UNIT-2

Date: ../../..
Page No. : .....

~~HDFS~~ Hadoop

→ Hadoop is an open source framework overseen by Apache software foundation which is written in Java

Hadoop

✱ UNIT → 1 ✱

Big data → Bigdata is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently.

Big data is also a data but with huge size.

✱ The three primary sources of Big data

→ Social data → Social data comes from the likes, tweets & Retweets, comments, video uploads, and general media that are uploaded and shared via world's favorite social media platforms.

→ Machine data → Machine data is defined as information which is generated by industrial equipment, sensors that are installed in machinery and even web logs which track user behavior.

→ Transactional data → Transactional data is generated from all the daily transactions that take place both online and offline. Invoices, payment orders, storage records, delivery receipts — all are characterized as transactional data yet data alone is almost meaningless.

Date : ..../..../....
Page No. : .....

BIG DATA PHASH 1

- Period: 1970 - 2000

> DMBS-based, Structured content:

→ RDBMS & data warehousing
→ Extract Transfer load
(OLAP) → Online Analytical Processing
→ Dashboard & scorecards
→ Data mining & Statistical analysis.

BIG DATA PHASE 2

- Period: 2000 - 2010

> web-based, unstructured content

→ Opinion mining
→ web analytics and web intelligence
→ social media analytics.
→ social Network analysis
→ Spatial - temporal analysis.

BIG DATA PHASE 3

- Period: 2010 - Present

> Mobile and sensor-based content

→ Location - aware analysis
→ Person - centered analysis
→ Context - relevant analysis
→ mobile visualization
→ Human computer interaction.

Example → The New york Stock Exchange generates about one terabyte of new trade data per day.

→ Social media: The statistic shows that 500+ terabytes of new data get ingested into the databases of social media site facebook, everyday. This data is mainly generated in terms of Photo and Video uploads, message exchanges putting comments etc.

→ Companies using big data are Amazon, Netflix, IBM American express, Google, oracle.

Date : ..../..../....
Page No. : .....

Why is Big data Important / Why we need Big data.

→ The importance of big data does NOT revolve around how much data a company has but how a company utilises the collected data. Every company uses data in its own way; the more efficiently a company uses its data, the more potential it has to grow.

The company can take data from any source and analyse it to find answers which will enable:

① Cost Saving ② Time Reductions
③ Understand the market conditions
④ Control online reputation
⑤ Using Big data Analytics to Boost Customer Acquisition and Retention

Applications of Big data.

→ In today's world, there are a lot of data, Big companies utilize those data for their business growth. By analyzing this data, the useful decision can be made in various cases as discussed below:-

① Tracking customer Spending Habit, Shopping Behavior:-
→ In big retails Store (like Amazon, Walmart, Big Bazar etc) management team has to keep data of customer's spending habit, shopping behavior, customer's most liked product. Which product is being searched / sold most, based on that data, production / collection rate of that product get fixed.

Date : ...../...../.....
Page No. : .....

2. Recommendation :→ By tracking customer spending habit, shopping behaviour, Big retails store provide a recommendation to the customer. E-commerce site like Amazon, walmart, flipkart does product recommendation.

3. Smart traffic system :→ Data about the condition of the traffic of different road, collected through camera kept beside the road, at entry and exit point of the city, gps device placed in the vehicle (Ola, Uber, cab, etc)

4. secure Air Traffic System → At various places of flight (like propeller etc) sensors present. These sensors capture data like the speed of flight, moisture, temperature, other environmental condition.

5. Auto Driving Car → Big data Analysis helps drive a car without human interpretation. on the various spot of car camera, a sensor placed, that gather data like the size of the surrounding car obstacle, distance from those etc.

6. Virtual personal Assistant Tool : Big data analysis helps virtual personal Assistant Tool (like Siri in Apple device, Cortana in windows, google Assistant in Android) to provide the Answer of the various questions asked by users.

Date : ../../..
Page No. : .....

① Education Sector → Online educational course conducting organization utilize big data to search candidate, interested in that course.

② Energy sector → Smart electric meter read consumed power every 15 minutes and sends this read data to the server, where data analyzed and it can be estimated what is the time in a day when the power load is less throughout the city.

③ media and Entertainment sector → media and entertainment service providing company like Netflix, Amazon Prime, Spotify do analysis on data collected from their users.

Security challenges

① Vulnerability to fake data generation
② Potential presence of untrusted inputs
③ Trouble of cryptographic protection
④ Possibility of sensitive information mining
⑤ Data provenance difficulties.
⑥ Absent security audits.

Date : ...../...../.....
Page No. : .....

5 V's of Big data / Characteristics of Big data

→ Big data is defined by the "5 V's. which are also termed as the characteristics of Big data as follows :-

- Volume
- Variety
- Velocity
- veracity
- value

* Volume :- Volume refers to the amount of data that you have. We measure the volume of our data in gigabytes, zettabytes (ZB), and yottbytes (YB). According to the industry trends, the volume of data will rise substantially for the coming years.

* Variety :- Variety refers to the different types of big data. it is among the biggest issues faced by the big data industry as it affects performance

* Velocity :- Velocity refers to the speed of the data processing. High velocity is crucial for the performance of any big data process. It consists of the rate of change, activity burst, and the linking of incoming data sets.

* Veracity :- Veracity refers to the accuracy of your data. It is among the most important Big data characteristics as low veracity can greatly damage the accuracy of your results.

Date : ../../..
Page No. : .....

* Value :- Value refers to the benefits that your organization derives from the data.
Does it enrich your organization's goals?
Does it help your organization enhance itself?
It is among the most important big data core characteristics.

Big data use cases

* Healthcare
→ Predictive analytics are being used to enhance palliative care, speed up the diagnostic process with AI-enabled chest x-rays, and reduce cases of end-stage renal disease by using predictive modeling to weigh the risks and benefits of kidney disease treatments.

* Retail
→ Data collected from loyalty programs, credit card transactions, website behaviour, social media and email engagement, IP addresses, mobiles applications, user logon, purchase histories, and more now give retailers a 360-degree view of the customer.

* Banking and finance
→ Financial decisions like investments and loans are now placed in the hands of AI, which uses machine learning technologies to process loan applications, evaluate potential investments, and calculate risk.
For example → Big data analytics can evaluate stock prices alongside social trends. That might impact the stock market.

Date : ...../...../.....
Page No. : .....

Education

→ In education sector, big data is being used to identify and improve teaching strategies to help students succeed academically.
Analytics are being used to measure teacher performance, track when student log into an online learning portal, how much time they spend on different pages, and how they progress through their coursework.

Manufacturing

→ Big data analytics in manufacturing allows organizations to gain end-to-end visibility into production processes, supply chain metrics, and environment conditions that impact productivity and deliverables.

Understanding Big data storage

→ Big data storage infrastructure that is designed specifically to store, manage and retrieve massive amounts of data, or big data.
Big data storage enables the storage and sorting of big data in such a way that it can easily be accessed, used and processed by applications and services working on big data. Big data storage is also able to flexible scale as required.

- Data sources → Internal data sources such as data from CRM system, ERP system, sales reports etc.
  External data sources such as government statistics and media channels.

- Data Storage → Big data storage software tools store, manage and retrieve massive amounts of data.
  → APACHE hadoop, mongo DB, Cassandra

- Data mining → Data mining tools allow businesses to extract usable data from a huge set of raw data to find relationships, patterns, and anomalies.
  → Rapidminer, SPSS modeler.

- Data Analytics → Although data mining tools incorporate data analysis, there are software designed specifically with advanced analytical capabilities.
  → Apache SPARK, Kafka, IBM analytics.

- Data visualization → Data visualization is also a type of data analytics tool.
  However, they are specifically designed to take the raw data and presenting it with beautiful and easy digestible visuals like graphs and charts.
  → Klipfolio, Looker

Best Practices for Big data Analytics

1) Define the Big data business goals.
2) Assess the strategize with partners.
3) Determine what you have and what you need in BD.
4) Keep continuous communication and assessment going.
5) Start slow, react fast in leveraging Big data
6) evaluate Big data technology requirements.
7) Align with Big data in the cloud
8) Manage your Big data experts, as you keep an eye on compliance and access issues.

UNIT-2

Date: ../../..
Page No.: .....

HDFS

HDFS Stand for Hadoop Distributed file System.
HDFS is utilized for storage permission. It is mainly designed for working on commodity Hardware devices (inexpensive devices), working on a distributed file system design.
HDFS is designed in such a way that it believes more in storing the data in a large chunk of blocks rather than storing small data blocks.

HDFS in Hadoop provides fault-tolerance and High availability to the storage layer and the other devices present in that Hadoop cluster.

features of HDFS (Hadoop Distributed file system)

→ it is suitable for the distributed storage and Processing.
→ Hadoop Provides a command interface to interact with HDFS
→ Streaming access to file system data.
→ HDFS Provides file permissions and authentication
→ The built-in servers of namenode and datanode help users to easily check the status of cluster.

Data storage Nodes in HDFS

- Name Node (master)
- Data Node (slave)

Date : ...../...../.....
Page No. : .....

* **NameNode** → The NameNode is the commodity hardware that contains the GNU/Linux operating system and the NameNode software. It is a software that can be run on commodity hardware.

The system having the namenodes acts as the master server and it does the following tasks—

- Manage the file system namespace.
- Regulates client's access to files
- It also executes file system operations such as renaming, closing, and opening files and directories.

* **Data Node** → The data Node is a commodity hardware having the GNU/LINUX operating system and data Node software. For every Node (Commodity hardware/system) in a cluster, there will be a data Node.

These Nodes manage the data storage of their system.

- DataNodes perform read write operation on the file system, as per Client Request.
- They also perform operations such as block creation, deletion, and replication according to the instruction of the Name Node.

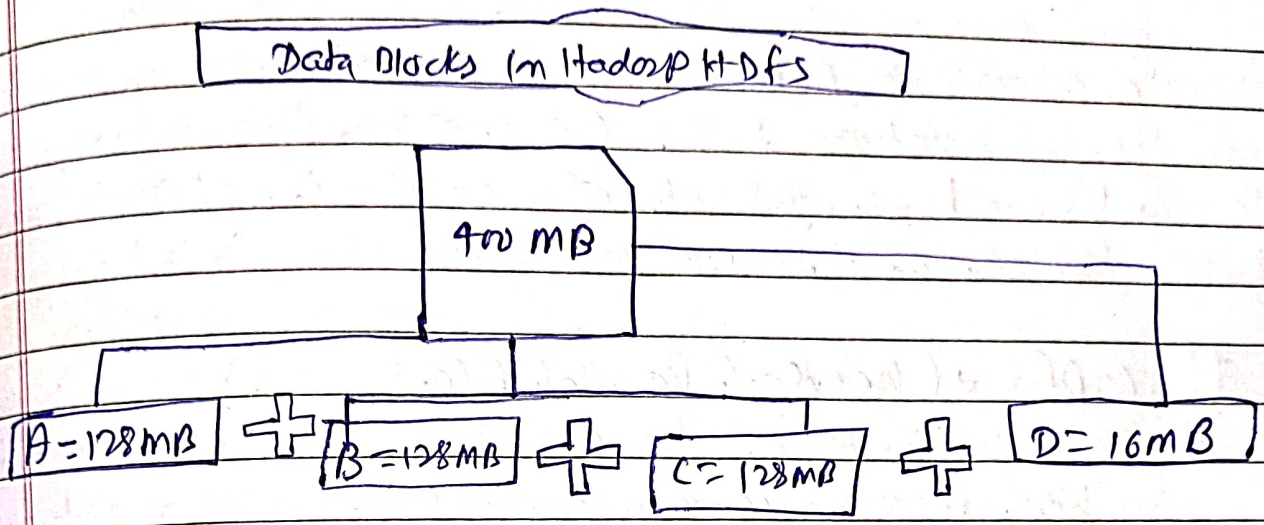
* **Block** → Generally the User data is stored in the files of HDFS. The file in a file system will be divided into one or more segments and stored in individual data Nodes.

These file segments are called as blocks.

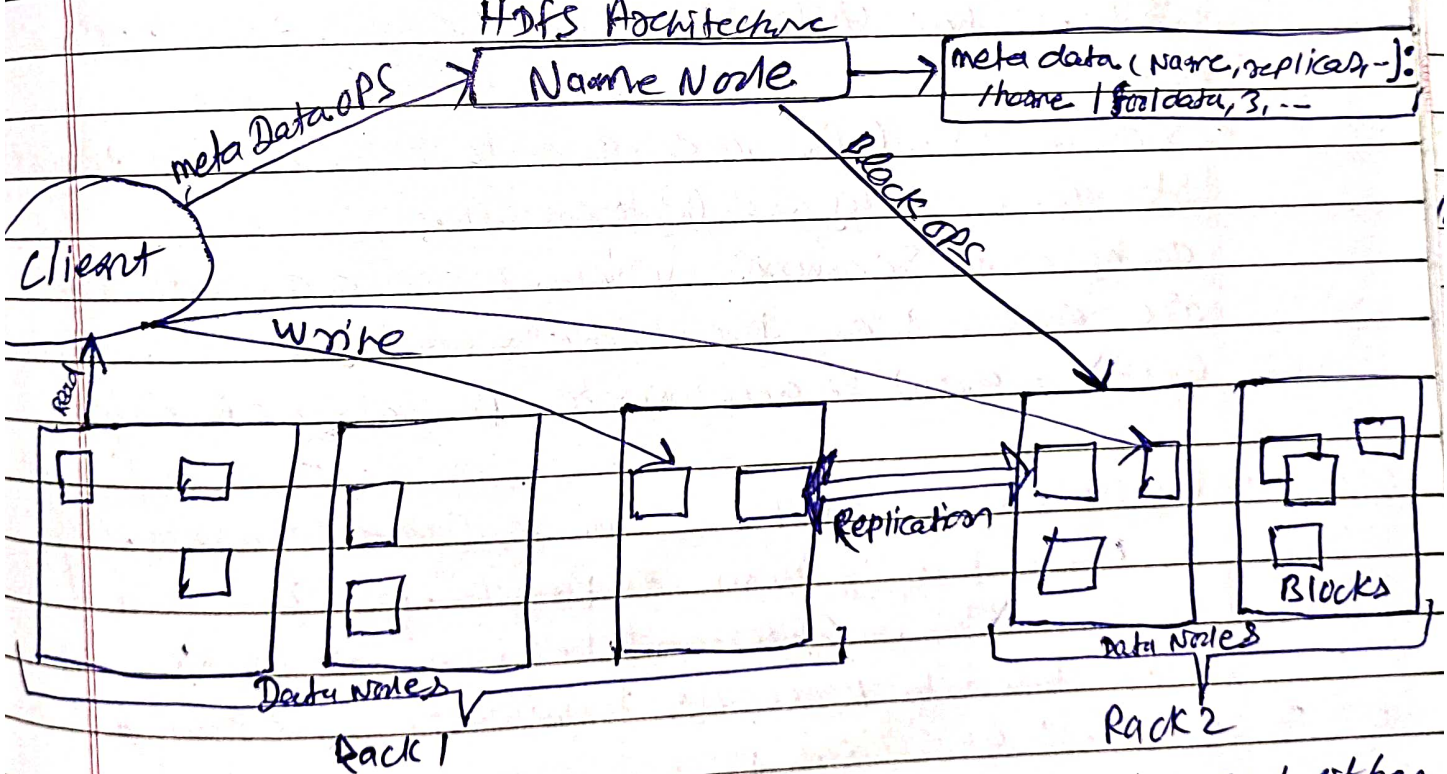
Another Words → The minimum amount of data that HDFS can read or write is called a Block. The default size is 64 MB, but it can be increase as per the need to change in HDFS configuration.

Date: ..../..../....
Page No.: .....

File Blocks in HDFS → Data in HDFS is always stored in form of blocks. So the single block of data is divided into multiple blocks of size 128 MB which is default and you can also change it manually.



HDFS Architecture



HDFS follows the master-slave architecture and it has following elements:

- Name Node
- Block
- Data Node

Goals of HDFS

- **Fault detection and recovery** → Since HDFS includes a large number of commodity hardware, failure of components is frequent. Therefore HDFS should have mechanisms for quick and Automatic fault detection and recovery.

- Huge datasets → HDFS should have hundreds of Nodes per Cluster to manage the application having huge datasets.

- Hardware at data → A Request task can be done efficiently, when the computation takes place near the data. Especially where huge datasets are involved, it reduces the network traffic and increases the throughput.

GPFS → (General Parallel file system)

→ GPFS which stands for General Parallel file system. is a high-performance clustered file system often used in big data and high performance computing environment.

→ Here's how GPFS is relevant in the context of big data :-

- Scalability :- GPFS is designed to scale both in terms of storage capacity and performance. In big data scenarios, where massive amounts of data are generated and processed, the ability to scale is crucial. GPFS can grow to accomodate petabytes of data.

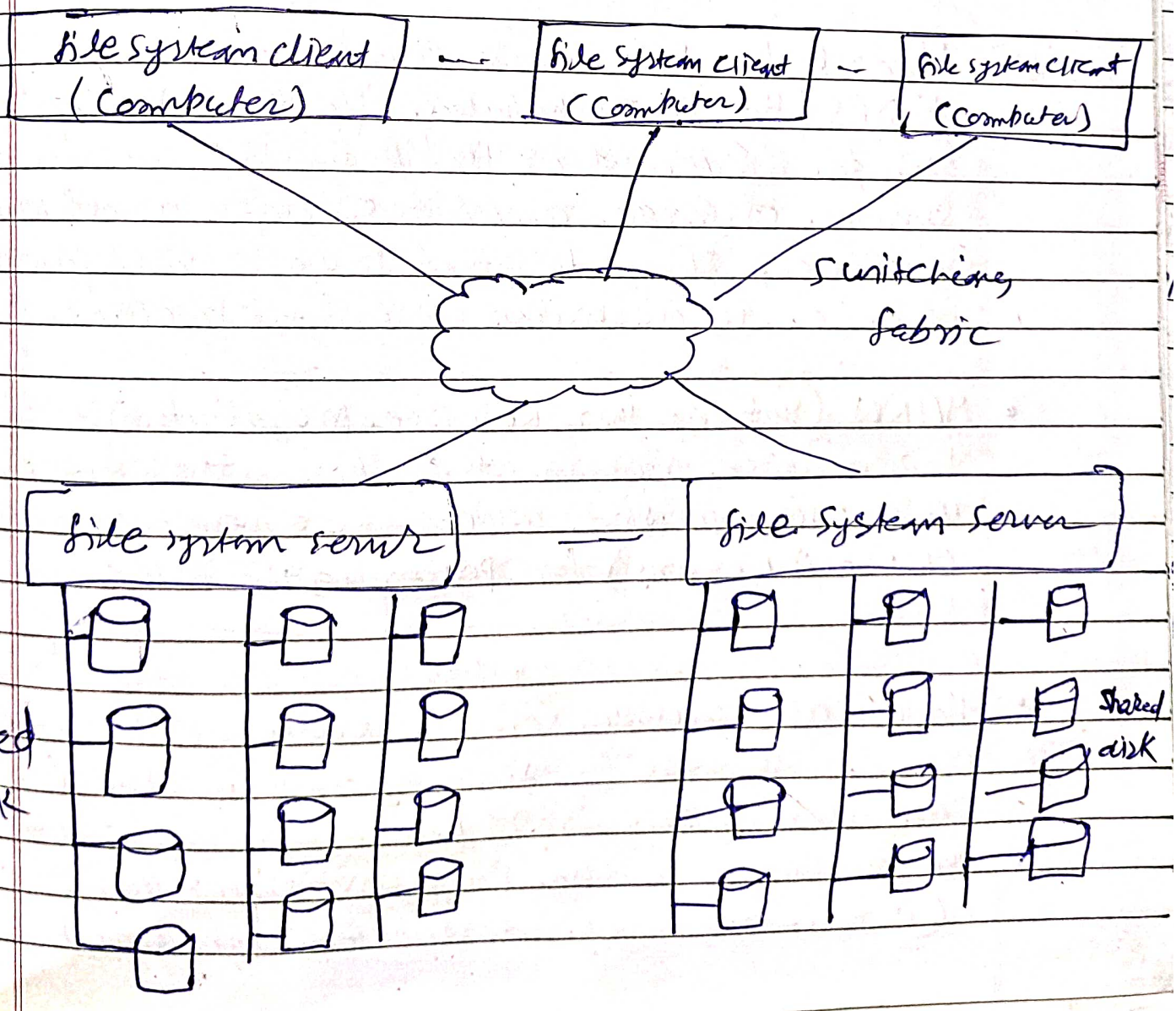
- Parallel access :- GPFS allows multiple servers to access the same file system in parallel. In big data clusters, this is important because it enables distributed data processing frameworks like Hadoop and Spark to access and process data concurrently.

- Data Redundancy : GPFS includes features for data redundancy and reliability. This is important for big data applications where data integrity is critical.

- High Throughput → GPFS is optimized for high throughput, which is essential for big data scenarios where large datasets need to be read and write quickly.
- Performance Tuning → GPFS provides tools for performance tuning, allowing administrators to optimize the file system's performance according to the specific requirements of their big data workloads.

Architecture of GPFS

→ The architecture of GPFS, also known as IBM Spectrum Scale, is designed to provide a high-performance and scalable clustered file system.



Hadoop Architecture

Date : ...../...../.....
Page No. : .....

Hadoop.

→ Hadoop is an open-source software framework that is used for storing and processing large amounts of data in a distributed computing environment. It is designed to handle big data and is based on the MapReduce programming model, which allows for the parallel processing of large datasets.

⇒ Hadoop is an open source software programming framework for storing a large amount of data and performing the computation. Its framework is based on Java programming with some native code in C and shell scripts.

Hadoop has two main components:

- HDFS (Hadoop Distributed File System):- This is the storage component of Hadoop, which allows for the storage of large amounts of data across multiple machines. It is designed to work with commodity hardware, which makes it cost-effective.

- YARN (Yet Another Resource Negotiator):- This is the resource management component of Hadoop, which manage the allocation of resources (such as CPU and memory) for processing the data stored in HDFS.

- Hadoop also include several additional modules that provide additional functionality, such as HIVE (a SQL-like query language), Pig (a high-level platform for creating MapReduce programs), and HBase (a non-relational, distributed database)

Date : ..../..../....
Page No. : .....

Features of hadoop.
→ It is fault tolerance.
→ It is highly available.
→ It's programming is easy.
→ It have huge flexible storage.
→ It is low cost.

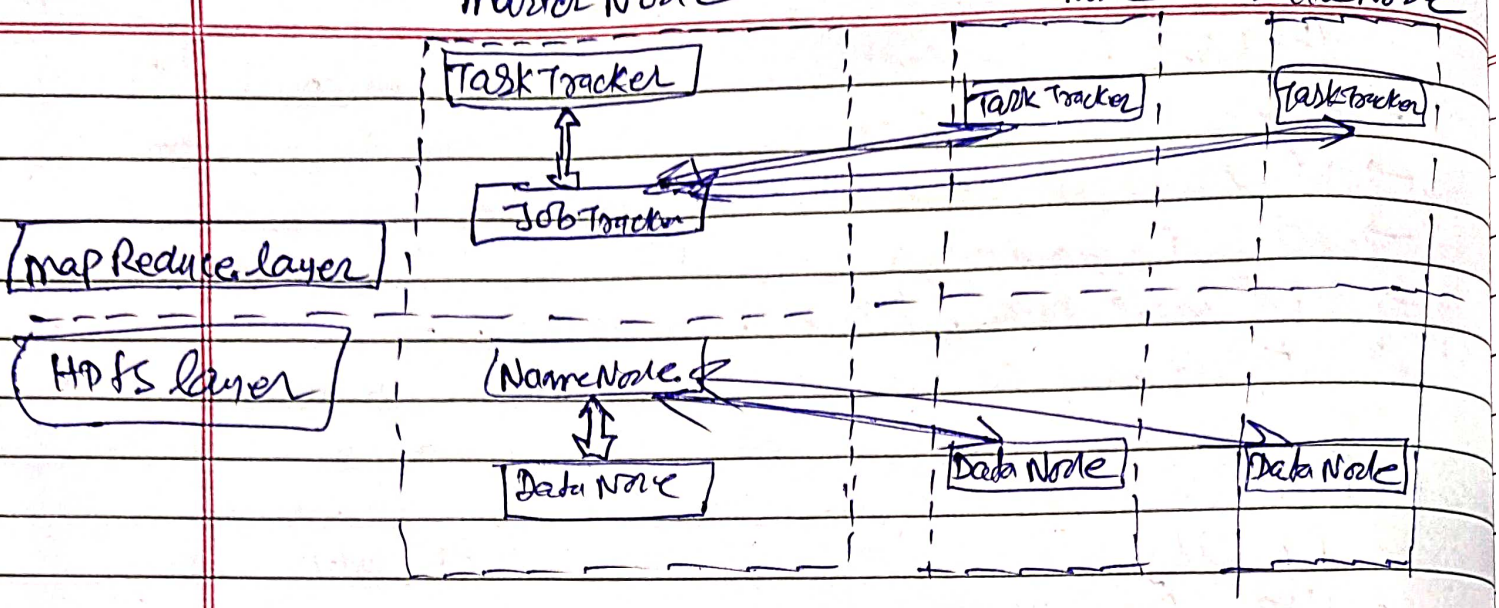
Advantages :-
- Ability to store a large amount of data.
- High flexibility
- cost effective
- High computional power
- Tasks are independent
- Linear scaling

Disadvantages :-
- Not very effective for small data.
- Hard cluster management.
- Has stability issues.
- Security concerns.

Hadoop Architecture
→ Hadoop Architecture is a package of the file system, map Reduce engine, and the HDFS (Hadoop Distributed File system). The MapReduce engine can be MapReduce/MR1 or YARN/MR2.

A Hadoop cluster consists of a single master and multiple slave nodes. The master node includes Job Tracker, Task Tracker, NameNode and Data Node whereas the slave node includes DataNode and Task Tracker.

High level Hadoop Architecture



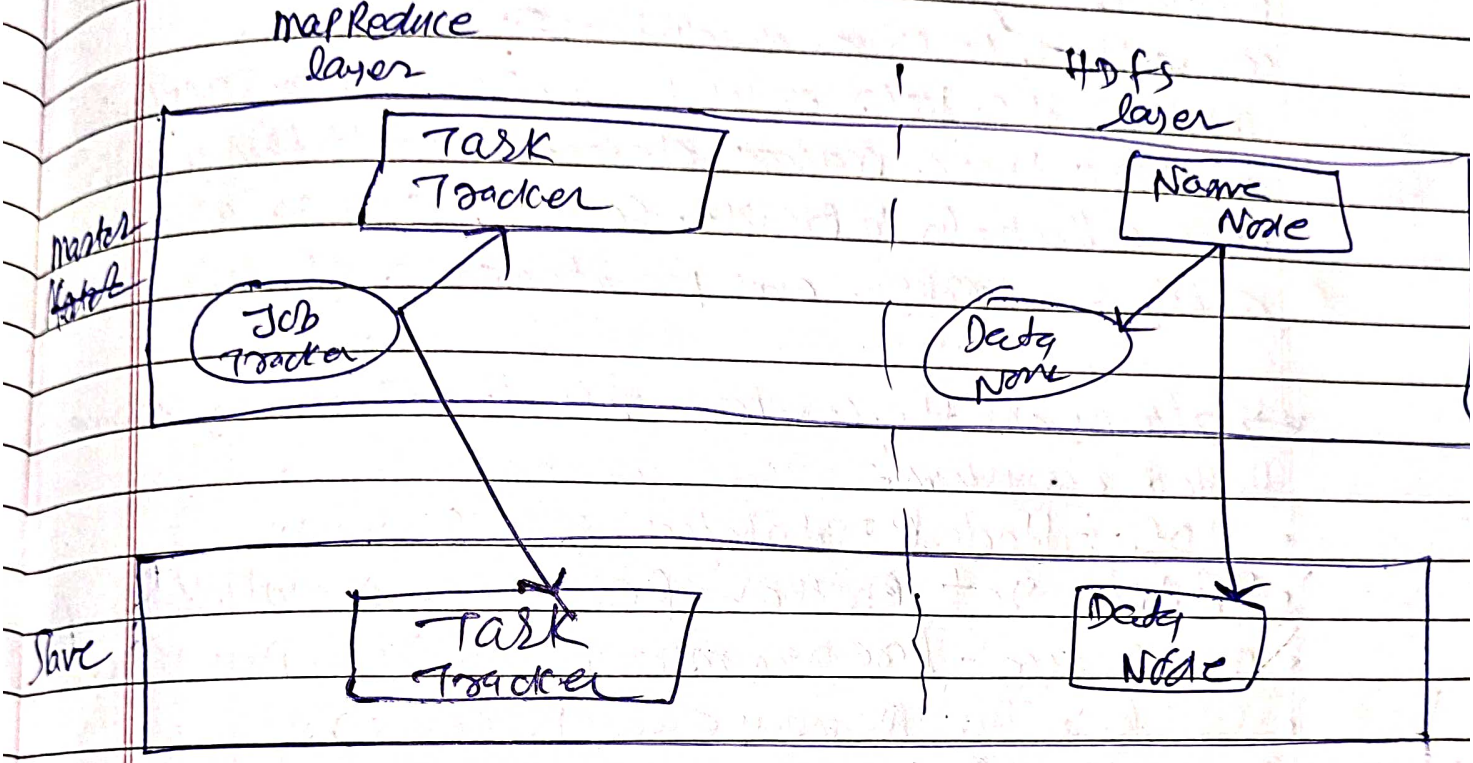
Hadoop has a master-slave Architecture for data storage and distributed data processing using mapreduce and HDFS methods.

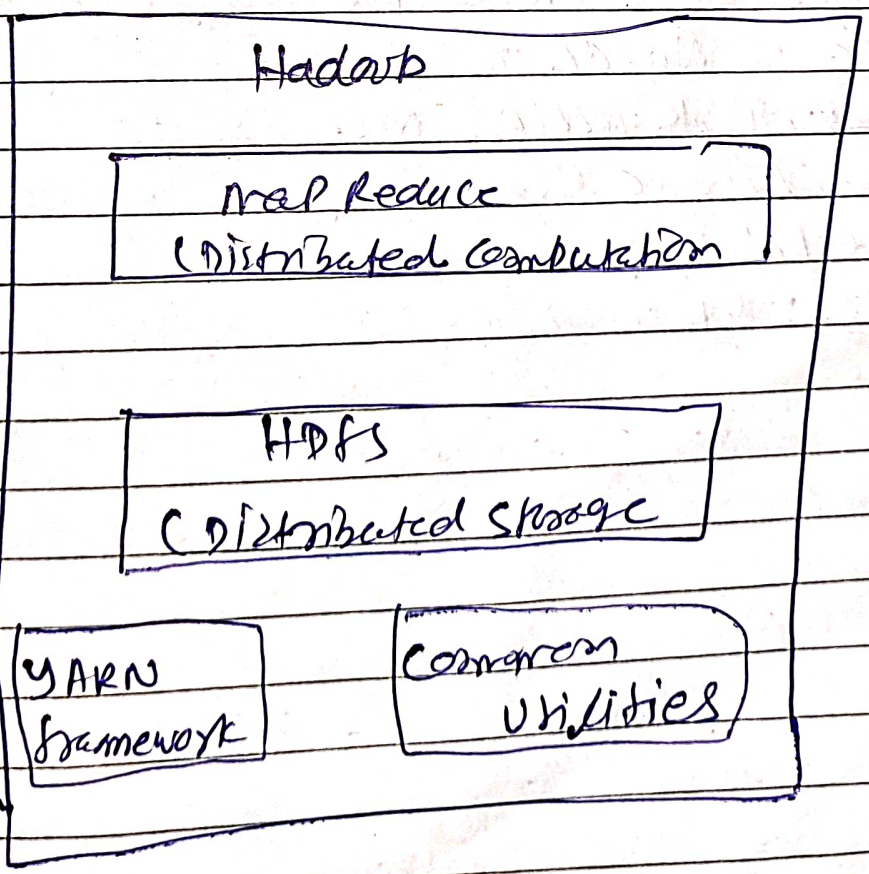
- NameNode :- Namenode represented every files and directory which is used in the namespace.
- Data Node :- Data Node helps you to manage the state of an HDFS node and allows you to interact with the blocks.
- Master Node → The master Node allows you to conduct parallel processing of data using Hadoop MapReduce.
- Slave Node :- The slave Node are the additional machines in the Hadoop cluster which allows you to store data to conduct complex calculations. Moreover, all the slave Node comes with Task Tracker and a Data Node.

In Hadoop, master or slave system can be set up in the cloud or on-Premise.

Date : ..../..../....
Page No. : .....

At its core, Hadoop has two major layers namely —
- Processing / Computation layer (MapReduce), and
- Storage layer (Hadoop Distributed file system).





Hadoop Ecosystem

Date : ..../..../....
Page No. : .....

Hadoop Ecosystem

→ Hadoop Ecosystem is a Platform or a suite which Provides various services to solve the big data problems. It includes Apache Projects and various commerical tools and solutions.
Most of the tools or solutions are used to supplement or support there major elements. All these tools work collectively to provide services such as absorption, analysis, storage and maintenance of data etc.

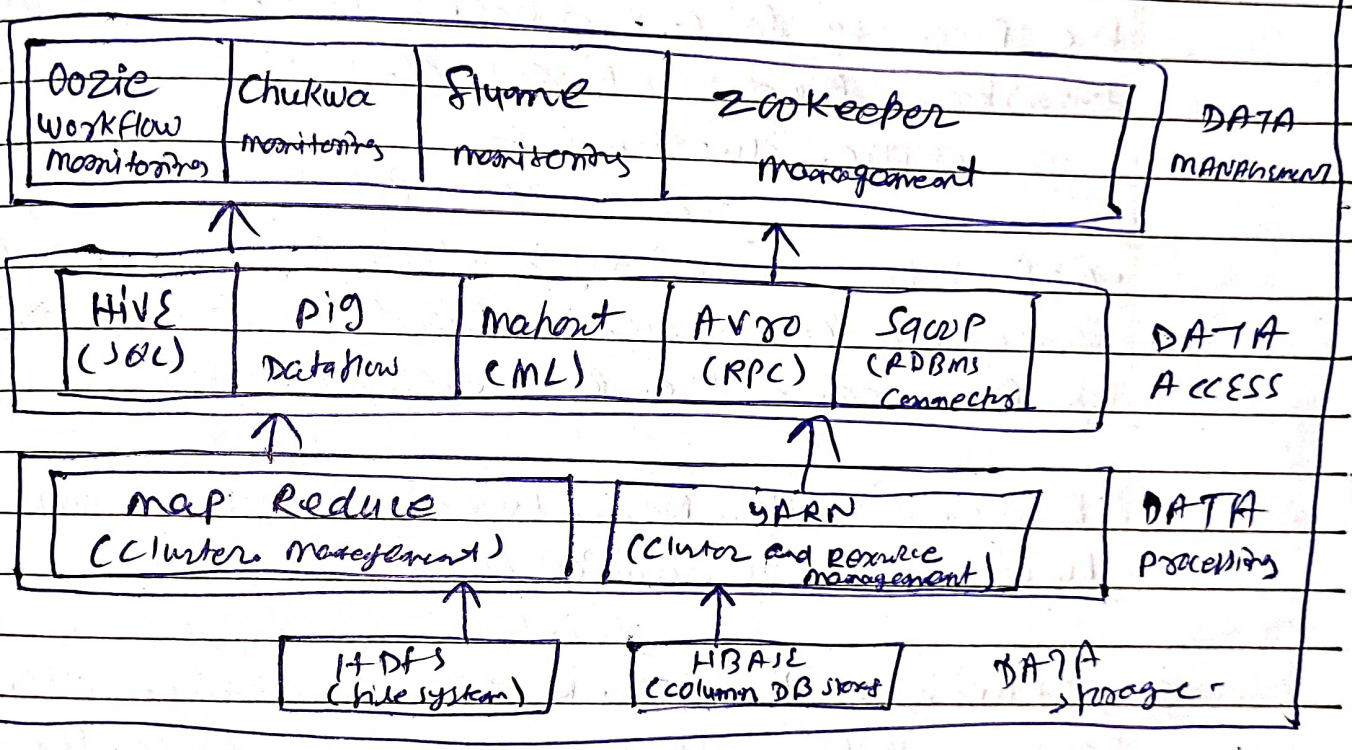
→ following are the components that collectively form a Hadoop ecosystem:-

- HDFS :- Hadoop Distributed File System.
- YARN :- Yet Another Resource Negotiator.
- Map Reduce :- Programming based Data Processing.
- Spark :- In-Memory data processing.
- Pig, HIVE :- Query based processing of data services.
- HBase :- NOSQL Database.
- Mahout, Spark MLLib :- ML algorithm libraries.
- Solar, Lucene :- Searching and Indexing.
- Zookeeper :- Managing Cluster.
- Oozie :- Job scheduling

Date : ..../..../....
Page No. : .....

Oozie
wo

Hadoop Ecosystem

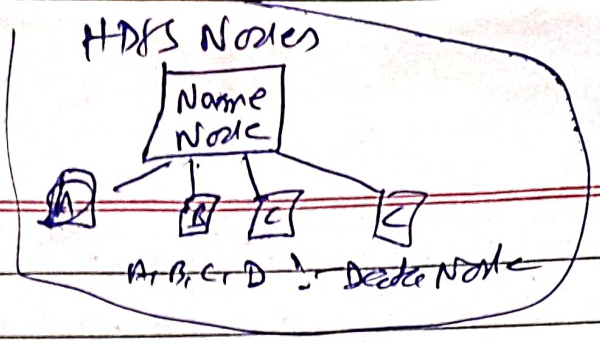


→ All these toolkits or components revolve around one term ie Data. That's the beauty of Hadoop that it revolves around data and hence making its synthesis easier.

HDFS → It is the primary or major components of Hadoop ecosystem and is responsible for storing ~~data~~ large data sets of structured or unstructured data across various Nodes

HDFS consists two core components

- NameNode
- Data Node



Date : ..../..../....
Page No. : .....

- HBASE → it is a NOSQL database which supports all kinds of data and thus capable of handling anything of Hadoop Database. It provides capabilities of Google's Big Table, thus able to work on Big data sets effectively.

- MapReduce :→ By making the use of distributed and parallel algorithm, MapReduce makes it possible to carry over the processing logic and help to write applications which transform big data sets into a manageable one.

  It use Two functions
  - Map & - Reduce.

- Map() performs sorting and filtering of data and generate a key-value pair based result which is later on processed by the Reduce() method.

- Reduce, as the Name suggests does the summarization by aggregating the mapped data.

- YARN :- Yet Another Resource Negotiator, as the Name implies, YARN is the one who helps to manage the resource across the clusters. It performs scheduling and resource allocation for the Hadoop system.

- Pig or HIVE (done)

- Mahout → Mahout, allows machine Learnability to a system or application. It provide various libraries or functionalities such as collaborative filtering, clustering and classification.

Date: / /
Page No.:

- Spark → It is a platform that handle all the process consumptive tasks like batch processing, interactive or iterative real-time processing, graph conversions, and visualization etc.

- Zookeeper → Zookeeper overcame all the problems by performing synchronization, inter-component based communication, grouping, and maintenance.

- Oozie → Oozie simply performs the task of scheduler, thus scheduling job and binding them together as a single unit. There kinds of jobs is
  Oozie workflow and Oozie coordinator jobs.

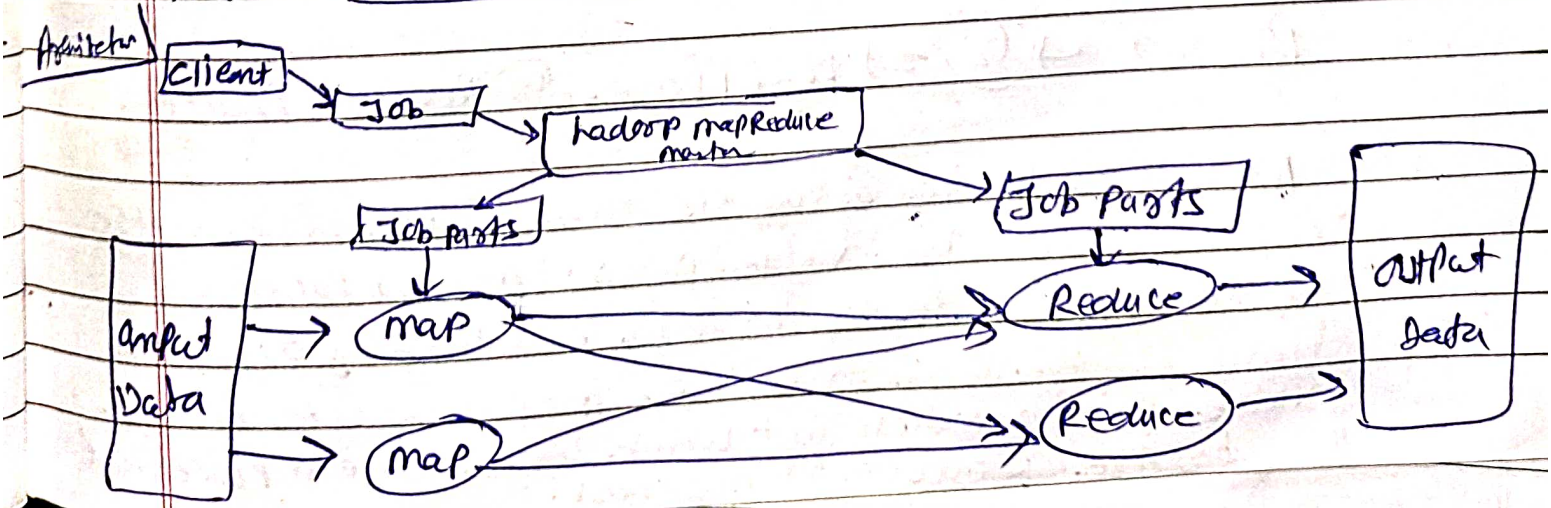
Map Reduce Architecture :-

⇒ MapReduce is a programming model used for efficient processing in parallel over large data-sets in a distributed manner.

The data is first split and then combined the final result.

The purpose of MapReduce in Hadoop is to map each of the jobs and then it will reduce it to equivalent task.

The map Reduce task is mainly divided into two phases Map Phase and Reduce Phase



Date : ..../..../....
Page No. : .....

Components of Map Reduce Architecture

(1) Client → The Map Reduce client is the one who brings the Job to the MapReduce for Processing. There can be multiple clients available that continuously send jobs for Processing to the Hadoop MapReduce Manager.

(2) Job → The map reduce job is the actual work that the client wanted to do which is comprised of so many smaller tasks that the client wants to process or execute.

(3) Hadoop Map Reduce Master → It divides the particular job into subsequent job parts.

(4) Job Parts :- The task or sub-jobs that are obtained after dividing the main job.
The result of all the job-parts combined to Produce the final output.

(5) Input Data → The data set that is fed to the MapReduce for Processing.

(6) Output Data → The final result is obtained after the processing.

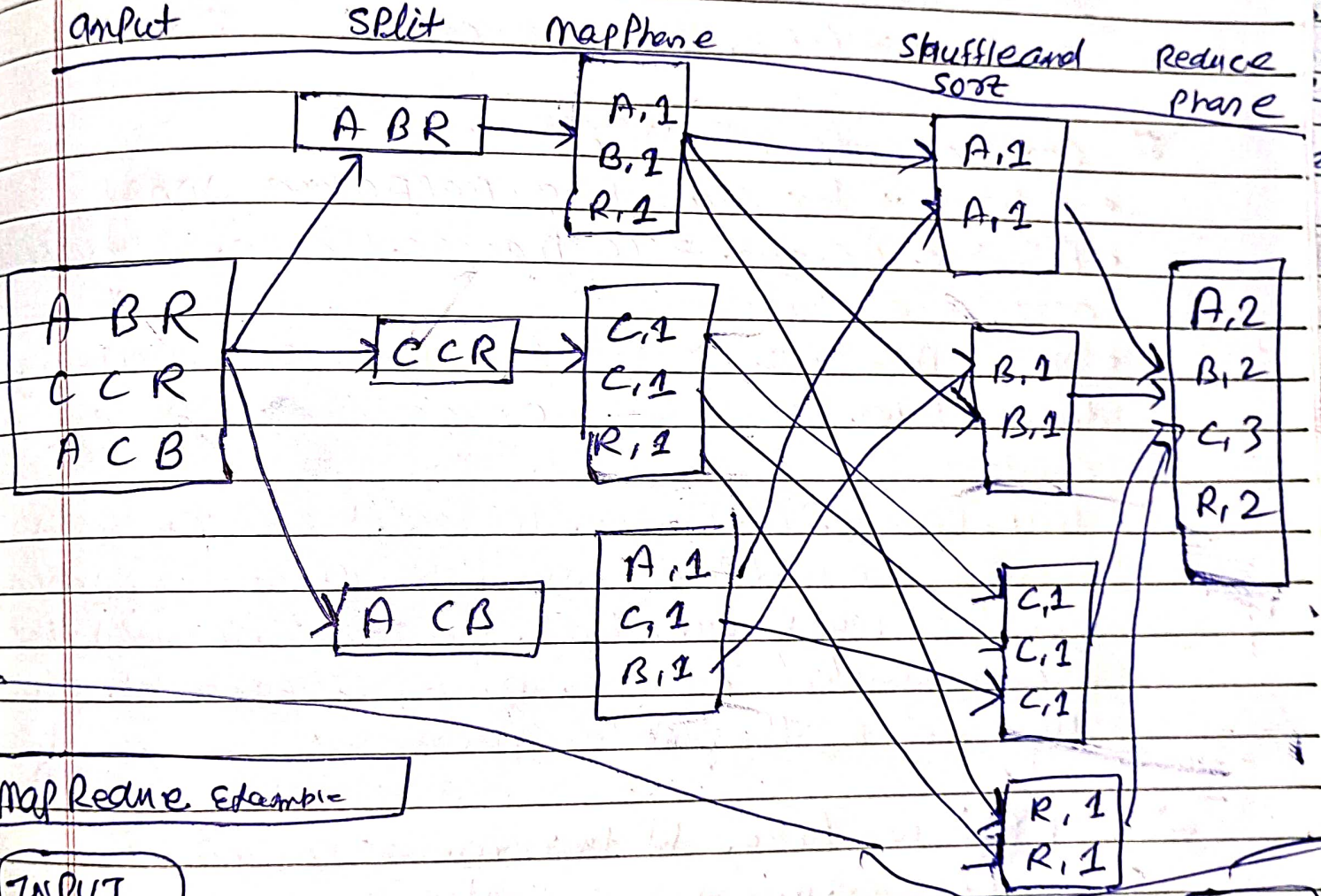
Phases → (2) ⇒ Map Phase (1) Reduce Phase.

(1) Map :- As the name suggests its main use is to map the input data in key-value pairs. Map() Performs sorting and filtering of data and thereby organizing them in the form of group. Map generates a Key-value pair based result which is later on processed by the Reduce() method.

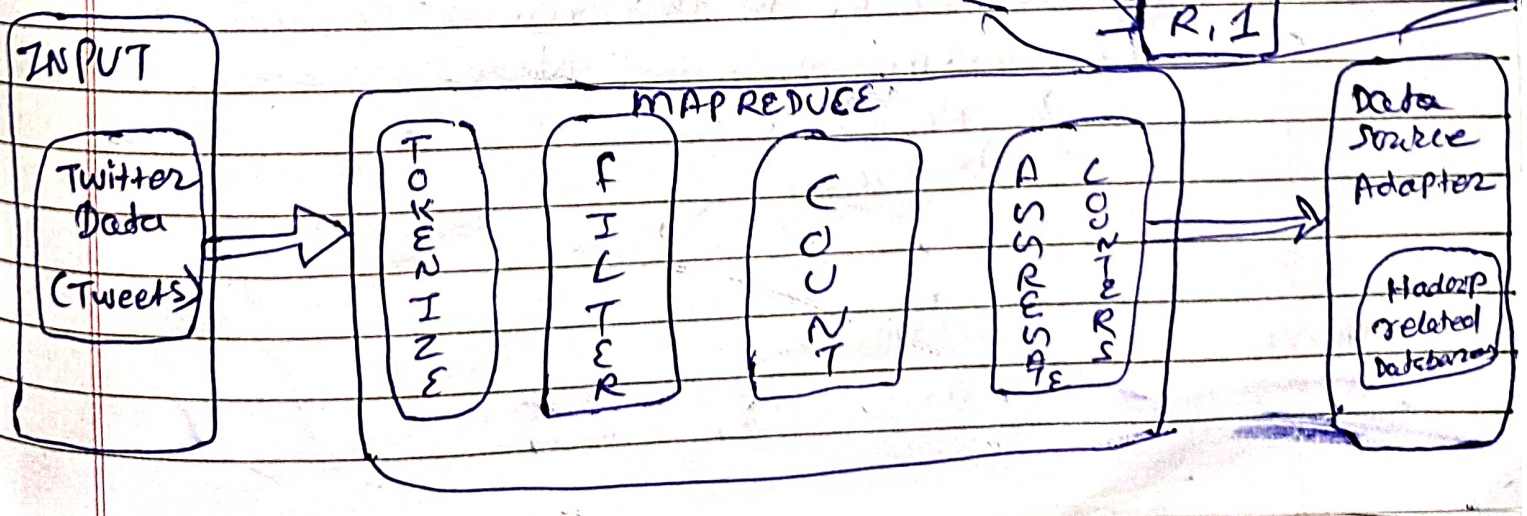
Date: / /
Page No.:

② Reduce → As the Name Suggests does the summarization by aggregating the mapped data.
An simple,
Reduce() takes the output generated by map() as input and combines those tuples into smaller set of tuples.

The two tasks ~~of Reduce~~ are Map & Reduce with the help of a diagram.



Map Reduce Example



Date : ..../..../....
Page No. : .....

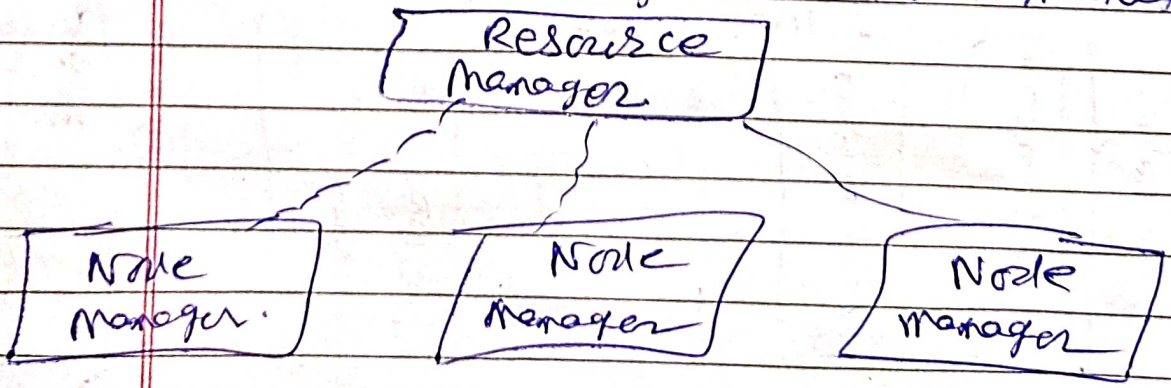
YARN

→ YARN is a framework on which MapReduce works. YARN Performs 2 operations that are Job scheduling and Resource management. The Purpose of Job scheduler is to divide a big task into small jobs so that each job can be assigned to various slave in a Hadoop cluster and Processing can be maximized. and the use of Resource manager is to manage all the resources that are made available for running a Hadoop cluster.

Components of YARN

- Client → for submitting mapReduce jobs
- Resource manager → TO manage the use of resources across the cluster.
- Node manager → for launching and monitoring the computer containers on machines in the cluster.
- Map Reduce Application master : Checks tasks running the mapReduce job. The application master and the mapReduce tasks run in containers that are scheduled by the resource manager and managed by the Node managers.

→ YARN comprises of two major components Resource manager and Node manager.



Date : ../../..
Page No. : .....

Resource Manager

→ it is a cluster-level component and runs on the master machine.

→ it manages resources and schedules applications running on top of YARN

→ it has two components: Scheduler & Application Manager.

- Scheduler is responsible for allocating resources to the various running Applications

- Application Manager is responsible for accepting job submissions.

Node Manager

→ it is a node-level component and runs on each slave machine.

→ it is responsible for managing containers and monitoring resource utilization in each container.

→ it continuously communicates with Resource Manager to remain up-to-date.

Benefits of YARN

- Scalability
- Utilization
- Multitenancy
- Compatibility.

8/

Page:
Date: / /

HIVE

HIVE is a data warehouse system which is used to analyze structured data. It is built on the top of Hadoop. It was developed by facebook.

HIVE provide the functionality of reading, writing, and managing large datasets residing in distributed storage.
It runs SQL like queries called HQL (Hive query language) which gets internally converted to map Reduce jobs.

features of HIVE

- Hive is fast and scalable
- It provide SQL-like queries ie HQL
- It uses indexing to accelerate queries
- It allows different storage types such as plain text, RC file and HBase
- It supports user defined functions (UDFs) where user can provide its functionality.

Limitations of HIVE.

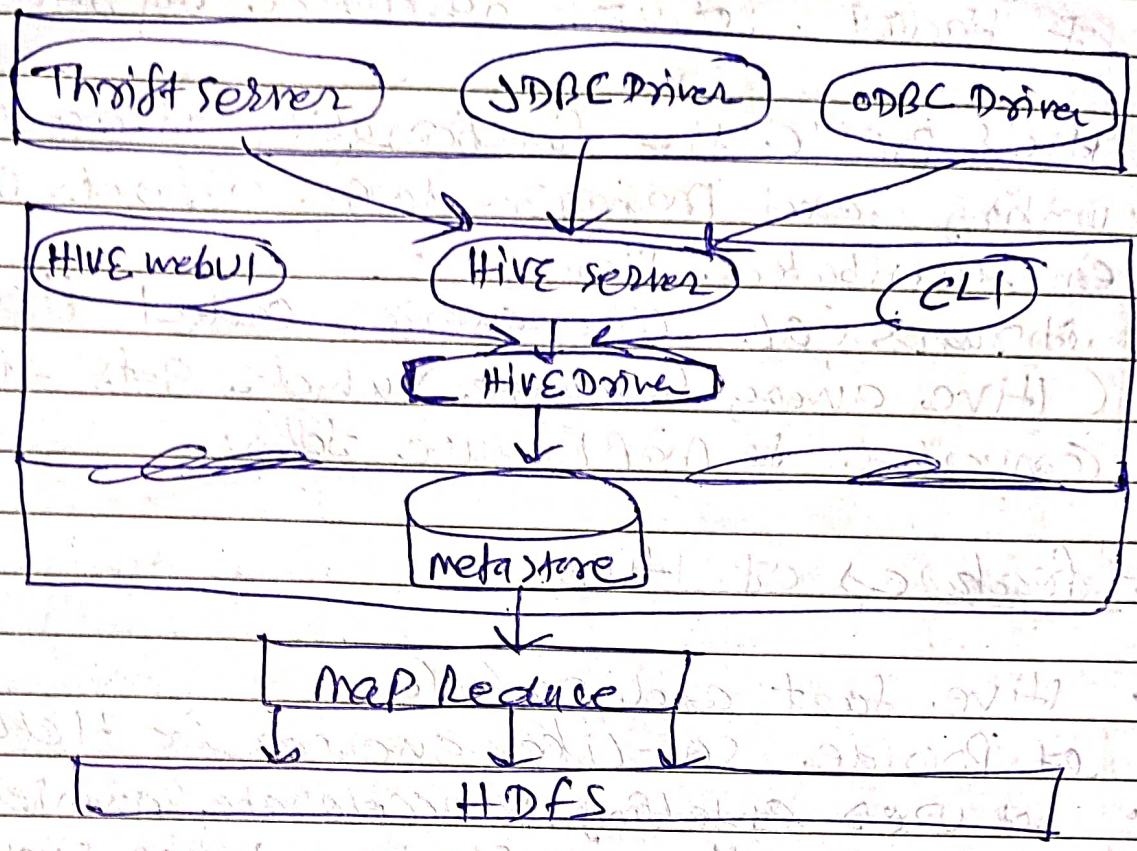
→ HIVE is not capable of handling real-time data.
→ it is not designed for online transaction processing.
→ ~~Hight~~ HIVE queries contain high latency.

→ pig →

Page :
Date : / /

HIVE Architecture

The following architecture explains the flow of submission of query into HIVE.



* HIVE Client → HIVE allows writing applications in various languages, including Java, Python, and C++. It supports different types of clients such as:

1) • Thrift server → It is a cross-language service provider platform that serves the request from all those programming languages that support Thrift.

2) • JDBC Driver → It is used to establish a connection b/w Hive and Java application.

3) • ODBC Driver → It allows the applications that support the ODBC Protocol to connect to HIVE.

~~8581858666~~ 1st ~~Aug~~ Sep 13,14

Page:
Date: / /

HIVE Services

The following are the services provided by HIVE—

- HIVE CLI → The HIVE CLI (Command line interface) is a shell where we can execute HIVE queries and commands.

- HIVE web user interface → it provide a web-based GUI for executing HIVE queries and commands.
  → it alternative of HIVE CLI.

2M
- HIVE metastore → it is a central repository that stores all the structure information of various tables and partitions in the warehouse.

- HIVE server → it is referred to as Apache Thrift server. it accepts the request from different client and provides it to Hive Driver.

- HIVE driver → it receives queries from different sources like web UI, CLI, Thrift, and JDBC/ODBC driver.

- HIVE Compiler → the purpose of compiler is to parse the query and perform semantic analysis on the different query blocks and expressions. it convert HIVEQL statements into MapReduce jobs.

- HIVE Execution Engine → optimizer generates the logical plan in the form of DAG of MapReduce tasks and HDFS tasks.

2M

Page :
Date : / /

* Pig :- Pig is a high-level data flow Platform for executing map reduce programs of Hadoop. It was developed by YAHOO. The language for Pig is Pig latin.

Pig can handle any type of data ie structured, semi-structured or unstructured and stores the corresponding results into Hadoop Data file System

Adv) • less code , • Reusability , Nested data types.

* YARN

→ YARN stands for "Yet Another Resource Negotiator YARN.

* Components of YARN

- Client → for submitting map Reduce jobs.
- Resource manager → To manage the use of resources across the cluster
- Node manager
- Map Reduce Application master

Benefits) Scalability , Utilization , Multitenancy.

* Characteristics of Big data

1. Volume
2. Veracity
3. Variety
4. Value
5. Velocity

} 5V's of Big Data.

Date : ...../...../.....
Page No. : .....

UNIT-3

* CLUSTERING *

Clustering

~~Clustering is the task of dividing~~

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled data set. It can be defined as:-

"A way of grouping the data points into different clusters consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."

→ The clustering technique is commonly used for Statistical data analysis.

USED IN:

(i) Market segmentation
(ii) Statistical data analysis
(iii) Social network analysis
(iv) Image Segmentation
(v) Anomaly detection

TYPES) (i) Partitioning Clustering
(ii) Density-Based Clustering
(iii) Distribution model-Based Clustering
(iv) Hierarchical Clustering
(v) Fuzzy clustering.

Page:
Date: / /

Clustering → Clustering or Cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as :-

"A way of grouping the data points into different clusters consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group.

→ The clustering technique is commonly used for Statistical data analysis.

Used in →
1. Market segmentation.
2. Statistical data analysis
3. Social network analysis
4. Image segmentation
5. Anomaly detection

Types
1. Partitioning Clustering
2. Density-Based clustering
3. Distribution Model-Based clustering
4. Hierarchical Clustering
5. Fuzzy Clustering.

(1) Partitioning Clustering → It is a type of clustering that divides the data into non-hierarchical groups. It is also known as the centroid-based method. The most common example of Partitioning clustering is the K-means clustering algorithm.
In this the dataset is divided into a set of K groups where K is used to define the number of pre-defined groups.

Page :
Date: / /

(2) Density Based Clustering

→ The Density Based clustering method connects the highly-dense areas into clusters, and the arbitrarily shaped distributions are formed as long as the dense region can be connected.

(3) Distribution Model - Based Clustering.

→ On the distribution model-Based clustering method, the data is divided based on the probability of how a dataset belongs to a particular distribution.

Example → Expectation - Maximization Clustering algorithm that uses Gaussian mixture models (GMM)

(4) Hierarchical Clustering

→ Hierarchical clustering can be used as an alternative for the Partitioned clustering as there is no requirement of Pre-specifying the number of clusters to be created.

In this technique, the dataset is divided into clusters to create a tree-like structure, which is also called a dendrogram.

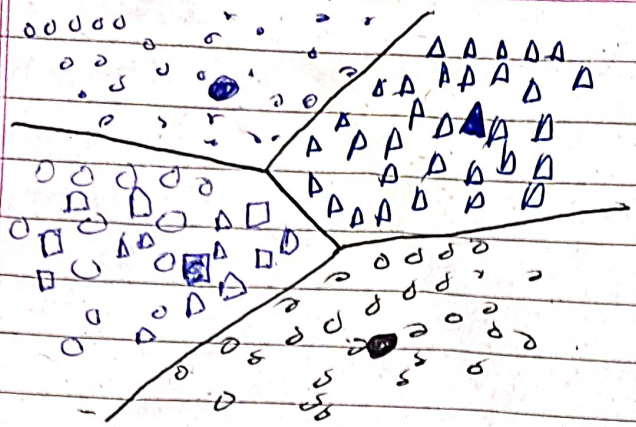
Example → Agglomerative Hierarchical algorithm

(5) Fuzzy Clustering → Fuzzy clustering is a type of soft method in which a data object may belong to more than one group or cluster.

Fuzzy C-means algorithm is the example of this type of clustering. also known as fuzzy K means algorithm.

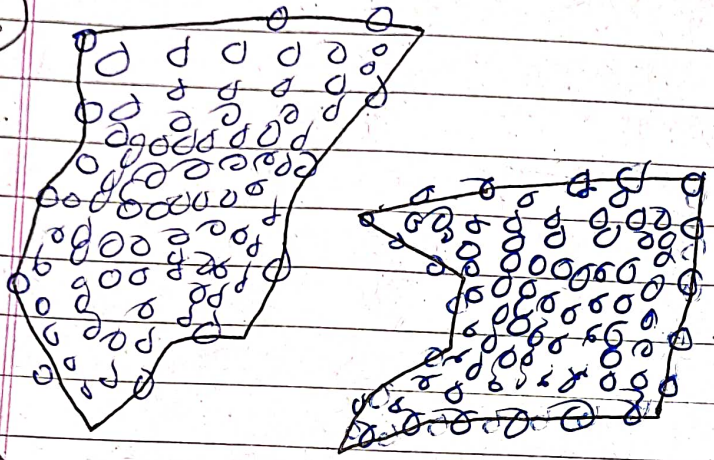
* Diagram of 2 Clustering types *

(1)



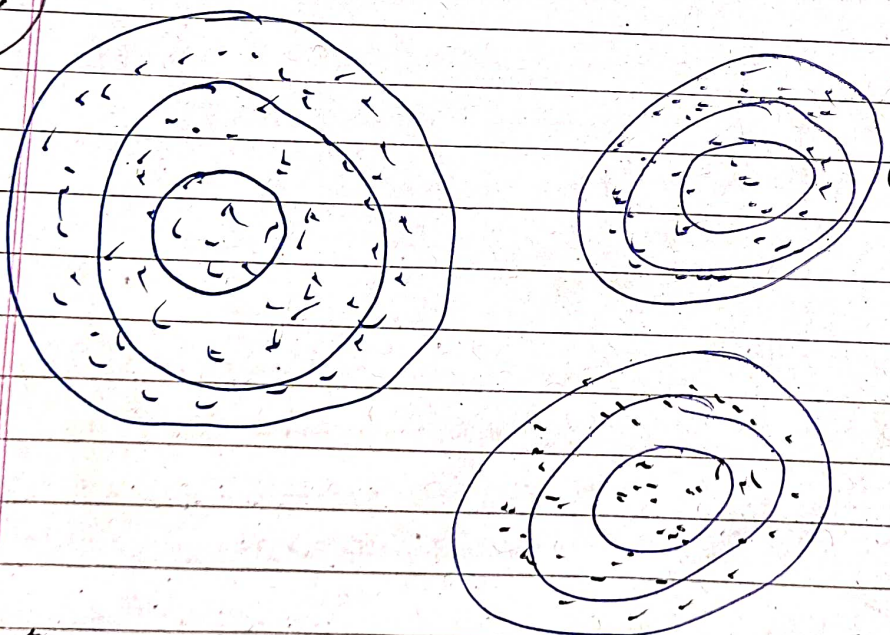
(1) Partitioning Clustering

(2)



(2) Density Based clustering.

(3)



(3) Distribution model Based clustering.

(4)

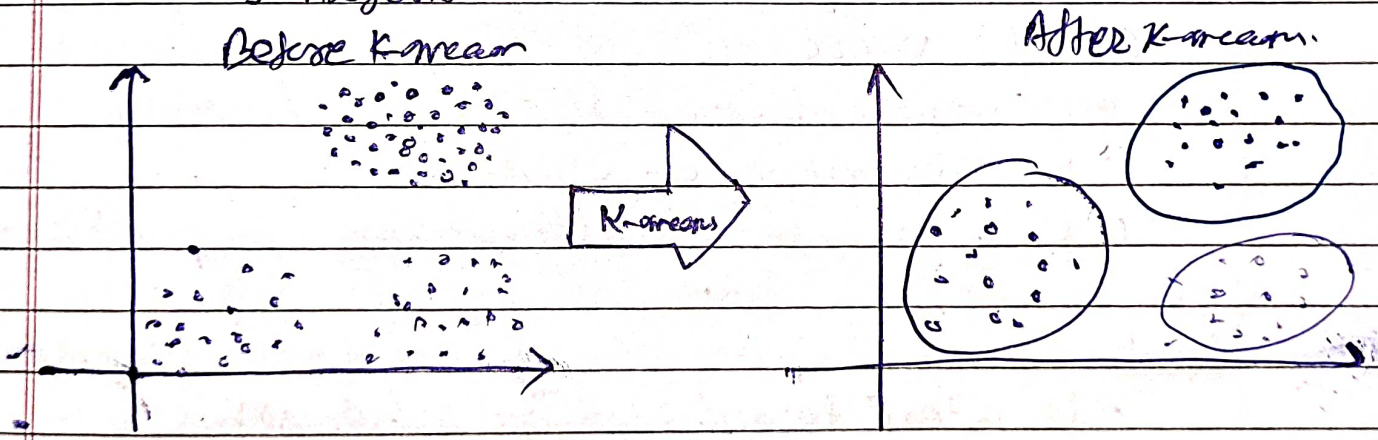
Date : ...../...../.....
Page No. : .....

K-means

K-means clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. "The main aim of this algorithm is to minimize the sum of distance between the data point and their corresponding clusters.

→ The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of K should be predetermined in this algorithm.

The below diagram explains the working of the K-means Clustering Algorithm



The K-means clustering algorithm mainly performs two tasks

- Determines the best value for K center points or centroids by an iterative process
- Assign each data points to its closest K-center.

Date : ..../..../....
Page No. : .....

How does the K-means Algorithm works

Step 1| Select the number K to decide the number of clusters.
Step 2| Select random K points or centroids
Step 3| Assign each data point to their closest centroid, which will form the predefined K clusters.
Step 4| Calculate the variance and place a new centroid of each cluster
Step 5| Repeat the third steps
Step 6| if any reassignment occurs, then go to Step-4 else go to FINISH.
Step 7| The model is ready.

Advantages:-

K-means clustering Algorithm offers the following advantages-

Point - 01 :-

It is relatively efficient with time complexity O(nkt) where-

- n = number of instances
- k = number of clusters
- t = number of iterations

Point - 02:

- It often terminates at local optimum.
- Techniques such as Simulated Annealing or Genetic Algorithms may be used to find the global optimum.

Date : ...../...../.....
Page No. : .....

* Disadvantages
- it requires to specify the number of clusters (k) in advance.
- it can not handle noisy data and outliers
- it is not suitable to identify clusters with non-convex shapes.

Use cases of K-means algorithm.

- Document classification

→ cluster documents in multiple categories based on tags, topics, and the content of the document. This is a standard classification problem and K-means is a highly suitable algorithm for this purpose.

- Delivery store optimization.

→ Optimize the process of good delivery using truck drones by using a combination of K-means to find the optimal number of launch locations and a genetic algorithm to solve the truck route as a traveling salesmen problem.

- Identifying crime localities.

→ with data related to crimes available in specific localities in a city, the category of crime, the area of the crime.

- Customer segmentation.

→ clustering helps marketers to improve their customer base, work on target areas, and segment customer based on purchase history, interests or activity monitoring.

- Insurance fraud Detection

→ Machine learning has a critical role to play in fraud detection and has numerous applications in automobile, healthcare, and insurance fraud detection.

- Call Record Detail Analysis

→ A call detail record (CDR) is the information captured by telecom companies during the call, sms, and internet activity of a customer.

- Automatic clustering of IT Alerts.

→ Large enterprise IT infrastructure technology components such as network, storage, or database generate large volumes of alert messages.

Clustering of data can provide insight into categories of alerts and mean time to repair and help in failure predictions.

Date : ...../...../.....
Page No. : .....

Segmentation

~~Classification~~

Data segmentation is the process of taking the data you hold and dividing it up and grouping similar data together based on the chosen parameters so that you can use it more efficiently within marketing and operations.

Example of Data segmentation could be :-

- Gender
- Customer vs Prospects
- Industry

Segmentation is a process of breaking a group of entities (Parent group) into multiple groups of entities (child group) so that entities in child group have higher homogeneity with in its entities.

* The key benefits of Data segmentation are :-

(1) you will be able to create messaging that is tailored and sophisticated to suit your target market.

(2) it allows you to easier conduct an analysis of your data stored in your database.

(3) Enables you to mass - personalise your marketing communications, reducing costs.

Date : ..../..../....
Page No. : .....

Linear Regression

→ Linear regression is the most basic and commonly used predictive analysis. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

for example → A modeler might want to relate the weight of individuals to their heights using linear regression model.

→ There are several linear regression analyses available to the researcher

- simple linear regression
  > one dependent variable (interval or ratio)
  > one independent variable (interval or ratio or dichotomous)
- multiple linear Regression
  > one dependent variable (interval or ratio)
  > Two or more independent variables (interval or ratio or dichotomous)
- Logistic regression
  > one dependent variable (binary)
  > Two or more independent variable(s) (interval or ratio or dichotomous)
- Ordinal regression
  > one dependent variable (ordinal)
  > one or more independent variable(s) (nominal or dichotomous)

Date : ...../...../.....
Page No. : .....

- Multinomial regression
  > One dependent Variable (nominal)
  > One or more independent Variable(s) (interval or ratio or dichotomous)

- Discriminant analysis
  > One dependent variable (nominal)
  > One or more independent variable(s) (interval or ratio).

Formula for Linear Regression is given by :-

$$y = a + bxy = a + bx$$

where

a = y-intercept of the line
b = slope of the line
x = values of the first data set
y = values of the second data set.

x and y are two variables on the regression line.

a and b are given by the following formulas :-

$$a(\text{intercept}) = (\Sigma y \Sigma x^2) - (\Sigma x \Sigma xy) / (n\Sigma x^2) - (\Sigma x)^2$$

$$b(\text{slope}) = n\Sigma xy - (\Sigma x)(\Sigma y) / (n\Sigma x^2) - (\Sigma x)^2$$

Example :-

Find the linear regression equation for the following two set of data

x	2	4	6	8
y	3	7	5	10

Date : / /
Page No. :

Given:

x	y	x^2	xy
2	3	4	6
4	7	16	28
6	5	36	30
8	10	64	80
Sum $x=20$	Sum $y=25$	Sum $x^2=120$	Sum $xy=144$

Using above formulas for calculating a and b

$b = 0.95$

$a = 1.5$

Linear regression is given by

$y = a + bx$

$y = 1.5 + 0.95x$

Date : ...../...../.....
Page No. : .....

Indexing

→ Indexing is a data structure that improves the speed of data retrieval operations on a database table at the cost of additional writes and storage space to maintain the indexed data structure.

A database query normally performs one of the three functions mentioned below:-

① Range request queries, to obtain a particular range. for example, finding the range of the number of users visiting a store on a weekend.

② Single value queries which may include finding a particular item at a store.

③ To check if a record exists in the database. for example, to check if an item is present in the store.

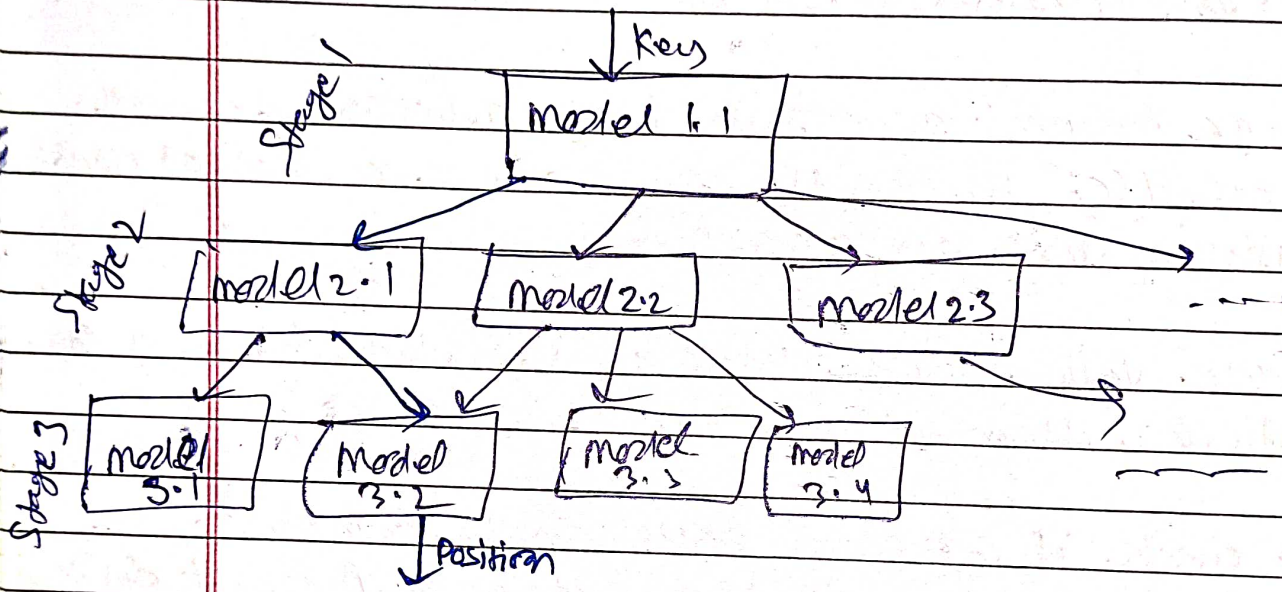
⇒ When we pass a query into our database, depending on the types of the query, different indexing method are used to obtain an answer. The three different types of indexing methods are as follows:-

① for range request queries, BTree-index is used
② To look up a record for a single key, Hash maps are used.
③ To determine if a record exists in the database Bit map index (Bloom filter) is used.

Date : ../../..
Page No. : .....

(i) Range Queries : On this case, the data is stored in sorted order. An index is built to find the starting point of the range. The process is very straightforward.

A B-Tree is used to carry out this function. B-Trees are used because they are memory efficient. However, processing a node of a B-Tree takes time.



{Staged model}

(ii) Single Value Queries

→ Point indexes or hash maps are used when a query is made to obtain a single value. Hash maps work by taking a function h such that $h(x) \rightarrow Pos$. Traditionally, a good hash function is one for which there is no correlation between the distribution of x and positions.

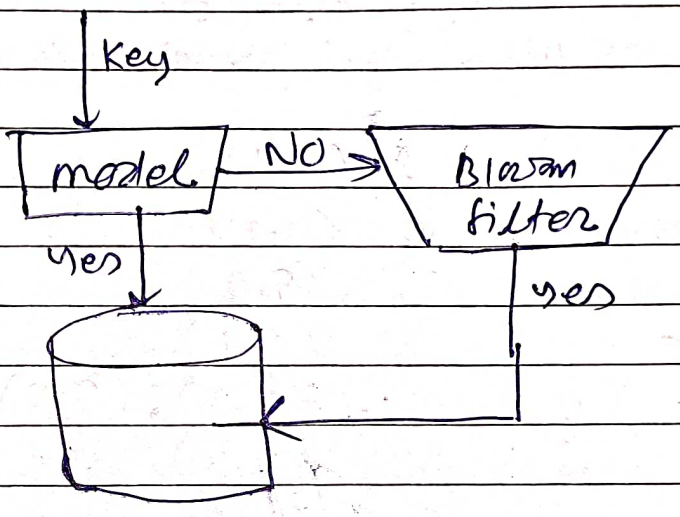
The authors consider placing N keys in an array with m positions. This is done because two keys can be mapped to the same position as suggested by the birthday paradox.

(11) Existence Index

→ Existence indexes are important to determine if a particular key is in our dataset, such as to confirm its in our dataset before retrieving data from cold storage.

Hence,

this task can be considered as a classification problem. A value either exists in the database or it doesn't. Bloom filters have been long used to find out if a value exists in the database or not.



Inverted Index

→ An inverted index is an index data structure storing a mapping from content, such as words or numbers, to its locations in a document or a set of documents.

In simple words.

it is a hashmap like data structure that directs you from a word to a document or a web page.

There are two types of inverted indexes:

- A record level inverted index
- A word-level inverted index

Date : ..../..../....
Page No. : .....

→ A record-level inverted index contains a list of references to documents for each word.

→ A word-level inverted index additionally contains the positions of each word within a document.

Suppose we want to search the texts
"hello everyone." "this article is based on inverted index," "which is hashmap like data structure".

If we index by (text, word within the text), the index with location in text is:

hello	(1,1)
everyone	(1,2)
this	(2,1)
article	(2,2)
is	(2,3); (3,2)
based	(2,4)
on	(2,5)
inverted	(2,6)
index	(2,7)
which	(3,1)
hashmap	(3,3)
like	(3,4)
data	(3,5)
structure	(3,6)

→ The word "hello" is in document 1 ("hello everyone.") starting at word 1, so has an entry (1,1) and word "is" is in document 2 and 3 at '3rd' and '2nd' positions respectively (here position is based on word).

The index may have weights, frequencies, or other indicators.

Date : ...../...../.....
Page No. : .....

Steps to build an inverted index
- fetch the document
- Stemming of Root word
- Record Document IDs

Example:

Words	Documents
ant	doc1
demo	doc2
world	doc1, doc2

Advantages of inverted index

- Inverted index is to allow fast full text searches.
- It is easy to develop.
- It is the most popular data structure used in document retrieval systems.

Inverted index also has disadvantages
- Large storage overhead and high maintenance costs on update, delete and insert.

Date : ..../..../....
Page No. : .....

Data Exploration

→ Data exploration is the initial step in data analysis, where user explore a large data set in an unstructured way to uncover initial patterns, characteristics, and points of interest.

→ Applications of Data Exploration

① In any situation where you have a massive set of information, data exploration can help cut it down to a manageable size and focus effort to optimize your analysis.

② Data exploration can also assist by reducing work time and finding more useful and actionable insights from the start alongside presenting clear paths to perform better analysis.

Date : ...../...../.....
Page No. : .....

Classification

Classification is types of supervised learning algorithms when the output variable is continuous, then it is a regression problem whereas when it contains discrete values, it is a classification problem.

Classification is another fundamental learning method that appears in applications related to data mining. In classification learning, A classifier is presented with a set of examples that are already classified.

Classification is the process of learning a model that elucidate different predetermined classes of data. It is a two-step process, comprised of a "learning step" and a "classification step".

In learning step, A classification model is constructed. and in classification step, the constructed model is used to prefigure the class labels for given data.

For example, in a banking application, the customer who applies for a loan may be classified as a safe and risky according to his/her age and salary. This type of activity is also called supervised learning.

→ The algorithm which implements the classification on a dataset is known as a classifier.

Two types

- Binary classifier
- Multi-class classifier

Date : ..../..../....
Page No. : .....

- Binary classifier → if the classification problem has only two possible outcomes, then it is called as Binary classifier.

Example → YES or NO, MALE or FEMALE, CAT or DOG
SPAM or NOT SPAM

- Multi-class classifier : if a classification problem has more than two outcomes, then it is called as multi-class classifier
Example → classification of types of crops
classification of types of music.

* Learners in classification problems.
In classification problem, there are two types of learner.

(1) Lazy Learners :- Lazy learner firstly stores the training dataset and wait until it receives the test dataset. In lazy learner case, classification is done on the basis of the most related data stored in the training dataset.

It takes less time in training but more time for predictions.
Ex → K-NN algorithm, Case-based reasoning.

(2) Eager learners → Eager learner develop a classification model based on a training dataset before receiving a test dataset.
Opposite to lazy learners, Eager learner takes more time in learning, and less time in Prediction
Example → Decision Tree, Naive Bayes, ANNs.

Date: ..../..../....
Page No.: .....

Types of ML classification Algorithms

- Linear models
  - > logistic Regression
  - > support vector machines
- Non-linear models
  - > K-nearest Neighbours
  - > Kernel SVM
  - > Naive Bayes
  - > Decision Tree classification
  - > Random forest classification

Decision Tree

→ Decision trees are a type of supervised learning algorithm where data will continuously be divided into different categories according to certain parameters.

Decision Tree is a flow like a tree structure that works on the principle of conditions. It is efficient and has strong algorithm used for Predictive analysis

it has mainly attributes that include internal Nodes, branches and a terminal node.

Every internal Node holds a "test" on an attribute, branches hold the conclusion of the test and every leaf Node means the class label.

It is used for both classification as well as Regression.

It is often termed as "CART" that means "classification and Regression Tree".

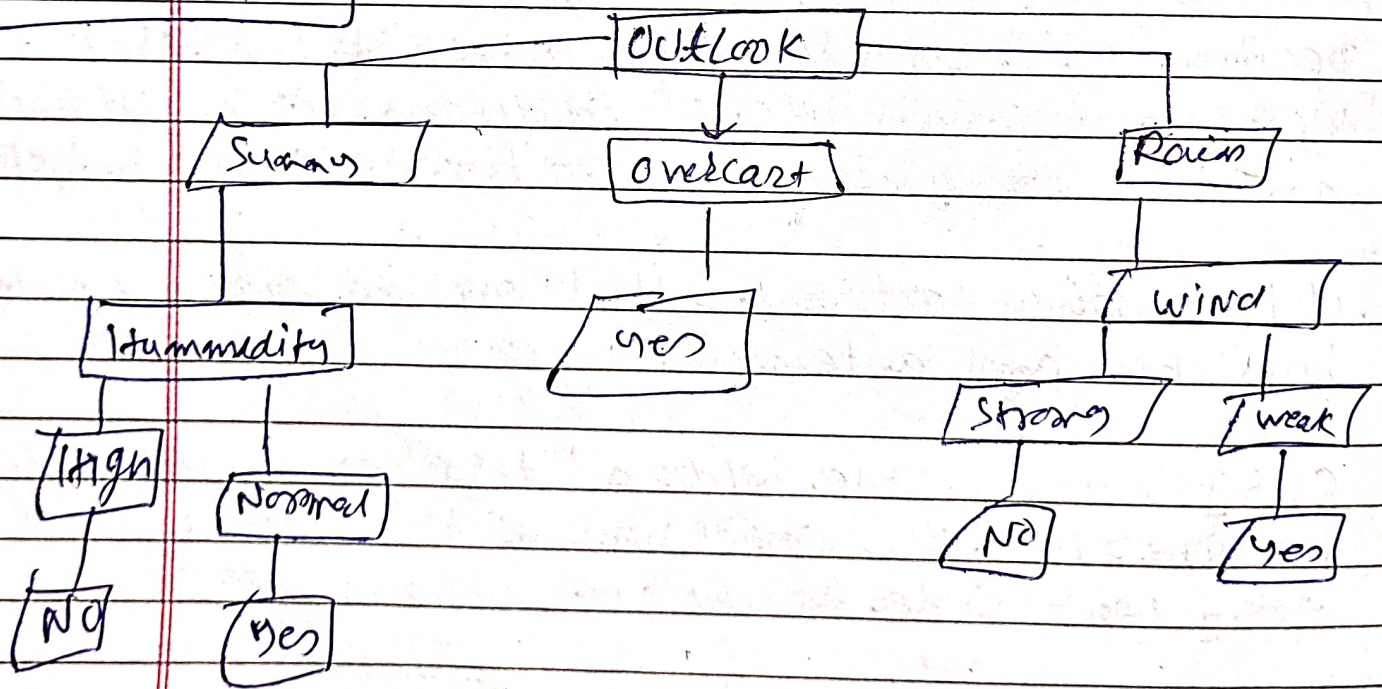
Tree algorithms are always preferred due to stability and reliability.

How can an algorithm be used to represent a tree.

→ Lets us see an Example of a basic decision Tree where it is to be decided on what conditions to Play cricket and on what conditions not to Play.

Day	Weather	Temp	Humidity	Wind	Play?
1	Sunny	Hot	High	weak	No
2	Cloudy	Hot	High	weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

Decision Tree

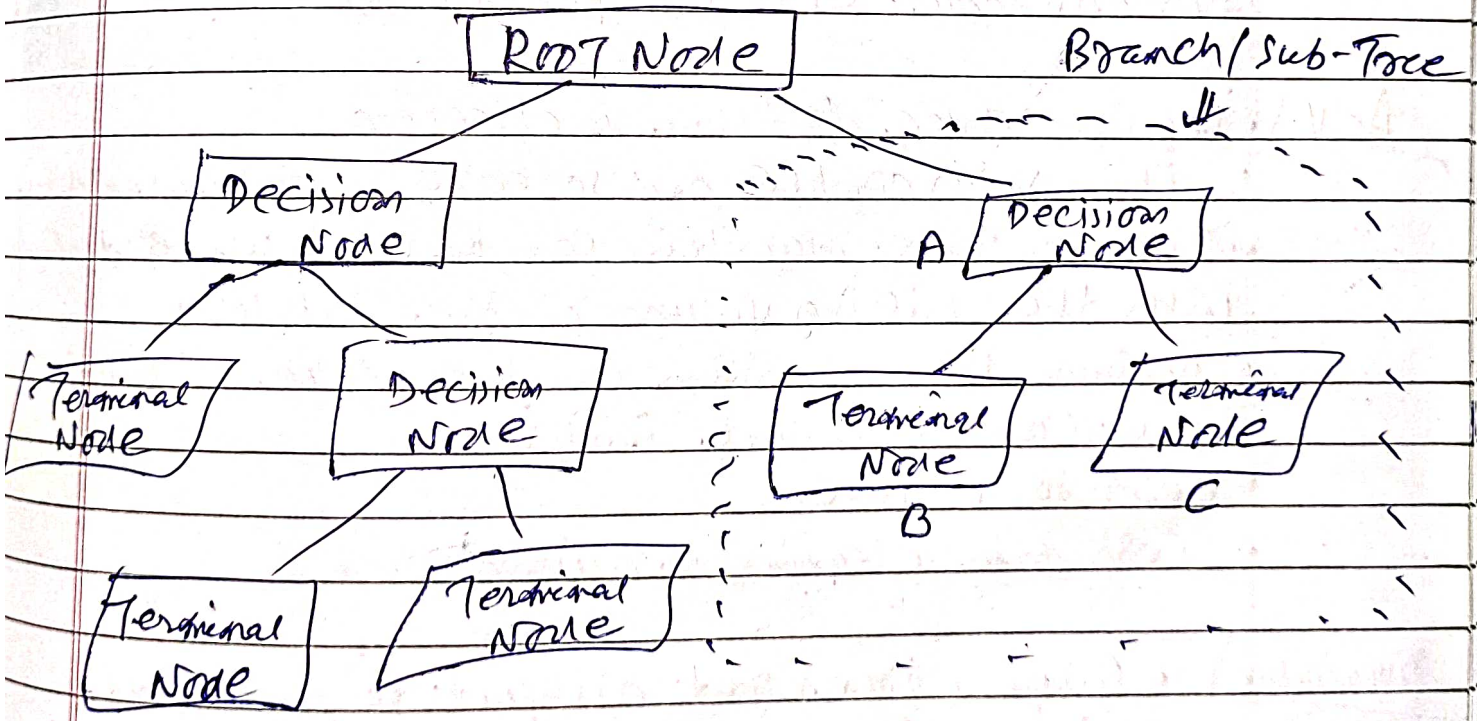


Decision Tree for playing cricket.

Date : ../../..
Page No. : .....

Q) The common terms used in Decision Tree are :-

- Branches :- Division of the whole tree is called Branches
- Root Node :- Represent the whole sample that is further divided.
- Splitting :- Division of Nodes is called splitting
- Terminal Node :- Node that does not split further is called a terminal Node.
- Decision Node :- it is a Node that also gets further divided into different sub-nodes being a sub node.
- Pruning :- Removal of subnodes from a decision Node.
- Parent and child Node → When a node gets divided further then that Node is termed as Parent Node whereas, the divided Nodes or the sub-nodes are termed as a child node of the parent Node.



Note :- A is Parent Node of B and C.

Date : ../../..
Page No. : .....

How Does Decision Tree Algo. Work.

→ it works on both the type of input & output that is categorical and continuous.
on classification problems, the decision tree asks questions, and based on their answers (Yes/No) it splits data into further sub branches.
it can also be used as a binary classification problem.

On a decision tree, the algorithm starts with a root node of a tree, then compares the values of different attributes and follows the next branch until it reaches the end leaf Node.
it uses different algorithms to check about the split and variable that allow the best homogeneous sets of population.

Adv.
- Decision Tree algorithm is effective.
- it is very simple and so easy to read and interpret.
- Decision Tree algorithm can be used while dealing with the missing values in the dataset.
- Decision Tree algorithm can take care of numeric as well as categorical features.
- Easy to Prepare.
- Less data cleaning required.

Disadvantage
- larger trees get difficult to interpret.
- Biased towards tree having more levels.

UNIT-4

Date : ..../..../....
Page No. : .....

Stream computing

→ Stream computing is a computing paradigm that reads data from collection of software or hardware sensors in a stream form and computes continuous data streams. where feedback results should be in a real-time data stream as well.

→ The word stream in stream computing is used to mean pulling in stream data, processing the data and streaming it back out as a single flow.

• Stream computing uses software algorithms that analyzes the data in real time as it streams into increase speed and accuracy when dealing with data handling and analysis.

Streaming analytics is the processing and analyzing of data records continuously rather than in batches.

Streaming analytics use cases → • News media • Utilities.

• ECommerce

→ Analyze user clickstreams to optimize the shopping experience with real-time pricing, promotion and inventory management.

• financial services

Analyze account activity to detect anomalous behaviour in the data stream and generate a security alert for abnormal behaviour

• Investment Services → Track market changes and adjust settings to customer platforms and enrich the data with demographic information to better services.

Date : ..../..../....
Page No. : .....

Challenges limitations of conventional systems.

Common challenges

→ It cannot work on unstructured data efficiently.
→ It is built on top of the relational data model.
→ It is batch oriented and we need to wait for nightly ETL (extract, transform and load) and transformation jobs to complete before the required insight is obtained.
→ Inadequate support of aggregated summaries of data.

Data challenges

- Volume, velocity, variety & veracity
- Data discovery and comprehensiveness
- Scalability
- Storage issues

Process challenges

- Capturing data
- Aligning data from different sources.
- Transforming data into suitable form for data analysis
- Modeling data (mathematically, simulation)
- Understanding output, visualizing results and display issues on mobile devices.

Management challenges

- Security, • Privacy, • Governance
- Ethical issues

Date : ..../..../....
Page No. : .....

(1) Volume
→ The volume of data, especially machine-generated data is exploding.
→ how fast data is growing every year, with new sources of data that are emerging.

(2) Processing
More than 80% of todays information is unstructured and it is typically too big to manage effectively.

(3) Management
→ A lot of this data is unstructured, or has complex structure that is hard to represent on rows and columns.

(4) Lack of Knowledge professionals
(5) Lack of Proper understanding of massive data
(6) Data Growth issues
(7) Confusion while Big data Tool selection
(8) Integrating Data from a spread of sources

Date : ..../..../....
Page No. : .....

Solving a real-time analytics problem using conventional system.

→ Today, companies typically store a copy of their operational data in a data lake, often built on Hadoop, where it is avaible for later analysis.
However.
for a growing number of digital transformations and omnichannel customer experiences, companies find they must run real-time analytics across their operational data set and a subset of the data in their data lake. For traditional infrastructures, real-time analytics proves challenging because of delays in accessing and processing data in a data lake and the difficulties in running federated queries across operational and archived data.

Mature in-memory computing (IMC) technology can help resolve these obstacles. The Technology offers real-time performance and massive scalability with built-in integrations of Polular data platforms.
The platforms are capable of running real-time analytics across operational and data lake data sets.

IMC ingests, processes and analyzes operational data with real-time performance and scalability to petabytes of in-memory data.

→ Some IMC platforms use built-in integrations to connect with Polular Streaming data platforms, such as "Apache Kafka", and data processing tools, such as "Apache Spark for connecting to Apache Hadoop.

Date : ../../..
Page No. : .....

- **Apache Kafka :-** Kafka builds the data pipelines and streaming apps that process encoming data in real time.

- **Apache Spark :** Spark is a unified analytics engine that performs large-scale data processing on data, such as powering federated queries and transferring data from a Hadoop-based data lake to an operational data store.

- **Apache Hadoop :** Hadoop includes a distributed file system that provides high-throughput access to application data.

Challenges TO be solved

SCALABILITY

→ Scalability is the property of a system to handle a growing amount of work by adding resources to the system.

A scalable data platform utilize added hardware or software to increase output and storage of data. When a company has a scalable data platform, it also prepared for the potential of growth in its data needs.

Types : SCALE IN
SCALE OUT

Date : / /
Page No. :

THREAD POOLING

→ A thread Pool reuses Previously created threads to execute current tasks and offers a solution to the Problem of thread cycle overhead and resource thrashing.
Since the thread is already existing when the request arrives, the delay introduced by thread creation is eliminated, making the application more responsive.

A thread pool is a collection of worker threads that efficiently execute asynchronous callbacks on behalf of the application.

Java Thread Pool Represents a group of worker threads that are waiting for the job and reused many times.

In the case of a thread pool, a group of fixed size threads is created. A thread from the thread pool is pulled out and assigned a job by the service provider. After completion of the job, the thread is contained in the thread pool again.

The thread pool is Primarily used to reduce the Number of application threads and Provide management of the worker threads.

* Risk in using Thread Pools.
(i) Deadlock
(ii) Thread leakage
(iii) Resource Thrashing.

Date : ..../..../....
Page No. : .....

* Deadlock → while deadlock can occur in any multi-threaded Programs, thread Pools introduce another case of deadlock,
one in which all the executing threads are waiting for the results from the blocked threads waiting in the queue due to the unavailability of threads for execution.

* Thread Leakage → Thread leakage occurs if a thread is removed from the pool to execute a task but NOT returned to it when the task completed.

* Resource Thrashing → if the thread pool size is very large then time is wasted in context switching between threads. Having more threads than the optimal number may cause starvation problem leading to resource thrashing as explained.

Date : ../../..
Page No. : .....

Benefits of Streaming Computing in Big data world.

→ Five Top Benefits of Stream Analytics are :-

- Improve operational efficiencies
→ Enables data and analytics team to understand ongoing events faster and act on them immediately.

- Reduce infrastructure cost
→ Streaming analytics and ingestion cloud services can help your organization manage workloads efficiently by auto-scaling the clusters to optimize costs.

- Provide faster insights and actions
→ An End-to-end, AI-powered streaming analytics solution can ingest any data from any source at any latency to process and operationalize it for faster and continuous insights to all users.

- Increase ROI
→ The ability to quickly collect, analyze and act on current data will give companies a competitive edge in their marketplace.
Real-time intelligence makes organizations more responsive to market trends, customer needs and business opportunities.

- Increase customer satisfaction.
→ Customer feedback is a valuable litmus test for what an organization is doing right and where it can improve.

- Reduce losses → Not only does data streaming support customer retention, but it prevents other losses as well.

→ Real-time intelligence can provide warnings of impending issues such as system outages, financial downturns, data breaches, and other issues that negatively effect business outcomes.

RTAP

Realtime Analytics Platforms

→ Realtime analytics permits business to get awareness and take action on data immediately or soon after the data enters their system. Real time apl analytics response queries within seconds. They grasp a large amount of data with high velocity and low reaction time.

for example :- ~~They~~ real-time big data analytics uses data in financial databases to ratify trading decisions.

Analytics can be on-demand or continuously.

On-demand ratifies results when the user requests it. Continuous renovation users as events happen and can be programmed to answer automatically to certain events.

for example, real-time web analytics might refurbish an administrator if the Page load presentation goes out of the present boundary.

They handle large amounts of data with high velocity and low response times.

Examples Examples of real-time customers analytics include on-

- viewing orders as they happen for better tracing and to identify fashion.

- choose customers with advancement as they shop for items in a store, affecting real-time decisions.

(RTAP)

Date : ..../..../....
Page No. : .....

Advantages of Real-time Analytics Platform

- Create custom interactive analytics tools.
- Share information through transparent dashboards.
- Customize monitoring of behavior.
- Make immediate changes when needed.
- Apply machine learning.

Real time Sentiment Analysis

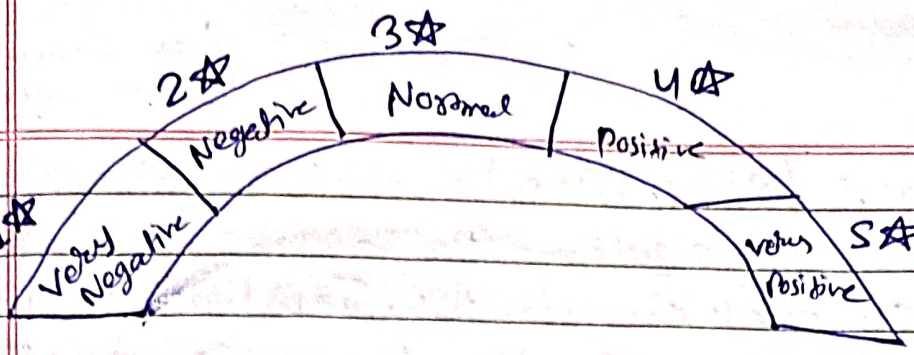
→ Sentiment analysis is a text analysis tool that uses machine learning with natural language Processing (NLP) to automatically read and classify text as Positive, Negative, Neutral, and everywhere in between. It can read all manner of text (online and elsewhere) for opinion and emotion - to understand the thoughts and feelings of the writer.

Sentiment Analysis (also known as opinion mining or emotion AI) is a sub-field of "NLP" that tries to identify and extract opinions within a given text across blogs, reviews, social media, forums, news etc.

Sentiment analysis looks at the emotion expressed in a text. It is commonly used to analyze customer feedback, survey response, and Product reviews.

Social media monitoring, reputation management, and customer experience are just a few areas that can benefit from Sentiment Analysis.

For Example :- Analyzing thousands of Product reviews can generate useful feedback on your pricing and Product features.



Date : ..../..../....
Page No. : .....

One easy way to do this with customer reviews is to rank 1-star reviews as "very Negative". 5-star reviews would be ranked as "very positive".

Benefits of Performing real-time sentiment Analysis.

- Marketing campaign success analysis.

→ Target your analysis to follow marketing campaign right as they launch and get a solid idea of how your messaging is working with current and potentially new customers.

- Stock market predictions

→ follow the real-time sentiment of any business as it rises and falls to get up-to-the-minute information on stock price changes.

- Process data at scale (more powerful)

→ Sentiment analysis helps businesses make sense of huge quantities of unstructured data. When you work with text, even 50 examples already can feel like Big data. Especially, when you deal with people's opinions in product reviews or on social media.

Date : ..../..../....
Page No. : .....

- Automation! (save time)

→ Sentiment analysis algorithms can analyze hundreds of megabytes of text in minutes. Instead of manually analyzing data in spreadsheets, you can now spend your time on more valuable activities.

- Real time analysis and insights. (Act faster)

→ Sentiment analysis is automated using machine learning. This means that businesses can get insights in real-time. This can be very helpful when identifying issues that need to be addressed right away.

How to Perform Sentiment Analysis in Real Time.

1. Set your goals
2. Gather your data
3. Clean your data
4. Analyze & visualize sentiments in real-time
5. Act on your results

There are two options when it comes to performing Sentiment Analysis :-

- build a model or
- convert in a SaaS tool.

→

Build a model can produce exceptional results, but it is time-consuming and costly.

→

SaaS tools are generally ready to put into use right away, much less expensive, and you can still train custom models to the specific language, needs, and criteria of your organisation.

Date : ..../..../....
Page No. : .....

3) MonkeyLearn's Powerful Saas Platform offers immediate access to sentiment analysis Tools and other text analytics techniques.

And with MonkeyLearn Studio, you can analyze and visualize your results in real time.

(1) Set your goals.

→ First, decide what you want to achieve. Do you want to compare sentiment toward your brand against that of your competition? Do you want to regularly online twitter or perform social listening to extract brand mentions and follow your brand sentiment from minute to minute?

2. Gather your data

→ There are a number of ways to get the data you need, from simply cutting and pasting, to using APIs.

APIs :-

- The Graph API is best for Pulling data directly from facebook
- Twitter's API allows users access to Public Twitter data.

3. Clean your data

→ Website, social media, and email data often have quite a bit of "noise". This can be repetitive text, banner ads, non-text symbols and emojis, email signatures, etc. You need to first remove this unnecessary data, or it will skew your results.

Date : ..../..../....
Page No. : .....

(4) Analyze & visualize sentiments in real time

→ MonkeyLearn Studio is an all-in-one real-time sentiment analysis and visualization tool. After a simple set-up, you just upload your data and visualize the results for powerful insights.

(5) Act on your results.

→ The results are in! With sentiment analysis and MonkeyLearn Studio, you can be confident your are making real-time, data-driven decisions.