Top 50 Big Data Interview Questions And Answers - Whizlabs

The era of big data has just begun. With more companies inclined towards big data to run their operations, the demand for talent at an all-time high. What does it mean for you? It only translates into better opportunities if you want to get employed in any of the big data positions. You can choose to become Data Analyst, Data Scientist, Database administrator, Big Data Engineer, Hadoop Big Data Engineer

and so on. In this article, we will go through top 25 big data interview questions related to Big Data.

Also, this article is equally useful for anyone who is preparing for Hadoop developer interview as a fresher or experienced.

Recommended Reading: Big Data Trends in 2018

50 Most Popular Big Data Interview Questions

To give your career an edge, you should be well-prepared for the big data interview. Before we start, it is important to understand that interview is a place where you and the interviewer interact only to understand each other, and not the other way around. Hence, you don't have to hide anything, just be honest and reply to the questions with honesty. If you feel confused or need more information, feel free to ask questions to the interviewer. Always be honest with your response, and ask questions when required.

Here are top Big Data interview questions with the detailed answers to the specific questions. For broader questions that's answer depends on your experience, we will share some tips on how to answer them.

Basic Big Data Interview Questions

Whenever you go for a Big Data interview, the interviewer may ask some basic level questions. Whether you are a fresher or experienced in the big data field, the basic knowledge is required. So, let's cover some frequently asked basic big data interview questions and answers to crack big data interview.

1. What do you know about the term "Big Data"?

Answer: Big Data is a term associated with complex and large datasets. A relational database cannot handle big data, and that's why special tools and methods are used to perform operations on a vast collection of data. Big data enables companies to understand their business better and helps them derive meaningful information from the unstructured and raw data collected on a regular basis. Big data also allows the companies to take better business decisions backed by data.

2. What are the five V's of Big Data?

Answer: The five V's of Big data is as follows:

- **Volume –** Volume represents the volume i.e. amount of data that is growing at a high rate i.e. data volume in Petabytes

- **Velocity –** Velocity is the rate at which data grows. Social media contributes a major role in the velocity of growing data.
- **Variety –** Variety refers to the different data types i.e. various data formats like text, audios, videos, etc.
- **Veracity –** Veracity refers to the uncertainty of available data. Veracity arises due to the high volume of data that brings incompleteness and inconsistency.
- **Value –**Value refers to turning data into value. By turning accessed big data into values, businesses may generate revenue.

5 V's of Big Data

Note: This is one of the basic and significant questions asked in the big data interview. You can choose to explain the five V's in detail if you see the interviewer is interested to know more. However, the names can even be mentioned if you are asked about the term "Big Data".

3. Tell us how big data and Hadoop are related to each other.

Answer: Big data and Hadoop are almost synonyms terms. With the rise of big data, Hadoop, a framework that specializes in big data operations also became popular. The framework can be used by professionals to analyze big data and help businesses to make decisions.

Note: This question is commonly asked in a big data interview. *Y*ou can go further to answer this question and try to explain the main components of Hadoop.

4. How is big data analysis helpful in increasing business revenue?

Answer: Big data analysis has become very important for the businesses. It helps businesses to differentiate themselves from others and increase the revenue. Through predictive analytics, big data analytics provides businesses customized recommendations and suggestions. Also, big data analytics enables businesses to launch new products depending on customer needs and preferences. These factors make businesses earn more revenue, and thus companies are using big data analytics. Companies may encounter a significant increase of 5-20% in revenue by implementing big data analytics. Some popular companies those are using big data analytics to increase their revenue is – Walmart, LinkedIn, Facebook, Twitter, Bank of America etc.

5. Explain the steps to be followed to deploy a Big Data solution.

Answer: Followings are the three steps that are followed to deploy a Big Data Solution –

i. Data Ingestion

The first step for deploying a big data solution is the data ingestion i.e. extraction of data from various sources. The data source may be a CRM like Salesforce, Enterprise Resource Planning System like SAP, RDBMS like MySQL or any other log files, documents, social media feeds etc. The data can be ingested either through batch jobs or real-time streaming. The extracted data is then stored in HDFS.

Steps of Deploying Big Data Solution

ii. Data Storage

After data ingestion, the next step is to store the extracted data. The data either be stored in HDFS or NoSQL database (i.e. HBase). The HDFS storage works well for sequential access whereas HBase for random read/write access.

iii. Data Processing

The final step in deploying a big data solution is the data processing. The data is processed through one of the processing frameworks like Spark, MapReduce, Pig, etc.

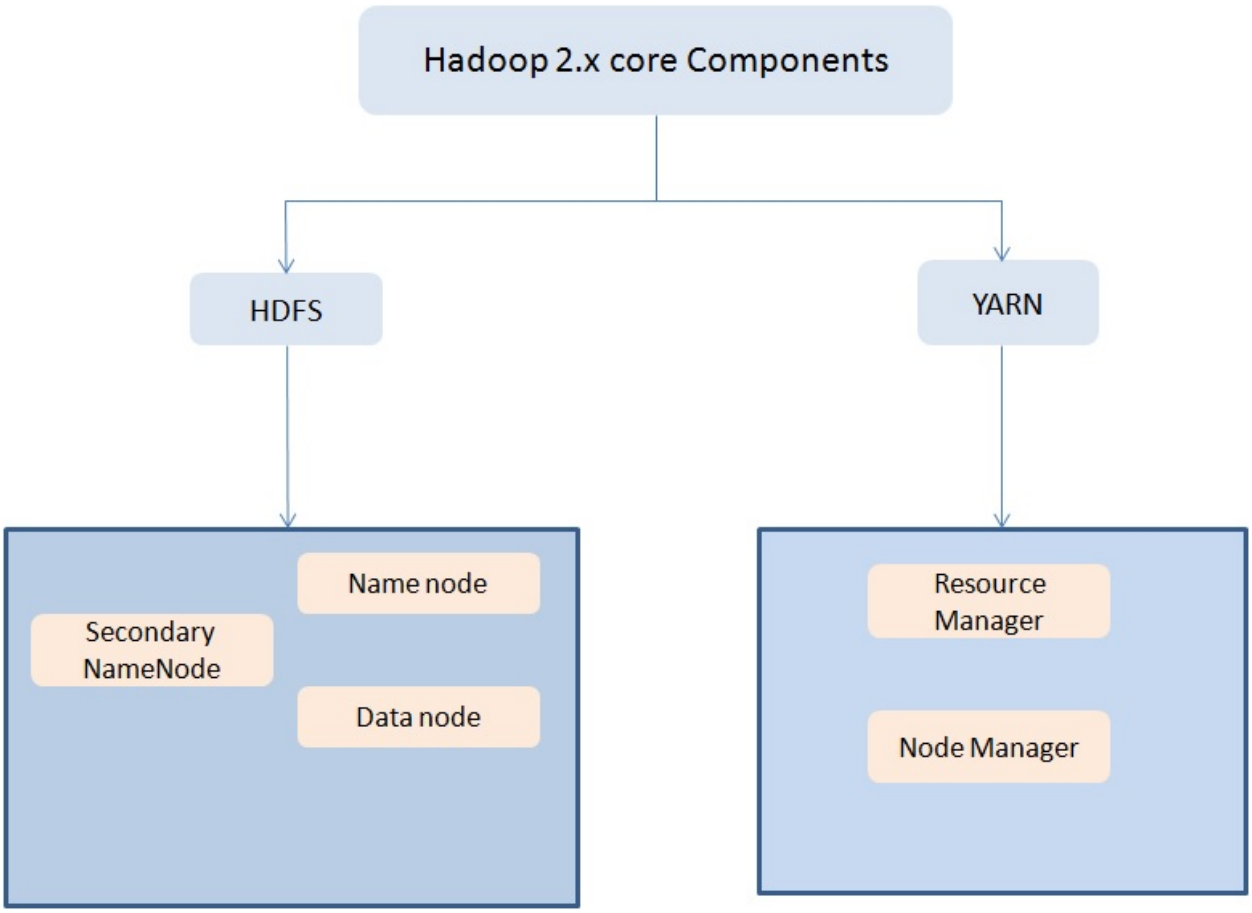
6. Define respective components of HDFS and YARN

Answer: The two main components of HDFS are-

- NameNode – This is the master node for processing metadata information for data blocks within the HDFS
- DataNode/Slave node – This is the node which acts as slave node to store the data, for processing and use by the NameNode

In addition to serving the client requests, the NameNode executes either of two following roles –

- CheckpointNode – It runs on a different host from the NameNode
- BackupNode- It is a read-only NameNode which contains file system metadata information excluding the block locations



The two main components of YARN are–

- ResourceManager– This component receives processing requests and accordingly allocates to respective NodeManagers depending on processing needs.
- NodeManager– It executes tasks on each single Data Node

7. Why is Hadoop used for Big Data Analytics?

Answer: Since data analysis has become one of the key parameters of business, hence, enterprises are dealing with massive amount of structured, unstructured and semi-structured data. Analyzing

unstructured data is quite difficult where Hadoop takes major part with its capabilities of

- Storage
- Processing
- Data collection

Moreover, Hadoop is open source and runs on commodity hardware. Hence it is a cost-benefit solution for businesses.

8. **What is fsck?**

Answer: fsck stands for File System Check. It is a command used by HDFS. This command is used to check inconsistencies and if there is any problem in the file. For example, if there are any missing blocks for a file, HDFS gets notified through this command.

9. What are the main differences between NAS (Network-attached storage) and HDFS?

Answer: The main differences between NAS (Network-attached storage) and HDFS –

- HDFS runs on a cluster of machines while NAS runs on an individual machine. Hence, data redundancy is a common issue in HDFS. On the contrary, the replication protocol is different in case of NAS. Thus the chances of data redundancy are much less.

- Data is stored as data blocks in local drives in case of HDFS. In case of NAS, it is stored in dedicated hardware.

10. What is the Command to format the NameNode?

Answer: $ hdfs namenode -format

> **"Big data is not just what you think, it's a broad spectrum. There are a number of career options in Big Data World. Here is an interesting and explanatory visual on Big Data Careers."**

Experience-based Big Data Interview Questions

If you have some considerable experience of working in Big Data world, you will be asked a number of questions in your big data interview based on your previous experience. These questions may be simply related to your experience or scenario based. So, get prepared with these best Big data interview questions and answers –

11. Do you have any Big Data experience? If so, please share it with us.

How to Approach: There is no specific answer to the question as it is a subjective question and the answer depends on your previous

experience. Asking this question during a big data interview, the interviewer wants to understand your previous experience and is also trying to evaluate if you are fit for the project requirement.

So, how will you approach the question? If you have previous experience, start with your duties in your past position and slowly add details to the conversation. Tell them about your contributions that made the project successful. This question is generally, the 2nd or 3rd question asked in an interview. The later questions are based on this question, so answer it carefully. You should also take care not to go overboard with a single aspect of your previous job. Keep it simple and to the point.

12. Do you prefer good data or good models? Why?

How to Approach: This is a tricky question but generally asked in the big data interview. It asks you to choose between good data or good models. As a candidate, you should try to answer it from your experience. Many companies want to follow a strict process of evaluating data, means they have already selected data models. In this case, having good data can be game-changing. The other way around also works as a model is chosen based on good data.

As we already mentioned, answer it from your experience. However, don't say that having both good data and good models is important as it is hard to have both in real life projects.

13. Will you optimize algorithms or code to make them run faster?

How to Approach: The answer to this question should always be "Yes." Real world performance matters and it doesn't depend on the data or model you are using in your project.

The interviewer might also be interested to know if you have had any previous experience in code or algorithm optimization. For a beginner, it obviously depends on which projects he worked on in the past. Experienced candidates can share their experience accordingly as well. However, be honest about your work, and it is fine if you haven't optimized code in the past. Just let the interviewer know your real experience and you will be able to crack the big data interview.

14. How do you approach data preparation?

How to Approach: Data preparation is one of the crucial steps in big data projects. A big data interview may involve at least one question based on data preparation. When the interviewer asks you this question, he wants to know what steps or precautions you take during data preparation.

As you already know, data preparation is required to get necessary data which can then further be used for modeling purposes. You should convey this message to the interviewer. You should also emphasize the type of model you are going to use and reasons behind choosing that

particular model. Last, but not the least, you should also discuss important data preparation terms such as transforming variables, outlier values, unstructured data, identifying gaps, and others.

15. How would you transform unstructured data into structured data?

How to Approach: Unstructured data is very common in big data. The unstructured data should be transformed into structured data to ensure proper data analysis. You can start answering the question by briefly differentiating between the two. Once done, you can now discuss the methods you use to transform one form to another. You might also share the real-world situation where you did it. If you have recently been graduated, then you can share information related to your academic projects.

By answering this question correctly, you are signaling that you understand the types of data, both structured and unstructured, and also have the practical experience to work with these. If you give an answer to this question specifically, you will definitely be able to crack the big data interview.

16. Which hardware configuration is most beneficial for Hadoop jobs?

Dual processors or core machines with a configuration of 4 / 8 GB RAM and ECC memory is ideal for running Hadoop operations. However, the hardware configuration varies based on the project-specific workflow and process flow and need customization accordingly.

17. What happens when two users try to access the same file in the HDFS?

HDFS NameNode supports exclusive write only. Hence, only the first user will receive the grant for file access and the second user will be rejected.

18. How to recover a NameNode when it is down?

The following steps need to execute to make the Hadoop cluster up and running:

1. Use the FsImage which is file system metadata replica to start a new NameNode.
2. Configure the DataNodes and also the clients to make them acknowledge the newly started NameNode.
3. Once the new NameNode completes loading the last checkpoint FsImage which has received enough block reports from the DataNodes, it will start to serve the client.

In case of large Hadoop clusters, the NameNode recovery process consumes a lot of time which turns out to be a more significant challenge in case of routine maintenance.

19. What do you understand by Rack Awareness in Hadoop?

It is an algorithm applied to the NameNode to decide how blocks and its replicas are placed. Depending on rack definitions network traffic is minimized between DataNodes within the same rack. For example, if we consider replication factor as 3, two copies will be placed on one rack whereas the third copy in a separate rack.

20. What is the difference between "HDFS Block" and "Input Split"?

The HDFS divides the input data physically into blocks for processing which is known as HDFS Block.

Input Split is a logical division of data by mapper for mapping operation.

Enhance your Big Data skills with the experts. Here is the Complete List of Big Data Blogs where you can find latest news, trends, updates, and concepts of Big Data.

Basic Big Data Hadoop Interview Questions

Hadoop is one of the most popular Big Data frameworks, and if you are going for a Hadoop interview prepare yourself with these basic level interview questions for Big Data Hadoop. These questions will be helpful for you whether you are going for a Hadoop developer or Hadoop Admin interview.

21. Explain the difference between Hadoop and RDBMS.

Answer: The difference between Hadoop and RDBMS is as follows –

Criteria	Hadoop	RDBMS
Schema	Based on 'Schema on Read'	Based on 'Schema on Write'
Data Types	Structured, Semi-structured and Unstructured data	Structured Data
Speed	Writes are Fast	Reads are Fast
Cost	Open source framework, free of cost	Licensed software, paid
Applications	Data discovery, Storage and processing of unstructured data	OLTP and complex ACID transactions

22. What are the common input formats in Hadoop?

Answer: Below are the common input formats in Hadoop –

- **Text Input Format –** The default input format defined in Hadoop is the Text Input Format.
- **Sequence File Input Format –** To read files in a sequence, Sequence File Input Format is used.
- **Key Value Input Format –** The input format used for plain text files (files broken into lines) is the Key Value Input Format.

23. Explain some important features of Hadoop.

Answer: Hadoop supports the storage and processing of big data. It is the best solution for handling big data challenges. Some important features of Hadoop are –

- **Open Source –** Hadoop is an open source framework which means it is available free of cost. Also, the users are allowed to change the source code as per their requirements.
- **Distributed Processing –** Hadoop supports distributed processing of data i.e. faster processing. The data in Hadoop HDFS is stored in a distributed manner and MapReduce is responsible for the parallel processing of data.
- **Fault Tolerance –** Hadoop is highly fault-tolerant. It creates three replicas for each block at different nodes, by default. This number can be changed according to the requirement. So, we can recover the data from another node if one node fails. The detection of node failure and recovery of data is done automatically.
- **Reliability –** Hadoop stores data on the cluster in a reliable manner that is independent of machine. So, the data stored in Hadoop environment is not affected by the failure of the machine.
- **Scalability –** Another important feature of Hadoop is the scalability. It is compatible with the other hardware and we can easily ass the new hardware to the nodes.
- **High Availability –** The data stored in Hadoop is available to access even after the hardware failure. In case of hardware failure, the data can be accessed from another path.

24. Explain the different modes in which Hadoop run.

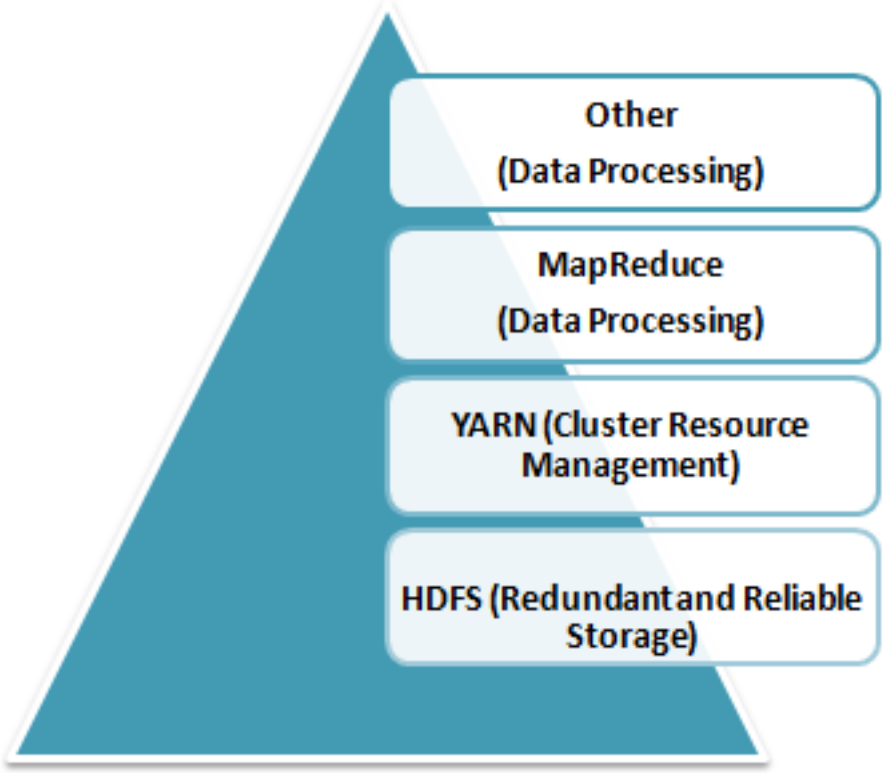
Answer: Apache Hadoop runs in the following three modes –

- **Standalone (Local) Mode –** By default, Hadoop runs in a local mode i.e. on a non-distributed, single node. This mode uses the local file system to perform input and output operation. This mode does not support the use of HDFS, so it is used for debugging. No custom configuration is needed for configuration files in this mode.
- **Pseudo-Distributed Mode –** In the pseudo-distributed mode, Hadoop runs on a single node just like the Standalone mode. In this mode, each daemon runs in a separate Java process. As all the daemons run on a single node, there is the same node for both the Master and Slave nodes.
- **Fully – Distributed Mode –** In the fully-distributed mode, all the daemons run on separate individual nodes and thus forms a multi-node cluster. There are different nodes for Master and Slave nodes.

25. Explain the core components of Hadoop.

Answer: Hadoop is an open source framework that is meant for storage and processing of big data in a distributed manner. The core components of Hadoop are –

- **HDFS (Hadoop Distributed File System) –** HDFS is the basic storage system of Hadoop. The large data files running on a cluster of commodity hardware are stored in HDFS. It can store data in a reliable manner even when hardware fails.



Core Components of Hadoop

- **Hadoop MapReduce –** MapReduce is the Hadoop layer that is responsible for data processing. It writes an application to process unstructured and structured data stored in HDFS. It is responsible for the parallel processing of high volume of data by dividing data into independent tasks. The processing is done in two phases Map and Reduce. The Map is the first phase of processing that specifies complex logic code and the Reduce is the second phase of processing that specifies light-weight operations.
- **YARN –** The processing framework in Hadoop is YARN. It is used for resource management and provides multiple data processing engines i.e. data science, real-time streaming, and batch processing.

26. What are the configuration parameters in a “MapReduce” program?

The main configuration parameters in “MapReduce” framework are:

- 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 which contains the map function
- The class which contains the reduce function
- JAR file which contains the mapper, reducer and the driver classes

27. What is a block in HDFS and what is its default size in Hadoop 1 and Hadoop 2? Can we change the block size?

Blocks are smallest continuous data storage in a hard drive. For HDFS, blocks are stored across Hadoop cluster.

- The default block size in Hadoop 1 is: 64 MB
- The default block size in Hadoop 2 is: 128 MB

Yes, we can change block size by using the parameter – **dfs.block.size** located in the hdfs-site.xml file.

28. What is Distributed Cache in a MapReduce Framework

Distributed Cache is a feature of Hadoop MapReduce framework to cache files for applications. Hadoop framework makes cached files available for every map/reduce tasks running on the data nodes. Hence, the data files can access the cache file as a local file in the designated job.

29. What are the three running modes of Hadoop?

The three running modes of Hadoop are as follows:

i. Standalone or local: This is the default mode and does not need any configuration. In this mode, all the following components of Hadoop uses local file system and runs on a single JVM –

- NameNode
- DataNode
- ResourceManager
- NodeManager

ii. Pseudo-distributed: *In this mode, all the master and slave Hadoop services are deployed and executed on a single node.*

iii. Fully distributed: In this mode, Hadoop master and slave services are deployed and executed on separate nodes.

30. Explain JobTracker in Hadoop

JobTracker is a JVM process in Hadoop to submit and track MapReduce jobs.

JobTracker performs the following activities in Hadoop in a sequence –

- JobTracker receives jobs that a client application submits to the job tracker
- JobTracker notifies NameNode to determine data node
- JobTracker allocates TaskTracker nodes based on available slots.
- it submits the work on allocated TaskTracker Nodes,
- JobTracker monitors the TaskTracker nodes.
- When a task fails, JobTracker is notified and decides how to reallocate the task.

Prepare yourself for the next Hadoop Job Interview with Top 50 Hadoop Interview Questions and Answers.

Hadoop Developer Interview Questions for Fresher

It is not easy to crack Hadoop developer interview but the preparation can do everything. If you are a fresher, learn the Hadoop concepts and prepare properly. Have a good knowledge of the different file systems, Hadoop versions, commands, system security, etc. Here are few questions that will help you pass the Hadoop developer interview.

31. What are the different configuration files in Hadoop?

Answer: The different configuration files in Hadoop are –

core-site.xml – This configuration file contains Hadoop core configuration settings, for example, I/O settings, very common for MapReduce and HDFS. It uses hostname a port.

mapred-site.xml – This configuration file specifies a framework name for MapReduce by setting mapreduce.framework.name

hdfs-site.xml – This configuration file contains HDFS daemons configuration settings. It also specifies default block permission and replication checking on HDFS.

yarn-site.xml – This configuration file specifies configuration settings for ResourceManager and NodeManager.

32. What are the differences between Hadoop 2 and Hadoop 3?

Answer: Following are the differences between Hadoop 2 and Hadoop 3 –

Criteria	Hadoop 2	Hadoop 3
Java Version	Minimum supported Java Version is Java 7	Minimum supported Java Version is Java 8
Fault Tolerance	Fault tolerance is handled by Replication that waste the space	Fault tolerance is handled by Erasure coding
Data Balancing	HDFS balancer is used for data balancing	Intra-datanode balancer is used for data balancing
Overhead in Storage Space	HDFS has 200% overhead in storage space	HDFS has 50% overhead in storage space
Default Ports	Some default ports are in ephemeral range, so fails to bind at start up	All the default ports are out of the ephemeral range, binds well at start up

33. How can you achieve security in Hadoop?

Answer: Kerberos are used to achieve security in Hadoop. There are 3 steps to access a service while using Kerberos, at a high level. Each step involves a message exchange with a server.

1. **Authentication –** The first step involves authentication of the client to the authentication server, and then provides a time-stamped TGT (Ticket-Granting Ticket) to the client.

2. **Authorization –** In this step, the client uses received TGT to request a service ticket from the TGS (Ticket Granting Server).
3. **Service Request –** It is the final step to achieve security in Hadoop. Then the client uses service ticket to authenticate himself to the server.

34. What is commodity hardware?

Answer: Commodity hardware is a low-cost system identified by less-availability and low-quality. The commodity hardware comprises of RAM as it performs a number of services that require RAM for the execution. One doesn't require high-end hardware configuration or supercomputers to run Hadoop, it can be run on any commodity hardware.

35. How is NFS different from HDFS?

Answer: There are a number of distributed file systems that work in their own way. NFS (Network File System) is one of the oldest and popular distributed file storage systems whereas HDFS (Hadoop Distributed File System) is the recently used and popular one to handle big data. The main differences between NFS and HDFS are as follows –

Criteria	NFS	HDFS
Data Size Support	NFS can store and process small amount of data	HDFS is mainly use to store and process big data
Data Storage	Data is stored on a single dedicated hardware	The data blocks are distributed on the local drives of hardware
Reliability	No reliability, data is not available in case of machine failure	Data is stored reliably, data is available even after machine failure
Data Redundancy	NFS runs on single machine, no chances of data redundancy	HDFS runs on a cluster of different machines, data redundancy may occur due to replication protocol

36. How do Hadoop MapReduce works?

There are two phases of MapReduce operation.

- Map phase – In this phase, the input data is split by map tasks. The map tasks run in parallel. These split data is used for analysis purpose.
- Reduce phase- In this phase, the similar split data is aggregated from the entire collection and shows the result.

37. What is MapReduce? What is the syntax you use to run a MapReduce program?

MapReduce is a programming model in Hadoop for processing large data sets over a cluster of computers, commonly known as HDFS. It is a parallel programming model.

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

38. What are the Port Numbers for NameNode, Task Tracker, and Job Tracker?

- **NameNode** – Port 50070
- **Task Tracker** – Port 50060
- **Job Tracker** – Port 50030

39.What are the different file permissions in HDFS for files or directory levels?

Hadoop distributed file system (HDFS) uses a specific permissions model for files and directories. Following user levels are used in HDFS –

- Owner
- Group
- Others.

For each of the user mentioned above following permissions are applicable –

- read (r)
- write (w)
- execute(x).

Above mentioned permissions work differently for files and directories.

For files –

- The **r** permission is for *reading* a file
- The **w** permission is for *writing* a file.

For directories –

- The **r** permission *lists the contents* of *a specific directory*.
- The **w** permission *creates or deletes a directory*.
- The **X** permission is for accessing a child directory.

40.What are the basic parameters of a Mapper?

The basic parameters of a Mapper are

- LongWritable and Text
- Text and IntWritable

Hadoop and Spark are the two most popular big data frameworks. But there is a commonly asked question – do

we need Hadoop to run Spark? Watch this video to find the answer to this question.

Hadoop Developer Interview Questions for Experienced

The interviewer has more expectations from an experienced Hadoop developer, and thus his questions are one-level up. So, if you have gained some experience, don't forget to cover command based, scenario-based, real-experience based questions. Here we bring some sample interview questions for experienced Hadoop developers.

41. How to restart all the daemons in Hadoop?

Answer: To restart all the daemons, it is required to stop all the daemons first. The Hadoop directory contains sbin directory that stores the script files to stop and start daemons in Hadoop.

Use stop daemons command /sbin/stop-all.sh to stop all the daemons and then use /sin/start-all.sh command to start all the daemons again.

42. What is the use of jps command in Hadoop?

Answer: The jps command is used to check if the Hadoop daemons are running properly or not. This command shows all the daemons running on a machine i.e. Datanode, Namenode, NodeManager, ResourceManager etc.

43. Explain the process that overwrites the replication factors in HDFS.

Answer: There are two methods to overwrite the replication factors in HDFS –

Method 1: On File Basis

In this method, the replication factor is changed on the basis of file using Hadoop FS shell. The command used for this is:

$hadoop fs – setrep –w2/my/test_file

Here, test_file is the filename that's replication factor will be set to 2.

Method 2: On Directory Basis

In this method, the replication factor is changed on directory basis i.e. the replication factor for all the files under a given directory is modified.

$hadoop fs –setrep –w5/my/test_dir

Here, test_dir is the name of the directory, the replication factor for the directory and all the files in it will be set to 5.

44. What will happen with a NameNode that doesn't have any data?

Answer: A NameNode without any data doesn't exist in Hadoop. If there is a NameNode, it will contain some data in it or it won't exist.

45. Explain NameNode recovery process.

Answer: The NameNode recovery process involves the below-mentioned steps to make Hadoop cluster running:

- In the first step in the recovery process, file system metadata replica (FsImage) starts a new NameNode.
- The next step is to configure DataNodes and Clients. These DataNodes and Clients will then acknowledge new NameNode.
- During the final step, the new NameNode starts serving the client on the completion of last checkpoint FsImage loading and receiving block reports from the DataNodes.

Note: Don't forget to mention, this NameNode recovery process consumes a lot of time on large Hadoop clusters. Thus, it makes routine maintenance difficult. For this reason, HDFS high availability architecture is recommended to use.

46. How Is Hadoop CLASSPATH essential to start or stop Hadoop daemons?

CLASSPATH includes necessary directories that contain jar files to start or stop Hadoop daemons. Hence, setting CLASSPATH is essential to start or stop Hadoop daemons.

However, setting up CLASSPATH every time is not the standard that we follow. Usually CLASSPATH is written inside */etc/hadoop/hadoop-env.sh* file. Hence, once we run Hadoop, it will load the CLASSPATH automatically.

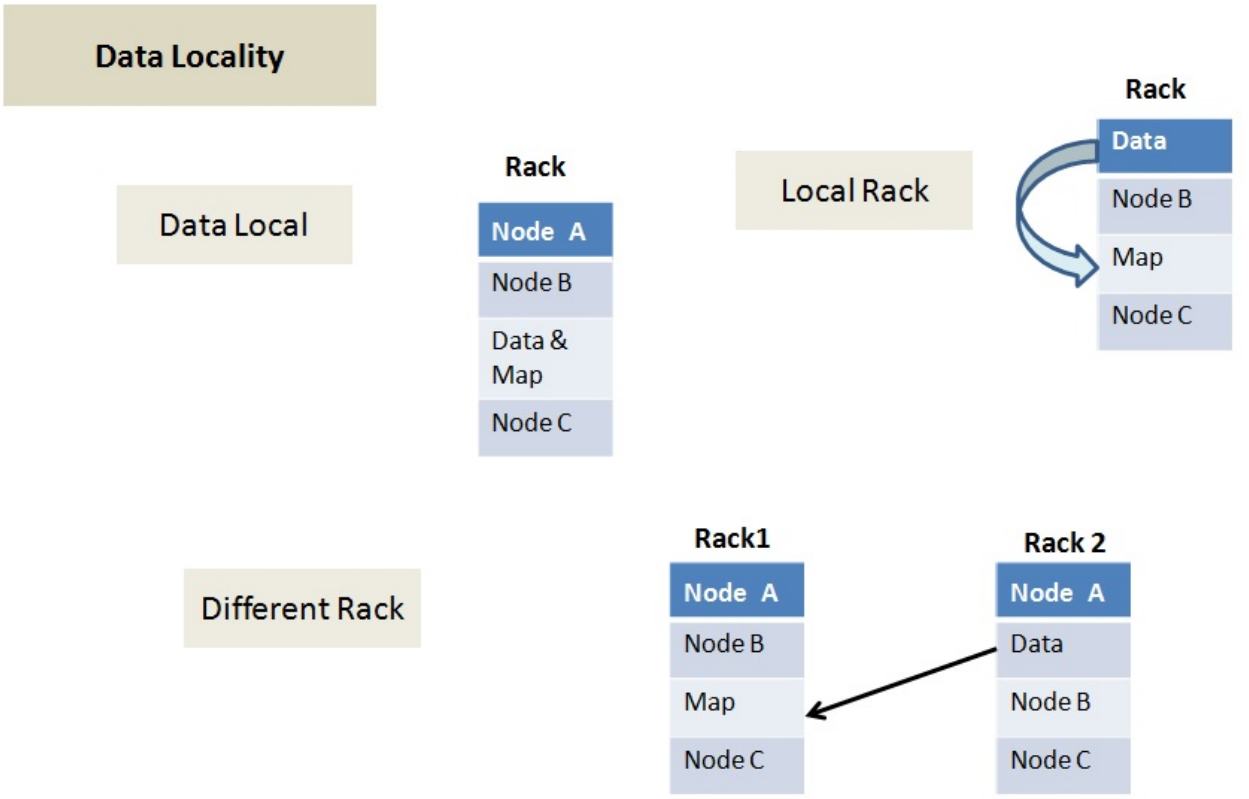
47. Why is HDFS only suitable for large data sets and not the correct tool to use for many small files?

This is due to the performance issue of NameNode. Usually, NameNode is allocated with huge space to store metadata for the large-scale file. The metadata is supposed to be a from a single file for optimum space utilization and cost benefit. In case of small size files, NameNode does not utilize the entire space which is a performance optimization issue.

48. Why do we need Data Locality in Hadoop? Explain.

Datasets in HDFS store as blocks in DataNodes the Hadoop cluster. During the execution of a MapReduce job the individual Mapper processes the blocks (Input Splits). If the data does not reside in the same node where the Mapper is executing the job, the data needs to be copied from the DataNode over the network to the mapper DataNode.

Now if a MapReduce job has more than 100 Mapper and each Mapper tries to copy the data from other DataNode in the cluster simultaneously, it would cause serious network congestion which is a big performance issue of the overall system. Hence, data proximity to the computation is an effective and cost-effective solution which is technically termed as Data locality in Hadoop. It helps to increase the overall throughput of the system.



Data locality can be of three types:

- **Data local –** In this type data and the mapper resides on the same node. This is the closest proximity of data and the most preferred scenario.
- **Rack Local –** In this scenarios mapper and data reside on the same rack but on the different data nodes.
- **Different Rack –** In this scenario mapper and data reside on the different racks.

49. DFS can handle a large volume of data then why do we need Hadoop framework?

Hadoop is not only for storing large data but also to process those big data. Though DFS(Distributed File System) too can store the data, but it lacks below features-

- It is not fault tolerant
- Data movement over a network depends on bandwidth.

50. What is Sequencefileinputformat?

Hadoop uses a specific file format which is known as Sequence file. The sequence file stores data in a serialized key-value pair. Sequencefileinputformat is an input format to read sequence files.

Final Words

Big Data world is expanding continuously and thus a number of opportunities are arising for the Big Data professionals. This top Big Data interview Q & A set will surely help you in your interview. However, we can't neglect the importance of certifications. So, if you want to demonstrate your skills to your interviewer during big data interview get certified and add a credential to your resume.

Looking for career growth? Get your career one level up with these Best Big Data Certifications in 2018. Choose your certification and prepare through our Big Data Training Courses to boost your career.

Good Luck with your interview!