R.M.K
GROUP OF ENGINEERING INSTITUTIONS

R.M.K
GROUP OF INSTITUTIONS

R.M.K
GROUP OF
INSTITUTIONS

Please read this disclaimer before proceeding:

This document is confidential and intended solely for the educational purpose of RMK Group of Educational Institutions. If you have received this document through email in error, please notify the system manager. This document contains proprietary information and is intended only to the respective group / learning community as intended. If you are not the addressee you should not disseminate, distribute or copy through e-mail. Please notify the sender immediately by e-mail if you have received this document by mistake and delete this document from your system. If you are not the intended recipient you are notified that disclosing, copying, distributing or taking any action in reliance on the contents of this information is strictly prohibited.

22AI901

Business Intelligence and Analytics

Batch/Year: 2022-2026/II

Created by:

Ms. S. Subi, Assistant Professor

Ms. J. Bhuvaneswari, Assistant Professor

R.M.K GROUP OF INSTITUTIONS

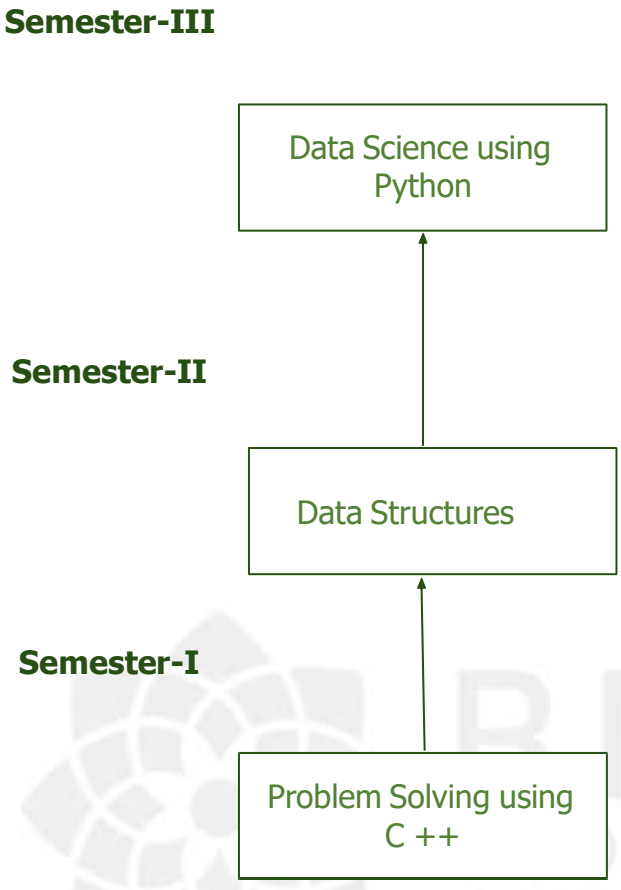
Table of Contents

R.M.K GROUP OF INSTITUTIONS

2. Course Objectives

- To understand the business intelligence (BI) methodology and concepts.
- To learn about descriptive, inferential statistics and data warehousing operations.
- To analyze wide range of applications of data mining.
- To analyze the various prescriptive analytics methods.
- To develop and deploy Business Analytic Models.

3. Pre-Requisites



4.SYLLABUS

22AI901 BUSINESS INTELLIGENCE AND ANALYTICS L T P C 2 0 2 3

UNIT I OVERVIEW OF BUSINESS INTELLIGENCE 6+6

Evolution of Computerized Decision Support to Analytics- A Framework for Business Intelligence - Analytics Overview - Analytics Examples- Introduction to Big Data Analytics- Overview of the Analytics Ecosystem.

List of Exercise/Experiments:

1.Perform Customer Segmentation, Classification using customer data of a certain organization. Analyze the data from the standpoint of paying capacity and purchasing pattern similarities among the company's clients.

2.Build a data model by taking an available data for a certain company and create a series of analysis and visualizations on various metrics related to the products of that company.

UNIT II DESCRIPTIVE ANALYTICS 6+6

The Nature of Data- Data Preprocessing- Statistical Modeling for Business Analytics- Regression Modeling for Inferential Statistics- Business Reporting- Data Visualization- Types of charts and graphs- Visual Analytics- Information Dashboards- Business Intelligence and Data Warehousing- Data Warehousing Process - Data Warehousing architecture - Data Integration and the Extraction, Transformation, and Load (ETL) Processes- Data Warehouse Development.

List of Exercise/Experiments:

1.Consider Groceries dataset for Market Basket Analysis and investigate customer's historical transactions. Focus on descriptive analytics of customer's purchase behavior, revealing interesting combinations of products that are frequently bought together, and creating valuable suggestions for the company.

2.Given Life Expectancy (WHO) dataset that provides information on both life expectancy and GDP per capita by year for different countries and regions, Explore and visualize the data using appropriate plots, and develop meaningful insights.

4.SYLLABUS

UNIT III PREDICTIVE ANALYTICS **6+6**

Data Mining Concepts – Data Mining Process – Data Mining Methods - Text Analytics and Text Mining – NLP – Applications – Process – Sentiment Analysis – Web Mining – Search Engines – Web Analytics – Social Analytics.

List of Exercise/Experiments:

1.Perform Customer Review Sentiment Analysis with text data extracted from customer reviews of a certain company and explore it using specialized statistical and linguistic tools to identify positive, negative, and neutral experiences and their strength and subjectivity.

2.Using Microsoft Stock Data/Amazon Stock Data or INTEL Stock Data, Explore the company's historical stock performance and find insights about the future.

UNIT IV PRESCRIPTIVE ANALYTICS **6+6**

Model-based Decision Making – Structure of Mathematical Models for Decision Support – Certainty, Uncertainty and Risk – Decision Modelling – Multiple Goals, Sensitivity Analysis, WhatIf Analysis and Goal Seeking – Decision Analysis – Introduction to Simulation – Location-based Analytics for Organizations – Impacts of Analytics in Organization. Case study: prepare a detailed report on applications of analytics in different industries.

List of Exercise/Experiments:

1.Perform Retail Price Optimization using dataset of price data for a retail company containing information such as product names, historical prices, product categories and characteristics, volume of sales, and time and geographic notations. Calculate the optimal selling prices for the products to create efficient, data-driven recommendations for the company.

2.Perform Credit Card Fraud Detection using online transactions dataset and analyze it for suspicious operations using statistical methods.

5. COURSE OUTCOMES

At the end of this course, the students will be able to:

COURSE OUTCOMES		K Level
CO1	Understand the business intelligence (BI) methodology and concepts.	K2
CO2	Learn about descriptive, inferential statistics and data warehousing operations.	K2
CO3	Analyze wide range of applications of data mining.	K3
CO4	Analyze the various prescriptive analytics methods.	K3
CO5	Develop and deploy Business Analytic Models.	K4

6.CO – PO /PSO Mapping Matrix

CO	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PS O1	PS O2	PS 03
1	3	2	1		1			1	1	1		1	2	3	3
2	3	2	1		3			1	1	1		1	2	3	3
3	3	2	1		3			3	3	3		3	2	3	3
4	3	3	2		3			3	3	3		3	2	3	3
5	3	2	2		3			3	3	3		3	2	3	3

R.M.K GROUP OF INSTITUTIONS

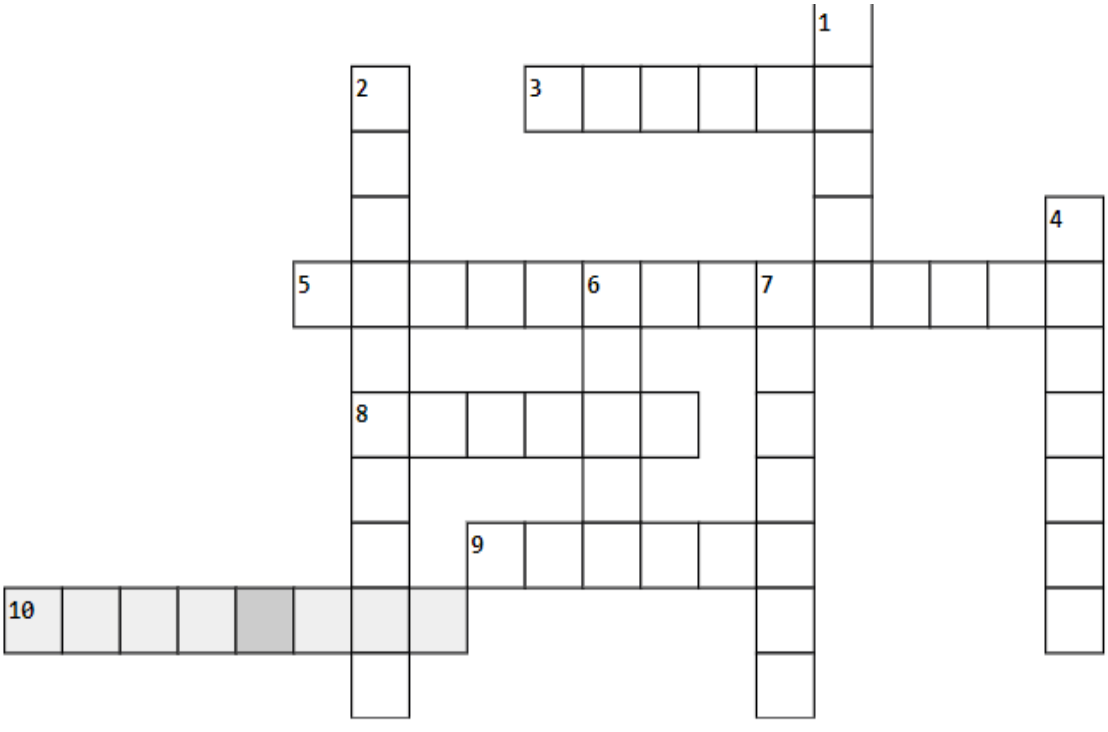
7. Lecture Plan – Unit 1

Sl. No.	Topic	Number of Periods	Proposed Date	Actual Lecture Date	CO	Taxonomy Level	Mode of Delivery
1	Evolution of Computerized Decision Support to Analytics	1	03.01.2024	03.01.2024	CO1	K3	PPT/Chalk & Talk
2	A Framework for Business Intelligence - Analytics Overview -	1	05.01.2024	05.01.2024	CO1	K3	PPT/Chalk & Talk
3	Analytics Examples-	1	06.01.2024	06.01.2024	CO1	K3	PPT/Chalk & Talk
4	Analytics Examples-	1	06.01.2024	06.01.2024	CO1	K3	PPT/Chalk & Talk
5	Introduction to Big Data Analytics	1	10.01.2024	10.01.2024	CO1	K3	PPT/Chalk & Talk
6	Overview of the Analytics Ecosystem	1	12.01.2024	12.01.2024	CO1	K3	PPT/Chalk & Talk

8. ACTIVITY BASED LEARNING

Crossword Puzzle

https://crosswordlabs.com/view/data-analysis-tools-2



Across

3. Used to streamline, model, visualize, and analyze data using its built-in data analytics tools.

5. A widely used spreadsheet software that offers basic data analysis and visualization capabilities.

8. A high-level programming language and environment primarily used for numerical and scientific computing, including data analysis and visualization.

9. An open-source data visualization and analysis tool that provides a visual programming interface for building data workflows.

10. Data visualization and business intelligence tool that offer associative data modeling and interactive analytics.

Down

1. A free, open source data analytics platform that supports data integration, processing, visualization, and reporting.

2. An open-source data science platform that provides tools for data preprocessing, modeling, and predictive analytics.

4. A data analytics and data blending platform that simplifies data preparation, blending, and advanced analytics for business users.

6. A software application for statistics and data analysis, commonly used in the social sciences and economics.

7. A powerful data visualization and business intelligence tool that allows users to create interactive and shareable dashboards and reports.

UNIT II DESCRIPTIVE ANALYTICS

The Nature of Data:

Data is the main ingredient for any BI, data science, and business analytics initiative.

- It can be viewed as the raw material.
- Once perceived as a big challenge to collect, store, and manage.
- Data is widely considered among the most valuable assets of an organization, with the potential to create invaluable insight to better understand customers, competitors, and business processes.

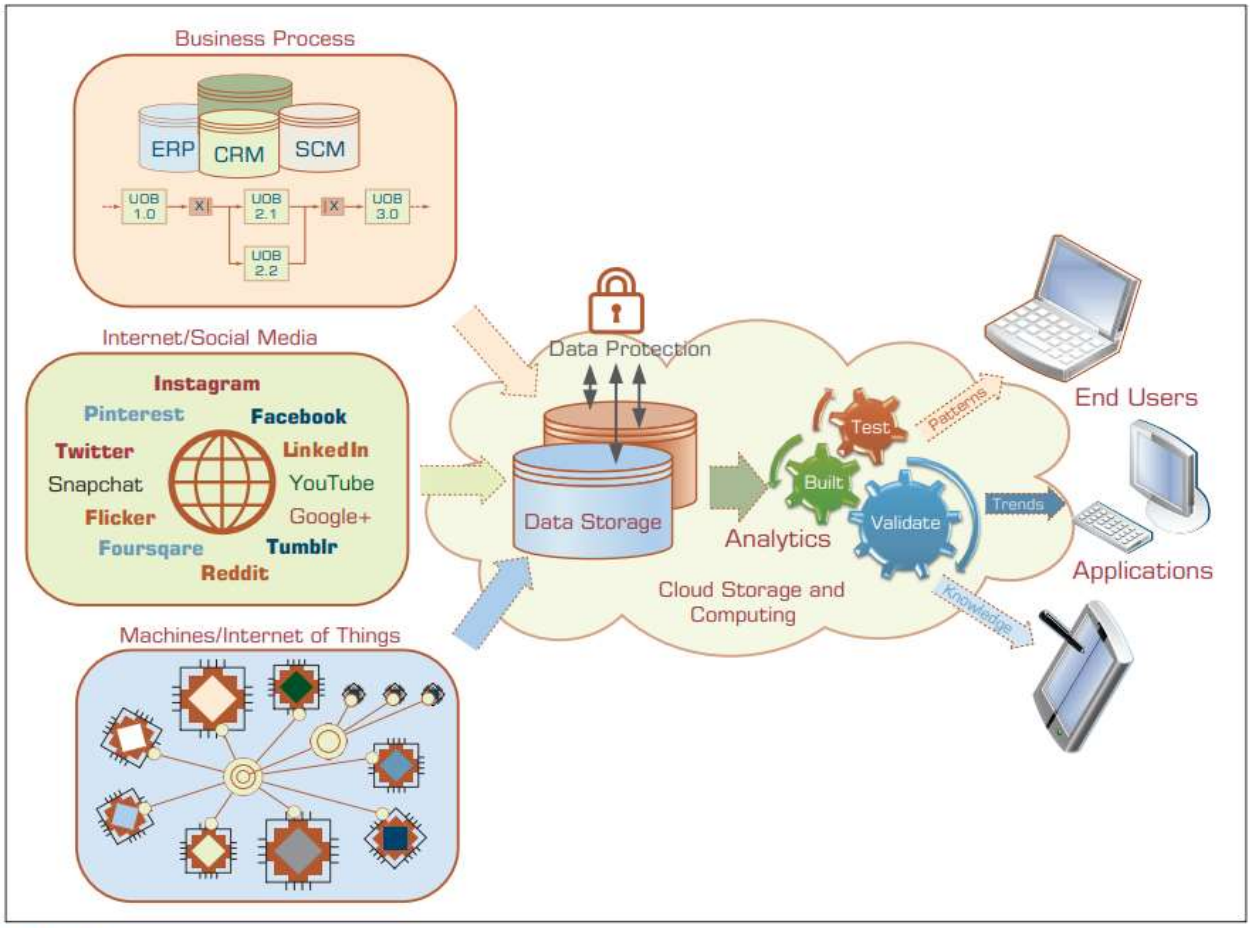


FIGURE 2.1 A Data to Knowledge Continuum.

UNIT II DESCRIPTIVE ANALYTICS

Following are some of the characteristics (metrics) that define the readiness level of data for an analytics study.

Data source reliability: refers to the originality and appropriateness of the storage medium where the data is obtained—answering the question of "Do we have the right confidence and belief in this data source?".

One should always look for the original source/creator of the data to eliminate/mitigate the possibilities of data misrepresentation and data transformation caused by the mishandling of the data.

Data content accuracy: means that data are correct and are a good match for the analytics problem—answering the question of "Do we have the right data for the job?"

The data should represent what was intended or defined by the original source of the data. For example, the customer's contact information recorded in a record within a database should be the same as what the patient said it was.

Data accessibility: means that the data are easily and readily obtainable—answering the question of "Can we easily get to the data when we need to?" Access to data may be tricky, especially if the data is stored in more than one location and storage medium and need to be merged/transformed while accessing and obtaining it.

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.

Increasing popularity in educational degrees and certificate programs for Information Assurance is an evidence to the criticality and the increasing urgency of this data quality metric. Any organization that maintains health records for individual patients must have systems in place that not only safeguard the data from unauthorized access.

Data richness means that all the required data elements are included in the data set.

- In essence, richness (or comprehensiveness) means that the available variables portray a rich enough dimensionality of the underlying subject matter for an accurate and worthy analytics study.
- It also means that the information content is complete (or near complete) to build a predictive and/or prescriptive analytics model.

Data consistency means that the data are accurately collected and combined/ merged.

- Consistent data represent the dimensional information (variables of interest) coming from potentially disparate sources but pertaining to the same subject.
- If the data integration/merging is not done properly, some of the variables of different subjects may find themselves in the same record—having two different patient records mixed up—for instance, it may happen while merging the demographic and clinical test result data records.

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.

- It also means that the data is recorded at or near the time of the event or observation so that the time-delay-related misrepresentation (incorrectly remembering and encoding) of the data is prevented.

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.

•If the data is aggregated, it may not contain the level of detail needed for an analytics algorithm to learn how to discern different records/cases from one another.

•For example, in a medical setting, numerical values for laboratory results should be recorded to the appropriate decimal place as required for the meaningful interpretation of test results and proper use of those values within an analytics algorithm.

Data validity is the term used to describe a match/mismatch between the actual and expected data values of a given variable. As part of the data definition, the acceptable values or value ranges for each data element must be defined. For example, a valid data definition related to gender would include three values: male, female, and unknown.

Data relevancy means that the variables in the data set are all relevant to the study being conducted. Relevancy is not a dichotomous measure (whether a variable is relevant or not); rather, it has a spectrum of relevancy from least relevant to most relevant.

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. It is often dirty, misaligned, overly complex, and inaccurate. A tedious and time-demanding process (so-called data preprocessing) is necessary to convert the raw real-world data into a well-refined form for analytics algorithms.

In the **first phase** of data preprocessing, the relevant data is **collected** from the identified sources, the necessary records and variables are selected (based on an intimate understanding of the data, the unnecessary information is filtered out), and the records coming from multiple data sources are integrated/merged.

In the **second phase** of data preprocessing, the data is **cleaned** (this step is also known as data scrubbing). Data in its original/raw/real-world form is usually dirty In this step, 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 (filled with a most probable value) or ignored; in other cases, the missing values are a natural part of the data set (e.g., the household income field is often left unanswered by people who are in the top income tier). In this step, the analyst should also identify noisy values in the data (i.e., the outliers) and smooth them out.

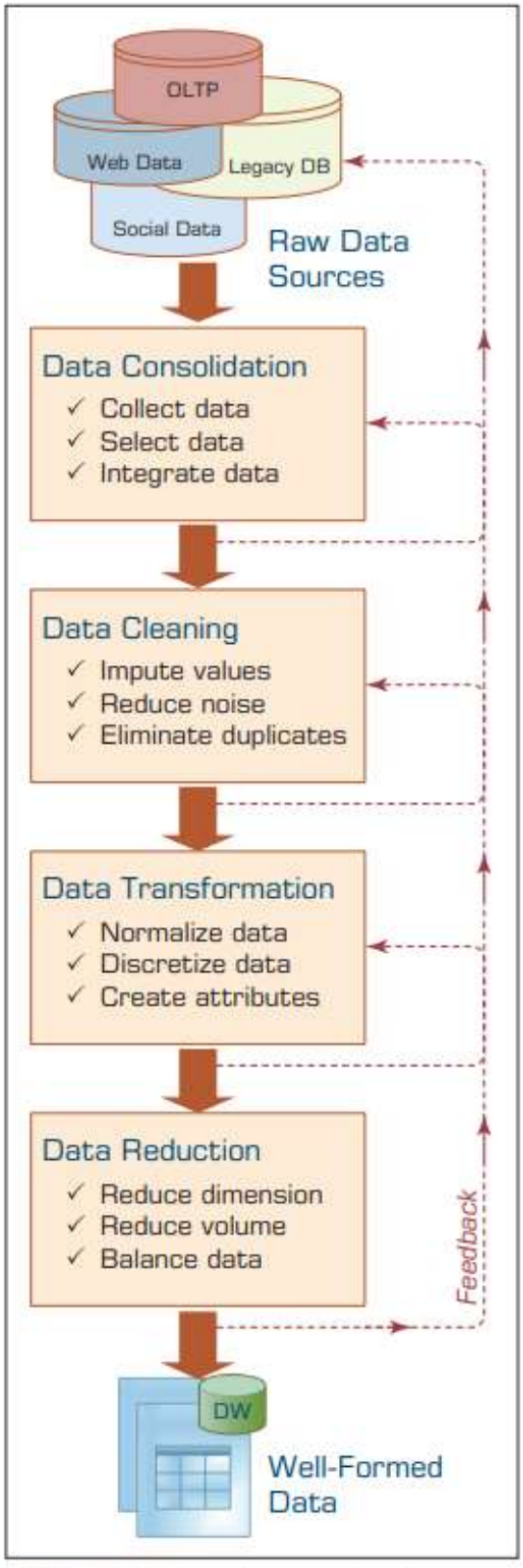


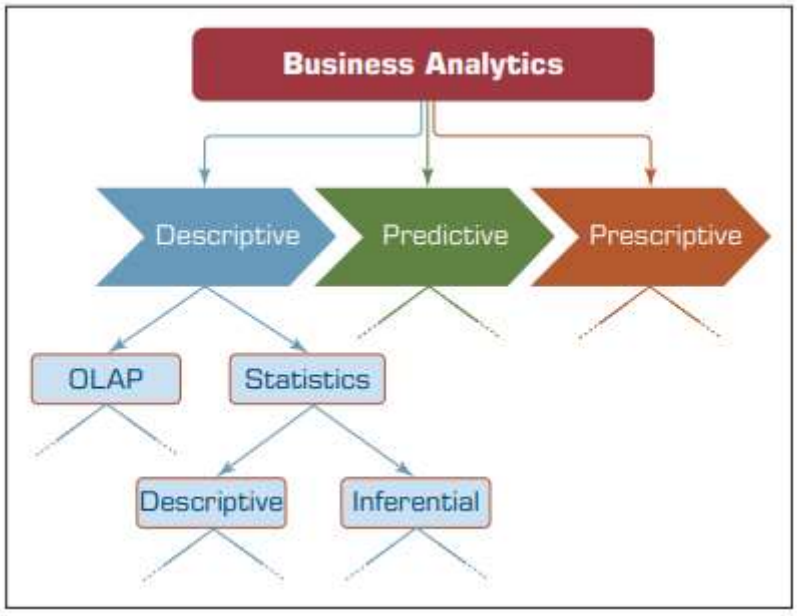
FIGURE 2.3 Data Preprocessing Steps.

In the **third phase** of data preprocessing, the data is **transformed** for better processing. For instance, in many cases the data is normalized between a certain minimum and maximum for all variables to mitigate the potential bias of one variable (having large numeric values, such as for household income) dominating other variables (such as number of dependents or years in service, which may potentially be more important) having smaller values.

The final phase of data preprocessing is data reduction. Even though data scientists (i.e., analytics professionals) like to have large data sets, too much data may also be a problem. In the simplest sense, one can visualize the data commonly used in predictive analytics projects as a flat file consisting of two dimensions: variables (the number of columns) and cases/records (the number of rows).

TABLE 2.1 A Summary of Data Preprocessing Tasks and Potential Methods

Main Task	Subtasks	Popular Methods
Data consolidation	Access and collect the data	SQL queries, software agents, Web services.
	Select and filter the data	Domain expertise, SQL queries, statistical tests.
	Integrate and unify the data	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	Normalize 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 normalization or scaling techniques.
	Discretize 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.



Statistical Modeling for Business Analytics

Statistics (statistical methods and underlying techniques) is usually considered as part of descriptive analytics in the above Figure. Some of the statistical methods can also be considered as part of predictive analytics such as discriminant analysis, multiple regression, logistic regression, and k-means clustering.

Descriptive analytics has two main branches: statistics and online analytics processing (OLAP). OLAP is the term used for analyzing, characterizing, and summarizing structured data stored in organizational databases (often stored in a data warehouse or in a data mart). The OLAP branch of descriptive analytics has also been called Business Intelligence. Statistics, on the other hand, helps to characterize the data either one variable at a time or multivariables all together, using either descriptive or inferential methods.

Statistics—a collection of mathematical techniques to characterize and interpret data—has been around for a very long time. Many methods and techniques have been developed to address the needs of the end users and the unique characteristics of the data being analyzed.

The main difference between descriptive and inferential statistics is the data used in these methods—whereas descriptive statistics is all about describing the sample data on hand, and inferential statistics is about drawing inferences or conclusions about the characteristics of the population.

Descriptive Statistics for Descriptive Analytics

Descriptive statistics, as the name implies, describes the basic characteristics of the data at hand, often one variable at a time. Using formulas and numerical aggregations, descriptive statistics summarizes the data in such a way that often meaningful and easily understandable patterns emerge from the study. Although it is very useful in data analytics and very popular among the statistical methods, descriptive statistics does not allow making conclusions (or inferences) beyond the sample of the data being analyzed.

In business analytics, descriptive statistics plays a critical role—it allows us to understand and explain/present our data in a meaningful manner using aggregated numbers, data tables, or charts/graphs. In essence, descriptive statistics helps us convert our numbers and symbols into meaningful representations for anyone to understand and use. Such an understanding not only helps business users in their decision-making processes, but also helps analytics professionals and data scientists to characterize and validate the data for other more sophisticated analytics tasks. Descriptive statistics allows analysts to identify data concertation, unusually large or small values (i.e., outliers), and unexpectedly distributed data values for numeric variables.

Measures of Centrality Tendency (May Also Be Called Measures of Location or Centrality)

Measures of centrality are the mathematical methods by which we estimate or describe central positioning of a given variable of interest. 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. It is by far the most popular and most commonly used measure of central tendency. It is used with continuous or discrete numeric data. For a given variable x, if we happen to have n values/observations (x1 , x2 , . . . , xn) , we can write the arithmetic mean of the data sample (x, pronounced as x-bar) as follows

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Median The median is the measure of center value in a given data set. It is the number in the middle of a given set of data that has been arranged/sorted in order of magnitude (either ascending or descending).

Mode The mode is the observation that occurs most frequently (the most frequent value in our data set). On a histogram it represents the highest bar in a bar chart, and hence, it may be considered as being the most popular option/value. The mode is most useful for data sets that contain a relatively small number of unique values.

Measures of Dispersion (May Also Be Called Measures of Spread or Decentrality)

Measures of dispersion are the mathematical methods used to estimate or describe the degree of variation in a given variable of interest. They are a representation of the numerical spread (compactness or lack thereof) of a given data set. To describe this dispersion, a number of statistical measures are developed;

Range The range is perhaps the simplest measure of dispersion. It is the difference between the largest and the smallest values in a given data set (i.e., variables). So we calculate range by simply identifying the smallest value in the data set (minimum), identifying the largest value in the data set (maximum), and calculating the difference between them (range = maximum – minimum).

Variance A more comprehensive and sophisticated measure of dispersion is the variance. It is a method used to calculate the deviation of all data points in a given data set from the mean. The larger the variance, the more the data are spread out from the mean and the more variability one can observe in the data sample.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

Standard Deviation The standard deviation is also a measure of the spread of values within a set of data. The standard deviation is calculated by simply taking the square root of the variations. The following formula shows the calculation of standard deviation from a

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}}$$

Mean Absolute Deviation In addition to variance and standard deviation, sometimes we also use mean absolute deviation to measure dispersion in a data set. It is a simpler way to calculate the overall deviation from the mean. Specifically, it is calculated by measuring the absolute values of the differences between each data point and the mean and summing them. It provides a measure of spread without being specific about the data point being lower or higher than the mean.

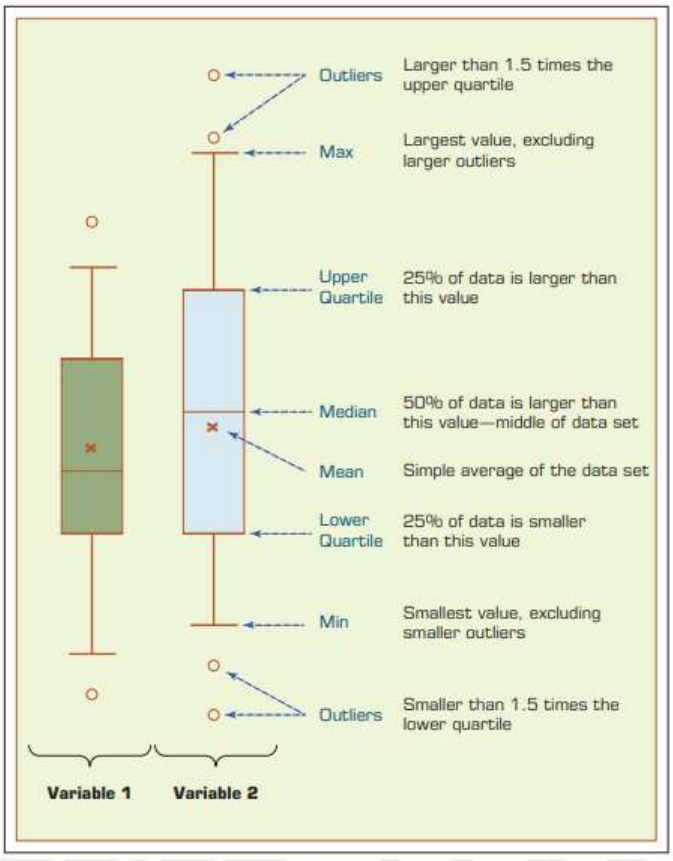
$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Quartiles and Interquartile Range Quartiles help us identify spread within a subset of the data. 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. Quartiles are a useful measure of dispersion because they are much less affected by outliers or a skewness in the data set than the equivalent measures in the whole data set.

A common way of expressing quartiles is as an interquartile range, which 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. The quartile-driven descriptive measures (both centrality and dispersion) are best explained with a popular plot called a box plot (or box-and-whiskers plot).

Box-and-Whiskers Plot

The box-and-whiskers plot (or simply a box plot) is a graphical illustration of several descriptive statistics about a given data set. They can be either horizontal or vertical, but vertical is the most common representation, especially in modern-day analytics software products. It is known to be first created and presented by John W. Tukey in 1969. Box plot is often used to illustrate both centrality and dispersion of a given data set (i.e., the distribution of the sample data) in an easy-to-understand graphical notation. Figure 2.8 shows a couple of box plots side by side, sharing the same y-axis.



Historically speaking, the box plot was not used widely and often enough (especially in areas outside of statistics), with the emerging popularity of business analytics, it is gaining fame in less-technical areas of the business world.

Its information richness and ease of understanding are largely to credit for its recent popularity. **The box plot shows the centrality (median and sometimes also mean) as well as the dispersion** (the density of the data within the middle half—drawn as a box between the first and third quartile), the minimum and maximum ranges (shown as extended lines from the box, looking like whiskers, 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.

The Shape of a Distribution

Although not as common as the centrality and dispersion, the shape of the data distribution is also a useful measure for the descriptive statistics. Before delving into the shape of the distribution we first need to define the distribution itself.

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. Because normal distribution is a perfectly symmetric unimodal distribution, it does not have skewness, that is, its skewness measure (i.e., the value of the coefficient of skewness) is equal to zero. The skewness measure/value can be either positive or negative.

$$Skewness = S = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)s^3}$$

where s is the standard deviation and n is the number of samples.

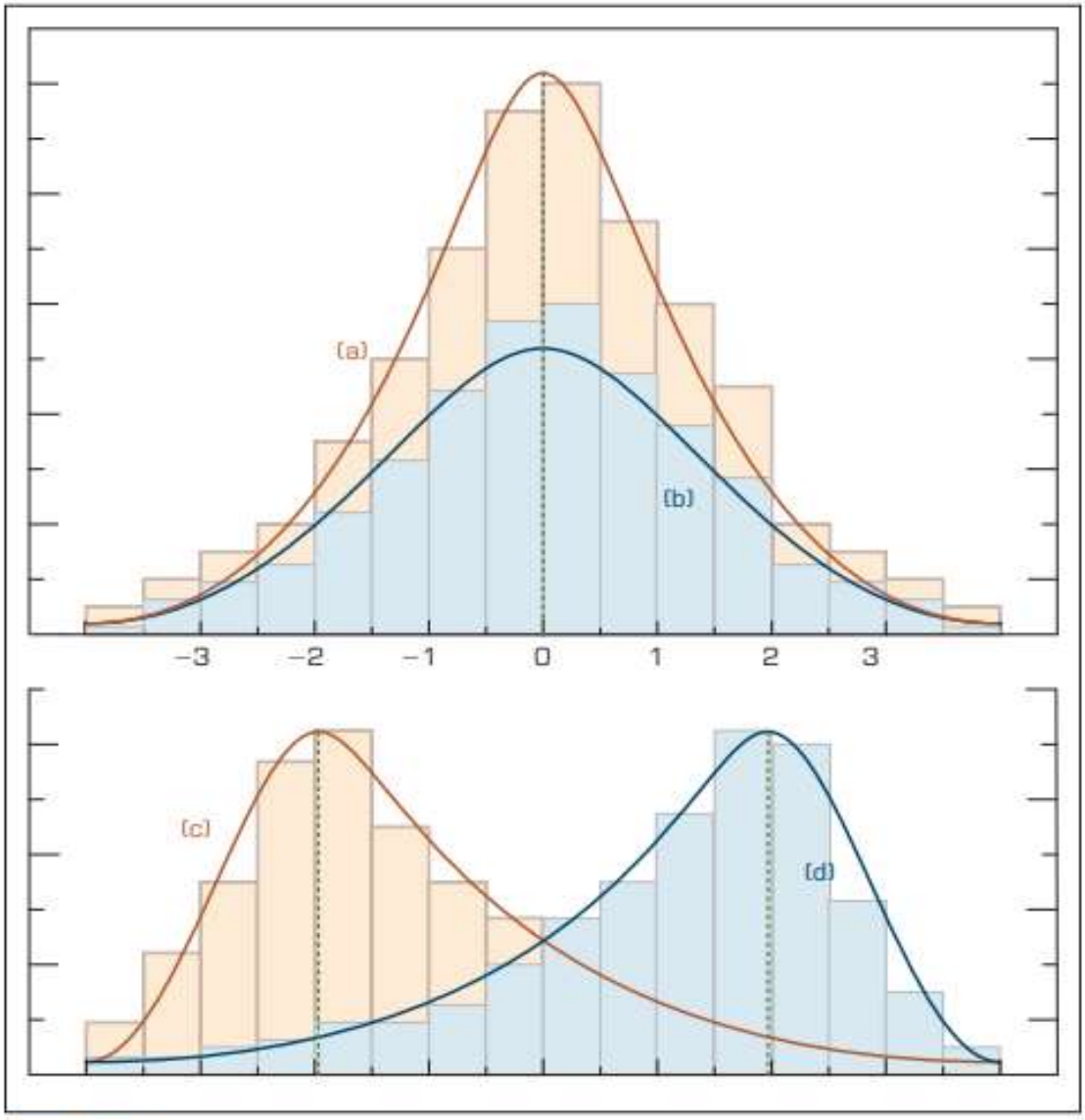


FIGURE 2.9 Relationship between Dispersion and Shape Properties.

Kurtosis is another measure to use in characterizing the shape of a unimodal distribution. As opposed to the sway in shape, kurtosis is more interested in characterizing the peak/tall/skinny nature of the distribution. Specifically, 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.

$$Kurtosis = K = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} - 3$$

Descriptive statistics (as well as inferential statistics) can easily be calculated using commercially viable statistical software packages (e.g., SAS, SPSS, Minitab, JMP, Statistica) or free/open source tools (e.g., R). Perhaps the most convenient way to calculate descriptive and some of the inferential statistics is to use Excel.

Regression Modeling for Inferential Statistics

Regression, especially linear regression, is perhaps the most widely known and used analytics technique in statistics. As popular as it is, essentially, 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. As popular as it is, essentially, 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. Regression aims to capture the functional relationship between and among the characteristics of the real world and describe this relationship with a mathematical model, which may then be used to discover and understand the complexities of reality—explore and explain relationships or forecast future occurrences.

Regression can be used for one of two purposes: hypothesis testing—investigating potential relationships between different variables, and prediction/forecasting—estimating values of a response variables based on one or more explanatory variables. These two uses are not mutually exclusive. In prediction, regression identifies additive mathematical relationships (in the form of an equation) between one or more explanatory variables and a response variable. Once determined, this equation can be used to forecast the values of the response variable for a given set of values of the explanatory variables.

CORRELATION VERSUS REGRESSION Because regression analysis originated from correlation studies, and because both methods attempt to describe the association between two (or more) variables, these two terms are often confused by professionals and even by scientists. 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. On the other hand, regression attempts to describe the dependence of a response variable on one (or more) explanatory variables where it implicitly assumes that there is a one-way causal effect from the explanatory variable(s) to the response variable, regardless of whether the path of effect is direct or indirect.

SIMPLE VERSUS MULTIPLE REGRESSION If the regression equation is built between one response variable and one explanatory variable, then it is called simple regression. For instance, the regression equation built to predict/explain the relationship between a height of a person (explanatory variable) and the weight of a person (response variable) is a good example of simple regression. Multiple regression is the extension of simple regression where the explanatory variables are more than one. For instance, in the previous example, if we were to include not only the height of the person but also other personal characteristics (e.g., BMI, gender, ethnicity) to predict the weight of a person, then we would be performing multiple regression analysis. In both cases, the relationship between the response variable and the explanatory variable(s) are linear and additive in nature. If the relationships are not linear, then we may want to use one of many other nonlinear regression methods to better capture the relationships between the input and output variables.

How Do We Develop the Linear Regression Model?

To understand the relationship between two variables, the simplest thing that one can do is to draw a scatter plot, where the y-axis represents the values of the response variable and the x-axis represents the values of the explanatory variable (see Figure 2.13). A scatter plot would show the changes in the response variable as a function of the changes in the explanatory variable. In the case shown in Figure 2.13, there seems to be a positive relationship between the two; as the explanatory variable values increase, so does the response variable.

Even though there are several methods/algorithms proposed to identify the regression line, the one that is most commonly used is called the **ordinary least squares (OLS)** method. The OLS method aims to minimize the sum of squared residuals (squared vertical distances between the observation and the regression point) and leads to a mathematical expression for the estimated value of the regression line (which are known as b parameters). For simple linear regression, relationship between the response variable (y) and the explanatory variable(s) (x) can be shown as a simple equation as follows: $y = \beta 0 + \beta 1x$

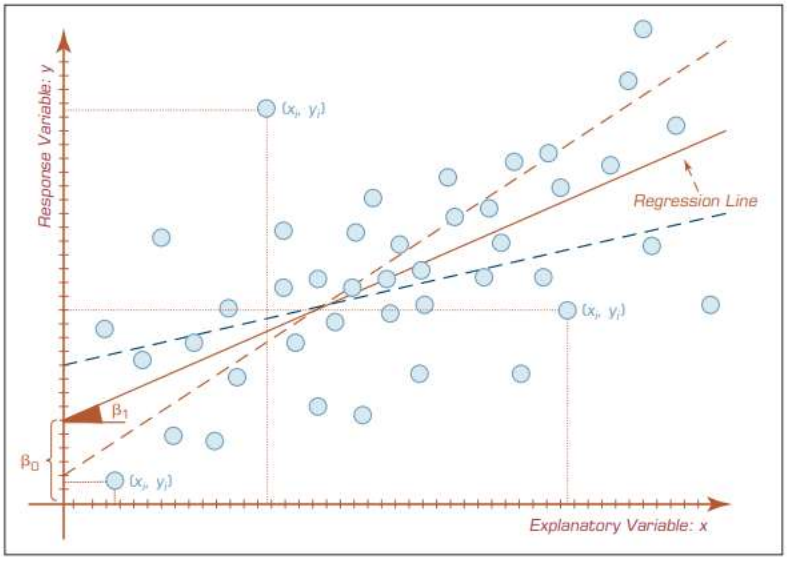


FIGURE 2.13 A Scatter Plot and a Linear Regression Line.

In this equation, b0 is called the intercept, and b1 is called the slope. Once OLS determines the values of these two coefficients, the simple equation can be used to forecast the values of y for given values of x. The sign and the value of b1 also reveal the direction and the strengths of relationship between the two variables. If the model is of a multiple linear regression type, then there would be more coefficients to be determined, one for each additional explanatory variable. As the following formula shows, the additional explanatory variable would be multiplied with the new bi coefficients and summed together to establish a linear additive representation of the response variable

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n$$

How Do We Know If the Model Is Good Enough? Because of a variety of reasons, sometimes models as representations of the reality do not prove to be good. Regardless of the number of explanatory variables included, there is always a possibility of not having a good model, and therefore the linear regression model needs to be assessed for its fit (the degree at which it represents the response variable).

For the numerical assessment, three statistical measures are often used in evaluating the fit of a regression model.

- R2 (R-squared),
- F-test
- root mean square error (RMSE).

Of the three, R2 has the most useful and understandable meaning because of its intuitive scale. The value of R2 ranges from zero to one (corresponding to the amount of variability explained in percentage) with zero indicating that the relationship and the prediction power of the proposed model is not good, and one indicating that the proposed model is a perfect fit that produces exact predictions (which is almost never the case). The good R2 values would usually come close to one, and the closeness is a matter of the phenomenon being modeled—whereas an R2 value of 0.3 for a linear regression model in social sciences can be considered good enough, an R2 value of 0.7 in engineering may be considered as not a good-enough fit.

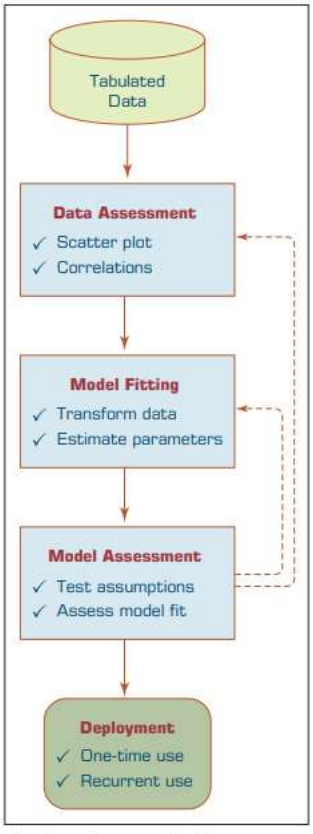


FIGURE 2.14 A Process Flow for Developing Regression Models.

What Are the Most Important Assumptions in Linear Regression?

Even though they are still the choice of many for data analyses (both for explanatory as well as for predictive modeling purposes), linear regression models suffer from several highly restrictive assumptions.

1. **Linearity.** This assumption states that the relationship between the response variable and the explanatory variables are linear. That is, the expected value of the response variable is a straight-line function of each explanatory variable, while holding all other explanatory variables fixed. Also, the slope of the line does not depend on the values of the other variables.
2. **Independence** (of errors). This assumption states that the errors of the response variable are uncorrelated with each other. This independence of the errors is weaker than actual statistical independence, which is a stronger condition and is often not needed for linear regression analysis.
3. **Normality** (of errors). This assumption states that the errors of the response variable are normally distributed. That is, they are supposed to be totally random and should not represent any nonrandom patterns.
4. **Constant variance** (of errors). This assumption, also called homoscedasticity, states that the response variables have the same variance in their error, regardless of the values of the explanatory variables. In practice this assumption is invalid if the response variable varies over a wide enough range/scale.
5. **Multicollinearity**. This assumption states that the explanatory variables are not correlated (i.e., do not replicate the same but provide a different perspective of the information needed for the model). Multicollinearity can be triggered by having two or more perfectly correlated explanatory variables presented to the model.

Logistic regression is a very popular, statistically sound, probability-based classification algorithm that employs supervised learning. It was developed in the 1940s as a complement to linear regression and linear discriminant analysis methods. It has been used extensively in numerous disciplines, including the medical and social sciences fields.

Logistic regression is similar to linear regression in that it also aims to regress to a mathematical function that explains the relationship between the response variable and the explanatory variables using a sample of past observations (training data). It differs from linear regression with one major point: its output (response variable) is a class as opposed to a numerical variable.

That is, whereas linear regression is used to estimate a continuous numerical variable, logistic regression is used to classify a categorical variable. Even though the original form of logistic regression was developed for a binary output variable (e.g., 1/0, yes/no, pass/fail, accept/reject), the present-day modified version is capable of predicting multiclass output variables (i.e., multinomial logistic regression). If there is only one predictor variable and one predicted variable, the method is called simple logistic regression (similar to calling linear regression models with only one independent variable as simple linear regression).

In predictive analytics, logistic regression models are used to develop probabilistic models between one or more explanatory/predictor variables (which may be a mix of both continuous and categorical in nature) and a class/response variable (which may be binomial/binary or multinomial/multiclass). Unlike ordinary linear regression, logistic regression is used for predicting categorical (often binary) outcomes of the response variable—treating the response variable as the outcome of a Bernoulli trial.

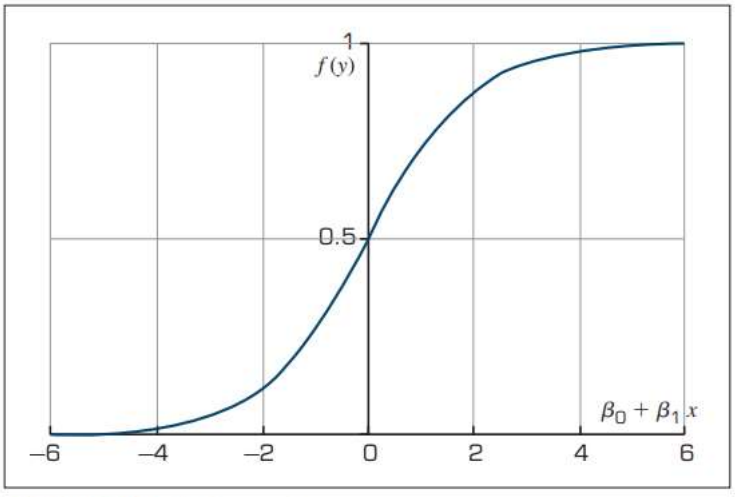


FIGURE 2.15 The Logistic Function.

The logistic function, f (y) in Figure 2.15, is the core of logistic regression, which can only take values between 0 and 1. The following equation is a simple mathematical representation of this function:

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

Business Reporting

Decision makers need information to make accurate and timely decisions. Information is essentially the contextualization of data. In addition to statistical means, information (descriptive analytics) can also be obtained using online analytics processing [OLTP] systems.

A report is any communication artifact prepared with the specific intention of conveying information in a digestible form to whoever needs it, whenever and wherever they may need it. It is usually a document that contains information (usually driven from data) organized in a narrative, graphic, and/or tabular form, prepared periodically (recurring) or on an as-needed (ad hoc) basis, referring to specific time periods, events, occurrences, or subjects. Business reports can fulfill many different (but often related) functions. Here are a few of the most prevailing ones:

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

Business reporting (also called OLAP or BI) is an essential part of the larger drive toward improved, evidence-based, optimal managerial decision making. The foundation of these business reports is various sources of data coming from both inside and outside the organization (online transaction processing [OLTP] systems). Creation of these reports involves ETL (extract, transform, and load) procedures in coordination with a data warehouse and then using one or more reporting tools (see Chapter 3 for a detailed description of these concepts).

Figure 2.18 shows the continuous cycle of data acquisition S information generation S decision making S 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.

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

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

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

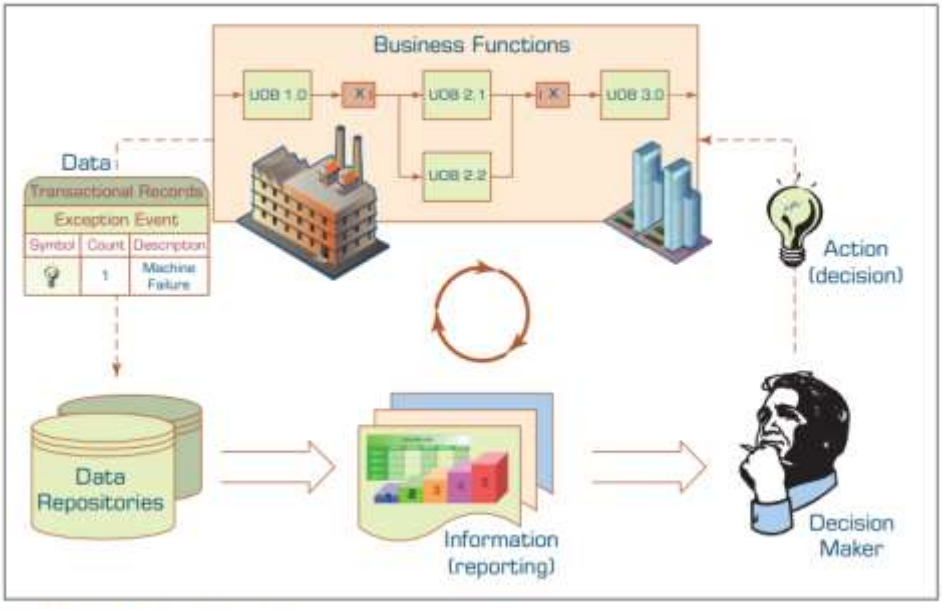


FIGURE 2.18 The Role of Information Reporting in Managerial Decision Making.

Data Visualization

Data visualization (or more appropriately, information visualization) has been defined as "the use of visual representations to explore, make sense of, and communicate data". It is also called information visualization, because information is the aggregation, summarization, and contextualization of data (raw facts), what is portrayed in visualizations is the information and not the data.

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

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

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

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

BAR CHART Bar charts are among the most basic visuals used for data representation. Bar charts are effective when you have nominal data or numerical data that splits nicely into different categories so you can quickly see comparative results and trends within your data. Bar charts are often used to compare data across multiple categories, such as percent of advertising spending by departments or by product categories

PIE CHART Pie charts are visually appealing, as the name implies, pie-looking charts. Because they are so visually attractive, they are often incorrectly used. Pie charts should only be used to illustrate relative proportions of a specific measure. For instance, they can be used to show the relative percentage of an advertising budget spent on different product lines, or they can show relative proportions of majors declared by college students in their sophomore year.

SCATTER PLOT Scatter plots are often used to explore the relationship between two or three variables (in 2-D or 2-D visuals). Because they are visual exploration tools, having more than three variables, translating them into more than three dimensions is not easily achievable. Scatter plots are an effective way to explore the existence of trends, concentrations, and outliers.

Bubble charts are often enhanced versions of scatter plots. Bubble charts, though, are not a new visualization type; instead, they should be viewed as a technique to enrich data illustrated in scatter plots (or even geographic maps). By varying the size and/or color of the circles, one can add additional data dimensions, offering more enriched meaning about the data.

Specialized Charts and Graphs

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

HISTOGRAM Graphically speaking, a histogram looks just like a bar chart. The difference between histograms and generic bar charts is the information that is portrayed. Histograms are used to show the frequency distribution of a variable or several variables. In a histogram, the x-axis is often used to show the categories or ranges, and the y-axis is used to show the measures/values/frequencies. Histograms show the distributional shape of the data. That way, one can visually examine if the data is normally or exponentially distributed.

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

PERT CHART PERT charts (also called network diagrams) are developed primarily to simplify the planning and scheduling of large and complex projects. They show precedence relationships among the project activities/tasks.

GEOGRAPHIC MAP When the data set includes any kind of location data (e.g., physical addresses, postal codes, state names or abbreviations, country names, latitude/longitude, or some type of custom geographic encoding), it is better and more informative to see the data on a map. Maps usually are used in conjunction with other charts and graphs, as opposed to by themselves. For instance, one can use maps to show distribution of customer service requests by product type (depicted in pie charts) by geographic locations.

BULLET Bullet graphs are often used to show progress toward a goal. A bullet graph is essentially a variation of a bar chart. Often they are used in place of gauges, meters, and thermometers in a dashboard to more intuitively convey the meaning within a much smaller space. Bullet graphs compare a primary measure (e.g., year-to-date revenue) to one or more other measures.

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

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

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

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

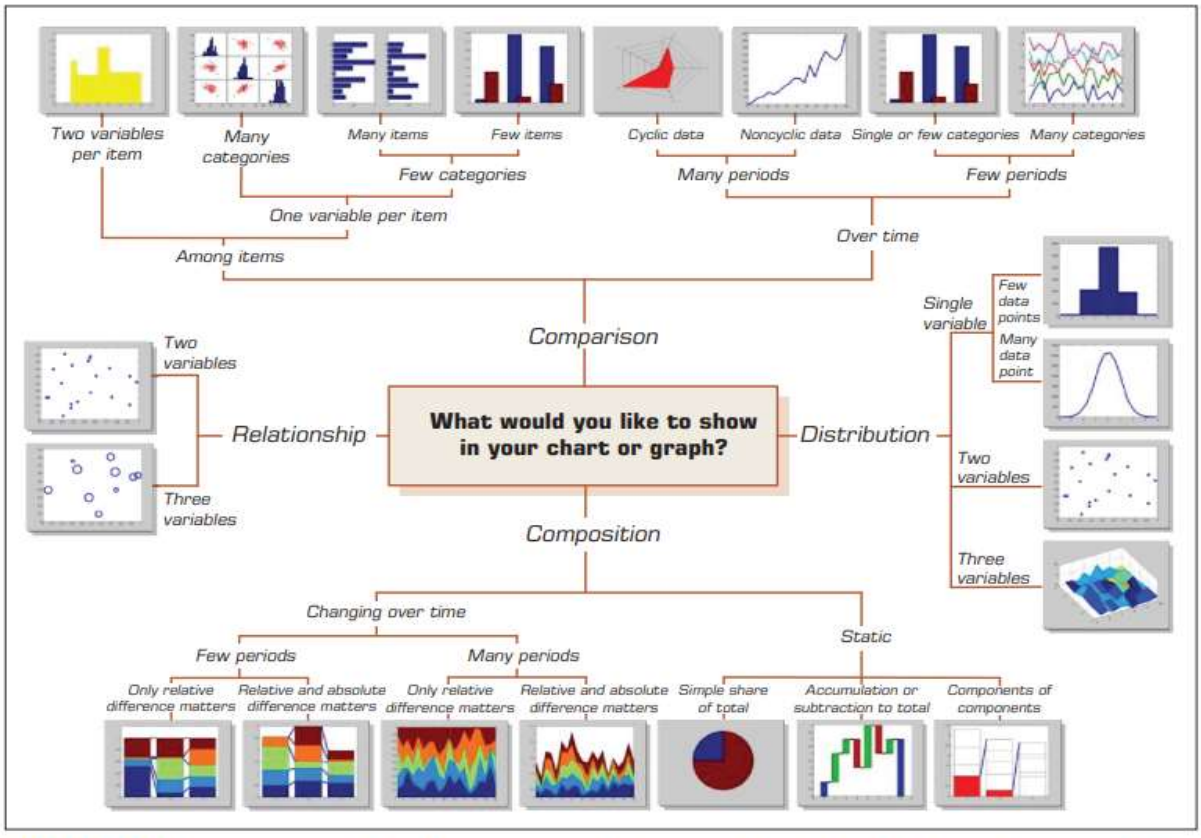


FIGURE 2.21 A Taxonomy of Charts and Graphs. *Source:* Adapted from Abela, A. (2008). Advanced presentations by design: Creating communication that drives action. New York: Wiley.

The Emergence of Visual Analytics As Seth Grimes has noted, there is a "growing palate" of data visualization techniques and tools that enable the users of business analytics and BI systems to better "communicate relationships, add historical context, uncover hidden correlations, and tell persuasive stories that clarify and call to action." The latest Magic Quadrant on Business Intelligence and Analytics Platforms released by Gartner in February 2016 further emphasizes the importance of data visualization in BI and analytics.

In BI and analytics, the key challenges for visualization have revolved around the intuitive representation of large, complex data sets with multiple dimensions and measures. For the most part, the typical charts, graphs, and other visual elements used in these applications usually involve two dimensions, sometimes three, and fairly small subsets of data sets. In contrast, the data in these systems reside in a data warehouse.

Visual Analytics Visual analytics is a recently coined term that is often used loosely to mean nothing more than information visualization. What is meant by visual analytics is the combination of visualization and predictive analytics. Whereas information visualization is aimed at answering, "What happened?" and "What is happening?" and is closely associated with BI (routine reports, scorecards, and dashboards), visual analytics is aimed at answering, "Why is it happening?" "What is more likely to happen?" and is usually associated with business analytics (forecasting, segmentation, correlation analysis).

High-Powered Visual Analytics Environments Due to the increasing demand for visual analytics coupled with fast-growing data volumes, there is an exponential movement toward investing in highly efficient visualization systems. Their new product, SAS Visual Analytics, is a very high-performance computing, in-memory solution for exploring massive amounts of data in a very short time (almost instantaneously). It empowers users to spot patterns, identify opportunities for further analysis, and convey visual results via Web reports or a mobile platform such as tablets and smartphones

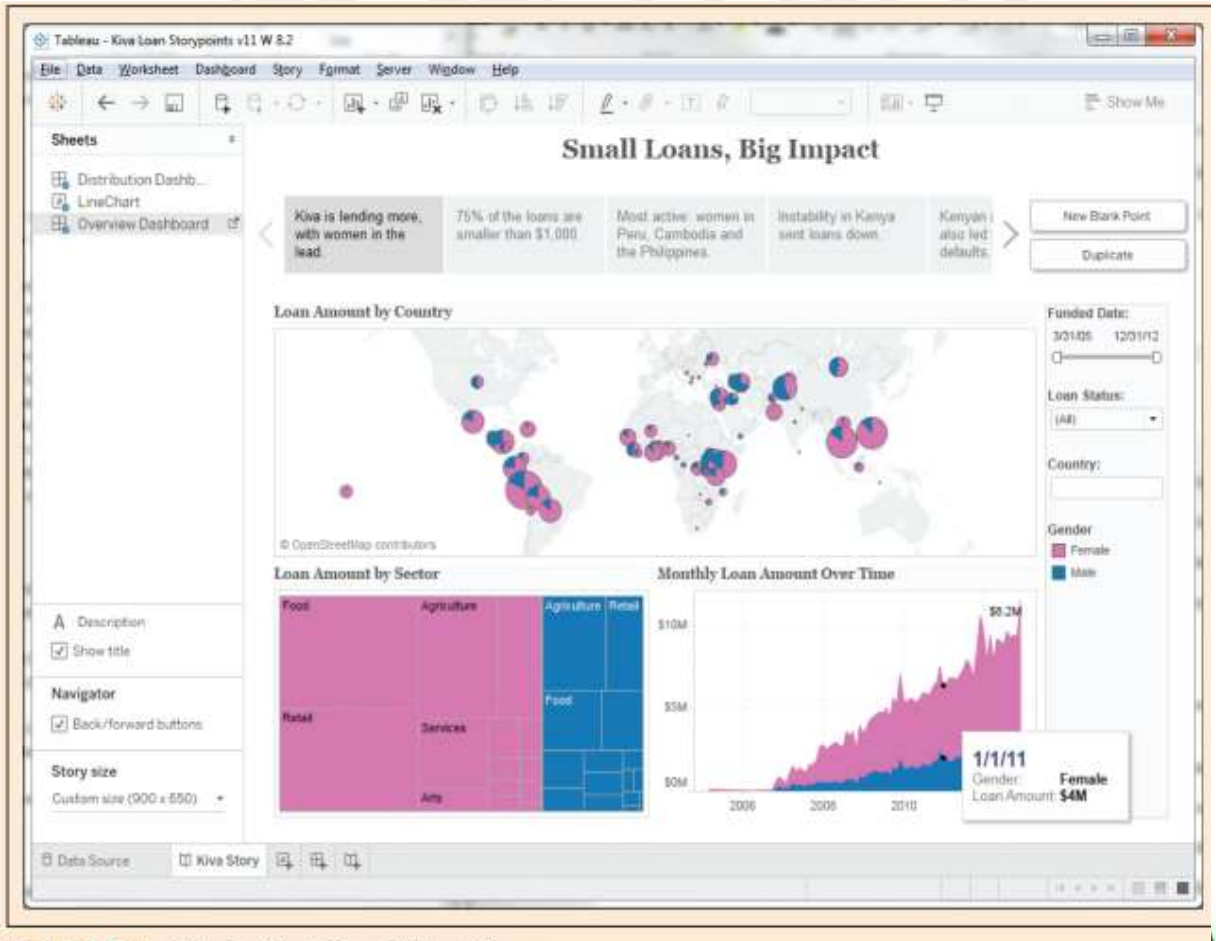


FIGURE 2.24 A Storyline Visualization in Tableau Software.

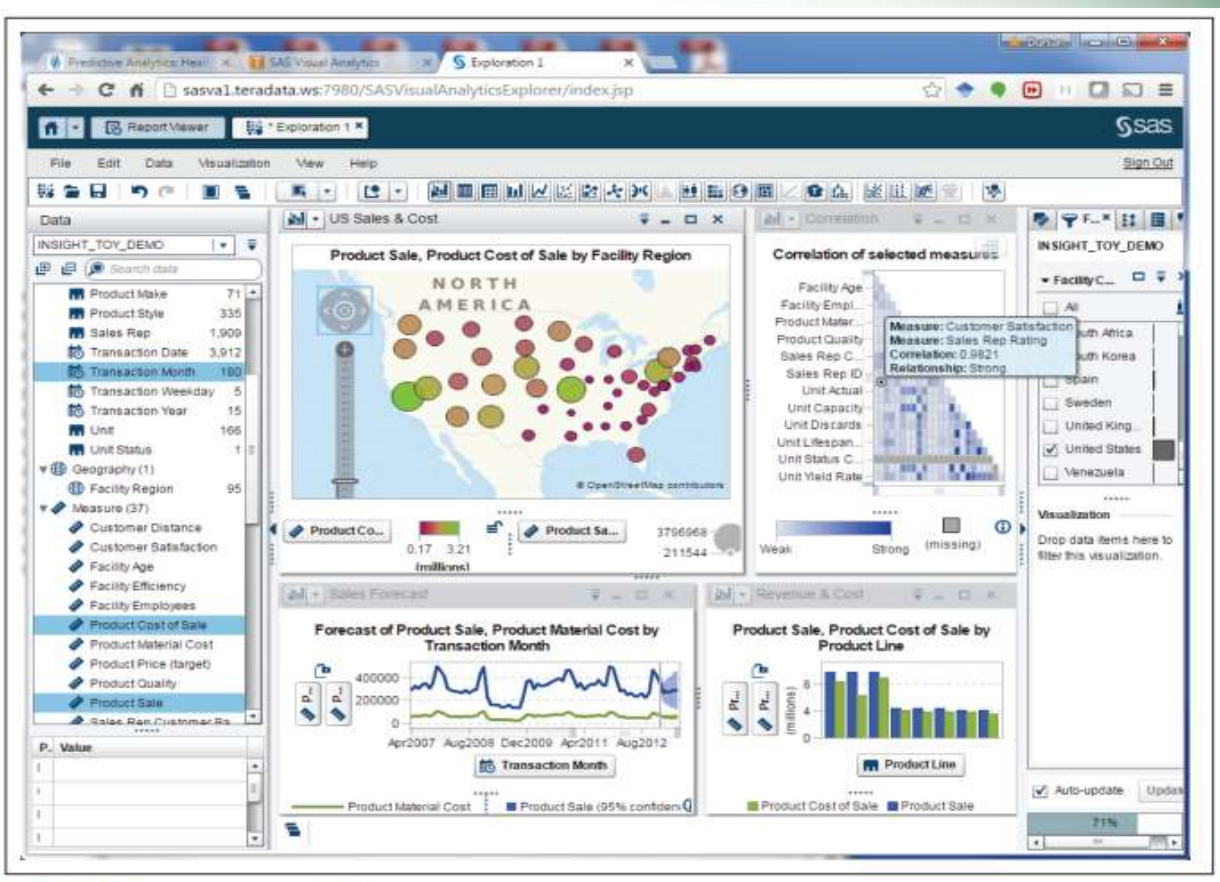


FIGURE 2.26 A Screenshot from SAS Visual Analytics. *Source:* SAS.com.

Information Dashboards

Information dashboards are common components of most, if not all, BI or business analytics platforms, business performance management systems, and performance measurement software suites. Dashboards provide visual displays of important information that is consolidated and arranged on a single screen so that information can be digested at a single glance and easily drilled in and further explored. A typical dashboard is shown in Figure 2.27.

This particular executive dashboard displays a variety of KPIs for a hypothetical software company called Sonatica (selling audio tools). This executive dashboard shows a high-level view of the different functional groups surrounding the products, starting from a general overview to the marketing efforts, sales, finance, and support departments.

FIGURE 2.27 A Sample Executive Dashboard. *Source:* dundas.com.

All of this is intended to give executive decision makers a quick and accurate idea of what is going on within the organization. On the left side of the dashboard, we can see (in a time series fashion) the quarterly changes in revenues, expenses, and margins, as well as the comparison of those figures to previous years' monthly numbers. On the upper-right side we see two dials with color-coded regions showing the amount of monthly expenses for support services (dial on the left) and the amount of other expenses (dial on the right).

As the color coding indicates, although the monthly support expenses are well within the normal ranges, the other expenses are in the red region, indicating excessive values. The geographic map on the bottom right shows the distribution of sales at the country level throughout the world. Behind these graphical icons there are variety of mathematical functions aggregating numerous data points to their highest level of meaningful figures. By clicking on these graphical icons, the consumer of this information can drill down to more granular levels of information and data.

Dashboard Design

Dashboards are not a new concept. Their roots can be traced at least to the executive information system of the 1980s. Today, dashboards are ubiquitous

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

1. **Monitoring**: Graphical, abstracted data to monitor key performance metrics.
2. **Analysis:** Summarized dimensional data to analyze the root cause of problems.
3. **Management:** Detailed operational data that identify what actions to take to resolve a problem. Because of these layers, dashboards pack a lot of information into a single screen.

"The fundamental challenge of dashboard design is to display all the required information on a single screen, clearly and without distraction, in a manner that can be assimilated quickly." To speed assimilation of the numbers, the numbers need to be placed in context. This can be done by comparing the numbers of interest to other baseline or target numbers, by indicating whether the numbers are good or bad, by denoting whether a trend is better or worse, and by using specialized display widgets or components to set the comparative and evaluative context.

What to Look for in a Dashboard?

They use visual components (e.g., charts, performance bars, sparklines, gauges, meters, stoplights) to highlight, at a glance, the data and exceptions that require action.

• They are transparent to the user, meaning that they require minimal training and are extremely easy to use.

• They combine data from a variety of systems into a single, summarized, unified view of the business.

• They enable drill-down or drill-through to underlying data sources or reports, providing more detail about the underlying comparative and evaluative context.

• They present a dynamic, real-world view with timely data refreshes, enabling the end user to stay up to date with any recent changes in the business.

• They require little, if any, customized coding to implement, deploy, and maintain.

Best Practices in Dashboard Design The real estate saying "location, location, location" makes it obvious that the most important attribute for a piece of real estate property is where it is located. For dashboards, it is "data, data, data." Even if a dashboard's appearance looks professional, is aesthetically pleasing, and includes graphs and tables created according to accepted visual design standards, it is also important to ask about the data: Is it reliable? Is it timely? Is any data missing? Is it consistent across all dashboards? Here are some of the experience-driven best practices in dashboard design.

1. Benchmark Key Performance Indicators with Industry Standards

Often when a report or a visual dashboard/scorecard is presented to business users, questions remain unanswered. The following are some examples:

- Where did you source this data from?
- While loading the data warehouse, what percentage of the data got rejected/encountered data quality problems?
- Is the dashboard presenting "fresh" information or "stale" information?
- When was the data warehouse last refreshed?
- When is it going to be refreshed next**?**

2. Validate the Dashboard Design by a Usability Specialist In most dashboard environments, the dashboard is designed by a tool specialist without giving consideration to usability principles. Even though it's a well-engineered data warehouse that can perform well, many business users do not use the dashboard, as it is perceived as not being user friendly, leading to poor adoption of the infrastructure and change management issues.

3. Prioritize and Rank Alerts/Exceptions Streamed to the Dashboard Because there are tons of raw data, it is important to have a mechanism by which important exceptions/behaviors are proactively pushed to the information consumers. A business rule can be codified, which detects the alert pattern of interest.

4. Enrich the Dashboard with Business-User Comments When the same dashboard information is presented to multiple business users, a small text box can be provided that can capture the comments from an end-user's perspective. This can often be tagged to the dashboard to put the information in context, adding perspective to the structured KPIs being rendered

5. Present Information in Three Different Levels Information can be presented in three layers depending on the granularity of the information: the visual dashboard level, the static report level, and the self-service cube level. When a user navigates the dashboard, a simple set of 8 to 12 KPIs can be presented, which would give a sense of what is going well and what is not.

6. Pick the Right Visual Construct Using Dashboard Design Principles In presenting information in a dashboard, some information is presented best with bar charts, some with time series line graphs, and when presenting correlations, a scatter plot is useful. Sometimes merely rendering it as simple tables is effective. Once the dashboard design principles are explicitly documented, all the developers working on the front end can adhere to the same principles while rendering the reports and dashboard.

7. Provide for Guided Analytics In a typical organization, business users can be at various levels of analytical maturity. The capability of the dashboard can be used to guide the "average" business user to access the same navigational path as that of an analytically savvy business user

The foundation for an important type of database, called a data warehouse, which is primarily used for decision support and provides the informational foundation for improved analytical capabilities.

Business intelligence (BI), as a term to describe evidence/fact-based managerial decision making, It is the descriptive analytics portion of the business analytics continuum, the maturity of which leads to advanced analytics—a combination of predictive and prescriptive analytics.

BI systems rely on a data warehouse as the information source for creating insight and supporting managerial decisions. A multitude of organizational and external data is captured, transformed, and stored in a data warehouse to support timely and accurate decisions through enriched business insight

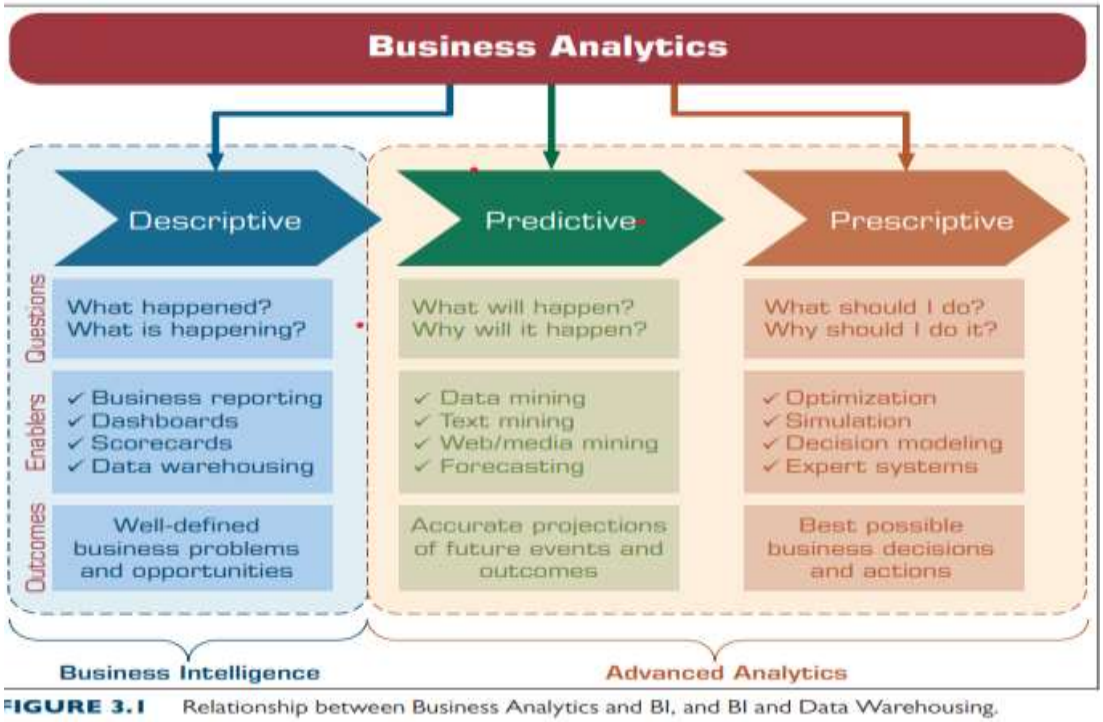


FIGURE 3.1 Relationship between Business Analytics and BI, and BI and Data Warehousing.

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

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, nonvolatile collection of data in support of management's decision-making process.

Characteristics of Data Warehousing:

a.**Subject oriented:** Data are organized by detailed subject, such as sales, products, or customers, containing only information relevant for decision support. Subject orientation enables users to determine not only how their business is performing and also for its purpose. It provides a more comprehensive view of the organization.

b. **Integrated**. Integration is closely related to subject orientation. Data warehouses must place data from different sources into a consistent format. To do so, they must deal with naming conflicts and discrepancies among units of measure. A data warehouse is presumed to be totally integrated.

c.**Time variant (time series).** A warehouse maintains historical data. The data do not necessarily provide current status (except in real-time systems). They detect trends, deviations, and long-term relationships for forecasting and comparisons, leading to decision making. Every data warehouse has a temporal quality. Time is the one important dimension that all data warehouses must support. Data for analysis from multiple sources contain multiple time points (e.g., daily, weekly, monthly views).

d.**Nonvolatile.** After data are entered into a data warehouse, users cannot change or update the data. Obsolete data are discarded, and changes are recorded as new data.

Some **additional characteristics** may include the following:

- **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.

TYPES OF DATA WAREHOUSES:

The three main types of data warehouses are data marts (DMs), operational data stores (ODS), and enterprise data warehouses (EDW)

Data marts: A DM is a subset of a data warehouse, typically consisting of a single subject area (e.g., marketing, operations). A DM can be either dependent or independent.

(i) Dependent data mart : It is a subset that is created directly from the data warehouse.

It support the concept of a single enterprise-wide data model, but the data warehouse must be constructed first. A dependent DM ensures that the end user is viewing the same version of the data that is accessed by all other data warehouse users.

Advantages : consistent data model and providing quality data.

Disadvantage: The high cost of data warehouses limits their use to large companies.

(ii) Independent data mart: As an alternative, many firms use a lower-cost, scaled-down version of a data warehouse referred to as an independent DM. An independent data mart is a small warehouse designed for a strategic business unit or a department, but its source is not an EDW.

2. Operational Data Stores An operational data store (ODS) provides a fairly recent form of customer information file. This type of database is often used as an interim staging area for a data warehouse. Unlike the static contents of a data warehouse, the contents of an ODS are updated throughout the course of business operations.

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.

An ODS is similar to short-term memory in that it stores only very recent information. In comparison, a data warehouse is like long-term memory because it stores permanent information.

An ODS consolidates data from multiple source systems and provides a near–real-time, integrated view of volatile, current data. The exchange, transfer, and load (ETL) processes for an ODS are identical to those for a data warehouse.

3. Enterprise Data Warehouses (EDW): An enterprise data warehouse (EDW) is a large-scale data warehouse that is used across the enterprise for decision support. The large-scale nature of an EDW provides integration of data from many sources into a standard format for effective BI and decision support applications.

EDWs are used to provide data for many types of decision support systems (DSS), including customer relationship management (CRM), supply chain management (SCM), business performance management (BPM), business activity monitoring, product life cycle management, revenue management, and sometimes even knowledge management systems.

Metadata : Metadata are data about data. It describes the structure of and some meaning about data, thereby contributing to their effective or ineffective use.

Types of Metadata in the point of pattern view:

1.syntactic metadata (i.e., data describing the syntax of data)

2. structural metadata (i.e., data describing the structure of the data)

3.semantic metadata (i.e., data describing the meaning of the data in a specific domain).

Data Warehousing Process:

Many organizations need to create data warehouses—massive data stores of time series data for decision support. Data are imported from various external and internal resources and are cleansed and organized in a manner consistent with the organization's needs.

After the data are populated in the data warehouse, DMs can be loaded for a specific area or department. Alternatively, DMs can be created first, as needed, and then integrated into an EDW. Often, though, DMs are not developed, but data are simply loaded onto PCs or left in their original state for direct manipulation using BI tools.

Components of DataWarehousing: The following are the components of Datawarehousing

1.Data sources
2.Data extraction and transformation.
3.Data loading
4. Comprehensive database
5.Metadata
6.Middleware tools.

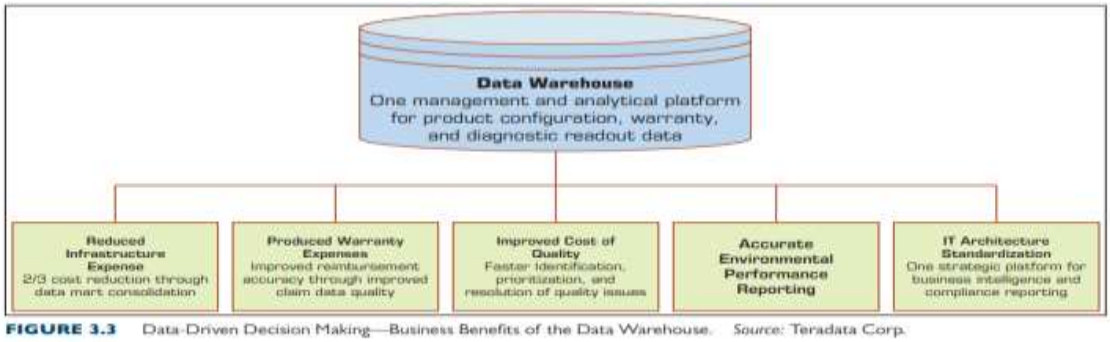


FIGURE 3.3 Data-Driven Decision Making—Business Benefits of the Data Warehouse. *Source:* Teradata Corp.

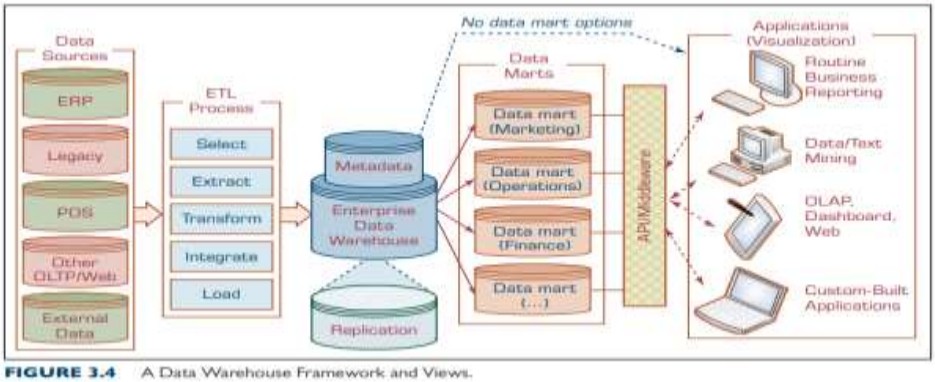


FIGURE 3.4 A Data Warehouse Framework and Views.

1. Data sources: Data are sourced from multiple independent operational "legacy" systems and possibly from external data providers (such as the U.S. Census). Data may also come from an OLTP or enterprise resource planning (ERP) system. Web data in the form of Web logs may also feed to a data warehouse.
2. Data extraction and transformation:Data are extracted and properly transformed using custom-written or commercial software called ETL.

3. Data loading: Data are loaded into a staging area, where they are transformed and cleansed. The data are then ready to load into the data warehouse and/or DMs.

4. Comprehensive database: Essentially, this is the EDW to support all decision analysis by providing relevant summarized and detailed information originating from many different sources.
5. Metadata:
6. Middleware tools. Middleware tools enable access to the data warehouse. Power users such as analysts may write their own SQL queries. Others may employ a managed query environment, such as Business Objects, to access data. There are many front-end applications that business users can use to interact with data stored in the data repositories, including data mining, OLAP, reporting tools, and data visualization tools.

Data Warehousing Architectures:

Several basic information system architectures can be used for data warehousing. Generally these architectures are commonly called client/server or n-tier architectures, of which they are two-tier and three-tier architectures .These types of multitiered architectures are known to be capable of serving the needs of large-scale, performance-demanding information systems such as data warehouses.

These n-tired architectures divide the data warehouse into three parts:
1.The data warehouse itself, which contains the data and associated software
2. Data acquisition (back-end) software, which extracts data from legacy systems and external sources, consolidates and summarizes them, and loads them into the data warehouse
3. Client (front-end) software, which allows users to access and analyze data from the warehouse (a DSS/BI/business analytics [BA] engine).
3-tier architecture:
In a three-tier architecture, operational systems contain the data and the software for data acquisition in one tier (i.e., the server), the data warehouse is another tier, and the third tier includes the DSS/BI/BA engine (i.e., the application server) and the client.

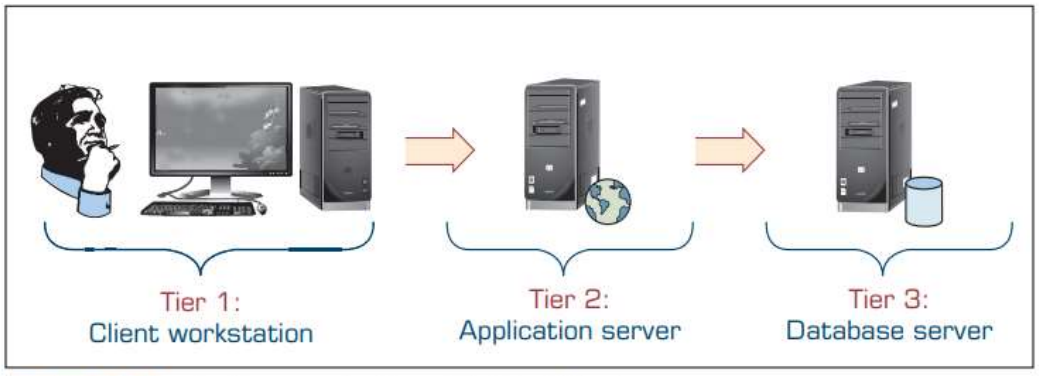


FIGURE 3.5 Architecture of a Three-Tier Data Warehouse.

Data from the warehouse are processed twice and deposited in an additional multidimensional database, organized for easy multidimensional analysis and presentation, or replicated in DMs.

Advantage : The separation of the functions of the data warehouse, which eliminates resource constraints and makes it possible to easily create DMs.

2-tier architecture: In a two-tier architecture, the DSS engine physically runs on the same hardware platform as the data warehouse .Therefore, it is more economical than the three-tier structure. The two-tier architecture can have performance problems for large data warehouses that work with data-intensive applications for decision support

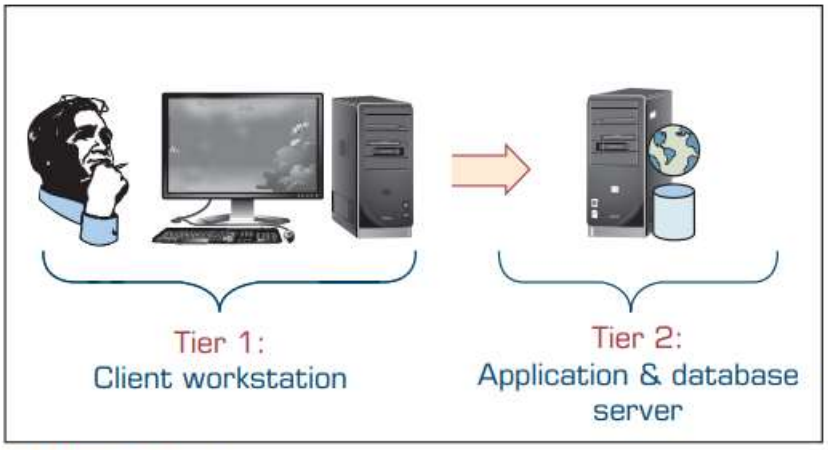


FIGURE 3.6 Architecture of a Two-Tier Data Warehouse.

Data warehousing and the Internet are two key technologies that offer important solutions for managing corporate data. The integration of these two technologies produces Web-based data warehousing.

Architecture of Web-based data warehousing:
The architecture is three-tiered and includes the PC client, Web server, and application server. On the client side, the user needs an Internet connection and a Web browser through the familiar graphical user interface (GUI). The Internet/intranet/extranet is the communication medium between client and servers.

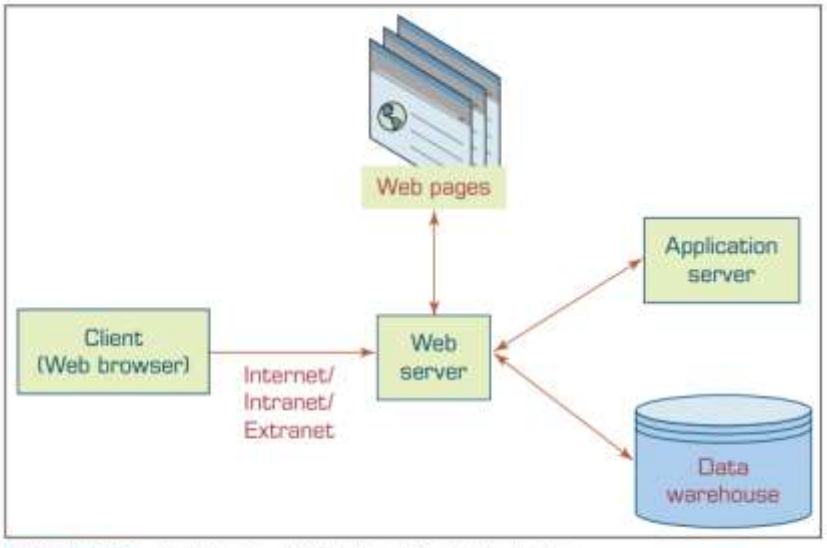


FIGURE 3.7 Architecture of Web-Based Data Warehousing.

On the server side, a Web server is used to manage the inflow and outflow of information between client and server. It is backed by both a data warehouse and an application server. Web-based data warehousing offers several compelling advantages, including ease of access, platform independence, and lower cost.

Web architectures for data warehousing are similar in structure to other data warehousing architectures, requiring a design choice for housing the Web data warehouse with the transaction server or as a separate server(s). Page-loading speed is an important consideration in designing Web-based applications; therefore, server capacity must be planned carefully.

Issues that must be considered when deciding which architecture:

. • Which database management system (DBMS) should be used? Most data warehouses are built using RDBMS. Oracle , SQL Server , and DB2 are the ones most commonly used. Each of these products supports both client/server and Web-based architectures.

• Will parallel processing and/or partitioning be used? Parallel processing enables multiple central processing units (CPUs) to process data warehouse query requests simultaneously and provides scalability. Data warehouse designers need to decide whether the database tables will be partitioned (i.e., split into smaller tables) for access efficiency and what the criteria will be.

• Will data migration tools be used to load the data warehouse? Moving data from an existing system into a data warehouse is a tedious and laborious task.

Depending on the diversity and the location of the data assets, migration may be a relatively simple procedure or a months-long project. The results of a thorough assessment of the existing data assets should be used to determine whether to use migration tools, and if so, what capabilities to seek in those commercial tools.

• What tools will be used to support data retrieval and analysis? Often it is necessary to use specialized tools to periodically locate, access, analyze, extract, transform, and load necessary data into a data warehouse.

A decision has to be made on (1) developing the migration tools in-house, (2) purchasing them from a third-party provider, or (3) using the ones provided with the data warehouse system.

Alternative Data Warehousing Architectures:

a. **Independent data marts:** This is arguably the simplest and the least costly architecture alternative. The DMs are developed to operate independent of each other to serve the needs of individual organizational units. Because of their independence, they may have inconsistent data definitions and different dimensions and measures, making it difficult to analyze data across the DMs.

b. **Data mart bus architecture**. This architecture is a viable alternative to the independent DMs where the individual marts are linked to each other via some kind of middleware. Because the data are linked among the individual marts, there is a better chance of maintaining data consistency across the enterprise. Even though it allows for complex data queries across DMs, the performance of these types of analysis may not be at a satisfactory level.

c. **Hub-and-spoke architecture.** This is perhaps the most famous data warehousing architecture today. Here the attention is focused on building a scalable and maintainable infrastructure (often developed in an iterative way, subject area by subject area) that includes a centralized data warehouse and several dependent DMs (each for an organizational unit). This architecture allows for easy customization of user interfaces and reports. On the negative side, this architecture lacks the holistic enterprise view and may lead to data redundancy and data latency.

d. **Centralized data warehouse.** It is similar to the hub-and-spoke architecture except that there are no dependent DMs; instead, there is a gigantic EDW that serves the needs of all organizational units. This centralized approach provides users with access to all data in the data warehouse instead of limiting them to DMs. In addition, it reduces the amount of data the technical team has to transfer or change, therefore simplifying data management and administration. If designed and implemented properly, this architecture provides a timely and holistic view of the enterprise to whoever, whenever, and wherever they may be within the organization.

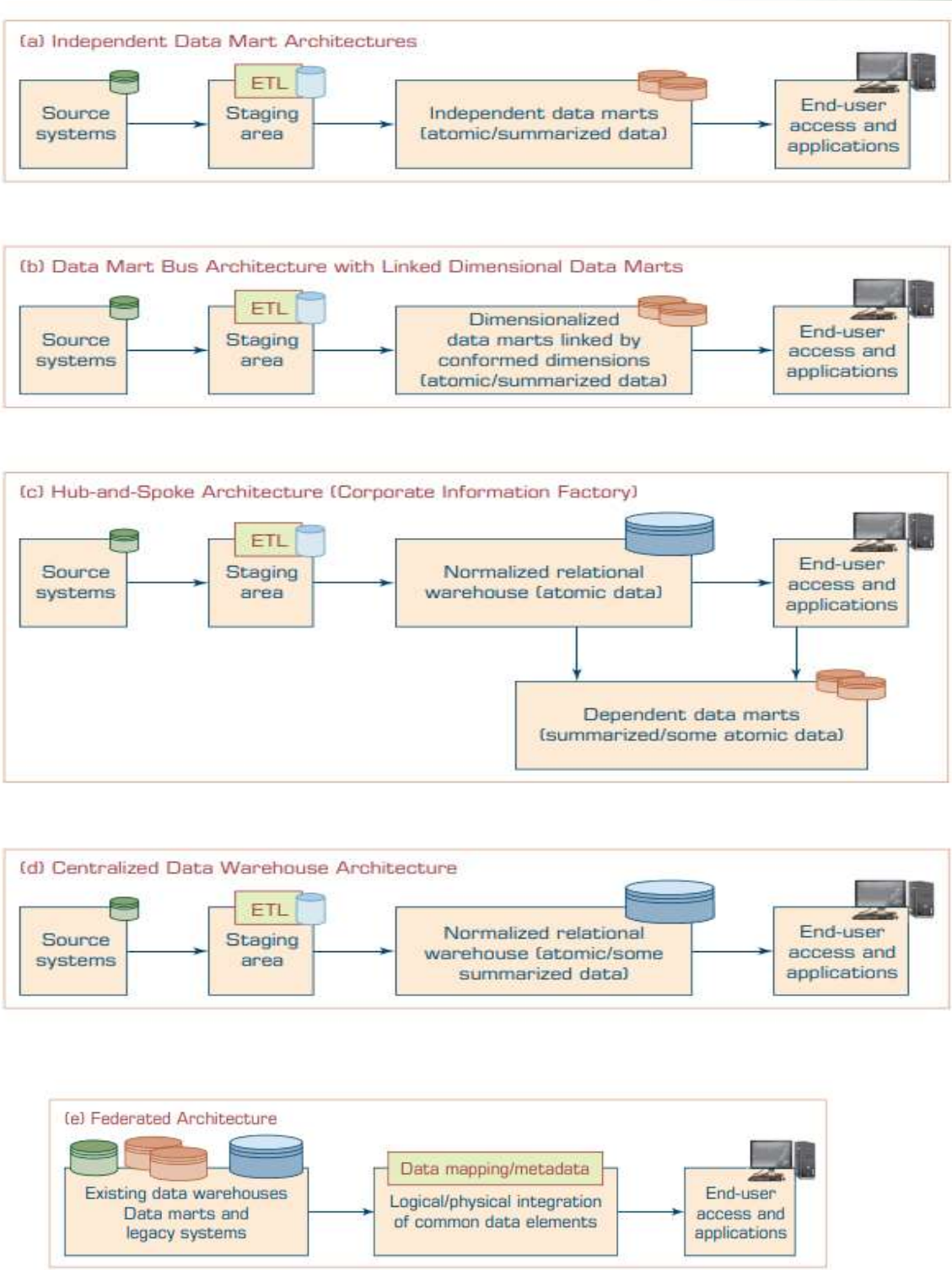


FIGURE 3.8 Alternative Data Warehouse Architectures. *Source:* Adapted from Ariyachandra, T., & Watson, H. (2006b). Which data warehouse architecture is most successful? *Business Intelligence Journal, 11*(1), 4–6.

e. **Federated data warehouse.**

It is a concession to the natural forces that finds the best plans for developing a perfect system. It uses all possible means to integrate analytical resources from multiple sources to meet changing needs or business conditions.

Essentially, the federated approach involves integrating disparate systems. In a federated architecture, existing decision support structures are left in place, and data are accessed from those sources as needed.

The federated approach is supported by middleware vendors that propose distributed query and join capabilities

Factors that potentially affect the architecture selection decision:

1. Information interdependence between organizational units
2. Upper management's information needs
3. Urgency of need for a data warehouse
4. Nature of end-user tasks
5. Constraints on resources
6. Strategic view of the data warehouse prior to implementation
7. Compatibility with existing systems
8. Perceived ability of the in-house IT staff
9. Technical issues
10. Social/political factors .

Which Architecture Is the Best?:

They used four measures to assess the success of the architectures: (1) information quality, (2) system quality, (3) individual impacts, and (4) organizational impacts. The questions used a 7-point scale, with the higher score indicating a more successful architecture.

TABLE 3.1 Average Assessment Scores for the Success of the Architectures

	Independent DMs	Bus Architecture	Hub-and-Spoke Architecture	Centralized Architecture (No Dependent DMs)	Federated Architecture
Information Quality	4.42	5.16	5.35	5.23	4.73
System Quality	4.59	5.60	5.56	5.41	4.69
Individual Impacts	5.08	5.80	5.62	5.64	5.15
Organizational Impacts	4.66	5.34	5.24	5.30	4.77

The results of the study indicate, independent DMs scored the lowest on all measures. This finding confirms the conventional wisdom that independent DMs are a poor architectural solution.

Next lowest on all measures was the federated architecture. Firms sometimes have disparate decision-support platforms resulting from mergers and acquisitions, and they may choose a federated approach, at least in the short term. The findings suggest that the federated architecture is not an optimal long-term solution.

However, is the similarity of the averages for the bus, hub-and-spoke, and centralized architectures. The differences are sufficiently small that no claims can be made for a particular architecture's superiority over the others, at least based on a simple comparison of these success measures. They also collected data on the domain and the size of the warehouses. They found that the hub-and-spoke architecture is typically used with more enterprise-wide implementations and larger warehouses.

They also investigated the cost and time required to implement the different architectures. Overall, the hub-and-spoke architecture was the most expensive and time-consuming to implement.

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

Data Integration : It comprises three major processes that, when correctly implemented, 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).

A major purpose of a data warehouse is to integrate data from multiple systems.

Integration technologies enable data and metadata integration are

- Enterprise application integration (EAI)
- Service-oriented architecture (SOA)
- Enterprise information integration (EII)
- Extraction, transformation, and load (ETL)

1. **Enterprise application integration (EAI) :** It 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.

2. Service-oriented architecture (SOA): Recently, EAI is accomplished by using 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.

3. Enterprise information integration (EII) : It is 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. EII tools use predefined metadata to populate views that make integrated data appear relational to end users.

4. Extraction, Transformation, and Load (ETL): ETL technologies, which have existed for some time, are instrumental in the process and use of data warehouses. The ETL process is an integral component in any data-centric project.

ETL process:

It 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).

Transformation occurs by using rules or lookup tables or by combining the data with other data. The three database functions are integrated into one tool to pull data out of one or more databases and place them into another, consolidated database or a 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).

ETL is extremely important for data integration as well as for data warehousing. **The purpose of the ETL process** is to load the warehouse with integrated and cleansed data. The data used in ETL processes can come from any source: a mainframe application, an ERP application, a CRM tool, a flat file, an Excel spreadsheet, or even a message queue.

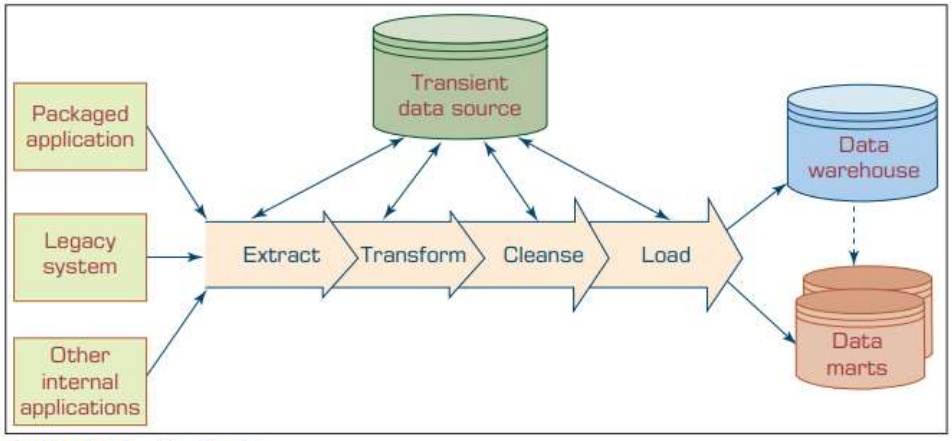


FIGURE 3.9 The ETL Process.

The process of migrating data to a data warehouse involves the extraction of data from all relevant sources. Data sources may consist of files extracted from OLTP databases, spreadsheets, personal databases (e.g., Microsoft Access), or external files. Typically, all the input files are written to a set of staging tables, which are designed to facilitate the load process.

A data warehouse contains numerous business rules that define such things as how the data will be used, summarization rules, standardization of encoded attributes, and calculation rules. Any data quality issues pertaining to the source files need to be corrected before the data are loaded into the data warehouse. One of the benefits of a well-designed data warehouse is that these rules can be stored in a metadata repository and applied to the data warehouse centrally. This differs from an OLTP approach, which typically has data and business rules scattered throughout the system.

The process of loading data into a data warehouse can be performed either through data transformation tools that provide a GUI to aid in the development and maintenance of business rules or through more traditional methods.

Issues affect whether an organization will purchase data transformation tools or build the transformation process itself:

- Data transformation tools are expensive.
- Data transformation tools may have a long learning curve.
- It is difficult to measure how the IT organization is doing until it has learned to use the data transformation tools.

The important criteria in selecting an ETL tool:

- Ability to read from and write to an unlimited number of data source architectures
- Automatic capturing and delivery of metadata
- A history of conforming to open standards
- An easy-to-use interface for the developer and the functional user

Performing extensive ETL may be a sign of poorly managed data and a fundamental lack of a coherent data management strategy.

Data Warehouse Development:

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

Direct benefits : • 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. A data warehouse permits information processing to be relieved from costly operational systems onto low-cost servers; therefore, many more end-user information requests can be processed more quickly.
- 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.

Indirect benefits: On the whole, these benefits enhance business knowledge, present a competitive advantage, improve customer service and satisfaction, facilitate decision making, and help in reforming business processes.

Data Warehouse Development Approaches:

Many organizations need to create the data warehouses used for decision support. Two competing approaches are employed.

The first approach is that of **Bill Inmon**, who is often called "the father of data warehousing." Inmon supports a top-down development approach that adapts traditional relational database tools to the development needs of an enterprise-wide data warehouse, also known as the EDW approach.

The second approach is that of **Ralph Kimball,** who proposed a bottom-up approach that employs dimensional modeling, also known as the DM approach.

THE INMON MODEL: THE EDW APPROACH Inmon's approach emphasizes top-down development, employing established database development methodologies and tools, such as entity-relationship diagrams (ERD) and an adjustment of the spiral development approach. The EDW approach does not preclude the creation of DMs. The EDW is the ideal in this approach because it provides a consistent and comprehensive view of the enterprise.

THE KIMBALL MODEL: THE DATA MART APPROACH Kimball's DM strategy is a "plan big, build small" approach. A DM is a subject-oriented or department-oriented data warehouse. It is a scaled-down version of a data warehouse that focuses on the requests of a specific department, such as marketing or sales. This model applies dimensional data modeling, which starts with tables. Kimball advocated a development methodology that entails a bottom-up approach, which in the case of data warehouses means building one DM at a time.

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

Representation of Data in Data Warehouse :

The design of data representation in the data warehouse has always been based on the concept of dimensional modeling.

Dimensional modeling is a retrieval-based system that supports high-volume query access. Representation and storage of data in a data warehouse should be designed in such a way that not only accommodates but also boosts the processing of complex multidimensional queries.

Often, the star schema and the snowflake schema are the means by which dimensional modeling is implemented in data warehouses.

Star schema:

The star schema (sometimes referenced as star join schema) is the most commonly used and the simplest style of dimensional modeling. 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). A fact table contains the descriptive attributes needed to perform decision analysis and query reporting, and foreign keys are used to link to dimension tables.

The decision analysis attributes consist of performance measures, operational metrics, aggregated measures (e.g., sales volumes, customer retention rates, profit margins, production costs, scrap rate), and all the other metrics needed to analyze the organization's performance

Surrounding the central fact tables (and linked via foreign keys) are dimension tables. The dimension tables contain classification and aggregation information about the central fact rows.

Dimension tables contain attributes that describe the data contained within the fact table; they address how data will be analyzed and summarized. Dimension tables have a one-to-many relationship with rows in the central fact table.

In querying, the dimensions are used to slice and dice the numerical values in the fact table to address the requirements of an ad hoc information need. The star schema is designed to provide fast query-response time, simplicity, and ease of maintenance for read-only database structures.

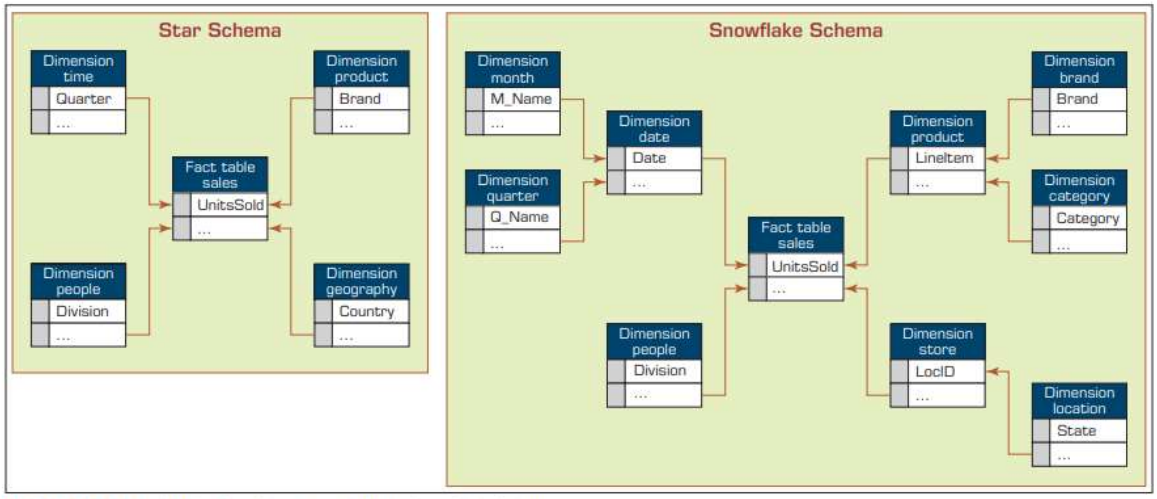


FIGURE 3.10 (a) The Star Schema, and (b) the Snowflake Schema.

Snowflake schema: It is a logical arrangement of tables in a multidimensional database in such a way that the entity-relationship diagram resembles a snowflake in shape. Closely related to the star schema, the snowflake schema is represented by centralized fact tables (usually only one), which are connected to multiple dimensions.

In the snowflake schema, however, dimensions are normalized into multiple related tables, whereas the star schema's dimensions are denormalized, with each dimension being represented by a single table.

10. Assignment

Toppers:

Medical Device Company Ensures Product Quality While Saving Money

A business scenario in which a data-rich medical device research and development company streamlined their analytics practices to have easy access to both the data and the analyses they need to continue the traditions of innovation and quality at the highest levels.

1. What were the main challenges for the medical device company? Were they market or technology driven? Explain.
2. What was the proposed solution?
3. What were the results? What do you think was the real return on investment (ROI)? Source:

   Source: Dell customer case study. Medical device company ensures product quality while saving hundreds of thousands of dollars.

Above average Improving Student Retention with Data-Driven Analytics:

A raw, readily available academic data within an educational organization is used to develop predictive models to better understand attrition and improve freshmen.

1. What is student attrition, and why is it an important problem in higher education?

2. What were the traditional methods to deal with the attrition problem?

3. List and discuss the data-related challenges within context of this case study.

4. What was the proposed solution? And, what were the results.

Predicting NCAA Bowl Game Outcomes:

how existing and readily available public data sources can be used to predict college football bowl game outcomes using both classification and regression-type prediction models.

1. What are the foreseeable challenges in predicting sporting event outcomes (e.g., college bowl games)?

2. How did the researchers formulate/design the prediction problem (i.e., what were the inputs and output, and what was the representation of a single sample—row of data)?

3. How successful were the prediction results? What else can they do to improve the accuracy?

Below average Flood of Paper Ends at FEMA:

To illustrate the power and the utility of automated report generation for a large (and, at a time of natural crisis, somewhat chaotic) organization like FEMA.

1. What is FEMA, and what does it do?

2. What are the main challenges that FEMA faces?

3. How did FEMA improve its inefficient reporting practices?

Slow performers - How to Calculate Descriptive Statistics in Microsoft Excel?

11. Part - A Question and Answers

1.List some of the characteristics of the nature of data. (CO1,K1)

- Data source reliability
- Data content accuracy
- Data accessibility
- Data security and data privacy
- Data richness
- Data consistency
- Data currency/data timeliness
- Data granularity
- Data validity
- Data relevancy

2.How do you describe the importance of data in analytics? Can we think of analytics without data? (CO1,K1)

Data is the main ingredient for any BI, data science, and business analytics initiative. In fact, it can be viewed as the raw material for what these popular decision technologies produce—information, insight, and knowledge. Without data none of these technologies could exist and be popularized.

3.Where does the data for business analytics come from? (CO1,K1)

Traditional ways to manually collect data (either via surveys or via human-entered business transactions) mostly left their places to modern day data collection mechanisms that use Internet and/or sensor/RFID-based computerized networks.

4.What is data? How does data differ from information and knowledge?(CO1,K1)

Data (datum in singular form) refers to a collection of facts usually obtained as the result of experiments, observations, transactions, or experiences. Data may consist of numbers, letters, words, images, voice recordings, and so on, as measurements of a set of variables.

5. What is categorical data. Give examples. (CO1,K1)

Categorical data represent the labels of multiple classes used to divide a variable into specific groups. Examples of categorical variables include race, sex, age group, and educational level.

6. What is Ordinal data. Give examples.(CO1,K1)

Ordinal data contain codes assigned to objects or events as labels that also represent the rank order among them. For example, the variable credit score can be generally categorized as (1) low, (2) medium, or (3) high. Similar ordered relationships can be seen in variables such as age group (i.e., child, young, middle-aged, elderly) and educational level (i.e., high school, college, graduate school).

7. What are the main data preprocessing steps? (CO1,K1)

- Data Consolidation
- Data Cleaning
- Data Transformation
- Data Reduction

8. What is measures of dispersion? (CO1,K1)

Measures of dispersion are the mathematical methods used to estimate or describe the degree of variation in a given variable of interest. They are a representation of the numerical spread (compactness or lack thereof) of a given data set. To describe this dispersion, a number of statistical measures are developed; the most notable ones are range, variance, and standard deviation.

9. What is Mean Absolute Deviation? (CO1,K1)

Measuring the absolute values of the differences between each data point and the mean and summing them. It provides a measure of spread without being specific about the data point being lower or higher than the mean.

$$MAD = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

10.Define Kurtosis? (CO1,K1)

Kurtosis is another measure to use in characterizing the shape of an unimodal distribution. As opposed to the sway in shape, kurtosis is more interested in characterizing the peak/tall/skinny nature of the distribution. Specifically, 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.

11. Difference between correlation and regression? (CO1,K1)

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. On the other hand, regression attempts to describe the dependence of a response variable on one (or more) explanatory variables where it implicitly assumes that there is a one-way causal effect from the explanatory variable(s) to the response variable, regardless of whether the path of effect is direct or indirect.

12.What is Ordinary Least Squares (OLS)? (CO1,K1)

OLS method aims to minimize the sum of squared residuals (squared vertical distances between the observation and the regression point) and leads to a mathematical expression for the estimated value of the regression line (which are known as b parameters)

13. **Define Business Reporting? What are the main characteristics of a good business report? (CO1,K1)**

A report is any communication artifact prepared with the specific intention of conveying information in a digestible form to whoever needs it, whenever and wherever they may need it. Business reports are various sources of data coming from both inside and outside the organization (online transaction processing [OLTP] systems). Creation of these reports involves ETL (extract, transform, and load) procedures in coordination with a data warehouse and then using one or more reporting tools. Key to any successful report are clarity, brevity, completeness, and correctness.

14. **What is Visual analytics** ? **(CO1,K1)**

Visual analytics is a recently coined term that is often used loosely to mean nothing more than information visualization. What is meant by visual analytics is the combination of visualization and predictive analytics. Whereas information visualization is aimed at answering, "What happened?" and "What is happening?" and is closely associated with BI (routine reports, scorecards, and dashboards), visual analytics is aimed at answering, "Why is it happening?" "What is more likely to happen?" and is usually associated with business analytics (forecasting, segmentation, correlation analysis)

15. **What are the most distinctive feature of a dashboard? (CO1,K1)**

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

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

16. What are the best practices in dashboard design? (CO1,K1)

- Benchmark Key Performance Indicators with Industry Standards
- Wrap the Dashboard Metrics with Contextual Metadata
- Validate the Dashboard Design by a Usability Specialist

- Prioritize 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.

17. Define Data Marts?

A data mart (DM) is usually smaller and focuses on a particular subject or department. A DM is a subset of a data warehouse, typically consisting of a single subject area (e.g., marketing, operations).

18. What is 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.

19.What is datawarehousing? What are its components?

A data warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision-making process.

Components: Data sources, Data extraction and transformation, Data loading**,** Comprehensive database, Metadata,Middleware tools.

20. Identify and discuss the role of middleware tools?

Middleware tools enable access to the data warehouse. Power users such as analysts may write their own SQL queries. Others may employ a managed query environment, such as Business Objects, to access data. There are many front-end applications that business users can use to interact with data stored in the data repositories, including data mining, OLAP, reporting tools, and data visualiza.

21. What are the key similarities and differences between a two-tiered architecture and a three-tiered architecture?

In a three-tier architecture, operational systems contain the data and the software for data acquisition in one tier (i.e., the server), the data warehouse is another tier, and the third tier includes the DSS/BI/BA engine (i.e., the application server) and the client. Data from the warehouse are processed twice and deposited in an additional multidimensional database, organized for easy multidimensional analysis and presentation, or replicated in DMs. 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 a two-tier architecture, the DSS engine physically runs on the same hardware platform as the data warehouse Therefore, it is more economical than the three-tier structure. The two-tier architecture can have performance problems for large data warehouses that work with data-intensive applications for decision support

21. What are the factors that potentially affect the architecture selection decision?

1. Information interdependence between organizational units 2. Upper management's information needs 3. Urgency of need for a data warehouse 4. Nature of end-user tasks 5. Constraints on resources 6. Strategic view of the data warehouse prior to implementation 7. Compatibility with existing systems 8. Perceived ability of the in-house IT staff 9. Technical issues 10. Social/political factors.

22. . Describe the three steps of the ETL process.

The three steps is 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). Transformation occurs by using rules or lookup tables or by combining the data with other data.

23. What is a cube? What do drill down, roll-up, slice, and dice mean?

The main operational structure in OLAP is based on a concept called cube. A cube in OLAP is a multidimensional data structure (actual or virtual) that allows fast analysis of data. It can also be defined as the capability of efficiently manipulating and analyzing data from multiple perspectives.

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. 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. The dice operation is a slice on more than two dimensions of a data cube.

24. Difference between OLTP and OLAP

An OLTP system addresses a critical business need, automating daily business transactions, and running real-time reports and routine analysis. But these systems are not designed for ad hoc analysis and complex queries that deal with a number of data items.

OLAP, on the other hand, is designed to address this need by providing ad hoc analysis of organizational data much more effectively and efficiently. OLAP and OLTP rely heavily on each other: OLAP uses the data captured by OLTP, and OLTP automates the business processes that are managed by decisions supported by OLAP.

25.What is dimensional modelling and how it is implemented in data warehouse?

Dimensional modeling is a retrieval-based system that supports high-volume query access. Representation and storage of data in a data warehouse should be designed in such a way that not only accommodates but also boosts the processing of complex multidimensional queries.

The star schema and the snowflake schema are the means by which dimensional modeling is implemented in data warehouses.

26.What is meant by star and snowflake schemas?

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). A fact table contains the descriptive attributes needed to perform decision analysis and query reporting, and foreign keys are used to link to dimension tables.

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. Dimensions are normalized into multiple related tables, whereas the star schema's dimensions are denormalized, with each dimension being represented by a single table.

12. Part - B Questions

1. Explain the steps of data preprocessing in detail (CO1,K1)
2. Explain Statistical modeling for Business Analytics in detail. (CO1,K1)
3. Explain the measures of Centrality Tendency and measures of dispersion in detail. (CO1,K1)
4. Explain Regression modeling for Inferential Statistics and the process flow for developing Regression models.(CO1,K1)
5. Explain Business reporting and its types?(CO1,K1)
6. Explain different Types of Charts and Graphs for Data visualization in detail (CO1,K1)
7. What is Data Warehouse and draw two-tier and three tier DW architecture and explain with an example.(CO1,K2)
8. Explain Data Integration and ETL Processes with a neat diagram.(CO2,K1)
9. What are the Warehouse development approaches. Give a list of Data Warehousing vendors and the products that they produce.(CO1,K1)
10. What are the various approaches of DataWarehouseDevelopment. development.(CO1,K1).
11. What are the representation of Data in DataWarehouse.Explain in detail.(CO1,K1)
12. What is a cube?What is its purpose in OLAP? What are commonly used operations in OLAP?Explain with neat diagram.

13. SUPPORTIVE ONLINE CERTIFICATION COURSES

NPTEL : https://onlinecourses.nptel.ac.in/noc20_mg11/preview

Coursera : https://in.coursera.org/specializations/business-analytics

Udemy : https://www.udemy.com/course/businessanalysis/

14. Real Time Applications

Fraud Detection in Financial Transactions:

Real-time analysis of financial transactions to identify and prevent fraudulent activities.

Dynamic Pricing in E-commerce:

Adjusting product prices in real-time based on market demand, competitor pricing, and other factors.

IoT Analytics for Smart Cities:

Monitoring and analyzing data from sensors, cameras, and other IoT devices to optimize city services, traffic flow, and resource allocation.

Healthcare Monitoring and Alerts:

Real-time monitoring of patient data, allowing for immediate alerts and interventions in critical situations.

Adaptive Streaming in Media Services:

Adjusting the quality of video streams based on network conditions and viewer preferences in real-time.

Supply Chain Optimization:

Tracking and analyzing data throughout the supply chain to optimize inventory levels, reduce delays, and enhance overall efficiency.

Energy Grid Monitoring:

Real-time monitoring of energy consumption and grid performance to optimize distribution and prevent outages.

Social Media Analytics:

Analyzing social media feeds in real-time to understand trends, sentiments, and engage with customers or address issues promptly.

Predictive Maintenance in Manufacturing:

Monitoring equipment sensors to detect anomalies and predict when maintenance is required to prevent downtime.

Real-time Analytics for Online Gaming:

Analyzing player behavior, preferences, and in-game actions to optimize gaming experiences and introduce dynamic challenges.

Network Security Monitoring:

Analyzing network traffic in real-time to detect and respond to security threats and breaches.

Weather Forecasting:

Processing real-time data from weather sensors to provide accurate and timely weather forecasts.

Airline Flight Optimization:

Analyzing real-time data on weather conditions, flight schedules, and airport operations to optimize flight routes and schedules.

Customer Interaction and Personalization:

Providing personalized recommendations and responses to customers in real-time based on their behavior and preferences.

15. Assessment Schedule

Sl. No.	ASSESSMENT	Proposed Date	Actual Date
1	FIRST INTERNAL ASSESSMENT	12-02-2024 to 17-02-2024	
2	SECOND INTERNAL ASSESSMENT	01-04-2024 to 06-04-2024	
3	MODEL EXAMINATION	20-04-2024 to 30-04-2024	
4	END SEMESTER EXAMINATION	11-05-2024	

16. PRESCRIBED TEXT BOOKS & REFERENCE BOOKS

TEXT BOOKS:

Ramesh Sharda, Dursun Delen, Efraim Turban, "Business Intelligence, Analytics, and Data Science: A Managerial Perspective", Pearson, 4th Edition, 2018.

Jesper Thorlund & Gert H.N. Laursen, "Business Analytics for Managers: Taking Business Intelligence beyond Reporting, Wiley, 2010.

REFERENCES:

Shmueli, Patel, and Bruce: Wiley, Data Mining for Business Intelligence, Concepts, Techniques and Applications, Wiley, 2010

R.N.Prasad and Seema Acharya, "Fundamentals of Business Analytics", 2nd Edition, Wiley, 2016.

17. Mini Project

Toppers – Design a dashboard on PowerBI to perform Market Basket Analysis

Above average – Employee Performance Analysis to be made for a certain organisation.

Average - Fraud Detection in Financial Transactions using Machine Learning. **Below average** – Product recommendation Analysis based on the customer preferences.

Slow performers – use machine learning algorithms and statistical analysis techniques to analyze medical data and help diagnose diseases or conditions. machine learning algorithms and

Thank you