ASSIGNMENT-1

HADOOP CS391

1.Explain Hadoop and discuss its history and issue with the traditional large-scale system.

ANS:-Big data refers to the large, diverse sets of information that grow at ever-increasing rates. It encompasses the volume of information, the velocity or speed at which it is created and collected, and the variety or scope of the data points being covered (known as the "three v's" of big data). Big data often comes from data mining and arrives in multiple formats.

SCOPE AND CHARACTERSTICS:-

Big data is a great quantity of diverse information that arrives in increasing volumes and with ever-higher velocity.

Big data can be structured (often numeric, easily formatted and stored) or unstructured (more free-form, less quantifiable).

Nearly every department in a company can utilize findings from big data analysis, but handling its clutter and noise can pose problems.

Big data can be collected from publicly shared comments on social networks and websites, voluntarily gathered from personal electronics and apps, through questionnaires, product purchases, and electronic check-ins.

Big data is most often stored in computer databases and is analyzed using software specifically designed to handle large, complex data sets.

---

2. Explain Hadoop and discuss its history and issue with the traditional large-scale system.

ANS:- Hadoop is an open source framework from Apache and is used to store process and analyze data which are very huge in volume. Hadoop is written in Java and is not OLAP (online analytical processing). It is used for batch/offline processing.It is being used by Facebook, Yahoo, Google, Twitter, LinkedIn and many more. Moreover it can be scaled up just by adding nodes in the cluster.

Modules of Hadoop

HDFS: Hadoop Distributed File System. Google published its paper GFS and on the basis of that HDFS was developed. It states that the files will be broken into blocks and stored in nodes over the distributed architecture.

Yarn: Yet another Resource Negotiator is used for job scheduling and manage the cluster.

Map Reduce: This is a framework which helps Java programs to do the parallel computation on data using key value pair. The Map task takes input data and converts it into a data set which can be computed in Key value pair. The output of Map task is consumed by reduce task and then the out of reducer gives the desired result.

Hadoop Common: These Java libraries are used to start Hadoop and are used by other Hadoop modules.

History of Hadoop

Let's focus on the history of Hadoop in the following steps: -

In 2002, Doug Cutting and Mike Cafarella started to work on a project, Apache Nutch. It is an open source web crawler software project.

While working on Apache Nutch, they were dealing with big data. To store that data they have to spend a lot of costs which becomes the consequence of that project. This problem becomes one of the important reason for the emergence of Hadoop.

In 2003, Google introduced a file system known as GFS (Google file system). It is a proprietary distributed file system developed to provide efficient access to data.

In 2004, Google released a white paper on Map Reduce. This technique simplifies the data processing on large clusters.

In 2005, Doug Cutting and Mike Cafarella introduced a new file system known as NDFS (Nutch Distributed File System). This file system also includes Map reduce.

In 2006, Doug Cutting quit Google and joined Yahoo. On the basis of the Nutch project, Dough Cutting introduces a new project Hadoop with a file system known as HDFS (Hadoop Distributed File System). Hadoop first version 0.1.0 released in this year.

Doug Cutting gave named his project Hadoop after his son's toy elephant.

In 2007, Yahoo runs two clusters of 1000 machines.

In 2008, Hadoop became the fastest system to sort 1 terabyte of data on a 900 node cluster within 209 seconds.

In 2013, Hadoop 2.2 was released.

In 2017, Hadoop 3.0 was released.

ISSUE WITH THE TRADITIONAL LARGE-SCALE SYSTEM

The reason traditional systems have a problem with big data is that they were not designed for it. Problem—Schema-On-Write: Traditional systems are schema-on-write. Schema-on-write requires the data to be validated when it is written. This means that a lot of work must be done before new data sources can be analyzed.

What was the main challenge in traditional database system?

High costs and failure rates

This is not just because of the complex architecture or technical challenges, but because these projects often fail to meet user requirements and needs. The same challenges exist when a data warehouse needs to be updated or changed to meet new reporting requirements or data needs

3.Compare:

a. Hadoop vs RDBMS

S.No.	Hadoop	RDBMS
1.	An open-source software used for storing data and running applications or processes concurrently.	Traditional row-column based databases, basically used for data storage, manipulation and retrieval.
2.	In this both structured and unstructured data is processed.	In this structured data is mostly processed.
3.	It is best suited for BIG data.	It is best suited for OLTP environment.
4.	It is highly scalable.	It is less scalable than Hadoop.
5.	Data normalization is not required in Hadoop.	Data normalization is required in RDBMS.
6.	It stores huge volume of data.	It stores transformed and aggregated data.

7.	It has some latency in response.	It has no latency in response.

b. Hadoop vs Big Data

Features	Hadoop	Big Data
Definition	Hadoop is a framework to handle and process this large volume of Big data.	Big Data refers to a large volume of both structured and unstructured data.
Significance	It is a tool that makes big data more meaningful by processing the data.	Big Data has no significance until it is processed and utilized to generate revenue.
Storage	Apache Hadoop HDFS is capable of storing big data.	It is very difficult to store big data because it comes

		in structured and unstructured form
Accessibility	Hadoop framework lets you access and process the data very fast when compared to other tools.	When it comes to accessing the big data, it is very difficult.

---

4. Explain types of Data with their source of origin.

ANS:- There are three types of data :-

1. Structured Data:- The term structured data refers to data that resides in a fixed field within a file or record. Structured data is typically stored in a relational database (RDBMS). It can consist of numbers and text, and sourcing can happen automatically or manually, as long as it's within an RDBMS

structure. It depends on the creation of a data model, defining what types of data to include and how to store and process it.

The programming language used for structured data is SQL (Structured Query Language). Developed by IBM in 1974, SQL handles relational databases. Typical examples of structured data are names, addresses, credit card numbers, geolocation, and so on.

2. Unstructured Data:- Unstructured data is more or less all the data that is not structured. Even though unstructured data may have a native, internal structure, it's not structured in a predefined way. There is no data model; the data is stored in its native format. Typical examples of unstructured data are rich media, text, social media activity, surveillance imagery, and so on.

3. Semi-structured Data:- Semistructured data is a third category that falls somewhere between the other two. It's a type of structured data that does not fit into the formal structure of a relational database. But while not matching the description of structured data entirely, it still employs tagging systems or other identifiable markers, separating different elements and enabling search. Sometimes, this is referred to as data with a self-describing structure.A typical example of semistructured data is smartphone photos. Every photo taken with a smartphone contains unstructured image content as well as the tagged time, location, and other identifiable (and structured) information. Semi-structured data formats include JSON, CSV, and XML file types.