Introduction to Data Analytics

* Data Analytics *

UNIT-I

Data: Data is a collection of information gathered by observations, measurements, research or analysis. They may consist of facts, numbers, names, figures or even description of things. Data is organised in the form of graphs, charts or tables.

Data Analytics: Data analytics is the process of collection, transformation, and organizing of data in order to draw conclusions, make predictions and drive informed decision making.
Data analytics helps you to make sense of the past and to predict future trends and behaviors rather than basing your decisions and strategies.

Why Data Analytics is important: (Need of DA):
Data analytics is very important in today scenario because it helps business growth and their performance. Data analytics is the process of analyzing raw data actionable insights, which are then used to inform and smart business decisions.
Data analytics help a business optimize its performance, perform more efficiently, maximize profit, or make more strategically-guided decisions. This is done by using software and other tools to gather and analyze raw data. Data analytics relies on a variety of software tools including spreadsheets, data visualization, data mining programs and open-source languages.

Scanned with OKEN Scanner

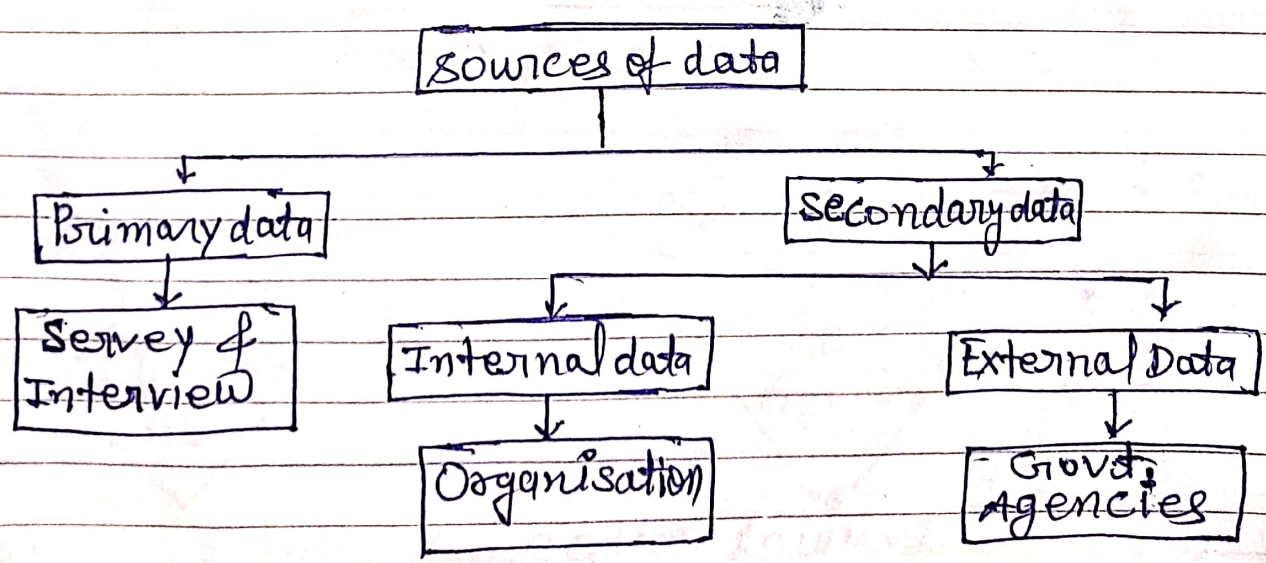
How to use Data Analytics:

- Reconize the business problem.
- Collect relevent data problem.
- Prepare the data for analysis.
- Analyze the data
- Establish analytics and models.
- Managed and optimize performance.

* Different sources of data for data Analysis:

* **Data Collection:** Data collection is the process of acquiring, collecting, extracting and storing the valuminous amount of data which may be in the structured or unstructured form like: text, video's, audio's, photo's XML-file records or other image files used in later stages of data analysis.

* Sources of data/Methods of data collection:



Methods of Data collection.

Methods of data Collection

- **Primary Data**
  - Personal Observation
  - Interview
  - Questionnaire/~~schedule~~ schedule
  - Survey
- **Secondary data**
  - Published data
  - Unpublished data.

• **Internal sources:** In an organization, when data is collected from its reports and records, it is known as internal sources of data.
Data collected and compiled through internal sources is called Internal data.
ex— A company publishes its annual report on profit and loss, total sales, Loan's etc.

• **External sources:** Refers to the data collected from ~~the~~ outside the organization.
Data collected and compiled through external sources is called external data.
ex— Tour and Travels company obtains information on "U.P Tourism" from U.P. transport corporation.
Data can be collected from primary and secondary sources:—

• **Primary Data:** The data which are collected for the first time by an individual or the group of individual from any institutions or organizations are known as primary sources of the data.

Scanned with OKEN Scanner

- **Secondary sources of data:** The data collected from any published or unpublished sources are called secondary sources of data.

Primary data:

- **observation Method:** Observation method is a method under which data from the field is collected with the help of observation by the observer or by the personally going to the field.

- **Interview Method:** The data collected during this process is through interviewing the target ~~audienc~~ audience by a person called the interviewer and the person who answers the interview is known as interview method.

Prepare Interview schedule
↓
Select subjects
↓
Conduct the interview
↓
Analyze and interpret data collected from the interview.

- **Questionnaire/schedule:** In this method, simple questions and their possible answers are written on plain paper and the respondents have to tick mark the possible answers from the given choices. Schedule very similar to questionnaire method. the main difference ~~between~~ is that that a schedule is filled by the enumerator who is specially appointed for the purpose.

Scanned with OKEN Scanner

- **Survey Method:** The survey method is a process of research where list of relevant questions is asked and answers are noted down in the form of text, audio or video. the survey method can be obtained in both online and offline mode. like through websites forms and emails.

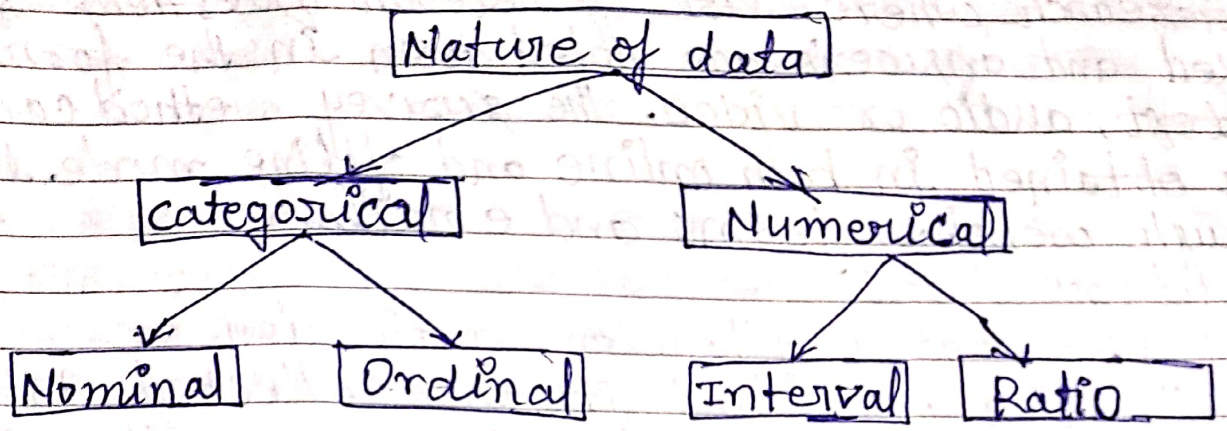
Secondary data:

- **Published data:** When an institution or organization published its own collected data (primary data) in public domain either in printed form or in electronic form, then these data are said to be secondary data in published form. example-
  - Government publications.
  - Semi-government publication.
  - International publication.
  - private publications
  - Newspaper and Magazine's.

- **Unpublished data:** Collected information in term of data or data observed through own experience by an individual or by an organization which is in unpublished form is known as unpublished source of secondary data. unpublished data is the statistical data obtained from several unpublished references. like:
  - Government Document
  - Semi government records.
  - private document.

Scanned with OKEN Scanner

Nature of data:



- **Categorical data:**

Categorical data represents 'characteristics. It is also called qualitative data, which means you can not express it as a numerical value. Therefore, you can not measure it. It involves variables such as words, symbols, pictures, which help you sort information based on category for example: gender, language etc.

1. **Nominal data:**

Nomial data is a type of qualitative data that is used to represent data into lable's based on different categories. they do not have any specific order. example:-

- Colours (red, blue, green etc.).
- Gender (Male, female etc.)
- Blood type (A, AB, O, B).
- language: (Hindi, English, Punjabi etc)

2. **Ordinal data:** Represents categories with a natural order or ranking, but the intervals between categories may not be equal such

as ratings or survey response. example:
Grades in exam [A, B, C, D --].
position in a competition [first, second, third --].
Times of day [Morning, Noon, Evening]

- Numerical Data:

Numerical data is a type of data that represents quantitative information that we can count and measure. they are also known as Quantitative data. It generall gives answers to "How many", "How much" etc example:-

- The number of apples.
- The number of students in a class.
- The height/weight of a person.

1. Interval data:

Interval data are measured and ordered with the nearest ~~value~~ item's but have no meaningfull zero point such as temperature measured in Celsius or Fahrenheit, time interval on a 12-hour clock (6am - 6pm).

2. Ratio data: Ratio Data represents numerical data, where there is a true zero point and both the intervals b/w values and Ratio of value are meaningful, such as weight, height or income.

[Classification of Data]

Data classification is the process of organizing data into categories that make it easy to retrieve, sort and store for future use. A well-planned data classification system makes essential data easy to find and retrieve.

Scanned with OKEN Scanner

Data can be classified into relevant categories which can be used or applied more efficiently. Data classification is important for data security and compilance and also for fulfilling different types of business or personal objectives.

- **Purpose of data classification:**

Some common business goals for these projects include the following:-

1. **Confidentiality:** A classification system safeguards highly sensitive data, such as customers, personal identifiable information (PII), including credit card numbers, social security numbers and other vulnerable data.
Establishing a classification system helps an organisation, focus on confidentiality and security policy requirements such as user permissions and encryption.

2. **Data Integrity:**
A system that focuses on data integrity will require more storage, user permissions and proper channels of access.
Integrity means that data is protected from unauthorised changes to ensure that it is reliable and correct.

3. **Data Availability:** Availability is important because it helps ensure that systems, application and data are functioning as they should and when they should. It also helps to ensure that authorised user have timely and reliable access to data and information services.

Scanned with OKEN Scanner

Features of data classification:

The main goal of organising data is to arrange the data in such a form that it becomes fairly available to user. So its basic features are following:-

- **Homogeneity:** The data items in a particular group should be similar to each other.
- **Clarity:** There must be no confusion in the positioning of any data item in a particular group.
- **Stability:** The data itemset must be stable, i.e any investigation should not affect the same set of a classifications.
- **Elastic:** One should be able to change the basis of classification as the purpose of classification changes.

[Types of data classification:]

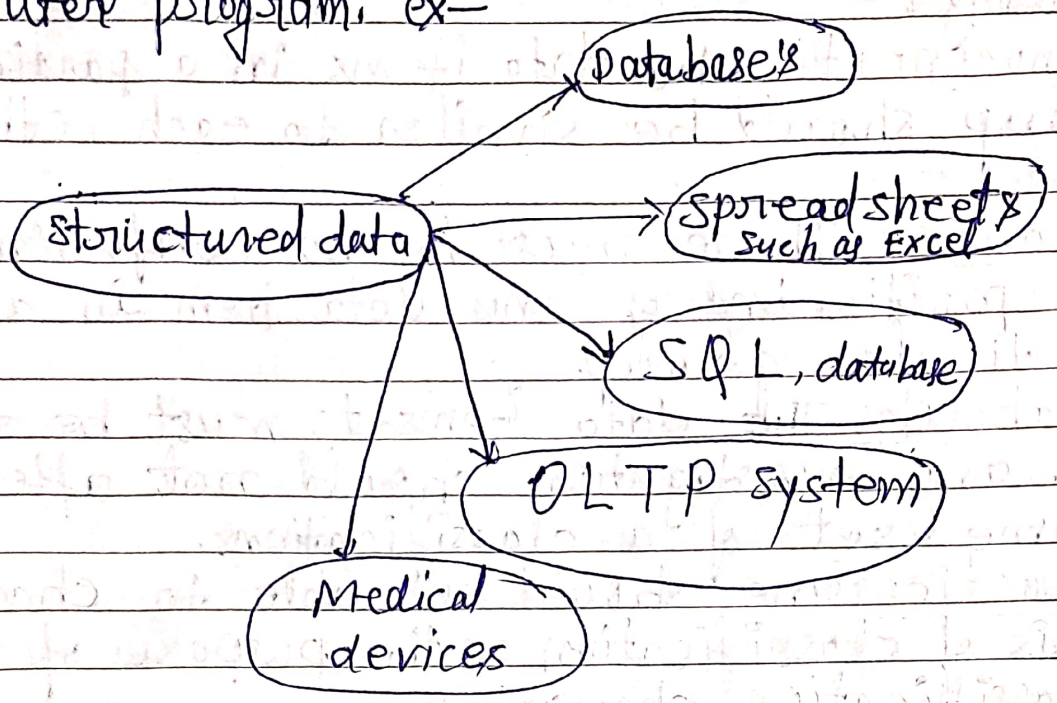
Data can be broadly classified into three types
- Structured data.
- Unstructured data.
- Semi-structured data.

- **Structured data:**

Structured data is created using a fixed schema and is maintained in a tabular format.
The elements in a structured data are addressable for effective analysis. It contains all the data which can be stored in the SQL database in a tabular format.

Scanned with OKEN Scanner

Structured data is the data, that conform's to a data model or schema, has a well-defined structure follow's a consistent order and can be easily accessed and used by a person or a computer program. ex-



"Sources of structured data"

Characteristics of structured data:

- Data conform's to a data model and has an easily identifiable structure.
- Data is stored in the form of row's and column's. Example- database.
- Entities in the same group have same attributes.
- Easy to access and query, so data can be easily used by other program's.
- Data elements are addressable, so efficient to analyse and process.

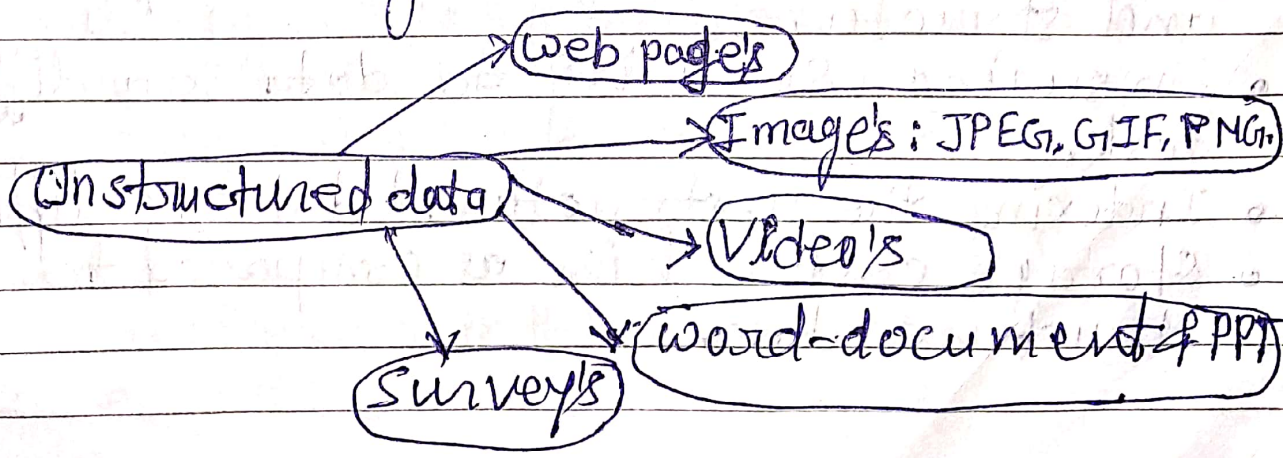
Scanned with OKEN Scanner

Advantages of structured data:

- Structured data has a well-defined structure that help's in easy storage and access of data.
- Data can be indexed based on text string as well as attributes. this makes search operation hassle-free.
- Data mining is easy i.e, knowledge can be easily extracted from data.
- Easily scalable in case, there is an increment of data.
- Ensuring security to data is easy.

Unstructured data:

It is defined as the data in which does not follow a pre-defined standard or you can say that it does not follow any organised format. This kind of data is also not fit for the relational database because in such a database you will see a pre-defined manner or organised way of data. About 80-90% data of an organization in this format this makes it very difficult and time-consuming to process and analyze unstructured data. examples:



Main source's of unstructured data:

- facebook,
- Instagram,
- Twitter
- Youtube.
- Whatsapp. etc.

* Characteristics of Unstructured data:

- Data neither conform's to a data model nor has any structure.
- Data can not be stored in the form of row's and column's as in database.
- Data does not follow any rules.
- Data lacks a particular format or sequences
- Data has no easily identifiable structure.

* Advantage's of Unstructured data:

- Data is very portable.
- Data is very scalable
- It can deal easily with the heterogeneity of sources.

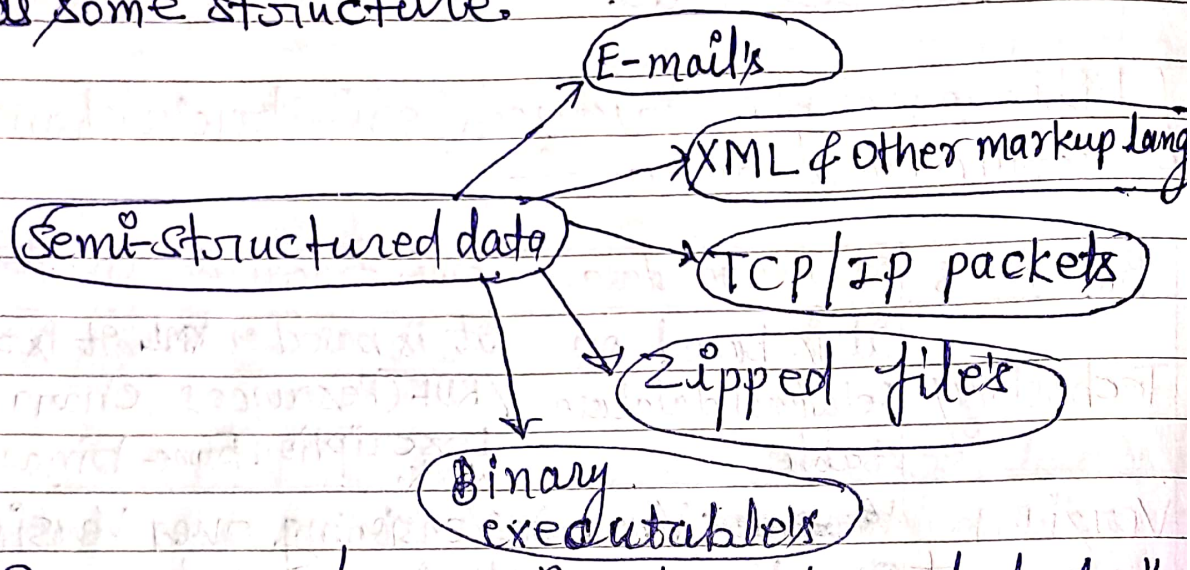
* Disadvantage's of unstructured data:

- It is difficult to store and manage unstructured data due to lack of schema and structure.
- Ensuring security to data is a difficult task.
- Indexing the unstructured data is difficult.
- Storage cost is high as compared to structure data.

Scanned with OKEN Scanner

• Semi-Structured data:

Semi-structured data is information that does not reside in a relational database but has some organisational properties that makes it easier to analyse. semi-structured data is data that does not conform's to a data model but has some structure.



"Sources of semi-structured data"

* Characteristics of semi-structured data:

- Data does not conform to a data model but has some structure.
- Data can not be stored in the form of row's and column's as in database.
- Due to lack of a well-defined structured, it can not use by computer programs easily.

* Advantages of semi-structured data:

- Flexible, i.e schema can be easily changed.
- Data is portable.
- It can deal easily with the heterogeneity of sources.

Scanned with OKEN Scanner

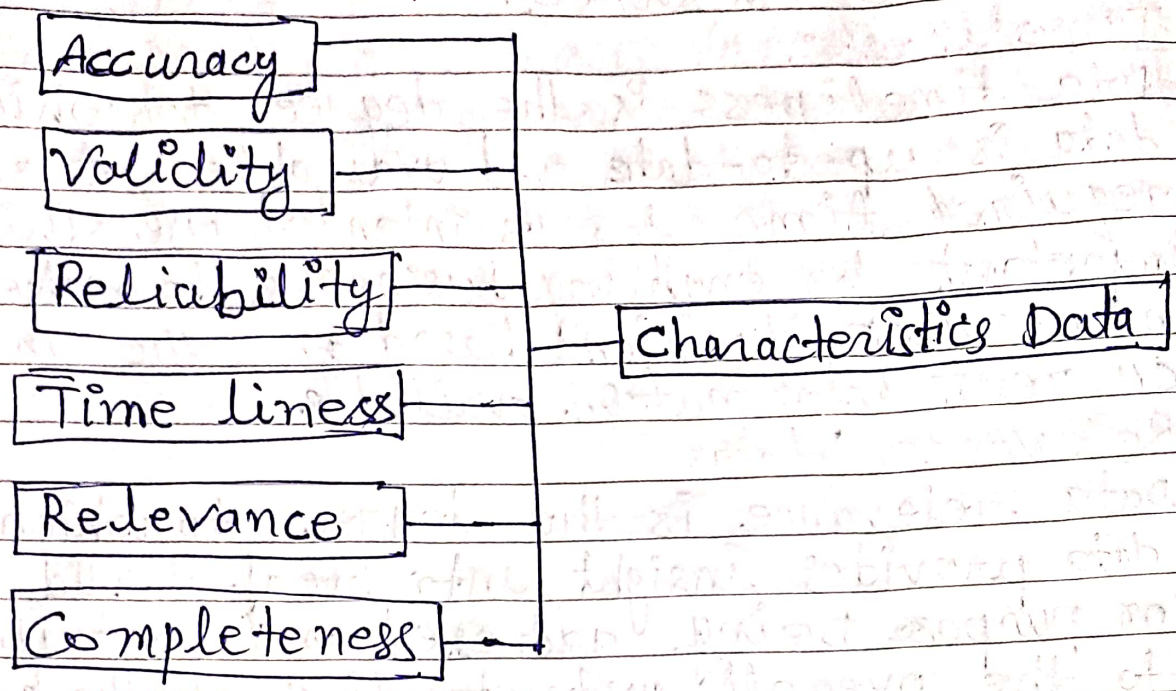
* Disadvantages of semi-structured data:

- Lack of fixed, rigid schema makes it difficult to store the data.
- Queries are less efficient as compared to structured data.
- Storage cost is high as compared to structured data.

[Differences b/w Structured, Semi-structured and Un-structured data :-]

Properties	Structured data	Semi-structured	Un-structured
Technology	It is based on Relational database table	It is based on XML / RDF (Resources Description Frame-work.	It is based on character and binary data.
Version Management	~~Versioning~~ Versioning Over tuples, row, tables.	Versioning over tuples or graph is possible.	Versioned as a whole.
Flexibility	It is schema dependent and less flexible	It is more flexible than structured data but less flexible than unstructured.	It is more flexible and there is absence of schema.
Scalability	It is very difficult to scale DB schema.	Its scalling is simpler than structured data	It is more scalable
Robustness	Very Robust	New, Technology, not very spread	—
Query Performance	structured query allow's complex joining.	Queries over anonymous nodes are possible.	Only textual queries are possible.

Characteristics of Data :



- **Accuracy:**
  Data should be sufficiently accurate for the intended uses and should be captured only once this characteristic refers to the exactness of the data. It can not have any erroneous elements and must convey the correct message without being misleading.
- **Validity:**
  Data should be recorded and used in ~~compliance~~ compliance with ~~requirement~~ the relevant requirements, including the correct application of any rules or definitions. this will ensure consistency between periods and with similar organizations, measuring what is intended to be measured.
- **Reliability:** Data reliability refers to the completeness and accuracy data as a measure of how well it can be counted.

Scanned with OKEN Scanner

on to consistent and free from errors across time and sources.

- **Timeliness:**

Data timeliness is the degree to which data is up-to-date and available at the required time for its intended use. this is important for enabling businesses to make quick and accurate decisions based on the most current information available.

- **Relevance data:**

Data relevance is the degree to which data provides insight into real-world problem or purpose being addressed and contributes to the overall understanding of the business.

- **Completeness:**

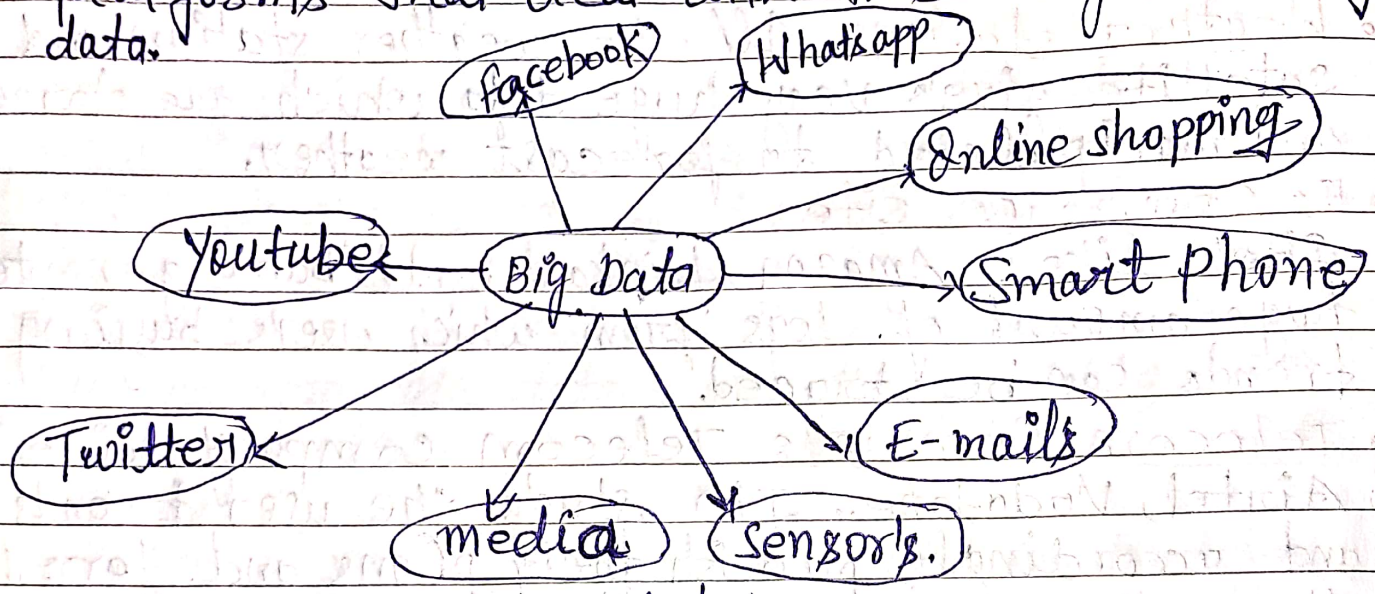
Data requirements should be clearly specified based on the organisational requirements and data collection processes matched to these requirements.

Introduction to Big Data Platform:

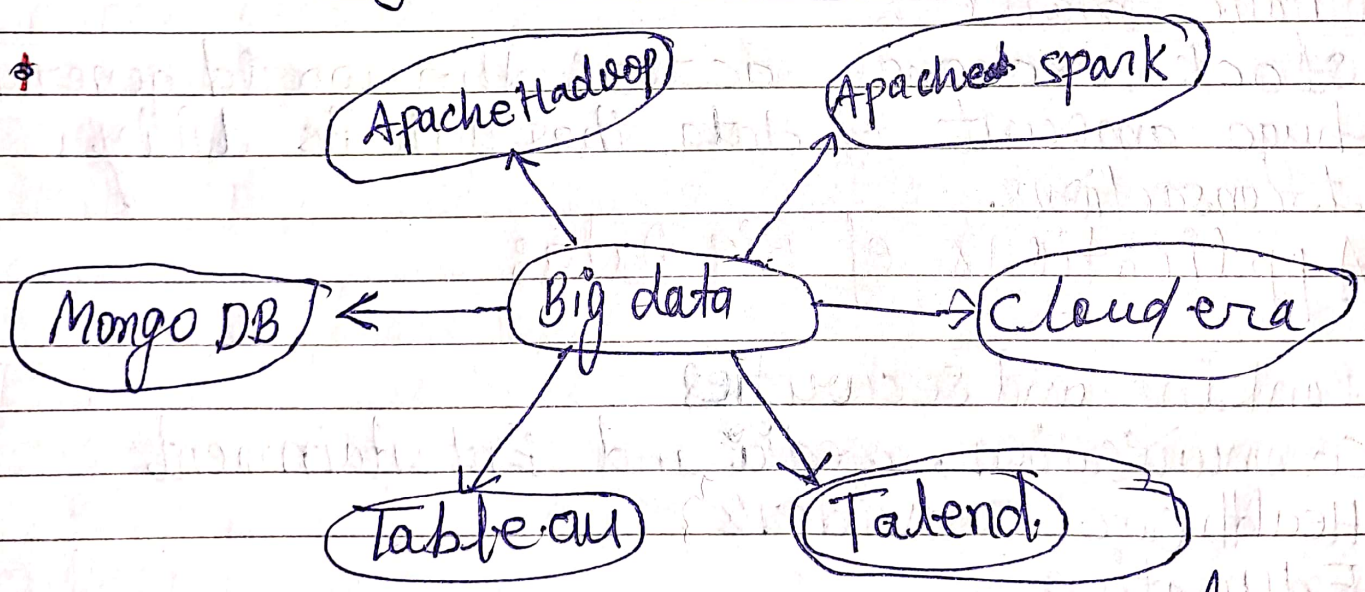
Big data is the collection of large and complex data which keeps growing exponentially with time and can not be processed using traditional methods. Data which are very large in size are called Big data.

Normally, we work on data of size KB, MB or maximum size GB and TB, But data which are petabyte (PB), exabyte (EB) i.e 10^{15} byte's.

size is called Big data. It is stated that almost 90% of today's data has been generated in the past three years. the size of large data can be range from several terabytes to petabytes and even Exabytes. thus, there is an inevitable need to develop platforms that deal with this large scale of data.



Big data platform



Big Data platform tools.

* Sources of Big data :

These data come's from many sources like :-

- **Social Networking sites:**
  facebook, Google, linkedIn all these sites generate huge amount of data on a day-to-day basis as they have billions of user's world-wide.

- **Weather station:** All the weather station and satellite give's very huge data which are stored and manipulated to forecast weather.

- **E- Commerce site:**
  Site's like: Amazon, flipkart, Alibaba generates huge amount of log's from which user's buying trends can be traced.

- **Telecom Company:** Telecom company like: Airtel, Vodafone, Jio study the user trends and accordingly publish their plans and for this they store the data of its millions users.

- **Share market:**
  Stock exchange across the world generates huge amount of data through its daily transactions.

* Application's of Big Data:

- Banking and Securities
- Communication, media and Entertainment
- Health care provider's
- Education
- Goverment
- Insurance
- Transportation.

Scanned with OKEN Scanner

Use of Big Data:

- Location tracking
- Precision Medicine
- Advertising
- Entertainment and Media.

Issues with Big Data:

there are mainly three issues with big data :-

- **Low Quality and Inaccurate data:**

Low-quality data or inaccurate data quality may lead to inaccurate results or predictions which does nothing apart from wasting time and effort effort of the individual.
To solve, to predict or to find new patterns from the data, the data must be of high quality and accurate.

- **Processing Large data sets:**

Due to a large amount of data, no traditional data management tool or software can directly process because the size of these large datasets are usually in terabytes which is difficult to process.

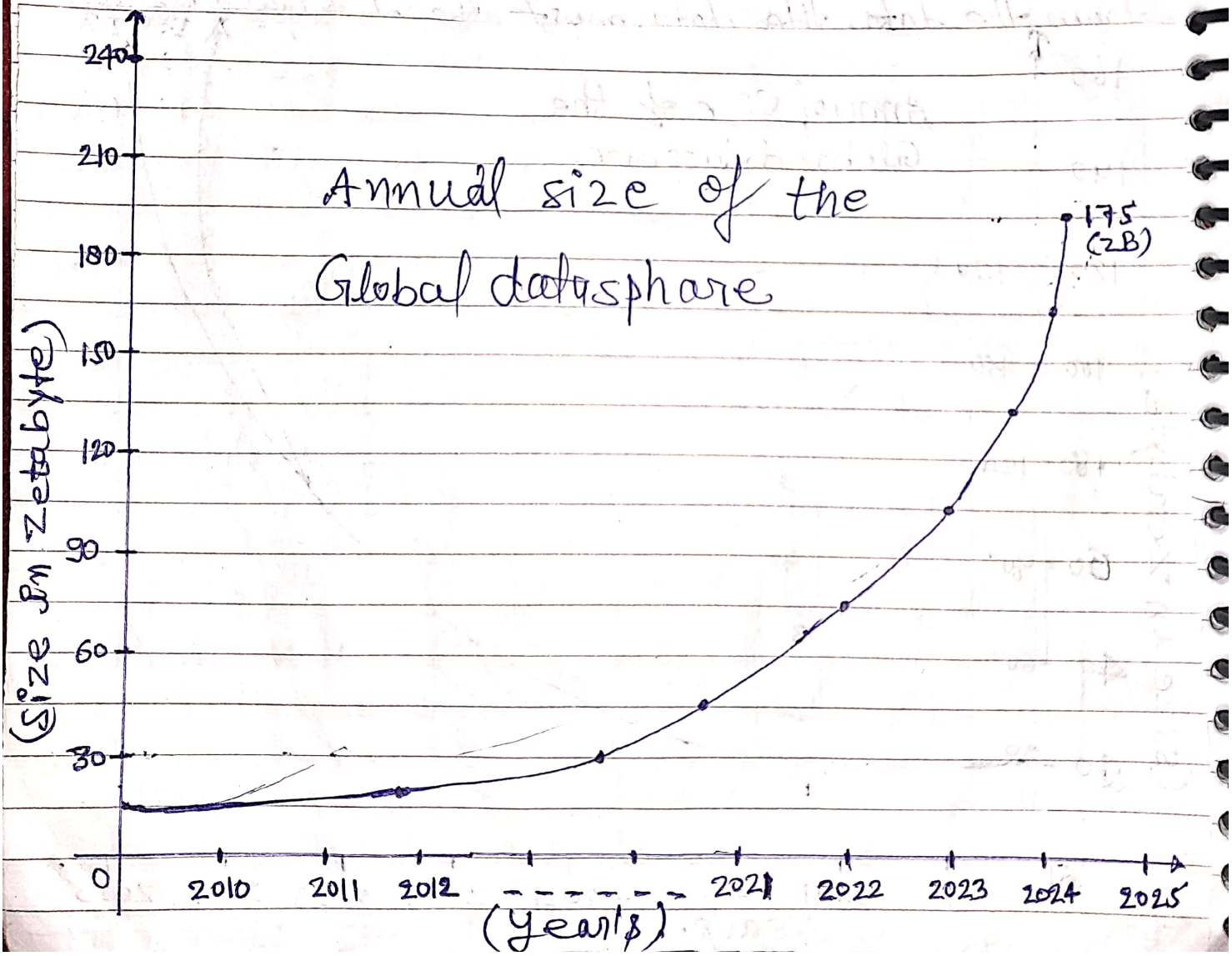
- **Integrating Data from a Variety of Sources:**

Data comes from various types of sources like social media, different websites, captured images/videos, customer log's, reports created by individuals, newspapers, e-mails etc. collecting and integrating various data which are of different types is basically a challenging task.

Scanned with OKEN Scanner

Characteristics of Big Data:

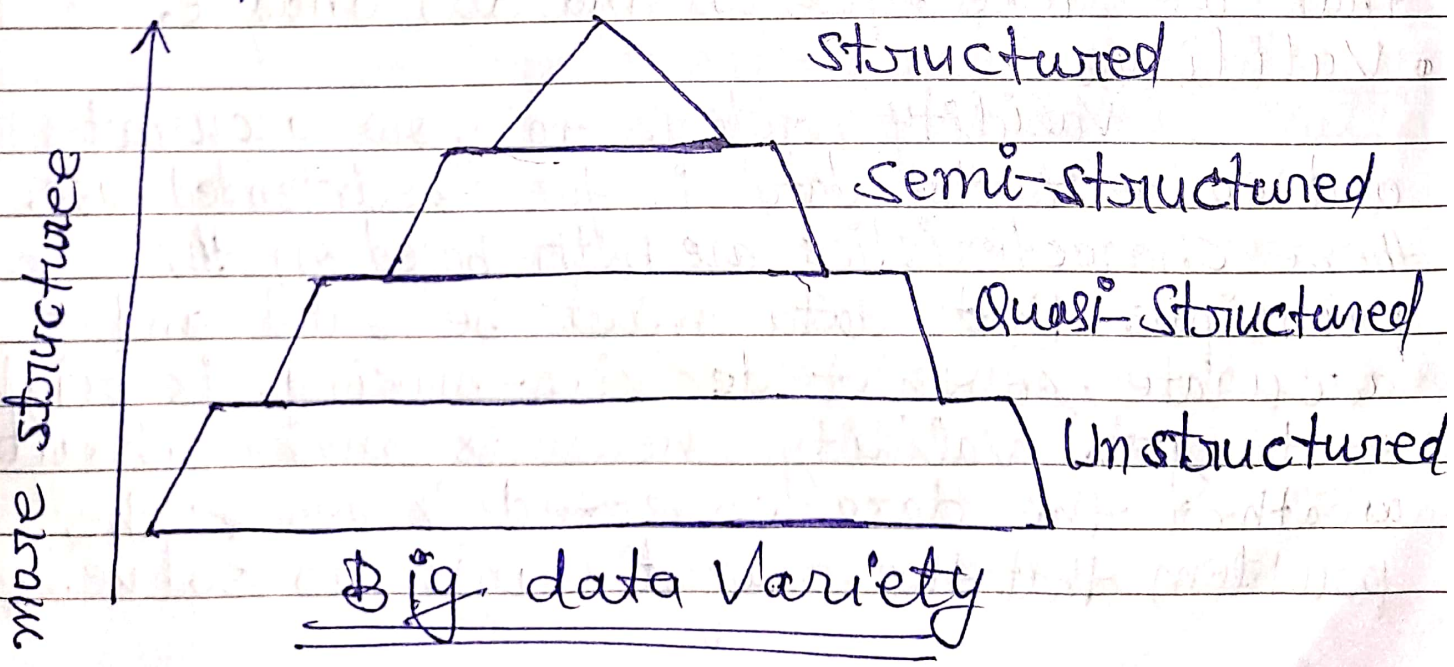
- **Volume:** Volume refers to the vast increase in the data growth. This is evident as more than 90% of the data, we encounter was produced recently. Big data is a vast 'volume' of data generated from many sources daily, such as business processes, machines, social media platforms, networks human interactions and many more. Volume is a huge amount of data. To determine the value of data, size of data, plays a very important role.



Scanned with OKEN Scanner

- **Velocity:** Velocity refers to the "high speed of accumulation of data". In Big data, velocity flows in the from sources like machine's, network's, social media, mobilephone etc. there is a massive and continuous 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 issue like velocity. Example:- There are more the 3.5 billion's user's searches per day on Google, also facebook. user's are increasing by (22%) (Approx.) year by year.

- **Variety:** Variety involves collecting data from various resources and in fuzzy and heterogeneous types. this include importing data is dissimilar formats namely structured (table reside in relational database- RDBMS, etc), semi-structured (e-mail, XML and other markup language etc) and unstructured (text, pictures, audio files, video and sensor's data etc.).



Scanned with OKEN Scanner

- **Veracity:** It refers to the provenance, accuracy and correctness of data. Multiple factors to ensure the veracity of Big data are as follows:-
  - Trustworthiness of data origin.
  - reliability and security of data store.
  - data availability
  - Correctness
  - Consistency
  - integrity
  - Completeness.

In simple term's Veracity mean's how much of the data is reliable. Veracity is the process of being able to handle and manage data efficiently.

- **Value:** Value refers to the benefits that your organisation derives from the data. Does it match your organization's goals? Does it help your organization enhance itself. Value is an essential characteristics of big data. It is valuable and reliable data that we store, process and also analyze.

- **Validity:**
  Validity refers to how accurate and correct the data is for its intended use. these characteristics are both based on the principle that data must be valid and accurate, so that decision-making is right on target. Validity here is more about weather the data collected is per the problem that the analyst wants to solve.

Scanned with OKEN Scanner

- **Variability:** Variability describes that data changes all the time. Different data can impact other characteristics such as the process and data transfer speed. It refers to the variance in meaning, number of inconsistencies, multiple of data dimension's and inconsistent data receiving speed.
- **Vulnerability:**
It refers to the security of the collected datasets that will be used for later analysis. It also denotes the flaw's in the system, which permits malicious activities to be conducted on the collected datasets. Example:
If a hacker breaches a banking website's, this puts sensitive information like customer's account details, credit card information and passwoard at risk. Companies set up data security teams to overcome this threat by setting up suitable guidelines to protect data.
- **Volatility:**
This is a measure of data's rate of change its lifetime. It define's how long you can store relevant data. Organization's need to understand how long a specific type of data is valid.
Example: some organizations may only keep the most recent year of customer data and Transaction's.
- **Visualisation:** It refers to the ability to present Big data into a visual context, such as diagrams, graphs, maps etc. It also assists people and organisation's to discover

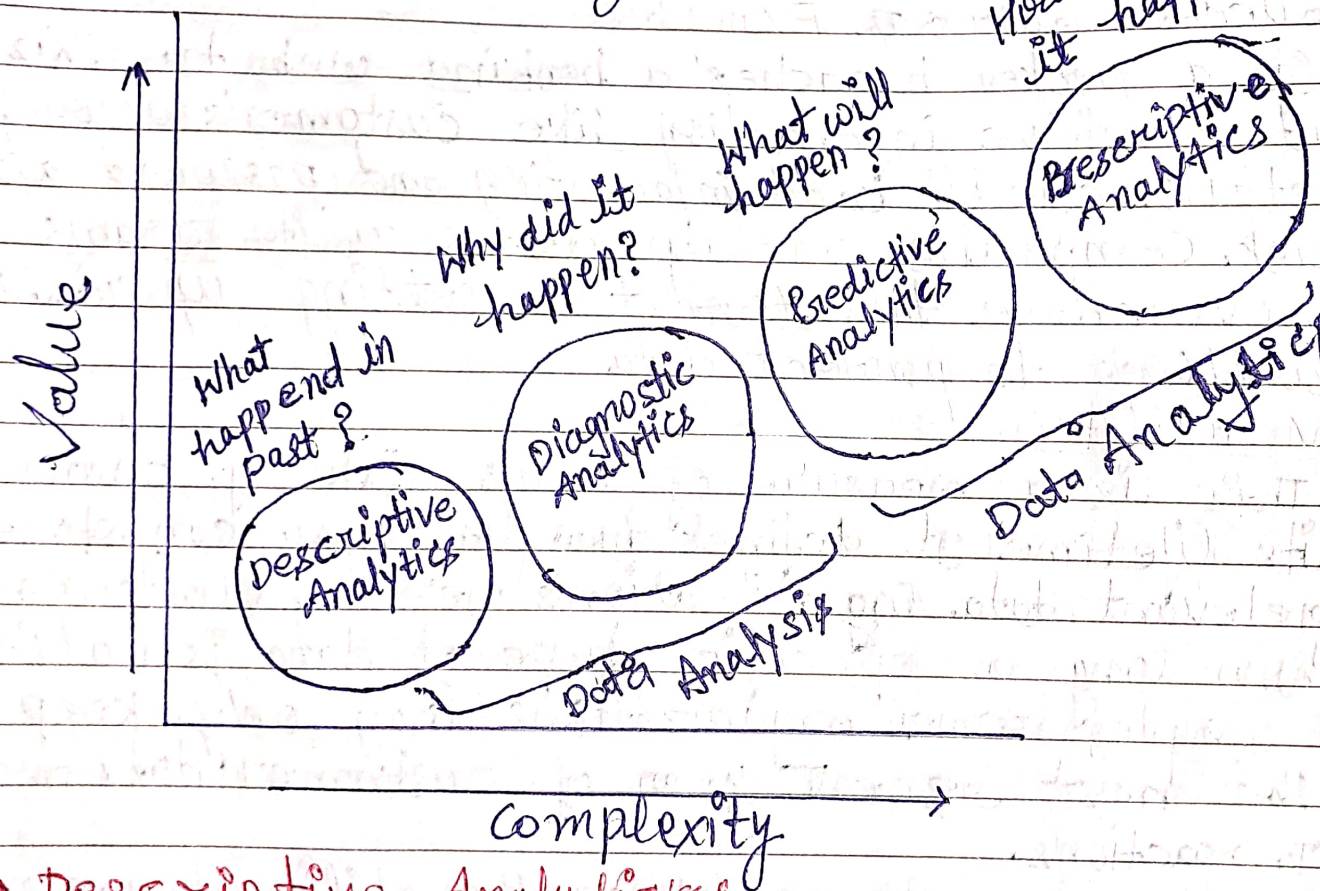
Scanned with OKEN Scanner

Patterns, correlations, trends and dependencies. Big data visualization is a powerful tool for decision-makers to access, evaluate and interpret massive data in even real-time and act upon it.

- **Data Analytics and its type:**

There are four types of data analytics:-

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics.



- **Descriptive Analytics:**

Descriptive analytics answer the Question "What happened". It seeks to provide a depiction or summary

view of facts and figures in an understandable format. This either inform or prepare data for further analysis. They can describe in detail about an event that has occurred in the past. A common example of descriptive analytics are company reports that simply provide a historic review like:-

- Data Queries
- Reports
- Descriptive statistics
- Data visualization.

- **Diagnostic Analytics:**

In this analysis, we generally use historical data over other data data to answer the question "Why did it happen? or for the solution of any problem. We try to find any dependency and pattern in the historical data of the particular problem. Common techniques used for Diagnostic Analytics are:-

- Data discovery
- Data mining
- Correlations.

- **Predictive Analytics:**

Predictive analytics help's to forecast trends based on the current events. predicting the probability of an event happening in future or estimating the accurate time. it will happen can all be determined with the help of predictive analytical models. Set of techniques that use model constructed from past data to predict the future or

ascertain impact of one variable on another.

- Linear regression.
- Time series analysis and forecasting.
- Data mining.

Prescriptive Analytics:

Prescriptive Analytics show's you how you can best take advantage of the future outcomes that have been predicted. What steps can you take to avoide a future problem. Prescritive analytics is without doubt, the most complex type of analysis involving algorithm's like:

- Machine Learning.
- statistical methods.
- Computation model.

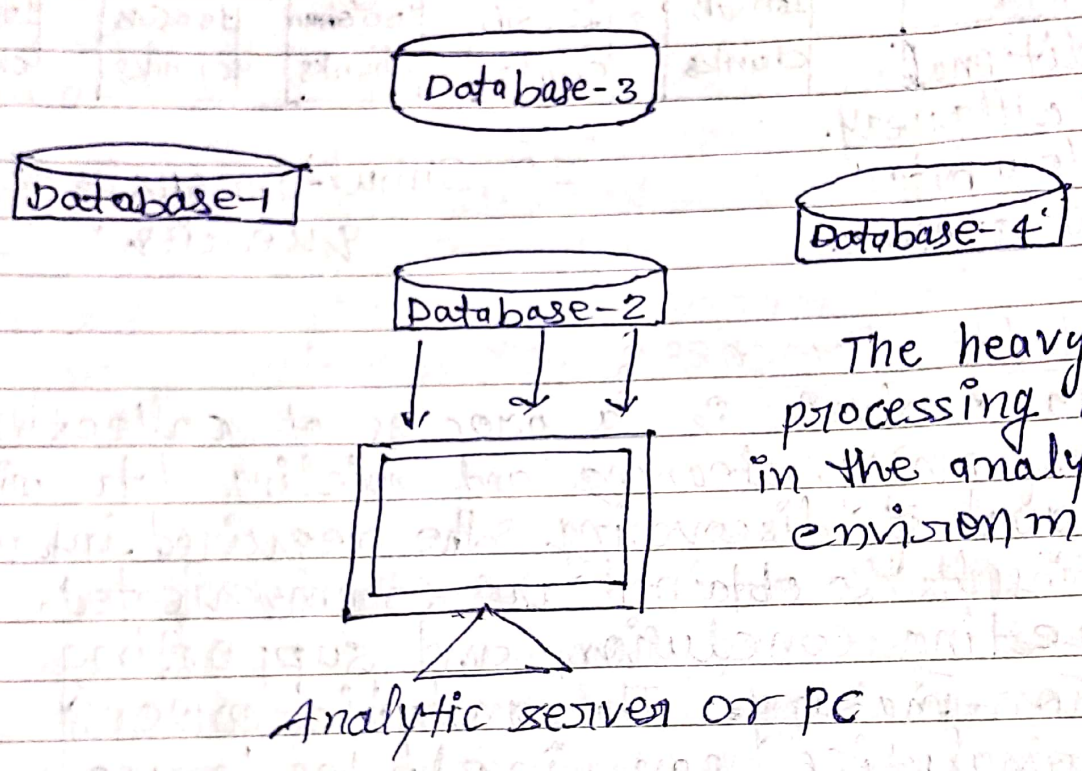
Evolution of Analytics Scalability:

Analytic scalability is the ability to use data to understand and solve a large variety of problems.

According to the definition of scalability, scalable data analysis refers to the ability of a hardware/software parallel system to exploit increasing computing resources effectively in the analysis of large datasets.

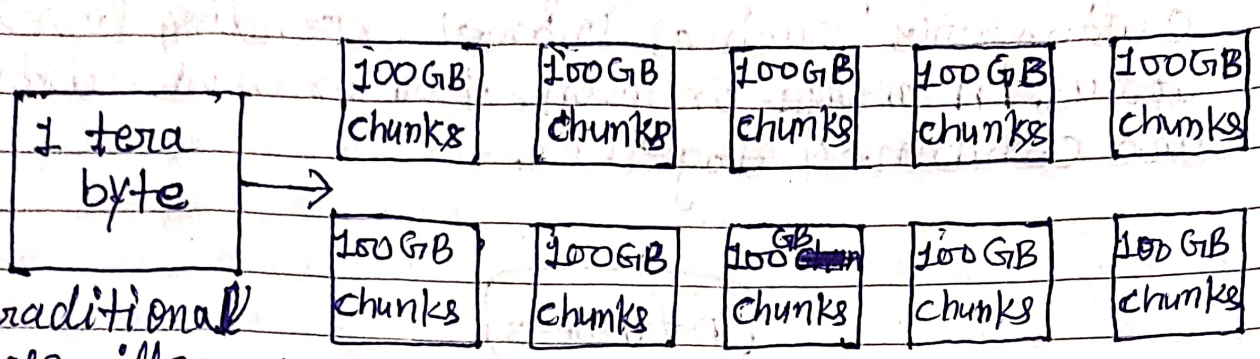
It is the ability of a system to handle the increasing ~~amount~~ amount of work required to perform its task. In analytic scalability we have to pull the data together in

a separate analytics environment and then start performing analysis.
Analyst do the merge operation on the datasets which contain rows and columns. The column's represent information about the customer's such as 'name', spending level, or status. In merge or join, two or more datasets are combined together.



• Massively Parallel Processing (MPP):
It is refers to the concept of using a large number of processing units or computing resources to solve a complex problem or perform a task simultaneously. Here, multiple processor's has its own operating system and memory. It breaks up the large datasets, because its difficult to

manage into easily workable chunks, assigning each to a processor. An MPP database breaks the data into independent pieces managed by independent storage and cpu resources.



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

10 Simultaneously 100-GB queries.

Analytic Process:

Data Analysis is a process of collecting, transforming, cleaning and modeling data with the goal of discovering the required information. The results so obtained are communicated, suggesting conclusions and supporting decision-making. The analytic process in data analytics can includes several steps, such as defining the problem:-

- Data Requirement specification
- Data collection
- Data Processing
- Data Cleaning
- Data Analysis
- Communication

Scanned with OKEN Scanner

- **Data Requirement Specification:**

The data required for analysis is based on a question or an experiment. Based on the requirements of those directing the analysis, the data neccessary as inputs to the the analysis is identified (e.g.- Population of people). Data may be numerical or categorical

- **Data Collection:** Data collection is the process of gathering information on targeted variables identified as data requirements. Data collection ensures that data gathered is accurate such that the related decision's are valid. Data is collected from various sources ranging from organisational database to the information in webpages.

- **Data Processing:**

After the collected data is required to be Data processing. the data that is collected must be processed or organised for analysis. Data processing is the process of transforming raw data into a usable format for analysis and used in data analytics. this include structuring the data as required for the relevant analysis tool's.

- **Data Cleaning:**

The processed and organised data may be incomplete, contain duplicates or contain errors. Data cleaning is the process of preventing and correcting these errors.

Scanned with OKEN Scanner

Data cleaning is not simply about erasing information to make space for new data.

- **Data Analysis:** Data that is processed, organised and cleaned would be ready for the analysis. Various data analysis techniques are available to understand, interpret and derive conclusions based on the requirement. Data analysis is the process of inspecting, cleaning, transforming and modeling data to discover useful information, draw conclusion and support decision-making.
- **Communication:** The result of the data analysis are to be reported in a format as required by the user's to support their decisions and further action. the feedback from the user's might result in additional analysis.

Modern data Analytic tools:

1. **Apache Hadoop:** Apache Hadoop is an open source, java-based framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.

Apachehadoop, a Big data analytics tool which is a java based free software framework.

It supports multiple languages and allows the developers to write applications in Java, Scala, R, or python. Analytics can be performed better as Spark has a rich set of SQL queries, machine learning algorithms and complex analytics etc.

2. KNIME: Konstanz Information Miner or most commonly known as KNIME is free and an open source data analytics tool to create data science applications and build machine learning models.
KNIME provides a GUI to create visual workflow using the drag and drop features. It supports multi-threaded in memory data processing.

3. Open Refine:
Open-Refine is a powerful tool for working with messy data, cleaning it, transforming it from one format into another.
Open-Refine tool is one of the efficient tool to work on the messy and large valume of data. It helps to explore large datasets easily.

4. Talend:
It is one of the most powerful data integration ETL tools available in the market and is developed in the Eclipse graphical development environment.
Talend allows quick deployment by means of automation. It offers open-source tools which can be downloaded for free. Talend enables you to integrate with many databases, and technologies.

Scanned with OKEN Scanner

5. Power BI: Power BI is a self-service business analytics tool developed by microsoft with the capability to analyze and visualize data, extract insights and share it, across various departments with your organization. It offers a wide range of detailed and attractive visualizations to create reports and dashboard. It supports multiple data sources such as excel, Text/csv, Oracle, pdf, xml. etc.

6. Tableau:
It is a market-leading Business Intelligence tool used to analyze and visualize data in an easy format. It can translate queries to visualization and import all ranges and size of data. Allows you to ask questions, spot trends and identify opportunities.

7. RapidMiner:
RapidMiner tool operates using visual programming and also it is much capable of manipulating, analyzing and modeling the data. RapidMiner allows developers to use the drag and drop option to operate on pieces of code when building models. RapidMiner provides more than 1500 operators for data analysis and transformations. It has various data exploration features which are used by users to gain visuable insights from the data.

8. R and Python programmings
R and Python are both is free open source, object oriented programming language. R and Python are completely free, hence its

easily available and can be used without any license. Both R and Python have good data handling capabilities.
Since, both R and Python are open in nature they get all the latest features without any delay.

9. Data wrapper: It is an online data visualization tool for making interactive charts. It uses data file in a csv, pdf or excel format. Data wrapper generate visualization in the form of bar, line, map etc. It can be embedded into any other websites as well.

Application of Data Analytics:

1. Security: Data analytics applications or more specifically, predictive analysis has also helped in dropping crime rates in certain areas.

2. Transportation: Data analytics is used in transportation to optimize logistics and improve customer services. It can be used especially in areas where we need to transport a large number of people to a specific area and require seamless transportation.

3. Marketing:
Data analytics is used to analyse customer behavior, performance and demographics to develop targeted marketing campaigns. This includes analyzing website traffic, social media engagement and email marketing.

Scanned with OKEN Scanner

4. Healthcare: Data analytics is used in healthcare to analyze patient data and improve patient outcomes. this includes analyzing electronic health records (EHRs) to identify disease patterns and improve treatment plans.

5. Human resources:
Data analytics is used in human resources to analyze employee data and identify ares for improvement. this includes analyzing employee performance, identifying training needs and predicting employee turnover.

6. Manufacturing:
Data analytics is used in manufacturing to optimize production processes and improve product quality. It also enables manufacturers to have real-time monitoring of production, which often translates into higher productivity and lowered costs.

7. Internet Searching:
Whe we use Google, we are using one of their many data analytics applications employed by the company. Most search engines like: google, Bing, Yahoo etc., use data analytics. these search engines use different algorithms to deliver the best result for a search query.

8. Media and entertainment industry:
The media and entertainment industry is one of the biggest users of big data analytics. As the number of users of different digital gadgets is increasing rapidly, media and entertainment companies are leveraging the power of data analytics.

Scanned with OKEN Scanner

Key Role of data Analytics Project:

There are certain key roles that are required for the complete and fulfilled functioning of the data science team to execute projects on analytics successfully the key roles are seven in number. Each key role plays a crucial role in developing a successful analytics projects.

1. Business user:

- The business user is one who understands the main area of the project and is also basically benefited from the results.
- This user gives advise and consults the team working on the projects about the value of the results obtained and how the operations on the output 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 issues.
- He generally provides the funds and measure's 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.

3. Project Manager:

- This person ensures that key milestone and

Scanned with OKEN Scanner

purpose of the project is met on the time and of the expected quality.

4. Business Intelligence Analyst:

- Business Intelligence Analyst provides business domain perfection based on a detailed and deep understanding of the data, key performance indicators, key matrix, and business intelligence from a reporting view of point.
- This person generally creates fascia and reports and knowns about the data feeds and sources.

5. Database Administrator (DBA):

- DBA facilitates and arranges the database environment to support the analytics need of the team working on a project.
- His responsibilities may includes providing permission to key database or tables and making sure that the appropriate security stages are in their correct places related to the repositories or not.

6. Data Engineer:

- Data engineer grasps deep technical skills to assist with tuning SQL queries for data management and data extraction and provides supports 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.

Scanned with OKEN Scanner

7. Data Scientist:

- Data scientist facilitates with the subject matter expertise for analytical techniques, data modeling and applying correct analytical techniques for 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.

Analysis Vs Reporting:

'Reporting' is the process of gathering and presenting data in a structured format such as graphs and tables. organizing information in predefined KPIs and metrics makes it easier for you to understand what is happening.
'Analytics' is the process of analyzing your data to identify patterns and gain insights. Using techniques such as predictive and prescriptive analytics helps your understand why things are happening and what to do next.
Reporting and analysis are both important for an organization to make informed decisions by presenting data in a format that is easy to understand.

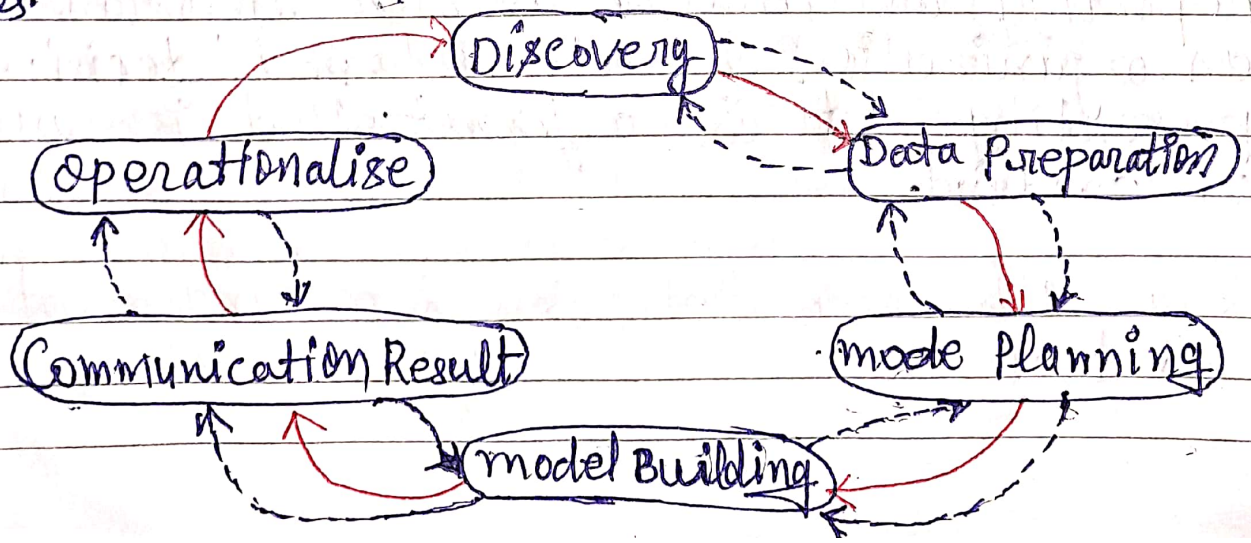
Key difference between Reporting and analysis:

Scanned with OKEN Scanner

Reporting	Analytics.
• It is done to show what's happening.	• Explains why something is happening.
• Summarize and present data for information purpose.	• Analytics is to gather insights for data enabled decision-making.
• It involves a person.	• Does not involve a person.
• The purpose of reporting is to organize the data into meaningful information	• The purpose of analytics is to draw conclusions based on data.
• Reporting is provided to the appropriate business leaders to perform effectively and efficiently within a firm.	• Analytics is used by data analysts scientists and business people to make effective decisions.

Data Analytics life cycle:

Data analytics lifecycle was designed to address Big data problems and data science projects. The process is repeated to show the real projects. Data analytics lifecycle defines the roadmap of how data is generated, collected, processed, used and analysed to achieve business goals.



1 Discoverys

This is the initial phase to set your project's objectives and find ways to achieve a complete data analytics life cycle. Start with defining your business domain (like- ML, AI, cloud etc) and ensure you have enough resources (like- Data, technology, people and time and framming of problem) to achieve your goals.

2. Data Preparation and Processings

The data preparation and processing phase involves collecting, processing, conditioning, cleaning and Extract Transform Load (ETL) data before moving to the model building process. One of the essential aspects of this phase is ensuring data availability for processing. You have to identify various data sources and analyse how much and what kind of data you can accumulate within a given timeframe. Evaluate the data structures, explore their attributes and acquire all the tools needed.

3. Model Plannings

In this phase one needs to analyze the quality of data and find a suitable model for your project. In this phase, data science team develop data sets for training, testing and production purposes. In this phase, several tools and techniques used for model planning like- clustering, Association rules, Regression, classification and tools are used like- MatLab, STASTICA.

Scanned with OKEN Scanner

4. Model Building:

In this phase, the team builds and executes the models meticulously as planned during the model planning phase. the data is tested and suitable answers are searched for to the given objectives. and also test the its parallel processing.

5. Communication Results:

In this phase aims to determine whether the project results are a success or failure and start collaborating with significant stakeholders. The team identifies the vital findings of their analysis, measure the associated business value and creates a summarised narrative to convey the stakeholders results.

6. Operationalise:

In this phase, The team will present the full in depth report with the briefing coding, key findings, and all the technical documents and papers to the stakeholders.

The results are closely monitored, ensuring they match with the expected goals. if the ~~findings~~ findings fit perfectly with the objectives, then the report can be finalised. the model can be then deployed and integrated into the business.

[End]

Scanned with OKEN Scanner