Hadoop Interview sample questions

If you had been wondering about how the interviews in Hadoop are held, then fear no more. Here we have given an array of questions and answers which has been framed by the best of skilled professionals from Intellipat who train the tricks for learning Hadoop Online. The entire write-up is going to give you the best idea about the correct answers to all the questions. Do let us know what you think about this. Happy job seeking to you all!

Best answers to the interview questions of Hadoop:

1. Give a clear comparison between Hadoop and Spark:

Answer 1: If we talk about storage facilities, then Hadoop has HDFS while Spark has none as such. On the other hand, the processing speed is mediocre in Hadoop while in case of Spark, it is excellent. Last but not the least, there are separate tools available in Hadoop in the Libraries section. While in case of Spark, they comprise of Spark Core, SQL, Graphx, Streaming, MLlib, etc.

2. What are the basic applications of Hadoop?

Answer 2: Hadoop, which is also rendered as Apache Hadoop, is one of the open source forums for the distribution of magnanimous amounts of data. It provides n enhanced performance as well as in-depth analysis of all the data (both structured and unstructured) that is generated on the digital platforms. Some of the works of Hadoop are being placed below:

- Traffic management in the roads
- The process of streaming
- Archive email management and web content management
- Processing of high functioning neuron signals by the use of Hadoop Computing cluster.
- The detection of frauds and the prevention of the same.
- Managing various posts and an array of media like photos and videos.
- Analysing the data of the customers
- Research in an array of fields like – defence, scientific research, military, etc.

3. Why is Hadoop different?

Answer 3: Hadoop is one of the best-distributed file systems which allows the storage of a huge amount of data on cloud technology. The data is stored in a distributed manner and that is the reason why the processing becomes easy as well. Specific nodes can process specific data in short periods of time. It is different because, in case of other computing methods, they are not as efficient in case of processing of the data.

4. What are the modes in which Hadoop can be run in?

Answer 4: Hadoop runs specifically in three basic modes:

- Standalone Mode: In this case, a local file system is used for the input and output operations and the main motto is to go for debugging.

- Single Node Cluster: Configuration of all the three files namely mapped-site.XML, core-site.xml and hdfs-site.xml.

- Fully distributed mode- In this case, the data is distributed in the cluster form and several nodes are allotted.

5. What is the difference between HDFS BLOCK and InputSplit?

Answer 5: Block is a physical representation of the data while split shows logical presentation. Also, the distribution can be done in the best possible manner with the help of Split.

6. Define Distributed Cache and give the advantages of the same

Answer 6: If a file is categorised into a specific job, then the distributed cache makes it easily accessible for distribution. It distributes simple data and tracks all the modifications so that the cache file can be read and used in the best possible manner.

7. What is the difference between Name Node, Checkpoint and Backup Node?

Answer 7: Name Node is the kind of data that consists of the files present in the HDFS system. Checkpoint Name node checks at regular intervals by downloading and analysing data at the local directory. The Backup Node is high functioning as a Checkpoint.

8. What are the common formats for Input in Hadoop?

The three common ones are namely- Text Input Format, Sequence File Input Format and Key Value Input Format.

9. Define DataNode and how does NameNode handle DataNode?

DataNode is the one that stores the data in HDFS. The work of NameNode is to replicate the data blocks from one node to the other.

10. Name the code methods that a Reducer Has:

There are three major steps- Setup, Reduce and Cleanup

11. What is the reason of using hadoop ? Or why we use hadoop?

Following are the four reasons because of which we use hadoop

- Hadoop allows us to run multiple data exploratory with full datasets .
- It is very difficult to find the large data set at low cost. Therefore we use hadoop as it offers a linear scalable storage capacity and high processing power. So that the large dataset can be stored using the RAW format which help us in building the accurate model.
- Pre processing of large scale Raw Data by using the tools like HIVE, PIG and by using other scripting languages like python.
- Redesign of schema is not difficult in hadoop as compare to the traditional RDBM system. Hadoop is also mentioned as “schema on read” that means it creates data agility.

12. Name the best hadoop technology companies of the world?

- IBM
- Cloudera Ask Bigger Questions
- MAPR Technologies
- Amazon Web Services
- DELL

13. When was Hadoop introduced?

The haddop was introduced on 10 December 2011.

14. Who is founder of hadoop ?

Hadoop was founded by Doug Cutting. At that time doug was working at yahoo and he named the project as hadoop after his son's toy elephant.

15. Is the Hadoop open source Platform?

Yes Hadoop is an open source platform which is basically build by using two technologies that is linux operating system and java programming language.

16. What are the limitations of hadoop for big data Analytics?

- Hadoop is designed for high capacity of large files therefore hadoop is not suitable for files that are small in size. Reading a small file is one of the major issues in HDFS.
- In hadoop data is distributed and then it is processed over the cluster in MapReduce which result in reducing the processing speed.

- The overall performance of hadoop is very low as it only supports the batch processing.
- Hadoop does not support the real time data processing.
- Hadoop does support the delta iteration as iterative processing is not so efficient in hadoop.
- It is very difficult to use hadoop because you need to code for every single operation.
- Security is the topic of concern while using hadoop as encryption in hadoop is missing at network and storage level.
- More chances of bugs because the code in hadoop is very lengthy. Because of that it takes a lot of time in solving the error and then execution of the code.