1.What is Bigdata?

Big data is technology which collect the massive amount of data among of cluster. Big data is a term that describes large, hard-to-manage volumes of data.

2. What is a cluster?

Cluster is the group of computer or server interconnected to perform a particular task. Which enable high availability and, in some cases, Load balancing and parallel processing.

3.What are frameworks & libraries?

Framework is a platform for developing software application and it is routine taste in generic module can be reusable. **Library**: Library is a piece of code which can be reusable.

4.what is Hadoop?

Hadoop is also one of the frameworks, used to store the data in the cluster.

It stores data in only file format

5.what are open source and distributions?

An Open-source database allows users to create a system based on their unique requirement and business needs. It is free and can also be shared.

Distribution means to spread the product throughout the marketplace such that many people can buy it. It also means the process of getting goods to consumers.

6. What are the core components of Big Data and Its purpose?

The core components of bigdata are 1. Name node

2.data node

3.standby node

4.secondary name node

5.standby name node

7.what are daemons?

Daemons mean **Process**. Hadoop Daemons are a set of processes that run on Hadoop in the background

Name node (master node), data node (subsidiary node)

8.advantages of Hadoop?

Cost efficient.

Scalability.

Flexibility

Speed

Fault Tolerance

9.what is scaling and its types?

Based on the requirement increasing the capacity for the nodes or clusters is called as scaling there are two types of scaling

1.Horizontal scaling

2.vertical scaling

10.differences between RDBMS and Hadoop

RDBMS is an information management system, which is based on a data model. It’s ensuring ACID properties. Suitable for OLTP. Less scalable than Hadoop. No latency.

Hadoop is an open- source software framework used for storing data and running application on a group of commodity hardware. Suitable for big data. Highly scalable. Some latency

11. what are 4v’s of bigdata?

1.value-value of the data we have and its details

2.variety-either structured data or semi structured or nonstructured data

3.volume-the total quantity of data which we have

4.velocity-the speed with which computation happens

12.types of files and each data?

1.csv (comma separated values)

2.text files

3. Json files

13.list some Hadoop efficient scenarios?

When there is massive amount of data Hadoop helps clusters to save data in a manner that it cannot be leaked or misused the clusters have data in a randomized manner

Hadoop becomes inefficient when

1.Low latency or near real-time data access

2. if you have many small files to be processes.

3. Multiple writes scenarios or scenarios requiring arbitrary writes or writes between the files.

14.what is name node and its purpose?

This is a node which consists of details of the data the cluster nodes are having it has naming inscribed where the data is sealed

15.what is data node and its purpose?

Data node is a node in a cluster there can be many data nodes in a cluster this consists of data which is so encrypted so there won’t be any data leaks

16.what is secondary name node?

If one name node fails, the data don’t know where to go and clusters don’t know where the particular data points are there what’s why Hadoop comes up with this backup secondary name node if one fails other one activates and gets the job done

17. what is standby name node and its purpose?

The Standby Name Node maintains enough state to provide a backup for some instance. For the Standby node to keep its state synchronized with the Active node, both nodes communicate through a group of separate daemons

18.what is replication factor?

The Replication Factor means the number of nodes where data (rows and partitions) are replicated. An RF of one means there is only one copy of a row in a cluster

19.block size and its default size?

Block is data container which consists of massive amount of data and there is some default size for each block the default size of block is 128mb

20.what is data node and name node communication?

Data node is the data constituent which transfers signals to the master or name node that the node is active so that it won’t activate backup nodes whenever the data node doesn’t communicate with name node within 3 seconds the name node understands that the data node failed

21.what is zookeeper and its uses?

Zookeepers have the information about configuration, naming. It is a centralized unit and using this information. Zookeeper maintains Hadoop as a Single Unit and is responsible for synchronizing Hadoop tasks

22.what is NAS and other name of NAS

Network-attached storage (NAS) is a file-level computer data storage server. NAS provides data access. HDFS distribute data across all the blocks in a Hadoop cluster.

23.what is edge /gateway node?

An edge node is a cluster that acts as an end user gateway for communication with other nodes in cluster computing. Edge nodes are also sometimes called gateway nodes or edge communication nodes.

24.what is data centers?

Data centers are the places where the large amounts of data connected to servers, and these helps us in computational efficiency in India we have data center in Mumbai

25.what is hive and its purpose?

Hive is a SQL data warehouse framework for big data. It is used to view text file as a table format. It is just like SQL where we can work on data

26.what is Sqoop?

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

Used to import data from relational databases such as MySQL, oracle to Hadoop HDFS, and export from Hadoop file system to relational database.

27.what is pig & HBase AND its purpose?

HBase: A scalable, distributed database that supports structured data storage for large tables.

Pig: A high-level data-flow language and execution framework for parallel computation

28.difference between batch and real time data process?

Batch is the quantity (certain amount) of data sent for computation

Real time data process means when the data is received form user the whole data is sent for processing or computation

29.what is MR?

MR (map reduce) the term "MapReduce" refers to two separate and distinct tasks that Hadoop programs perform.

it is a programming paradigm that enables massive scalability across hundreds or thousands of servers or clusters in Hadoop.

30.Explain MR daemons and its function?

Mr does consists of mappers and reducers

Name node data node are daemons

31.what is YARN and its features?

YARN, it is part of the bigdata architecture the Hadoop operating system, enables you to manage resources and schedule jobs in Hadoop. YARN allows you to use various data processing engines for batch, interactive processing of data stored in HDFS

32. list YARN daemons?

1. Resource Manager

2.Node manager

3.application master

33.What is Resource Manager and its functions?

There is one resource manager for one cluster

1.Controls application startup

2.Schedule resources on the slave nodes

3.Resource manager task is to collect resource information from the load manager and after getting the data and Resource manager releases the container to perform the task.

34.What is Node Manager and its functions

Node manager manages all the nodes in the cluster

1. There will be one node manager forone slave node

2.Starts all processes for a running application

Manges resources on the slave nodes

Node manager sends the resource status to resource manager and Node manager monitor the container resource utilization.

35.what is application master?

Application master establish each node and monitor the container task status and again reports to resource manager.

Once task completed the application master will vanished

36.what is container and its functions?

A container is like a storage space, including all its dependencies, libraries and other binaries, and the configuration files needed to run it, bundled into a single package that can be moved, in total, from one computing environment to another.

37.Explain step by step process of daemons functions?

1.resource manager: it controls the startup of applications and schedules task for slave nodes

2.node manager: there will be one node manager for each slave node which are subsidiaries for a resource node

3.application master: Application master establish each node and monitor the container task status and again reports to resource manager.

38. Hadoop architecture diagram?

Master node(resource manager, name node)

Master node

39.MR dataflow diagram?

output

Input split1

Record reader

Mapper

partition

Input split2

Record reader

Mapper

partition

Input file

Shuffle and sort happen here

40.yarn architecture?

Nodes

mappers

reducers

Client node

Nodes

Mappers

reducers

Resource manager

Client node

Nodes

Mappers

reducers

Client node