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

	Blogs	Videos	Interview Questions
⌂ (https://www.edureka.co/blog/all)	(https://www.edureka.co/blog/)	(https://www.edureka.co/blog/videos/)	(https://www.edureka.co/blog/interview-questions/)

Top Hadoop Interview Questions To Prepare In 2018 – HDFS

🔖 Recommended by 248 users

Ashish Bakshi (https://www.edureka.co/blog/author/ashishbedureka-co/) | Jul 26,2018 |

Add to Bookmark (https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hdfs-2/)

99.2K   (https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hdfs-2/#comments-wrapper)   34
(https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hdfs-2/#disqus_thread)

📧 Email this Post   👁

Well, if you are looking for Hadoop HDFS interview questions and aspire to become a Hadoop Developer or Hadoop Admin, you have come to the right place. This list of interview questions on HDFS will prepare you to match the expectations of the employer. Before moving ahead in this Hadoop HDFS Interview Questions blog, let us know about the trends and demands in the field of Hadoop. You can check out *this skill report (https://www.edureka.co/skill-report)* which talks about the top technical skills to master in 2018.

Big Data and Hadoop Job Trends:

According to Forrester's prediction, the Big Data market will grow at nearly **13% rate** over the next **5 years**, which is more than **twice** as compared to the predicted growth of IT in general. Hadoop has proven itself to be a de facto in solving the problems associated with Big Data. Therefore, it is not surprising to see the huge demand for Hadoop professionals in the IT industry. Now, let us have a glance at the following image which shows the rising trend of the Hadoop jobs posting over the past few years:

Source: *indeed.com*

From the above image you can clearly infer that the Hadoop is the most sought after skill in the employment, not only today but, in the coming years as well. Therefore, it's high time to pull up your sleeves and prepare for the coming Hadoop interviews.

Hadoop Interview Questions and Answers | Edureka

Here, I have provided you a list of the most frequent asked Hadoop HDFS interview questions that will help you to grab the opportunities being generated in the field of Big Data and Hadoop.

Get Hadoop Certified Now!

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

hadoop)

Hadoop HDFS Interview Questions

1. What are the core components of Hadoop?

1. What are the core components of Hadoop?

Hadoop Core Components

Component	Description
HDFS	*Hadoop (https://www.edureka.co/blog/hadoop-tutorial/)* Distributed file system or *HDFS (https://www.edureka.co/blog/hdfs-tutorial)* is a Java-based distributed file system that allows us to store Big data across multiple nodes in a Hadoop cluster.
YARN	YARN is the processing framework in Hadoop that allows multiple data processing engines to manage data stored on a single platform and provide Resource management.

2. What are the key features of HDFS?

♣**Tip**: You should also explain the features briefly while listing different HDFS features.

Some of the prominent features of HDFS are as follows:

- **Cost effective and Scalable:** HDFS, in general, is deployed on a commodity hardware. So, it is very economical in terms of the cost of ownership of the project. Also, one can scale the cluster by adding more nodes.
- **Variety and Volume of Data:** HDFS is all about storing huge data i.e. Terabytes & Petabytes of data and different kinds of data. So, I can store any type of data into HDFS, be it structured, unstructured or semi structured.
- **Reliability and Fault Tolerance:** HDFS divides the given data into data blocks, replicates it and stores it in a distributed fashion across the Hadoop cluster. This makes HDFS very reliable and fault tolerant.
- **High Throughput:** Throughput is the amount of work done in a unit time. HDFS provides high throughput access to application data.

3. Explain the HDFS Architecture and list the various HDFS daemons in HDFS cluster?

While listing various HDFS daemons, you should also talk about their roles in brief. Here is how you should answer this question:

Apache Hadoop *HDFS Architecture* (https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/) follows a Master/Slave topology where a cluster comprises a single NameNode (Master node or daemon) and all the other nodes are DataNodes (Slave nodes or daemons). Following daemon runs in HDFS cluster:

- **NameNode:** It is the master daemon that maintains and manages the data block present in the DataNodes.
- **DataNode:** DataNodes are the slave nodes in HDFS. Unlike NameNode, DataNode is a commodity hardware, that is responsible of storing the data as blocks.
- **Secondary NameNode:** The Secondary NameNode works concurrently with the primary NameNode as a *helper daemon*. It performs checkpointing.

4. What is checkpointing in Hadoop?

Checkpointing is the process of combining the Edit Logs with the FsImage (File system Image). It is performed by the Secondary NameNode.

5. What is a NameNode in Hadoop?

The *NameNode (https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/)* is the master node that manages all the DataNodes (slave nodes). It records the metadata information regarding all the files stored in the cluster (on the DataNodes), e.g. The location of blocks stored, the size of the files, permissions, hierarchy, etc.

6. What is a DataNode?

DataNodes (https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/) are the slave nodes in HDFS. It is a commodity hardware that provides storage for the data. It serves the read and write request of the HDFS client.

7. Is Namenode machine same as DataNode machine as in terms of hardware?

Unlike the DataNodes, a NameNode is a highly available server that manages the File System Namespace and maintains the metadata information. Therefore, NameNode requires higher RAM for storing the metadata information corresponding to the millions of HDFS files in the memory, whereas the DataNode needs to have a higher disk capacity for storing huge data sets.

8. What is the difference between NAS (Network Attached Storage) and HDFS?

Here are the key differences between NAS and HDFS:

- 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 a service for storing and accessing files. Whereas Hadoop Distributed File System (HDFS) is a distributed file system 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.

9. What is the difference between traditional RDBMS and Hadoop?

This question seems to be very easy, but in an interview these simple questions matter a lot. So, here is how you can answer the very question:

	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.	RDBMS is based on 'schema on write' where schema validation is done before loading the	On the contrary, Hadoop follows the schema on

Write	schema validation is done before loading the data.	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 Trasanctional Processing) system.	Hadoop is used for Data discovery, data analytics or OLAP system.

10. What is throughput? How does HDFS provides good throughput?

Throughput is the amount of work done in a unit time. HDFS provides good throughput because:

- The HDFS is based on Write Once and Read Many Model, it simplifies the data coherency issues as the data written once can't be modified and therefore, provides high throughput data access.

- In Hadoop, the computation part is moved towards the data which reduces the network congestion and therefore, enhances the overall system throughput.

11. What is Secondary NameNode? Is it a substitute or back up node for the NameNode?

Here, you should also mention the function of the Secondary NameNode while answering the later part of this question so as to provide clarity:

A **Secondary NameNode** (https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/) is a helper daemon that performs checkpointing in HDFS. No, it is not a backup or a substitute node for the NameNode. It periodically, takes the edit logs (meta data file) from NameNode and merges it with the FsImage (File system Image) to produce an updated FsImage as well as to prevent the Edit Logs from becoming too large.

12. What do you mean by meta data in HDFS? List the files associated with metadata.

The metadata in HDFS represents the structure of HDFS directories and files. It also includes the various information regarding HDFS directories and files such as ownership, permissions, quotas, and replication factor.

♣**Tip:** While listing the files associated with metadata, give a one line definition of each metadata file.

There are two files associated with metadata present in the NameNode:

- **FsImage:** It contains the complete state of the file system namespace since the start of the NameNode.

- **EditLogs:** It contains all the recent modifications made to the file system with respect to the recent FsImage.

13. What is the problem in having lots of small files in HDFS?

As we know, the NameNode stores the metadata information regarding 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 much of files will lead to the generation of too much meta data and storing these meta data in the RAM will become a challenge. As a thumb rule, metadata for a file, block or directory takes 150 bytes.

14. What is a heartbeat in HDFS?

Heartbeats in HDFS are the signals that are sent by DataNodes to the NameNode to indicate that it is functioning properly (alive). By default, the heartbeat interval is 3 seconds, which can be configured using dfs.heartbeat.interval in hdfs-site.xml.

15. How would you check whether your NameNode is working or not?

There are many ways to check the status of the NameNode. Most commonly, one uses the **jps command** to check the status of all the daemons running in the HDFS. Alternatively, one can visit the NameNode's Web UI for the same.

16. What is a block?

You should begin the answer with a general definition of a block. Then, you should explain in brief about the blocks present in HDFS and also mention their default size.

Blocks (https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/) are the smallest continuous location on your hard drive where data is stored. HDFS stores each file as blocks, and distribute it across the Hadoop cluster. The default size of a block in HDFS is 128 MB (Hadoop 2.x) and 64 MB (Hadoop 1.x) which is much larger as compared to the Linux system where the block size is 4KB. The reason of having this huge block size is to minimize the cost of seek and reduce the meta data information generated per block.

17. Suppose there is file of size 514 MB stored in HDFS (Hadoop 2.x) using default block size configuration and default replication factor. Then, how many blocks will be created in total and what will be the size of each block?

Default block size in Hadoop 2.x is 128 MB. So, a file of size 514 MB will be divided into 5 blocks (514 MB/128 MB) where the first four blocks will be of 128 MB and the last block will be of 2 MB only. Since, we are using the default replication factor i.e. 3, each block will be replicated thrice. Therefore, we will have 15 blocks in total where 12 blocks will be of size 128 MB each and 3 blocks of size 2 MB each.

18. How to copy a file into HDFS with a different block size to that of existing block size configuration?

♣**Tip:** You should start the answer with the command for changing the block size and then, you should explain the whole procedure with an example. This is how you should answer this question:

Yes, one can copy a file into HDFS with a different block size by using '-Ddfs.blocksize=block_size' where the block_size is specified in Bytes.

Let me explain it with an example: Suppose, I want to copy a file called test.txt of size, say of 120 MB, into the HDFS and I want the block size for this file to be 32 MB (33554432 Bytes) instead of the default (128 MB). So, I would issue the following command:

hadoop fs -Ddfs.blocksize=33554432 -copyFromLocal /home/edureka/test.txt /sample_hdfs

Now, I can check the HDFS block size associated with this file by:

hadoop fs -stat %o /sample_hdfs/test.txt

Else, I can also use the NameNode web UI for seeing the HDFS directory.

♣**Tip:** You can go through the blog on **Hadoop Shell Commands** (https://www.edureka.co/blog/hdfs-commands-hadoop-shell-command) where you will find various Hadoop commands, explained with an example.

19. Can you change the block size of HDFS files?

Yes, I can change the block size of HDFS files by changing the default size parameter present in hdfs-site.xml. But, I will have to restart the cluster for this property change to take effect.

20. What is a block scanner in HDFS?

Block scanner runs periodically on every DataNode to verify whether the data blocks stored are correct or not. The following steps will occur when a corrupted data block is detected by the block scanner:

- First, the DataNode will report about the corrupted block to the NameNode.
- Then, NameNode will start the process of creating a new replica using the correct replica of the corrupted block present in other DataNodes.
- The corrupted data block will not be deleted until the replication count of the correct replicas matches with the replication factor (3 by default).

This whole process allows HDFS to maintain the integrity of the data when a client performs a read operation. One can check the block scanner report using the DataNode's web interface- **localhost:50075/blockScannerReport** as shown below:
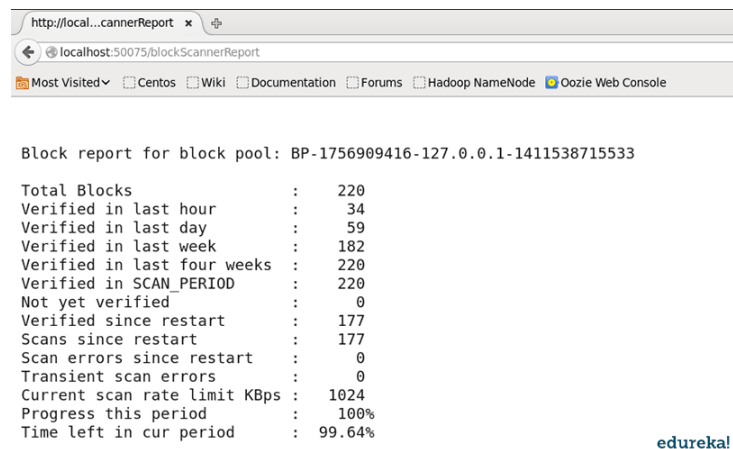
Fig. – *Block Scanner Report – Hadoop HDFS Interview Question*

Learn Hadoop from Industry Experts

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

hadoop)

21. HDFS stores data using commodity hardware which has higher chances of failures. So, How HDFS ensures the Fault Tolerance capability of the system?

♣**Tip:** Basically, this question is regarding replication of blocks in Hadoop and how it helps in providing fault tolerance.

HDFS provides fault tolerance by replicating the data blocks and distributing it among different DataNodes across the cluster. By default, this replication factor is set to 3 which is configurable. So, if I store a file of 1 GB in HDFS where the replication factor is set to default i.e. 3, it will finally occupy a total space of 3 GB because of the replication. Now, even if a DataNode fails or a data block gets corrupted, I can retrieve the data from other replicas stored in different DataNodes.

22. Replication causes data redundancy and consume a lot of space, then why is it pursued in HDFS?

Replication is pursued in HDFS to provide the fault tolerance. And, yes, it will lead to the consumption of a lot of space, but one can always add more nodes to the cluster if required. By the way, in practical clusters, it is very rare to have free space issues as the very first reason to deploy HDFS was to store huge data sets. Also, one can change the replication factor to save HDFS space or use different codec provided by Hadoop to compress the data.

23. Can we have different replication factor of the existing files in HDFS?

♣**Tip:** You should always answer such type of questions by taking an example to provide clarity.

Yes, one can have different **replication factor (https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/)** for the files existing in HDFS. Suppose, I have a file named test.xml stored within the sample directory in my HDFS with the replication factor set to 1. Now, the command for changing the replication factor of text.xml file to 3 is:

hadoop fs -setrwp -w 3 /sample/test.xml

Finally, I can check whether the replication factor has been changed or not by using following command:

hadoop fs -ls /sample

or

hadoop fsck /sample/test.xml -files

24. What is a rack awareness algorithm and why is it used in Hadoop?

Rack Awareness (https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/) algorithm in Hadoop ensures that all the block replicas are not stored on the same rack or a single rack. Considering the replication factor is 3, the Rack Awareness Algorithm says that the first replica of a block will be stored on a local rack and the next two replicas will be stored on a different (remote) rack but, on a different DataNode within that (remote) rack. There are two reasons for using Rack Awareness:

- **To improve the network performance:** In general, you will find *greater network bandwidth* between machines in the same rack than the machines residing in different rack. So, the Rack Awareness helps to reduce write traffic in between different racks and thus provides a better write performance.
- **To prevent loss of data:** I don't have to worry about the data even if an entire rack fails because of the switch failure or power failure. And if one thinks about it, it will make sense, as it is said that *never put all your eggs in the same basket.*

25. How data or a file is written into HDFS?

The best way to answer this question is to take an example of a client and list the steps that will happen while performing the write without going into much of the details:

Suppose a client wants to write a file into HDFS. So, the following steps will be performed internally during the whole HDFS write process:

- The client will divide the files into blocks and will send a write request to the NameNode.
- For each block, the NameNode will provide the client a list containing the IP address of DataNodes (depending on replication factor, 3 by default) where the data block has to be copied eventually.
- The client will copy the first block into the first DataNode and then the other copies of the block will be replicated by the DataNodes themselves in a sequential manner.

♣**Tip:** I would suggest you to go through the blog on **HDFS Read/Write Architecture (https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/)** where the whole process of HDFS Read/Write has been explained in detail with images.

Do you prefer watching a video tutorial to understand & prepare yourself for your Hadoop interview? Here is our video on the top 50 Hadoop interview questions.

26. Can you modify the file present in HDFS?

No, I cannot modify the files already present in HDFS, as HDFS follows Write Once Read Many model. But, I can always append data into the existing HDFS file.

27. Can multiple clients write into an HDFS file concurrently?

No, multiple clients can't write into an HDFS file concurrently. HDFS follows single writer multiple reader model. The client which opens a file for writing is granted a lease by the NameNode. Now suppose, in the meanwhile, some other client wants to write into that very file and asks NameNode for the write permission. At first, the NameNode will check whether the lease for writing into that very particular file has been granted to someone else or not. Then, it will reject the write request of the other client if the lease has been acquired by someone else, who is currently writing into the very file.

29. Does HDFS allow a client to read a file which is already opened for writing?

Basically, the intent of asking this question is to know about the constraints associated with reading a file which is currently being written by some client. You may answer this question in following manner:

Yes, one can read the file which is already opened. But, the problem in reading a file which is currently being written lies in the consistency of the data i.e. HDFS does not provide the surety that the data which has been written into the file will be visible to a new reader before the file has been closed. For this, one can call the hflush operation explicitly which will push all the data in the buffer into the write pipeline and then the hflush operation will wait for the acknowledgements from the DataNodes. Hence, by doing this the data that has been written into the file before the hflush operation will be visible to the readers for sure.

30. Define Data Integrity? How does HDFS ensure data integrity of data blocks stored in HDFS?

Data Integrity talks about the correctness of the data. It is very important for us to have a guarantee or assurance that the data stored in HDFS is correct. However, there is always a slight chance that the data will get corrupted during I/O operations on the disk. HDFS creates the checksum for all the data written to it and verifies the data with the checksum during read operation by default. Also, each DataNode runs a block scanner periodically, which verifies the correctness of the data blocks stored in the HDFS.

31. What do you mean by the High Availability of a NameNode? How is it achieved?

NameNode used to be single point of failure in Hadoop 1.x where the whole Hadoop cluster becomes unavailable as soon as NameNode is down. In other words, **High Availability (https://www.edureka.co/blog/how-to-set-up-hadoop-cluster-with-hdfs-high-availability/)** of the NameNode talks about the very necessity of a NameNode to be active for serving the requests of Hadoop clients.

To solve this Single Point of Failure problem of NameNode, HA feature was intorduced in Hadoop 2.x where we have two NameNode in our HDFS cluster in an active/passive configuration. Hence, if the active NameNode fails, the other passive NameNode can take over the responsibility of the failed NameNode and keep the HDFS up and running.

32. Define Hadoop Archives? What is the command for archiving a group of files in HDFS.

Hadoop Archive was introduced to cope up with the problem of increasing memory usage of the NameNode for storing the metadata information because of too many small files. Basically, it allows us to pack a number of small HDFS files into a single archive file and therefore, reducing the metadata information. The final archived file

follows the .har extension and one can consider it as a layered file system on top of HDFS.

The command for archiving a group of files:

hadoop archive –archiveName edureka_archive.har /input/location /output/location

33. How will you perform the inter cluster data copying work in HDFS?

One can perform the inter cluster data copy by using distributed copy command given as follows:

hadoop distcp hdfs://<source NameNode> hdfs://<target NameNode>

I hope you find this blog on Hadoop HDFS Interview questions very helpful and informative. For becoming a Hadoop expert, you need to work on various Hadoop related projects as well, apart from having sound theoretical knowledge. We have designed a curriculum which covers all the aspects of the Hadoop framework along with lots of hands on experience. You will be taught by the Industry Experts who will be sharing their precious experience of various Hadoop related projects.

Get Project Based Hadoop Training

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

hadoop)

Last but not the least, I would like to say that HDFS is only one of the many components present in Hadoop framework. We have created dedicated blogs for important interview questions corresponding to each Hadoop component. Therefore, I would suggest you to follow the links provided below and enjoy the readings:

- ***Top 50 Hadoop Interview Questions*** (https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/)
- ***Hadoop Cluster Interview Questions*** (https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hadoop-cluster/)
- ***Hadoop MapReduce Interview Questions*** (https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-mapreduce/)
- ***Pig Interview Questions (https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-pig/)***
- ***Hive Interview Questions*** (https://www.edureka.co/blog/interview-questions/hive-interview-questions/)
- ***HBase Interview Questions*** (https://www.edureka.co/blog/interview-questions/hbase-interview-questions/)

Got a question for us? Please mention it in the comments section of this Hadoop HDFS Interview Questions and we will get back to you.

About Ashish Bakshi (11 Posts (https://www.edureka.co/blog/author/ashishbedureka-co/))

Share on

(https://www.facebook.com/sharer/sharer.php?
u=https://www.edureka.co/blog/interview-
questions/hadoop-
interview-
questions-
hdfs-
2/)

◁ PREVIOUS   NEXT ▷

Got your brain cells running?
Stay tuned to latest technology updates

Enter your Email Address

SUBSCRIBE

Related Posts

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

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

Hadoop Developer-Job Responsibilities & Skills
👁 56K

(https://www.edureka.co/blog/hadoop-tutorial/)
(https://www.edureka.co/blog/10-reasons-why-big-data-analytics-is-the-best-career-move)
(https://www.edureka.co/blog/hadoop-developer-job-responsibilities-skills/)

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   **34 Comments**

34 Comments   **https://www.edureka.co/blog/**   ● **Rajiv Chaudhuri** ▾

♡ Recommend   ▾ Tweet   f Share   Sort by Best ▾

             Join the discussion…

     Apurv Krishan • 4 years ago
     This is a very useful list. Thank you :)
     1 ∧ | ∨ • Reply • Share ›

             EdurekaSupport Mod ↗ Apurv Krishan • 4 years ago
             You are welcome, Apurv. Feel free to go through our other blog posts as well.
             ∧ | ∨ • Reply • Share ›

     Tushar Bhoyar • 3 months ago
     When you are adding datanode, what is the impact of new blocks created by running jobs?
     ∧ | ∨ • Reply • Share ›

     RANJIT M • a year ago
     Command wrong at point 24. It should be setrep -w
     ∧ | ∨ • Reply • Share ›

     3JorgeSchuldt • 2 years ago
     Practical blog post ! I learned a lot from the facts - Does someone know where I could possibly get a sample MD GCAAR 1204 form to fill out ?
     ∧ | ∨ • Reply • Share ›

     Srikantha • 3 years ago
     Really a great initiative. Thanks a lot for providing such a good knowledge
     ∧ | ∨ • Reply • Share ›

             EdurekaSupport Mod ↗ Srikantha • 3 years ago
             Thanks, Srikantha! Check out the other posts on Hadoop Interview Questions as well.
             ∧ | ∨ • Reply • Share ›

     samsammy • 3 years ago
     What is the difference between input split and blocks and how these affect each other?
     ∧ | ∨ • Reply • Share ›

     bikash • 4 years ago
     How to join two table in hadoop?
     ∧ | ∨ • Reply • Share ›

             EdurekaSupport Mod ↗ bikash • 3 years ago
             Hi Bikash, For joining tables in Hadoop, we have mapsidejoin in hive. You can refer to post to know about it: http://www.edureka.co/blog/...
             ∧ | ∨ • Reply • Share ›

     Praveen • 4 years ago
     What actually is difference between Secondary Name Node and Passive Name Node?? or both are same???

What actually is difference between Secondary Name Node and Passive Name Node??? or both are same???
∧ | ∨ • Reply • Share ›

EdurekaSupport Mod → Praveen • 4 years ago

Hi Praveen,
the secondary namenode periodically pulls these two(Edits nd fsimage) files and the namenode starts writing changes to a new edits file. Then, the secondary namenode merges the changes from the edits file with the old snapshot from the fsimage file and creates an updated fsimage file. This updated fsimage file is then copied to the namenode.

FailOver NameNode or Passive Namenode(Only if HA enabled) is a follows. Prior to Hadoop 2.0.0, the NameNode was a single point of failure (SPOF) in an HDFS cluster. Each cluster had a single NameNode, and if that machine or process became unavailable, the cluster as a whole would be unavailable until the NameNode was either restarted or brought up on a separate machine. the above problems is addressed by providing the option of running two redundant NameNodes in an HA cluster, the Standby/Passive NameNode also performs checkpoints of the namespace state, and thus it is not necessary to run a Secondary nameNode, CheckpointNode, or BackupNode in an HA cluster. In fact, to do so would be an error.
∧ | ∨ • Reply • Share ›

sangeetha.r44 • 4 years ago

Thank you Edureka. You guys are doing great, i will spread the word.
∧ | ∨ • Reply • Share ›

EdurekaSupport Mod → sangeetha.r44 • 4 years ago

Thanks Sangeetha!!
∧ | ∨ • Reply • Share ›

Shobhit • 4 years ago

Thanks Edureka Support team, for providing very good to the point answers...
It would be helpful for us , if you can provide a bit more detail on Active and Passive Namenodes (their way of functioning).

Thanks once again.
∧ | ∨ • Reply • Share ›

EdurekaSupport Mod → Shobhit • 4 years ago

Hello Shobhit, we are glad that you found the points useful. We will consider your suggestions. Meanwhile, you can go through our other blog posts.
∧ | ∨ • Reply • Share ›

karuna Devanagavi • 4 years ago

The interview questions and answers are excellent.......If possible please upload the interview questions of Hive, HBase and Cassandra...... Thank you.
∧ | ∨ • Reply • Share ›

EdurekaSupport Mod → karuna Devanagavi • 4 years ago

Thanks Karuna! Will consider your suggestion. Mean while please go through our other blog posts as well.
∧ | ∨ • Reply • Share ›

jana • 4 years ago

Does hadoop has its own data type? If so why can't it use the java data types
∧ | ∨ • Reply • Share ›

EdurekaSupport Mod → jana • 4 years ago

Hi Jana, Below are the datatypes in Hadoop:
Primitive Datatypes: IntWritable, LongWritable , BooleanWritable,
FloatWritable, ByteWritable

Built-in data types : Text , ByteWritable, VIntWritable, VLongWritable,
Nullwritable

Hadoop use the datatypes in java writable versions.
javadatatype -> hadoop datatype
int -------->IntWritable
long -------->LongWritable
boolean -------->BooleanWritable
float -------->FloatWritable
byte --------> ByteWritable
∧ | ∨ • Reply • Share ›

Siva Ambati • 4 years ago

Hi Support Team, when i click for 'Hadoop Interview Questions – MapReduce!' i am getting the second list of questions, but not third list.

Please help me in this regard.
Thanks.
Siva
∧ | ∨ • Reply • Share ›

EdurekaSupport Mod → Siva Ambati • 4 years ago

Hi Siva, check the link again. It's directing to the 3rd list. Or you can click this link: http://www.edureka.in/blog/...
∧ | ∨ • Reply • Share ›

Siva Ambati → EdurekaSupport • 4 years ago
Hi Team

Hi Team,
Thank you so much for quick turnaround. Excellent. it is directing to third list now. Thank you so much.

21 ∧ | ∨ • Reply • Share ›

vignesh • 4 years ago

Hai,Excellent for that interviews questions this reviews very useful for my job interview preparation.

∧ | ∨ • Reply • Share ›

EdurekaSupport Mod → vignesh • 4 years ago

Thanks Vignesh!!

∧ | ∨ • Reply • Share ›

Yupeng • 5 years ago

[ ] to spend? Running a Hadoop cluster in-house or on EMR isn't cheap, and avdoiing some of a0its hassles is worth something too. The answer varied depending on the vendor, but word on the street was [ ]

∧ | ∨ • Reply • Share ›

NILESH NAYAK • 5 years ago

Nice

∧ | ∨ • Reply • Share ›

Krish Na • 5 years ago

thanks

∧ | ∨ • Reply • Share ›

Srimanth Babu • 5 years ago

thanks

∧ | ∨ • Reply • Share ›

PRAVEEN • 5 years ago

Hi,
I would like to congratulate and thank edureka for providing such crisp info about the emerging hadoop technology.
It would be great if someone can provide us the single node and multiple node cluster documents.

Thanks in Advance

Praveen

∧ | ∨ • Reply • Share ›

hadoop interview questions • 5 years ago

excellent piece of information, I had come to know about your website from my friend kishore, pune,i have read atleast 8 posts of yours by now, and let me tell you, your site gives the best and the most interesting information. This is just the kind of information that i had been looking for, i'm already your rss reader now and i would regularly watch out for the new posts, once again hats off to you! Thanx a lot once again, Regards, hadoop interview questions

∧ | ∨ • Reply • Share ›

Sunil • 5 years ago

The answers for the below two questions seem to be contradictitng each other, please check.

What is a Namenode?
Namenode is the master node on which job tracker runs............

Are Namenode and job tracker on the same host?
No, in practical environment, Namenode is on a separate host and job tracker is on a separate host.

∧ | ∨ • Reply • Share ›

Abdul Salam • 5 years ago

My 2 months doubts are cleared masha allha...
Really am apriciating the contents posted here...
Simple English .. suited for all...
I belived that edureka will shine more and more insha allha....

Am from chennai.. u can call me for any help to assist more..

Sr.Etl Developer

∧ | ∨ • Reply • Share ›

Deep • 4 years ago

how can we merge the data of two file written in diffrent data format like one is seprated by pipe and other one is by comma ...located at diffrent location.

∧ | ∨ • Reply • Share ›

ALSO ON **HTTPS://WWW.EDUREKA.CO/BLOG/**

Machine Learning with R for Beginners with Example

4 comments • 4 months ago

Avatar 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 …

Spring Boot Microservices: Building Microservices Application Using Spring Boot

135 comments • 5 months ago

Avatar Aditya C — Can you please send me the code to oraclesoatrainer@gmail.com

Top 10 Reasons Why You Should Learn Selenium

1 comment • 6 months ago

Avatar **Partho** — Since there is no Global Certification in Selenium what is the benchmark then.....

Python Pandas Tutorial : Learn Pandas for Data Analysis

1 comment • 6 months ago

Avatar **Amit Arora** — More than 100 Pandas Exercises Solutions Tricks Issues Examples (http://www.pythonprogrammin...

✉ Subscribe   Ⓓ Add Disqus to your siteAdd DisqusAdd   🔒 Disqus' Privacy PolicyPrivacy PolicyPrivacy

---

Subscribe
to our newsletter

Enter your Email Address

SUBSCRIBE

Related Blogs

Top 50 Hadoop Interview Questions You Must Prepare In 2018 (https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/)

(https://www.
questions/top
50-hadoop-
interview-
questions-
2016/)

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

(https://www.
questions/had
interview-
questions-
mapreduce/)
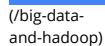
Hadoop Tutorial: All you need to know about Hadoop! (https://www.edureka.co/blog/hadoop-tutorial/)

(https://www.
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)

(https://www.
reasons-why-
big-data-
analytics-is-
the-best-
career-move)

Big Data Analytics Courses

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

(/big-data-
and-hadoop)

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

(/pyspark-
certification-
training)

Hadoop Administration Certification Training (/hadoop-admin)

(/hadoop-
admin)

Apache Kafka Certification Training (/apache-kafka)

(/apache-

Edureka

About us
(https://www.edureka.co/about-
us)
News & Media
(https://www.edureka.co/allmedia)
Contact us
(https://www.edureka.co/contact-
us)
Blog
(https://www.edureka.co/blog/all)
Community
(https://www.edureka.co/community)

Work with us

Careers
(https://www.edureka.co/careers)
Become an Instructor
(https://www.edureka.co/instructors)
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

f   in

(https://www.facebook.com/edurekaIN)(https://twitter.com/edurekaIN)(https://www.linkedin.com/company/edureka)(https://www.youtube.com/user/edurekaIN)

Learn on
the GO!

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

(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.