This list of Hadoop interview questions is prepared with extensive inputs from industry experts to give you an advantage in your job interviews. You will understand what Hadoop applications are, how Hadoop is different from other parallel processing engines, Hadoop running modes, NameNode, DataNode, JobTracker, TaskTracker, and more.

Basic Big Data Hadoop Interview Questions

1. What do you mean by the term or concept of Big Data?

Big Data means a set or collection of large datasets that keeps on growing exponentially. It is difficult to manage Big Data with traditional data management tools. Examples of Big Data include the amount of data generated by Facebook or Stock Exchange Board of India on a daily basis. There are three types of Big Data:

- Structured Big Data
- Unstructured Big Data
- Semi-structured Big Data

2. What are the characteristics of Big Data?

The characteristics of Big Data are as follows:

- Volume
- Variety
- Velocity
- Variability

Where,

Volume means the size of the data, as this feature is of utmost importance while handling Big Data solutions. The volume of Big Data is usually high and complex.

Variety refers to the various sources from which data is collected. Basically, it refers to the types, structured, unstructured, and semi-structured, and heterogeneity of Big Data.

Velocity means how fast or slow the data is getting generated. Basically, Big Data velocity deals with the speed at which the data is generated from business processes, operations, application logs, etc.

Variability, as the name suggests, means how differently the data behaves in different situations or scenarios in a given period of time.

3. What are the various steps involved in deploying a Big Data solution?

Deploying a Big Data solution includes the following steps:

- **Data Ingestion:** As a first step, the data is drawn out or extracted from various sources so as to feed it to the system.
- **Data Storage:** Once data ingestion is completed, the data is stored in either HDFS or NoSQL database.

- **Data Processing:** In the final step, the data is processed through frameworks and tools such as Spark, MapReduce, Pig, etc.

4. What is the reason behind using Hadoop in Big Data analytics?

Businesses generate a lot of data in a single day and the data generated is unstructured in nature. Data analysis with unstructured data is difficult as it renders traditional big data solutions ineffective. Hadoop comes into the picture when the data is complex, large and especially unstructured. Hadoop is important in Big Data analytics because of its characteristics:

- **Data storage**
- **Data processing**
- **Collection plus extraction of data**

5. What do you understand by fsck in Hadoop?

fsck stands for file system check in Hadoop, and is a command that is used in HDFS. fsck checks any and all data inconsistencies. If the command detects any inconsistency, HDFS is notified regarding the same.

6. Can you explain some of the important features of Hadoop?

Some of the important features of Hadoop are:

Fault Tolerance: Hadoop has a high-level of fault tolerance. To tackle faults, Hadoop, by default, creates three replicas for each block at different nodes. This number can be modified as per the requirements. This helps to recover the data from another node if one node has failed. Hadoop also facilitates automatic recovery of data and node detection.

Open Source: One of the best features of Hadoop is that it is an open-source framework and is available free of cost. Hadoop also allows its users to change the source code as per their requirements.

Distributed Processing: Hadoop stores the data in a distributed manner in HDFS. Distributed processing implies fast data processing. Hadoop also uses MapReduce for the parallel processing of the data.

Reliability: One of the benefits of Hadoop is that the data stored in Hadoop is not affected by any kind of machine failure, which makes Hadoop a reliable tool.

Scalability: Scalability is another important feature of Hadoop. Hadoop's compatibility with other hardware makes it a preferred tool. You can also easily add new hardware to the nodes in Hadoop.

High Availability: Easy access to the data stored in Hadoop makes it a highly preferred Big Data management solution. Not only this, the data stored in Hadoop can be accessed even if there is a hardware failure as it can be accessed from a different path.

7. What is Hadoop and what are its components?

Apache Hadoop is the solution for dealing with Big Data. Hadoop is an open-source framework that offers several tools and services to store, manage, process, and analyze Big Data. This allows organizations to make significant business decisions in an effective and efficient manner, which was not possible with traditional methods and systems.

There are 3 main components of Hadoop. They are :

- HDFS
- YARN
- MapReduce

HDFS

It is a system that allows you to distribute the storage of big data across a cluster of computers. Italso maintains the redundant copies of data.So, if one of your computers happens to randomly burst into flames or if some technical issues occur, HDFS can actually recover from that by creating a backup from a copy of the data that it had saved automatically, and you won't even know if anything happened.

YARN

Next in the Hadoop ecosystem is YARN (Yet Another Resource Negotiator). It is the place where the data processing of Hadoop comes into play. YARN is a system that manages the resources on your computing cluster. It is the one that decides who gets to run the tasks, when and what nodes are available for extra work, and which nodes are not available to do so.

MapReduce

MapReduce, the next component of the Hadoop ecosystem, is just a programming model that allows you to process your data across an entire cluster. It basically consists of Mappers and Reducers that are different scripts, which you might write, or different functions you might use when writing a MapReduce program.

8. Explain Hadoop Architecture.

The Hadoop Architecture comprises of the following :

- Hadoop Common
- HDFS
- MapReduce
- YARN

Hadoop Common

Hadoop Common is a set of utilities that offers support to the other three components of Hadoop. It is a set of Java libraries and scripts that are required by MapReduce, YARN, and HDFS to run the Hadoop cluster.

HDFS

HDFS stands for Hadoop Distributed File System. It stores data in the form of small memory blocks and distributes them across the cluster. Each data is replicated multiple times to ensure data availability. It has two daemons. One for master node — NameNode and their for slave nodes —DataNode.

NameNode and DataNode : The NameNode runs on the master server. It manages the Namespace and regulates file access by the client. The DataNode runs on slave nodes. It stores the business data.

MapReduce

It executes tasks in a parallel fashion by distributing the data as small blocks. The two most important tasks that the Hadoop MapReduce carries out are Mapping the tasks and Reducing the tasks.

YARN

It allocates resources which in turn allow different users to execute various applications without worrying about the increased workloads.

9. In what all modes can Hadoop be run?

Hadoop can be run in three modes:

- **Standalone Mode:** The default mode of Hadoop, standalone mode uses a local file system for input and output operations. This mode is mainly used for debugging purposes, and it does not support the use of HDFS. Further, in this mode, there is no custom configuration required for mapred-site.xml, core-site.xml, and hdfs-site.xml files. This mode works much faster when compared to other modes.
- **Pseudo-distributed Mode (Single-node Cluster):** In the case of pseudo-distributed mode, you need the configuration for all the three files mentioned above. All daemons are running on one node; thus, both master and slave nodes are the same.
- **Fully distributed mode (Multi-node Cluster):** This is the production phase of Hadoop, what it is known for, where data is used and distributed across several nodes on a Hadoop cluster. Separate nodes are allotted as master and slave nodes.

10. Name some of the major organizations globally that use Hadoop?

Some of the major organizations globally that are using Hadoop as a Big Data tool are as follows:

- Netflix
- Uber
- The National Security Agency (NSA) of the United States
- The Bank of Scotland
- Twitter

11. What are the real-time industry applications of Hadoop?

Hadoop, well known as Apache Hadoop, is an open-source software platform for scalable and distributed computing of large volumes of data. It provides rapid, high-performance, and cost-effective analysis of structured and unstructured data generated on digital platforms and within the organizations. It is used across all departments and sectors today.

Here are some of the instances where Hadoop is used:

- Managing traffic on streets
- Streaming processing
- Content management and archiving emails
- Processing rat brain neuronal signals using a Hadoop computing cluster
- Fraud detection and prevention
- Advertisements targeting platforms are using Hadoop to capture and analyze clickstream, transaction, video, and social media data
- Managing content, posts, images, and videos on social media platforms

- Analyzing customer data in real-time for improving business performance
- Public sector fields such as intelligence, defense, cyber security, and scientific research
- Getting access to unstructured data such as output from medical devices, doctor’s notes, lab results, imaging reports, medical correspondence, clinical data, and financial data

12. What is HBase?

Apache HBase is a distributed, open-source, scalable, and multidimensional database of NoSQL. HBase is based on Java; it runs on HDFS and offers Google-Bigtable-like abilities and functionalities to Hadoop. Moreover, HBase’s fault-tolerant nature helps in storing large volumes of sparse datasets. HBase gets low latency and high throughput by offering faster access to large datasets for read or write functions.

13. What is a Combiner?

A combiner is a mini version of a reducer that is used to perform local reduction processes. The mapper sends the input to a specific node of the combiner, which later sends the respective output to the reducer. It also reduces the quantum of the data that needs to be sent to the reducers for improving the efficiency of MapReduce.

14. Is it okay to optimize algorithms or codes to make them run faster? If yes, why?

Yes, it is always suggested and recommended to optimize algorithms or codes to make them run faster. The reason for this is that optimized algorithms are pretrained and have an idea about the business problem. The higher the optimization, the higher the speed.

15. What is the difference between RDBMS and Hadoop?

Following are some of the differences between RDBMS (Relational Database Management) and Hadoop based on various factors:

	RDBMS	Hadoop
Data Types	It relies on structured data and the data schema is always known.	Hadoop can store structured, unstructured, and semi-structured data.
Cost	Since it is licensed, it is paid software.	It is a free open-source framework.
Processing	It offers little to no capabilities for processing.	It supports data processing for data distributed in a parallel manner across the cluster.
Read vs Write Schema	It follows ‘schema on write’, allowing the validation of schema to be done before data loading.	It supports the policy of schema on read.
Read/Write Speed	Reads are faster since the data schema is known.	Writes are faster since schema validation does not take place during HDFS write.
Best Use Case	It is used for Online Transactional Processing (OLTP) systems.	It is used for data analytics, data discovery, and OLAP systems.

16. What is Apache Spark?

Apache Spark is an open-source framework engine known for its speed and ease of use in Big Data processing and analysis. It also provides built-in modules for graph processing, machine learning, streaming, SQL, etc. The execution engine of Apache Spark supports in-memory computation and cyclic data flow. It can also access diverse data sources such as HBase, HDFS, Cassandra, etc.

17. Can you list the components of Apache Spark?

The components of the Apache Spark framework are as follows:

- Spark Core Engine
- **Spark Streaming**
- Mllib
- GraphX
- Spark SQL
- Spark R

One thing that needs to be noted here is that it is not necessary to use all Spark components together. But yes, the Spark Core Engine can be used with any of the other components listed above.

18. What are the differences between Hadoop and Spark?

Criteria	Hadoop	Spark
Dedicated storage	HDFS	None
Speed of processing	Average	Excellent
Libraries	Separate tools available	Spark Core, SQL, Streaming, MLlib, and GraphX

19. What is Apache Hive?

Apache Hive is an open-source tool or system in Hadoop; it is used for processing structured data stored in Hadoop. Apache Hive is the system responsible for facilitating analysis and queries in Hadoop. One of the benefits of using Apache Hive is that it helps SQL developers to write Hive queries almost similar to the SQL statements that are given for analysis and querying data.

20. Does Hive support multiline comments?

No. Hive does not support multiline comments. It only supports single-line comments as of now.

21. Explain the major difference between HDFS block and InputSplit

In simple terms, HDFS block is the physical representation of data, while InputSplit is the logical representation of the data present in the block. InputSplit acts as an intermediary between the block and the mapper.

Suppose there are two blocks:

Block 1: ii nntteell

Block 2: li ppaatt

Now considering the map, it will read Block 1 from ii to ll but does not know how to process Block 2 at the same time. InputSplit comes into play here, which will form a logical group of Block 1 and Block 2 as a single block.

It then forms a key-value pair using InputFormat and records the reader and sends the map for further processing with InputSplit. If you have limited resources, then you can increase the split size to limit the number of maps. For instance, if there are 10 blocks of 640 MB, 64 MB each, and limited resources, then you can assign the split size as 128 MB. This will form a logical group of 128 MB, with only five maps executing at a time.

However, if the split size property is set to false, then the whole file will form one InputSplit and will be processed by a single map, consuming more time when the file is bigger.

Intermediate Big Data Hadoop Interview Questions

22. What is the Hadoop Ecosystem?

Hadoop Ecosystem is a bundle or a suite of all the services that are related to the solution of Big Data problems. It is precisely speaking, a platform consisting of various components and tools that function jointly to execute Big Data projects and solve the issues therein. It consists of Apache projects and various other components that together constitute the Hadoop Ecosystem.

23. What is Hadoop Streaming?

Hadoop Streaming is one of the ways that are offered by Hadoop for non-Java development. Hadoop Streaming helps you to write MapReduce program in any language which can write to standard output and read standard input.The primary mechanisms are Hadoop Pipes which gives a native C++ interface to Hadoop and Hadoop Streaming which permits any program that uses standard input and output to be used for map tasks and reduce tasks. With the help of Hadoop Streaming, one can create and run MapReduce jobs with any executable or script as the mapper and/or the reducer.

24. How is Hadoop different from other parallel computing systems?

Hadoop is a distributed file system that lets you store and handle large amounts of data on a cloud of machines, handling data redundancy.

The primary benefit of this is that since the data is stored in several nodes, it is better to process it in a distributed manner. Each node can process the data stored on it, instead of spending time moving the data over the network.

On the contrary, in the relational database computing system, you can query the data in real-time, but it is not efficient to store the data in tables, records, and columns, when the data is large.

Hadoop also provides a scheme to build a column database with Hadoop HBase for runtime queries on rows.

Listed below are the main components of Hadoop:

- **HDFS:** HDFS is Hadoop’s storage unit.
- **MapReduce:** MapReduce the Hadoop’s processing unit.
- **YARN:** YARN is the resource management unit of Apache Hadoop.

25. Can you list the limitations of Hadoop?

Hadoop is considered a very important Big Data management tool. However, like other tools, it also has some limitations of its own. They are as below:

- In Hadoop, you can configure only one NameCode.
- Hadoop is suitable only for the batch processing of a large amount of data.
- Only map or reduce jobs can be run by Hadoop.
- Hadoop supports only one Name No and One Namespace for each cluster.
- Hadoop does not facilitate horizontal scalability of NameNode.
- Hourly backup of MetaData from NameNode needs to be given to the Secondary NameNode.
- Hadoop can support only up to 4000 nodes per cluster.
- In Hadoop, the JobTracker, one and only single component, performs a majority of the activities such as managing Hadoop resources, job schedules, job monitoring, rescheduling jobs, etc.
- Real-time data processing is not possible with Hadoop.
- Due to the preceding reason, JobTracker is the only possible single point of failure in Hadoop.

26. What is distributed cache? What are its benefits?

Distributed cache in Hadoop is a service by MapReduce framework to cache files when needed.

Once a file is cached for a specific job, Hadoop will make it available on each DataNode both in the system and in the memory, where map and reduce tasks are executed. Later, you can easily access and read the cache files and populate any collection, such as an array or hashmap, in your code.

The benefits of using distributed cache are as follows:

- It distributes simple, read-only text/data files and/or complex files such as jars, archives, and others. These archives are then un-archived at the slave node.
- Distributed cache tracks the modification timestamps of cache files, which notify that the files should not be modified until a job is executed.

27. Name the different configuration files in Hadoop

Below given are the names of the different configuration files in Hadoop:

- mapred-site.xml
- core-site.xml
- hdfs-site.xml
- yarn-site.xml

28. Can you skip the bad records in Hadoop? How?

In Hadoop, there is an option where sets of input records can be skipped while processing map inputs. This feature is managed by the applications through the SkipBadRecords class.

The SkipBadRecords class is commonly used when map tasks fail on input records. Please note that the failure can occur due to faults in the map function. Hence, the bad records can be skipped in Hadoop by using this class.

29. What are the various components of Apache HBase?

There are three main components of Apache HBase that are mentioned below:

- **HMaster:** It manages and coordinates the region server just like NameNode manages DataNodes in HDFS.
- **Region Server:** It is possible to divide a table into multiple regions and the region server makes it possible to serve a group of regions to the clients.
- **ZooKeeper:** ZooKeeper is a coordinator in the distributed environment of HBase. ZooKeeper communicates through the sessions to maintain the state of the server in the cluster.

30. What is the syntax to run a MapReduce program?

The syntax used to run a MapReduce program is hadoop_jar_file.jar /input_path /output_path.

31. Which command will you give to copy data from the local system onto HDFS?

Untitled

32. What are the components of Apache HBase's Region Server?

The following are the components of HBase's region server:

- **BlockCache:** It resides on the region server and stores data in the memory, which is read frequently.
- **WAL:** Write ahead log or WAL is a file that is attached to each region server located in the distributed environment.
- **MemStore:** MemStore is the write cache that stores the input data before it is stored in the disk or permanent memory.
- **HFile:** HDFS stores the HFile that stores the cells on the disk.

33. What are the various schedulers in YARN?

Mentioned below are the numerous schedulers that are available in YARN:

- **FIFO Scheduler:** The first-in-first-out (FIFO) scheduler places all the applications in a single queue and executes them in the same order as their submission. As the FIFO scheduler can block short applications due to long-running applications, it is less efficient and desirable for professionals.
- **Capacity Scheduler:** A different queue makes it possible to start executing short-term jobs as soon as they are submitted. Unlike in the FIFO scheduler, the long-term tasks are completed later in the capacity scheduler.
- **Fair Scheduler:** The fair scheduler, as the name suggests, works fairly. It balances the resources dynamically between all the running jobs and is not required to reserve a specific capacity for them.

34. What are the main components of YARN? Can you explain them?

The main components of YARN are explained below:

- **Resource Manager:** It runs on a master daemon and is responsible for controlling the **resource allocation** in the concerned cluster.
- **Node Manager:** It is responsible for executing a task on every single data node. Node manager also runs on the slave daemons in Hadoop.
- **Application Master:** It is an important component of YARN as it controls the user job life cycle and the resource demands of single applications. The application master works with the node manager to monitor the task execution.
- **Container:** It is like a combination of the Hadoop resources, which may include RAM, network, CPU, HDD, etc., on one single node.

35. Explain the difference among NameNode, Checkpoint NameNode, and Backup Node

- NameNode is the core of HDFS. NameNode manages the metadata. In simple terms, NameNode is the data about the data being stored. It supports a directory tree-like structure consisting of all the files present in HDFS on a Hadoop cluster. NameNode uses the following files for namespace:
  - **fsimage file:** It keeps track of the latest checkpoint of the namespace.
  - **edits file:** It is a log of changes that have been made to the namespace since the checkpoint.
- Checkpoint NameNode has the same directory structure as NameNode. Checkpoint NameNode creates checkpoints for namespace at regular intervals by downloading the fsimage and editing files and margining them within the local directory. The new image after merging is then uploaded to NameNode. There is a similar node to Checkpoint, commonly known as the Secondary Node, but it does not support the upload-to-NameNode functionality.
- Backup Node executes the online streaming of the File system edits transaction in the Primary Namenode. It is also responsible for implementing the Checkpoint functionality and acts as the dynamic backup for the Filesystem Namespace (Metadata) in the Hadoop system.

36. What are the most common input formats in Hadoop?

There are three most common input formats in Hadoop:

- **Text Input Format:** Default input format in Hadoop
- **Key-value Input Format:** Used for plain text files where the files are broken into lines
- **Sequence File Input Format:** Used for reading files in sequence

37. What are the most common output formats in Hadoop?

The following are the commonly used output formats in Hadoop:

- **Textoutputformat:** TextOutputFormat is by default the output format in Hadoop.
- **Mapfileoutputformat:** Mapfileoutputformat writes the output as map files in Hadoop.
- **DBoutputformat:** DBoutputformat writes the output in relational databases and Hbase.
- **Sequencefileoutputformat:** Sequencefileoutputformat is used in writing sequence files.
- **SequencefileAsBinaryoutputformat:** SequencefileAsBinaryoutputformat is used in writing keys to a sequence file in binary format.

38. How to execute a Pig script?

The three methods listed below enable users to execute a Pig script:

- Grunt shell
- Embedded script
- Script file

39. What is Apache Pig and why is it preferred over MapReduce?

Apache Pig is a Hadoop-based platform that allows professionals to analyze large sets of data and represent them as data flows. Pig reduces the complexities that are required while writing a program in MapReduce, giving it an edge over MapReduce.

The following are some of the reasons why Pig is preferred over MapReduce:

- While Pig is a language for high-level data flow, MapReduce is a paradigm for low-level data processing.
- Without the need to write complex Java code in MapReduce, a similar result can easily be achieved in Pig.
- Pig approximately reduces the code length by 20 times, reducing the time taken for development by about 16 times than MapReduce.
- Pig offers built-in functionalities to perform numerous operations, including sorting, filters, joins, ordering, etc., which are extremely difficult to perform in MapReduce.
- Unlike MapReduce, Pig provides various nested data types such as bags, maps, and tuples.

40. What are the components of the Apache Pig architecture?

The components of the Apache Pig architecture are as follows:

- **Parser:** It is responsible for handling Pig scripts and checking the syntax of the script.
- **Optimizer:** Its function is to carry out the logical optimization such as projection pushdown, etc. It is the optimizer that receives the logical plan (DAG).
- **Compiler:** It is responsible for the conversion of the logical plan into a series of MapReduce jobs.
- **Execution Engine:** In the execution engine, MapReduce jobs get submitted in Hadoop in a sorted manner.
- **Execution Mode:** The execution modes in Apache Pig are local, and MapReduce modes and their selection entirely depends on the location where the data is stored and the place where you want to run the Pig script.

41. Mention some commands in YARN to check application status and to kill an application.

The YARN commands are mentioned below as per their functionalities:

Untitled

This command allows professionals to check the application status.

Untitled

The command mentioned above enables users to kill or terminate a particular application.

43. What are the commands to restart NameNode and all the daemons in Hadoop?

The following commands can be used to restart NameNode and all the daemons:

- NameNode can be stopped with the **./sbin /Hadoop-daemon.sh** stop NameNode command. The NameNode can be started by using the **./sbin/Hadoop-daemon.sh** start NameNode command.
- The daemons can be stopped with the **./sbin /stop-all.sh** The daemons can be started by using the **./sbin/start-all.sh** command.

44. Define DataNode. How does NameNode tackle DataNode failures?

DataNode stores data in HDFS; it is a node where actual data resides in the file system. Each DataNode sends a heartbeat message to notify that it is alive. If the NameNode does not receive a message from the DataNode for 10 minutes, the NameNode considers the DataNode to be dead or out of place and starts the replication of blocks that were hosted on that DataNode such that they are hosted on some other DataNode. A BlockReport contains a list of all blocks on a DataNode. Now, the system starts to replicate what was stored in the dead DataNode.

The NameNode manages the replication of the data blocks from one DataNode to another. In this process, the replication data gets transferred directly between DataNodes such that the data never passes the NameNode.

45. What is the significance of Sqoop's eval tool?

The eval tool in Sqoop enables users to carry out user-defined queries on the corresponding database servers and check the outcome in the console.

46. Can you name the default file formats for importing data using Apache Sqoop?

Commonly, there are two file formats in Sqoop to import data. They are:

- Delimited Text File Format
- Sequence File Format

47. What is the difference between relational database and HBase?

The difference between relational database and HBase are mentioned below:

Relational Database	HBase
It is schema-based.	It has no schema.
It is row-oriented.	It is column-oriented.
It stores normalized data.	It stores denormalized data.
It consists of thin tables.	It consists of sparsely populated tables.
There is no built-in support or provision for automatic partitioning.	It supports automated partitioning.

48. What is the jps command used for?

The jps command is used to know or check whether the Hadoop daemons are running or not. The active or running status of all Hadoop daemons, which are namenode, datanode, resourcemanager, nodemanager, are displayed by this command.

49. What are the core methods of a reducer?

The three core methods of a reducer are as follows:

- **setup():** This method is used for configuring various parameters such as input data size and distributed cache.public void setup (context)
- **reduce():** This method is the heart of the reducer and is always called once per key with the associated reduced task.public void reduce(Key, Value, context)
- **cleanup():** This method is called to clean the temporary files, only once at the end of the task.public void cleanup (context)

50. What is Apache Flume? List the components of Apache Flume

Apache Flume is a tool or system, in Hadoop, that is used for assembling, aggregating, and carrying large amounts of streaming data. This can include data such as record files, events, etc. The main function of Apache Flume is to carry this streaming data from various web servers to HDFS.

The components of Apache Flume are as below:

- Flume Channel
- Flume Source
- Flume Agent
- Flume Sink
- Flume Event

Advanced Big Data Hadoop Interview Questions

51. What are the differences between MapReduce and Pig?

The differences between MapReduce and Pig are mentioned below:

MapReduce	Pig
It has more lines of code as compared to Pig.	It has fewer lines of code.
It is a low-level language that makes it difficult to perform operations such as join.	It is a high-level language that makes it easy to perform join and other similar operations.
Its compiling process is time-consuming.	During execution, all the Pig operators are internally converted into a MapReduce job.
A MapReduce program that is written in a particular version of Hadoop may not work in others.	It works in all Hadoop versions.

52. List the configuration parameters in a MapReduce program

The configuration parameters in MapReduce are given below:

- Input locations of Jobs in the distributed file system
- Output location of Jobs in the distributed file system
- The input format of data
- The output format of data
- The class containing the map function
- The class containing the reduce function
- JAR file containing the classes—mapper, reducer, and driver

53. What is the default file size of an HDFS data block?

Hadoop keeps the default file size of an HDFS data block as 128 mb.

54. Why are the data blocks in HDFS so huge?

The reason behind the large size of the data blocks in HDFS is that the transfer happens at the disk transfer rate in the presence of large-sized blocks. On the other hand, if the size is kept small, there will be a large number of blocks to be transferred, which will force the HDFS to store too much metadata, thus increasing traffic.

55. What is a SequenceFile in Hadoop?

Extensively used in MapReduce I/O formats, SequenceFile is a flat-file containing binary key-value pairs. The map outputs are stored as SequenceFile internally. It provides reader, writer, and sorter classes. The three SequenceFile formats are as follows:

- Uncompressed key-value records
- Record compressed key-value records—only values are compressed here
- Block compressed key-value records—both keys and values are collected in blocks separately and compressed. The size of the block is configurable

56. What do you mean by WAL in HBase?

WAL is otherwise referred to as a write ahead log. This file is attached to each Region Server present inside the distributed environment. WAL stores the new data, which is yet to be kept in permanent storage. WAL is often used to recover datasets in case of any failure.

57. List the two types of metadata that are stored by the NameNode server

The NameNode server stores metadata in disk and RAM. The two types of metadata that the NameNode server stores are:

- EditLogs
- FsImage

58. Explain the architecture of YARN and how it allocates various resources to applications?

There is an application, API, or client that communicates with the ResourceManager, which then deals with allocating resources in the cluster. It has an awareness of the resources present with each node manager. There are two internal components of the ResourceManager, application manager and scheduler. The scheduler is responsible for allocating resources to the numerous applications running in parallel based on their requirements. However, the scheduler does not track the application status.

The application manager accepts the submission of jobs and manages and reboots the application master if there is a failure. It manages the applications' demands for resources and communicates with the scheduler to get the needed resources. It interacts with the NodeManager to manage and execute the tasks that monitor the jobs running. Moreover, it also monitors the resources utilized by each container.

A container consists of a set of resources, including CPU, RAM, and network bandwidth. It allows the applications to use a predefined number of resources.

The ResourceManager sends a request to the NodeManager to keep a few resources to process as soon as there is a job submission. Later, the NodeManager assigns an available container to carry out the processing. The ResourceManager then starts the application master to deal with the execution and it runs in one of the given containers. The rest of the containers available are used for the execution process. This is the overall process of how YARN allocates resources to applications via its architecture.

59. What are the differences between Sqoop and Flume?

The following are the various differences between Sqoop and Flume:

Sqoop	Flume
It works with NoSQL databases and RDBMS for importing and exporting data.	It works with streaming data, which is regularly generated in the Hadoop environment.
In Sqoop, loading data is not event-driven.	In Flume, loading data is event-driven.
It deals with data sources that are structured, and Sqoop connectors help in extracting data from them.	It extracts streaming data from application or web servers.
It takes data from RDBMS, imports it to HDFS, and exports it back to RDBMS.	Data from multiple sources flows into HDFS.

60. What is the role of a JobTracker in Hadoop?

A JobTracker's primary role is resource management, managing the TaskTrackers, tracking resource availability, and task life cycle management, tracking the tasks' progress and fault tolerance.

- JobTracker is a process that runs on a separate node, often not on a DataNode.
- JobTracker communicates with the NameNode to identify the data location.
- JobTracker finds the best TaskTracker nodes to execute the tasks on the given nodes.
- JobTracker monitors individual TaskTrackers and submits the overall job back to the client.
- JobTracker tracks the execution of MapReduce workloads local to the slave node.

61. Can you name the port numbers for JobTracker, NameNode, and TaskTracker

JobTracker: The port number for JobTracker is Port 50030

NameNode: The port number for NameNode is Port 50070

TaskTracker: The port number for TaskTracker is Port 50060

62. What are the components of the architecture of Hive?

- **User Interface:** It requests the execute interface for the driver and also builds a session for this query. Further, the query is sent to the compiler in order to create an execution plan for the same.
- **Metastore:** It stores the metadata and transfers it to the compiler to execute a query.
- **Compiler:** It creates the execution plan. It consists of a DAG of stages wherein each stage can either be a map, metadata operation, or reduce an operation or job on HDFS.
- **Execution Engine:** It bridges the gap between Hadoop and Hive and helps in processing the query. It communicates with the metastore bidirectionally in order to perform various tasks.

63. Is it possible to import or export tables in HBase?

Yes, tables can be imported and exported in HBase clusters by using the commands listed below:

For export:

Untitled

64. Why does Hive not store metadata in HDFS?

Hive stores the data of HDFS and the metadata is stored in the RDBMS or it is locally stored. HDFS does not store this metadata because the read or write operations in HDFS take a lot of time. This is why Hive uses RDBMS to store this metadata in the megastore rather than HDFS. This makes the process faster and enables you to achieve low latency.

65. What are the significant components in the execution environment of Pig?

The main components of a Pig execution environment are as follows:

- **Pig Scripts:** They are written in Pig with the help of UDFs and built-in operators and are then sent to the execution environment.
- **Parser:** It checks the script syntax and completes type checking. Parser's output is a directed acyclic graph (DAG).
- **Optimizer:** It conducts optimization with operations such as transform, merges, etc., to minimize the data in the pipeline.
- **Compiler:** It automatically converts the code that is optimized into a MapReduce job.
- **Execution Engine:** The MapReduce jobs are sent to these engines in order to get the required output.

66. What is the command used to open a connection in HBase?

The command mentioned below can be used to open a connection in HBase:

Untitled

67. What is the use of RecordReader in Hadoop?

Though InputSplit defines a slice of work, it does not describe how to access it. This is where the RecordReader class comes into the picture; it takes the byte-oriented data from its source and converts it into record-oriented key-value pairs such that it is fit for the Mapper task to read it. Meanwhile, InputFormat defines this Hadoop RecordReader instance.

68. How does Sqoop import or export data between HDFS and RDBMS?

The steps followed by Sqoop to import and export data, using its architecture, between HDFS and RDBMS are listed below:

- Search the database to collect metadata.
- Sqoop splits the input dataset and makes use of the respective map jobs to push these splits to HDFS.
- Search the database to collect metadata.
- Sqoop splits the input dataset and makes use of respective map jobs to push these splits to RDBMS. Sqoop exports back the Hadoop files to the RDBMS tables.

69. What is speculative execution in Hadoop?

One limitation of Hadoop is that by distributing the tasks on several nodes, there are chances that a few slow nodes limit the rest of the program. There are various reasons for the tasks to be slow, which are sometimes not easy to detect. Instead of identifying and fixing the slow-running tasks, Hadoop tries to detect when the task runs slower than expected and then launches other equivalent tasks as a backup. This backup mechanism in Hadoop is speculative execution.

Speculative execution creates a duplicate task on another disk. The same input can be processed multiple times in parallel. When most tasks in a job come to completion, the speculative execution mechanism schedules duplicate copies of the remaining tasks, which are slower, across the nodes that are free currently. When these tasks are finished, it is intimated to the JobTracker. If other copies are executing speculatively, then Hadoop notifies the TaskTrackers to quit those tasks and reject their output.

Untitled

70. What is Apache Oozie?

Apache Oozie is nothing but a scheduler that helps to schedule jobs in Hadoop and bundles them as a single logical work. Oozie jobs can largely be divided into the following two categories:

- **Oozie Workflow:** These jobs are a set of sequential actions that need to be executed.
- **Oozie Coordinator:** These jobs are triggered as and when there is data available for them, until which, it rests.

71. What happens if you try to run a Hadoop job with an output directory that is already present?

It will throw an exception saying that the output file directory already exists.

To run the MapReduce job, it needs to be ensured that the output directory does not exist in the HDFS.

To delete the directory before running the job, shell can be used:

Untitled

Or the Java API:

Untitled

72. How can you debug Hadoop code?

First, the list of MapReduce jobs currently running should be checked. Next, it needs to be ensured that there are no orphaned jobs running; if yes, the location of RM logs needs to be determined.

- Run:

Untitled

Look for the log directory in the displayed result. Find out the job ID from the displayed list and check if there is an error message associated with that job.

- On the basis of RM logs, identify the worker node that was involved in the execution of the task.
- Now, log in to that node and run the below-mentioned code:

Untitled

- Then, examine the NodeManager The majority of errors come from the user-level logs for each MapReduce job.

73. How to configure the replication factor in HDFS?

The hdfs-site.xml file is used to configure HDFS. Changing the dfs.replication property in hdfs-site.xml will change the default replication for all the files placed in HDFS.

The replication factor on a per-file basis can also be modified by using the following:

Untitled

The replication factor of all the files under a directory can also be changed.

Untitled

74. How to compress a mapper output not touching reducer output?

To achieve this compression, the following should be set:

Untitled

75. What are the basic parameters of a mapper?

Given below are the basic parameters of a mapper:

- LongWritable and Text
- Text and IntWritable

76. What is the difference between map-side join and reduce-side join?

Map-side join is performed when data reaches the map. A strict structure is needed for defining map-side join.

On the other hand, reduce-side join, or repartitioned join, is simpler than map-side join since the input datasets in reduce-side join need not be structured. However, it is less efficient as it will have to go through sort and shuffle phases, coming with network overheads.

77. How can you transfer data from Hive to HDFS?

By writing the query:

Untitled

Write the query for the data to be imported from Hive to HDFS. The output received will be stored in part files in the specified HDFS path.

78. Which companies use Hadoop?

- Yahoo! – It is the biggest contributor to the creation of Hadoop; its search engine uses Hadoop
- Facebook – developed Hive for analysis
- Amazon
- Netflix
- Adobe
- eBay
- Spotify
- Twitter