Business Intelligence Summary

written by

KNiemeijer

The Marketplace to Buy and Sell your Study Material

On Stuvia you will find the most extensive lecture summaries written by your fellow students. Avoid resits and get better grades with material written specifically for your studies.

www.stuvia.com

BUSINESS INTELLIGENCE

BUSINESS INTELLIGENCE, ANALYTICS, AND DATA SCIENCE – A MANAGERIAL PERSPECTIVE

Ramesh Sharda, Dursun Delen, and Efraim Turban

Fourth Edition

Summary of Business Intelligence 2018 course (INFOMBIN) at Utrecht University

Koen Niemeijer

Contents

1. An Overview of Business Intelligence, Analytics, and Data Science

1.1 OPENING VIGNETTE: Sports Analytics—An Exciting Frontier for Learning and Understanding Applications of Analytics

Examples illustrating possible use of business intelligence (BI). Interesting read, but not much to learn from. This applies to the rest of the opening vignettes as well and will henceforth be skipped.

1.2 Changing Business Environments and Evolving Needs for Decision Support and Analytics

Because of fast-paced changing business environments, information technology is vital to support decision making. Some developments have clearly contributed to facilitating growth of decision support and analytics in a number of ways, including the following:

- **Group communication and collaboration**: Information systems can improve the collaboration process of a group and enable its members to be at different locations (saving travel costs) so that decisions can be made together.
- **Improved data management**: Data for these can be stored in different databases anywhere in the organisation and even possibly outside the organisation and systems can search, store, and transmit needed data quickly, economically, securely, and transparently.
- **Managing giant data warehouses and Big Data**: Data warehouses (DWs) contain enormous amounts of data. Big Data has enabled massive data coming from a variety of sources and in many different forms, which allows a very different view into organisational performance that was not possible in the past.
- **Analytical support**: Analysis technologies allow for quicker and better execution of processes at a reduced cost. Expertise can even be derived directly from analytical systems.
- **Overcoming cognitive limits in processing and storing information**: People can only store and processes so much. This is known as their cognitive limits. Computerised systems enable people to overcome their cognitive limits by quickly accessing and processing vast amounts of stored information.
- **Knowledge management**: Gathered formal and informal sources support decision making in organisations.
- **Anytime, anywhere support:** Ubiquitous support increases speed at which information needs to be processed and converted into decisions.

1.3 Evolution of Computerized Decision Support to Analytics/Data Science

Data analytics started simple with **management information systems (MIS)** and **decision support systems (DSS)**. As systems became more advanced, mature **operations research (OR**) models and **rule-based expert systems (ES)** became

prevalent. The 1980s was all about structuring data into centralised solutions such as **enterprise resource planning (ERP)** systems and **relational database management (RDBM)** systems. In the 1990s, the need for more versatile reporting led to the development of **executive information systems (EISs)**. To make this highly versatile reporting possible while keeping the transactional integrity of the business information systems intact, it was necessary to create a middle data tier known as a **data warehouse (DW)** as a repository to specifically support business reporting and decision making. In the 2000s, these were called **business intelligence (BI)** systems. Because of the globalised competitive marketplace, decision makers needed to acquire data more quickly. This led to the development of **right-time data warehousing**. Nowadays, **Big Data** drives decision making with **social networking/social media** as an interesting data source.

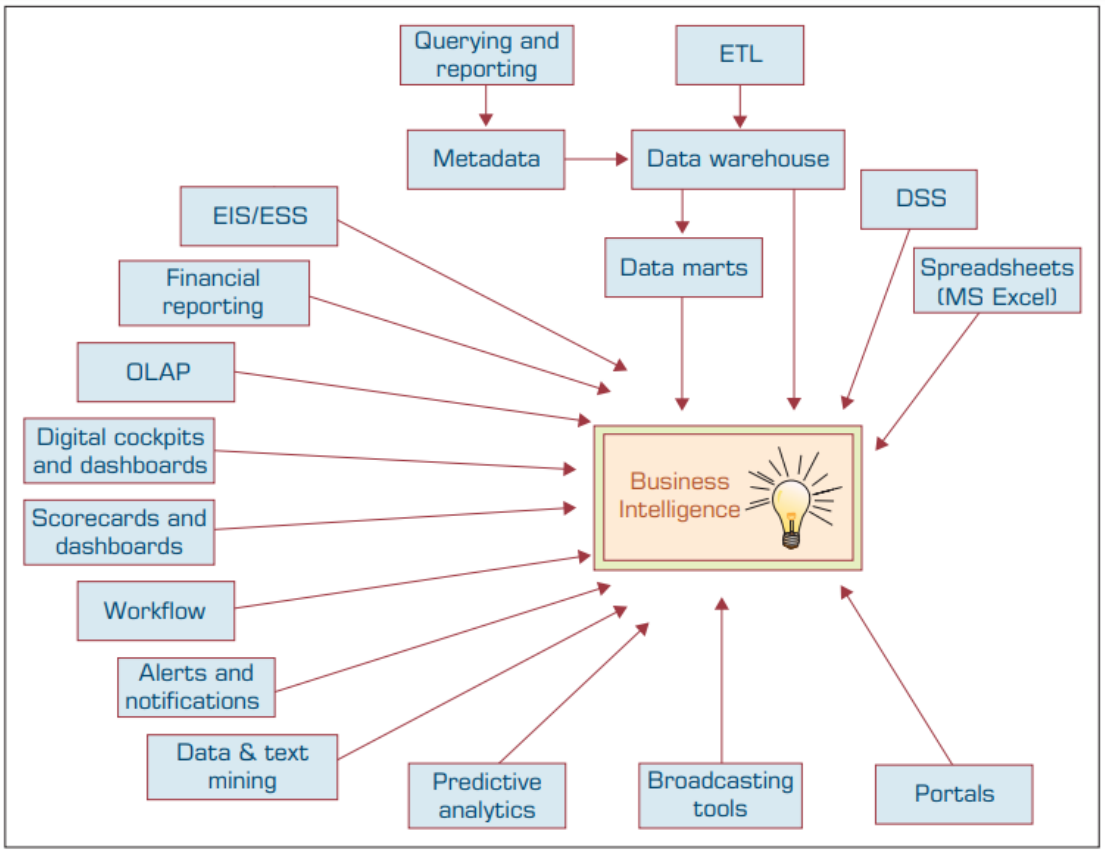


FIGURE 1.9 Evolution of Business Intelligence (BI).

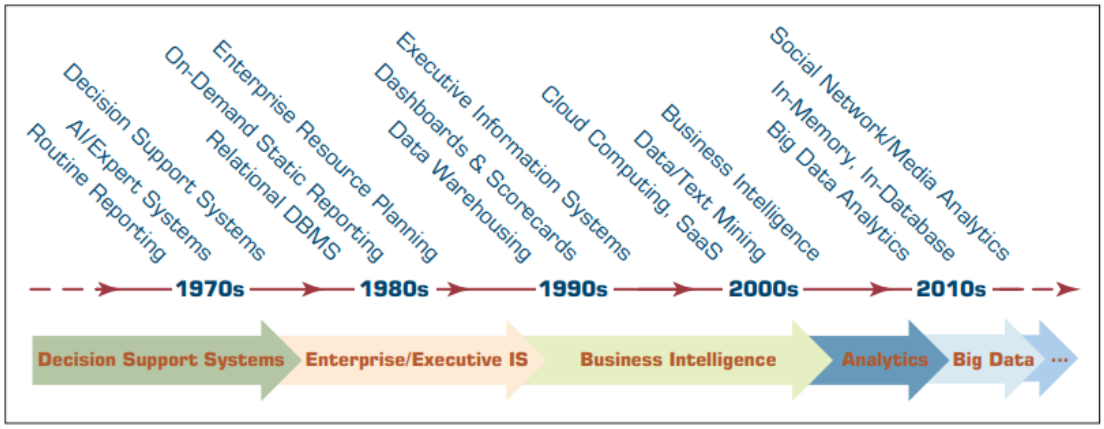


FIGURE 1.8 Evolution of Decision Support, Business Intelligence, and Analytics.

1.4 A Framework for Business Intelligence

Business intelligence (BI) is an umbrella term that combines architectures, tools, databases, analytical tools, applications, and methodologies. BI's major objective is to enable interactive access (sometimes in real time) to data, to enable manipulation of data, and to give business managers and analysts the ability to conduct appropriate analyses. The process of BI is based on the **transformation** of data to information, then to decisions, and finally to actions.

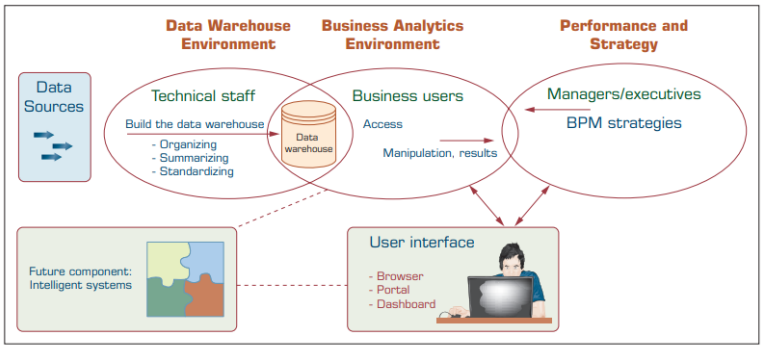


FIGURE 1.10 A High-Level Architecture of BI. (*Source:* Based on W. Eckerson, *Smart Companies in the 21st Century: The Secrets of Creating Successful Business Intelligent Solutions.* The Data Warehousing Institute, Seattle, WA, 2003, p. 32, Illustration 5.)

A BI system has four major components: a **DW**, with its source data; **business analytics**, a collection of tools for manipulating, mining, and analysing the data in the DW; **Business Performance Management (BPM)** for monitoring and analysing performance; and a **user interface** (e.g., a dashboard).

BI is not **transaction processing**. That is, processing transactions such as payments, withdrawals., or calculations of sales. These are handled by **online transaction processing (OLTP)**. DW contain data used in analysis. The results of these analyses can then be used for decision making. DWs are intended to work with informational data used for **online analytical processing (OLAP)** systems. Most ERP and **supply chain management (SCM)** systems are stored in OTLP where each request is considered to be a transaction.

In order to implement BI in an organisation, it must align with some business goal. A framework developed by Gartner (2004) decomposes planning and execution into **business**, **organisation**, **functionality**, and **infrastructure** components. That is, strategic and operational objectives must be defined while considering the available organisational skills to achieve those objectives, plans and assessments need to be made to make the change, a company

must be amenable to change, and various systems must be integrated for successful implementation. When everything is in place, BI should be started and a BI Competency Centre should be established within the company.

Another important success factor of BI is its ability to facilitate a real-time, on-demand agile environment. Whereas traditional DWs use static data, real-time, on-demand BI uses near-real-time data. One approach is the DW model of traditional BI systems. Products from innovative BI platform providers provide a service-oriented, near-real-time solution that populates the DW much faster than the typical nightly extract/ transfer/load batch update does. Another approach, **business activity management (BAM)**, bypasses the DW entirely and uses **Web services** or other monitoring means to discover key business events. These software monitors (or **intelligent agents**) can be placed on server or application to trigger events themselves.

1.5 Analytics Overview

Analytics is the process of developing actionable decisions or recommendations for actions based on insights generated from historical data. This idea of looking at all the data to understand what is happening, what will happen, and how to make the best of it has also been encapsulated by the Institute for Operations Research and Management Science (INFORMS). These three levels are identified as **descriptive**, **predictive**, and **prescriptive**. in proposing three levels of analytics.

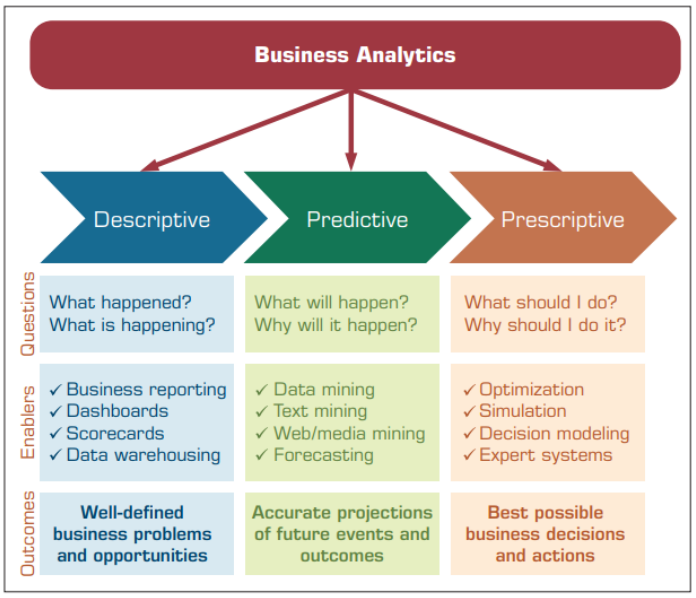


FIGURE 1.11 Three Types of Analytics.

Descriptive (or reporting) analytics refers to knowing what is happening in the organisation and understanding some underlying trends and causes of such occurrences. This involves data consolidation for developing reports and queries, for example, but also data visualisation.

Predictive analytics aims to determine what is likely to happen in the future. This analysis is based on statistical techniques as well as other more recently developed techniques that fall under the general category of **data mining**. The goal of these techniques is to be able to predict if the customer is likely to switch to a competitor ("**churn**") or to predict other behavioural patterns using, for example, logistic regression, clustering algorithms, or association mining techniques.

The goal of **prescriptive analytics** is to recognise what is going on as well as the likely forecast and make decisions to achieve the best performance possible. The goal here is to provide a decision or a recommendation for a specific action. The decisions may be presented to a decision maker in a report or may be used directly in an automated decision rules. Thus, these types of analytics can also be termed **decision** or **normative analytics**.

The word *analytics* may be attached to any specific industry or type of data. Literally, any systematic analysis of data in a specific sector is being labelled as "(fill-in-blanks)" analytics, such as sport analytics, market analytics, or health analytics. There are unique issues within each vertical segment (type of analytics) that influence how the data may be collected, processed, analysed, and the applications implemented. Thus, the differentiation of analytics based on a vertical focus is good for the overall growth of the discipline.

Analytics can also be called **data science**. In this field, there are **data analysts** who are concerned with data compilation, cleaning, reporting, and perhaps some visualisation (descriptive or reporting analytics) and **data scientists** responsible for predictive analysis, statistical analysis, and more advanced analytical tools and algorithms (predictive and prescriptive analytics).

1.6 Analytics Examples in Selected Domains

This sections describes an example of a health organisation and a generic retail value chain. The case study of the health organisation is about predicting client behaviour and defining rigorous metrics and predictive models (PMs). After that, it is explained that retail organisations need analytics to cope with enormous amounts of data and to stay ahead of competition by predicting continuously changing customer preferences.

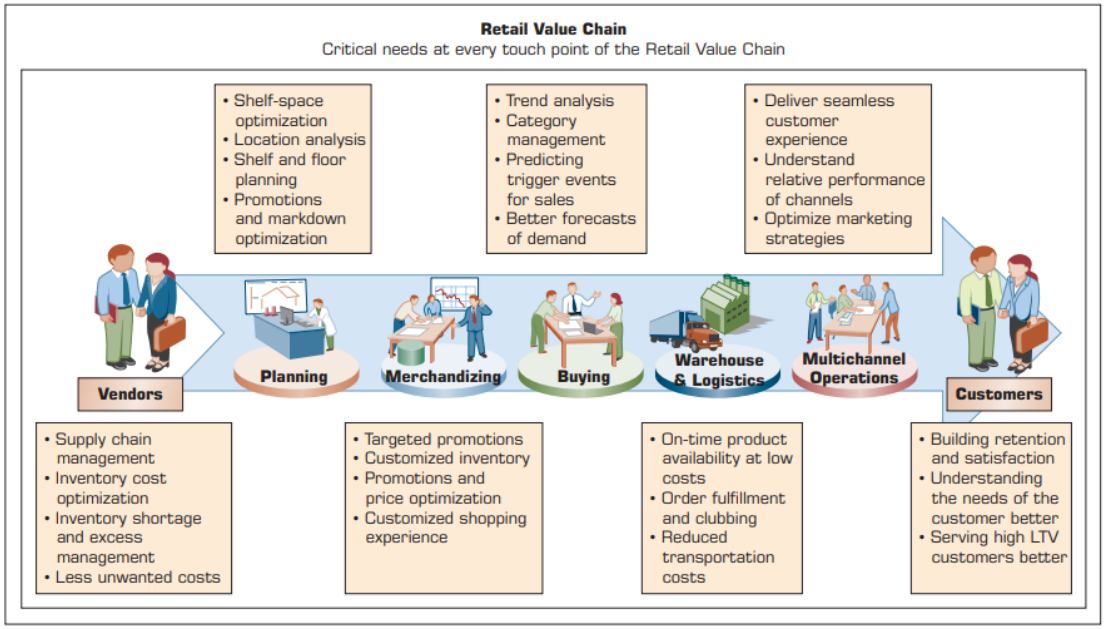


FIGURE 1.12 Example of Analytics Applications in a Retail Value Chain. Contributed by Abhishek Rathi, CEO, vCreaTek.com

TABLE 1.1 Examples of Analytics Applications in the Retail Value Chain

Analytic Application	Business Question	Business Value
Inventory Optimization	1. Which products have high demand? 2. Which products are slow moving or becoming obsolete?	1. Forecast the consumption of fast-moving products and order them with sufficient inventory to avoid a stock-out scenario. 2. Perform fast inventory turnover of slow-moving products by combining them with one in high demand.
Price Elasticity	1. How much net margin do I have on the product? 2. How much discount can I give on this product?	1. Markdown prices for each product can be optimized to reduce the margin dollar loss. 2. Optimized price for the bundle of products is identified to save the margin dollar.
Market Basket Analysis	1. What products should I combine to create a bundle offer? 2. Should I combine products based on slow-moving and fast-moving characteristics? 3. Should I create a bundle from the same category or different category line?	1. The affinity analysis identifies the hidden correlations between the products, which can help in following values: a) Strategize the product bundle offering based on focus on inventory or margin. b) Increase cross-sell or up-sell by creating bundle from different categories or the same categories, respectively.
Shopper Insight	1. Which customer is buying what product at what location?	1. By customer segmentation, the business owner can create personalized offers resulting in better customer experience and retention of the customer.
Customer Churn Analysis	1. Who are the customers who will not return? 2. How much business will I lose? 3. How can I retain them? 4. What demography of customer is my loyal customer?	1. Businesses can identify the customer and product relationships that are not working and show high churn. Thus can have better focus on product quality and reason for that churn. 2. Based on the customer lifetime value (LTV), the business can do targeted marketing resulting in retention of the customer.

1.7 A Brief Introduction to Big Data Analytics

Big Data is data that cannot be stored in a single storage unit. Big Data typically refers to data that comes in many different forms: structured, unstructured, in a stream, and so forth. There are two aspects to managing data on this scale: **storing** and **processing**. A solution was proposed that involved storing this data in chunks on different machines connected by a network—putting a copy or two of this chunk in different locations on the network, both logically and physically.

Data is worthless if it does not provide business value, and for it to provide business value, it has to be analysed. Passing all computation to one powerful computer does not work; this scale would create a huge overhead on such a powerful computer. Another ingenious solution was proposed: Push computation to the data, instead of pushing data to a computing node. This was a new paradigm and gave rise to a whole new way of processing data. This is what we know today as the **MapReduce** programming paradigm.

1.8 An Overview of the Analytics Ecosystem

In order to know where you fit in as a BI professional, it is important to understand the **analytics ecosystem**.

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

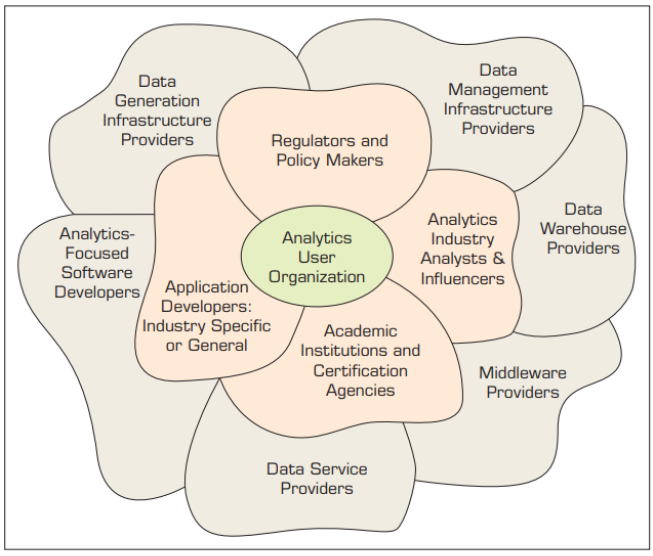


FIGURE 1.13 Analytics Ecosystem.

- **Data generation infrastructure providers** are a group of companies that enable generating and collection of data that may be used for developing analytical insights. The primary focus is on enabling an organisation to develop new insights into its operations as opposed to running its core operations.
- **Data management infrastructure providers** includes all of the major organisations that provide hardware and software targeting the basic foundation for all data management solutions.
- **Data warehouse providers** provide technology and services aimed toward integrating data from multiple sources, thus enabling organisations to derive and deliver value from its data assets.
- The general goal of the **middleware industry** is to provide easy-to-use tools for reporting or descriptive analytics, which forms a core part of BI or analytics employed at organisations.
- Several companies realised the opportunity to develop specialised data collection, aggregation, and distribution mechanisms. These **data service providers** typically focus on a specific industry sector and build on their existing relationships in that industry through their niche platforms and services for data collection.
- **Analytics-focused software developers** have developed analytics software for general use with data that has been collected in a DW or is available through one of the platforms identified earlier (including Big Data). Major industry player can be identified using the three types of analytics:
    - **Reporting** or **descriptive analytics** is enabled by the tools available from the middleware industry players identified earlier, or unique capabilities offered by focused providers.
    - **Predictive analytics**.

    - **Prescriptive analytics** software providers offer modelling tools and algorithms for optimization of operations usually called management science/operations research software.
- **Application developers** use their industry knowledge, analytical expertise, solutions available from the data infrastructure, DW, middleware, data aggregators, and analytics software providers to develop custom solutions for a specific industry. Thus, this industry group makes it possible for analytics technology to be used in a specific industry.
- **Analytics industry analysts and influencers** consist of three types of organisations and professionals:
    - Professional organisations that provide advice to the analytics industry providers and users (Gartner, McKinsey).
    - Professional societies or organisations that also provide some of the same services but are membership based and organised (INFORMS)
    - Analytics ambassadors, influencers, or evangelists. These analysts have presented their enthusiasm for analytics through their seminars, books, and other publications (Steve Baker, Tom Davenport).
- **Universities and academic programmes** that prepare professionals for the industry. It includes various components of business schools such as information systems, marketing, management sciences, and so on. Another group of players called **certification programmes** assist with developing competency in analytics that award a certificate of expertise in specific software.
- **Regulators and policy makers** are responsible for defining rules and regulations for protecting employees, customers, and shareholders of the analytics organisations.
- **Analytic user organisation** are the economic engine of the whole analytics industry. Without users there would be no analytics industry. Goals are to explore opportunities in a sector and to try and gain/retain a competitive advantage.

2. Descriptive Analytics I: Nature of Data, Statistical Modelling, and Visualisation

2.2 The Nature of Data

Data can be viewed as the raw material for what these popular decision technologies produce—information, insight, and **knowledge**. Data – often called Big Data –is nowadays widely considered amongst the most valuable assets of an organisation, with the potential to create invaluable insight to better understand customers, competitors, and the business processes. Big Data is characterised by its quantity, if it is structured or not, and method of collection (automatic/manual and continuously or in one batch). Automated

data collection allows us to collect more volumes of data but also enhancing the **data quality** and integrity.

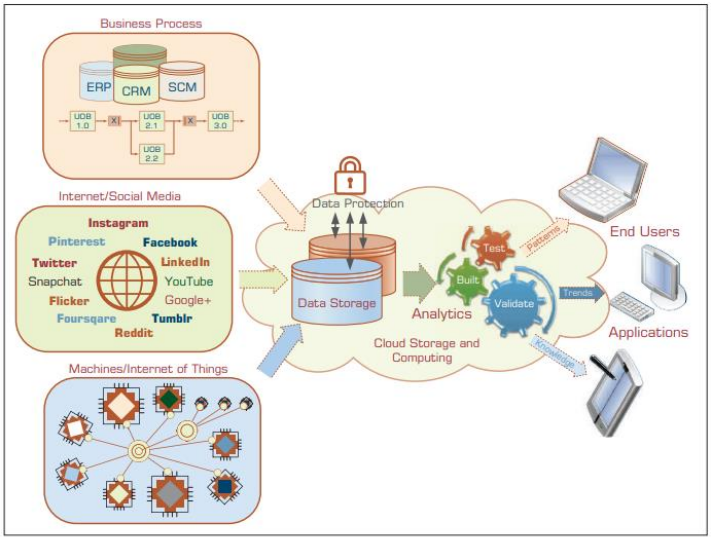


FIGURE 2.1 A Data to Knowledge Continuum.

Depending on the goal of the analytics, data need to be made **analytics ready**. That is, relevant metrics need to be defined, the data may need to be transformed, or data needs to comply with the quality and quantity requirements. Metrics that define the readiness level of data for an analytics study include:

- **Data source reliability** refers to the originality and appropriateness of the storage medium where the data is obtained.
- **Data content accuracy** means that data are correct and are a good match for the analytics problem.
- **Data accessibility** means that the data are easily and readily obtainable.
- **Data security and data privacy** means that the data is secured to only allow those people who have the authority and the need to access it and to prevent anyone else from reaching it.
- **Data richness** means that all the required data elements are included in the data set.
- **Data consistency** means that the data are accurately collected and combined/ merged.
- **Data currency/data timeliness** means that the data should be up-to-date (or as recent/new as it needs to be) for a given analytics model.
- **Data granularity** requires that the variables and data values be defined at the lowest (or as low as required) level of detail for the intended use of the data.
- **Data validity** is the term used to describe a match/mismatch between the actual and expected data values of a given variable.
- **Data relevancy** means that the variables in the data set are all relevant to the study being conducted.

2.3 A Simple Taxonomy of Data

Data (**datum** in singular form) refers to a collection of facts usually obtained as the result of experiments, observations, transactions, or experiences. It can be in different levels of abstraction. At the highest level, data can be classified as structured or unstructured data. **Unstructured/semi structured** data is composed of any combination of textual, imagery, voice, and Web content. **Structured data** is what data mining algorithms use and can be classified as

categorical or numeric. Structured data can be divided into numerical and categorical data.

- **Categorical (discrete) data** represent the labels of multiple classes used to divide a variable into specific groups (e.g. Gender).
- **Nominal data** contain measurements of simple codes assigned to objects as labels, which are not measurements.
- **Ordinal data** contain codes assigned to objects or events as labels that also represent the rank order among them (e.g. (1) Low, (2) medium, etc.).
- **Numeric (continuous) data** represent the numeric values of specific variables. Numeric values representing a variable can be integer (taking only whole numbers) or real (taking also the fractional number).
- **Interval data** are variables that can be measured on interval scales.
- **Ratio data** include measurement variables commonly found in the physical sciences and engineering. Informally, the distinguishing feature of a ratio scale is the possession of a nonarbitrary zero value.

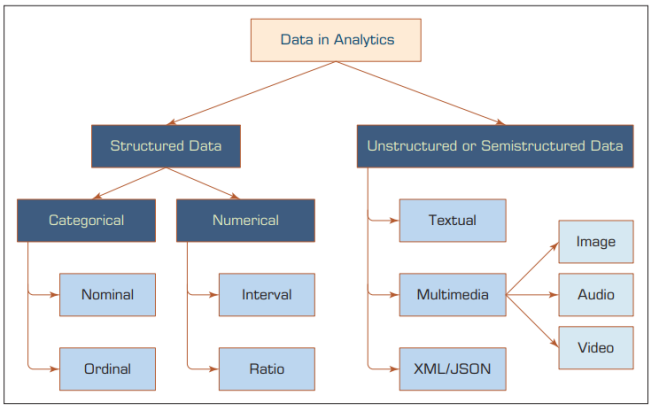


FIGURE 2.2 A Simple Taxonomy of Data.

2.4 The Art and Science of Data Preprocessing

Data in its original form (i.e., the real-world data) is not usually ready to be used in analytics tasks. Therefore, **data preprocessing** is necessary, that is to convert the raw real-world data into a well-refined form for analytics algorithms.

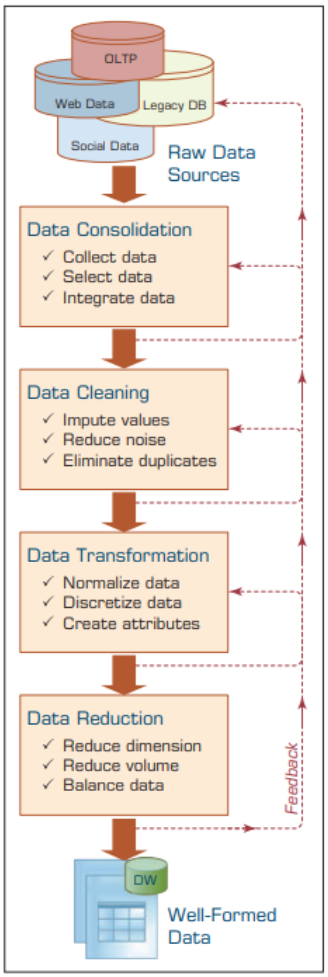


FIGURE 2.3 Data Preprocessing Steps.

In the first phase:

- The relevant data is **collected** from the identified sources
- The necessary records and variables are **selected**
- The records coming from multiple data sources are **integrated/merged**.

In the second phase, the data is cleaned:

- The values in the data set are identified and dealt with. In some cases, missing values are an anomaly in the data set, in which case they need to be **imputed** or **ignored**; in other cases, the missing values are a natural part of the data set.
- **Inconsistencies** in the data should be handled using domain knowledge and/or expert opinion.

In the third phase, the data is transformed for better processing:

- The data is **normalised** between a certain minimum and maximum for all variables to mitigate the potential bias of one variable dominating other variables having smaller values.
- **Discretisation** to distinguish between information and/or aggregation to gain an overview.

The final phase of data preprocessing is data reduction:

- Predictive analytics projects as a flat file consisting of two dimensions: variables and cases/records.
- Treat variables as different dimensions that describe the phenomenon from different perspectives. This is called **dimensional reduction** (or **variable selection**). To reduce the data set, data should be carefully sampled.

TABLE 2.1 A Summary of Data Preprocessing Tasks and Potential Method

Main Task	**Subtasks**	**Popular Methods**
Data consolidation	Access and collect the data Select and filter the data Integrate and unify the data	SQL queries, software agents, Web services. Domain expertise, SQL queries, statistical tests. SQL queries, domain expertise, ontology-driven data mapping.
Data cleaning	Handle missing values in the data	Fill in missing values (imputations) with most appropriate values (mean, median, min/ max, mode, etc.); recode the

		missing values with a constant such as "ML"; remove the record of the missing value; do nothing.
	Identify and reduce noise in the data	Identify the outliers in data with simple statistical techniques (such as averages and standard deviations) or with cluster analysis; once identified, either remove the outliers or smooth them by using binning, regression, or simple averages.
	Find and eliminate erroneous data	Identify the erroneous values in data (other than outliers), such as odd values, inconsistent class labels, odd distributions; once identified, use domain expertise to correct the values or remove the records holding the erroneous values.
Data transformation	Normalise the data	Reduce the range of values in each numerically valued variable to a standard range (e.g., 0 to 1 or –1 to +1) by using a variety of normalisation or scaling techniques.
	Discretise or aggregate the data	If needed, convert the numeric variables into discrete representations using range or frequency-based binning techniques; for categorical variables, reduce the number of values by applying proper concept hierarchies.
	Construct new attributes	Derive new and more informative variables from the existing ones using a wide range of mathematical functions (as simple as addition and multiplication or as complex as a hybrid combination of log transformations).
Data reduction	Reduce number of attributes	Principal component analysis, independent component analysis, chi-square testing, correlation analysis, and decision tree induction.
	Reduce number of records	Random sampling, stratified sampling, expert-knowledge-driven purposeful sampling.
	Balance skewed data	Oversample the less represented or undersample the more represented classes.

Application Case 2.2 describes an example of applying data mining and machine learning algorithms to determine which students will drop out. The process is as depicted on the right. One of the problems faced was that dropout/no dropout was imbalanced, i.e. only 20% of the students dropped out. This leads to a high accuracy but low precision (due to a high number of false negatives). This class imbalance was solved by taking samples from the 'No' class to move the data to 50/50.

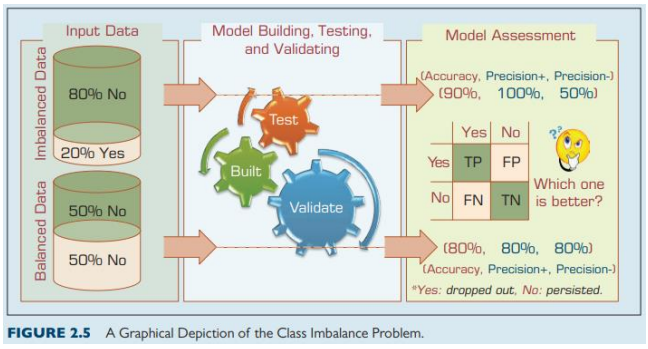


FIGURE 2.5 A Graphical Depiction of the Class Imbalance Problem.

(1) Rebalancing can amplify structural bias. (2) If you leave out students that continue (under-sampling), you throw away a lot of information. (3) If you rebalance it to 50/50 drop-out/continuing, the classifier will say 50% will drop-out, because that is what you trained on.

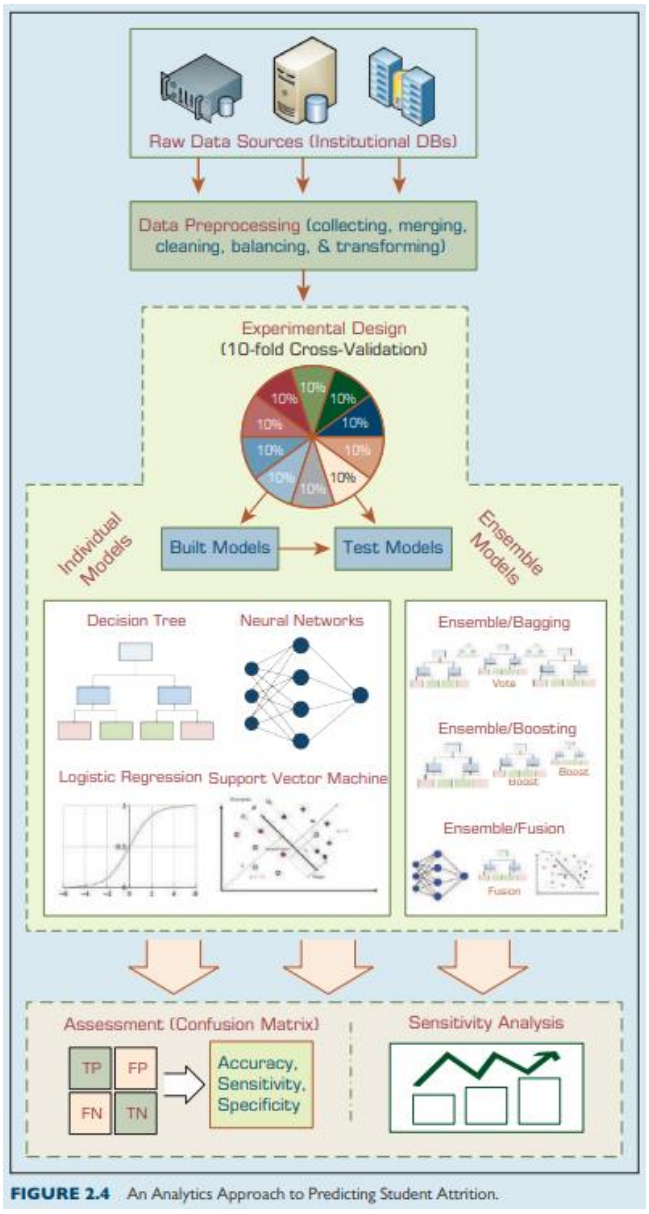


FIGURE 2.4 An Analytics Approach to Predicting Student Attrition.

2.5 Statistical Modelling for Business Analytics

Statistics is usually considered a part of descriptive analytics. Some of the statistical methods can also be considered as part of predictive analytics. Descriptive analytics has two main branches: statistics and **online analytics processing (OLAP)**. OLAP (or **business intelligence**) is the term used for analysing, characterising, and summarising structured data stored in organisational databases using cubes (i.e. multidimensional data structures that are created to extract a subset of data values to answer a specific business question). **Statistics**—a collection of mathematical techniques to characterise and interpret data—helps to characterise the data either one variable at a time or multivariables altogether, using either descriptive or inferential methods. Furthermore, **descriptive statistics** is all about describing the sample data on

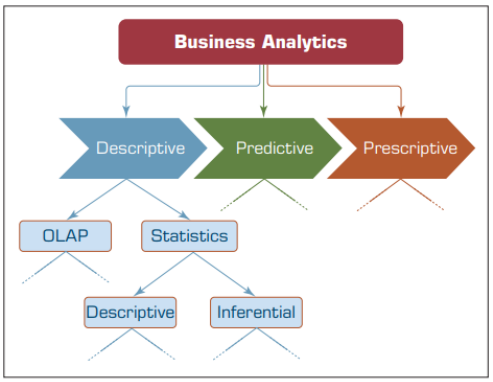


FIGURE 2.7 Relationship between Statistics and Descriptive Analytics.

hand, and **inferential statistics** is about drawing inferences or conclusions about the characteristics of the population.

In business analytics, descriptive statistics allows us to understand and explain/present our data in a meaningful manner using aggregated numbers, data tables, or charts/graphs. For this we use measures of central tendency and measures of dispersion.

Measures of central tendency: A measure of central tendency is a single numerical value that aims to describe a set of data by simply identifying or estimating the central position within the data.

- **Arithmetic mean**: The arithmetic mean (or simply mean or average) is the sum of all the values/observations divided by the number of observations in the data set.
  - $\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n}$ or $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
  - Sensitive to outliers
- **Median**: The number in the middle of a given set of data that has been arranged/sorted in order of magnitude (either ascending or descending).
  - Not so much affected by outliers.
- **Mode**: The observation that occurs most frequently (the most frequent value in our data set).
  - Not useful when data is precise/has too many unique values.

Measures of dispersion: The mathematical methods used to estimate or describe the degree of variation in a given variable of interest.

- **Range**: The difference between the largest and the smallest values in a given data set (i.e., variables).
  - Range = maximum - minimum
- **Variance**: A method used to calculate the deviation of all data points in a given data set from the mean.
  - $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
  - Differences are squared so that larger deviations from the mean contribute significantly to the value of variance.
  - Because difference is squared, the numbers that represent deviation/variance become somewhat meaningless.
- **Standard deviation**: a measure of the spread of values within a set of data.
  - $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$
- **Mean absolute deviation**: A simpler way to calculate the overall deviation from the mean by measuring the absolute values of the differences between each data point and the mean and summing them.
  - $MAD = \frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n}$

- **Quartiles and interquartile range**: A **quartile** is a quarter of the number of data points given in a data set. Quartiles are determined by first sorting the data and then splitting the sorted data into four disjoint smaller data sets. The **interquartile range** describes the difference between the third quartile (Q3) and the first quartile (Q1), telling us about the range of the middle half of the scores in the distribution.
    - Not so much affected by outliers or skewness.
- **Box-and-whiskers plot (box plot)**: A graphical illustration of several descriptive statistics about a given data set. The box plot shows the centrality as well as the dispersion, the minimum and maximum ranges (that are calculated as 1.5 times the upper or lower end of the quartile box) along with the outliers that are larger than the limits of the whiskers. It also shows whether the data is symmetrically distributed with respect to the mean or it sways one way or another.

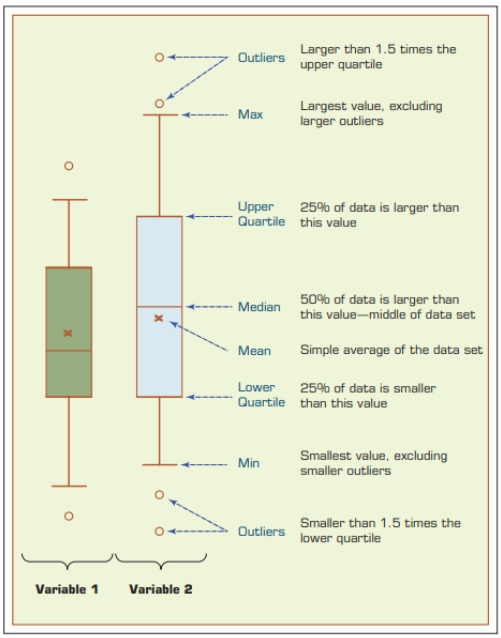


FIGURE 2.8 Understanding the Specifics about Box-and-Whiskers Plots.

- **The shape of a distributor:** A **distribution** is the frequency of data points counted and plotted over a small number of class labels or numerical ranges. As the dispersion of a data set increases, so does the standard deviation, and the shape of the distribution looks wider. Two measures of shape characteristics: **Skewness** and **kurtosis**.
    - **Skewness** is a measure of asymmetry (sway) in a distribution of the data that portrays a unimodal structure—only one peak exists in the distribution of the data. In Figure 2.9, (c) represents a positively skewed distribution, whereas (d) represents a negatively skewed distribution. In the same figure, both (a) and (b) represent perfect symmetry and hence zero measure for skewness.

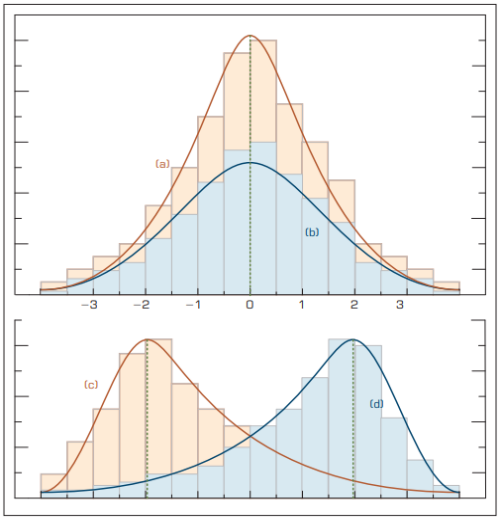


FIGURE 2.9 Relationship between Dispersion and Shape Properties.

        - $\text{Skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{\frac{3}{2}}} * \frac{n^2}{(n-1)(n-2)}$
    - **Kurtosis** measures the degree to which a distribution is more or less peaked than a normal distribution. Whereas a positive kurtosis indicates a relatively peaked/tall distribution, a negative kurtosis indicates a relatively flat/short distribution.
        - $\text{Kurtosis} = K = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{ns^4} - 3$

2.6 Regression Modelling for Inferential Statistics

Regression is a relatively simple statistical technique to model the dependence of a variable (response or output variable) on one (or more) explanatory (input) variables. Once identified, this relationship between the variables can be formally represented as a linear/additive function/equation. Regression can be used for one of two purposes: **hypothesis testing** and **prediction/forecasting**. Often confused with regression, **correlation** makes no a priori assumption of whether one variable is dependent on the other(s) and is not concerned with the relationship between variables; instead it gives an estimate on the degree of association between the variables.

If the regression equation is built between one response variable and one explanatory variable, then it is called **simple regression**. **Multiple regression** is the extension of simple regression where the explanatory variables are more than one.

To develop a linear regression model, one can simple create a scatterplot to look at the relationship between the response and input variable. Simple regression analysis aims to find the signature of a straight line passing through right between the plotted dots in such a way that it minimises the distance between the dots and the line. The **ordinary least squares (OLS)** method aims to minimise the sum of squared residuals (squared vertical distances between the observation and the regression point).

$$y = \beta_0 + \beta_1 x$$

In this equation, β_0 is called the intercept, and β_1 is called the slope. If it is a multiple linear regression, coefficient get added to the equation to establish a linear additive representation of the response variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

In order to assess a model's fit (the degree at which it represents the response variable), three statistical measures are often used: R^2, the overall F-test, and the root mean square error (RMSE). R^2 is the most useful because of its intuitive scale with 0 indicating no explained variance and 1 indicating exact prediction.

Linear regression models suffer from several highly restrictive assumptions:

1. **Linearity**: The relationship between the response variable and the explanatory variables are linear and additive.
2. **Independence (of errors)**: The errors of the response variable are uncorrelated with each other.
3. **Normality (of errors)**: The errors of the response variable are normally distributed.
4. **Constant variance (of errors)**: also called **homoscedasticity**, states that the response

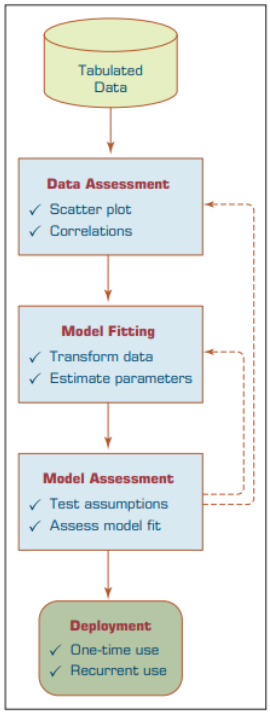


FIGURE 2.14 A Process Flow for Developing Regression Models.

variables have the same variance in their error, regardless of the values of the explanatory variables.

5. **Multicollinearity**: the explanatory variables are not correlated.

Logistic regression is a probability-based classification algorithm that employs supervised learning. It differs from linear regression with one major point: its output (response variable) is a class as opposed to a numerical variable.

$$f(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Unlike linear regression with normally distributed residuals, it is not possible to find a closed-form expression for the coefficient values that maximises the likelihood function, so an iterative process must be used instead.

If there are no distinctly identifiable explanatory variables or there are too many of them in a complex relationship, a time-series can be developed instead. A **time-series** is a sequence of data points of the variable of interest, measured and represented at successive points in time spaced at uniform time intervals. **Time-series forecasting** is the use of mathematical modelling to predict future values of the variable of interest based on previously observed values. Time-series forecasting assumes all the explanatory variables are aggregated and consumed in the response variable's time-variant behaviour and therefore are focused on extrapolating on their time-varying behaviour to estimate the future values by analysing patterns such as random variations, time trends, and seasonal cycles.

2.7 Business Reporting

Information provided by analytics should be provided to those who need it. A **report** is any communication artefact prepared with the specific intention of conveying information in a digestible form to whoever needs it, whenever and wherever they may need it. The foundation of these **business reports** is various sources of data coming from both inside and outside the organisation. Creation of these reports involves **ETL (Extract, transform, and load)** procedures in coordination with a data warehouse and then using one or more reporting tools.

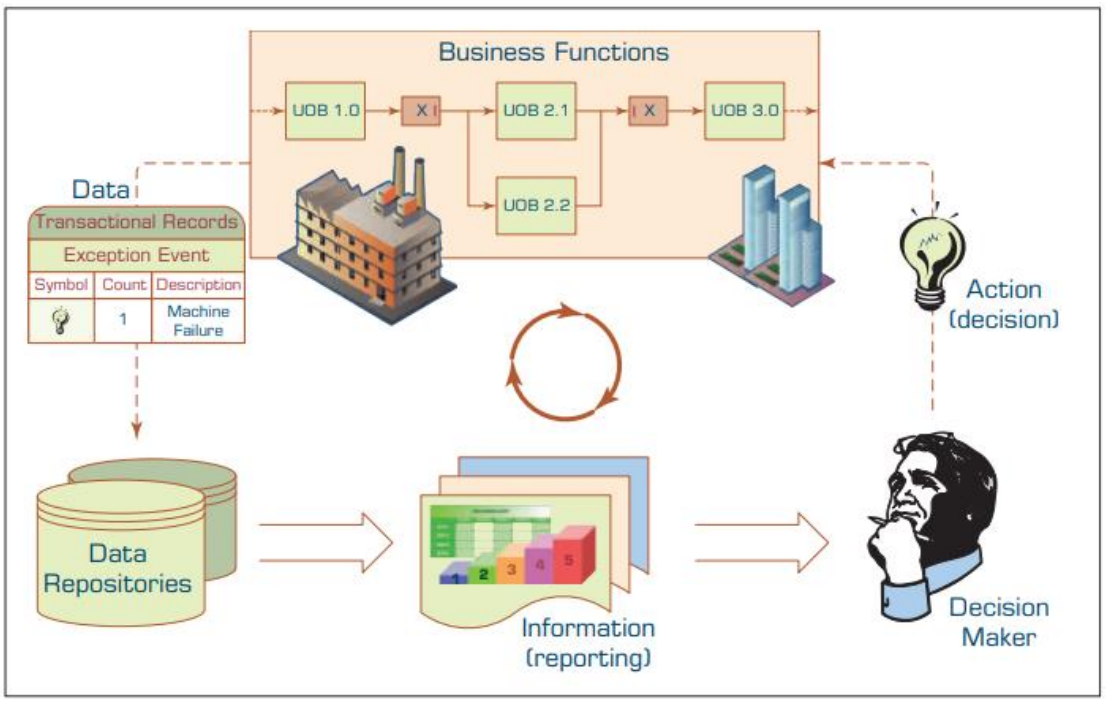


FIGURE 2.18 The Role of Information Reporting in Managerial Decision Making.

Figure 2.18 shows the continuous cycle of data acquisition → information generation → decision making → business process management. Perhaps the most critical task in this cyclical process is the reporting (i.e., information generation)— converting data from different sources into actionable information.

There are three categories of business reports for managerial purposes:

- **Metric management reports**: **Key performance indicators (KPIs)** that measure the business performance through outcome-oriented agreements.
- **Dashboard-type reports**: Present different performance indicators like a dashboard in a car.
- **Balanced scorecard-type reports**: An integrated view of success in an organisation by displaying financial performance, customer, business process, and learning and growth perspectives.

2.8 Data Visualisation

Data visualisation (or more appropriately, **information visualisation**) has been defined as "the use of visual representations to explore, make sense of, and communicate data". Data visualisation has been around since 200AD, but most developments occurred in the last two and a half centuries. The Internet emerged as a new medium for visualisation and brought with it a whole lot of new tricks and capabilities. The future of data visualisation is hard to predict, but we can extrapolate that there will be more three-dimensional visualisation, more immersive experience with multidimensional data in a virtual reality environment, and holographic visualisation of information.

2.9 Different Types of Charts and Graphs

Basic charts and graphs that are commonly used for information visualisation:

- **Line chart**: Line charts (or a line graphs) show the relationship between two variables by sequentially connect individual data points; they are most often used to track changes or trends over time.
- **Bar chart**: Bar charts are effective when you have nominal data or numerical data that splits nicely into different categories so you can quickly see comparative results and trends within your data.
- **Pie chart**: Pie charts are visually appealing, as the name implies, pie-looking charts. Pie charts should only be used to illustrate relative proportions of a specific measure.
- **Scatter plot**: Scatter plots are often used to explore the relationship between two or three variables (in 2-D or 2-D visuals). Scatter plots are an effective way to explore the existence of trends, concentrations, and outliers.
- **Bubble chart**: Bubble charts are often enhanced versions of scatter plots. By varying the size and/or colour of the circles, one can add additional data dimensions, offering more enriched meaning about the data.

Special charts derived from the basic charts as special cases and charts relatively new and specific to a problem type and/ or an application area:

- **Histogram**: Graphically similar to a bar chart. Histograms are used to show the frequency distribution of a variable or several variables where the x-axis is often used to show the categories or ranges, and the y-axis is used to show the measures/values/frequencies.
- **Gantt chart**: Gantt charts are a special case of horizontal bar charts that are used to portray project timelines, project tasks/activity durations, and overlap among the tasks/ activities.
- **Pert chart**: PERT charts (also called network diagrams) are developed primarily to simplify the planning and scheduling of large and complex projects. They show precedence relationships among the project activities/tasks.
- **Geographic map**: When the data set includes any kind of location data, it is better and more informative to see the data on a map. Maps usually are used in conjunction with other charts and graphs, as opposed to by themselves.
- **Bullet**: Bullet graphs are often used to show progress toward a goal. A bullet graph is essentially a variation of a bar chart in the form of gauges, meters, or thermometers.
- **Heat map**: Illustrate the comparison of continuous values across two categories using colour. The goal is to help the user quickly see where the intersection of the categories is strongest and weakest in terms of numerical values of the measure being analysed.
- **Highlight table**: Similar to heat maps but also showing how data intersects by using colour, highlight tables add a number on top to provide additional details.

- **Tree map**: Tree maps display hierarchical (tree-structured) data as a set of nested rectangles.

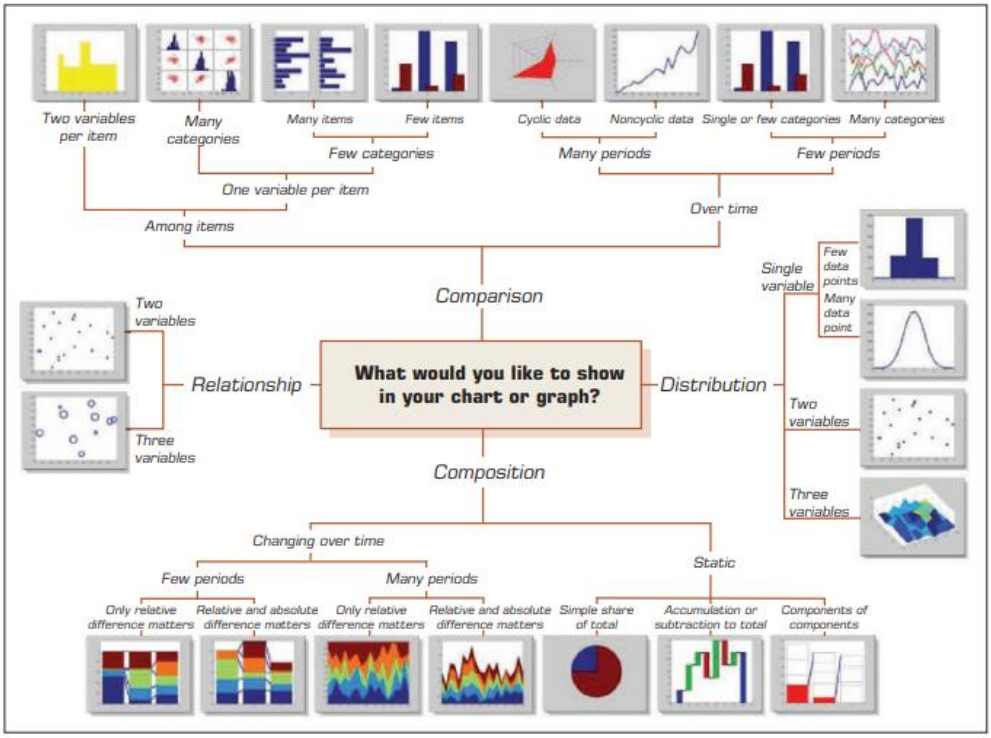


FIGURE 2.21 A Taxonomy of Charts and Graphs. *Source:* Adapted from Abela, A. (2008). Advanced presentations by design: Creating communication that drives action. New York: Wiley.

2.10 The Emergence of Visual Analytics

Visual analytics is the combination of visualisation and predictive analytics. Visual analytics is aimed at answering, "Why is it happening?" "What is more likely to happen?" and is usually associated with business analytics (forecasting, segmentation, correlation analysis). Due to fast-growing volumes of data, high-performance computing is getting more important. **High-performance computing** is an in-memory solution for exploring massive amounts of data in a very short time (almost instantaneously). It empowers users to spot patterns, identify opportunities for further analysis, and convey visual results via Web reports or a mobile platform such as tablets and smartphones.

2.11 Information Dashboards

Dashboards are common in most BI or business analytics platforms. **Dashboards** provide visual displays of important information that is consolidated and arranged on a single screen so that information can be digested at a single glance and easily drilled in and further explored.

The most distinctive feature of a dashboard is its three layers of information:

1. **Monitoring**: Graphical, abstracted data to monitor key performance metrics.
2. **Analysis**: Summarised dimensional data to analyse the root cause of problems.
3. **Management**: Detailed operational data that identify what actions to take to resolve a problem.

The fundamental challenge of dashboard design is to display all the required information on a single screen, clearly and without distraction, in a manner that can be assimilated quickly. Furthermore, numbers need to be placed into context (i.e. show if the numbers are good or bad) by comparing them to past results, benchmarks, or forecasts.

All dashboards have things in common, for example that the architecture of the dashboard fits in with the larger architecture of the BI system. Also:

- They use visual components to highlight, at a glance, the data and exceptions that require action.
- They are transparent to the user, meaning that they require minimal training and are extremely easy to use.
- They combine data from a variety of systems into a single, summarised, unified view of the business.

Best practices in dashboard design:

- Benchmark key performance indicators with industry standard.
- Wrap the dashboard metrics with contextual metadata.
- Validate the dashboard design by a usability specialist.
- Prioritise and rank alerts/exceptions streamed to the dashboard.
- Enrich the dashboard with business-user comments.
- Present information in three different levels.
- Pick the right visual construct using dashboard design principles.
- Provide for guided analytics.

3. Descriptive Analytics II: Business Intelligence and Data Warehousing

3.2 Business Intelligence and Data Warehousing

BI systems rely on a data warehouse as the information source for creating insight and supporting managerial decisions. A **data warehouse (DW)** is a pool of data produced to support decision making; it is also a repository of current and historical data of potential interest to managers throughout the organisation. Data are usually structured to be available in a form ready for analytical processing activities (i.e., **online analytical processing [OLAP],** data mining, querying, reporting, and other decision support applications). A data warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's **decision-making process**.

In the early 1900s, people were using data (though mostly via manual methods) to formulate trends to help business users make informed decisions. In the 1970s, real business data-processing applications, the ones run on the corporate mainframes, had complicated file structures using **early-generation databases** in which they stored data. Although these applications did a decent job of performing routine transactional data-processing functions, the data created as a result of these functions was **locked away** in the depths of the files and

databases. The 1980s were the decade of personal computers and minicomputers. That led to a portentous problem called **islands of data**. The solution to this problem led to a new type of software, called a **distributed database management system**, which would pull, consolidate, and sort the requested data across the organisation. However, it wasn't very efficient. In 1983, Teradata shipped out the first **RDBMS (relational database management system)**. The 1990s philosophy involved going back to the 1970s, in which data from those places was copied to another location. Since 2010, the big buzz has been **Big Data**.

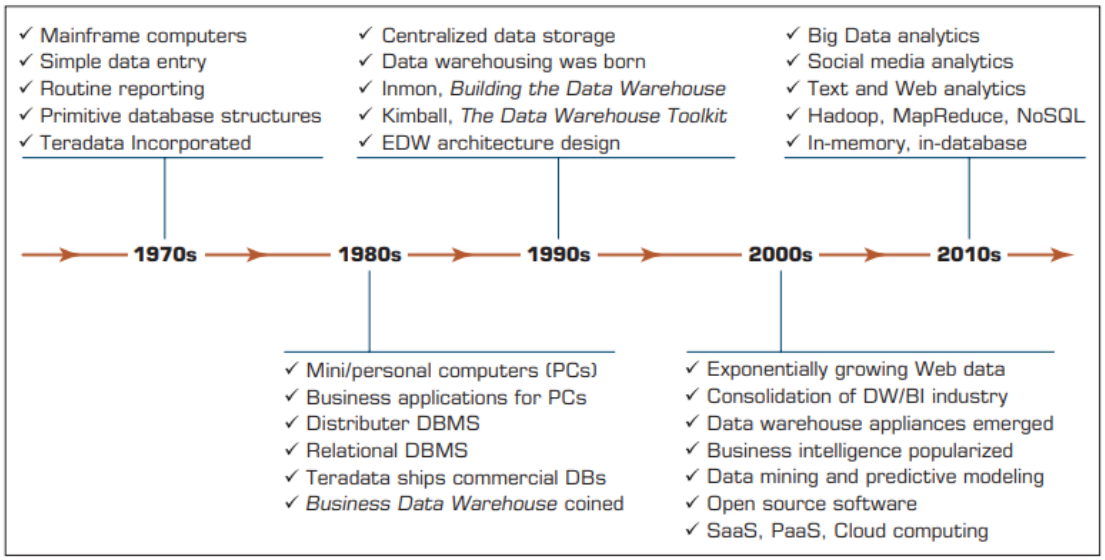


FIGURE 3.2 A List of Events That Led to Data Warehousing Development.

Characteristics of data warehousing:

- **Subject oriented**: Data are organized by detailed subject.
- **Integrated**: Placing data from different sources into a consistent format.
- **Time variant (time-series)**: The temporal quality of a data warehouse to detect trends, deviations, and long-term relationships for forecasting and comparisons, leading to decision making.
- **Non-volatile**: After data are entered into a data warehouse, users cannot change or update the data.
- **Web based**: Data warehouses are typically designed to provide an efficient computing environment for Web-based applications.
- **Relational/multidimensional**: A data warehouse uses either a relational structure or a multidimensional structure.
- **Client/server**: A data warehouse uses the client/server architecture to provide easy access for end users.
- **Real time**: Newer data warehouses provide real-time, or active, data-access and analysis capabilities.
- **Include metadata**: A data warehouse contains metadata (data about data) about how the data are organized and how to effectively use them.

Data warehousing is a discipline that results in applications that provide **decision support capability**, allows **ready access to business information**, and **creates business insight**. Three main types:

- **Data marts (DM)**: A DM is a subset of a data warehouse, typically consisting of a single subject area.
    - Dependent or independent.
        - A **dependent data mart** is a subset that is created directly from the data warehouse. It has the advantages of using a consistent data model and providing quality data.
        - An **independent data mart** is a small warehouse designed for a strategic business unit or a department, but its source is not an EDW.
- **Operational data stores (ODS)**: An ODS is used for short-term decisions involving mission-critical applications rather than for the medium- and long-term decisions associated with an EDW.
    - **Oper marts** are created when operational data needs to be analysed multidimensionally. The data for an oper mart come from an ODS.
- **Enterprise data warehouses (EDW):** An EDW is a large-scale data warehouse that is used across the enterprise for decision support. It integrates data from many sources into a standard format for effective BI and decision support applications.

Finally, **metadata** are data about data. Metadata describe the structure of and some meaning about data, thereby contributing to their effective or ineffective use.

3.3 Data Warehousing Process

Working with multiple databases, either integrated in a data warehouse or not, has become an extremely difficult task requiring considerable expertise, but it can provide immense benefits far exceeding its cost. As an illustrative example, Figure 3.3 shows business benefits of the EDW built by Teradata for a major automobile manufacturer.

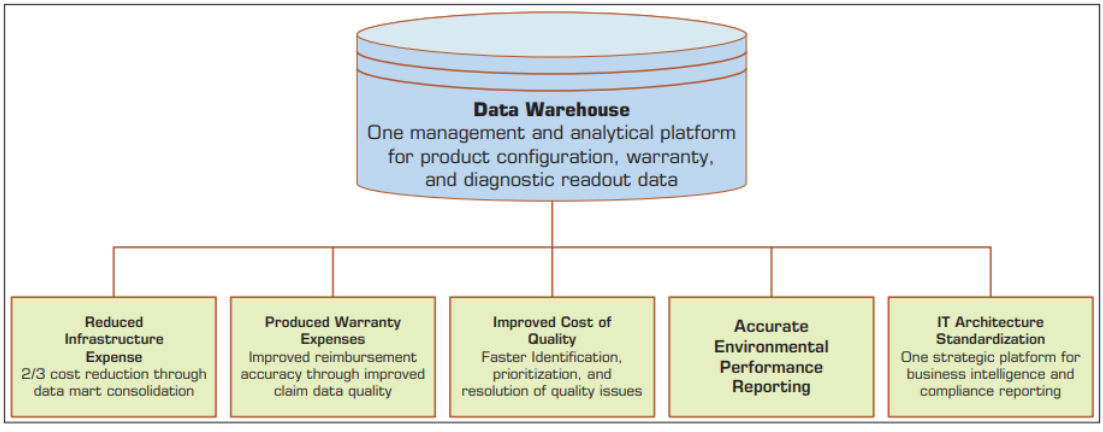


FIGURE 3.3 Data-Driven Decision Making—Business Benefits of the Data Warehouse. *Source:* Teradata Corp.

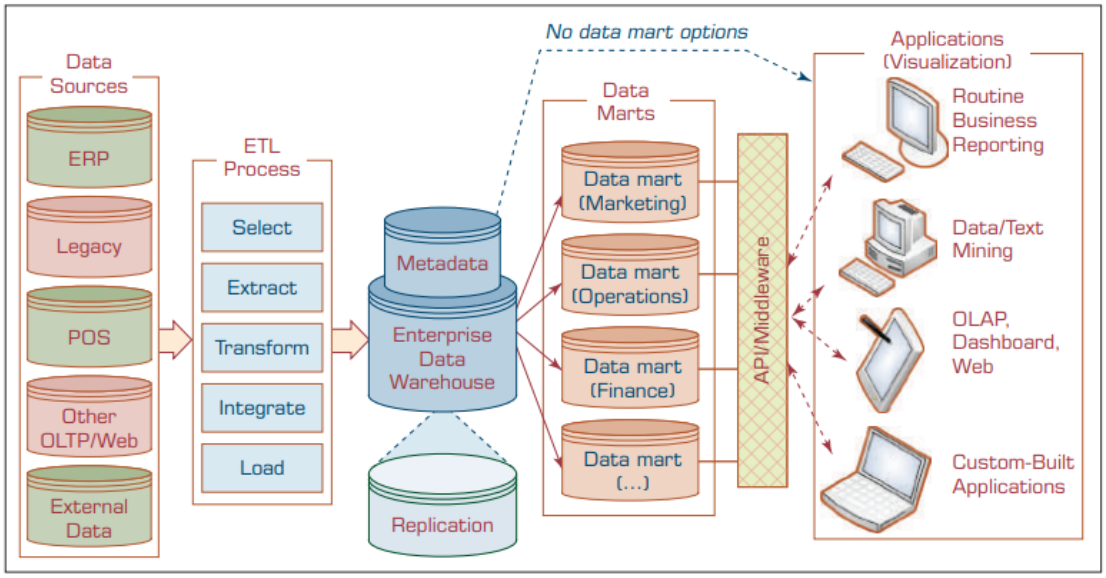


FIGURE 3.4 A Data Warehouse Framework and Views.

Major components of the data warehousing process:

- **Data sources**: Data are sourced from multiple independent operational "legacy" systems and possibly from external data providers.
- **Data extraction and transformation**: Data are extracted and properly transformed using custom-written or commercial software called ETL.
- **Data loading**: Data are loaded into a staging area, where they are transformed and cleansed.
- **Comprehensive database**: The EDW to support all decision analysis by providing relevant summarised and detailed information originating from many different sources.
- **Metadata**: Metadata are maintained so that they can be assessed by IT personnel and users.
- **Middleware tools**: Middleware tools enable access to the data warehouse, such as SQL or Business Objects.

3.4 Data Warehousing Architectures

Data warehouse architectures can either be **client/server** or **n-tier architectures**, with two-tier or three-tier architectures being the most common. Data warehouses can be divided into three parts:

- The data warehouse itself
- Data acquisition (back-end) software
- Client (front-end) software

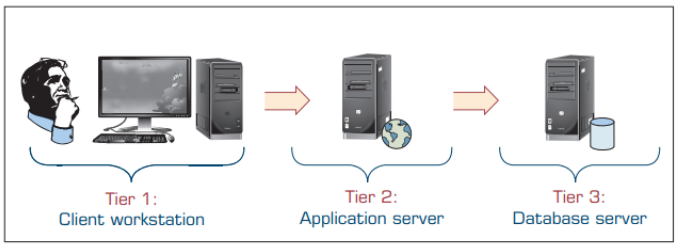


FIGURE 3.5 Architecture of a Three-Tier Data Warehouse.

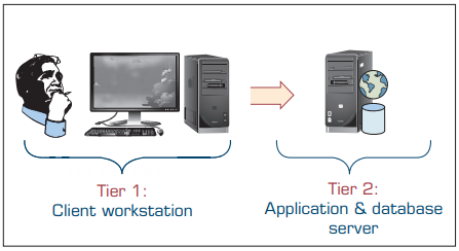


FIGURE 3.6 Architecture of a Two-Tier Data Warehouse.

The advantage of the three-tier architecture is its separation of the functions of the data warehouse, which eliminates resource constraints and makes it possible to easily create DMs. In contrast, a two-tier architecture is more economical since it requires less hardware. The best architecture depends on the situation of the organisation.

The integration of the Internet and data warehousing produces Web-based data warehousing. A Web-based data warehousing architecture is three-tiered. Web-based data warehousing offers several compelling advantages, including ease of access, platform independence, and lower cost.

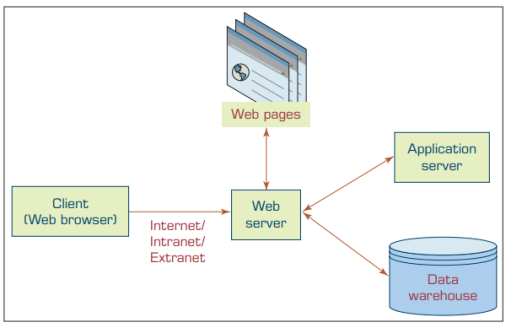
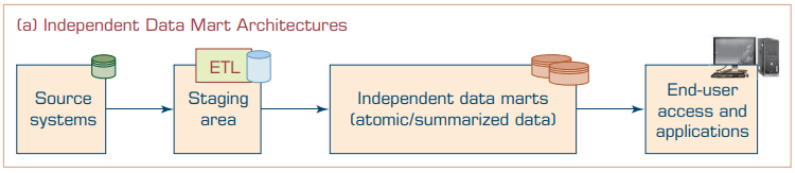


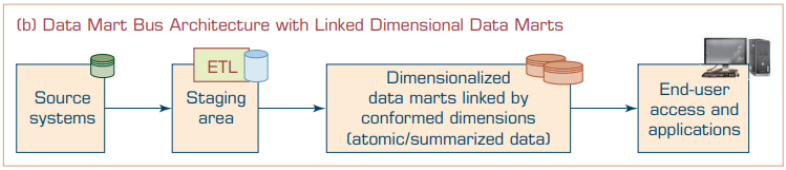
FIGURE 3.7 Architecture of Web-Based Data Warehousing.

At the highest level, data warehouse architecture design viewpoints can be categorised into enterprise-wide data warehouse (EDW) design and DM design. Alternatives:

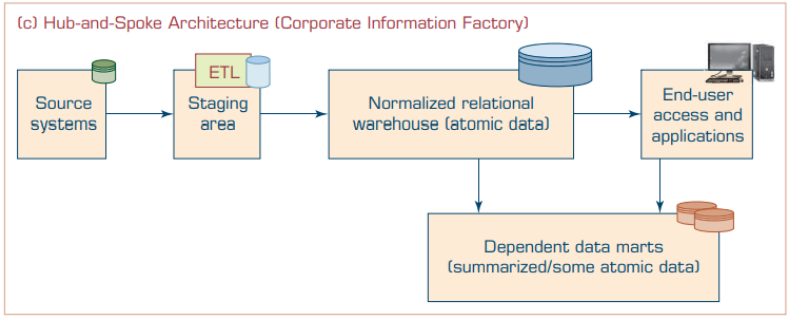
- **Independent data marts**
  - Simplest and least costly architecture.
  - DMs operate independently → Inconsistent data.



- **Data mart bus architecture**
  - DMs are linked to each other through middleware → Higher consistency than independent DMs.
  - Performance may be unsatisfactory.

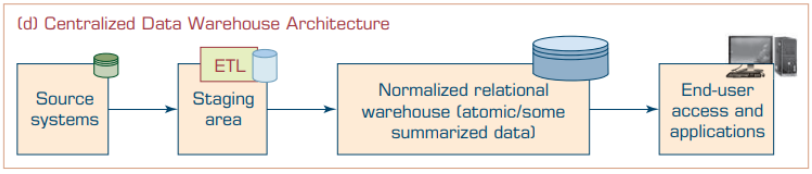


- **Hub-and-spoke architecture**
  - Focused on building a scalable and maintainable infrastructure.
  - Allows for easy customization of user interfaces and reports.
  - Lacks the holistic enterprise view and may lead to data redundancy and data latency.

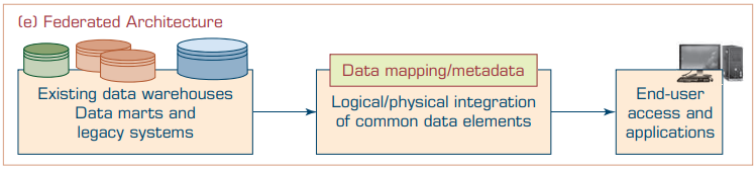


- **Centralised data warehouse**

- Instead of independent DMs (like hub-and-spoke), EDW serves the needs of all organisational units.
- Reduces the amount of data the technical team has to transfer or change → Simpler data management and administration.
- Timely and holistic view of the enterprise to whoever, whenever, and wherever they may be within the organisation.



- **Federated data warehouse**
  - Involves integrating disparate systems.
  - Existing decision support structures are left in place, and data are accessed from those sources as needed supported by middleware.



3.5 Data Integration and the Extraction, Transformation, and Load (ETL) Processes

Integrating, consolidating, and giving users access to a data warehouse used to be a laborious task. **Data integration** comprises three major processes that permit data to be accessed and made accessible to an array of ETL and analysis tools and the data warehousing environment:

- **data access** (i.e. the ability to access and extract data from any data source),
- **data federation** (i.e. the integration of business views across multiple data stores), and
- **change capture** (based on the identification, capture, and delivery of the changes made to enterprise data sources).

Various integration technologies enable data and metadata integration:

- **Enterprise application integration (EAI)**: provides a vehicle for pushing data from source systems into the data warehouse. It involves integrating application functionality and is focused on sharing functionality (rather than data) across systems, thereby enabling flexibility and reuse. For this it uses SOA.
- **Service-oriented architecture (SOA)** coarse-grained services (a collection of business processes or functions) that are well defined and documented. Using Web services is a specialized way of implementing an SOA.
- **Enterprise information integration (EII)**: an evolving tool space that promises real-time data integration from a variety of sources, such as relational databases, Web services, and multidimensional databases. It

is a mechanism for pulling data from source systems to satisfy a request for information using XML.

- **Extraction, transformation, and load (ETL)**

The **ETL process** consists of

- **extraction** (i.e. reading data from one or more databases),
- **transformation** (i.e. converting the extracted data from its previous form into the form in which it needs to be so that it can be placed into a data warehouse or simply another database), and
- **load** (i.e. putting the data into the data warehouse).

ETL tools also

- **transport** data between sources and targets,
- **document** how data elements (e.g., metadata) change as they move between source and target,
- **exchange** metadata with other applications as needed, and
- **administer** all runtime processes and operations (e.g., scheduling, error management, audit logs, statistics).

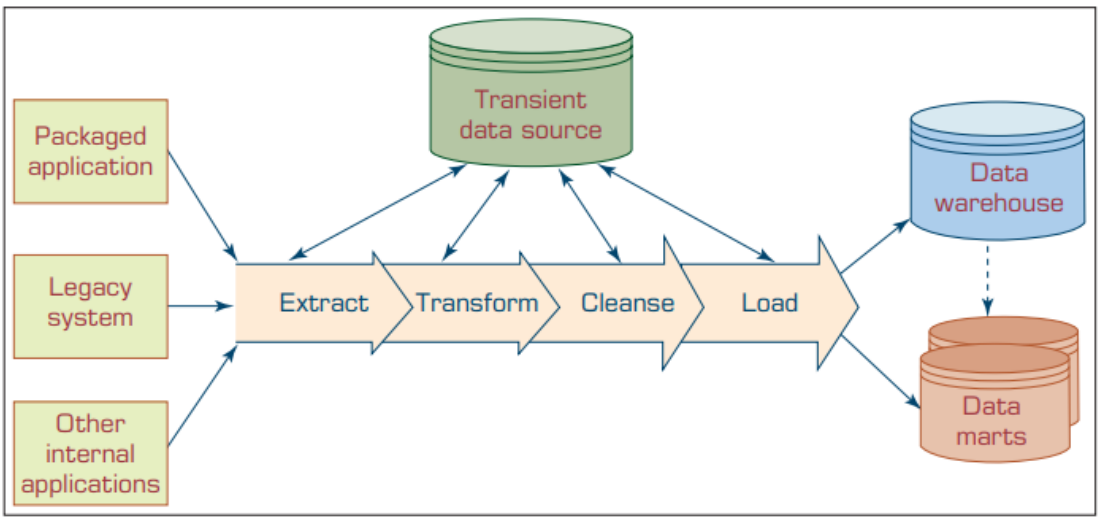


FIGURE 3.9 The ETL Process.

3.6 Data Warehouse Development

A data warehouse provides several benefits that can be classified as direct and indirect:

- End users can perform **extensive analysis** in numerous ways.
- A **consolidated view** of corporate data (i.e., a single version of the truth) is possible.
- Better and more **timely information** is possible.
- **Enhanced system performance** can result. A data warehouse frees production processing because some operational system reporting requirements are moved to DSS.
- **Data access is simplified**.

Before building a data warehouse to receive these benefits, costs need to be taken into account. An ROI approach can be distinguished in:

- **keepers** (i.e. money saved by improving traditional decision support functions)
- **gatherers** (i.e. money saved due to automated collection and dissemination of information)
- **users** (i.e. money saved or gained from decisions made using the data warehouse)

Costs include those related to hardware, software, network bandwidth, internal development, internal support, training, and external consulting.

Three approaches exist to creating a DW:

- The **Inmon model:** The EDW approach . Top-down development. Use established database development methodologies and tools to create a data warehouse that provides a consistent and comprehensive view of the enterprise.
- The **Kimball model**: The Data Mart approach. Bottom-up development. Start building DMs for each department according to their needs and work up from there using dimensional modelling.
- A **hosted data warehouse** has nearly the same, if not more, functionality as an on-site data warehouse, but it does not consume computer resources on client premises. A hosted data warehouse offers the benefits of BI minus the cost of computer upgrades, network upgrades, software licenses, in-house development, and in-house support and maintenance.

TABLE 3.3 Contrasts between the DM and EDW Development Approaches

Effort	DM Approach	EDW Approach
Scope	One subject area	Several subject areas
Development time	Months	Years
Development cost	$10,000 to $100,000+	$1,000,000+
Development difficulty	Low to medium	High
Data prerequisite for sharing	Common (within business area)	Common (across enterprise)
Sources	Only some operational and external systems	Many operational and external systems
Size	Megabytes to several gigabytes	Gigabytes to petabytes
Time horizon	Near-current and historical data	Historical data
Data transformations	Low to medium	High
Update frequency	Hourly, daily, weekly	Weekly, monthly
Technology		
Hardware	Workstations and departmental servers	Enterprise servers and mainframe computers
Operating system	Windows and Linux	Unix, Z/OS, OS/390
Databases	Workgroup or standard database servers	Enterprise database servers
Usage		
Number of simultaneous users	10s	100s to 1,000s
User types	Business area analysts and managers	Enterprise analysts and senior executives
Business spotlight	Optimizing activities within the business area	Cross-functional optimization and decision making

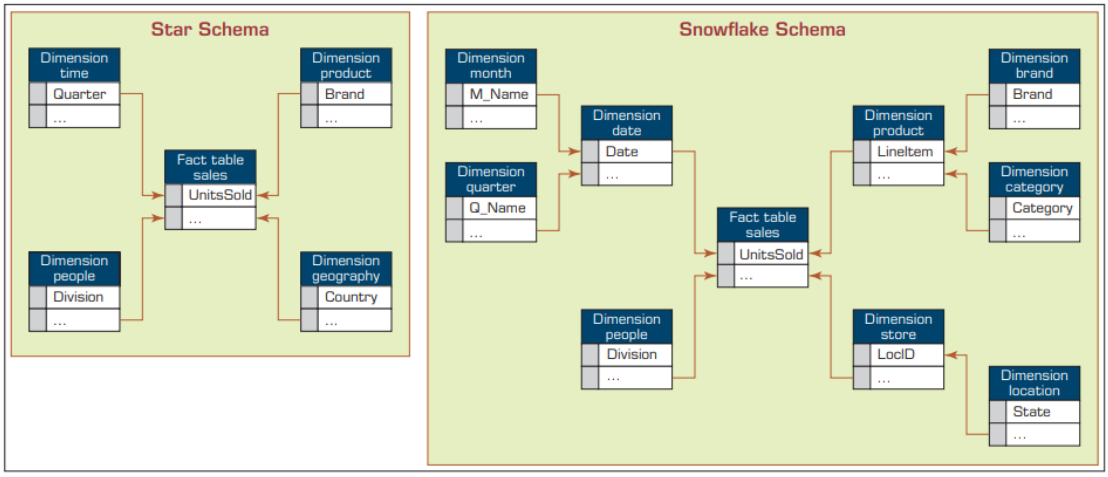
TABLE 3.4 Essential Differences between Inmon's and Kimball's Approaches

Characteristic	Inmon	Kimball
Methodology and Architecture		
Overall approach	Top-down	Bottom-up
Architecture structure	Enterprise-wide (atomic) data warehouse "feeds" departmental databases	DMs model a single business process, and enterprise consistency is achieved through a data bus and conformed dimensions
Complexity of the method	Quite complex	Fairly simple
Comparison with established development methodologies	Derived from the spiral methodology	Four-step process; a departure from RDBMS methods
Discussion of physical design	Fairly thorough	Fairly light
Data Modeling		
Data orientation	Subject or data driven	Process oriented
Tools	Traditional (entity-relationship diagrams [ERD], data flow diagrams [DFD])	Dimensional modeling; a departure from relational modeling
End-user accessibility	Low	High
Philosophy		
Primary audience	IT professionals	End users
Place in the organization	Integral part of the corporate information factory	Transformer and retainer of operational data
Objective	Deliver a sound technical solution based on proven database methods and technologies	Deliver a solution that makes it easy for end users to directly query the data and still get reasonable response times

If a company does not want to deal with the hardware or software issues when developing a DW, they can **outsource** it.

Dimensional modelling is a retrieval-based system that supports high-volume query access. A **star schema** contains a central fact table surrounded by and connected to several **dimension tables**. The **fact table** contains a large number of rows that correspond to observed facts and external links (i.e. **foreign keys**) to link to dimension tables. The attributes of fact tables consists of **performance measures** to addresses what the data warehouse supports for decision analysis. The **dimension tables** contain classification and aggregation information (i.e. the data itself) about the central fact rows. Dimension tables have a one-to-many relationship with rows in the central fact table.

The **snowflake schema** is a logical arrangement of tables in a multidimensional database in such a way that the entity-relationship diagram resembles a snowflake in shape. In the snowflake schema, however, dimensions are **normalised** into multiple related tables.



OLAP is an approach to quickly answer ad hoc questions by executing **multidimensional analytical queries** against organisational data repositories (i.e., data warehouses, DMs). **OLTP (online transaction processing system)** is a term used for a transaction system that is primarily responsible for **capturing** and **storing** data related to day-to-day business functions such as ERP, CRM, SCM, and POS. An OLTP system addresses a **critical business need**, **automating daily business transactions**, and running **real-time reports** and **routine analysis**. OLAP uses the data captured by OLTP, and OLTP automates the business processes that are managed by decisions supported by OLAP.

TABLE 3.5 A Comparison between OLTP and OLAP

Criteria	OLTP	OLAP
Purpose	To carry out day-to-day business functions	To support decision making and provide answers to business and management queries
Data source	Transaction database (a normalized data repository primarily focused on efficiency and consistency)	Data warehouse or DM (a nonnormalized data repository primarily focused on accuracy and completeness)
Reporting	Routine, periodic, narrowly focused reports	Ad hoc, multidimensional, broadly focused reports and queries
Resource requirements	Ordinary relational databases	Multiprocessor, large-capacity, specialized databases
Execution speed	Fast (recording of business transactions and routine reports)	Slow (resource intensive, complex, large-scale queries)

OLAP is based on a cube. A **cube** in OLAP is a multidimensional data structure (actual or virtual) that allows fast analysis of data. Relational databases are not well suited for near instantaneous analysis of large amounts of data. Instead, they are better suited for manipulating records that represent a series of transactions.

Different types of operations exist:

- **Slice**. A slice is a subset of a multidimensional array (usually a two-dimensional representation) corresponding to a single value set for one

(or more) of the dimensions not in the subset. A simple slicing operation on a three-dimensional cube is shown in Figure 3.11.

- **Dice**. The dice operation is a slice on more than two dimensions of a data cube.
- **Drill Down/Up**. Drilling down or up is a specific OLAP technique whereby the user navigates among levels of data ranging from the most summarised (up) to the most detailed (down).
- **Roll-up**. A roll-up involves computing all the data relationships for one or more dimensions. To do this, a computational relationship or formula might be defined.
- **Pivot**. This is used to change the dimensional orientation of a report or ad hoc query-page display.

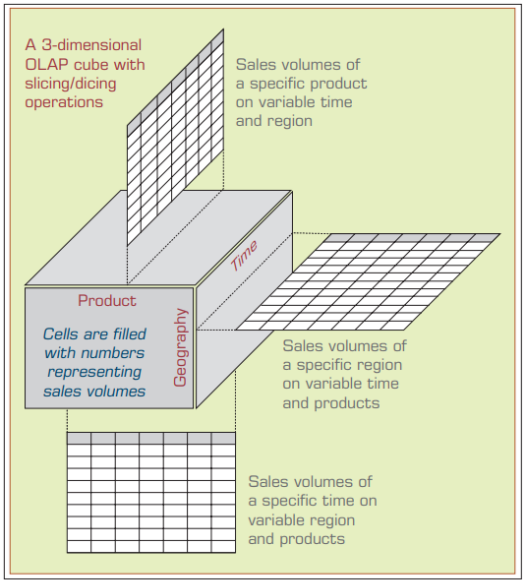


FIGURE 3.11 Slicing Operations on a Simple Three-Dimensional Data Cube.

3.7 Data Warehousing Implementation Issues

Because company want to compare their BI practices, a maturity model has been developed consisting of six stages: **prenatal**, **infant**, **child**, **teenager**, **adult**, and **sage**. Business value rises as the data warehouse progresses through each succeeding stage.

Risks when developing a data warehouse:

- Starting with the wrong sponsorship chain.
- Setting expectations that you cannot meet.
- Engaging in politically naive behaviour.
- Loading the warehouse with information just because it is available (data landfill).
- Believing that data warehousing database design is the same as transactional database design.
- Choosing a data warehouse manager who is technology oriented rather than user oriented.
- Focusing on traditional internal record-oriented data and ignoring the value of external data and of text, images, and, perhaps, sound and video.
- Delivering data with overlapping and confusing definitions.
- Believing promises of performance, capacity, and scalability.
- Believing that your problems are over when the data warehouse is up and running.
- Focusing on ad hoc data mining and periodic reporting instead of alerts.

Furthermore, you also need support from senior management and user participation in the development of data and access modelling.

A DW needs to support scalability. A DW must scale both horizontally and vertically. Good scalability means that queries and other data-access functions will grow (ideally) linearly with the size of the warehouse.

3.8 Data Warehouse Administration, Security Issues, and Future Trends

Data warehouses need strong monitoring to sustain satisfactory efficiency and productivity. A **data warehouse administrator (DWA)** should be familiar with high-performance software, hardware, and networking technologies and must possess excellent communication skills.

Security in a data warehouse should focus on four main areas:

1. Establishing effective corporate and security **policies and procedures**.
2. Implementing logical security **procedures and techniques** to restrict access.
3. Limiting **physical access** to the data centre environment.
4. Establishing an effective **internal control review process** with an emphasis on security and privacy.

Future of data warehouses:

- **Web, social media, and Big Data**: Social media is a rich source of data. Because of the sheer volume, velocity, and variety of the data it is also called Big Data.
- **Open source software**: (1) Recession is driving up the interest for low-cost software; (2) open source tools are coming into a new level of maturity; (3) open source software augments traditional enterprise software without replacing it.
- **SaaS (software as a service)**: SaaS is a way of licensing software to users as a service on their own or their consumers' servers. Finding SaaS providers for data warehouse consumers can be challenging but will improve as software offerings become more agile.
- **Cloud computing**: Software resources are pooled and virtualised, so that they can be freely allocated to applications and software platforms as resources are needed. The dynamic allocation of a cloud is particularly useful when the data volume of the warehouse varies unpredictably, making capacity planning difficult.
- **Data lakes**: A data lake is a large storage location that can hold huge quantities of data (structured, unstructured, or semi-structured) in its native/raw format for a potential future use. A data lake is not a replacement for a data warehouse but complements it. The main commonality between a data lake and data warehouse is that they are both data storage mechanisms, and conversely, the main difference is that one is all about structured/tabular data and the other is about all kinds of data (i.e., Big Data).

TABLE 3.6 A Simple Comparison between a Data Warehouse and a Data Lake

Dimension	Data Warehouse	Data Lake
The nature of data	Structured, processed	Any data in raw/native format
Processing	Schema-on-write (SQL)	Schema-on-read (NoSQL)
Retrieval speed	Very fast	Slow
Cost	Expensive for large data volumes	Designed for low-cost storage
Agility	Less agile, fixed configuration	Highly agile, flexible configuration
Novelty/newness	Not new/matured	Very new/maturing
Security	Well-secured	Not yet well-secured
Users	Business professionals	Data scientists

- **Columnar**: A column-oriented database management system (also commonly called a columnar database) is a system that stores data tables as sections of columns of data rather than as rows of data (which is the way most RDBMS do it). *Column-oriented organisations* are more efficient when (1) an aggregate needs to be computed over many rows but only for a notably smaller subset of all columns of data; and (2) new values of a column are supplied for all rows at. *Row-oriented organisations* are more efficient when (1) many columns of a single row are required at the same time; and (2) writing a new row if all of the column data is supplied at the same time. In addition, a columnar database is better at comparisons of columns.
- **Real-time data warehousing**: Real-time data warehousing (RDW) implies that the refresh cycle of an existing data warehouse to update the data is more frequent (near-real time).
- **Data warehouse appliances (all-in-one solutions to DW)**: A data warehouse appliance consists of an integrated set of servers, storage, operating system(s), database management systems, and software optimised for data warehousing. They improve performance by parallel processing at a low cost of ownership. They reduce administrations for day-to-day operations, setup, and integration.
- **Data management solutions and practices**: Practices are needed to update data management tool: **master data management (MDM)**. (1) tighter integration with operational systems demands MDM; (2) most data warehouses still lack MDM and data quality functions; and (3) regulatory and financial reports must be perfectly clean and accurate.
- **In-database processing technology (putting the algorithms where the data is)**: In-database processing (also called in-database analytics) refers to the integration of the algorithmic extent of data analytics into data warehousing. This makes for significant efficiency and performance improvements.
- **In-memory storage technology (moving the data in the memory for faster processing)**: To use RAM instead of a hard disk to increate speed. It is expensive.

- **New database management systems**: A database management system (DBMS) is a basic part of a database and must therefore also evolve.
- **Advanced analytics**: The use of more advanced methods than OLAP such as artificial intelligence.

3.9 Business Performance Management

The term **business performance management (BPM)** refers to the business processes, methodologies, metrics, and technologies used by enterprises to measure, monitor, and manage business performance. It encompasses three key components:

1. A set of **integrated, closed-loop management and analytic processes** (supported by technology) that addresses financial as well as operational activities.
2. Tools for businesses to **define** strategic goals and then **measure** and **manage performance** against those goals.
3. A **core set of processes**, including financial and operational planning, consolidation and reporting, modelling, analysis, and monitoring of key performance indicators (KPIs), linked to organisational strategy.

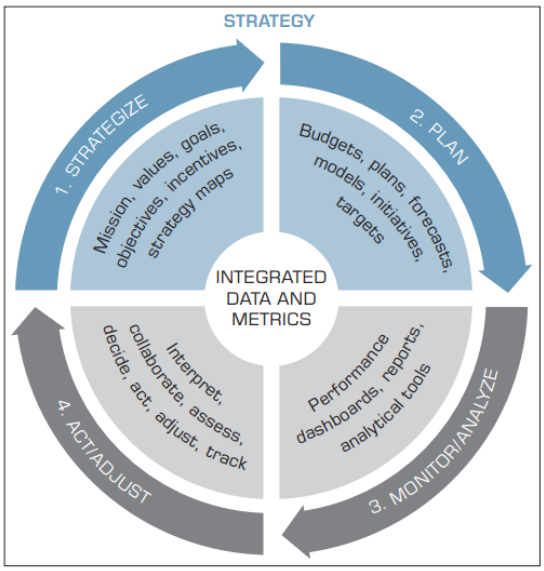


FIGURE 3.12 Closed-Loop BPM Cycle.

BPM encompasses a closed-loop set of processes that link strategy to execution to **optimise business performance** (see Figure 3.12).

1. **Strategise: *Where do we want to go?*** A high-level plan of action, encompassing a long period of time (often several years) to achieve a defined goal. More specifically, it is the process of identifying and stating the organisation's mission, vision, and objectives, and developing plans (at different levels of granularity—strategic, tactical, and operational) to achieve these objectives.
2. **Plan: *How do we get there?*** What tactics and initiatives will be pursued to meet the performance targets established by the strategic plan? What are the expected financial results of executing the tactics?
    a. An **operational plan** translates an organisation's strategic objectives and goals into a set of well-defined tactics and initiatives, resource requirements, and expected results for some future time period, usually, but not always, a year. The key to successful operational planning is **integration**. Strategy drives tactics, and tactics drive results.
    b. The **financial planning** and **budgeting process** has a logical structure that typically starts with those tactics that generate some form of revenue or income.

3. **Monitor/Analyse: *How are we doing?*** A comprehensive framework for monitoring performance should address two key issues: what to monitor and how to monitor. An organisation has to monitor specific issues in the form of indicators or measures, usually key performance indicators (KPIs). One way of selecting these KPIs is using the balanced scorecard method.
4. **Act and Adjust: *What do we need to do differently?*** Virtually all strategies depend on new projects. Because these projects often fail, the answer to the question of "What do we need to do differently?" becomes a vital issue.

3.10 Performance Measurement

Performance measurement systems assist managers in tracking the implementations of business strategy by comparing actual results against strategic goals and objectives. Since raw numbers are rarely of value, results need to be compared against other numbers. A **KPI** represents a strategic objective and measures performance against a goal. KPIs are multidimensional, i.e. have a variety of distinguishing features:

- **Strategy**: KPIs embody a strategic objective.
- **Targets**: KPIs measure performance against specific target.
- **Ranges**: Targets have performance ranges.
- **Encodings**: Ranges are encoded in software, enabling the visual display of performance.
- **Time frames**: Targets are assigned time frames by which they must be accomplished.
- **Benchmarks**: Targets are measured against a baseline or benchmark.

Outcome KPIs—sometimes known as **lagging indicators**—measure the output of past activity (e.g. revenues).
Driver KPIs—sometimes known as **leading indicators** or **value drivers**—measure activities that have a significant impact on outcome KPIs (e.g., sales leads).

Four examples of operational areas for KPIs:

- Customer performance: customer satisfaction, speed, and accuracy.
- Service performance: Service-call resolution rates, service renewal rates.
- Sales operations: New pipeline accounts, sales meetings secured.
- Sales/plan forecast: Price-to-purchase accuracy, order-to-fulfilment ratio.

Any performance management system has a performance measurement system but not the other way around.

3.11 Balanced Scorecards

The **balanced scorecard** suggests that we view the organisation from four perspectives—customer, financial, internal business processes, and learning and growth—and to develop objectives, measures, targets, and initiatives relative to each of these perspectives.

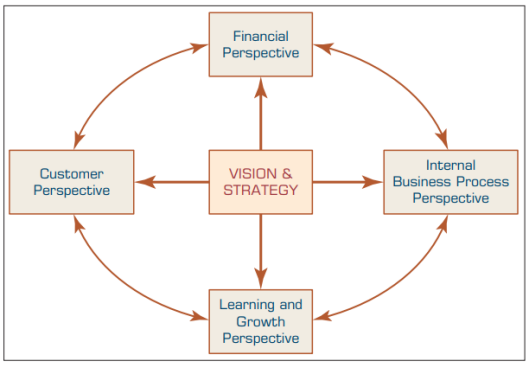


FIGURE 3.13 Four Perspectives in Balanced Scorecard Methodology.

- **The customer perspective**: Focusses on customer focus and customer satisfaction. Metrics for this are kinds of customers and the kinds of processes for which we are providing a product or service to those customer groups.
- **The financial perspective**: Focusses on timely and accurate funding data possibly streamlined by a corporate database so that it is centralised an automated. The current emphasis on financials leads to the "unbalanced" situation with regard to other perspectives.
- **The learning and growth perspective**: "To achieve our vision, how will we sustain our ability to change and improve?" It includes employee training, knowledge management, and corporate cultural characteristics related to both individual and corporate-level improvement.
- **The internal business process perspective**: Metrics focus on how well the internal business processes and functions are running and whether the outcomes of these processes meet and exceed the customer requirements.

3.12 Six Sigma as a Performance Measurement System

Six Sigma provides the means to measure and monitor key processes related to a company's profitability and to accelerate improvement in overall business performance. Sigma σ is used to measure **variability** in processes, which is in the quality arena synonymous with the number of defects. Six Sigma is the highest (best) achievable level. Six Sigma depends on **DMAIC**, a closed-loop business improvement model:

1. **Define**: Define the goals, objectives, and boundaries of the improvement activity.
2. **Measure**: Establish quantitative measures to measure the existing system.
3. **Analyse**: Analyse the system to identify ways to eliminate the gap between the current performance of the system or process and the desired goal.
4. **Improve**: Initiate actions to eliminate the gap by finding ways to do things better, cheaper, or faster.
5. **Control**: Realise this better system by modifying compensation and incentive systems, policies, procedures, etc.

TABLE 3.7 Comparison of the Balanced Scorecard and Six Sigma

Balanced Scorecard	Six Sigma
Strategic management system	Performance measurement system
Relates to the longer-term view of the business	Provides snapshot of business's performance and identifies measures that drive performance toward profitability
Designed to develop a balanced set of measures	Designed to identify a set of measurements that impact profitability
Identifies measurements around vision and values	Establishes accountability for leadership for wellness and profitability
Critical management processes are to clarify vision/strategy, communicate, plan, set targets, align strategic initiatives, and enhance feedback	Includes all business processes—management and operational
Balances customer and internal operations without a clearly defined leadership role	Balances management and employees' roles; balances costs and revenue of heavy processes
Emphasizes targets for each measurement	Emphasizes aggressive rate of improvement for each measurement, irrespective of target
Emphasizes learning of executives based on feedback	Emphasizes learning and innovation at all levels based on process feedback; enlists all employees' participation
Focuses on growth	Focuses on maximizing profitability
Heavy on strategic content	Heavy on execution for profitability
Management system consisting of measures	Measurement system based on process management

4. Predictive Analytics I: Data Mining Process, Methods, and Algorithms

4.2 Data Mining Concepts and Applications

Data mining is a term used to describe discovering or "mining" knowledge from large amounts of data. It can also be defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases," where the data are organised in records structured by categorical, ordinal, and continuous variables.

FIGURE 4.1 Data Mining Is a Blend of Multiple Disciplines.

In general, data mining seeks to identify four major types of patterns:

1. **Associations** find the commonly co-occurring groupings of things.
2. **Predictions** tell the nature of future occurrences of certain events based on what has happened in the past.
3. **Clusters** identify natural groupings of things based on their known characteristics.
4. **Sequential relationships** discover time-ordered events.

Generally speaking, data mining tasks can be classified into three main categories. These learning algorithms in these categories can be either supervised or unsupervised. With **supervised learning** algorithms, the training data includes both the descriptive attributes (i.e., independent variables or decision variables) as well as the class attribute (i.e., output variable or result variable). In contrast, with **unsupervised learning** the training data includes only the descriptive attributes.

- **Prediction** is commonly referred to as the act of telling about the future. Forecasting is another term for prediction.
- The objective of **classification** is to analyse the historical data stored in a database and automatically generate a model that can predict future behaviour into classes. Common classification tools are:
    - **Neural networks** involve the development of mathematical structures that have the capability to learn from past experiences presented in the form of well-structured data sets.
        - Advantages: Effective when the number of variables involved is rather large and the relationships among them are complex and imprecise.
        - Disadvantages: They need considerable training, cannot be trained on very large datasets, and hard to give a rationale for a decision.
    - **Decision trees** classify data into a finite number of classes based on the values of the input variables. They consists of if-else statements and are therefore faster than neural networks. However, they require **discretisation**, i.e. converting continuous valued numerical variables to ranges and categories.
    - **Rule induction**: Same as decision trees, only if-then statements are induced from the training data directly, and they need not be hierarchical in nature.

  Statistics-based classification techniques make unrealistic assumptions about the data, such as independence and normality—which limit their use in classification-type data mining projects.
- **Clustering** partitions a collection of things into segments whose members share similar characteristics. Unlike in classification, in clustering the class labels are unknown.
- **Associations**, or association rule learning in data mining, is a popular and well-researched technique for discovering interesting relationships among variables in large databases. Two commonly used derivatives of association rule mining are **link analysis** and **sequence mining**. With link analysis, the linkage among many objects of interest is discovered automatically. With sequence mining, relationships are examined in terms of their order of occurrence to identify associations over time.
- **Visualisation and time-series forecasting**. Visualisation can be used in conjunction with other data mining techniques to gain a clearer understanding of underlying relationships. In time-series forecasting, the data consists of values of the same variable that is captured and stored over time in regular intervals.

Data mining is different from statistics when it comes to the size of the data set (large vs small) and the starting situation (loosely vs well-defined proposition and hypothesis).

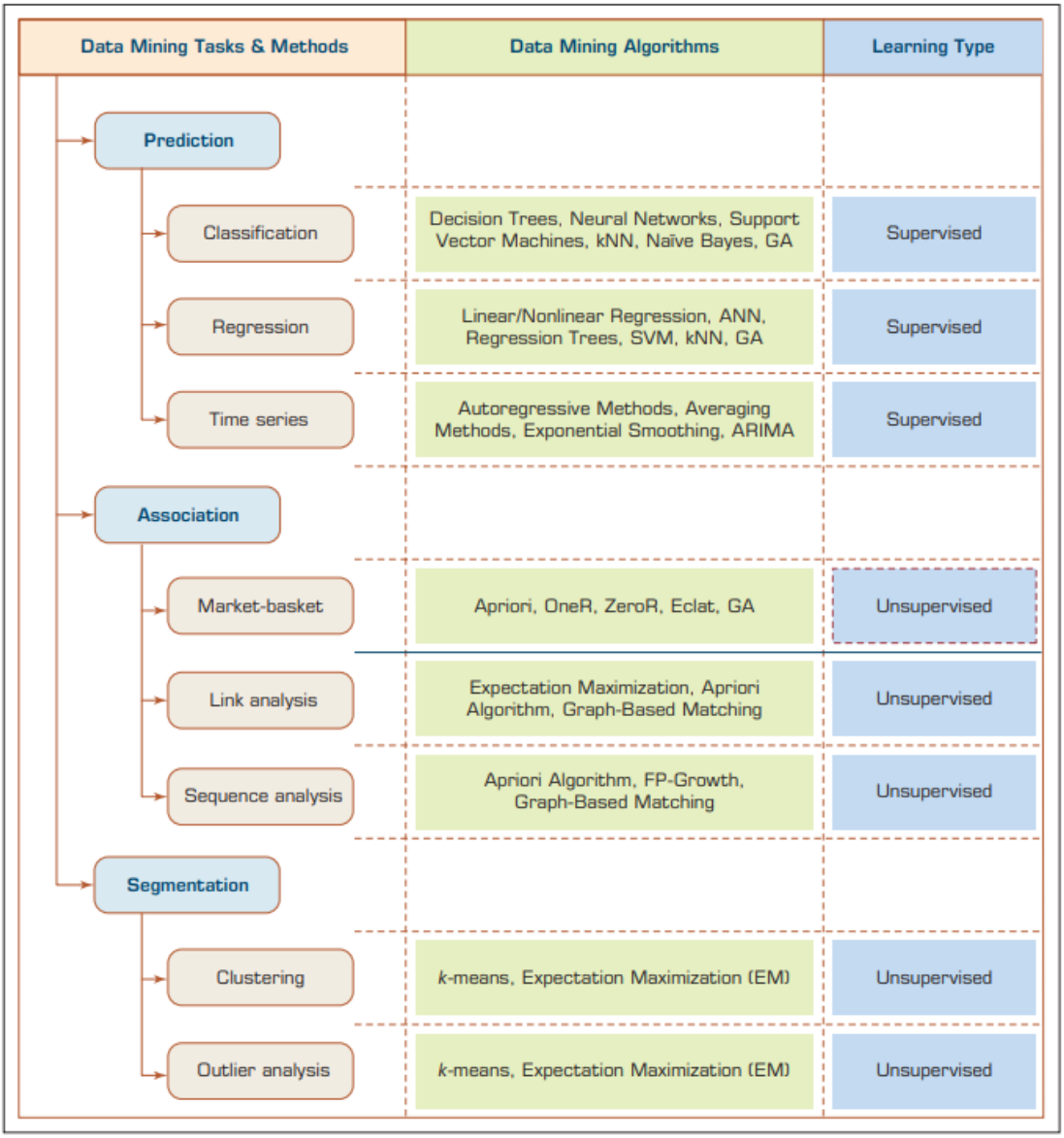


FIGURE 4.2 A Simple Taxonomy for Data Mining Tasks, Methods, and Algorithms.

4.3 Data Mining Applications

This section describes the usage of data mining in various domains. Below I list a few examples I selected from the complete list.

- Customer relationship management. The goal of CRM is to create one-on-one relationships with customers by developing an intimate understanding of their needs and wants. Root cause analysis of customer attrition,, customer profiling.
- Banking: Automating a loan application process, detecting fraudulent transactions.
- Retailing and logistics: Market-basket analysis, forecasting of consumption levels.

- Manufacturing and production: Predict machinery failures, identify anomalies and commonalities in production systems.
- Brokerage and securities trading: Predict bond prices, identify and prevent fraudulent activities in securities trading.
- Insurance: Forecast claim amounts for properties, identify and prevent incorrect claim payments and fraudulent activities.
- Computer hardware and software: Predict disk failures, identify potentially unsecure software products.
- Government and defence: forecast the cost of moving military personnel and equipment, predict resource consumption for better planning and budgeting.
- Travel industry: Predict sales of different services in order to optimally price services to maximize revenues as a function of time-varying transactions (yield management), forecast demand at different locations.
- Healthcare: forecast the level and the time of demand at different service locations, understand the underlying reasons for customer and employee attrition.
- Medicine: identify novel patterns to improve survivability of patients with cancer, identify the functions of different genes in the human chromosome.
- Entertainment industry: analyse view data, predict financial success of films.
- Homeland security and law enforcement: identify patterns of terrorist behaviours, discover crime patterns.
- Sports

4.4 Data Mining Process

To systematically carry out data mining projects and thereby maximise chances of success, a standardised process such as **CRISP-DM** can be follow. CRISP-DM, or Cross-Industry Standard Process for Data Mining consists of six steps:

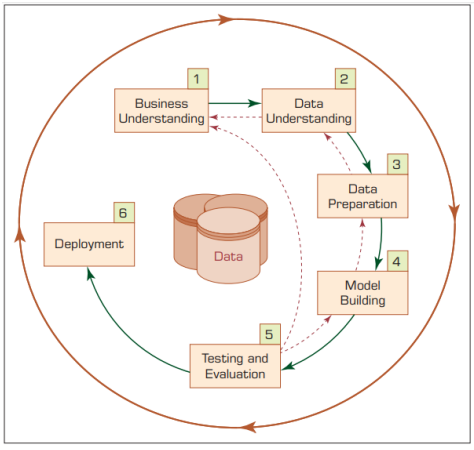


FIGURE 4.3 The Six-Step CRISP-DM Data Mining Process.

1. Business understanding: Understand the goal of the study.
2. Data understanding: Select and analyse appropriate data.
3. Data preparation: Data preprocessing. Select and transform raw data into final data.
4. Model building: Select and implement models.
5. Testing and evaluation: Assess and evaluate models for accuracy and generality. Then, decide what value the models provide for the business.
6. Deployment: Put the models to use (e.g. reports).

In addition to CRISP-DM, there also is **SEMMA** which stands for sample, explore, modify, model, and assess. SEMMA is an iterative experimental cycle used to design and evaluate models. One key difference between SEMMA and CRISP-DM is that CRISP-DM take a more comprehensive approach by also looking at the business side, whereas SEMMA assumes that the data mining project's goals and objectives along with the appropriate data sources have been identified and understood.

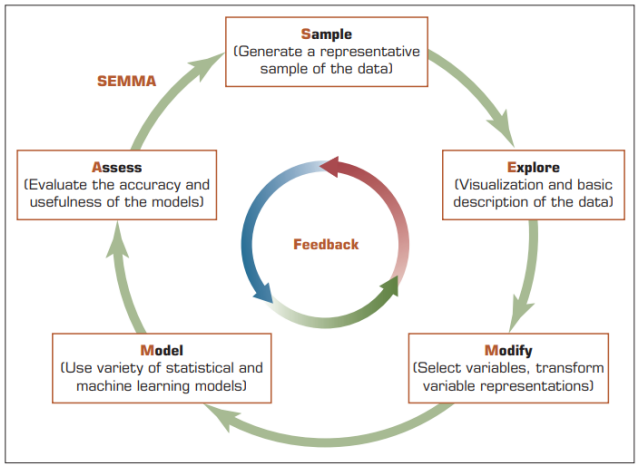


FIGURE 4.5 SEMMA Data Mining Process.

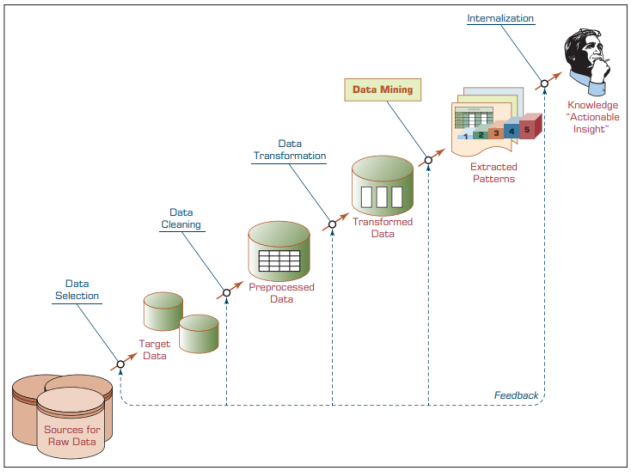


FIGURE 4.6 KDD (Knowledge Discovery in Databases) Process.

Knowledge discovery in databases (KDD) (previously used as a term for data mining) is the process of using data mining methods to find useful information and patterns in the data, as opposed to data mining, which involves using algorithms to identify patterns in data derived through the KDD process.

4.5 Data Mining Methods

Classification

Classification learns patterns from past data to place new instances (with unknown labels) into their respective groups or classes. If what is being predicted is a class label, the prediction problem is called a classification, whereas if it is a numeric value, the prediction problem is called a **regression**. Even though both clustering and classification determine groups, their difference is that classification looks at the relationship between the input and output variables (supervised learning), whereas clustering only looks at groups from the input variables (unsupervised learning). Factors to assess models are:

- Predictive accuracy: Calculate the accuracy rate based on Figure 4.8.
- Speed
- Robustness
- Scalability
- Interpretability

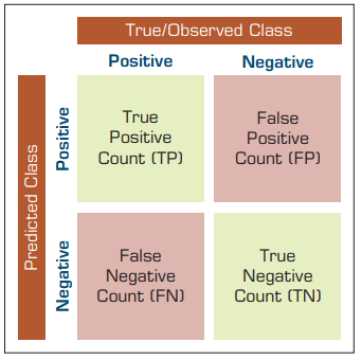


FIGURE 4.8 A Simple Confusion Matrix for Tabulation of Two-Class Classification Results.

Estimating the True Accuracy

The primary source for accuracy estimation is the **confusion matrix** (also called a **classification matrix** or a **contingency table**). From this table, we can calculate various accuracy metrics.

TABLE 4.1 Common Accuracy Metrics for Classification Models

Metric	Description
$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	The ratio of correctly classified instances (positives and negative) divided by the total numbers of instances
$\text{True Positive Rate} = \frac{TP}{TP + FN}$	(a.k.a. Sensitivity) The ratio of correctly classified positives divided by the total positive count (i.e., hit rate or recall)
$\text{True Negative Rate} = \frac{TN}{TN + FP}$	(a.k.a. Specificity) The ratio of correctly classified negatives divided by the total negative count (i.e., false alarm rate)
$\text{Precision} = \frac{TP}{TP + FP}$	The ratio of correctly classified positives divided by the sum of correctly classified positives and incorrectly classified positives
$\text{Recall} = \frac{TP}{TP + FN}$	Ratio of correctly classified positives divided by the sum of correctly classified positives and incorrectly classified negatives

Accuracy metrics are useful because it allows you to know the prediction accuracy of a model and allows you to draw comparisons between models.

Popular estimation methodologies used for classification-type data mining models:

- **Simple split**: Split the data set into two parts: 2/3rd for the training set and 1/3rd for the test set. The training set is used to train the model and the test set is used to test the model. One criticism about this method is that the two data sets are the same, i.e. they have the same properties.

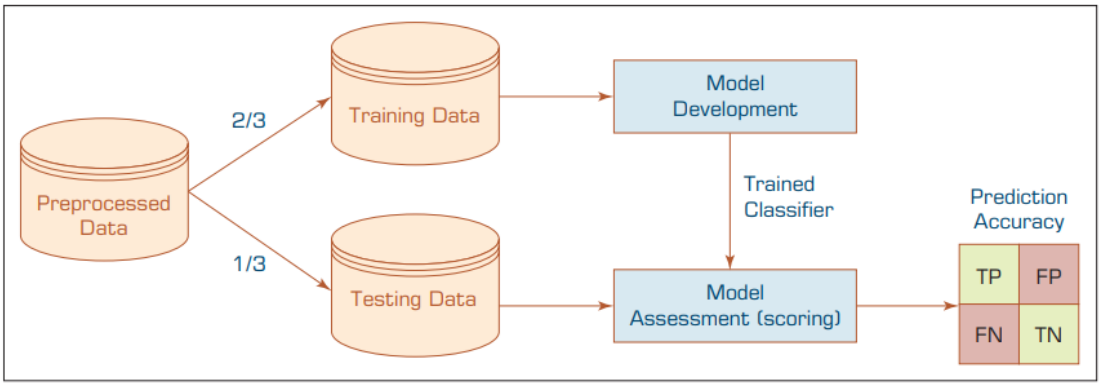


FIGURE 4.9 Simple Random Data Splitting.

- **K-fold cross validation**: In k-fold cross-validation, also called rotation estimation, the complete data set is randomly split into k mutually exclusive subsets of approximately equal size. The classification model is trained and tested k times. Each time it is trained on all but one fold and then tested on the remaining single fold. The cross-validation estimate of the overall accuracy of a model is calculated by simply averaging the k individual accuracy measures, as shown in the following equation:

$$CVA = \frac{1}{k}\sum_{i=1}^{k} A_i$$

where CVA stands for cross-validation accuracy, k is the number of folds used, and A is the accuracy measure of each fold.

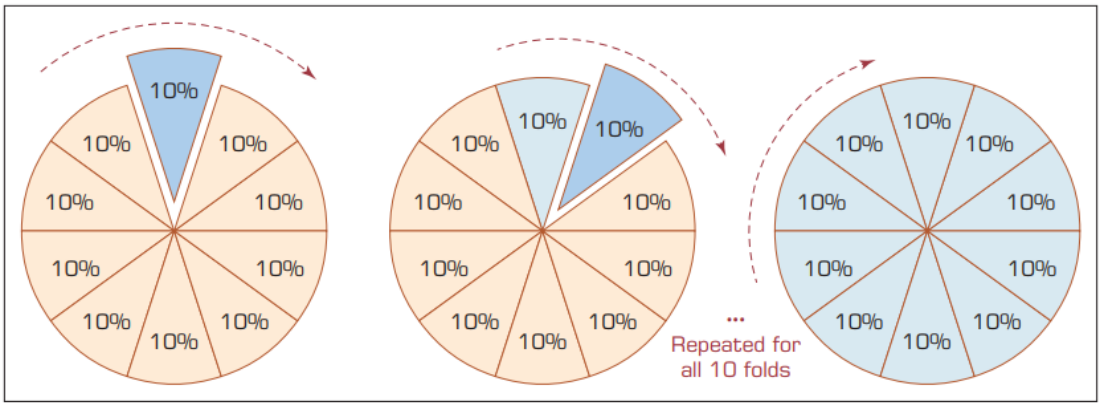


FIGURE 4.10 A Graphical Depiction of *k*-Fold Cross-Validation.

- **Leave-one-out**: Similar to K-fold cross-validation where each data point is a fold. This is a time consuming methodology, but sometimes for small data sets it is a viable option.
- **Bootstrapping**: Data are sampled (with replacement) for training, and the rest is used for testing.
- **Jackknifing**: Similar to leave-one-out, with jackknifing the accuracy is calculated by leaving one sample out at each iteration of the estimation process.
- **Area under the ROC curve (AUC)**: A graphical assessment technique where the true positive rate is plotted on the y-axis and the false positive rate is plotted on the x-axis.

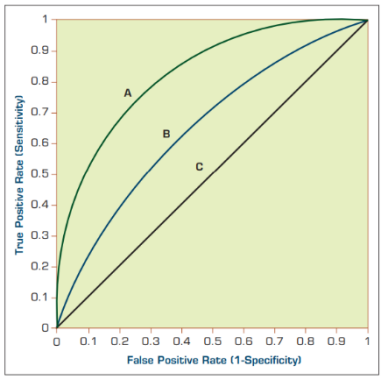


FIGURE 4.11 A Sample ROC Curve.

A number of techniques (or algorithms) are used for classification modelling, including the following:

- **Decision tree analysis**
- **Statistical analysis**: Techniques such as logistic regression and discriminant analysis that make assumptions about the relationship between the variables.
- **Neural networks**
- **Case-based reasoning**: This approach uses historical cases to recognise commonalities to assign a new case into the most probable category.
- **Bayesian classifiers**
- **Genetic algorithms**
- **Rough sets**: This method takes into account the partial membership of class labels to predefined categories in building models (collection of rules) for classification problems.

Creating **ensembles** is the process of intelligently combining the forecasts or predictions created and provided by two or more prediction models to improve the accuracy and robustness s of information outcomes. These combinations can be either **homogeneous** (i.e. the same type of models), for

example boosting or bagging, or **heterogeneous** (i.e. different types of models). One of the advantages of heterogeneous ensembles is that because the models are different, they look at the data from different perspectives. They can be combined by giving each model a single vote or by weighted voting.

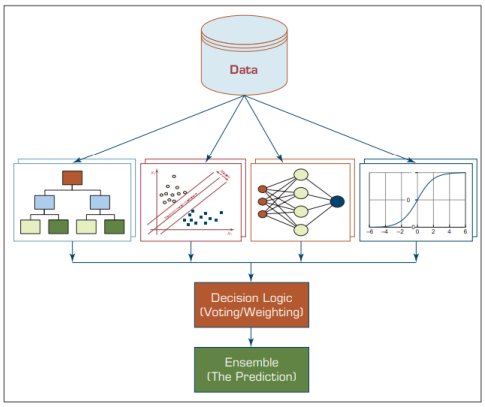


FIGURE 4.12 Graphical Illustration of a Heterogeneous Ensemble.

A **decision tree** contains input variables (**attributes**) and the resulting output is a class label. A **branch** represents the outcome of a test to classify a pattern one of the attributes. A **leaf node** at the end represents the final class choice for a pattern. The data set is recursively split until each division consists entirely or primarily from one class. Each non-leaf node of the tree contains a **split point**, which is a test on one or more attributes and determines how the data are to be divided further. Trees are built from the bottom up and then pruned to increase its generalisation. Attributes are split based on their optimal value. A general algorithm for building a decision tree is as follows:

1. Create a root node and assign all of the training data to it.
2. Select the best splitting attribute.
3. Add a branch to the root node for each value of the split. Split the data into mutually exclusive (nonoverlapping) subsets along the lines of the specific split and move to the branches.
4. Repeat steps 2 and 3 for each and every leaf node until the stopping criteria is reached (e.g., the node is dominated by a single class label).

One way to measure the optimal value of a split is the **Gini index**:

$$\text{gini}(S) = 1 - \sum_{j=1}^{n} p_j^2$$

where p_j is the relative frequency of class j in S. The attribute/split combination that provides the smallest $\text{gini}_{\text{split}}(S)$ is chosen to split the node.

Information gain or **entropy** measures the extent of uncertainty or randomness in a data set. If all the data in a subset belong to just one class, there is no uncertainty or randomness in that data set, so the entropy is zero. The amount of information gained by splitting attribute A in the set S with values p and n is:

$$E(A) = \sum_{i=1}^{n} \frac{p_i + n_1}{p + n} I(p_i, n_i)$$

In other words, the information gained by splitting attribute A would be

$$\text{Gain}(A) = I(p, n) - E(A)$$

where $I(p, n)$ is the amount of information before the split.

Cluster Analysis for Data Mining

Clustering is an exploratory data analysis tool for classifying items, events, or concepts into common groupings called **clusters**, so that the degree of association is strong among members of the same cluster and weak among members of different clusters. Each cluster describes the class to which its members belong. Most clustering algorithms need a fixed number of clusters to work with. Ways of determining the number of clusters is to look at the variance explained as a function of the number of clusters, look at the Akaike information criterion (AIC), which is a measure of the goodness of fit (based on the concept of entropy), or at the Bayesian information criterion (BIC), which is a model-selection criterion (based on maximum likelihood estimation). Clustering analysis methods can be either

- **Divisive**: With divisive classes, all items start in one cluster and are broken apart.
- **Agglomerative**: With agglomerative classes, all items start in individual clusters, and the clusters are joined together.

Most clustering algorithms also need a **distance measure**, such as Euclidian distance (the distance you would measure with a ruler) or Manhattan distance (rectilinear distance).

The **k-means** algorithm (where k stands for the predetermined number of clusters) assigns each data point to the cluster whose centre (also called the centroid) is the nearest. The centre is calculated as the average of all the points in the cluster; that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.

Initialization step: Choose the number of clusters (i.e., the value of k).
Step 1: Randomly generate k random points as initial cluster centres.
Step 2: Assign each point to the nearest cluster centre.
Step 3: Recompute the new cluster centres.
Repetition step: Repeat steps 2 and 3 until some convergence criterion is met (usually that the assignment of points to clusters becomes stable).

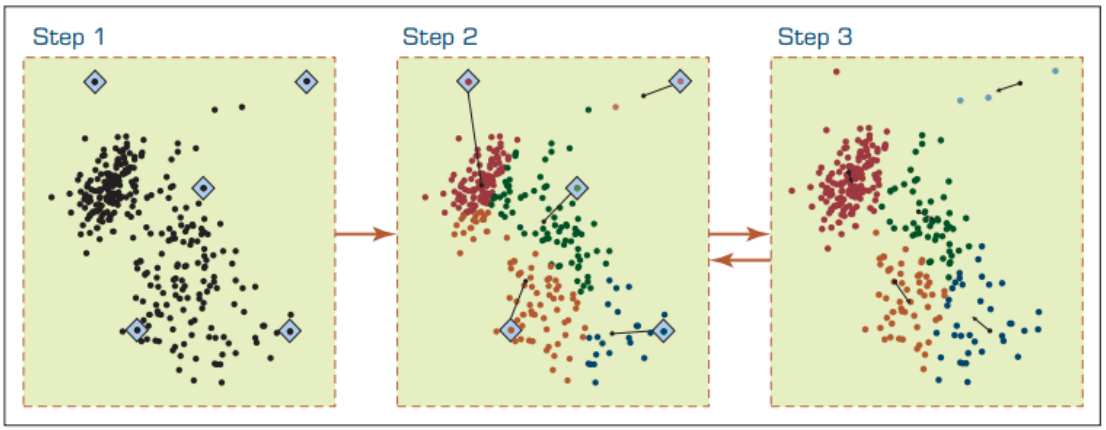


FIGURE 4.13 A Graphical Illustration of the Steps in the *k*-Means Algorithm.

Association Rule Mining

Association rule mining (also known as affinity analysis or market-basket analysis) aims to find interesting relationships (affinities) between variables (items) in large databases. Because of its successful application to retail business problems, it is commonly called market-basket analysis. s. The main idea in market-basket analysis is to identify strong relationships among different products (or services) that are usually purchased together. Association rule mining uses two common metrics: support, and confidence and lift. An association rule looks like this:

$$X \Rightarrow Y\ [Supp(\%), Conf(\%)$$

$$Support(S) = Supp(X \Rightarrow Y) = \frac{number\ of\ baskets\ that\ contains\ both\ X\ and\ Y}{total\ number\ of\ baskets}$$

The support (S) of a collection of products is the measure of how often these products and/or services appear together in the same transaction; that is, the proportion of transactions in the data set that contain all of the products and/or services mentioned in a specific rule.

$$Confidence(C) = Conf(X \Rightarrow Y) = \frac{Supp(X \Rightarrow Y)}{Supp(X)}$$

The confidence of a rule is the measure of how often the products and/or services on the right-hand side (RHS) (consequent) go together with the products and/or services on the left-hand side (LHS) (antecedent), that is, the proportion of transactions that include LHS while also including the RHS. In other words, it is the conditional probability of finding the RHS of the rule present in transactions where the LHS of the rule already exists.

$$\text{Lift}(\text{X} \Rightarrow \text{Y}) = \frac{\text{Conf}(\text{X} \Rightarrow \text{Y})}{\text{Expected Conf}(\text{X} \Rightarrow \text{Y})} = \frac{\text{S(X)}}{\frac{\text{S(X)} * \text{S(Y)}}{\text{S(X)}}} = \frac{\text{S(X} \Rightarrow \text{Y)}}{\text{S(X)} * \text{S(Y)}}$$

The lift value of an association rule is the ratio of the confidence of the rule and the expected confidence of the rule. The expected confidence of a rule is defined as the product of the support values of the LHS and the RHS divided by the support of the LHS.

The **Apriori algorithm** is the most commonly used algorithm to discover association rules. Given a set of **itemsets**, the algorithm attempts to find subsets that are common to at least a minimum number of the itemsets (i.e., complies with a minimum support, e.g. 50% of total transactions). Apriori uses a bottom-up approach, where frequent subsets are extended one item at a time (a method known as **candidate generation**, whereby the size of frequent subsets increases from one-item subsets to two-item subsets, then three-item subsets, etc.), and groups of candidates at each level are tested against the data for

minimum support. The algorithm terminates when no further successful extensions are found.

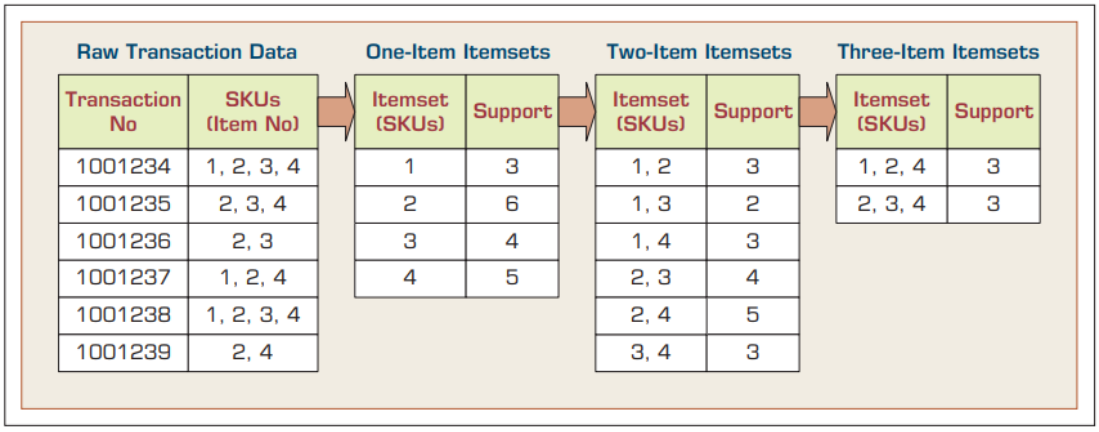


FIGURE 4.14 Identification of Frequent Itemsets in the Apriori Algorithm.

4.6 Data Mining Software Tools

Most data mining tools are developed by the well-established statistical software companies because statistics is the foundation of data mining, and these companies have the means to cost-effectively develop them into full-scale data mining systems. Commercial tools are IBM SPSS Modeler, Dell Statistica, and SAP Infinite Sight. Additionally, free tools with an intuitive user interface are **WEKA**, **KNIME**, and **RapidMiner** (though RapidMiner was commercialised). Commercial tools are usually more efficient. Furthermore, **Microsoft SQL Server** has become increasingly popular as a BI suite. There is also a free educational version called **Microsoft Enterprise Consortium**. From a large survey, it became apparent that R, Python, SQL, and Excel are the most popular tools for data mining.

4.7 Data Mining Privacy Issues, Myths, and Blunders

Data should be de-identified before applying any data mining algorithm, so that privacy is ensured and data cannot be traced back to a specific individual. There are a number of myths and blunders in the field of data mining that should be avoided.

TABLE 4.6 Data Mining Myths

Myth	Reality
Data mining provides instant, crystal-ball-like predictions.	Data mining is a multistep process that requires deliberate, proactive design and use.
Data mining is not yet viable for mainstream business applications.	The current state of the art is ready to go for almost any business type and/or size.
Data mining requires a separate, dedicated database.	Because of the advances in database technology, a dedicated database is not required.
Only those with advanced degrees can do data mining.	Newer Web-based tools enable managers of all educational levels to do data mining.
Data mining is only for large firms that have lots of customer data.	If the data accurately reflect the business or its customers, any company can use data mining.

Some blunders/pitfalls:

1. Selecting the wrong problem for data mining.
2. Ignoring what your sponsor thinks data mining is and what it really can and cannot do.
3. Beginning without the end in mind.
4. Define the project around a foundation that your data can't support.
5. Leaving insufficient time for data preparation
6. Looking only at aggregated results and not at individual records.
7. Being sloppy about keeping track of the data mining procedure and results.
8. Using data from the future to predict the future.
9. Ignoring suspicious findings and quickly moving on.
10. Starting with a high-profile complex project that will make you a superstar.
11. Running data mining algorithms repeatedly and blindly.
12. Ignore the subject matter experts.
13. Believing everything you are told about the data.
14. Assuming that the keepers of the data will be fully on board with cooperation.
15. Measuring your results differently from the way your sponsor measures them.
16. If you build it, they will come: don't worry about how to serve it up.

6. Prescriptive Analytics: Optimisation and Simulation

6.2 Model-Based Decision Making

- **Environmental scanning and analysis**: is the monitoring, scanning, and interpretation of collected information. Before decision making, the problem has to be understood.
- **Variable identification** of the model (e.g. decision, result, uncontrollable) is critical, as are the relationships among the variables.
- **Forecasting (predictive analytics)**: prerequisite of prescriptive analytics is knowing what has happened and what is likely to happen; is essential for construction and manipulating models, because when a decision is implemented, the results usually occur in the future.
- The following table lists the **categories** of models:

TABLE 6.1 Categories of Models

Category	Process and Objective	Representative Techniques
Optimization of problems with few alternatives	Find the best solution from a small number of alternatives	Decision tables, decision trees, analytic hierarchy process
Optimization via algorithm	Find the best solution from a large number of alternatives, using a step-by-step improvement process	Linear and other mathematical programming models, network models
Optimization via an analytic formula	Find the best solution in one step, using a formula	Some inventory models
Simulation	Find a good enough solution or the best among the alternatives checked, using experimentation	Several types of simulation
Heuristics	Find a good enough solution, using rules	Heuristic programming, expert systems
Predictive models	Predict the future for a given scenario	Forecasting models, Markov analysis
Other models	Solve a what-if case, using a formula	Financial modeling, waiting lines

- **Model management:** Models, like data, must be managed to maintain their integrity, and thus their applicability. Such management is done with the aid of model-based management systems, which are analogous to database management systems (DBMS).
- **Knowledge-based modelling:** DSS uses mostly quantitative models, whereas expert systems use qualitative, knowledge-based models in their applications.

6.3 Structure of Mathematical Models for Decision Support

Quantitative models are typically made up of four basic components:

Result (outcome) variables: level of effectiveness of a system; dependent variables.
Decision variables: describe alternative courses of action, e.g. the amount to invest in shares is a decision variable.

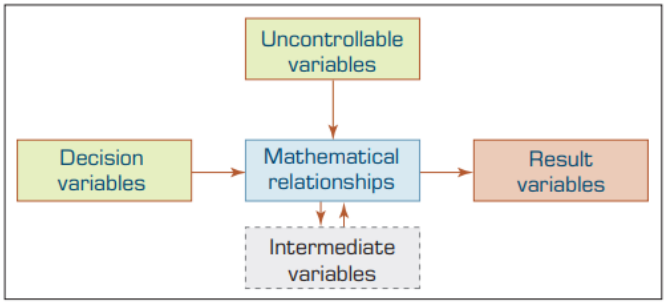


FIGURE 6.1 The General Structure of a Quantitative Model.

Uncontrollable variables/parameters: if these are fixed then they are as stated, if they 'vary', they are called variables; e.g. interest rate.

Intermediate result variables: intermediate outcomes in mathematical model.

(Almost) Complete example: Employee salaries: it constitutes a **decision variable** for management: it determines employee satisfaction (**intermediate outcome**), which determines the productivity level (**final result**).

6.4 Certainty, Uncertainty, and Risk

There are **three levels of decision making** (listed with decreasing knowledge):

- **Certainty**: assumed that complete knowledge is available; decision maker knows exactly what the outcome of each course of action will be (although not 100%). E.g. investing in U.S. Treasury bills.
- **Risk:** a decision in which the decision maker must consider several possible outcomes for each alternative, each with a given probability of occurrence. **Risk analysis** is a decision-making method that analyses the risk associated with different alternatives.
- **Uncertainty**: the decision maker considers situations in which several outcomes are possible for each course of action. In contrast to the risk situation, in this case, the decision maker does not know, or cannot estimate, the probability of occurrence of the possible outcomes.

6.6 Mathematical Programming Optimisation

Mathematical programming is a family of tools designed to help solve managerial problems in which the decision maker must allocate scarce resources among competing activities to optimise a measurable goal. **Linear programming (LP)** is perhaps the best-known optimisation model. It deals with the optimal allocation of resources among competing activities. The problem is to find the values of the decision variables X_1, X_2 , and so on, such that the value of the result variable Z is maximised, subject to a set of linear constraints that express the technology, market conditions, and other uncontrollable variables. In linear programming (LP), all relationships among the variables are all linear equations and inequalities.

LP allocation problems usually display the following characteristics:

- A limited quantity of economic resources is available for allocation.
- The resources are used in the production of products or services.
- There are two or more ways in which the resources can be used. Each is called a **solution** or a **program**.
- Each activity (product or service) in which the resources are used yields a return in terms of the stated goal.
- The allocation is usually restricted by several limitations and requirements, called **constraints**.

Even though there can be infinite solution, at least one is the best solution, i.e. the degree of goal attainment associated with it is the highest. This is called the **optimal solution**.

Every LP model is composed of:

- **decision variables** (whose values are unknown and are searched for),
- an **objective function** (a linear mathematical function that relates the decision variables to the goal, measures goal attainment, and is to be optimised),
- **objective function coefficients** (unit profit or cost coefficients indicating the contribution to the objective of one unit of a decision variable),

- **constraints** (expressed in the form of linear inequalities or equalities that limit resources and/or requirements; these relate the variables through linear relationships),
- **capacities** (which describe the upper and sometimes lower limits on the constraints and variables), and
- **input/output (technology) coefficients** (which indicate resource utilisation for a decision variable).

There are three aspects to LP that should be answered:

- What do we control?
- What do we want to achieve?
- What constrains us?

6.7 Multiple Goals, Sensitivity Analysis, What-If Analysis, and Goal Seeking

A **sensitivity analysis** attempts to assess the impact of a change in the input data or parameters on the proposed solution (i.e., the result variable). it allows flexibility and adaptation to changing conditions and to the requirements of different decision-making situations, provides a better understanding of the model and the decision making situation it attempts to describe, and permits the manager to input data to increase the confidence in the model. It tests relationships such as the following:

- The impact of changes in external (uncontrollable) variables and parameters on the outcome variable(s)
- The impact of changes in decision variables on the outcome variable(s)
- The effect of uncertainty in estimating external variables
- The effects of different dependent interactions among variables
- The robustness of decisions under changing conditions

They are used for:

- Revising models to eliminate too-large sensitivities
- Adding details about sensitive variables or scenarios
- Obtaining better estimates of sensitive external variables
- Altering a real-world system to reduce actual sensitivities
- Accepting and using the sensitive (and hence vulnerable) real world, leading to the continuous and close monitoring of actual results

Automatic sensitivity analysis is performed in standard quantitative model implementations such as LP, in a fast manner. E.g. it reports the range within which a certain input variable or parameter value (e.g. unit cost) can vary without having any significant impact on the proposed solution. However, they are limited to one change at a time and only for certain variables

Trial-and-error sensitivity analysis determines the impact of changes in any variable or in several variables; i.e. you change some input data and then solve the problem again. Better solutions may be discovered.

What-If Analysis is structured as "What will happen to the solution if an input variable, an assumption, or a parameter value is changed?" An appropriate UI can easily let managers ask these questions (with immediate answers).

Goal Seeking calculates the values of the inputs necessary to achieve a desired level of an output (goal); i.e. a backward solution approach. E.g. What annual R&D budget is needed for an annual growth rate of 15% by 2018? You can compute a **break-even point.**

6.8 Decision Analysis with Decision Tables and Decision Trees

Decision tables conveniently organise information and knowledge in a systematic, tabular manner to prepare it for **(decision) analysis**. For example, say that an investment company is considering investing in one of three alternatives: bonds, stocks, or certificates of deposit (CDs). The company is interested in one goal: maximizing the yield on the investment after 1 year. If it were interested in other goals, such as safety or liquidity, the problem would be classified as one of **multicriteria decision analysis.**

For **Treating uncertainty,** several methods exist:

- for example the **optimistic approach** assumes that the best possible outcome of each alternative will occur and then selects the best of the best.
- The **pessimistic approach** assumes that the worst possible outcome for each alternative will occur and selects the best of these.
- Another simply assumes that all states of nature are equally possible.

Treating risk is about selecting the alternative with the greatest expected value. A **decision tree** shows the relationships of the problem graphically and can handle complex situations in a compact form. With many alternatives it can be cumbersome.

7. Big Data Concepts and Tools

7.2 Definition of Big Data

Big Data has become a popular term to describe the exponential growth, availability, and use of information, both structured and unstructured.

The three "V"s, plus some additional "V"s.

- **Volume**
- **Variety**
- **Velocity**
- **Veracity:** accuracy, quality, truthfulness, or trustworthiness of the data.
- **Variability**: highly inconsistent with periodic peak (e.g. trending on social media).
- **Value proposition**: organisations can gain greater business value with big data.

7.3 Fundamentals of Big Data Analytics

The next picture describes the **keys to success** concerning Big Data:

These techniques are collectively called **high-performance computing:**

- **In-memory analytics:** Solves complex problems in near real time with highly accurate insights by allowing analytical computations and Big Data to be processed in-memory and distributed across a dedicated set of nodes.
- **In-database analytics:** Speeds time to insights and enables better data governance by performing data integration and analytic functions inside the database so you won't have to move or convert data repeatedly.
- **Grid computing:** Promotes efficiency, lower cost, and better performance by processing jobs in a shared, centrally managed pool of IT resources.
- **Appliances**: Brings together hardware and software in a physical unit that is not only fast but also scalable on an as-needed basis.

FIGURE 7.4 Critical Success Factors for Big Data Analytics.

7.7 Big Data and Stream Analytics

Stream analytics (also called data-in-motion analytics and real-time data analytics, among others) is a term commonly used for the analytic process of extracting actionable information from continuously flowing/streaming data. The data elements in a stream are often called **tuples**. In a relational database sense, a tuple is similar to a row of data (a record, an object, an instance). However, in the context of semi-structured or unstructured data, a tuple is an abstraction that represents a package of data, which can be characterized as a set of attributes for a given object.

Perpetual analytics, on the other hand, evaluates every incoming observation against all prior observations, where there is no window size. Recognising how the new observation relates to all prior observations enables the discovery of real-time insight.

Critical event processing is a method of capturing, tracking, and analysing streams of data to detect events (out of normal happenings) of certain types that are worthy of the effort. Complex event processing is an application of stream analytics that combines data from multiple sources to infer events or patterns of interest either before they actually occur or as soon as they happen.

Data stream mining, as an enabling technology for stream analytics, is the process of extracting novel patterns and knowledge structures from continuous, rapid data records.

8. Future Trends, Privacy and Managerial Considerations in Analytics

8.5 Issues of Legality, Privacy, and Ethics

Data scientists should be aware of the various legal, privacy, and ethical that may arise due to technology.

Legal issues may arise due to evolving technology that raises questions such as 'who is liable if an enterprise finds itself bankrupt as a result of using the advice of an analytic application?'. Will the enterprise itself, the auditing and accounting firms, or the software developers be held accountable for this? Typically, insights **derived** from public data are safe because they have not been directly **acquired** from insider sources.

Privacy is the right to be left alone and the right to be free from unreasonable personal intrusions. Even though privacy is an important right, there are two things that should be noted: (1) the right of privacy is not absolute. Privacy must be balanced against the needs of society. (2) The public's right to know is superior to the individual's right to privacy. More specifically, a number of privacy issues arise due to technology:

- **Collecting information about individuals** has become easier due to technology. In previous times, such connections could not be made but they are now possible due to, for example, the internet.
- **Mobile user privacy** has declined GPS tracking from phones.
- **Homeland securi**ty monitors individuals on a global scale, thereby analysing a lot of private information.
- **Internet companies monetise their users** by selling their data or by creating profiles based on their data by means of clustering or association rule mining.

Another question that arises is about who owns the data. For example, a modern car is equipped with many sensors that provide to be a 'gold mine' of data to a company. However, there is debate about if the data belongs to the company or to the user.

Finally, **ethics in decision making and support** is also a hot topic. The study of ethical issues is complex because of its **multidimensionality**. Who is the agent? What action was actually taken or is being contemplated? One story that made many users upset (although it was not illegal) some time back was Facebook's experiment to present different News Feeds to the users and monitor their emotional reactions as measured by replies, likes, sentiment analysis, and so on. Most companies, including technology companies, run user testing to identify the features most liked or disliked and fine-tune their product offerings. Because Facebook is so large, running this experiment without the users' informed consent was viewed as unethical.

8.6 Impacts of Analytics in Organisations: An Overview

The impact of analytics on organisations will be much faster than in previous revolutions (e.g. industrial revolution).

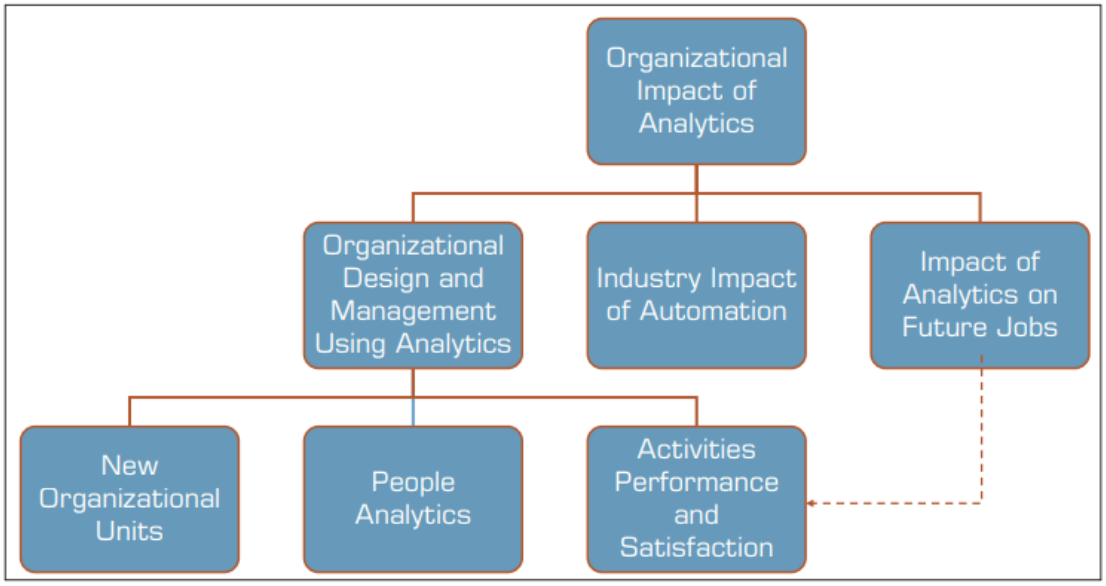


FIGURE 8.7 Impact of Analytics on Organizations.

- **New organisational units**: New units in organisation have to be formed specifically tasked with performing analytics. Next to that, new units within IT providers have been formed as well.
- **Redesign of an organisation through the user of analytics**: The behaviour of people in an organisation change due to technology. For example, LinkedIn is being used by HR departments to identify new candidates. Another example is the use of GPS sensors in badges to follow the movement and interactions of employees.
- **Analytics impact on managers' activities, performance, and job satisfaction**: Research found that employees working with automated systems (which perform a part of their job) were more satisfied with their job. Additionally, analytics help managers make better decisions more efficiently. Finally, leadership qualities could be changed because interaction is more and more digital, leading physical appearance to become less attractive.
- **Industrial restructuring**: Technology leads to industrialisation (automation) in companies, but may not work well for all jobs just yet.
- **Automation's impact on jobs**: While some jobs will be automated, this does not necessarily lead to a loss of jobs. Some argue that '**polarisation**' of the labour market will happen, which means there will be significant job growth in the top and bottom tier in the market, but not in the middle. This is because low-skilled jobs may not be easily replaceable (janitor, personal care), which is also true for very high-skilled jobs.
- **Unintended effects of analytics**: Analytical models may not be transparent due to their complexity, possibly leading to undesired effects.

Keywords

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W