UPES

UNIVERSITY OF TOMORROW

BIG DATA

MID - SEM

SUBMITTED BY:

AYUSH KUMAR JHA
500086400
B.C.A-IOT (2020-2023)

SUBMITTED TO:

SHELLY MA'AM

ANSWER-1:

1. Lack of knowledge Professionals

To run these modern technologies and large Data tools, companies need skilled data professionals. These professionals will include data scientists, data analysts, and data engineers to work with the tools and make sense of giant data sets. One of the Big Data Challenges that any Company face is a drag of lack of massive Data professionals. This is often because data handling tools have evolved rapidly, but in most cases, the professionals haven't. Actionable steps got to be taken to bridge this gap.

2. Lack of proper understanding of Massive Data

Companies fail in their Big Data initiatives, all thanks to insufficient understanding. Employees might not know what data is, its storage, processing, importance, and sources. Data professionals may know what's happening, but others might not have a transparent picture. For example, if employees don't understand the importance of knowledge storage, they could not keep the backup of sensitive data. They could not use databases properly for storage. As a result, when this important data is required, it can't be retrieved easily.

3. Data Growth Issues

One of the foremost pressing challenges of massive Data is storing these huge sets of knowledge properly. the quantity of knowledge being stored in data centers and databases of companies is increasing rapidly. As these data sets grow exponentially with time, it gets challenging to handle. Most of the info is unstructured and comes from documents, videos, audio, text files, and other sources. This suggests that you cannot find them in the database.

4. Confusion while Big Data Tool selection

Companies often get confused while selecting the simplest tool for giant Data analysis and storage. Is HBase or Cassandra the simplest technology for data storage? Is Hadoop MapReduce ok, or will Spark be a far better option for data analytics and storage? These questions bother companies, and sometimes they're unable to seek out the answers. They find themselves making poor decisions and selecting inappropriate technology. As a result, money, time, efforts, and work hours are wasted.

5. Integrating Data from a Spread of Sources

Data in a corporation comes from various sources, like social media pages, ERP applications, customer logs, financial reports, e-mails, presentations, and reports created by employees. Combining all this data to organize reports may be a challenging task. This is a neighborhood often neglected by firms. Data integration is crucial for analysis, reporting, and business intelligence, so it's perfect.

6. Securing Data

Securing these huge sets of knowledge is one of the daunting challenges of massive Data. Often companies are so busy in understanding, storing, and analyzing their data sets that they push data security for later stages. This is often not a sensible move as unprotected data repositories can become breeding grounds for malicious hackers. Companies can lose up to $3.7 million for a stolen record or a knowledge breach.

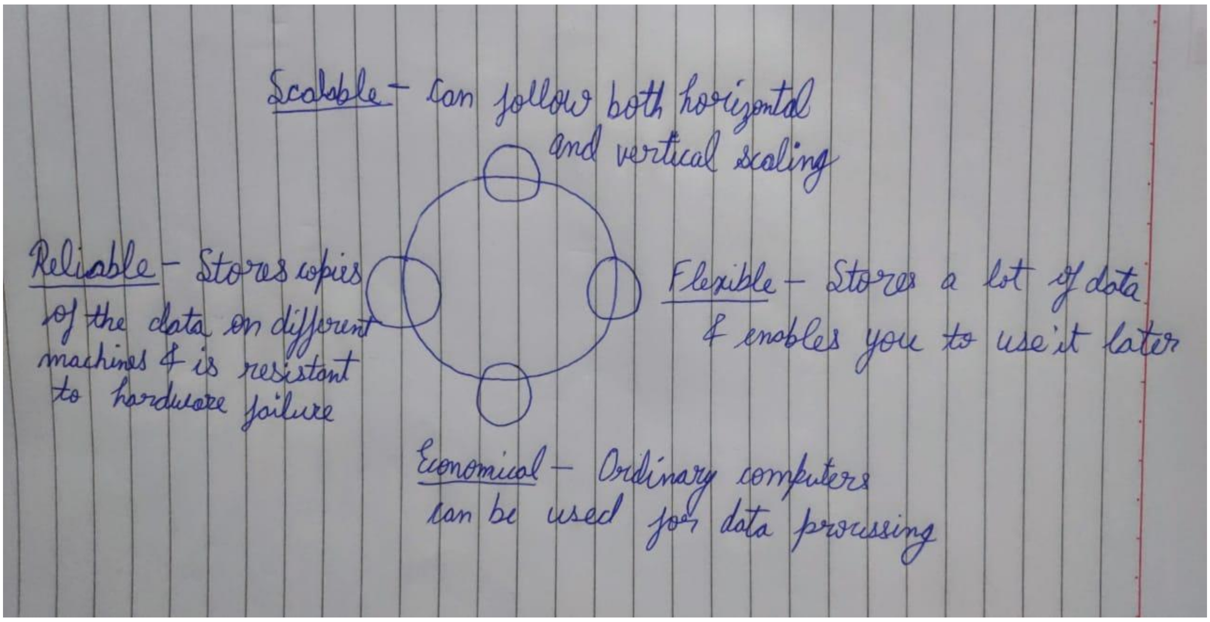
ANSWER-2:

This is the basic difference :-
Traditional Database Systems
Data is stored in a central location and sent to the processor at run time.
Traditional Database Systems cannot be used to process and store a large amount of data (big data).
Traditional RDBMS is used to manage only structured and semi-structured data. It cannot be used to manage unstructured data.

Where as Hadoop is
In Hadoop, the program goes to the data. It initially distributes the data to multiple systems and later runs the computation wherever the data is located.
Hadoop works better when the data size is big. It can process and store a large amount of data easily and effectively.
Hadoop has the ability to process and store a variety of data, whether it is structured or unstructured.

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. The problem with traditional Relational databases is that storing the massive volume of data is not cost-effective, so the company's started to remove the Raw data. which may not result in the correct scenario of their business. Means Hadoop provides us 2 main benefits with the cost one is it's open-source means free to use and the other is that it uses commodity hardware which is also inexpensive.

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. This means it can easily process any kind of data independent of its structure which makes it highly flexible. It is very much useful for enterprises as they can process large datasets easily, so the businesses can use Hadoop to analyse valuable insights of data from sources like social media, email, etc. With this flexibility, Hadoop can be used with log processing, Data Warehousing, Fraud detection, etc.



Scalable - Can follow both horizontal and vertical scaling

Reliable - Stores copies of the data on different machines & is resistant to hardware failure

Flexible - Stores a lot of data & enables you to use it later

Economical - Ordinary computers can be used for data processing

ANSWER-3:

Components of Hadoop Ecosystem

HDFS is a storage layer of Hadoop suitable for distributed storage and processing.

- It provides file permissions, authentication, and streaming access to file system data.

HBase is a NoSQL database or non-relational database that stores data in HDFS.

- It provides support to high volume of data and high throughput.
- It is used when you need random, real-time read/write access to your big data.

Sqoop is a tool designed to transfer data between Hadoop and relational database servers.

- It is used to import data from relational databases such as Oracle and MySQL to HDFS and export data from HDFS to relational databases.

Flume is a distributed service for ingesting streaming data suited for event data from multiple systems.

- It has a simple and flexible architecture based on streaming data flows.
- It is robust and fault tolerant and has tunable reliability mechanisms.
- It uses a simple extensible data model that allows for online analytic application.

Spark is an open-source cluster computing framework that supports Machine learning, Business intelligence, Streaming, and Batch processing. Spark solves similar problems as Hadoop MapReduce does but has a fast in-memory approach and a clean functional style API.

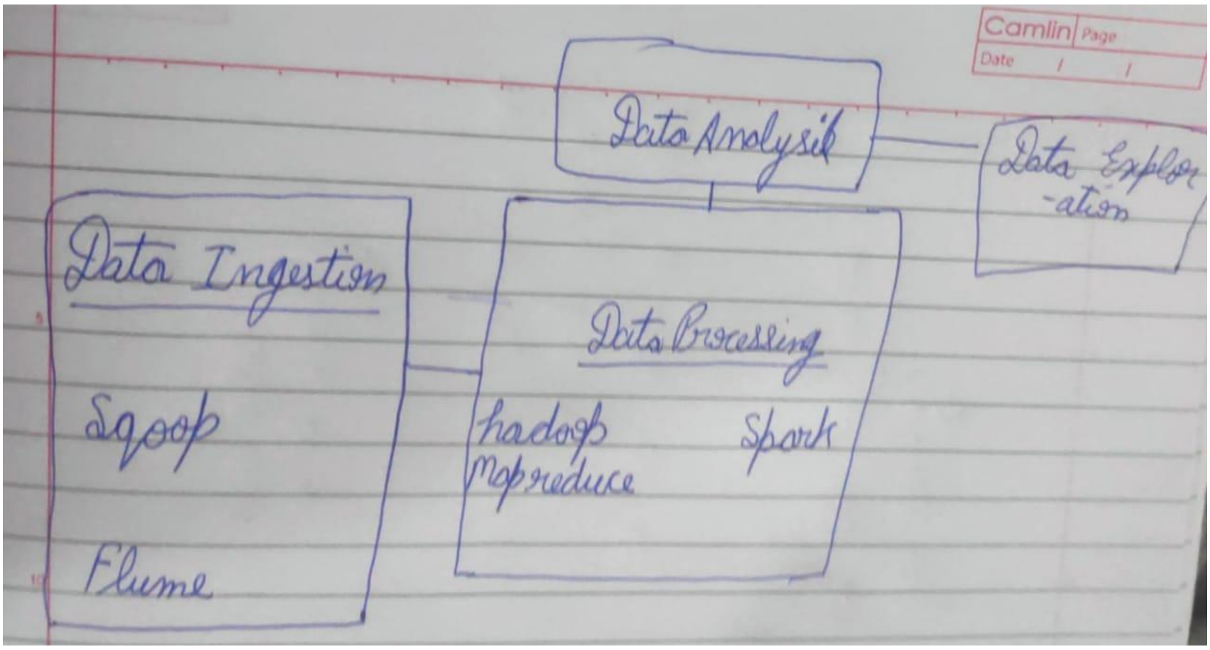
Spark is an open-source cluster computing framework that supports Machine learning, Business intelligence, Streaming, and Batch processing. Spark solves similar problems as Hadoop MapReduce does but has a fast in-memory approach and a clean functional style API.

Once the data is processed, it is analyzed using an open-source high-level dataflow system called Pig. • Pig converts its scripts to Map and Reduce code to reduce the effort of writing complex map-reduce programs.

- Ad-hoc queries like Filter and Join, which are difficult to perform in MapReduce, can be easily done using Pig.

PIG: Once the data is processed, it is analyzed using an open-source high-level dataflow system called Pig. Pig converts its scripts to Map and Reduce code to reduce the effort of writing complex map-reduce programs. Ad-hoc queries like Filter and Join, which are difficult to perform in MapReduce, can be easily done using Pig.

- IMPALA It is an open-source high performance SQL engine that runs on the Hadoop cluster. It is ideal for interactive analysis and has very low latency, which can be measured in milliseconds. Impala supports a dialect of SQL, so data in HDFS is modeled as a database table.



ANSWER-4:

In the traditional system, storing and retrieving volumes of data had three major issues:
Speed: Search and analysis is time-consuming
Reliability: Fetching data is difficult
Cost: 10,000 to $14,000 per terabyte

HDFS resolves all major issues of the traditional file system.
Speed: Search and analysis is time-consuming
Cost: 10,000 to $14,000 per terabyte
Cost: 10,000 to $14,000 per terabyte

HDFS is a distributed file system that provides access to data across Hadoop clusters. It manages and supports analysis of very large volumes of Big Data.
HDFS has high fault-tolerance
HDFS has high throughput
HDFS is economical

HDFS Work:-
A patron gifts his books to a college library.
The librarian decides to arrange the books on a small rack.
The librarian then distributes multiple copies of each book on other racks based on the category.

ANSWER-5:

1. Simplify Human behaviour and element

Understanding human behavior, nature, and habits is not an easy task. It consumes a lot of time to observe, research, and the list goes on. But Artificial Intelligence and Big Data make it work in minutes for us, as they give you insight in a short period. Even you can research better by embracing AI and Big Data technology.

2. Exclude spam

No one likes when an unnecessary post or message comes on their feed on Instagram. Everyone wants to see what they like and follow; here, Artificial Intelligence comes to the rescue for such actions. Because technology helps social media to identify fake messages and get rid of fake accounts. It enables blocking spam messages and inappropriate accounts. Instagram runs in several languages such as English, Spanish, Hindi, Arabic, and many more so, it became easy to detect spam.

3. Emphasize Advertising

As we mentioned earlier, Instagram is not just a platform for photos and video; but gradually turning into a business platform. However, it is not merely that; it blends photos, videos, likes, comments, shares, business, influence, shopping, and many others. Therefore, one platform has multiple usages, and it helps to promote. Targeting Advertising has become sleek on Instagram by users showing interest in products and brands.

4. Explore and Search Function

One of the most used features of Instagram is exploration and search. People search for what they like, accounts, and learn about more new similar interest accounts on exploring. With this function, you will get to know about trending tags, information, content, and more. Artificial Intelligence and Big Data gather all data across the world and list out the most used data and activity known as Instagram trends.

6. Personalization and Ranking System

Social media users show what they like and according to their preference and choice. After understanding customer behavior and culture, Instagram uses Artificial Intelligence to make feed personalize for them. The users will be shown their interest or possible interest post and stories. It will help both users and the platform to connect, engage, and context more.