211720118030

Hadoop Architecture and HDFS

Homework 2

Q1. What are HDFS and YARN?

HDFS stands for Hadoop Distributed File System. HDFS operates as a distributed file system designed to run on commodity hardware.

YARN allows the data stored in HDFS to be processed and run by various data processing engines such as batch processing, stream processing, interactive processing, graph processing and many more

Q2. What are the various Hadoop daemons and their roles in a Hadoop cluster?

Various Hadoop daemons:

>Name Node

>Secondary Name Node

>Data Node

Roles:

>Hadoop is a framework written in Java, so all these processes are Java Processes

>Node Manager works on this System which can manage the memory resource within the node and the memory disk

Q3. Why does one remove or add nodes in a Hadoop cluster frequently?

Hadoop cluster is a manager node will be deployed on a reliable hardware with high configurations, the Slave node's will be deployed on commodity hardware and chance of data node crashing is more. So more frequently we'll see admin remove and add new data node in a cluster

Q4. What happens when two clients try to access the same file in the HDFS?

Two clients can't write into HDFS file at the same time. When a client is granted a permission to write data on data node block, the block gets locked till the completion of a write operation. If some another client request to write on the same block of the same file then it is not permitted to do so.

Q5. How does Name node tackle Data node failures?

Data blocks on the failed Data node are replicated on other Data nodes based on the specified replication factor in HDFS site xml file. Once the failed data nodes come back the Name node will manage the replication factor again. This is how Name node handles the failure of data node.

Q6. What will you do when Name node is down?

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

Q7. How is HDFS fault tolerant?

The HDFS is highly fault tolerant so if any machine fails, the other machine containing the copy of that data automatically become active.

Q8. Why do we use HDFS for applications having large data sets and not when there are a 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 because the Name node performs storage of metadata for the file system in RAM, the amount of memory limits the number of files in HDFS file system

Q9. How do you define “block” in HDFS? What is the default block size in Hadoop 1 and in Hadoop 2? Can it be changed?

HDFS splits huge file into small modules that are called Blocks. These are the smallest unit of data in file system The default size of a block in HDFS is 64 MB for Hadoop 1. X and 128 MB for Hadoop 2. x and we can change the Hadoop block size.