By OnlineInterviewQuestions.com

Hadoop Interview Questions

Best Hadoop Interview Questions and Answers

Hadoop is a framework for distributed processing of large data sets across the clusters of commodity computers.

Hadoop developers have a great market to capture as every other MNC is looking to recruit one. You can now build a career as a Hadoop developer and get placed in one of your dream companies. You can prepare for your interview and clear the biggest barrier. If you are looking for Hadoop HDFS questions and answers and seek to become a Hadoop Developer or Hadoop Admin, you have come to the right place. We have provided a list of the Hadoop Interview Questions that will prove to be useful. These are the most widely recognized and prominently asked **Big Data Hadoop Interview Questions**, which you will undoubtedly get in huge information interviews.

Getting ready through these **Hadoop Interview Questions and Answers** will without a doubt give you an edge in this competitive time.

Q1. What does HDFS mean?

The HDFS is one of the storage systems of the Hadoop structure. It is a circulated file structure that can helpfully keep running on item equipment for processing unstructured information. Because of this functionality of HDFS that is worked to keep running on commodity equipment, it can be very fault tolerant. Similar information is kept in numerous areas and in the case of one storage area neglecting to give the required information; similar information can be effortlessly fetched from another area.

Q2. What are the key features of HDFS?

Various key features of HDFS are as follows:

HDFS is a profoundly versatile and reliable storage system for big data stage Hadoop. Working intimately with Hadoop YARN for data handling and information analytics, it enhances the information administration layer of the Hadoop bunch making it sufficiently productive to process enormous information simultaneously. HDFS additionally works in close coordination with HBase. Here are some of the features, which make this technology quite special:

- Storage of bulk data
- Least intervention
- Computing
- Scaling out

- Rollback
- Information integrity

Q3. What is check pointing in Hadoop?

Check pointing is a fundamental part of keeping up and holding on file system metadata in HDFS. It’s urgent for proficient Name-Node recovery and restart, and is a vital indicator of general cluster health. However, check pointing can also because of confusion for operators of Apache Hadoop clusters. Check pointing is a procedure that takes a fsimage and alters log and compacts them into another fsimage. Along these lines, rather than replaying a possibly unbounded alter log, the NameNode can load the final in-memory state straightforwardly from the fsimage.

Q4. What does Name-Node mean in Hadoop?

The Name-Node is the center part of an HDFS document framework. It keeps the directory tree of all records in the file system and tracks where over the cluster the document information is kept. It doesn’t store the information of these records itself. The Name-Node is a Single mode of Failure for the HDFS Cluster. HDFS isn’t as of now a High Availability structure. At the point when the Name-Node goes down, the document structure goes disconnected. There is another optional Secondary NameNode that can be facilitated on a different machine. It just makes checkpoints of the namespace by merging the edits document into the fsimage document and does not give any genuine repetition.

Q5. What did mean by Data-Node?

Data-Node stores information in HDFS; it is a node where the real information resides in the document. Each data node sends a pulse message to notify that it is alive. If the name-node does not get a message from data-node for 10 minutes, it is considered to be dead or out of place and begins replication of obstructs that were facilitated on that information node with the end goal that they are facilitated on some other information node. A Block-Report consists of the list rundown of all blocks on a Data-Node.

Q6. What does heartbeat in HDFS means?

A heartbeat is an indication of the signal that it is alive. A data-node sends a pulse to Name-node and task tracker will send its heartbeat to job tracker. If the Name-node or job tracker does not get a heartbeat, then they will choose that there is some issue in data-node or task tracker is unable to perform the assigned task.

Q7. What does block mean?

Smallest consistent location on your hard drive where information is stored is known as a block. HDFS stores each document as blocks, and appropriate it over the Hadoop cluster. The default size of a square in HDFS is 128 MB (Hadoop 2.x) and 64 MB (Hadoop 1.x), which is considerably bigger when contrasted with the Linux system where the block size is 4KB. The reason of having this enormous square size is to limit the cost of look for and diminish the Meta information data created per block.

Q8. What does rack awareness algorithm means and why is it utilized as a part of Hadoop?

Rack Awareness algorithm in Hadoop guarantees that all the block copies are not stored on a similar rack or a solitary rack. Considering the reproduction factor is 3, the Rack Awareness Algorithm says that the primary replica of a block will be socked on a local rack and the following two replicas will be put away on an alternate (remote) rack at the same time, on an alternate Data-Node inside that (remote) rack. There are two purposes for utilizing Rack Awareness:

- **To enhance the network performance-:** You will discover more prominent system data transmission between machines in a similar rack than the machines living in various racks. In this way, the Rack Awareness helps to compose movement in the middle of various racks and therefore gives a superior write performance.
- **To prevent loss of data:** You need not worry about the information even if the whole rack fails because of switch failure or electrical power failure.

Q9. Characterize Data Integrity? How does HDFS ensure information integrity of data blocks squares kept in HDFS?

Data Integrity discusses the accuracy of the information. It is essential for us to have a guarantee or assurance that the information kept in HDFS is right. However, there is dependably a slight chance that the information will be corrupted during I/O tasks on the disks. HDFS makes the checksum for all of the information kept in touch with it and confirmed the information with the checksum during the read activity of course. Additionally, each Data-Node runs a block scanner occasionally, which checks the accuracy of the information blocks kept in the HDFS.

Q10. What is the throughput? How does HDFS give great throughput?

Throughput is the measure of work done in a unit time. HDFS gives great throughput because of the followings:

- The HDFS depends on Write Once and Read Many Model, it improves the information coherency issues as the information written once can’t be altered and consequently, gives high throughput data access.
- In Hadoop, the calculation part is moved towards the information which decreases the system blockage and in this way, improves the overall system throughput.

Q11. Clarify the difference between NAS and HDFS.

Here are some of the differences between NAS and HDFS:

- NAS keeps running on a single machine, and in this way, there is no probability of information repetition though HDFS keeps running on a cluster of machines subsequently there is information excess due to the replication convention.
- Hadoop HDFS is intended to work with Map Reduce structure. In Map Reduce structure calculation move to the data rather than Data to the calculation. NAS isn't appropriate for Map Reduce, as it stores information independently from the calculations.
- In HDFS, data blocks are appropriated over all of the machines in a group. Though in NAS, data is put away on devoted hardware.
- HDFS utilizes ware hardware, which is financially effective, though a NAS is a high-end storage gadget which incorporates high cost.

Q12. What does secondary name-node means?

Secondary Name-Node in Hadoop is a particularly devoted node in HDFS group whose primary function is to take checkpoints of the document structure metadata present on name-node. It's anything but a checkpoints name-node. It just checkpoints name node's file framework namespace. The Secondary NameNode is a helping hand to the primary Name-Node but not substitute for primary name-node.

Q13. What do you mean by Meta information in HDFS? List the documents related to metadata.

Name-Node Metadata stores the record for Block mapping, locations of blocks on DataNodes, dynamic data nodes, and much more metadata are altogether stored in memory on the Name-Node. When we check the Name-Node status site, basically the greater part of that data is kept in memory somewhere.

The main thing stored on disk is the fsimage, edit log, and status logs. Name-Node never truly utilizes these records on disk, aside from when it begins. The fsimage and edits record practically exist to have the capacity to bring the Name-Node back up if it should be halted or it crashes.

- **1) fsimage –** An fsimage document contains the entire state of the file system at a point in time. Each document system modification is doled out in a unique process, monotonically expanding transaction ID. A fsimage document speaks to the file system state after all alterations up to a particular transaction ID.
- **2) Edits –** An edited file is a log that lists each document system change (record creation, deletion or alteration) that was made after the latest fsimage. When a document is put into HDFS, it is converted into blocks (of configurable size).

Q14. Would you be able to change the block size of HDFS files?

Hadoop Distributed File System (HDFS) stores documents as information blocks and circulates these blocks over the whole cluster. As HDFS was intended to be fault tolerant and to keep running on ware equipment, blocks are replicated in various circumstances to guarantee high data accessibility.

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

Q15. What do you mean by block scanner in HDFS?

Block Scanner is fundamentally used to recognize corrupt data-node Block. During a writing task, when a data node writes into the HDFS, it confirms a checksum for that information. This checksum helps in confirming the information corruptions during the information transmission.

At the point when similar information is perused from the HDFS, the customer confirms the checksum returned by the data-node against the checksum it figures against the information to check the information corruption that may have caused by the information node that may have happened during the shortage of information in the data node.

Q16. Can we change the document present in HDFS?

No, we can’t change the documents, which are present in HDFS, as HDFS takes after Write Once Read Many models.

Q17. What does the High Availability of a Name-Node means? How is it accomplished?

To make the HDFS high accessible means, it must be accessible constantly. So we can accomplish HDFS HA by making the name-node high accessible with the goal that it could serve HDFS related demands and queries whenever it is needed.

To settle this Single Point of Failure issue of Name-Node, HA highlight was introduced in Hadoop 2. X where we have two Name-Node in our HDFS cluster in a functioning/passive configuration. Consequently, if the active Name-Node fails down, the other inactive Name-Node can assume control over the obligation of the failed NameNode and keep the HDFS running.

Q18. Does HDFS enable a customer to peruse a record, which is already opened for writing?

Yes, one can read the document, which is as of already opened. However, the issue in perusing a document which is right now being composed lies in the consistency of the information, i.e. HDFS does not give the surety that the information which has been built into the document will be visible to another reader before the document has been closed down. For this, one can call the hflush activity explicitly which will drive all of the information in the cushion into the composed pipeline and afterward the hflush task will wait for the affirmations from the DataNodes.

Q19. Will various customers write into an HDFS record simultaneously?

No, multiple customers can't write into an HDFS document simultaneously. HDFS takes after single writer multiple readers model.

Q20. What do you think about the Speculative Execution?

In Hadoop, Speculative Execution is a procedure that happens during the slower execution of an errand at a node. In this procedure, the master node begins executing another occurrence of that same task on the other hub. Furthermore, the errand, which is done first is acknowledged, and killing that halts the execution of other.

Q21. How to use Combiner in Hadoop ?

A combiner is an Optional Component or Class and it can be Specified via **ob.**setcombinerclass**(Class name**), to perform the local aggregation of the intermediate outputs, which helps to cut down the amount of data transferred from the Mapper to the reducer.

Q22. Is it possible to create multiple table in hive for same data?

Initially, it looks not possible as I am a regular RDBMS user like other programmers, but then I tried to connect in Hive context. I found it is possible as Hive creates schema and append on top of an existing data file. One can have multiple schema for one data file, schema would be saved in hive's metastore and data will not be parsed read or serialized to disk in given schema. When s/he will try to retrieve data schema will be used. Lets say if my file have 5 column (Id, Name, Class, Section, Course) we can have multiple schema by choosing any number of column

Q23. What type of Data Hadoop Can Handle ?

Hadoop can able to handle all types of data like structured, Un-Structured, pictures, videos, telecom communications records, log files etc

Q24. What is Hadoop mapreduce?

Q25. What is oozie in Hadoop ?

Oozie is a workflow scheduler for Hadoop developed by Apache which runs the workflow of dependent jobs. In Oozie, users are permitted to create DAG's (Directed Acyclic Graphs) of workflows, which can be run in parallel and sequentially in Hadoop.

Oozie consists of two parts; workflow engine and coordinator engine. It is scalable and can handle the timely execution of thousands of workflows (each consists of dozens of jobs) in a Hadoop cluster. Moreover, Oozie is very much adjustable, as well. One can easily start, quit, suspend and rerun works. It makes it very easy to rerun broken workflows.