Module 2: Big Data Analysis & Technology Concepts

INTRODUCTION **4**

Official Supplement: Examples of Big Data Mechanisms 4

Mind Map Poster 5

PART I: THE BIG DATA ANALYSIS LIFECYCLE **6**

Lifecycle Stages 6

Stage 1: Business Case Evaluation 7

Stage 2: Data Identification 8

Stage 3: Data Acquisition & Filtering 9

Stage 4: Data Extraction 11

Stage 5: Data Validation & Cleansing 13

Stage 6: Data Aggregation & Representation 15

Stage 7: Data Analysis 17

Stage 8: Data Visualization 19

Stage 9: Utilization of Analysis Results 20

Optional Reading 21

Exercise 2.1: Organize the Lifecycle Stages 22

PART II: BIG DATA ANALYSIS CONCEPTS **26**

STATISTICAL ANALYSIS **28**

A/B Testing 28

Correlation 29

Regression 31

Regression vs. Correlation 33

VISUAL ANALYSIS **37**

Heat Maps 38

Time Series Analysis 40

Network Analysis 41

Spatial Data Analysis 42

MACHINE LEARNING **47**

Law of Large Numbers 48

Law of Diminishing Marginal Utility 48

Classification 49
Clustering 50
Outlier Detection 51
Filtering 52
SEMANTIC ANALYSIS 56
Natural Language Processing 57
Text Analytics 58
Sentiment Analysis 59
Analysis Topic Mapping 59
Exercise 2.2: Map Problem Statements to Analysis Techniques 61
PART III: BIG DATA TECHNOLOGY CONCEPTS 69
BIG DATA TECHNOLOGY CONSIDERATIONS 70
Clusters 70
File Systems 71
Distributed File Systems 71
NoSQL 72
Parallel Data Processing 72
Distributed Data Processing 73
Processing Workloads 73
Processing Workload: Batch 73
Processing Workload: Transactional 74
Cloud Computing 75
BIG DATA TECHNOLOGY MECHANISMS 80
Examples of Big Data Mechanisms Supplement 80
Storage Device 80
Processing Engine 81
Resource Manager 82
Data Transfer Engine 83
Query Engine 85
Analytics Engine 86
Workflow Engine 87
Coordination Engine 88
Exercise 2.3: Fill in the Blanks 90
EXERCISE ANSWERS 96

EXERCISE 2.1 ANSWERS 96
EXERCISE 2.2 ANSWERS 96
EXERCISE 2.3 ANSWERS 97
EXAM B90.02 98
MODULE 2 SELF-STUDY KIT 98
CONTACT INFORMATION AND RESOURCES 99
GENERAL PROGRAM INFORMATION 99
GENERAL INFORMATION ABOUT COURSE MODULES AND SELF-STUDY KITS 99
PEARSON VUE EXAM INQUIRIES 99
PUBLIC INSTRUCTOR-LED WORKSHOP SCHEDULE 99
PRIVATE INSTRUCTOR-LED WORKSHOPS 100
BECOMING A CERTIFIED TRAINER 100
GENERAL BDSCP INQUIRIES 100
AUTOMATIC NOTIFICATION 100
FEEDBACK AND COMMENTS 100

Introduction

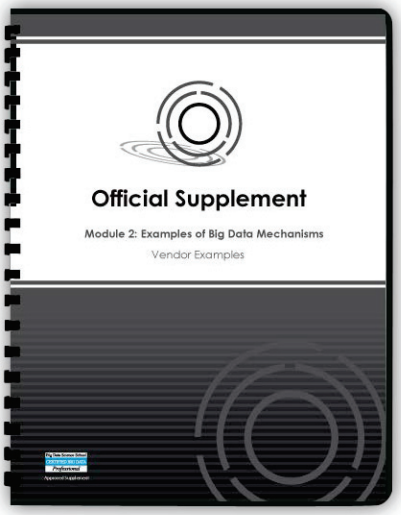
This is the official course booklet for the Big Data Science Certified Professional Course **Module 2: Big Data Analysis & Technology Concepts** and the corresponding Pearson VUE **Exam B90.02**.

This document is comprised of the following three primary parts:

- Part I: The Big Data Analysis Lifecycle
- Part II: Big Data Analysis Concepts
  - Statistical Analysis Techniques
  - Semantic Analysis Techniques
  - Machine Learning Techniques
  - Visual Analysis Techniques
  - Analysis Topic Mapping
- Part III: Big Data Technology Concepts
  - Big Data Technology Considerations
  - Big Data Technology Mechanisms

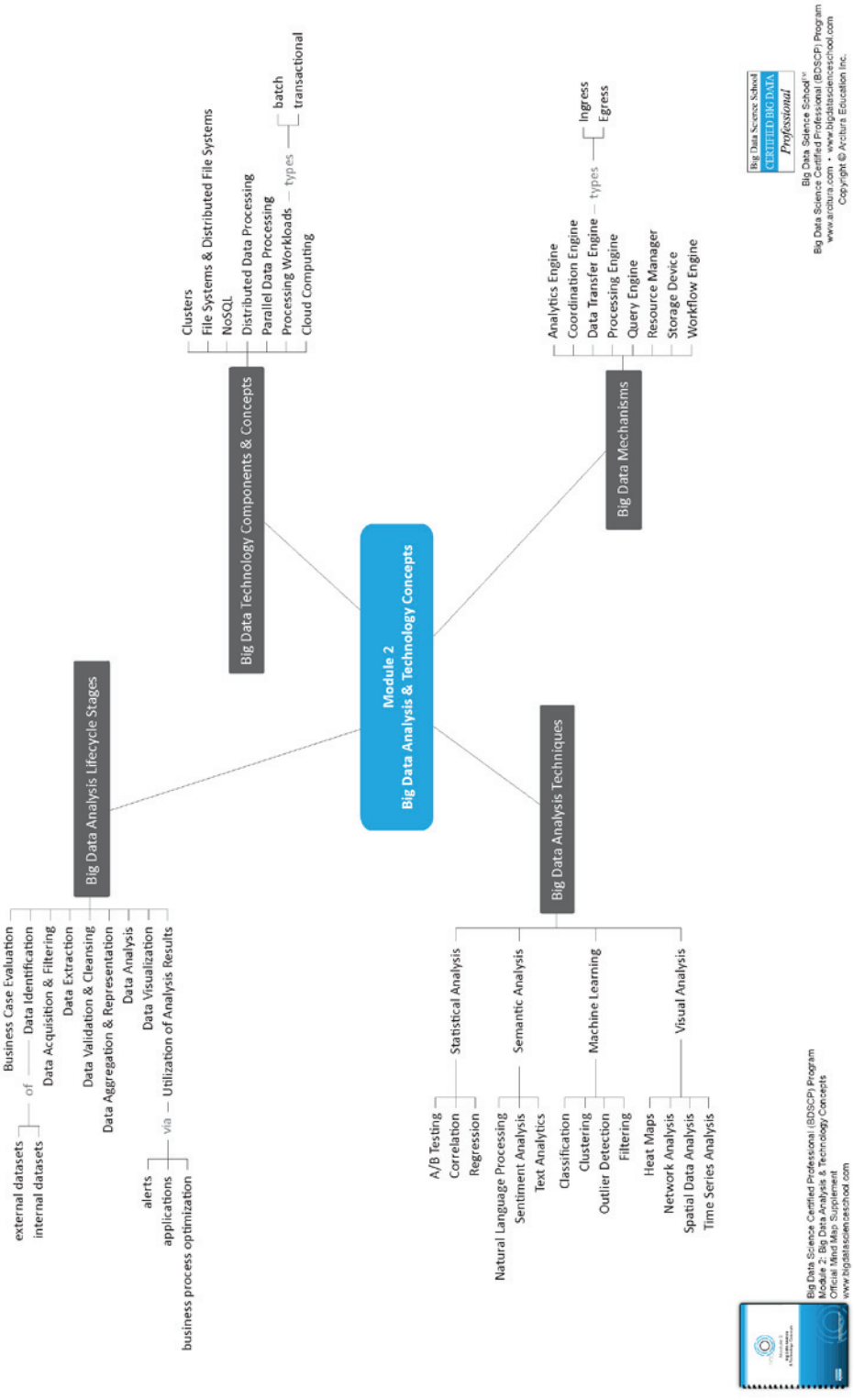
Official Supplement: Examples of Big Data Mechanisms

This supplement maps each of the Big Data mechanisms covered in this course module to one or more open source and/or vendor products or technologies.



Mind Map Poster

The *BDSCP Module 2: Mind Map Poster* that accompanies this course booklet provides an alternative visual representation of the topics covered in this course.



Part I: The Big Data Analysis Lifecycle

Big Data analysis differs from traditional data analysis primarily because of the volume, velocity, and variety characteristics of the data it processes.

To address the distinct circumstances and requirements of carrying out Big Data analysis and analytics, a fundamental step-by-step process is needed to organize the tasks involved with retrieving, processing, producing, and repurposing data.

The upcoming sections explore the Big Data analysis lifecycle that demonstrates these tasks.

Lifecycle Stages

The Big Data analysis lifecycle can be divided into the following nine stages, as shown in Figure 2.1:

1. Business Case Evaluation
2. Data Identification
3. Data Acquisition & Filtering
4. Data Extraction
5. Data Validation & Cleansing
6. Data Aggregation & Representation
7. Data Analysis
8. Data Visualization
9. Utilization of Analysis Results

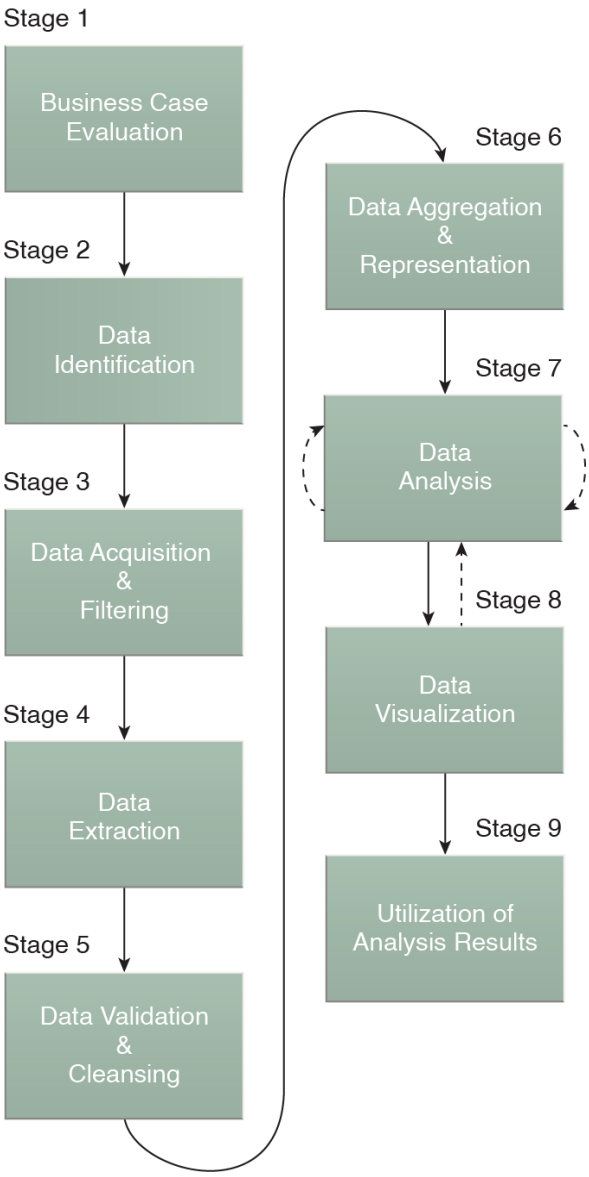


Figure 2.1- The Big Data Analysis Lifecycle

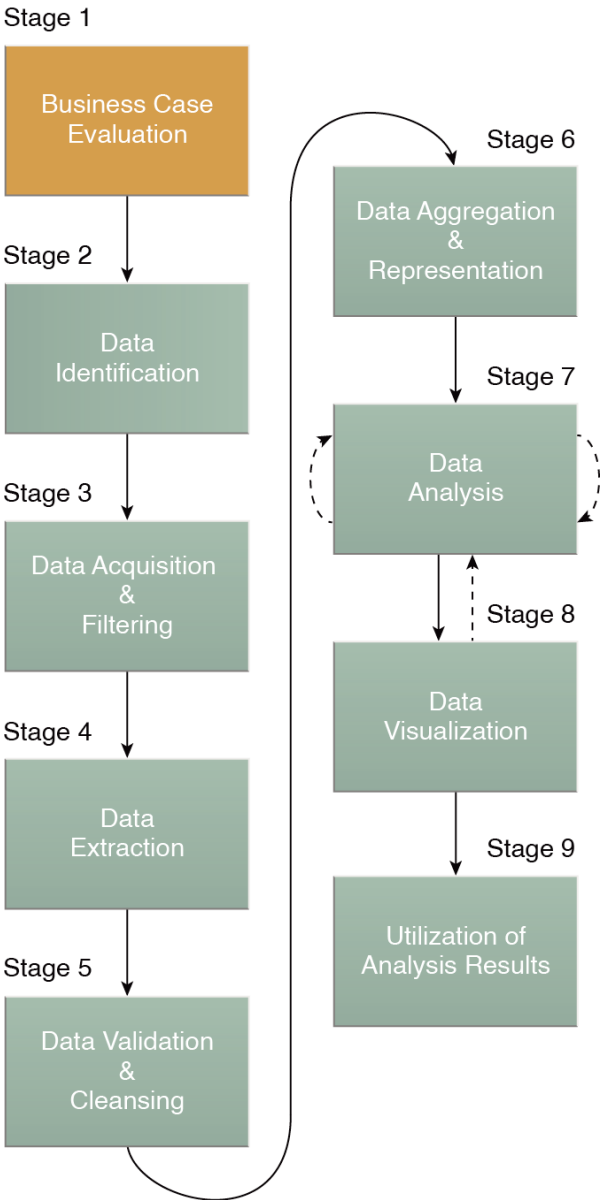
Stage 1: Business Case Evaluation

Each Big Data analysis lifecycle must begin with a well-defined business scope and a clear understanding of the justification, motivation, and goals of carrying out the analysis. The Business Case Evaluation stage requires that a business case be created, assessed, and approved prior to proceeding with the actual hands-on analysis tasks.

An evaluation of a Big Data analysis business case helps decision-makers understand the business resources that will need to be utilized, and which business challenges the analysis will tackle. The further identification of KPIs during this stage helps determine how closely the data analysis outcome needs to meet the identified goals and objectives.

Based on the business requirements documented in the business case, it can be determined whether the business problems being addressed are really Big Data problems. In order to qualify as a Big Data problem, a business problem needs to be directly related to one or more of the Big Data characteristics of volume, velocity, or variety.

Note also that another outcome of this stage is the determination of the underlying budget required to carry out the analysis project. Any required purchase of tools, hardware, training, etc. must be understood in advance so that the anticipated investment can be weighed against the expected benefits of achieving the goals. Initial iterations of the Big Data analysis lifecycle will require more up-front investment of Big Data technologies, products, and training compared to later iterations where these earlier investments can be repeatedly leveraged.



Stage 2: Data Identification

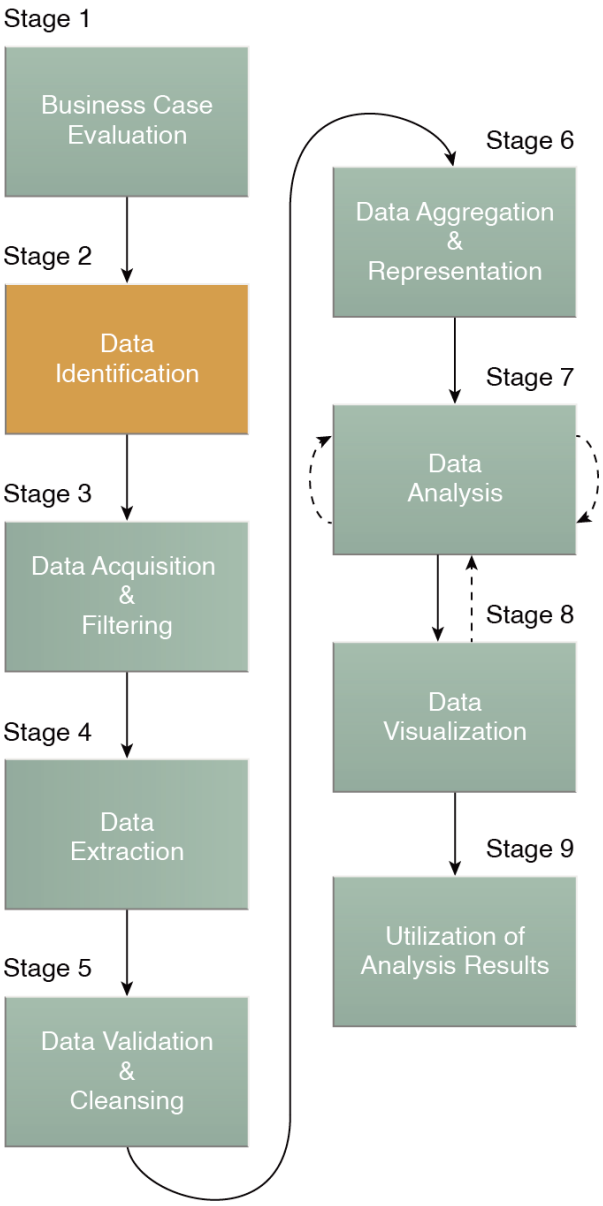
The Data Identification stage is dedicated to identifying the datasets required for the analysis project, and their sources.

Identifying a wider variety of data sources may increase the probability of finding hidden patterns and correlations. For example, it can be beneficial to identify as many types of related data sources and insights as possible, especially when unsure of exactly what to look for.

Depending on the business scope of the analysis project and nature of the business problems being addressed, the required datasets and their sources can be internal and/or external to the enterprise.

In the case of internal datasets, a list of available datasets from internal sources, such as data marts and operational systems, are typically compiled and matched against a pre-defined dataset specification.

In the case of external datasets, a list of possible third-party data providers, such as data markets and publicly available datasets, are generally compiled. Some forms of external data may be embedded within blogs or other types of content-based Web sites, in which case they may need to be harvested via automated tools.



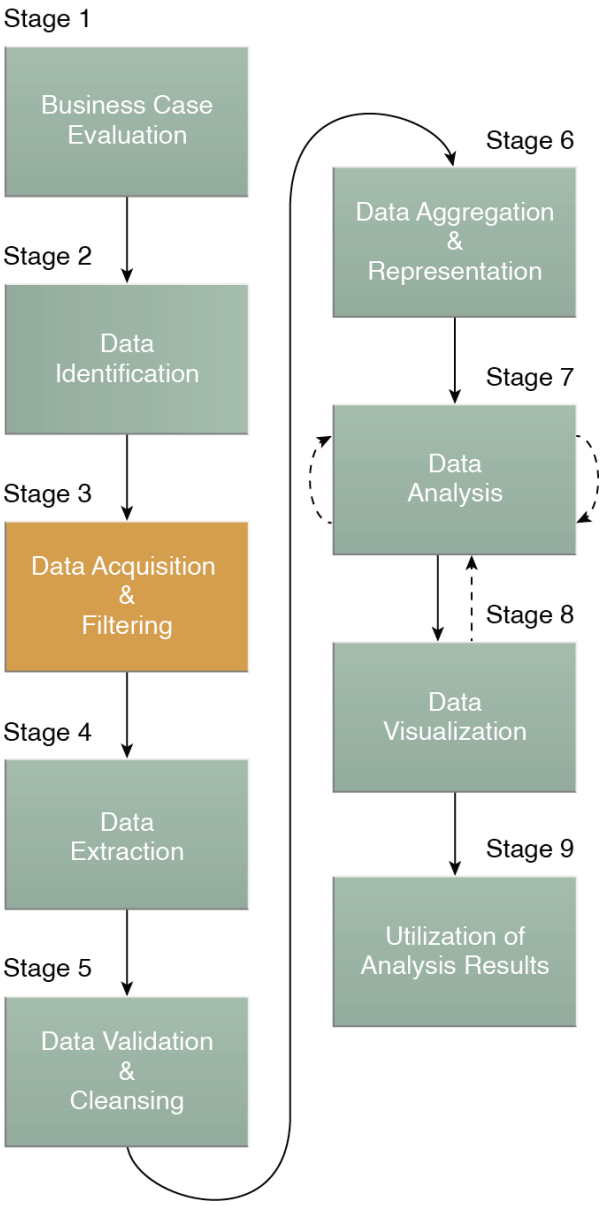
Stage 3: Data Acquisition & Filtering

During the Data Acquisition and Filtering stage, the data is gathered from all of the data sources that were identified during the previous stage, and is then subjected to the automated filtering of corrupt data or data that has been deemed to have no value to the analysis objectives.

Depending on the type of data source, data may come as a dump of files, such as data purchased from a third-party data provider, or may require API integration, such as with Twitter. In many cases, especially where external, unstructured data is concerned, some or most of the acquired data may be irrelevant (noise) and can be discarded as part of the filtering process.

Data classified as "corrupt" can include records with missing or nonsensical values or invalid data types. Data that is filtered out for one analysis may possibly be valuable for a different type of analysis. Therefore, it is advisable to store a verbatim copy of the original dataset before proceeding with the filtering. To save on required storage space, the verbatim copy is compressed before storage.

Both internal and external data needs to be persisted once it gets generated or enters the enterprise boundary. For batch analytics, this data is persisted to disk prior to analysis. In the case of realtime analytics, the data is analyzed first and then persisted to disk.



Metadata can be added via automation to data from both internal and external data sources, as shown in Figure 2.2, to improve the classification and querying. Examples of appended metadata can include dataset size and structure, source information, date and time of creation or collection, and language-specific information. It is vital that metadata be machine-readable and passed forward along subsequent analysis stages. This helps maintain data provenance throughout the Big Data analysis lifecycle, which helps establish and preserve data accuracy and quality.

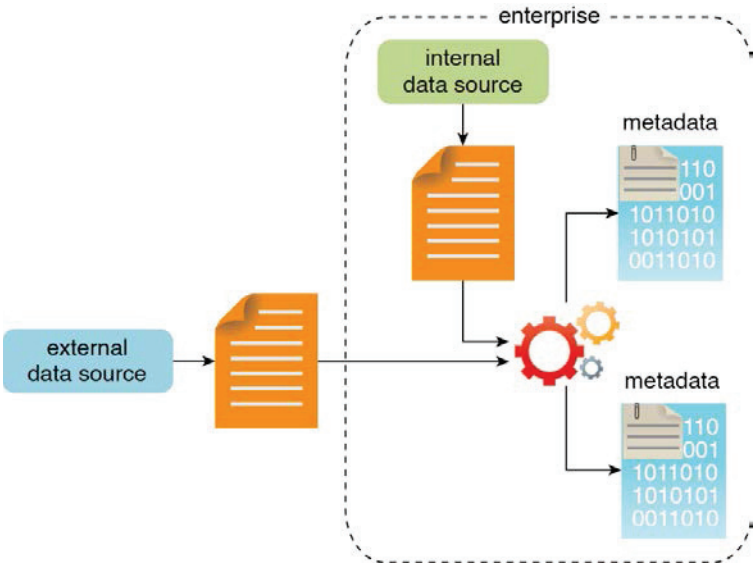


Figure 2.2 – Metadata is added to data from internal and external sources.

Stage 4: Data Extraction

Some of the data identified as input for the analysis may arrive in a format incompatible with the Big Data solution. The need to address disparate types of data is more likely with data from external sources. The Data Extraction lifecycle stage is dedicated to extracting disparate data and transforming it into a format that the underlying Big Data solution can use for the purpose of the data analysis.

The extent of extraction and transformation required depends on the types of analytics and capabilities of the Big Data solution. For example, extracting the required fields from delimited textual data, such as with Web server log files, may not be necessary if the underlying Big Data solution can already directly process those files.

Similarly, extracting text for text analytics, which requires scans of whole documents, will not be necessary if the underlying Big Data solution can already read the document in its native format directly.

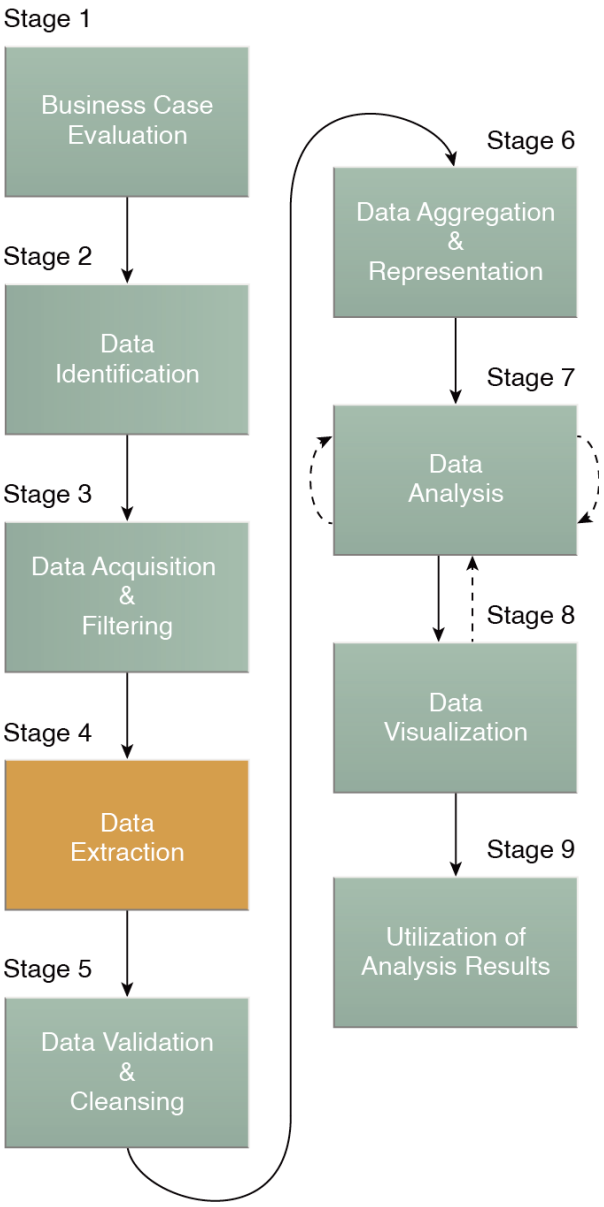


Figure 2.3 illustrates the extraction of comments and a user ID embedded within an XML document without the need for further transformation.

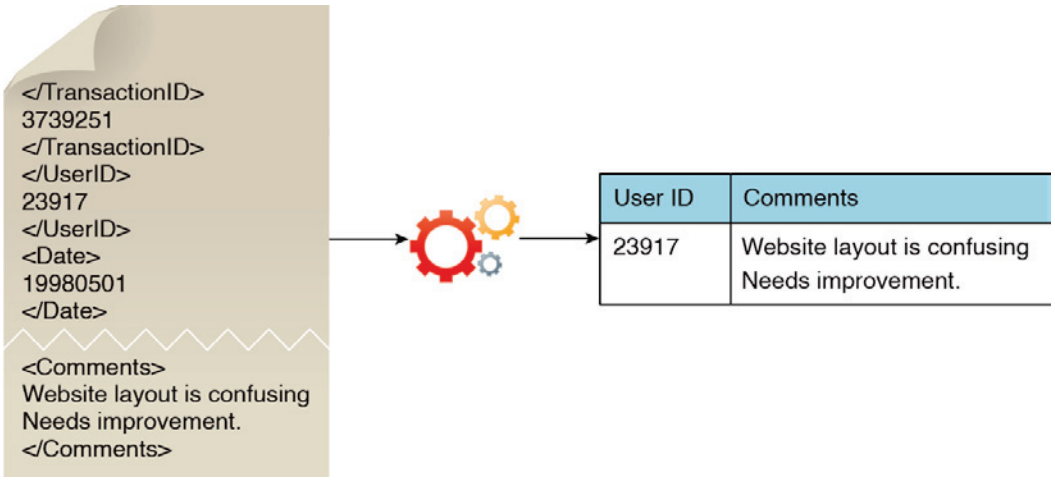


Figure 2.3 – Comments and user IDs are extracted from an XML document.

Figure 2.4 demonstrates the extraction of the latitude and longitude coordinates of a user from a single JSON field.

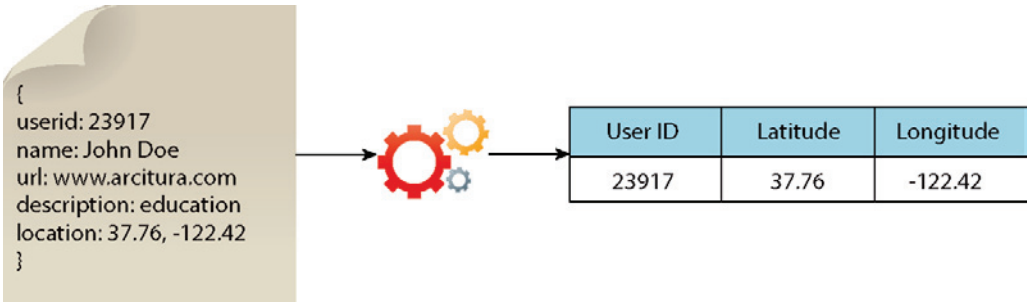


Figure 2.4 – The user ID and coordinators of a user are extracted from a single JSON field.

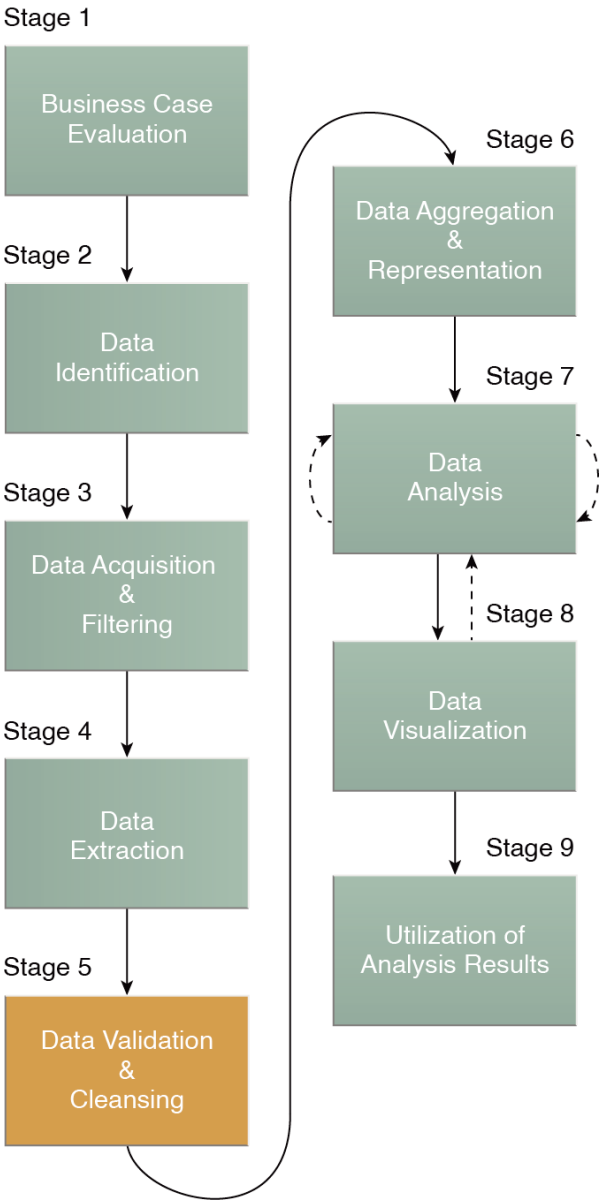
Further transformation is needed in order to separate the data into two separate fields as required by the Big Data solution.

Stage 5: Data Validation & Cleansing

Invalid data can skew and falsify analysis results. Unlike traditional enterprise data, where the data structure is pre-defined and data is pre-validated, data input into Big Data analyses can be unstructured without any indication of validity. Its complexity can further make it difficult to arrive at a set of suitable validation constraints.

The Data Validation and Cleansing stage is dedicated to establishing often complex validation rules and removing any known invalid data.

Big Data solutions often receive redundant data across different datasets. This redundancy can be exploited to explore interconnected datasets in order to assemble validation parameters and fill in missing valid data.



For example, as illustrated in Figure 2.5:

1. The first value in Dataset B is validated against its corresponding value in Dataset A.
2. The second value in Dataset B is not validated against its corresponding value in Dataset A.
3. If a value is missing, it is inserted from Dataset A.

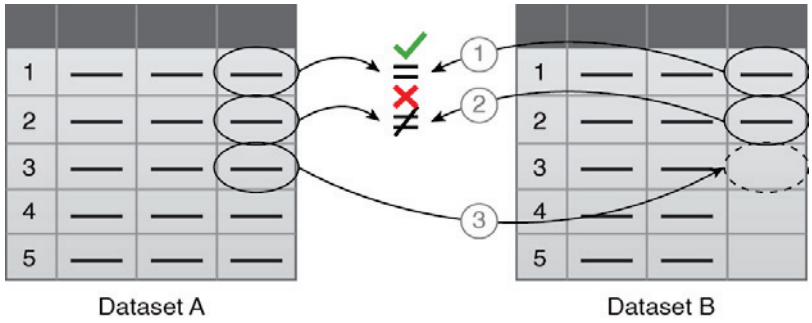


Figure 2.5 – Data validation can be used to examine interconnected datasets in order to fill in missing valid data.

For batch analytics, data validation and cleansing can be achieved via an offline ETL operation. For realtime analytics, a more complex in-memory system is required to validate and cleanse the data at the source. Provenance can play an important role in determining the accuracy and quality of questionable data. Data that appears to be invalid may still be valuable in that it may possess hidden patterns and trends, as shown in Figure 2.6.

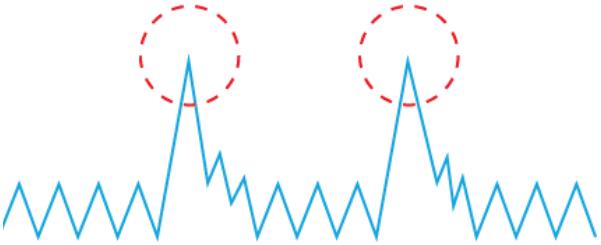


Figure 2.6 – The presence of invalid data is resulting in spikes. Although the data appears abnormal, it may be indicative of a new pattern.

Stage 6: Data Aggregation & Representation

Data may be spread across multiple datasets, requiring that datasets be joined together via common fields, for example date or ID. In other cases, the same data fields may appear in multiple datasets, such as date of birth. Either way, a method of data reconciliation is required or the dataset representing the correct value needs to be determined.

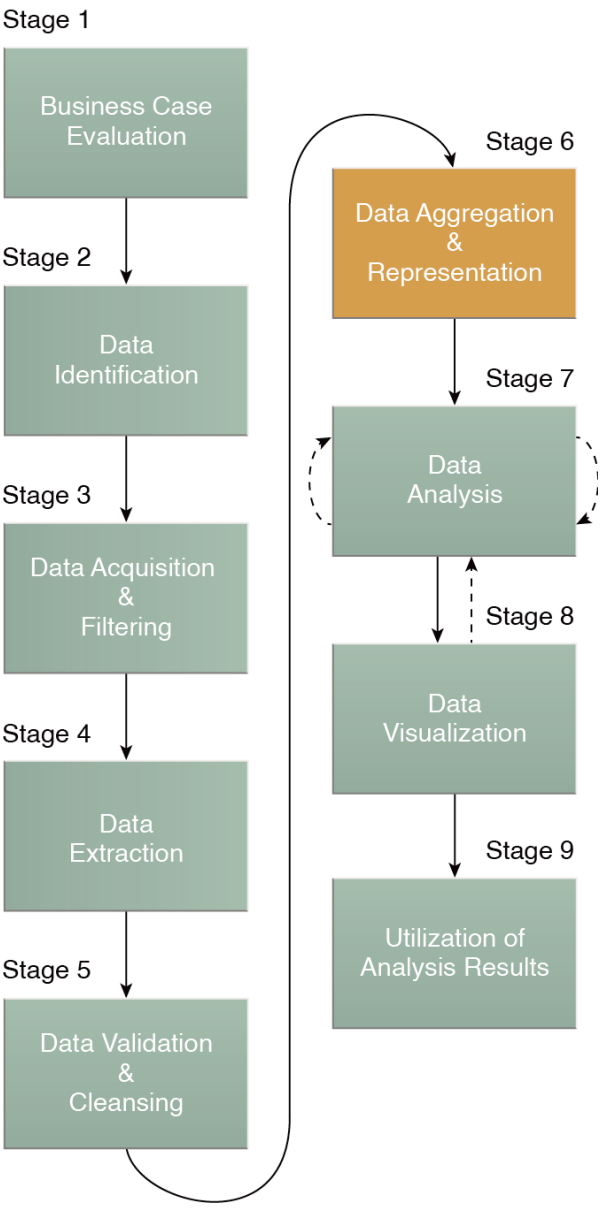
The Data Aggregation & Representation stage is dedicated to integrating multiple datasets together to arrive at a unified view.

Performing this stage can become complicated because of differences in:

- Data Structure – Although the data format may be the same, the data model may be different.
- Semantics – A value that is labeled differently in two different datasets may mean the same thing, for example “surname” and “last name”.

Reconciling these differences can require complex logic that is executed automatically without the need for human intervention. The large volumes processed by Big Data solutions can make data aggregation a time and effort-intensive operation.

Future data analysis requirements need to be considered during this stage to help foster data reusability. Whether data aggregation is required or not, it is important to understand that the same data can be stored in many different forms. One form may be better suited for a particular type of analysis than another. For example, data stored as a BLOB would be of little use if the analysis requires access to individual data fields.



A data structure standardized by the Big Data solution can act as a common denominator that can be used for a range of analysis techniques and projects. This can require establishing a central, standard analysis repository, such as a NoSQL database, as shown Figure 2.7.

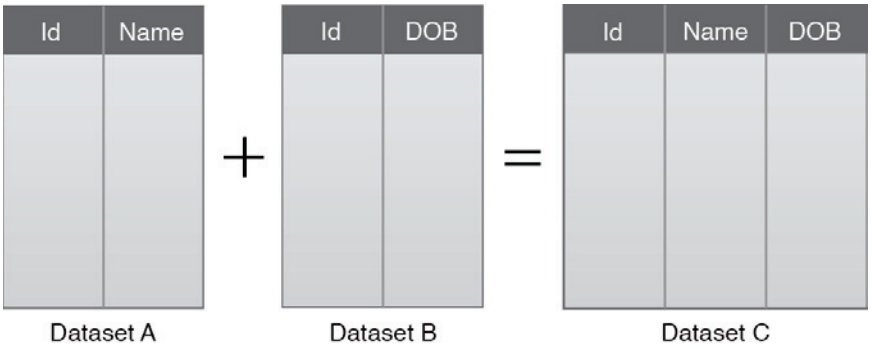


Figure 2.7 – A simple example of data aggregation where two datasets are aggregated together using the Id field.

Figure 2.8 shows the same piece of data stored in two different formats. Dataset A contains the desired piece of data, but it is part of a BLOB that is not readily accessible for querying. Dataset B contains the same piece of data organized in column-based storage, enabling each field to be queried individually.

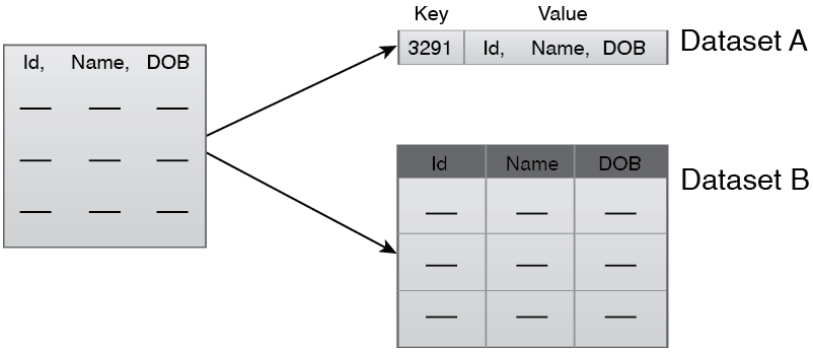
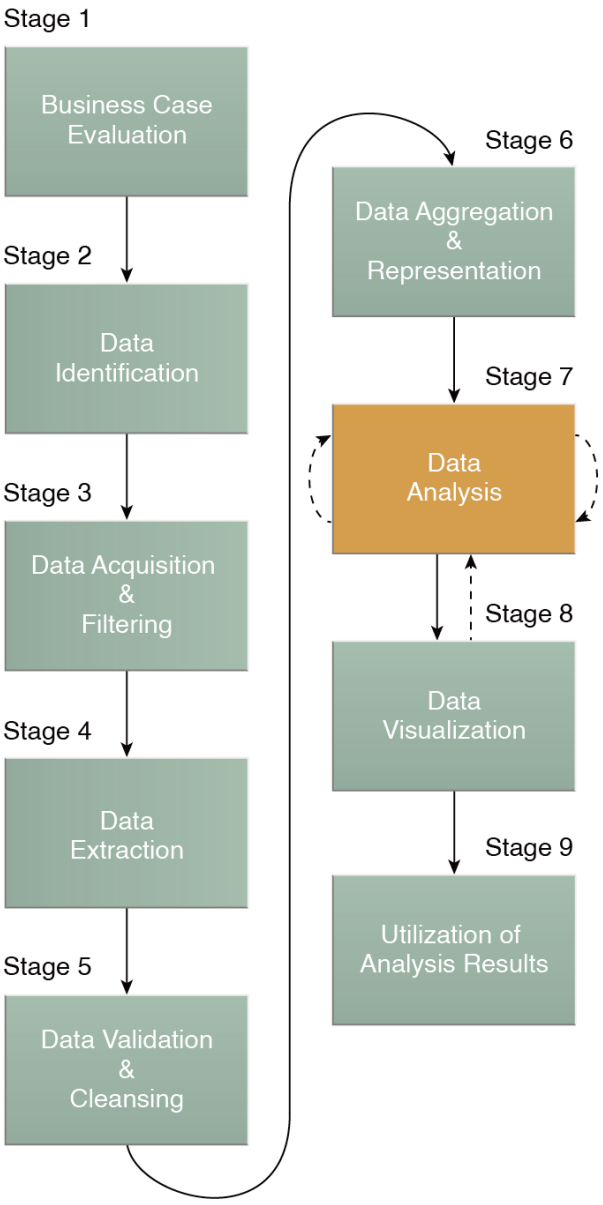


Figure 2.8 – Dataset A and B can be combined to create a standardized data structure with a Big Data solution.

Stage 7: Data Analysis

The Data Analysis stage is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics. This stage can be iterative in nature, especially if the data analysis is exploratory so that analysis is repeated until the appropriate pattern or correlation is uncovered. The exploratory analysis approach will be explained shortly, along with confirmatory analysis.

Depending on the type of analytics required, this stage can be as simple as querying a dataset to compute an aggregation for comparison. On the other hand, it can be as challenging as combining data mining and complex statistical analysis techniques to discover patterns and anomalies or to generate a statistical or mathematical model to depict relationships between variables.



The approach taken when carrying out this stage can be classified as confirmatory analysis or exploratory analysis, the latter of which is linked to data mining, as shown in Figure 2.9.

Confirmatory data analysis is a deductive approach where the cause of the phenomenon being investigated is proposed beforehand. The proposed cause or assumption is called a hypothesis. The data is then analyzed to prove or disprove the hypothesis and provide definitive answers to specific questions. Data samples are typically used. Unexpected findings or anomalies are usually ignored since a predetermined cause was assumed.

Exploratory data analysis is an inductive approach that is closely associated to data mining. No hypothesis or predetermined assumptions are generated. Instead, the data is explored through analysis to develop an understanding of the cause of the phenomenon. Although it may not provide definitive answers, this method provides a general direction that can facilitate the discovery of patterns or anomalies. Large amounts of data and visual analysis are typically used.

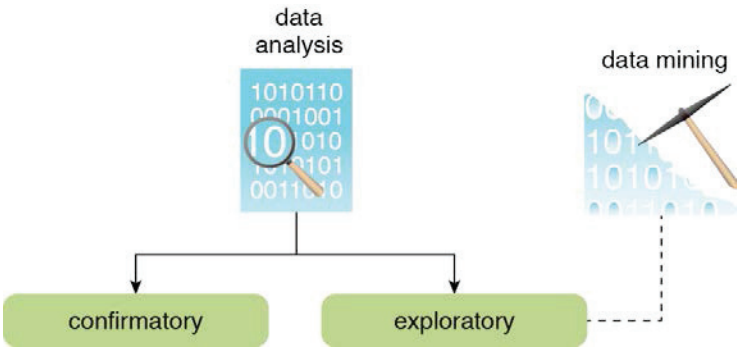


Figure 2.9 – Data analysis can be carried out as confirmatory or exploratory analysis.

Stage 8: Data Visualization

The ability to analyze massive amounts of data and find useful insights carries little value if the only ones that can interpret the results are the analysts.

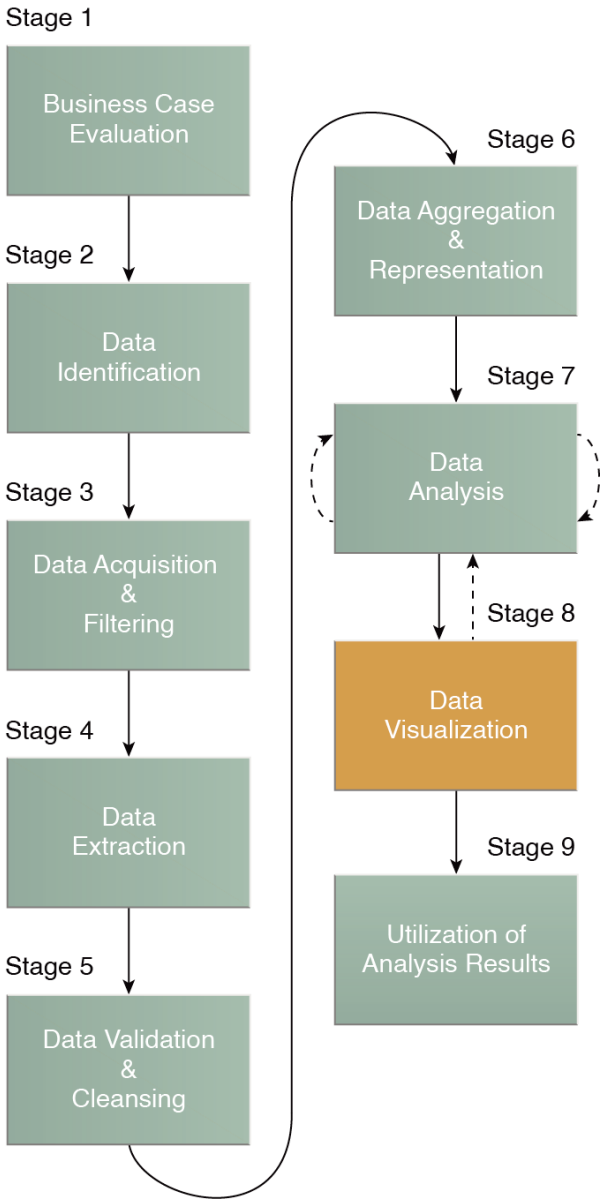
The Data Visualization stage is dedicated to using data visualization techniques and tools to graphically communicate the analysis results for effective interpretation by business users.

Business users need to be able to understand the results in order to obtain value from the analysis and subsequently have the ability to provide feedback, as indicated by the dashed line leading from Stage 8 back to Stage 7.

The results of completing the Data Visualization stage provide users with the ability to perform visual analysis, allowing for the discovery of answers to questions that users have not yet even formulated. Visual analysis is covered later in this workbook.

The same results may be presented in a number of different ways, which can influence the interpretation of the results. Consequently, it is important to use the most suitable visualization technique by keeping the business domain in context.

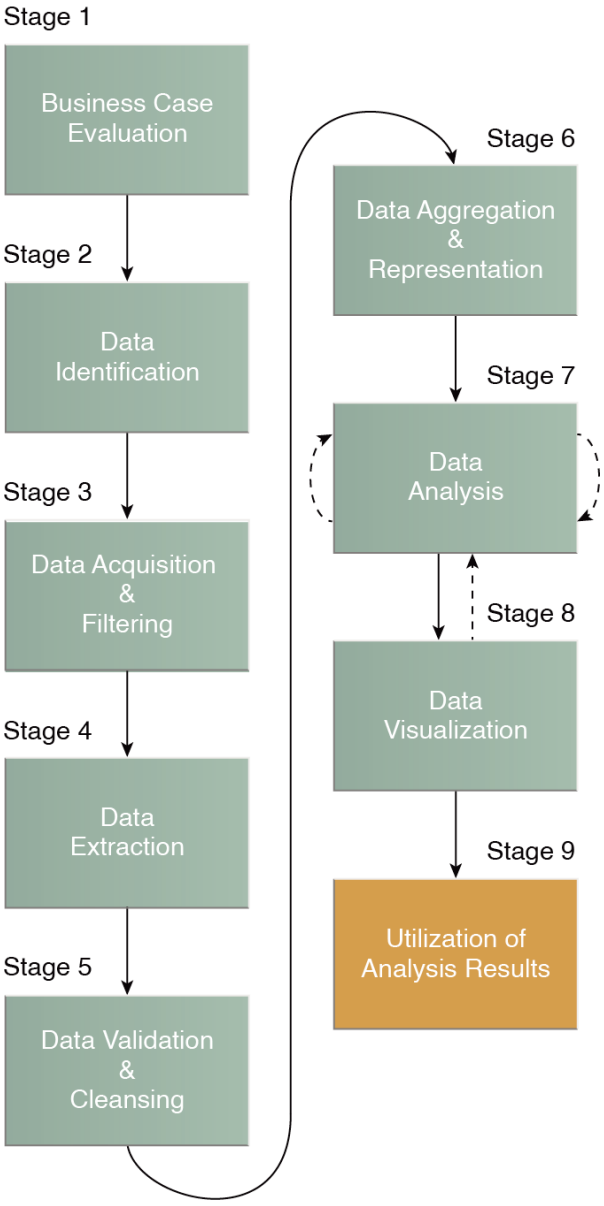
Another aspect to keep in mind is that providing a method of drilling down to comparatively simple statistics is crucial, in order for users to understand how the statistics were generated.



Stage 9: Utilization of Analysis Results

Subsequent to analysis results being made available to business users to support business decision-making, such as via dashboards, there may be further opportunities to utilize the analysis results. The Utilization of Analysis Results stage is dedicated to determining how and where processed analysis data can be further leveraged.

Depending on the nature of the analysis problems being addressed, it is possible for the analysis results to produce "models" that encapsulate new insights and understandings about the nature of the patterns and relationships that exist within the data that was analyzed. A model may look like a mathematical equation or a set of rules. Models can be used to improve business process logic and application system logic, and can form the basis of a new system or software program.



Common areas that are explored during this stage include the following:

Input for Enterprise Systems – The data analysis results may be automatically or manually fed directly into enterprise systems to enhance and optimize their behavior and performance. For example, an online store can be fed processed customer-related analysis results that may impact how it generates product recommendations. New models may be used to improve the programming logic within existing enterprise systems or may form the basis of new systems.

Business Process Optimization – The identified patterns, correlations, and anomalies discovered during the data analysis are used to refine business processes. An example is consolidating transportation routes as part of a supply chain process. Models may also lead to opportunities to improve business process logic.

Alerts – Data analysis results can be used as input for existing alerts or may form the basis of new alerts. For example, alerts may be created to inform users via e-mail or SMS text about an event that requires them to take corrective action.

Optional Reading

The *Big Data Analytics* book that is included with this module discusses the Big Data Analysis Lifecycle further in *Chapter 10: Hands-on Big Data.*

Exercise 2.1: Organize the Lifecycle Stages

Organize the following Big Data analysis lifecycle stages into the correct order:

Data Visualization	1. _____
Data Extraction	2. _____
Data Identification	3. _____
Data Validation & Cleansing	4. _____
Data Acquisition & Filtering	5. _____
Utilization of Analysis Results	6. _____
Data Aggregation & Representation	7. _____
Business Case Evaluation	8. _____
Data Analysis	9. _____

Exercise answers are provided at the end of this booklet.

Notes

Notes / Sketches

Part II: Big Data Analysis Concepts

Each of these analyses will be supplemented in this section:

- Statistical Analysis
- Visual Analysis
- Machine Learning
- Semantic Analysis

These data analysis techniques that can be applied in the Data Analysis lifecycle stage are grouped into the following four primary categories:

Statistical Analysis

- A/B Testing
- Correlation
- Regression

Visual Analysis

- Heat Maps
- Time Series Analysis
- Network Analysis
- Spatial Data Analysis

Machine Learning

- Classification
- Clustering
- Outlier Detection
- Filtering

Semantic Analysis

- Natural Language Processing
- Text Analytics
- Sentiment Analysis

NOTE

With the exception of the machine learning techniques, each of the following analysis techniques is supplemented with a simple example based on the ice cream sales scenario that was introduced in Module 1.

The machine learning techniques introduce a new scenario based on the analysis requirements of a bank.

Statistical Analysis

Statistical analysis uses statistical methods based on mathematical formulas as a means for analyzing data. This type of analysis is commonly used to describe datasets via summarization, such as providing the mean, median, or mode of statistics associated with the dataset. It can also be used to infer patterns and relationships within the dataset, such as regression and correlation.

This section describes the following types of statistical analysis:

- A/B Testing
- Correlation
- Regression

A/B Testing

A/B testing, also known as split or bucket testing, compares two versions of an element to determine which version is superior based on a pre-defined metric. The element can be a range of things. For example, it can be content, such as a Web page, or an offer for a product or service, such as deals on electronic items. The current version of the element is called the control version, whereas the modified version is called the treatment. Both versions are subjected to an experiment simultaneously. The observations are recorded to determine which version is more successful.

For example, in order to determine the best possible layout for an ice cream ad on Company A's Web site, two different versions of the ad are used. Version A is an existing ad (the control) while Version B is slightly altered (the treatment) as compared to Version A's layout. Both versions are then simultaneously shown to different users:

- Version A to Group A
- Version B to Group B

The analysis of the results reveals that Version B of the ad resulted in more sales as compared to Version A. Although A/B testing can be implemented in almost any domain, it is most often used in marketing. Generally, the objective is to gauge human behavior with the goal of increasing sales, as per the upcoming example.

In other areas such as the scientific domains, the objective may simply be to observe which version works better in order to improve a process or product. Figure 2.10 provides an example of A/B testing in two different e-mail versions sent simultaneously.

E-Mail A E-Mail B

Figure 2.10 – Two different e-mail versions are sent out simultaneously as part of a marketing campaign to see which version brings in more prospective customers.

Sample questions can include:

- *Is the new version of a drug better than the old one?*
- *Will this new formula for an anti-dandruff shampoo be more effective than the old formula?*
- *Is the newly designed homepage of the Web site generating more user traffic?*

Correlation

Correlation is an analysis technique used to determine whether two variables are related to each other. If they are found to be related, the next step is to determine what their relationship is. For example, the value of Variable A increases whenever the value of Variable B increases. We may be further interested in discovering how closely Variables A and B are related, which means we may also want to analyze the extent to which Variable B increases in relation to Variable A's increase.

The use of correlation helps to develop an understanding of a dataset and find relationships that can assist in explaining a phenomenon. Correlation is therefore commonly used for data mining where the identification of relationships between variables in a dataset leads to the discovery of patterns and anomalies. This can reveal the nature of the dataset or the cause of a phenomenon.

When two variables are considered to be correlated they are considered to be aligned based on a linear relationship. This means that when one variable changes, the other variable also changes proportionally and constantly.

Correlation is expressed as a decimal number between -1 to +1, which is known as the correlation coefficient. The degree of relationship changes from being strong to weak when moving from -1 to 0 or +1 to 0.

Figure 2.11 shows a correlation of +1, which suggests that there is a strong positive relationship between the two variables.

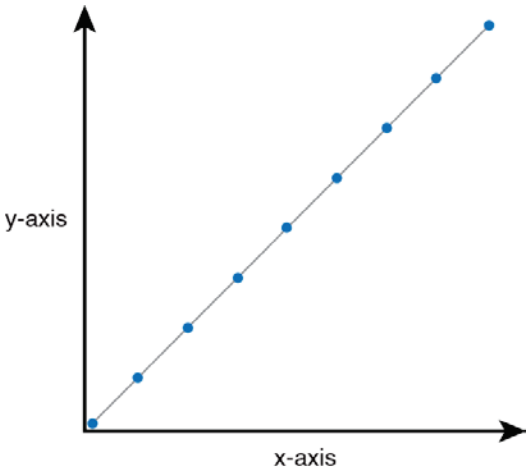


Figure 2.11 – When one variable increases, the other also increases and vice versa.

Figure 2.12 shows a correlation of 0, which suggests that there is no relationship at all between the two variables.

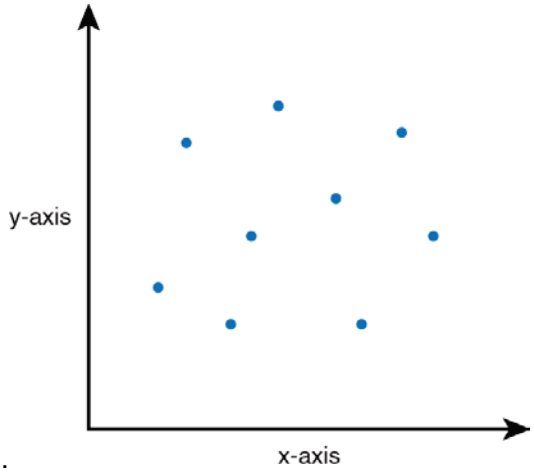


Figure 2.12 – When one variable increases, the other may stay the same, or increase or decrease arbitrarily.

In Figure 2.13, a slope of -1 suggests that there is a strong negative relationship between the two variables.

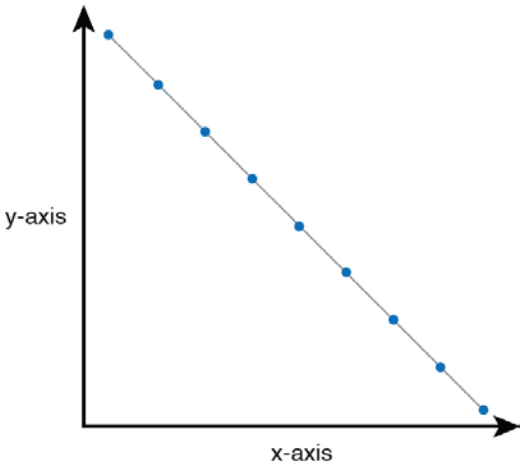


Figure 2.13 – When one variable increases, the other decreases, and vice versa.

For example, managers believe that ice cream stores need to stock more ice cream for hot days, but don't know how much extra to stock. To determine if a relationship actually exists between temperature and ice cream sales, the analysts first apply correlation to the number of ice creams sold and the recorded temperature readings. A value of +0.75 suggests that there exists a strong relationship between the two. This relationship indicates that as temperature increases, more ice creams are sold.

Further sample questions addressed by correlation can include:

- *Does distance from the sea affect the temperature of a city?*
- *Do students who perform well at elementary school perform equally well at high school?*
- *To what extent is obesity linked with overeating?*

Regression

The analysis technique of regression explores how a dependent variable is related to an independent variable within a dataset. As a sample scenario, regression could help determine the type of relationship that exists between temperature, the independent variable, and crop yield, the dependent variable.

Applying this technique helps determine how the value of the dependent variable changes in relation to changes in the value of the independent variable. When the independent variable increases, for example, does the dependent variable also increase? If yes, is the increase in a linear or non-linear proportion?

For example, in order to determine how much extra stock each ice cream store needs to have, the analysts apply regression by feeding in the values of temperature readings. These values are based on the weather forecast as an independent variable and the number of ice creams as

the dependent variable. What the analysts discover is that 15% of additional stock is required for every 5-degree increase in temperature. More than one independent variable can be tested at the same time.

However, in such cases, only one independent variable may change. The others are kept constant. Regression can help enable a better understanding of what a phenomenon is and why it occurred. It can also be used to make predictions about the values of the dependent variable while it is still unknown.

Linear regression represents a constant rate of change, as shown in Figure 2.14.

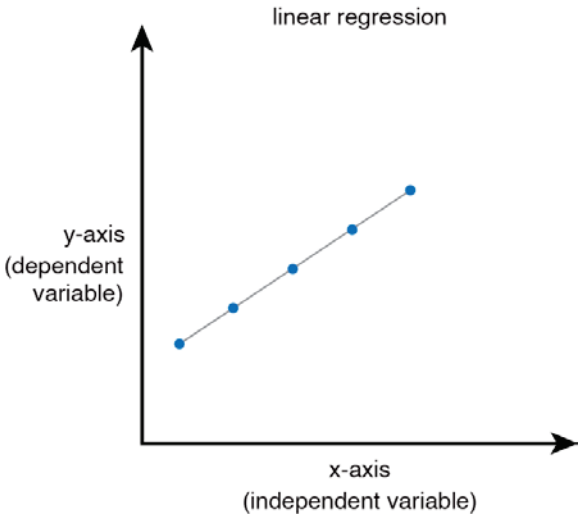


Figure 2.14 – Linear Regression

Non-linear regression represents the variable rate of change, as shown in Figure 2.15.

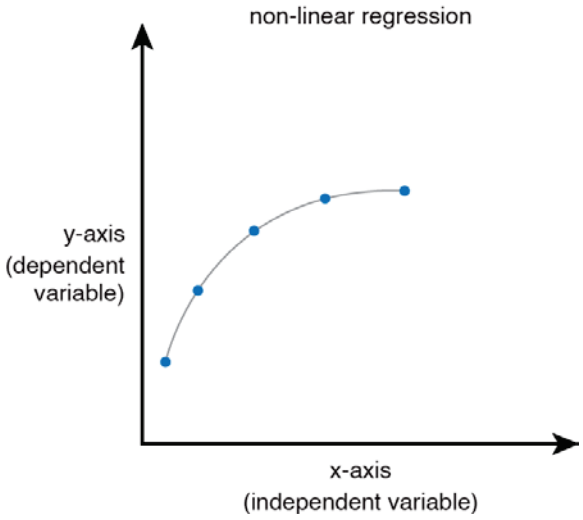


Figure 2.15 – Non-linear Regression

Sample questions can include:

- *What will be the temperature of a city that is 250 miles away from the sea?*
- *What will be the grades of a student studying at a high school based on her primary school grades?*
- *What are the chances that a person will be obese based on the amount of his food intake?*

Regression vs. Correlation

Regression and correlation have a number of important differences. Correlation does not imply causation. The change in the value of one variable may not be responsible for the change in the value of the second variable, although both may change at the same rate. Correlation assumes that both variables are independent.

Regression, on the other hand, deals with already identified dependent and independent variables, and implies that there is a degree of causation between the dependent and independent variables that may be direct or indirect.

Within Big Data, correlation can first be applied to discover if a relationship exists. Regression can then be applied to further explore the relationship and predict the values of the dependent variable, based on the known values of the independent variable.

Notes

Notes / Sketches

Visual Analysis

Visual analysis is a form of data analysis that involves the graphic representation of data to enable or enhance its visual perception. Based on the premise that humans can understand and draw conclusions from graphics more quickly than from text, visual analysis acts as a discovery tool in the field of Big Data.

The objective is to use graphic representations to develop a deeper understanding of the data being analyzed. Specifically, it helps identify and highlight hidden patterns, correlations, and anomalies. Visual analysis is also directly related to exploratory data analysis as it encourages the formulation of questions from different angles.

This section describes the following types of visual analysis:

- Heat Maps
- Time Series Analysis
- Network Analysis
- Spatial Data Analysis

Heat Maps

Heat maps are an effective visual analysis technique for expressing patterns, data compositions via part-whole relations, and geographic distributions of data. They also facilitate the identification of areas of interest and the discovery of extreme (high/low) values within a dataset.

For example, in order to identify the top- and worst-selling regions for ice cream sales, the ice cream sales data is plotted using a heat map. Green is used to highlight the best performing regions, while red is used to highlight worst performing regions.

The heat map itself is a visual, color-coded representation of data values. Each value is given a color according to its type or the range that it falls under. For example, a heat map may assign the values of 0 – 3 to the color red, 4 – 6 to amber, and 7 – 10 to green.

A heat map can be in the form of a chart or a map. A chart represents a matrix of values in which each cell is color-coded according to the value, as shown in Figure 2.16. It can also represent hierarchical values by using color-coded nested rectangles.

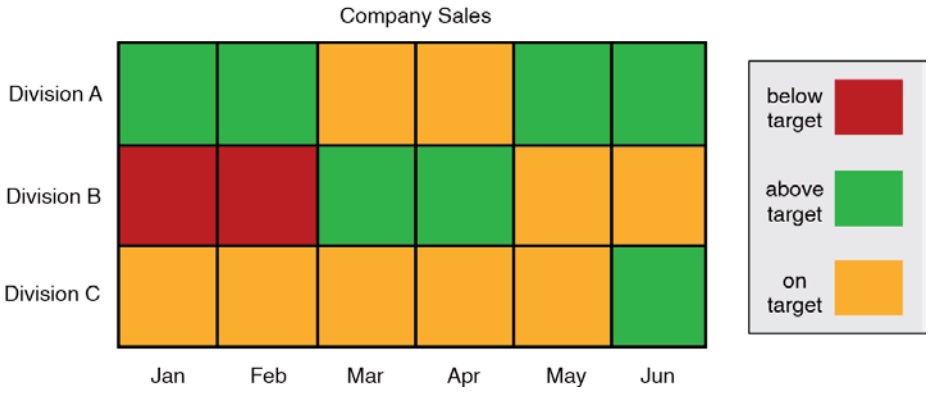


Figure 2.16 – This chart heat map depicts the sales of three divisions within a company over a period of six months.

In Figure 2.17, a map represents a geographic measure by which different regions are color-coded according to a certain theme. Instead of coloring the whole region, the map may be superimposed by a layer made up of collections of colored points relating to various regions, or colored shapes representing various regions.

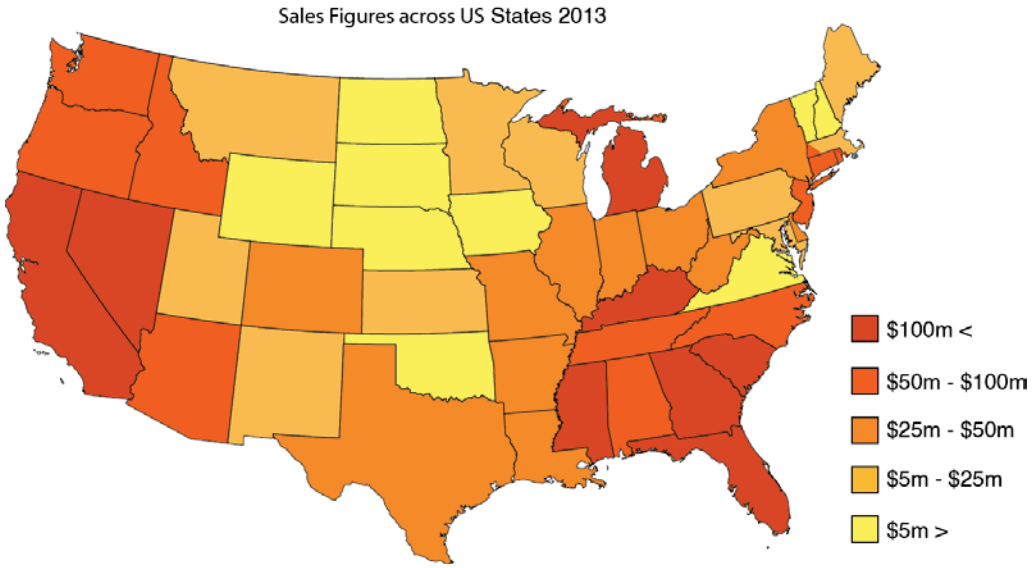


Figure 2.17 – A heat map of the US Sales Figures from 2013

Sample questions can include:

- *How can I visually identify any patterns related to carbon emissions across a large number of cities around the world?*
- *How can I see if there are any patterns of different types of cancers in relation to different ethnicities?*
- *How can I analyze soccer players according to their strengths and weaknesses?*

Time Series Analysis

Time series analysis is the analysis of data that is recorded over periodic intervals of time. This type of analysis makes use of time series, which is a time-ordered collection of values recorded over regular time intervals. An example is a time series that is represented by sales figures that are recorded at the end of each month.

Time series analysis helps to uncover patterns within data that are time-dependent. Once identified, the pattern can be extrapolated for future predictions. For example, to identify seasonal sales patterns, monthly ice cream sales figures are plotted as a time series which further helps to forecast sales figures for the next season.

Time series analyses are usually used for forecasting by identifying long-term trends, seasonal periodic patterns and irregular short-term variations in the dataset. Unlike other types of analyses, time series analysis always includes time as a comparison variable, and the data collected is always time-dependent.

A time series is generally expressed using a line chart, with time plotted on the x-axis and the recorded data value plotted on the y-axis.

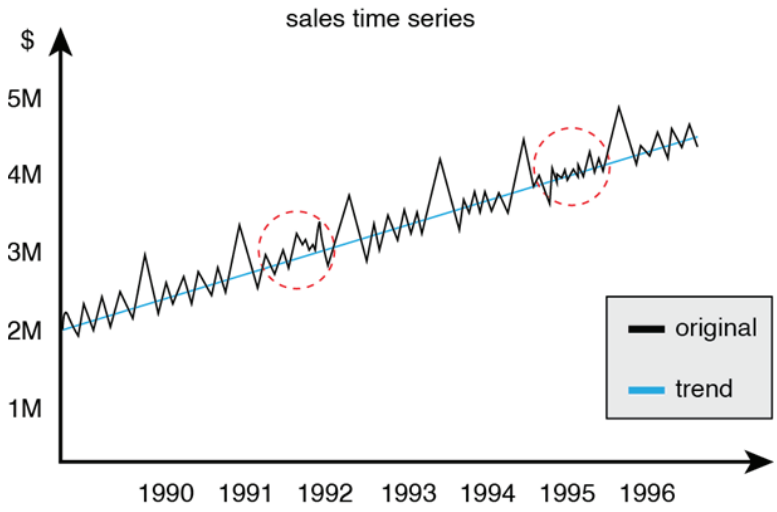


Figure 2.18 – A line chart depicts a sales time series from 1990 to 1996.

The time series presented in Figure 2.18 spans seven years. The evenly spaced peaks towards the end of each year show seasonal periodic patterns, for example Christmas sales. The dotted red circles represent short-term irregular variations. The blue line shows an upward trend, indicating an increase in sales.

Sample questions can include:

- *How much yield should the farmer expect based on historical yield data?*
- *What is the expected increase in population in the next 5 years?*
- *Is the current decrease in sales a one-off occurrence or does it occur regularly?*

Network Analysis

Within the context of visual analysis, a network is an interconnected collection of entities. An entity can be a person, a group, or some other business domain object such as a product. Entities may be connected with one another directly or indirectly. Some connections may only be one-way, so that traversal in the reverse direction is not possible.

Network analysis is a technique that focuses on analyzing relationships between entities within the network. It involves plotting entities as nodes and connections as edges between nodes. There are specialized variations of network analysis, including:

- route optimization
- social network analysis
- spread prediction, such as the spread of a contagious disease

The following is a simple example based on ice cream sales for the application of network analysis for route optimization:

Some ice cream store managers are complaining about the time it takes for delivery trucks to drive between the central warehouse and stores in remote areas. On hotter days, ice creams delivered from the central warehouse to the remote stores melt and can't be sold. Network analysis is used to find the shortest routes between the central warehouse and the remote stores in order to minimize the durations of deliveries.

Consider the social network in Figure 2.19 for a simple example of social network analysis:

- John has many friends, whereas Alice only has one friend.
- The results of a social network analysis reveal that Alice will most likely befriend John and Katie, since they have a common friend named Oliver.

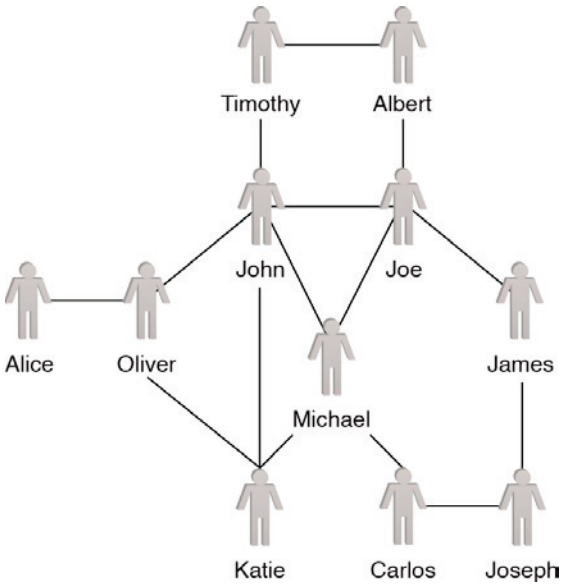


Figure 2.19 – A Social Network

Sample questions can include:

- *How can I identify influencers within a large group of users?*
- *Are two individuals related to each other via a long chain of ancestry?*
- *How can I identify interaction patterns among a very large number of protein-to-protein interactions?*

Spatial Data Analysis

Spatial data analysis is focused on analyzing location-based data in order to find different geographic relationships and patterns between entities. Spatial or geospatial data is commonly used to identify the geographic location of individual entities.

Spatial data is manipulated through a geographical information system (GIS) that plots spatial data on a map generally using its longitude and latitude coordinates. With the ever-increasing availability of location-based data, such as sensor and social media data, spatial data can be analyzed to gain location insights.

For example, as part of a corporate expansion, more ice cream stores are planned to be opened. There is a requirement that no two stores can be within a distance of 5 kilometers of each other to prevent the stores from competing with each other. Spatial data analysis is used to plot existing store locations, and to then identify optimal locations for new stores at least 5 kilometers away from existing stores.

Applications of spatial data analysis include operations and logistic optimization, environmental sciences, and infrastructure planning. Data used as input for spatial data analysis can either contain exact locations, such as longitude and latitude, or the information required to calculate locations, such as zip codes or IP addresses. Spatial data analysis provides analysis features more sophisticated than those of heat maps.

Furthermore, spatial data analysis can be used to determine the number of entities that fall within a certain radius of another entity. For example, a supermarket is using spatial analysis for targeted marketing, as shown in Figure 2.20. Locations are extracted from the users' social media messages. Personalized offers are sent out in realtime based on the proximity of the user.

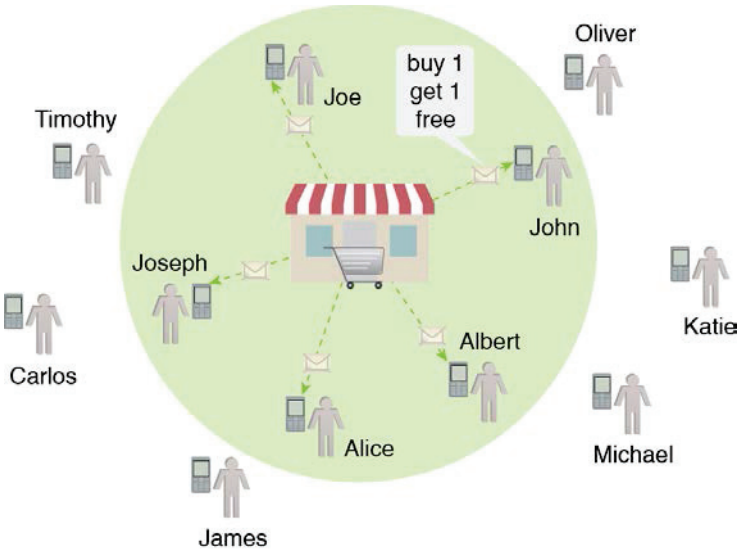


Figure 2.20 – Spatial data analysis can be used for targeted marketing.

Sample questions can include:

- *How many houses will be affected due to a road widening project?*
- *How far do customers have to commute in order to get to a supermarket?*
- *Where are the high and low concentrations of a particular mineral based on readings taken from a number of sample locations within an area?*

Notes

Notes / Sketches

Machine Learning

Humans are good at spotting patterns and relationships within data. Unfortunately, we can't process large amounts of data very fast. Machines, on the other hand, are very adept at processing large amounts of data quickly, but only if they know how.

If human knowledge can be combined with the processing speed of machines, machines will be able to process large amounts of data without requiring much human intervention. This is the basic concept of machine learning. Machine learning and its relationship to data mining were introduced in Module 1. In this section, these topics are explored in further detail through coverage of the following types of machine learning techniques:

- Classification
- Clustering
- Outlier Detection
- Filtering

Before we can proceed to these techniques, we first need to establish the two fundamental laws that pertain to machine learning:

- Law of Large Numbers, applies to Big Data
- Law of Diminishing Marginal Utility, does not apply to Big Data

Law of Large Numbers

The law of large numbers states that the confidence with which predictions can be made increases with the size of the data that is being analyzed. In other words, the accuracy and applicability of the patterns and relationships that are found in a large dataset will be higher than that of a smaller dataset. This means that the greater the amount of data available for analysis, the better we become at making correct decisions.

Law of Diminishing Marginal Utility

In the context of traditional data analysis, the law of diminishing marginal utility states that, starting with a reasonably large sample size, the value obtained from the analysis of additional data decreases as more data is successively added to the original sample. This is a traditional data analysis principle that claims that data held in a reasonably sized dataset provides the maximum value.

The law of diminishing marginal utility does not apply to Big Data. The greater the volume and variety of data that Big Data solutions can process allows for each additional batch of data to carry greater potential of unearthing new patterns and anomalies. Therefore, the value of each additional batch does not diminish value; rather, it provides more value.

Classification

Classification is a supervised learning technique by which data is classified into relevant, previously learned categories. It consists of two steps:

1. The system is fed data that is already categorized or labeled, so that it can develop an understanding of the different categories.
2. The system is fed unknown but similar data for classification, based on the understanding it developed.

A common application of this technique is for the filtering of e-mail spam. Note that classification can be performed for two or more categories. In a simplified classification process, the machine is fed labeled data during training that builds its understanding of the classification, as shown in Figure 2.21. The machine is then fed unlabeled data, which it classifies itself.

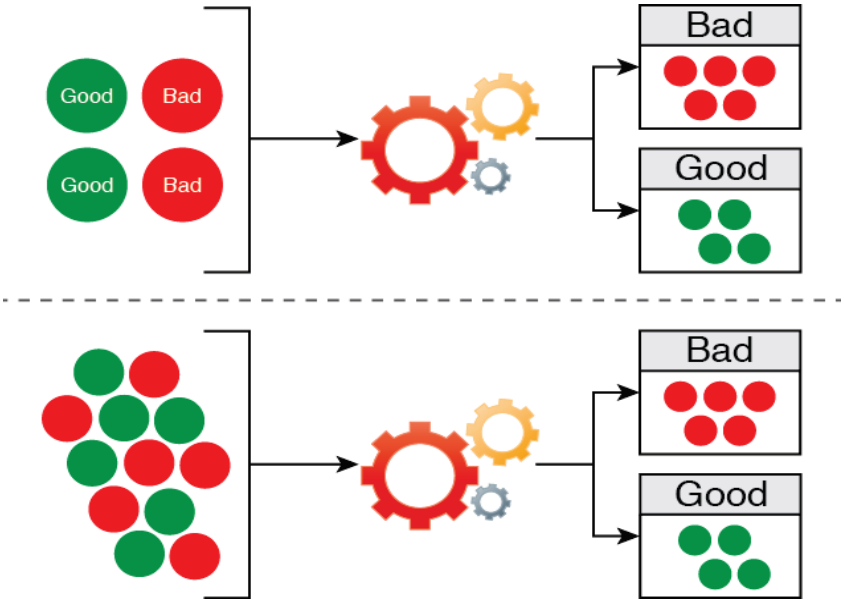


Figure 2.21 – Machine learning can be used to automatically classify datasets.

For example, a bank wants to find out which of its customers is likely to default on loan payments. Based on old data, a training dataset is compiled that contains tagged examples of customers that have or have not previously defaulted. This training data is fed to a classification algorithm that is used to develop an understanding of "good" and "bad" customers. Finally, new untagged customer data is fed in order to find out whether a given customer belongs to the defaulting category.

Sample questions can include:

- *Should an applicant's credit card application be accepted or rejected based on other accepted or rejected applications?*
- *Is tomato a fruit or a vegetable based on the known examples of fruit and vegetables?*
- *Does a fingerprint belong to a suspect based on a record of his previous fingerprints?*

Clustering

Clustering is an unsupervised learning technique by which data is divided into different groups so that the data in each group has similar properties. There is no prior learning of categories required. Instead, categories are implicitly generated based on the data groupings. How the data is grouped depends on the type of algorithm used. Each algorithm uses a different technique to identify clusters.

Clustering is generally used in data mining to get an understanding of the properties of a given dataset. After developing this understanding, classification can be used to make better predictions about similar but new or unseen data.

Clustering can be applied to the categorization of unknown documents and to personalized marketing campaigns by grouping together customers with similar behavior. A scatter graph provides a visual representation of how clustering works in Figure 2.22.

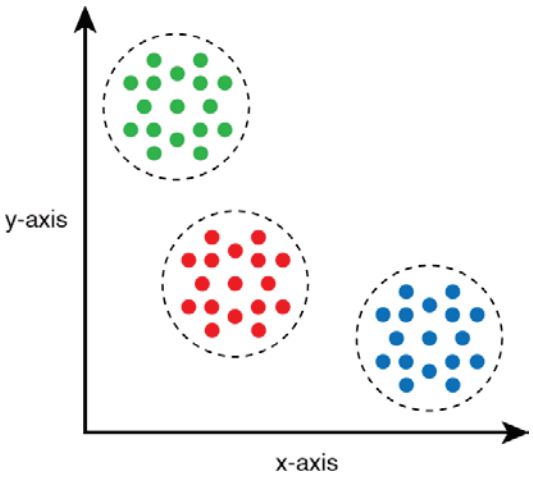


Figure 2.22 – A scatter graph summarizes the results of clustering.

For example, the bank wants to introduce its existing customers to a range of new financial products based on the customer profiles it has on record. The analysts categorize customers into multiple groups using clustering. Each group is then introduced to one or more financial products most suitable to the characteristics of the overall profile of the group.

Sample questions can include:

- *How many different species of trees exist based on the similarity between trees?*
- *How many different categories of elements are there in the periodic table?*
- *What are the different groups of viruses based on their characteristics?*

Outlier Detection

Outlier detection is the process of finding data that is significantly different from or inconsistent with the rest of the data within a given dataset. This machine learning technique is used to identify anomalies, abnormalities, and deviations that can be advantageous, such as opportunities, or disadvantageous, such as risks.

Outlier detection is closely related to the concept of classification and clustering, although its algorithms focus on finding abnormal values. It can be based on either supervised or unsupervised learning. Applications for outlier detection include fraud detection, medical diagnosis, network data analysis, and sensor data analysis. A scatter graph can visually identify outliers, as shown in Figure 2.23.

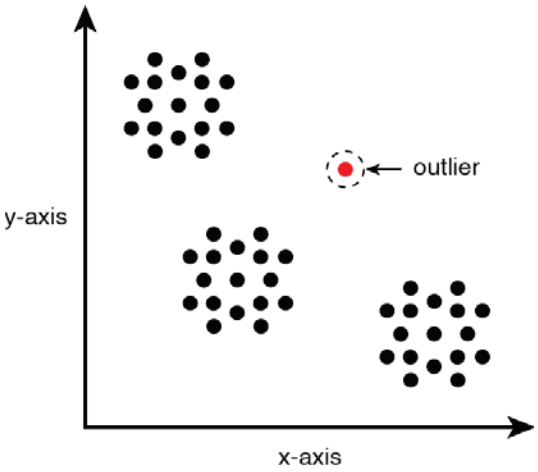


Figure 2.23 – A scatter graph highlights an outlier.

For example, in order to find out whether or not a transaction is likely to be fraudulent, the bank's IT team builds a system employing an outlier detection technique that is based on supervised learning. A set of known fraudulent transactions is first fed into the outlier detection algorithm. After training the system, unknown transactions are then fed into the outlier detection algorithm to predict if they are fraudulent or not.

Sample questions can include:

- *Is a player using performance enhancing drugs?*
- *Are there any wrongly identified fruits and vegetables in the training dataset used for classification task?*
- *Is there a particular strain of virus that does not respond to medication?*

Filtering

Filtering is the automated process of finding relevant items from a pool of items. Items can be filtered either based on a user's own behavior or by matching the behavior of multiple users. Filtering is generally applied via the following two approaches:

- collaborative filtering
- content-based filtering

A common medium by which filtering is implemented is via the use of a recommender system. Collaborative filtering is an item filtering technique based on the collaboration, or merging, of a user's past behavior. A target user's past behavior, including their likes, ratings, purchase history, and more, is collaborated with the behavior of similar users. Based on the similarity of the users' behavior, items are filtered for the target user.

Collaborative filtering is solely based on the similarity between users' behavior, and requires a large amount of user behavior data in order to accurately filter items. It is an example of the application of the law of large numbers.

Content-based filtering is an item filtering technique focused on the similarity between users and items. A user profile is created based on the user's past behavior, for example their likes, ratings, purchase history, etc. The similarities identified between the user profile and the attributes of various items, lead to items being filtered for the user. Contrary to collaborative filtering, content-based filtering is solely dedicated to individual user preferences and does not require data about other users.

A recommender system predicts user preferences and generates suggestions for the user accordingly. Suggestions commonly pertain to recommending items, such as movies, books, Web pages, people, etc. A recommender system typically uses either collaborative filtering or content-based filtering to generate suggestions. It may also be based on a hybrid of both collaborative filtering and content-based filtering to fine-tune the accuracy and effectiveness of generated suggestions.

For example, in order to realize cross-selling opportunities, the bank builds a recommender system that uses content-based filtering. Based on matches found between financial product purchased by customers and the properties of similar financial products, the recommender system automates suggestions for potential financial products that customers may also be interested in.

Sample questions can include:

- *How can only the news articles that a user is interested in be displayed?*
- *Which holiday destinations can be recommended based on the travel history of a holidaymaker?*
- *Which other new users can be suggested as friends based on the current profile of a person?*

Notes

Notes / Sketches

Semantic Analysis

A fragment of text or speech data can carry different meanings in different contexts, whereas a complete sentence may retain its meaning, even if structured in different ways. In order for the machines to extract valuable information, text and speech data needs to be understood by the machines in the same way as humans do. Semantic analysis represents practices for extracting meaningful information from textual and speech data.

This section describes the following types of semantic analysis:

- Natural Language Processing (NLP)
- Text Analytics
- Sentiment Analysis

Natural Language Processing

Natural language processing is a computer's ability to comprehend human speech and text as naturally understood by humans. This allows computers to perform a variety of useful tasks, such as full-text searches.

In order to increase the quality of customer care, the ice cream company employs natural language processing to transcribe customer calls into textual data that are then mined for the most commonly recurring reasons of customer discontent.

Instead of hard-coding the required learning rules, either supervised or unsupervised machine learning is applied to develop the computer's understanding of the natural language. In general, the more learning data the computer has, the more correctly it can decipher human text and speech.

Natural language processing includes both text and speech recognition. For speech recognition, the system attempts to comprehend the speech and then performs an action, such as transcribing text.

Sample questions can include:

- *How can an automated phone exchange system that can recognize the correct department extension as dictated verbally by the caller be developed?*
- *How can grammatical mistakes be automatically identified?*
- *How can a system that can correctly understand different accents of English language be designed?*

Text Analytics

Unstructured text is generally much more difficult to analyze and search in comparison to structured text. Text analytics is the specialized analysis of text through the application of data mining, machine learning, and natural language processing techniques to extract value out of unstructured text. Text analytics essentially provides the ability to discover text rather than just search it.

Useful insights from text-based data can be gained by helping businesses develop an understanding of the information that is contained within a large body of text. As a continuation of the preceding NLP example, the transcribed textual data is further analyzed using text analytics to extract meaningful information about the common reasons behind customer discontent.

The basic tenet of text analytics is to turn unstructured text into data that can be searched and analyzed. As the amount of digitized documents, e-mails, social media posts, and log files increases, businesses have an increasing need to leverage any value that can be extracted from these forms of semi-structured and unstructured data. Solely analyzing operational (structured) data may cause businesses to miss out on cost-saving or business expansion opportunities, especially those that are customer-focused.

Applications include document classification and search, as well as building a 360-degree view of a customer by extracting information from a CRM system.

Text analytics generally involves two steps:

1. Parsing text within documents to extract:
    - Named Entities – person, group, place, company
    - Pattern-Based Entities – social insurance number, zip code
    - Concepts – an abstract representation of an entity
    - Facts – relationship between entities
2. Categorization of documents using these extracted entities and facts.

The extracted information can be used to perform a context-specific search on entities, based on the type of relationship that exists between the entities. Figure 2.24 shows a simplified representation of text analysis.

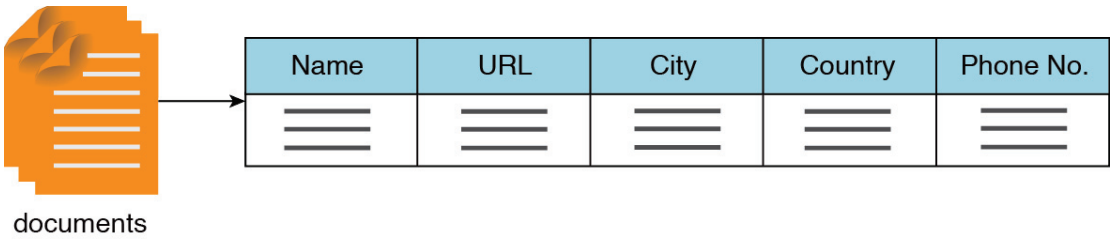


Figure 2.24 – Entities are extracted from text files using semantic rules and structured so that they can be searched.

Sample questions can include:

- *How can I categorize Web sites based on the content of their Web pages?*
- *How can I find the books that contain content which is relevant to the topic that I am studying?*
- *How can I identify contracts that contain confidential company information?*

Sentiment Analysis

Sentiment analysis is a specialized form of text analysis that focuses on determining the bias or emotions of individuals. This form of analysis determines the attitude of the author of the text by analyzing the text within the context of the natural language. Sentiment analysis not only provides information about how individuals feel, but also the intensity of their feeling. This information can then be integrated into the decision-making process. Common applications for sentiment analysis include early identification of customer satisfaction or dissatisfaction, gauging product success or failure, and spotting new trends.

For example, the ice cream company would like to learn about which of its ice cream flavors are most liked by children. Sales data alone does not provide this information because the children that consume the ice cream are not necessarily the purchasers of the ice cream. Sentiment analysis is applied to archived customer feedback left on the ice cream company's Web site to extract information specifically regarding children's preferences for certain ice cream flavors over other flavors.

Sample questions can include:

- *How can customer reactions to the new packaging of the product be gauged?*
- *Which contestant is a likely winner of a singing contest?*
- *Are customers defecting to a competitor?*

Analysis Topic Mapping

The following analysis and analytics topics were covered in Module 1:

- Quantitative Analysis
- Qualitative Analysis
- Data Mining
- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics
- Supervised Learning
- Unsupervised Learning

Most of these analysis-related practices can be applied by, or are in some way related to, several of the preceding analysis techniques. The following briefly describes how these topic areas can be related.

- Quantitative Analysis – Correlation and regression are examples of quantitative analyses. A/B testing can make use of quantitative analysis techniques for results comparison.
- Qualitative Analysis – NLP, text analytics, and sentiment analysis can be used in support of qualitative analysis.
- Data Mining – Data mining can be carried out via the use of or supported by correlation, heat maps, time series analysis, network analysis, spatial data analysis, clustering, outlier detection, natural language processing, and text analytics.
- Descriptive Analytics – A/B testing, heat maps, and spatial data analysis are considered forms of descriptive analytics.
- Diagnostic Analytics – Correlation, regression, time series analysis, network analysis, and spatial data analysis are considered forms of diagnostic analytics.
- Predictive Analytics – Correlation, regression, time series analysis, classification, clustering, outlier detection, filtering, natural language processing, text analytics, and sentiment analysis are considered forms of predictive analytics.
- Prescriptive Analytics – Prescriptive analytics are based on predictive analytics techniques and therefore are associated with the same analysis techniques as predictive analytics. Additionally, prescriptive analytics may utilize heat maps, network analysis, and spatial data analysis to graphically show various outcomes.
- Supervised Learning – Classification, outlier detection, filtering, natural language processing, text analytics, and sentiment analysis can utilize supervised learning.
- Unsupervised Learning – Clustering, outlier detection, filtering, natural language processing, text analytics, and sentiment analysis can utilize unsupervised learning.

Exercise 2.2: Map Problem Statements to Analysis Techniques

Correctly identify which analysis technique can be applied to each of the following problem statements. *Exercise answers are provided at the end of this booklet.*

Text Analytics	Spatial Data Analysis
Time Series Analysis	NLP
Clustering	Sentiment Analysis
Filtering	Correlation
Network Analysis	A/B Testing
Regression	Outlier Detection
Classification	Heat Maps

1. A new startup company is planning to launch a smartphone application that recommends products for its users based on subjective reviews received across multiple social media Web sites and blogs. Which semantic analysis technique can the startup use to interpret textual comments in order to understand why a user makes a product recommendation?

   _____

2. A recruitment agency has acquired two other agencies as part of a corporate expansion. To minimize operating costs, upper management is planning to decrease the number of recruitment agents. However, some managers have indicated that this may result in longer turnaround times, as far as scanning and matching the right candidate is concerned. One of the more technology-aware managers believes that this task can be automated. Which semantic analysis technique can the IT team use to scan a large number of candidate applications automatically?

   _____

3. Jane is a biologist who believes that the rate of photosynthesis is somehow linked to the intensity of light. She has gathered data on light intensity and the rate of photosynthesis. Which statistical analysis technique does Jane need to use to prove or disprove her hypothesis?

_____

4. Jane has proven that the rate of photosynthesis is linked to the intensity of light. She now wants to pinpoint the optimal intensity of light for yielding the highest rate of photosynthesis. Which statistical analysis technique does Jane need to use?

_____

5. An environmental agency wants to measure the extent of damage caused by the spillage of hazardous liquids resulting from accidents involving commercial trucks. The agency has acquired data related to the configuration of interconnected drainage pipes running alongside highways. Which visual analysis technique should be used to analyze the flow of liquid through the drainage pipes?

_____

6. A social media Website provides users with the ability to personalize content for viewing. Users can currently only personalize the content's categories. However, the actual delivered content, such as news or games, is the same for all of the users who have selected a particular category. To further enhance the user online experience, the Web developers want to personalize the content down to the individual user level. Which machine learning technique should be used in this scenario?

_____

7. Roger works as an analyst for a large hotel chain. He has been asked to report on the daily room occupancy levels of each hotel across the country. Roger obtains the data for each hotel on a daily basis. In order to simplify his report, Roger plans to use descriptive values of low, medium, and high to represent the room occupancy level. Which visual analysis technique does Roger need to use to enable management to easily spot the hotels with a low level of room occupancy?

_____

8. Joe works for a law enforcement agency and has collected crime figures for the past 10 years. He identifies a possible pattern while browsing through the data. However, the crime figures in question only date back to a few months and he is unsure of the kind of pattern he has identified. Joe believes that correctly identifying trends can proactively reduce crime. Which visual analysis technique does Joe need to apply?

_____

9. Alice is a medical research student who is currently studying the relationship between the demographics of patients and the patients' diseases. She has gathered several sample known cases that identify salient factors whose presence results in the contracting of certain diseases. Which machine learning technique should Alice use to predict whether a new patient will contract a particular disease?

_____

10. Henry is a botanist who has been collecting data on plant species during his research in the Amazon. Some of these plant species have common characteristics, while others are vastly different. However, Henry cannot identify the common characteristics. Henry's initial attempt at grouping together similar plant species is unsuccessful. Which machine learning technique should Henry use to group similar plant species together?

_____

11. As recent national elections have resulted in rigging allegations made by the opposition party, a commission is established for investigation. The analysts working alongside the commission have access to voting figures across all polling stations. Which machine learning technique should the analysts use to identify the polling stations that produced unusual patterns of casted votes?

_____

12. John has been conducting research on people who suffer from smoke-related diseases. Initial research shows that there is a direct connection between the factories that produce smoke and their distance from the affected people's residences. As part of compiling a report, John wants to determine what can be recommended as a safe distance between factories and residential neighborhoods. Which visual analysis technique should be applied, if the location data on the affected people and factories is available?

_____

13. Adam is a Web developer who is making changes to the homepage of a retail Web site to increase user-friendliness and improve navigation. However, Adam is not sure which layout is most suitable. Which statistical analysis technique does Adam need to use to determine the most suitable layout?

_____

14. A library has recently digitized many of its books to enable its members to access books in digital format over the Internet. However, some of its elderly members have complained that they can't read off of the screen for too long because of eyesight issues. The library decides to make the existing digitalized books available as audio books. Which semantic analysis technique can be used to convert digital books into audio books?

_____

Notes

Notes / Sketches

Part III: Big Data Technology Concepts

This portion of the workbook is divided into the following sections:

- Big Data Technology Considerations
- Big Data Technology Mechanisms

Big Data Technology Considerations

This section introduces the following key technology components and concepts that are relevant to Big Data solutions and mechanisms:

- Clusters
- File Systems & Distributed File Systems
- NoSQL
- Distributed Data Processing
- Parallel Data Processing
- Processing Workloads
- Cloud Computing

Clusters

Within computing, a cluster is a tightly coupled collection of servers, or nodes. These servers usually have the same hardware specifications and are connected together via a network to work as a single unit, as shown in Figure 2.25. Each node in the cluster has its own dedicated resources, such as memory and hard drive, and runs its own operating system just like a desktop computer. A cluster can be used to execute a task based on distributed/parallel data processing frameworks.

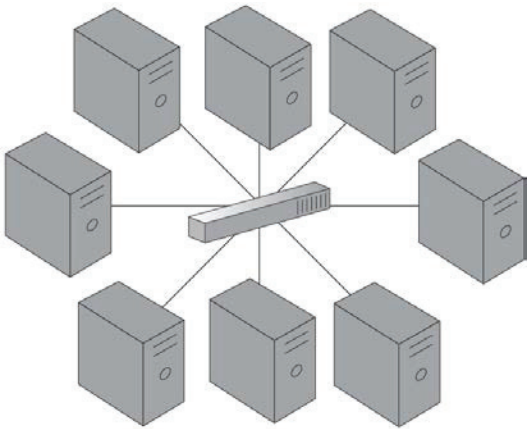
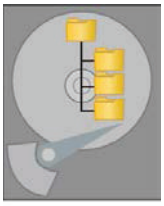


Figure 2.25 – The symbol used to represent a cluster.

File Systems

A file system is a method of storing and organizing data on a storage medium, such as flash drives, DVDs, and hard drives, pictured in Figure 2.26. A file is an atomic unit of storage used by the file system to store data. Files are organized inside of a directory.

A file system provides a logical view of the data stored on the storage medium as a tree structure of files and directories. Operating systems employ file systems for data storage. Each operating system provides support for one or more file systems, like NTFS for Windows and ext for Linux.



hard drive

Figure 2.26 – The symbol used to represent a hard drive.

Distributed File Systems

Within computing, a cluster is a tightly coupled collection of servers or nodes. These servers are connected together via a network to work as a single unit. A distributed file system is a file system that can store large files spread across a cluster, as illustrated in Figure 2.27.

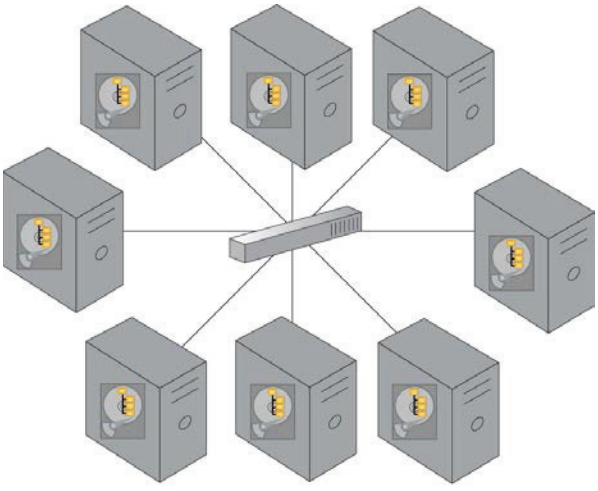


Figure 2.27 – The symbol used to represent distributed file systems.

To the client, a file appears local and can be accessed via multiple locations. Examples include the Google File System (GFS) and Hadoop Distributed File System (HDFS).

NoSQL

A not-only SQL (NoSQL) database is a non-relational database that is highly scalable, fault-tolerant, and specifically designed to house unstructured data. A NoSQL database generally provides an API-based query interface rather than the SQL interface. However, some NoSQL databases may also provide an SQL-like query interface, as shown in Figure 2.28.

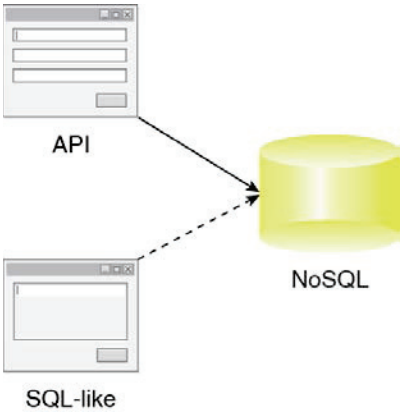


Figure 2.28 – A NoSQL database can provide an API or SQL-like query interface.

Parallel Data Processing

Parallel data processing involves the simultaneous execution of multiple sub-tasks that collectively comprise a larger task. The premise is to reduce the execution time by dividing a single larger task into multiple smaller tasks.

Although parallel data processing can be achieved through multiple networked machines, it is more typically achieved within the confines of a single machine with multiple processors or cores, as shown in Figure 2.29.

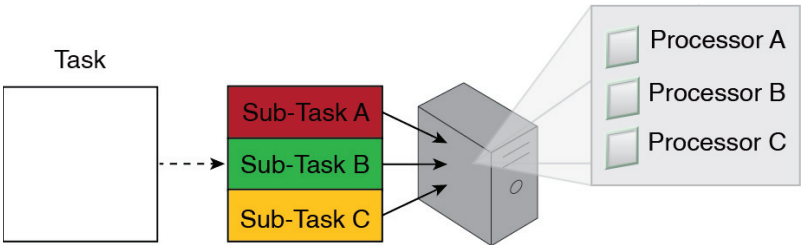


Figure 2.29 – A task can be divided into three sub-tasks that are executed in parallel on three different processors within the same machine.

Distributed Data Processing

Distributed data processing is closely related to parallel data processing in how the same principle of "divide-and-conquer" is applied. However, distributed data processing is always achieved through physically separate machines that are networked together as a cluster. In Figure 2.30, a task is divided into three sub-tasks that are then executed on three different machines sharing one physical switch.

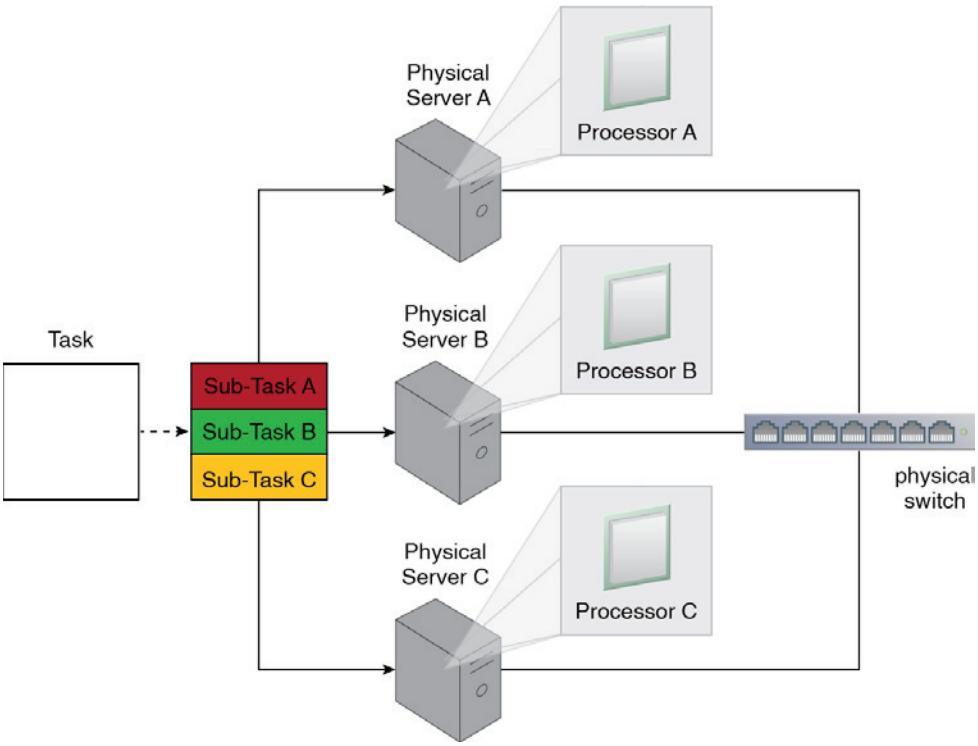


Figure 2.30 – An example of distributed data processing.

Processing Workloads

A processing workload in Big Data is defined as the amount and nature of data that is processed within a certain amount of time. Workloads are usually divided into two types:

- Batch
- Transactional

Processing Workload: Batch

Also known as offline processing, batch workload processing involves processing data in batches and usually imposes delays, resulting in high-latency responses. Batch workloads typically involve large quantities of data with sequential read/writes, and comprise groups of read or write queries.

Queries can be complex and involve multiple joins. OLAP systems commonly process workloads in batches. Strategic BI and analytics fall in this category as they are highly read-intensive tasks involving large volumes of data. As shown in Figure 2.31, a batch workload comprises grouped read/writes with a larger data footprint consisting of complex joins and high-latency responses.

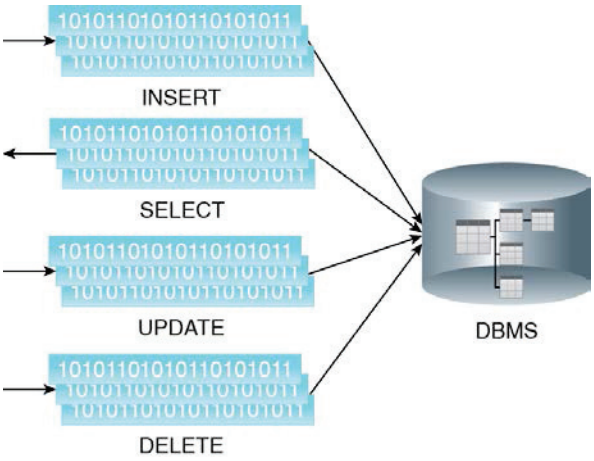


Figure 2.31 – A batch workload can include grouped read/writes to INSERT, SELECT, UPDATE, and DELETE.

Processing Workload: Transactional

Also known as online processing, transactional workload processing follows an approach whereby data is processed interactively without delay, resulting in low-latency responses. Transaction workloads involve small amounts of data with random read/writes.

OLTP and operational systems, which are write-intensive, as well as operational BI and analytics, which are read-intensive, both fall within this category. Although these workloads contain a mix of read/write queries, they are generally more write-intensive than read-intensive.

Transactional workloads comprise random read/writes that involve fewer joins and require low-latency responses with a smaller data footprint, as shown in Figure 2.32.

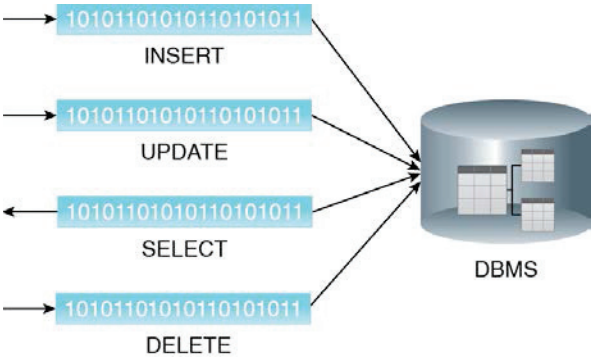


Figure 2.32 – Transactional Workloads have few joins and have lower latency responses than batch workloads.

Cloud Computing

Cloud computing is a specialized form of distributed computing that introduces utilization models for remotely provisioning scalable and measured IT resources. Big Data solutions can be partially or fully deployed in clouds in order to leverage the storage and computing resources that are available from the cloud provider.

The clustered processing resources required by Big Data solutions can benefit from the highly scalable and elastic IT resources available on cloud-based environments. Hadoop's batch-based data processing fully lends itself to the pay-per-use model of cloud computing, which can reduce operational costs since a typical Hadoop cluster size can range from a few to a few thousand nodes.

It makes sense for enterprises already using cloud computing to reuse the cloud for their Big Data initiatives because:

- IT already possesses the required cloud computing skills
- the input data already exists in the cloud

Migrating to the cloud is logical for enterprises planning to run analytics on datasets that are available via data markets, as most data markets store their data in a cloud such as Amazon S3.

In short, cloud computing provides the three ingredients required for a Big Data solution: input data, computing, and storage.

Notes

Notes / Sketches

Big Data Technology Mechanisms

Big Data solutions require a distributed processing environment that can accommodate large-scale data volumes, velocity, and variety. This type of environment is provided by a platform that is comprised of a set of distributed storage and processing technologies.

Big Data mechanisms represent the primary common components of Big Data solutions, regardless of the open source or vendor products used for implementation. This section explores these mechanisms and how they act as the moving parts within Big Data solutions to provide the features and functions required to carry out the Big Data analysis lifecycle.

The following fundamental Big Data mechanisms are covered:

- Storage Device
- Processing Engine
- Resource Manager
- Data Transfer Engine
- Query Engine
- Analytics Engine
- Workflow Engine
- Coordination Engine

At minimum, any given Big Data solution needs to contain the processing engine, storage device, and resource manager mechanisms in order to effectively process large datasets in support of the Big Data analysis lifecycle.

Examples of Big Data Mechanisms Supplement

The mechanisms covered in this section are intentionally documented as abstract components with common and fundamental feature-sets. The *Examples of Big Data Mechanisms* supplement booklet that accompanies this course provides real-world examples of open source products for each of the mechanisms.

Each product implementation of a mechanism will be distinct and may include proprietary features and technologies. As explained in the supplement, the Hadoop open source framework that was briefly introduced in Module 1 establishes a solution environment comprised of multiple mechanisms.

Storage Device

Storage devices provide the underlying data storage environment for persisting the datasets that are processed by Big Data solutions. A storage device can exist as a distributed file system or a database.

Figure 2.33 - The symbol used to represent the storage device.

Distributed file systems can be used for persisting immutable data that is intended for streaming access or batch processing. Databases, such as NoSQL repositories, can be used for structured and unstructured storage and read/write data access, as shown in Figure 2.34. Note that distributed file systems and databases are both on-disk storage devices.

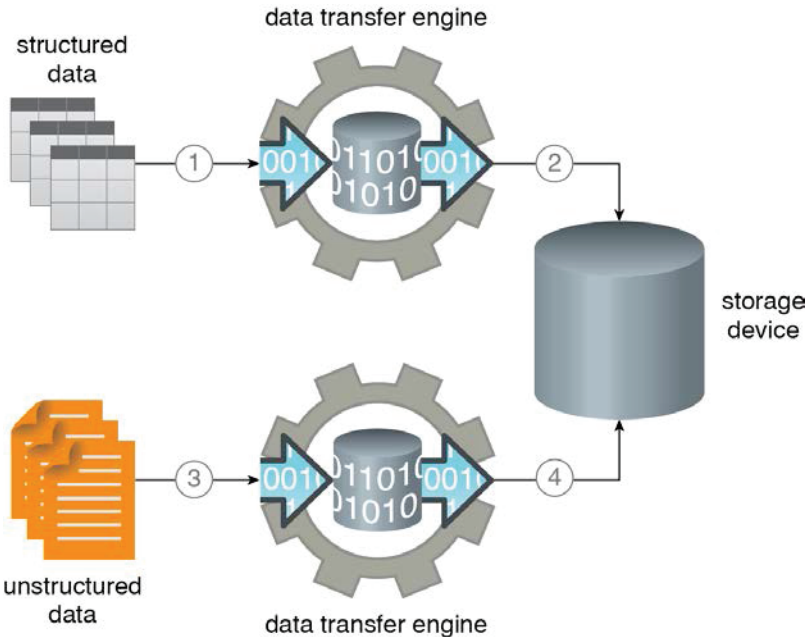


Figure 2.34 – Structured data is imported into a storage device (1) using a data transfer engine (2). Unstructured data is imported (3) using another type of data transfer engine (4).

Processing Engine

The processing engine is responsible for processing data, usually retrieved from storage devices, based on pre-defined logic, in order to produce a result. Any data processing that is requested by the Big Data solution is fulfilled by the processing engine.

Figure 2.35 - The symbol used to represent the processing engine.

A Big Data processing engine utilizes a distributed parallel programming framework that enables it to process very large amounts of data distributed across multiple nodes. It requires processing resources that they request from the resource manager.

Processing engines generally fall into two categories:

- A batch processing engine that provides support for batch data processing, where processing tasks can take anywhere from minutes to hours to complete. This type of processing engine is considered to have high latency.
- A realtime processing engine that provides support for realtime data processing with sub-second response times. This type of processing engine is considered to have low latency.

The Big Data solution's processing requirements dictate the type of processing engine that is used. Figure 2.36 provides an example where a processing job is forwarded to a processing engine via the resource manager.

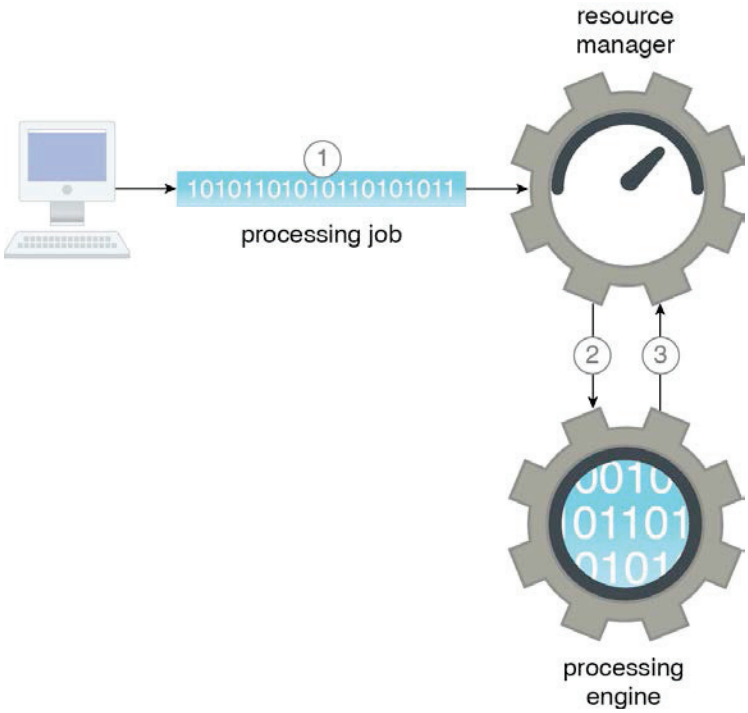


Figure 2.36 – A processing job is submitted to the resource manager (1). The resource manager then allocates an initial set of resources and forwards the job to the processing engine (2), which then requests further resources from the resource manager (3).

Resource Manager

Data that is held in storage can be processed in a variety of ways by a given Big Data solution, and all data processing requests require the allocation of processing resources.

Users of Big Data solutions can make numerous data processing requests, each of which can have different processing workload requirements.

Figure 2.37 - The symbol used to represent the resource manager.

A resource manager acts as a scheduler that schedules and prioritizes processing requests according to individual processing workload requirements. The resource manager essentially acts as a resource arbitrator that manages and allocates available resources, as shown in the example in Figure 2.38.

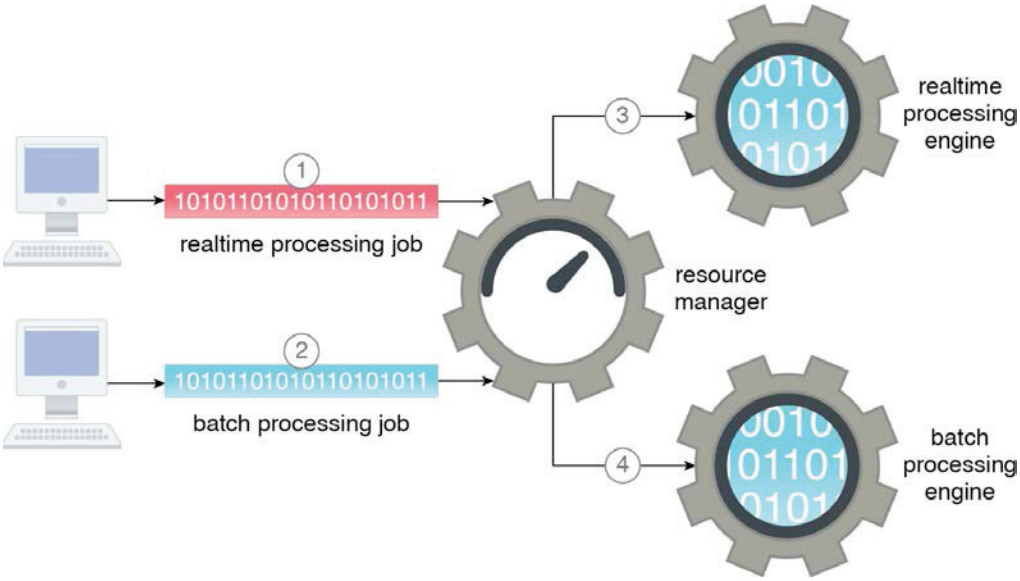


Figure 2.38 – A realtime processing job (1) and a batch processing job (2) are submitted for execution. The resource manager allocates resources according to the job workload requirements and then schedules the jobs on a realtime processing engine (3) and a batch processing engine (4) respectively.

Data Transfer Engine

Data needs to be imported before it can be processed by the Big Data solution. Similarly, processed data may need to be exported to other systems before it can be used outside of the Big Data solution.

Figure 2.39 - The symbol used to represent the data transfer engine.

A data transfer engine enables data to be moved in or out of Big Data solution storage devices. Unlike other data processing systems, where input data conforms to a schema and is mostly structured, data sources for a Big Data solution tend to include a mix of structured and unstructured data.

A given data transfer engine may support either data ingress or egress functions, in which case it can be further qualified as follows:

- data transfer ingress engine
- data transfer egress engine

Data transfer ingress and egress functionality can be further grouped into the following categories:

- event (ingress only)
- file (ingress and egress)
- relational (ingress and egress)

A data transfer engine generally provides only one of the listed functions. It is common for multiple different data transfer engines to be part of a Big Data solution to facilitate a range of import and export requirements for different types of data.

Event-based data transfer ingress engines generally use a publish-subscribe model based on the use of a queue to ensure high reliability and availability. These engines may provide the agent-based processing of inflight data, which enables various data cleansing and transformation activities to be performed in realtime.

Data transfer engines enable the substitution of data that is distributed across a range of sources residing in multiple systems outside of the Big Data solution. A data transfer engine may internally use a processing engine to process multiple large datasets in parallel. This allows large amounts of data to be imported or exported within a short period of time. A workflow engine may provide integration with a data transfer engine to enable the automated import and export of data. Figure 2.40 provides an example of a data transfer engine.

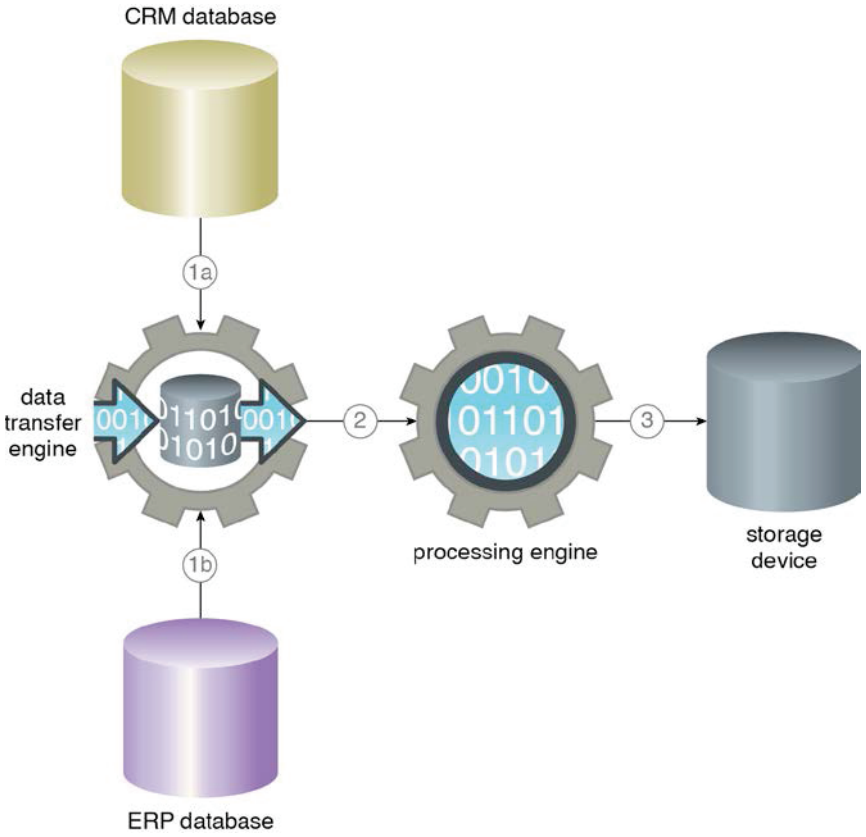


Figure 2.40 – A data transfer engine imports data from two different databases (1a,1b). However, the actual import jobs are run by the processing engine (2), which executes the import jobs and then persists the imported data to the storage device (3).

Query Engine

The processing engine enables data to be queried and manipulated in other ways, but to implement this type of functionality requires custom programming. Analysts working with Big Data solutions are not expected to know how to program processing engines.

A query engine abstracts the processing engine from end-users by providing a front-end user interface that can be used to query underlying data, along with features for creating query execution plans.

Figure 2.41 - The symbol used to represent the query engine.

Languages that are more familiar and easier to work with, such as SQL, can be used by non-technical users to perform ETL tasks and run ad-hoc queries for data analysis activities.

Common processing functions performed by a query engine include sum, average, group by, join, and sort.

Under the hood, the query engine seamlessly transforms user queries into the relevant low-level code that can be used by the processing engine. The use of query engines can reduce development time and enable the manipulation of large datasets without the need to write complex programming logic. Figure 2.42 provides an example of a query engine.

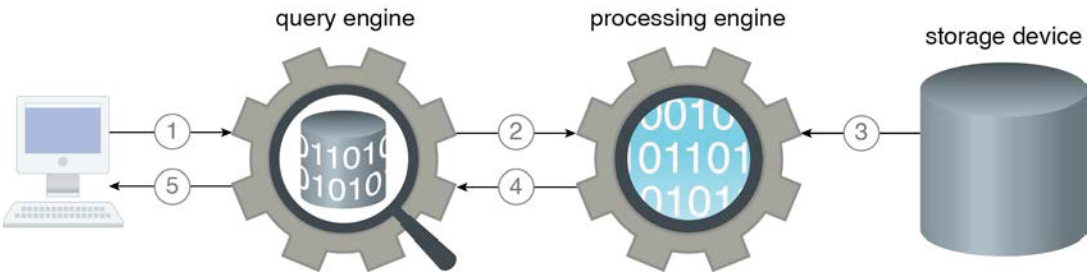


Figure 2.42 – A client performs a simple aggregation query on the data persisted in the storage device (1). The query engine creates a query execution plan and creates jobs that need to be executed on the processing engine (2). The processing engine retrieves the required data from the storage device (3) and then executes the required jobs. The results are then forwarded to the query engine (4), which sends the results back to the client after further processing (5).

Analytics Engine

An analytics engine is able to process advanced statistical and machine learning algorithms in support of analytics processing requirements, including the identification of patterns and correlations. It generally uses a processing engine to run algorithms on large datasets.

Figure 2.43 - The symbol used to represent the analytics engine.

An analytics engine is employed when the comparatively simple data manipulation functions of a query engine are insufficient. Some proprietary analytics engines also provide specialized data analysis features, such as text analytics and machine log analysis processing. Figure 2.43 provides an example of an analytics engine.

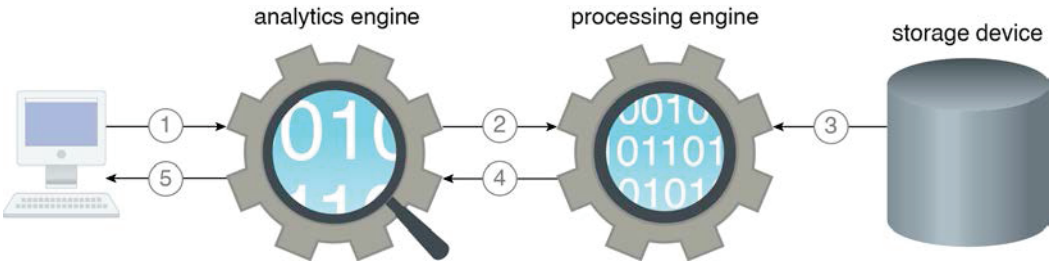


Figure 2.43 – A client performs an advanced statistical operation on the data persisted in the storage device (1). The analytics engine creates jobs that need to be executed on the processing engine (2). The processing engine gets the required data from the storage device (3) and then executes the required jobs. The results are then forwarded to the analytics engine (4) which sends the results back to the client after further processing.

Workflow Engine

The ability to query data and perform ETL operations via the query engine is useful for ad-hoc data analysis. However, performing the same set of operations in a particular order repeatedly is often required in order to obtain up-to-date results based on the latest data. A workflow engine provides the ability to design and process a complex sequence of operations that can be triggered either at set time intervals or when data becomes available.

Figure 2.44 - The symbol used to represent the workflow engine.

The workflow logic processed by a workflow engine can involve the participation of other Big Data mechanisms, as shown in Figure 2.45. For example, a workflow engine can execute logic that collects relational data from multiple databases at regular intervals via the data transfer engine, applies a set of ETL operations via the processing engine, and finally persists the results to a NoSQL storage device.

The defined workflows are analogous to a flowchart with control logic, such as decisions, forks, joins, and generally rely on a batch-style processing engine for execution. The output of one workflow can become the input of another workflow.

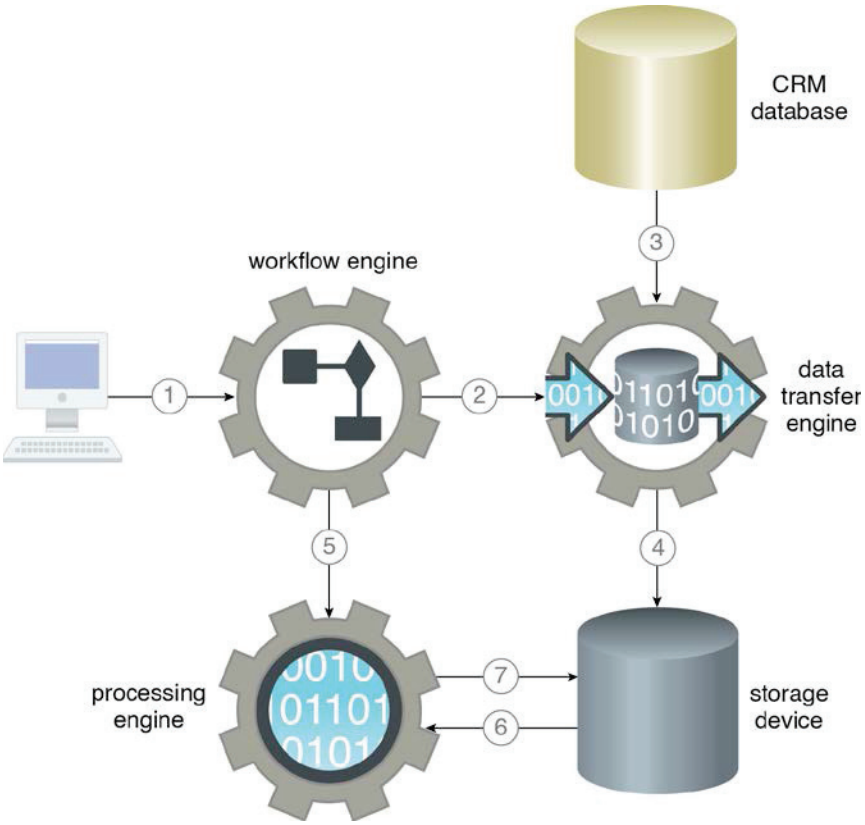


Figure 2.45 – A client first creates a workflow job using the workflow engine (1). As the first step of the configuration job, the workflow engine triggers a data ingress job (2), which is executed by the data transfer engine in the form of a data import from a CRM database (3). The imported data is then persisted in the storage device (4). As part of the second step of the configuration job, the workflow engine then triggers the processing engine for the execution of a data processing job (5). In response, the processing engine retrieves the required data from the storage device (6), executes the data processing job, and then persists the results back to the storage device (7).

Coordination Engine

A distributed Big Data solution that needs to run on multiple servers relies on the coordination engine to ensure operational consistency across all of the participating servers. Coordination engines make it possible to develop highly reliable, highly available distributed Big Data solutions that can be deployed in a cluster.

Figure 2.46 - The symbol used to represent the coordination engine.

The processing engine will often use the coordination engine to coordinate data processing across a large number of servers. This way, the processing engine does not require its own coordination logic.

The coordination engine can also be used for the following purposes, as shown in Figure 2.47:

- to support distributed locks
- to support distributed queues
- to establish a highly available registry for obtaining configuration information
- for reliable asynchronous communication between processes that are running on different servers

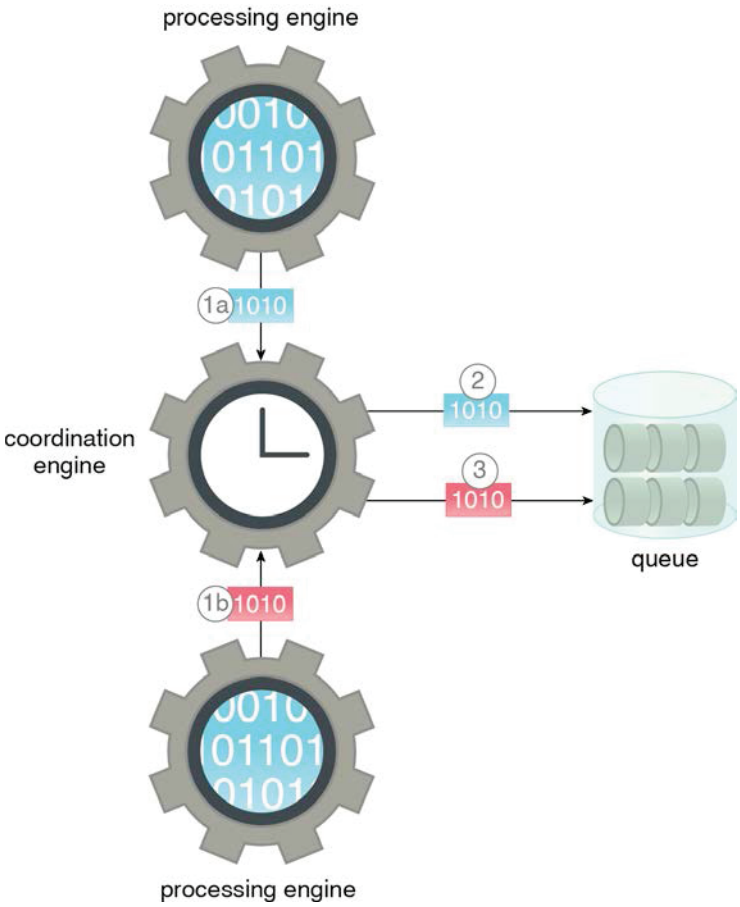


Figure 2.47 – Two nodes in a cluster need to write to a shared queue as part of executing a job, and both send a write request at the same time (1a, 1b). The write request is coordinated by the coordination engine. One request is sent to the queue (2) before the other request is sent in a serialized manner (3).

Exercise 2.3: Fill in the Blanks

1. The _____ is limited in how it can usually only perform simple data operations, such as average, join, and sort. The _____ is more advanced and can perform difficult machine learning and statistical algorithms.

2. The functionalities of data transfer _____ engines and data transfer _____ engines can be categorized into event, file, and relational functionalities.

3. Batch processing engines are _____, meaning they can take up to a few hours to finish a data processing task. _____, on the other hand, provides sub-second response times.

4. The _____ mechanism can be used to process complex sequences of operations at set times or when data becomes available.

5. A _____ schedules data processing requests and acts as a resource arbitrator that allocates resources to different applications. This mechanism, along with the _____ and _____ mechanisms, is mandatory in Big Data platforms to achieve interoperability between large datasets.

6. _____ and _____ are examples of the storage device mechanism.

7. The _____ mechanism is used to maintain operational consistency across multiple _____ in a distributed Big Data

solution. This mechanism is typically used by the _____ mechanism to evenly distribute data processing as well.

Exercise answers are provided at the end of this booklet.

Notes

Notes / Sketches

Exercise Answers

Exercise 2.1 Answers

1. Business Case Evaluation
2. Data Identification
3. Data Acquisition & Filtering
4. Data Extraction
5. Data Validation & Cleansing
6. Data Aggregation & Representation
7. Data Analysis
8. Data Visualization
9. Utilization of Analysis Results

Exercise 2.2 Answers

1. Sentiment Analysis
2. Text Analytics
3. Correlation
4. Regression
5. Network Analysis
6. Filtering
7. Time Series Analysis
8. Heat Maps
9. Classification
10. Clustering
11. Outlier Detection
12. Spatial Data Analysis
13. A/B Testing

14. Natural Language Processing (NLP)

Exercise 2.3 Answers

1. The **query engine** is limited in how it can usually only perform simple data operations, such as average, join, and sort. The **analytics engine** is more advanced and can perform difficult machine learning and statistical algorithms.

2. The functionalities of data transfer **ingress** engines and data transfer **egress** engines can be categorized into event, file and relational functionalities.

3. Batch processing engines are **high latency**, meaning they can take up to a few hours to finish a data processing task. **Realtime processing engines**, on the other hand, provides sub-second response times.

4. The **workflow engine** mechanism can be used to process complex sequences of operations at set times or when data becomes available.

5. A **resource manager** schedules data processing requests and acts as a resource arbitrator that allocates resources to different applications. This mechanism, along with the **storage device** and **processing engine** mechanisms, is mandatory in Big Data platforms to achieve interoperability between large datasets.

6. **Distributed file systems** and **databases** are examples of the storage device mechanism.

7. The **coordination engine** mechanism is used to maintain operational consistency across multiple **servers** in a distributed Big Data solution. This mechanism is typically used by the **processing engine** mechanism to evenly distribute data processing as well.

Exam B90.02

The course you just completed corresponds to Exam B90.02, which is an official exam that is part of the Big Data Science Certified Professional (BDSCP) program.

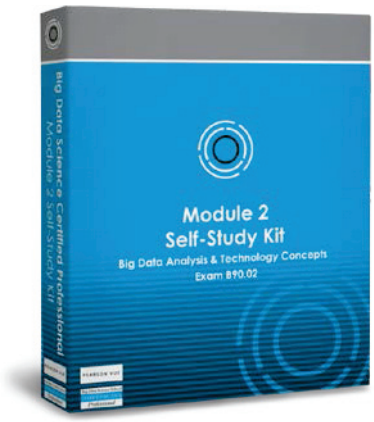
This exam can be taken at Pearson VUE testing centers worldwide or via Pearson VUE Online Proctoring, which enables you to take exams from your home or office workstation with a live proctor. For more information, visit:

www.bigdatascienceschool.com/exams/

www.pearsonvue.com/arcitura/

www.pearsonvue.com/arcitura/op/ (Online Proctoring)

Module 2 Self-Study Kit



An official BDSCP Self-Study Kit is available for this module, providing additional study aids and resources, including a separate self-study guide, Audio Tutor CDs and flash cards.

Note that versions of this self-study kit are available with and without a Pearson VUE exam voucher for Exam B90.02.

For more information, visit:
www.bigdataselfstudy.com

Contact Information and Resources

AITCP Community

Join the growing international Arcitura IT Certified Professional (AITCP) community by connecting on official social media platforms: LinkedIn, Twitter, Facebook, and YouTube.

Social media and community links are accessible at:

- www.arcitura.com/community
- www.servicetechbooks.com/community

General Program Information

For general information about the BDSCP program and Certification requirements, visit:

www.bigdatascienceschool.com and www.bigdatascienceschool.com/matrix/

General Information about Course Modules and Self-Study Kits

For general information about BDSCP Course Modules and Self-Study Kits, visit:

www.bigdatascienceschool.com and www.bigdataselfstudy.com

Pearson VUE Exam Inquiries

For general information about taking BDSCP Exams at Pearson VUE testing centers or via Pearson VUE Online Proctoring, visit:

www.pearsonvue.com/arcitura/
www.pearsonvue.com/arcitura/op/ (Online Proctoring)

Public Instructor-Led Workshop Schedule

For the latest schedule of instructor-led BDSCP workshops open for public registration, visit:

www.bigdatascienceschool.com/workshops

Private Instructor-Led Workshops

Certified trainers can deliver workshops on-site at your location with optional on-site proctored exams. To learn about options and pricing, contact:

info@arcitura.com

or

1-800-579-6582

Becoming a Certified Trainer

If you are interested in attaining the Certified Trainer status for this or any other Arcitura courses or programs, learn more by visiting:

www.arcitura.com/trainerdevelopment/

General BDSCP Inquiries

For any other questions relating to this Course or any Module, Exam, or Certification that is part of the BDSCP program, contact:

info@arcitura.com

or

1-800-579-6582

Automatic Notification

To be automatically notified of changes or updates to the BDSCP program and related resource sites, send a blank message to:

notify@arcitura.com

Feedback and Comments

Help us improve this course. Send your feedback or comments to:

info@arcitura.com