1. List and define 5V's of Big Data. list different challenges and issues in Big Data.

→ The 5V's of Big Data are as follows -

(1) volume

→ The name 'Big Data' itself is related to a size which is enormous. volume is a huge amount of data. To determine the volume of data, size of data plays a crucial role. If the volume of data is very large then it is actually considered as 'Big Data'.

(2) velocity

→ It refers to the high speed of accumulation of data. In Big Data velocity data flows in from sources like machines, networks, social media, mobile phones etc

(3) variety

→ It refers to nature of data that is structured, semi-structured and unstructured also

(4) veracity

→ It refers to inconsistencies and uncertainity in data which is available can sometimes get messy and quality and accuracy are ~~different~~ difficult to control.

(5) value

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

Teachers Sign

→ some of the Big Data challenges are -
(1) sharing and accessing data
(2) Privacy and security
(3) Analytical challenges
(4) Technical challenges

2. List and explain different components of Hadoop
→ (1) HDFS
→ Hadoop distributed file system is the backbone of Hadoop which runs on java language and stores data in Hadoop applications. They act as a command interface to interact with Hadoop. The two components of HDFS are - Data Node and Name Node

→ Name Node manages file system and operates all data nodes and maintains records of meta data updating. In case of deletion of data, they automatically record it in Edit log

→ Data Node requires vast storage space due to reading and writing operations. They work according to the instructions of the Name node.

→ (2) HBASE
→ It is an opensource framework storing all types of data and doesn't support the SQL database. They run on top of HDFS and written in java language.

→ The two major components of HBASE are :- HBASE master and regional server. HBASE master is responsible for load balancing in a Hadoop cluster and controls the failover. The regional server's role would be a worker node and responsible for reading, writing data in cache.

→ (3) YARN
→ It is an important component in the ecosystem and called an operating system in Hadoop which provides resource management and job scheduling task. The components are resource and Node manager, application manager and container. They also act as guards.

→ (4) Sqoop
→ It is a tool that helps in data transfer between HDFS and MySQL and gives hand on to import and export data, they have a connector for fetching and connecting data

→ (5) Apache Spark
→ It is an open source cluster computing framework for data analytic and an essential data processing engine

→ (6) Apache Flume
→ It is distributed service collecting a large amount of data from the source and moves back to its origin and transferred to HDFS.

→ (7) Hadoop Map Reduce
→ It is responsible for data processing and acts as a core component of Hadoop

→ (8) Apache Pig
→ Data manipulation of Hadoop is performed by Apache Pig. It helps in the reuse of code and easy to read and write code.

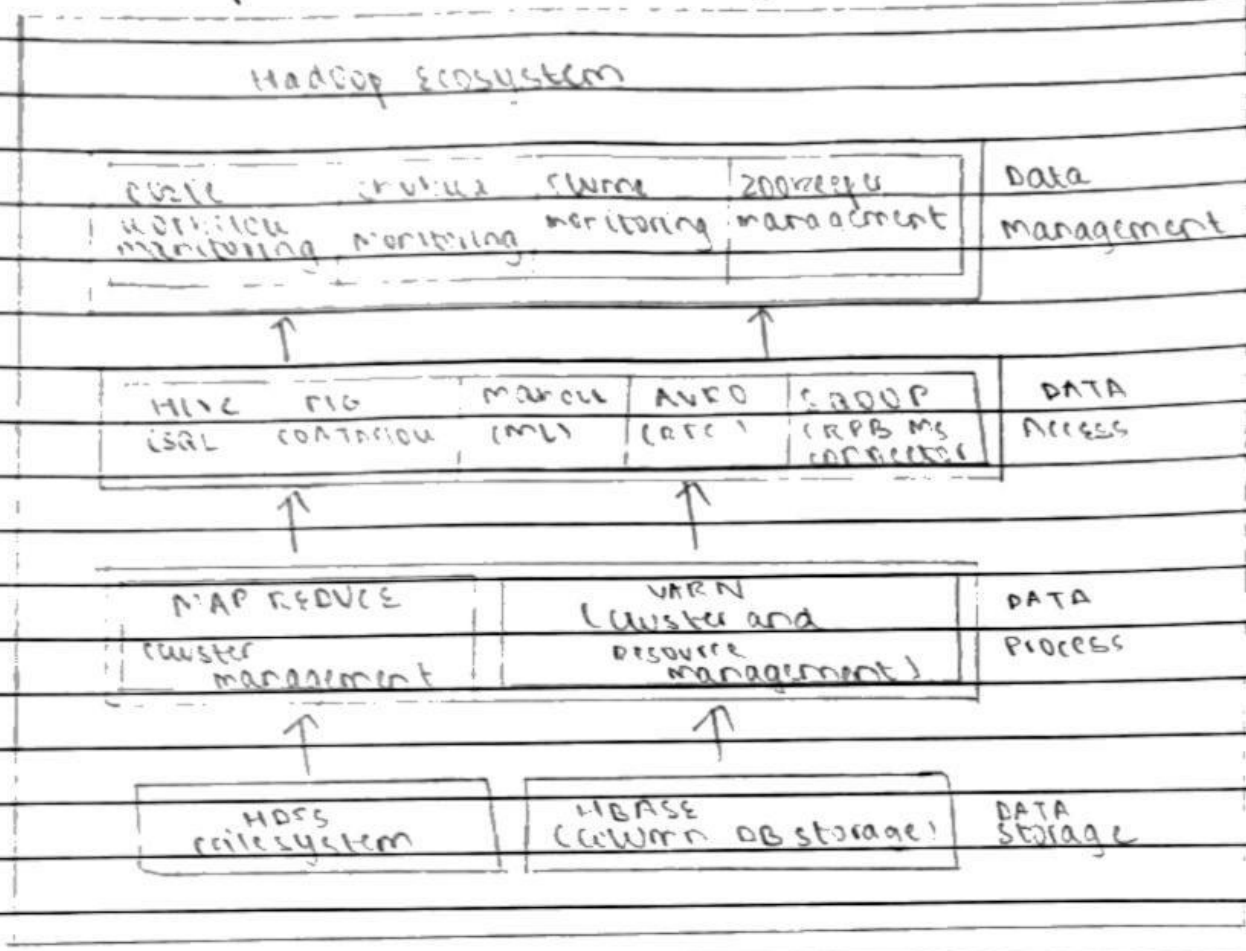
→ (9) Hive
→ It is an open source platform for performing data warehousing concepts; it manages to query large data sets stored in HDFS

→ (10) Apache Drill
→ open source SQL engine which process non-relational databases and file systems

→ (11) Apache zookeeper
→ It is an API that helps in distributed coordination.

→ (12) Oozie
→ It is a java web application that maintains many workflows in a Hadoop cluster

3 Draw and explain the 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 commercial tools and solutions

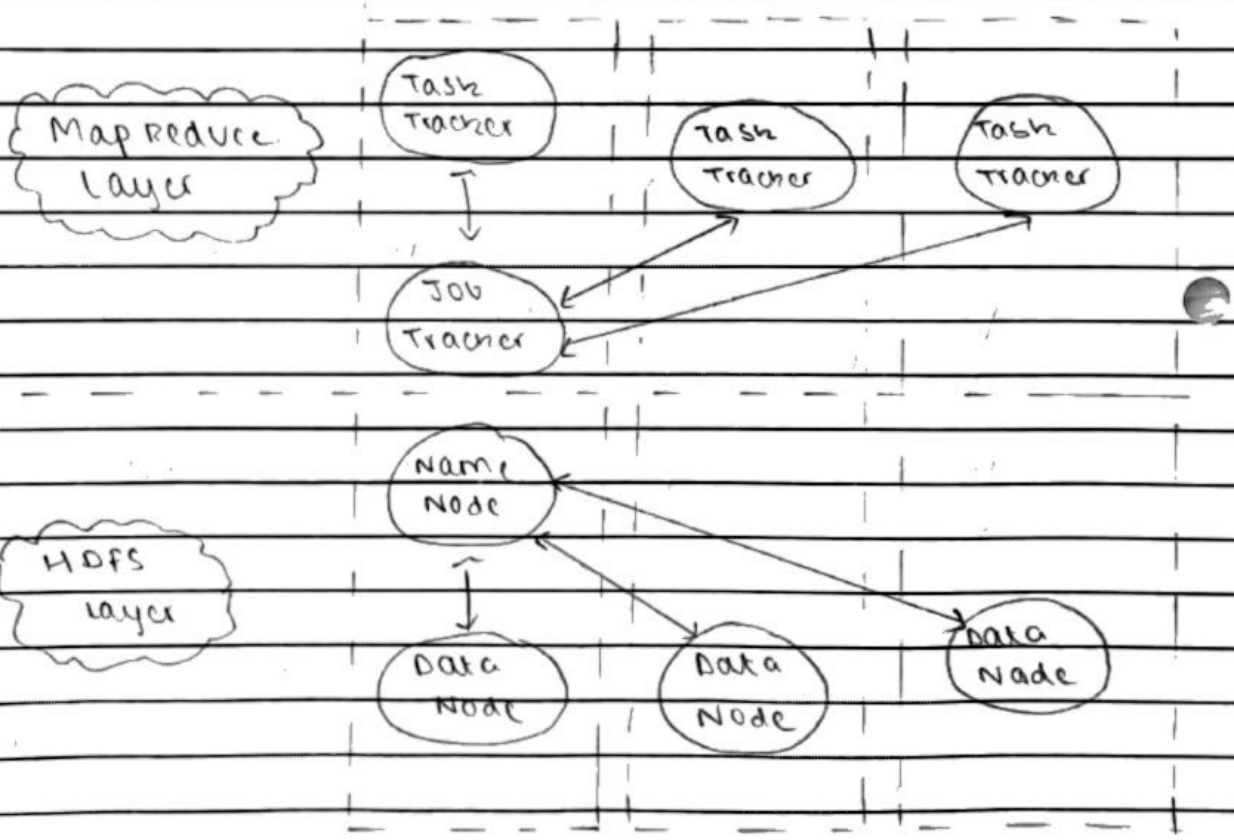
→ There are 4 major elements of Hadoop - HDFS, Map Reduce, YARN and Hadoop common. Most of the tools work collectively to provide services such as absorption, analysis, storage and maintenance of data

Teachers Sign

→ Components of Hadoop -

1. HDFS → Hadoop distributed file system
2. YARN → Yet Another resource negotiator
3. MapReduce → Programming based data process
4. Spark → In memory data processing
5. PIG HIVE → Query based processing on data service
6. HBASE → NO SQL Data Base
7. Mahout → ML algorithm libraries
8. Solar, Lucence → searching and Indexing
9. Zookeeper → Managing cluster
10. Oozie → Job scheduling

4. Draw and explain architecture of Hadoop. List limitations of Hadoop



→ Name node
It represents every files and directory which is used in the ~~##~~ namepace

→ Data Node
It helps you to manage the state of an HDFS node and allows you to interact with the blocks

→ Master Node
It allows you to conduct parallel processing of data using Hadoop MapReduce

→ Slave Node
Allows you to store data for complex calculations.

→ limitations of Hadoop →
(1) Problem with small files
(2) vulnerability
(3) low performance in small data surroundings
(4) lack of security
(5) High up processing
(6) Batch ~~##~~ processing is only supported

5. What is MapReduce Engine? Explain its principle execution

→ It is a processing technique and a program model for distribution

→ The MapReduce algorithm contains two main tasks, namely Map and Reduce

→ Map takes a set of data and converts it into another set of data, where individual elements are broken down

→ Reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after map job.

→ The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes.

6 Advantages of NoSQL

→ Big data capability

→ No single point failure

→ Easy replication

→ Fast performance and horizontal scalability

→ Can handle unstructured data as well

→ Multiple vendor choices are available

→ Easy and human readable query

18dcs007

TOES

Name .....
Subject..... Chapter.....
Date ..... Page No. .....

7. List the different NOSQL data architectural patterns? Explain with diagram and give one example of each

→ (1) Key-value store database

→ This model is one of the most basic models of NOSQL database. As the name suggests, the data is stored in form of key-value pairs. The value is typically linked on co-related to the key. The key value pair storage database generally store data as a hash table where each key is unique

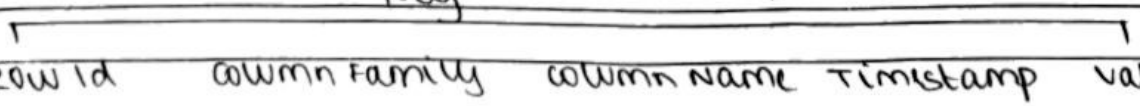
Eg - DyonamoDB, Berkeley DB

key	value
k_1	v_1
k_2	v_2
:	:
k_m	v_m

→ (2) Column store database

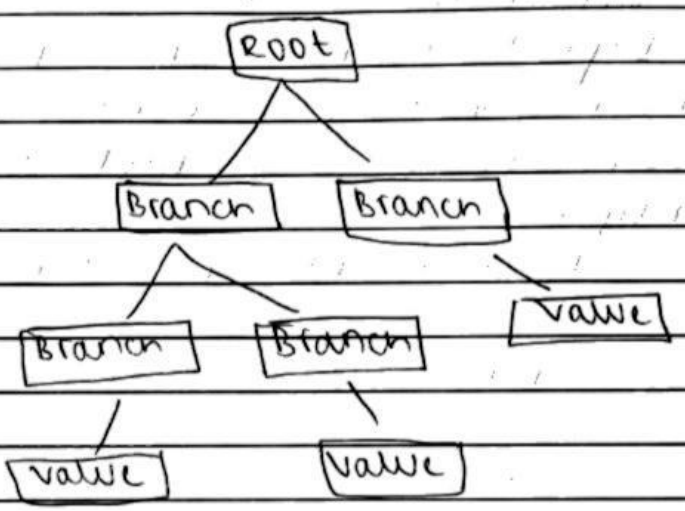
→ The data is stored in individual cells which are further grouped into columns. Column oriented databases work only on columns. They store large amount of data

Eg - HBASE, Bigtable by google

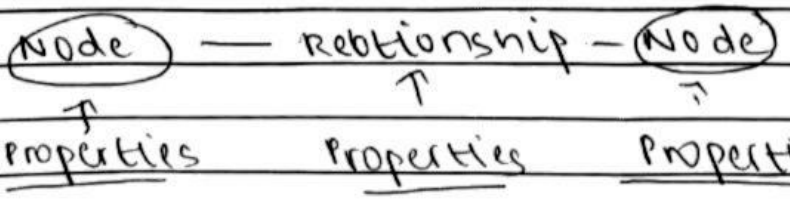


Teachers Sign

→ (3) Document database
→ It fetches and accumulates data in form of key value pairs but values are called documents. It has tree like structure and can be complex
Eg - Mango DB, Couch DB



→ (4) Graph database
→ It deals with storage and ~~Manu~~ management of data in graphs. Objects are called as nodes and are joined together by relationship called edges
Eg - Neo4J, Flock DB



8. Hamming distance
→ calculates distance between two binary vectors
→ Mostly used when dealing with one-hot-encode categorial columns
Hamming distance = sum for i to N abs

Euclidean distance
→ calculates distance between two real valued vectors
→ Mostly used for columns with numerical values like int and float
Euclidean distance = sqrt (sum for i to N (v1[i] - v2[i])²)
$= \sqrt{\sum (x_1 - y_1)^2}$

Manhattam distance
→ More useful to vectors that describe objects on uniform grid like chessboard
Manhattam distance = $|x_1 - x_2| + |y_1 + y_2|$ for point y_1, y_2, x_1, x_2

Minkowski distance
→ It is generalization of euclidean and Manhattam distance
Minkowski distance = $\sqrt[p]{\sum |x_1 - y_1|^p}$

9. Steam data
→ stream data is data that continously generated by different sources.
characteristics - Time sensitive
- continuous
- Heterogenous
- Imperfect
- volatile

Teachers Sign

1rd15007

challenges - choosing data formats
- Algorithm testing, life cycle management
- scaling and performance

→ sampling methods - Bernoulli sampling
- Resorvoir sampling
- stratified sampling
- Based reservoir sampling

10.	Traditional Data Approach	Big Data Approach
→	Data integration is easy	Data integration is difficult
→	centralized and managed in centralized form	Managed in distributed form
→	volume ranges from GB to TB	volume ranges from petabytes to zettabytes
*	Document store	Graph store
→	stores the data in form of document as value	stores the data in node connected by edges
→	eg Mango DB	eg Neo4J
*	Job tracker	Task tracker
→	Runs on seperate node	Runs on all nodes
→	Replaced by resource manger	Replaced by Node manager in MR2
→	receives requests for execution from client	Assigned mapper and reducer tasks to execute

*	Name Node	Data Node
→	centerpiece of HDFS	Responsible for storing actual data
→	known as Master	known as slave
→	Single point of failure	If its down, it doesnt affect availability of data

*	DBMS	NOSQL
→	Data stored as file system	Non-relational database system
→	supports single user only	Multiple user
→	Mix of open source	Open source only