1. What is big data?

Big Data is a technology that processes data across the cluster. The process involves storing and retrieving.

1. What is Cluster?

General Definition: It is a group of similar objects

It is a group of computers or servers that are interconnected to perform a particular task.

Advantages of Clusters: High Availability

Load Balancing

Parallel Processing

1. Framework and Libraries?

Framework: It is a platform for developing software application and its routine task in generic module can be reusable.

Libraries: It is code or function that is readily available and reusable.

1. What is Hadoop?

It is an open-source framework that is used to store data but not process or analyze data.

1. What is Open source and Distribution and example for Distributions

Open-Source: It is a software or code that can accessed and modified by all.

Distribution: It a version of the of the open-source software for which support is provided. Hadoop has 2 distributors Cloudera and Hortonworks.

1. What are the core components of Big Data and Its purpose?
2. What are daemons and its purpose?

Daemons are background running processes or background running servers. Their main process is to perform processes in the background.

1. Hadoop advantages?

Robust, Scalable, Reliable, Flexible.

1. What are scaling and its types?

Scaling is used to increase or decrease the size of the clusters.

Two types of Scaling: Horizontal and Vertical Scaling

Horizontal: Adding computers or servers to increase the clusters

It has two types: Scale Up and Scale Down

Scale Up: Adding computers to increase the size of the clusters.

Its an easy process

Does not require down time

Scale Down: It is the opposite of scale up it is done to reduce the size of the clusters.

Vertical: Adding hardware components to the current computers or servers.

1. List Different between RDBMS Vs Big Data

|  |  |
| --- | --- |
| RDBMS | Big Data |
| Handles data till GB (Giga Byte) | Handles Petabyte of Data |
| Mostly Proprietary | Open-Source Framework |
| Designed for Client Server Architecture | Designed to support distributed architecture |
| High Usage requires High end Server | Runs on Commodity Hardware |
| Needs Structured Data | Very good support for unstructured data |
| Stable Product | Still Evolving |

1. Explains 4V's of Big Data?

Volume: Size of the data

Variety: Types of data

Velocity: speed of transmission or generation of data

Veracity: Untrusted/Uncleaned data

1. Types of Files and each data / file producers’ details
2. List some Hadoop efficient and inefficient scenarios

Inefficient:

* Low latency or near real time data access
* When there re large number of small files to be processed
* Multiple write scenarios

Efficient:

* Analytics
* Data Retention
* Log file Processing
* Analysis of text, image, video content.

1. What is Name node, and its purpose

Name node is the master node that contains the meta data info of the data stored in the data node in HDFS. It controls the data node.

1. What is Data node, and its purpose

Data node is the slave node that contains the actual data in HDFS.

1. What is Secondary name node and its purpose

Secondary Name node is used for bookkeeping purpose for the Name node. It prevents overloading of the Name node.

Secondary Name node on receiving the data creates an edit log file and file system image and maps them together to create a FS image Checkpoint. This checkpoint is transmitted to Name node.

It sends the checkpoint to both Name Node and Standby node.

1. What is Standby name node and its purpose

When the name node fails Standby node acts as a Name node to prevent data loss. When name node comes alive, and control is transferred to the Name node. Secondary Name node sends the checkpoint to both Name Node and Standby node.

1. What is Replication factor (Default?)

To prevent loss of data in case of data node failure each block received is replicated and stored in each node. Generally, the replication factor is 3. So, for storing 1Gb of data 3Gb clusters are required.

1. What is Block and default block size?

Blocks are the containers that divide the data received equally to the clusters. In Hadoop 1 the default size is 64 Mb and in Hadoop 2 its 128 Mb, this size can be reduced by admin.

1. Explain Data node and Name node communication. (Each Seconds details)

Every 3 seconds the data node sends heartbeat signal to Name node to signal its functioning. Every 30 seconds the data node sends reports about the blocks to the Name node. If the Data node is dead the Name node waits for 180 seconds resilient time and considers the node as a dead node. Till the resilient time the data to the node is stored in queue to prevent data loss and once considered dead the data in queue is distributed equally to the rest of the nodes. Once the data node comes back alive, a refresh script is run to distribute the data equally to all the data nodes.

1. What is Zookeeper and its use?
2. What is NAS the other name for NAS?

NAS stands for network access storage. It is also known as Local the client uploads the data to the Local from which it is sent to HAAS.

1. What is Edge / Gateway node?

It is node containing the IP address so that Hadoop can be accessed.

1. What is Commodity Hardware & Data Centers?
2. What is Hive and its purpose

Hive is a SQL or Datawarehouse Framework for big data which is used to view flat file in table format.

1. What is Sqoop and its purpose
2. What is Pig & HBase and its purpose
3. What is the diff between Batch and Real time Data Process?
4. What is MR?

Map Reduce is an algorithmic framework is used to process(send) data into Hadoop.

1. Explain MR daemons and its functions (Hadoop 1.x)

There are two daemons for Map Reduce: Job Tracker and Task Tracker.

Job Tracker is the master; it controls the resource and the job.

Task Tracker is the one that does the actual task.

1. What is YARN and its features (Hadoop 2.x)

YARN stands for Yet Another Resource Negotiator. It is used for resource management and job scheduling.

1. List YARN daemons?

There are 4 daemons: Resource Manager

Node Manager

Application Master

Job History Server

1. What is Resource Manager and its functions

The resource manager is the master in YARN it collects the resource from the Node manager and sends the container to perform the task.

1. What is Node manager and its functions

Node manager sends the resource to the Resource Manager and monitors the container resource utilization. If the container resource exceeds it destroys the container.

1. What is Application Master and its functions

It is a short-term daemon that is created when the task is received. It records the container task status and reports it to the resource manager. Once task is completed it vanishes.

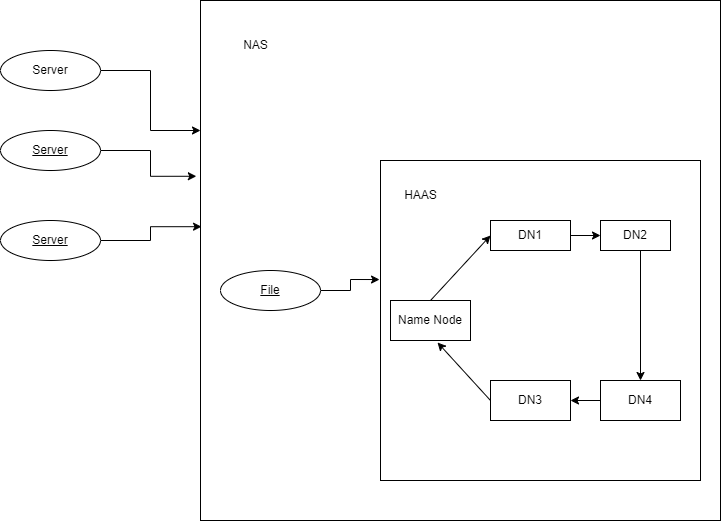
1. What is Container and its functions.

It receives the resource and performs the task.

1. Explain step by step process of all the above daemons’ functions

When the client sends the task to the resource manager, it gets the resource from the Node manager and sends the container to perform the task. It is a short-term daemon that is created when the task is received. It records the container task status and reports it to the resource manager. Node manager sends the resource to the Resource Manager and monitors the container resource utilization. If the container resource exceeds it destroys the container.

1. Hadoop Architecture Diagram



1. MR Data flow Diagram

Graphical user interface, diagram, Teams

Description automatically generated

1. Yarn Demons Diagram

Diagram

Description automatically generated