Data Analytics

(KCS-051)

Unit : 1

Introduction to Data Analytics

Syllabus

Introduction to data analytics: Sources and nature of data, classification of data (structured, semi structured, unstructured), characteristic of data, Introduction to big data platform, Need of data analytics, Evolution of analytic scalability, Analytic process and tools, analysis vs reporting, mordern data analytic tools, application of data analytics.

Data Analytics Lifecycle: Need, Key roles for successful analytic projects, various phases of data analytics life cycle - discovery, data preparation, model planning, model building, communication results, operationalization.

Introduction to data Analytics.

Sources and nature of data

Data

→ Data is a fact or figures obtained from experiments or surveys, used as a basis for making calculations or drawing conclusions.

→ For example, text, images and sound in a form that is suitable for storage in or processing by a computer.

Sources of data

There are two types of sources of data available.

1. Primary source of data
2. Secondary source of data

Primary source of data :

* The data which is raw, original and extracted directly from the official sources is known as primary data.

* This type of data is directly collected by performing techniques such as questionaries, interviews and surveys.

* Data collected must be according to the demand and requirements of the target audience on which analysis is performed otherwise it would be a burden in the data processing.

Methods for collecting primary data

→ Interview method.
→ Survey method.
→ Observation method
→ Experimental method.

Experimental method : The experimental method is the process of collecting data through performing experiments, research and investigation. The most frequently used experiment methods are CRD, RBD, LSD, FD.

CRD - Completely Randomized Design

CRD is a simple experimental design used in data analytics which is based on randomization and replication. It is mostly used for comparing the experiments.

RBD - Randomized Block Design

RBD is an experimental design in which the experiments is divided into small units called blocks. Random experiments are performed on each of the blocks and results are drawn using a technique known as Analysis of variance (ANOVA). RBD is originated from the agriculture sector.

LSD - Latin Square Design

LSD is an experimental design that is similar to CRD and RBD blocks but contains rows and columns. It is an arrangement of N X N squares with an equal amount of rows and columns which contains letter that occurs only once in a row. Hence the differences can be easily found with fewer errors in the experiments. Sudoku puzzle is an example of Latin square design.

Factorial Design - FD

FD is an experimental design where each experiment has two factors each with possible values and on performing trail other combinational factors are derived.

Secondary Source of data

* Secondary data is the data which has already been collected and reused again for some valid purpose.
* This type of data is previously recorded from primary data and it has two types of sources named internal source and external source.

Internal source

These types of data can easily be found within the organization such as market record, sales record, transaction customer data, account resources, etc. The cost and time consumption is less obtaining internal sources.

External source

This data which can't be found at internal organization and can be gained through external third party resources is external source of data. The cost and time consumption is more because this contains a huge amount of data.

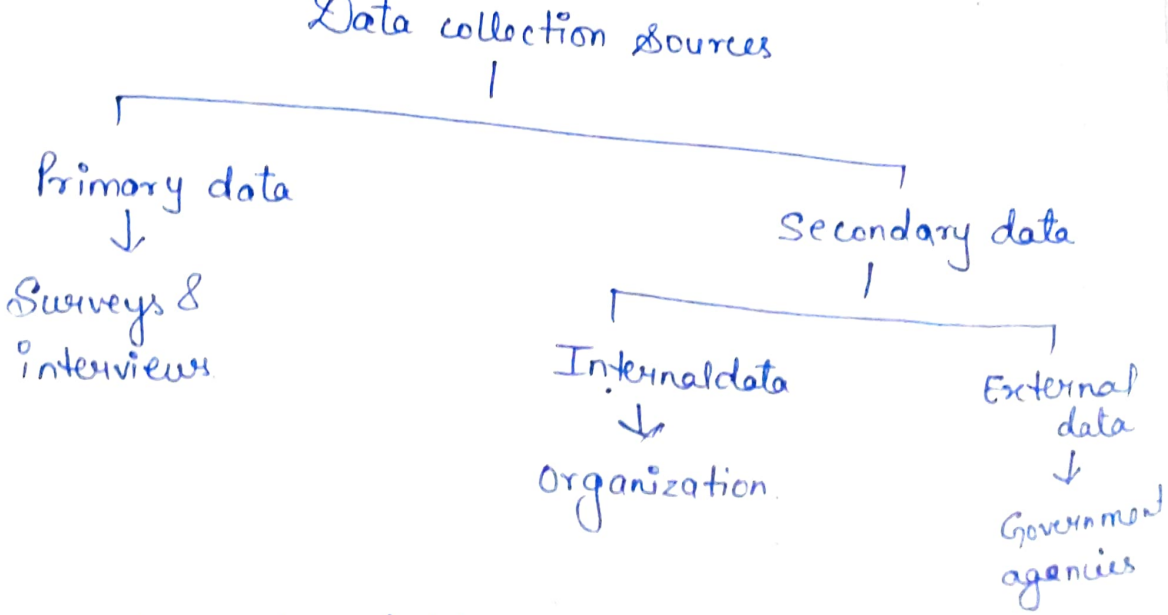
Examples - Government publications, News publications, Syndicate services and other non-governmental publications.

Other sources of data

→ Sensor data : With the advancement of IoT devices, the sensors of these devices collect data which can be used for sensor data analytics to track the performance and usage of products.

→ Satellites data : Satellites collect a lot of images and data in terabytes on daily basis through surveillance cameras which can be used to collect useful information.

→ Web traffic : Due to fast and cheap internet facilities many formats of data which can be uploaded by users on different platforms can be predicted and collected with their permission for data analysis. The search engines also provide there data through keywords and queries searched mostly.



classification of data collection sources.

Nature of data

The nature of data is classified into four categories.

→ Nominal data
→ Ordinal data
→ Interval data
→ Ratio data

Nominal data

The nominal scale is used for assigning numbers as the identification of individual unit. For example, the classification of journals according to the discipline they belong to may be considered as nominal data. If numbers are assigned to describe the categories, the numbers represent only the name of the category.

Ordinal data

It indicates the ordered or graded relationship among the numbers assigned to the observation made. These numbers cannot ranks of different categories having a relationship in a definite order.

for example, to study the responsiveness of a library staff a researcher may assign '1' to indicate poor, '2' to indicate average, '3' to indicate good and '4' to indicate excellent. The numbers 1, 2, 3, 4 in this case are set of ordinal data.

The ordinal data show the direction of the difference and not the exact amount of difference.

Interval data

Interval data are ordered categories of data and the differences between various categories are of equal measurements.

for example, we can measure the IQ of a group of children. After assigning numerical value to the IQ of each child, the data can be grouped with the interval of 10 like 0 to 10, 10 to 20 and so on.

→ Ratio data

Ratio data are the quantative measurement of a variable in terms of magnitude. In ratio data, we can say that one thing is twice or thrice of another..

For example, ~~measurents~~ measurements involving weight, distance, price etc.

Classification of data

Collection of information stored in a particular file represented as form data.

The data are classified into three forms of data.

→ Structured form
→ Unstructured form
→ Semi structured form.

Structured form : Any form of relational database structure where relation between attributes is possible. That there exists a relation between rows and columns in the database with a table structure. Eg. using database programming languages. (sql, oracle, mysql etc)

Unstructured form : Any form of data that does not have predefined structure is represented as unstructured form of data. Eg: video, images, comments, posts, few websites such as blogs and wikipedia.

Semistructured data : ~~Data~~ Does not have form tabular data similar to RDBMS. ~~Re~~ Predefined organized formats available. Eg: csv, xml, json, txt file with tab seperator etc.

Characteristics of Data

→ Accuracy
→ Completeness
→ Reliability
→ Relevance
→ Timeliness.

Accuracy → Data accuracy refers to error-free records that can be used as a reliable source of information.

Completeness → Data completeness refers to the comprehensiveness or wholeness of data. There should be no gaps or missing information for data to be truly complete.

Reliability → Data reliability means that data is complete and accurate and it is a crucial foundation for building data trusts across the organization.

Relevance → Data relevance assesses whether the information can serve its purpose in a particular contexts.

Timeliness → Data timeliness refers to the how up to date the information is.

Introduction to Big Data Platform.

Big data

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deals with data sets that are too large, or complex to be deal with by tradilional data processing application software.

Types of big data

→ Structured
→ Unstructured
→ Semi-Structured

Characteristics of big data

The are 5 characteristics of big data are as follows:-

→ Volume
→ Variety
→ Veracity
→ Value
→ Velocity

Volume

* The name 'big data' itself is related to a size which is enormous.
* Volume is a huge amount of data.
* To determine the value of value, size of data plays a very crucial role. If the volume of data is very large then it is actually considered as a 'big data'. This means whether a particular data can be actually be considered as a big data or not, is dependent upon volume of data.

* Example: In the year 2016, the estimated global mobile traffic was 6.2 exabytes (6.2 billion GB) per month. Also by the year 2020 we will have almost 40,000 Exabytes of data.

Velocity

* Velocity refers to the high speed of accumulation of data.
* In big data velocity data flows in from sources like machines, networks, social media, mobile phone etc.

* There is a massive and continous flow of data. This determines the potential of data how fast the data is generated and processed to meet the demands.

* Sampling data can help in dealing with the issue like 'velocity'.

* Example. There are more than 3.5 billion searches per day are made on google. Also, facebook users are increasing by 22% (approx) year by year.

* Variety
  * It refers to nature of data that is structured, semistructured, unstructured data.
  * It also refers to heterogeneous sources.
  * Variety is basically the arrival of data from new sources that are both inside and outside of an enterprise. It can be structured., semi-structured and unstructured.

Veracity (Truthful)

* It refers to inconsistencies and uncertainity in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.
* Big data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources.
* Example, data in bulk would create confusions whereas less amount of data could convey half or incomplete information.

Value

* The bulk of data having no value is no good to the company, unless you turn it into something useful.
* Data in itself is of no use or importance but it needs to be converted into something valuable to extract information.

Advantages of big data

→ Opportunities to make better decisions
→ Increasing productivity and Efficiency.
→ Reducing costs
→ Improving customer service and customer experience.

→ Fraud and Anomaly detection
→ Greater Agility and speed to market

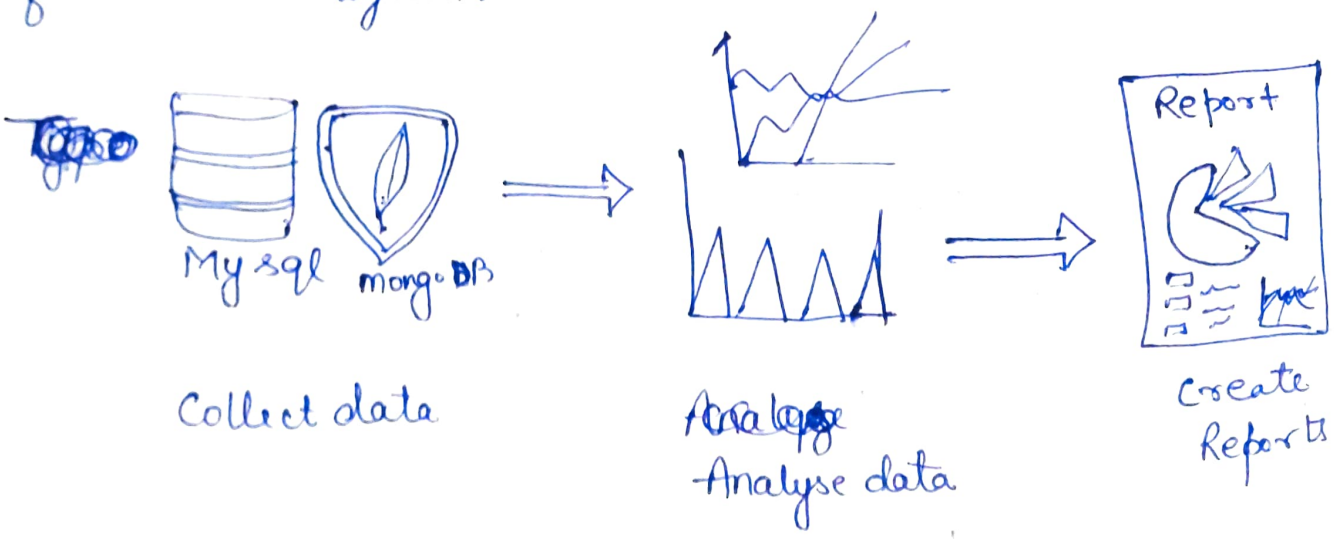
Disadvantages of Big Data

→ Questionable data quality
→ Heightend security risks
→ Compliance headaches
→ Costs and infrastructure issues
→ Big data skills shortage.

Data Analytics.

Data Analytics

Data Analytics is the science of analyzing raw data in order to make conclusions about that information. This information can be used to optimize processes to increase the overall efficiency of business or system.



Data Analytics.

Types of Data Analytics

1. Descriptive analytics
2. Predictive analytics
3. Prescriptive analytics.
4. Diagonistic analytics.

Descriptive analytics : In descriptive analytics the result is always going lead with the probability among 'n' numbers of options where each option has an equal chance of probability.

Eg: (observation, case-study, surveys).

Predictive analytics : This type of analytics deals with predicting past data to make decisions based on certain algorithms. In case of a doctor the doctor questions the patient about the past to correct his illness through already existing procedures.

Eg: healthcare, sports, weather, insurance, social media analysis.

Prescriptive analytics : Prescriptive analytics works with predictive analytics, which uses data to determine near-term outcomes. Prescriptive analytics make use of machine learning to help businesses decide a course of action based on computer program's predictions.

Eg: healthcare, banking.

Diagonistic analytics : This focusses more on why something happened. This involves more diverse data inputs and bit of hypothesizing.

Need of Data Analytics.

→ Gather hidden insights
→ Generate Reports
→ Perform market analysis
→ Improve business requirements.

Gather hidden insights : Hidden insights from data are gathered and then analyzed with respect to business requirements.

Generate Reports : Reports are generated from the data and are passed to the respective teams and individual to deal with further actions for high rise in business.

Perform Market Analysis : Market analysis can be performed to understand the strengths and the weaknesses of competitors.

Improve business requirements : Analysis of data improving business to customer requirements and experiences.

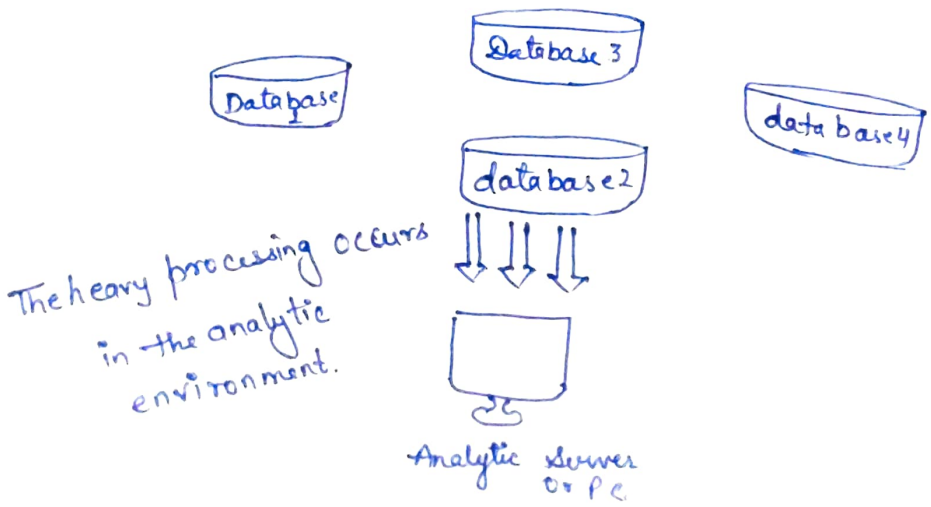
Evolution of Analytic Scalability

Scalability : The ability of a system to handle increasing amount of work required to perform its task.

- The increase in data storage ability has grown in recent years to accomodate the need for big data

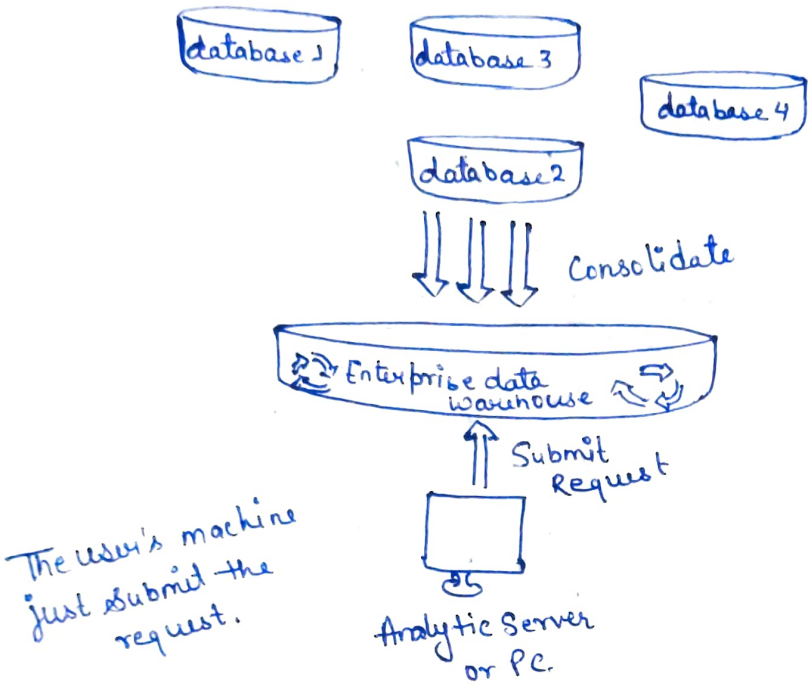
Traditional Analytic Architecture

- we had to pull all together into a separate analytics environment to do analysis.



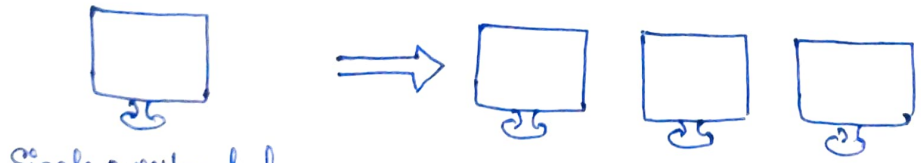
Mordern in Database Architecture

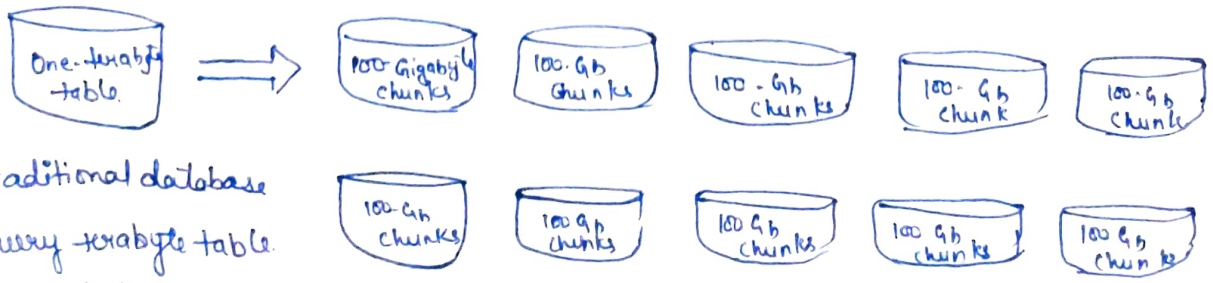
- The processing stays in the database where the data has been consolidated



Massively Parallel Processing (MPP)

- An MPP database breaks the data into independent chunks with independent disk and CPU.



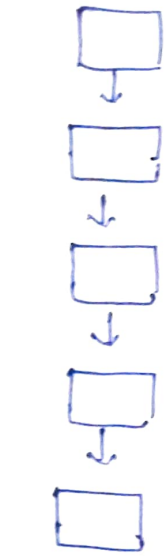


A traditional database will query terabyte table one row at time.

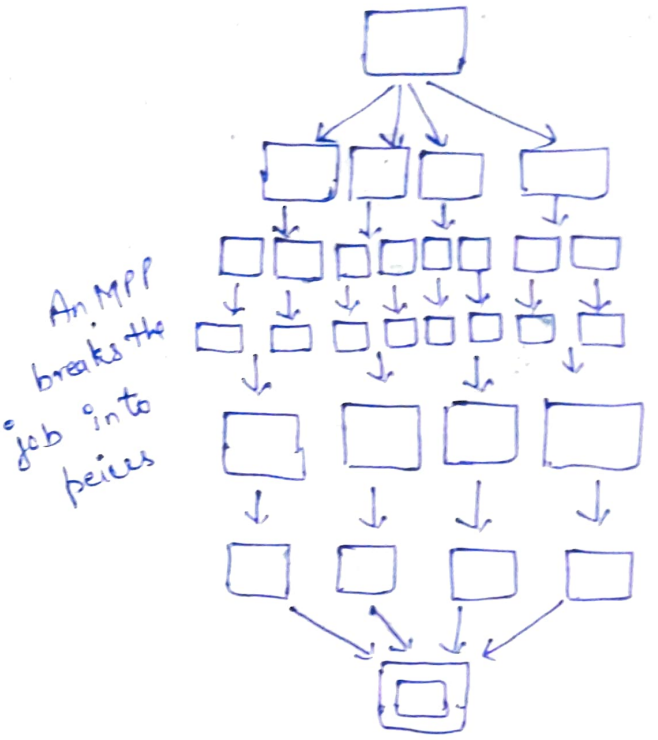
10 Simultaneous 100-gigabyte queries

Concurrent Processing

- An MPP system allows the different sets of CPU and disk to run the process concurrently.



Single threaded process



Parallel process

- MPP system build in redundency to make recovery easy.
- MPP system have resource management tools
  → Manage the CPU and disk space
  → Query optimizer.

Cloud Computing

- McKinsey and Company paper from 2009
  - Mask the underlying infrastructure from the user
  - Be elastic to scale on demand
  - On a pay-per-use-basis.

- National Institute of Standards and Technology (NIST)
  - On demand self service
  - Resource pooling
  - Rapid elasticity
  - Measured service

Two types of cloud Environment

1. Public cloud

→ The services and infrastructure are provided off-site over the internet.
→ Greatest level of efficiency in shared resources.
→ Less secured and more vulnerable than private clouds.

2. Private cloud

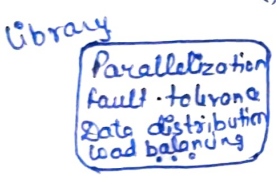
→ Infrastructure operated solely for single organization.
→ The same features of a public cloud
→ Offer the greatest level of security and control
→ Necessary to purchase and own the entire cloud infrastructure.

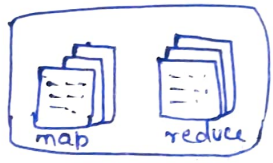
Grid Computing

- The federation of computer resources to reach a common goal.

Map Reduce

- A parallel programming framework.





→ Map function
- Processing a key/value pairs to generate a set of intermediate key/value pairs.

→ Reduce function
- Merging all intermediate values associated with the same immediate key.

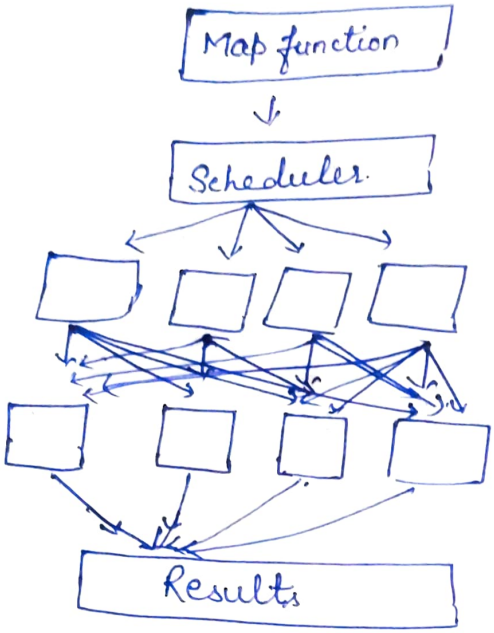
How map reduce works

- Let's assume there are 20 terabytes of data and 20 mapreduce server nodes for a project

1. Distribute a Terabyte to each of the 20 nodes using a simple file copy process.
2. Submit two programs (Map, Reduce) to the scheduler.

3. The map program finds the data on disk and executes the logic it contains

4. The result of the map step are then passed to the reduce process to summarize and aggregate the final answers.



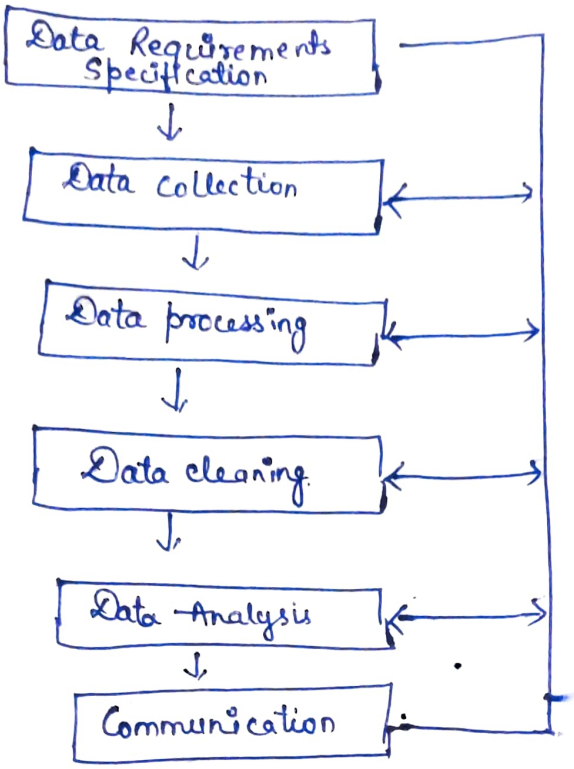
Analytic process and Analytic tool

Analytic process:

Data Analytic process is a process of collecting, transforming, cleaning and modelling data with the goal of discovering the required information. The results so obtained are communicated, suggesting conclusions and supporting decision making.

Data Analysis process consists of the following phases that are iterative in nature -

- → Data Requirements Specification
- → Data collection
- → Data processing
- → Data cleaning
- → Data Analysis
- → Communication



Analytic tools

1. Spreadsheet
   - → Microsoft excel
   - → Spreadsheets

2. Database
   - → Relational
   - → Column
   - → Document
   - → Graph.

3. Programming Languages
   - → R and Python

4. Self service data virtualization
   - → Tableau
   - → Power BI
   - → Qlik sense
   - → AWS Quick Sight

5. Big data tools
→ Hadoop
→ Data lakes
→ Apache spark

6. Cloud
→ AWS
→ Google cloud
→ Microsoft Edge.

Analysis vs Reporting

Parameters	Reporting	Analytics
Purpose	Shows what is happening	Explain why it is happening
Tasks	• Organizing • Formatting • Summarizing	• Questioning • Interpreting • Exploring
Results	Results are pushed to user for Review	Users pull results to answer questions.
Value	Translates data into information	offers recommendation to drive actions.

Application of data analytics

1. Security
2. Transportation
3. Fraud and Risk detection
4. Manage Risks
5. Proper spending
6. Customer interactions
7. City planning
8. Healthcare
9. Travel
10. Energy management
11. Internet /websearch
12. Digital advertisement

Data Analytics Lifecycle:

Phases of data analytics lifecycle.

The data analytic life cycle is designed for big data problems and data science projects. The cycle is iterative to represent real project.

The phases which involved in data analytics life cycle are:-

1. Discovery
2. Data preparation
3. Model planning
4. Model building
5. Communication Results
6. Operationalize

Discovery

→ The data science team learn and investigate the problem
→ Develop context and understanding
→ Come to know about data sources needed and available for the project.
→ The team formulates initial hypothesis that can be later tested with data.

Data preparation

→ Steps to explore, preprocess and condition data prior to modelling and analysis,
→ It requires the presence of an analytic sandbox, the team execute, load and transform, to get into the sandbox.
→ Data preparation tasks are likely to be performed multiple times and not in predefined order.
→ Several tools commonly used for this phase are-Hadoop, Alpine miner, OpenRefine etc.

Model planning

→ Team explores data to learn about relationship between variables. Subsequently, select keys variables and the most suitable models.

→ In this phase, data science team develop data sets for training, testing and production purposes.

→ Team builds and executes models based on the work done in the model planning phase.

→ Several tools commonly used for this phase are - Matlab, STASS

Model building

→ Team develops datasets for testing, training and production purposes.

→ Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.

→ Free or open-source tools - Rand PL/R, Octave, WEKA.

→ Commercial tools - Matlab, STASTICA

Communication Results

→ After executing a model team need to compare outcomes of modelling to criteria established for success and failure.

→ Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning assumptions.

→ Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders

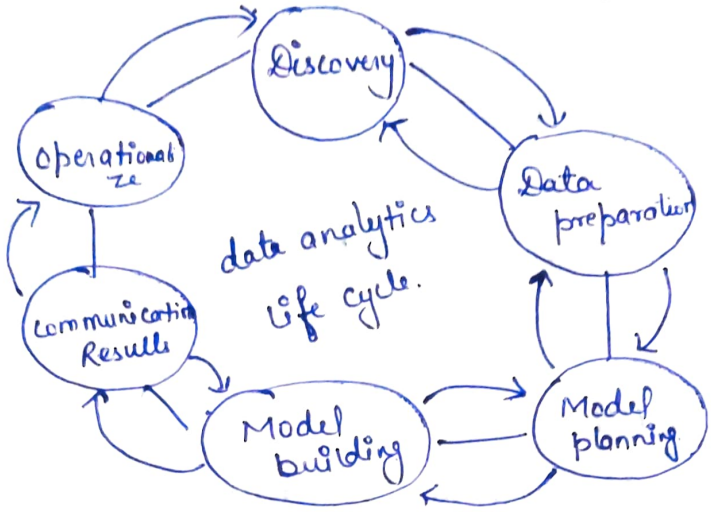
Operationalize

→ The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.

→ This approach enables team to learn about performance related constraints of model in production environment on small scale & nbsp, and make adjustments before full deployment.

→ The team delivers final reports, briefing, codes.

→ Free or open source tools - octave, weka, sql, Madlib.



Key roles for successful analytic project

1. Business User

→ The business user is the one who understands the main area of the project and is also basically benefited from the results.

→ This user gives advice and consult the team working on the project about the value of the results obtained and how the operation on the outputs are done.

→ The business manager, line manager, or deep subject matter expert in the project mains fulfills this role.

2. Project Sponsor

→ The project sponsor is the one who is responsible to initiate the project. Project sponsor provides the actual requirements for the project and presents the basic business issue.

→ He generally provides the funds and measures the degree of value from the final output of the team working on the project.

→ This person introduce the prime concern and brooms the desired output

Project Manager

→ This person ensures that key milestone and purpose of the project is met on time and of the expected quality.

Business Intelligence Analyst

→ Business intelligence analyst provides business domain perfection based on a detailed and deep understanding of the data, key performance indicators (KPIs); key matrise and business intelligence from a reporting point of view.

→ This person generally creates fascia and reports and knows about the data feeds and sources.

Database Administrator (DBA)

→ DBA facilitates and arrange the database environment to support the analytics need of the team working on a project.

→ His responsibilities may include providing permission to key databases or tables and making sure that the appropriate security stages are in their correct places related to the data repositories or not.

Data Engineer

→ Data engineer grasp deep technical skills to assist with tuning SQL queries for data management and data extraction and provides support for data intake into the analytic sandbox.

→ The data engineer works jointly with the data scientist to help build data in correct ways for analysis.

Data scientist

→ Data scientist facilitates with the subject matter expertise for analytical techniques, data modelling and applying correct analytical techniques for a given business issues.

→ He ensures overall analytical objectives are met.

→ Data scientist outline and apply analytical methods and proceed towards the data available for the concerned project.