# 50 Must-Read Hadoop Interview Questions & Answers

Whizlabs Dec 29th, 2017 - Big Data

Are you planning to land a job with big data and data analytics? Are you worried about cracking the Hadoop job interview? We have put together a list of Hadoop Interview Questions that will come in handy. You might have sound knowledge regarding the software framework, but all of it can't be tested in a short 15 minutes interview session. So the interviewer will ask you some specific questions they think are apt to judge your knowledge in the subject matter.

## Top 50 Hadoop Interview Questions and Answers

Currently, jobs related to Big Data are on the rise. One out of every five big companies is moving to Big Data Analytics, and hence it is high time to start applying for jobs in this field. So without further delay, we present Top 50 Hadoop Interview Questions and Answers that will help you to crack the interview**.**

# Basic Hadoop Interview Questions

These are the most common and popularly asked Hadoop Interview Questions which you are bound to face in big data interviews. Preparing through these Hadoop Interview Questions will undoubtedly give you an edge over the competition.

To start off the list, we will be focusing on the common and Basic Hadoop Interview Questions that people come across when applying for a Hadoop related job, irrespective of position.

## 1. What are the concepts used in the Hadoop Framework?

Answer: The Hadoop Framework functions on two core concepts:

- **HDFS**: Abbreviation for Hadoop Distributed File System, it is a Java-based file system for scalable and reliable storage of large datasets. HDFS itself works on the Master-Slave Architecture and stores all its data in the form of blocks.
- **MapReduce**: This is the programming model and the associated implementation for processing and generating large data sets. The Hadoop jobs are basically divided into two different tasks job. The map job breaks down the data set into key-value pairs or tuples. And then the reduce job takes the output of the map job and combines the data tuples into a smaller set of tuples.

***Preparing for MapReduce Interview? Here' 10 Most Popular MapReduce Interview Questions***

## 2. What is Hadoop? Name the Main Components of a Hadoop Application.

Answer: Hadoop is what evolved as the solution to the "Big Data" problem. Hadoop is described as the framework that offers a number of tools and services in order to store and process Big Data. It also plays an important role in the analysis of big data and to make efficient business decisions when it is difficult to make the decision using the traditional method.

Hadoop offers a vast toolset that makes it possible to store and process data very easily. Here are all the main components of the Hadoop:

- Hadoop Common
- HDFS
- Hadoop MapReduce

- YARN
- PIG and HIVE – The Data Access Components.
- HBase – For Data Storage
- Apache Flume, Sqoop, Chukwa – The Data Integration Components
- Ambari, Oozie and ZooKeeper – Data Management and Monitoring Component
- Thrift and Avro – Data Serialization components
- Apache Mahout and Drill – Data Intelligence Components

## 3. How many Input Formats are there in Hadoop? Explain.

Answer: There are following three input formats in Hadoop –

1. Text Input Format: The text input is the default input format in Hadoop.
2. Sequence File Input Format: This input format is used to read files in sequence.
3. Key Value Input Format: This input format is used for plain text files.

## 4. What do you know about YARN?

Answer: YARN stands for Yet Another Resource Negotiator, it is the Hadoop processing framework. YARN is responsible to manage the resources and establish an execution environment for the processes.

## 5. Why do the nodes are removed and added frequently in a Hadoop cluster?

Answer: The following features of Hadoop framework makes a Hadoop administrator to add (commission) and remove (decommission) Data Nodes in a Hadoop clusters –

1. The Hadoop framework utilizes commodity hardware, and it is one of the important features of Hadoop framework. It results in a frequent DataNode crash in a Hadoop cluster.
2. The ease of scale is yet another important feature of the Hadoop framework that is performed according to the rapid growth of data volume.

## 6. What do you understand by "Rack Awareness"?

Answer: In Hadoop, Rack Awareness is defined as the algorithm through which NameNode determines how the blocks and their replicas are stored in the Hadoop cluster. This is done via rack definitions that minimize the traffic between DataNodes within the same rack. Let's take an example – we know that the default value of replication factor is 3. According to the "Replica Placement Policy" two copies of replicas for every block of data will be stored in a single rack whereas the third copy is stored in the different rack.

## 7. What do you know about the Speculative Execution?

Answer: In Hadoop, Speculative Execution is a process that takes place during the slower execution of a task at a node. In this process, the master node starts executing another instance of that same task on the other node. And the task which is finished first is accepted and the execution of other is stopped by killing that.

## 8. State some of the important features of Hadoop.

Answer: The important features of Hadoop are –

- Hadoop framework is designed on Google MapReduce that is based on Google's Big Data File Systems.
- Hadoop framework can solve many questions efficiently for Big Data analysis.

## 9. Do you know some companies that are using Hadoop?

Answer: Yes, I know some popular names that are using Hadoop.

Yahoo – using Hadoop

Facebook – developed Hive for analysis

Amazon, Adobe, Spotify, Netflix, eBay, and Twitter are some other well-known and established companies that are using Hadoop.

## 10. How can you differentiate RDBMS and Hadoop?

Answer: The key points that differentiate RDBMS and Hadoop are –

1. RDBMS is made to store structured data, whereas Hadoop can store any kind of data i.e. unstructured, structured, or semi-structured.
2. RDBMS follows "Schema on write" policy while Hadoop is based on "Schema on read" policy.
3. The schema of data is already known in RDBMS that makes Reads fast, whereas in HDFS, writes no schema validation happens during HDFS write, so the Writes are fast.
4. RDBMS is licensed software, so one needs to pay for it, whereas Hadoop is open source software, so it is free of cost.
5. RDBMS is used for Online Transactional Processing (OLTP) system whereas Hadoop is used for data analytics, data discovery, and OLAP system as well.

# Hadoop Architecture Interview Questions

Up next we have some Hadoop interview questions based on Hadoop architecture. Knowing and understanding the Hadoop architecture helps a Hadoop professional to answer all the Hadoop Interview Questions correctly.

## 11. What are the differences between Hadoop 1 and Hadoop 2?

Answer: The following two points explain the difference between Hadoop 1 and Hadoop 2:

In Hadoop 1.X, there is a single NameNode which is thus the single point of failure whereas, in Hadoop 2.x, there are Active and Passive NameNodes. In case, the active NameNode fails, the passive NameNode replaces the active NameNode and takes the charge. As a result, high availability is there in Hadoop 2.x.

In Hadoop 2.x, the YARN provides a central resource manager that share a common resource to run multiple applications in Hadoop whereas data processing is a problem in Hadoop 1.x.

## 12. What do you know about active and passive NameNodes?

Answer: In high-availability Hadoop architecture, two NameNodes are present.

**Active NameNode –** The NameNode that runs in Hadoop cluster, is the Active NameNode.

**Passive NameNode –** The standby NameNode that stores the same data as that of the Active NameNode is the Passive NameNode.

On the failure of active NameNode, the passive NameNode replaces it and takes the charge. In this way, there is always a running NameNode in the cluster and thus it never fails.

## 13. What are the Components of Apache HBase?

Answer: Apache HBase Consists of the following main components:

- **Region Server**: A Table can be divided into several regions. A group of these regions gets served to the clients by a Region Server.
- **HMaster**: This coordinates and manages the Region server.

- **ZooKeeper**: This acts as a coordinator inside HBase distributed environment. It functions by maintaining server state inside of the cluster by communication in sessions.

## 14. How is the DataNode failure handled by NameNode?

Answer: NameNode continuously receives a signal from all the DataNodes present in Hadoop cluster that specifies the proper function of the DataNode. The list of all the blocks present on a DataNode is stored in a block report. If a DataNode is failed in sending the signal to the NameNode, it is marked dead after a specific time period. Then the NameNode replicates/copies the blocks of the dead node to another DataNode with the earlier created replicas.

## 15. Explain the NameNode recovery process.

Answer: The process of NameNode recovery helps to keep the Hadoop cluster running, and can be explained by the following steps –

Step 1: To start a new NameNode, utilize the file system metadata replica (FsImage).

Step 2: Configure the clients and DataNodes to acknowledge the new NameNode.

Step 3: Once the new Name completes the loading of last checkpoint FsImage and receives block reports from the DataNodes, the new NameNode start serving the client.

## 16. What are the different schedulers available in Hadoop?

Answer: The different available schedulers in Hadoop are –

COSHH – It schedules decisions by considering cluster, workload, and using heterogeneity.

FIFO Scheduler – It orders the jobs on the basis of their arrival time in a queue without using heterogeneity.

Fair Sharing – It defines a pool for each user that contains a number of maps and reduce slots on a resource. Each user is allowed to use own pool for the execution of jobs.

## 17. Can DataNode and NameNode be commodity hardware?

Answer: DataNodes are the commodity hardware only as it can store data like laptops and personal computers, these are required in large numbers. Instead, NameNode is the master node; it stores metadata about all the blocks stored in HDFS. It needs high memory space, thus works as a high-end machine with great memory space.

## 18. What are the Hadoop daemons? Explain their roles.

Answer: The Hadoop daemons are NameNode, Secondary NameNode, DataNode, NodeManager, ResourceManager, JobHistoryServer. The role of different Hadoop daemons is –

**NameNode –** The master node, responsible for metadata storage for all directories and files is known as the NameNode. It also contains metadata information about each block of the file and their allocation in Hadoop cluster.

**Secondary NameNode –** This daemon is responsible to merge and store the modified Filesystem Image into permanent storage. It is used in case the NameNode fails.

**DataNode –** The slave node containing actual data is the DataNode.

**NodeManager –** Running on the slave machines, the NodeManager handles the launch of application container, monitoring resource usage and reporting same to the ResourceManager.

**ResourceManager –** It is the main authority responsible to manage resources and to schedule applications running on the top of YARN.

**JobHistoryServer –** It is responsible to maintain every information about the MapReduce jobs when the Application Master stops to work (terminates).

### 19. Define "Checkpointing". What is its benefit?

Answer: Checkpointing is a procedure to that compacts a FsImage and Edit log into a new FsImage. In this way, the NameNode handles the loading of the final in-memory state from the FsImage directly, instead of replaying an edit log. The secondary NameNode is responsible to perform the checkpointing process.

**Benefit of Checkpointing**

Checkpointing is a highly efficient process and decreases the startup time of the NameNode.

## Hadoop Administrator Interview Questions

The Hadoop Administrator is responsible to handle that Hadoop cluster is running smoothly. To crack the Hadoop Administrator job interview, you need to go through Hadoop Interview Questions related to Hadoop environment,  cluster etc. The common Hadoop Administrator interview questions are as follows:

### 20. When deploying Hadoop in a production environment, what are the Important Hardware considerations?

Answer: **Memory System's memory requirements**: This will vary between the worker services and management services based on the application.

**Operating System:** A 64-bit OS is preferred as it avoids any such restrictions on the amount of memory that can be used on the worker nodes.

**Storage**: A Hadoop Platform should be designed by moving the computing activities to data and thus achieving scalability and high performance.

**Capacity**: Large Form Factor disks will cost less and allow for more storage.

**Network:** Two TOR switches per rack is ideal to avoid any chances for redundancy.

## 21. What should you consider while deploying a secondary NameNode?

Answer: A secondary NameNode should always be deployed on a separate Standalone system. This prevents it from interfering with the operations of the primary node.

## 22. Name the modes in which Hadoop code can be run.

Answer: There are different modes to run Hadoop code –

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

## 23. Name the operating systems supported by Hadoop deployment.

Answer: Linux is the main operating system that is used for Hadoop. However, it can also e deployed on Windows operating system with the help of some additional software.

## 24. Why is HDFS used for the applications with large data sets, and not for the multiple small files?

Answer: 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.

## 25. What are the important properties of hdfs-site.xml?

Answer: There are three important properties of hdfs-site.xml:

- data.dr – identify the location of the storage of data.
- name.dr – identifies the location of metadata storage and specify whether DFS is located on disk or the on the remote location.

- checkpoint.dir – for Secondary NameNode.

### 26. What are the essential Hadoop tools that enhance the performance of Big Data?

Answer: Some of the essential Hadoop tools that enhance the performance of Big Data are –

Hive, HDFS, HBase, Avro, SQL, NoSQL, Oozie, Clouds, Flume, SolrSee/Lucene, and ZooKeeper

### 27. What do you know about SequenceFile?

Answer: SequenceFile is defined as the flat file that contains binary key or value pairs. It is mainly used in Input/Output format of the MapReduce. The map outputs are stored internally as SequenceFile.

Different formats of SequenceFile are –

**Record compressed key/value records –** In this format, values are compressed.

**Block compressed key/value records –** In this format, both the values and keys are separately stored in blocks and then compressed.

**Uncompressed key/value records –** In this format, neither values nor keys are compressed.

### 28. Explain the functions of Job Tracker.

Answer: In Hadoop, the Job Tracker performs various functions, that are followings –

1. It manages resources, tracks availability of resources, and manages the life cycle of tasks.
2. It is responsible to identify the location of data by communicating with NameNode.
3. It executes the tasks on given nodes by finding the best task tracker node.
4. Job Tracker manages to monitor the all task trackers individually and then submit the overall job to the client.
5. It is responsible to track the MapReduce workloads execution from local to the slave node.

---

## Hadoop HDFS Interview Questions

Hadoop Distributed File System (HDFS) is the main storage system used by Hadoop. For a Hadoop professional, it is required to have the knowledge of HDFS, its components, and its working. So, here are some HDFS based Hadoop Interview Questions that will help you to go through Hadoop interview. It is recommended to first read the basic Hadoop interview questions before these HDFS related Hadoop interview questions for better understanding.

**29. How is HDFS different from NAS?**

Answer: The following points differentiates HDFS from NAS –

- Hadoop Distributed File System (HDFS) is a distributed file system that stores data using commodity hardware whereas Network Attached Storage (NAS) is just a file level server for data storage, connected to a computer network.
- HDFS stores data blocks in the distributed manner on all the machines present in a cluster whereas NAS stores data on a dedicated hardware.
- HDFS stores data using commodity hardware that makes it cost-effective while NAS stores data on high-end devices that includes high expenses.
- HDFS work with MapReduce paradigm while NAS does not work with MapReduce as data and computation are stored separately.

## 30. Is HDFS fault-tolerant? If yes, how?

Answer: Yes, HDFS is highly fault-tolerant. Whenever some data is stored on HDFS, the NameNode replicates (copies) that data to multiple DataNode. The value of default replication factor is 3 that can be changed as per your requirements. In case a DataNode goes down, the NameNode takes the data from replicas and copies it to another node, thus makes the data available automatically. In this way, HDFS has fault tolerance feature and known as fault tolerant.

## 31. Differentiate HDFS Block and Input Split.

Answer: The main difference between HDFS Block and the Input Split is that the HDFS Block is known to be the physical division of data whereas the Input Split is considered as the logical division of the data. For processing, HDFS first divides data into blocks and then stores all the blocks together, while the MapReduce first divides the data into input split and then assign this input split to the mapper function.

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

Answer: Note that HDFS is known to support exclusive writes (processes one write request for a file at a time) only.

Wh the n first client contacts the NameNode to open the file to write, the NameNode provides a lease to the client to create this file. When the second client sends a request to open that same file to write, the NameNode find that the lease for that file has already been given to another client, and thus reject the second client’s request.

## 33. What is the block in HDFS?

Answer: The smallest site or say, location on the hard drive that is available to store data, is known as the block. The data in HDFS is stored as blocks and then it is distributed over the Hadoop cluster. The whole file is first divided into small blocks and then stored as separate units.

---

# Hadoop Developer Interview Questions

A Hadoop developer is responsible for the development of Hadoop applications while working in the big data domain. The Hadoop interview questions are simply based on the understanding of Hadoop ecosystem and its components. Here are the Hadoop interview questions that will help you with Hadoop developer interview.

### 34. What is Apache Yarn?

Answer: YARN stands for Yet Another Resource Negotiator. It is a Hadoop Cluster resource management system. It was introduced in Hadoop 2 to help MapReduce and is the next generation computation and resource management framework in Hadoop. It allows Hadoop to support more varied processing approaches and a broader array of applications.

### 35. What is Node Manager?

Answer: Node Manager is the YARN equivalent of the Tasktracker. It takes in instructions from the ResourceManager and manages resources available on a single node. It is responsible for containers and also monitors and reports their resource usage to the ResourceManager. Every single container processes that run on a slave node gets initially provisioned, monitored and tracked by the Node Manager daemon corresponding to that slave node.

### 36. What is the RecordReader in Hadoop used for?

In Hadoop, RecordReader is used to read the split data into a single record. It is important to combine data as Hadoop splits the data into various blocks. For example, if the input data is split like –

Row1: Welcome to

Row 2: the Hadoop world

Using RecordReader, it will be read as “Welcome to the Hadoop world”.

### 37. What is the procedure to compress mapper output without affecting reducer output?

In order to compress the mapper output without affecting reducer output, set the following:

Conf.set(“mapreduce.map.output.compress” , true)

Conf.set(“mapreduce.output.fileoutputformat.compress” , false)

## 38. Explain different methods of a Reducer.

The different methods of a Reducer are as follows:

1. Setup() – It is used to configure different parameters such as input data size.

Syntax: public void setup (context)

1. Cleanup() – It is used for cleaning all the temporary files at the end of the task.

Syntax: public void cleanup (context)

1. Reduce() – This method is known as the heart of the reducer. It is regularly used once per key with the associated reduce task.

Syntax: public void reduce (Key, Value, context)

## 39. How can you configure the replication factor in HDFS?

For the configuration of HDFS, hdfs-site.xml file is used. In order to change the default value of replication factor for all the files stored in HDFS, following property is changed in hdfs-site.xml

dfs.replication

## 40. What is the use of “jps” command?

The “jps” command is used to check whether the Hadoop daemons are in running state. This command will list all the Hadoop daemons running on the machine i.e. namenode, nodemanager, resourcemanager, datanode etc.

## 41. What is the procedure to restart “NameNode” or all other daemons in Hadoop?

There are different methods to restart NameNode and all other daemons in Hadoop –

**Method to restart NameNode:** First, stop the NameNode using the command /sbin/hadoop-daemon.sh stop namenode and then start the NameNode again using the command /sbin/hadoop-daemon.sh start namenode

**Method to restart all the daemons:** Use the command /sbin/stop-all.sh to stop all the daemons at a time and then use the command /sbin/start-all.sh to start all the stopped daemons at the same time.

## 42. What is the query to transfer data from Hive to HDFS?

The query to transfer data from Hive to HDFS is –

hive> insert overwrite directory ‘/ ‘ select * from emp;

The output of this query will be stored in the part files at the specified HDFS path.

### 43. What are the common Hadoop shell commands used for Copy operation?

The common Hadoop shell commands for Copy operation are –

fs –copyToLocal

fs –put

fs –copyFromLocal

---

## Scenario-Based Hadoop Interview Questions

Often you will be asked some tricky questions regarding particular scenarios and how you will handle them. The reason for asking such Hadoop Interview Questions is to check your Hadoop skills. These Hadoop interview questions specify how you implement your Hadoop knowledge and approach to solve given big data problem. These Scenario-based Hadoop interview questions will give you an idea.

### 44. You have a directory XYZ that has the following files – Hadoop123Training.txt,_Spark123Training.txt,#DataScience123Training.txt, .Salesforce123Training.txt. If you pass the XYZ directory to the Hadoop MapReduce jobs, how many files are likely to be processed?

Answer: Hadoop123Training.txt and #DataScience123Training.txt are the only files that will be processed by MapReduce jobs. This happens because we need to confirm that none of the files has a hidden file prefix such as "_" or ."" while processing a file in Hadoop using a FileInputFormat. MapReduce FileInputFormat will use HiddenFileFilter class by default to ignore all such files. However, we can create our custom filter to eliminate such criteria.

### 45. We have a Hive partitioned table where the country is the partition column. We have 10 partitions and data is available for just one country. If we want to copy the data for other 9 partitions, will it be reflected with a command or manually?

Answer: In the above case, the data will only be available for all the other partitions when the data will be put through command, instead of copying it manually.

## 46. What is the difference between -put, -copyToLocal, and and –copyFromLocal commands?

These three commands can be differentiated on the basis of what they are used for –

-put: This command is used to copy the file from a source to the destination

-copyToLocal: This command is used to copy the file from Hadoop system to the local file system.

-copyFromLocal: This command is used to copy the file from the local file system to the Hadoop System.

## 47. What is the difference between Left Semi Join and Inner Join

The Left Semi Join will return the tuples only from the left-hand table while the Inner Join will return the common tuples from both the tables (i.e. left-hand and right-hand tables) depending on the given condition.

## 48. What are the values of default block size in Hadoop 1 and Hadoop 2? Is it possible to change the block size?

Answer: The default value of block size in Hadoop 1 is 64 MB.

The default value of block size in Hadoop 2 is 128 MB.

Yes, it is possible to change the block size from the default value. The following parameter is used hdfs-site.xml file to change and set the block size in Hadoop –

dfs.block.size

## 49. How will you check if NameNode is working properly with the use of jps command?

Answer: The following status can be used to check it NameNode is working with the use of jps command

/etc/init.d/hadoop-0.20-namenode

## 50. MapReduce jobs are getting failed on a recently restarted cluster while these jobs were working well before the restart. What can be the reason for this failure?

Depending on the size of data, the replication of data will take some time. Hadoop cluster requires to copy/replicate all the data. So, the clear reason for job failure is the big data size, and thus the replication process is being delayed. It can take even few minutes to some hours to take place and thus, for the jobs to work properly.

## Conclusion

Hadoop is a constantly growing field that opens a large number of jobs every year for freshers as well as experienced ones. The best way to prepare for a Hadoop job is to answer all the Hadoop Interview Questions you find your way. We created this list of Hadoop interview questions for you, that we will keep regularly updating. These are the Hadoop interview questions that have been asked in recent Hadoop interviews, and thus will be helpful for you.

*If you want any other information about Hadoop, just leave a comment below and our Hadoop expert will get in touch with you. In case, you are looking for Big Data certification (HDPCA/HDPCD) online training, click here.*