***Big Data Interview Questions & Answers***

**1. Define Big Data and explain the Vs of Big Data.**

This is one of the most introductory yet important Big Data interview questions. The answer to this is quite straightforward:

Big Data can be defined as a collection of complex unstructured or semi-structured data sets which have the potential to deliver actionable insights.

**The four Vs of Big Data are –**  
**Volume –** Talks about the amount of data  
**Variety –** Talks about the various formats of data  
**Velocity –** Talks about the ever increasing speed at which the data is growing  
**Veracity –** Talks about the degree of accuracy of data available  
  
[Big Data Tutorial for Beginners: All You Need to Know](https://upgrad.com/blog/big-data-tutorial-for-beginners/)

**2. How is Hadoop related to Big Data?**

When we talk about Big Data, we talk about Hadoop. So, this is another Big Data interview question that you will definitely face in an interview.

Hadoop is an open-source framework for storing, processing, and analyzing complex unstructured data sets for deriving insights and intelligence.

**3. Define HDFS and YARN, and talk about their respective components.**

Now that we’re in the zone of Hadoop, the next Big Data interview question you might face will revolve around the same.

The HDFS is Hadoop’s default storage unit and is responsible for storing different types of data in a distributed environment.

**HDFS has the following two components:**

**NameNode –** This is the master node that has the metadata information for all the data blocks in the HDFS.  
**DataNode –** These are the nodes that act as slave nodes and are responsible for storing the data.  
**YARN, short for Yet Another Resource Negotiator**, is responsible for managing resources and providing an execution environment for the said processes.  
**The two main components of YARN are –**  
**ResourceManager –** Responsible for allocating resources to respective NodeManagers based on the needs.  
**NodeManager –** Executes tasks on every DataNode.  
[7 Interesting Big Data Projects You Need To Watch Out](https://upgrad.com/blog/interesting-big-data-projects/)

**4. What do you mean by commodity hardware?**

This is yet another Big Data interview question you’re most likely to come across in any interview you sit for.

Commodity Hardware refers to the minimal hardware resources needed to run the Apache Hadoop framework. Any hardware that supports Hadoop’s minimum requirements is known as ‘Commodity Hardware.’

**5. Define and describe the term FSCK.**

FSCK stands for Filesystem Check. It is a command used to run a Hadoop summary report that describes the state of HDFS. It only checks for errors and does not correct them. This command can be executed on either the whole system or a subset of files.

**6. What is the purpose of the JPS command in Hadoop?**

The JPS command is used for testing the working of all the Hadoop daemons. It specifically tests daemons like NameNode, DataNode, ResourceManager, NodeManager and more.  
(In any Big Data interview, you’re likely to find one question on JPS and its importance.)  
[Big Data: Must Know Tools and Technologies](https://upgrad.com/blog/big-data-must-know-tools-and-technologies/)

**7. Name the different commands for starting up and shutting down Hadoop Daemons.**

This is one of the most important Big Data interview questions to help the interviewer gauge your knowledge of commands.

**To start all the daemons:**  
./sbin/start-all.sh  
  
**To shut down all the daemons:**  
./sbin/stop-all.sh

**8. Why do we need Hadoop for Big Data Analytics?**

This Hadoop interview questions test your awareness regarding the practical aspects of Big Data and Analytics.

In most cases, Hadoop helps in exploring and analyzing large and unstructured data sets. Hadoop offers storage, processing and data collection capabilities that help in analytics.

**9. Explain the different features of Hadoop.**

Listed in many Big Data Interview Questions and Answers, the best answer to this is –

**Open-Source –** Hadoop is an open-sourced platform. It allows the code to be rewritten or modified according to user and analytics requirements.  
**Scalability –** Hadoop supports the addition of hardware resources to the new nodes.  
**Data Recovery –** Hadoop follows replication which allows the recovery of data in the case of any failure.  
**Data Locality –** This means that Hadoop moves the computation to the data and not the other way round. This way, the whole process speeds up.

**10. Define the Port Numbers for NameNode, Task Tracker and Job Tracker.**

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

**11. What do you mean by indexing in HDFS?**

HDFS indexes data blocks based on their sizes. The end of a data block points to the address of where the next chunk of data blocks get stored. The DataNodes store the blocks of data while NameNode stores these data blocks.  
[Big Data Applications in Pop-Culture](https://upgrad.com/blog/big-data-applications-in-the-pop-culture/)

**12. What are Edge Nodes in Hadoop?**

Edge nodes refer to the gateway nodes which act as an interface between Hadoop cluster and the external network. These nodes run client applications and cluster management tools and are used as staging areas as well. Enterprise-class storage capabilities are required for Edge Nodes, and a single edge node usually suffices for multiple Hadoop clusters.

**13. What are some of the data management tools used with Edge Nodes in Hadoop?**

This Big Data interview question aims to test your awareness regarding various tools and frameworks.

Oozie, Ambari, Pig and Flume are the most common data management tools that work with Edge Nodes in Hadoop.

**14. Explain the core methods of a Reducer.**

**There are three core methods of a reducer. They are-**  
  
**setup() –** This is used to configure different parameters like heap size, distributed cache and input data.  
**reduce() –** A parameter that is called once per key with the concerned reduce task  
**cleanup() –** Clears all temporary files and called only at the end of a reducer task.

**15. Talk about the different tombstone markers used for deletion purposes in HBase.**

This Big Data interview question dives into your knowledge of HBase and its working.  
There are three main tombstone markers used for deletion in HBase. They are-

**Family Delete Marker –** For marking all the columns of a column family.  
**Version Delete Marker –** For marking a single version of a single column.  
**Column Delete Marker –** For marking all the versions of a single column.  
[Big Data Engineers: Myths vs. Realities](https://upgrad.com/blog/big-data-engineers-myths-vs-realities/)

**16. How can Big Data add value to businesses?**

One of the most common big data interview question. In the present scenario, Big Data is everything. If you have data, you have the most powerful tool at your disposal. Big Data Analytics helps businesses to transform raw data into meaningful and actionable insights that can shape their business strategies. The most important contribution of Big Data to business is data-driven business decisions. Big Data makes it possible for organizations to base their decisions on tangible information and insights.

Furthermore, Predictive Analytics allows companies to craft customized recommendations and marketing strategies for different buyer personas. Together, Big Data tools and technologies help boost revenue, streamline business operations, increase productivity, and enhance customer satisfaction. In fact, anyone who’s not leveraging Big Data today is losing out on an ocean of opportunities.

**17. How do you deploy a Big Data solution?**

You can deploy a Big Data solution in three steps:

* **Data Ingestion** – This is the first step in the deployment of a Big Data solution. You begin by collecting data from multiple sources, be it social media platforms, log files, business documents, anything relevant to your business. Data can either be extracted through real-time streaming or in batch jobs.
* **Data Storage** – Once the data is extracted, you must store the data in a database. It can be HDFS or HBase. While HDFS storage is perfect for sequential access, HBase is ideal for random read/write access.
* **Data Processing** – The last step in the deployment of the solution is data processing. Usually, data processing is done via frameworks like Hadoop, Spark, MapReduce, Flink, and Pig, to name a few.

**18. How is NFS different from HDFS?**

The Network File System (NFS) is one of the oldest distributed file storage systems, while Hadoop Distributed File System (HDFS) came to the spotlight only recently after the upsurge of Big Data.

The table below highlights some of the most notable differences between NFS and HDFS:

|  |  |
| --- | --- |
| **NFS** | **HDFS** |
| It can both store and process small volumes of data. | It is explicitly designed to store and process Big Data. |
| The data is stored in dedicated hardware. | Data is divided into data blocks that are distributed on the local drives of the hardware. |
| In the case of system failure, you cannot access the data. | Data can be accessed even in the case of a system failure. |
| Since NFS runs on a single machine, there’s no chance for data redundancy. | HDFS runs on a cluster of machines, and hence, the replication protocol may lead to redundant data. |

**19. List the different file permissions in HDFS for files or directory levels.**

One of the common big data interview questions. The Hadoop distributed file system (HDFS) has specific permissions for files and directories. There are three user levels in HDFS – Owner, Group, and Others. For each of the user levels, there are three available permissions:

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

These three permissions work uniquely for files and directories.

For files –

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

Although there’s an execute(x) permission, you cannot execute HDFS files.

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.

**20. Elaborate on the processes that overwrite the replication factors in HDFS.**

In HDFS, there are two ways to overwrite the replication factors – on file basis and on directory basis.

**On File Basis**

In this method, the replication factor changes according to the file using Hadoop FS shell. The following command is used for this:

$hadoop fs – setrep –w2/my/test\_file

Here, test\_file refers to the filename whose replication factor will be set to 2.

**On Directory Basis**

This method changes the replication factor according to the directory, as such, the replication factor for all the files under a particular directory, changes. The following command is used for this:

$hadoop fs –setrep –w5/my/test\_dir

Here, test\_dir refers to the name of the directory for which the replication factor and all the files contained within will be set to 5.

**21. Name the three modes in which you can run Hadoop.**

One of the most common question in any big data interview. The three modes are:

* **Standalone mode** – This is Hadoop’s default mode that uses the local file system for both input and output operations. The main purpose of the standalone mode is debugging. It does not support HDFS and also lacks custom configuration required for mapred-site.xml, core-site.xml, and hdfs-site.xml files.
* **Pseudo-distributed mode** – Also known as the single-node cluster, the pseudo-distributed mode includes both NameNode and DataNode within the same machine. In this mode, all the Hadoop daemons will run on a single node, and hence, the Master and Slave nodes are the same.
* **Fully distributed mode** – This mode is known as the multi-node cluster wherein multiple nodes function simultaneously to execute [Hadoop](https://intellipaat.com/blog/what-is-hadoop/" \t "_blank) jobs. Here, all the Hadoop daemons run on different nodes. So, the Master and Slave nodes run separately.

**22. Explain “Overfitting.”**

Overfitting refers to a modeling error that occurs when a function is tightly fit (influenced) by a limited set of data points. Overfitting results in an overly complex model that makes it further difficult to explain the peculiarities or idiosyncrasies in the data at hand. As it adversely affects the generalization ability of the model, it becomes challenging to determine the predictive quotient of overfitted models. These models fail to perform when applied to external data (data that is not part of the sample data) or new datasets.

Overfitting is one of the most common problems in Machine Learning. A model is considered to be overfitted when it performs better on the training set but fails miserably on the test set. However, there are many methods to prevent the problem of overfitting, such as cross-validation, pruning, early stopping, regularization, and assembling.

**23. What is Feature Selection?**

Feature selection refers to the process of extracting only the required features from a specific dataset. When data is extracted from disparate sources, not all data is useful at all times – different business needs call for different data insights. This is where feature selection comes in to identify and select only those features that are relevant for a particular business requirement or stage of data processing.

The main goal of feature selection is to simplify ML models to make their analysis and interpretation easier. Feature selection enhances the generalization abilities of a model and eliminates the problems of dimensionality, thereby, preventing the possibilities of overfitting. Thus, feature selection provides a better understanding of the data under study, improves the prediction performance of the model, and reduces the computation time significantly.

Feature selection can be done via three techniques:

* **Filters method**

In this method, the features selected are not dependent on the designated classifiers. A variable ranking technique is used to select variables for ordering purposes. During the classification process, the variable ranking technique takes into consideration the importance and usefulness of a feature. The Chi-Square Test, Variance Threshold, and Information Gain are some examples of the filters method.

* **Wrappers method**

In this method, the algorithm used for feature subset selection exists as a ‘wrapper’ around the induction algorithm. The induction algorithm functions like a ‘Black Box’ that produces a classifier that will be further used in the classification of features. The major drawback or limitation of the wrappers method is that to obtain the feature subset, you need to perform heavy computation work. Genetic Algorithms, Sequential Feature Selection, and Recursive Feature Elimination are examples of the wrappers method.

* **Embedded method**

The embedded method combines the best of both worlds – it includes the best features of the filters and wrappers methods. In this method, the variable selection is done during the training process, thereby allowing you to identify the features that are the most accurate for a given model. L1 Regularisation Technique and Ridge Regression are two popular examples of the embedded method.

**24. Define “Outliers.”**

An outlier refers to a data point or an observation that lies at an abnormal distance from other values in a random sample. In other words, outliers are the values that are far removed from the group; they do not belong to any specific cluster or group in the dataset. The presence of outliers usually affects the behavior of the model – they can mislead the training process of ML algorithms. Some of the adverse impacts of outliers include longer training time, inaccurate models, and poor outcomes.

However, outliers may sometimes contain valuable information. This is why they must be investigated thoroughly and treated accordingly.

**25. Name some outlier detection techniques.**

Again, one of the most important big data interview questions. Here are six outlier detection methods:

* **Extreme Value Analysis** – This method determines the statistical tails of the data distribution. Statistical methods like ‘z-scores’ on univariate data are a perfect example of extreme value analysis.
* **Probabilistic and Statistical Models** – This method determines the ‘unlikely instances’ from a ‘probabilistic model’ of data. A good example is the optimization of Gaussian mixture models using ‘expectation-maximization’.
* **Linear Models** – This method models the data into lower dimensions. Proximity-based Models – In this approach, the data instances that are isolated from the data group are determined by Cluster, Density, or by the Nearest Neighbor Analysis.
* **Information-Theoretic Models** – This approach seeks to detect outliers as the bad data instances that increase the complexity of the dataset.
* **High-Dimensional Outlier Detection** – This method identifies the subspaces for the outliers according to the distance measures in higher dimensions.

**26. Explain Rack Awareness in Hadoop.**

Rack Awareness is one of the popular big data interview questions. Rach awareness is an algorithm that identifies and selects DataNodes closer to the NameNode based on their rack information. It is applied to the NameNode to determine how data blocks and their replicas will be placed. During the installation process, the default assumption is that all nodes belong to the same rack.

**Rack awareness helps to:**

* Improve data reliability and accessibility.
* Improve cluster performance.
* Improve network bandwidth.
* Keep the bulk flow in-rack as and when possible.
* Prevent data loss in case of a complete rack failure.

**27. Can you recover a NameNode when it is down? If so, how?**

Yes, it is possible to recover a NameNode when it is down. Here’s how you can do it:

* Use the FsImage (the file system metadata replica) to launch a new NameNode.
* Configure DataNodes along with the clients so that they can acknowledge and refer to newly started NameNode.
* When the newly created NameNode completes loading the last checkpoint of the FsImage (that has now received enough block reports from the DataNodes) loading process, it will be ready to start serving the client.

However, the recovery process of a NameNode is feasible only for smaller clusters. For large Hadoop clusters, the recovery process usually consumes a substantial amount of time, thereby making it quite a challenging task.

**28. Name the configuration parameters of a MapReduce framework.**

The configuration parameters in the MapReduce framework include:

* The input format of data.
* The output format of data.
* The input location of jobs in the distributed file system.
* The output location of jobs in the distributed file system.
* The class containing the map function
* The class containing the reduce function
* The JAR file containing the mapper, reducer, and driver classes.

**29. What is a Distributed Cache? What are its benefits?**

Any Big Data Interview Question and Answers guide won’t complete without this question. Distributed cache in Hadoop is a service offered by the MapReduce framework used for caching files. If a file is cached for a specific job, Hadoop makes it available on individual DataNodes both in memory and in system where the map and reduce tasks are simultaneously executing. This allows you to quickly access and read cached files to populate any collection (like arrays, hashmaps, etc.) in a code.

Distributed cache offers the following benefits:

* It distributes simple, read-only text/data files and other complex types like jars, archives, etc.
* It tracks the modification timestamps of cache files which highlight the files that should not be modified until a job is executed successfully.

**30. What is a SequenceFile in Hadoop?**

In Hadoop, a SequenceFile is a flat-file that contains binary key-value pairs. It is most commonly used in MapReduce I/O formats. The map outputs are stored internally as a SequenceFile which provides the reader, writer, and sorter classes.

There are three SequenceFile formats:

* Uncompressed key-value records
* Record compressed key-value records (only ‘values’ are compressed).
* Block compressed key-value records (here, both keys and values are collected in ‘blocks’ separately and then compressed).

**31. Explain the role of a JobTracker.**

One of the common big data interview questions. The primary function of the JobTracker is resource management, which essentially means managing the TaskTrackers. Apart from this, JobTracker also tracks resource availability and handles task life cycle management (track the progress of tasks and their fault tolerance).

Some crucial features of the JobTracker are:

* It is a process that runs on a separate node (not on a DataNode).
* It communicates with the NameNode to identify data location.
* It tracks the execution of MapReduce workloads.
* It allocates TaskTracker nodes based on the available slots.
* It monitors each TaskTracker and submits the overall job report to the client.
* It finds the best TaskTracker nodes to execute specific tasks on particular nodes.

**32. Name the common input formats in Hadoop.**

Hadoop has three common input formats:

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

**33. What is the need for Data Locality in Hadoop?**

One of the important big data interview questions. In HDFS, datasets are stored as blocks in DataNodes in the Hadoop cluster. When a  MapReduce job is executing, the individual Mapper processes the data blocks (Input Splits). If the data does is not present in the same node where the Mapper executes the job, the data must be copied from the DataNode where it resides over the network to the Mapper DataNode.

When a MapReduce job has over a hundred Mappers and each Mapper DataNode tries to copy the data from another DataNode in the cluster simultaneously, it will lead to network congestion, thereby having a negative impact on the system’s overall performance. This is where Data Locality enters the scenario. Instead of moving a large chunk of data to the computation, Data Locality moves the data computation close to where the actual data resides on the DataNode. This helps improve the overall performance of the system, without causing unnecessary delay.

**34. What are the steps to achieve security in Hadoop?**

In Hadoop, Kerberos – a network authentication protocol – is used to achieve security. Kerberos is designed to offer robust authentication for client/server applications via secret-key cryptography.

When you use Kerberos to access a service, you have to undergo three steps, each of which involves a message exchange with a server. The steps are as follows:

* **Authentication** – This is the first step wherein the client is authenticated via the authentication server, after which a time-stamped TGT (Ticket Granting Ticket) is given to the client.
* Authorization – In the second step, the client uses the TGT for requesting a service ticket from the TGS (Ticket Granting Server).
* Service Request – In the final step, the client uses the service ticket to authenticate themselves to the server.

**35. How can you handle missing values in Big Data?**

Missing values refer to the values that are not present in a column. It occurs when there’s is no data value for a variable in an observation. If missing values are not handled properly, it is bound to lead to erroneous data which in turn will generate incorrect outcomes. Thus, it is highly recommended to treat missing values correctly before processing the datasets. Usually, if the number of missing values is small, the data is dropped, but if there’s a bulk of missing values, data imputation is the preferred course of action.

In Statistics, there are different ways to estimate the missing values. These include regression, multiple data imputation, listwise/pairwise deletion, maximum likelihood estimation, and approximate Bayesian bootstrap.

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

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.

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

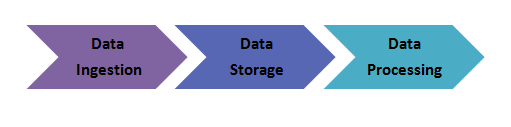
**Answer:** Big data analysis has become very important for the businesses. It helps businesses to differentiate themselves from others and increase the revenue. Through predictive analytics, big data analytics provides businesses customized recommendations and suggestions. Also, big data analytics enables 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.

**40.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.

[](https://www.whizlabs.com/wp-content/uploads/2017/11/Deploying-Big-Data-Solution.png)**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.

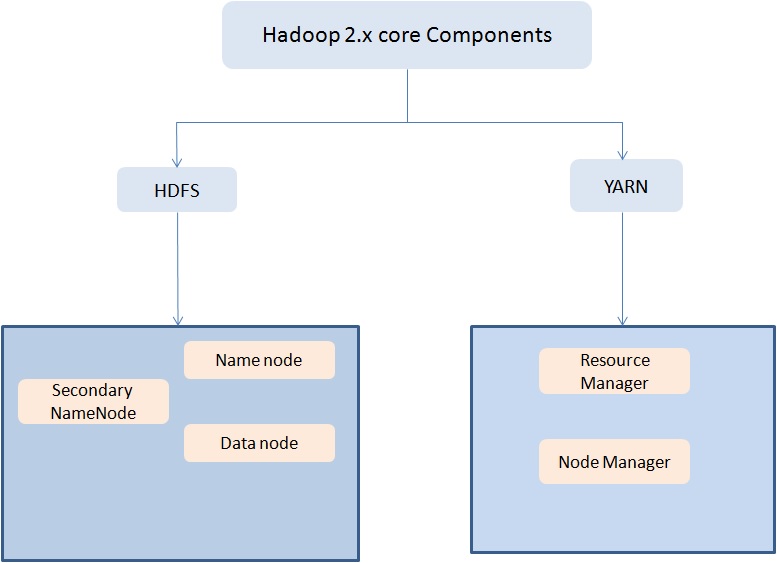
**43. Define respective components of HDFS and YARN**

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

**44. Why is Hadoop used for Big Data Analytics?**

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.

**45. What is fsck?**

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.

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

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.

**47. 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*](https://www.whizlabs.com/blog/best-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 –

**48. 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.

**49. 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.

**50. 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.

**51. 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.

**52. 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.

**53. 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.

**54. 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.

**55. 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.

**56. 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.

**57. 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.

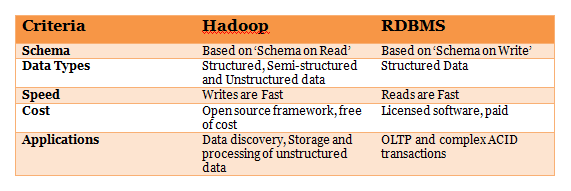
Input Split is a logical division of data by mapper for mapping operation.

*Enhance your Big Data skills with the experts. Here is the*[*Complete List of Big Data Blogs*](https://www.whizlabs.com/blog/a-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.

**58. Explain the difference between Hadoop and RDBMS.**

The difference between Hadoop and RDBMS is as follows –[](https://www.whizlabs.com/wp-content/uploads/2017/11/Hadoop-vs-RDBMS.png)22. **59.What are the common input formats in Hadoop?**

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.

**60. Explain some important features of Hadoop.**

 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.

**61. Explain the different modes in which Hadoop run.**

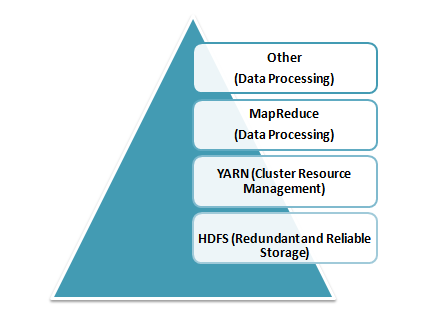
Apache Hadoop runs in the following three modes –

* **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.

**62. Explain the core components of Hadoop.**

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.

[](https://www.whizlabs.com/wp-content/uploads/2017/11/Core-Components-of-Hadoop.png)**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.

**63. 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

**64. 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.

**65. 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.

**66. 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.*](https://www.whizlabs.com/blog/top-50-hadoop-interview-questions/)

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.

**67. What are the different configuration files in Hadoop?**

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.

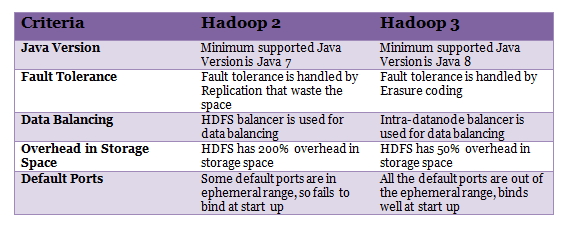
**mapred-site.xml –**This configuration file specifies a framework name for MapReduce by setting mapreduce.framework.name

**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.

**68. What are the differences between Hadoop 2 and Hadoop 3?**

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

[](https://www.whizlabs.com/wp-content/uploads/2017/11/Hadoop2-and-Hadop3.png)33. **69.How can you achieve security in Hadoop?**

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.

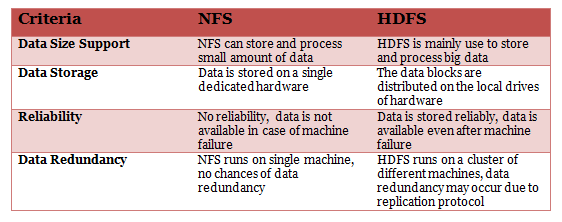
1. **Authentication –** The first step involves authentication of the client to the authentication server, and then provides a time-stamped TGT (Ticket-Granting Ticket) to the client.
2. **Authorization –** In this step, the client uses received TGT to request a service ticket from the TGS (Ticket Granting 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.

**70. What is commodity hardware?**

 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.

**71. How is NFS different from HDFS?**

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 –

[](https://www.whizlabs.com/wp-content/uploads/2017/11/NFS-vs-HDFS.png)

**72. 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.

**73. 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

**74.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.

**75.What are the basic parameters of a Mapper?**

The basic parameters of a Mapper are

* LongWritable and Text
* Text and IntWritable