Big Data Analysis Techniques

A Primer for Data Science Terms

ABSTRACT

Big Data analysis blends traditional statistical data analysis approaches with computational ones. This document will serve as an introductory material to understand the various terms and techniques associated with Big Data Analytics.

Suriya Priya Asaithambi

Big Data Engineering for Analytics

Institute of Systems Science
National University of Singapore
25, Heng Mui Keng Terrace, Singapore - 119615

Table of Contents

1.Data Sets

Collections or groups of related data are generally referred to as datasets. Each group or dataset member (datum) shares the same set of attributes or properties as others in the same dataset. Some examples of datasets are:

- ✓ tweets stored in a flat file
- ✓ a collection of image files in a directory
- ✓ an extract of rows from a database table stored in a CSV formatted file
- ✓ historical weather observations that are stored as XML files

2.Data Analysis

Data analysis is the process of examining data to find facts, relationships, patterns, insights and/or trends. The overall goal of data analysis is to support better decision-making. Carrying out data analysis helps establish patterns and relationships among the data being analysed.

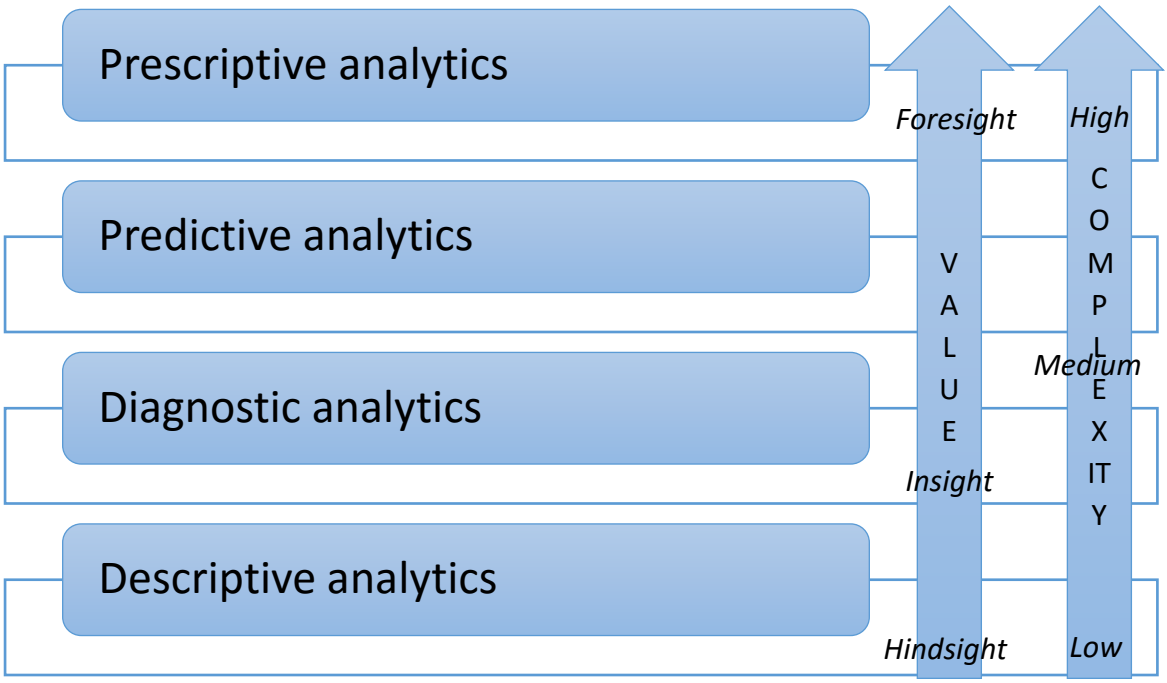
3.Data Analytics

Data analytics is a broader term that encompasses data analysis. Data analytics are a discipline which includes management of the complete data lifecycle which includes collecting, cleansing, organizing, storing, analysing and governing data. The term includes the development of analysis methods, scientific techniques and automated tools.

In Big Data environments, data analytics have developed methods that allow data analysis to occur through the use of highly scalable distributed technologies and frameworks that are capable of analysing large volumes of data from different sources. Data analytics enable data-driven decision-making with scientific backing so that decisions can be based on factual data and not simply on past experience or intuition alone. There are four general categories of analytics that are distinguished by the results they produce:

1. descriptive analytics
2. diagnostic analytics
3. predictive analytics
4. prescriptive analytics

The different analytics types leverage different techniques and analysis algorithms. This implies that there may be varying data, storage and processing requirements to facilitate the delivery of multiple types of analytic results.



a.Descriptive Analytics

Descriptive analytics are carried out to answer questions about events that have already occurred. This form of analytics contextualizes data to generate information.

b.Diagnostic Analytics

Diagnostic analytics aim to determine the cause of a phenomenon that occurred in the past using questions that focus on the reason behind the event. The goal of this type of analytics is to determine what information is related to the phenomenon in order to enable answering questions that seek to determine why something has occurred.

c.Predictive Analytics

Predictive analytics are carried out in attempt to determine the outcome of an event that might occur in the future. With predictive analytics, information is enhanced with meaning to generate knowledge that conveys how that information is related. The strength and magnitude of the associations are used to generate future predictions based upon past events. It is important to understand that the models used for predictive analytics have implicit dependencies on the conditions under which the past events occurred. If these underlying conditions change, then the models that make predictions need to be updated.

d. Prescriptive Analytics

Prescriptive analytics build upon the results of predictive analytics by prescribing actions that should be taken. The focus is not only on which prescribed option is best to follow, but why. In other words, prescriptive analytics provide results that can be reasoned about because they embed elements of situational understanding. Thus, this kind of analytics can be used to gain an advantage or mitigate a risk.

4.Big Data Characteristics

For a dataset to be considered Big Data, it must possess one or more characteristics (traits) that require accommodation in the solution design and architecture of the Big Data engineering environment. The five Big Data traits are commonly referred to as the Five Vs:

1. volume
2. velocity
3. variety
4. veracity
5. value

a.Volume

The anticipated volume of data that is processed by Big Data solutions is substantial and ever-growing. High data volumes impose distinct data storage and processing demands, as well as additional data preparation, curation and management processes.

b.Velocity

Big Data can arrive at fast speeds, and enormous datasets can accumulate within very short periods of time. From an enterprise's point of view, the velocity of data translates into the amount of time it takes for the data to be processed once it enters the enterprise's perimeter. Coping with the fast inflow of data requires the enterprise to design highly elastic and available processing solutions and corresponding data storage capabilities. Depending on the data source, velocity may not always be high.

c. Variety

Data variety refers to the multiple formats and types of data that need to be supported by Big Data solutions. Data variety brings challenges for enterprises in terms of data integration, transformation, processing, and storage.

d.Veracity

Veracity refers to the quality or fidelity of data. Data that enters Big Data environments needs to be assessed for quality, which can lead to data processing activities to resolve invalid data and remove noise. In relation to veracity, data can be part of the signal or noise of a dataset.

e.Value

Value is defined as the usefulness of data for an enterprise. The value characteristic is intuitively related to the veracity characteristic in that the higher the data fidelity, the more value it holds for the business. Value is also dependent on how long data processing takes because analytics results have a shelf-life.

5.Big Data Analysis Techniques

There are seven basic types of big data analysis techniques that are popular.

a.Quantitative Analysis

Quantitative data analysis technique uses scientific methods. This may include generation of data models, theories and hypotheses. It instruments manipulation of data variables.

Quantitative analysis collects, models and analyses empirical data. Quantitative analysis technique that focuses on quantifying the patterns and correlations found in the data. Based on statistical practices, this technique involves analysing a large number of observations from a dataset. Since the sample size is large, the results can be applied in a generalized manner to the entire dataset. Quantitative analysis results are absolute in nature and can therefore be used for numerical comparisons.

b.Qualitative Analysis

Qualitative analysis is the examination, analysis and interpretation of observations for the purpose of discovering underlying meanings and patterns of relationships. Qualitative analysis focuses on describing various data qualities using words. It involves analysing a smaller sample in greater depth compared to quantitative data analysis. These analysis results cannot be generalized to an entire dataset due to the small sample size. They also cannot be measured numerically or used for numerical comparisons. The analysis results state only that the figures were "not as high as," and do not provide a numerical difference. The output of qualitative analysis is a description of the relationship using words.

c.Data Mining

Data mining, also known as data discovery, is a specialized form of data analysis that targets large datasets. In relation to Big Data analysis, data mining generally refers to automated, software-based techniques that sift through massive datasets to identify patterns and trends. Specifically, it involves extracting hidden or unknown patterns in the data with the intention of identifying previously unknown patterns. Data mining forms the basis for predictive analytics.

d.Statistical Analysis

Statistical analysis uses statistical methods based on mathematical formulas as a means for analysing data. Statistical analysis is most often quantitative, but can also be qualitative. 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. There are three types of statistical analysis. They are described below.

1. **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. A/B testing is applicable in many domains particularly in marketing.
2. **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. 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.

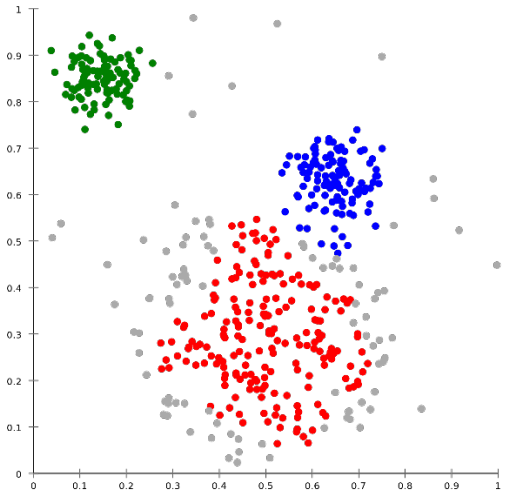
3. **Regression**: The analysis technique of regression explores how a dependent variable is related to an independent variable within a dataset. Applying this technique helps determine how the value of the dependent variable changes in relation to changes in the value of the independent variable. Regression 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.

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.

e.Machine learning

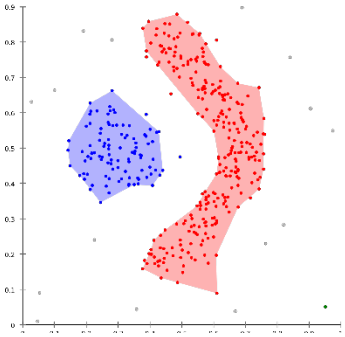
Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed. Machine learning analysis explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions expressed as outputs, rather than following strictly static program instructions. Four of the popular machine learning techniques are:

1. **Classification**: Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of pattern recognition.

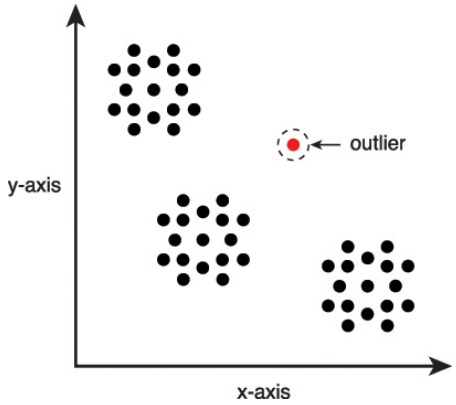


2. **Clustering**: Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense

or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis.



3. **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 unfavourable, such as risks. 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 visually highlights data points that are outliers



4. **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 behaviour or by matching the behaviour of multiple users. Filtering is generally applied via the following two approaches: (1) *Collaborative filtering* is an item filtering technique based on the collaboration, or merging, of a user's past behaviour with the behaviours of others. Collaborative filtering is solely based on the similarity between users' behaviour. It requires a large amount of user behaviour data in order to accurately filter items. It is an example of the application of big data. (2) *Content-based filtering* is an item filtering technique focused on the similarity between users and items. 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.

f.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. Semantic analysis represents practices for extracting meaningful information from textual and speech data. Three popular semantic analysis examples are:

1. **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. 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.
2. **Text Analytics**: Unstructured text is generally much more difficult to analyse 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.
3. **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 analysing 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 identifying customer satisfaction or dissatisfaction early, gauging product success or failure, and spotting new trends.

g.Visual analysis

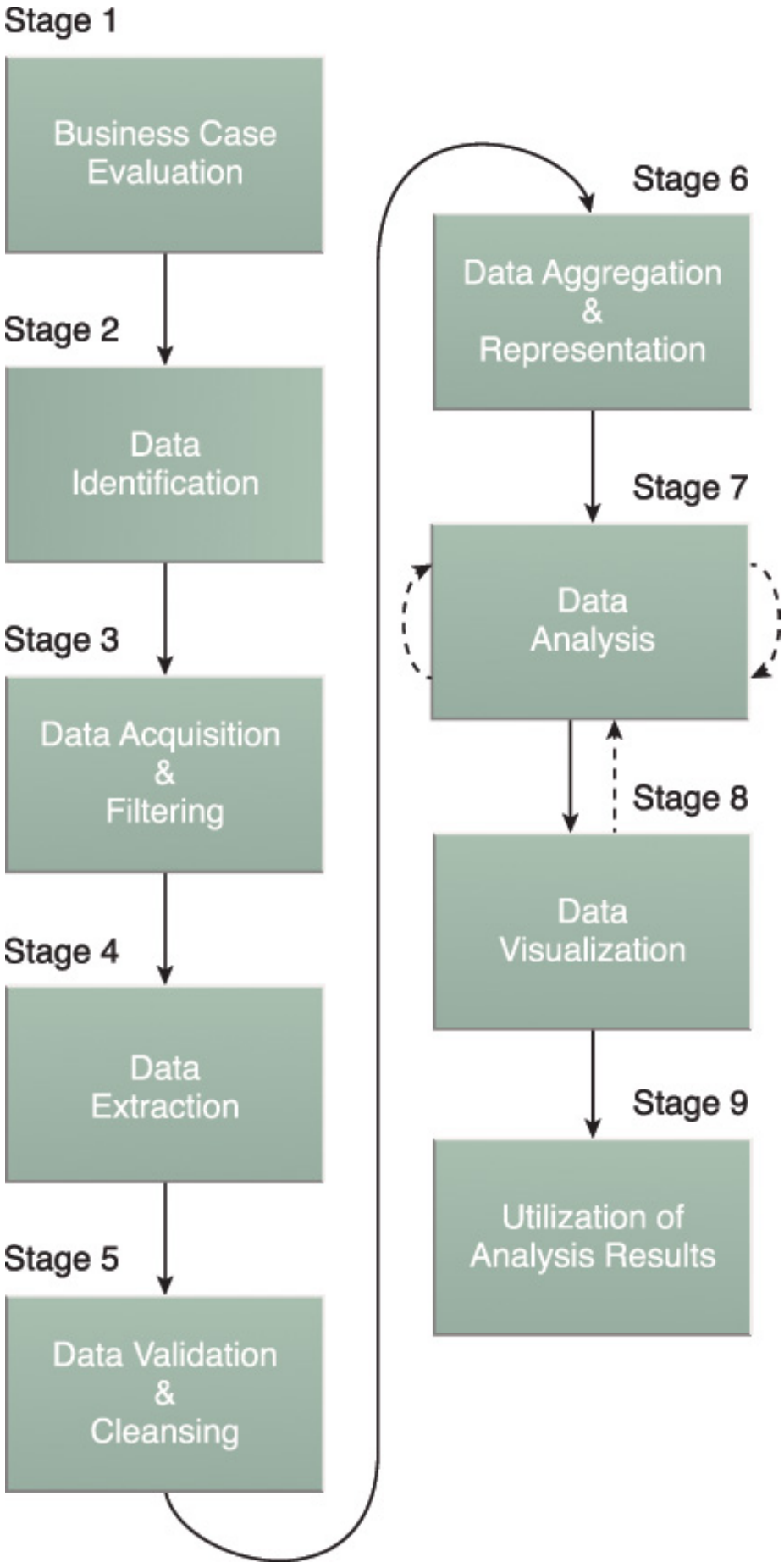
Visual analysis is a form of data analysis that involves the graphic representation of data to enable or enhance its visual perception. Based on the premise that humans can understand and draw conclusions from graphics more quickly than from text, visual analysis acts as a discovery tool in the field of Big Data. The objective is to use graphic representations to develop a deeper understanding of the data being analysed. 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. Some examples of visual analysis are heat maps, time series plots, network graphs and spatial data mapping.

6.Big Data Analytics Lifecycle

Big Data analysis differs from traditional data analysis due to the volume, velocity and variety characteristics of the data being processes. The capability to manage and analyse big data (petabytes) enables companies to deal with clusters of information that could have an impact on the business. This requires analytical engines that can manage this highly distributed data and provide results that can be optimized to solve a business problem. Analytics can get quite complex with big data. The Big Data analytics lifecycle can be divided into the following nine stages.

1. **Business Case Evaluation**: Each Big Data analytics lifecycle must begin with a well-defined business case that presents a clear understanding of the justification, motivation and goals of carrying out the analysis.
2. **Big Data Identification:** Identification stage is dedicated to identifying the datasets required for the analysis project and their sources.
3. **Big 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. The acquired data is then subjected to automated filtering for the removal of corrupt data or data that has been deemed to have no value to the analysis objectives.
4. **Big Data Extraction:** Some of the data identified as input for the analysis may arrive in a format incompatible for processing. 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 useful processing format.
5. **Big 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.
6. **Big Data Aggregation & Representation**: The Data Aggregation and Representation stage, is dedicated to integrating multiple datasets together to arrive at a unified view.
7. **Big Data Analysis:** The Data Analysis stage is dedicated to carrying out the actual analysis task, which typically involves one or more types of analytics.
8. **Big Data Visualization:** 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.
9. **Big Data Utilization of Analysis Results:** Utilization of Analysis Results stage, is dedicated to determining how and where processed analysis data can be further leveraged.

The stages are serially pipelined and each stage depends on its previous stage. The nine stages of Big Data Analytics lifecycle is shown in the diagram overleaf.


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