INFO5060 – Data Analytics and Business Intelligence

General Information

This is an intensive unit which will be completed over a span of 4 weeks. An estimated of 40 hours per week is needed through the semester. The mark distribution for the unit is as follows:

Type	Description	Weight	Due
Quizzes	Quiz based on the topic though every week. It will be a one-time attempt, around 5 questions, and will be available for only 48 hours. Time limit for this is 20 minutes.	10%	Every week
Assignment (Submission 1)	Group assignment to create a Dashboard Solution. Only one dataset can be chosen per group.	10%	3rd July
Presentation	BI Solution presentation	10%	14th July
Peer Review	Group assignment peer review feedback	5%	8th July
Assignment (Submission 2)	Group assignment to revise the dashboard	15%	15th July
Final Exam	For all week's content. It is a mix of MCQ, Short answer and long answer questions. It is an offline exam, closed book.	50%	17th of July 2023

Weekly Schedule

WK	Topic	Learning activity	Learning outcomes
Week 01	Course Overview & Admin	Lecture (2 hr)	LO1 LO2 LO3 LO4 LO5 LO6
	Introduction to Business Intelligence & Descriptive Analysis I	Lecture (2.5 hr)	LO1 LO2 LO3 LO4 LO5 LO6
	Course Overview & Admin	Practical (2.5 hr)	LO1 LO2 LO3 LO4 LO5 LO6
	Introduction to Business Intelligence & Descriptive Analysis I	Practical (2 hr)	LO1 LO2 LO3 LO4 LO5 LO6
Week 02	Descriptive Analysis II	Lecture (2 hr)	LO1 LO2 LO3 LO4 LO5 LO6
	Predictive Analytics I & Project discussion	Lecture (2.5 hr)	LO1 LO2 LO3 LO4 LO5 LO6
	Descriptive Analysis II	Practical (2.5 hr)	LO1 LO2 LO3 LO4 LO5 LO6
	Predictive Analytics I & Project discussion	Practical (2 hr)	LO1 LO2 LO3 LO4 LO5 LO6
Week 03	Project discussion & Predictive Analytics II	Lecture (2 hr)	LO1 LO2 LO3 LO4 LO5 LO6
	Predictive Analytics II & Analytics at Scale	Lecture (2.5 hr)	LO1 LO2 LO3 LO4 LO5 LO6
	Project discussion & Predictive Analytics II	Practical (2.5 hr)	LO1 LO2 LO3 LO4 LO5 LO6
	Predictive Analytics II & Analytics at Scale	Practical (2 hr)	LO1 LO2 LO3 LO4 LO5 LO6
Week 04	Analytics at Space & Evaluation Success	Lecture (2 hr)	LO1 LO2 LO3 LO4 LO5 LO6
	Unit review	Lecture (2.5 hr)	LO1 LO2 LO3 LO4 LO5 LO6
	Analytics at Space & Evaluation Success	Practical (2.5 hr)	LO1 LO2 LO3 LO4 LO5 LO6
	Unit review	Practical (2 hr)	LO1 LO2 LO3 LO4 LO5 LO6

Readings

- ~~Business Intelligence, Analytics, and Data Science: A Managerial Perspective~~
- ~~With extracts from Fekete, A. Information Visualization and Charts 2021~~
- ~~Data Analytics Made Accessible, by Anil K. Maheshwari, 2015~~
- ~~Data mining: concept and techniques, 2nd edition, Han M. and Kamber M:, chapter 1, p 6~~

Examination Information

- MCQ
- Short Answer Questions
- Long Answer Questions

Week 1

Introduction

In this unit, we will see how business works in the corporate world and how being a data analyst one can handle data from the business intelligence point of view. From the data analytics point of view, there are multiple roles such as Data Analyst, Data Scientist, and Business Analyst. Data Scientist work with programming languages to create machine learning models, etc. Business Analysts prepare all the information for the Data Analyst to deliver the work. This includes all the preparatory works, the requirements, etc. The Data Engineer works in the architecture side wherein they use tools such as SQL to access the data efficiently to prepare visuals for the higher authorities.

One needs to understand the data completely. What does the data mean? What questions can be asked to the data? The more efficiently one knows the data, the easier it is to use tools and visualize the data.

Types of Analytics

- Descriptive analytics

  It is the first step of analytics. It gives us insight into the simplest to the most complex type of questions. The outcome of Descriptive analytics are well defined business problems and opportunities. It is the examination of data or content and is usually the answer to a question. This type of analytics is enabled through business reporting, dashboards, scorecards, and data warehousing. Descriptive analysis provides a retrospective analysis of historic data. A few example questions:
  - What happened?
  - What is happening?
- Predictive analytics

  This is the second step of analytics. It leverages machine learning models to predict the state of things in the future based on past events or historical data. The outcome of predictive analytics are accurate projections of future events and outcomes. This type of analytics is enabled by data mining and forecasting. Predictive analytics is the process of using data to forecast future outcomes. A few example questions:
  - What will happen?
  - Why will it happen?
- Prescriptive analytics

  This is the third step of analytics. It is the use of advanced processes and tools to analyse data and content to recommend the optimal course of action or strategy moving forward. The outcomes of prescriptive analytics are the best possible business

decision and action. This type of analytics is enabled through optimization, simulation, decision modelling, and expert systems. A few example questions:

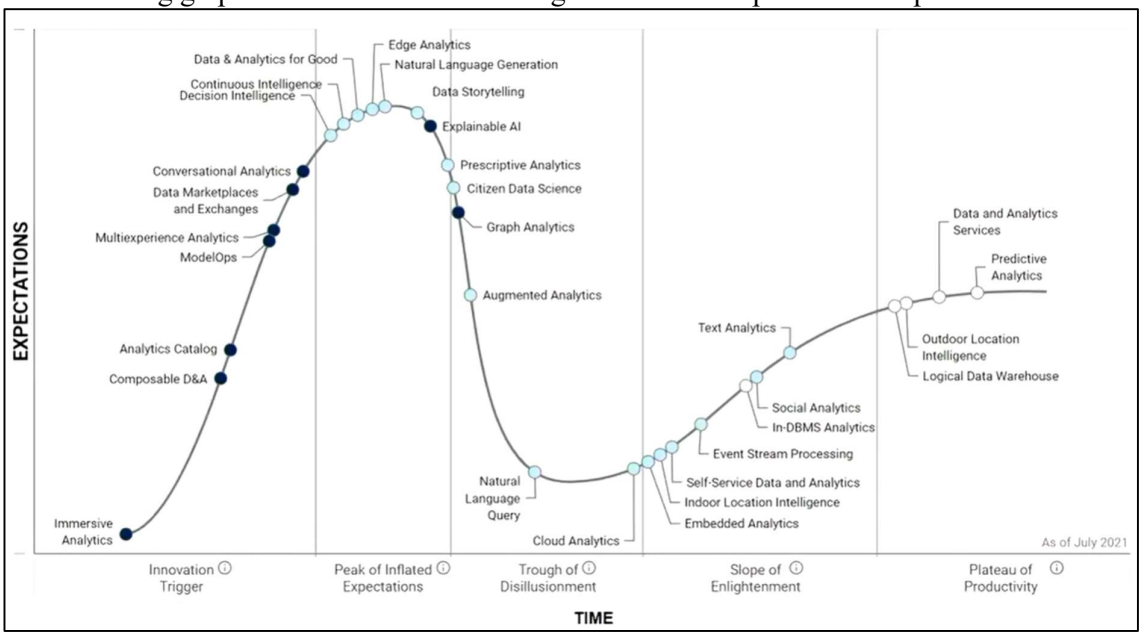
- What should I do?
- Why should I do it?

Apache Hadoop and Spark

Apache Hadoop is an open-source framework that allows for the distributed storage and processing of large datasets across clusters of computers using simple programming models. Hadoop is a cluster-based ecosystem. Similarly, Apache Spark is also an open-source framework that is a distributed processing system used for big data workloads. It uses optimized query execution for fast analytic queries against data of any size. Apache Spark is also built on Hadoop.

Introduction to Business Intelligence

Business Intelligence is a software that ingests business data and presents it in user friendly views such as reports, dashboards, charts, and graphs. BI greatly enhances how a company approaches its decision-making by using data to answer questions of the company's past and present. By providing real-time, hyper-accurate reports, and by assisting organizations to better understand the information being detailed, BI helps eliminate the need for guesswork. The following graph is the timeline of the usage of BI with respect to the expectations.



Every innovation takes us to the peak of expectations over time, these expectations slowly die out and then flattens out as a plateau. In the above image one can see the emergence of BI as an innovation trigger. This goes into the peak of inflated expectations and after a certain period the expectations decline. This brings the graph into the trough of disillusionment after which gradually increases into the slope of enlightenment and then plateaus at the plateau of productivity.

The most common considerations that organisations face is demand, supply, cost, sale, products, process, quality, competitions, etc. Questions can be asked to solve business problems and these questions need to be answered through BI.

Data is present everywhere. It comes from various sources and will be mostly unstructured. Through Business Intelligence, we work with data, information, intelligence, and analytics. Data is nothing but facts and statistics collected for reference or analysis. Information are facts provided or learned about something or someone. Intelligence is the ability to acquire and apply knowledge and skills. Analytics is the systematic computational analysis of data or statistics. Most of the time, data analysts face with a lot of jargons. These are words which are specific to a particular industry and is usually bad for communication. Data on its own is unprocessed and can tell us nothing more than the state of being it was observed in.

Information itself is processed data. It is data that has been collected and processed into a more meaningful form. It can answer more meaningful questions about the entities the entire dataset is about. Intelligence is the information that has been leveraged into a decision-making process. Thus, business intelligence is the information that has been leveraged into a decision-making process of a business. Analytics is the tools, techniques, and processes that move things from data to information to intelligence.

Data

Data is generally categorized into 2 types, structured and unstructured data. Structured data is tabulated data or data which has a uniform type. Unstructured data is a mix of different types of data such as images, videos, text messages, etc. Data cleaning is an important task which is required to use data to gain insights. Only when the data is cleaned, and noise removed can one gain some valuable information from the data. Data can be obtained from various sources in different formats and these data needs to be converted into a simple and uniform data format. We use spreadsheets for this.

Spreadsheets are not data storages. Data storages are traditionally SQL databases such as MySQL, SQL Server, Oracle, etc. Recently, NoSQL is also gaining popularity such as MongoDB, HBase, etc. Followed by data storage, one needs some application engine to handle computation of data (data processing) and followed by which data presentation which is usually done through some form of intranet or reports portal.

Today we have got lots of data mined and collected by organisations. This data is called big data and most of this data is discarded. A lot of big data is based on some contributing ideas such as the most valuable data is already being used, there is some value in the data not being used and that there is enough data that is not being used that some value there can become huge value if enough data is used.

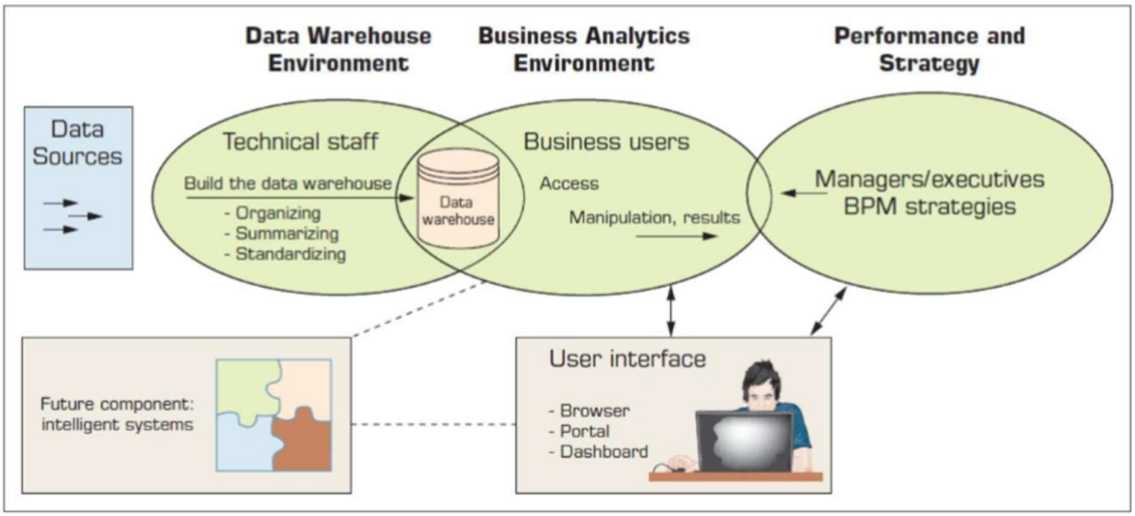
OLTP

OLTP is short for Online Transaction Processing. These work in operational databases such as ERP, CRM, etc. and their main goal is to capture data.

OLAP

OLAP is short for Online Analytical Processing. These work in data warehouses, data lakes, etc. The goal of OLAP is to support decisions.

High level architecture of Business Intelligence



Data is brought into the workspace through data sources. This data will be prepared by the technical staff and loaded into the Data warehouse. From this data warehouse business users will access the data and manipulate it to get their desired results. These results are accessed by the managers, executives, or BPM strategists to make informed decisions for the organisation.

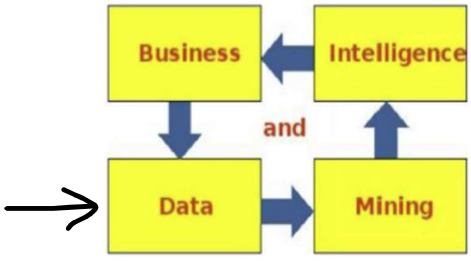
Dashboards

Dashboards are a basic visualization tool. It is a combination of data visualization tool, analytical/processing engine, and the supporting tools to allow data ingress and egress, deployment, access control, etc.

Week 2

Data Analytics

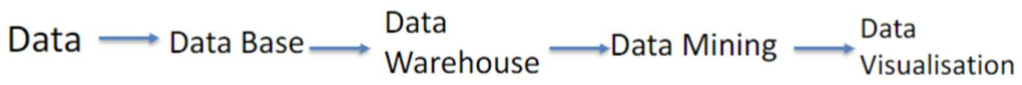
Data can be analysed and minded using special tools and techniques to generate patterns and intelligence which reflects how the business is functioning. Data comes from different sources, can be structured or unstructured data. This needs to be processed and visualized. A few techniques which need to be applied are data mining (to obtain and analyse data), descriptive analytics (Descriptive analytics are of 2 types, the first type deals with the data processing steps.



Data Analytics Made Accessible, Chapter 1, page 11

Data Preprocessing

Data can be modelled and stored in a database. It is extracted from the operational data stores and stored in a data warehouse. The warehouse can be combined with other sources of data and mined using data mining techniques to generate new insights. The insights visualized and communicated to the right audience in real time for competitive advantage. Data engineers are the one who manage the data warehouses.

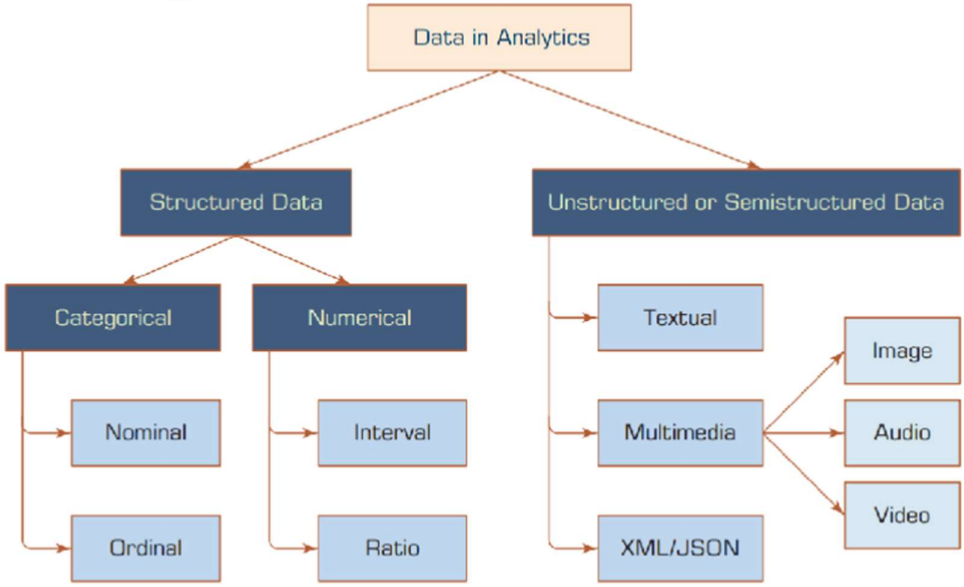


The progression of data processing activities

The collected data is messy, has missing bits, may be heterogenous (different), disparate, and big. A huge part of analytics occurs before the analysing step begins; this step is data preprocessing. Data is ready if:

- Data source is **trusted**, creates insights, and can be used for decision making.
- Data source is **accessible**, meaning that it enables to move quickly through the data and build an informed culture where data leads to better decision and action.
- Data can be **protected**, and access can be restricted through the usage of data protection tools and policies.
- Data is **complete** and usable for particulate use of users.
- Data is **consistent** as it moves over a network or between different applications on a computer.
- Data is **available** and accessible not only from the source but also from any other applications.
- Data is **granular** in nature. Meaning that there are multiple levels of details of the data.
- Data is **consistent** within a defined domain and hence is valid for the usage.
- Data is **relevant** which means that it includes completeness, consistency, accuracy, and timeliness between data content and area of interest of the users.

Data is of different types and a classification is given below:



Examples for each type of data is given below:

- Structured
  - Categorical
    - Nominal: Male/Female, Sydney/Canberra/Wollongong
    - Ordinal: Good/Better/Best, Child/Teen/Adult/Elderly
  - Numerical
    - Interval: year, temperature, etc.
    - Ratio: count, distance, etc.

- Unstructured
    - Textual: comments, reviews, etc.
    - Multimedia: gifs, images, videos, etc.
    - XML/JSON: web pages, etc.

When visualizing data, the processed data is either dimensional or measures. Dimensions are usually categorical in nature; Measures are usually numerical in nature. Data preprocessing has further steps, these are:

- Data consolidation
    - It relates to the collection of data, selection of the data from different sources, and the integration of the datasets into one usable dataset. Usually this data may be static, but in case it is dynamic, the data will constantly change over time. Data can even be ion different forms such as one temperature data being in Celsius or another being in Fahrenheit, etc.
- Data cleaning
    - It relates to the removal of outliers, imputations, noise reduction and duplicate eliminations. There can also be human errors or recording errors which need to be processed. Some actions that can be used other than the previously stated are using default data, omitting the cells, omitting the rows, etc.
- Data transformation
    - Normalizing the data, discretization, and creation of attributes in the data. This step is done when the data is not structured conveniently. Continuous data I placed into groups or aggregates are combined, etc.
- Data reduction
    - This relates to the reduction of dimension, volume reduction, and make sure that the data is balanced. Not all big data is good, as it can be slow and expensive. Hence, removing data which is not required and using only those needed, is an important step for big data processing.

Data Validation

It is the process of checking, cleaning, and ensuring the accuracy and consistency of the data. It also involves the reporting and decision making. It maintains the data integrity as it helps identify and correct errors, inconsistencies, and inaccuracies. Data validation is applicable in various stages of the data life cycle. Data validation is particularly crucial when integrating data from multiple sources.

Data validation is very important as it improves decision making, increases efficiency, enhances data security, and makes sure that the data is regulatory compliant. There are different types of validations:

- Syntax validation
  It refers to the process of checking whether the data conforms to the specified syntax or format. It involves verifying whether the data follows the rules and patterns defined for its structure, such as data types, lengths, characters, and formats.
- Semantic validation
  It refers to the process of checking the meaning and logical consistency of the data. Unlike syntax validation, which focuses on the format and structure of the data, semantic validation evaluates the content and context of the data to ensure it is meaningful and coherent within the given domain or system.
- Business rule validation

It refers to the process of checking whether the data adheres to the specific rules and constraints defined by the business or organization. Business rules are guidelines or conditions that govern how data should be structured, processed, or used within a particular business context.

- Comparison validation
  It refers to the process of comparing data values or properties against specific criteria or reference values. It involves evaluating the relationship between different data elements to ensure they meet certain conditions or constraints.

There are different methods of validation methods as well, these are:

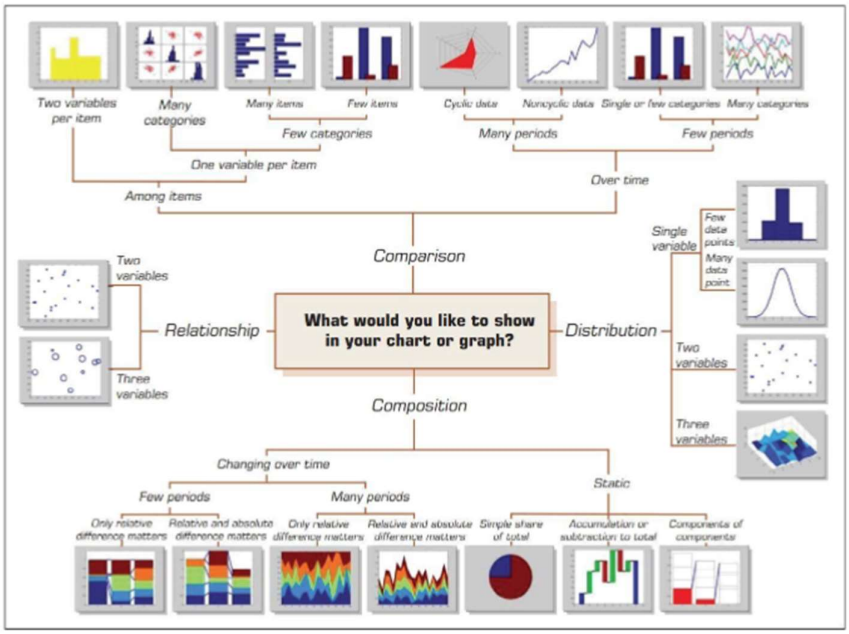
- Field level validation
  Field-level validation refers to the process of validating individual data fields or attributes to ensure they meet specific criteria or constraints. It involves checking the integrity and correctness of data within each field independently.
- Form level validation
  Form-level validation, also known as cross-field validation, refers to the process of validating data and relationships across multiple fields within a form or data entry interface. It involves checking the coherence and consistency of data across different fields, ensuring that they meet specific criteria or constraints when considered together.
- Data saving validation.
  When data is being saved or stored, validation processes are commonly performed to ensure the integrity, accuracy, and consistency of the data being saved.
- Search criteria validation
  Search criteria validation refers to the process of validating the input or parameters used for conducting a search or query in a search system or database. It involves verifying the validity and coherence of the search criteria to ensure accurate and meaningful search results.

Data Visualization

The primary guideline for having a better visualization is given below:

- Show the data.
- Reduce the clutter. (LESS IS MORE)
- Integrate the graphics and text.
- Avoid the spaghetti chart.
- Start with grey. (GREY IS DEFAULT, REMEMBER COLOUR THEORY)

There are various types of charts that can be used to visualize data. A few examples are line charts, bar charts, pie charts, scatter plot, bubble charts, histograms, geographic maps, heat maps, etc. Choosing the type of visual is also very important. The image given below helps in choosing the graph:



Some key issues with data visualization lies with biases, misrepresentation, distortion, obfuscation, but also attractiveness.

Statistical Modelling for Business Analytics

Statistics is a collection of mathematical techniques to characterize and interpret data. Descriptive statistics is describing sample data on hand. Inferential statistics is drawing inferences about the population based on sample data. Through statistics we use measures of central tendency such as arithmetic mean, median, mode, measures of dispersion such as dispersion, range, variance, standard deviation, etc. Shape of distribution such as histogram, skewness, kurtosis, etc.

Inferential statistics is used for hypothesis testing, forecasting, and to assess the strength of the relationship between variables. Regression is a part of inferential statistics which is one of the most widely known and used analytical technique. It is used to characterize relationship between explanatory variables (input) and response variables (output). Regression analysis is based on these inferential statistics. There are different types of common regression analysis, these are:

- Linear regression analysis
    - Univariable linear regression studies the linear relationship between the dependent variable Y and a single independent variable X.
- Multiple regression analysis
    - Multiple regression is a statistical technique that can be used to analyse the relationship between a single dependent variable and several independent variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value.
- Logistic regression analysis
    - It is a complex and difficult regression analysis. A huge amount of data is required to produce an outcome from this type of regression analysis. In this kind of regression analysis, the target variable is a binomial variable. It is a classification algorithm which employs supervised learning. Logistic regression

is a predictive analysis and can be used to describe the data as well. It explains the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables.

Descriptive Analytics

It involves breaking down the data and summarizing its main features and characteristics. Descriptive Analytics presents what has happened in the past without exploring why or how. Descriptive analytics uses basic descriptive statistics such as measures of distribution (frequency or count), central tendency (mean, median, mode), and variability (variance, standard deviation). Descriptive analytics often presents reports, pivot tables and visualizations such as histograms, line graphs, pie charts, and box and whisker plots.

The main advantage of Descriptive analysis is given below:

- Complex data is presented in an easily digestible format.
- It provides a direct measure of the incidence of key data points.
- It is inexpensive and only requires basic math skills.
- It is faster to carry out, especially with excel and python.
- It relies on data that organizations have already access to.
- It looks at the complete population than a sample of it, making it more accurate.

The main disadvantages of Descriptive analysis are given below:

- It is possible to summarize the datasets but does not give us a complete story.
- We cannot use descriptive analysis to test a hypothesis or understand why data is present the way it is.
- It can't be used to predict what may happen in the future.
- It is not possible to generalize the findings to a broader population.
- It tells us nothing about the data collection methodology, meaning the data may include errors.

Predictive Analysis

It is a modelling technique to make predictions about future outcomes and performance. It makes predictions about certain unknowns in the future. Predictive analysis and Machine learning are two completely different disciplines. Predictive analysis can be used for forecasting, marketing, supply chain, human resources, etc. There are different models involved with predictive analysis such as decision trees, regression, neural networks, cluster models, etc.

Data warehousing

Data warehouse is a pool of data produced to support decision making. There are multiple different concepts involved with Data warehousing and ETL. A few of these are:

- **Data Store:** It is any mechanism for the storage of electronic data.
- **Operational Data Store:** It is related to live data before it is processed into a data warehouse, providing real time analysis rather thana static historical data storage for the data warehouse.
- **Database:** Software that manages data storage.
- **Data Lake:** A mass data storage infrastructure that stores un-purposed data. It is usually unprocessed data as the purpose is unknown for the data.

- **Data Ocean:** Data Lake but bigger.
- **Data Warehouse:** A mass data storage infrastructure that stores data for specific purposes and has pre-processed that data for that purpose.
- **Data Mart:** A subset of a data warehouse intended to store data on a single aspect of an enterprise such as a single department. These are of 2 types:
    - **Dependent Data Mart:** A subset of a data warehouse. Ensures consistency, requires the expensive DW to exist.
    - **Independent Data Mart:** The same data as a dependent data mart would hold, but without the backing of a data warehouse. Cheaper to operate but sacrifices guaranteed consistency.
- **Data Pool:** A data combination infrastructure, intended to unify different formats and structures, usually for the purposes of cross-organisation communication.

A data warehouse is pre-processed and ready for use data storage. Data warehouses have data stored in them which has a purpose. This can be multi-purpose as data warehouse for single purpose use is way too expensive! A few characteristics of Data warehousing are:

- Data accessibility

  Most platforms are either web - based, or at least web-enabled. Thus, data warehouses are too.
- Data timeliness

  Data warehouses are typically relatively up to date but are not intended for the most rapid analyses.
- Meta data

  Data warehouses will include metadata to enable the decisions embedded in the data to be understood, any issues with reliability, accuracy, validity etc to be considered by users.

ETL

ETL is short for Extract, Transform, and Load. It is a fancy term for data preprocessing into and out of data warehouses and other data stores. Extract means reading the data in some form from the data stores. Transform means mutating the data in some way to make sure the analytics fits the target. Loading means putting the data into the target data store or analytics engine.

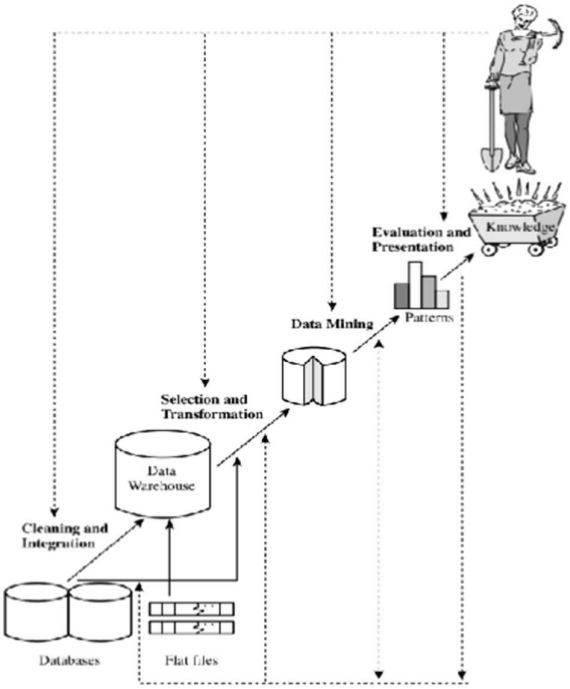
Week 3

Data Mining

Data is collected from everywhere in the world. It is collected daily, and a huge amount of data is collected every minute. This data in its raw form is useless, and hence such data needs to be turned into useful information and knowledge.

Data mining refers to extracting or mining knowledge from large amounts of data. It simply means knowledge discovery from data. Data mining turns a large collection of data into knowledge. This is usually done by finding trends and patterns across the data. The evolution of Database System Technology started from data collection and database creation in the 1960s and earlier. This was followed by the database management systems which was created in 1970s. This evolved into Advanced Database systems, Advanced data analysis, web-based databases, and much more. All of this has been integrated into the new generation of integrated data and information systems.

Concepts of Data Mining



The data is all collected into the databases. This data is cleaned and integrated and stored in data warehouses. Through selection and transformation, selected data is processed into data mining and trends and patterns emerge from this. Using different techniques one can make sense of this trend and patterns and through evaluation and presentation, create knowledge. The following are some characteristics of data mining:

- It is a process:
  Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories.
- Nontrivial:
  Data here if previously unknown can be useful information in the future. More calculation is involved that the number of actions taken around those results.
- Valid:
  The conclusions must be based on mathematically/statistically sound principles, can be reproduced on new data with influence conclusions, etc.
- Novel:
  The results are previously unknown, can be interpreted as the creation of knowledge that was not present in the data store previously.
- Potentially useful
  This starts to leave the idea of data mining and enters the idea of business intelligence. It requires an appreciation of what value means in that context.

Data Mining functionalities

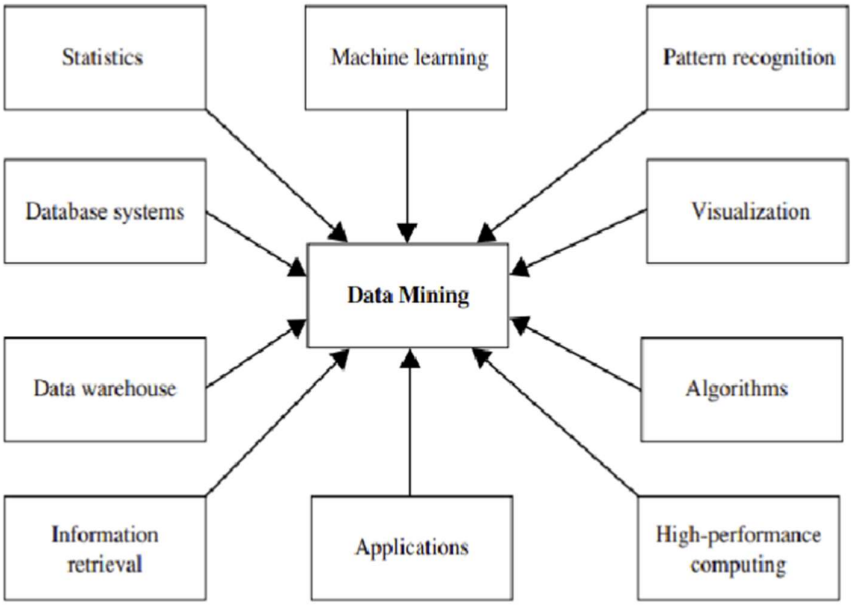
This refers to the representation of types of patterns that must be discovered in data mining tasks. These tasks are classified into two types:

- Descriptive mining
  Defining the common features of the data in the database.
- Predictive mining
  Inferencing on the current information to develop predictions.

There are different functionalities involved in data mining such as data characterization, data discrimination, association analysis, classification, prediction, clustering, outlier analysis, etc. Data mining process usually involves the following steps:

- Business understanding
  Understand the problem that needs to be solved with defining project objectives.
- Data understanding
  Explore data for better understanding of its structure, quality, and content.
- Data preparation
  Cleaning, inconsistency removal, accuracy and etc
- Modelling
  Appropriate algorithm, train model with data and evaluate model performance.
- Evaluation
  Performance of the model and predict outcomes on new data.
- Deployment
  Integrating the model into existing systems.

Techniques used in Data Mining



A few more techniques are:

- Labelling/Classification
  Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts.
- Grouping/Segmentation
  Clustering can be used to generate class labels data. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.
- Prediction/Forecasting
  Values are unknown and uncertain most often but not necessarily. Using prediction or forecasting, one can predict future values based on the past values.

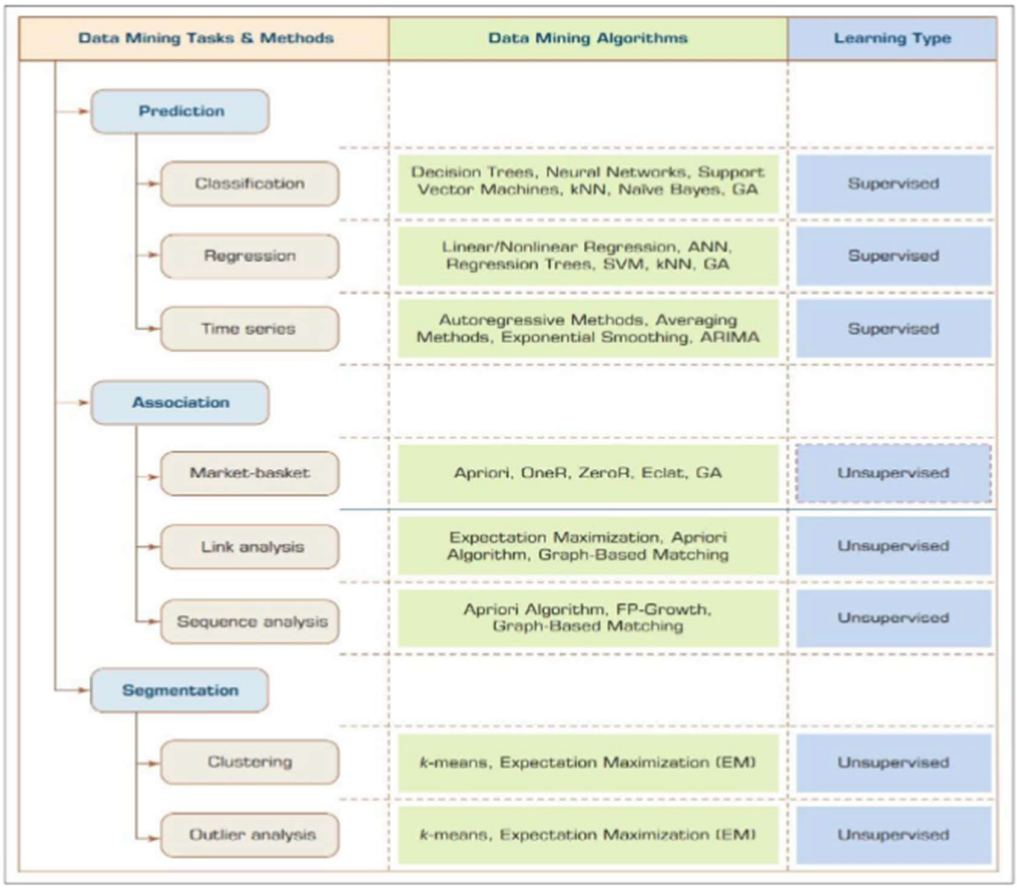
There are several types of confidence that can be had in the mining output. These are:

- Calculation confidence

  It is expressed numerically and calculated specifically based on known error ranges.
- Model confidence

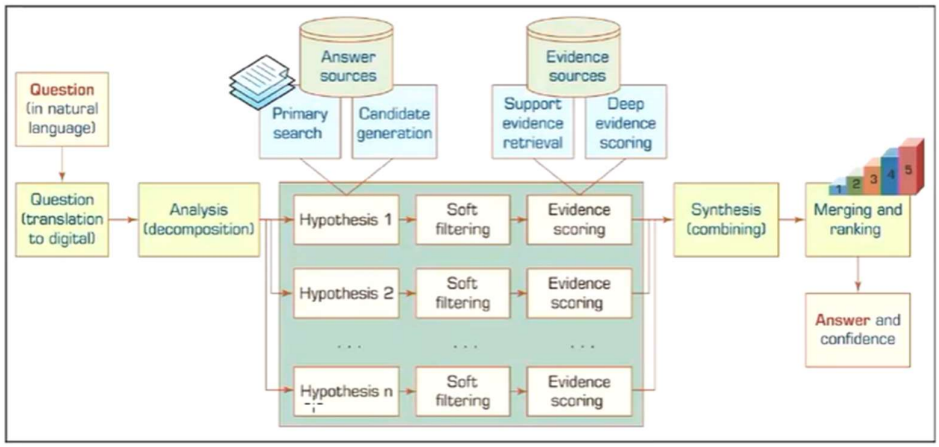
  It is determined by an expert based on the quality of the model developed, the appropriateness of the calculations involved and many other factors.
- User confidence

  It is critical to the business/organisational intelligence. It is usually decided by the stakeholders and is usually meant for aligning with business principles.



Data Preprocessing

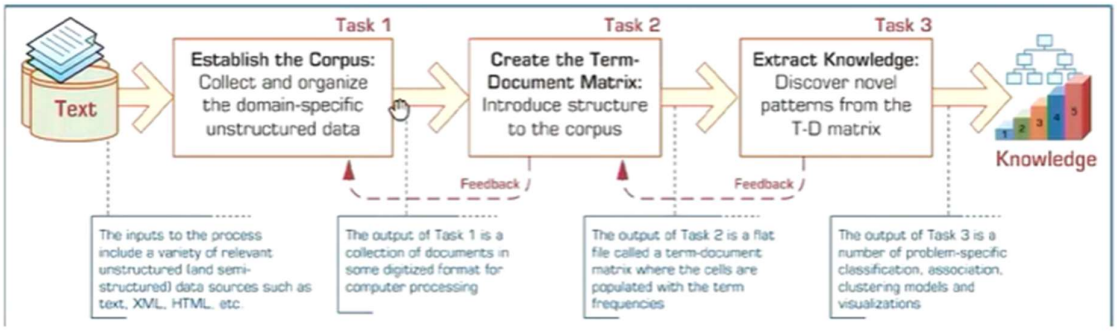
Analysing the unstructured data and converting them into structural data as per requirements. Data is of different types, broadly, it is categorised into structured data and unstructured data. Unstructured data usually consists of data of different types of data such as text, images, videos, pdfs, etc. The process of analysing unstructured data is given below:



A question is asked to the system in its natural language, this is translated into the machine language. Once translated, the question is analysed (decomposed) to form various hypothesis. This hypothesis comes from the primary search of various answers from the answer source. After soft filtering the remaining information is scored based on evidence from the evidence sources. These answers are then combined to merge and rank the hypotheses after which the answer and its confidence is received.

Text Data Mining

Text mining is a component of data mining. It deals with unstructured text data. It involves use of NLP techniques to transform unstructured text data into structured data. It can be defined as: Text mining is the procedure of synthesis information, by analysis relations, patterns, and rules among textual data. To perform text mining, we need to first impose structure to the data and then mine the structured data. The process of knowledge gain from text mining is given below:



The steps are:

1. Establish the corpus:
    a. Collect all relevant unstructured data.
    b. Digitize, standardize the collection by converting it into ASCII files.
    c. Place the collection in a common place.
2. Create the Term Documents Matrix
    a. Introduce structure to the corpus by keyword matching.
    b. Create a matrix of this information which would be a sparse matrix.
3. Extract patterns/knowledge
    a. Classification, clustering, association, trend analysis, etc.

A few techniques which can improve keyword matching:

- Stemming
  This technique identifies words with a common root form. For example: Swim, swimmer, swimming, swam, etc.
- Synonyms
  This technique identifies words with the same meaning. For example: throw, toss, hurl, cast, etc.
- Homonyms
  This technique identifies words that are written the same way but mean different. For example: Bark (Tree), Bark (Dog), Mean (personality), Mean (mathematical), etc.
- Contextual analysis
  Using the context of words to discern their meaning. Here we look for parts of speech such as nouns, verbs, adjectives, etc.
- True NLP
  This is a work in progress. One of the end goals of NLP is to make computers code as humans can.

Key applications of Text data mining:

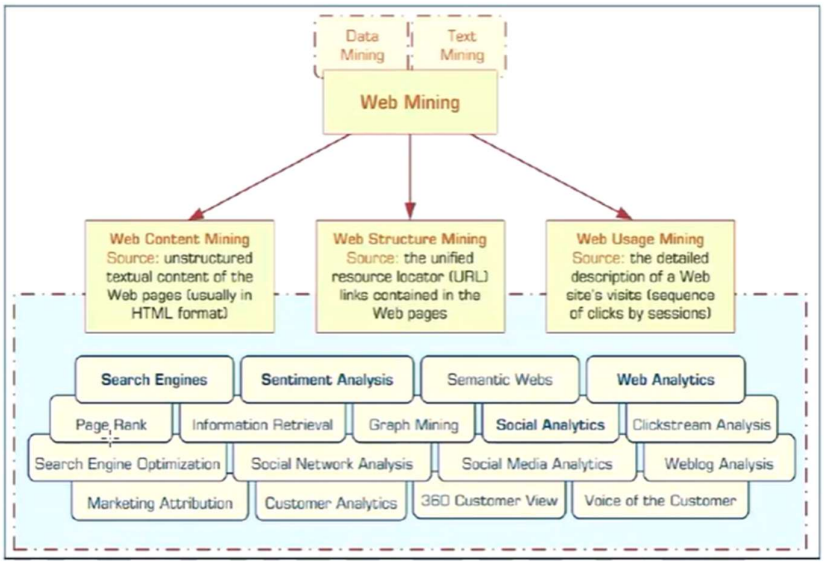
- Information extraction: Identification of key phrases and relationship within texts by looking for predefined objects and sequences in texts by way of pattern matching.
- Topic tracking: Based on user profile and documents that a user views; predictions can be made what other documents or topics might be of interest to the viewer.
- Summarization: Summarizing a document to save time for a reader.
- Categorization: Identifying the main theme of a document and categorizing it correctly based on the predefined categories based on those themes.
- Clustering: Grouping similar documents without having predefined set of categories.
- Concept linking: Connect related documents by identifying their shared concepts and help user find information that might not have been found through traditional approaches.
- Question answer: Finding the best answer to given question through knowledge driven pattern matching.

Sentiment Analysis is the process of analysing digital text to determine if the emotional tone of the message is positive, negative, or neutral. For sentiment analysis, samples from textual data are checked for the presence of any sentiment, if there is none, it goes back to the data source. If there is any sentiment present, the sentiment polarity is calculated and recorded. The target for the sentiment is identified and then its tabulated.

Social Network Data Mining

This type of data mining usually leverages web mining which is a specialization of text mining. An image is given below:

Social Network Data Mining



Social Network Data Mining is the monitoring, analysing, measuring, and interpreting of digital interactions and relationships of people, topics, ideas, and content. There are broadly 3 main categories in Social Network Data Mining which are:

- Social Network Analysis
  Identifying the networks that exist between people. Places, products, etc. for the purpose of analysis.
- Content Analysis
  Perform text mining on social network and social media content.
- Multimedia Analysis
  It deals with analysis of different types of data such as video, audio, etc. This data can be structured, unstructured, or semi-structured.

Week 4

Prescriptive Analytics

Prescriptive analytics is the use of advances processes and tools to analyse data and content to recommend the optimal course of action or strategy moving forward. To do prescriptive analytics one needs to consider information about possible situations or scenarios, available resources, past performances, and current performances and then suggest a course of action or strategy. Prescriptive analytics enables better decision making.

It has the following advantages:

- Prevent fraud, limit risks, increase efficiency, etc.
- Simulate the probability of various outcomes and show the probability of each, helping organizations to better understand the level of risk and uncertainty, etc.

It has the following disadvantages:

- It is only effective if organizations know what questions to ask and how to react to answers.
- It is only suitable for short term solutions.
- It is unreliable if more time is needed.

Big Data

Data is constantly being created and is done so by multiple sources. Big data has 3 attributes that stand out, these are:

- Huge **volume** of data: Big data can be billions and trillions of row than thousands or millions.
- **Complexity** of data types and structures: Big data reflects the variety of new data sources, formats, structures, etc.
- **Speed** of new data creation and growth: Big data can describe high velocity data, it has a rapid data ingestion and near real time analysis.

Big data can come in multiple forms. It can be structured or unstructured. Contrary to much of the traditional data analysis techniques, big data requires different tools to process and analyse. Usually, distributed computing environments and massively parallel processing architectures are preferred to process such complex data. Under Big data, there are different types of data, these are:

- Structured data: These data have a defined data type, format, and structure. For example: transaction data, OLAP data cubes, traditional RDBMS, etc.
- Semi-structured data: These data have a discernible pattern that enables parsing data files that are self-describing and defined by an XML.
- Quasi-structured data: Erratic data formats than be formatted with effort, tools, and time.
- Unstructured data:
  This type of data has no structure and may include text document. For examples: PDFs, images, etc.

Big Data is described by the 5Vs as their characteristic traits:

- Volume: The more data, the bigger the big data.
- Variety: Well, formed, clean, entirely structured data is good.
- Velocity: How quickly the data is coming to you.

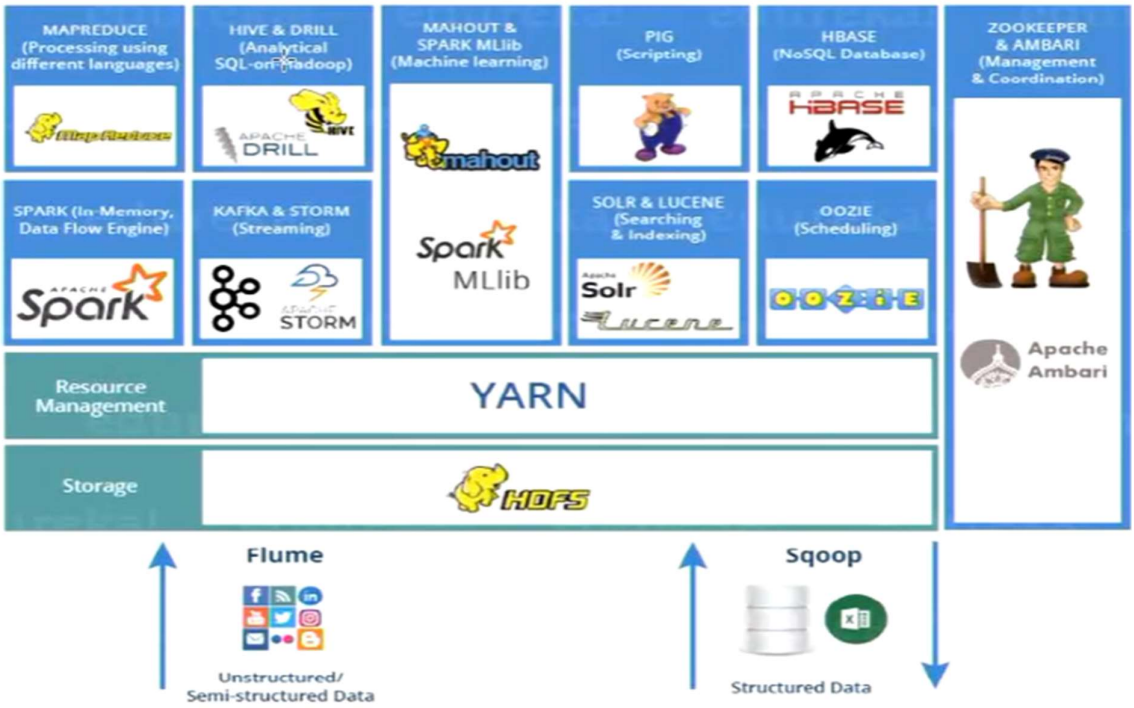
A challenge for big data is scalability. The traditional approach is to scale with the increasing load, but this is a very expensive approach. The databases are being overwhelmed with the data they are facing and hence data scaling is being implemented. Data scaling can help in improving performance by:

- Scaling down the amount of data processed, or the resources needed to perform the processing.
- Scaling up the computing resources on a node using parallel processing and faster memory or storage.
- Scaling out through computing via distributed nodes in a cluster/cloud.

HDFS and Hadoop Map Reduce

Hadoop is an open-source framework for storing and analysing massive amounts of distributed, unstructured data. Hadoop clusters run on inexpensive commodity hardware so projects can scale out inexpensively. Hadoop is part of Apache Software and can work with several things. It can be used for indexing the web for search, graph analysis, text analysis, machine learning, etc. Hadoop breaks up the data into parts and then load it into a file system made up of multiple nodes running on commodity hardware using the HDFS (Hadoop Distributed File System). A node acts as the facilitator and another node acts as a job tracker.

The Hadoop ecosystem is given below:



- **MapReduce:** It is a programming model withing the Hadoop framework that is used to access big data stored in the Hadoop File System.
- **Hive & Drill:** Analytical SQL type of query in Hadoop which can be used to look for information from millions of rows.
- **Mahout:** It is a platform where one can learn all the ML algorithms. It has its own library which can be used to run ML models.
- **PIG:** It is a scripting language just like Python.
- **HBASE:** It is a NoSQL database.
- **Zookeeper & Ambari:** It is a management and coordination system which can enable one to manage all the platforms of Hadoop.
- **Spark:** It is a data processing engine which is one of the fastest processing engines. It is 100 times faster than MapReduce.
- **Kafka & Storm:** It is used for streaming data meaning that the continuous inflow of data is managed using Kafka.
- **Oozie:** It is a scheduling tool which allows cluster administrators to build complex data transformations out of multiple component tasks (Job Scheduling).
- **Yarn:** It is the resource manager of Hadoop which is responsible for allocating system resources to various applications running in a Hadoop cluster and scheduling tasks to be executed on different cluster nodes.
- **HDFS:** It is the storage system of Hadoop. It is short for Hadoop Distributed File System.
- **Flume:** It can manage all the unstructured and semi-structured data preprocessing.
- **Sqoop:** It can store all the structured data.
- **Solar, Lucence:** Used for searching and Indexing.

Features of Hadoop Distributed File System:

- It is suitable for the distributed storage and processing.

- Hadoop provides a (LINUX) command interface to interact with HDFS.
- The built in servers of Namenode and Data Node help users to easily check the status of the cluster.
- Streaming access to the file system data.
- HDFS provides file permissions and authentications.

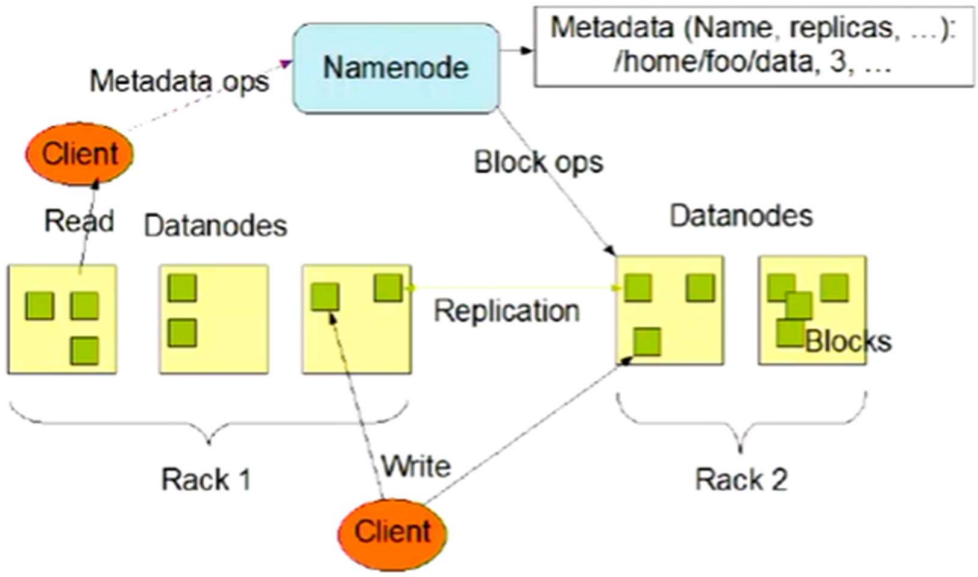
Goals of HDFS:

- HDFS includes many commodities hardware which causes failure of components to be more frequent. HDFS should have mechanisms for quick and automatic fault detection and recovery. **(Fault detection and recovery)**
- HDFS should have hundreds of nodes per cluster to manage the applications having huge datasets. **(Huge datasets)**
- Where huge datasets are involved, it reduces the network traffic and increases the throughput. **(Hardware at data)**

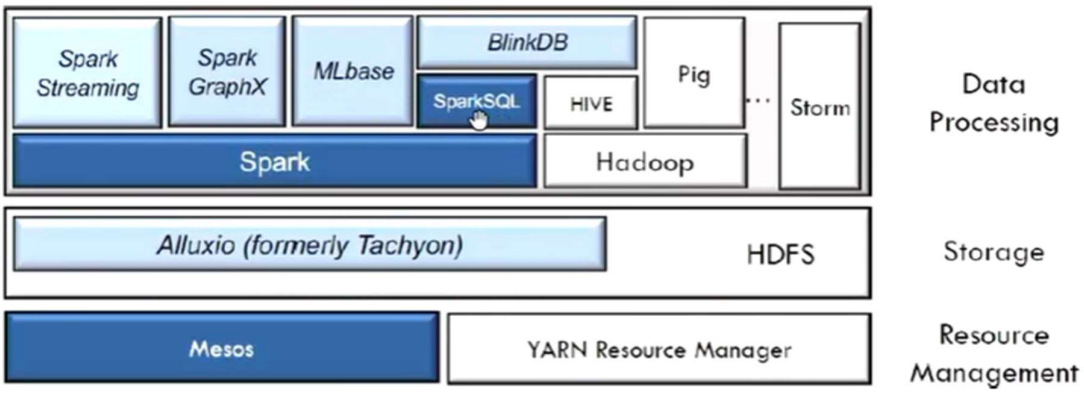
Advantages of Hadoop

- It can capture, store and process huge volumes of data in a timely manner.
- It can combine data quickly and at a reasonable cost.
- It can process data quickly as it is captured.
- It has good data governance with security, privacy, and access controls.

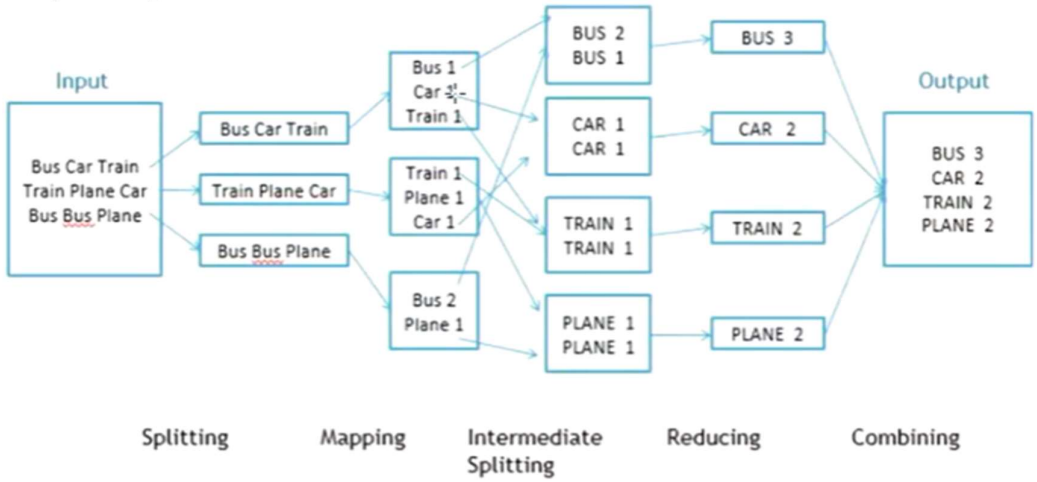
Hadoop Architecture



Information from the File system comes in through the Name node (Also called Parent Node, Cluster Node). Namenode send this information to the Data node (Also called the Child node). There can be only 1 Name node and multiple data nodes. The Name node will contain the metadata and requires less storage and high computational resources. It is the main node, and it doesn't store the actual data. The Data Node creates a separate block inside each node and replicates in into different blocks to prevent loss of data. Each block can of different sizes. There are certain number of Data nodes which can be allocated to the racks. When a user is searching for a particular information, the nearest block containing the information will be returned.



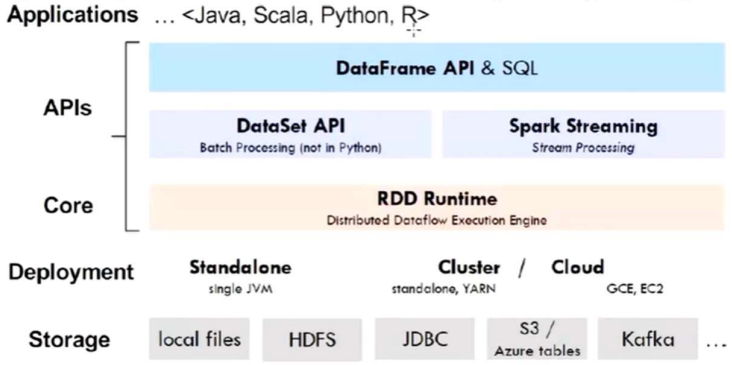
Working of MapReduce



The input is split into smaller components, mapped, has another intermediate splitting based on index numbers, reduced, and then combined. MapReduce has the capability of reducing the data based on the labelling of the data.

Apache Spark

It is a new framework that can quickly perform processing tasks on very large datasets and can also distribute data processing tasks across multiple computers. The execution model supports batch, streaming, and interactive computations. It is easy to develop sophisticated algorithms using python and Scala shell scripts. It provides high level abstractions for graph-based models and ML algorithms as well. It is compatible with the existing open-source ecosystem (Hadoop/HDFS) as well. The architecture of Apache Spark is given below:



Spark creates an environment/platform (called spark context) inside the temporary memory and retrieves the data and processes it inside the memory. This makes the process very fast. Apache Spark allows for a standalone (single) deployment or a cluster/cloud (multiple) deployment as well. There are 3 ways to deploy Apache Spark:

- Standalone deployment

  The cluster is deployed on a single PC. Spark occupies the place on top of HDFS and space is allocated for HDFS. Spark and MapReduce will run side by side to cover all spark jobs on the cluster.
- Hadoop Yarn deployment

  Spark runs on Yarn without any pre-installation or root access. Integrates spark into Hadoop ecosystem and allows other components to run on top of the stack.
- Spark in MapReduce deployment

  Spark job in addition to standalone deployment. User can start spark and uses its shell without admin access.

Apache Spark components:

- **Spark SQL:** Built on top of Spark core to support structured data. To query SQL and Hive query language.
- **Spark Streaming:** Makes the system scalable and fault tolerant. It provides for fast scheduling.
- **Mlib Machine Learning:** It contains Machine Learning algorithms and is 9 times faster than disk-based systems.
- **GraphX Graph Processing:** It can manipulate graphs and performs graph parallel computations. It supports subgraphs, join vertices, and aggregate messages.
- **Spark Core:** The heart of Spark as it performs the core functions of task scheduling, fault recovery, interacting with storage and memory, etc.

Some Questions that can come in exams (Based on what he said in lectures)

1. How many types of downfalls and starting points, starting from the beginning, exists in the Gartner Hype Cycle for Analytics and Business Intelligence?
   a. Explain all the points and trends in the Gartner Hype Cycle.
2. What is the high-level architecture of Business Intelligence in an Organisation?
   a. Explain the high-level architecture of Week 1 page 31.
3. Who are the audience of the analysed/visualized data?
   a. All the stakeholders starting from low level to high level such as team leaders, managers, executive directors, CEO, Stakeholders, etc.
4. Who manages the data warehouse?
   a. Data engineers as the manage and prepare the data.
5. What are the advantages of descriptive analytics?
   a. Given in Page 8 of Thursday 1pm to 3:30pm lecture slides of Week 2.
6. What are the disadvantages of descriptive analytics?
   a. Given in Page 9 of Thursday 1pm to 3:30pm lecture slides of Week 2.
7. How can data mining turn a large amount of data into knowledge that can help meet a current global challenge?
   a. Hinted in Page 5 of Thursday 10am to 12pm lecture slides of Week 3.
8. What are the different processing steps involved in Data mining process to get meaningful information from unstructured data?
   a. Answer is Page 7 of Thursday 10am to 12pm lecture slides of Week 3.
9. What are the 3 ways of Spark Deployment?
   a. Standalone, Hadoop Yarn, Spark in MapReduce.

10. What is the main functionality of Spark standalone?
    a. Spark and MapReduce run side by side to cover all spark jobs. It is placed on top of HDFS and space is allocated for HDFS.
11. What is the main functionality of Spark Hadoop Yarn?
    a. It runs on yarn without pre-installation or root access. It integrates spark into Hadoop ecosystem and allows other components to run on top of the stack.
12. What is the main functionality of Spark in MapReduce?
    a. Spark job in addition to standalone deployment. User can start spark and uses its shell without any admin access.