Hadoop Developer Interview Questions

-----------------------------------------

1) Explain how Hadoop is different from other parallel computing solutions.

2) What are the modes Hadoop can run in?

3) What will a Hadoop job do if developers try to run it with an output directory that is already present?

4) How can you debug your Hadoop code?

5) Did you ever built a production process in Hadoop? If yes, what was the process when your Hadoop job fails due to any reason? (Open Ended Question)

6) Give some examples of companies that are using Hadoop architecture extensively.

Hadoop Admin Interview Questions

-----------------------------------------

7) If you want to analyze 100TB of data, what is the best architecture for that?

8) Explain about the functioning of Master Slave architecture in Hadoop?

9) What is distributed cache and what are its benefits?

10) What are the points to consider when moving from an Oracle database to Hadoop clusters? How would you decide the correct size and number of nodes in a Hadoop cluster?

11) How do you benchmark your Hadoop Cluster with Hadoop tools?

Hadoop Interview Questions on HDFS

12) Explain the major difference between an HDFS block and an InputSplit.

13) Does HDFS make block boundaries between records?

14) What is streaming access?

15) What do you mean by “Heartbeat” in HDFS?

16) If there are 10 HDFS blocks to be copied from one machine to another. However, the other machine can copy only 7.5 blocks, is there a possibility for the blocks to be broken down during the time of replication?

17) What is Speculative execution in Hadoop?

18) What is WebDAV in Hadoop?

19) What is fault tolerance in HDFS?

20) How are HDFS blocks replicated?

21) Which command is used to do a file system check in HDFS?

22) Explain about the different types of “writes” in HDFS.

Hadoop MapReduce Interview Questions

-----------------------------------------

23) What is a NameNode and what is a DataNode?

24) What is Shuffling in MapReduce?

25) Why would a Hadoop developer develop a Map Reduce by disabling the reduce step?

26) What is the functionality of Task Tracker and Job Tracker in Hadoop? How many instances of a Task Tracker and Job Tracker can be run on a single Hadoop Cluster?

27) How does NameNode tackle DataNode failures?

28) What is InputFormat in Hadoop?

29) What is the purpose of RecordReader in Hadoop?

30) What is InputSplit in MapReduce?

31)In Hadoop, if custom partitioner is not defined then, how is data partitioned before it is sent to the reducer?

32) What is replication factor in Hadoop and what is default replication factor level Hadoop comes with?

33) What is SequenceFile in Hadoop and Explain its importance?

34) If you are the user of a MapReduce framework, then what are the configuration parameters you need to specify?

35) Explain about the different parameters of the mapper and reducer functions.

36) How can you set random number of mappers and reducers for a Hadoop job?

37) How many Daemon processes run on a Hadoop System?

38) What happens if the number of reducers is 0?

39) What is meant by Map-side and Reduce-side join in Hadoop?

40) How can the NameNode be restarted?

41) Hadoop attains parallelism by isolating the tasks across various nodes; it is possible for some of the slow nodes to rate-limit the rest of the program and slows down the program. What method Hadoop provides to combat this?

42) What is the significance of conf.setMapper class?

43) What are combiners and when are these used in a MapReduce job?

44) How does a DataNode know the location of the NameNode in Hadoop cluster?

45) How can you check whether the NameNode is working or not?

Pig Interview Questions

-----------------------------------------

46) When doing a join in Hadoop, you notice that one reducer is running for a very long time. How will address this problem in Pig?

47) Are there any problems which can only be solved by MapReduce and cannot be solved by PIG? In which kind of scenarios MR jobs will be more useful than PIG?

48) Give an example scenario on the usage of counters.

Hive Interview Questions

-----------------------------------------

49) Explain the difference between ORDER BY and SORT BY in Hive?

50) Differentiate between HiveQL and SQL.

More

---------------

1. Why can't we use Java primitive data types in Map Reduce?

2. Explain how do you decide between Managed & External tables in hive

3. Can we change the default location of Managed tables

4. What are the factors that we consider while creating a hive table

5. What are the compression techniques and how do you decide which one to use

6. Co group in Pig

7. How to include partitioned column in data - Hive

9. What hadoop -put command do exactly

10. What is the limit on Distributed cache size?

11. Handling skewed data

12. What are the Different joins in hive?

13. Explain about SMB join in Hive

1.What is biggest Hadoop cluster you have worked on?

2.Give us the details of your cluster – node configuration etc.?

3.List your involvement in modeling, ingestion, transformations, aggregation and data access layer.

4.What are the biggest challenges you have faced during implementation of Hadoop projects?

5.What kind of HDFS file formats have you used?

6.What was the selection criteria in deciding the file formats and explain the alternatives considered?

7.What is your experience with traditional data warehouse/marts development? Be specific on activities you were involved?

8.Why did your organization choose to use Hadoop?

9.Describe a complex problem you have solved in Hive, Pig or Java MapReduce?

10.What is your selection process in selecting a tool from the Hadoop ecosystem?

11.Describe one or two performance issues you have resolved in Hadoop?

12.Explain your troubleshooting process.