**INTERVIEW QUESTIONS**

|  |
| --- |
|  |
|  | **1. Hadoop? – open source / reliable/ scalable framework used for storage and analysis of data. It came out as solution to big data.** |
|  | **For storage we have HDFS and analysis we have map reduce.**  **2.What is big data? – large amount of data. It depends on organization.** |
|  | **3.Commands – u have a file of 50 GB how will you upload on HDFS.** |
|  | **By put command** |
|  | **Get – HDFS to local** |
|  | **Ls- show the contents of the directory** |
|  | **Pwd- print working directory** |
|  | **4. What is map reduce?** |
|  | **Intermediately stages – we do not have directly mapper and reducer.** |
|  | **5. How do u submit Hadoop job or job or monitor these jobson cluster?** |
|  | **Hadoop is batch processing map reduce is batch processing. One using Yarn and one using hadoop. In yarn we have 2 daemons – resource manager and node manager.** |
|  | **Yarn –jar pathOfJar main class with package name inputPath outputPath** |
|  | **To monitor the job – one application with job ID is available – we can click and check map reducer and all details** |
|  |  |
|  | **6. What is block? – size of memory in which file is divided. Default block size is 128 MB. Previous version it is 64 MB. Why is it 128 MB? – one of the most optimized no – for optimization. This info is provided from documentation.** |
|  | **7. Can you tell difference between block and split?** |
|  | **Command to check details of block and replication factor –** |
|  | **hdfs fsck /newdir/largedeck.txt –block** |
|  | **gives all info abt replication factor and file size** |
|  |  |
|  | **8. How do you monitor jobs in cluster?** |
|  | **9. How do we change block size?** |
|  | **We can by changing general configuration by hdfs-site.xml file** |
|  | **10. HDFS architecture** |
|  | **Master slave architecture. Master stores all metadata.** |
|  | **11. What is replication factor?** |
|  | **No of times each and every data block is replicated. Block is replicated to provide fault tolerance. The default replication factor is 3. We can change this using –setrep command** |
|  | **12. Is secondary node a backup of Namenode?** |
|  | **No. It is used to maintain metadata to name node. By merging edit log and fsimage. It is just a checkpoint. Not a replacement of namenode.** |
|  | **Cluster will switch to new name node if active namenode dies.** |
|  | **13. What is hive?** |
|  | **Hive is a Java based tool for interactive query. It is an open source framework to efficiently store and process large datasets. Ultimately its map reduce.** |
|  | **14. What is internal table and external table? And what are their differences?** |
|  | **Internal table is a directory structure. External table is just logical. Hive doesn’t create real table. When we drop an internal table, it drops the data, and it also drops the metadata. When we drop an external table, it only drops the meta data.** |
|  | **15. Where in the internal directory data is moved to? How can I change this default path?** |
|  | **16. What is diff between Hiive and My SQL?** |
|  | **Hive is used to maintain data warehouse but RDBMS is used to maintain database. In RDBMS normalized data is stored. Hive stores both – normalized and de-normalized. RDBMS does not support partitioning but Hive supports partitioning.** |
|  | **17. What is Sqoop?** |
|  | **Sqqop is one of the Hadoop ecosystem tool, it’s a java based tool used to import and export RDBMS to hdfs/hdfs to rdbms.** |
|  |  |
|  | **18. What is join in database ? - Retrieving data from multiple tables is joining** |
|  | **19. what are the type of join?** |
|  | **20. What is sub-query? subquery is also known as inner query. after main query we can have inner query** |
|  | **21. avg, min, max** |
|  | **22. What are the aggregate functions?** |
|  | **23. What is group by?** |
|  | **24. How do you organize SQL query?** |
|  | **25. What is the difference between have and having?** |
|  | **26. what is different between where clause and having clause ?** |
|  | **27. what are UDFs?** |
|  | **28. What is diff bet function and processor?** |
|  | **29. what is constraints ?** |
|  | **30. What are the data types you've worked on? diff formats of data?** |
|  | **31. what is Data types in hive ?** |
|  | **32. What is delete ,drop, and truncate?** |
|  | **33. what is foreign key?** |
|  | **34. how do you calculate second highest salary?** |
|  | **35. what is OOP** |
|  | **36. difference between class and object** |
|  | **37. difference between class and object** |
|  | **38. What are constructors – automatically gets executed on object creation.** |
|  | **39. Inheritance, abstraction, polymorphism,encapsulation** |
|  | **40. What are getters and setters- they are used to access private members of class.** |
|  | **41. Access specifiers.** |
|  | **42. Overloading and overriding.** |
|  | **43. Static keyword.** |
|  | **44. Difference between static and final.** |
|  | **45. Diff Between Tuple, list, Array and Dictionary -- In Python.**  **46. difference between namenode and datanode.**  **47. if datanode fails what happens to namenode.**  **48. difference between left join and right join.**  **49. highest second salary for employees.**  **50. what are namspaces in python.**  **51. scope of local variable and global variable.** |