SINDHU V
211720104141
3rd yr CSE-C

1)what are HDFS and YARN

Hadoop Distributed File System (HDFS) is the primary data storage system used by Hadoop applications. HDFS employs a NameNode and DataNode architecture to implement a distributed file system that provides high-performance access to data across highly scalable Hadoop clusters

YARN (Yet Another Resource Negotiator) – YARN is the program execution system that enhanced MapReduce (MR). YARN is a more generalized program queueing, scheduling, and execution management system than the old MR system which YARN subsumed. Much of the execution on a Hadoop v. 2 system is still of the MR paradigm, such as Hive (SQL on Hadoop) processing, and YARN manages that in its paradigm.

2)what are the various hadoop daemons and their roles in a hadoop

cluster? Apache Hadoop consists of the following Daemons:

NameNode

DataNode

Secondary Name Node

Resource Manager

Node Manager

Roles:

1. NameNode

NameNode works on the Master System. The primary purpose of Namenode is to manage all the MetaData. Metadata is the list of files stored in HDFS(Hadoop Distributed File System).

2. DataNode

DataNode works on the Slave system. The NameNode always instructs DataNode for storing the Data.

3. Secondary NameNode

Secondary NameNode is used for taking the hourly backup of the data. In case the Hadoop cluster fails, or crashes, the secondary Namenode will take the hourly backup or checkpoints of that data and store this data into a file name fsimage

4. Resource Manager

Resource Manager is also known as the Global Master Daemon that works on the Master System

5. Node Manager

The Node Manager works on the Slaves System that manages the memory resource within the Node and Memory Disk

SINDHU V
211720104141
3rd yr CSE-C

3)why does one remove or add nodes in a hadoop cluster frequently?
In a Hadoop cluster a Manager node will be deployed on a reliable hardware with high configurations, the Slave node's will be deployed on commodity hardware. So chance's of data node crashing is more . So more frequently you will see admin's remove and add new data node's in a cluster.

4)what happens when two clients try to access the same file in the HDFS?

HDFS provides support only for exclusive writes so when one client is already writing the file, the other client cannot open the file in write mode. When the client requests the NameNode to open the file for writing, NameNode provides lease to the client for writing to the file. So, if another client requests for lease on the same it will be rejected.

5)how does namenode tackle datanode failures?

Data blocks on the failed Datanode are replicated on other Datanodes based on the specified replication factor in hdfs-site. Xml file. Once the failed datanodes comes back the Name node will manage the replication factor again. This is how Namenode handles the failure of data node.

6)what will you do when namenode is down?

When the NameNode goes down, the file system goes offline. There is an optional SecondaryNameNode that can be hosted on a separate machine. It only creates checkpoints of the namespace by merging the edits file into the fsimage file and does not provide any real redundancy.

7)how is HDFS fault tolerant?

Fault Tolerance in HDFS. Fault tolerance in Hadoop HDFS refers to the working strength of a system in unfavorable conditions and how that system can handle such a situation. HDFS is highly fault tolerant. Before Hadoop 3, it handles faults by the process of replica creation.

8)why do we use hdfs for application having large data sets and not when there are lot of small files?

HDFS is more efficient for a large number of data sets, maintained in a single file as compared to the small chunks of data stored in multiple files. As the NameNode performs storage of metadata for the file system in RAM, the amount of memory limits the number of files in HDFS file system. In simple words, more files will generate more metadata, that will, in turn, require more memory (RAM). It is recommended that metadata of a block, file, or directory should take 150 bytes

9)how do you define "block" in HDFS?what is the default block size in hadoop 1 and in hadoop2 ?can it be changed?

The Hadoop Distributed File System (HDFS) stores files in block-sized chunks called data blocks. These blocks are then stored as independent units and are restricted to 128 MB blocks by default. However, they can be adjusted by the user according to their requirements. Users can adjust block size through the dfs.