1a

Big Data Characteristics

Big Data contains a large amount of data that is not being processed by traditional data storage or the processing unit. It is used by many **multinational companies** to **process** the data and business of many **organizations**.

- **Volume**
- **Variety**
- **Velocity**

Volume :

- The name Big Data itself is related to an enormous size. Big Data is a vast 'volumes' of data generated from many sources daily, such as **business processes, machines, social media platforms, networks, human interactions,** and many more.
- **Facebook** can generate approximately a **billion** messages, **4.5 billion** times that the "**Like**" button is recorded, and more than **350 million** new posts are uploaded each day. Big data technologies can handle large amounts of data.

Variety

Big Data can be **structured, unstructured, and semi-structured** that are being collected from different sources. Data will only be collected from **databases** and **sheets** in the past, But these days the data will comes in array forms, that are **PDFs, Emails, audios, SM posts, photos, videos,** etc.

The data is categorized as below:

A Structured data: In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.

a. **Semi-structured:** In Semi-structured, the schema is not appropriately defined, e.g., **JSON, XML, CSV, TSV**, and **email**. OLTP (**Online Transaction Processing**) systems are built to work with semi-structured data. It is stored in relations, i.e., **tables**.

b. **Unstructured Data**: All the **unstructured files, log files, audio files**, and **image** files are included in the unstructured data. Some organizations have

much data available, but they did not know how to **derive** the value of data since the data is raw

Velocity

Velocity plays an important role compared to others. Velocity creates the speed by which the data is created in **real-time**. It contains the linking of incoming **data sets speeds, rate of change**, and **activity bursts**. The primary aspect of Big Data is to provide demanding data rapidly.

Big data velocity deals with the speed at the data flows from sources like **application logs, business processes, networks, and social media sites, sensors, mobile devices,** etc.

1b

A **Data Warehousing** (DW) is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyze business data from heterogeneous sources. The data warehouse is the core of the BI system which is built for data analysis and reporting.

Hadoop :

It is an open-source software program framework for storing information and strolling applications on clusters of commodity hardware. It offers large storage for any sort of data, extensive processing strength, and the potential to deal with actually limitless concurrent duties or jobs.

S.No.	Data Warehouse	Hadoop
1.	In this, we first analyze the data and then further do the processing.	It can process various types of data such as Structured data, unstructured data, or raw data.
2.	It is convenient for storing a small volume of data.	It deals with a large volume of data.
3.	It uses schema-for-write logic to process the data.	It deals with schema-for-read logic to process the data.
4.	It is very less agile as compared to Hadoop.	It is more agile as compared to Data Warehouse.
5.	It is of fixed configuration.	It can be configured or reconfigured, accordingly.
6.	It has high security for storing different data.	Security is a great concern and It is improving and working on it.
7.	It is mainly used by business professionals.	It mainly deals with Data Engineering and Data Science.

2a

Features of Hadoop

1. Open Source:

Hadoop is open-source, which means it is free to use. Since it is an open-source project the source-code is available online for anyone to understand it or make some modifications as per their industry requirement.

2. Highly Scalable Cluster:

Hadoop is a highly scalable model. A large amount of data is divided into multiple inexpensive machines in a cluster which is processed parallelly.the number of these machines or nodes can be increased or decreased as per the enterprise's requirements.

3. Fault Tolerance is Available:

Hadoop uses commodity hardware(inexpensive systems) which can be crashed at any moment. In Hadoop data is replicated on various DataNodes in a Hadoop cluster which ensures the availability of data if somehow any of your systems got crashed.

4. High Availability is Provided:

Fault tolerance provides High Availability in the Hadoop cluster. High Availability means the availability of data on the Hadoop cluster. Due to fault tolerance in case if any of the DataNode goes down the same data can be retrieved from any other node where the data is replicated.

5. Cost-Effective:

Hadoop is open-source and uses cost-effective commodity hardware which provides a cost-efficient model, unlike traditional Relational databases that require expensive hardware and high-end processors to deal with Big Data.

6. Hadoop Provide Flexibility:

Hadoop is designed in such a way that it can deal with any kind of dataset like structured(MySql Data), Semi-Structured(XML, JSON), Un-structured (Images and Videos) very efficiently.

7. Easy to Use:

Hadoop is easy to use since the developers need not worry about any of the processing work since it is managed by the Hadoop itself. Hadoop ecosystem is also very large comes up with lots of tools like Hive, Pig, Spark, HBase, Mahout, etc.

8. Hadoop uses Data Locality:

The concept of Data Locality is used to make Hadoop processing fast. In the data locality concept, the computation logic is moved near data rather than moving the data to the computation logic

9. Provides Faster Data Processing:

Hadoop uses a distributed file system to manage its storage i.e. HDFS(Hadoop Distributed File System). In DFS(Distributed File System) a large size file is broken into small size file blocks then distributed among the Nodes available in a Hadoop

Mild stones in Hadoop :

Milestone 1. You need to decide whether the solution is deployed on-premises or in the cloud?

Milestone 2. You need to decide whether it is vanilla Hadoop or a Hadoop distribution?

Milestone 3. You need to figure out the required size and structure of Hadoop clusters

Milestone 4. Combine all factors of the architecture

2b

RDMS (Relational Database Management System): RDBMS is an information management system, which is based on a data model.In RDBMS tables are used for information storage.

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important 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 into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples.

MapReduce Vs RDBMS

1.	MapReduce suits in an application where the data is written once and read many times like in your Facebook profile you post your photo once and that picture of your seen by your friends many times, whereas RDBMS good for data sets that are continuously updated.
2.	The RDBMS is suits for an application where data size is limited like it's in GBs,whereas MapReduce suits for an application where data size is in Petabytes.
3.	The RDBMS accessed data in interactive and batch mode, whereas MapReduce access the data in batch mode.
4.	The RDBMS schema structure is static, whereas MapReduce schema is dynamic.
5.	The RDBMS suits with structure data sets, whereas MapReduce suits with un-structure data sets.
6.	The RDBMS scaling is nonlinear, whereas MapReduce is linear.

4a

Grid Computing
It is a process architecture that combines different computing resources from multiple locations to achieve desired and common goal.
It distributes workload across multiple systems and allow computers to contribute their individual resources to common goal.
It makes better use of existing resources, address rapid fluctuations in customer demands, improve computational capabilities, provide flexibility, etc.
It mainly focuses on sharing computing resources.
It is of three types i.e., computational grid, data grid, and collaborative grid.
It is used in ATMs, back-end infrastructures, marketing research, etc.
Its main purpose is to integrate usage of computer resources from cooperating partners in form of VO (Virtual Organizations).
Its characteristics include resource coordination, transparent access, dependable access, etc.

Volunteer Computing

When people first hear about Hadoop and MapReduce, they often ask, "How is it different from SETI@home?" SETI, the Search for Extra-Terrestrial Intelligence, runs a project called SETI@home in which volunteers donate CPU time from their otherwise idle computers to analyze radio telescope data for signs of intelligent life outside earth. SETI@home is the most well-known of many *volunteer computing* projects; others include the Great Internet Mersenne Prime Search (to search for large prime numbers) and Folding@home (to understand protein folding and how it relates to disease).

Volunteer computing projects work by breaking the problem they are trying to solve into chunks called *work units*, which are sent to computers around the world to be analyzed. For example, a SETI@home work unit is about 0.35 MB of radio telescope data, and takes hours or days to analyze on a typical home computer. When the analysis is completed, the results are sent back to the server, and the client gets another work unit. As a precaution to combat cheating, each work unit is sent to three different machines and needs at least two results to agree to be accepted.

Although SETI@home may be superficially similar to MapReduce (breaking a problem into independent pieces to be worked on in parallel), there are some significant differences. The SETI@home problem is very CPU-intensive, which makes it suitable for running on hundreds of thousands of computers across the world,[8] since the time to transfer the work unit is dwarfed by the time to run the computation on it. Volunteers are donating CPU cycles, not bandwidth.

MapReduce is designed to run jobs that last minutes or hours on trusted, dedicated hardware running in a single data center with very high aggregate bandwidth interconnects. By contrast, SETI@home runs a perpetual computation on untrusted machines on the Internet with highly variable connection speeds and no data locality.

4b

Types of Big Data Technologies:

Big Data Technology is mainly classified into two types:

1. **Operational Big Data Technologies**
2. **Analytical Big Data Technologies**

Firstly, The Operational Big Data is all about the normal day to day data that we generate. This could be the **Online Transactions, Social Media,** or the data from a **Particular Organisation** etc. You can even consider this to be a kind of Raw Data which is used to feed the **Analytical Big Data Technologies.**

A few examples of **Operational Big Data Technologies** are as follows:

Online ticket bookings, which includes your Rail tickets, Flight tickets, movie tickets etc.

Online shopping which is your Amazon, Flipkart, Walmart, Snap deal and many more.

Data from social media sites like Facebook, Instagram, what's app and a lot more.

The employee details of any Multinational Company.

Analytical Big Data is like the advanced version of Big Data Technologies. It is a little complex than the Operational Big Data. In short, Analytical big data is where the actual performance part comes into the picture and the crucial real-time business decisions are made by analyzing the Operational Big Data.

Few examples of **Analytical Big Data Technologies** are as follows:

* Stock marketing

- Carrying out the Space missions where every single bit of information is crucial.
- Weather forecast information.
- Medical fields where a particular patients health status can be monitored.

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 four major elements of Hadoop i.e. HDFS, MapReduce, YARN, and Hadoop Common. Most of the tools or solutions are used to supplement or support these 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

MapReduce: Programming based Data Processing

Spark: In-Memory data processing

PIG, HIVE: Query based processing of data services

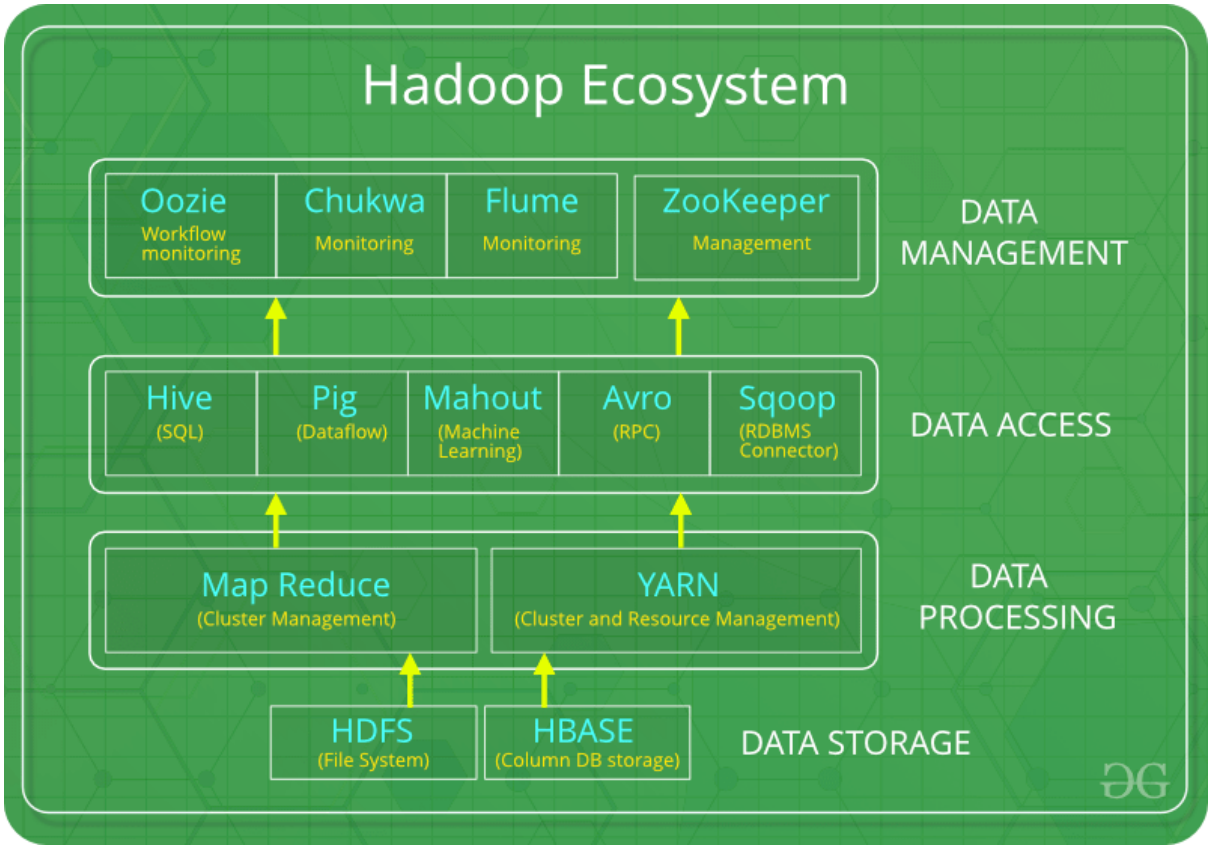
HBase: NoSQL Database

Mahout, Spark MLLib: Machine Learning algorithm libraries

Solar, Lucene: Searching and Indexing

Zookeeper: Managing cluster

Oozie: Job Scheduling



HDFS:

HDFS is the primary or major component of Hadoop ecosystem and is responsible for storing large data sets of structured or unstructured data across various nodes and thereby maintaining the metadata in the form of log files.

HDFS consists of two core components i.e.

Name node

Data Node

YARN:

Yet Another Resource Negotiator, as the name implies, YARN is the one who helps to manage the resources across the clusters. In short, it performs scheduling and resource allocation for the Hadoop System.

Consists of three major components i.e.

Resource Manager

Nodes Manager

Application Manager

MapReduce:

By making the use of distributed and parallel algorithms, MapReduce makes it possible to carry over the processing's logic and helps to write applications which transform big data sets into a manageable one.

MapReduce makes the use of two functions i.e. Map() and Reduce()

PIG:

Pig was basically developed by Yahoo which works on a pig Latin language, which is Query based language similar to SQL.

HIVE:

With the help of SQL methodology and interface, HIVE performs reading and writing of large data sets. However, its query language is called as HQL (Hive Query Language).

Mahout:

Mahout, allows Machine Learnability to a system or application. Machine Learning, as the name suggests helps the system to develop itself based on some patterns, user/environmental interaction or on the basis of algorithms.

Apache Spark:

It's a platform that handles all the process consumptive tasks like batch processing, interactive or iterative real-time processing, graph conversions, and visualization, etc.

Apache HBase:

It's a NoSQL database which supports all kinds of data and thus capable of handling anything of Hadoop Database. It provides capabilities of Google's BigTable, thus able to work on Big Data sets effectively.

---

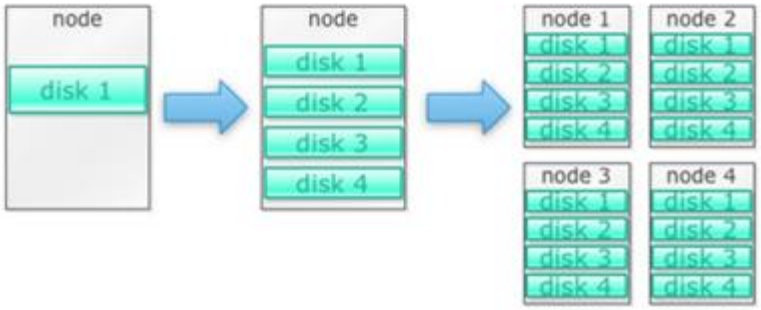
6 Describe the challenges with Big Data and how Hadoop is providing solution.

How Hadoop Solves the Big Data Problem

1 Hadoop is built to run on a cluster of machines

Let's say that we need to store lots of photos. We will start with a single disk. When we exceed a single disk, we may use a few disks stacked on a machine. When we max out all the disks on a single machine, we need to get a bunch of machines, each with a bunch of disks.

This is exactly how Hadoop is built. Hadoop is designed to run on a cluster of machines from the get go.



Hadoop clusters scale horizontally

More storage and compute power can be achieved by adding more nodes to a Hadoop cluster. This eliminates the need to buy more and more powerful and expensive hardware.

Hadoop can handle unstructured/semi-structured data

Hadoop doesn't enforce a schema on the data it stores. It can handle arbitrary text and binary data. So Hadoop can digest any unstructured data easily.

Hadoop clusters provides storage and computing

We saw how having separate storage and processing clusters is not the best fit for big data. Hadoop clusters, however, provide storage and distributed computing all in one.

The Business Case for Hadoop

Hadoop provides storage for big data at reasonable cost

Storing big data using traditional storage can be expensive. Hadoop is built around commodity hardware, so it can provide fairly large storage for a reasonable cost. Hadoop has been used in the field at petabyte scale.

Hadoop allows for the capture of new or more data

Sometimes organizations don't capture a type of data because it was too cost prohibitive to store it. Since Hadoop provides storage at reasonable cost, this type of data can be captured and stored.

With Hadoop, you can store data longer

To manage the volume of data stored, companies periodically purge older data. For example, only logs for the last three months could be stored,

while older logs were deleted. With Hadoop it is possible to store the historical data longer.

Hadoop provides scalable analytics

There is no point in storing all this data if we can't analyze them. Hadoop not only provides distributed storage, but also distributed processing as well, which means we can crunch a large volume of data in parallel.