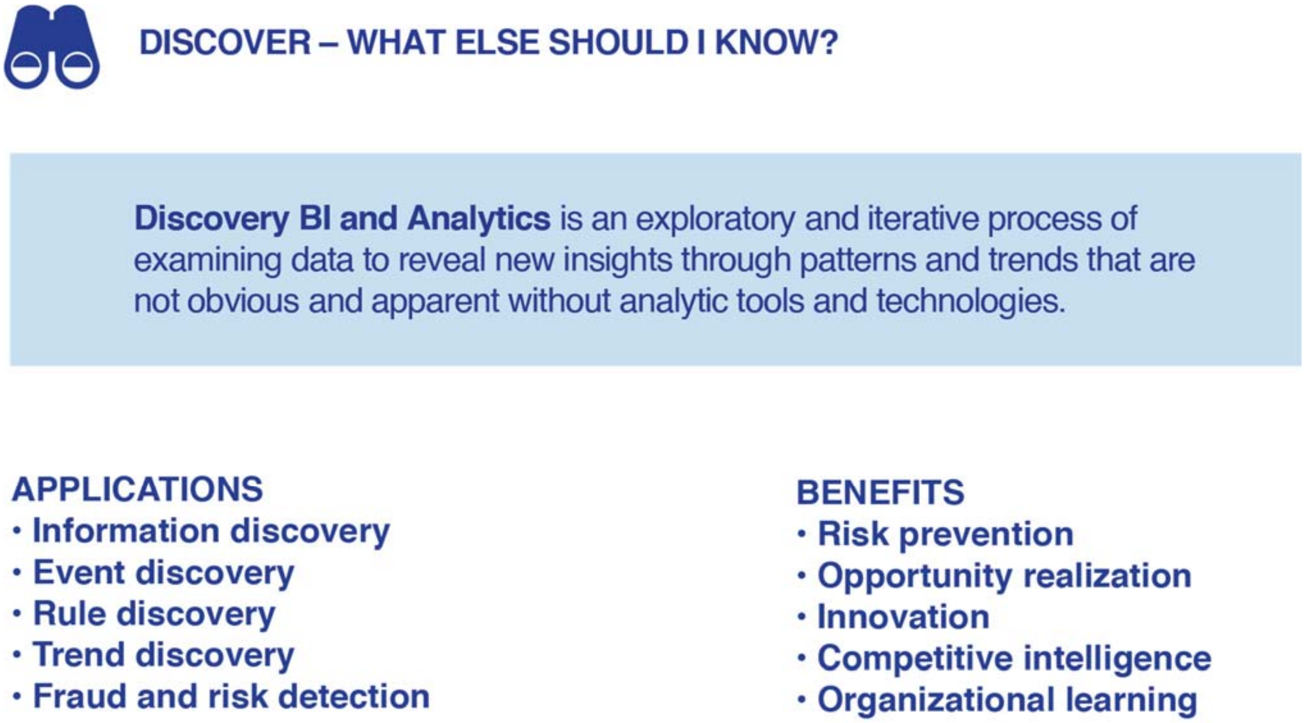
VALUE

Business and technical benefits from diagnostic business intelligence and analytics include:

- Enhanced slice-and-dice analysis
- Causal analysis
- Process improvement

Business Capabilities

Discovery BI and Analytics



DISCOVER – WHAT ELSE SHOULD I KNOW?

Discovery BI and Analytics is an exploratory and iterative process of examining data to reveal new insights through patterns and trends that are not obvious and apparent without analytic tools and technologies.

APPLICATIONS

- **Information discovery**
- **Event discovery**
- **Rule discovery**
- **Trend discovery**
- **Fraud and risk detection**

BENEFITS

- **Risk prevention**
- **Opportunity realization**
- **Innovation**
- **Competitive intelligence**
- **Organizational learning**

Business Capabilities

Discovery BI and Analytics

WHAT ELSE SHOULD I KNOW?

The purpose of discovery business intelligence and analytics is to gain new knowledge by using data and analytics techniques to create previously unknown information and insights. Discovery techniques are applied to learn new things both about the data and about the business.

BUSINESS APPLICATIONS

Through the application of various techniques, such as advanced data visualization, data and text mining, classification and clustering, linear and logistic regression, and time series analysis, improved understanding about information, events, rules, and trends is acquired.

Discovery is an iterative process where techniques are applied in a sequence of explore, discover, verify, operationalize, and repeat.

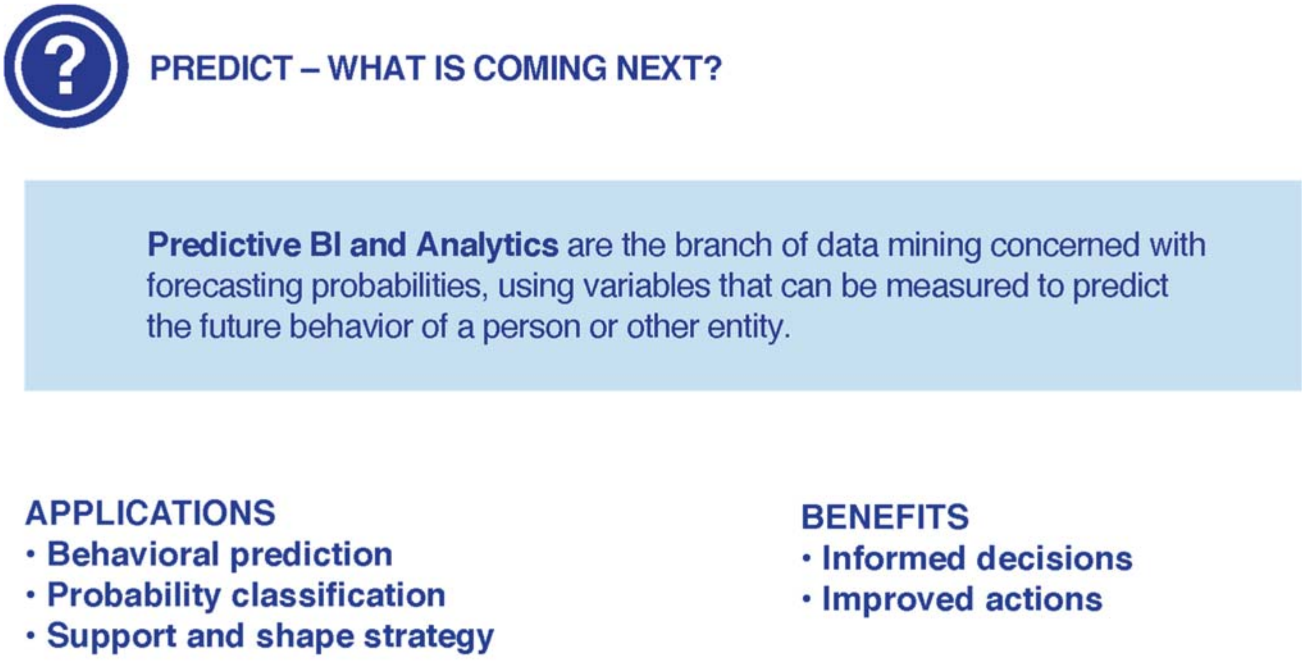
VALUE

Business and technical benefits from discovery business intelligence and analytics include:

- Risk Prevention
- Opportunity Realization
- Innovation
- Competitive Intelligence
- Organizational Learning

Business Capabilities

Predictive BI and Analytics



PREDICT – WHAT IS COMING NEXT?

Predictive BI and Analytics are the branch of data mining concerned with forecasting probabilities, using variables that can be measured to predict the future behavior of a person or other entity.

APPLICATIONS

- **Behavioral prediction**
- **Probability classification**
- **Support and shape strategy**

BENEFITS

- **Informed decisions**
- **Improved actions**

Business Capabilities

Predictive BI and Analytics

WHAT NEXT?

Predictive business intelligence and analytics look into the future to forecast probabilities of future outcomes. They answer questions about what is likely to happen by estimating the probability of events and conditions. Predictive modeling is highly dependent on the results of discovery analytics.

It is important to understand that predictive models do not demand well-understood cause and effect relationships to provide useful predictions. Input variables called predictive or explanatory variables are used as predictors. They need only to be sufficiently correlated to a predicted variable to be useful. Predictive analytics is not concerned with knowing why. That is the domain of diagnostic business intelligence and analytics.

BUSINESS APPLICATIONS

Predictive BI and analytics are specific subsets of data mining that employ the common mining techniques of classification, segmentation, association, sequencing, and forecasting to build predictive models. Two common types of models (from *Introduction to Modeling Techniques in Predictive Analytics*, Thomas W. Miller) are:

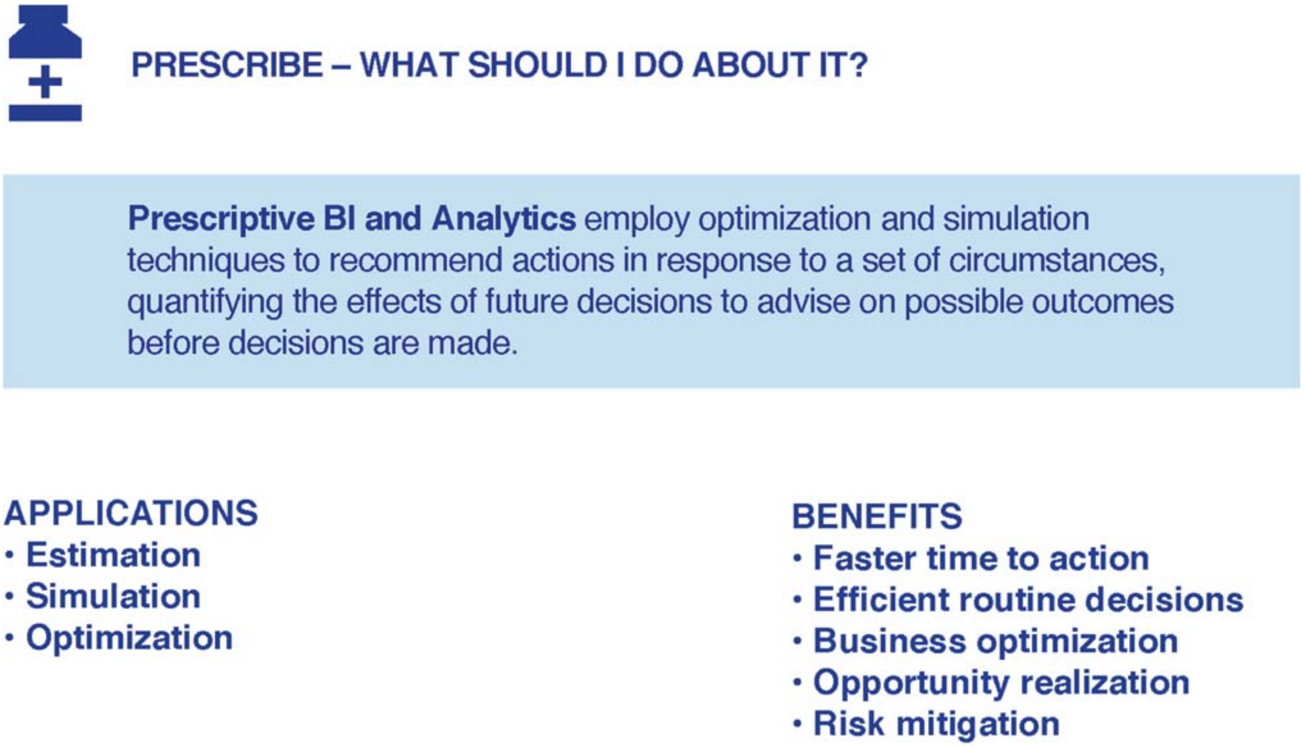
- *Regression:* This involves “predicting a response with meaningful magnitude, such as quantity sold, stock price, or return on investment.”
- *Classification:* This involves “predicting a categorical response. Which brand will be purchased? Will the consumer buy the product or not? Will the account holder pay off or default on the loan? Is this bank transaction true or fraudulent?”

VALUE

Some business applications and key benefits enabled by predictive analytics are presented on the facing page. There are substantial strategic benefits provided by predictive analytics. When you consider that strategy is the work of shaping your future, what could be more valuable than analytics that help to see into that future?

Business Capabilities

Prescriptive BI and Analytics



PRESCRIBE – WHAT SHOULD I DO ABOUT IT?

Prescriptive BI and Analytics employ optimization and simulation techniques to recommend actions in response to a set of circumstances, quantifying the effects of future decisions to advise on possible outcomes before decisions are made.

APPLICATIONS

- **Estimation**
- **Simulation**
- **Optimization**

BENEFITS

- **Faster time to action**
- **Efficient routine decisions**
- **Business optimization**
- **Opportunity realization**
- **Risk mitigation**

Business Capabilities

Prescriptive BI and Analytics

WHAT TO DO?

Prescriptive business intelligence and analytics enable decision makers to explore, evaluate, and select a course of action in response to existing or expected problems, needs, or circumstances. This type of analysis prescribes decisions and actions to be taken given a set of assumptions, options, and scenarios.

The purpose of prescriptive business intelligence and analytics is to guide decision makers to a feasible course of action based on simulation techniques, an optimal course of action based on optimization techniques, or a recommended course of action based on a rules engine.

BUSINESS APPLICATIONS

Simulation and optimization are core techniques as described in the purpose statement above. Estimation is also used in prescriptive analytics, especially when recommendations need to quantify measurable outcomes of each alternative decision.

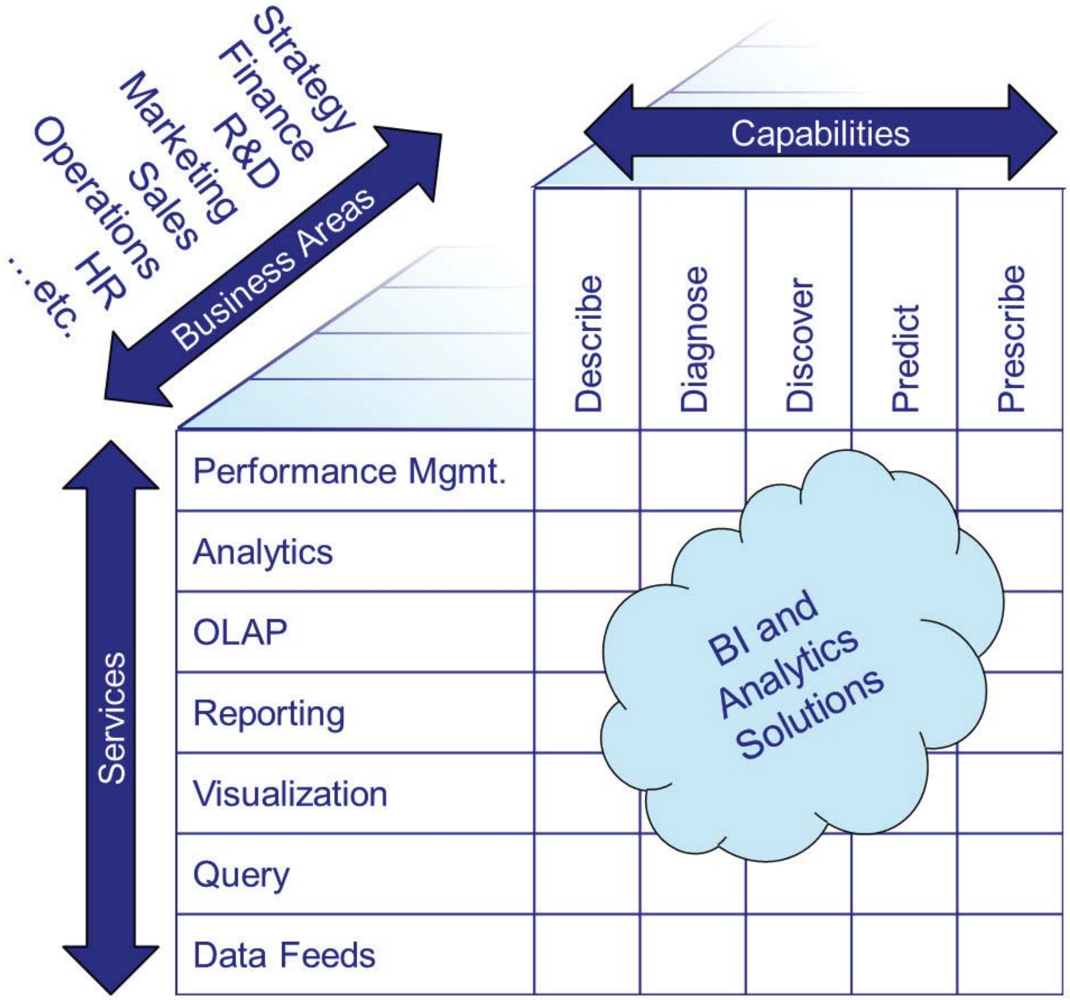
BUSINESS VALUE

Business and technical benefits from prescriptive business intelligence and analytics include:

- Faster Time to Action
- Efficient Routine Decisions
- Business Optimization
- Opportunity Realization
- Risk Mitigation

Business Capabilities

Capabilities Through Services



Business Capabilities

Capabilities Through Services

DELIVERING VALUE

There is not a one-to-one mapping between a business capability and a business intelligence solution.

For example, OLAP services may provide descriptive or diagnostic capabilities; performance management services may provide descriptive or prescriptive capabilities.

SERVICES FOCUS

As the illustration suggests, business applications are solutions built for specific business functions. They provide capabilities through one or more services.

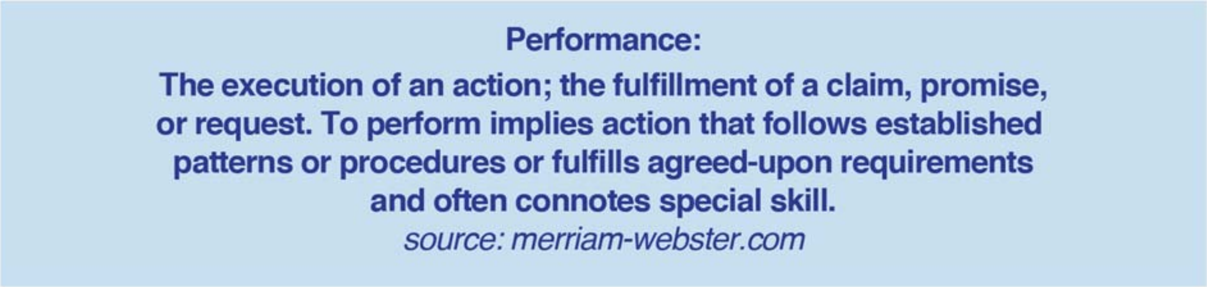
Services include:

- *Performance management services* identify and track key performance indicators and associated goals, surfacing them through dashboards and scorecards.
- *Analytics services* leverage analysis, statistics, and data mining to generate insights that are deeper than what can be revealed by surface-level inspection.
- *OLAP services* provide the ability to view historical data from various perspectives and levels of aggregation. Often referred to as "slice and dice analysis."
- *Reporting services* provide information needed on a regular and predictable basis at strategic, tactical, and operational levels of the business.
- *Visualization and storytelling services* facilitate communication of BI and analytics insights, helping channel them into actions and business impact.
- *Query services* allow businesspeople to begin their own inquiries into data resources through ad hoc and managed access environments.
- *Self-service BI and analytics services* enable businesspeople to bring their own data resources into an analytics environment.

The remainder of this module will explore the first two services in this list. The rest are explored in *Module 3: OLAP and Other Information Services.*

Performance Management

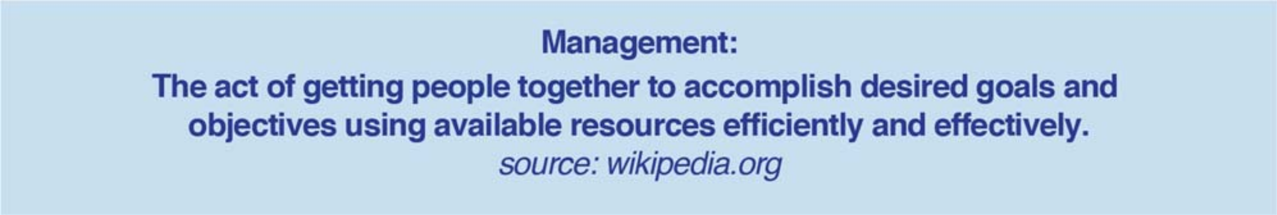
Definition and Concepts



Performance:

The execution of an action; the fulfillment of a claim, promise, or request. To perform implies action that follows established patterns or procedures or fulfills agreed-upon requirements and often connotes special skill.

source: merriam-webster.com



Management:

The act of getting people together to accomplish desired goals and objectives using available resources efficiently and effectively.

source: wikipedia.org

Performance Management

Performance Management: Definition and Concepts

PERFORMANCE DEFINED

The Merriam-Webster online dictionary defines performance as “the execution of an action; the fulfillment of a claim, promise, or request.” Performance is the doing of something.

The definition continues, “to perform implies action that follows established patterns or procedures or fulfills agreed-upon requirements and often connotes special skill.”

You’ll see several key words and phrases from this definition repeated throughout this course:

- Execution
- Action
- Procedures
- Agreed-upon requirements
- Skills

MANAGEMENT DEFINED

Management is defined as “the act of getting people together to accomplish desired goals and objectives using available resources efficiently and effectively.” (source: wikipedia.org)

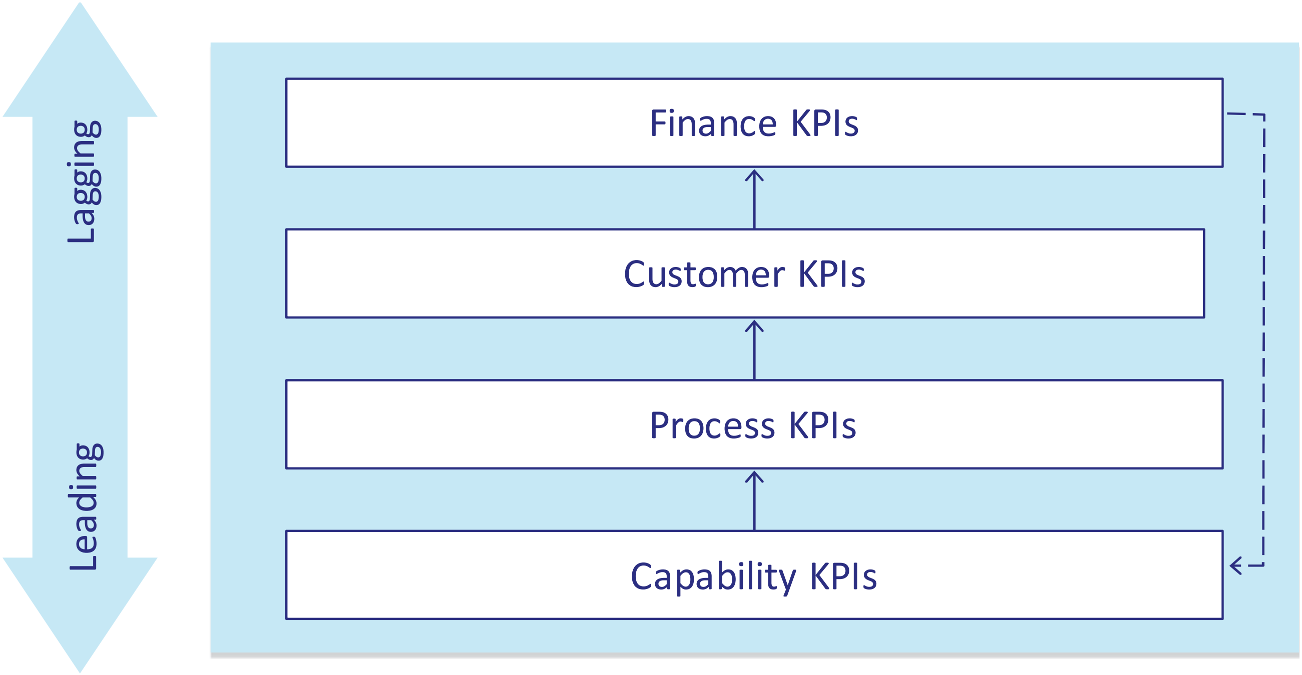
This definition also contains key words and phrases that represent recurring concepts throughout the course:

- Goals and objectives
- Resources
- Efficiency and effectiveness

The reference to people is also an important concept in this definition. People are at the core of performance management.

Performance Management

Key Performance Indicators



Performance Management

Key Performance Indicators

KPI DEFINED

Key performance indicators are "financial and non-financial metrics used to help an organization define and measure progress toward organizational goals." [1] This definition contains several key concepts:

- KPIs include both financial and non-financial metrics.
- KPIs help to define organizational goals.
- KPIs help to measure progress toward goals.

A metric refers to a measurement of business activity. However, in a performance management system, we want to do more than just measure business activity; we want to measure performance aligned with business strategy.

THE BALANCED SCORECARD

KPIs are the cornerstone of the Balanced Scorecard (BSC) introduced by Robert Kaplan and David Norton (Harvard Business School) in 1992. Since that time it has evolved substantially and become the de facto standard for strategic business scorecards.

The BSC approach is founded on these principles:

- *Financial performance* metrics are lagging indicators
- *Customer outcomes* drive financial performance
- *Internal process excellence* drives customer performance
- *Organizational capability* (learning and growth) drives process performance

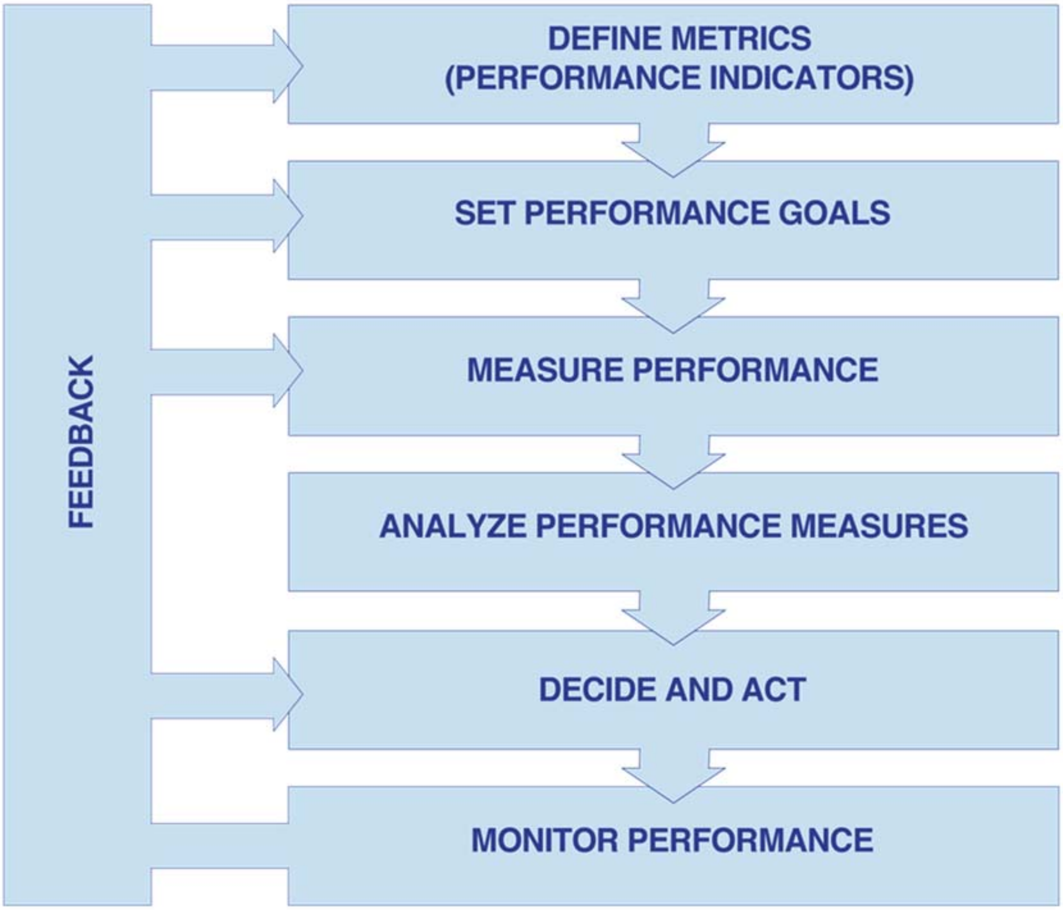
These four categories can be tailored to a business as a whole, to a function or department, or even to an individual employee.

In each category, a set of metrics are selected that align with business goals and correspond with the activities and focus of the target audience. In combination with goals, these metrics can be monitored and presented using dashboards and scorecards, enabling businesspeople to take actions to have positive business impact.

[1] *Business Dashboards: A Visual Catalog for Design and Deployment*, pp. 24, Rasmussen, Chen, and Bansal

Performance Management

Metrics, Measures, and Monitoring



Performance Management

Metrics, Measures, and Monitoring

THREE M'S OF PERFORMANCE MANAGEMENT

Metrics, measurement, and monitoring are the core elements that make performance management work. *Metrics* are the quantitative criteria that are used to manage performance. *Measurement* is the process of collecting metrics. *Monitoring* is the act of tracking measures against metrics-based goals.

METRICS AND GOALS

Performance management begins by selecting and defining the right performance indicators. A balanced performance management program includes both effectiveness and efficiency indicators. Well-defined indicators determine the metrics and measures that are needed.

Goal setting is the next step in the process. Each performance indicator needs to have corresponding goals that express both the level of performance to be achieved and the time frame of the goal.

MEASURING PERFORMANCE

Measurement collects the data needed to calculate actual performance values for each indicator. Measurement includes quantitative data, reference data to associate measures with dimensions, and identity data needed to trace performance indicators to their sources. Measurements are made available to participants through dashboards and scorecards.

MONITORING, ACTION, AND FEEDBACK

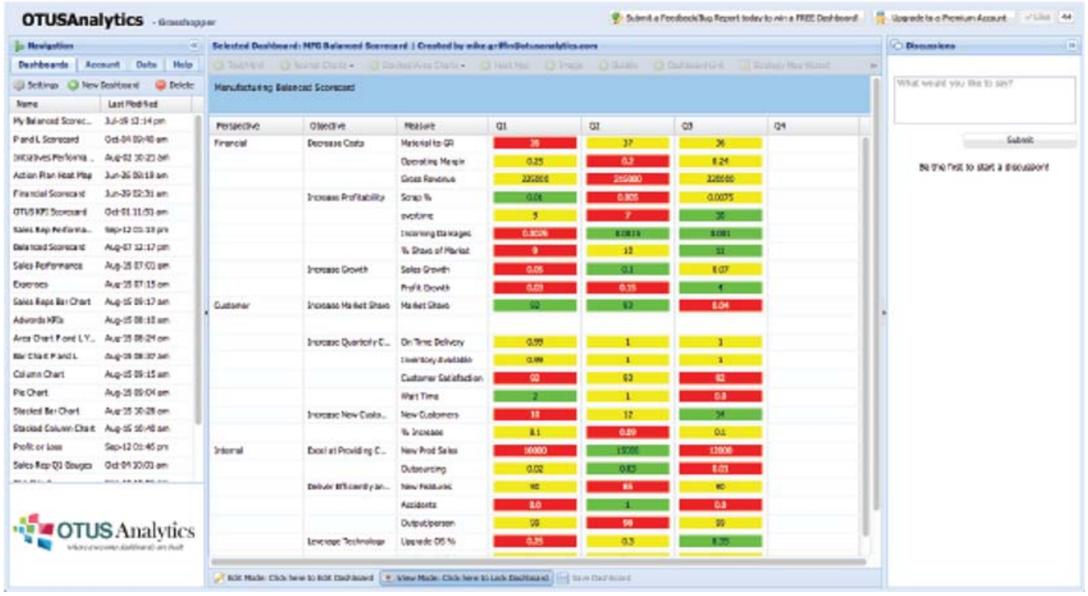
The ultimate value of measurement and analysis is achieved through monitoring and feedback. When performance is measured and monitored feedback loops can be established:

- The impact of decisions and actions is known quickly and quantitatively, not eventually and anecdotally. When the results of decisions and actions are not what is expected, the feedback occurs rapidly enough to reconsider the decision and make changes.
- Feedback to measurement processes and systems provides valuable input needed to calibrate—to improve the precision and accuracy of the measures.

Performance Management

Scorecards and Dashboards

SCORECARD



OTUSAnalytics - Grasshopper
Submit a Feedback/Bug Report today to win a FREE Dashboard!
Upgrade to a Premium Account
Navigation
Dashboards
Account
Data
Help
Settings
New Dashboard
Delete
Discussions
Submit
OTUS Analytics
Edit Mode: Click here to Edit Dashboard
View Mode: Click here to Lock Dashboard
Save Dashboard

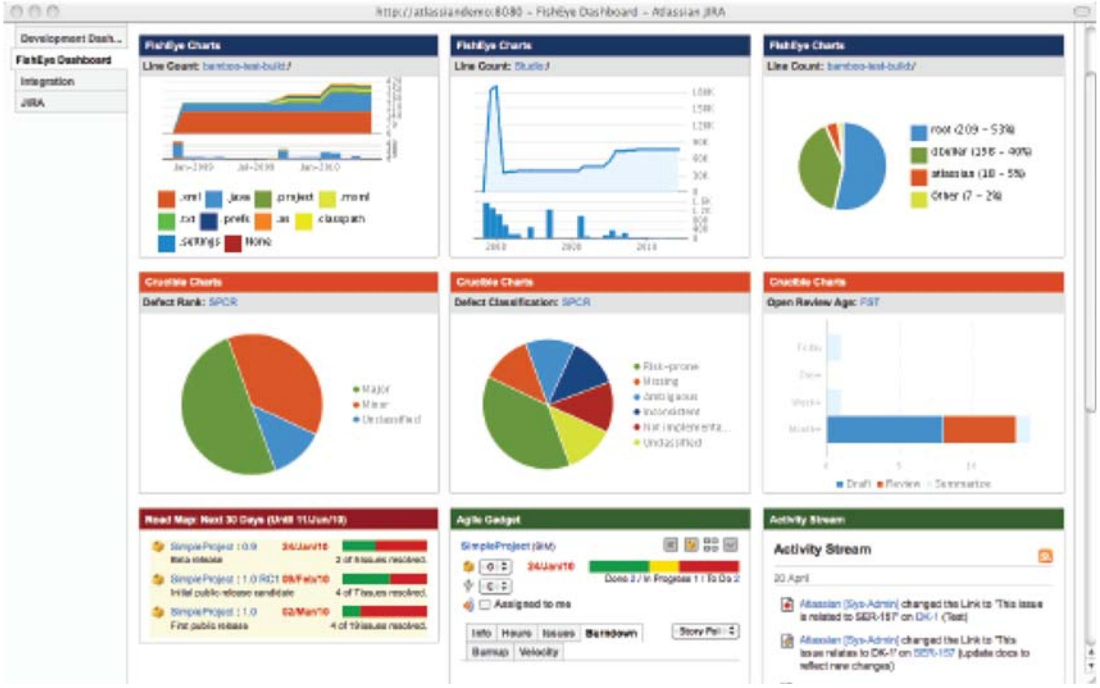
Name	Last Modified
My Balanced Scorec...	Jul-19 12:14 pm
P and L Scorecard	Oct-04 09:40 am
Initiatives Performa...	Aug-02 10:23 am
Action Plan Heat Map	Jun-26 09:18 am
Financial Scorecard	Jun-29 02:31 am
OTUS KPI Scorecard	Oct-01 11:51 am
Sales Rep Performa...	Sep-12 01:18 pm
Balanced Scorecard	Aug-07 12:17 pm
Sales Performance	Aug-25 07:01 am
Expenses	Aug-25 07:15 am
Sales Reps Bar Chart	Aug-16 09:17 am
Adwords KPIs	Aug-15 08:18 am
Area Chart P and L Y...	Aug-25 08:24 am
Bar Chart P and L	Aug-15 08:37 am
Column Chart	Aug-15 09:15 am
Pie Chart	Aug-25 09:04 am
Stacked Bar Chart	Aug-25 10:28 am
Stacked Column Chart	Aug-16 10:48 am
Profit or Loss	Sep-12 01:45 pm
Sales Rep Q1 Gauges	Oct-04 10:01 am

Selected Dashboard: MFG Balanced Scorecard | Created by mike.griffin@otusanalytics.com

Manufacturing Balanced Scorecard

Perspective	Objective	Measure	Q1	Q2	Q3	Q4
Financial	Decrease Costs	Material to GR	35	37	36	
		Operating Margin	0.25	0.2	0.24	
		Gross Revenue	225000	215000	226000	
	Increase Profitability	Scrap %	0.01	0.005	0.0075	
		overtime	5	7	20	
		Incoming Damaged	0.0025	0.0018	0.001	
		% Share of Market	9	12	11	
	Increase Growth	Sales Growth	0.05	0.1	0.07	
		Profit Growth	0.03	0.15	4	
Customer	Increase Market Share	Market Share	52	53	0.04	
	Increase Quarterly C...	On Time Delivery	0.99	1	1	
		Inventory Available	0.99	1	1	
		Customer Satisfaction	80	83	82	
		Wait Time	2	1	0.8	
	Increase New Custo...	New Customers	10	12	24	
		% Increase	8.1	0.09	0.1	
Internal	Excel at Providing C...	New Prod Sales	10000	15000	12000	
		Outsourcing	0.02	0.03	0.01	
	Deliver Efficiently an...	New Features	90	85	80	
		Accidents	3.0	1	0.8	
		Output/person	99	98	99	
	Leverage Technology	Upgrade OS %	0.25	0.3	0.35	

DASHBOARD



SimpleProject : 0.9 24/Jan/10 Beta release 2 of 8 issues resolved.

SimpleProject : 1.0 RC1 08/Feb/10 Initial public release candidate 4 of 7 issues resolved.

SimpleProject : 1.0 02/Mar/10 First public release 4 of 19 issues resolved.

SimpleProject (SIM) 24/Jan/10 Done 2 / In Progress 1 / To Do 2

□ Assigned to me

Activity Stream

20 April

Atlassian [Sys-Admin] changed the Link to 'This issue is related to SER-157' on DK-1 (Test)

Atlassian [Sys-Admin] changed the Link to 'This issue relates to DK-1' on SER-157 (update docs to reflect new changes)

Performance Management

Scorecards and Dashboards

PERFORMANCE SCORECARDS

A performance scorecard is a visual representation of progress of a business or business unit over time, illustrating achievement as compared to goals or targets. The core concepts of scorecards are targets and KPIs. Each KPI is typically shown with metrics information showing the actual value for the KPI, the target value, variance, and directional trend. Scorecards often combine tabular presentation of data—rows for each KPI and columns for each attribute—with graphical presentation such as sparklines representing directional trend for the KPI.

PERFORMANCE DASHBOARDS

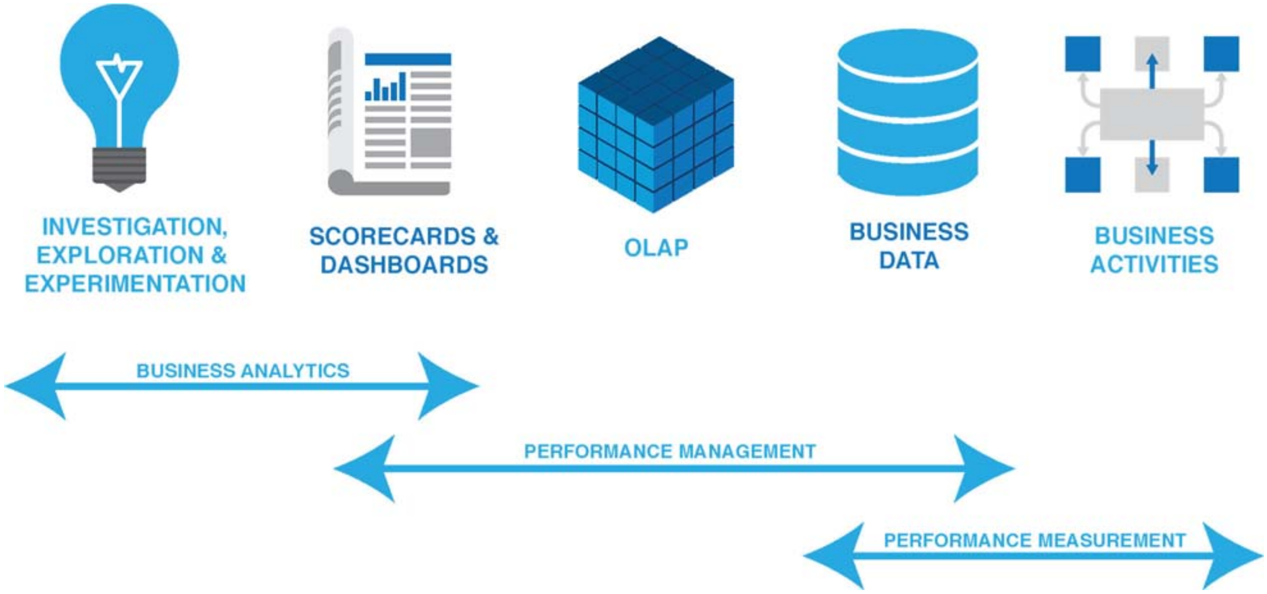
When describing business dashboards, many people rely on the metaphor of an automobile dashboard. The dashboard is necessary to know the state of the automobile in real time. A dashboard puts all the vital information of an operation—whether driving a car or driving a business—on a single screen that can be viewed at a glance. Where the scorecard shows progress over time with comparative context, the dashboard shows the current state in real time. Where scorecards rely largely on tabular presentation, dashboards are almost entirely graphical.

WORKING TOGETHER

Effective performance management applies dashboards and scorecards working together. KPIs on scorecards typically roll up to higher-level KPIs on dashboards. A dashboard, for example, may visually present enterprise KPIs that aggregate several business unit indicators.

Business Analytics

Continuum



Business Analytics

Continuum

BUSINESS ANALYTICS VS. PERFORMANCE MANAGEMENT

A 2009 survey found that more than 60 percent of respondents associated business analytics with query and analysis, reporting, and dashboard tools. (*Computerworld,* February 2009) Dashboards and scorecards are simply ways to visualize and present business measures and metrics. They may be used to deliver measures and metrics that are derived through analytics processes, but they are more commonly found in performance management applications. These technologies are better suited to the predictable and scheduled nature of performance management than to the on-demand and volatile nature of business analytics.

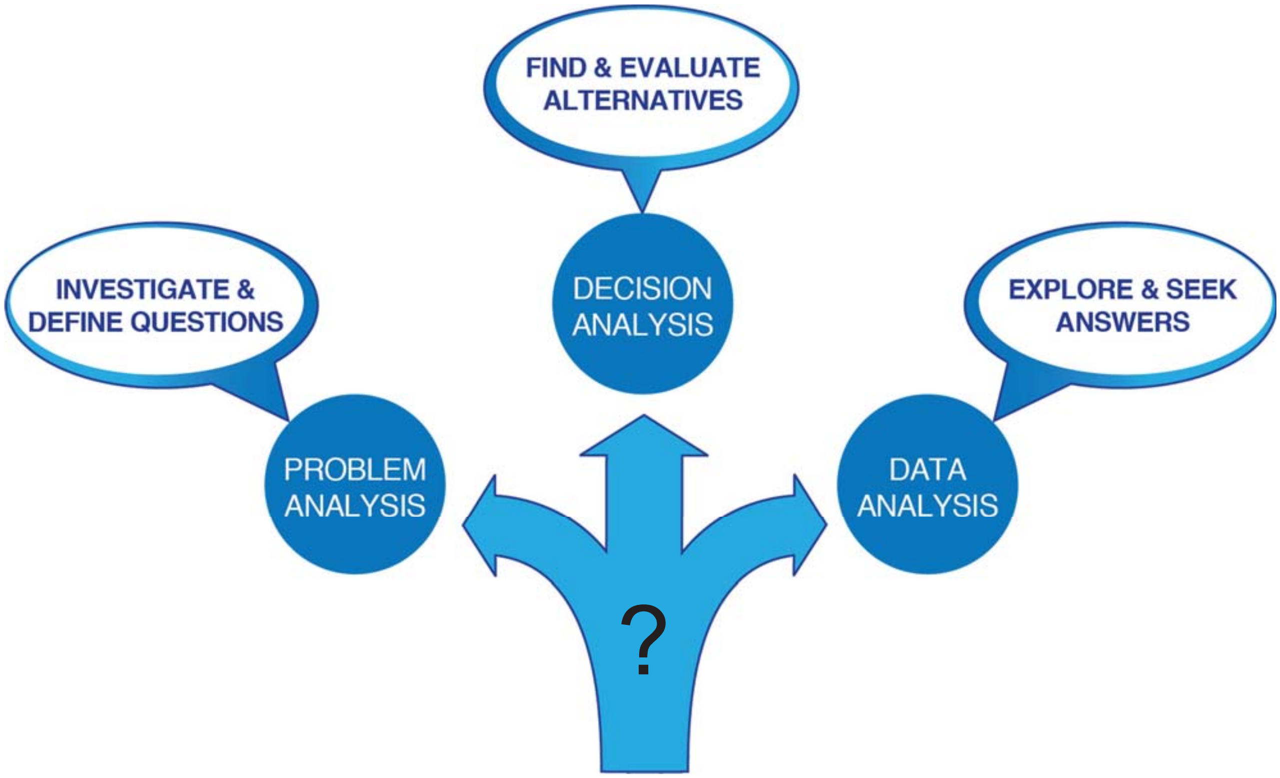
OLAP offers a bit more of the adaptability that analytics demand. Cubes may be good data sources and offer opportunities for data exploration, but OLAP has only a limited place in the business analytics toolset.

THE NATURE OF ANALYTICS

Business analytics needs are more dynamic than the common metrics applications of performance management. Forward-looking analysis raises new questions as quickly as it answers current questions. The real value is derived not from the data, but from the conversations and ideas driven by looking at data. Wikipedia’s definition captures some of the essence with the phrase “continuous iterative exploration.” Good analytics processes are characterized largely by a sense of searching—of continuously seeking answers, understanding, insight, and foresight.

Business Analytics

Analysis Types



Business Analytics

Analysis Types

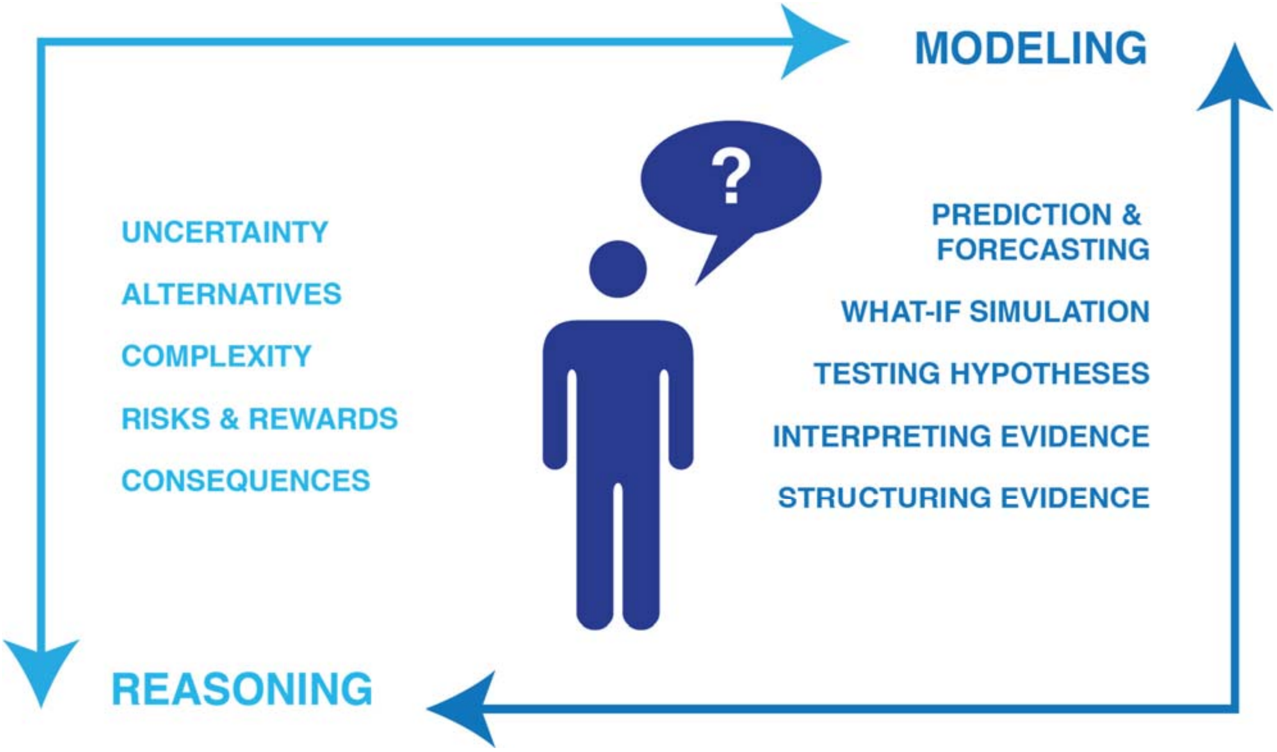
THREE KINDS OF ANALYSIS

The purpose of business intelligence and analytics is to support business analysis processes, which typically fall into three categories:

- *Problem analysis* is used to understand a poorly defined or ill-structured problem. The purpose of problem analysis is to add definition and structure—to increase understanding of a problem and identify questions that must be answered to resolve the problem. Deeper understanding of the problem is rarely found in data, so don't start problem analysis with the data. Once the problem is well defined, then data analysis and solution seeking can begin.
- *Decision analysis* is used to identify and evaluate alternative courses of action. Decision analysis uses data to develop and confirm understanding of cause and effect.
- *Data analysis* is the activity of exploring the data to seek answers for a well-defined and well-structured problem. Data analysis is a natural step in the process of decision analysis and in the solution-seeking activities of problem analysis.

Business Analytics

Analytic Modeling



MODELING

PREDICTION & FORECASTING

WHAT-IF SIMULATION

TESTING HYPOTHESES

INTERPRETING EVIDENCE

STRUCTURING EVIDENCE

REASONING

UNCERTAINTY

ALTERNATIVES

COMPLEXITY

RISKS & REWARDS

CONSEQUENCES

Business Analytics

Analytic Modeling

MODELING OUTCOMES

Every analytics model has specific purposes that are described by the outcomes to be produced by the model. These purposes describe what the model is intended to do—a sort of internal view of modeling purpose. Modeling outcomes include:

- Structuring evidence to organize the facts and reduce complexity.
- Interpreting evidence to find meaning hidden in the facts.
- Testing hypotheses to seek confirming or contrary evidence.
- Prediction to forecast future events or conditions.
- Simulation of what-if scenarios by manipulating variables and observing the results.

MODELING APPLICATION

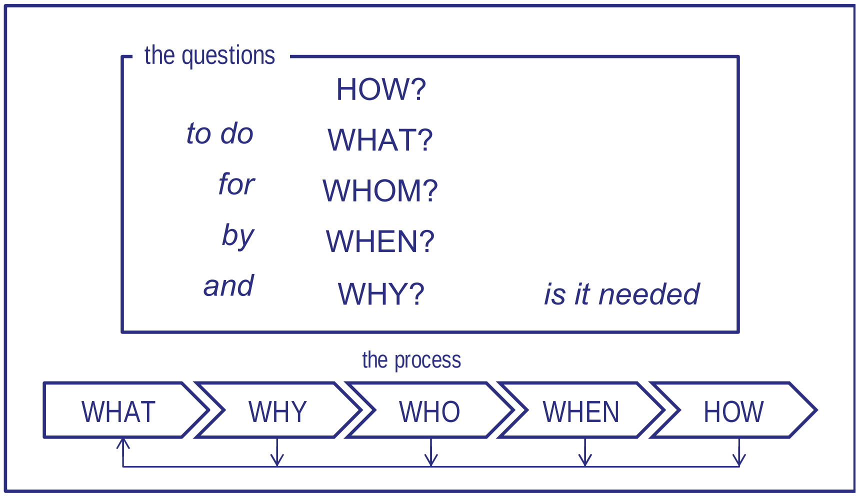
The real value of modeling is derived from application of the models. The business use of a model—the external view of purpose—is to support human reasoning and decision-making processes. From this perspective the purpose of modeling includes:

- Reducing uncertainty in decision processes.
- Identifying and evaluating alternative decisions or actions.
- Reducing complex situations to improve understandability.
- Evaluating risks and rewards of various alternatives.
- Identifying potential consequences and side effects of decisions and actions.

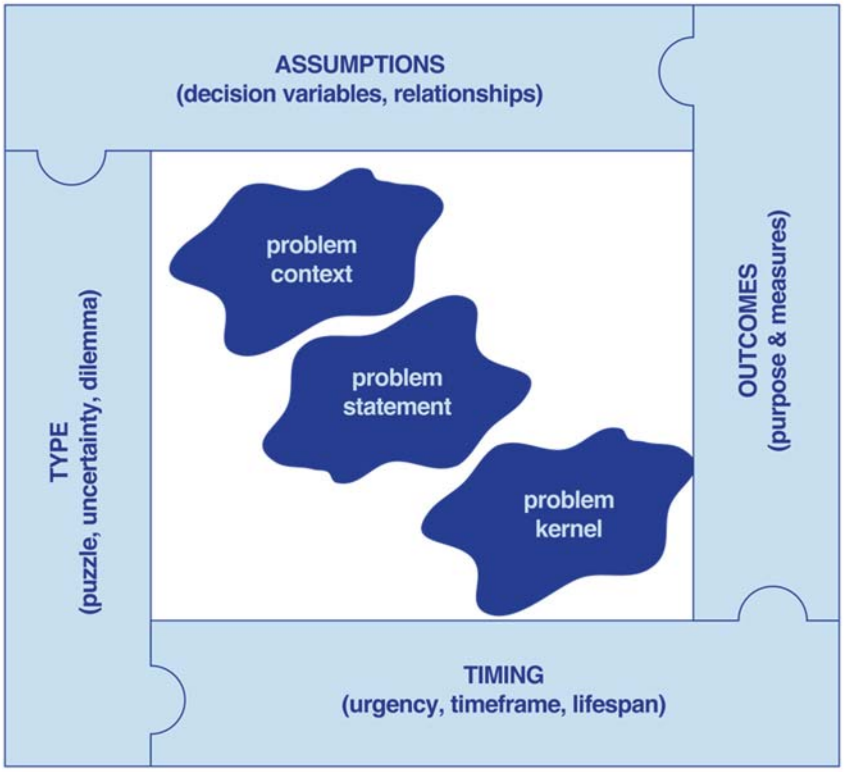
Business Analytics

Framing Models

Questioning Model for Problem Framing



Kernel Seeking Model for Problem Framing


ASSUMPTIONS
(decision variables, relationships)
TYPE
(puzzle, uncertainty, dilemma)
problem context
problem statement
problem kernel
OUTCOMES
(purpose & measures)
TIMING
(urgency, timeframe, lifespan)

Business Analytics

Framing Models

PROBLEM FRAMING

Most analytics problems begin with some uncertainty about the nature of the problem. Framing is the means to remove vagueness by adding specifics to our understanding of the problem. Problem framing is an important first step to avoid "blind alleys" and unnecessary costs by stating a clear and well-defined purpose for the data analysis.

FROM VAGUE TO SPECIFIC

The facing page illustrates two kinds of framing models—questioning and kernel seeking.

The questioning model asks

How to do *what* by *when* for *whom*, and *why* is it needed?

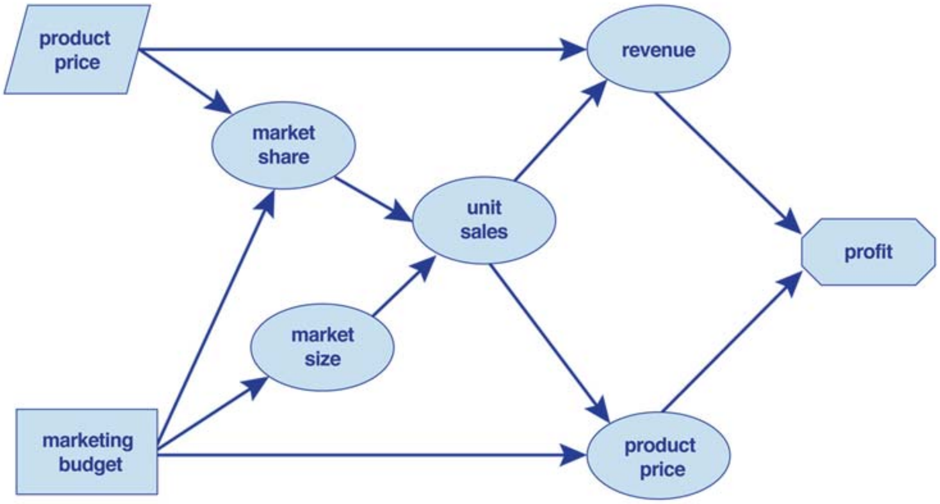
The sequence in which questions are addressed, however, is likely to vary from the sequence expressed above. It is common to ask: What needs to be done? Why is it needed? Who needs it? By when? How to make it happen?

The kernel seeking model works from context, through problem statement, to express the essence of the problem as a single simple sentence.

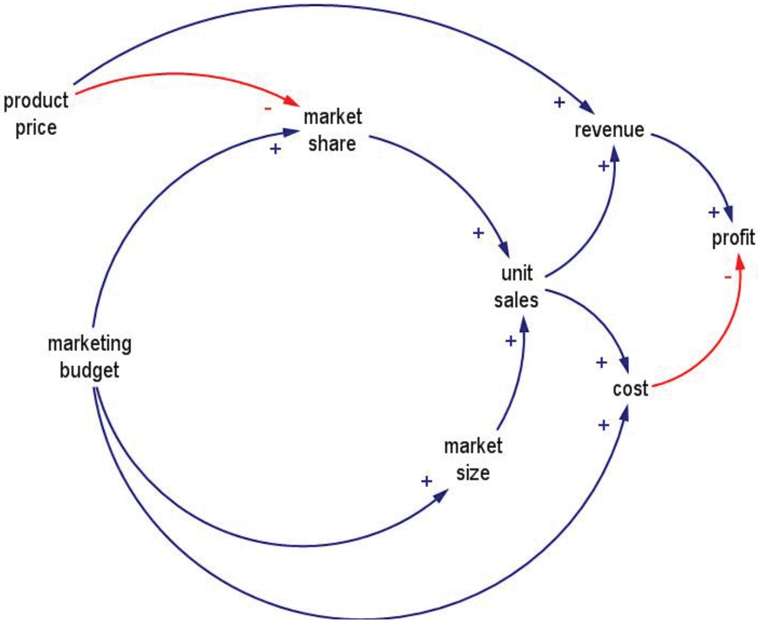
Business Analytics

Causal Models

Influence Diagram for Problem Analysis



Causal Loop Diagram for Problem Analysis



Business Analytics

Causal Models

PROBLEM MODELING

Problem modeling is an analysis activity that is used to understand a poorly defined or ill-structured problem—to fully grasp the business needs. The purpose of problem analysis is to add definition and structure—to increase understanding of a problem and identify questions that must be answered to resolve the problem. Deeper understanding of the problem is rarely found in data, so don't start problem analysis with the data. Once the problem is well defined, then data analysis and solution seeking can begin.

UNDERSTANDING SYSTEM DYNAMICS

Not surprisingly, problem analysis is strongly connected with cause and effect. Causal analytics applies a systematic approach to uncover the root cause(s) for a problem or opportunity.

There are several common approaches to causal modeling. Examples include:

- Causal loop diagrams
- Influence models
- Fishbone diagrams
- Pareto analysis

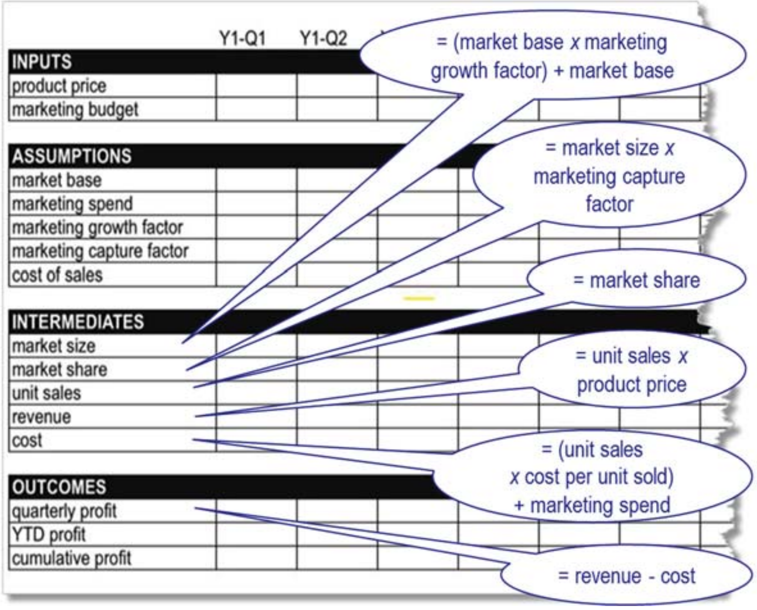
A common aspect of all of these techniques is that they go beyond the symptoms (the visible problem) and hunt for the actions or conditions that were major contributors to the problem.

The facing page shows two types of models that are used to understand the dynamics—cause and effect relationships—inherent in a system under analysis.

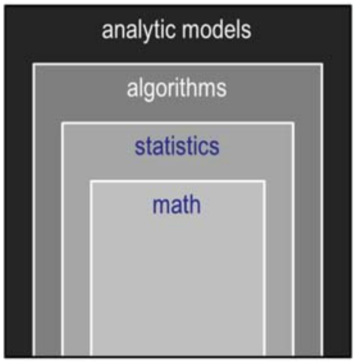
Data Analytics

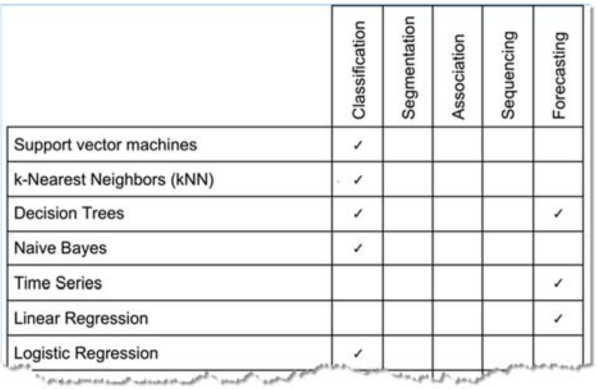
Solution Models

Formula-Based Models



Algorithm-Based Models





	Classification	Segmentation	Association	Sequencing	Forecasting
Support vector machines	✓				
k-Nearest Neighbors (kNN)	✓				
Decision Trees	✓				✓
Naive Bayes	✓				
Time Series					✓
Linear Regression					✓
Logistic Regression	✓				

Naïve Bayes – Conditional Probability with Normal Distribution

$$\hat{P}(X_j \mid C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2} \right)$$

μ_{ji} : mean (average) of feature values X_j of examples for which $C = c_i$

σ_{ji} : standard deviation of feature values X_j of examples for which $C = c_i$

Data Analytics

Solution Models

UNDERSTANDING THE CHOICES

Solution models deliver information to support informed decision making, answering what, why, and what-if questions to guide the choices that lead to positive business outcomes. Solution modeling is the activity that most people think of when they hear the term “analytic modeling.”

ANALYZING DATA

Solution modeling applies mathematics, statistics, and algorithms to data and communicates findings as data visualizations. Two types of solution modeling are common: formula-based modeling and algorithm-based modeling.

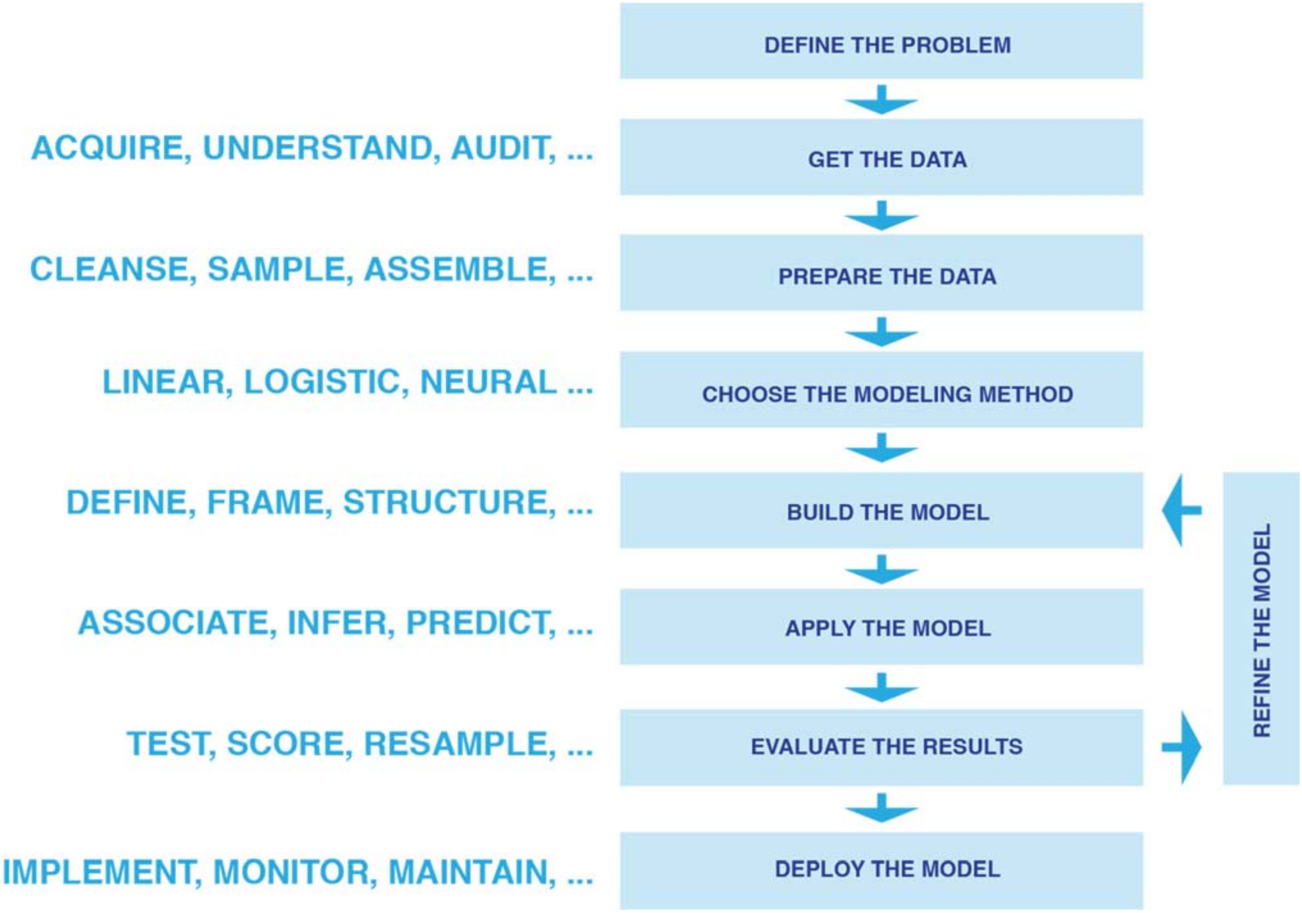
- *Formula-based models* can be built using basic tools such as Excel. These models typically define input variables and output measures that are needed to perform an analytics activity such as simulation. A sequence of formulas builds the path from input to output. Formula-based models can also be built using statistical analysis software.

- *Algorithm-based models* are built using statistical tools and data mining software. A few examples of the many tools include SAS, RapidMiner, Alteryx, and Predixion. These models apply algorithms to find patterns in data, performing activities such as data mining and prediction. Many algorithms are available to perform tasks such as classification, clustering, association, sequencing, and forecasting.

In the age of data science, much attention is given to algorithmic models and the continuing development of new algorithms for advanced analytics and machine learning. Algorithmic models are the heart of analytics and are fundamental for data mining and predictive analytics. Many basic analysis problems, however, don’t need the complexities of algorithmic modeling. Much useful analysis can be done with formula-based models built in spreadsheets and similar tools.

Data Analytics

Modeling Process


DEFINE THE PROBLEM
ACQUIRE, UNDERSTAND, AUDIT, ...
GET THE DATA
CLEANSE, SAMPLE, ASSEMBLE, ...
PREPARE THE DATA
LINEAR, LOGISTIC, NEURAL ...
CHOOSE THE MODELING METHOD
DEFINE, FRAME, STRUCTURE, ...
BUILD THE MODEL
ASSOCIATE, INFER, PREDICT, ...
APPLY THE MODEL
TEST, SCORE, RESAMPLE, ...
EVALUATE THE RESULTS
IMPLEMENT, MONITOR, MAINTAIN, ...
DEPLOY THE MODEL
REFINE THE MODEL

Data Analytics

Modeling Process

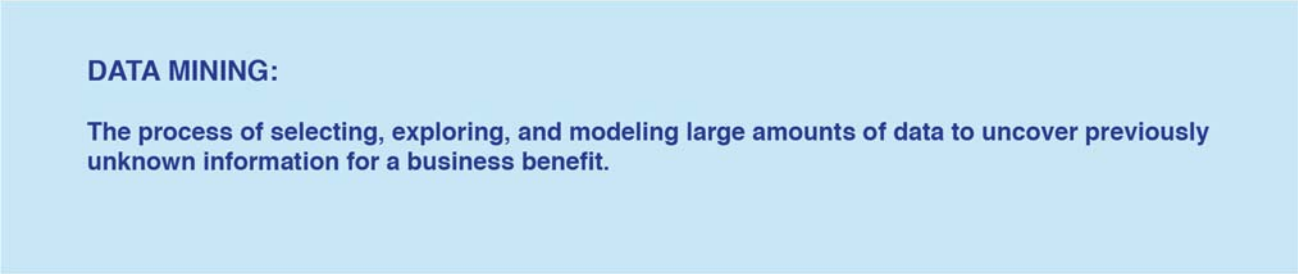
DEVELOPING AND USING ANALYTICS MODELS

The range of skills needed to develop and apply analytics models is quite extensive. The diagram on the facing page illustrates a typical modeling process and identifies some of the skill areas at each step of the process:

- Define the problem
  - Knowledge of how the business works
  - Problem framing—who, what, and where
  - Influence diagramming
- Get the data
  - Data acquisition
  - Data profiling
  - Data quality assessment
- Prepare the data
  - Variables determination
  - Data cleansing and quality improvement
  - Data sampling—bias, statistical significance, etc.
  - Data structures, migration, and manipulation
- Choose the modeling technique
  - Knowledge of statistical methods
  - Application of various methods
- Build the model
  - Model purpose definition
  - Model framing—variables, functions, constraints, etc.
  - Model structure—spreadsheets, visualizations, outputs, etc.
- Apply the model
  - Association and inference
  - Prediction, forecasting, simulation
  - Hypothesis affirmation or contradiction
- Evaluate the model
  - Model testing—reality and test cases
  - Model scoring—precision, accuracy, reliability
- Refine the model
  - Functional refinement
  - Calibration and confidence
- Deploy the model
  - Production implementation for recurring execution needs
  - Change monitoring and model maintenance

Data Analytics

Data Mining



DATA MINING:

The process of selecting, exploring, and modeling large amounts of data to uncover previously unknown information for a business benefit.

Data Analytics

Data Mining

"LISTENING" TO THE DATA

Data mining is the process of selecting, exploring, and modeling large amounts of data to uncover previously unknown information for business benefit. The Gartner Group further defines it as "the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques."

Note that the word *automatic* doesn't appear in either definition. Most leading data mining vendors do not advocate a black-box approach to exploratory data mining. Successful data mining is a human skill that is assisted by tools and technology. Some suggest that data mining is simply the new term for statistics, and indeed data mining borrows heavily from statistics.

Data mining is complemented by a variety of preprocessing functions, statistics, and data visualization tools. As the technology has become easier to use, it has moved from an esoteric application for a select few to a broader user community with a variety of data mining needs.

Data Analytics

Geospatial Analytics

Data Analytics

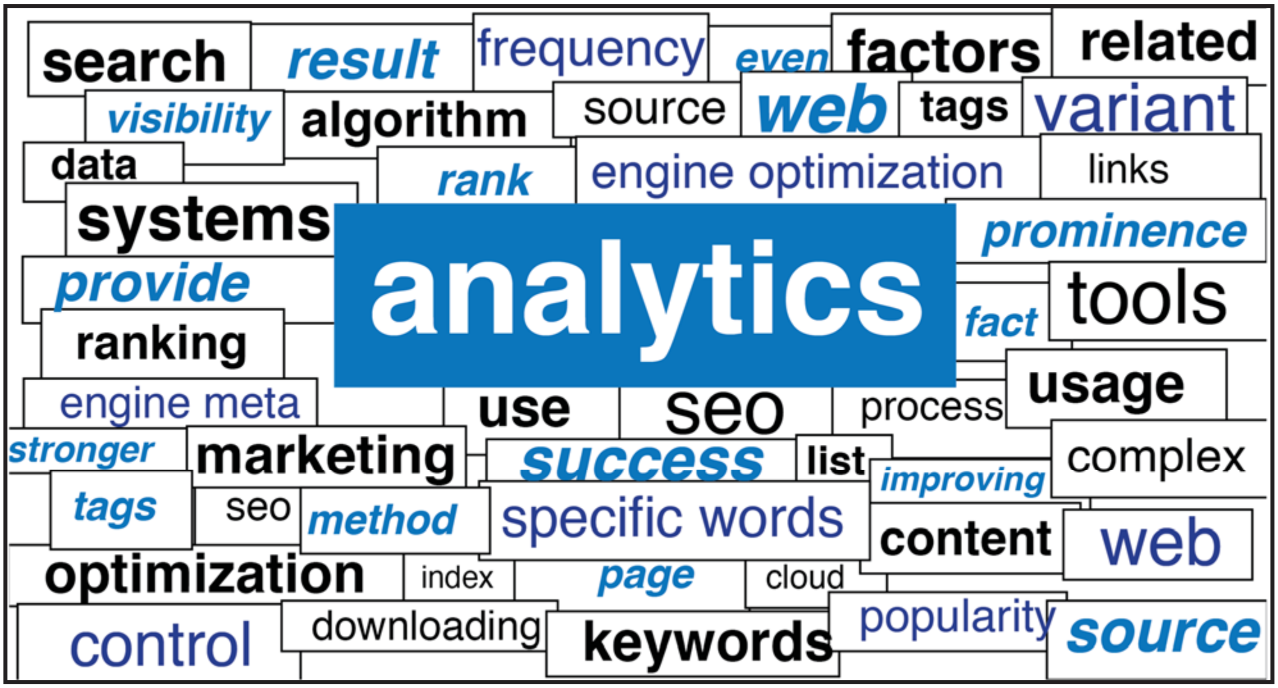
Geospatial Analytics

INSIGHT THROUGH LOCATION

Geospatial analytics is the technique of applying statistical analysis and analytic modeling to data that has geographical or geospatial attributes or dimensions. Geospatial analytics enables location intelligence by mapping other business data with location-identifying data. Spatial mapping makes it possible to transform large amounts of data into color-coded visual representations where location-based trends, patterns, and anomalies can be easily viewed and understood.

Data Analytics

Text Analytics



Data Analytics

Text Analytics

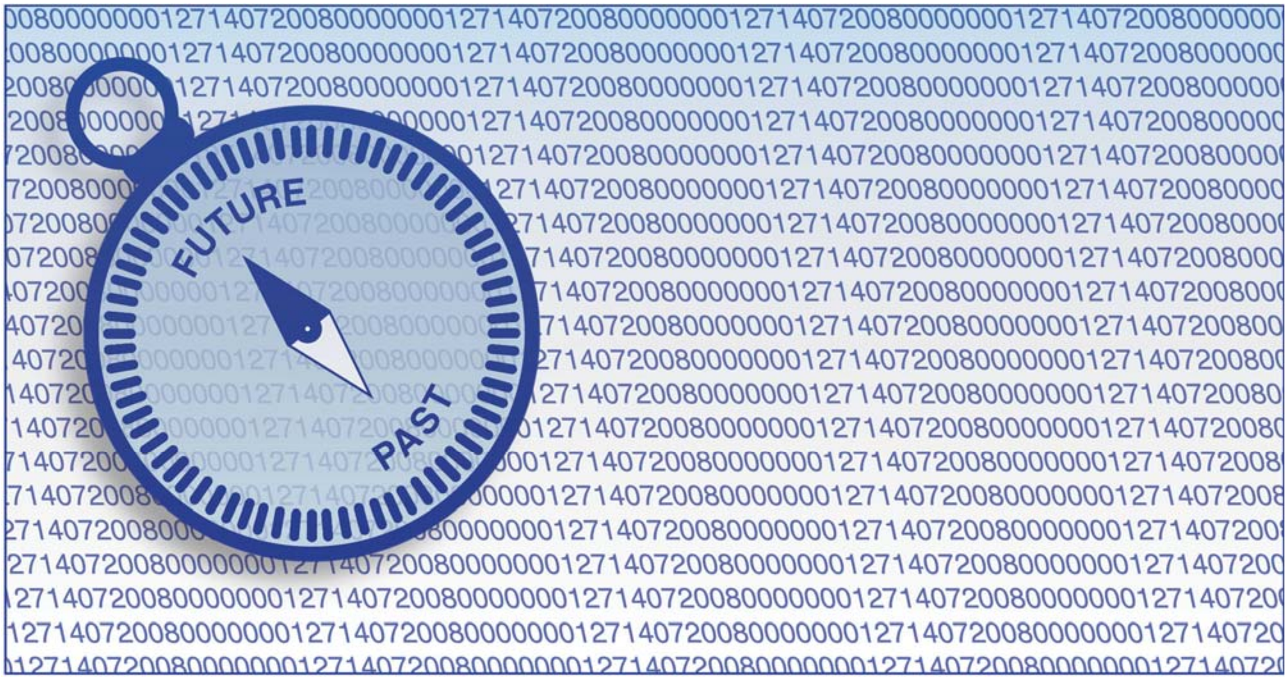
FINDING INSIGHT IN UNSTRUCTURED TEXT

Text analysis is a process of extracting information and developing insight from textual data. The extraction process is sometimes called text mining. Text mining involves syntactical analysis to study text content and to parse text into phrases and fragments for further analysis. Insight development involves lexical analysis—applying a domain-specific lexicon to infer meaning of phrases and fragments. Analysis of word frequency distributions, word patterns, and word associations are all involved when turning text into data for analysis.

One of the most common applications of text analytics in BI is customer sentiment analysis. Examination of text found in customer reviews and consumer comments from social media, email, and similar sources produces insight into the feelings and beliefs of consumers related to specific products, services, or brands. When sentiment data is analyzed in conjunction with other structured data—timing, geography, pricing, sales channels, consumer demographics, etc.—it becomes possible to infer cause-and-effect relationships. With causal understanding businesses can take action to gain favorable sentiment and minimize unfavorable impressions.

Data Analytics

Forecasting and Prediction



Data Analytics

Forecasting and Prediction

SEEING INTO THE FUTURE

Both forecasting and prediction estimate future probabilities based on predictor values, but the two techniques are different.

FORECASTING

Forecasts provide probabilities of future outcomes based on past and present data. Common applications include the analysis of trends. Sales trends, for example, are used along with other data to predict what the sales will be for future periods.

PREDICTING

Predictions provide probabilities on behavior that will result from potential actions. For example, a company may try to understand what a group of customers with certain characteristics may do in response to a sales campaign or other company action.

Data Analytics

Simulation and Optimization

Data Analytics

Simulation and Optimization

SIMULATION

Simulation techniques enable the analyst to manipulate decision variables (input) to explore their impact on performance variables (output). Examples of simulation techniques include:

- *Discrete Event Simulation* to analyze service levels in restaurants, traffic intersections, etc.

- *Continuous System Simulation* to analyze physical equipment and processes

- *Monte Carlo Simulation* for analyzing risk and uncertainty

Key considerations include valid information sources, selection of the key characteristics and behaviors to be included, appropriate approximations and assumptions, and validation of the outcomes.

OPTIMIZATION

Optimization techniques enable the analyst to frame a problem in a manner that determines the “best” settings for decision variables to optimize the value of an output variable. Examples of optimization techniques include:

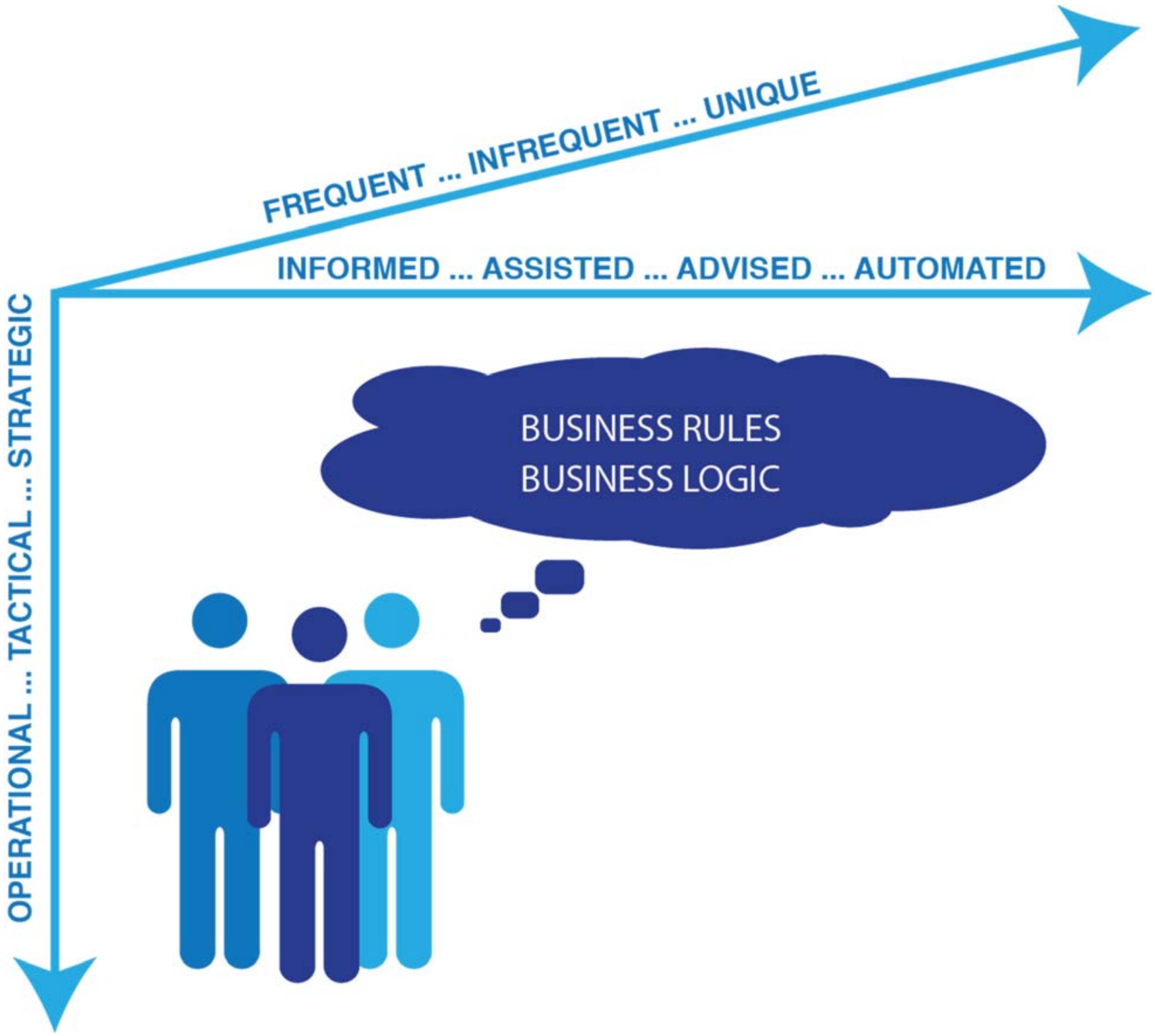
- *Linear Programming* calculates optimum decision variables to maximize or minimize an outcome variable. The objective function and all constraints are assumed to be linear.

- *Integer Programming* is an alternative to linear programming when the decision variables are constrained to be integers. Similar to linear programming, the objective function and the constraints are also linear.

- *Evolutionary Programming*, also known as genetic algorithms, is used to provide a search algorithm that determines the optimum setting of decision variables. The approach is based on principles from the theory of evolution. This technique can support nonlinear problems. It is more complex than being constrained to use linear assumptions.

Data Analytics

Decision Management


FREQUENT ... INFREQUENT ... UNIQUE
INFORMED ... ASSISTED ... ADVISED ... AUTOMATED
OPERATIONAL ... TACTICAL ... STRATEGIC
BUSINESS RULES
BUSINESS LOGIC

Data Analytics

Decision Management

ACTION FROM ANALYTICS

Decision management systems are among the most advanced forms of business analytics—putting analytics to work because value is created through action, not through information alone.

DECISION TYPES

Business decisions may be strategic, tactical, or operational. Some decisions are frequent and recurring, while others are infrequent and occasional or unique. Operational decisions tend to be more repeatable—frequent and recurring—than strategic and tactical decisions.

DECISION-MAKING TYPES

Decision-making processes rely on BI in a variety of ways. They may be:

- Informed—BI provides relevant information to decision makers.
- Assisted—BI provides relevant information to and simulates outcomes of alternatives for decision makers.
- Advised—BI provides relevant information, predictive analysis, and recommendations to decision makers.
- Automated—A decision management system (a component of BI) is the decision maker.

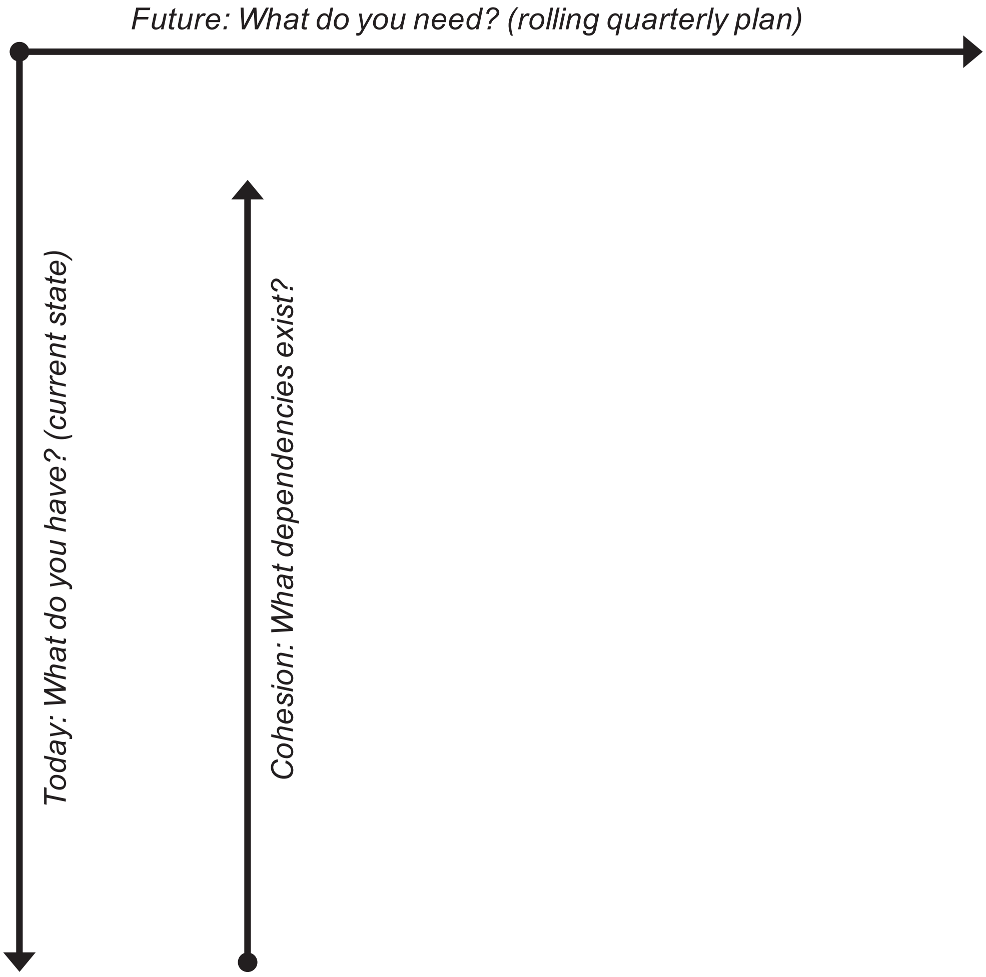
DECISION MANAGEMENT SYSTEMS

Decision management systems target recurring operational decisions that are specific to a customer or a transaction. Analysis and modeling of a group of similar decisions identifies the logic, business rules, and data points that shape the decisions. Decision services are implemented to encapsulate logic and business rules, and to access and apply data to the decision processes. Logic, rules, and data are applied to determine which of the decision-making types is applied for each individual decision.

According to James Taylor, CEO of Decision Management Solutions, “Decision Management Systems are agile, analytic, and adaptive. They are agile so they can be rapidly changed to cope with new regulations or business conditions. They are analytic, putting an organization’s data to work improving the quality and effectiveness of decisions. They are adaptive, learning from what works and what does not work to continuously improve over time.” (Decision Management Systems Platform Technologies Report Version 2, Update 4, January 4, 2013 by James Taylor, © 2012 Decision Management Solutions)

The BI and Analytics Roadmap

Capabilities, Metrics, and Analytics


Future: What do you need? (rolling quarterly plan)
Today: What do you have? (current state)
Cohesion: What dependencies exist?

The BI and Analytics Roadmap

Capabilities, Metrics, and Analytics

MAPPING CAPABILITIES AND SERVICES

The first section of the BI and analytics roadmap relates to program organization topics covered in the previous module, as well as business capabilities discussed in this module.

- Program organization
  - Program management and planning
  - Organizational structure(s) for BI and analytics responsibilities
- Business capabilities
  - Descriptive
  - Diagnostic
  - Discovery
  - Predictive
  - Prescriptive

The second section illustrates performance management and analytics services. (Additional services will be covered in the next module.)

- Performance management
  - Business KPIs
  - Goal planning
  - Dashboards and scorecards
- Analytics
  - Business analytics (framing models, causal models)
  - Data analytics (formula-based models, algorithmic models)

PROJECTS

Once you determine where you need to be, use that as the basis for an action plan for getting there. Map these into your roadmap, being sure to identify dependencies with other systems and planned activities. Projects may include:

- Develop new systems
- Extend / enhance
- Maintain / modify
- Decommission / retire

Be specific. For example, list *predictive capability for risk management via policy risk model*, rather than *predictive analytics*. Consider priorities and dependencies as you plot each project on the timeline.

Mistakes to Avoid

In Predictive Analytics Efforts

- FAILURE TO FRAME THE BUSINESS PROBLEM
- FAILURE TO CHANGE FROM A REACTIVE TO PROACTIVE MINDSET
- FAILURE TO CONTROL OUTPUT
- FAILURE TO GET TRAINING
- FAILURE TO UPDATE AND MANAGE MODELS
- FAILURE TO TAKE ACTION BASED ON RESULTS
- FAILURE TO THINK BEYOND STRUCTURED DATA
- FAILURE TO PREPARE THE DATA
- FAILURE TO SECURE EXECUTIVE BUY-IN EARLY IN YOUR PROJECT
- FAILURE TO CONSIDER CULTURE AND PROCESS ISSUES

Source: *Ten Mistakes to Avoid in Predictive Analytics Efforts* by Fern Halper. © TDWI

Mistakes to Avoid

In Predictive Analytics Efforts

TEN MISTAKES TO AVOID

Ten Mistakes to Avoid in Predictive Analytics Efforts by Fern Halper.

Predictive analytics—a statistical or data mining solution consisting of algorithms and techniques used on both structured and unstructured data to determine outcomes—is becoming a mainstream analytics technology, and organizations are realizing its competitive value.

Business intelligence is typically reactive and can't estimate targets (called outcomes) of interest, such as: Who will respond to a promotion? Who will drop my service? When will a piece of equipment fail? Predictive analytics, however, is proactive.

Predictive analytics is one of the most popular kinds of advanced analytics. Although many companies are excited about the possibility of utilizing predictive analytics, there are a number of interrelated themes about what not to do when it comes to predictive analytics projects.

Mistakes to avoid are shown on the facing page.

Mistakes to Avoid

When Deploying BI and Analytics

- FOCUSING ON TECHNOLOGY RATHER THAN IMPROVING THE TOTAL USER EXPERIENCE
- NEGLECTING TO TAKE ADVANTAGE OF THE DATA STORYTELLING POTENTIAL OF VISUAL ANALYTICS
- NEGLECTING TO BUILD ON TECHNOLOGY’S POPULARITY TO CREATE AN ANALYTICS CULTURE
- NOT PRIORITIZING DATA GOVERNANCE AND IMPROVEMENT IN BUSINESS-IT COLLABORATION
- FAILING TO EVALUATE THE POTENTIAL OF ENABLING DATA INTERACTION WITH HADOOP SYSTEMS
- IGNORING SELF-SERVICE DATA PREPARATION TECHNOLOGIES
- FAILING TO INTEGRATE VISUAL BI AND ANALYTICS WITH APPLICATIONS AND DECISION MANAGEMENT
- NEGLECTING TO APPLY VISUAL ANALYTICS TO PERFORMANCE MANAGEMENT
- TURNING A BLIND EYE TO THE POTENTIAL OF IN-MEMORY AND CLOUD COMPUTING

Source: *Ten Mistakes to Avoid When Deploying Visual BI and Analytics for Business Decisions* by David Stodder. © TDWI

Mistakes to Avoid

When Deploying BI and Analytics

TEN MISTAKES TO AVOID

Ten Mistakes to Avoid When Deploying Visual BI and Analytics for Business Decisions by David Stodder.

Few things are hotter today than visual business intelligence (BI) and analytics tools and applications. Across organizations, business-side executives, managers, and other personnel want to strengthen the role of data and analytics in their daily decision making. Data visualization, data discovery, and analytics technologies are rapidly maturing and are becoming easier to use so that these “nontechnical” users can do more in a self-service fashion.

However, success with these technologies is not as simple as just turning them on. What should be a promising new chapter in users’ interaction with data could become a big disappointment if some key considerations are not addressed. The mistakes on the facing page identify important factors in the success or failure of visual BI and analytics deployment and recommend practices for addressing them.

Discussion

Business Metrics and Analytics

LET'S TALK ABOUT IT!

- Does your business monitor performance through key performance indicators?
- Is there a goal-setting process?
- What business areas are currently leveraging analytics capabilities?
- What forms of business analytics are currently applied? Problem framing? Problem modeling? Causal modeling?
- What forms of data analytics are currently in place? Formula-based models? Algorithm-based models?

Discussion

Business Metrics and Analytics

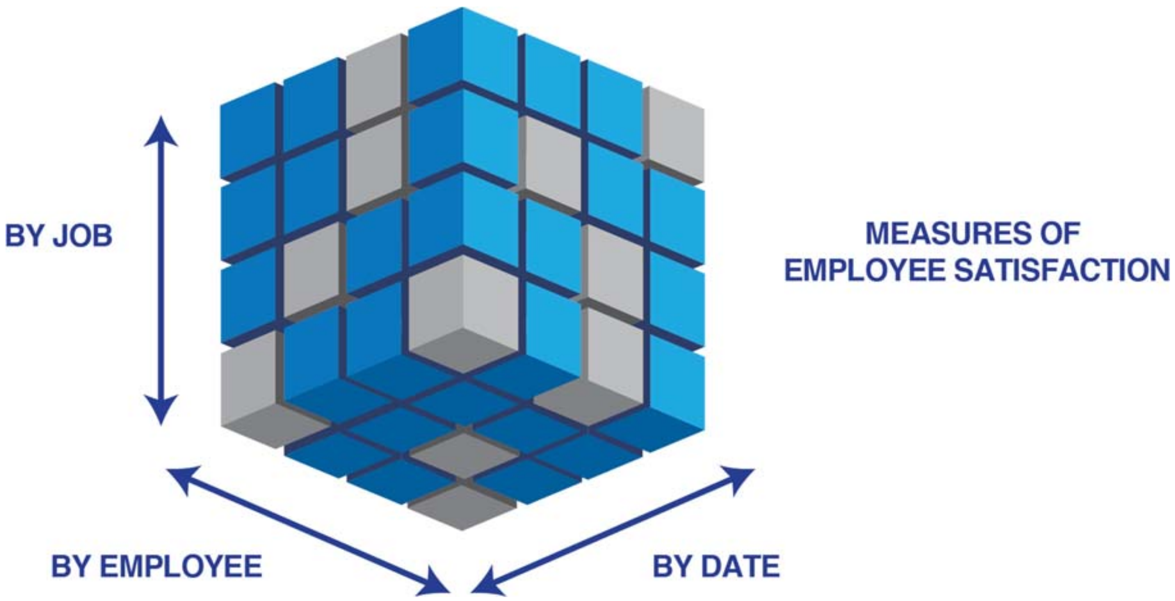
Notes:

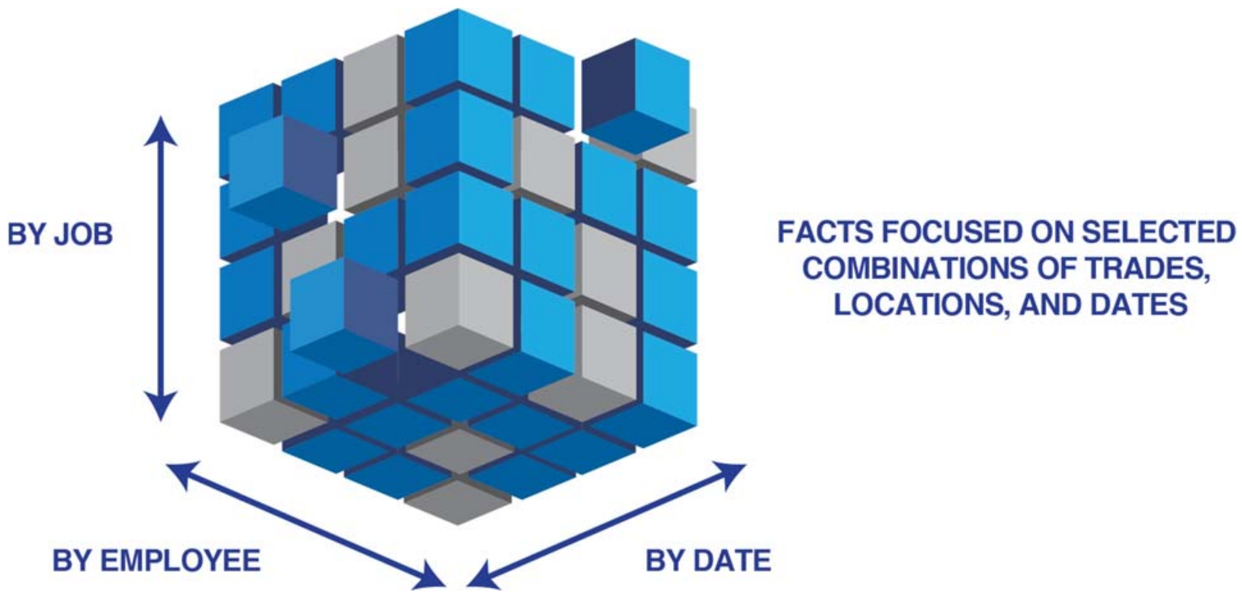
Module 3

OLAP and Other Information Services

OLAP Services

Online Analytical Processing




BY JOB
FACTS FOCUSED ON SELECTED COMBINATIONS OF TRADES, LOCATIONS, AND DATES
BY EMPLOYEE
BY DATE

OLAP Services

Online Analytical Processing

OLAP DEFINED

OLAP is a data access and manipulation technology that allows users to selectively access data and view it from different perspectives. OLAP tools are used to analyze different dimensions of multidimensional data. The data typically represents business measures that can be aggregated along dimension hierarchies, selected and limited by elements of dimensions, and viewed at the intersections of multiple dimensions.

OLAP is among the most common business analysis tools currently in use. OLAP is powerful; the tools are relatively easy to learn and use, and they are a widely accepted and integral part of business analysis.

THE OLAP INTERFACE

Selecting which dimensions are in view and on which axis is achieved using a pivot-table-like interface. The OLAP cube (or star schema) with this interface supports several operations:

- *Slice*—Slicing selects a subset of the data by limiting data to a single value for one dimension. Slicing produces a two-dimensional array.

- *Dice*—Dicing selects a subset of the data by limiting the values that are active in two or more dimensions. Dicing produces a three-dimensional array that is in effect a smaller cube.

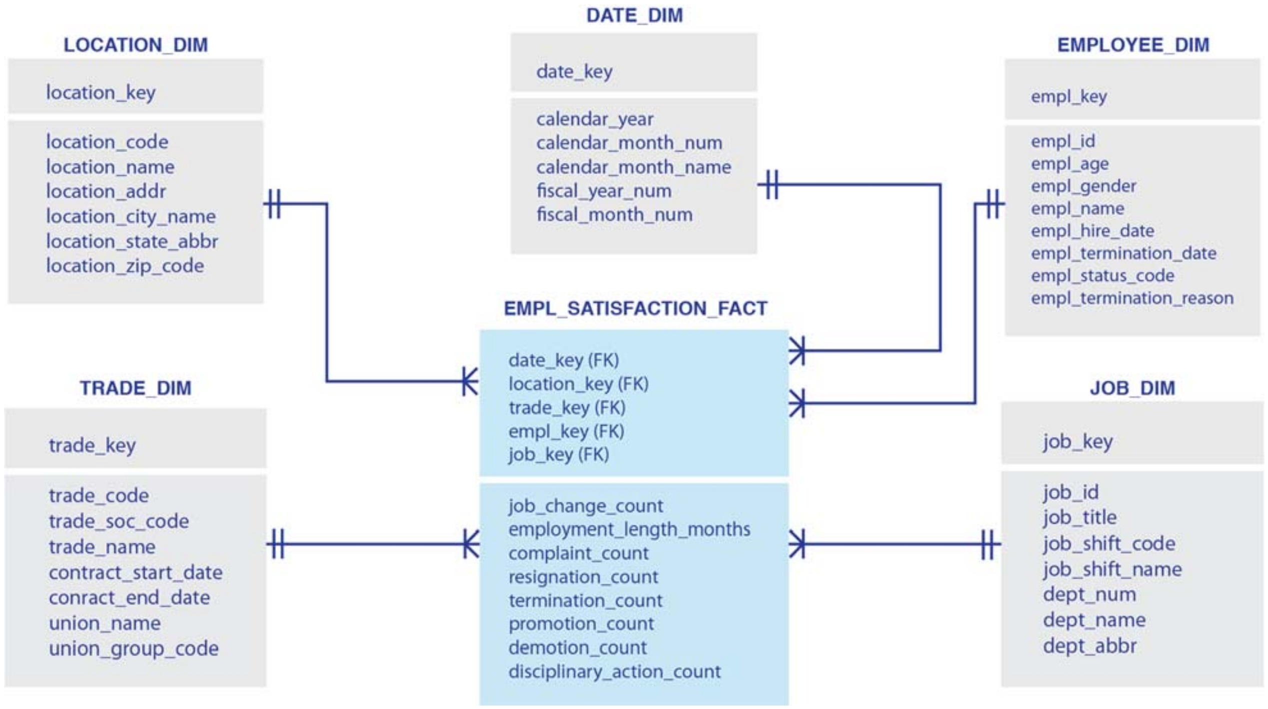
- *Drill Down*—Drilling down moves from summary to greater detail, resulting in finer grain for the data that is in view. Drill is an important analysis function as most analysis begins by looking at summary data then drilling to detail where questions arise.

- *Roll Up*—Rolling up summarizes data to show totals at varying levels along a hierarchical dimension.

- *Pivot*—Pivot rotates a cube in space to change which dimensions are positioned on the x, y, and z axes of a three-dimensional view.

OLAP Services

Dimensional Data Marts and Star Schema



OLAP Services

Dimensional Data Marts and Star Schema

DIMENSIONAL DATA MARTS

Dimensional data marts are the data foundation for OLAP. A data mart is a set of data tailored to support the analysis and reporting requirements of a business unit, business function, or work group. A dimensional data mart organizes the data as a collection of related facts (typically business measures) that are associated with analysis dimensions.

FACTS, DIMENSIONS, AND CONFORMANCE

Key components of the dimensional model are *facts*, which are metrics, and *dimensions*, which are reference data.

In the example on the facing page *empl_satisfaction_fact* contains facts about employee satisfaction. Those facts can be selected, grouped, and summarized by any combination of date, location, employee, trade, and job as supported by the dimensions.

A given data mart may contain multiple dimensional models that reference the same reference data. A human resources data mart, for example, may have models for employee satisfaction, job actions, professional development, benefit participation, and payroll.

Higher order questions require combining data from multiple models. The principle of *conformance* states that reference data across models should be consistent. For example, questions about whether professional development improves employee satisfaction will require a consistent definition of employee across the models for satisfaction and professional development.

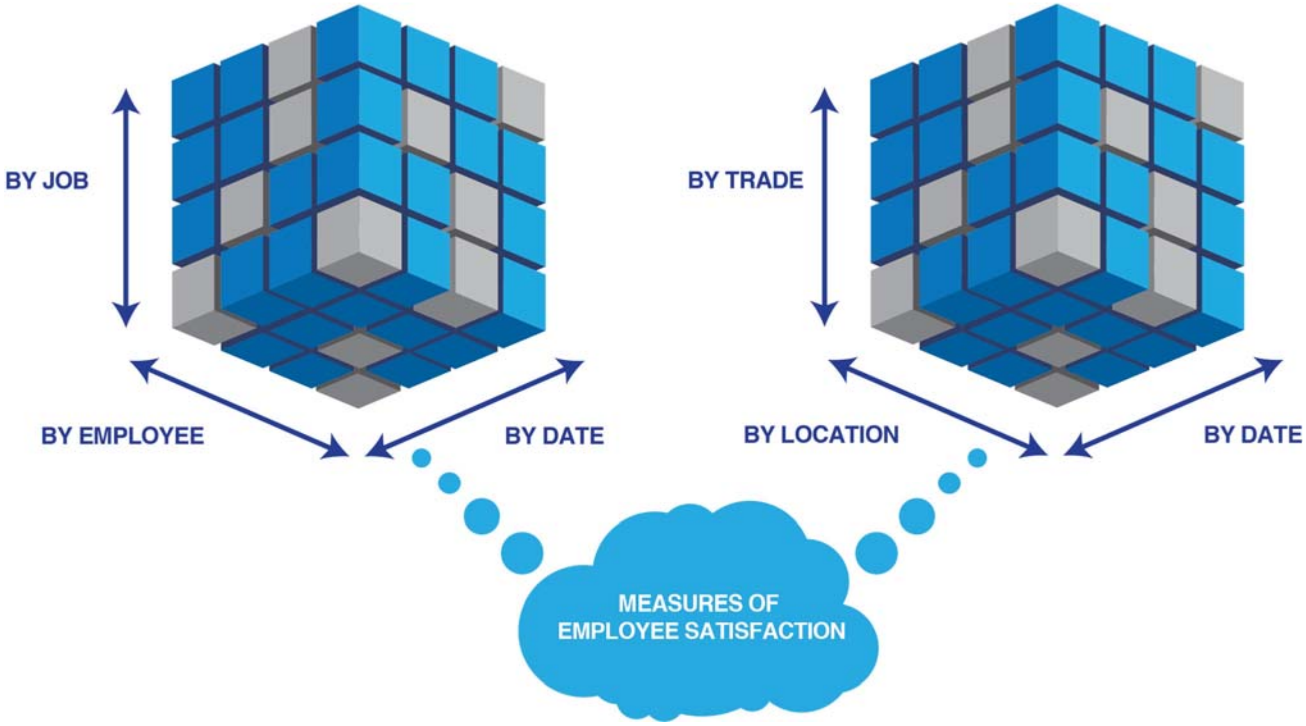
The need for consistent reference data, both within data marts and between them, requires significant effort in understanding business requirements and data integration. The most powerful implementations follow enterprise-level modeling efforts, but this is not always feasible.

STAR SCHEMA

A star schema is the physical implementation of relational tables to collect and store multidimensional data. A star schema consists of a single fact table surrounded by a set of dimension tables. The example on the facing page has one fact table—*empl_satisfaction_fact*—that is associated with five dimension tables: *date_dim, location_dim, employee_dim, trade_dim,* and *job_dim.* The schema can be implemented using RDBMS, but foreign key relationships are constrained to those that associate the fact table with each of the dimension tables.

OLAP Services

The OLAP Cube



OLAP Services

The OLAP Cube

DIMENSIONAL DATA STORAGE

Dimensional models may also be implemented in a multidimensional database (MDB) where the primary data structure is known as a cube.

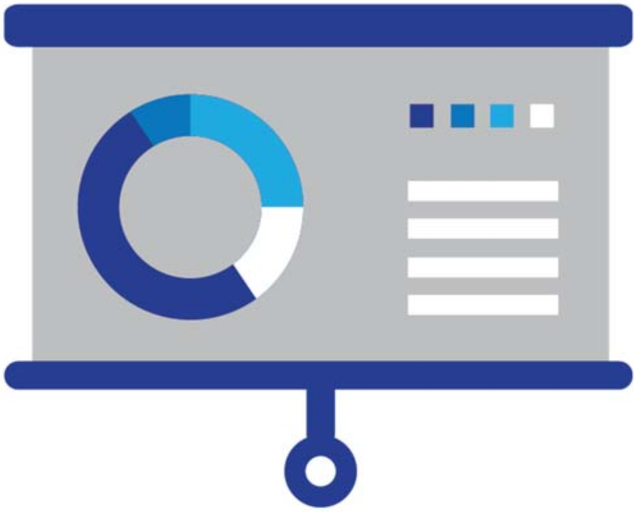
An *OLAP cube* is a data store in which contains both granular data and pre-summarized aggregations. When a cube is generated, summaries are pre-calculated and the data is stored together with the detail.

When OLAP operations are performed against star schema data in a relational database, summaries must be calculated at query execution time, which may result in performance concerns. With a cube, responses are faster, but scalability issues may arise due to pre-summarization.

In geometry a cube has exactly three dimensions of the same size. For OLAP more than three dimensions are allowed and the equal size constraint is removed. The cube analogy fits, however, because people are generally not able to view more than three dimensions simultaneously and OLAP technology provides the ability to change which dimensions are in view at any time.

BI Reporting

Enterprise and Operational Reporting



- ✓ PUBLISH & SUBSCRIBE
- ✓ ENTERPRISE REPORTING
- ✓ OPERATIONAL REPORTING

BI Reporting

Enterprise and Operational Reporting

PUBLISH AND SUBSCRIBE

The *publish-and-subscribe* model of reporting provides a catalog of reports that are published at regular frequency with an interface that allows business users to subscribe to reports of interest. Data reported to each subscriber may be limited by security constraints and other business rules.

ENTERPRISE REPORTING

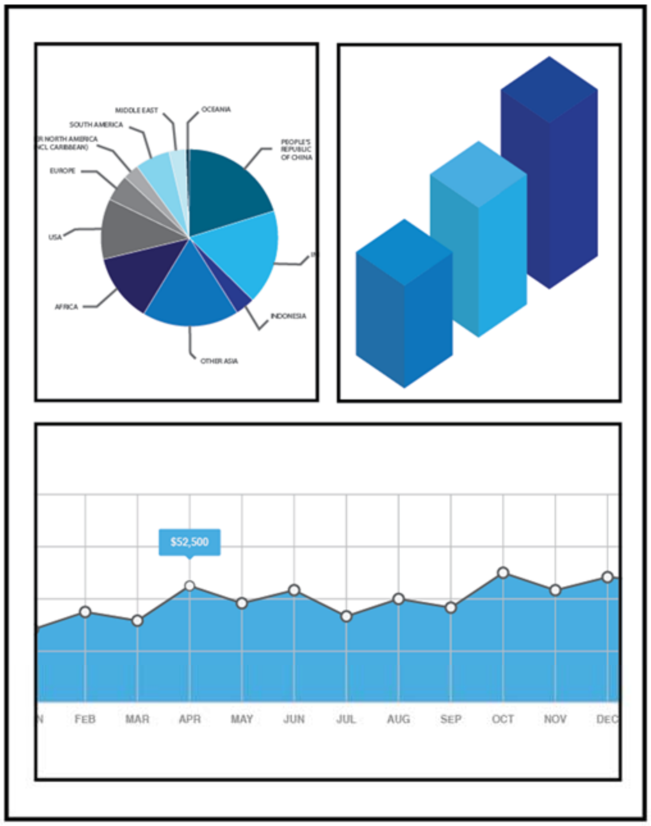
Enterprise reporting is the regular provision of information to decision makers throughout the enterprise to support them in their work. The reports typically contain summarized information, and they are often formatted as a combination of graphs, text, and tables. They may be delivered as dashboards, scorecards, financial statements, metrics monitors, etc. Enterprise reports are typically distributed through an intranet as a set of regularly updated Web pages or an enterprise reporting portal.

OPERATIONAL REPORTING

Operational reporting is more detailed and narrowly focused than enterprise reporting. Where enterprise reporting is summarized, operational reporting is detailed. Where enterprise reporting is enterprisewide, operational reports focus on specific business processes and day-to-day activities. In BI, operational reporting is a component of operational BI that is used to provide very low-latency feedback to operational processes to inform very short-term and transactional decisions.

BI Reporting

On-Demand Reporting



- ✓ AD HOC REPORTING
- ✓ PARAMETERIZED REPORTING

BI Reporting

On-Demand Reporting

AD HOC REPORTING

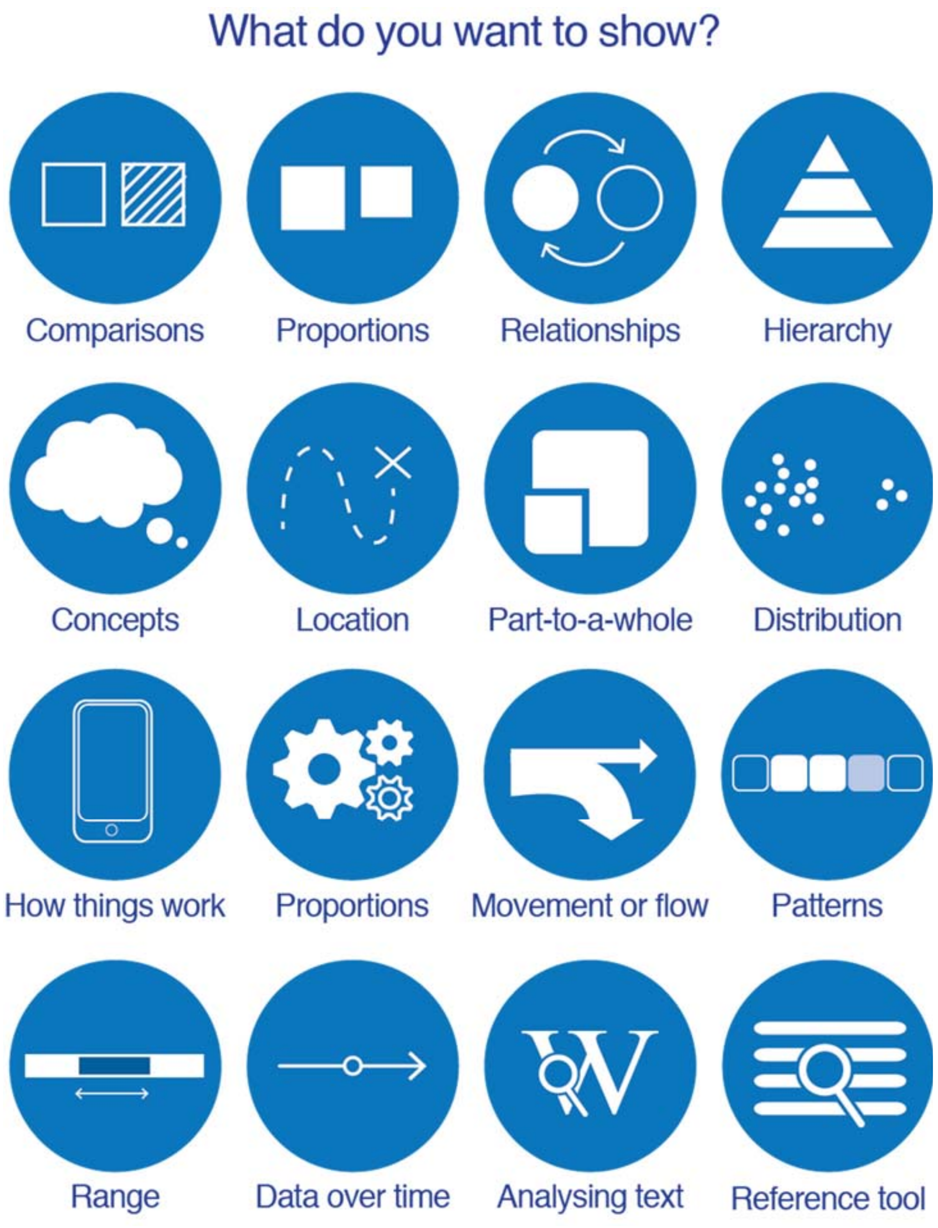
Ad hoc reporting is similar to ad hoc query—user initiated and customized to meet specific requirements. These reports are created using advanced query- and report-generating tools to select data, sequencing, formatting, totals, etc.

PARAMETERIZED REPORTING

Parameterized reporting is similar to parameterized query with user-supplied values at execution time for a limited set of variables. More rigid in formatting, sequencing, and totaling than ad hoc reporting, the most common parameters for these reports are for user selection of data with some ability to control sequence and totals.

Visualization and Storytelling

Communicating Insights



Source: www.datavizcatalogue.com

Visualization and Storytelling

Communicating Insights

A NEW LITERACY

As the facility to work with data takes hold in business, so to do the complexities of the insights confirmed or discovered. Some analytics confirm what is already known. In other cases, analytics bring insights that contradict conventional wisdom or are simply surprising.

Communicating the insights drawn from BI and analytics processes requires crafting a message that is comprehensible, memorable, and drives action. This has led to an increasing emphasis on the importance of data visualization.

The data visualization catalogue (www.datavizcatalogue.com) is a valuable and free online resource to help find the best-suited chart to communicate your intended message and to understand the anatomy and use of each chart type.

At the time of this writing the catalog has details for 60 types of charts and graphs. The website opens to a master index of the chart types, and also includes capability to:

- Search by function—find graphing methods suitable for comparisons, proportions, relationships, hierarchy, concepts, location, part-to-whole relationships, distribution, how things work, processes and methods, movement and flow, patterns, ranges, data over time, text analysis, and reference tools.

- View by list—where the list is categorized by graphs and plots, diagrams, tables, maps, and other visuals.

The facing image shows the structure of the search by function Web page.

Visualization and Storytelling

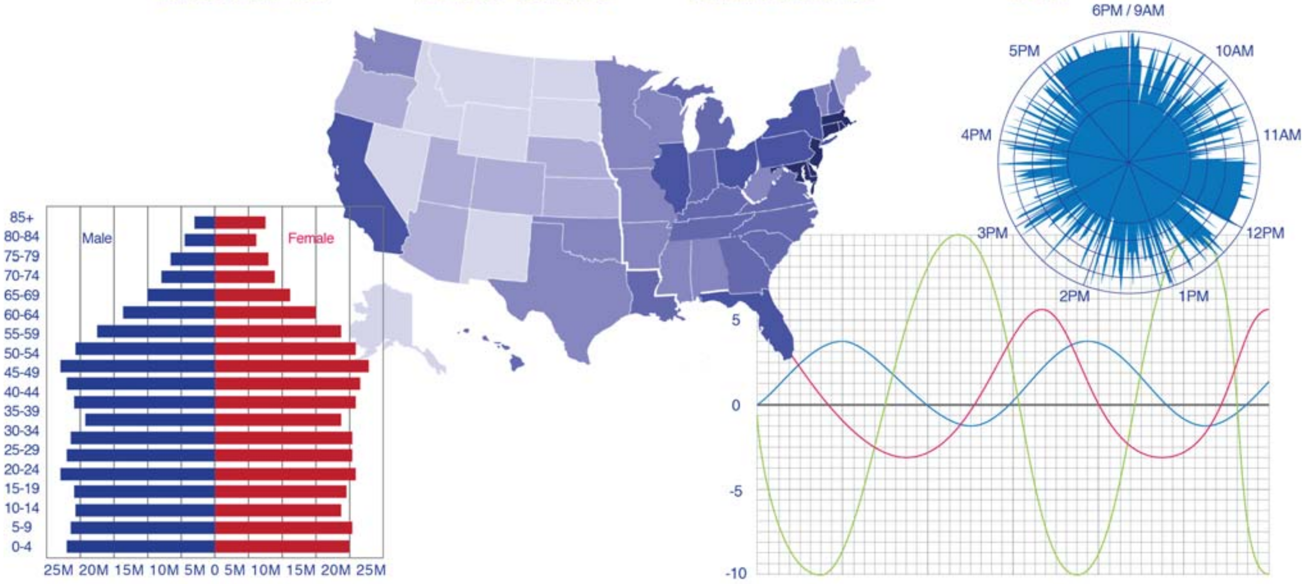
Data Visualization

LINE GRAPHS
COLUMN GRAPHS
BAR GRAPHS

PICTOGRAPHS
PIE CHARTS
DONUT CHARTS

COSMOGRAPHS
SCATTER GRAPHS
AREA GRAPHS

SURFACE GRAPHS
BUBBLE GRAPHS
ETC.



Visualization and Storytelling

Data Visualization

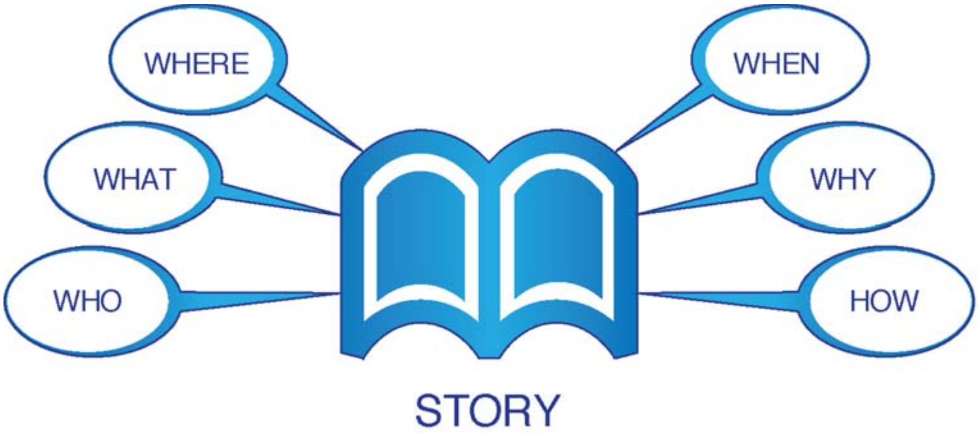
SEEING THE DATA

Data visualization encompasses the processes, techniques, and tools to present data in visual formats that allow information consumers to see data in graphical and non-tabular formats. Visualization can highlight patterns, trends, outliers, and anomalies in data that are not easily seen when the data is presented in columnar reporting formats. It also enables high-density data presentation (lots of information in a small physical space) that is especially important for performance dashboards and mobile BI.

Visualization goes well beyond the familiar bar charts, pie charts, and line graphs that are typical of spreadsheets. Dials, bullet graphs, heat maps, geographic maps, pictographs, and more are included in the visualization toolkit.

Visualization and Storytelling

Data Storytelling



STATISTICAL TERMS	STORYTELLING TERMS
Average, median, mode	Usual, typical, customary
Standard deviation, variance	Unusual, exceptional, dissimilar
Probability	Chance, odds
Population, sample	Everyone, instance, specimen

From *Ten Mistakes to Avoid in Data Storytelling*,
© David L. Wells. Used with permission.

Visualization and Storytelling

Data Storytelling

WHY STORIES?

Data stories group several related visualizations and connect them with narrative that expresses the analyst/storyteller's interpretation of their meaning. Storytelling makes analytics more real and personal than individual charts and graphs without interpretation. A good data story connects the data and the visuals with cause and effect, elicits personal responses, and drives conversation and interaction

Good stories break through the everyday noise in business by speaking to deep human needs. Storytelling is useful in presenting information or results related to virtually any type of analysis. It is also useful at other points during the development cycle, such as during requirements discovery and verification.

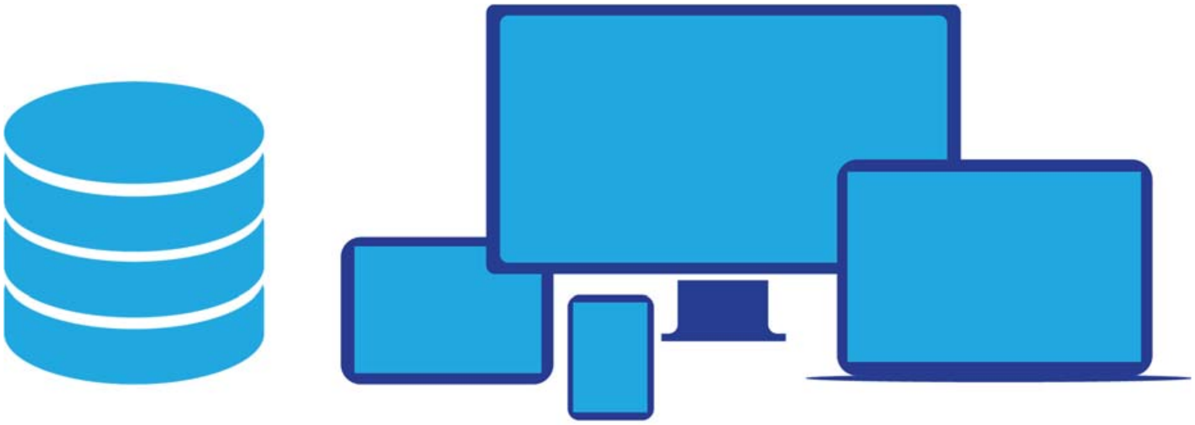
DIFFERENCES IN COMMUNICATION

Statistics are used to convey information, but they may not be the best way to convey findings and communications. A data story connects the cause and effect better than mere data points, tables, and graphs. These also help to drive communication and elicit personal responses.

John Allen Paulos describes the relationship between stories and statistics. (http://opinionator.blogs.nytimes.com/2010/10/24/stories-vs-statistics/?php=true&_type=blog&_r-l&_r=0).

Data Access and Delivery

Query Services



- ✓ AD HOC QUERY
- ✓ MANAGED QUERY
- ✓ PARAMETERIZED QUERY

Data Access and Delivery

Query Services

FROM QUESTIONS TO DATA

Query services are the means by which data consumers access databases to obtain data and answer simple questions. When querying relational databases, a technical person might code SQL queries, but that isn't ideal for the typical business user. Query services are implemented using graphical user interfaces and query-builder wizards to create SQL queries without the need to understand SQL coding. Query services can also be connected to dimensional data, and provided in conjunction with OLAP services.

Query services generally support three kinds of queries:

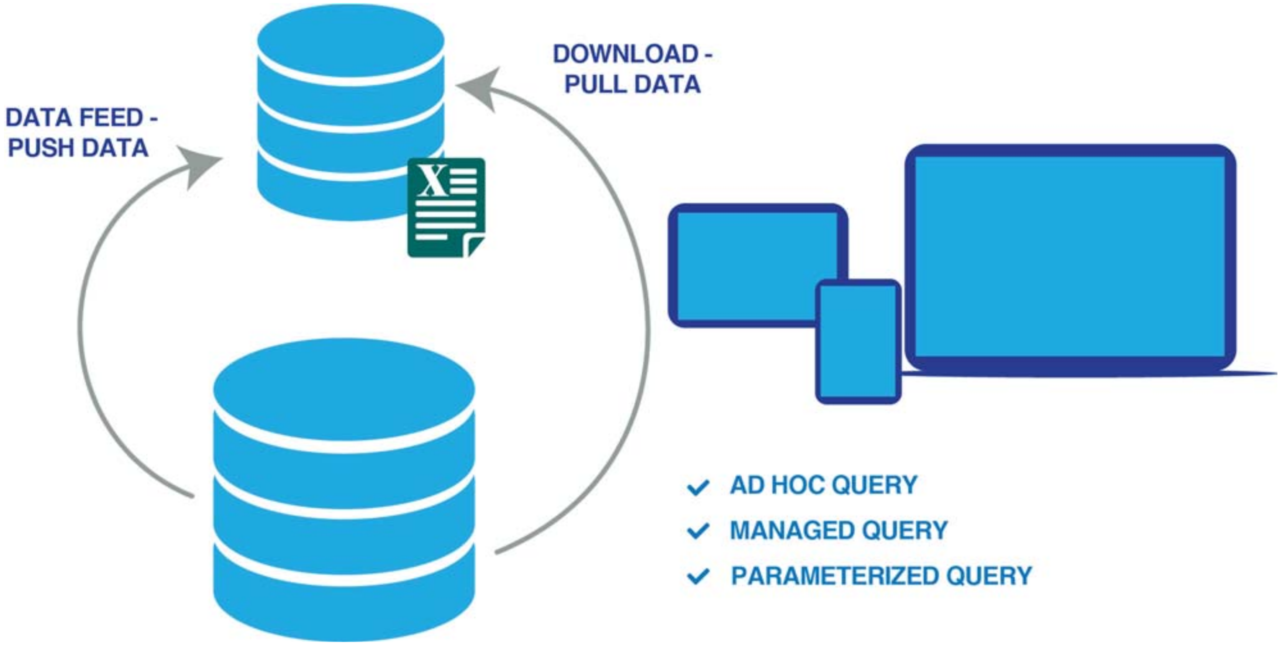
- *Ad hoc queries* where users create specific, customized queries for their unique needs.

- *Managed queries* in a managed query environment (MQE) where users access and apply preconfigured query structures including stored procedures to meet their data access needs. MQE is less prone to performance and security challenges than ad hoc query.

- *Parameterized queries* in which the user supplies the values of some variables at execution time. Variables used for selection and grouping are most common among user-supplied variables.

Data Access and Delivery

Data Feeds and Downloads



Data Access and Delivery

Data Feeds and Downloads

JUST THE DATA, PLEASE

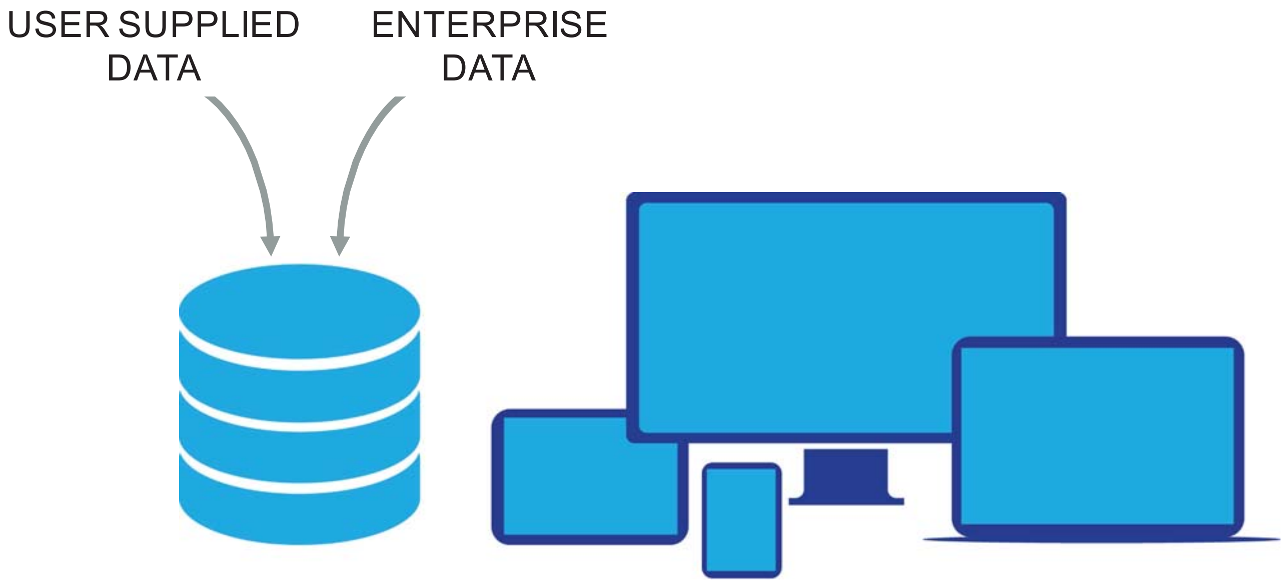
Sometimes business people just want data. They load it into spreadsheets, blend it with their local data, build their own reports, perform their own analysis, and use it to feed end-user systems and databases. Data distribution occurs in two common ways:

- Data feeds "push" data to those who need it using a predefined and prescribed format. A publish/subscribe capability is sometimes used to distribute data feeds.

- Downloads deliver data that is "pulled" by users through query services. Downloads offer more formatting flexibility than data feeds but place greater responsibility on the user to get the right data at the right time.

Self-Service

Evolving Service Models



USER SUPPLIED DATA

ENTERPRISE DATA

- ✓ DATA IMPORT
- ✓ PROFILING
- ✓ DATA BLENDING
- ✓ EXPLORATION

Self-Service

Evolving Service Models

BRING YOUR OWN DATA

Self-service is one of the latest trends in BI and analytics. A self-service environment allows someone to import their own data sets into the data management environment and then perform their own data exploration activities. These may include:

- Data *import (*or *ingestion),* often supported by a *data lake* (to be discussed in Module 4)
- Data *profiling* (to be discussed in Module 5)
- Data *blending* (or *integration*) with other enterprise data (to be discussed in Module 4)
- Data exploration
- Visualization and reporting

OTHER SERVICE MODELS

Self-service is one extreme of a spectrum of service models.

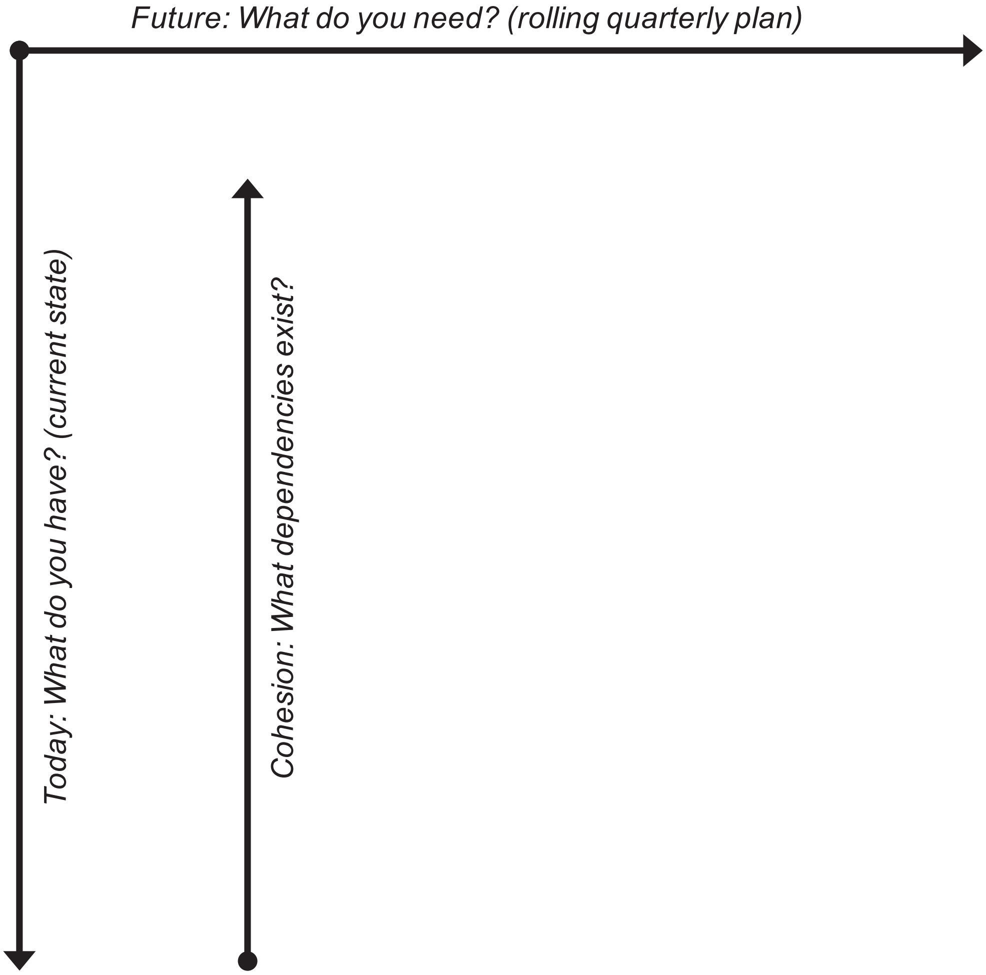
Central services are the "we build it for you" model that works well for standard reports and routinely published information. In the central services model, standards, processes, and technology are prescribed. A single centralized team is responsible for development, deployment, and management of information services. This model works well when goals are exceptional consistency, strong governance, rapid delivery, and managed costs. The central services model may be challenged to scale up to meet high demand for services.

Shared services is the "we build the Legos" model where a central team builds and publishes reusable data components that are accessed, configured, and assembled by distributed teams to meet their local needs. With published interfaces it is practical for local data to be appended to or integrated with central data. The shared services model defines processes, standardizes architecture, and maintains a centralized team for shared work, but much project and process work occurs in individual project teams and distributed business units. The blend of centralized and decentralized resources achieves good efficiency of resource utilization.

Hybrid services. As a practical matter, many organizations evolve to a mix-and-match hybrid of service models. Good guidelines and clear understanding of the criteria by which projects and service models are matched is important to ensure appropriate use of each level.

The BI and Analytics Roadmap

OLAP and Other Information Services


Future: What do you need? (rolling quarterly plan)
Today: What do you have? (current state)
Cohesion: What dependencies exist?

The BI and Analytics Roadmap

OLAP and Other Information Services

MAPPING INFORMATION SERVICES

The facing page illustrates parts of the BI and analytics roadmap that relate to information services. Review the current state inventory and identify future state requirements for the items listed below. Here you're identifying needs for business application systems to provide information services.

When defining the future state, be sure to include information services that are needed to support or enable business capabilities, metrics, and analytics that are already mapped.

- OLAP
    - Dimensional data marts
    - OLAP services
- Reporting
    - Enterprise reporting
    - Operational reporting
    - Publish and subscribe
- Data Access
    - Query services
    - Data feeds and downloads
- Service models
    - Central services
    - Shared services
    - Self-service

PROJECTS

What kind of projects are required to reach your future state? Map these into your roadmap, being sure to identify dependencies with other systems and planned activities. Projects may include:

- Develop new systems
- Extend / enhance
- Maintain / modify
- Decommission / retire

Consider priorities and dependencies as you plot each project on the timeline. Be specific about the items on the timeline. For example, list *enterprise sales performance reporting*, not simply *enterprise reporting*.

Mistakes to Avoid

In Dimensional Modeling

- IGNORING THE STRATEGIC VALUE OF DIMENSIONAL DESIGN
- NOT USING DIMENSIONAL DESIGN TO MANAGE SCOPE
- DOING DESIGN WORK AT DESIGN TIME
- LEAVING DESIGN TO DESIGNERS
- RELYING ON YOUR BUSINESS KNOWLEDGE
- NOT CONSIDERING YOUR DATABASE MANAGEMENT SYSTEM AND BUSINESS INTELLIGENCE TOOLS
- SACRIFICING DETAIL
- SHORTCUTTING HISTORY
- REUSING UNIQUE IDENTIFIERS
- TRYING TO SAVE SPACE

Source: *Ten Mistakes to Avoid in Dimensional Design* by Chris Adamson. © TDWI

Mistakes to Avoid

In Dimensional Modeling

TEN MISTAKES TO AVOID

Ten Mistakes to Avoid in Dimensional Design by Chris Adamson

In virtually every data warehouse implementation, you can find the products of dimensional design: the star schema, the snowflake, or the cube. Despite this near-universal acceptance, the basic principles of dimensional design are commonly misunderstood and misapplied.

Many mistakes are errors in approach, committed before design work begins. Chief among these is the failure to exploit strategic functions offered by a dimensional model. Others include waiting until the design stage to do dimensional design, then leaving the design work to designers.

Success is often thwarted by common technical errors, such as sacrificing operational detail or taking shortcuts with historical data. Designers also doom their solutions by failing to adapt them to the software tools that comprise the data warehouse.

Avoiding the ten mistakes on the facing page will help you ensure successful implementations in any data warehouse architecture, including those advocated by Ralph Kimball and W.H. Inmon.

Mistakes to Avoid

In Data Storytelling

- UNDERESTIMATING THE POWER OF THE NARRATIVE
- FOCUSING ON DATA BEFORE AUDIENCE
- MISSING THE CAUSE AND EFFECT
- USING THE LANGUAGE OF STATISTICS
- FAILURE TO SHOW YOUR VOICE
- PROVIDING TOO MUCH INFORMATION
- TELLING A ONE-SIZE-FITS-ALL STORY
- TELLING WITHOUT LISTENING
- DELIVERING STORIES WITHOUT ACTORS
- CHOOSING THE WRONG MEDIA

Source: *Ten Mistakes to Avoid in Data Storytelling* by Dave Wells. © TDWI

Mistakes to Avoid

In Data Storytelling

TEN MISTAKES TO AVOID

Ten Mistakes to Avoid in Data Storytelling by Dave Wells.

All too frequently, analytics activity ends with producing and publishing data visualizations—charts and graphs that illustrate findings from data but fail to interpret those findings, communicate insight, or inspire action. Although data visualization is valuable, the viewers of data visuals are left to draw their own conclusions that may be incorrect, out of context, or inconsistent.

Data storytelling takes the next step beyond data visualization by connecting multiple visuals with narrative, offering interpretation, and inviting conversation. Data storytelling is not just a way to share analytical insights. It is also a highly effective way to validate and refine insights. Data storytelling is the new horizon of business analytics.

A well-told story that is interesting and convincing may appear quite easy on the surface, but crafting a good story is challenging. Becoming a skilled storyteller is not easy and may be particularly challenging for those who are naturally inclined to the disciplines of structure, statistics, and sciences. Ten mistakes to avoid are provided on the facing page.

Discussion

OLAP and Other Information Services

LET’S TALK ABOUT IT!

- Does your organization currently provide OLAP services? For which business areas?
- To what degree is reference data consistent across data marts?
- What forms of reporting are a part of your architecture, and what is needed in the future?
- What is your current policy regarding self-service? Will this change in the foreseeable future?
- Has your organization developed literacies in data visualization and data storytelling? If not, what would be required to do so?

Discussion

OLAP and Other Information Services

Notes:

tdwi

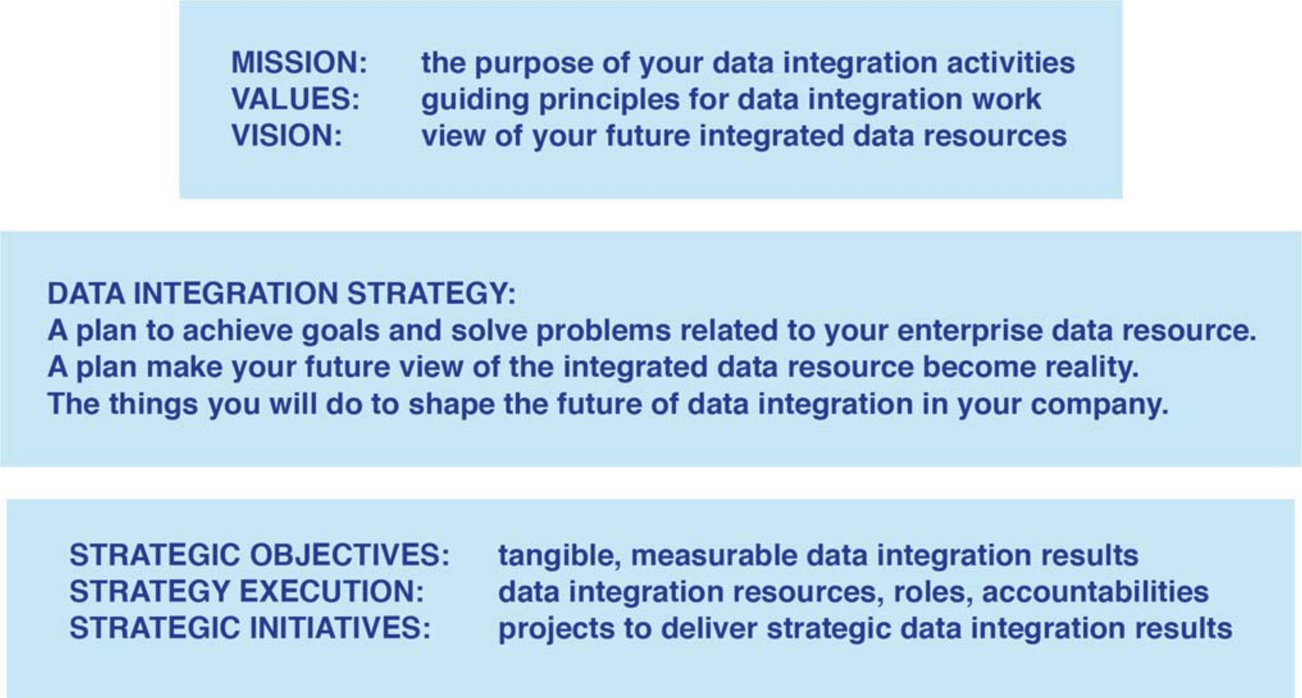
Transforming Data
With Intelligence™

Module 4

Data Integration

Data Integration Architecture

Integration Strategy



MISSION:	the purpose of your data integration activities
VALUES:	guiding principles for data integration work
VISION:	view of your future integrated data resources

DATA INTEGRATION STRATEGY:
A plan to achieve goals and solve problems related to your enterprise data resource.
A plan make your future view of the integrated data resource become reality.
The things you will do to shape the future of data integration in your company.

STRATEGIC OBJECTIVES:	tangible, measurable data integration results
STRATEGY EXECUTION:	data integration resources, roles, accountabilities
STRATEGIC INITIATIVES:	projects to deliver strategic data integration results

Data Integration Architecture

Integration Strategy

DEFINITION FOCUSED ON DATA INTEGRATION

Data integration strategy is a plan to shape the future of your integrated data resource. To develop that plan you need to know:

- Mission—the purpose of integrating data. Many state the purpose as "a single version of the truth"—a statement that may be too shallow or too much a cliché to effectively drive data integration vision.
- Values—guiding principles for data integration. What are the top three to five things that really matter here: availability? reliability? quality? timeliness? trustworthiness? data sharing? business impact?
- Vision—your future view of data integration. What current problems will be solved? What unfulfilled needs will be satisfied? What goals will be attained?

To achieve the desired future you need to identify:

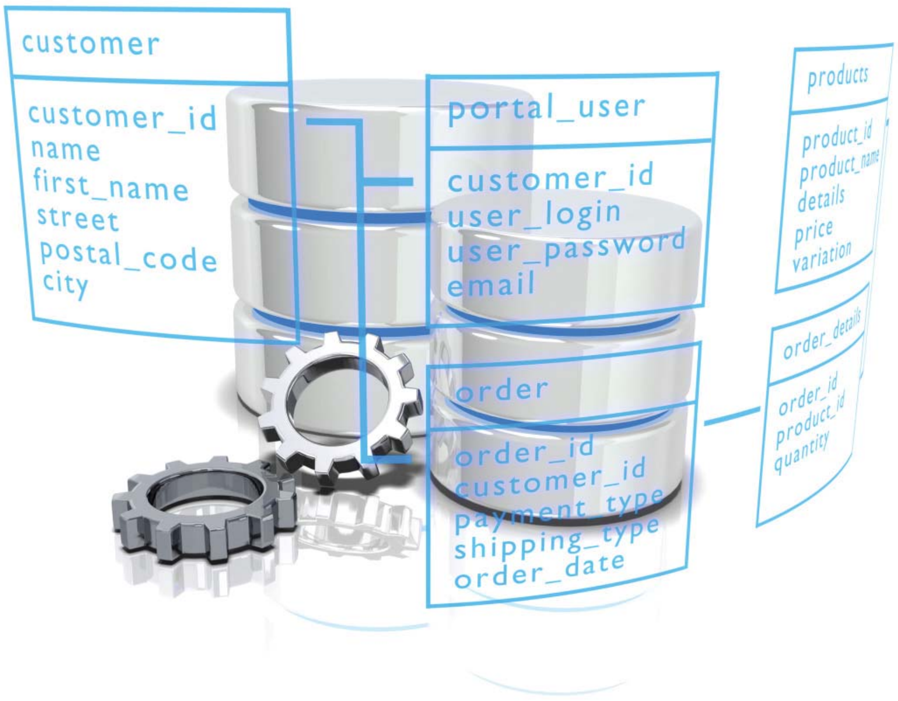
- Objectives—tangible, measurable data integration results that support stated values. The objectives should solve problems, meet needs, and achieve goals described by the vision.
- Execution—the who, what, and how to get the job done. Action is difficult to achieve without designated resources and roles, and it is difficult to sustain when accountability is not clearly designated.
- Initiatives—the projects, activities, and actions that will be undertaken to meet the objectives.

A FIRST STEP TO ARCHITECTURE

Strategy is a future view; the purpose of strategy is to shape your own future. To build a sustainable architecture, begin by looking at the future from a strategic perspective. For data integration the goal is to realize substantial benefits from well-managed, conscientiously integrated data. Commonly cited benefits include business agility, data asset value, and information quality.

Data Integration Architecture

The Purpose of Architecture



Well suited to intended purpose and use

Fit gracefully into the environment

Structurally sound

Compliant with codes and regulations

Sustainable through expected lifespan

Aesthetically pleasing

Data Integration Architecture

The Purpose of Architecture

DEFINITION

Architecture defines the roles, structure, relationships, and rules by which a collection of components constitute a cohesive whole—the glue that bonds individual parts in a system. Architecture is an early-stage design activity that precedes detailed design, specification, and construction.

ROLES OF ARCHITECTURE

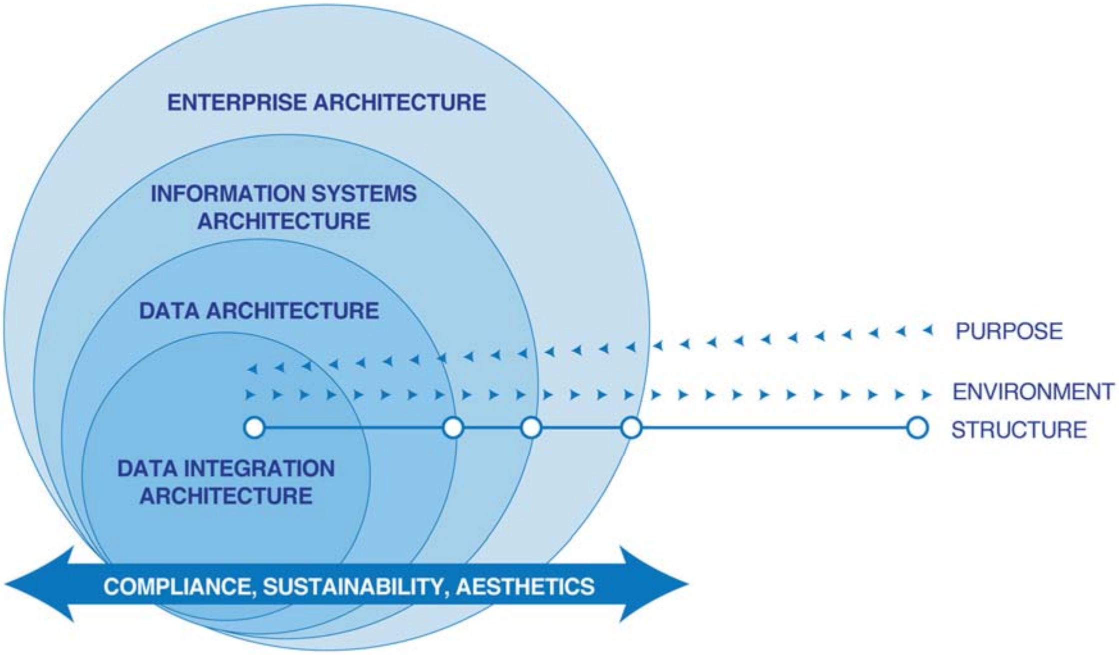
Effective architecture ensures that the things we build:

- Are suited to the purposes for which they are intended
- Fit gracefully into their environment
- Are structurally sound
- Comply with codes, regulations, and standards
- Are sustainable through their expected lifespan
- Are aesthetically pleasing

These principles hold true for the architecture of many things—including buildings, bridges, and information systems.

Data Integration Architecture

Integration and Data


ENTERPRISE ARCHITECTURE
INFORMATION SYSTEMS ARCHITECTURE
DATA ARCHITECTURE
DATA INTEGRATION ARCHITECTURE
PURPOSE
ENVIRONMENT
STRUCTURE
COMPLIANCE, SUSTAINABILITY, AESTHETICS

Data Integration Architecture

Integration and Data

DATA ARCHITECTURE

Data architecture defines the roles, structure, relationships, and rules to manage the data assets of an enterprise. Data architecture is a subset of information systems architecture, which is in turn a subset of enterprise architecture.

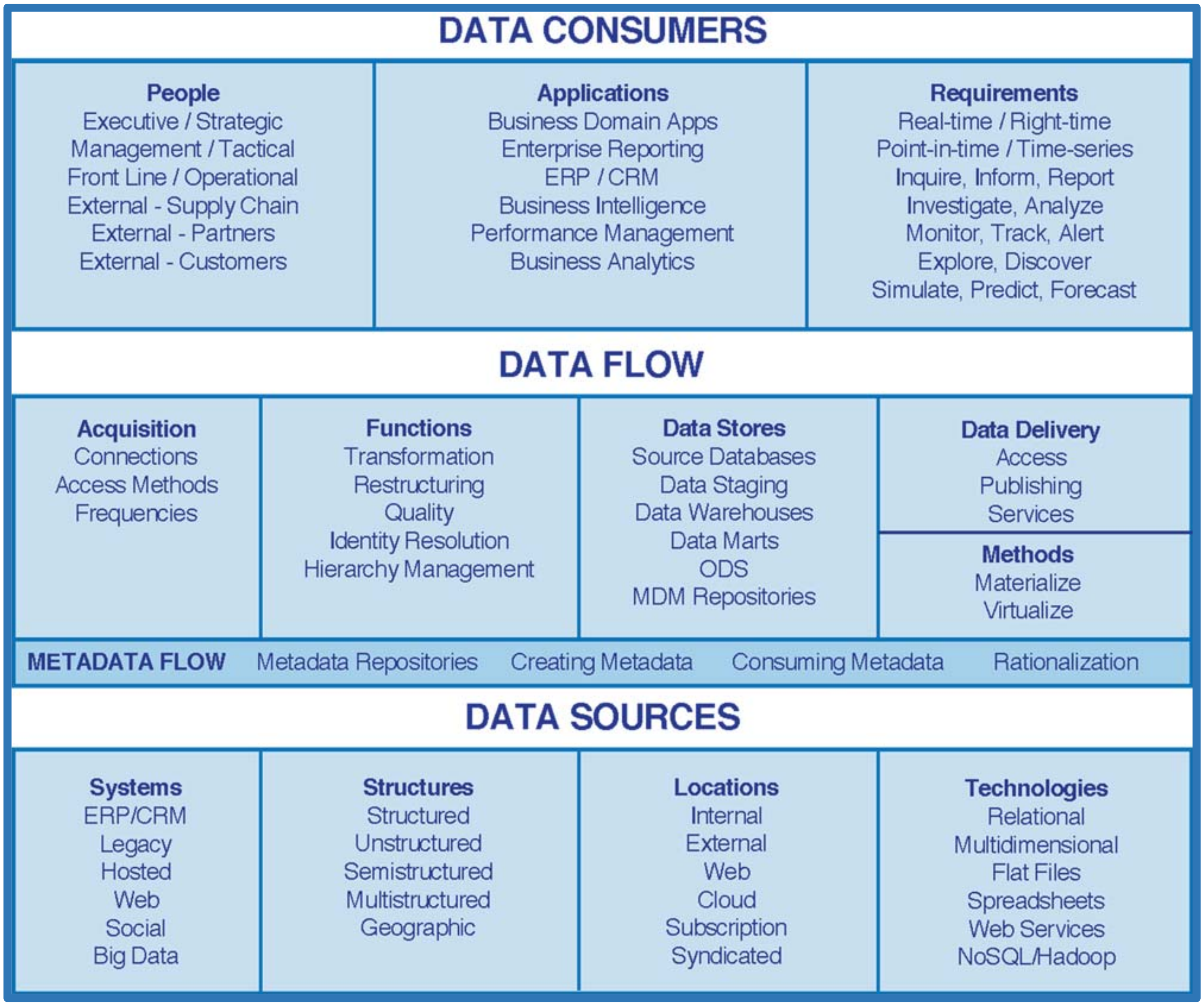
DATA INTEGRATION ARCHITECTURE

Data integration architecture is a subset of data architecture. The three supersets in which it is contained partially define the purpose of data integration and comprise the data integration environment. Architectural structure may be independent but compatible at each level. Compliance, sustainability, and aesthetics are consistent themes across all levels.

Data integration architecture defines the roles, structure, relationships, and rules to aggregate a collection of data integration components into a data integration system.

Data Integration Architecture

Components and Structures



DATA CONSUMERS

People
Executive / Strategic
Management / Tactical
Front Line / Operational
External - Supply Chain
External - Partners
External - Customers

Applications
Business Domain Apps
Enterprise Reporting
ERP / CRM
Business Intelligence
Performance Management
Business Analytics

Requirements
Real-time / Right-time
Point-in-time / Time-series
Inquire, Inform, Report
Investigate, Analyze
Monitor, Track, Alert
Explore, Discover
Simulate, Predict, Forecast

DATA FLOW

Acquisition
Connections
Access Methods
Frequencies

Functions
Transformation
Restructuring
Quality
Identity Resolution
Hierarchy Management

Data Stores
Source Databases
Data Staging
Data Warehouses
Data Marts
ODS
MDM Repositories

Data Delivery
Access
Publishing
Services

Methods
Materialize
Virtualize

METADATA FLOW Metadata Repositories Creating Metadata Consuming Metadata Rationalization

DATA SOURCES

Systems
ERP/CRM
Legacy
Hosted
Web
Social
Big Data

Structures
Structured
Unstructured
Semistructured
Multistructured
Geographic

Locations
Internal
External
Web
Cloud
Subscription
Syndicated

Technologies
Relational
Multidimensional
Flat Files
Spreadsheets
Web Services
NoSQL/Hadoop

Data Integration Architecture

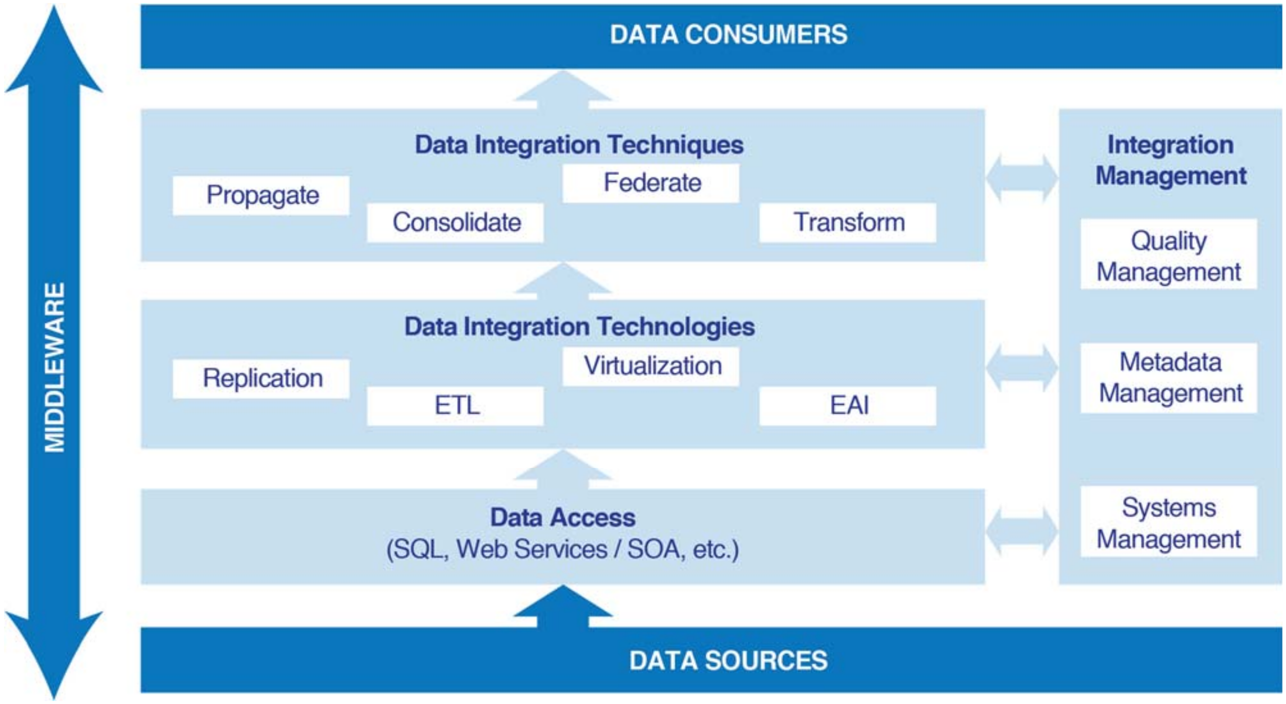
Components and Structures

CONNECTING BUSINESS WITH DATA

The purpose of data integration is to connect the business people and systems that consume data with the data that is needed in a way that reconciles inconsistencies among disparate data sources, represents all of the important relationships contained in the data, and provides the necessary metadata and interfaces to make the data understandable and accessible. The two end-points are data sources and data consumers. Both end-points are diverse and complex. The middleware that connects them is equally complex with many components to acquire, transform, store, describe, and deliver data.

Data Integration Architecture

Integration Techniques and Technologies



Data Integration Architecture

Integration Techniques and Technologies

TECHNIQUES

Several techniques are combined to achieve data integration goals:

- *Propagation* creates copies of data according to a set of rules. Source data from an operational database, for example, might be propagated to a staging area in preparation for data warehouse processing.
- *Federation* creates integrated views of data that is not physically integrated. Overlap and inconsistency among disparate data sources is rationalized and reconciled as views are created. Federated data sources are able to operate autonomously yet appear to be integrated. Federation might be used, for example, to build a quick on-the-fly data mart without the time and cost of physical integration.
- *Consolidation* builds a physical data store of integrated data that is a distinct separate copy of the original data sources. Overlap and inconsistency among disparate data sources is rationalized and reconciled as data is copied and moved. Data warehouses consist primarily of consolidated data.
- *Transformation* is the processing that changes data during federation and consolidation. The primary role of consolidation is to rationalize and reconcile inconsistencies among data sources.

TECHNOLOGIES

Data integration technologies support the techniques described above. Replication technology creates copies of data and is used in propagation. Virtualization technology is abstraction- and view-based data integration that is used to federate data. Extract, transform, and load (ETL) technology is used to consolidate data. Enterprise application integration (EAI) is a message-bus means of communicating among systems that can be used for extended capabilities in propagation, federation, and consolidation.

ACCESS

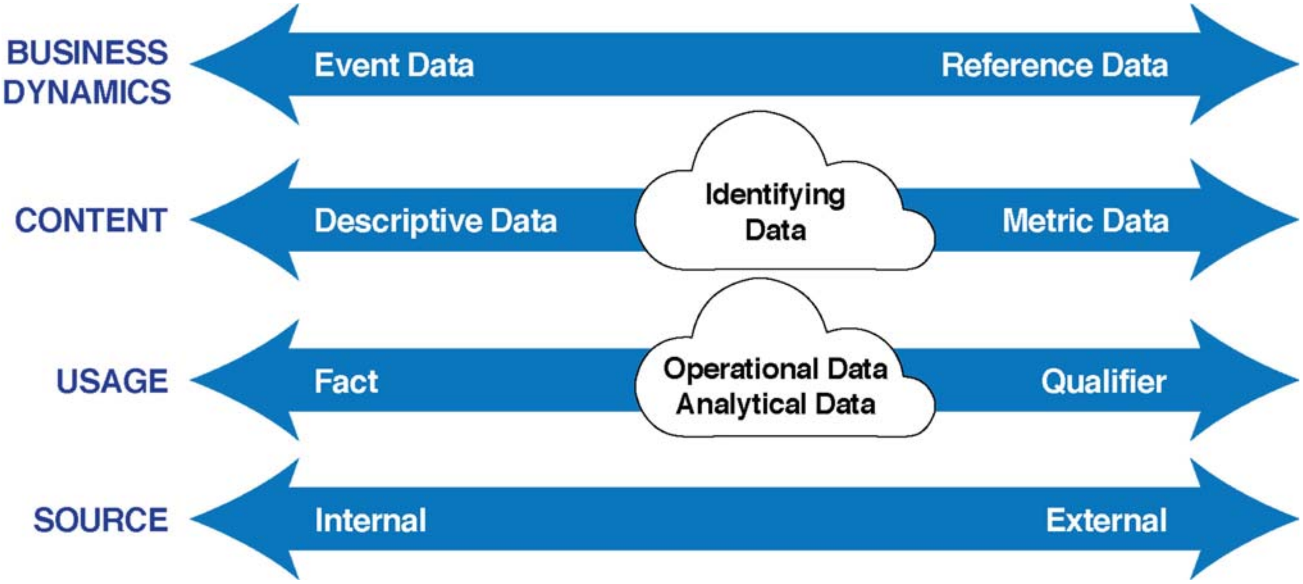
Data access includes all of the technologies, formats, and protocols that are necessary to connect with and acquire data from sources. The variety of query languages, connectors, and services depends on the kind of data sources that are accessed.

MANAGEMENT

Data integration is achieved through complex systems that require management of quality, of metadata, and of the systems themselves. Data integration systems are particularly sensitive to change—change in data sources, in target destinations of data, in business requirements, and in technology. The management components are essential to sustainable data integration systems.

Data Types and Sources

Data Properties



Data Types and Sources

Data Properties

PROPERTIES

Business intelligence and analytics work with many different types of data, and it is important to understand the types. A single data item generally has characteristics in more than one category. Not all categories are comprised of mutually exclusive data types. One set of data types can be viewed based on the data properties.

BUSINESS DYNAMICS

Business dynamics refers to how the data contributes to recording business activities. *Event Data* is data whose values are determined by a business occurrence. *Reference Data* is data whose values provide the context for business events—these values are not determined by the events for which they provide context.

CONTENT

Content classifies the nature of domain values of a data item. *Descriptive Data* is data that records the non-quantitative or non-measurement properties of things. *Identifying Data* is a subset of descriptive data that distinguishes among unique occurrences of an item. *Metric Data* is data that records the quantifiable facts that may be used as business measures.

BUSINESS USAGE

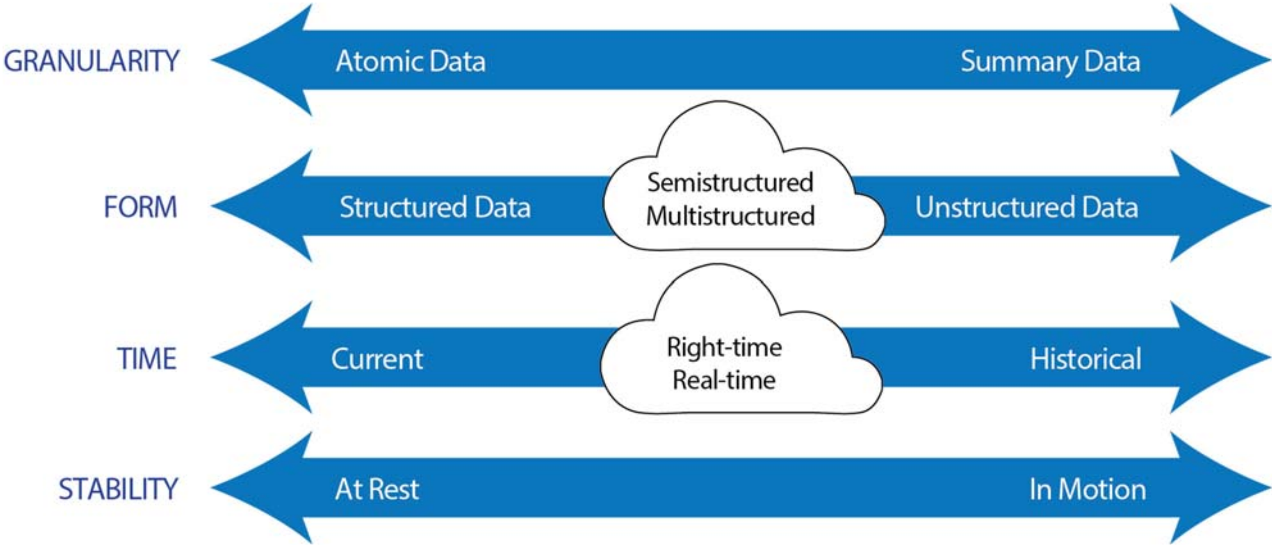
Business usage indicates how data is applied in specific business scenarios. A *fact* is a discrete item of business information of interest to a business person seeking information. A *qualifier* is a criterion by which information is accessed, sorted, grouped, summarized, and presented by a business person seeking information. *Operational Data* is data used in day-to-day business activities. *Analytical Data* is data used to analyze and understand what is happening and why it is happening.

SOURCE

Source indicates the origin of the data. *Internal Data* is data that is collected by the enterprise, managed by its systems, and stored within its databases, regardless of whether or not it is stored on site. It includes data in transactional databases as well as data in decision support databases, intranet content, email repositories, spreadsheets, etc. *External Data* is data that is not collected, managed, or owned by the enterprise and is acquired from sources such as data syndication and subscription services, including postal service databases, social media feeds, etc., to enrich internal data.

Data Types and Sources

Data Characteristics



Data Types and Sources

Data Characteristics

CHARACTERISTICS Another set of data types can be viewed based on data characteristics.

GRANULARITY Granularity refers to the level of detail of the data. *Atomic Data* is data at the lowest level of detail available from any source. *Summary Data* represents a higher level (e.g., monthly summary) than the atomic structures.

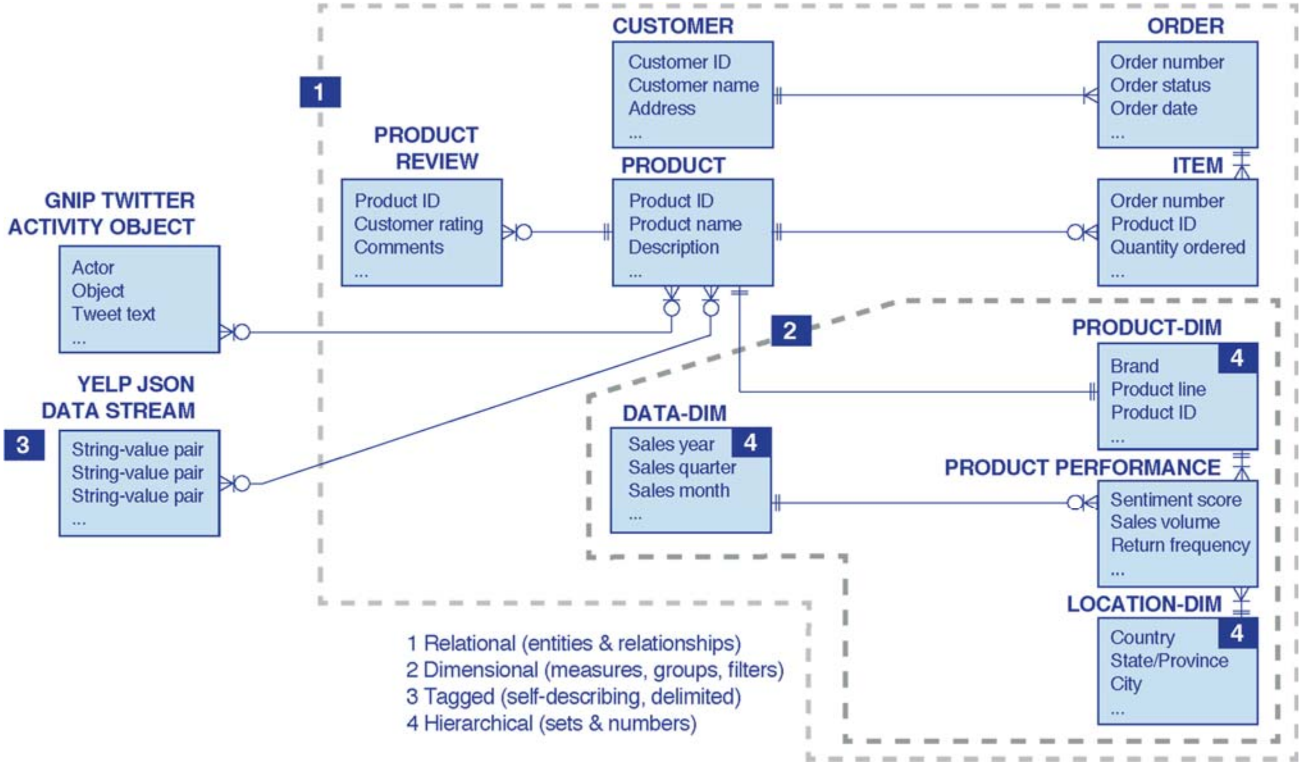
FORM Form indicates the degree and basis of the data structure. *Structured Data* is data that can be described using the relational data model, such as flat files, spreadsheets, relational databases, multidimensional databases, etc. *Unstructured Data* may actually have structure, but it is data that does not have a defined data model or does not fit the relational model. Examples include text and images. *Semistructured Data* is also known as self-describing data because the expression of the structure is embedded within the data itself. The best known example is XML. Unlike structured data, with semistructured data each instance of an entity may have different attributes and data types. *Multistructured Data* is data that contains a variety of data formats and types in a single database or data store. Email is an example of complex, multistructured data.

TIME Time refers to the currency and repeatability of the data. *Current Data* is data for which there is only one value retained, regardless of the age of the data. *Historical Data* is data for which multiple values exist and are retained over time. *Real-Time Data* is data that is available for usage as soon as the business event takes place. *Right-Time Data* reflects a compromise between the business value of having real-time (or near-real-time) data and the cost and complexity of providing it.

STABILITY Stability refers to whether or not the data is moving at the time of interest. *Data at Rest* is data that is physically in any digital form. It is used sometime after the event it is describing takes place. *Data in Motion* (also known as data in transit) is data that is being used while the event is taking place. It may be sourced internally from a real-time system capturing the information, or externally, for example from the Internet.

Data Types and Sources

Data Structure



Data Types and Sources

Data Structure

DATA STRUCTURE Data structure refers to how the data is physically organized in a data structure.

RELATIONAL DATA *Relational data* is organized logically as entities, relationships among the entities, and attributes of entities. It is physically implemented as tables that are organized into rows and columns. The data model on the facing page illustrates relational data as everything contained within the boundary that is labeled 1.

DIMENSIONAL DATA *Dimensional data* is organized logically as a collection of related business measures associated with dimensions that are used to group and filter the measures. It is implemented physically as relational tables in star schema configuration and sometimes as an OLAP cube that is generated from the star schema. Dimensional data is a subset of relational data. The data model on the facing page illustrates dimensional data as everything that is contained within the boundary labeled 2.

TAGGED DATA *Tagged data* is self-describing (semistructured) data that is delimited with characters that enable parsing separate tags and their associated values. The *Yelp JSON data stream* labeled 3 is an example of tagged data that contains social media reviews of products.

HIERARCHICAL DATA *Hierarchical data* is structured data organized as sets and subsets—data with parent/child relationships. In the example on the facing page each dimension (labeled 4) of the star schema is hierarchical: *brands* contain *product lines* which contain *products*; *years* contain *quarters* which contain *months*, etc. The hierarchies shown here are collapsed into a single table to conform to star schema standards. In other instances each level of hierarchy may be implemented as a separate table with one-to-many relationships among the tables.

Data Types and Sources

Big Data Defined

Data sets with sizes beyond the ability of commonly-used software tools to capture, integrate, manage, and process within a reasonable amount of time.

Massive volumes of both structured and unstructured data that are so large that they're difficult to process with traditional database and software techniques.

Data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.

(The McKinsey Global Institute, 2011)

Data Types and Sources

Big Data Defined

DESCRIPTION

The term *big data* has become popular to describe rapid growth in the volume, variety, and velocity of data that is now available in business—unstructured data, semistructured data, social media data, location data, radio frequency data, and more. These types of data tend to yield data sets that are too large, complex, or unwieldy to work with traditional data management and analytics technologies.

BIG DATA TYPES AND TECHNOLOGY

Structured big data can be readily stored in tabular forms. It is typically used for analysis and is stored and manipulated using columnar databases and high-speed analytics and data warehouse appliances based on multiple parallel processing (MPP) hardware architectures.

Unstructured and semistructured big data includes formats such as text, social media content, multimedia content, and Web logs. The technologies to work with unstructured big data are emerging and evolving. Today’s most commonly used technologies include Hadoop, MapReduce, and NoSQL.

Hadoop is a distributed file system designed to work with large data collections that combine structured data with more complex types of data. MapReduce is a programming framework to write applications that work with Hadoop data sets.

NoSQL describes a class of database management systems whose common characteristics are that they are not based on a relational model and they do not use SQL as a programming language. NoSQL databases are optimized for distributed storage and retrieval of very large volumes of data, either structured or unstructured. NoSQL supports append and read functions. Insert, delete, update, and join are not typically supported.

Data Types and Sources

Big Data Sources

- ✓ Web and Social Media
- ✓ Machine to Machine (M2M)
- ✓ Big Transaction Data
- ✓ Biometrics
- ✓ Human-Generated Data
- ✓ Publicly Available Data
- ✓ Legacy Documents

Data Types and Sources

Big Data Sources

WEB AND SOCIAL DATA

Web and social data is among the most frequently used big data. It includes clickstream data captured when people interact with websites, the content that is collected by Web crawlers (Web scraping), and the social media data that can be acquired from Facebook, Twitter, LinkedIn, blogs, and similar sites.

SENSORS AND M2M DATA

Sunil Soares describes *machine-to-machine data* as acquired through "technologies that allow both wireless and wired systems to communicate with other devices. M2M uses a device such as a sensor or meter to capture an event such as speed, temperature, pressure, flow, or salinity." (*Big Data Governance* by Sunil Soares, p.11)

BIG TRANSACTION DATA

Big transaction data does not mean data about high volume or high dollar value transactions. This data records especially large volumes of transactions such as healthcare billings, telecommunication call detail records, and utility billing.

BIOMETRICS

Biometrics "refers to the automatic identification of a person based on anatomical or behavioral characteristics or traits. Anatomical data is created from the physical characteristics of a person including a fingerprint, an iris, a retina, a face ... Behavioral data includes handwriting and keystroke analysis ... Biometric data is increasingly available in the commercial arena where it can be combined with other types of data such as social media." (*Big Data Governance* by Sunil Soares, p.11)

HUMAN-GENERATED DATA

Human-generated data includes unstructured and semistructured data such as call center agents' notes, voice recordings, email, paper documents, surveys, and electronic medical records. Much of human-generated data is text and is processed using text analytics.

PUBLICLY AVAILABLE DATA

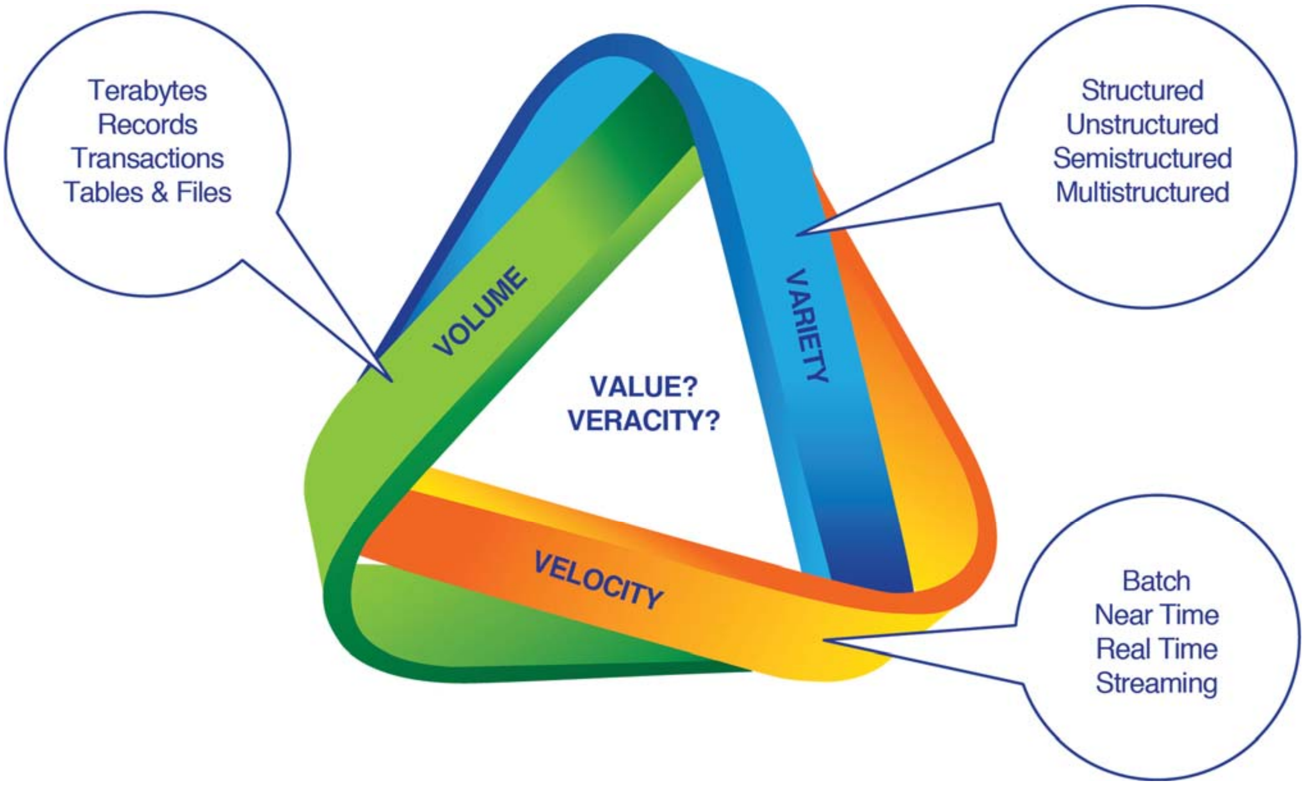
Publicly available data comes in many forms and from a variety of sources including government agencies and data services providers. Weather data, map data, census data, address data, and much more is available often at little or no cost.

LEGACY DOCUMENTS

Policies and procedures, legal documents, and text- and form-based records exist in virtually every enterprise, often in machine-readable forms. Research, legal discovery, and similar disciplines can find great value in these big data sources.

Data Types and Sources

Big Data Characteristics



Data Types and Sources

Big Data Characteristics

THREE V'S

Gartner describes big data as "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

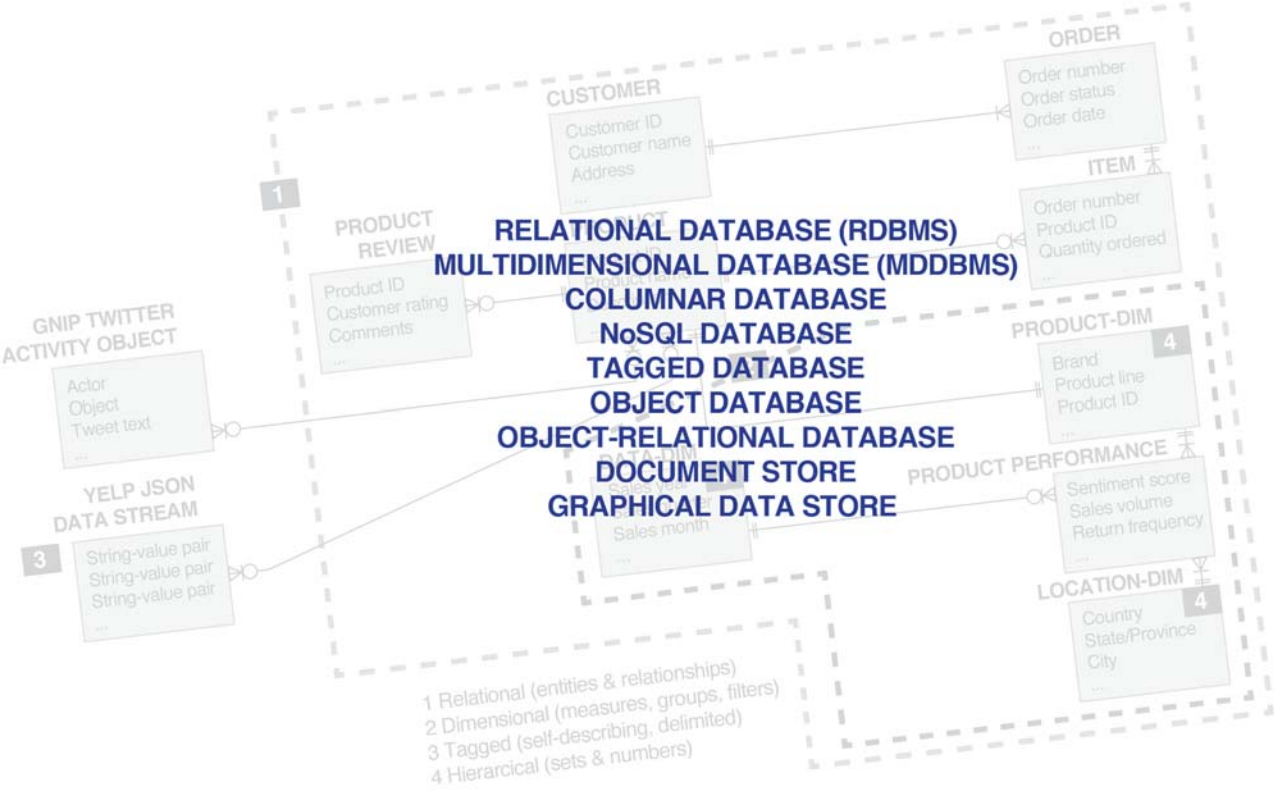
Volume, velocity, and variety are often repeated as the defining characteristics of big data. Volume describes the amount of data. Variety addresses the many types of data that are available—structured, unstructured, social, spatial, etc. Velocity describes the accelerating rate at which data becomes available.

PLUS TWO MORE

The three V's of big data are quite widely recognized today. Some organizations include the additional V's of veracity and value. Veracity is concerned with quality and reliability of data, and value with business impact. Although the two additional V's are not necessarily defining characteristics of big data, they may define *useful* big data.

Data Types and Sources

Physical Storage



RELATIONAL DATABASE (RDBMS)
MULTIDIMENSIONAL DATABASE (MDDBMS)
COLUMNAR DATABASE
NoSQL DATABASE
TAGGED DATABASE
OBJECT DATABASE
OBJECT-RELATIONAL DATABASE
DOCUMENT STORE
GRAPHICAL DATA STORE

Data Types and Sources

Physical Storage

ROW-BASED

Row-based databases present data as two-dimensional arrays of rows and columns where each row represents one occurrence of an entity and the columns represent attributes of the entities. Data is physically stored as records or strings where each record contains the data of one row. This is the most common form of data storage for relational data.

COLUMNAR

Columnar databases work with relational data similar to row-based databases. They present data as two-dimensional arrays and physically store data as records, but there is a significant difference between row-based and column-based databases. Where row-based stores each row as a record, columnar databases store each column as a record. Columnar databases are optimized for analysis where typically many rows but only a few columns are used in an analysis process. Eliminating the need to access all of the columns for a large number of rows accelerates data access, reduces the volume of data retrieved, and enables in-memory data handling.

MULTI-DIMENSIONAL

Multidimensional databases are optimized for OLAP processing. They are the means to store OLAP cubes with summaries pre-calculated along dimensions and combinations of dimensions. Where relational databases store data as strings or records, multidimensional databases store complex arrays. Where relational databases are queried using structured query language (SQL), multidimensional databases use multidimensional expressions (MDX) language.

NO-SQL

NoSQL defines a class of databases that are optimized to manage and access exceptionally large volumes of data. The variety of NoSQL databases is large and the common characteristics are that data is not stored as tables and not accessed with SQL. Flat files, key-value pairs, arrays, and combinations of these methods can be found in various NoSQL databases.

TAGGED

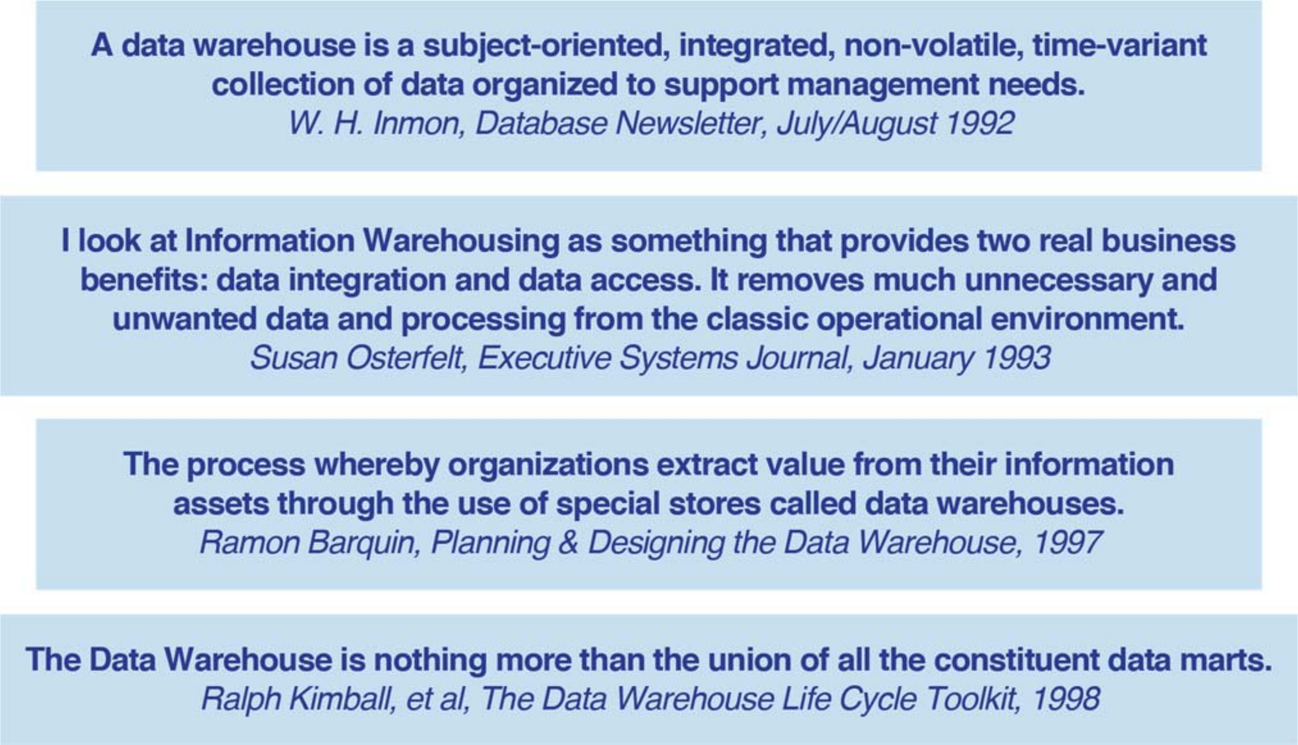
Tagged databases are used to store semistructured data such as XML-based data.

DOCUMENTS AND OBJECTS

A variety of specialty databases—object, object-relational, document store, and graphical data store among them—are used to store various types of unstructured and multistructured data.

Data Warehousing

Definitions



A data warehouse is a subject-oriented, integrated, non-volatile, time-variant collection of data organized to support management needs.
W. H. Inmon, Database Newsletter, July/August 1992

I look at Information Warehousing as something that provides two real business benefits: data integration and data access. It removes much unnecessary and unwanted data and processing from the classic operational environment.
Susan Osterfelt, Executive Systems Journal, January 1993

The process whereby organizations extract value from their information assets through the use of special stores called data warehouses.
Ramon Barquin, Planning & Designing the Data Warehouse, 1997

The Data Warehouse is nothing more than the union of all the constituent data marts.
Ralph Kimball, et al, The Data Warehouse Life Cycle Toolkit, 1998

Data Warehousing

Definitions

CONSENSUS DEFINITIONS

Multiple and sometimes conflicting definitions of data warehousing terms do exist. Still, there is some consensus—or at least intended consensus—among the varied definitions. Common themes are: integrated, subject-oriented, time-variant, nonvolatile, accessible, meets business information needs, and a process that turns data into information.

INTEGRATED

Warehousing provides a single comprehensive source of information for and about the business. Answering a business question does not require accessing multiple sources across a variety of technology platforms with potentially inconsistent data.

SUBJECT-ORIENTED

Data and information are organized and presented as business subjects aligned with information needs, not as computer files designed for transactional processing needs.

TIME-VARIANT

The warehouse contains a history of the business, as well as relatively current business information. Structures and intervals are kept consistent across time, allowing time-specific analytics such as trend analysis.

NON-VOLATILE

The warehouse provides stable information. Business data, once written to the warehouse, is not overwritten. The body of data grows through regular addition of new data in a way that maintains accurate historical records.

ACCESSIBLE

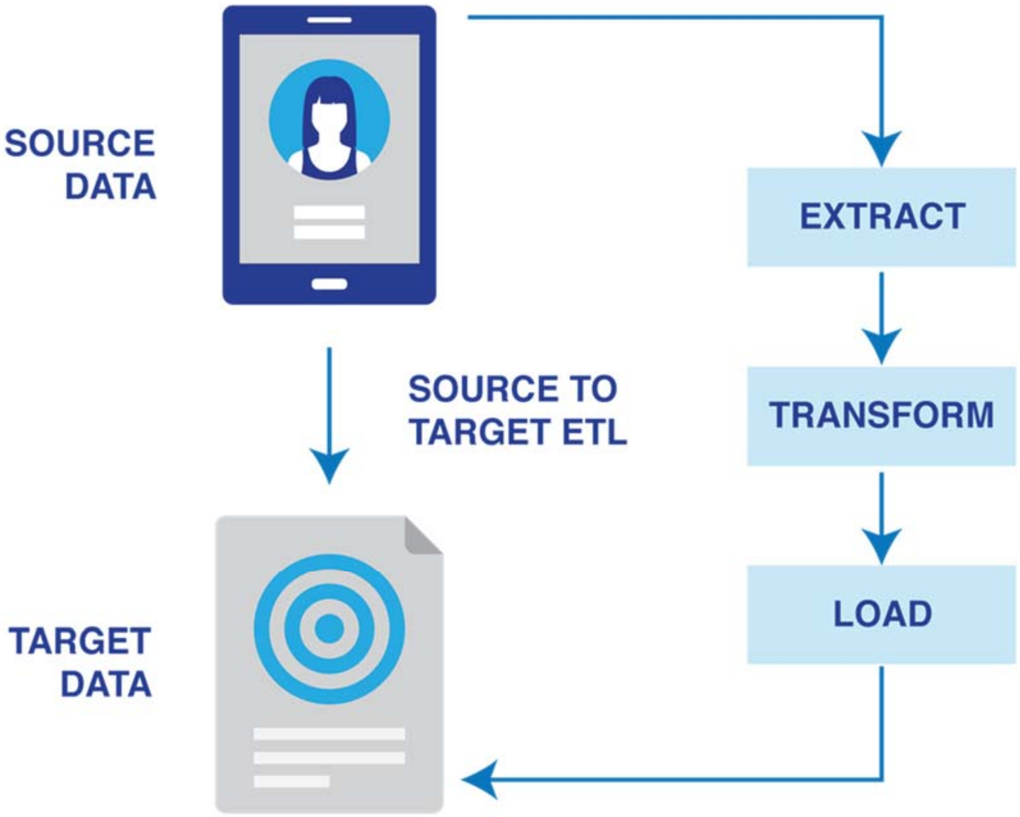
The primary purpose of a data warehouse is to provide readily accessible information to business people. The data is organized for easy access.

MEETS BUSINESS INFORMATION NEEDS

Warehousing provides an organized data resource against which a variety of standard tools can be applied by business knowledge workers to manipulate, analyze, and generate answers to business questions and support decision making. It is foundational to business intelligence and analytics.

Data Warehousing

Applied Data Integration



Data Warehousing

Applied Data Integration

THE DATA WAREHOUSING SYSTEM

The data warehousing system begins with data sources—initial sources are primarily operational data of the business—and transforms the data into valuable information.

THE BUILDING BLOCKS

Data warehousing processes migrate data from original sources to readily accessible information through a sequence of activities commonly known as ETL—extraction, transformation, and loading.

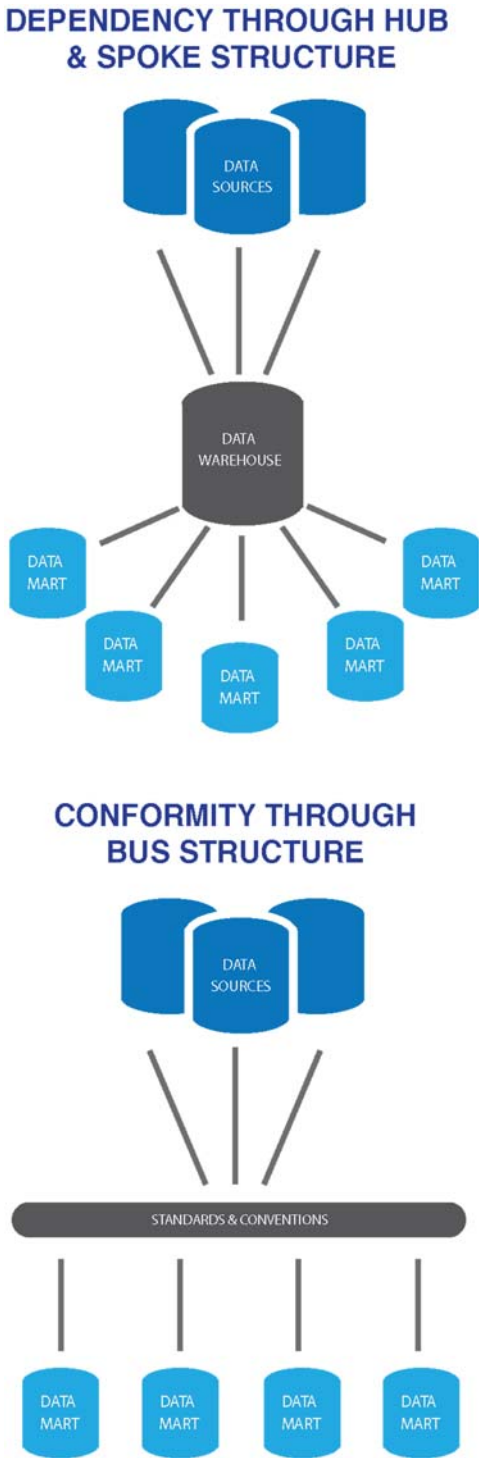
- *Extraction* is the process of acquiring a copy of data from a source. The extraction step may capture all data or only that data which has changed since a previous extraction. Extraction may be performed in a variety of ways, ranging from sequential batch processing of the source to transaction logging and data replication.

- *Transformation* changes the nature of the data. The purpose of data transformation is to imbue the data with the qualities desirable in a data warehouse—making it integrated, subject-oriented, non-volatile, time-variant, and accessible.

- *Loading* is the step that populates tables in warehousing databases. Appending new rows to existing tables is a common form of database loading. Other techniques include dropping or truncating and rebuilding tables and deleting and replacing changed rows. Directly updating columns of business data in warehousing tables is generally not a good practice.

Data Warehousing

Data Warehouse Architecture



DEPENDENCY THROUGH HUB & SPOKE STRUCTURE

CONFORMITY THROUGH BUS STRUCTURE

Data Warehousing

Data Warehouse Architecture

INTEGRATION DEBATES

Integration is one of the fundamental concepts of data warehousing. Inmon's early definition of a data warehouse identifies "integrated" as one of the four defining characteristics of a data warehouse. Integration discussions center around how to achieve that integration, with particular focus on data marts:

- Should data marts **depend** on a single integrated source?
- Should data marts **conform** to integration standards?
- Should data marts be **independent** in process and structure?

HUB-AND-SPOKE DEPENDENCY

Dependent data marts are architected to be populated from a single integrated data source—typically a data warehouse. When designing dependent data marts, integration is the responsibility of the data warehouse. The warehouse serves as an "integration hub" in a hub-and-spoke relationship between warehouse and data marts. Common arguments for and against dependency are shown below:

PROS	CONS
• Integration work is done only once. • Integration is consistent across all marts. • Warehouse data is reusable by many marts.	• Too much up-front analysis and modeling. • Takes too long to design. • Warehouse grows large and hard to maintain.

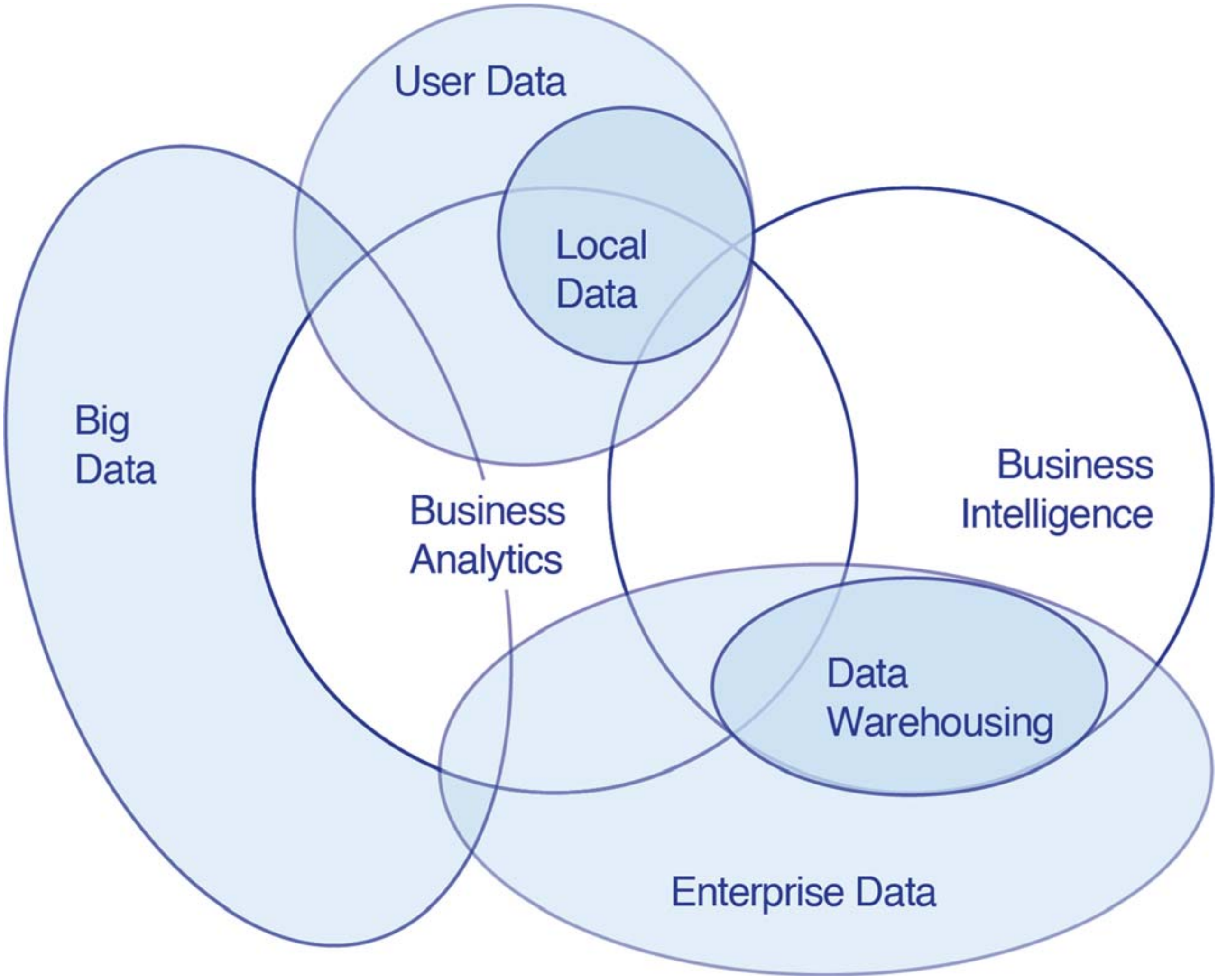
INTEGRATION BUS CONFORMITY

Conformed data marts are populated directly from original sources and achieve the goals of integration by complying with established data definition and data structure standards. Conforming facts and dimensions are the interface standards to support mart-to-source and mart-to-mart consistency in a bus architecture. The common arguments are:

PROS	CONS
• Integration is done where business needs it. • Doesn't require lots of up-front modeling. • You can achieve it incrementally.	• Too easy to get around standards. • Standards don't have enterprise view. • May have to rework existing marts.

Data Stores

Diversity of Data Sources



Data Stores

Diversity of Sources

DATA ARCHITECTURE

Today's enterprise information architecture comprises multiple categories of information. Enterprise information assets include repositories for business analytics and business intelligence, as well as user data, local data, and various big data sources.

Data stores in the BI and analytics environment may include:

- Data staging
- Data lakes
- Analytics sandboxes
- Data warehouses
- Data marts
- Master data repositories

Additional repositories may be necessary to perform utility functions, such as data standardization, quality assessment, and data cleansing. These topics will be discussed in *Module 5: Data Management.*

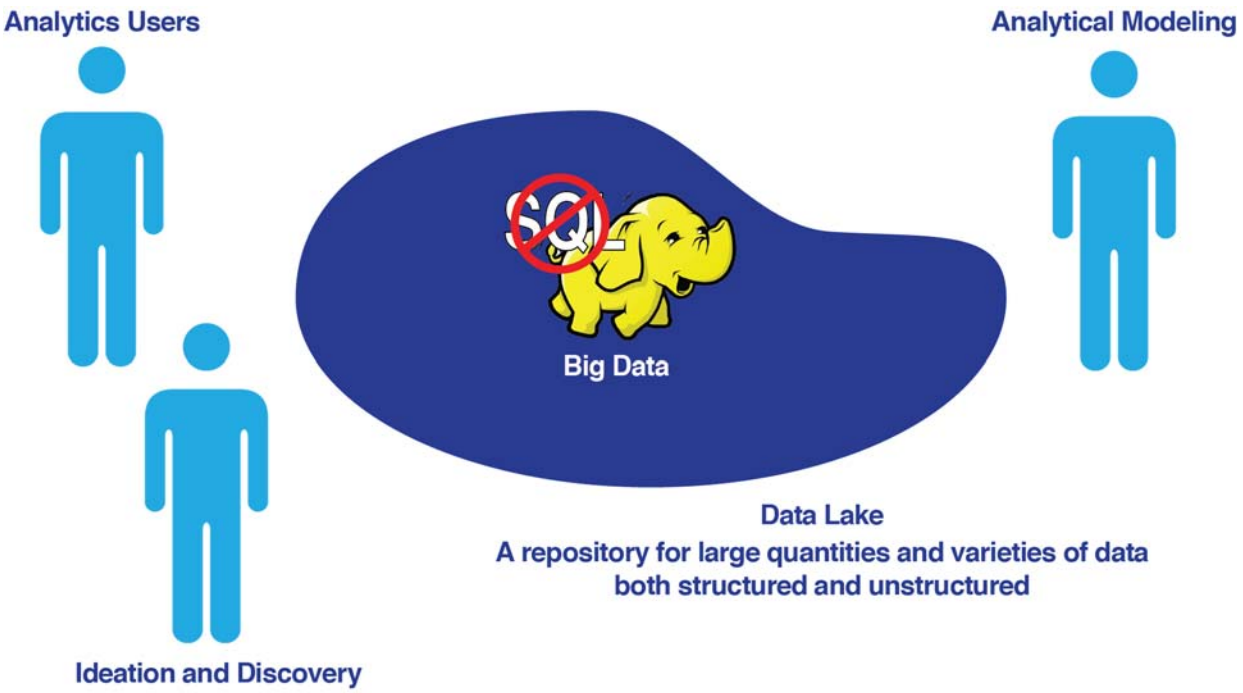
IMPACT OF BIG DATA

Big data has expanded the number and types of data sources that can be used to enrich the analytics process. Including data from Web searches, online shopping, email, text messaging, social media activity, machine-to-machine communications, sensor data, and much more expands analysis opportunities. Big data is not a new or standalone silo of data.

In this environment, integration tasks may involve integrating big data with master data to achieve more robust views of master data entities—a true 360-degree view of customers, for example. It may also involve integrating big data with a data warehouse. For both instances—master data and warehouse data—it is practical (especially with data virtualization technology) to create row/column views of unstructured data. Data integration becomes easier if both unstructured and structured data are viewed as rows and columns.

Data Stores

The Data Lake


Analytics Users
Analytical Modeling
SQL
Big Data
Data Lake
A repository for large quantities and varieties of data both structured and unstructured
Ideation and Discovery

Data Stores

The Data Lake

FIRST STOP FOR NEW DATA

A *data lake* is defined as a large repository for large quantities and varieties of data, both structured and unstructured[1]. The data lake is used to ingest data from various sources and can maintain the original format of the data.

For many potential data sets, not much is known up front about their structure or content. In a relational system, this would be a problem because a model is required before data can be added to the database. This paradigm is called *schema on write*. It is not convenient for new data sources—because development of a model requires detailed analysis.

Data lakes typically leverage nonrelational storage. These NoSQL databases do not require a model in order to store data. This is convenient because new sources are not understood. However, by the time information is accessed programmatically, the data formats must be understood by developers. This paradigm is called *schema on read*.

Data lakes use commodity cluster computing techniques and scalable, low-cost storage. Hadoop is the primary approach used with data lakes for a flexible, cost-effective data processing model which can scale easily as data volumes grow. Hadoop is a distributed file system designed to work with large data collections that combine structured data with more complex types of data. MapReduce is a programming framework to write applications that work with Hadoop data sets.

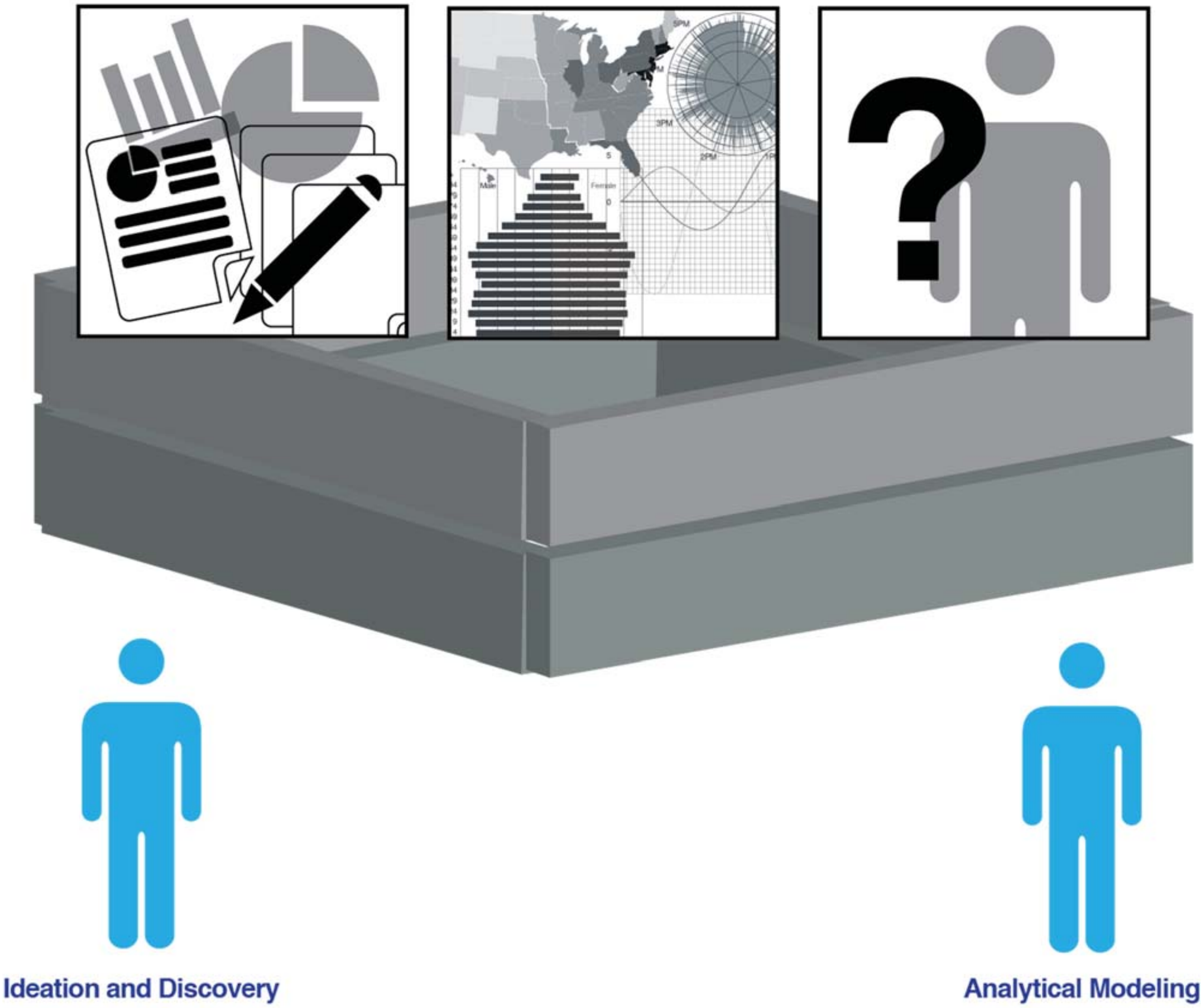
OTHER USES

Data lakes have many uses beyond intake of new data sources. These including supporting exploration, modeling, and analytics. Data lakes may also be used as a staging area for a data warehouse or a data mart.

[1] *Data lakes: An emerging approach to cloud-based big data* http://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/features/data-lakes.html, Curran.

Data Stores

Analytics Sandboxes



Data Stores

Analytics Sandboxes

ANALYTICS PURPOSE AND DISCOVERY

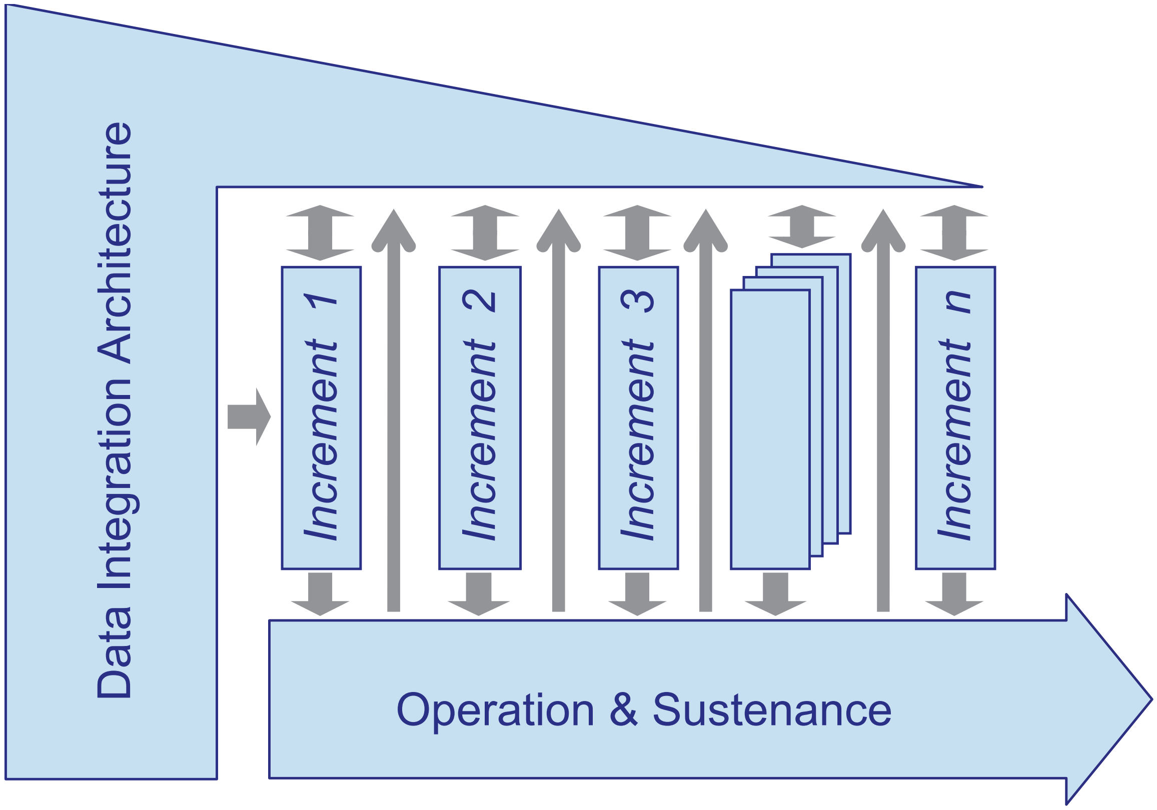
The analytics sandbox is a space where multiple data sets can be combined and explored, supporting the processes of discovery and analytics. A sandbox may be a part of the data lake, or it may be a separate data store.

A key objective of the analytics sandbox is to test a variety of hypotheses about data and analytics. Analytics sandboxes are often referred to as discovery environments and are typically “non-production” environments where users have the ability to ingest, store, process, and model data as needed to support the discovery process. Analytical modeling, ideation, root cause analysis, and exploratory data analysis are some of the activities typically completed in an analytics sandbox.

The scope of data integration in the sandbox will depend on the scope of data discovery or analytics purpose. Data integration can be one-time or temporary until the analytics objective is achieved. As a result of the process, data sources may be evaluated for value and long-term use. Additionally, data integration, transformation, and cleansing rules may be discovered.

Implementation

Process



Data Integration Architecture

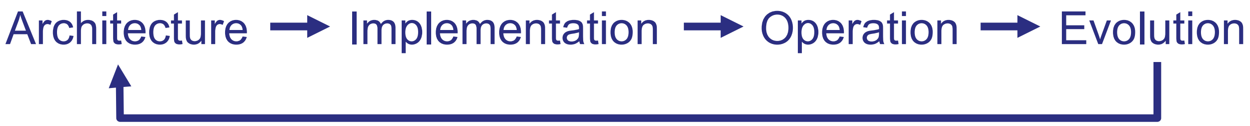
Increment 1

Increment 2

Increment 3

Increment n

Operation & Sustenance



Architecture → Implementation → Operation → Evolution

Implementation

Process

DATA ARCHITECTURE

Data architecture establishes the framework, standards, and procedures for data stores at an enterprise level. Architecture is an early-stage design task that takes place prior to the design and development of any specific data store.

INCREMENTAL CONSTRUCTION

Data stores are typically added to the environment incrementally. Each increment may be a new capability, such as a data lake, or an instance of a data store that supports a new application, such as an analytics sandbox or a data mart.

Incremental implementation is a pragmatic approach to building an enterprise-level BI and analytics environment in a segmented, evolutionary fashion. The incremental development approach is, by nature, evolutionary. The first increment delivers a subset of the data environment that meets a limited set of information needs. As each increment is added, the environment becomes more complete and is able to meet a larger set of information needs.

OPERATION

Operational tasks attend to administrative issues, such as:

- Ensuring that refreshes occur
- Monitoring, managing, and tuning
- Responding to change
- Delivering data and information services

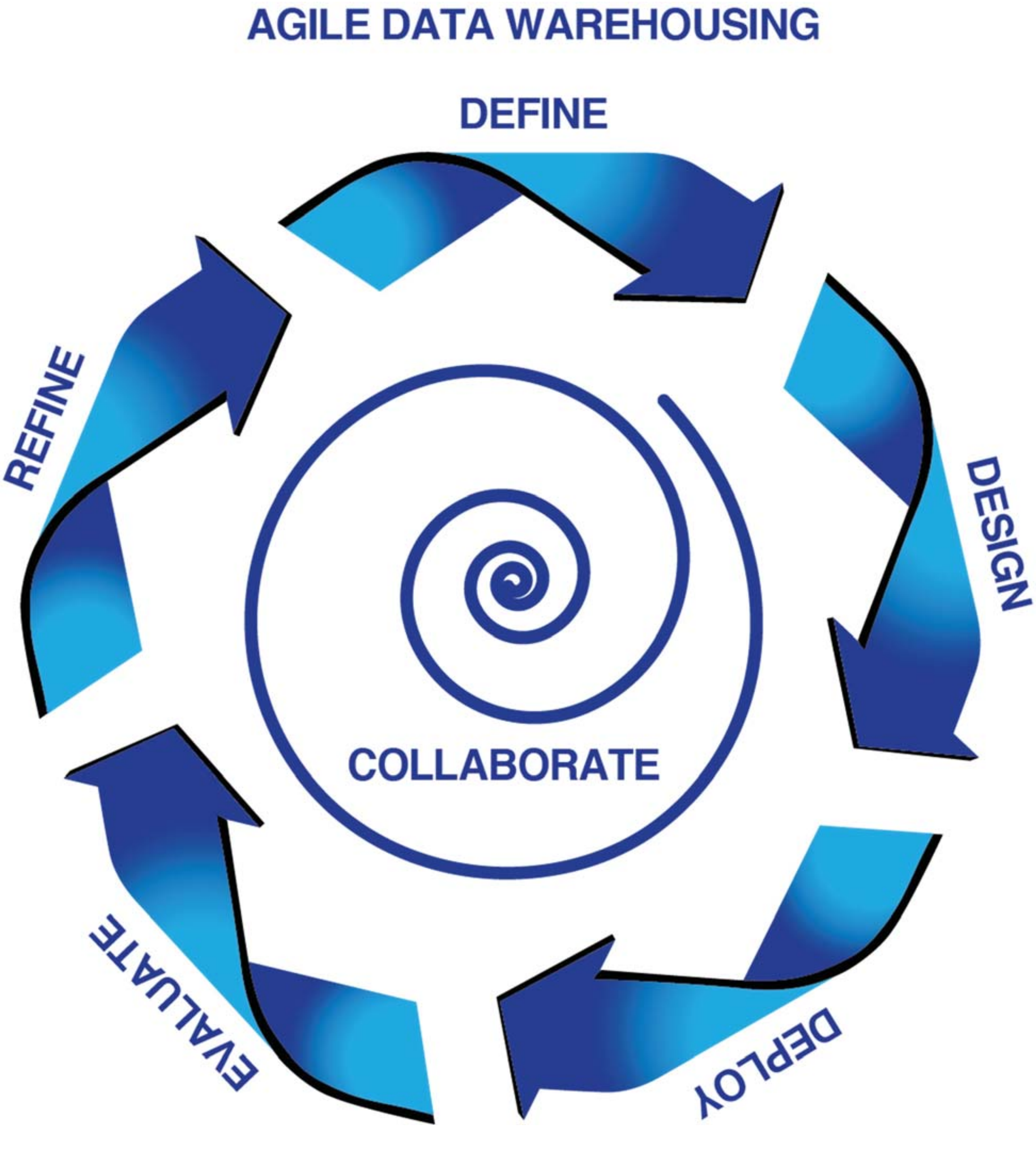
Although these concepts apply to all components of the data architecture for BI and analytics, the operation phase is often referred to as warehouse operation.

EVOLVING ARCHITECTURE

Incremental development also offers an opportunity to learn and to minimize the impact of mistakes. It is improbable that anyone will develop the “perfect” data architecture before building a first increment. Both incremental development activities and those of operating the environment provide valuable feedback that helps to refine the architecture.

Implementation

Agile Development



Implementation

Agile Development

WHAT IS AGILE?

If you look in the dictionary you'll find that agile means "moving quickly and lightly" or "mentally quick." Synonyms include nimble and spry. Yet the term has quickly taken on particular meaning in systems and software development. Agile development has become synonymous with concepts of business collaboration and test-driven development.

THE AGILE MANIFESTO

Several of the pioneers and leading practitioners of agile development (too many to list here) collectively authored the following *Manifesto for Agile Software Development.* Following these principles will make us quick and nimble in systems development whether called "agile" or not.

"We are uncovering better ways of developing software by doing it and helping others do it. Through this work we have come to value:

- *Individuals and interactions over processes and tools*
- *Working software over comprehensive documentation*
- *Customer collaboration over contract negotiation*
- *Responding to change over following a plan*

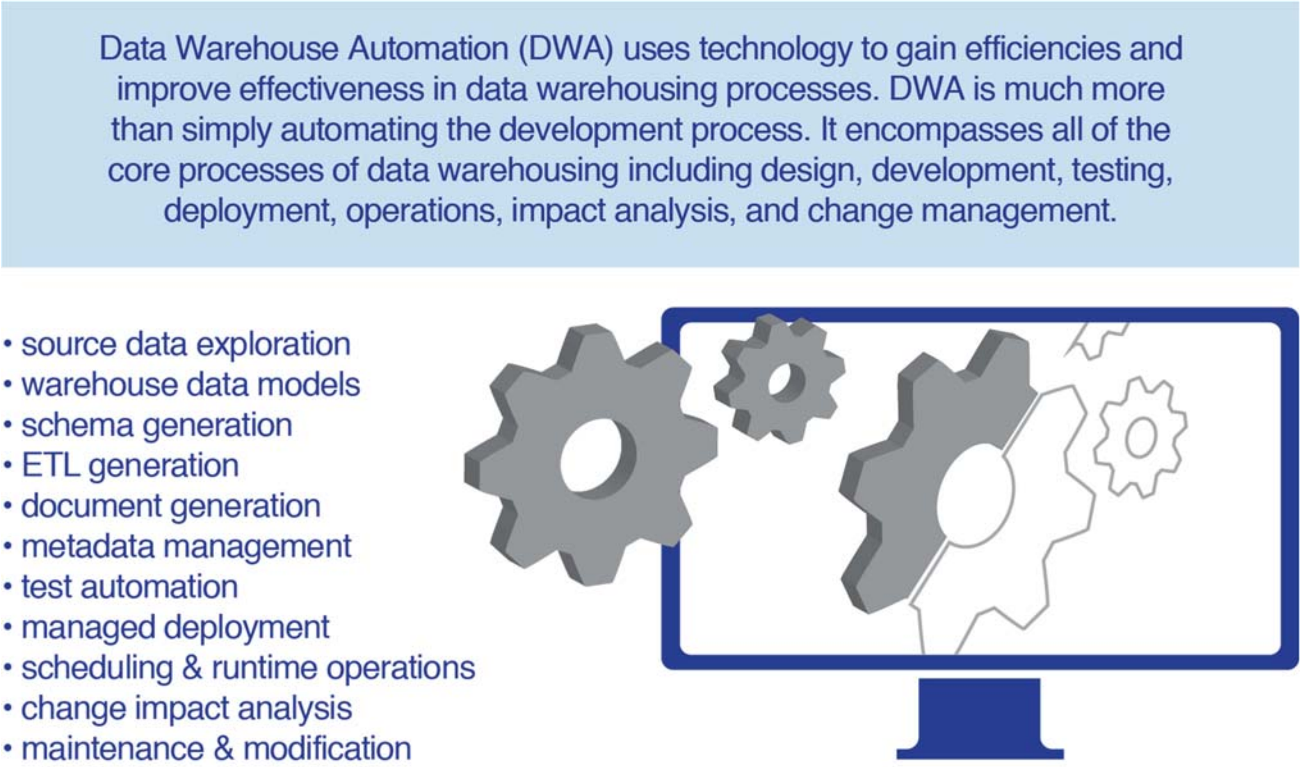
That is, while there is value in the items on the right, we value the items on the left more." (http://agilemanifesto.org)

AGILE DEVELOPMENT

Agile business intelligence and analytics development applies the concepts and practices of agile development to BI and analytics implementation. Collaboration with business and data subject experts, combined with rapidly repeated test-driven development is used to understand source data, map source data to target data, discover and apply data transformation logic, evolve the data model, etc.

Implementation

Data Warehouse Automation



Data Warehouse Automation (DWA) uses technology to gain efficiencies and improve effectiveness in data warehousing processes. DWA is much more than simply automating the development process. It encompasses all of the core processes of data warehousing including design, development, testing, deployment, operations, impact analysis, and change management.

- source data exploration
- warehouse data models
- schema generation
- ETL generation
- document generation
- metadata management
- test automation
- managed deployment
- scheduling & runtime operations
- change impact analysis
- maintenance & modification

Implementation

Data Warehouse Automation

DATA WAREHOUSE IN CONTEXT

The data warehouse provides a major portion of the foundation for business intelligence and analytics. Traditionally, it is developed through a series of manual steps. There are tools that can be used to reduce the manual effort and quicken the process.

DATA WAREHOUSE AUTOMATION DEFINED

Data warehouse automation is more than simply automation of ETL development or even automation of the entire development process. It encompasses the entire data warehousing life cycle from planning, analysis, and design through development, extending into operations, maintenance, and change management.

CHANGING BEST PRACTICES

Adoption of data warehouse automation changes the way that we think about building data warehouses. The widely accepted best practice of extensive up-front analysis, design, and modeling can be left behind as the mindset changes from “get it right the first time” to “develop fast and develop frequently”—an approach that is aligned with today’s agile development practices.

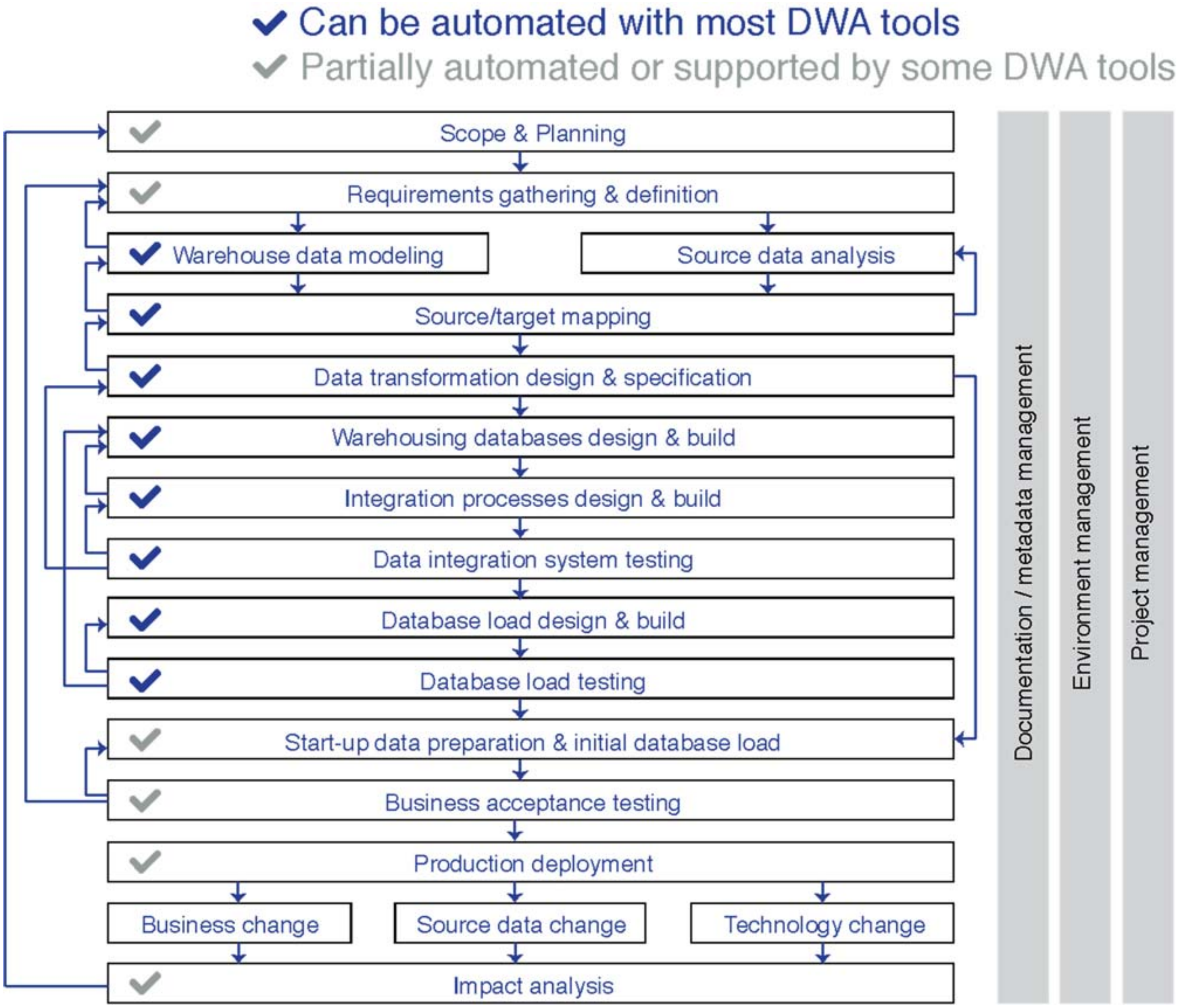
BENEFITS

Automation in data warehousing has many of the same benefits as in manufacturing:

- Increased productivity and speed of production
- Reduction of manual effort
- Improved quality and consistency
- Better controls and process optimization opportunities

Implementation

What Can You Automate?


✓ Can be automated with most DWA tools
✓ Partially automated or supported by some DWA tools
Scope & Planning
Requirements gathering & definition
Warehouse data modeling
Source data analysis
Source/target mapping
Data transformation design & specification
Warehousing databases design & build
Integration processes design & build
Data integration system testing
Database load design & build
Database load testing
Start-up data preparation & initial database load
Business acceptance testing
Production deployment
Business change
Source data change
Technology change
Impact analysis
Documentation / metadata management
Environment management
Project management

Implementation

What Can You Automate?

LABOR INTENSITY

A typical data warehouse development process is burdened by labor-intensive activities, a high degree of complexity and dependency among deliverables, and volatility of data and requirements. Some of the data warehousing pain that can be relieved by automation includes:

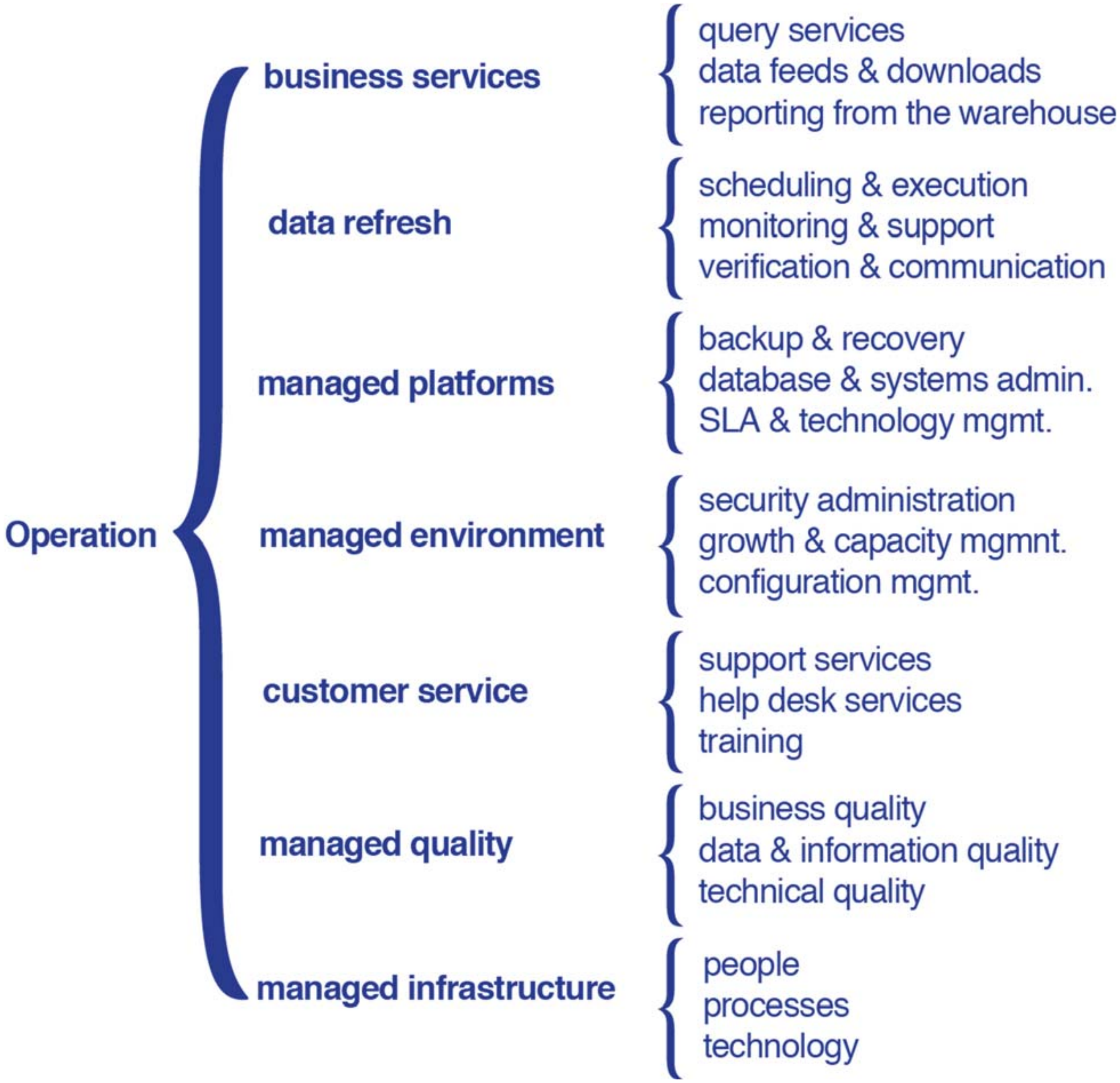
- Slow, laborious, and difficult to get right requirement gathering
- Interdependency between source data analysis and warehouse data modeling
- Manually mapping sources to targets
- Detailed specification of data transformation logic
- Translating data models into schema and building databases
- Hand coding and manual testing of ETL processing
- Preparing and loading initial start-up data
- Acceptance testing and deployment to production

USING AUTOMATION

Data warehouse automation is capable of relieving much of the pain. Nearly everything in the diagram on the facing page can be automated. All automation tools support core activities such as data modeling and data integration. The most robust automation products support all of the activities.

Operation

Components



- Operation
  - business services
    - query services
    - data feeds & downloads
    - reporting from the warehouse
  - data refresh
    - scheduling & execution
    - monitoring & support
    - verification & communication
  - managed platforms
    - backup & recovery
    - database & systems admin.
    - SLA & technology mgmt.
  - managed environment
    - security administration
    - growth & capacity mgmnt.
    - configuration mgmt.
  - customer service
    - support services
    - help desk services
    - training
  - managed quality
    - business quality
    - data & information quality
    - technical quality
  - managed infrastructure
    - people
    - processes
    - technology

Operation

Components

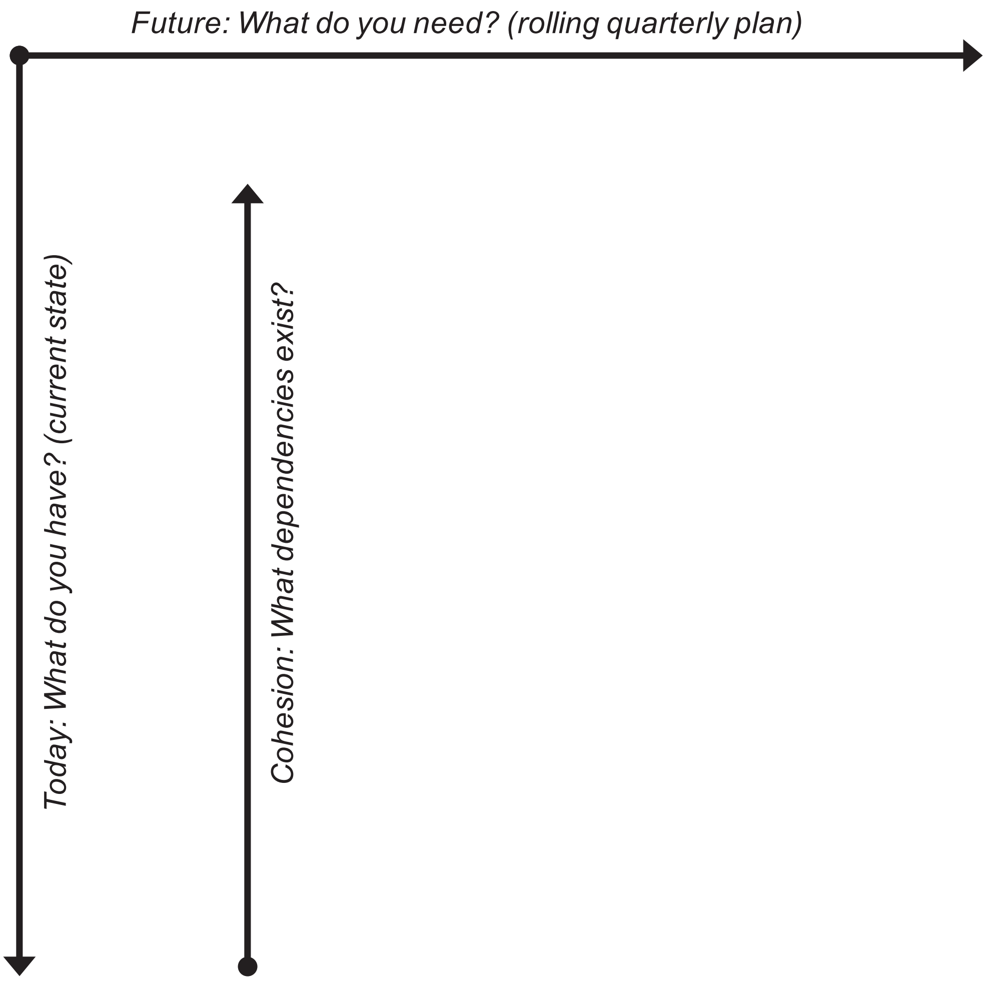
DAY-TO-DAY MANAGEMENT

Managing day-to-day operation of the ecosystem involves many responsibilities.

- Business services—Sustaining the data warehouse demands a commitment to delivering reliable and valuable business services in an environment of high-frequency change.
- Data refresh—The first and most frequent responsibility in data warehouse operation is that of regular data refresh. It is essential to get new data into the warehouse and marts on a timely basis and as expected by the business.
- Managed platforms—Platform management includes backup and recovery responsibility, DBMS maintenance and support, server and network support, performance monitoring and tuning, and technology administration.
- Managed environment—Security administration assures that warehousing data is accessed in accordance with security and privacy policies for the data. Capacity management includes monitoring growth, measuring capacities, and forecasting resource needs. Configuration management assures platform reliability in an environment of frequent technology change.
- Customer service and support—Customer support services may be as unique and varied as warehouse users and their information needs. These services aid business people in getting value from the data warehouse, covering the entire range of assistance needs from determining requirements to understanding and applying information received.
- Managed quality—Warehouse quality is much more than simple data quality. Business quality directly affects the business value derived from the data warehouse. Information quality is based on trust in the data and belief in its value. Technical quality affects reliability, availability, and performance of the warehouse.
- Managed infrastructure—Putting it all together, the people, processes, and technology must be managed in a cohesive way.

The BI and Analytics Roadmap

Data Integration


Future: What do you need? (rolling quarterly plan)
Today: What do you have? (current state)
Cohesion: What dependencies exist?

The BI and Analytics Roadmap

Data Integration

MAPPING DATA SYSTEMS

The facing page adds the parts of the BI and analytics roadmap that relate to data integration. Review the current state inventory and identify future state requirements for the items listed below. Here you're identifying needs for architectural evolution, new data sources, and new or modified data integration systems and data stores.

- Integration architecture components
  - Integration requirements
  - Integration techniques
- Data sources
  - Internal and external
  - Enterprise, departmental, user
  - Database, log, machine-generated
  - Relational, NoSQL
  - Structured, unstructured, semistructured
- BI and analytics data stores
  - Data lake
  - Data warehouse
  - Data mart
  - Analytics sandbox
- Data integration lifecycle
  - Architecture
  - Implementation
  - Operation

PROJECTS

What kind of projects are required to reach your future state? Map these into your roadmap, being sure to identify dependencies with other systems and planned activities. Projects may include:

- Develop new systems
- Extend / enhance
- Maintain / modify
- Decommission / retire

Be specific when adding projects to the roadmap. For example, list *build analytics sandbox for fraud detection*, not *create sandbox*.

Mistakes to Avoid

When Using Data Federation

- FAILING TO UNDERSTAND THE ROLE OF DATA FEDERATION IN DATA MANAGEMENT ARCHITECTURE
- USING DATA FEDERATION TO CREATE A VIRTUAL DATA WAREHOUSE
- USING DATA FEDERATION WITH POOR-QUALITY, HIGHLY COMPLEX DATA SOURCES
- NOT USING DATA FEDERATION FOR PROTOTYPING
- NOT MONITORING THE IMPACT OF DATA FEDERATION ON SOURCE SYSTEMS
- USING DATA FEDERATION FOR REAL-TIME ANALYTICS
- USING DATA FEDERATION ONLY WITH RELATIONAL DATA
- FAILING TO EVALUATE THE ARCHITECTURE AND FEATURES OF A VENDOR’S DATA FEDERATION SOLUTION
- FAILING TO DETERMINE DATA FEDERATION INFRASTRUCTURE REQUIREMENTS
- MISSING THE OPPORTUNITY TO DEFINE A SHARED BUSINESS VIEW OF SOURCE DATA

Source: *Ten Mistakes to Avoid When Using Data Federation Technology* by Claudia Imhoff and Colin White.

Mistakes to Avoid

When Using Data Federation

TEN MISTAKES TO AVOID FOR FEDERATION AND VIRTUALIZATION

Ten Mistakes to Avoid When Using Data Federation Technology by Claudia Imhoff and Colin White.

Data federation is a relatively new form of data integration, but it has achieved a significant role in today's data management strategies. Data federation allows data integration teams to quickly create virtually integrated sets of data for many purposes—business intelligence (BI), customer relationship management (CRM), master data management (MDM), and so on. It can combine data without having to actually move the data from the original sources, greatly accelerating the data integration process.

This "data virtualization" has proven to be a boon to data integration, but it does have its limits. There are many situations where data federation is not a good idea; the difficulty comes in understanding these situations so you can determine the best times to use data federation.

Mistakes to Avoid

In Your Big Data Implementation

- LACK OF BUSINESS CASE
- DATA THAT LACKS RELEVANCE
- LACK OF DATA QUALITY
- INSUFFICIENT DATA GRANULARITY
- MISSING DATA CONTEXTUALIZATION
- NOT UNDERSTANDING DATA COMPLEXITY
- POOR DATA PREPARATION
- ORGANIZATIONAL IMMATURITY
- LACK OF DATA GOVERNANCE
- BELIEVING TECHNOLOGY IS A SILVER BULLET

Source: *Ten Mistakes to Avoid in Your Big Data Implementation* by Krish Krishnan.

Mistakes to Avoid

In Your Big Data Implementation

TEN MISTAKES TO AVOID IN YOUR BIG DATA IMPLEMENTATION

Ten Mistakes to Avoid in Your Big Data Implementation by Krish Krishnan.

Big data is the biggest buzzword in the industry today. Every organization—big or small—is looking into understanding and deploying a big data program. Big data doesn't just refer to having larger volumes of data. We must consider the source(s) of the data.

One purpose of a big data implementation is to incorporate additional data sets into the current data infrastructure to help the enterprise answer more types of questions with its data. Although the possibility of accomplishing this goal seems realistic with the evolution of technology and commoditization of an enterprise's infrastructure, there are several critical pitfalls to avoid. The facing page lists the most common mistakes that occur when implementing a big data program.

Discussion

Data Integration

LET’S TALK ABOUT IT!

- Does your organization have a formal data architecture for BI and analytics? A formal integration architecture?
- What techniques are you using for data integration?
- How is your organization currently leveraging big data sources?
- Does your architecture include a data lake? What is its function?
- How are analytics sandboxes implemented in your environment?
- Do you employ an iterative process for the introduction of new analytics data stores?

Discussion

Data Integration

Notes:

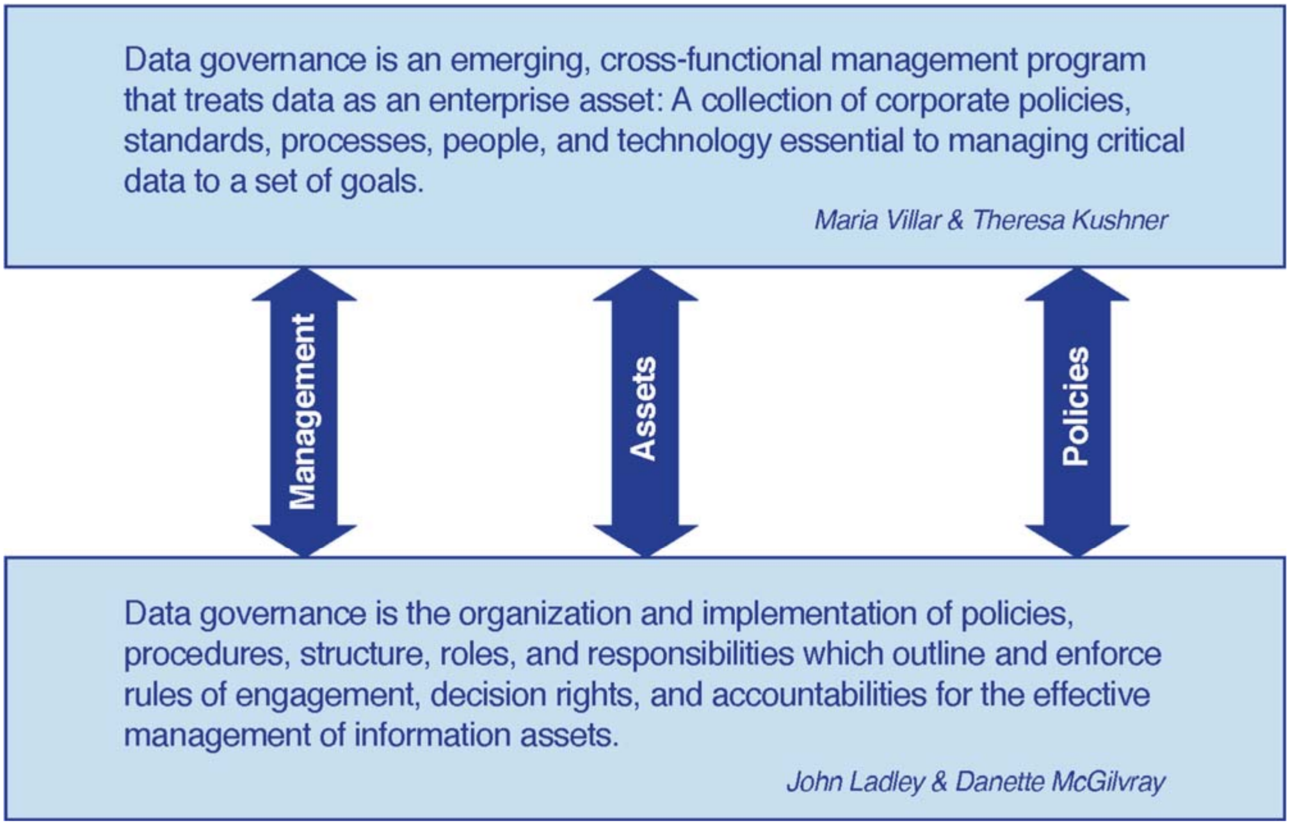
Transforming Data With Intelligence™

Module 5

Data Management

Data Governance

Data Governance Concepts



Data governance is an emerging, cross-functional management program that treats data as an enterprise asset: A collection of corporate policies, standards, processes, people, and technology essential to managing critical data to a set of goals.

Maria Villar & Theresa Kushner

Data governance is the organization and implementation of policies, procedures, structure, roles, and responsibilities which outline and enforce rules of engagement, decision rights, and accountabilities for the effective management of information assets.

John Ladley & Danette McGilvray

Data Governance

Data Governance Concepts

MANAGING A VALUABLE ASSET

Maria Villar and Theresa Kushner define *data governance* as a "management program that treats data as an enterprise asset: a collection of corporate policies, standards, processes, people, and technology …"[1]

John Ladley and Danette McGilvray define data governance as "policies, procedures, structure, roles and responsibilities which outline and enforce rules of engagement, decision rights, and accountabilities for the effective management of information assets."[2]

Both definitions are informative and thought provoking. Considering them together provides a good sense of the core of data governance as an asset management practice with attention to data-related policies.

ADDITIONAL PERSPECTIVE

No single standard definition exists, but several other information management practitioners define data governance in a variety of ways that add depth to the overall understanding of the subject:

Gwen Thomas defines it as "execution and enforcement of authority over the management of data and data-related processes."[3]

Alex Berson and Larry Dubov define it as "a process focused on managing the quality, consistency, usability, security, and availability of information."[4]

David Loshin describes data governance as "a program for defining information policies that relate to the constraints of the business …" [5]

ASSETS & VALUE MANAGEMENT

Jonathan Geiger says, "data governance recognizes that data is an important enterprise asset and applies the same rigor to managing this asset is it does for any other asset."[6] When considering asset management for financial or property assets it is generally true that value is a key consideration. It follows, then, that data value is key in data governance.

[1] Source: *Data Governance Fundamentals,* www.elearningcurve.com. Villar & Kushner are authors of *Managing Your Business Data*.

[2] Source: *Executing Data Quality Projects.* McGilvray is the author; Ladley is a well-known EIM consultant.

[3] Source: *Data Governance Defined* by John Ladley (www.enterprisedatajournal.com/article/data-governance-defined.html).

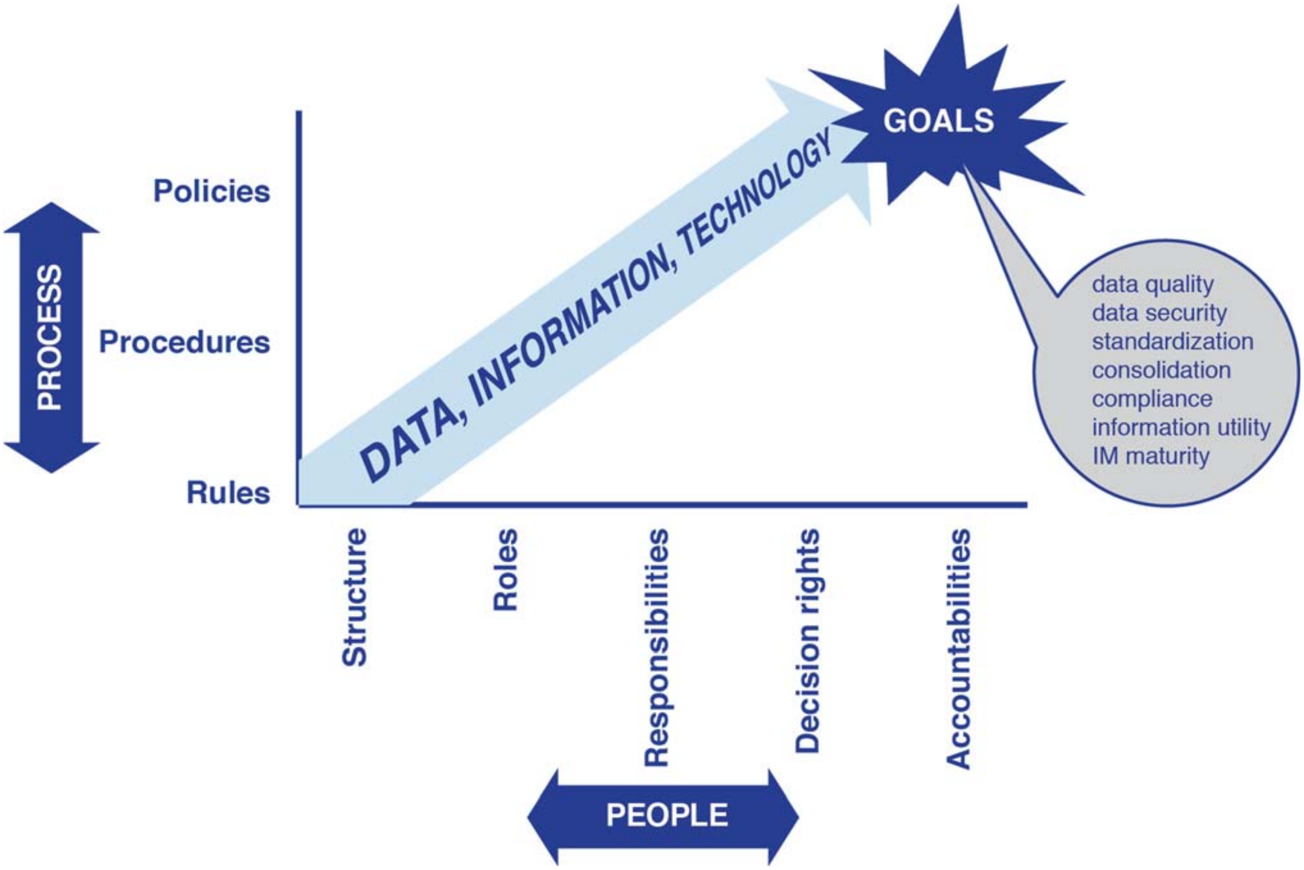
[4] Source: *Master Data Management and Customer Data Integration for a Global Enterprise* by Berson and Dubov.

[5] Source: *Master Data Management* by Loshin.

[6] Source: *Data Governance Defined* by John Ladley (www.enterprisedatajournal.com/article/data-governance-defined.html). Geiger is executive vice president of Intelligent Solutions and a well-known speaker and consultant.

Data Governance

Data Governance Concepts



Data Governance

Data Governance Concepts

GOVERNANCE DIMENSIONS

Data governance is a program of managing information assets to achieve defined information management goals. Governance establishes the processes that are needed and designates the responsibilities of people to achieve the goals.

The process dimension of data governance includes policies, procedures, and rules. The people dimension of data governance includes organizational structure, roles, responsibilities, decision rights, and accountabilities.

These dimensions create a management framework within which data and information are managed and technologies are employed to achieve specific information management goals.

GOVERNANCE GOALS

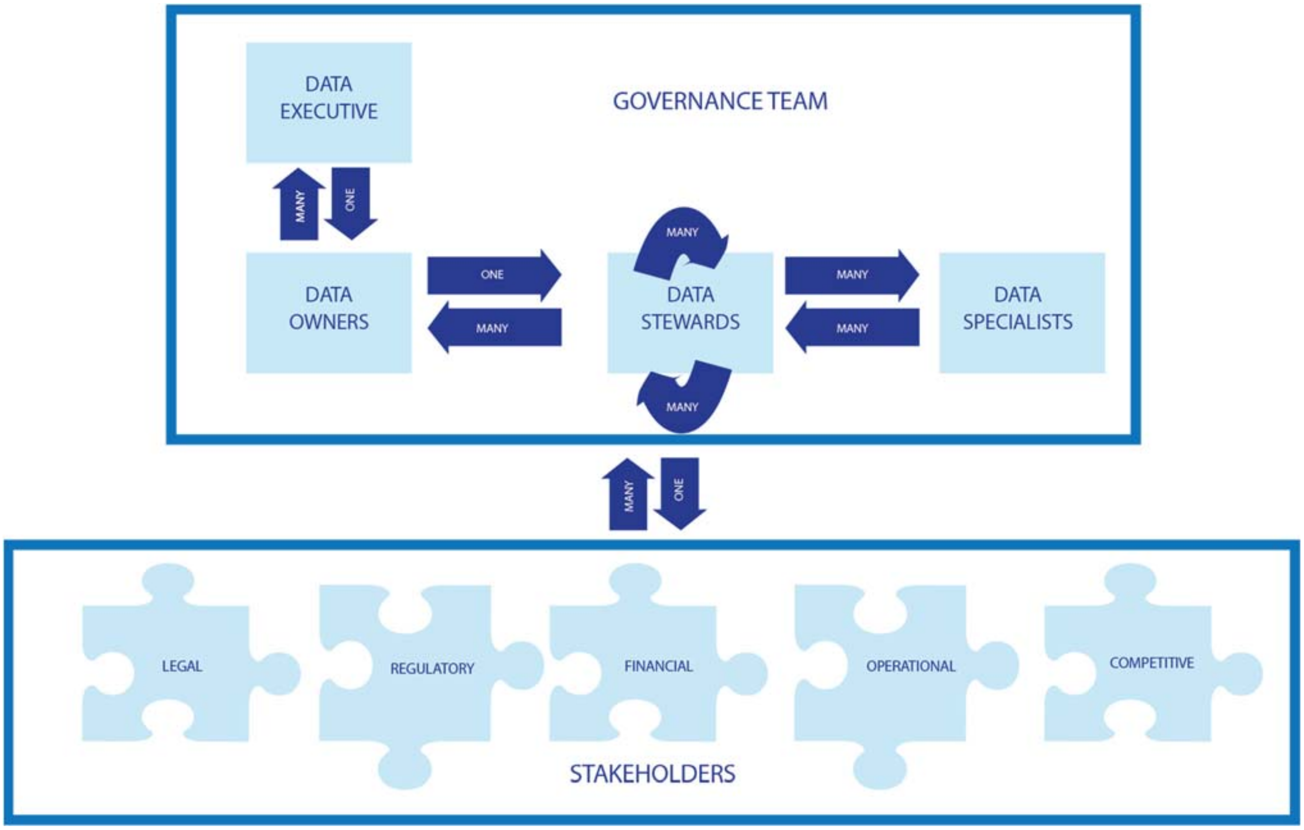
Goals are the driving force of data governance—the reasons to govern data and the foundations upon which governance processes are built. Common goals include such things as:

- data quality
- data security
- data standardization
- data consolidation
- regulatory compliance
- information utility
- information management maturity

As with any program, data governance goals are not static. They change over time as the business evolves and the governance program matures.

Data Governance

Data Governance Roles and Responsibilities



Data Governance

Data Governance Roles and Responsibilities

THE GOVERNANCE TEAM

A typical data governance organization includes four roles:

- A *data executive* designated as the person who provides overall leadership of a data governance program.
- *Data owners* responsible for access, distribution, retention, etc.
- *Data stewards* who facilitate consensus data definitions and foster sound data quality, data usage, and data security practices.
- *Data specialists* such as data architects, data modelers, database developers, and database administrators who have custodial responsibility for data.

THE DATA EXECUTIVE

The data executive role is handled differently in various organizations and data governance programs. In some instances the data executive is a C-level position. The CIO or the CFO may be designated as having executive responsibility for data. Security and compliance-driven programs may designate a senior security or compliance officer. In other instances a data governance program manager is created as a senior management position that provides a bridge for engagement of director and CxO positions. Whatever means makes sense for your organization's size, needs, governance goals, and culture, it is important that an executive perspective is included in the data governance team.

DATA OWNERS

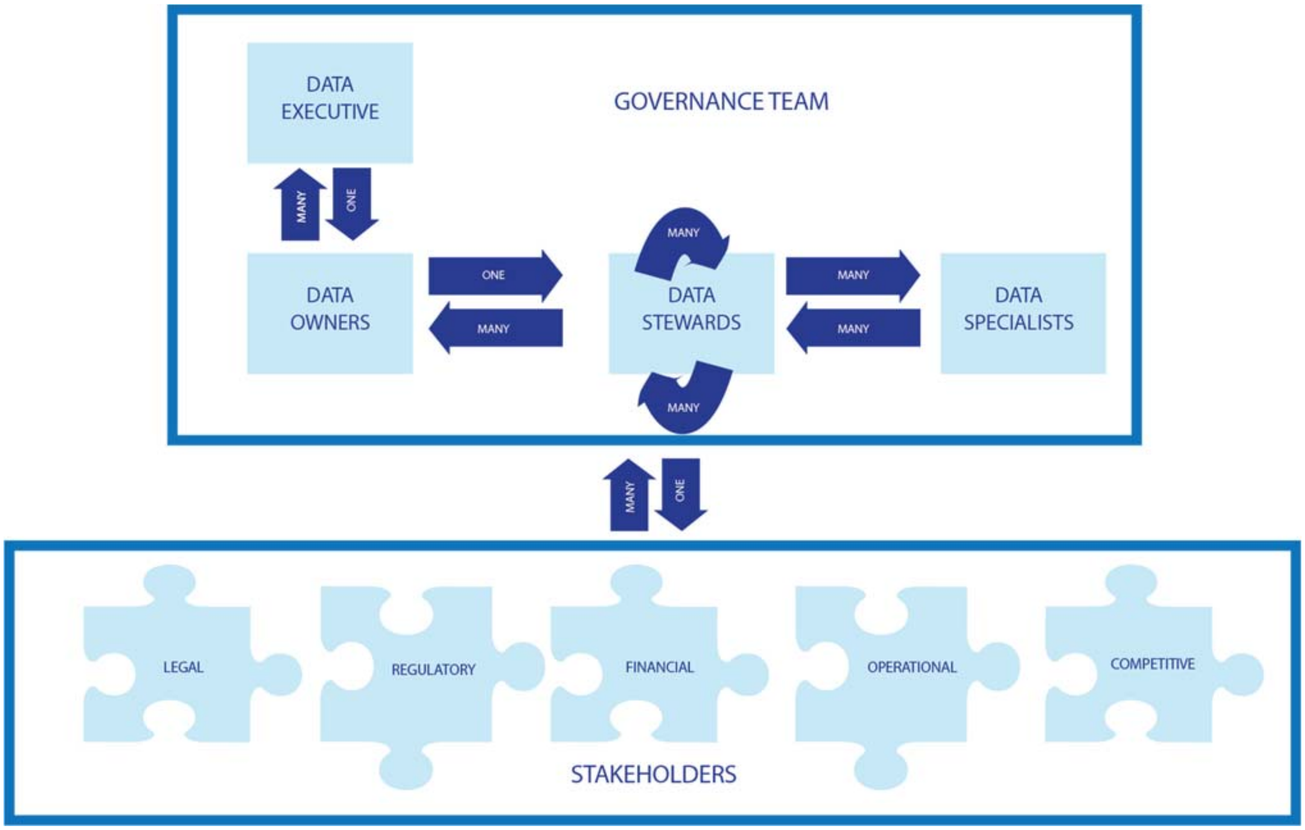
A data owner has responsibility for and related authority to make decisions about quality, access, distribution, retention, business definitions, etc. A good data owner:

- Knows the regulations, policies, and laws governing data privacy
- Knows the data
- Knows the business (and underlying business rules)
- Has a focused view of business objectives

Ideally data owners are senior managers from the business functional areas with the greatest dependency on the data. When a data domain crosses multiple business areas, designation of a single business owner can be based on a combination of knowledge, interest, and affinity of business area with all of the ownership responsibilities.

Data Governance

Data Governance Roles and Responsibilities (continued)



Data Governance

Data Governance Roles and Responsibilities (continued)

DATA STEWARDS

The data steward is responsible for facilitating consensus data definitions, for guiding data quality and usage practices, and for making recommendations about access, security, distribution, and retention. A good data steward:

- Knows the data and understands the business
- Has facilitation skills
- Takes a global view
- Communicates concerns about data quality

Typically stewards are from business organizations, not from IT groups. There are, however, many kinds of data stewards (discussed on the following page), and some are appropriate IT roles. An effective steward has a stake in the quality and utility of the data. The roles and responsibilities of data stewards should be formally assigned and communicated throughout the organization.

DATA SPECIALISTS (CUSTODIANS)

Data custodians include all of the various data specialists—architects, modelers, DBAs, etc.—who are responsible for housing the data; ensuring archiving, backup, and recoverability; and preventing corruption and/or loss. A good data custodian:

- Has the necessary technical knowledge and skills
- Understands good data management practices
- Is aware of regulations, policies, and standards governing data
- Understands and applies best practices for data quality and security

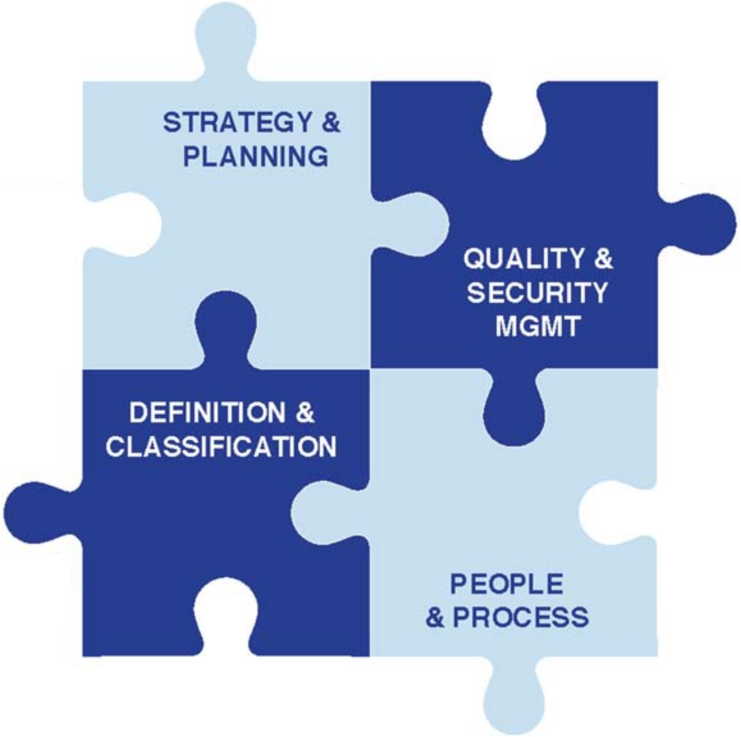
Ideally, data custodians are from IT organizations. They are typically from IT units responsible for managing IT infrastructure, including data storage, disaster recovery, and systems security.

Data Governance

Data Stewardship

- Business data requirements
- Priorities and roadmap
- Data-to-business alignment
- Data mgmt. policies and practices

- Quality goals and measures
- Security policies and regulations
- Quality and security monitoring
- Communication and education
- Detect root cause analysis



- Clear unambiguous definitions
- Applied data naming standards
- Consistency of use
- Declared system of record
- Known origins and uses

- Work with business and IT
- Teamwork among all stewards
- Facilitation and consensus building
- Core of data governance council

Data Governance

Data Stewardship

ROLE CRITICALITY

Data governance involves four roles—executive, ownership, stewardship, and custodianship. Stewardship, however, differs from the other roles in some significant ways. Data stewardship needs to be explored in greater depth because:

- Stewardship is a distinctly new and different role. Ownership recognizes and formalizes a set of responsibilities of business managers but does not redefine the job of business management. Similarly, custodianship recognizes and formalizes responsibilities of data specialists but doesn't redefine their jobs.

- Data stewardship is the nexus of data governance. It links owners of different but related data subjects, and it connects business rules and requirements with data models, database design, information systems implementation, and day-to-day management and administration of data.

SCOPE OF RESPONSIBILITY

Data stewardship responsibilities fit into four major categories:

- *Strategy and planning* identify business requirements for data and information, help to set priorities, and maintain a roadmap of activities to meet requirements. Continuously aligning data with business needs and data management policies and practices with business goals are primary strategy and planning objectives.

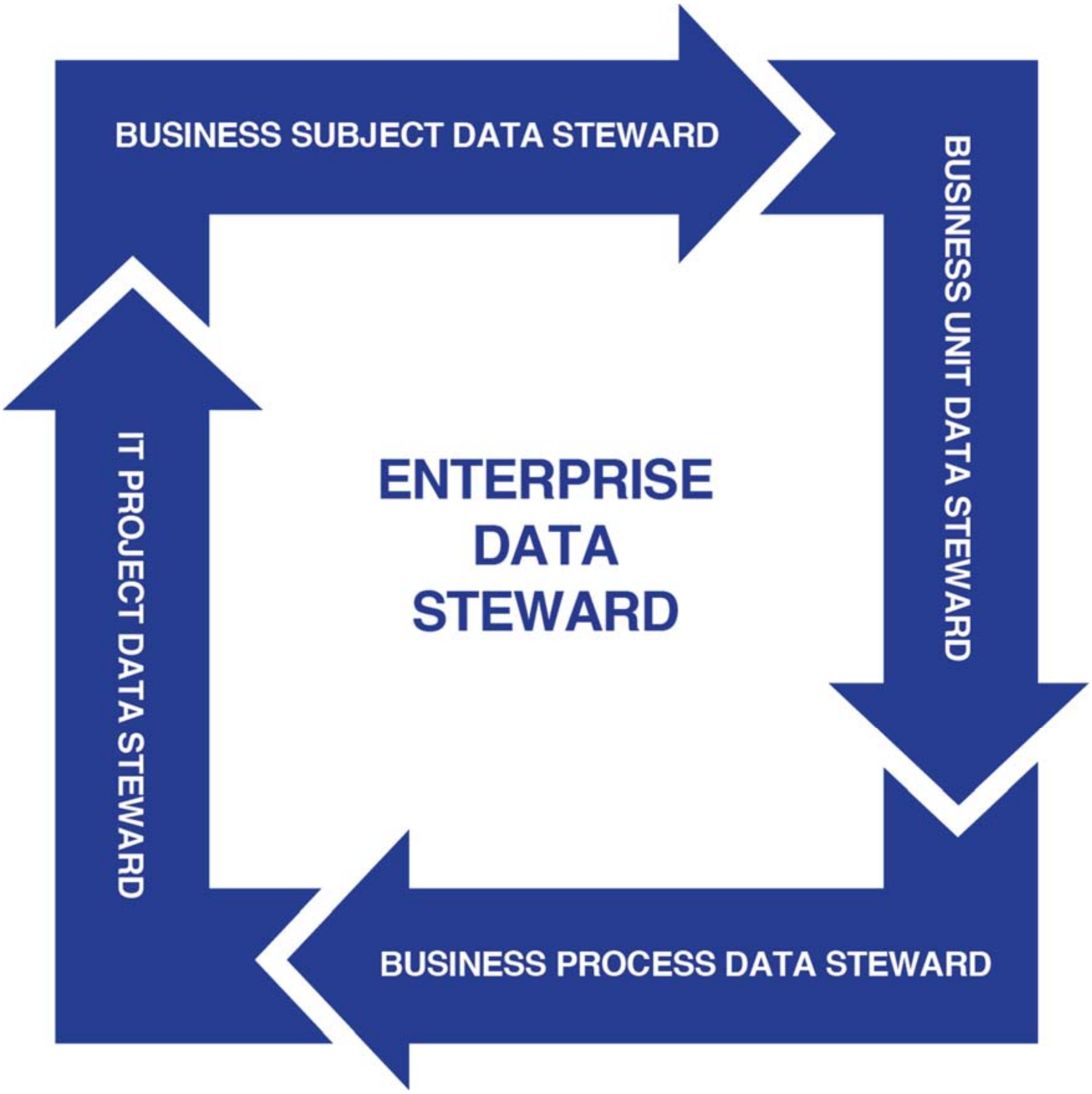
- *Definition and classification* address the many definitional and metadata topics previously discussed—data definitions, data naming, consistent use of data, lineage and traceability of data, etc.

- *Quality and security management* have a direct connection to the goals of—and motivations for—governance. Areas of responsibility include policies, regulations, goals, measures, monitoring, communication, education, and root cause analysis.

- *People and process* responsibilities are among the most important data steward responsibilities. Through teamwork, facilitation, and consensus building they form the core of a governance organization.

Data Governance

Data Stewardship


BUSINESS SUBJECT DATA STEWARD
BUSINESS UNIT DATA STEWARD
BUSINESS PROCESS DATA STEWARD
IT PROJECT DATA STEWARD
ENTERPRISE DATA STEWARD

Data Governance

Data Stewardship

MANY KINDS OF DATA STEWARDS

It is not realistic to assign the responsibility for all your data to a single data steward. Every organization is sure to have many stewards.

Claudia Imhoff recognized the need for multiple stewards very early in the emergence of the discipline. Imhoff wrote in 1997, "A typical corporate Data Stewardship function should have one Data Steward assigned to each major data subject area. These subject areas consist of the critical data entities or subjects such as Customer, Order, Product, Market Segment, Employee, Organization, Inventory, etc. Usually, there are about 15-20 major subject areas in any corporation."[1]

Imhoff describes what is known as a *Business Subject Data Steward*. As the discipline has evolved several other kinds of data stewards have emerged, including:

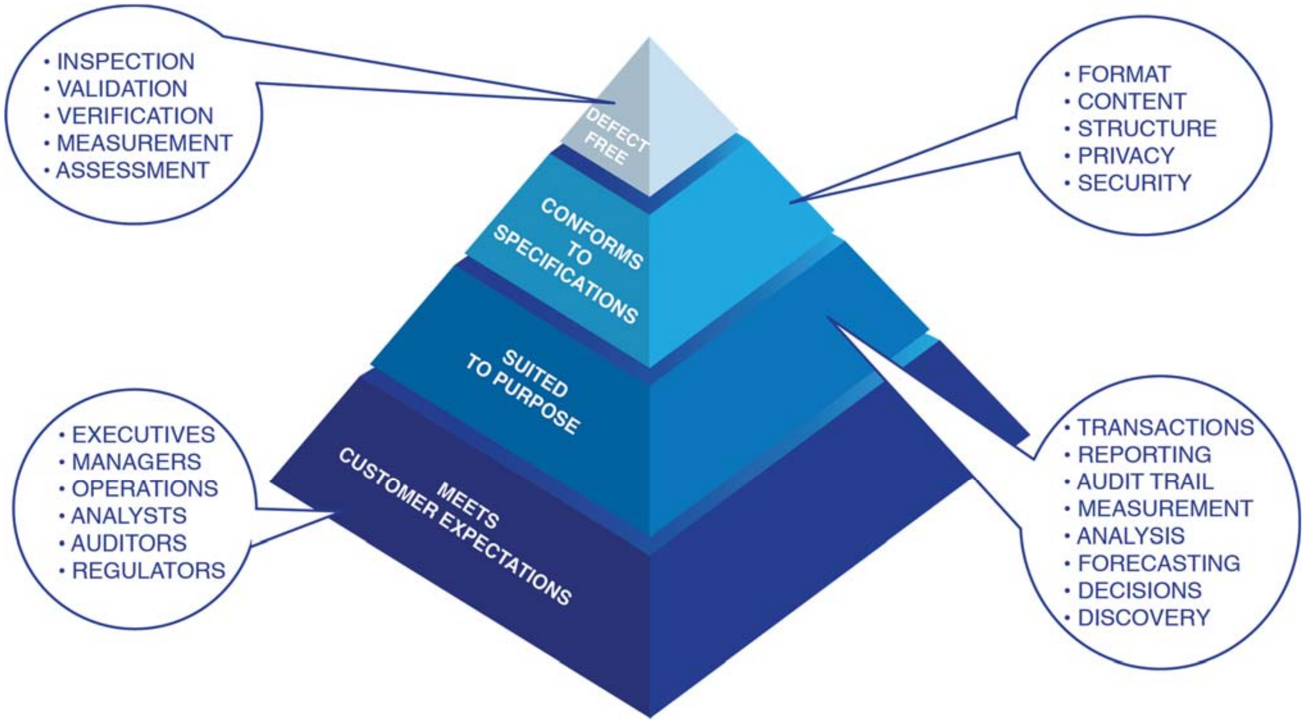
- *Business Unit Data Stewards* are responsible for the data needs of a particular business department or business function. This approach may augment subject stewardship for departments with a high level of data dependency and need for data management.
- *Business Process Data Stewards* provide data oversight from a process perspective and may supplement subject stewardship when a business process depends upon critical data flows from many sources.
- *Business Location Data Stewards* represent a location-specific data view in global and multinational corporations.
- *IT Project Data Stewards* represent a data management perspective on data-centric projects such as those to implement ERP, MDM, or data warehousing systems.

You may also need to identify an *Enterprise Data Steward*—also known as *Lead Data Steward* or *Chief Data Steward*—as a clarifying and coordinating role among many stewards. According to Imhoff, "The lead Data Steward's responsibility is to determine and control the domain of each Data Steward. These domains can become muddy and unclear, especially where subject areas intersect. Political battles can develop between the Data Stewards if their domains are not clearly established."[1]

[1] Source: "Data Stewardship: Process for Achieving Data Integrity," The Data Administration Newsletter (www.tdan.com/view-articles/4196). Imhoff is founder and president of Intelligent Solutions and a well-known speaker and author in business intelligence.

Data Quality

Data Quality Concepts



Data Quality

Data Quality Concepts

QUALITY DEFINITIONS

The Merriam-Webster dictionary defines quality as “degree of excellence.” The important point here is that quality is not an absolute, but something that exists in degrees. One common definition describes high quality as *defect free*. This interpretation comes from the community of quality practitioners who base their practice on the principle of zero defects. They define quality as *conformance to specifications* and defects as variance from specifications. Another widely used definition states that quality is *suitability to purpose*—a thing is of high quality when it is well suited to its intended purpose, and it is of poor quality when badly suited to its purpose. The principles of Total Quality Management (TQM) define quality as consistently *meeting customer expectations*. This principle promotes the idea that quality doesn’t reside within a product; it can only be judged in relation to the expectations of the customer using the product.

DATA AND DEFECTS

Defect-free data requires identification of the things that are data defects (more about this later), after which you can manage data by inspecting it to find defects, by validating and verifying data as free of defects, and by measuring defects as part of data quality assessment.

DATA AND SPECIFICATIONS

Conformance to specifications requires formal data specifications, which may address any or all of data format, content, and structure as well as usage-oriented specifications such as those for data privacy and security. Data quality management will test data against specifications.

DATA AND PURPOSE

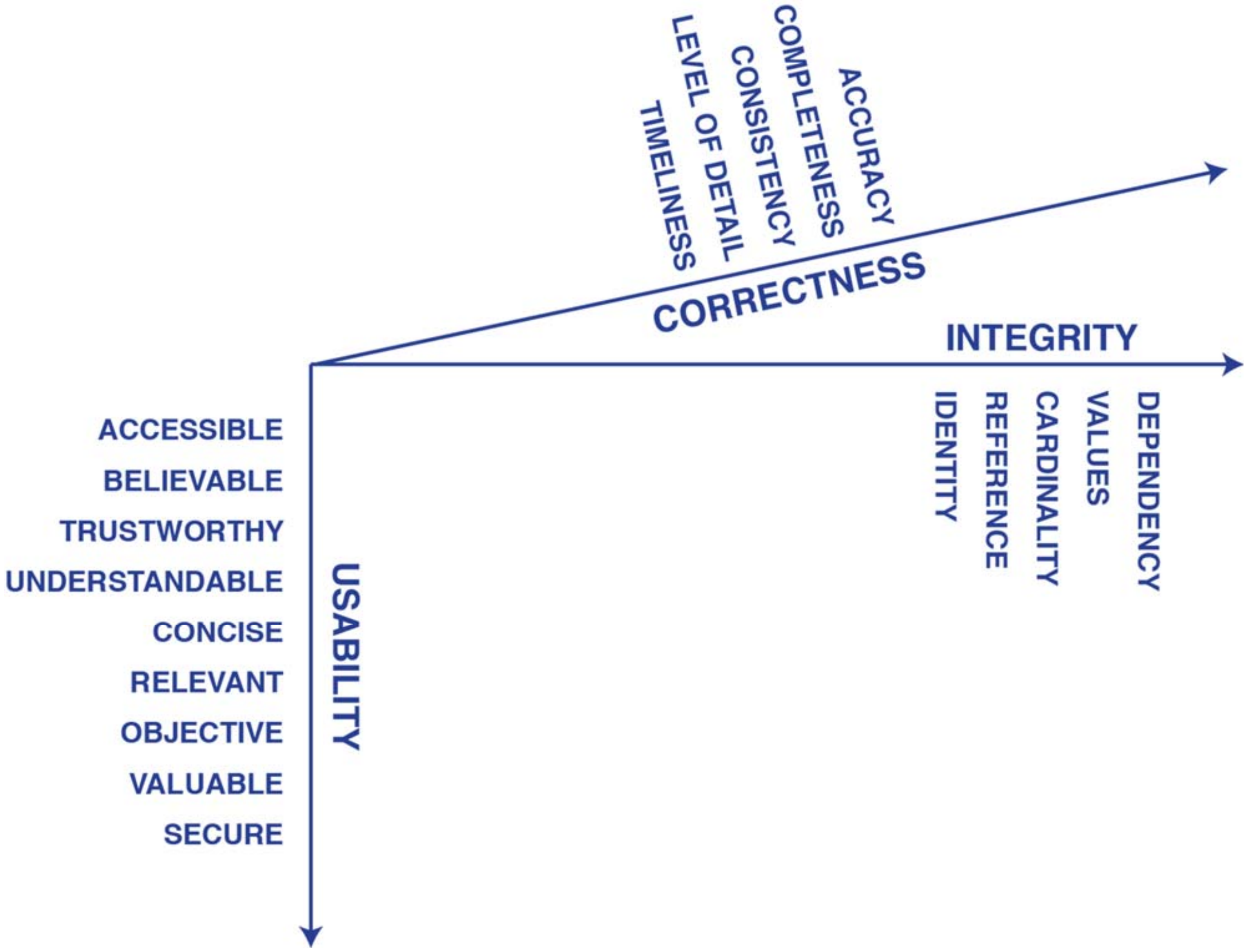
Suitability to purpose must consider all purposes for which data is used, ranging from business transactions and operational reporting to BI and analytics. Expect the quality criteria to vary widely among the different uses. Variations in quality criteria increase the level of difficulty in data quality management, but attention to them makes quality management efforts more effective and far-reaching.

DATA AND EXPECTATIONS

Data quality as meeting customer expectations must consider the wide range of data and information consumers. Expect wide variation in the expectations through the range of consumers, both internal and external. The quality management implications of varied expectations are much like those for varied purpose—greater complexity and greater impact.

Data Quality

Data Quality Concepts



Data Quality

Data Quality Concepts

DIMENSIONS OF DATA QUALITY

Data quality can be viewed in three dimensions asking:

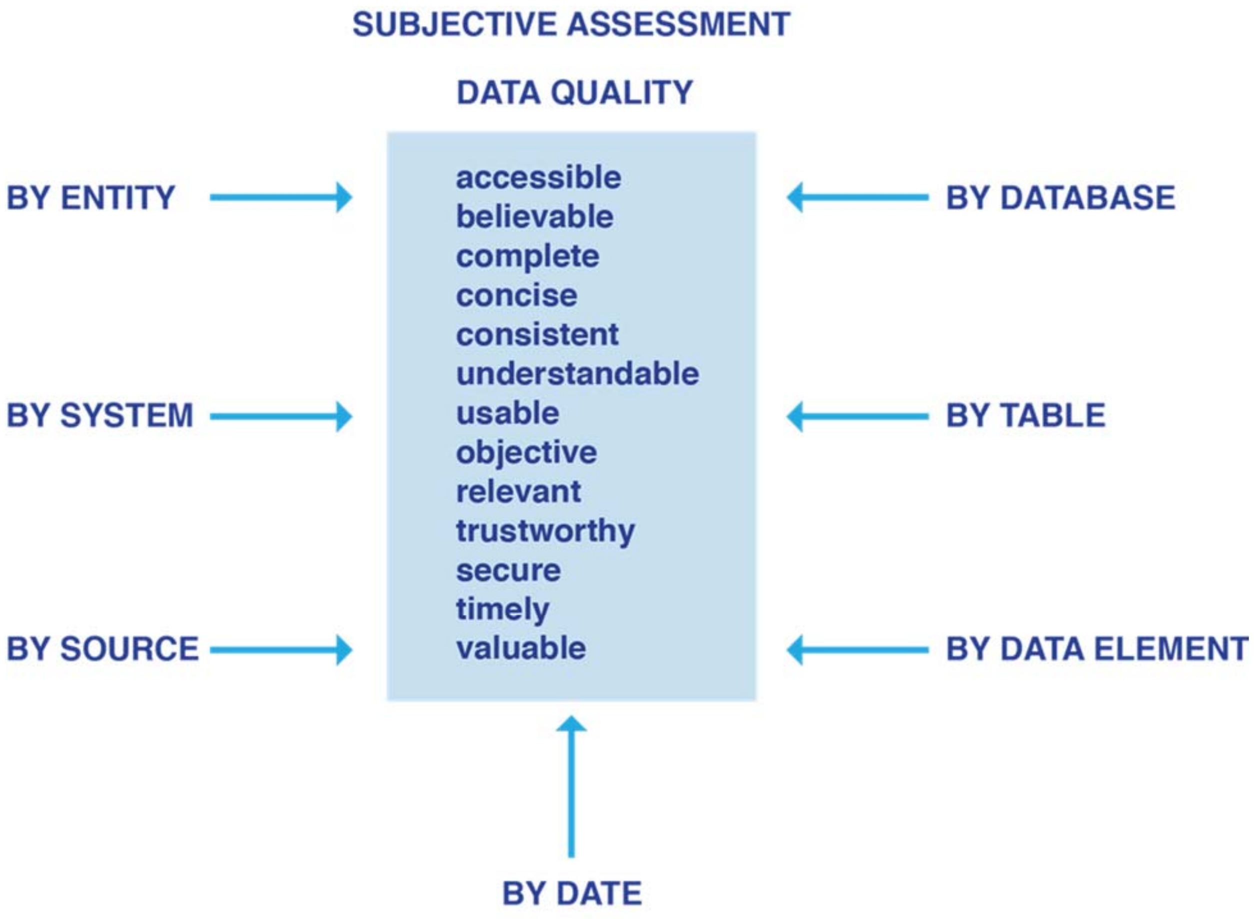
- Is the data *correct*?
  Attributes of correctness include
  - Accuracy—Does it represent the truth?
  - Completeness—Is anything missing?
  - Consistency—Is it without conflict and ambiguity?
  - Level of detail—Is it sufficiently granular and precise?
  - Timeliness—Is it fresh enough? Does it contain needed history?

- Does the data have *integrity*?
  Attributes of integrity include
  - Identity—Is each occurrence of an entity uniquely identified?
  - Reference—Are all relationships complete without dead ends?
  - Cardinality—Do relationships comply with business rules governing the numbers of each entity in a relationship?
  - Values—Are all values within the allowed domain of values?
  - Dependency—Does the data comply with all business rules for dependency among attributes and relationships?

- Is the data *usable*?
  Attributes of usability include
  - Accessibility—Can the data be accessed when and where it is needed?
  - Believability—Do data consumers believe it is high quality?
  - Trustworthiness—Do data consumers trust the data?
  - Understandability—Is the data easy to understand without risk of misinterpretation, misuse, and misinformation?
  - Conciseness—Is the data free from unnecessary and unwanted complexity and noise that make it difficult to use?
  - Relevance—Does the data match business needs and provide answers to business questions?
  - Objectivity—Is the data free from bias and manipulation?
  - Value—Can the data be used to achieve favorable business outcomes?
  - Security—Is the data adequately protected from intrusion, corruption, and loss?

Data Quality

Data Quality Assessment



Data Quality

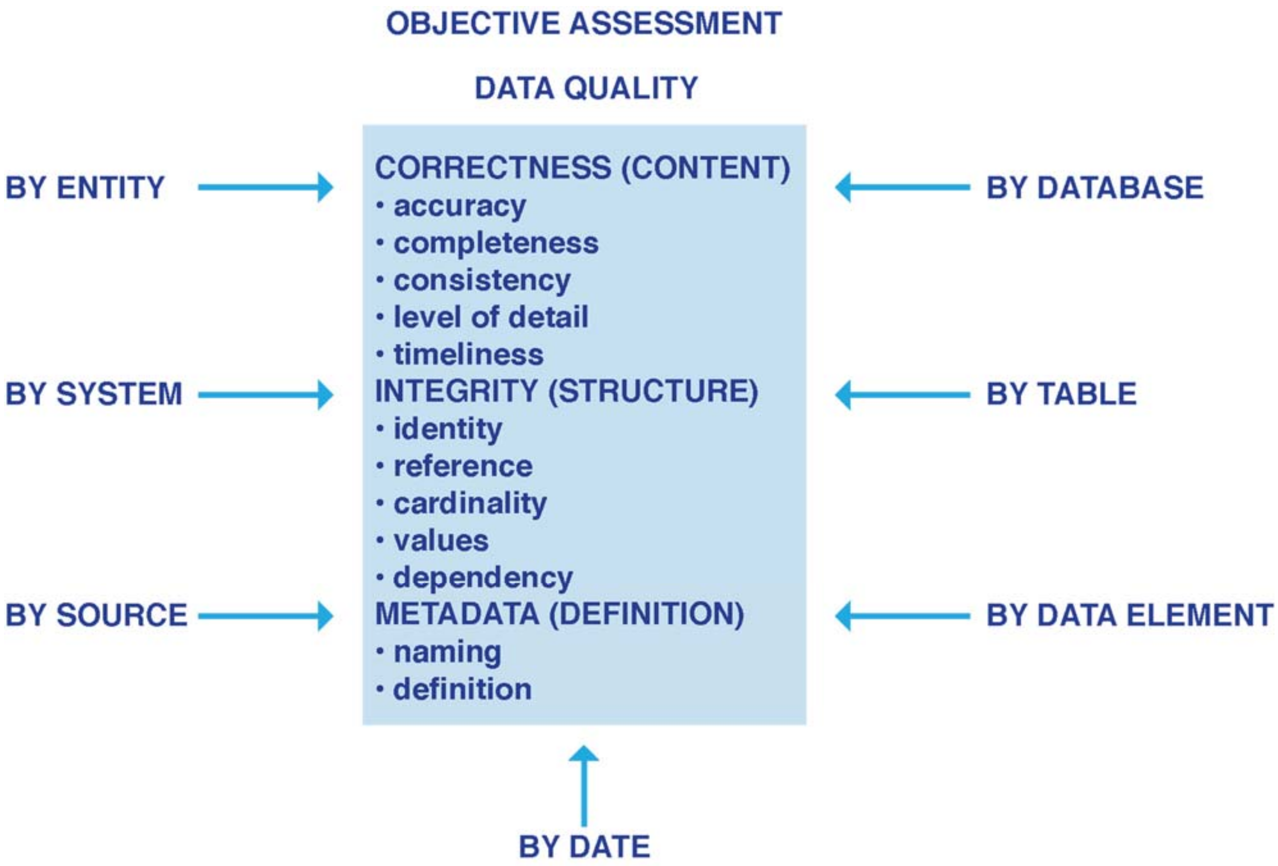
Data Quality Assessment

SUBJECTIVE ASSESSMENT

Subjective assessment seeks to understand and to quantify the perceptions and beliefs of people who work with data. Surveys are the most common method of subjective assessment, asking questions to measure the perceived level to which data satisfies subjective criteria such as believability, usability, and relevance. A well-designed survey collects sufficient demographic and contextual data to view the results in multiple dimensions. Multidimensional analysis of data quality by variables such as entity, database, system, source, etc., is valuable when planning the next steps after assessment.

Data Quality

Data Quality Assessment



OBJECTIVE ASSESSMENT

DATA QUALITY

CORRECTNESS (CONTENT)
- accuracy
- completeness
- consistency
- level of detail
- timeliness

INTEGRITY (STRUCTURE)
- identity
- reference
- cardinality
- values
- dependency

METADATA (DEFINITION)
- naming
- definition

BY ENTITY

BY SYSTEM

BY SOURCE

BY DATABASE

BY TABLE

BY DATA ELEMENT

BY DATE

Data Quality

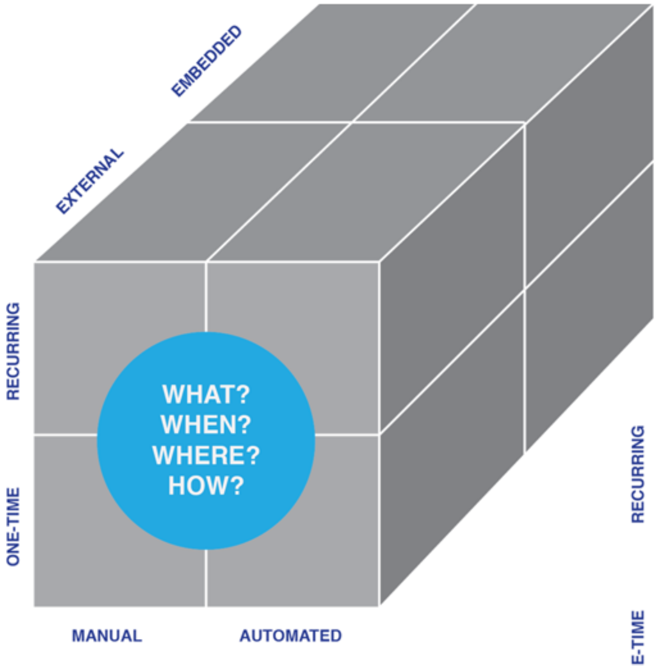
Data Quality Assessment

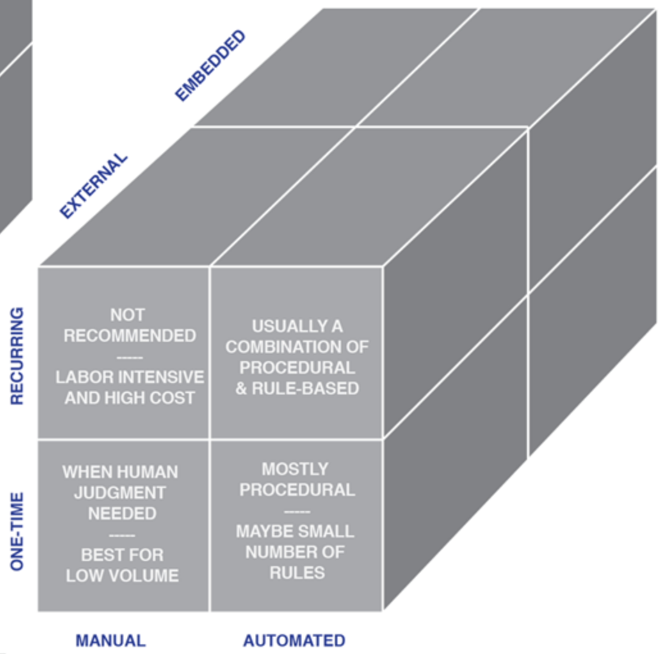
OBJECTIVE ASSESSMENT

Objective assessment seeks to quantify very specific and tangible characteristics of the data. Where subjective assessment focuses on what people say, objective assessment asks: what does the data say? Objective assessment is performed by processing test data against data quality rules for criteria related to content correctness and structural integrity. As with subjective assessment, a multidimensional view of data quality measures is a valuable data quality management capability.

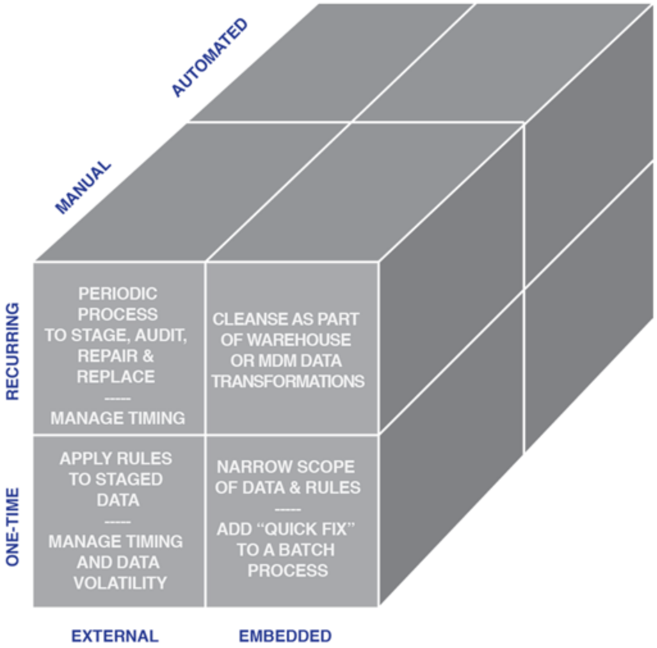
Data Quality

Data Quality Improvement

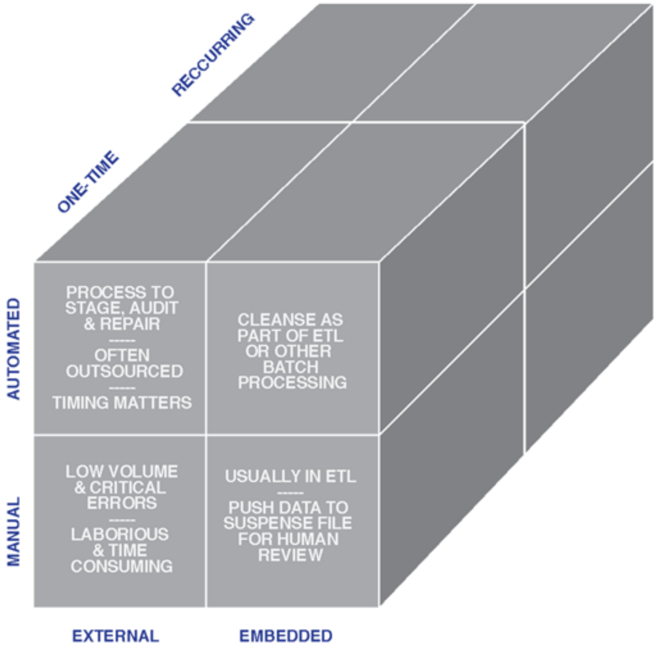




	MANUAL	AUTOMATED
RECURRING	NOT RECOMMENDED ----- LABOR INTENSIVE AND HIGH COST	USUALLY A COMBINATION OF PROCEDURAL & RULE-BASED
ONE-TIME	WHEN HUMAN JUDGMENT NEEDED ----- BEST FOR LOW VOLUME	MOSTLY PROCEDURAL ----- MAYBE SMALL NUMBER OF RULES



	EXTERNAL	EMBEDDED
RECURRING	PERIODIC PROCESS TO STAGE, AUDIT, REPAIR & REPLACE ----- MANAGE TIMING	CLEANSE AS PART OF WAREHOUSE OR MDM DATA TRANSFORMATIONS
ONE-TIME	APPLY RULES TO STAGED DATA ----- MANAGE TIMING AND DATA VOLATILITY	NARROW SCOPE OF DATA & RULES ----- ADD "QUICK FIX" TO A BATCH PROCESS



	EXTERNAL	EMBEDDED
AUTOMATED	PROCESS TO STAGE, AUDIT & REPAIR ----- OFTEN OUTSOURCED ----- TIMING MATTERS	CLEANSE AS PART OF ETL OR OTHER BATCH PROCESSING
MANUAL	LOW VOLUME & CRITICAL ERRORS ----- LABORIOUS & TIME CONSUMING	USUALLY IN ETL ----- PUSH DATA TO SUSPENSE FILE FOR HUMAN REVIEW

Data Quality

Data Quality Improvement

DEFINITION AND CONCEPTS

Data cleansing is the act of detecting and correcting or removing corrupt or inaccurate records from a record set, table, or database. It is a process of finding and removing data quality defects. Cleansing may involve removing defective data from the collection, obtaining correct data from an alternate source, or adjusting defective data to comply with data quality rules.

Data cleansing may be:

- Manual (performed by people) or automated (performed by computer)
- One-time (a single-instance repair) or recurring (periodic processing)
- Embedded (integrated into existing processes) or external (performed as a standalone process).

These options combine in some interesting ways—embedded, automated, recurring for example; or external, manual, one-time. A complete data cleansing solution typically mixes-and-matches several options. High-level questions for each cleansing activity include:

- What to cleanse—which data and which defects?
- When to cleanse—at what point in business and systems schedules?
- Where to cleanse—at what point in the flow of data and processes?
- How to cleanse—using what methods and workflow?

CLEANSING VARIATIONS

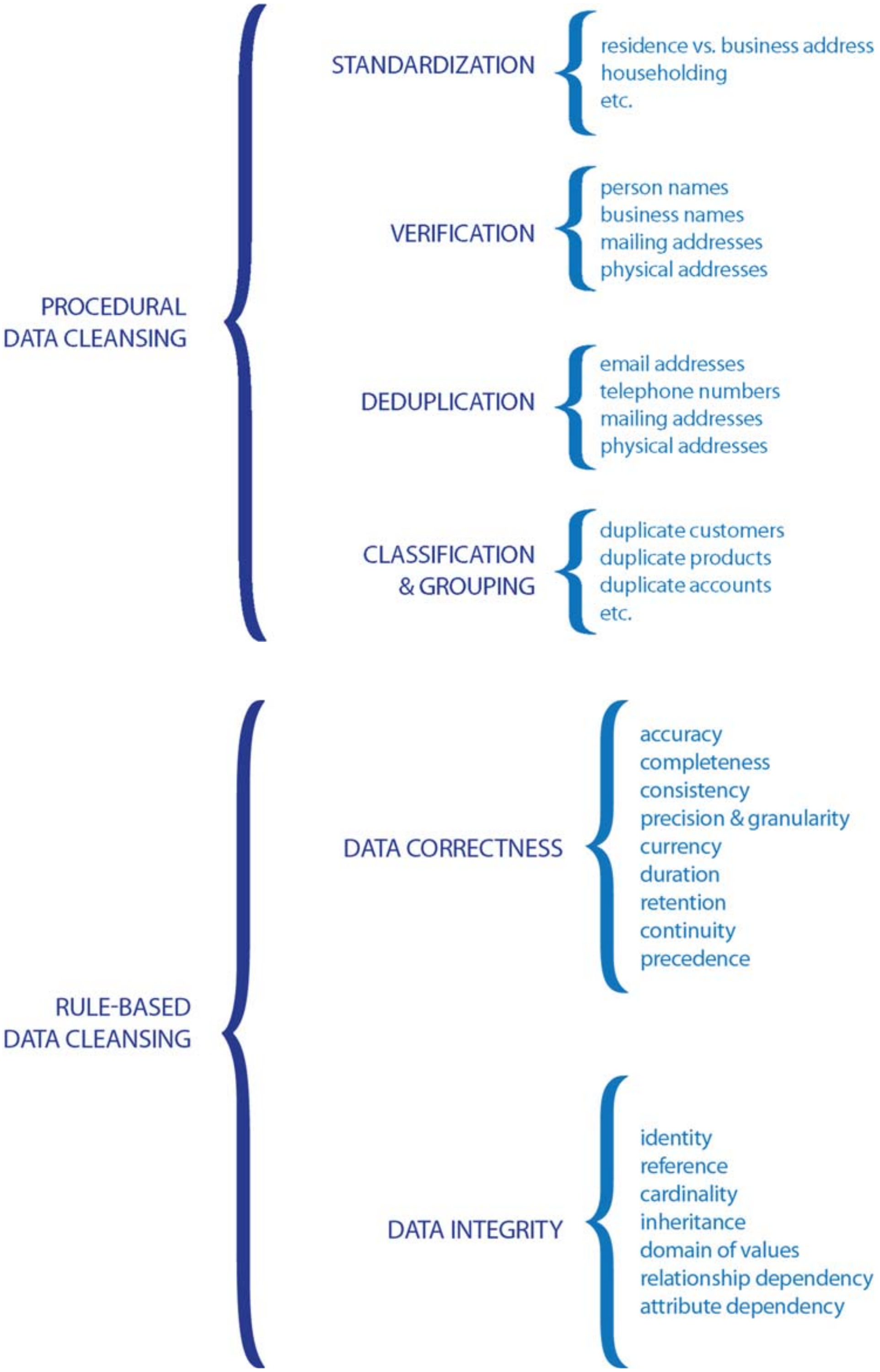
Using two-dimensional views for a closer look at combinations offers some guidelines for the use of each. One-time/manual/external, for example, yields these guidelines:

- Effective for low-volume cleansing where human judgment is needed.
- Stage the data separately from production data, and then apply cleansing rules.
- Be conscious of the time that manual processing takes and consider the matching issues that may occur when you're ready to merge staged and cleansed data back into a volatile production database.
- Manual processing is best suited to low-volume, critical errors because it is laborious and time-consuming.

Collectively the guidelines indicate that one-time/manual/external works for a small number of critical errors needing human judgment to correct. There is risk if production data changes while a copy is separately staged for cleansing.

Data Quality

Data Quality Improvement



Data Quality

Data Quality Improvement

CLEANSING TECHNIQUES

Data quality techniques fit into two categories—procedural and rule based. *Procedural DQ* is an algorithmic approach to cleansing or standardization of data where predictable and consistent patterns exist and where the same cleansing logic can be applied across many instances of similar data types. *Rule-based DQ* applies inspection and logic to data where the patterns and criteria are not predictable and consistent across a large population of data, but are specialized to the enterprise and often to the database.

PROCEDURAL DATA CLEANSING

Procedural data cleansing processes are used to standardize or verify specific types of data.

Standardization uses procedures to apply patterns and criteria that are typically global, geographical/cultural, industry, or enterprise standards. Standardizing person names, for example, to get consistent name patterns could be an enterprise-level standard. Standardizing salutations and designations—Mr., Mrs., Ms., Dr., etc.—for consistent recording and abbreviation is an example of a global standard. Standardizing U.S. mailing addresses to conform to U.S. Postal Service conventions for formatting and use of abbreviations is a geographic example.

Verification applies to data elements and groups such as email addresses, telephone numbers, and mailing addresses to confirm that the things represented by the data actually exist. Verification is usually performed by checking with external databases or subscription data services.

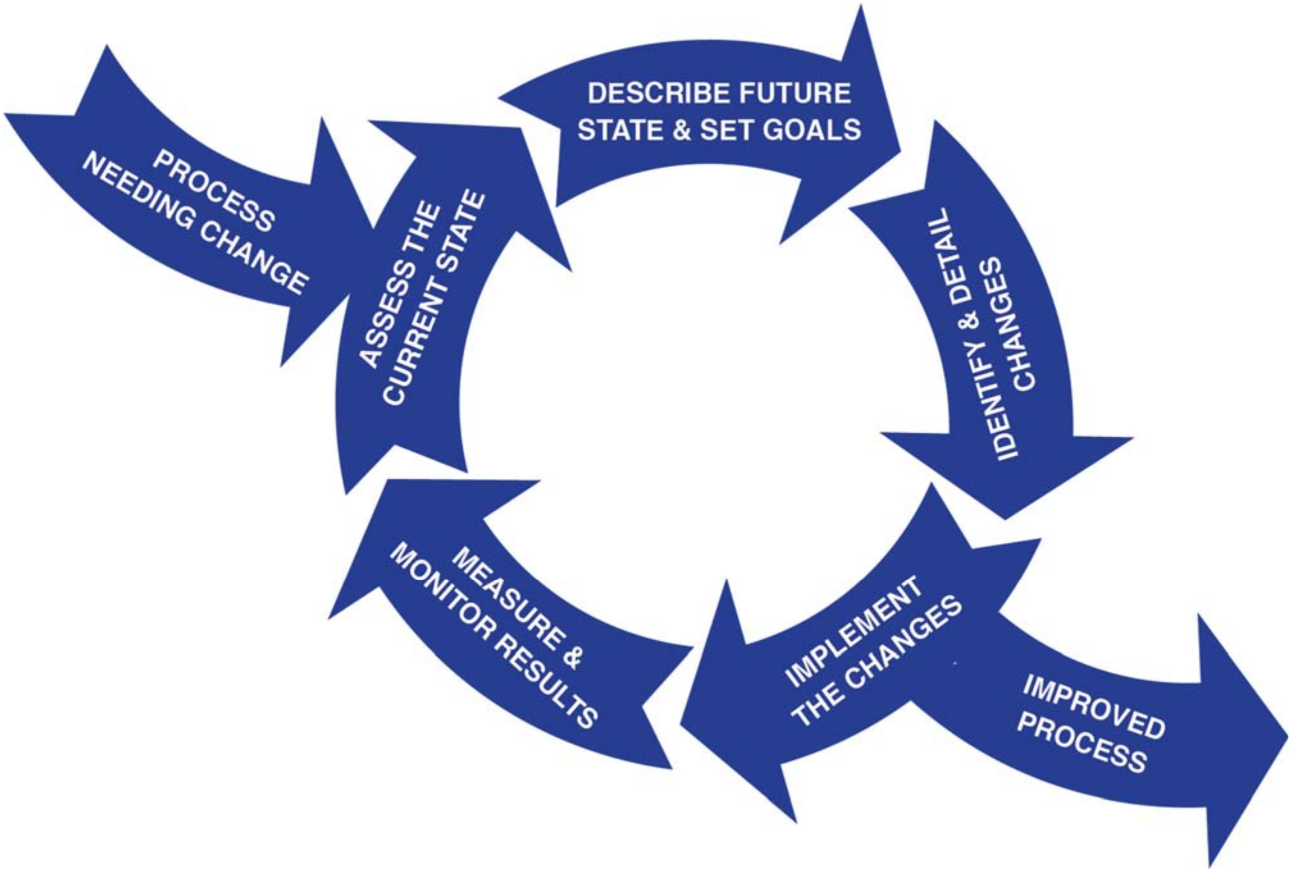
RULE-BASED DATA CLEANSING

Rule-based cleansing picks up where procedural techniques stop. Procedural DQ works with algorithms. Rule-based techniques use data quality rules that express constraints upon the qualities of the data. These are the rules regarding the integrity and correctness of the data.

Combining a data quality rule with a data transformation rule produces a data cleansing rule. The data quality rule is the means by which DQ defects are detected. The data transformation rule is the means to correct defects.

Data Quality

Data Quality Improvement


PROCESS NEEDING CHANGE
ASSESS THE CURRENT STATE
DESCRIBE FUTURE STATE & SET GOALS
IDENTIFY & DETAIL CHANGES
IMPLEMENT THE CHANGES
MEASURE & MONITOR RESULTS
IMPROVED PROCESS

Data Quality

Data Quality Improvement

DEFECT PREVENTION

Root causes of data quality defects are often found in processes. (This was described in *Module 2: Supporting the Organization.*) Once the cause is found, the most effective preventive actions are typically process changes to remove the causes. The principles and techniques of process improvement can be applied effectively.

PROCESS IMPROVEMENT DEFINED

Process improvement is the work of preventing the occurrence of future defects. In data quality, as with any other product, causes of defects fall into two broad categories—defective materials and process deficiencies. Process improvement focuses on correcting process deficiencies to eliminate the causes of defects.

PROCESS IMPROVEMENT CYCLES

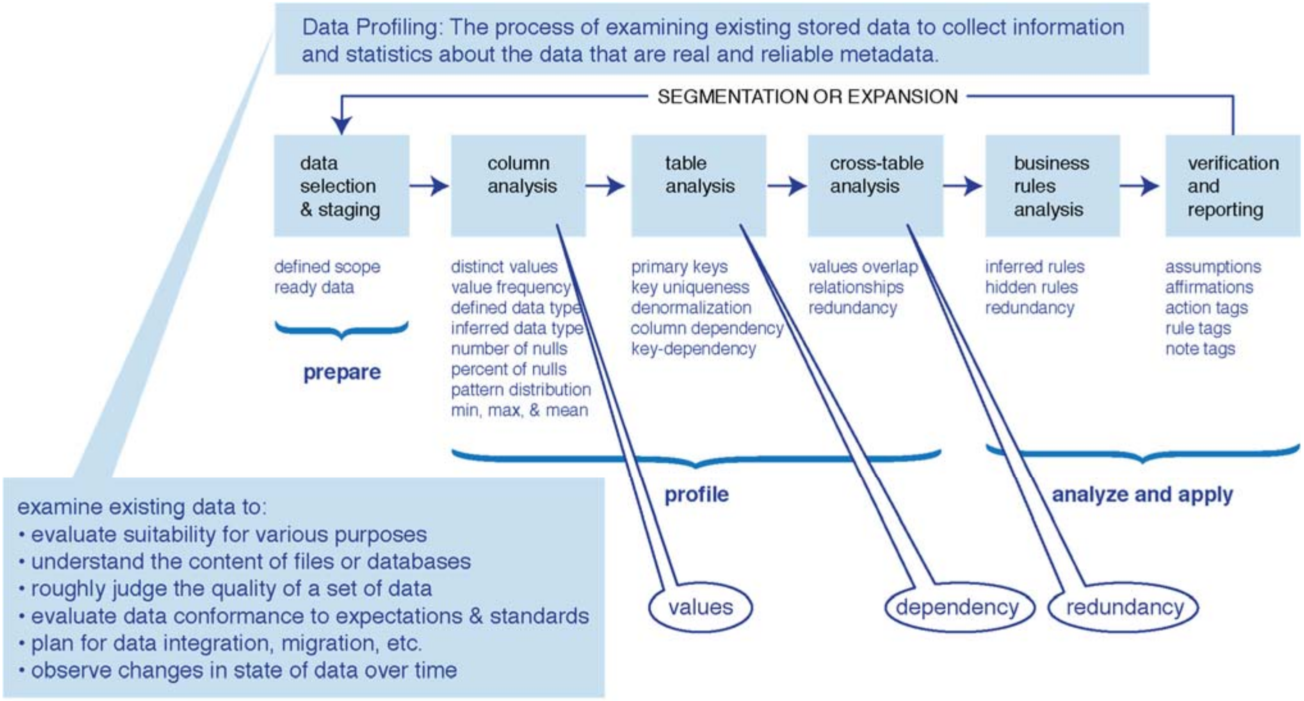
Process improvement begins with recognition of a process needing to change and ends with implementation of an improved process. Between the beginning and the end is a cyclical process of the following:

- Assess the current state—know where you are objectively
- Describe the future state and set goals—know where you want to go and make it measurable
- Identify and detail changes—build an action plan
- Implement the changes—execute the action plan
- Measure and monitor results—check progress against goals

Repeat the cycle until the process is optimized.

Data Profiling

Purpose and Processes



Data Profiling: The process of examining existing stored data to collect information and statistics about the data that are real and reliable metadata.

SEGMENTATION OR EXPANSION

data selection & staging → column analysis → table analysis → cross-table analysis → business rules analysis → verification and reporting

data selection & staging: defined scope, ready data

column analysis: distinct values, value frequency, defined data type, inferred data type, number of nulls, percent of nulls, pattern distribution, min, max, & mean

table analysis: primary keys, key uniqueness, denormalization, column dependency, key-dependency

cross-table analysis: values overlap, relationships, redundancy

business rules analysis: inferred rules, hidden rules, redundancy

verification and reporting: assumptions, affirmations, action tags, rule tags, note tags

prepare

profile

analyze and apply

values

dependency

redundancy

examine existing data to:

- evaluate suitability for various purposes
- understand the content of files or databases
- roughly judge the quality of a set of data
- evaluate data conformance to expectations & standards
- plan for data integration, migration, etc.
- observe changes in state of data over time

Data Profiling

Purpose and Processes

EXTRACTING METADATA FROM STORED DATA

Data profiling is the process of examining existing data to collect statistics about that data. The statistics provide "real and true metadata" that is used to understand data content and structure. Profiling helps with subjective assessment of data quality, but more importantly, it provides valuable input to find and define data quality rules. Profiling provides real knowledge about the current state of data.

WHY PROFILE?

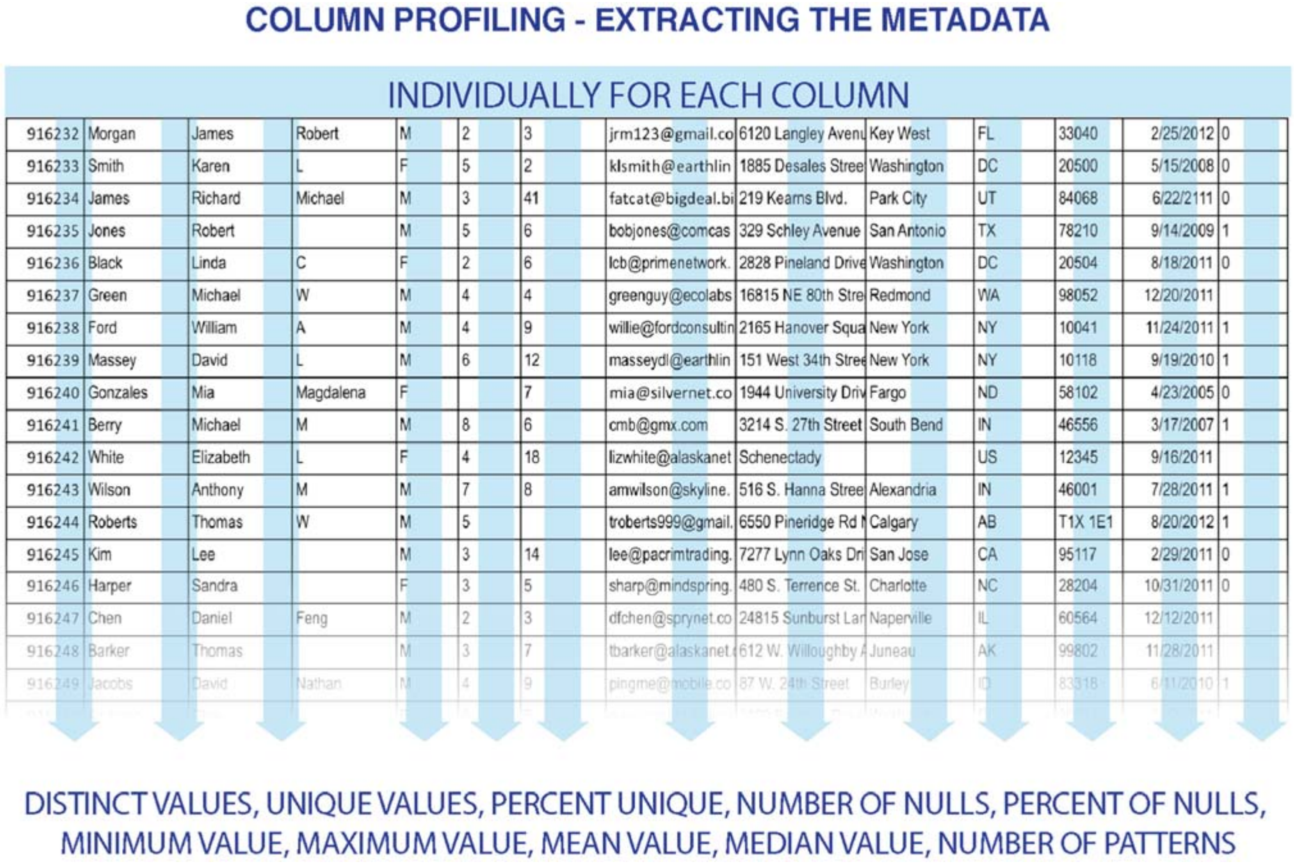
Data profiling is the work of looking at the data to understand it. Although looking at the data may seem an obvious necessity, it is often overlooked. The tendency to review data models, descriptions, definitions, and program code causes many to overlook the obvious, and those who do look at the data often do so in an unstructured way that leads to seeing only that which is expected.

Data is examined to:

- evaluate suitability for various purposes
- understand the content of files or databases
- roughly judge the quality of a set of data
- evaluate data conformance to expectations and standards
- plan for data integration, migration, etc.
- observe changes in the state of data over time

Data Profiling

Profiling Techniques



COLUMN PROFILING - EXTRACTING THE METADATA

INDIVIDUALLY FOR EACH COLUMN

916232	Morgan	James	Robert	M	2	3	jrm123@gmail.co	6120 Langley Avenu	Key West	FL	33040	2/25/2012	0
916233	Smith	Karen	L	F	5	2	klsmith@earthlin	1885 Desales Stree	Washington	DC	20500	5/15/2008	0
916234	James	Richard	Michael	M	3	41	fatcat@bigdeal.bi	219 Kearns Blvd.	Park City	UT	84068	6/22/2111	0
916235	Jones	Robert		M	5	6	bobjones@comcas	329 Schley Avenue	San Antonio	TX	78210	9/14/2009	1
916236	Black	Linda	C	F	2	6	lcb@primenetwork.	2828 Pineland Drive	Washington	DC	20504	8/18/2011	0
916237	Green	Michael	W	M	4	4	greenguy@ecolabs	16815 NE 80th Stre	Redmond	WA	98052	12/20/2011	
916238	Ford	William	A	M	4	9	willie@fordconsultin	2165 Hanover Squa	New York	NY	10041	11/24/2011	1
916239	Massey	David	L	M	6	12	masseydl@earthlin	151 West 34th Stree	New York	NY	10118	9/19/2010	1
916240	Gonzales	Mia	Magdalena	F		7	mia@silvernet.co	1944 University Driv	Fargo	ND	58102	4/23/2005	0
916241	Berry	Michael	M	M	8	6	cmb@gmx.com	3214 S. 27th Street	South Bend	IN	46556	3/17/2007	1
916242	White	Elizabeth	L	F	4	18	lizwhite@alaskanet	Schenectady		US	12345	9/16/2011	
916243	Wilson	Anthony	M	M	7	8	amwilson@skyline.	516 S. Hanna Stree	Alexandria	IN	46001	7/28/2011	1
916244	Roberts	Thomas	W	M	5		troberts999@gmail.	6550 Pineridge Rd N	Calgary	AB	T1X 1E1	8/20/2012	1
916245	Kim	Lee		M	3	14	lee@pacrimtrading.	7277 Lynn Oaks Dri	San Jose	CA	95117	2/29/2011	0
916246	Harper	Sandra		F	3	5	sharp@mindspring.	480 S. Terrence St.	Charlotte	NC	28204	10/31/2011	0
916247	Chen	Daniel	Feng	M	2	3	dfchen@sprynet.co	24815 Sunburst Lar	Naperville	IL	60564	12/12/2011	
916248	Barker	Thomas		M	3	7	tbarker@alaskanet.(	612 W. Willoughby A	Juneau	AK	99802	11/28/2011	
916249	Jacobs	David	Nathan	M	4	9	pingme@mobile.co	87 W. 24th Street	Burley	ID	83318	6/11/2010	1

DISTINCT VALUES, UNIQUE VALUES, PERCENT UNIQUE, NUMBER OF NULLS, PERCENT OF NULLS, MINIMUM VALUE, MAXIMUM VALUE, MEAN VALUE, MEDIAN VALUE, NUMBER OF PATTERNS

Data Profiling

Profiling Techniques

FREQUENCY OF VALUES

Frequency of values, or frequency distribution, is a statistical concept that is central to data profiling. Frequency distribution is the tabulation of the number of occurrences of each unique value in a data set; in profiling it tabulates unique values in a column of data. Frequency distribution is the foundation from which other column metadata is determined.

BUILDING ON THE FOUNDATION

Statistics are a necessary foundation from which much more metadata can be derived. Profiling a column of data produces obvious statistics such as min, max, mean, frequency, and a list of distinct values. That metadata, however, is only the beginning. It can be applied to derive and infer much more:

- For a single column it is possible to infer a data type from the set of unique values, and to compare the inferred data type with the data type that is defined for the column. You might, for example, profile a column defined as VCHAR and find that all of the values in that column appear to be DATE data.

- Additional inference is possible when profiling multiple columns of a table. A column has unique values, for example, when frequency distribution is flat and the number of distinct values is equal to the count of all values. Probability is high that the column is a primary key for the table.

- Patterns of correspondence among columns in a single table lead to inference of column dependency, discovery of multivalued columns, evidence of repeating groups (embedded tables), and other forms of denormalization.

- Profiling of columns across multiple tables exposes overlapping domains of values which are indicators of table-to-table relationships and of data redundancy.

Data Profiling

Profiling Techniques

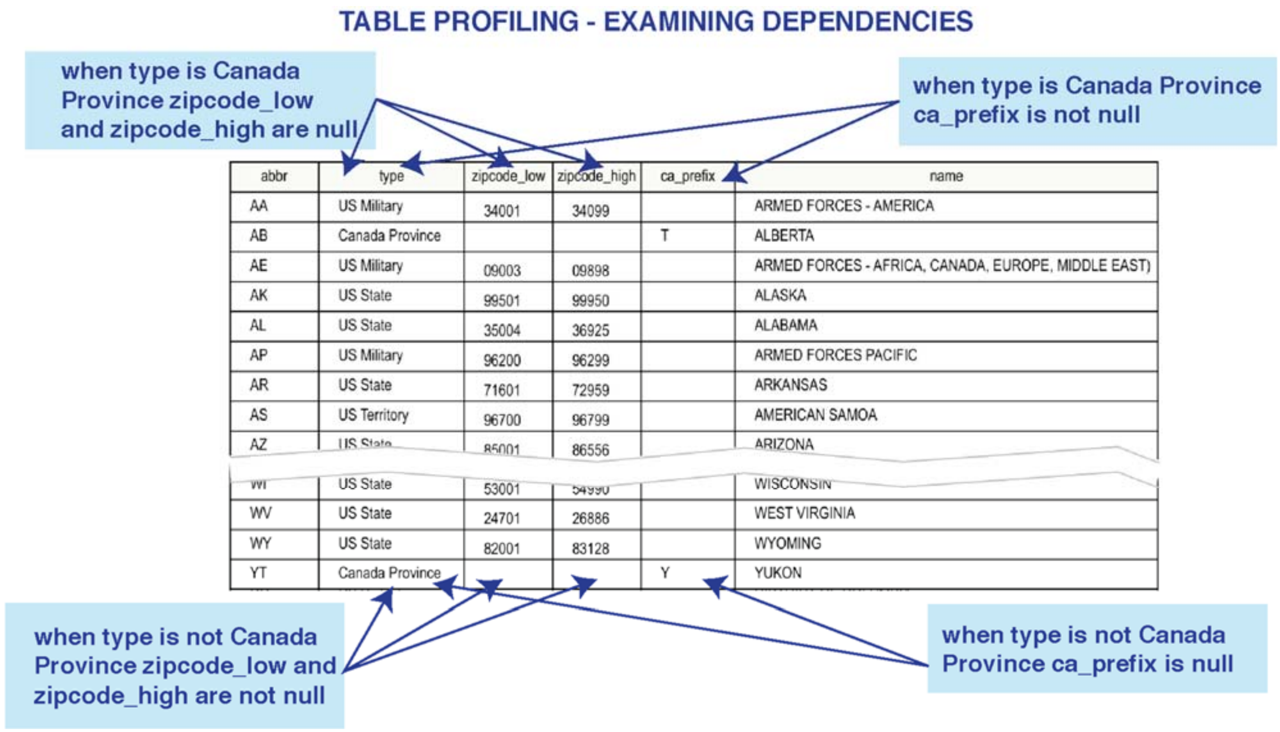


TABLE PROFILING - EXAMINING DEPENDENCIES

when type is Canada Province zipcode_low and zipcode_high are null

when type is Canada Province ca_prefix is not null

abbr	type	zipcode_low	zipcode_high	ca_prefix	name
AA	US Military	34001	34099		ARMED FORCES - AMERICA
AB	Canada Province			T	ALBERTA
AE	US Military	09003	09898		ARMED FORCES - AFRICA, CANADA, EUROPE, MIDDLE EAST)
AK	US State	99501	99950		ALASKA
AL	US State	35004	36925		ALABAMA
AP	US Military	96200	96299		ARMED FORCES PACIFIC
AR	US State	71601	72959		ARKANSAS
AS	US Territory	96700	96799		AMERICAN SAMOA
AZ	US State	85001	86556		ARIZONA
WI	US State	53001	54990		WISCONSIN
WV	US State	24701	26886		WEST VIRGINIA
WY	US State	82001	83128		WYOMING
YT	Canada Province			Y	YUKON

when type is not Canada Province zipcode_low and zipcode_high are not null

when type is not Canada Province ca_prefix is null

Data Profiling

Profiling Techniques

PATTERNS AMONG COLUMNS

Column dependencies are exhibited as persistent and recurring patterns among the values of two or more columns. The postal table examples illustrated here include:

- When type is *Canada Province*, then zipcode_low and zipcode_high are null.
- When type is *Canada Province*, then ca_prefix is not null.
- When type is not *Canada Province,* then zipcode_low and zipcode_high are not null.
- When type is not *Canada Province*, then ca_prefix is null.

Several other dependencies, though not explicitly designated on the facing page, can be found in this table:

- When zipcode_low is null, then zipcode_high is null.
- When zipcode_high is null, then zipcode_low is null.
- The value of zipcode_low is always less than the value of zipcode_high.
- When type begins with the letters *US*, then zipcode_high and zipcode_low are not null.
- When type begins with the letters *US,* then ca_prefix is null.

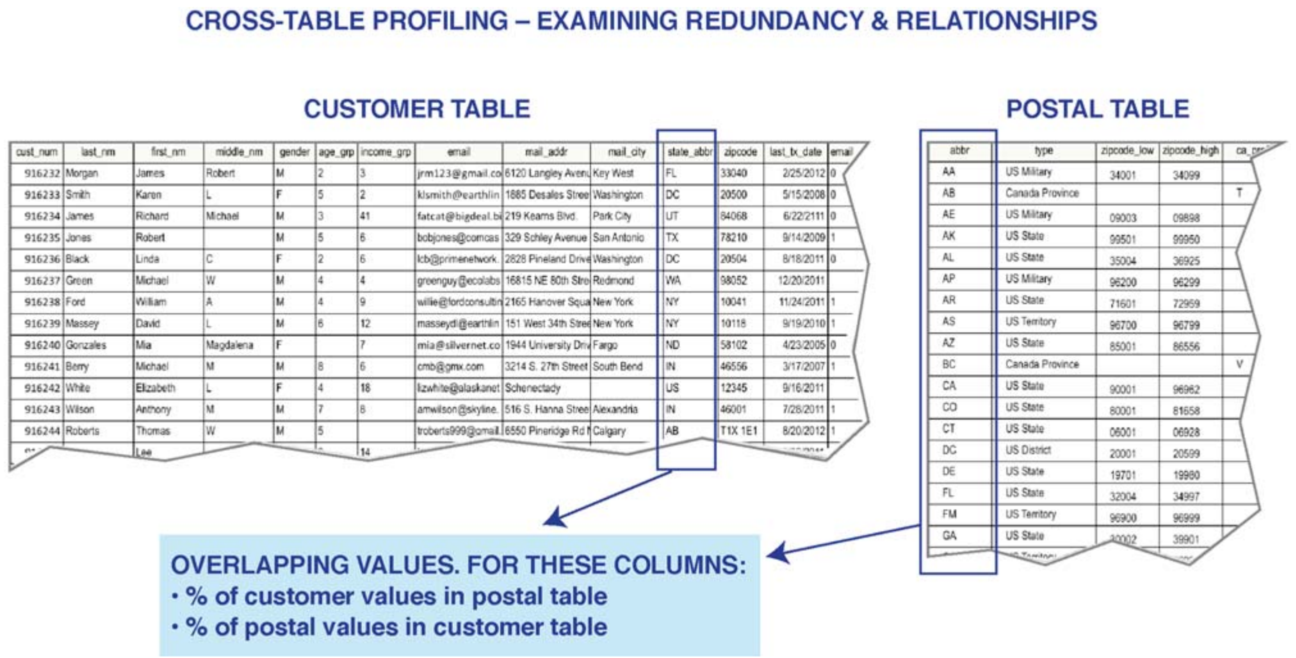
DOMAIN CONSTRAINTS

Column dependencies are readily translated into data quality rules that limit the allowable values of a data element. These rules generally fall into three categories:

- Constrained domain of values, where the set of allowed values for a specific data element is limited to a subset of the full domain set.
- Derived values, where a single correct value is determined by the values of one or more other data elements.
- Null rules, where allowance or requirement of a null value is based on the values of other data elements.

Data Profiling

Profiling Techniques



CROSS-TABLE PROFILING – EXAMINING REDUNDANCY & RELATIONSHIPS

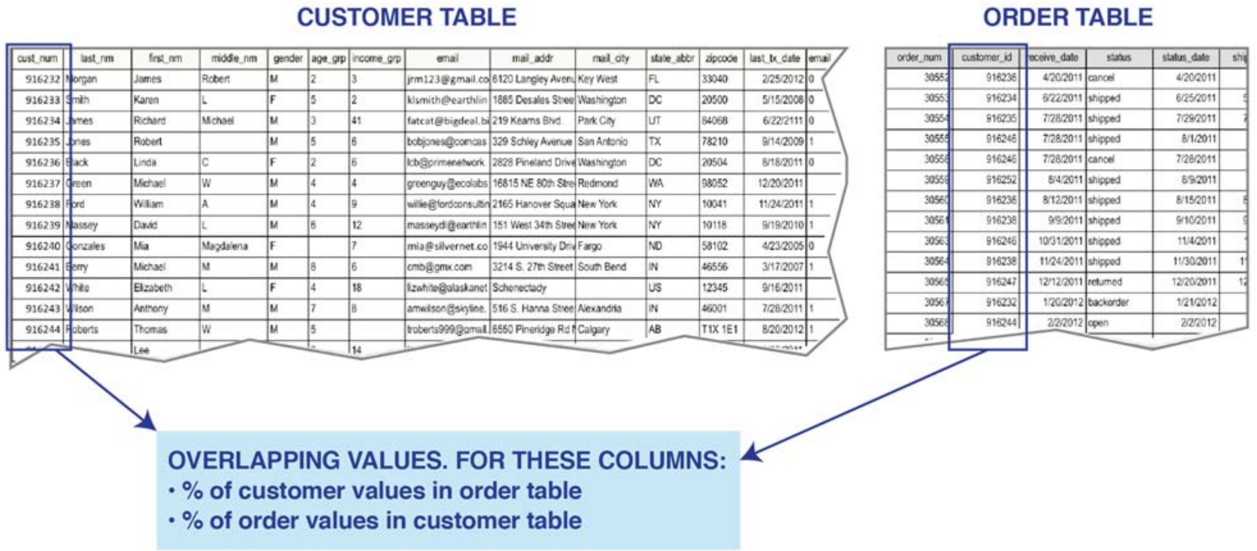
CUSTOMER TABLE

cust_num	last_nm	first_nm	middle_nm	gender	age_grp	income_grp	email	mail_addr	mail_city	state_abbr	zipcode	last_tx_date	email
916232	Morgan	James	Robert	M	2	3	jrm123@gmail.co	6120 Langley Avenu	Key West	FL	33040	2/25/2012	0
916233	Smith	Karen	L	F	5	2	klsmith@earthlin	1885 Desales Stree	Washington	DC	20500	5/15/2008	0
916234	James	Richard	Michael	M	3	41	fatcat@bigdeal.bi	219 Kearns Blvd.	Park City	UT	84068	6/22/2111	0
916235	Jones	Robert		M	5	6	bobjones@comcas	329 Schley Avenue	San Antonio	TX	78210	9/14/2009	1
916236	Black	Linda	C	F	2	6	lcb@primenetwork.	2828 Pineland Drive	Washington	DC	20504	8/18/2011	0
916237	Green	Michael	W	M	4	4	greenguy@ecolabs	16815 NE 80th Stre	Redmond	WA	98052	12/20/2011	
916238	Ford	William	A	M	4	9	willie@fordconsultin	2165 Hanover Squa	New York	NY	10041	11/24/2011	1
916239	Massey	David	L	M	6	12	masseydl@earthlin	151 West 34th Stree	New York	NY	10118	9/19/2010	1
916240	Gonzales	Mia	Magdalena	F		7	mia@silvernet.co	1944 University Driv	Fargo	ND	58102	4/23/2005	0
916241	Berry	Michael	M	M	8	6	cmb@gmx.com	3214 S. 27th Street	South Bend	IN	46556	3/17/2007	1
916242	White	Elizabeth	L	F	4	18	lizwhite@alaskanet	Schenectady		US	12345	9/16/2011	
916243	Wilson	Anthony	M	M	7	8	amwilson@skyline.	516 S. Hanna Stree	Alexandria	IN	46001	7/28/2011	1
916244	Roberts	Thomas	W	M	5		troberts999@gmail.	6550 Pineridge Rd N	Calgary	AB	T1X 1E1	8/20/2012	1
		Lee				14							

POSTAL TABLE

abbr	type	zipcode_low	zipcode_high	ca_pr
AA	US Military	34001	34099	
AB	Canada Province			T
AE	US Military	09003	09898	
AK	US State	99501	99950	
AL	US State	35004	36925	
AP	US Military	96200	96299	
AR	US State	71601	72959	
AS	US Territory	96700	96799	
AZ	US State	85001	86556	
BC	Canada Province			V
CA	US State	90001	96962	
CO	US State	80001	81658	
CT	US State	06001	06928	
DC	US District	20001	20599	
DE	US State	19701	19980	
FL	US State	32004	34997	
FM	US Territory	96900	96999	
GA	US State	30002	39901	

OVERLAPPING VALUES. FOR THESE COLUMNS:
- % of customer values in postal table
- % of postal values in customer table



CUSTOMER TABLE

cust_num	last_nm	first_nm	middle_nm	gender	age_grp	income_grp	email	mail_addr	mail_city	state_abbr	zipcode	last_tx_date	email
916232	Morgan	James	Robert	M	2	3	jrm123@gmail.co	6120 Langley Avenu	Key West	FL	33040	2/25/2012	0
916233	Smith	Karen	L	F	5	2	klsmith@earthlin	1885 Desales Stree	Washington	DC	20500	5/15/2008	0
916234	James	Richard	Michael	M	3	41	fatcat@bigdeal.bi	219 Kearns Blvd.	Park City	UT	84068	6/22/2111	0
916235	Jones	Robert		M	5	6	bobjones@comcas	329 Schley Avenue	San Antonio	TX	78210	9/14/2009	1
916236	Black	Linda	C	F	2	6	lcb@primenetwork.	2828 Pineland Drive	Washington	DC	20504	8/18/2011	0
916237	Green	Michael	W	M	4	4	greenguy@ecolabs	16815 NE 80th Stre	Redmond	WA	98052	12/20/2011	
916238	Ford	William	A	M	4	9	willie@fordconsultin	2165 Hanover Squa	New York	NY	10041	11/24/2011	1
916239	Massey	David	L	M	6	12	masseydl@earthlin	151 West 34th Stree	New York	NY	10118	9/19/2010	1
916240	Gonzales	Mia	Magdalena	F		7	mia@silvernet.co	1944 University Driv	Fargo	ND	58102	4/23/2005	0
916241	Berry	Michael	M	M	8	6	cmb@gmx.com	3214 S. 27th Street	South Bend	IN	46556	3/17/2007	1
916242	White	Elizabeth	L	F	4	18	lizwhite@alaskanet	Schenectady		US	12345	9/16/2011	
916243	Wilson	Anthony	M	M	7	8	amwilson@skyline.	516 S. Hanna Stree	Alexandria	IN	46001	7/28/2011	1
916244	Roberts	Thomas	W	M	5		troberts999@gmail.	6550 Pineridge Rd N	Calgary	AB	T1X 1E1	8/20/2012	1
		Lee				14							

ORDER TABLE

order_num	customer_id	receive_date	status	status_date	shi
30552	916236	4/20/2011	cancel	4/20/2011	
30553	916234	6/22/2011	shipped	6/25/2011	5
30554	916235	7/28/2011	shipped	7/29/2011	7
30555	916246	7/28/2011	shipped	8/1/2011	
30558	916246	7/28/2011	cancel	7/28/2011	
30559	916252	8/4/2011	shipped	8/9/2011	
30560	916236	8/12/2011	shipped	8/15/2011	8
30561	916238	9/9/2011	shipped	9/10/2011	9
30563	916246	10/31/2011	shipped	11/4/2011	1
30564	916238	11/24/2011	shipped	11/30/2011	11
30565	916247	12/12/2011	returned	12/20/2011	12
30567	916232	1/20/2012	backorder	1/21/2012	
30568	916244	2/2/2012	open	2/2/2012	

OVERLAPPING VALUES. FOR THESE COLUMNS:
- % of customer values in order table
- % of order values in customer table

Data Profiling

Profiling Techniques

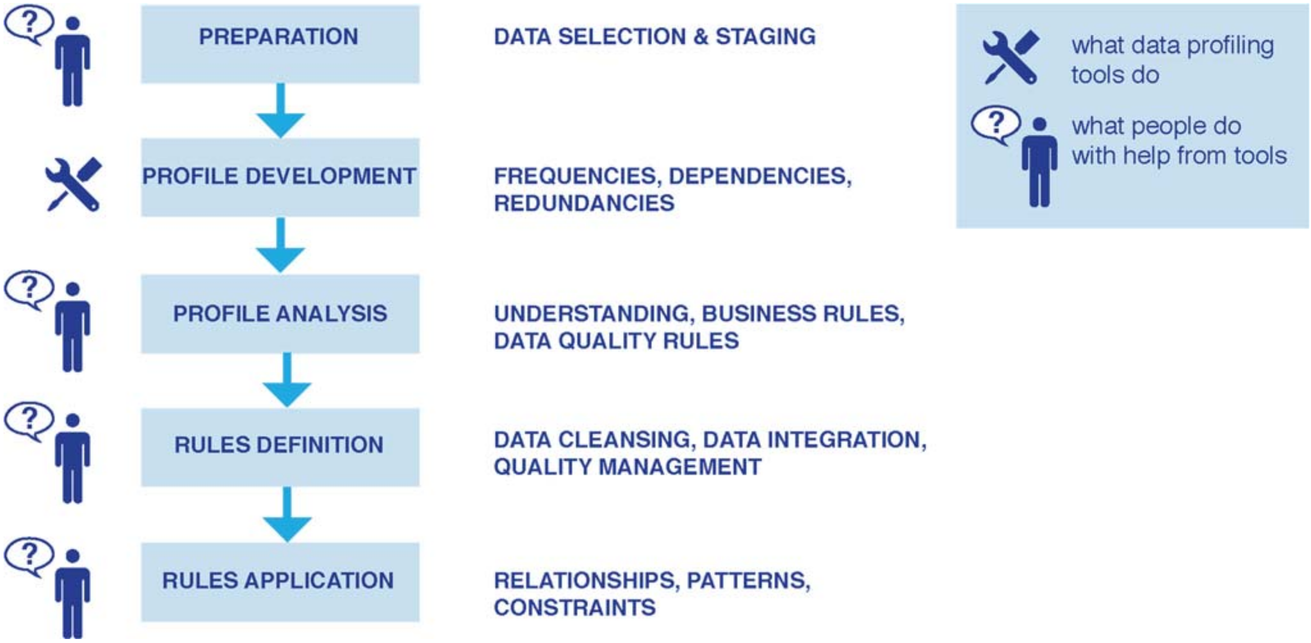
VALUES OVERLAP ACROSS TABLES

Cross-table profiling examines relationships between columns in different tables. Through cross-table profiling you may find foreign key relationships, redundancy, inconsistency, synonymous but differently named columns, similarly named columns with circumstantial differences, and more.

Apparent redundancy of data across multiple tables may indicate repetitive data, similar information under different circumstances, or relationships among tables.

Data Profiling

Tools and Technology



❓	PREPARATION	DATA SELECTION & STAGING
🛠	PROFILE DEVELOPMENT	FREQUENCIES, DEPENDENCIES, REDUNDANCIES
❓	PROFILE ANALYSIS	UNDERSTANDING, BUSINESS RULES, DATA QUALITY RULES
❓	RULES DEFINITION	DATA CLEANSING, DATA INTEGRATION, QUALITY MANAGEMENT
❓	RULES APPLICATION	RELATIONSHIPS, PATTERNS, CONSTRAINTS

🛠 what data profiling tools do

❓ what people do with help from tools

Data Profiling

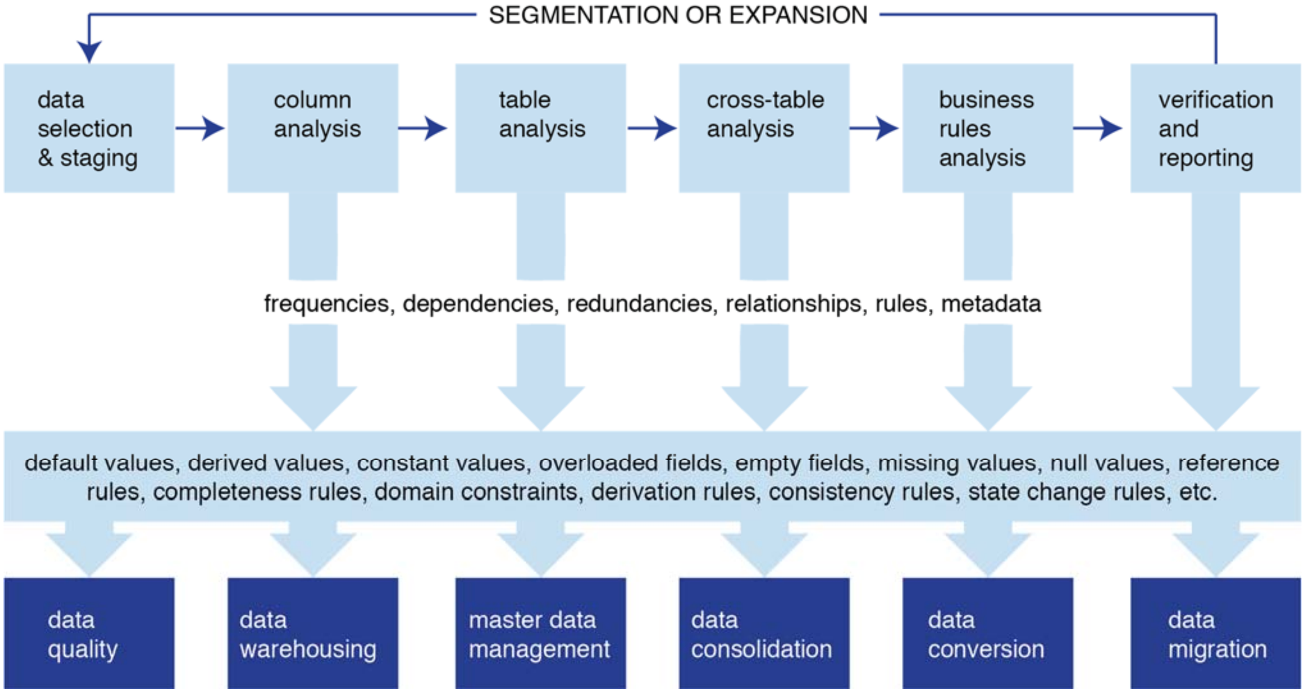
Tools and Technology

PEOPLE AND TECHNOLOGY

Profiling without data profiling tools is difficult, but tools don't perform data profiling. Tools are valuable because they produce the profile metadata—the column profiles, frequency tables, and dependency profiles. However, the metadata is only part of the story. People profile data with help from tools; people are the critical component. They prepare the data, analyze the profiles, interpret the results, and find and apply the rules.

Data Profiling

Application


SEGMENTATION OR EXPANSION
data selection & staging
column analysis
table analysis
cross-table analysis
business rules analysis
verification and reporting
frequencies, dependencies, redundancies, relationships, rules, metadata
default values, derived values, constant values, overloaded fields, empty fields, missing values, null values, reference rules, completeness rules, domain constraints, derivation rules, consistency rules, state change rules, etc.
data quality
data warehousing
master data management
data consolidation
data conversion
data migration

Data Profiling

Application

NEXT STEPS

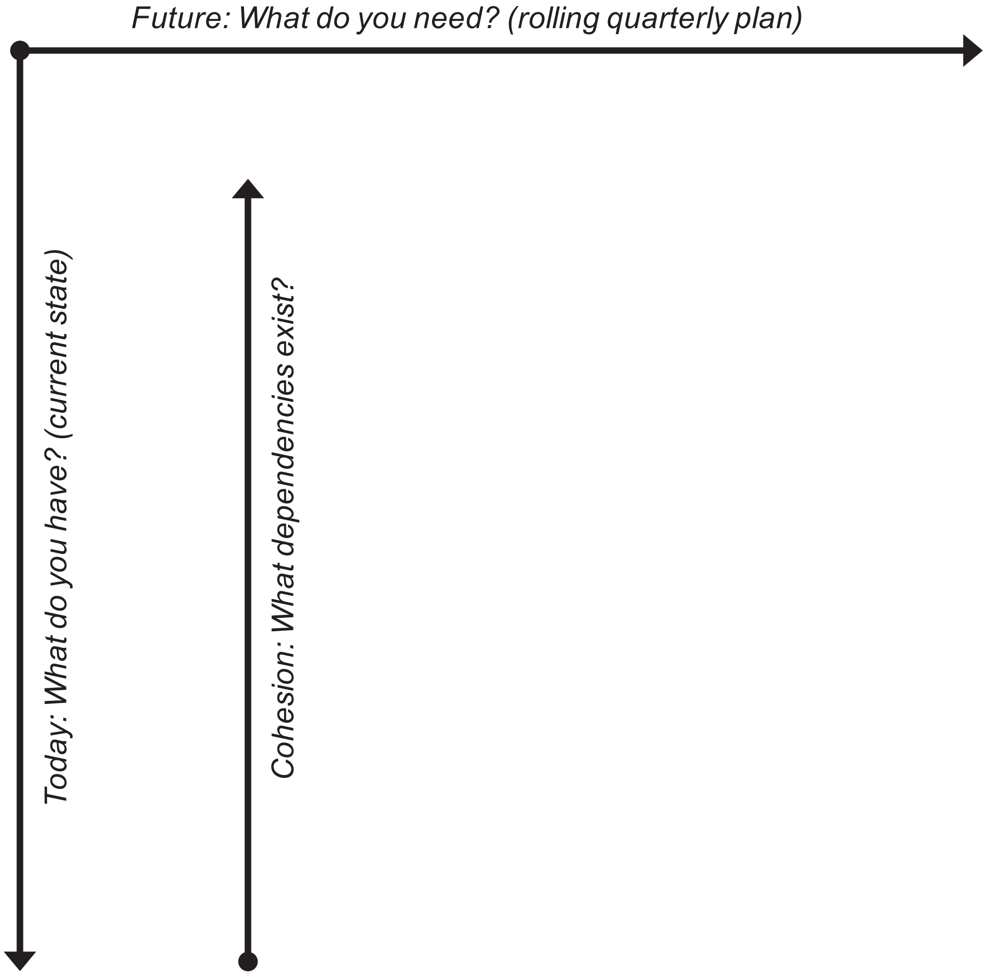
Rarely do we undertake data management projects where data profiling doesn't contribute value. Profiling has direct impact on data quality through data quality projects. It has less direct but very real impact with many other kinds of projects including:

- Data warehousing
- Master data management
- Data consolidation
- Data conversion
- Data migration

Many data quality practitioners believe that consolidation, conversion, and migration are key areas where we introduce data quality defects. If that is true, it is because we perform the work without fully understanding the data. Data profiling can make a real difference to these projects.

The BI and Analytics Roadmap

Data Management


Future: What do you need? (rolling quarterly plan)
Today: What do you have? (current state)
Cohesion: What dependencies exist?

The BI and Analytics Roadmap

Data Management

MAPPING DATA MANAGEMENT FUNCTIONS

The next sections of the BI and analytics roadmap relate to data management. Review your current state inventory and identify future state requirements for the items listed below. Once you determine where you need to be, use that as the basis for an action plan for getting there.

When defining the future state, be sure to include data management needs to support or enable business capabilities, metrics and analytics, information services, and data integration needs that have already been mapped.

Then consider priorities and dependencies as you plot each future state item on the timeline. Be specific about items on the timeline. For example, list *customer data deduplication*, not simply *customer data quality*.

- Data Governance
  - Governance program management
  - Data ownership
  - Data stewardship
- Data Quality
  - Definition
  - Assessment
  - Error Correction
  - Error Prevention
  - Process Improvement
- Data Profiling
  - Source data profiling
  - Profiling of BI and analytics data stores

Mistakes to Avoid

When Creating Your Data Strategy

- THINKING OF YOUR DATA STRATEGY AS A PLAN RATHER THAN A PROCESS
- TREATING DATA STRATEGY AS ASSET MANAGEMENT
- BELIEVING YOUR DATA STRATEGY IS ALL ABOUT IT
- ASSUMING YOUR DATA STRATEGY IS INDEPENDENT FROM TECHNOLOGY
- CREATING A DATA STRATEGY FOCUSED ONLY ON DATA IN ENTERPRISE DATABASES
- VIEWING DATA STRATEGY AS A BUSINESS INTELLIGENCE AND ANALYTICS
- EQUATING DATA STRATEGY WITH DATA MANAGEMENT
- FOCUSING ONLY ON CURRENT REQUIREMENTS
- ATTEMPTING TO EXECUTE THE STRATEGY ALL AT ONCE
- IGNORING CHANGE MANAGEMENT

Source: *Ten Mistakes to Avoid When Creating Your Data Strategy* by Mark Madsen.

Mistakes to Avoid

When Creating Your Data Strategy

TEN MISTAKES TO AVOID FOR DATA MANAGEMENT

Ten Mistakes to Avoid When Creating Your Data Strategy by Mark Madsen.

Strategy is essentially an exercise that connects problems and opportunities with solutions. In today's world, the solutions almost always involve an information technology component, which means using or creating new sources of data. With the advent of so many new sources and the ability to take advantage of them, it's no surprise that it has become fashionable to discuss data strategy.

Data strategy focuses on how data can be used as a resource to further the goals of a business strategy. This means building capabilities: treating data as an asset, organizing to make better use of it, and building the necessary management and technology infrastructures.

There are many ways to build capabilities. Choices impose constraints and trade-offs, which are the essence of crafting a set of policies, procedures, and plans that make up a data strategy. Ten mistakes to avoid when crafting a data strategy are provided on the facing page.

Mistakes to Avoid

When Building a Data Quality Program

- USING EVENTS AND ANECDOTES AS THE SOLE BUSINESS DRIVERS
- APPLYING VALUE JUDGMENTS TO INFORMATION
- FAILING TO EVOLVE FROM A REACTIVE TO A PROACTIVE ENVIRONMENT
- BUYING SOFTWARE FIRST
- IGNORING THE DATA
- NOT ACCOUNTING FOR ORGANIZATIONAL BEHAVIOR
- FAILING TO STANDARDIZE AND MANAGE MASTER REFERENCE DATA
- ISOLATING DATA QUALITY IN THE IT DEPARTMENT
- NOT SECURING THE PROPER EXPERTISE FOR KNOWLEDGE TRANSFER
- FAILING TO BUILD AN ENTERPRISE DATA QUALITY CENTER OF EXCELLENCE

Source: *Ten Mistakes to Avoid When Building a Data Quality Program* by David Loshin. © TDWI

Mistakes to Avoid

When Building a Data Quality Program

TEN MISTAKES TO AVOID FOR DATA QUALITY

Ten Mistakes to Avoid When Building a Data Quality Program by David Loshin.

Enterprise quality improvement programs are rapidly becoming more visible as more reports and articles describe the value placed on high-quality information. C-level executives, concerned with regulatory compliance, are finding themselves personally accountable for both the levels and processes associated with data governance and quality assurance. Once a tedious chore relegated to the back office, data quality is now viewed as an organizational necessity.

Data quality improvement involves more than just name and postal address correction. The complexity and impact of the data quality conundrum grows in proportion to the amount of data we capture, store, manage, review, aggregate, summarize, etc. Yet data quality initiatives are frequently doomed when mistakes are ignored. Ten mistakes to avoid are presented on the facing page.

Discussion

Data Management

LET’S TALK ABOUT IT!

- To what extent does your organization govern data assets?
- Do business areas participate in data governance? Are there assigned data stewards?
- How is your organization assessing its data quality?
- How is data profiling in use within your organization?

Discussion

Data Management

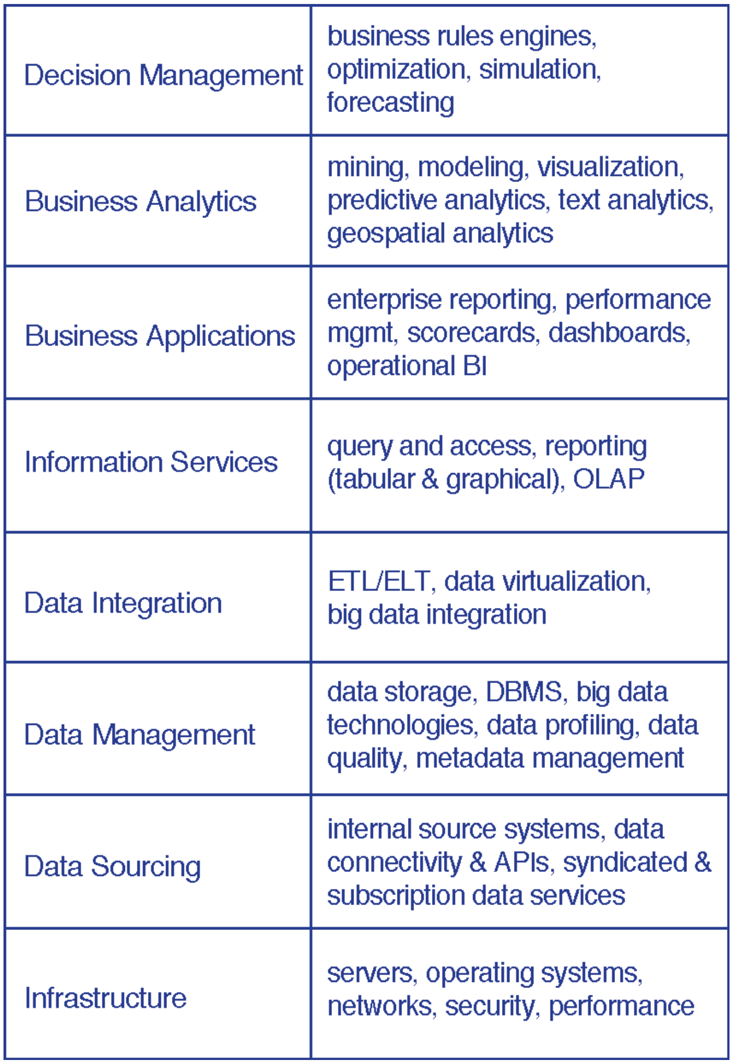
Notes:

Module 6

BI and Analytics Technology

The Technology Stack

Technology Layers



Decision Management	business rules engines, optimization, simulation, forecasting
Business Analytics	mining, modeling, visualization, predictive analytics, text analytics, geospatial analytics
Business Applications	enterprise reporting, performance mgmt, scorecards, dashboards, operational BI
Information Services	query and access, reporting (tabular & graphical), OLAP
Data Integration	ETL/ELT, data virtualization, big data integration
Data Management	data storage, DBMS, big data technologies, data profiling, data quality, metadata management
Data Sourcing	internal source systems, data connectivity & APIs, syndicated & subscription data services
Infrastructure	servers, operating systems, networks, security, performance

The Technology Stack

Technology Layers

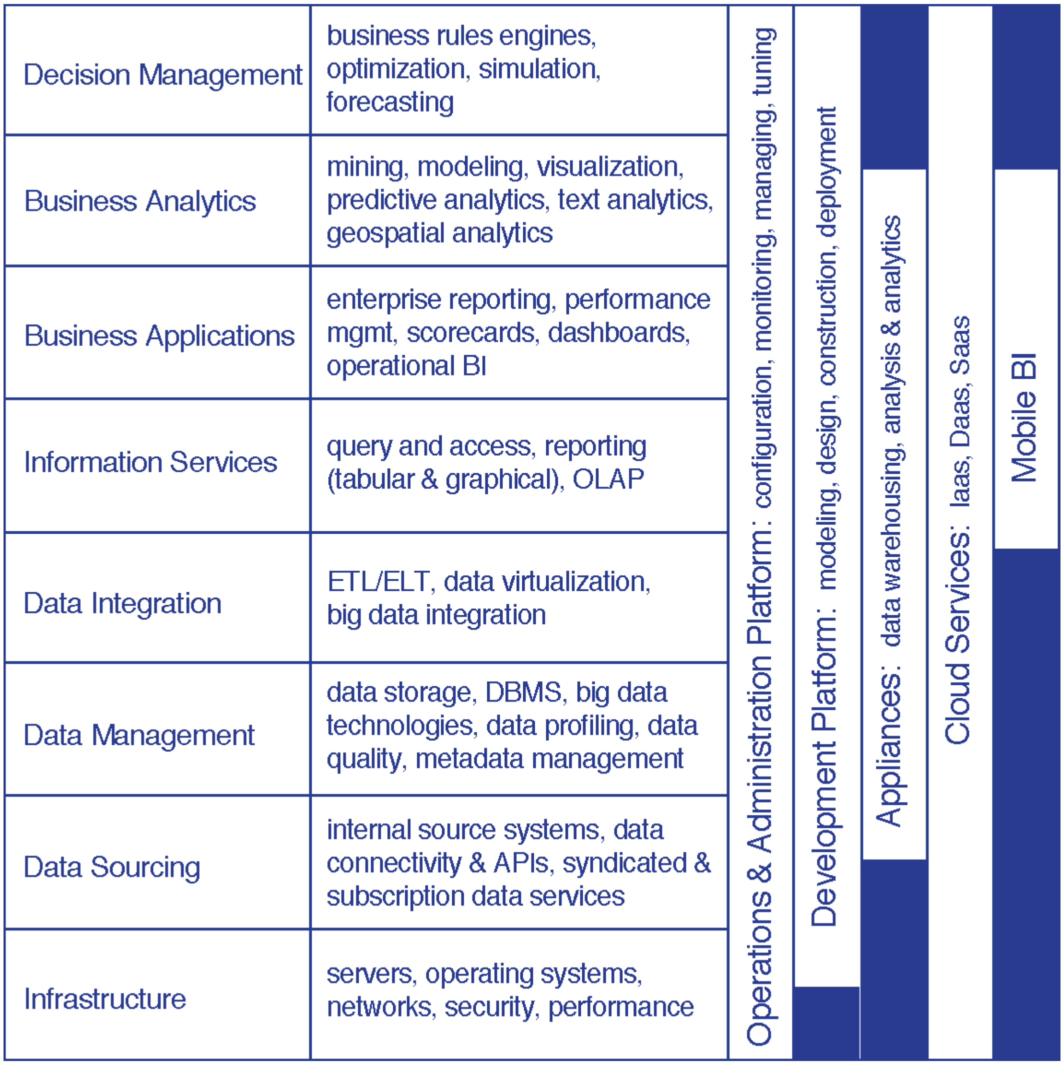
FROM TECHNICAL FOUNDATION TO BUSINESS VALUE

The diagram on the facing page illustrates the BI technology stack. Working from the bottom of the diagram to the top follows a progression from foundation technology to value through business capabilities.

- *Infrastructure* includes all of the hardware and foundation software needed to enable and implement higher-level layers in the stack—the servers, operating systems, networks, security management, and performance optimization.
- *Data sourcing* includes the systems and databases from which data is obtained as well as the languages, protocols, services, and connectors that provide access to data sources.
- *Data management* includes data storage and database management systems, big data technologies such as NoSQL, data profiling tools, data quality and cleansing tools, and metadata management systems and repositories.
- *Data integration* includes ETL and variations such as ELT (extract, load, then transform) for data consolidation, data virtualization tools for federation and unstructured data integration, and technologies for big data integration.
- *Information services technologies* range from query languages to GUI-based query and reporting tools and OLAP technologies.
- *Business applications* technologies enable enterprise reporting, performance management systems, dashboards, scorecards, and operational BI with real-time feedback.
- *Business analytics technologies* are used for data mining, analytics modeling, data visualization, and advanced analytics methods such as text analysis and spatial analysis.
- *Decision management technologies* are needed to implement decision management systems. They include business rules engines and tools to support simulation, forecasting, and process optimization.

The Technology Stack

Functions and Services



Decision Management	business rules engines, optimization, simulation, forecasting
Business Analytics	mining, modeling, visualization, predictive analytics, text analytics, geospatial analytics
Business Applications	enterprise reporting, performance mgmt, scorecards, dashboards, operational BI
Information Services	query and access, reporting (tabular & graphical), OLAP
Data Integration	ETL/ELT, data virtualization, big data integration
Data Management	data storage, DBMS, big data technologies, data profiling, data quality, metadata management
Data Sourcing	internal source systems, data connectivity & APIs, syndicated & subscription data services
Infrastructure	servers, operating systems, networks, security, performance

Operations & Administration Platform: configuration, monitoring, managing, tuning

Development Platform: modeling, design, construction, deployment

Appliances: data warehousing, analysis & analytics

Cloud Services: Iaas, Daas, Saas

Mobile BI

The Technology Stack

Functions and Services

USABLE TECHNOLOGY

Each of the technologies in the stack must provide specific features and functions that make it usable and support people working with the technology.

Required features and services include the following:

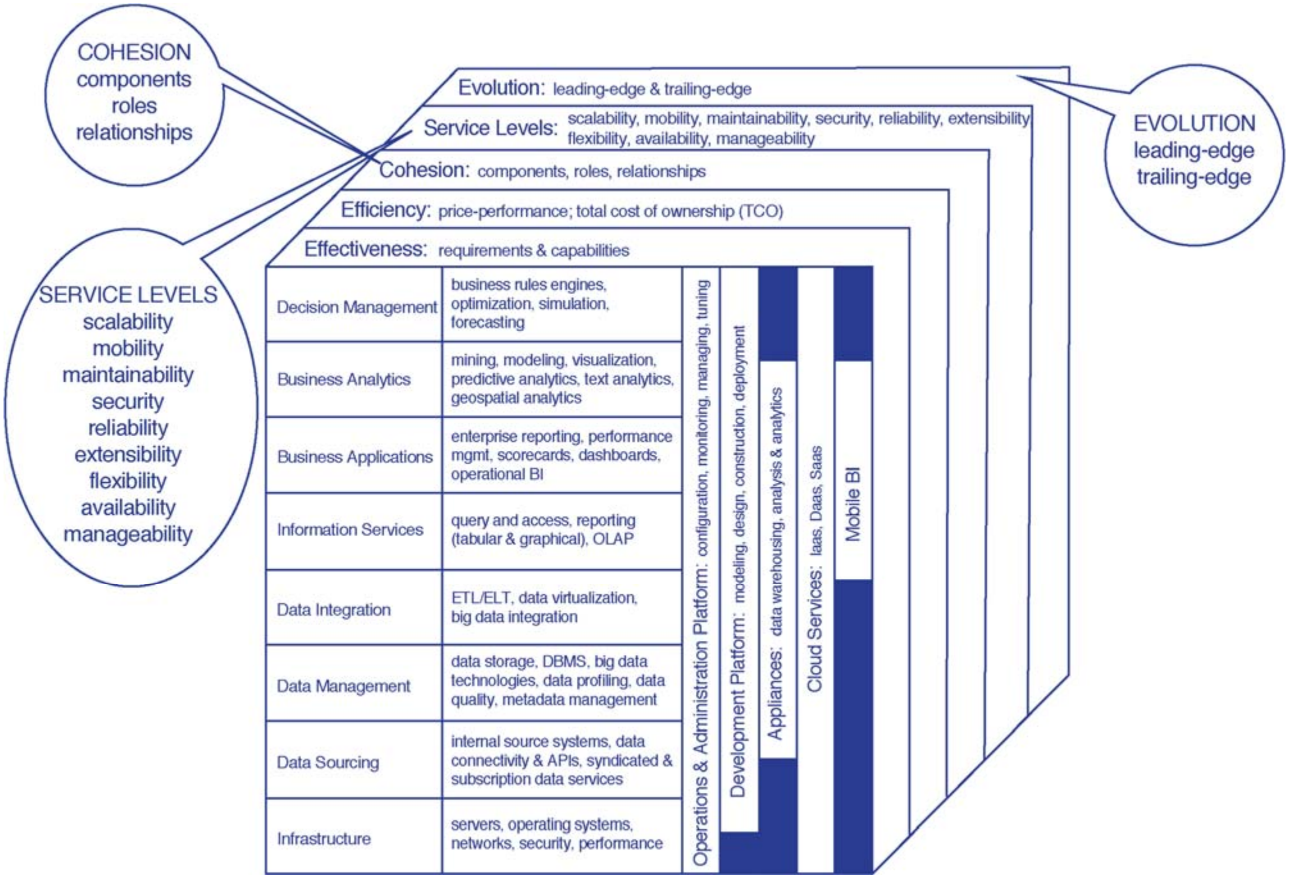
- Operations and administration platforms with functions to monitor, manage, tune, and configure systems and technology at every layer of the stack.
- A development platform (or multiple platforms) that includes functions for modeling, design, construction, and deployment of applied technology systems spanning from data sourcing to decision management.

Desired features and services may include:

- Appliances for data warehousing and/or analytics that provide an integrated set of servers, storage, operating systems, and database management preconfigured for easy installation and high performance. Analytics appliances integrate software for analytics capabilities into the configuration.
- Cloud services to support hosting of infrastructure-as-a-service (IaaS) for servers and operating systems, data-as-a-service (DaaS) for storage and DBMS, or software-as-a-service (SaaS) for specific information services and business applications.
- Mobile BI with communications, compression, visualization, and other features needed to deliver information services and business applications to smartphones and tablets.

Technology Architecture

The Right Technology—Present and Future



COHESION
components
roles
relationships

SERVICE LEVELS
scalability
mobility
maintainability
security
reliability
extensibility
flexibility
availability
manageability

EVOLUTION
leading-edge
trailing-edge

Evolution: leading-edge & trailing-edge

Service Levels: scalability, mobility, maintainability, security, reliability, extensibility flexibility, availability, manageability

Cohesion: components, roles, relationships

Efficiency: price-performance; total cost of ownership (TCO)

Effectiveness: requirements & capabilities

Decision Management	business rules engines, optimization, simulation, forecasting
Business Analytics	mining, modeling, visualization, predictive analytics, text analytics, geospatial analytics
Business Applications	enterprise reporting, performance mgmt, scorecards, dashboards, operational BI
Information Services	query and access, reporting (tabular & graphical), OLAP
Data Integration	ETL/ELT, data virtualization, big data integration
Data Management	data storage, DBMS, big data technologies, data profiling, data quality, metadata management
Data Sourcing	internal source systems, data connectivity & APIs, syndicated & subscription data services
Infrastructure	servers, operating systems, networks, security, performance

Operations & Administration Platform: configuration, monitoring, managing, tuning

Development Platform: modeling, design, construction, deployment

Appliances: data warehousing, analysis & analytics

Cloud Services: Iaas, Daas, Saas

Mobile BI

Technology Architecture

The Right Technology—Present and Future

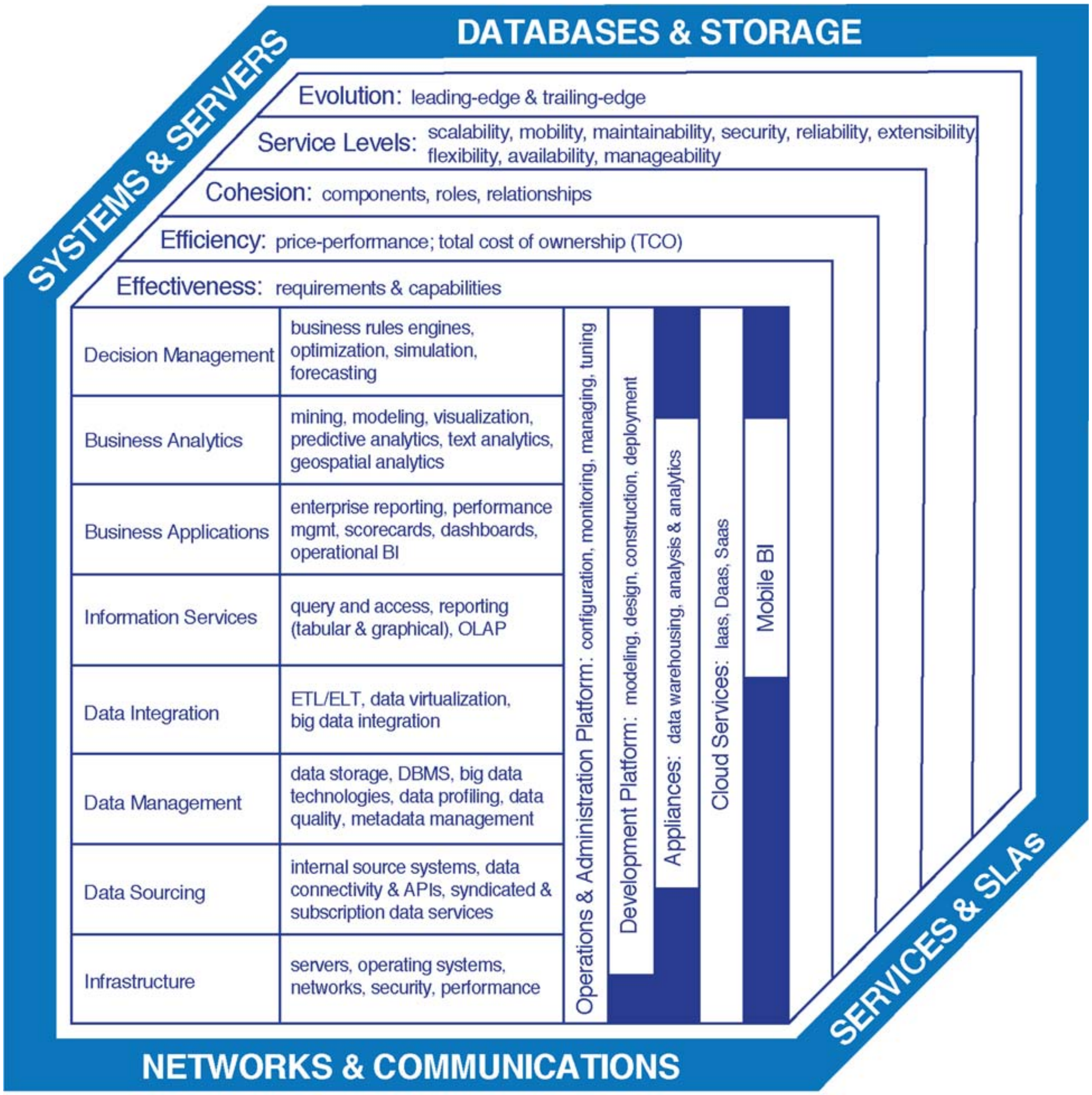
SUSTAINABLE TECHNOLOGY

Technology architecture is needed to ensure that the technology stack will adapt to change and can be sustained throughout the lifespan of enterprise BI and analytics systems. Architectural considerations and responsibilities include:

- Effectiveness of technology—the ability to meet BI requirements and enable needed business capabilities.
- Efficiency of technology including price/performance ratio and total cost of ownership (TCO).
- Cohesion of technology that is well-integrated or compatible with the right components, performing in the right roles, and interconnected in the right ways.
- Service levels committed and delivered for scalability, mobility, maintainability, security, reliability, extensibility, flexibility, availability, manageability, and more. Wherever a business expectation for quality of service in BI exists, a corresponding service level responsibility exists.
- Evolution of the technology stack with attention both to extending the leading edge of technology and picking up the trailing edge. When we fail to pull the trailing edge forward, the range of technologies expands and the degree of integration and cohesion erodes. Deteriorating cohesion of technology ultimately affects service levels, user satisfaction, and business value.

Technology Management

Reliable Platforms



DATABASES & STORAGE

SYSTEMS & SERVERS

SERVICES & SLAs

NETWORKS & COMMUNICATIONS

Evolution: leading-edge & trailing-edge

Service Levels: scalability, mobility, maintainability, security, reliability, extensibility flexibility, availability, manageability

Cohesion: components, roles, relationships

Efficiency: price-performance; total cost of ownership (TCO)

Effectiveness: requirements & capabilities

Decision Management	business rules engines, optimization, simulation, forecasting
Business Analytics	mining, modeling, visualization, predictive analytics, text analytics, geospatial analytics
Business Applications	enterprise reporting, performance mgmt, scorecards, dashboards, operational BI
Information Services	query and access, reporting (tabular & graphical), OLAP
Data Integration	ETL/ELT, data virtualization, big data integration
Data Management	data storage, DBMS, big data technologies, data profiling, data quality, metadata management
Data Sourcing	internal source systems, data connectivity & APIs, syndicated & subscription data services
Infrastructure	servers, operating systems, networks, security, performance

Operations & Administration Platform: configuration, monitoring, managing, tuning

Development Platform: modeling, design, construction, deployment

Appliances: data warehousing, analysis & analytics

Cloud Services: Iaas, Daas, Saas

Mobile BI

Technology Management

Reliable Platforms

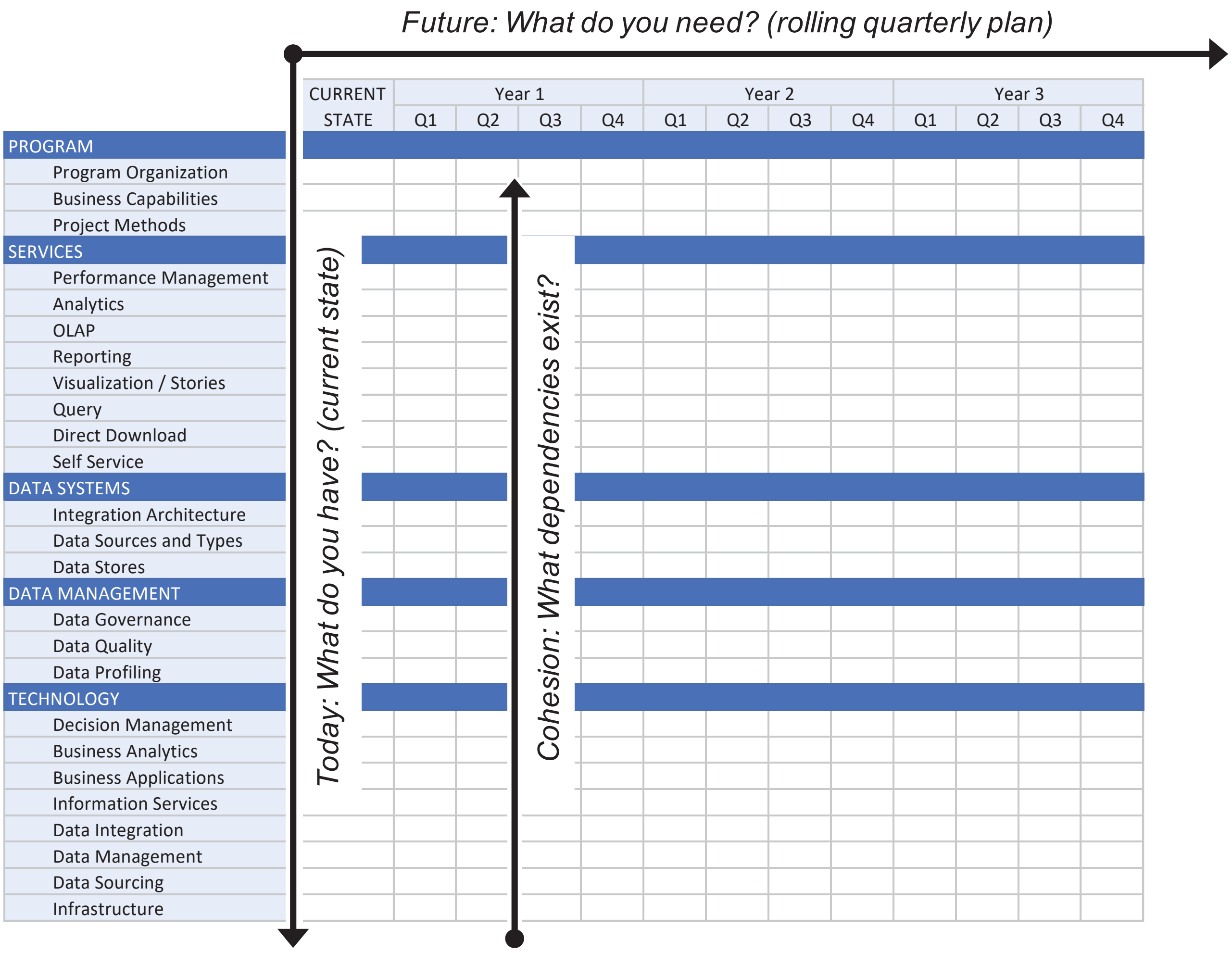
KEEPING THE LIGHTS ON

Although architecture is essential to sustainable technology, day-to-day management is equally important. Reliable BI and analytics platforms require regular and routine maintenance of:

- Systems and servers—Systems administration and systems management activities for servers, operating systems, etc., including capacity and growth management.
- Networks and communications—Network administration and management activities for communications hardware and software including capacity and growth management.
- Databases and storage—Database administration and storage management including capacity and growth management.
- Services and SLAs—Active monitoring and oversight of internal service level agreements; relationship management with external service providers; provisioning and service management for cloud services, subscription services, and syndicated data services, etc.

The BI and Analytics Roadmap

Technology



Future: What do you need? (rolling quarterly plan)

Today: What do you have? (current state)

Cohesion: What dependencies exist?

	CURRENT	Year 1				Year 2				Year 3			
	STATE	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
PROGRAM													
Program Organization													
Business Capabilities													
Project Methods													
SERVICES													
Performance Management													
Analytics													
OLAP													
Reporting													
Visualization / Stories													
Query													
Direct Download													
Self Service													
DATA SYSTEMS													
Integration Architecture													
Data Sources and Types													
Data Stores													
DATA MANAGEMENT													
Data Governance													
Data Quality													
Data Profiling													
TECHNOLOGY													
Decision Management													
Business Analytics													
Business Applications													
Information Services													
Data Integration													
Data Management													
Data Sourcing													
Infrastructure													

The BI and Analytics Roadmap

Technology

MAPPING THE TECHNOLOGY

The facing page illustrates the parts of the BI and analytics roadmap that relate to technology. Review the current state inventory and identify future state requirements for the entire technology stack. Plot technology on the roadmap as needed to support planned capabilities, services, systems, and projects.

The BI and analytics roadmap now represents a logical progression from need for business capabilities through layers of services, systems, and the technology that is needed. It is a plan for continuous and incremental evolution of BI and analytics systems and growth of BI maturity.

Mistakes to Avoid

When Adopting New Technologies in BI

- CONFUSING PRODUCT COMBINATIONS WITH EMERGING TECHNOLOGIES
- FAILING TO ARTICULATE VALUE TO THE BUSINESS SPONSORS
- NOT UNDERSTANDING THE ROLE OF TECHNOLOGY IN THE ECOSYSTEM
- ADOPTING AN EMERGING TECHNOLOGY BECAUSE OF MARKETING HYPE
- UNDERESTIMATING THE MATURITY AND LEVEL OF EFFORT
- FAILING TO RECOGNIZE THAT EMERGING TECHNOLOGY SKILLS MAY BE SCARCE
- NOT PERFORMING COMPANY BACKGROUND DUE DILIGENCE
- FAILING TO CAREFULLY DESIGN A PROCESS FOR EVALUATION AND IMPLEMENTATION
- NOT PLANNING FOR TECHNOLOGY INTEGRATION AND IMPACT
- NOT BALANCING NEW TECHNOLOGY EVALUATIONS WITH ONGOING DELIVERY

Source: *Ten Mistakes to Avoid When Adopting New Technologies in BI* by John O’Brien. © TDWI

Mistakes to Avoid

When Adopting New Technologies in BI

TEN MISTAKES TO AVOID FOR NEW TECHNOLOGY

Ten Mistakes to Avoid When Adopting New Technologies in BI by John O'Brien.

Research has shown that the top-performing companies in nearly every industry share the characteristics of being innovative and adopting emerging technologies into the fabric of their business DNA. Companies that can effectively adopt and leverage emerging technologies can create innovative new products and become more efficient in managing costs and processes. These organizations realize tangible value that increases their competitive advantage and secures their position as leaders in their markets. However, you should be aware of the risks and mistakes commonly encountered when evaluating and adopting a new technology. A few wrong choices or bad decisions can cost your business a great deal of time, money, and resources, cause undue distraction, and lower morale.

Awareness and insights to mitigate and overcome common mistakes will prepare you for navigating emerging technology options and determining best-fit solutions for your organization's needs.

Mistakes to Avoid

In Hadoop Implementations

- THINKING TECHNOLOGY IS A SILVER BULLET
- ADOPTING A “STORE FIRST, THEN ANALYZE” WORKFLOW
- FAILURE TO PREVENT TRANSFORMATION OVERLOAD
- FAILURE TO SECURE EXECUTIVE SPONSORSHIP
- FAILURE TO ESTABLISH GOALS
- IMPROPER INFRASTRUCTURE PLANNING
- FAILURE TO PROPERLY CONFIGURE AND MANAGE SEMANTIC DATA
- MAPREDUCE IMPLEMENTATION
- FAILURE TO PROVIDE BIG DATA GOVERNANCE
- USING HADOOP AS AN ENTERPRISE DATA REPOSITORY

Source: *Ten Mistakes to Avoid in Hadoop Implementations* by Krish Krishnan. © TDWI

Mistakes to Avoid

In Hadoop Implementations

TEN MISTAKES TO AVOID IN HADOOP IMPLEMENTATIONS

Ten Mistakes to Avoid in Hadoop Implementations by Krish Krishnan.

Data management and analytics are foundational requirements for creating, managing, and executing a successful business. From an infrastructure perspective, however, it is a struggle to build an integrated data platform that can support the information architecture required by an enterprise data repository and analytics hub.

In the past decade, we have seen a successful set of distributed processing architectures—including Google and Nutch—that inspired us to bring distributed data processing architecture with Hadoop and its ecosystem of projects. Enterprises have explored Hadoop since 2009, and many are now focusing on that ecosystem.

Today, this infrastructure distribution is being implemented as the enterprise hub for all data; some implementations are successful, but many others are abysmal failures. When inspecting failures and listening to companies and teams, we see that fundamental steps have been missed or ignored, including end-user management, data security, performance tuning, infrastructure configuration, and sizing. From the Hadoop infrastructure perspective, simply applying workarounds to implementation doesn't work. Ten mistakes to avoid are provided on the facing page.

Discussion

BI and Analytics Technology

LET’S TALK ABOUT IT!

- What technologies do you currently have in place?
- How well do they support the other layers of the roadmap (data management, data integration, BI and analytics services, business capabilities)?
- To what extent is your architecture defined?
- How are you ensuring the sustainability of your ecosystem?

Discussion

BI and Analytics Technology

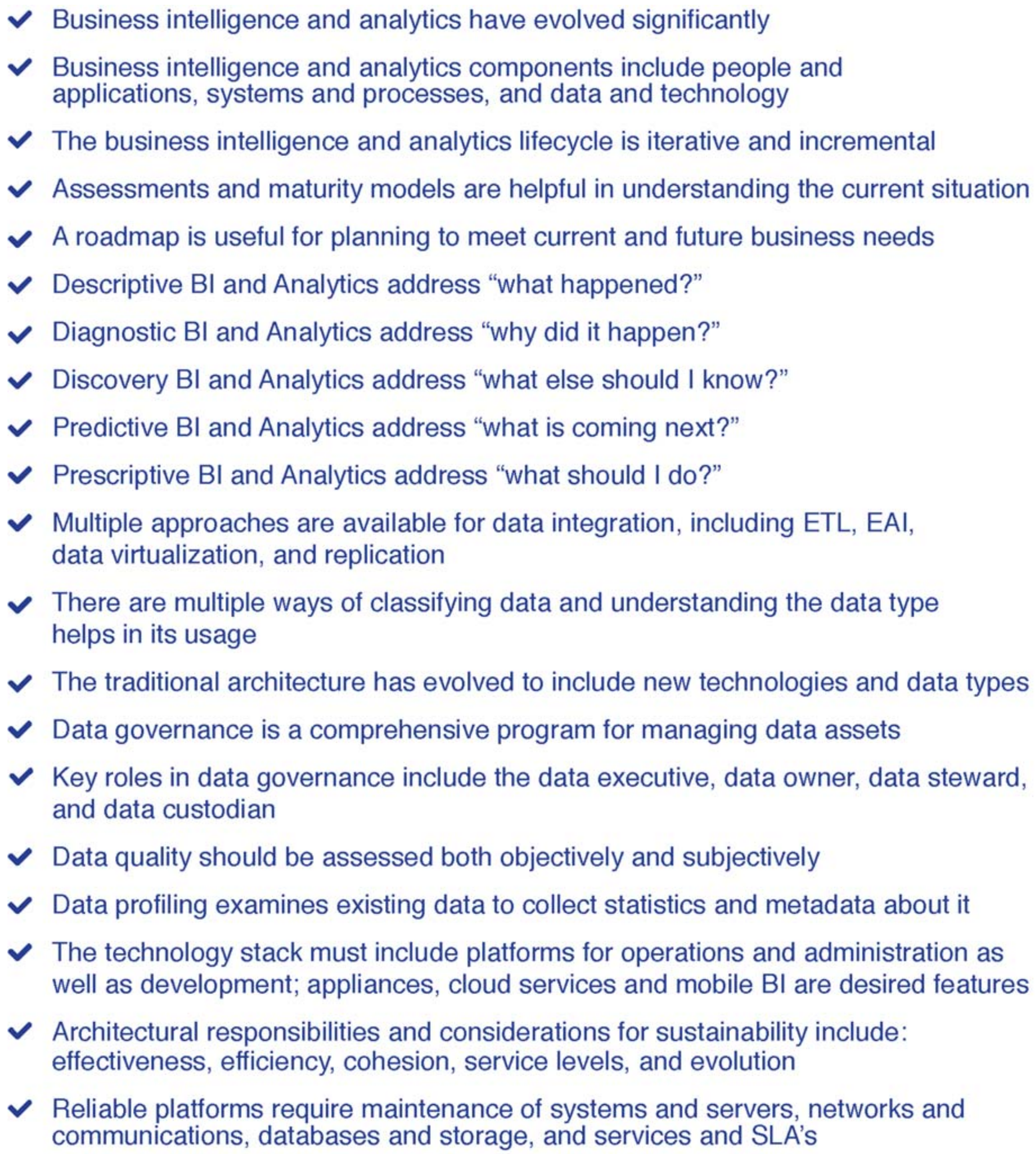
Notes:

Module 7

Summary

Summary

Key Points



✓ Business intelligence and analytics have evolved significantly

✓ Business intelligence and analytics components include people and applications, systems and processes, and data and technology

✓ The business intelligence and analytics lifecycle is iterative and incremental

✓ Assessments and maturity models are helpful in understanding the current situation

✓ A roadmap is useful for planning to meet current and future business needs

✓ Descriptive BI and Analytics address “what happened?”

✓ Diagnostic BI and Analytics address “why did it happen?”

✓ Discovery BI and Analytics address “what else should I know?”

✓ Predictive BI and Analytics address “what is coming next?”

✓ Prescriptive BI and Analytics address “what should I do?”

✓ Multiple approaches are available for data integration, including ETL, EAI, data virtualization, and replication

✓ There are multiple ways of classifying data and understanding the data type helps in its usage

✓ The traditional architecture has evolved to include new technologies and data types

✓ Data governance is a comprehensive program for managing data assets

✓ Key roles in data governance include the data executive, data owner, data steward, and data custodian

✓ Data quality should be assessed both objectively and subjectively

✓ Data profiling examines existing data to collect statistics and metadata about it

✓ The technology stack must include platforms for operations and administration as well as development; appliances, cloud services and mobile BI are desired features

✓ Architectural responsibilities and considerations for sustainability include: effectiveness, efficiency, cohesion, service levels, and evolution

✓ Reliable platforms require maintenance of systems and servers, networks and communications, databases and storage, and services and SLA’s

Appendix A

Bibliography and References

Bibliography and References

Business Dashboards: A Visual Catalog for Design and Deployment, Rasmussen, Chen, and Bansal
John Wiley & Sons, (2009)

Business Intelligence, Second Edition: The Savvy Manager's Guide, Loshin
Morgan Kaufman, (2013)

Business Intelligence Success Factors, Parr Rud
John Wiley & Sons, (2009)

Corporate Information Factory, Inmon, Imhoff, and Sousa
John Wiley & Sons, (2000)

Data Architecture: From Zen to Reality, Tupper
Elseveir, (2011)

The Data Governance Imperative: A Business Strategy for Corporate Data, Sarsfield
IT Governance Publishing, (2009)

Data Mining Techniques, Second Edition, Berry and Linoff
John Wiley & Sons, (2004)

Data Strategy, Adelman, Moss, and Abai
Addison-Wesley, (2005)

Data Virtualization: Going Beyond Traditional Data Integration to Achieve Business Agility,
Davis and Eve
Composite Software, (2011)

The Data Warehouse Lifecycle Toolkit, Kimball, Reeves, Ross, and Thornthwaite
John Wiley & Sons, (1998)

Decision Management Systems: A Practical Guide to Using Business Rules and Predictive Analytics, Taylor
Pearson Education, (2011)

Decision Management Systems Platform Technologies Report Version 2, Update 4, Taylor
Decision Management Solutions, (2012)

Bibliography and References

Decision Making, Rowe
Harvard Business School Press, (2006)

Executing Data Quality Projects, McGilvray
Morgan Kaufman, (2008)

Information Dashboard Design: The Effective Visual Communication of Data, Few
O'Reilly Media, (2006)

A Manager's Guide to Data Warehousing, Reeves
John Wiley & Sons, (2009)

Managing Your Business Data, Kushner and Villar
Racom Books, (2009)

Modeling for lnsight, Powell and Batt
Wiley, (2008)

Performance Dashboards: Measuring, Monitoring and Managing Your Business, Eckerson
John Wiley & Sons, (2006)

Performance Leadership: The Next Practices to Motivate Your People, Align Stakeholders, and Lead Your Industry, Buytendijk
McGraw-Hill, (2009)

Performance Management: Integrating Strategy Execution, Methodologies, Risk, and Analytics, Cokins
McGraw-Hill, (2009)

Tapping Into Unstructured Data, Inmon and Nesavich
Prentice Hall, (2007)

Three Dimensional Analysis: Data Profiling Techniques, Lindsey
Data Profiling LLC, (2008)

This page intentionally left blank.