Top 25 Hadoop Admin Interview Questions and Answers

1) What daemons are needed to run a Hadoop cluster?

DataNode, NameNode, TaskTracker, and JobTracker are required to run Hadoop cluster.

2) Which OS are supported by Hadoop deployment?

The main OS use for Hadoop is Linux. However, by using some additional software, it can be deployed on Windows platform.

3) What are the common Input Formats in Hadoop?

Three widely used input formats are:

1. Text Input: It is default input format in Hadoop.
2. Key Value: It is used for plain text files
3. Sequence: Use for reading files in sequence

4) What modes can Hadoop code be run in?

Hadoop can be deployed in

1. Standalone mode
2. Pseudo-distributed mode
3. Fully distributed mode.

5) What is the main difference between RDBMS and Hadoop?

RDBMS is used for transactional systems to store and process the data whereas Hadoop can be used to store the huge amount of data.

6) What are the important hardware requirements for a Hadoop cluster?

There are no specific requirements for data nodes.

However, the namenodes need a specific amount of RAM to store filesystem image in memory. This depends on the particular design of the primary and secondary namenode.

7) How would you deploy different components of Hadoop in production?

You need to deploy jobtracker and namenode on the master node then deploy datanodes on multiple slave nodes.

8) What do you need to do as Hadoop admin after adding new datanodes?

You need to start the balancer for redistributing data equally between all nodes so that Hadoop cluster will find new datanodes automatically. To optimize the cluster performance, you should start rebalancer to redistribute the data between datanodes.

9) What are the Hadoop shell commands can use for copy operation?
The copy operation command are:

fs –copyToLocal

fs –put

fs –copyFromLocal.

10) What is the Importance of the namenode?

The role of namenonde is very crucial in Hadoop. It is the brain of the Hadoop. It is largely responsible for managing the distribution blocks on the system. It also supplies the specific addresses for the data based when the client made a request.

11) Explain how you will restart a NameNode?

The easiest way of doing is to run the command to stop running sell script.

Just click on stop.all.sh. then restarts the NameNode by clocking on start-all-sh.

12) What happens when the NameNode is down?

If the NameNode is down, the file system goes offline.

13) Is it possible to copy files between different clusters? If yes, How can you achieve this?

Yes, we can copy files between multiple Hadoop clusters. This can be done using distributed copy.

14) Is there any standard method to deploy Hadoop?

No, there are now standard procedure to deploy data using Hadoop. There are few general requirements for all Hadoop distributions. However, the specific methods will always different for each Hadoop admin.

15) What is distcp?

Distcp is a Hadoop copy utility. It is mainly used for performing MapReduce jobs to copy data. The key challenges in the Hadoop environment is copying data across various clusters, and distcp will also offer to provide multiple datanodes for parallel copying of the data.

16) What is a checkpoint?

Checkpointing is a method which takes a FsImage. It edits log and compacts them into a new FsImage. Therefore, instead of replaying an edit log, the NameNode can be load in the final in-memory state directly from the FsImage. This is surely more efficient operation which reduces NameNode startup time.

17) What is rack awareness?

It is a method which decides how to put blocks base on the rack definitions. Hadoop will try to limit the network traffic between datanodes which is present in the same rack. So that, it will only contact remote.

18) What is the use of 'jps' command?

The 'jps' command helps us to find that the Hadoop daemons are running or not. It also displays all the Hadoop daemons like namenode, datanode, node manager, resource manager, etc. which are running on the machine.

19) Name some of the essential Hadoop tools for effective working with Big Data?

"Hive," HBase, HDFS, ZooKeeper, NoSQL, Lucene/SolrSee, Avro, Oozie, Flume, Clouds, and SQL are some of the Hadoop tools that enhance the performance of Big Data.

20) How many times do you need to reformat the namenode?

The namenode only needs to format once in the beginning. After that, it will never formated. In fact, reformatting of the namenode can lead to loss of the data on entire the namenode.

21) What is speculative execution?

If a node is executing a task slower then the master node. Then there is needs to redundantly execute one more instance of the same task on another node. So the task finishes first will be accepted and the other one likely to be killed. This process is known as “speculative execution.”

22) What is Big Data?

Big data is a term which describes the large volume of data. Big data can be used to make better decisions and strategic business moves.

23) What is Hadoop and its components?

When “Big Data” emerged as a problem, Hadoop evolved as a solution for it. It is a framework which provides various services or tools to store and process Big Data. It also helps to analyze Big Data and to make business decisions which are difficult using the traditional method.

24) What are the essential features of Hadoop?

Hadoop framework has the competence of solving many questions for Big Data analysis. It's designed on Google MapReduce which is based on Google’s Big Data file systems.

25) What is the main difference between an "Input Split" and "HDFS Block"?

"Input Split" is the logical division of the data while The "HDFS Block" is the physical division of the data.