**1)What is BigData?**

**2)What is Hadoop?**

**3)How should hadoop solving the big data problems?**

A) [Read Hadoop Mapreduce Interview Question](http://www.hadooptpoint.com/bigdata-hadoop-interview-questions/)

4) What is SSH ?

A) Secure Shell also called as Secure Socket Shell

5) What is the Use of SSH in Hadoop ?

A) We should use SSH in Hadoop because SSH is a built-in username and password schema that can be used for secure access to a remote host; it is a more secure alternative to rlogin and telnet .

6) How will format the HDFS ?

A) $hadoop namenode -format

**7)What is MapReduce?**

**8)What are different types of filesystem?**

**9)What is HDFS?**

A) [Read Hadoop Mapreduce Interview Question](http://www.hadooptpoint.com/bigdata-hadoop-interview-questions/)

10) What are the most commonly used Input Formats in Hadoop ?

* TextInputFormat
* KeyValueInputFormat
* SequenceFileInputFormat

11) Which one is default InputFormat in Hadoop ?

A) TextInputFormat

**12)What is the default block size in hdfs?**

64MB

13) What is Hadoop Custom partitioner ?

A) [Read About Hadoop Custom Partitioner in hadoop](http://www.hadooptpoint.com/hadoop-custom-partitioner-in-mapreduce-example/)

14) Write a Mapreduce Program for Character Count ?

A) [Character Count Program In Mapreduce](http://www.hadooptpoint.com/character-count-example-in-mapreduce-java/)

[DeZyre - live hands on training](https://www.dezyre.com/)

* Call us 1-844-696-6465 (US Toll Free)

* [Home](https://www.dezyre.com/)

* [Courses](https://www.dezyre.com/article/top-100-hadoop-interview-questions-and-answers-2016/159)

* [Blog](https://www.dezyre.com/blog)

* [Tutorials](https://www.dezyre.com/tutorial)

* [Interview Questions](https://www.dezyre.com/blog/interview-questions)

* [Online Hackathons](https://www.dezyre.com/hackerday)

* [Student Portfolios](https://www.dezyre.com/student-portfolios)

* [Sign In](https://www.dezyre.com/user/login)
* Build Projects, Learn Skills, Get Hired

* [REQUEST INFO](https://www.dezyre.com/article/top-100-hadoop-interview-questions-and-answers-2016/159)

**Top 100 Hadoop Interview Questions and Answers 2016**

21 Aug 2015

*Latest Update made on October 28, 2016.*

Big Data and Hadoop is a constantly changing field which required people to quickly upgrade their skills, to fit the requirements for Hadoop related jobs. If you are applying for a Hadoop job role, it is best to be prepared to answer any Hadoop interview question that might come your way. We will keep updating this list of [Hadoop Interview](http://www.dezyre.com/article/-dezyre-insync-interview-tips-to-get-hired-by-big-data-hadoop-companies/136)questions, to suit the current industry standards.

*If you would like more information about Big Data and Hadoop Training, please click the orange "Request Info" button on top of this page.*

With more than 30,000 open Hadoop developer jobs, professionals must familiarize themselves with the each and every component of the Hadoop ecosystem to make sure that they have a deep understanding of what Hadoop is so that they can form an effective approach to a given big data problem. To help you get started, DeZyre presented a [comprehensive list of Top 50 Hadoop Developer Interview Questions](https://www.dezyre.com/article/top-50-hadoop-interview-questions-/92)asked during recent Hadoop job interviews.

With the help of DeZyre’s Hadoop Instructors, we have put together a detailed list of Hadoop latest interview questions based on the different [components of the Hadoop Ecosystem](https://www.dezyre.com/article/big-data-and-hadoop-training-hadoop-components-and-architecture/114) such as MapReduce, [Hive](https://www.dezyre.com/hadoop-course/hive), HBase, [Pig](https://www.dezyre.com/hadoop-course/pig), YARN, [Flume](https://www.dezyre.com/hadoop-course/flume), Sqoop, HDFS, etc.

[](https://www.dezyre.com/Hadoop-Training-online/19?from=a159bgs)

We had to spend lots of hours researching and deliberating on what are the best possible answers to these interview questions. We would love to invite people from the industry – hadoop developers, hadoop admins and architects to kindly help us and everyone else – with answering the unanswered questions.



Here are top Hadoop Developer Interview Questions and Answers based on different components of the Hadoop Ecosystem-

**1)**[**Hadoop Basic Interview Questions**](https://www.dezyre.com/article/top-100-hadoop-interview-questions-and-answers-2016/159#basicAnswers)

**2)**[**Hadoop HDFS Interview Questions**](https://www.dezyre.com/article/top-100-hadoop-interview-questions-and-answers-2016/159#hdfs)

**3) [MapReduce Interview Questions](https://www.dezyre.com/article/top-100-hadoop-interview-questions-and-answers-2016/159" \l "mapReduce)**

**4)**[**Hadoop HBase Interview Questions**](https://www.dezyre.com/article/top-100-hadoop-interview-questions-and-answers-2016/159#hbase)

**5)**[**Hadoop Sqoop Interview Questions**](https://www.dezyre.com/article/top-100-hadoop-interview-questions-and-answers-2016/159#sqoop)

**6)**[**Hadoop Flume Interview Questions**](https://www.dezyre.com/article/top-100-hadoop-interview-questions-and-answers-2016/159#flume)

**7)**[**Hadoop Zookeeper Interview Questions**](https://www.dezyre.com/article/top-100-hadoop-interview-questions-and-answers-2016/159#zookeeper)

**8)**[**Pig Interview Questions**](https://www.dezyre.com/article/top-100-hadoop-interview-questions-and-answers-2016/159#pig)

**9)**[**Hive Interview Questions**](https://www.dezyre.com/article/top-100-hadoop-interview-questions-and-answers-2016/159#hive)

**10)**[**Hadoop YARN Interview Questions**](https://www.dezyre.com/article/top-100-hadoop-interview-questions-and-answers-2016/159#yarn)

**Big Data Hadoop Interview Questions and Answers**

These are Hadoop Basic Interview Questions and Answers for freshers and experienced.

**1. What is Big Data?**[Click here to Tweet](http://ctt.ec/7bp0j)

Big data is defined as the voluminous amount of structured, unstructured or semi-structured data that has huge potential for mining but is so large that it cannot be processed using traditional database systems. Big data is characterized by its high velocity, volume and variety that requires cost effective and innovative methods for information processing to draw meaningful business insights. More than the volume of the data – it is the nature of the data that defines whether it is considered as Big Data or not.

Here is an interesting and explanatory visual on “What is Big Data?”

**2. What do the four V’s of Big Data denote?** [Click here to Tweet](http://ctt.ec/8oUMr)

IBM has a nice, simple explanation for the four critical features of big data:  
a) Volume –Scale of data  
b) Velocity –Analysis of streaming data  
c) Variety – Different forms of data  
d) Veracity –Uncertainty of data

Here is an explanatory video on the four V’s of Big Data

**3. How big data analysis helps businesses increase their revenue? Give example.**[Click here to Tweet](http://ctt.ec/vchus)

Big data analysis is helping businesses differentiate themselves – for example Walmart the world’s largest retailer in 2014 in terms of revenue - is using big data analytics to increase its sales through better predictive analytics, providing customized recommendations and launching new products based on customer preferences and needs. Walmart observed a significant 10% to 15% increase in online sales for $1 billion in incremental revenue. There are many more companies like Facebook, Twitter, LinkedIn, Pandora, JPMorgan Chase, Bank of America, etc. using big data analytics to boost their revenue.

Here is an interesting video that explains how various industries are leveraging big data analysis to increase their revenue

**4. Name some companies that use Hadoop.** [Click here to tweet this question](http://ctt.ec/941dz)

Yahoo (One of the biggest user & more than 80% code contributor to Hadoop)  
Facebook  
Netflix  
Amazon  
Adobe  
eBay  
Hulu  
Spotify  
Rubikloud  
Twitter

What companies are you applying to for Hadoop job roles?



Click on this link to view a detailed [list of some of the top companies using Hadoop](https://www.dezyre.com/article/top-10-industries-using-big-data-and-121-companies-who-hire-hadoop-developers/69).

**5. Differentiate between Structured and Unstructured data.**[Click here to Tweet](http://ctt.ec/40Hif)

Data which can be stored in traditional database systems in the form of rows and columns, for example the online purchase transactions can be referred to as Structured Data. Data which can be stored only partially in traditional database systems, for example, data in XML records can be referred to as semi structured data. Unorganized and raw data that cannot be categorized as semi structured or structured data is referred to as unstructured data. Facebook updates, Tweets on Twitter, Reviews, web logs, etc. are all examples of unstructured data.

**6. On what concept the Hadoop framework works?** [Click here to Tweet](http://ctt.ec/Hi6cx)

Hadoop Framework works on the following two core components-

1)HDFS – [Hadoop Distributed File System](https://www.dezyre.com/hadoop-course/hdfs) is the java based file system for scalable and reliable storage of large datasets. Data in HDFS is stored in the form of blocks and it operates on the Master Slave Architecture.

2)Hadoop MapReduce-This is a java based programming paradigm of Hadoop framework that provides scalability across various Hadoop clusters. [MapReduce](https://www.dezyre.com/hadoop-course/mapreduce" \o "All you need to know about Hadoop MapReduce" \t "_blank)distributes the workload into various tasks that can run in parallel. Hadoop jobs perform 2 separate tasks- job. The map job breaks down the data sets into key-value pairs or tuples. The reduce job then takes the output of the map job and combines the data tuples to into smaller set of tuples. The reduce job is always performed after the map job is executed.

Here is a visual that clearly explain the HDFS and Hadoop MapReduce Concepts-

**7) What are the main components of a Hadoop Application?** [Click here to Tweet](http://ctt.ec/2614o)

Hadoop applications have wide range of technologies that provide great advantage in solving complex business problems.

Core components of a Hadoop application are-

1) Hadoop Common

2) HDFS

3) Hadoop MapReduce

4) YARN

Data Access Components are - Pig and Hive

Data Storage Component is - HBase

Data Integration Components are - Apache Flume, Sqoop, Chukwa

Data Management and Monitoring Components are - Ambari, Oozie and Zookeeper.

Data Serialization Components are - Thrift and Avro

Data Intelligence Components are - Apache Mahout and Drill.

**8. What is Hadoop streaming?** [Click here to Tweet](http://ctt.ec/bVeo5)

Hadoop distribution has a generic application programming interface for writing Map and Reduce jobs in any desired programming language like Python, Perl, Ruby, etc. This is referred to as Hadoop Streaming. Users can create and run jobs with any kind of shell scripts or executable as the Mapper or Reducers.

**9. What is the best hardware configuration to run Hadoop?** [Click here to Tweet](http://ctt.ec/yzs9b)

The best configuration for executing Hadoop jobs is dual core machines or dual processors with 4GB or 8GB RAM that use ECC memory. Hadoop highly benefits from using ECC memory though it is not low - end. ECC memory is recommended for running Hadoop because most of the Hadoop users have experienced various checksum errors by using non ECC memory. However, the hardware configuration also depends on the workflow requirements and can change accordingly.

**10. What are the most commonly defined input formats in Hadoop?** [Click here to Tweet](http://ctt.ec/BF0Xb)

The most common Input Formats defined in Hadoop are:

* Text Input Format- This is the default input format defined in Hadoop.
* Key Value Input Format- This input format is used for plain text files wherein the files are broken down into lines.
* Sequence File Input Format- This input format is used for reading files in sequence.

We have further categorized Big Data Interview Questions for Freshers and Experienced-

* Hadoop Interview Questions and Answers for Freshers - Q.Nos- 1,2,4,5,6,7,8,9
* Hadoop Interview Questions and Answers for Experienced - Q.Nos-3,8,9,10

For a detailed PDF report on Hadoop Salaries - [**CLICK HERE**](https://docs.google.com/forms/d/1LFuWEKQKCLR231qR9WE5PZakJj77fTDIW6ox5328HFM/viewform)

**Hadoop HDFS Interview Questions and Answers**

**1. What is a block and block scanner in HDFS?** [Click here to Tweet](http://ctt.ec/lbfcd)

Block - The minimum amount of data that can be read or written is generally referred to as a “block” in HDFS. The default size of a block in HDFS is 64MB.

Block Scanner - Block Scanner tracks the list of blocks present on a DataNode and verifies them to find any kind of checksum errors. Block Scanners use a throttling mechanism to reserve disk bandwidth on the datanode.

**2. Explain the difference between NameNode, Backup Node and Checkpoint NameNode.** [Click here to Tweet](http://ctt.ec/JG2aa)

**NameNode**: NameNode is at the heart of the HDFS file system which manages the metadata i.e. the data of the files is not stored on the NameNode but rather it has the directory tree of all the files present in the HDFS file system on a hadoop cluster. NameNode uses two files for the namespace-

fsimage file- It keeps track of the latest checkpoint of the namespace.

edits file-It is a log of changes that have been made to the namespace since checkpoint.

**Checkpoint Node-**

Checkpoint Node keeps track of the latest checkpoint in a directory that has same structure as that of NameNode’s directory. Checkpoint node creates checkpoints for the namespace at regular intervals by downloading the edits and fsimage file from the NameNode and merging it locally. The new image is then again updated back to the active NameNode.

**BackupNode:**

Backup Node also provides check pointing functionality like that of the checkpoint node but it also maintains its up-to-date in-memory copy of the file system namespace that is in sync with the active NameNode.

**3. What is commodity hardware?** [Click here to Tweet](http://ctt.ec/15x4e)

Commodity Hardware refers to inexpensive systems that do not have high availability or high quality. Commodity Hardware consists of RAM because there are specific services that need to be executed on RAM. Hadoop can be run on any commodity hardware and does not require any super computer s or high end hardware configuration to execute jobs.

**4. What is the port number for NameNode, Task Tracker and Job Tracker?** [Click here to Tweet](http://ctt.ec/wf0Uv)

NameNode 50070

Job Tracker 50030

Task Tracker 50060

**5. Explain about the process of inter cluster data copying.** [Click here to Tweet](http://ctt.ec/3XL43)

HDFS provides a distributed data copying facility through the DistCP from source to destination. If this data copying is within the hadoop cluster then it is referred to as inter cluster data copying. DistCP requires both source and destination to have a compatible or same version of hadoop.

**6. How can you overwrite the replication factors in HDFS?** [Click here to Tweet](http://ctt.ec/xu7YN)

The replication factor in HDFS can be modified or overwritten in 2 ways-

1)Using the Hadoop FS Shell, replication factor can be changed per file basis using the below command-

$hadoop fs –setrep –w 2 /my/test\_file (test\_file is the filename whose replication factor will be set to 2)

2)Using the Hadoop FS Shell, replication factor of all files under a given directory can be modified using the below command-

3)$hadoop fs –setrep –w 5 /my/test\_dir (test\_dir is the name of the directory and all the files in this directory will have a replication factor set to 5)

Attend a [Hadoop Interview](https://www.dezyre.com/hadoop-interview-questions-preparation/45?from=top100hadoopinterviewblog) session with experts from the industry!

**7. Explain the difference between NAS and HDFS.** [Click here to Tweet](http://ctt.ec/4ta3x)

* NAS runs on a single machine and thus there is no probability of data redundancy whereas HDFS runs on a cluster of different machines thus there is data redundancy because of the replication protocol.
* NAS stores data on a dedicated hardware whereas in HDFS all the data blocks are distributed across local drives of the machines.
* In NAS data is stored independent of the computation and hence Hadoop MapReduce cannot be used for processing whereas HDFS works with Hadoop MapReduce as the computations in HDFS are moved to data.

What technologies are you working on currently? (Java, Datawarehouse, Business Intelligence, ETL, etc.)



**8. Explain what happens if during the PUT operation, HDFS block is assigned a replication factor 1 instead of the default value 3.** [Click here to Tweet](http://ctt.ec/dxU2s)

Replication factor is a property of HDFS that can be set accordingly for the entire cluster to adjust the number of times the blocks are to be replicated to ensure high data availability. For every block that is stored in HDFS, the cluster will have n-1 duplicated blocks. So, if the replication factor during the PUT operation is set to 1 instead of the default value 3, then it will have a single copy of data. Under these circumstances when the replication factor is set to 1 ,if the DataNode crashes under any circumstances, then only single copy of the data would be lost.

**9. What is the process to change the files at arbitrary locations in HDFS?** [Click here to Tweet](http://ctt.ec/OGf26)

HDFS does not support modifications at arbitrary offsets in the file or multiple writers but files are written by a single writer in append only format i.e. writes to a file in HDFS are always made at the end of the file.

**10. Explain about the indexing process in HDFS.** [Click here to Tweet](http://ctt.ec/y88ua)

Indexing process in HDFS depends on the block size. HDFS stores the last part of the data that further points to the address where the next part of data chunk is stored.

**11. What is a rack awareness and on what basis is data stored in a rack?** [Click here to Tweet](http://ctt.ec/157VY)

All the data nodes put together form a storage area i.e. the physical location of the data nodes is referred to as Rack in HDFS. The rack information i.e. the rack id of each data node is acquired by the NameNode. The process of selecting closer data nodes depending on the rack information is known as Rack Awareness.

The contents present in the file are divided into data block as soon as the client is ready to load the file into the hadoop cluster. After consulting with the NameNode, client allocates 3 data nodes for each data block. For each data block, there exists 2 copies in one rack and the third copy is present in another rack. This is generally referred to as the Replica Placement Policy.

**12. What happens to a NameNode that has no data?**

There does not exist any NameNode without data. If it is a NameNode then it should have some sort of data in it.

**13. What happens when a user submits a Hadoop job when the NameNode is down- does the job get in to hold or does it fail.**

The Hadoop job fails when the NameNode is down.

**14. What happens when a user submits a Hadoop job when the Job Tracker is down- does the job get in to hold or does it fail.**

The Hadoop job fails when the Job Tracker is down.

**15. Whenever a client submits a hadoop job, who receives it?**

NameNode receives the Hadoop job which then looks for the data requested by the client and provides the block information. JobTracker takes care of resource allocation of the hadoop job to ensure timely completion.

We have further categorized Hadoop HDFS Interview Questions for Freshers and Experienced-

* Hadoop Interview Questions and Answers for Freshers - Q.Nos- 2,3,7,9,10,11,13,14
* Hadoop Interview Questions and Answers for Experienced - Q.Nos- 1,2, 4,5,6,7,8,12,15

Here are few more frequently asked [Hadoop HDFS Interview Questions and Answers for Freshers and Experienced](https://www.dezyre.com/article/hdfs-interview-questions-and-answers-for-2016/263)

[Click here to know more about our IBM Certified Hadoop Developer course](https://www.dezyre.com/Hadoop-Training-online/19)

**Hadoop MapReduce Interview Questions and Answers**

**1. Explain the usage of Context Object.** [Click here to Tweet](http://ctt.ec/y_NYW)

Context Object is used to help the mapper interact with other Hadoop systems. Context Object can be used for updating counters, to report the progress and to provide any application level status updates. ContextObject has the configuration details for the job and also interfaces, that helps it to generating the output.

**2. What are the core methods of a Reducer?** [Click here to Tweet](http://ctt.ec/e2_Z3)

The 3 core methods of a reducer are –

1)setup () – This method of the reducer is used for configuring various parameters like the input data size, distributed cache, heap size, etc.

Function Definition- *public void setup (context)*

2)reduce () it is heart of the reducer which is called once per key with the associated reduce task.

Function Definition -public void reduce (Key,Value,context)

3)cleanup () - This method is called only once at the end of reduce task for clearing all the temporary files.

Function Definition -public void cleanup (context)

**3. Explain about the partitioning, shuffle and sort phase** [Click here to Tweet](http://ctt.ec/cNUfx)

**Shuffle Phase-**Once the first map tasks are completed, the nodes continue to perform several other map tasks and also exchange the intermediate outputs with the reducers as required. This process of moving the intermediate outputs of map tasks to the reducer is referred to as Shuffling.

**Sort Phase**- Hadoop MapReduce automatically sorts the set of intermediate keys on a single node before they are given as input to the reducer.

**Partitioning Phase-**The process that determines which intermediate keys and value will be received by each reducer instance is referred to as partitioning. The destination partition is same for any key irrespective of the mapper instance that generated it.

**4. How to write a custom partitioner for a Hadoop MapReduce job?** [Click here to Tweet](http://ctt.ec/dKygh)

Steps to write a Custom Partitioner for a Hadoop MapReduce Job-

* A new class must be created that extends the pre-defined Partitioner Class.
* getPartition method of the Partitioner class must be overridden.
* The custom partitioner to the job can be added as a config file in the wrapper which runs Hadoop MapReduce or the custom partitioner can be added to the job by using the set method of the partitioner class.

We have further categorized Hadoop MapReduce Interview Questions for Freshers and Experienced-

* Hadoop Interview Questions and Answers for Freshers - Q.Nos- 2
* Hadoop Interview Questions and Answers for Experienced - Q.Nos- 1,3,4,

Here are a few more frequently asked [Hadoop MapReduce Interview Questions and Answers](https://www.dezyre.com/article/mapreduce-interview-questions-and-answers-for-2016/248)

**Hadoop HBase Interview Questions and Answers**

**1. When should you use HBase and what are the key components of HBase?**

HBase should be used when the big data application has –

1)A variable schema

2)When data is stored in the form of collections

3)If the application demands key based access to data while retrieving.

Key components of HBase are –

Region- This component contains memory data store and Hfile.

Region Server-This monitors the Region.

HBase Master-It is responsible for monitoring the region server.

Zookeeper- It takes care of the coordination between the HBase Master component and the client.

Catalog Tables-The two important catalog tables are ROOT and META.ROOT table tracks where the META table is and META table stores all the regions in the system.

**2. What are the different operational commands in HBase at record level and table level?**

Record Level Operational Commands in HBase are –put, get, increment, scan and delete.

Table Level Operational Commands in HBase are-describe, list, drop, disable and scan.

**3. What is Row Key?**

Every row in an HBase table has a unique identifier known as RowKey. It is used for grouping cells logically and it ensures that all cells that have the same RowKeys are co-located on the same server. RowKey is internally regarded as a byte array.

**4. Explain the difference between RDBMS data model and HBase data model.**

RDBMS is a schema based database whereas HBase is schema less data model.

RDBMS does not have support for in-built partitioning whereas in HBase there is automated partitioning.

RDBMS stores normalized data whereas HBase stores de-normalized data.

**5. Explain about the different catalog tables in HBase?**

The two important catalog tables in HBase, are ROOT and META. ROOT table tracks where the META table is and META table stores all the regions in the system.

**6. What is column families? What happens if you alter the block size of ColumnFamily on an already populated database?**

The logical deviation of data is represented through a key known as column Family. Column families consist of the basic unit of physical storage on which compression features can be applied. In an already populated database, when the block size of column family is altered, the old data will remain within the old block size whereas the new data that comes in will take the new block size. When compaction takes place, the old data will take the new block size so that the existing data is read correctly.

**7. Explain the difference between HBase and Hive.**

HBase and Hive both are completely different hadoop based technologies-Hive is a data warehouse infrastructure on top of Hadoop whereas HBase is a NoSQL key value store that runs on top of Hadoop. Hive helps SQL savvy people to run MapReduce jobs whereas HBase supports 4 primary operations-put, get, scan and delete. HBase is ideal for real time querying of big data where Hive is an ideal choice for analytical querying of data collected over period of time.

**8. Explain the process of row deletion in HBase.**

On issuing a delete command in HBase through the HBase client, data is not actually deleted from the cells but rather the cells are made invisible by setting a tombstone marker. The deleted cells are removed at regular intervals during compaction.

**9. What are the different types of tombstone markers in HBase for deletion?**

There are 3 different types of tombstone markers in HBase for deletion-

1)Family Delete Marker- This markers marks all columns for a column family.

2)Version Delete Marker-This marker marks a single version of a column.

3)Column Delete Marker-This markers marks all the versions of a column.

**10. Explain about HLog and WAL in HBase.**

All edits in the HStore are stored in the HLog. Every region server has one HLog. HLog contains entries for edits of all regions performed by a particular Region Server.WAL abbreviates to Write Ahead Log (WAL) in which all the HLog edits are written immediately.WAL edits remain in the memory till the flush period in case of deferred log flush.

We have further categorized Hadoop HBase Interview Questions for Freshers and Experienced-

* Hadoop Interview Questions and Answers for Freshers - Q.Nos-1,2,4,5,7
* Hadoop Interview Questions and Answers for Experienced - Q.Nos-2,3,6,8,9,10

Here are few more [HBase Interview Questions and Answers](https://www.dezyre.com/article/hbase-interview-questions-and-answers-for-2016/281" \o "HBase Interview Questions and Answers" \t "_blank)

**Hadoop Sqoop Interview Questions and Answers**

**1. Explain about some important Sqoop commands other than import and export.**

**Create Job (--create)**

Here we are creating a job with the name my job, which can import the table data from RDBMS table to HDFS. The following command is used to create a job that is importing data from the employee table in the db database to the HDFS file.

$ Sqoop job --create myjob \

--import \

--connect jdbc:mysql://localhost/db \

--username root \

--table employee --m 1

**Verify Job (--list)**

‘--list’ argument is used to verify the saved jobs. The following command is used to verify the list of saved Sqoop jobs.

$ Sqoop job --list

**Inspect Job (--show)**

‘--show’ argument is used to inspect or verify particular jobs and their details. The following command and sample output is used to verify a job called myjob.

$ Sqoop job --show myjob

**Execute Job (--exec)**

‘--exec’ option is used to execute a saved job. The following command is used to execute a saved job called myjob.

$ Sqoop job --exec myjob

**2. How Sqoop can be used in a Java program?**

The Sqoop jar in classpath should be included in the java code. After this the method Sqoop.runTool () method must be invoked. The necessary parameters should be created to Sqoop programmatically just like for command line.

**3. What is the process to perform an incremental data load in Sqoop?**

The process to perform incremental data load in Sqoop is to synchronize the modified or updated data (often referred as delta data) from RDBMS to Hadoop. The delta data can be facilitated through the incremental load command in Sqoop.

Incremental load can be performed by using Sqoop import command or by loading the data into hive without overwriting it. The different attributes that need to be specified during incremental load in Sqoop are-

1)Mode (incremental) –The mode defines how Sqoop will determine what the new rows are. The mode can have value as Append or Last Modified.

2)Col (Check-column) –This attribute specifies the column that should be examined to find out the rows to be imported.

3)Value (last-value) –This denotes the maximum value of the check column from the previous import operation.

**4. Is it possible to do an incremental import using Sqoop?**

Yes, Sqoop supports two types of incremental imports-

1)Append

2)Last Modified

To insert only rows Append should be used in import command and for inserting the rows and also updating Last-Modified should be used in the import command.

**5. What is the standard location or path for Hadoop Sqoop scripts?**

/usr/bin/Hadoop Sqoop

**6. How can you check all the tables present in a single database using Sqoop?**

The command to check the list of all tables present in a single database using Sqoop is as follows-

**Sqoop list-tables –connect jdbc: mysql: //localhost/user;**

**7. How are large objects handled in Sqoop?**

Sqoop provides the capability to store large sized data into a single field based on the type of data. Sqoop supports the ability to store-

1)CLOB ‘s – Character Large Objects

2)BLOB’s –Binary Large Objects

Large objects in Sqoop are handled by importing the large objects into a file referred as “LobFile” i.e. Large Object File. The LobFile has the ability to store records of huge size, thus each record in the LobFile is a large object.

**8. Can free form SQL queries be used with Sqoop import command? If yes, then how can they be used?**

Sqoop allows us to use free form SQL queries with the import command. The import command should be used with the –e and – query options to execute free form SQL queries. When using the –e and –query options with the import command the –target dir value must be specified.

**9. Differentiate between Sqoop and distCP.**

DistCP utility can be used to transfer data between clusters whereas Sqoop can be used to transfer data only between Hadoop and RDBMS.

**10. What are the limitations of importing RDBMS tables into Hcatalog directly?**

There is an option to import RDBMS tables into Hcatalog directly by making use of –hcatalog –database option with the –hcatalog –table but the limitation to it is that there are several arguments like –as-avrofile , -direct, -as-sequencefile, -target-dir , -export-dir are not supported.

We have further categorized Hadoop Sqoop Interview Questions for Freshers and Experienced-

* Hadoop Interview Questions and Answers for Freshers - Q.Nos- 4,5,6,9
* Hadoop Interview Questions and Answers for Experienced - Q.Nos- 1,2,3,6,7,8,10

Here are few more frequently asked [Sqoop Interview Questions and Answers for Freshers and Experienced](https://www.dezyre.com/article/sqoop-interview-questions-and-answers-for-2016/274)

**Hadoop Flume Interview Questions and Answers**

**1) Explain about the core components of Flume.**

The core components of Flume are –

Event- The single log entry or unit of data that is transported.

Source- This is the component through which data enters Flume workflows.

Sink-It is responsible for transporting data to the desired destination.

Channel- it is the duct between the Sink and Source.

Agent- Any JVM that runs Flume.

Client- The component that transmits event to the source that operates with the agent.

**2) Does Flume provide 100% reliability to the data flow?**

Yes, Apache Flume provides end to end reliability because of its transactional approach in data flow.

**3) How can Flume be used with HBase?**

Apache Flume can be used with HBase using one of the two HBase sinks –

* HBaseSink (org.apache.flume.sink.hbase.HBaseSink) supports secure HBase clusters and also the novel HBase IPC that was introduced in the version HBase 0.96.
* AsyncHBaseSink (org.apache.flume.sink.hbase.AsyncHBaseSink) has better performance than HBase sink as it can easily make non-blocking calls to HBase.

**Working of the HBaseSink –**

In HBaseSink, a Flume Event is converted into HBase Increments or Puts. Serializer implements the HBaseEventSerializer which is then instantiated when the sink starts. For every event, sink calls the initialize method in the serializer which then translates the Flume Event into HBase increments and puts to be sent to HBase cluster.

**Working of the AsyncHBaseSink-**

AsyncHBaseSink implements the AsyncHBaseEventSerializer. The initialize method is called only once by the sink when it starts. Sink invokes the setEvent method and then makes calls to the getIncrements and getActions methods just similar to HBase sink. When the sink stops, the cleanUp method is called by the serializer.

**4) Explain about the different channel types in Flume. Which channel type is faster?**

The 3 different built in channel types available in Flume are-

MEMORY Channel – Events are read from the source into memory and passed to the sink.

JDBC Channel – JDBC Channel stores the events in an embedded Derby database.

FILE Channel –File Channel writes the contents to a file on the file system after reading the event from a source. The file is deleted only  after the contents are successfully delivered to the sink.

MEMORY Channel is the fastest channel among the three however has the risk of data loss. The channel that you choose completely depends on the nature of the big data application and the value of each event.

**5) Which is the reliable channel in Flume to ensure that there is no data loss?**

FILE Channel is the most reliable channel among the 3 channels JDBC, FILE and MEMORY.

**6) Explain about the replication and multiplexing selectors in Flume.**

Channel Selectors are used to handle multiple channels. Based on the Flume header value, an event can be written just to a single channel or to multiple channels. If a channel selector is not specified to the source then by default it is the Replicating selector. Using the replicating selector, the same event is written to all the channels in the source’s channels list. Multiplexing channel selector is used when the application has to send different events to different channels.

**7) How multi-hop agent can be setup in Flume?**

Avro RPC Bridge mechanism is used to setup Multi-hop agent in Apache Flume.

**8) Does Apache Flume provide support for third party plug-ins?**

Most of the data analysts use Apache Flume has plug-in based architecture as it can load data from external sources and transfer it to external destinations.

**9) Is it possible to leverage real time analysis on the big data collected by Flume directly? If yes, then explain how.**

Data from Flume can be extracted, transformed and loaded in real-time into Apache Solr servers using *MorphlineSolrSink*

**10) Differentiate between FileSink and FileRollSink**

The major difference between HDFS FileSink and FileRollSink is that HDFS File Sink writes the events into the Hadoop Distributed File System (HDFS) whereas File Roll Sink stores the events into the local file system.

**Hadoop Flume Interview Questions and Answers for Freshers - Q.Nos- 1,2,4,5,6,10**

**Hadoop Flume Interview Questions and Answers for Experienced- Q.Nos-  3,7,8,9**

**Hadoop Zookeeper Interview Questions and Answers**

**1) Can Apache Kafka be used without Zookeeper?**

It is not possible to use Apache Kafka without Zookeeper because if the Zookeeper is down Kafka cannot serve client request.

**2) Name a few companies that use Zookeeper.**

Yahoo, Solr, Helprace, Neo4j, Rackspace

**3) What is the role of Zookeeper in HBase architecture?**

In HBase architecture, ZooKeeper is the monitoring server that provides different services like –tracking server failure and network partitions, maintaining the configuration information, establishing communication between the clients and region servers, usability of ephemeral nodes to identify the available servers in the cluster.

**4) Explain about ZooKeeper in Kafka**

Apache Kafka uses ZooKeeper to be a highly distributed and scalable system. Zookeeper is used by Kafka to store various configurations and use them across the hadoop cluster in a distributed manner. To achieve distributed-ness, configurations are distributed and replicated throughout the leader and follower nodes in the ZooKeeper ensemble. We cannot directly connect to Kafka by bye-passing ZooKeeper because if the ZooKeeper is down it will not be able to serve the client request.

**5) Explain how Zookeeper works**

ZooKeeper is referred to as the King of Coordination and distributed applications use ZooKeeper to store and facilitate important configuration information updates. ZooKeeper works by coordinating the processes of distributed applications. ZooKeeper is a robust replicated synchronization service with eventual consistency. A set of nodes is known as an ensemble and persisted data is distributed between multiple nodes.

3 or more independent servers collectively form a ZooKeeper cluster and elect a master. One client connects to any of the specific server and migrates if a particular node fails. The ensemble of ZooKeeper nodes is alive till the majority of nods are working. The master node in ZooKeeper is dynamically selected by the consensus within the ensemble so if the master node fails then the role of master node will migrate to another node which is selected dynamically. Writes are linear and reads are concurrent in ZooKeeper.

**6) List some examples of Zookeeper use cases.**

* Found by Elastic uses Zookeeper comprehensively for resource allocation, leader election, high priority notifications and discovery. The entire service of Found built up of various systems that read and write to   Zookeeper.
* Apache Kafka that depends on ZooKeeper is used by LinkedIn
* Storm that relies on ZooKeeper is used by popular companies like Groupon and Twitter.

**7) How to use Apache Zookeeper command line interface?**

ZooKeeper has a command line client support for interactive use. The command line interface of ZooKeeper is similar to the file and shell system of UNIX. Data in ZooKeeper is stored in a hierarchy of Znodes where each znode can contain data just similar to a file. Each znode can also have children just like directories in the UNIX file system.

Zookeeper-client command is used to launch the command line client. If the initial prompt is hidden by the log messages after entering the command, users can just hit ENTER to view the prompt.

**8) What are the different types of Znodes?**

There are 2 types of Znodes namely- Ephemeral and Sequential Znodes.

* The Znodes that get destroyed as soon as the client that created it disconnects are referred to as Ephemeral Znodes.
* Sequential Znode is the one in which sequential number is chosen by the ZooKeeper ensemble and is pre-fixed when the client assigns name to the znode.

**9) What are watches?**

Client disconnection might be troublesome problem especially when we need to keep a track on the state of Znodes at regular intervals. ZooKeeper has an event system referred to as watch which can be set on Znode to trigger an event whenever it is removed, altered or any new children are created below it.

**10) What problems can be addressed by using Zookeeper?**

In the development of distributed systems, creating own protocols for coordinating the hadoop cluster results in failure and frustration for the developers. The architecture of a distributed system can be prone to deadlocks, inconsistency and race conditions. This leads to various difficulties in making the hadoop cluster fast, reliable and scalable. To address all such problems, Apache ZooKeeper can be used as a coordination service to write correct distributed applications without having to reinvent the wheel from the beginning.

**Hadoop ZooKeeper Interview Questions and Answers for Freshers - Q.Nos- 1,2,8,9**

**Hadoop ZooKeeper Interview Questions and Answers for Experienced- Q.Nos-3,4,5,6,7, 10**

**Hadoop Pig Interview Questions and Answers**

**1) What are different modes of execution in Apache Pig?**

Apache Pig runs in 2 modes- one is the “Pig (Local Mode) Command Mode” and the other is the “Hadoop MapReduce (Java) Command Mode”. Local Mode requires access to only a single machine where all files are installed and executed on a local host whereas MapReduce requires accessing the Hadoop cluster.

**2) Explain about co-group in Pig.**

COGROUP operator in Pig is used to work with multiple tuples. COGROUP operator is applied on statements that contain or involve two or more relations. The COGROUP operator can be applied on up to 127 relations at a time. When using the COGROUP operator on two tables at once-Pig first groups both the tables and after that joins the two tables on the grouped columns.

We have further categorized Hadoop Pig Interview Questions for Freshers and Experienced-

* Hadoop Interview Questions and Answers for Freshers - Q.No-1
* Hadoop Interview Questions and Answers for Experienced - Q.No- 2

Here are a few more frequently asked [Pig Hadoop Interview Questions and Answers for Freshers and Experienced](https://www.dezyre.com/article/pig-interview-questions-and-answers-for-2016/244)

**Hadoop Hive Interview Questions and Answers**

**1) Explain about the SMB Join in Hive.**

In SMB join in Hive, each mapper reads a bucket from the first table and the corresponding bucket from the second table and then a merge sort join is performed. Sort Merge Bucket (SMB) join in hive is mainly used as there is no limit on file or partition or table join. SMB join can best be used when the tables are large. In SMB join the columns are bucketed and sorted using the join columns. All tables should have the same number of buckets in SMB join.

**2) How can you connect an application, if you run Hive as a server?**

When running Hive as a server, the application can be connected in one of the 3 ways-

ODBC Driver-This supports the ODBC protocol

JDBC Driver- This supports the JDBC protocol

Thrift Client- This client can be used to make calls to all hive commands using different programming language like PHP, Python, Java, C++ and Ruby.

**3) What does the overwrite keyword denote in Hive load statement?**

Overwrite keyword in Hive load statement deletes the contents of the target table and replaces them with the files referred by the file path i.e. the files that are referred by the file path will be added to the table when using the overwrite keyword.

**4) What is SerDe in Hive? How can you write your own custom SerDe?**

SerDe is a Serializer DeSerializer. Hive uses SerDe to read and write data from tables. Generally, users prefer to write a Deserializer instead of a SerDe as they want to read their own data format rather than writing to it. If the SerDe supports DDL i.e. basically SerDe with parameterized columns and different column types, the users can implement a Protocol based DynamicSerDe rather than writing the SerDe from scratch.

We have further categorized Hadoop Hive Interview Questions for Freshers and Experienced-

Hadoop Hive Interview Questions and Answers for Freshers- Q.Nos-3

Hadoop Hive Interview Questions and Answers for Experienced- Q.Nos-1,2,4

Here are a few more frequently asked  [Hadoop Hive Interview Questions and Answers for Freshers and Experienced](https://www.dezyre.com/article/hive-interview-questions-and-answers-for-2016/246)

**Hadoop YARN Interview Questions and Answers**

**1)What are the stable versions of Hadoop?**

Release 2.7.1 (stable)

Release 2.4.1

Release 1.2.1 (stable)

**2) What is Apache Hadoop YARN?**

YARN is a powerful and efficient feature rolled out as a part of Hadoop 2.0.YARN is a large scale distributed system for running big data applications.

**3) Is YARN a replacement of Hadoop MapReduce?**

YARN is not a replacement of Hadoop but it is a more powerful and efficient technology that supports MapReduce and is also referred to as Hadoop 2.0 or MapReduce 2.

**4) What are the additional benefits YARN brings in to Hadoop?**

* Effective utilization of the resources as multiple applications can be run in YARN all sharing a common resource.In Hadoop MapReduce there are seperate slots for Map and Reduce tasks whereas in YARN there is no fixed slot. The same container can be used for Map and Reduce tasks leading to better utilization.
* YARN is backward compatible so all the existing MapReduce jobs.
* Using YARN, one can even run applications that are not based on the MaReduce model

**5)  How can native libraries be included in YARN jobs?**

There are two ways to include native libraries in YARN jobs-

1)  By setting the -Djava.library.path on the command line  but in this case there are chances that the native libraries might not be loaded correctly and there is possibility of errors.

2) The better option to include native libraries is to the set the LD\_LIBRARY\_PATH in the .bashrc file.

**6) Explain the differences between Hadoop 1.x and Hadoop 2.x**

* In Hadoop 1.x, MapReduce is responsible for both processing and cluster management whereas in Hadoop 2.x processing is taken care of by other processing models and YARN is responsible for cluster management.
* Hadoop 2.x scales better when compared to Hadoop 1.x with close to 10000 nodes per cluster.
* Hadoop 1.x has single point of failure problem and whenever the NameNode fails it has to be recovered manually. However, in case of Hadoop 2.x StandBy NameNode overcomes the SPOF problem and whenever the NameNode fails it is configured for automatic recovery.
* Hadoop 1.x works on the concept of slots whereas Hadoop 2.x works on the concept of containers and can also run generic tasks.

**7) What are the core changes in Hadoop 2.0?**

Hadoop 2.x provides an upgrade to Hadoop 1.x in terms of resource management, scheduling and the manner in which execution occurs. In Hadoop 2.x the cluster resource management capabilities work in isolation from the MapReduce specific programming logic. This helps Hadoop to share resources dynamically between multiple parallel processing frameworks like Impala and the core MapReduce component. Hadoop 2.x Hadoop 2.x allows workable and fine grained resource configuration leading to efficient and better cluster utilization so that the application can scale to process larger number of jobs.

**8) Differentiate between NFS, Hadoop NameNode and JournalNode.**

HDFS is a write once file system so a user cannot update the files once they exist either they can read or write to it. However, under certain scenarios in the enterprise environment like file uploading, file downloading, file browsing or data streaming –it is not possible to achieve all this using the standard HDFS. This is where a distributed file system protocol Network File System (NFS) is used. NFS allows access to files on remote machines just similar to how local file system is accessed by applications.

Namenode is the heart of the HDFS file system that maintains the metadata and tracks where the file data is kept across the Hadoop cluster.

StandBy Nodes and Active Nodes communicate with a group of light weight nodes to keep their state synchronized. These are known as Journal Nodes.

**9) What are the modules that constitute the Apache Hadoop 2.0 framework?**

Hadoop 2.0 contains four important modules of which 3 are inherited from Hadoop 1.0 and a new module YARN is added to it.

1. Hadoop Common – This module consists of all the basic utilities and libraries that required by other modules.
2. HDFS- Hadoop Distributed file system that stores huge volumes of data on commodity machines across the cluster.
3. MapReduce- Java based programming model for data processing.
4. YARN- This is a new module introduced in Hadoop 2.0 for cluster resource management and job scheduling.

[CLICK HERE](https://www.dezyre.com/article/hadoop-2-0-yarn-framework-the-gateway-to-easier-programming-for-hadoop-users/84) to read more about the YARN module in Hadoop 2.x.

**10) How is the distance between two nodes defined in Hadoop?**

Measuring bandwidth is difficult in Hadoop so network is denoted as a tree in Hadoop. The distance between two nodes in the tree plays a vital role in forming a Hadoop cluster  and is defined by the network topology and java interface DNStoSwitchMapping. The distance is equal to the sum of the distance to the closest common ancestor of both the nodes. The method getDistance(Node node1, Node node2) is used to calculate the distance between two nodes with the assumption that the distance from a node to its parent node is always 1.

We have further categorized Hadoop YARN Interview Questions for Freshers and Experienced-

* Hadoop Interview Questions and Answers for Freshers - Q.Nos- 2,3,4,6,7,9
* Hadoop Interview Questions and Answers for Experienced - Q.Nos- 1,5,8,10

What other questions do you have regarding your Hadoop career?



**Hadoop Interview FAQ’s – An Interviewee Should Ask an Interviewer**

For many hadoop job seekers, the question from the interviewer – “Do you have any questions for me?” indicates the end of a Hadoop developer job interview. It is always enticing for a Hadoop job seeker to immediately say “No” to the question for the sake of keeping the first impression intact.However, to land a hadoop job or any other job, it is always preferable to fight that urge and ask relevant questions to the interviewer.

Asking questions related to the Hadoop technology implementation, shows your interest in the open hadoop job role and also conveys your interest in working with the company.Just like any other interview, even hadoop interviews are a two-way street- it helps the interviewer decide whether you have the desired hadoop skills they in are looking for in a hadoop developer, and helps an interviewee decide if that is the kind of big data infrastructure and hadoop technology implementation you want to devote your skills for foreseeable future growth in the big data domain.

Candidates should not be afraid to ask questions to the interviewer. To ease this for hadoop job seekers, DeZyre has collated few hadoop interview FAQ’s that every candidate should ask an interviewer during their next hadoop job interview-

**1) What is the size of the biggest hadoop cluster a company X operates?**

Asking this question helps a hadoop job seeker understand the hadoop maturity curve at a company.Based on the answer of the interviewer, a candidate can judge how much an organization invests in Hadoop and their enthusiasm to buy big data products from various vendors. The candidate can also get an idea on the hiring needs of the company based on their hadoop infrastructure.

**2) For what kind of big data problems, did the organization choose to use Hadoop?**

Asking this question to the interviewer shows the candidates keen interest in understanding the reason for hadoop implementation from a business perspective. This question gives the impression to the interviewer that the candidate is not merely interested in the hadoop developer job role but is also interested in the growth of the company.

**3) Based on the answer to question no 1, the candidate can ask the interviewer why the hadoop infrastructure is configured in that particular way, why the company chose to use the selected big data tools and how workloads are constructed in the hadoop environment.**

Asking this question to the interviewer gives the impression that you are not just interested in maintaining the big data system and developing products around it but are also seriously thoughtful on how the infrastructure can be improved to help business growth and make cost savings.

**4) What kind of data the organization works with or what are the HDFS file formats the company uses?**

The question gives the candidate an idea on the kind of big data he or she will be handling if selected for the hadoop developer job role. Based on the data, it gives an idea on the kind of analysis they will be required to perform on the data.

**5) What is the most complex problem the company is trying to solve using Apache Hadoop?**

Asking this question helps the candidate know more about the upcoming projects he or she might have to work and what are the challenges around it. Knowing this beforehand helps the interviewee prepare on his or her areas of weakness.

**6) Will I get an opportunity to attend big data conferences? Or will the organization incur any costs involved in taking advanced hadoop or big data certification?**

This is a very important question that you should be asking these the interviewer. This helps a candidate understand whether the prospective hiring manager is interested and supportive when it comes to professional development of the employee.

**Hadoop Interview FAQ’s – Interviewer Asks an Interviewee**

So, you have cleared the technical interview after preparing thoroughly with the help of the Hadoop Interview Questions shared by DeZyre. After an in-depth technical interview, the interviewer might still not be satisfied and would like to test your practical experience in navigating and analysing big data. The expectation of the interviewer is to judge whether you are really interested in the open position and ready to work with the company, regardless of the technical knowledge you have on hadoop technology.

There are quite a few on-going debates in the hadoop community, on the advantages of the various components in the hadoop ecosystem-- for example what is better MapReduce, Pig or Hive or Spark vs. Hadoop or when should a company use MapReduce over other alternative? etc. Interviewee and Interviewer should both be ready to answer such hadoop interview FAQs, as there is no right or wrong answer to these questions.The best possible way to answer these Hadoop interview FAQs is to explain why a particular interviewee favours an option. Answering these hadoop interview FAQs with practical examples as to why the candidate favours an option, demonstrates his or her understanding of the business needs and helps the interviewer judge the flexibility of the candidate to use various big data tools in the hadoop ecosystem.

Here are a few hadoop interview FAQs that are likely to be asked by the interviewer-

**1) If you are an experienced hadoop professional then you are likely to be asked questions like  –**

* **The number of nodes you have worked with in a cluster.**
* **Which hadoop distribution have you used in your recent project.**
* **Your experience on working with special configurations like High Availability.**
* **The data volume you have worked with in your most recent project.**
* **What are the various tools you used in the big data and hadoop projects you have worked on?**

Your answer to these interview questions will help the interviewer understand your expertise in Hadoop based on the size of the hadoop cluster and number of nodes. Based on the highest volume of data you have handled in your previous projects, interviewer can assess your overall experience in debugging and troubleshooting issues involving huge hadoop clusters.

The number of tools you have worked with help an interviewer judge that you are aware of the overall hadoop ecosystem and not just MapReduce. To be selected, it all depends on how well you communicate the answers to all these questions.

**2) What are the challenges that you faced when implementing hadoop projects?**

Stay Tuned to the blog for more updates on Hadoop Interview FAQ's!!!

We hope that these Hadoop Interview Questions and Answers have pre-charged you for your next Hadoop Interview.Get the Ball Rolling and share your hadoop interview experiences in the comments below.Please do! It's all part of our shared mission to ease Hadoop Interviews for all prospective Hadoopers.We invite you to get involved.

[Become a Certified Hadoop Developer in 30 Days](https://www.dezyre.com/Hadoop-Training-online/19)

**Related Posts**

[How much Java is required to learn Hadoop?](https://www.dezyre.com/article/-how-much-java-is-required-to-learn-hadoop/103)

[Top 50 Hadoop Interview Questio](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92)

**[Answers](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)**

**[Currently have 142 answers](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)**

**[Q: What companies are you applying to for Hadoop job roles?](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)**

[[](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92)](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)

**[Himani](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)** [Facebook   
Nov 01 2016, 09:04 AM](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)

[[](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92)](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)

**[Jabeen Fatima](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)** [Tech Mahindra   
Oct 28 2016, 09:48 PM](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)

[[](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92)](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)

**[Hadoop developer](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)** [Digimatrix solutions pvt ltd   
Oct 28 2016, 02:01 PM](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)

[[](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92)](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)

**[Kalyani](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)** [Cognizant   
Oct 26 2016, 01:28 PM](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)

[[](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92)](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)

**[anand kishore](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)** [I AM SEEKING A JOB OF DATA ANALYST IN COMPANIES LIKE ALTER RESOURCE AND GLOBAL ANALYTICS   
Oct 21 2016, 02:57 PM](https://www.dezyre.com/article/top-50-hadoop-interview-questions-for-2016/92" \t "_blank)

View 59 more answers

**Q: What technologies are you working on currently? (Java, Datawarehouse, Business Intelligence, ETL, etc.)**



**suresh** Java/J2ee   
Nov 04 2016, 11:00 PM



**surya** datawarehouse   
Nov 03 2016, 12:29 AM



**Kondalu** java   
Oct 31 2016, 12:10 PM



**adddgfdh** datawarehouse   
Oct 22 2016, 10:40 PM



**Mohini** Java, Machine Learning   
Oct 21 2016, 02:54 PM

View 69 more answers

**Q: What other questions do you have regarding your Hadoop career?**



**Kalu** How good is Hadoo pcareer   
Aug 02 2016, 05:19 PM



**Sachin** How can Spark be used with Hadoop   
Jul 28 2016, 06:31 PM



**Anonymous** Do I really need Java to sit for these interviews since I am not so proficient in this?   
Apr 15 2016, 03:25 PM



**Anonymous** What hands-on projects should I be working on to get the right jobs.   
Apr 15 2016, 03:25 PM

Follow

[comments powered by Disqus](http://disqus.com/)

**Upcoming Live Online Hadoop Training**

**06**

**Nov**

**Sun to Thurs (3 weeks)**

**6:30 PM - 8:30 PM PST**

**$399**

[LEARN MORE](https://www.dezyre.com/Hadoop-Training-online/19)

**16**

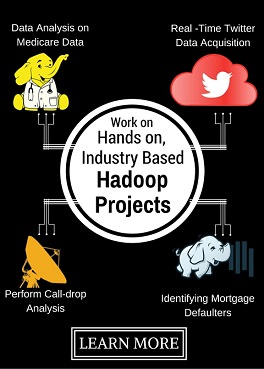
**Nov**

**Sun to Thurs (3 weeks)**

**6:30 PM - 8:30 PM PST**

**$399**

[LEARN MORE](https://www.dezyre.com/Hadoop-Training-online/19)

* [](https://www.dezyre.com/hadoop-hands-on-training/47?utm_source=dezyre&utm_campaign=demo_hadoop_oct29&utm_medium=blog_banner)
* [](https://www.dezyre.com/hackerday?from=blogimage)
* [](https://www.dezyre.com/Hadoop-Training-online/19)

Relevant Courses

* [Hadoop Online Training](https://www.dezyre.com/Hadoop-Training-online/19)
* [Apache Spark Training](https://www.dezyre.com/apache-spark-scala-training-online/38)
* [Data Science in Python Training](https://www.dezyre.com/data-science-in-python-online-training/36)
* [Data Science in R Language Training](https://www.dezyre.com/data-science-in-R-programming-online-training/37)
* [Salesforce Certification Training](https://www.dezyre.com/Salesforce-Certifications-ADM-201-DEV-401/29)
* [NoSQL Database Training](https://www.dezyre.com/NoSQL-Databases/27)
* [Hadoop Admin Training](https://www.dezyre.com/Hadoop-Administration/28)

Blog Categories

* [Big Data](https://www.dezyre.com/blog-category/big-data)
* [CRM](https://www.dezyre.com/blog-category/crm)
* [Data Science](https://www.dezyre.com/blog-category/data-science)
* [Mobile App Development](https://www.dezyre.com/blog-category/mobile-app-development)
* [NoSQL Database](https://www.dezyre.com/blog-category/nosql-database)
* [Web Development](https://www.dezyre.com/blog-category/web-development)

You might also like

* [Recap of Data Science News for October](https://www.dezyre.com/article/recap-of-data-science-news-for-october/316)
* [Recap of Apache Spark News for October](https://www.dezyre.com/article/recap-of-apache-spark-news-for-october/315)
* [Recap of Hadoop News for October](https://www.dezyre.com/article/recap-of-hadoop-news-for-october/314)
* [Scenario-Based Hadoop Interview Questions to prepare for in 2016](https://www.dezyre.com/article/scenario-based-hadoop-interview-questions-to-prepare-for-in-2016/313)
* [Top Data Science Certifications to choose from in 2016](https://www.dezyre.com/article/top-data-science-certifications-to-choose-from-in-2016/312)
* [Best Hadoop Certification: Cloudera vs Hortonworks](https://www.dezyre.com/article/best-hadoop-certification-cloudera-vs-hortonworks/311)
* [Recap of Data Science News for September](https://www.dezyre.com/article/recap-of-data-science-news-for-september/310)
* [Recap of Apache Spark News for September](https://www.dezyre.com/article/recap-of-apache-spark-news-for-september/309)
* [Recap of Hadoop News for September](https://www.dezyre.com/article/recap-of-hadoop-news-for-september/308)
* [How to build a Data Science portfolio that will get you hired?](https://www.dezyre.com/article/how-to-build-a-data-science-portfolio-that-will-get-you-hired/307)

Tutorials

* [Hadoop Online Tutorial – Hadoop HDFS Commands Guide](https://www.dezyre.com/hadoop-tutorial/hadoop-hdfs-commands)
* [MapReduce Tutorial–Learn to implement Hadoop WordCount Example](https://www.dezyre.com/hadoop-tutorial/hadoop-mapreduce-wordcount-tutorial)
* [Hadoop Hive Tutorial-Usage of Hive Commands in HQL](https://www.dezyre.com/hadoop-tutorial/hive-commands)
* [Hive Tutorial-Getting Started with Hive Installation on Ubuntu](https://www.dezyre.com/hadoop-tutorial/install-hive)
* [Learn Java for Hadoop Tutorial: Inheritance and Interfaces](https://www.dezyre.com/hadoop-tutorial/java-for-hadoop-inheritance-and-interfaces)
* [Learn Java for Hadoop Tutorial: Classes and Objects](https://www.dezyre.com/hadoop-tutorial/java-for-hadoop-classes-and-objects)
* [Learn Java for Hadoop Tutorial: Arrays](https://www.dezyre.com/hadoop-tutorial/java-for-hadoop-arrays)
* [Apache Spark Tutorial–Run your First Spark Program](https://www.dezyre.com/apache%20spark-tutorial/spark-tutorial)
* [PySpark Tutorial-Learn to use Apache Spark with Python](https://www.dezyre.com/apache%20spark-tutorial/pyspark-tutorial)
* [R Tutorial- Learn Data Visualization with R using GGVIS](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/ggvis)
* [Neural Network Training Tutorial](https://www.dezyre.com/data%20science%20in%20python-tutorial/neural-network-training-tutorial)
* [Python List Tutorial](https://www.dezyre.com/data%20science%20in%20python-tutorial/python-list-tutorial)
* [MatPlotLib Tutorial](https://www.dezyre.com/data%20science%20in%20python-tutorial/matplotlib-tutorial)
* [Decision Tree Tutorial](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/decision-tree-tutorial)
* [Neural Network Tutorial](https://www.dezyre.com/data%20science%20in%20python-tutorial/neural-network-tutorial)
* [Performance Metrics for Machine Learning Algorithms](https://www.dezyre.com/data%20science%20in%20python-tutorial/performance-metrics-for-machine-learning-algorithm)
* [R Tutorial: Data.Table](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/r-data-table-tutorial)
* [SciPy Tutorial](https://www.dezyre.com/data%20science%20in%20python-tutorial/scipy-introduction-tutorial)
* [Step-by-Step Apache Spark Installation Tutorial](https://www.dezyre.com/apache%20spark-tutorial/apache-spark-installation-tutorial)
* [Introduction to Apache Spark Tutorial](https://www.dezyre.com/apache%20spark-tutorial/tutorial-introduction-to-apache-spark)
* [R Tutorial: Importing Data from Web](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/r-tutorial-importing-data-from-web)
* [R Tutorial: Importing Data from Relational Database](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/r-tutorial-importing-data-from-relational-database)
* [R Tutorial: Importing Data from Excel](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/tutorial-importing-data-from-excel-to-r)
* [Introduction to Machine Learning Tutorial](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/introduction-to-machine-learning-tutorial)
* [Machine Learning Tutorial: Linear Regression](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/linear-regression-tutorial)
* [Machine Learning Tutorial: Logistic Regression](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/logistic-regression-tutorial)
* [Support Vector Machine Tutorial (SVM)](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/support-vector-machine-tutorial)
* [K-Means Clustering Tutorial](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/k-means-clustering-techniques-tutorial)
* [dplyr Manipulation Verbs](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/dplyr-manipulations-verbs)
* [Introduction to dplyr package](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/introduction-to-dplyr-package)
* [Importing Data from Flat Files in R](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/importing-data-from-flat-files-in-r)
* [Principal Component Analysis Tutorial](https://www.dezyre.com/data%20science%20in%20python-tutorial/principal-component-analysis-tutorial)
* [Pandas Tutorial Part-3](https://www.dezyre.com/data%20science%20in%20python-tutorial/pandas-introductory-tutorial-part-3)
* [Pandas Tutorial Part-2](https://www.dezyre.com/data%20science%20in%20python-tutorial/pandas-introductory-tutorial-part-2)
* [Pandas Tutorial Part-1](https://www.dezyre.com/data%20science%20in%20python-tutorial/pandas-introductory-tutorial-part-1)
* [Tutorial- Hadoop Multinode Cluster Setup on Ubuntu](https://www.dezyre.com/hadoop-tutorial/hadoop-multinode-cluster-setup)
* [Data Visualizations Tools in R](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/data-visualizations-tools-r)
* [R Statistical and Language tutorial](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/r-statistical-language-tutorial)
* [Introduction to Data Science with R](https://www.dezyre.com/data%20science%20in%20python-tutorial/introduction-to-data-science-r)
* [Apache Pig Tutorial: User Defined Function Example](https://www.dezyre.com/hadoop-tutorial/apache-pig-tutorial-udf)
* [Apache Pig Tutorial Example: Web Log Server Analytics](https://www.dezyre.com/hadoop-tutorial/pig-tutorial-web-log-server-analytics)
* [Impala Case Study: Web Traffic](https://www.dezyre.com/hadoop-tutorial/impala-case-study-web-traffic)
* [Impala Case Study: Flight Data Analysis](https://www.dezyre.com/hadoop-tutorial/impala-case-study-flight-data-analysis)
* [Hadoop Impala Tutorial](https://www.dezyre.com/hadoop-tutorial/hadoop-impala-tutorial)
* [Apache Hive Tutorial: Tables](https://www.dezyre.com/hadoop-tutorial/apache-hive-tutorial-tables)
* [Flume Hadoop Tutorial: Twitter Data Extraction](https://www.dezyre.com/hadoop-tutorial/flume-hadoop-twitter-data-extraction)
* [Flume Hadoop Tutorial: Website Log Aggregation](https://www.dezyre.com/hadoop-tutorial/flume-hadoop-website-log-aggregation)
* [Hadoop Sqoop Tutorial: Example Data Export](https://www.dezyre.com/hadoop-tutorial/hadoop-sqoop-tutorial-example-data-export)
* [Hadoop Sqoop Tutorial: Example of Data Aggregation](https://www.dezyre.com/hadoop-tutorial/hadoop-sqoop-tutorial-data-aggregation)
* [Apache Zookepeer Tutorial: Example of Watch Notification](https://www.dezyre.com/hadoop-tutorial/apache-zookeeper-watch-notification)
* [Apache Zookepeer Tutorial: Centralized Configuration Management](https://www.dezyre.com/hadoop-tutorial/apache-zookeeper-tutorial-centralize-configuration)
* [Hadoop Zookeeper Tutorial](https://www.dezyre.com/hadoop-tutorial/zookeeper-tutorial)
* [Hadoop Sqoop Tutorial](https://www.dezyre.com/hadoop-tutorial/hadoop-sqoop-tutorial)
* [Hadoop PIG Tutorial](https://www.dezyre.com/hadoop-tutorial/pig-tutorial)
* [Hadoop Oozie Tutorial](https://www.dezyre.com/hadoop-tutorial/oozie-tutorial)
* [Hadoop NoSQL Database Tutorial](https://www.dezyre.com/hadoop-tutorial/nosql-database-tutorial)
* [Hadoop Hive Tutorial](https://www.dezyre.com/hadoop-tutorial/hive-tutorial)
* [Hadoop HDFS Tutorial](https://www.dezyre.com/hadoop-tutorial/hdfs-tutorial)
* [Hadoop hBase Tutorial](https://www.dezyre.com/hadoop-tutorial/hbase-tutorial)
* [Hadoop Flume Tutorial](https://www.dezyre.com/hadoop-tutorial/flume-tutorial)
* [Hadoop 2.0 YARN Tutorial](https://www.dezyre.com/hadoop-tutorial/yarn-tutorial)
* [Hadoop MapReduce Tutorial](https://www.dezyre.com/hadoop-tutorial/hadoop-mapreduce-tutorial-)
* [Big Data Hadoop Tutorial for Beginners- Hadoop Installation](https://www.dezyre.com/hadoop-tutorial/big-data-hadoop-tutorial)

Online Courses

* [Hadoop Training](https://www.dezyre.com/Hadoop-Training-online/19)
* [Spark Certification Training](https://www.dezyre.com/apache-spark-scala-training-online/38)
* [Data Science in Python](https://www.dezyre.com/data-science-in-python/36)
* [Data Science inR](https://www.dezyre.com/data-science-in-R-programming/37)
* [Data Science Training](https://www.dezyre.com/data-science-training-online/)
* [Hadoop Training in California](https://www.dezyre.com/hadoop-course/hadoop-training-in-california)
* [Hadoop Training in New York](https://www.dezyre.com/hadoop-course/hadoop-training-in-newyork)
* [Hadoop Training in Texas](https://www.dezyre.com/hadoop-course/hadoop-training-in-texas)
* [Hadoop Training in Virginia](https://www.dezyre.com/hadoop-course/hadoop-training-in-virginia)
* [Hadoop Training in Washington](https://www.dezyre.com/hadoop-course/hadoop-training-in-washington)
* [Hadoop Training in New Jersey](https://www.dezyre.com/hadoop-course/hadoop-training-in-newjersey)

Courses

* [Big Data and Hadoop Certification Training](https://www.dezyre.com/Hadoop-Training-online/19)
* [Hadoop Project based Training](https://www.dezyre.com/hadoop-hands-on-training/47)
* [Apache Spark Certification Training](https://www.dezyre.com/apache-spark-scala-training-online/38)
* [Data Science Training Course](https://www.dezyre.com/data-science-training-course-online/36)
* [Data Science in R Programming](https://www.dezyre.com/data-science-in-R-programming-online-training/37)
* [Salesforce Certifications - ADM 201 and DEV 401](https://www.dezyre.com/Salesforce-Certifications-ADM-201-DEV-401/29)
* [Hadoop Administration for Big Data](https://www.dezyre.com/Hadoop-Administration/28)
* [Certificate in NoSQL Databases for Big Data](https://www.dezyre.com/NoSQL-Databases/27)
* [Advanced MS Excel with Macro, VBA and Dashboards](https://www.dezyre.com/Advanced-Excel/30)

[https://s3.amazonaws.com/files.dezyre.com/dezyre2.0/images/comodo_secure_113x59_transp.png](javascript:if(window.open('https://secure.comodo.com/ttb_searcher/trustlogo?v_querytype=W&v_shortname=SC5&v_search=https://www.dezyre.com/article/top-100-hadoop-interview-questions-and-answers-2016/159&x=6&y=5','tl_wnd_credentials'+(new%20Date()).getTime(),'toolbar=0,scrollbars=1,location=1,status=1,menubar=1,resizable=1,width=374,height=660,left=60,top=120'))%7b%7d;tLlB(tLTB);)

[EV SSL Certificate](https://ssl.comodo.com/ev-ssl-certificates.php)

About DeZyre

* [About Us](https://www.dezyre.com/index/aboutus)
* [Contact Us](https://www.dezyre.com/index/contactus)
* [DeZyre Reviews](https://www.dezyre.com/reviews/dezyre-hadoop-training-reviews/19)
* [Blog](https://www.dezyre.com/blog)
* [Tutorials](https://www.dezyre.com/tutorial)
* [Webinar](https://www.dezyre.com/free-webinar)
* [Online Hackathons](https://www.dezyre.com/hackerday)
* [Student Portfolios](https://www.dezyre.com/student-portfolios)
* [Privacy Policy](https://www.dezyre.com/index/privacy)
* [Disclaimer](https://www.dezyre.com/index/disclaimer)

Connect with us

**Dezyre Online**

Copyright 2016 Iconiq Inc. All rights reserved. All trademarks are property of their respective owners.

Share to LinkedIn

, Number of shares21

Share to TwitterShare to Facebook

, Number of shares

Share to Reddit

, Number of shares1

Share to Google+

[DeZyre - live hands on training](https://www.dezyre.com/)

* Call us 1-844-696-6465 (US Toll Free)

* [Home](https://www.dezyre.com/)

* [Courses](https://www.dezyre.com/article/mapreduce-interview-questions-and-answers-for-2016/248)

* [Blog](https://www.dezyre.com/blog)

* [Tutorials](https://www.dezyre.com/tutorial)

* [Interview Questions](https://www.dezyre.com/blog/interview-questions)

* [Online Hackathons](https://www.dezyre.com/hackerday)

* [Student Portfolios](https://www.dezyre.com/student-portfolios)

* [Sign In](https://www.dezyre.com/user/login)
* Build Projects, Learn Skills, Get Hired

* [REQUEST INFO](https://www.dezyre.com/article/mapreduce-interview-questions-and-answers-for-2016/248)

**MapReduce Interview Questions and Answers for 2016**

28 Apr 2016

Hadoop job interview is a tough road to cross with many pitfalls, that can make good opportunities fall off the edge. One, often over-looked part of Hadoop job interview is - thorough preparation. So, here’s how DeZyre helps you get ready for your interview for a Hadoop developer job role.

Hadoop Interview Questions: Everyone has gone through one set of interview questions. However, everyone wishes they had prepared some better ones. So, industry experts at DeZyre present you with a collection some of the best and most frequently asked hadoop interview questions.

After a series of blogs on the most frequently asked Hadoop Interview Questions-

[Top 100 Hadoop Interview Questions and Answers](https://www.dezyre.com/article/-top-100-hadoop-interview-questions-and-answers-2015/159)

[Hadoop Developer Interview Questions at Top Tech Companies](https://www.dezyre.com/article/hadoop-developer-interview-questions-at-top-tech-companies/243),

[Top Hadoop Admin Interview Questions and Answers](https://www.dezyre.com/article/top-hadoop-admin-interview-questions-and-answers/238)

[Top 50 Hadoop Interview Questions](https://www.dezyre.com/article/top-50-hadoop-interview-questions-/92)

[Hadoop Pig Interview Questions and Answers](https://www.dezyre.com/article/pig-interview-questions-and-answers-for-2016/244)

[Hadoop Hive Interview Questions and Answers](https://www.dezyre.com/article/hive-interview-questions-and-answers-for-2016/246)

The next in the series of articles highlighting the most commonly asked Hadoop Interview Questions related to each of the tools in the Hadoop ecosystem is - MapReduce Interview Questions and Answers for 2016.

*If you would like more information about Big Data careers, please click the orange "Request Info" button on top of this page.*

[Java MapReduce](https://www.dezyre.com/hadoop-course/mapreduce) is the heart of the elephant in the big data room, created by Google, that allows developers to write applications for processing huge amounts of data in parallel on commodity hardware in a reliable manner. MapReduce divides a large data processing job into smaller tasks. For instance, a MapReduce job can be considered as the equivalent of constructing a house. Each job of building a house is broken down into several small tasks like laying the foundation, putting up the drywall, etc.  and then it is assigned to individual workers i.e. mappers and reducers. The combined output results in the construction of a new house. The major advantage of using MapReduce programming paradigm is that data processing is easily scalable when there are multiple computing nodes. After writing a MapReduce program, scaling it to thousands of machines in a Hadoop cluster, merely requires changing the configuration.

For the complete list of big data companies and their salaries- [**CLICK HERE**](https://docs.google.com/forms/d/1LFuWEKQKCLR231qR9WE5PZakJj77fTDIW6ox5328HFM/viewform)

Without much ado, let’s charge you for your next hadoop job interview with commonly asked Hadoop MapReduce Interview Questions and Answers-

**Hadoop MapReduce Interview Questions and Answers**

**1) What is identity Mapper and identity reducer?**

IdentityMapper is the default Mapper class in Hadoop. This mapper is executed when no mapper class is defined in the MapReduce job.

IdentityReducer is the default Reducer class in Hadoop. This mapper is executed when no reducer class is defined in the MapReduce job. This class merely passes the input key value pairs into the output directory.

**2) Explain about the basic parameters of mapper and reducer function.**

**Mapper Function Parameters**

The basic parameters of a mapper function are LongWritable, text, text and IntWritable.

LongWritable, text- Input Parameters

Text, IntWritable- Intermediate Output Parameters

Here is a sample code on the usage of Mapper function with basic parameters –

public static class Map extends MapReduceBase implements Mapper {  
private final static IntWritable one = new IntWritable (1);   
private Text word = new Text () ;}

**Reducer Function Parameters**

The basic parameters of a reducer function are text, IntWritable, text, IntWritable

First two parameters Text, IntWritable represent Intermediate Output Parameters

The next two parameters Text, IntWritable represent Final Output Parameters

**3) How data is spilt in Hadoop?**

The InputFormat used in the MapReduce job create the splits. The number of mappers are then decided based on the number of splits. Splits are not always created based on the HDFS block size. It all depends on the programming logic within the getSplits () method of InputFormat.

**4) What is the fundamental difference between a MapReduce Split and a HDFS block?**

MapReduce split is a logical piece of data fed to the mapper. It basically does not contain any data but is just a pointer to the data. HDFS block is a physical piece of data.

**5) When is it not recommended to use MapReduce paradigm for large scale data processing?**

It is not suggested to use MapReduce for iterative processing use cases, as it is not cost effective, instead Apache Pig can be used for the same.

**6) What happens when a DataNode fails during the write process?**

When a DataNode fails during the write process, a new replication pipeline that contains the other DataNodes opens up and the write process resumes from there until the file is closed. NameNode observes that one of the blocks is under-replicated and creates a new replica asynchronously.

**7) List the configuration parameters that have to be specified when running a MapReduce job.**

* Input and Output location of the MapReduce job in HDFS.
* Input and Output Format.
* Classes containing the Map and Reduce functions.
* JAR file that contains driver classes and mapper, reducer classes.

**8) Is it possible to split 100 lines of input as a single split in MapReduce?**

Yes this can be achieved using Class NLineInputFormat

**9) Where is Mapper output stored?**

The intermediate key value data of the mapper output will be stored on local file system of the mapper nodes. This directory location is set in the config file by the Hadoop Admin. Once the Hadoop job completes execution, the intermediate will be cleaned up.

**10) Explain the differences between a combiner and reducer.**

Combiner can be considered as a mini reducer that performs local reduce task. It runs on the Map output and produces the output to reducers input. It is usually used for network optimization when the map generates greater number of outputs.

* Unlike a reducer, the combiner has a constraint that the input or output key and value types must match the output types of the Mapper.
* Combiners can operate only on a subset of keys and values i.e. combiners can be executed on functions that are commutative.
* Combiner functions get their input from a single mapper whereas reducers can get data from multiple mappers as a result of partitioning.

**11) When is it suggested to use a combiner in a MapReduce job?**

Combiners are generally used to enhance the efficiency of a MapReduce program by aggregating the intermediate map output locally on specific mapper outputs. This helps reduce the volume of data that needs to be transferred to reducers. Reducer code can be used as a combiner, only if the operation performed is commutative. However, the execution of a combiner is not assured.

**12) What is the relationship between Job and Task in Hadoop?**

A single job can be broken down into one or many tasks in Hadoop.

**13)  Is it important for Hadoop MapReduce jobs to be written in Java?**

It is not necessary to write Hadoop MapReduce jobs in Java but users can write MapReduce jobs in any desired programming language like Ruby, Perl, Python, R, Awk, etc. through the Hadoop Streaming API.

**14) What is the process of changing the split size if there is limited storage space on Commodity Hardware?**

If there is limited storage space on commodity hardware, the split size can be changed by implementing the “Custom Splitter”. The call to Custom Splitter can be made from the main method.

**15)  What are the primary phases of a Reducer?**

The 3 primary phases of a reducer are –

1) Shuffle

2) Sort

3) Reduce

**16) What is a TaskInstance?**

The actual Hadoop MapReduce jobs that run on each slave node are referred to as Task instances. Every task instance has its own JVM process. For every new task instance, a JVM process is spawned by default for a task.

**17) Can reducers communicate with each other?**

Reducers always run in isolation and they can never communicate with each other as per the Hadoop MapReduce programming paradigm.

**18) What is the difference between Hadoop and RDBMS?**

* In RDBMS, data needs to be pre-processed being stored, whereas Hadoop requires no pre-processing.
* RDBMS is generally used for OLTP processing whereas Hadoop is used for analytical requirements on huge volumes of data.
* Database cluster in RDBMS uses the same data files in shared storage whereas in Hadoop the storage is independent of each processing node.

**19) Can we search files using wildcards?**

Yes, it is possible to search for file through wildcards.

**20) How is reporting controlled in hadoop?**

The file hadoop-metrics.properties file controls reporting.

**21) What is the default input type in MapReduce?**

Text

**22) Is it possible to rename the output file?**

Yes, this can be done by implementing the multiple format output class.

**23) What do you understand by compute and storage nodes?**

Storage node is the system, where the file system resides to store the data for processing.

Compute node is the system where the actual business logic is executed.

**24) When should you use a reducer?**

It is possible to process the data without a reducer but when there is a need to combine the output from multiple mappers – reducers are used. Reducers are generally used when shuffle and sort are required.

**25) What is the role of a MapReduce partitioner?**

MapReduce is responsible for ensuring that the map output is evenly distributed over the reducers. By identifying the reducer for a particular key, mapper output is redirected accordingly to the respective reducer.

* [Big Data](https://www.dezyre.com/blog-category/big-data)
* [CRM](https://www.dezyre.com/blog-category/crm)
* [Data Science](https://www.dezyre.com/blog-category/data-science)
* [Mobile App Development](https://www.dezyre.com/blog-category/mobile-app-development)
* [NoSQL Database](https://www.dezyre.com/blog-category/nosql-database)
* [Web Development](https://www.dezyre.com/blog-category/web-development)

You might also like

* [Recap of Data Science News for October](https://www.dezyre.com/article/recap-of-data-science-news-for-october/316)
* [Recap of Apache Spark News for October](https://www.dezyre.com/article/recap-of-apache-spark-news-for-october/315)
* [Recap of Hadoop News for October](https://www.dezyre.com/article/recap-of-hadoop-news-for-october/314)
* [Scenario-Based Hadoop Interview Questions to prepare for in 2016](https://www.dezyre.com/article/scenario-based-hadoop-interview-questions-to-prepare-for-in-2016/313)
* [Top Data Science Certifications to choose from in 2016](https://www.dezyre.com/article/top-data-science-certifications-to-choose-from-in-2016/312)
* [Best Hadoop Certification: Cloudera vs Hortonworks](https://www.dezyre.com/article/best-hadoop-certification-cloudera-vs-hortonworks/311)
* [Recap of Data Science News for September](https://www.dezyre.com/article/recap-of-data-science-news-for-september/310)
* [Recap of Apache Spark News for September](https://www.dezyre.com/article/recap-of-apache-spark-news-for-september/309)
* [Recap of Hadoop News for September](https://www.dezyre.com/article/recap-of-hadoop-news-for-september/308)
* [How to build a Data Science portfolio that will get you hired?](https://www.dezyre.com/article/how-to-build-a-data-science-portfolio-that-will-get-you-hired/307)

Tutorials

* [Hadoop Online Tutorial – Hadoop HDFS Commands Guide](https://www.dezyre.com/hadoop-tutorial/hadoop-hdfs-commands)
* [MapReduce Tutorial–Learn to implement Hadoop WordCount Example](https://www.dezyre.com/hadoop-tutorial/hadoop-mapreduce-wordcount-tutorial)
* [Hadoop Hive Tutorial-Usage of Hive Commands in HQL](https://www.dezyre.com/hadoop-tutorial/hive-commands)
* [Hive Tutorial-Getting Started with Hive Installation on Ubuntu](https://www.dezyre.com/hadoop-tutorial/install-hive)
* [Learn Java for Hadoop Tutorial: Inheritance and Interfaces](https://www.dezyre.com/hadoop-tutorial/java-for-hadoop-inheritance-and-interfaces)
* [Learn Java for Hadoop Tutorial: Classes and Objects](https://www.dezyre.com/hadoop-tutorial/java-for-hadoop-classes-and-objects)
* [Learn Java for Hadoop Tutorial: Arrays](https://www.dezyre.com/hadoop-tutorial/java-for-hadoop-arrays)
* [Apache Spark Tutorial–Run your First Spark Program](https://www.dezyre.com/apache%20spark-tutorial/spark-tutorial)
* [PySpark Tutorial-Learn to use Apache Spark with Python](https://www.dezyre.com/apache%20spark-tutorial/pyspark-tutorial)
* [R Tutorial- Learn Data Visualization with R using GGVIS](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/ggvis)
* [Neural Network Training Tutorial](https://www.dezyre.com/data%20science%20in%20python-tutorial/neural-network-training-tutorial)
* [Python List Tutorial](https://www.dezyre.com/data%20science%20in%20python-tutorial/python-list-tutorial)
* [MatPlotLib Tutorial](https://www.dezyre.com/data%20science%20in%20python-tutorial/matplotlib-tutorial)
* [Decision Tree Tutorial](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/decision-tree-tutorial)
* [Neural Network Tutorial](https://www.dezyre.com/data%20science%20in%20python-tutorial/neural-network-tutorial)
* [Performance Metrics for Machine Learning Algorithms](https://www.dezyre.com/data%20science%20in%20python-tutorial/performance-metrics-for-machine-learning-algorithm)
* [R Tutorial: Data.Table](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/r-data-table-tutorial)
* [SciPy Tutorial](https://www.dezyre.com/data%20science%20in%20python-tutorial/scipy-introduction-tutorial)
* [Step-by-Step Apache Spark Installation Tutorial](https://www.dezyre.com/apache%20spark-tutorial/apache-spark-installation-tutorial)
* [Introduction to Apache Spark Tutorial](https://www.dezyre.com/apache%20spark-tutorial/tutorial-introduction-to-apache-spark)
* [R Tutorial: Importing Data from Web](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/r-tutorial-importing-data-from-web)
* [R Tutorial: Importing Data from Relational Database](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/r-tutorial-importing-data-from-relational-database)
* [R Tutorial: Importing Data from Excel](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/tutorial-importing-data-from-excel-to-r)
* [Introduction to Machine Learning Tutorial](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/introduction-to-machine-learning-tutorial)
* [Machine Learning Tutorial: Linear Regression](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/linear-regression-tutorial)
* [Machine Learning Tutorial: Logistic Regression](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/logistic-regression-tutorial)
* [Support Vector Machine Tutorial (SVM)](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/support-vector-machine-tutorial)
* [K-Means Clustering Tutorial](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/k-means-clustering-techniques-tutorial)
* [dplyr Manipulation Verbs](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/dplyr-manipulations-verbs)
* [Introduction to dplyr package](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/introduction-to-dplyr-package)
* [Importing Data from Flat Files in R](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/importing-data-from-flat-files-in-r)
* [Principal Component Analysis Tutorial](https://www.dezyre.com/data%20science%20in%20python-tutorial/principal-component-analysis-tutorial)
* [Pandas Tutorial Part-3](https://www.dezyre.com/data%20science%20in%20python-tutorial/pandas-introductory-tutorial-part-3)
* [Pandas Tutorial Part-2](https://www.dezyre.com/data%20science%20in%20python-tutorial/pandas-introductory-tutorial-part-2)
* [Pandas Tutorial Part-1](https://www.dezyre.com/data%20science%20in%20python-tutorial/pandas-introductory-tutorial-part-1)
* [Tutorial- Hadoop Multinode Cluster Setup on Ubuntu](https://www.dezyre.com/hadoop-tutorial/hadoop-multinode-cluster-setup)
* [Data Visualizations Tools in R](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/data-visualizations-tools-r)
* [R Statistical and Language tutorial](https://www.dezyre.com/data%20science%20in%20r%20programming-tutorial/r-statistical-language-tutorial)
* [Introduction to Data Science with R](https://www.dezyre.com/data%20science%20in%20python-tutorial/introduction-to-data-science-r)
* [Apache Pig Tutorial: User Defined Function Example](https://www.dezyre.com/hadoop-tutorial/apache-pig-tutorial-udf)
* [Apache Pig Tutorial Example: Web Log Server Analytics](https://www.dezyre.com/hadoop-tutorial/pig-tutorial-web-log-server-analytics)
* [Impala Case Study: Web Traffic](https://www.dezyre.com/hadoop-tutorial/impala-case-study-web-traffic)
* [Impala Case Study: Flight Data Analysis](https://www.dezyre.com/hadoop-tutorial/impala-case-study-flight-data-analysis)
* [Hadoop Impala Tutorial](https://www.dezyre.com/hadoop-tutorial/hadoop-impala-tutorial)
* [Apache Hive Tutorial: Tables](https://www.dezyre.com/hadoop-tutorial/apache-hive-tutorial-tables)
* [Flume Hadoop Tutorial: Twitter Data Extraction](https://www.dezyre.com/hadoop-tutorial/flume-hadoop-twitter-data-extraction)
* [Flume Hadoop Tutorial: Website Log Aggregation](https://www.dezyre.com/hadoop-tutorial/flume-hadoop-website-log-aggregation)
* [Hadoop Sqoop Tutorial: Example Data Export](https://www.dezyre.com/hadoop-tutorial/hadoop-sqoop-tutorial-example-data-export)
* [Hadoop Sqoop Tutorial: Example of Data Aggregation](https://www.dezyre.com/hadoop-tutorial/hadoop-sqoop-tutorial-data-aggregation)
* [Apache Zookepeer Tutorial: Example of Watch Notification](https://www.dezyre.com/hadoop-tutorial/apache-zookeeper-watch-notification)
* [Apache Zookepeer Tutorial: Centralized Configuration Management](https://www.dezyre.com/hadoop-tutorial/apache-zookeeper-tutorial-centralize-configuration)
* [Hadoop Zookeeper Tutorial](https://www.dezyre.com/hadoop-tutorial/zookeeper-tutorial)
* [Hadoop Sqoop Tutorial](https://www.dezyre.com/hadoop-tutorial/hadoop-sqoop-tutorial)
* [Hadoop PIG Tutorial](https://www.dezyre.com/hadoop-tutorial/pig-tutorial)
* [Hadoop Oozie Tutorial](https://www.dezyre.com/hadoop-tutorial/oozie-tutorial)
* [Hadoop NoSQL Database Tutorial](https://www.dezyre.com/hadoop-tutorial/nosql-database-tutorial)
* [Hadoop Hive Tutorial](https://www.dezyre.com/hadoop-tutorial/hive-tutorial)
* [Hadoop HDFS Tutorial](https://www.dezyre.com/hadoop-tutorial/hdfs-tutorial)
* [Hadoop hBase Tutorial](https://www.dezyre.com/hadoop-tutorial/hbase-tutorial)
* [Hadoop Flume Tutorial](https://www.dezyre.com/hadoop-tutorial/flume-tutorial)
* [Hadoop 2.0 YARN Tutorial](https://www.dezyre.com/hadoop-tutorial/yarn-tutorial)
* [Hadoop MapReduce Tutorial](https://www.dezyre.com/hadoop-tutorial/hadoop-mapreduce-tutorial-)
* [Big Data Hadoop Tutorial for Beginners- Hadoop Installation](https://www.dezyre.com/hadoop-tutorial/big-data-hadoop-tutorial)

Online Courses

* [Hadoop Training](https://www.dezyre.com/Hadoop-Training-online/19)
* [Spark Certification Training](https://www.dezyre.com/apache-spark-scala-training-online/38)
* [Data Science in Python](https://www.dezyre.com/data-science-in-python/36)
* [Data Science inR](https://www.dezyre.com/data-science-in-R-programming/37)
* [Data Science Training](https://www.dezyre.com/data-science-training-online/)
* [Hadoop Training in California](https://www.dezyre.com/hadoop-course/hadoop-training-in-california)
* [Hadoop Training in New York](https://www.dezyre.com/hadoop-course/hadoop-training-in-newyork)
* [Hadoop Training in Texas](https://www.dezyre.com/hadoop-course/hadoop-training-in-texas)
* [Hadoop Training in Virginia](https://www.dezyre.com/hadoop-course/hadoop-training-in-virginia)
* [Hadoop Training in Washington](https://www.dezyre.com/hadoop-course/hadoop-training-in-washington)
* [Hadoop Training in New Jersey](https://www.dezyre.com/hadoop-course/hadoop-training-in-newjersey)

Courses

* [Big Data and Hadoop Certification Training](https://www.dezyre.com/Hadoop-Training-online/19)
* [Hadoop Project based Training](https://www.dezyre.com/hadoop-hands-on-training/47)
* [Apache Spark Certification Training](https://www.dezyre.com/apache-spark-scala-training-online/38)
* [Data Science Training Course](https://www.dezyre.com/data-science-training-course-online/36)
* [Data Science in R Programming](https://www.dezyre.com/data-science-in-R-programming-online-training/37)
* [Salesforce Certifications - ADM 201 and DEV 401](https://www.dezyre.com/Salesforce-Certifications-ADM-201-DEV-401/29)
* [Hadoop Administration for Big Data](https://www.dezyre.com/Hadoop-Administration/28)
* [Certificate in NoSQL Databases for Big Data](https://www.dezyre.com/NoSQL-Databases/27)
* [Advanced MS Excel with Macro, VBA and Dashboards](https://www.dezyre.com/Advanced-Excel/30)

**Dezyre Online**

Copyright 2016 Iconiq Inc. All rights reserved. All trademarks are property of their respective owners.

Share to LinkedIn

, Number of shares6

Share to TwitterShare to Facebook

, Number of shares1

Share to Reddit

, Number of shares

Share to Google+

**Flat 20% & upto 50% off + Free additional Courses. Hurry up!**

[Intellipaat](https://intellipaat.com/)

[Intellipaat](https://intellipaat.com/)

* [Resources](https://intellipaat.com/interview-question/big-data-hadoop-interview-questions/)
* [Blog](https://intellipaat.com/blog/)
* [All Courses](https://intellipaat.com/all-courses/big-data/)
* [Login](https://intellipaat.com/interview-question/big-data-hadoop-interview-questions/#login)
* [0](https://intellipaat.com/all-courses/)

[**Browse Courses**](https://intellipaat.com/interview-question/big-data-hadoop-interview-questions/)

Top of Form



Bottom of Form

**Popular Search:**[Big Data Hadoop](https://intellipaat.com/big-data-hadoop-training/)**,**[Tableau](https://intellipaat.com/tableau-training/)**,**[Data Science](https://intellipaat.com/data-science-training/)**,**[Salesforce](https://intellipaat.com/salesforce-training/)**,**[Qlikview](https://intellipaat.com/qlikview-training/)

[Home](https://intellipaat.com/) » **Hadoop Interview Questions**

Top of Form



Bottom of Form



**INTERVIEW QUESTIONS**

* [Top Splunk Interview Questions And Answers](https://intellipaat.com/interview-question/splunk-interview-questions/)
* [Top Apache Solr Interview Questions And Answers](https://intellipaat.com/interview-question/apache-solr-interview-questions/)
* [Top Apache Storm Interview Questions And Answers](https://intellipaat.com/interview-question/apache-storm-interview-questions/)
* [Top Apache Spark Interview Questions And Answers](https://intellipaat.com/interview-question/apache-spark-interview-questions/)
* [Top Mapreduce Interview Questions And Answers](https://intellipaat.com/interview-question/map-reduce-interview-questions/)
* [Top HDFS Interview Questions And Answers](https://intellipaat.com/interview-question/hdfs-interview-questions/)
* [Top PIG Interview Questions – Most Asked](https://intellipaat.com/interview-question/pig-interview-questions/)
* [Top Hive Interview Questions – Most Asked](https://intellipaat.com/interview-question/hive-interview-questions/)

[**Browse All Interview Questions**](https://intellipaat.com/interview-question/)

**POPULAR TUTORIALS**

* [Hadoop Tutorial](https://intellipaat.com/tutorial/hadoop-tutorial/)

[**Browse All Tutorials**](https://intellipaat.com/tutorial/)

Top Hadoop Interview Questions And Answers

Here are top 20 objective type sample Hadoop Interview questions and their answers are given just below to them. These sample questions are framed by experts from Intellipaat who train for [Learn Hadoop Online](https://intellipaat.com/big-data-hadoop-training/) to give you an idea of type of questions which may be asked in interview. We have taken full care to give correct answers for all the questions. Do comment your thoughts. Happy Job Hunting!

Top Answers to Hadoop Interview Questions

**1.What are real-time industry applications of Hadoop?**

Hadoop, well known as Apache Hadoop, is an open-source software platform for scalable and distributed computing of large volumes of data. It provides rapid, high performance and cost-effective analysis of structured and unstructured data generated on digital platforms and within the enterprise. It is used in almost all departments and sectors today.

Some of the instances where Hadoop is used:

• Managing traffic on streets  
• Streaming processing  
• Content Management and Archiving Emails  
• Processing Rat Brain Neuronal Signals using a Hadoop Computing Cluster  
• Fraud detection and Prevention  
• Advertisements Targeting Platforms are using Hadoop to capture and analyze click stream, transaction, video and social media data  
• Managing content, posts, images and videos on social media platforms  
• Analyzing customer data in real-time for improving business performance  
• Public sector fields such as intelligence, defense, cyber security and scientific research  
• Financial agencies are using Big Data Hadoop to reduce risk, analyze fraud patterns, identify rogue traders, more precisely target their marketing campaigns based on customer segmentation, and improve customer satisfaction  
• Getting access to unstructured data like output from medical devices, doctor’s notes, lab results, imaging reports, medical correspondence, clinical data, and financial data. *Read this log to find out*[*how Big Data is transforming real estate now*](https://intellipaat.com/blog/how-big-data-is-transforming-real-estate/)*.*

***Are you interested in learning Hadoop? Well, we have the comprehensive***[*Hadoop Online Course*](https://intellipaat.com/big-data-hadoop-training/#curriculum)***to give you a head start in your career.***

**2.How is Hadoop different from other parallel computing systems?**

Hadoop is a distributed file system, which lets you store and handle massive amount of data on a cloud of machines, handling data redundancy. Go through this [HDFS content to know how the distributed file system works](https://intellipaat.com/tutorial/hadoop-tutorial/hdfs-overview/). The primary benefit is that since data is stored in several nodes, it is better to process it in distributed manner. Each node can process the data stored on it instead of spending time in moving it over the network.  
On the contrary, in Relational database computing system, you can query data in real-time, but it is not efficient to store data in tables, records and columns when the data is huge. [Learn about Oracle DBA](https://intellipaat.com/tutorial/oracle-dba-tutorial/) now.  
Hadoop also provides a scheme to build a Column Database with Hadoop HBase, for runtime queries on rows. Learn more in this [HBase Tutorial](https://intellipaat.com/tutorial/hbase-tutorial/" \t "_blank)**.**

**3.What all modes Hadoop can be run in?**

Hadoop can run in three modes:  
1. **Standalone Mode:** Default mode of Hadoop, it uses local file stystem 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. Much faster when compared to other modes.  
2. **Pseudo-Distributed Mode (Single Node Cluster):** 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 node are the same.  
3. **Fully Distributed Mode (Multiple Cluster Node):** This is the production phase of Hadoop (what Hadoop is known for) where data is used and distributed across several nodes on a Hadoop cluster. Separate nodes are allotted as Master and Slave.

***Learn more about Hadoop in this***[*Hadoop Certification course*](https://intellipaat.com/big-data-hadoop-training/#certification)***to get ahead in your career!***

**4.Explain the major difference between HDFS block and InputSplit.**

In simple terms, block is the physical representation of data while split is the logical representation of data present in the block. Split acts a s an intermediary between block and mapper. Suppose we have two blocks:  
**Block 1: ii nntteell  
Block 2: Ii ppaatt**  
Now, considering the map, it will read first block from ii till ll, but does not know how to process the second block at the same time. Here comes Split into play, which will form a logical group of Block1 and Block 2 as a single block. It then forms key-value pair using inputformat and records reader and sends map for further processing With inputsplit, if you have limited resources, you can increase the split size to limit the number of maps. For instance, if there are 10 blocks of 640MB (64MB each) and there are limited resources, you can assign ‘split size’ as 128MB. This will form a logical group of 128MB, with only 5 maps executing at a time.  
However, if the ‘split size’ property is set to false, whole file will form one inputsplit and is processed by single map, consuming more time when the file is bigger.

**5.What is distributed cache and what are its benefits?**

Distributed Cache, in Hadoop, is a service by MapReduce framework to cache files when needed. Learn more in this [MapReduce Tutorial](https://intellipaat.com/tutorial/mapreduce-tutorial/" \t "_blank) now. Once a file is cached for a specific job, hadoop will make it available on each data node both in system and in memory, where map and reduce tasks are executing.Later, you can easily access and read the cache file and populate any collection (like array, hashmap) in your code.  
**Benefits of using distributed cache are:**  
• It distributes simple, read only text/data files and/or complex types like jars, archives and others. These archives are then un-archived at the slave node.  
• Distributed cache tracks the modification timestamps of cache files, which notifies that the files should not be modified until a job is executing currently.

***Give your career a big boost by going through our***[*Hadoop Online Training Videos*](https://intellipaat.com/big-data-hadoop-training/#sample-video)***now!***

**6.Explain the difference between NameNode, Checkpoint NameNode and BackupNode.**

**NameNode** is the core of HDFS that manages the metadata – the information of what file maps to what block locations and what blocks are stored on what datanode. In simple terms, it’s the data about the data being stored. NameNode supports a directory tree-like structure consisting of all the files present in HDFS on a Hadoop cluster. It uses following files for namespace:  
fsimage file- It keeps track of the latest checkpoint of the namespace.  
edits file-It is a log of changes that have been made to the namespace since checkpoint.  
**Checkpoint NameNode** has the same directory structure as NameNode, and creates checkpoints for namespace at regular intervals by downloading the fsimage and edits file and margining them within the local directory. The new image after merging is then uploaded to NameNode.  
There is a similar node like Checkpoint, commonly known as Secondary Node, but it does not support the ‘upload to NameNode’ functionality.  
**Backup Node** provides similar functionality as Checkpoint, enforcing synchronization with NameNode. It maintains an up-to-date in-memory copy of file system namespace and doesn’t require getting hold of changes after regular intervals. The backup node needs to save the current state in-memory to an image file to create a new checkpoint.

***Learn about the various Hadoop components in this***[*Big Data Hadoop Video Tutorial*](https://intellipaat.com/big-data-hadoop-video-tutorial/)***.***

**7.What are the most common Input Formats in Hadoop?**

There are three most common input formats in Hadoop:  
**• Text Input Format:** Default input format in Hadoop.  
**• Key Value Input Format:** used for plain text files where the files are broken into lines  
**• Sequence File Input Format:** used for reading files in sequence

**8.Define DataNode and how does NameNode tackle DataNode failures?**

DataNode stores data in HDFS; it is a node where actual data resides in the file system. Each datanode sends a heartbeat message to notify that it is alive. If the namenode does noit receive a message from datanode for 10 minutes, it considers it to be dead or out of place, and starts replication of blocks that were hosted on that data node such that they are hosted on some other data node.A BlockReport contains list of all blocks on a DataNode. Now, the system starts to replicate what were stored in dead DataNode.  
The NameNode manages the replication of data blocksfrom one DataNode to other. In this process, the replication data transfers directly between DataNode such that the data never passes the NameNode.

**9.What are the core methods of a Reducer?**

The three core methods of a Reducer are:  
1. **setup():** this method is used for configuring various parameters like input data size, distributed cache.  
public void setup (context)  
2. **reduce():** heart of the reducer always called once per key with the associated reduced task  
public void reduce(Key, Value, context)  
3. **cleanup():** this method is called to clean temporary files, only once at the end of the task  
public void cleanup (context)

**10.What is SequenceFile in Hadoop?**

Extensively used in MapReduce I/O formats, SequenceFile is a flat file containing binary key/value pairs. The map outputs are stored as SequenceFile internally. It provides Reader, Writer and Sorter classes. The three SequenceFile formats are:  
1. Uncompressed key/value records.  
2. Record compressed key/value records – only ‘values’ are compressed here.  
3. Block compressed key/value records – both keys and values are collected in ‘blocks’ separately and compressed. The size of the ‘block’ is configurable.

**11.What is Job Tracker role in Hadoop?**

Job Tracker’s primary function is resource management (managing the task trackers), tracking resource availability and task life cycle management (tracking the taks progress and fault tolerance).  
• It is a process that runs on a separate node, not on a DataNode often  
• Job Tracker communicates with the NameNode to identify data location  
• Finds the best Task Tracker Nodes to execute tasks on given nodes  
• Monitors individual Task Trackers and submits the overall job back to the client.  
• It tracks the execution of MapReduce workloads local to the slave node.

**12.What is the use of RecordReader in Hadoop?**

Since Hadoop splits data into various blocks, RecordReader is used to read the slit data into single record. For instance, if our input data is split like:  
Row1: Welcome to  
Row2: Intellipaat  
It will be read as “Welcome to Intellipaat” using RecordReader.

**13.What is Speculative Execution in Hadoop?**

One limitation of Hadoop is that by distributing the tasks on several nodes, there are chances that few slow nodes limit the rest of the program. Tehre are various reasons for the tasks to be slow, which are sometimes not easy to detect. Instead of identifying and fixing the slow-running tasks, Hadoop tries to detect when the task runs slower than expected and then launches other equivalent task as backup. This backup mechanism in Hadoop is Speculative Execution.  
It creates a duplicate task on another disk. The same input can be processed multiple times in parallel. When most tasks in a job comes to completion, the speculative execution mechanism schedules duplicate copies of remaining tasks (which are slower) across the nodes that are free currently. When these tasks finish, it is intimated to the JobTracker. If other copies are executing speculatively, Hadoop notifies the TaskTrackers to quit those tasks and reject their output.  
Speculative execution is by default true in Hadoop. To disable, set mapred.map.tasks.speculative.execution and mapred.reduce.tasks.speculative.execution  
JobConf options to false.

**14.What happens if you try to run a Hadoop job with an output directory that is already present?**

It will throw an exception saying that the output file directory already exists. To run the MapReduce job, you need to ensure that the output directory does not exist before in the HDFS.  To delete the directory before running the job, you can use shell:Hadoop fs –rmr /path/to/your/output/Or via the Java API: FileSystem.getlocal(conf).delete(outputDir, true);

Prepare yourself for the [MapReduce Interview questions and answers](https://intellipaat.com/interview-question/map-reduce-interview-questions/" \t "_blank)Now

**15.How can you debug Hadoop code?**

First, check the list of MapReduce jobs currently running. Next, we need to see that there are no orphaned jobs running; if yes, you need to determine the location of RM logs.  
1. Run: “ps –ef | grep –I ResourceManager”  
and look for log directory in the displayed result. Find out the job-id from the displayed list and check if there is any error message associated with that job.  
2. On the basis of RM logs, identify the worker node that was involved in execution of the task.  
3. Now, login to that node and run – “ps –ef | grep –iNodeManager”  
4. Examine the Node Manager log. The majority of errors come from user level logs for each map-reduce job.

**16.How to configure Replication Factor in HDFS?**

hdfs-site.xml is used to configure HDFS. Changing the dfs.replication property in hdfs-site.xml will change the default replication for all files placed in HDFS.  
You can also modify the replication factor on a per-file basis using the Hadoop FS Shell:[training@localhost ~]$ hadoopfs –setrep –w 3 /my/fileConversely, you can also change the replication factor of all the files under a directory.  
[training@localhost ~]$ hadoopfs –setrep –w 3 -R /my/dir

Go through [Hadoop Administration Training](https://intellipaat.com/hadoop-administration-training/) to learn about Replication Factor In HDFS now!

**17.How to compress mapper output but not the reducer output?**

To achieve this compression, you should set:  
conf.set(“mapreduce.map.output.compress”, true)  
conf.set(“mapreduce.output.fileoutputformat.compress”, false)

**18.What is the difference between Map Side join and Reduce Side Join?**

Map side Join at map side is performed data reaches the map. You need a strict structure for defining map side join. On the other hand, Reduce side Join (Repartitioned Join) is simpler than map side join since the input datasets need not be structured. However, it is less efficient as it will have to go through sort and shuffle phases, coming with network overheads.

**19.How can you transfer data from Hive to HDFS?**

By writing the query: **hive> insert overwrite directory ‘/’ select \* from emp;**  
You can write your query for the data you want to import from Hive to HDFS. The output you receive will be stored in part files in the specified HDFS path.

**20.What companies use Hadoop, any idea?**

[Learn how Big Data and Hadoop have changed the rules of the game](https://intellipaat.com/blog/big-data-changing-disruptive-innovation/) in this blog post. Yahoo! (the biggest contributor to the creation of Hadoop) – Yahoo search engine uses Hadoop, Facebook – Developed Hive for analysis , Amazon, Netflix, Adobe, eBay, Spotify, Twitter, Adobe.