**Hadoop Cluster Interview Questions**

Looking out for Hadoop Cluster Interview Questions that are frequently asked by employers? Here is the second list of Hadoop Cluster Interview Questions which covers setting up a Hadoop Cluster. I hope you must not have missed the first part of our Hadoop Interview Questions series which covers the [***top 50 Hadoop interview questions***](https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/).

Always keep in mind that, only theoretical knowledge is not enough to crack an interview. Employers expects from the candidate to have practical knowledge and hands-on experience on Hadoop as well. So, this Hadoop Cluster Interview Questions will help you to gain practical knowledge of Hadoop framework.

If you are interested in installing Hadoop Cluster on your own system, you can go through our blog on [***Single Node Hadoop Cluster Setup***](https://www.edureka.co/blog/install-hadoop-single-node-hadoop-cluster) and [***Multi node Hadoop Cluster setup***](https://www.edureka.co/blog/setting-up-a-multi-node-cluster-in-hadoop-2.X).

**1. Which are the modes in which Hadoop can run?**

We have three modes in which Hadoop can run and that are:

* ***Standalone (local) mode***: Default mode of Hadoop, it uses the local file system for input and output operations. This mode is mainly used for debugging purpose, and it does not support the use of HDFS. Further, in this mode, there is no custom configuration required for mapred-site.xml, core-site.xml, hdfs-site.xml files.
* ***Pseudo-distributed mode:*** In this case, you need configuration for all the three files mentioned above. In this case, all daemons are running on one node and thus, both Master and Slave nodes are on the same machine.
* ***Fully distributed mode:*** This is the production phase of Hadoop where data is distributed across several nodes on a Hadoop cluster. Separate nodes are allotted as Master and Slaves.

**2. What are the features of Standalone (local) mode?**

In stand-alone mode, there are no daemons, everything runs on a single JVM. It has no DFS and it utilizes the local file system. Stand-alone mode is suitable only for running MapReduce programs during development for testing. It is one of the most least used environments.

**3. What are the features of Pseudo mode?**

Pseudo mode is used in both for development and in the testing environment. In the Pseudo mode, all the daemons run on the same machine.

**4. What are the features of Fully Distributed mode?**

This is an important question as Fully Distributed mode is used in the production environment, where we have ‘n’ number of machines forming a Hadoop cluster. Hadoop daemons run on a cluster of machines. There is one host onto which Namenode is running and other hosts on which Datanodes are running. NodeManagers are installed on every DataNode and it is responsible for execution of the task on every single DataNode. All these NodeManagers are managed by ResourceManager, which receives the processing requests, and then passes the parts of requests to corresponding NodeManagers accordingly.

**5. What is configured in /etc/hosts and what is its role in setting Hadoop cluster?**

This is a technical question which challenges your basic concept. /etc/hosts file contains the hostname and their IP address of that host. It maps the IP address to the hostname. In Hadoop cluster, we store all the hostnames (master and slaves) with their IP addresses in /etc/hosts so, that we can use hostnames easily instead of IP addresses.

**6. What are the default port numbers of NameNode, ResourceManager & MapReduce JobHistory Server?**

You are expected to remember basic server port numbers if you are working with Hadoop. The port number for corresponding daemons are as follows:

Namenode – ’50070’

ResourceManager – ’8088’

MapReduce JobHistory Server – ’19888’.

**7. What are the main Hadoop configuration files?**

*♣ Tip: Generally, approach this question by telling the 4 main configuration files in Hadoop and giving their brief descriptions to show your expertise.*

* **core-site.xml:**core-site.xml informs Hadoop daemon where NameNode runs on the cluster. It contains configuration settings of Hadoop core such as I/O settings that are common to HDFS & MapReduce.
* **hdfs-site.xml:** hdfs-site.xml contains configuration settings of HDFS daemons (i.e. NameNode, DataNode, Secondary NameNode). It also includes the replication factor and block size of HDFS.
* **mapred-site.xml**: mapred-site.xml contains configuration settings of the MapReduce framework like number of JVM that can run in parallel, the size of the mapper and the reducer, CPU cores available for a process, etc.
* **yarn-site.xml:**yarn-site.xml contains configuration settings of ResourceManager and NodeManager like application memory management size, the operation needed on program & algorithm, etc.

These files are in the *conf/hadoop/*directory inside Hadoop directory.

**8. How does Hadoop CLASSPATH plays vital role in starting or stopping in Hadoop daemons?**

*♣ Tip: To check your knowledge on Hadoop the interviewer may ask you this question.*

CLASSPATH includes all the directories containing jar files required to start/stop Hadoop daemons. The CLASSPATH is set inside ***/etc/hadoop/hadoop-env.sh*** file.

**9. What is a spill factor with respect to the RAM?**

*♣ Tip: This is a theoretical question, but if you add a practical taste to it, you might get a preference.*

The map output is stored in an in-memory buffer; when this buffer is almost full, then spilling phase starts in order to move data to a temp folder.

Map output is first written to buffer and buffer size is decided by *mapreduce.task.io.sort.mb* property .By default, it will be 100 MB.

When the buffer reaches certain threshold, it will start spilling buffer data to disk. This threshold is specified in*mapreduce.map.sort.spill.percent .*

**10. What is command to extract the compressed file in tar.gz format?**

This is an easy question,***tar -xvf /file\_location/filename.tar.gz***command will extract the tar.gz compressed file.

**11. How will you check Java and Hadoop is installed on your system?**

By using the following commands we can check whether Java and Hadoop are installed and their paths are set inside .bashrc file:

For checking Java –***java -version***

For checking Hadoop –***hadoop version***

**12. What is the default replication factor and how will you change it?**

The default replication factor is 3.

*♣ Tip: Default Replication Factor could be changed in three ways. Answering all the three ways will show your expertise.*

1. By adding this property to ***hdfs-site.xml***:

|  |  |
| --- | --- |
| 1  2  3  4  5 | <property>  <name>dfs.replication</name>  <value>5</value>  <description>Block Replication</description>  </property> |

1. Or you can change the replication factor on per file basis using following command:

***hadoop fs –setrep –w 3 /file\_location***

1. Or you can change replication factor for all the files in a directory using the following command:

***hadoop fs –setrep –w 3 -R /directory\_location***

**13. What is the full form of fsck?**

The full form of fsck is *File System Check.*HDFS supports the fsck (filesystem check) command to check for various inconsistencies. It is designed for reporting the problems with the files in HDFS, for example, missing blocks of a file or under-replicated blocks.

**14. Which are the main hdfs-site.xml properties?**

The three main hdfs-site.xml properties are:

1. ***dfs.name.dir***gives you the location where NameNode stores the metadata (FsImage and edit logs) and where DFS is located – on the disk or onto the remote directory.
2. ***dfs.data.dir***which gives you the location of DataNodes, where the data is going to be stored.
3. ***fs.checkpoint.dir*** is the directory on the filesystem where the Secondary NameNode stores the temporary images of edit logs, which is to be merged and the FsImage for backup.

**15. What happens if you get a ‘connection refused java exception’ when you type hadoop fsck /?**

If you get a ‘connection refused java exception’ when you type hadoop fsck, it could mean that the NameNode is not working on your VM.

**16. How can we view the compressed files via HDFS command?**

We can view compressed files in HDFS using ***hadoop fs -text /filename****command*.

**17. What is the command to move into safe mode and exit safe mode?**

*♣ Tip: Approach this question by first explaining safe mode and then moving on to the commands.*

Safe Mode in Hadoop is a maintenance state of the NameNode during which NameNode doesn’t allow any changes to the file system. During Safe Mode, HDFS cluster is read-only and doesn’t replicate or delete blocks.

* To know the status of safe mode, you can use the command: ***hdfs dfsadmin -safemode get***
* To enter safe mode: ***hdfs dfsadmin -safemode enter***
* To exit safe mode: ***hdfs dfsadmin -safemode leave***

**18. What does ‘jps’ command does?**

***jps*** command is used to check all the Hadoop daemons like NameNode, DataNode, ResourceManager, NodeManager etc. which are running on the machine.

**19. How can I restart Namenode?**

This question has two answers, answering both will give you a plus point. We can restart NameNode by following methods:

1. You can stop the NameNode individually using***.*** ***/sbin /hadoop-daemon.sh stop namenode***command and then start the NameNode using***.*** ***/sbin/hadoop-daemon.sh start namenode***
2. Use .***/sbin/stop-all.sh***and and then use .***/sbin/start-all.sh*** command which will stop all the daemons first and then start all the daemons.

**20. How can we check whether NameNode is working or not?**

To check whether NameNode is working or not, use the ***jps*** command, this will show all the running Hadoop daemons and there you can check whether NameNode daemon is running or not**.**

**21. How can we look at the Namenode in the web browser UI?**

If you want to look for NameNode in the browser, the port number for NameNode web browser UI is ***50070***. We can check in web browser using *http://master:50070/dfshealth.jsp*.

**22. What are the different commands used to startup and shutdown Hadoop daemons?**

*♣ Tip: Explain all the three ways of stopping and starting Hadoop daemons, this will show your expertise.*

1. ***./sbin/start-all.sh*** to start all the Hadoop daemons and .***/sbin/stop-all.sh*** to stop all the Hadoop daemons.
2. Then you can start all the dfs daemons together using***.*** ***/sbin/start-dfs.sh***, yarn daemons together using***./sbin/start-yarn.sh***and MR job history server using***.*** ***/sbin/mr-jobhistory-daemon.sh start historyserver***. To stop these daemons similarly we can use***.*** ***/sbin/start-yarn.sh***, ***./sbin/start-yarn.sh****&****./sbin/mr-jobhistory-daemon.sh stop historyserver***.
3. The last way is to start all the daemons individually and stop them individually:

***./sbin/hadoop-daemon.sh start namenode***

***./sbin/hadoop-daemon.sh start datanode***

***./sbin/yarn-daemon.sh start resourcemanager***

***./sbin/yarn-daemon.sh start nodemanager***

***./sbin/mr-jobhistory-daemon.sh start historyserver***

and stop them similarly.

**23. What do slaves file consist of?**

Slaves file consists of a list of hosts, one per line and the list contains DataNode location on which Node Manager servers run.

**24. What do masters file consist of?**

The masters file contains Secondary NameNode server location.

**25. What does hadoop-env.sh do?**

***hadoop-env.sh*** provides the environment for Hadoop to run. For example, JAVA\_HOME, CLASSPATH etc. are set over here.

**26. Where is hadoop-env.sh file present?**

As we discussed earlier, where all the configuration files reside, thus hadoop-env.sh file is present in the  */etc/hadoop* directory.

**27. In Hadoop\_PID\_DIR, what does PID stands for? What does it do?**

PID stands for ‘Process ID’. This directory stores the Process ID of the servers that are running.

**28. What does hadoop-metrics.properties file do?**

*♣ Tip: As this file is configured manually only in special cases, so answering this question will impress the interviewer indicating your expertise about configuration files.*

***hadoop-metrics.properties*** is used for ‘*Performance* *Reporting*‘ purposes. It controls the reporting for Hadoop. The API is abstract so that it can be implemented on top of a variety of metrics client libraries. The choice of client library is a configuration option, and different modules within the same application can use different metrics implementation libraries. This file is stored inside ***/etc/hadoop***.

**29. What are the network requirements for Hadoop?**

You should answer this question as, the Hadoop core uses Shell (SSH) for communication with salve and to launch the server processes on the slave nodes. It requires a *password-less* SSH connection between the master and all the slaves and the secondary machines, so every time it does not have to ask for authentication as master and slave requires rigorous communication.

**30. Why do we need a password-less SSH in Fully Distributed environment?**

We need a *password-less* SSH in a Fully-Distributed environment because when the cluster is live and running in Fully Distributed environment, the communication is too frequent. The DataNode and the NodeManager should be able to send messages quickly to master server.

**31. Does this lead to security issues?**

No, not at all. Hadoop cluster is an isolated cluster and generally, it has nothing to do with the internet. It has a different kind of a configuration. We needn’t worry about that kind of a security breach, for instance, someone hacking through the internet, and so on. Hadoop has a very secured way to connect to other machines to fetch and to process data.

**32. On which port does SSH work?**

SSH works on Port No. **22,**though it can be configured. **22** is the default Port number.

**33. Can you tell us more about SSH?**

SSH is nothing but a secure shell communication, it is a kind of a protocol that works on a Port No. 22, and when you do an SSH, what you really require is a password, to connect to the other machine. SSH is not only between masters and slaves, but can be between two hosts.

**34. What happens to a NameNode, when ResourceManager is down?**

When a ResourceManager is down, it will not be functional (for submitting jobs) but NameNode will be present. So, the cluster is accessible if NameNode is working, even if the ResourceManager is not working.

**35. How can we format HDFS?**

*♣ Tip: Attempt this question by starting with the command to format the HDFS and then exlain what this command does.*

Hadoop distributed file system(HDFS) can be formatted using ***bin/hadoop namenode -format***command. This command formats the HDFS via NameNode. This command is only executed for the first time. Formatting the file system means initializing the directory specified by the dfs.name.dir variable. If you run this command on existing filesystem, you will lose all your data stored on your NameNode. Formatting a Namenode will not format the DataNode. It will format the FsImage and edit logs data stored on the NameNode and will lose the data about the location of blocks stored in HDFS.

Never format, up and running Hadoop filesystem. You will lose all your data stored in the HDFS.

**36. Can we use Windows for Hadoop?**

Red*Hat Linux* and *Ubuntu* are the best Operating Systems for Hadoop. Windows is not used frequently for installing Hadoop as there are many support problems attached with Windows. Thus, Windows is not a preferred environment for Hadoop.

I hope these Hadoop Cluster Interview Questions were helpful for you. This is just a beginning of our Hadoop Interview Question series. I would suggest you to go through the whole series, to get in-depth knowledge on Hadoop Interview Questions. It’s never too late to strengthen your basics. Learn Hadoop from industry experts while working with real-life use cases.