By OnlineInterviewQuestions.com

Hadoop Mapreduce Interview Questions

Hadoop MapReduce is one of the software structured for effectively writing an application for preparing a large amount of information in parallel or on a vast cluster of a commodity. As it deals with preparing data, it is probably going to be asked in **Hadoop Map Reduce Interview Questions** and Answers. There is an enormous demand for the Map-Reduce experts in the market.

It doesn't make a difference, if you are a beginner or experienced one or the one who has re-applied for another job position, experiencing the most prevalent Hadoop Map Reduce questions and answers can assist you to get prepared for the Map-Reduce interview. This blog contains usually asked Hadoop Map-reduce questions and answers, which will make you more confident while going through an interview. Hope these **Hadoop Map Reduce questions** will assist you to get selected in Hadoop interview.

Q1. What is Hadoop Map Reduce?

Map Reduce is the core of Hadoop. It is one of the programming paradigms that acknowledge into consideration enormous adaptability across a thousand's of the server in a Hadoop cluster. It is a processing layer of Hadoop. Map Reduce is a programming model intended for preparing large volumes of information in parallel by isolating the work into the arrangement of chunks. We need to compose the business logic; at that point rest work will be taken care of by the system.

Q2. What is the need of Map Reduce?

Map Reduce is information handling paradigms in itself. This was one of its kind information handlings and has been transformative. While utilizing Map Reduce, we are moving the calculation to information, which is less expensive when compared with information, is moved to the calculation.

Q3. Clarify what is shuffling in Map Reduce?

The procedure by which the framework lays out the sort and transfers the map outputs to the reducer as sources of information is known as the shuffling.

Q4. What is Mapper in Map Reduce?

Mapper is the client characterize program, which controls the info split in (key, value) combines according to the code design. Regularly Mapper is the base class, which needs to reach out by a software engineer to compose their logic according to the requirement. While broadening mapper, the programmer needs to specify information and output type under mapper class arguments.

Q5. Clarify what JobTracker is in Hadoop? What are the activities followed by Hadoop?

In Hadoop for submitting and following Map Reduce occupations, Job Tracker is utilized. Job Tracker is a basic service which cultivates out all MapReduce tasks to the different nodes in the group, preferably to those nodes which as of now contain the information, or at very least are situated in the same rack from nodes containing the information.

Job Tracker performs following activities in Hadoop:

- Client application presents jobs to the job tracker
- Job Tracker imparts to the Name mode to decide data area
- Near the data or with accessible openings Job Tracker finds Task Tracker nodes
- On choosing Task Tracker Nodes, it submits the work
- When a task fails, Job tracker notifies and chooses what to do then.
- Job Tracker observes the Task Tracker nodes

Q6. Clarify what combiners are and when you should utilize a combiner in a Map Reduce Job?

To enhance the effectiveness of Map Reduce Program, Combiners are utilized. The amount of information can be lessened with the assistance of combiner's that should be exchanged across to the reducers. If the task performed is commutative and affiliated you can utilize your reducer code as a combiner.

Q7. What is Speculative Execution?

In Hadoop, Map Reduce breaks jobs into various tasks, and these tasks run parallel rather than going for consecutive, in this manner decreases overall execution time. This model of execution is delicate to moderate tasks as they slow down the general execution of a job. There might be different explanations behind the slowdown of tasks, including hardware debasement or programming misconfiguration, yet it might be difficult to identify causes since the tasks still complete effectively, although additional time is taken than the normal time. Hadoop doesn't attempt to analyze and settle moderate running tasks; rather, it endeavors to recognize them and runs reinforcement tasks for them. This is called speculative execution in Hadoop.

Q8. What are the four essential parameters of a mapper?

The four essential parameters of a mapper are Long, Writable, text, text and Int-Writable. The initial two represents the input parameters and the second two speak about intermediate output parameters.

Q9. Clarify what WebDAV is in Hadoop?

Web Distributed Authoring and Versioning (WebDAV) is an expansion of the Hypertext Transfer Protocol (HTTP) that enables customers to perform remote Web content composing tasks. The WebDAV protocol gives a structure to clients to make, change and move reports on a server, normally a web server or web share. On most working framework WebDAV shares can be mounted as file systems, so it is conceivable to get to HDFS as a standard file system by uncovering HDFS over WebDAV.

Q10. Clarify what is sqoop in Hadoop?

Sqoop is a device intended to exchange information between Hadoop and social database servers. It is utilized to import information from social databases, for example, MySQL, Oracle to Hadoop HDFS, and export from the Hadoop file framework to social databases.

Q11. Clarify how Job Tracker schedules an assignment?

The task tracker conveys heart messages to Job tracker generally like clockwork to ensure that Job Tracker is active and working. The message also informs Job Tracker about the number of accessible slots, so the Job Tracker can stay updated with wherein the cluster work can be appointed.

Q12. Specify what the information segments utilized by Hadoop are?

Various data components, which are utilized by Hadoop are as follows:

- Spark
- Hive
- Pig
- Hbase
- Oozie
- Sqoop

Q13. Clarify how ordering in HDFS is finished?

Hadoop rose as an answer to the “Enormous Data” issues. It is an open source programming structure for distributed storage and circulated preparing of large data sets. Apache Hadoop has a unique method for Indexing. As, Hadoop structure store the information according to the Data Bock size, HDFS will continue storing the last piece of the information which will state where the following part of the information will be.

Q14. What does rack awareness mean?

In a large cluster of Hadoop, keeping in mind the end goal to enhance the network traffic while perusing/composing HDFS file, name-node picks the data node which is nearer to a similar rack or close-by rack to Read/Write ask. Name node accomplishes rack data by keeping up the rack id’s of each data node. This idea that picks nearer data nodes, which are based on rack data is called Rack Awareness in Hadoop. Rack awareness consists of the knowledge of Cluster topology or more specifically how the different information nodes are conveyed over the racks of a Hadoop cluster. Default Hadoop installation expects that all data nodes belong to with a similar rack.

Q15. What happens if the quantity of the reducer is 0 in MapReduce?

If we set the quantity of Reducer to 0 at that point, no reducer will execute, and no accumulation will occur. In such a case, we must go for “Map only job” in Hadoop.

Q16. Clarify what a Task Tracker is in Hadoop?

A Task Tracker in Hadoop is a slave node daemon in the cluster that acknowledges tasks from a Job Tracker. It also conveys the heartbeat messages to the Job Tracker, at regular intervals, to confirm that the Job Tracker is yet alive.

Q17. What is the most widely recognized info formats characterized in Hadoop?

In Hadoop, Input records store the information for a Map-Reduce work. Input files, which store information regularly reside in HDFS. Hence, in Map-Reduce, Input Format characterizes how these information files split and read. **Input Format** does **Input split**.
Most common Input Format is as follows:

- FileInputFormat
- TextInputFormat
- KeyValueTextInputFormat

Q18. Clarify what is Sequence file input format?

Hadoop Sequence documents are one of the Apache Hadoop specific file formats, which store information in the serialized key-value combine. Hadoop Sequence File is utilized as a part of Map Reduce as input/output formats. By default Mapper output is stored on local document framework, which is in Mapper node. Outputs of Maps are put away utilizing Sequence File. Inside Hadoop utilizes Sequence File organize for the Mapper which is stored in the local document system. In general, Apache Hadoop supports text records which are normally utilized for keeping and storing the information, other than the text documents it additionally supports binary documents and one of these binary formats are called Sequence records.

Q19. What are the fundamental configurations parameters specified in Map Reduce?

To work appropriately, Map Reduce needs some design parameters to be set accurately. Without them set accurately, the map and reduce jobs won't work appropriately. The configuration parameters that should be set effectively are as per the following:

- Job's input area in HDFS.
- Job's output area in HDFS.
- Input and Output format.
- Classes that contain the map and decrease capacities.
- Lastly jar file for reducer, mapper and driver classes.

Q20. What happens when a Data Node fails?

As large information processing is data and time delicate, there are backup processes if DataNode fails. Once a DataNode fails, another replication pipeline is made. The pipeline assumes control over the compose procedure and resumes from where it fizzled. Name Node, which continually watches if any of the blocks is under-repeated, administers the entire procedure or not.

Q21. What is MapReduce?

MapReduce is a programming model based on the Java programming language designed for distributed computing. Mapreduce framework. It is used to process parallel problems on a huge dataset across a number of distributed nodes. Using MapReduce, it is easy to scale data processing over multiple computing nodes. Many libraries have been written in different programming languages for the MapReduce programming model. But the popular open-source library is the Apache Hadoop

Q22. List the main components of MapReduce Job?

The two main components of the MapReduce Job are the JobTracker and TaskTracker.

JobTracker - It is the master that creates and runs the job in the MapReduce. It runs on the name node and allocates the job to TaskTrackers. The JobTracker first receives the request from the client and talks to the name node to determine the location of the data. Then, it finds the best TaskTracker node to execute the tasks based on the data locality. JobTracker also monitors the individual TaskTrackers and submits the overall status of the job to the client.

TaskTracker - It is the slave that runs on the data node. TaskTracker runs the job sent by the JobTracker and reports the status of the task back to it. The TaskTracker will be assigned with the Mapper and Reducer tasks to execute the job.

Q23. List operations of the MapReduce framework?

The two main operations executed by the MapReduce framework is the **map** and **reduce task**.

In the map operation, the data is split and mapped into different nodes for processing.

In the reduce task, the processed data shuffled and reduced.

Q24. Explain Shuffling and Sorting in MapReduce?

Shuffling is the process of transferring the data. It transfers the data from the mappers to the reducers. The output data from the map is sent as input to the reducer. This process is necessary for the reducers or they would not have any input.

Sorting operation sorts the keys generated by the mapper. It is done to easily distinguish when a new reduce task should start. When a new key in the sorted input data is different from the previous, then a new reduce task starts.

Q25. What is Mapper in MapReduce?

Mapper is one of the functions in the **MapReduce** that is used to process the input data. It processes the data by creating several small chunks of it. The mapper function takes the input in the form of a key/value pairs and processes the data into several chunks. The input data is specified to the Mapper function by the InooutFormat that defines the location of the input data. The RecordReader objects present in the InputFormat are used to extract the key/value from the input source. Then, the Mapper processes the input and creates an intermediate output that is sent o the Reducer.

Q26. What is Identity Mapper and Chain Mapper?

The Identity Mapper is one of the pre-defined mapper class that can be used with any key/value pairs of data. It is a generic class and also the default mapper class provided by Hadoop. When no mapper class is specified in the MR Driver class, the Identity Mapper class is invoked automatically when a MapReduce job is assigned.

The ChainMapper is also one of the pre-defined mapper classes that allows using multiple mapper classes within a single Map task. All the mappers are run in a chain fashion, that is the output of the first mapper becomes the input of the second mapper and so on. The output of the last mapper class is written to the intermediate files.

Q27. What is JobConf in MapReduce?

JonConf in the Hadoop MapReduce is the primary interface used by the user to describe a map-reduce job to the Hadoop framework for execution. The job parameters set in the JobConf cannot be altered if they are marked final by the administrators. Also, some parameters in the JobConf are easy to set while some parameters are relatively more complex for the user to control finely. You can also define other advanced facets of the job such as comparators to be used, files to be put in the DistributedCache, etc in the JobConf.

Q28. What is Combiner in MapReduce?

A Combiner or a semi-reducer is just an optional class in the **MapReduce**. It mainly summarizes the map output records with the same key. The combiner accepts the input from the map class. The output of the combiner will be sent over the network to the reducer task as the input. So the combiner works between the Map and Reduce class to reduce the amount of data transfer between them. For the combiner to work properly, the key-value type should be the same between the mapper and the reducer class. Also, the combiner doesn’t have a predefined interface as it implements the reduce() method from the interface of the reducer.

Q29. What is Reducer in MapReduce?

As the name suggests, the **reducer reduces** the set of **intermediate values**. The intermediate values are reduced to the smaller set of values that share a key.

The Reducer has three phases. They are the **shuffle**, **sort**, and the **reduce phase**.

The shuffle and sort phase occur concurrently to get the input from the mapper phase and organize it. The reduce phase occurs after the shuffle. and sort to aggregate the key-value pairs. After it reduces the data, the output is written to the FileSystem.

Q30. What is LazyOutputFormat in MapReduce?

The LazyOutputFormat is just a wrapper output format. It ensures that the output file is created only when the given partition of the data exists. It won't create any empty files in the directories in the MapReduce job if there are no records for the specific partition.

Q31. What is Counter in MapReduce?

The Counters in the Hadoop MapReduce provides a way to measure the number of operations that occur in the **MapReduce job**. It is really useful in diagnosing the problem and gathering statistics about the MapReduce job. Counters are defined either by the MapReduce framework or the applications. Two types of Counters are the Built-In counters and User-Defined counters. Each counter in the Hadoop has along for the value and is named by an Enum.

Q32. What is distributed Cache in MapReduce?

The Distributed cache in the Hadoop MapReduce framework is used to cache the files that are needed by the applications.

It is used to cache read-only text files, archives, jar files, etc. The cached file in the MapReduce distributed cache is available on each data node where the map/reduce tasks are running. Distributed cache provides many benefits such as storing complex data, eliminating a single point of failure, and ensuring data consistency.

Q33. What is the default input type in MapReduce?

There are many **input formats** available in the MapReduce such as the **TextInputFormat**, **KeyValueTextInputFormat**, **FileInputFormat**, etc. But the default InputFormat in the MapReduce is the TextInput Format. It treats each line of the input file as a separate record. This format is useful for unformatted data or line-based records like the log files.

Q34. What is reducer mapper?

Mapper is the first phase in the **MapReduce model** to solve the problem. In this phase, the data is distributed and computed on each node. Here, the concept of parallel processing is executed for fast processing.

The next and the final phase is the Reducer in the MapReduce model. It reduces the data from the mapper and stores the data into the file. It also performs tasks such as shuffle and sort.

Q35. What is input split in MapReduce?

InputSplit in the MapReduce is used to represent the data logically that is used by the mapper process. So the number of InputSplits are equal to the number of map tasks. Every InputSplit has a storage location and the length of the InputSplits is measured in bytes. The important thing to note is that the InputSplit just references the data and it doesn't actually contain the data. The Input format in the Hadoop is responsible for creating the InputSplits. The split size based on the size of the data in the MapReduce program can be user-defined.