Hadoop Interview Questions

There is given hadoop interview questions and answers that has been asked in many companies. Let's see the list of top hadoop interview questions.

1) What is Hadoop?

Hadoop is a distributed computing platform. It is written in Java. It consist of the features like Google File System and MapReduce.

2) What platform and Java version is required to run Hadoop?

Java 1.6.x or higher version are good for Hadoop, preferably from Sun. Linux and Windows are the supported operating system for Hadoop, but BSD, Mac OS/X and Solaris are more famous to work.

3) What kind of Hardware is best for Hadoop?

Hadoop can run on a dual processor/ dual core machines with 4-8 GB RAM using ECC memory. It depends on the workflow needs.

4) What are the most common input formats defined in Hadoop?

These are the most common input formats defined in Hadoop:

1. TextInputFormat
2. KeyValueInputFormat
3. SequenceFileInputFormat

TextInputFormat is a by default input format.

5) What is InputSplit in Hadoop? Explain.

When a hadoop job runs, it splits input files into chunks and assign each split to a mapper for processing. It is called InputSplit.

6) What is testinformat?

In textinputformat, each line in the text file is a record. Value is the content of the line while Key is the byte offset of the line. For instance, Key: longWritable, Value: text

7) What is the sequencefileinputformat in Hadoop?

In Hadoop, Sequencefileinputformat is used to read files in sequence. It is a specific compressed binary file format which passes data between the output of one MapReduce job to the input of some other MapReduce job.

8) How many InputSplits is made by a Hadoop Framework?

Hadoop will make 5 splits as following:

* One split for 64K files
* Two splits for 65MB files, and
* Two splits for 127MB files

9) What is the use of RecordReader in Hadoop?

InputSplit is assigned with a work but doesn't know how to access it. The record holder class is totally responsible for loading the data from its source and convert it into keys pair suitable for reading by the Mapper. The RecordReader's instance can be defined by the Input Format.

10) What is JobTracker in Hadoop?

JobTracker is a service within Hadoop which runs MapReduce jobs on the cluster.

11) What is WebDAV in Hadoop?

WebDAV is a set of extension to HTTP which is used to support editing and uploading files. On most operating system WebDAV shares can be mounted as filesystems , so it is possible to access HDFS as a standard filesystem by exposing HDFS over WebDAV.

12) What is sqoop in Hadoop?

Sqoop is a tool used to transfer data between Relational Database Management System (RDBMS) and Hadoop HDFS. By using Sqoop, you can transfer data from RDBMS like MySQL or Oracle into HDFS as well as exporting data from HDFS file to RDBMS.

13) What are the functionalities of JobTracker?

These are the main tasks of JobTracker:

* To accept jobs from client.
* To communicate with the NameNode to determine the location of the data.
* To locate TaskTracker Nodes with available slots.
* To submit the work to the chosen TaskTracker node and monitors progress of each tasks.

14) Define TaskTracker.

TaskTracker is a node in the cluster that accepts tasks like MapReduce and Shuffle operations from a JobTracker.

15) What is Map/Reduce job in Hadoop?

Map/Reduce is programming paradigm which is used to allow massive scalability across the thousands of server.

Actually MapReduce refers two different and distinct tasks that Hadoop performs. In the first step maps jobs which takes the set of data and converts it into another set of data and in the second step, Reduce job. It takes the output from the map as input and compress those data tuples into smaller set of tuples.

16) What is "map" and what is "reducer" in Hadoop?

**Map:** In Hadoop, a map is a phase in HDFS query solving. A map reads data from an input location, and outputs a key value pair according to the input type.

**Reducer:** In Hadoop, a reducer collects the output generated by the mapper, processes it, and creates a final output of its own.

17) What is shuffling in MapReduce?

Shuffling is a process which is used to perform the sorting and transfer the map outputs to the reducer as input.

18) What is NameNode in Hadoop?

NameNode is a node, where Hadoop stores all the file location information in HDFS (Hadoop Distributed File System). We can say that NameNode is the centrepiece of an HDFS file system which is responsible for keeping record of all the files in the file system, and tracks the file data across the cluster or multiple machines.

19) What is heartbeat in HDFS?

Heartbeat is a signal which is used between a data node and name node, and between task tracker and job tracker. If the name node or job tracker doesn't respond to the signal then it is considered that there is some issue with data node or task tracker.

20) How indexing is done in HDFS?

There is a very unique way of indexing in hadoop. Once the data is stored as per the block size, the HDFS will keep on storing the last part of the data which specifies the location of the next part of the data.

21) What happens when a data node fails?

If a data node fails the job tracker and name node will detect the failure. After that all tasks are re-scheduled on the failed node and then name node will replicate the user data to another node.

22) What is Hadoop Streaming?

Hadoop streaming is a utility which allows you to create and run map/reduce job. It is a generic API that allows programs written in any languages to be used as Hadoop mapper.

23) What is a combiner in Hadoop?

A Combiner is a mini-reduce process which operates only on data generated by a Mapper. When Mapper emits the data, combiner receives it as input and sends the output to reducer.

24) What are the Hadoop's three configuration files?

Following are the three configuration files in Hadoop:

* core-site.xml
* mapred-site.xml
* hdfs-site.xml

25) What are the network requirements for using Hadoop?

Following are the network requirement for using Hadoop:

* Password-less SSH connection.
* Secure Shell (SSH) for launching server processes.

26) What do you know by storage and compute node?

**Storage node:** Storage node is the machine or computer where your file system resides to store the processing data.

**Compute Node:** Compute node is a machine or computer where your actual business logic will be executed.

27) Is it necessary to know java to learn Hadoop?

If you have a background in any programming language like C, C++, PHP, Python, Java etc. It may be really helpful, but if you are nil in java, it is necessary to learn Java and also get the basic knowledge of SQL.

28) How to debug Hadoop code?

There are many ways to debug Hadoop codes but the most popular methods are:

* By using Counters.
* By web interface provided by Hadoop framework.

29) Is it possible to provide multiple inputs to Hadoop? If yes, explain.

Yes, It is possible. The input format class provides methods to insert multiple directories as input to a Hadoop job.

30) What is the relation between job and task in Hadoop?

In Hadoop, A job is divided into multiple small parts known as task.

31) What is the difference between Input Split and HDFS Block?

Logical division of data is called Input Split and physical division of data is called HDFS Block.

32) What is the difference between RDBMS and Hadoop?

|  |  |
| --- | --- |
| **RDBMS** | **Hadoop** |
| RDBMS is relational database management system. | Hadoop is node based flat structure. |
| RDBMS is used for OLTP processing. | Hadoop is used for analytical and for big data processing. |
| In RDBMS, the database cluster uses the same data files stored in shared storage. | In Hadoop, the storage data can be stored independently in each processing node. |
| In RDBMS, preprocessing of data is required before storing it. | In Hadoop, you don't need to preprocess data before storing it. |

33) What is the difference between HDFS and NAS?

HDFS data blocks are distributed across local drives of all machines in a cluster whereas, NAS data is stored on dedicated hardware.

34) What is the difference between Hadoop and other data processing tools?

Hadoop facilitates you to increase or decrease the number of mappers without worrying about the volume of data to be processed.

35) What is distributed cache in Hadoop?

Distributed cache is a facility provided by MapReduce Framework. It is provided to cache files (text, archives etc.) at the time of execution of the job. The Framework copies the necessary files to the slave node before the execution of any task at that node.

36) What commands are used to see all jobs running in the Hadoop cluster and kill a job in LINUX?

Hadoop job - list

Hadoop job - kill jobID

37) What is the functionality of JobTracker in Hadoop? How many instances of a JobTracker run on Hadoop cluster?

JobTracker is a giant service which is used to submit and track MapReduce jobs in Hadoop. Only one JobTracker process runs on any Hadoop cluster. JobTracker runs it within its own JVM process.

Functionalities of JobTracker in Hadoop:

* When client application submits jobs to the JobTracker, the JobTracker talks to the NameNode to find the location of the data.
* It locates TaskTracker nodes with available slots for data.
* It assigns the work to the chosen TaskTracker nodes.
* The TaskTracker nodes are responsible to notify the JobTracker when a task fails and then JobTracker decides what to do then. It may resubmit the task on another node or it may mark that task to avoid.

38) How JobTracker assign tasks to the TaskTracker?

The TaskTracker periodically sends heartbeat messages to the JobTracker to assure that it is alive. This messages also inform the JobTracker about the number of available slots. This return message updates JobTracker to know about where to schedule task.

39) Is it necessary to write jobs for Hadoop in Java language?

No, There are many ways to deal with non-java codes. HadoopStreaming allows any shell command to be used as a map or reduce function.

40) Which data storage components are used by Hadoop?

HBase data storage component is used by Hadoop.