





Home (/) > Big Data (/Big-data-and-analytics) > Big Data Analytics (<https://www.edureka.co/blog/category/big-data-analytics/>) > Top 50 Hadoop Interview Qu...


[\(https://www.edureka.co/blog/all/\)](https://www.edureka.co/blog/all/)


[Blogs \(https://www.edureka.co/blog/\)](https://www.edureka.co/blog/)


[Videos \(https://www.edureka.co/blog/videos/\)](https://www.edureka.co/blog/videos/)




[Interview Questions \(https://www.edureka.co/blog/interview-questions/\)](https://www.edureka.co/blog/interview-questions/)



Top 50 Hadoop Interview Questions You Must Prepare In 2018

Recommended by 512 users

Shubham Sinha (<https://www.edureka.co/blog/author/shubham-sinha/>) | Jul 17, 2018

<https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/>
<https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/#comments-wrapper>

 Add to Bookmarks (<https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/>)
  Email this Post (shubham.sinha@edureka.co)

 185.4K
  (https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/#disqus_thread)

Top 50 Hadoop Interview Questions for 2018

In this Hadoop interview questions blog, we will be covering all the frequently asked questions that will help you ace the interview with their best solutions. But before that, let me tell you how the demand is continuously increasing for Big Data and Hadoop experts. You can check out **this skill report** (<https://www.edureka.co/skill-report>) which talks about the top technical skills to master in 2018.

Following are a few stats that reflect the growth in the demand for Big Data & Hadoop quite accurately:

- Big Data will drive \$48.6 billion in annual spending by 2019- IDC.
- McKinsey predicts that by 2018 there will be a shortage of 1.5M data experts
- Average salary of a Big Data Hadoop developer in the US is \$135k- Indeed.com
- Average annual salary in the United Kingdom is £66,250-£66,750- itjobswatch.co.uk

I would like to draw your attention towards the Big Data revolution. Earlier, organizations were only concerned about operational data, which was less than 20% of the whole data. Later, they realized that analyzing the whole data will give them better business insights & decision-making capability. That was the time when big giants like Yahoo, Facebook, Google, etc. started adopting Hadoop & Big Data related technologies. In fact, nowadays one of every fifth company is moving to Big Data analytics. Hence, the demand for jobs in Big Data Hadoop is rising like anything. Therefore, if you want to boost your career, Hadoop and Spark are just the technology you need. This would always give you a good start either as a fresher or experienced.

Prepare with these top Hadoop interview questions to get an edge in the burgeoning Big Data market where global and local enterprises, big or small, are looking for the quality Big Data and Hadoop experts. This definitive list of top Hadoop interview questions will take you through the questions and answers around **Hadoop Cluster** (<https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hadoop-cluster/>), **HDFS** (<https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hdfs-2/>), **MapReduce** (<https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-mapreduce/>), **Pig** (<https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-pig/>), **Hive** (<https://www.edureka.co/blog/interview-questions/hive-interview-questions/>), **HBase** (<https://www.edureka.co/blog/interview-questions/hbase-interview-questions/>). This blog is the gateway to your next Hadoop job.

In case you have come across a few difficult questions in a Hadoop interview and are still confused about the best answer, kindly put those questions in the comment section below. We will be happy to answer them.

Hadoop Interview Questions and Answers | Edureka

In the meantime, you can maximize the Big Data Analytics career opportunities that are sure to come your way by taking Hadoop online training with Edureka. Click below to know more.

Get Skilled in Hadoop

(https://www.edureka.co/big-data-and-hadoop/)

1. What are the basic differences between relational database and HDFS?

Here are the key differences between HDFS and relational database:

RDBMS vs. Hadoop

	RDBMS	Hadoop
Data Types	RDBMS relies on the structured data and the schema of the data is always known.	Any kind of data can be stored into Hadoop i.e. Be it structured, unstructured or semi-structured.
Processing	RDBMS provides limited or no processing capabilities.	Hadoop allows us to process the data which is distributed across the cluster in a parallel fashion.
Schema on Read Vs. Write	RDBMS is based on 'schema on write' where schema validation is done before loading the data.	On the contrary, Hadoop follows the schema on read policy.
Read/Write Speed	In RDBMS, reads are fast because the schema of the data is already known.	The writes are fast in HDFS because no schema validation happens during HDFS write.
Cost	Licensed software, therefore, I have to pay for the software.	Hadoop is an open source framework. So, I don't need to pay for the software.
Best Fit Use Case	RDBMS is used for OLTP (Online Transactional Processing) system.	Hadoop is used for Data discovery, data analytics or OLAP system.

2. Explain "Big Data" and what are five V's of Big Data?

"Big data" is the term for a collection of large and complex data sets, that makes it difficult to process using relational database management tools or traditional data processing applications. It is difficult to capture, curate, store, search, share, transfer, analyze, and visualize Big data. Big Data has emerged as an opportunity for companies. Now they can successfully derive value from their data and will have a distinct advantage over their competitors with enhanced business decisions making capabilities.

♣ *Tip: It will be a good idea to talk about the 5Vs in such questions, whether it is asked specifically or not!*

- **Volume:** The volume represents the amount of data which is growing at an exponential rate i.e. in Petabytes and Exabytes.
- **Velocity:** Velocity refers to the rate at which data is growing, which is very fast. Today, yesterday's data are considered as old data. Nowadays, social media is a major contributor to the velocity of growing data.
- **Variety:** Variety refers to the heterogeneity of data types. In another word, the data which are gathered has a variety of formats like videos, audios, csv, etc. So, these various formats represent the variety of data.
- **Veracity:** Veracity refers to the data in doubt or uncertainty of data available due to data inconsistency and incompleteness. Data available can sometimes get messy and may be difficult to trust. With many forms of big data, quality and accuracy are difficult to control. The volume is often the reason behind for the lack of quality and accuracy in the data.
- **Value:** It is all well and good to have access to big data but unless we can turn it into a value it is useless. By turning it into value I mean, Is it adding to the benefits of the organizations? Is the organization working on Big Data achieving high ROI (Return On Investment)? Unless, it adds to their profits by working on Big Data, it is useless.

As we know Big Data is growing at an accelerating rate, so the factors associated with it are also evolving. To go through them and understand it in detail, I recommend you to go through **Big Data Tutorial** (<https://www.edureka.co/blog/big-data-tutorial/>) blog.

3. What is Hadoop and its components.

When "Big Data" emerged as a problem, Apache Hadoop evolved as a solution to it. Apache Hadoop is a framework which provides us various services or tools to store and process Big Data. It helps in analyzing Big Data and making business decisions out of it, which can't be done efficiently and effectively using traditional systems.

♣ *Tip: Now, while explaining Hadoop, you should also explain the main components of Hadoop, i.e.:*

- **Storage unit**– HDFS (NameNode, DataNode)
- **Processing framework**– YARN (ResourceManager, NodeManager)

4. What are HDFS and YARN?

HDFS (Hadoop Distributed File System) is the storage unit of Hadoop. It is responsible for storing different kinds of data as blocks in a distributed environment. It follows master and slave topology.

♣ *Tip: It is recommended to explain the HDFS components too i.e.*

- **NameNode:** NameNode is the master node in the distributed environment and it maintains the metadata information for the blocks of data stored in HDFS like block location, replication factors etc.
- **DataNode:** DataNodes are the slave nodes, which are responsible for storing data in the HDFS. NameNode manages all the DataNodes.

YARN (Yet Another Resource Negotiator) is the processing framework in Hadoop, which manages resources and provides an execution environment to the processes.

♣ *Tip: Similarly, as we did in HDFS, we should also explain the two components of YARN:*

- **ResourceManager:** It receives the processing requests, and then passes the parts of requests to corresponding NodeManagers accordingly, where the actual processing takes place. It allocates resources to applications based on the needs.
- **NodeManager:** NodeManager is installed on every DataNode and it is responsible for the execution of the task on every single DataNode.

If you want to learn in detail about HDFS & YARN go through **Hadoop Tutorial** (<https://www.edureka.co/blog/hadoop-tutorial/>) blog.

5. Tell me about the various Hadoop daemons and their roles in a Hadoop cluster.

Generally approach this question by first explaining the HDFS daemons i.e. NameNode, DataNode and Secondary NameNode, and then moving on to the YARN daemons i.e. ResourceManager and NodeManager, and lastly explaining the JobHistoryServer.

i.e. ResourceManager and NodeManager, and lastly explaining the JobHistoryServer.

- **NameNode:** It is the master node which is responsible for storing the metadata of all the files and directories. It has information about blocks, that make a file, and where those blocks are located in the cluster.
- **Datanode:** It is the slave node that contains the actual data.
- **Secondary NameNode:** It periodically merges the changes (edit log) with the FsImage (Filesystem Image), present in the NameNode. It stores the modified FsImage into persistent storage, which can be used in case of failure of NameNode.
- **ResourceManager:** It is the central authority that manages resources and schedule applications running on top of YARN.
- **NodeManager:** It runs on slave machines, and is responsible for launching the application's containers (where applications execute their part), monitoring their resource usage (CPU, memory, disk, network) and reporting these to the ResourceManager.
- **JobHistoryServer:** It maintains information about MapReduce jobs after the Application Master terminates.

Hadoop Installation Interview Questions

(<https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hadoop-cluster/>)

Hadoop HDFS Interview Questions

6. Compare HDFS with Network Attached Storage (NAS).

In this question, first explain NAS and HDFS, and then compare their features as follows:

- Network-attached storage (NAS) is a file-level computer data storage server connected to a computer network providing data access to a heterogeneous group of clients. NAS can either be a hardware or software which provides services for storing and accessing files. Whereas Hadoop Distributed File System (HDFS) is a distributed filesystem to store data using commodity hardware.
- In HDFS Data Blocks are distributed across all the machines in a cluster. Whereas in NAS data is stored on a dedicated hardware.
- HDFS is designed to work with MapReduce paradigm, where computation is moved to the data. NAS is not suitable for MapReduce since data is stored separately from the computations.
- HDFS uses commodity hardware which is cost-effective, whereas a NAS is a high-end storage devices which includes high cost.

7. List the difference between Hadoop 1 and Hadoop 2.

This is an important question and while answering this question, we have to mainly focus on two points i.e. Passive NameNode and YARN architecture.

- In Hadoop 1.x, "NameNode" is the single point of failure. In Hadoop 2.x, we have Active and Passive "NameNodes". If the active "NameNode" fails, the passive "NameNode" takes charge. Because of this, high availability can be achieved in Hadoop 2.x.
- Also, in Hadoop 2.x, YARN provides a central resource manager. With YARN, you can now run multiple applications in Hadoop, all sharing a common resource. MRV2 is a particular type of distributed application that runs the MapReduce framework on top of YARN. Other tools can also perform data processing via YARN, which was a problem in Hadoop 1.x.

Hadoop 1.x vs. Hadoop 2.x

	Hadoop 1.x	Hadoop 2.x
Passive NameNode	NameNode is a Single Point of Failure	Active & Passive NameNode
Processing	MRV1 (Job Tracker & Task Tracker)	MRV2/YARN (ResourceManager & NodeManager)

8. What are active and passive "NameNodes"?

In HA (High Availability) architecture, we have two NameNodes – Active "NameNode" and Passive "NameNode".

- Active "NameNode" is the "NameNode" which works and runs in the cluster.
- Passive "NameNode" is a standby "NameNode", which has similar data as active "NameNode".

When the active "NameNode" fails, the passive "NameNode" replaces the active "NameNode" in the cluster. Hence, the cluster is never without a "NameNode" and so it never fails.

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

One of the most attractive features of the Hadoop framework is its *utilization of commodity hardware*. However, this leads to frequent "DataNode" crashes in a Hadoop cluster. Another striking feature of Hadoop Framework is the *ease of scale* in accordance with the rapid growth in data volume. Because of these two reasons, one of the most common task of a Hadoop administrator is to commission (Add) and decommission (Remove) "Data Nodes" in a Hadoop Cluster.

Read this blog to get a detailed understanding on **commissioning and decommissioning nodes** (<https://www.edureka.co/blog/commissioning-and-decommissioning-nodes-in-a-hadoop-cluster/>) in a Hadoop cluster.

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

HDFS supports exclusive writes only.

When the first client contacts the "NameNode" to open the file for writing, the "NameNode" grants a lease to the client to create this file. When the second client tries to open the same file for writing, the "NameNode" will notice that the lease for the file is already granted to another client, and will reject the open request for the second client.

11. How does NameNode tackle DataNode failures?

NameNode periodically receives a Heartbeat (signal) from each of the DataNode in the cluster, which implies DataNode is functioning properly.

A block report contains a list of all the blocks on a DataNode. If a DataNode fails to send a heartbeat message, after a specific period of time it is marked dead.

The NameNode replicates the blocks of dead node to another DataNode using the replicas created earlier.

12. What will you do when NameNode is down?

The NameNode recovery process involves the following steps to make the Hadoop cluster up and running:

1. Use the file system metadata replica (FsImage) to start a new NameNode.
2. Then, configure the DataNodes and clients so that they can acknowledge this new NameNode, that is started.
3. Now the new NameNode will start serving the client after it has completed loading the last checkpoint FsImage (for metadata information) and received enough block reports from the DataNodes.

Whereas, on large Hadoop clusters this NameNode recovery process may consume a lot of time and this becomes even a greater challenge in the case of the routine maintenance. Therefore, we have HDFS High Availability Architecture which is covered in the **HA architecture** (<https://www.edureka.co/blog/how-to-set-up-hadoop-cluster-with-hdfs-high-availability/>) blog.

13. What is a checkpoint?

In brief, "Checkpointing" is a process that takes an FsImage, edit log and compacts them into a new FsImage. Thus, instead of replaying an edit log, the NameNode can load the final in-memory state directly from the FsImage. This is a far more efficient operation and reduces NameNode startup time. Checkpointing is performed by Secondary NameNode.

14. How is HDFS fault tolerant?

When data is stored over HDFS, NameNode replicates the data to several DataNode. The default replication factor is 3. You can change the configuration factor as per your need. If a DataNode goes down, the NameNode will automatically copy the data to another node from the replicas and make the data available. This provides fault tolerance in HDFS.

15. Can NameNode and DataNode be a commodity hardware?

The smart answer to this question would be, DataNodes are commodity hardware like personal computers and laptops as it stores data and are required in a large number. But from your experience, you can tell that, NameNode is the master node and it stores metadata about all the blocks stored in HDFS. It requires high memory (RAM) space, so NameNode needs to be a high-end machine with good memory space.

16. Why do we use HDFS for applications having large data sets and not when there are a lot of small files?

HDFS is more suitable for large amounts of data sets in a single file as compared to small amount of data spread across multiple files. As you know, the NameNode stores the metadata information regarding the file system in the RAM. Therefore, the amount of memory produces a limit to the number of files in my HDFS file system. In other words, too many files will lead to the generation of too much metadata. And, storing these metadata in the RAM will become a challenge. As a thumb rule, metadata for a file, block or directory takes 150 bytes.

Check Out Our Hadoop Course

(<https://www.edureka.co/big-data-and-hadoop>)

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

Blocks are the nothing but the smallest continuous location on your hard drive where data is stored. HDFS stores each as blocks, and distribute it across the Hadoop cluster. Files in HDFS are broken down into block-sized chunks, which are stored as independent units.

- Hadoop 1 default block size: 64 MB
- Hadoop 2 default block size: 128 MB

Yes, blocks can be configured. The dfs.block.size parameter can be used in the hdfs-site.xml file to set the size of a block in a Hadoop environment.

18. What does 'jps' command do?

The 'jps' command helps us to check if the Hadoop daemons are running or not. It shows all the Hadoop daemons i.e namenode, datanode, resourcemanager, nodemanager etc. that are running on the machine.

19. How do you define "Rack Awareness" in Hadoop?

Rack Awareness is the algorithm in which the "NameNode" decides how blocks and their replicas are placed, based on rack definitions to minimize network traffic between "DataNodes" within the same rack. Let's say we consider replication factor 3 (default), the policy is that "for every block of data, two copies will exist in one rack, third copy in a different rack". This rule is known as the "Replica Placement Policy".

To know rack awareness in more detail, refer to the **HDFS architecture** (<https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/>) blog.

20. What is "speculative execution" in Hadoop?

If a node appears to be executing a task slower, the master node can redundantly execute another instance of the same task on another node. Then, the task which finishes first will be accepted and the other one is killed. This process is called "speculative execution".

21. How can I restart "NameNode" or all the daemons in Hadoop?

This question can have two answers, we will discuss both the answers. We can restart NameNode by following methods:

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

These script files reside in the sbin directory inside the Hadoop directory.

22. What is the difference between an “HDFS Block” and an “Input Split”?

The “HDFS Block” is the physical division of the data while “Input Split” is the logical division of the data. HDFS divides data in blocks for storing the blocks together, whereas for processing, MapReduce divides the data into the input split and assign it to mapper function.

23. Name the three modes in which Hadoop can run.

The three modes in which Hadoop can run are as follows:

1. **Standalone (local) mode:** This is the default mode if we don't configure anything. In this mode, all the components of Hadoop, such as NameNode, DataNode, ResourceManager, and NodeManager, run as a single Java process. This uses the local filesystem.
2. **Pseudo-distributed mode:** A single-node Hadoop deployment is considered as running Hadoop system in pseudo-distributed mode. In this mode, all the Hadoop services, including both the master and the slave services, were executed on a single compute node.
3. **Fully distributed mode:** A Hadoop deployment in which the Hadoop master and slave services run on separate nodes, are stated as fully distributed mode.

Check out more questions on HDFS

(<https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hdfs-2/>)

Hadoop MapReduce Interview Questions**24. What is “MapReduce”? What is the syntax to run a “MapReduce” program?**

It is a framework/a programming model that is used for processing large data sets over a cluster of computers using parallel programming. The syntax to run a MapReduce program is **hadoop_jar_file.jar /input_path /output_path**.

If you have any doubt in MapReduce or want to revise your concepts you can refer this **MapReduce tutorial** (<https://www.edureka.co/blog/mapreduce-tutorial/>).

25. What are the main configuration parameters in a “MapReduce” program?

The main configuration parameters which users need to specify in “MapReduce” framework are:

- Job's input locations in the distributed file system
- Job's output location in the distributed file system
- Input format of data
- Output format of data
- Class containing the map function
- Class containing the reduce function
- JAR file containing the mapper, reducer and driver classes

26. State the reason why we can't perform “aggregation” (addition) in mapper? Why do we need the “reducer” for this?

This answer includes many points, so we will go through them sequentially.

- We cannot perform “aggregation” (addition) in mapper because sorting does not occur in the “mapper” function. Sorting occurs only on the reducer side and without sorting aggregation cannot be done.
- During “aggregation”, we need the output of all the mapper functions which may not be possible to collect in the map phase as mappers may be running on the different machine where the data blocks are stored.
- And lastly, if we try to aggregate data at mapper, it requires communication between all mapper functions which may be running on different machines. So, it will consume high network bandwidth and can cause network bottlenecking.

27. What is the purpose of “RecordReader” in Hadoop?

The “InputSplit” defines a slice of work, but does not describe how to access it. The “RecordReader” class loads the data from its source and converts it into (key, value) pairs suitable for reading by the “Mapper” task. The “RecordReader” instance is defined by the “Input Format”.

28. Explain “Distributed Cache” in a “MapReduce Framework”.

Distributed Cache can be explained as, a facility provided by the MapReduce framework to cache files needed by applications. Once you have cached a file for your job, Hadoop framework will make it available on each and every data nodes where you map/reduce tasks are running. Then you can access the cache file as a local file in your Mapper or Reducer job.

29. How do “reducers” communicate with each other?

This is a tricky question. The “MapReduce” programming model does not allow “reducers” to communicate with each other. “Reducers” run in isolation.

30. What does a “MapReduce Partitioner” do?

A “MapReduce Partitioner” makes sure that all the values of a single key go to the same “reducer”, thus allowing even distribution of the map output over the “reducers”. It redirects the “mapper” output to the “reducer” by determining which “reducer” is responsible for the particular key.

31. How will you write a custom partitioner?

Custom partitioner for a Hadoop job can be written easily by following the below steps:

- Create a new class that extends Partitioner Class
- Override method – getPartition, in the wrapper that runs in the MapReduce.
- Add the custom partitioner to the job by using method set Partitioner or add the custom partitioner to the job as a config file.

32. What is a “Combiner”?

A “Combiner” is a mini “reducer” that performs the local “reduce” task. It receives the input from the “mapper” on a particular “node” and sends the output to the https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/?utm_source=hd&utm_campaign=lms_hd_120616&utm_m... 5/13

A Combiner is a mini reducer that performs the local reduce task. It receives the input from the mapper on a particular node and sends the output to the "reducer". "Combiners" help in enhancing the efficiency of "MapReduce" by reducing the quantum of data that is required to be sent to the "reducers".

33. What do you know about "SequenceFileInputFormat"?

"SequenceFileInputFormat" is an input format for reading within sequence files. It is a specific compressed binary file format which is optimized for passing the data between the outputs of one "MapReduce" job to the input of some other "MapReduce" job.

Sequence files can be generated as the output of other MapReduce tasks and are an efficient intermediate representation for data that is passing from one MapReduce job to another.

More Questions on MapReduce

(<https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-mapreduce/>)

Apache Pig Interview Questions

34. What are the benefits of Apache Pig over MapReduce?

Apache Pig is a platform, used to analyze large data sets representing them as data flows developed by Yahoo. It is designed to provide an abstraction over MapReduce, reducing the complexities of writing a MapReduce program.

- Pig Latin is a high-level data flow language, whereas MapReduce is a low-level data processing paradigm.
- Without writing complex Java implementations in MapReduce, programmers can achieve the same implementations very easily using Pig Latin.
- Apache Pig reduces the length of the code by approx 20 times (according to Yahoo). Hence, this reduces the development period by almost 16 times.
- Pig provides many built-in operators to support data operations like joins, filters, ordering, sorting etc. Whereas to perform the same function in MapReduce is a humongous task.
- Performing a Join operation in Apache Pig is simple. Whereas it is difficult in MapReduce to perform a Join operation between the data sets, as it requires multiple MapReduce tasks to be executed sequentially to fulfill the job.
- In addition, pig also provides nested data types like tuples, bags, and maps that are missing from MapReduce.

35. What are the different data types in Pig Latin?

Pig Latin can handle both atomic data types like int, float, long, double etc. and complex data types like tuple, bag and map.

Atomic data types: Atomic or scalar data types are the basic data types which are used in all the languages like string, int, float, long, double, char[], byte[].

Complex Data Types: Complex data types are Tuple, Map and Bag.

To know more about these data types, you can go through our **Pig tutorial** (<https://www.edureka.co/blog/pig-tutorial/>) blog.

36. What are the different relational operations in "Pig Latin" you worked with?

Different relational operators are:

1. for each
2. order by
3. filters
4. group
5. distinct
6. join
7. limit

37. What is a UDF?

If some functions are unavailable in built-in operators, we can programmatically create User Defined Functions (UDF) to bring those functionalities using other languages like Java, Python, Ruby, etc. and embed it in Script file.

Check out Pig Interview Questions

(<https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-pig/>)

Apache Hive Interview Questions

38. What is "SerDe" in "Hive"?

Apache Hive is a data warehouse system built on top of Hadoop and is used for analyzing structured and semi-structured data developed by Facebook. Hive abstracts the complexity of Hadoop MapReduce.

The "SerDe" interface allows you to instruct "Hive" about how a record should be processed. A "SerDe" is a combination of a "Serializer" and a "Deserializer". "Hive" uses "SerDe" (and "FileFormat") to read and write the table's row.

To know more about Apache Hive, you can go through this **Hive tutorial** (<https://www.edureka.co/blog/hive-tutorial/>) blog.

39. Can the default "Hive Metastore" be used by multiple users (processes) at the same time?

"Derby database" is the default "Hive Metastore". Multiple users (processes) cannot access it at the same time. It is mainly used to perform unit tests.

40. What is the default location where "Hive" stores table data?

The default location where Hive stores table data is inside HDFS in /user/hive/warehouse.

Check out Hive Interview Questions

(<https://www.edureka.co/blog/interview-questions/hive-interview-questions/>)

Apache HBase Interview Questions

41. What is Apache HBase?

HBase is an open source, multidimensional, distributed, scalable and a NoSQL database written in Java. HBase runs on top of HDFS (Hadoop Distributed File System) and provides BigTable (Google) like capabilities to Hadoop. It is designed to provide a fault-tolerant way of storing the large collection of sparse data sets. HBase achieves high throughput and low latency by providing faster Read/Write Access on huge datasets.

To know more about HBase you can go through our **HBase tutorial** (<https://www.edureka.co/blog/hbase-tutorial>) blog.

42. What are the components of Apache HBase?

HBase has three major components, i.e. HMaster Server, HBase RegionServer and Zookeeper.

- **Region Server:** A table can be divided into several regions. A group of regions is served to the clients by a Region Server.
- **HMaster:** It coordinates and manages the Region Server (similar as NameNode manages DataNode in HDFS).
- **ZooKeeper:** Zookeeper acts like as a coordinator inside HBase distributed environment. It helps in maintaining server state inside the cluster by communicating through sessions.

To know more, you can go through this **HBase architecture** (<https://www.edureka.co/blog/hbase-architecture/>) blog.

43. What are the components of Region Server?

The components of a Region Server are:

- **WAL:** Write Ahead Log (WAL) is a file attached to every Region Server inside the distributed environment. The WAL stores the new data that hasn't been persisted or committed to the permanent storage.
- **Block Cache:** Block Cache resides in the top of Region Server. It stores the frequently read data in the memory.
- **MemStore:** It is the write cache. It stores all the incoming data before committing it to the disk or permanent memory. There is one MemStore for each column family in a region.
- **HFile:** HFile is stored in HDFS. It stores the actual cells on the disk.

44. Explain "WAL" in HBase?

Write Ahead Log (WAL) is a file attached to every Region Server inside the distributed environment. The WAL stores the new data that hasn't been persisted or committed to the permanent storage. It is used in case of failure to recover the data sets.

45. Mention the differences between "HBase" and "Relational Databases"?

HBase is an open source, multidimensional, distributed, scalable and a NoSQL database written in Java. HBase runs on top of HDFS and provides BigTable like capabilities to Hadoop. Let us see the differences between HBase and relational database.

HBase vs. Relational Database

HBase	Relational Database
It is schema-less	It is schema-based database
It is column-oriented data store	It is row-oriented data store
It is used to store de-normalized data	It is used to store normalized data
It contains sparsely populated tables	It contains thin tables
Automated partitioning is done in HBase	There is no such provision or built-in support for partitioning

Check out HBase Interview Questions

(<https://www.edureka.co/blog/interview-questions/hbase-interview-questions/>)

Apache Spark Interview Questions

46. What is Apache Spark?

The answer to this question is, Apache Spark is a framework for real-time data analytics in a distributed computing environment. It executes in-memory computations to increase the speed of data processing.

It is 100x faster than MapReduce for large-scale data processing by exploiting in-memory computations and other optimizations.

Yes, one can build "Spark" for a specific Hadoop version. Check out this blog to learn more about **building YARN and HIVE on Spark** (<https://www.edureka.co/blog/yarn-hive-get-electrified-by-spark/>).

RDD is the acronym for Resilient Distribution Datasets – a fault-tolerant collection of operational elements that run parallel. The partitioned data in RDD are immutable and distributed, which is a key component of Apache Spark.

(<https://www.edureka.co/blog/interview-questions/top-apache-spark-interview-questions-2016/>)

Apache ZooKeeper coordinates with various services in a distributed environment. It saves a lot of time by performing synchronization, configuration maintenance, grouping and naming.

- **Oozie Workflow:** These are the sequential set of actions to be executed. You can assume it as a relay race. Where each athlete waits for the last one to complete his part.
- **Oozie Coordinator:** These are the Oozie jobs which are triggered when the data is made available to it. Think of this as the response-stimuli system in our body. In the same manner, as we respond to an external stimulus, an Oozie coordinator responds to the availability of data and it rests otherwise.

"Oozie" is integrated with the rest of the Hadoop stack supporting several types of Hadoop jobs such as "Java MapReduce", "Streaming MapReduce", "Pig", "Hive" and "Sqoop".

Feeling overwhelmed with all the questions the interviewer might ask in your Hadoop interview? Now it is time to go through a series of Hadoop interview questions which covers different aspects of the Hadoop framework. It's never too late to strengthen your basics. Learn Hadoop from industry experts while working with real-life use cases.

(<https://www.edureka.co/big-data-and-hadoop>)



Shubham Sinha is a Big Data and Hadoop expert working as a Research Analyst at Edureka. He is keen to work with Big Data related technologies such as Hadoop, Spark, Flink and Storm and web development technologies including Angular, Node.js & PHP.



(https://www.facebook.com/hadoopcommunity/)
u=https://www.facebook.com/hadoopcommunity/

Share on

NEXT >

Enter your Email Address

SUBSCRIBE

Related Posts



Hadoop Tutorial: All you need to know about Hadoop!
👁 97K

(<https://www.edureka.co/blog/hadoop-tutorial/>)



10 Reasons Why Big Data Analytics is the Best Career Move
👁 256.5K

(<https://www.edureka.co/blog/10-reasons-why-big-data-analytics-is-the-best-career-move/>)



Big Data In Healthcare: How Hadoop Is Revolutionizing Healthcare Analytics
👁 4.3K

(<https://www.edureka.co/blog/hadoop-big-data-in-healthcare/>)

Browse Categories

Big Data NoSQL (<https://www.edureka.co/blog/category/big-data-nosql/>)

Blockchain (<https://www.edureka.co/blog/category/blockchain/>)

Business Intelligence (<https://www.edureka.co/blog/category/business-intelligence/>)

Cloud Computing (<https://www.edureka.co/blog/category/cloud-computing/>)

Cyber Security (<https://www.edureka.co/blog/category/cyber-security/>)

Deep Learning (<https://www.edureka.co/blog/category/deep-learning/>)

Finance (<https://www.edureka.co/blog/category/finance/>)

Frameworks (<https://www.edureka.co/blog/category/frameworks/>)

Marketing (<https://www.edureka.co/blog/category/marketing/>)

Mobile Development (<https://www.edureka.co/blog/category/mobile-development/>)

Operations (<https://www.edureka.co/blog/category/operations/>)

Programming (<https://www.edureka.co/blog/category/programming/>)

Project Management (<https://www.edureka.co/blog/category/project-management/>)

Robotic Process Automation (<https://www.edureka.co/blog/category/robotic-process-automation/>)

Success Story (<https://www.edureka.co/blog/category/success-story/>)

Systems & Architecture (<https://www.edureka.co/blog/category/systems-architecture/>)

Systems Engineering (<https://www.edureka.co/blog/category/systems-engineering/>)

Testing (<https://www.edureka.co/blog/category/testing/>)

Comments

26 Comments

26 Comments <https://www.edureka.co/blog/>

Rajiv Chaudhuri ▾

👍 Recommend 6 🐦 Tweet 📄 Share

Sort by Best ▾



Join the discussion...

Jignesh Solanki • 2 years ago ▾ ▹

Sincerely Thank you Edureka !! It is great compilation of the key points in the form of interview question / answers. It is really very useful and handy, It will serve as anytime reference point :) Enjoyed reading it.

46 ^ | ▾ • Reply • Share ▹

EdurekaSupport Mod ➔ Jignesh Solanki • 2 years ago ▾ ▹

Hey Jignesh, thanks for the wonderful feedback! We're glad we could help. :) Do subscribe to our blog to stay updated on upcoming posts and do spread the word. Cheers!

^ | ▾ • Reply • Share ▹

Md Ansari • a month ago ▾ ▹

thanks so much

1 ^ | ▾ • Reply • Share ▹

vaijayanthi mala • 7 months ago ▾ ▹

Thanks a lot for effort and time in coming up with all this content. The way it is written with focus on interview in mind is much appreciated. .

1 ^ | ▾ • Reply • Share ▹

EdurekaSupport Mod ➔ vaijayanthi mala • 6 months ago ▾ ▹

Hey Vaijayanthi, thank you for appreciating our work. Do browse through our channel to find more such tutorials! Cheers :)

^ | ▾ • Reply • Share ▹

Jignesh Solanki • 2 years ago ▾ ▹

Sincerely Thank you Edureka !! It is great compilation of the key points in the form of interview question / answers. It is really very useful and handy, It will serve as anytime reference point :) Enjoyed reading it.

1 ^ | v • Reply • Share ›

EdurekaSupport Mod → Jignesh Solanki • 2 years ago

Hey Jignesh, thanks for checking out our blog. We're glad you found the compilation useful! You can check out more interview questions on Hive, HDFS, MapReduce, Pig and HBase here: <https://www.edureka.co/blog/....> Hope this helps. Cheers!

^ | v • Reply • Share ›

Kanha Shukla • 2 years ago

Thank you so much . I spend the whole day on this blog in order to go through all of its content properly, Really great piece of work. thanks a lot. please keep up the practice.
some more questions on spark and GOGGLE DREMEL will be a real great amendment.

sincere thanks anyway

1 ^ | v • Reply • Share ›

EdurekaSupport Mod → Kanha Shukla • 2 years ago

Hey Kanha, thanks for checking out the blog and for the wonderful feedback! We're glad you found it useful. We have communicated your feedback to the relevant team and will incorporate it soon. Meanwhile, do check out this blog: <http://www.edureka.co/blog/....> We thought you might find it relevant. Cheers!

^ | v • Reply • Share ›

Kanha Shukla → EdurekaSupport • 2 years ago

Sure and Thanks , But that would be great if you can really find me a recruiter who is willing to hire a fresher provided I come up to his mark.

41 ^ | v • Reply • Share ›

EdurekaSupport Mod → Kanha Shukla • 2 years ago

Hey Kanha, we do not provide placement services. Having said that, we can assure you that since our Big Data and Hadoop certification course is widely recognized in the industry, you can definitely get a leg up by completing the course. Please take a look: <http://www.edureka.co/big-d....>

^ | v • Reply • Share ›

Ashish Jain • 2 years ago

Thanks, Its a good selection. I wish more interview questions on Spark.

1 ^ | v • Reply • Share ›

EdurekaSupport Mod → Ashish Jain • 2 years ago

Hey Ashish, thanks for checking out the blog! We're glad you found it useful. We will definitely come up with more Spark-related interview questions. Do subscribe to our blog to stay posted. Cheers!

^ | v • Reply • Share ›

vinodh • 3 years ago

Thanks

1 ^ | v • Reply • Share ›

**D Lusk** • 2 years ago

I am beginning learning hadoop, and this will help me with my studies

1 ^ | v • Reply • Share ›

EdurekaSupport Mod → D Lusk • 2 years ago

+D Lusk, thanks for checking out our blog. We're glad we could help. Here's another blog that will help you get the basics of Hadoop right: <https://goo.gl/i80FqY>. Please feel free to write to us if you have any questions. Cheers!

^ | v • Reply • Share ›

Kokila • a year ago

useful interview preparation questions and answers. Really thanks for this article. keep share still more tips. <http://zenrays.com/data-ana...>

^ | v • Reply • Share ›

sasanka ghosh • a year ago

Thanks for doing the hard work . Kudos . But some glaring error in concepts between RDBMS and BIG Data . Long Long way to go for current big data technology to challenge RDBMS in 300-400 TB range OLAP and ACID OLTP. Current RDBMS can sclae up as well as scale OUT in OLTP also

^ | v • Reply • Share ›

santhosh kumar • 2 years ago

Thanks for the info, will this cover entire hadoop framework ? if not please share the link it will be helpfull.

^ | v • Reply • Share ›

EdurekaSupport Mod → santhosh kumar • 2 years ago

Hey Santhosh, thanks for checking out our blog. Could you please elaborate on your query? Do you mean to ask if our course covers the entire Hadoop framework? If that's what you mean to ask, yes, our course covers HDFS, Hadoop MapReduce, Yarn, Pig, Hive, HBase, Oozie, and Spark (intro). You can check out more details here: <https://goo.gl/DaL5Ym>. Storm and Kafka are full- fledged courses which we also offer. Hope this helps. Cheers!

^ | v • Reply • Share ›

**S S Goutham** • 2 years ago

Thanks for your great article...

I have a question on Hive.. I need to insert 10,000 rows from un-partitioned table into partition table with two partition columns..To perform this task it is taking more time.. My Question is there any way to increase the mappers for that job to make the process fast as normal one...

^ | v • Reply • Share ›


EdurekaSupport Mod  S S Goutham • 2 years ago

Hey Goutham, thanks for checking out our blog. To answer your query, we can set/increase the number of mappers in mapred-site.xml Or we can set manually in program by using the below property.
`conf.setNumMapTasks(int num);`
 Any one can increase the mappers - either developer or admin - but, that is totally depends on the cluster and cpu cores. For more information on this, you can refer to the below given links.
<http://ask.fclose.com/375/h...>
<http://wiki.apache.org/hado...>

Hope this helps. Cheers!

  • Reply • Share ›**ronny** • 2 years ago

I Am 28 Now!! I Have worked in an small it company as a java developer!! Then i have prepared for ibps, so now any chances for me to get a big data job if i trained from any institute!! Or year gap of 4 Years makes obstacles for big data job

  • Reply • Share ›**EdurekaSupport** Mod  ronny • 2 years ago


Hey Ronny, thanks for checking out the blog! Your age and experience will not be an obstacle if you have the right skill sets. You can get a good start with the Edureka Hadoop course which not only equips you with industry relevant skills but also trains you in practical components. Also, once your live project is complete, you will be awarded with a course completion certificate that is well recognized in the industry. You can check out the course details here:
<http://www.edureka.co/big-d....> Please write to us if you have any further questions. Cheers!

  • Reply • Share ›**Pradeep Reddy** • 2 years ago

Very nice collection of questions, thank you.

  • Reply • Share ›**EdurekaSupport** Mod  Pradeep Reddy • 2 years ago

We are happy we could help. Thanks for taking the time out to check out our blog. Do keep coming back as we put up new blogs every week on all your favorite topics.

  • Reply • Share ›ALSO ON [HTTPS://WWW.EDUREKA.CO/BLOG/](https://www.edureka.co/blog/)**Machine Learning with R for Beginners with Example**

4 comments • 4 months ago

purushottam kumar — Hi , your blog is very good and easy to understand. can u please send me or give me the link of above classification data set ...

How To Install Kubernetes Cluster On Ubuntu 16.04

9 comments • 4 months ago

Pramod Lawate — I got below error while executing below command# kubeadm init --apiserver-advertise-address=198.168.56.100 ...

K-Nearest Neighbors Algorithm Using Python

1 comment • 2 months ago

mehru — thank you so much. This article really helped...i want to implement this on mnist dataset..can you help

How Blockchain Technology Works? Step by Step Guide for Beginners

2 comments • 6 months ago


ICO Development — Hey, great write up this is exactly true. Thanks for sharing blog, I think most people unaware about blockchain and its ...


 [Subscribe](#)  [Add Disqus to your site](#)[Add Disqus](#)[Add](#)  [Disqus' Privacy Policy](#)[Privacy Policy](#)[Privacy Policy](#)


Subscribe
to our newsletter


Enter your Email Address

SUBSCRIBE

- 


Top Hadoop Interview Questions To Prepare In 2018 – HDFS (<https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hdfs-2/>)
- 


Hadoop MapReduce Interview Questions In 2018 (<https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-mapreduce/>)
- 


Hadoop Tutorial: All you need to know about Hadoop! (<https://www.edureka.co/blog/hadoop-tutorial/>)
- 


10 Reasons Why Big Data Analytics is the Best Career... (<https://www.edureka.co/blog/10-reasons-why-big-data-analytics-is-the-best-career-move>)

Big Data Analytics Courses

- 

Big Data Hadoop Certification Training (</big-data-and-hadoop>)
- 

Python Spark Certification Training using PySpark (</pyspark-certification-training>)
- 

Hadoop Administration Certification Training (</hadoop-admin>)
- 

Apache Kafka Certification Training (</apache-kafka>)

Edureka

About us
(<https://www.edureka.co/about-us>)

News & Media
(<https://www.edureka.co/all-media>)

Contact us
(<https://www.edureka.co/contact-us>)

Blog
(<https://www.edureka.co/blog/>)

Work with us

Careers
(<https://www.edureka.co/careers>)

Become an Instructor
(<https://www.edureka.co/instructor>)

Become an Affiliate
(<https://www.edureka.co/affiliate-program>)

Hire from Edureka
(<https://www.edureka.co/hire-from-edureka>)

Useful Links





Reviews
(<https://www.edureka.co/reviews>)

Terms & conditions
(<https://www.edureka.co/terms-and-conditions>)


Privacy policy
(<https://www.edureka.co/privacy-policy>)


Sitemap
(<https://www.edureka.co/sitemap>)

Follow us on



Learn on the GO!


(<https://itunes.apple.com/in/app/edureka/id1033145415?mt=8>)



Community
(<https://www.edureka.co/community>)

(<https://play.google.com/store/apps/details?id=co.edureka.app>)

edureka!
(<https://www.edureka.co>)
© 2014 Brain4ce Education Solutions Pvt. Ltd. All rights Reserved.

"PMP®", "PMI®", "PMI-ACP®" and "PMBOK®" are registered marks of the Project Management Institute, Inc.
MongoDB®, Mongo and the leaf logo are the registered trademarks of MongoDB, Inc.