

Looking out for Hadoop MapReduce Interview Questions that are frequently asked by employers?

I hope you have not missed the previous blog in this interview questions blog series that contains the most frequently asked **Top 50 Hadoop Interview Questions** (<https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/>) by the employers. Now, before moving ahead in this Hadoop MapReduce Interview Questions blog, let us have a brief understanding of MapReduce framework and its working:

MapReduce is a programming framework that allows us to perform distributed and parallel processing on large data sets in a distributed environment.

- MapReduce framework also follows Master/Slave Topology where the master node (Resource Manager) manages and tracks various MapReduce jobs being executed on the slave nodes (Node Managers).
- Resource Manager consists of two main components:
 - **Application Master:** It accepts job-submissions, negotiates the container for ApplicationMaster and handles failures while executing MapReduce jobs.
 - **Scheduler:** Scheduler allocates resources that is required by various MapReduce application running on the Hadoop cluster.

- As the name MapReduce suggests, reducer phase takes place after the mapper phase has been completed.
- So, the first is the map job, where a block of data is read and processed to produce key-value pairs as intermediate outputs.
- The reducer receives the key-value pair from multiple map jobs.
- Then, the reducer aggregates those intermediate data tuples (intermediate key-value pair) into a smaller set of tuples or key-value pairs which is the final output.

"A mind troubled by doubt cannot focus on the course to victory."

– Arthur Golden

The above quote reflects the importance of having your fundamentals clear before appearing for an interview as well as while going through this Hadoop MapReduce Interview Question blog. Therefore, I would suggest you to go through **MapReduce Tutorial** (<https://www.edureka.co/blog/mapreduce-tutorial/>) blog to brush up your basics.

Learn MapReduce from Industry Experts

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

Here, is the list of *Hadoop MapReduce Interview Questions* that will help you to stand up to the expectation of the employers.

Hadoop Interview Questions and Answers | Edureka

1. What are the advantages of using MapReduce with Hadoop?

Advantages of MapReduce

Advantage	Description
Flexible	Hadoop MapReduce programming can access and operate on different types of structured and unstructured
Parallel Processing	MapReduce programming divides tasks for execution in parallel
Resilient	Is fault tolerant that quickly recognizes the faults & then apply a quick recovery solution implicitly
Scalable	Hadoop is a highly scalable platform that can store as well as distribute large data sets across plenty of servers
Cost-effective	High scalability of Hadoop also makes it a cost-effective solution for ever-growing data storage needs
Simple	It is based on a simple programming model

Simple	It is based on a simple programming model.
Secure	Hadoop MapReduce aligns with HDFS and HBase security for security measures.
Speed	It uses the distributed file system for storage that processes even the large sets of unstructured data in minutes.

2. What do you mean by data locality?

- **Data locality** (https://www.edureka.co/blog/mapreduce-tutorial/#data_locality) talks about moving computation unit to data rather than data to the computation unit.
- MapReduce framework achieves data locality by processing data locally
- Which means processing of the data happens in the very node by Node Manager where data blocks are present.

3. Is it mandatory to set input and output type/format in MapReduce?

No, it is not mandatory to set the input and output type/format in MapReduce. By default, the cluster takes the input and the output type as 'text'.

4. Can we rename the output file?

Yes, we can rename the output file by implementing *multiple format output class*.

5. What do you mean by shuffling and sorting in MapReduce?

Shuffling and sorting takes place after the completion of map task where the input to the every reducer is sorted according to the keys. Basically, the process by which the system sorts the key-value output of the map tasks and transfer it to the reducer is called shuffle.

6. Explain the process of spilling in MapReduce?

The output of a map task is written into a circular memory buffer (RAM). The default size of buffer is set to 100 MB which can be tuned by using `mapreduce.task.io.sort.mb` property. Now, spilling is a process of copying the data from memory buffer to disc when the content of the buffer reaches a certain threshold size. By default, a background thread starts spilling the contents from memory to disc after 80% of the buffer size is filled. Therefore, for a 100 MB size buffer the spilling will start after the content of the buffer reach a size of 80 MB.

Note: One can change this spilling threshold using `mapreduce.map.sort.spill.percent` which is set to 0.8 or 80% by default.

7. What is a distributed cache in 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. Therefore, one can access the cache file as a local file in your Mapper or Reducer job.

8. What is a combiner and where you should use it?

Combiner is like a mini reducer function that allow us to perform a local aggregation of map output before it is transferred to reducer phase. Basically, it is used to optimize the network bandwidth usage during a MapReduce task by cutting down the amount of data that is transferred from a mapper to the reducer.

9. Why the output of map tasks are stored (spilled) into local disc and not in HDFS?

The outputs of map task are the intermediate key-value pairs which is then processed by reducer to produce the final aggregated result. Once a MapReduce job is completed, there is no need of the intermediate output produced by map tasks. Therefore, storing these intermediate output into HDFS and replicate it will create unnecessary overhead.

10. What happens when the node running the map task fails before the map output has been sent to the reducer?

In this case, map task will be assigned a new node and whole task will be run again to re-create the map output.

11. What is the role of a MapReduce Partitioner?

A partitioner divides the intermediate key-value pairs produced by map tasks into partition. The total number of partition is equal to the number of reducers where each partition is processed by the corresponding reducer. The partitioning is done using the hash function based on a single key or group of keys. The default partitioner available in Hadoop is *HashPartitioner*.

12. How can we assure that the values regarding a particular key goes to the same reducer?

By using a partitioner we can control that a particular key – value goes to the same reducer for processing.

13. What is the difference between Input Split and HDFS block?

HDFS block defines how the data is physically divided in HDFS whereas input split defines the logical boundary of the records required for processing it.

14. What do you mean by InputFormat?

InputFormat describes the input-specification for a MapReduce job. The MapReduce framework relies on the InputFormat of the job to:

- Validate the input-specification of the job.
- Split-up the input file(s) into logical InputSplit instances, each of which is then assigned to an individual Mapper.
- Provide the RecordReader implementation used to read records from the logical InputSplit for processing by the Mapper.

15. What is the purpose of TextInputFormat?

TextInputFormat is the default input format present in the MapReduce framework. In TextInputFormat, an input file is produced as keys of type LongWritable (byte offset of the beginning of the line in the file) and values of type Text (content of the line).

of the beginning of the line in the file) and values of type TEXT (content of the line).

16. What is the role of RecordReader in Hadoop MapReduce?

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".

17. What are the various configuration parameters required to run a MapReduce job?

The main **configuration parameters** (https://www.edureka.co/blog/mapreduce-tutorial/#explanation_of_mapreduce_program) 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

18. When should you use 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.

19. What is an identity Mapper and Identity Reducer?

Identity mapper is the default mapper provided by the Hadoop framework. It runs when no mapper class has been defined in the MapReduce program where it simply passes the input key – value pair for the reducer phase.

Like Identity Mapper, Identity Reducer is also the default reducer class provided by the Hadoop, which is automatically executed if no reducer class has been defined. It also performs no computation or process, rather it just simply write the input key – value pair into the specified output directory.

20. What is a map side join?

Map side join is a process where two data sets are joined by the mapper.

21. What are the advantages of using map side join in MapReduce?

The advantages of using map side join in MapReduce are as follows:

- Map-side join helps in minimizing the cost that is incurred for sorting and merging in the shuffle and reduce stages.
- Map-side join also helps in improving the performance of the task by decreasing the time to finish the task.

22. What is reduce side join in MapReduce?

As the name suggests, in the reduce side join, the reducer is responsible for performing the join operation. It is comparatively simple and easier to implement than the map side join as the sorting and shuffling phase sends the values having identical keys to the same reducer and therefore, by default, the data is organized for us.

♣ **Tip:** I would suggest you to go through a dedicated blog on **reduce side join** (<https://www.edureka.co/blog/mapreduce-example-reduce-side-join/>) in MapReduce where the whole process of reduce side join is explained in detail with an example.

23. What do you know about NLineInputFormat?

NLineInputFormat splits 'n' lines of input as one split.

24. Is it legal to set the number of reducer task to zero? Where the output will be stored in this case?

Yes, It is legal to set the number of reduce-tasks to *zero* if there is no need for a reducer. In this case the outputs of the map task is directly stored into the HDFS which is specified in the setOutputPath(Path).

25. Is it necessary to write a MapReduce job in Java?

No, MapReduce framework supports multiple languages like Python, Ruby etc.

26. How do you stop a running job gracefully?

One can gracefully stop a MapReduce job by using the command: *hadoop job -kill JOBID*

27. How will you submit extra files or data (like jars, static files, etc.) for a MapReduce job during runtime?

The distributed cache is used to distribute large read-only files that are needed by map/reduce jobs to the cluster. The framework will copy the necessary files from a URL on to the slave node before any tasks for the job are executed on that node. The files are only copied once per job and so should not be modified by the application.

28. How does inputsplit in MapReduce determines the record boundaries correctly?

RecordReader is responsible for providing the information regarding record boundaries in an input split.

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. Define Speculative Execution

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

Check out our Hadoop Course

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

I hope you find this blog on Hadoop MapReduce Interview Questions to be informative and helpful. You are welcome to mention your doubts and feedback in the comment section given below. In this blog, I have covered the interview questions for MapReduce only. To save your time in visiting several sites for interview questions related to each Hadoop component, we have prepared a series of interview question blogs that covers all the components present in Hadoop framework. Kindly, refer to the links given below to explore all the Hadoop related interview question and strengthen your fundamentals:

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



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

f t in G+

(<https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hadoop-cluster/>)
 u=<https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hdfs-2/>
 question=<https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-pig/>
 interview=<https://www.edureka.co/blog/interview-questions/hive-interview-questions/>
 question=<https://www.edureka.co/blog/interview-questions/hbase-interview-questions/>
 mapreduce/)

Share on

◀ PREVIOUS

NEXT ▶

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

Enter your Email Address

SUBSCRIBE

Related Posts



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/>)



MapReduce Tutorial – Fundamentals of MapReduce with MapReduce Example
👁 74.4K

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

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

14 Comments

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

Rajiv Chaudhuri ▾

 Recommend
  Tweet
  Share

Sort by Best ▾



Join the discussion...

bharadwaj • 2 years ago

can you explain in detail about custom input format..?...



 Reply
  Share

EdurekaSupport Mod → bharadwaj • 2 years ago

Hey Bharadwaj, thanks for checking out the blog. With regard to your query, custom input format can be implemented as per specific requirement. Please have a look into some below input formats available in MapReduce.

The default InputFormat is the TextInputFormat. This treats each line of each input file as a separate record, and performs no parsing. This is useful for unformatted data or line-based records like log files.

A more interesting input format is the KeyValueInputFormat. This format also treats each line of input as a separate record. While the TextInputFormat treats the entire line as the value, the KeyValueInputFormat breaks the line itself into the key and value by searching for a tab character. This is particularly useful for reading the output of one MapReduce job as the input to another.

Finally, the SequenceFileInputFormat reads special binary files that are specific to Hadoop. These files include many features designed to allow data to be rapidly read into Hadoop mappers. Sequence files are block-compressed and provide direct serialization and deserialization of several arbitrary data types (not just text). 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.

Hope this helps. Please get in touch if you have any other queries.



 Reply
  Share


bala • 3 years ago

what generic InputSplit class?



 Reply
  Share


Sande • 3 years ago

what data structure used in Hadoop?



 Reply
  Share

EdurekaSupport Mod → Sande • 3 years ago

Hi Sande, HDFS is the default underlying storage platform of Hadoop. Its like any other file system in the sense that it does not care what structure the files have. It only ensures that files will be saved in a redundant fashion and available for retrieval quickly.

So it is totally up to you the user, to store files with whatever structure you like inside them.

A MapReduce program simply gets the file data fed to it as an input. Not necessarily the entire file, but parts of it depending on InputFormats etc. The Map program then can make use of the data in whatever way it wants to.



 Reply
  Share


Awanish • 5 years ago

very nice post, thanks a lot!!
very helpful.


 Reply
  Share

Karthik • 2 years ago

What is custom key? and How can i implement custom key?



 Reply
  Share

EdurekaSupport Mod → Karthik • 2 years ago

Hey Karthik, thanks for checking out the blog. Here's a brief explanation about custom key and its implementation.

– In Hadoop, data types to be used as key must implement WritableComparable interface and data types to be used as value must implement Writable interface.

– if your custom key / value are of the same type then you can write one custom datatype for both the key / value which implements WritableComparable, otherwise you need to implement two different data types. One for key which implements WritableComparable and second for value which implements Writable

```
interface.  
//Custom Data-Type  
public class MyCustomKey implements WritableComparable  
{  
    //Create Mapper with Custom Key  
    public class MyMapper extends Mapper  
    {  
    }  
}
```

^ | v • Reply • Share ›

Karthik → EdurekaSupport • 2 years ago

Thank you..

1 ^ | v • Reply • Share ›

AMIT RAJPUT • 3 years ago

In hadoop framewrok, who decide input split?

^ | v • Reply • Share ›

sulthan syedibrahim → AMIT RAJPUT • 3 years ago

The input split can be set by three property settings

i. split.minsize

ii.split.maximumsize and

iii. by default as block size

usually developers define the split size as block size. if you have data and the data should be processed within single mapper at the time you can mention the size of the split much higher than the file size.

^ | v • Reply • Share ›

ALSO ON [HTTPS://WWW.EDUREKA.CO/BLOG/](https://www.edureka.co/blog/)

AI and IoT in FIFA: Smart Sports

3 comments • 3 months ago

Apoorva Verma — The digital transformation of football which began with Goal line technology (GLT), is going beyond GLT towards Internet of ...

Top 55 Blockchain Interview Questions You Must Prepare In 2018

4 comments • 7 months ago

Maximilian Fischer — A blockchain (database) only contains a record of all changes that were ever made to it. Value is ascribed to the fact of actually being ...

Top Technical Skills to Secure Jobs of the Future

4 comments • 3 months ago


Gvsm Chaithanya — gvsmc1996@gmail.com

What Is Microservices – Introduction To Microservice Architecture

1 comment • 8 months ago

Nadeem Inamdar — Good blog! Hate to be a grammar nazi .. but What ARE Microservices!

 [Subscribe](#)

 [Add Disqus to your site](#)[Add Disqus](#)[Add](#)


 [Disqus' Privacy Policy](#)[Privacy](#)[Policy](#)[Privacy](#)


Subscribe
to our newsletter


Enter your Email Address


SUBSCRIBE

Related Blogs

- 


(<https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/>)
- 


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


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


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

Big Data Analytics Courses

- 

(</big-data-and-hadoop>)
- 

(</pyspark-certification-training>)
- 

(</hadoop-admin>)
- 

(</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/>)

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

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




(<https://www.facebook.com/edureka1N>)
(<https://twitter.com/edureka1N>)
(<https://www.linkedin.com/company/edureka1N>)
(<https://www.youtube.com/channel/UC8q331454157mt=8>)

Learn on the GO!

(<https://itunes.apple.com/in/app/edureka/id10331454157mt=8>)



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



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

"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

https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-mapreduce/?utm_source=hd&utm_campaign=lms_hd_120616&utm_me...

7/8

