MID - 1

BIG DATA ANALYTICS PREV Q&A

1. consider the wordcount mapreduce application we covered in class. Let:

. The main class be WordCount

. The jar file be wordcount.jar

. The input be stored in a directory called wc-in under the users home directory in HDFS.

. The output directory be wc-out, to be created under the users home directory in HDFS.

1) what is the YARN command to submit and run the job in Hadoop? Ans: yarn jar wordcount.jar WordCount wc-in wc-out

2) what is the HDFS command to show the whole output in the terminal? Ans: hdfs dfs -cat wc-out/*

2. The HDFS command to display the content of the root directory in HDFS is hdfs dfs -ls / space after -ls

3. HDFS is an efficient distributed filesystem for reading and writing large amount of data in parallel. False TRUE

4. The Quorum Journal Manager (QJM) is a dedicated HDFS implementation used in HDFS HA to provide a highly available shared edit log. It is the recommended choice for most HDFS installations.

5. The mapreduce application will use zero reducers if you skip setting the reducer class in the Driver. False

6. Setting the mapping and Reducer classes for mapreduce applications in the driver code is optional True False

7. The command to start YARN in the vm is start-yarn.sh OR yarn start
The command will start the resource manager and node manager daemons.

8. When we submit an application to YARN. The application will use the first allocated container to launch and run its Application Master process.

9. In the java code of the Drivers of mapreduce applications, we invoke the method job.waitForCompletion() on the job object to submit the job and wait for the job to finish. The name of the other method that can submit the job without blocking the Drivers is job.submit()

10.

11. In HDFS HA, graceful failover is a failover that is initiated manually by the administrators usually for routine maintenance True False

12. In Hadoop, the idea of running the computation where the data resides is called Data Locality

13. Fig. that shows 3 case a, c, 0, 2, 4

14. To use HDFS services we have to start the Name Node and the Data Node daemons on the master and worker machines respectively.

15. Among YARN available schedulers, The FIFO scheduler is the one recommended for shared clusters False

16. Data locality in mapreduce applies only to the mappers rather than the mappers and reducers True True

17. Hadoop distcp is much more efficient than HDFS cp command for copying data to and from Hadoop filesystems in parallel. True

18. The only programming language that can be used in MapReduce applications is Java. False

19. In mapreduce 1 (before YARN) we used containers to manage and allocate resources among jobs False

True

20. Hadoop is meant for Batch processing rather than real time processing. True

21. In contrast to mapreduce 1, YARN Resource manager will keep track of the available resources and the running jobs, while each application will track its own progress. True

22. memory becomes a limiting factor for scaling if we keep using a single namenode. HDFS Federation allows the cluster to scale by adding namenodes, each of which manages a portion of the filesystem namespace.

23. Before YARN was added to Hadoop, we had to start the JobTracker and the Task Tracker daemons to allow jobs to run in the cluster.

24.

25. In HDFS high availability (HA), in the event of failure of the active namenode the secondary namenode takes over its duties. False

26.

27. One of the common Hadoop processing workloads is ETL. ETL stands for Extract, Transform & Load.

28. The HDFS command to upload stocks.csv file, stored locally in the current working directory, to the users home directory in HDFS is hdfs dfs -put stocks.csv stocks.csv

hdfs dfs -put stocks.csv ~/

29. The default scheduling policy within each queue is fair under the fair scheduler, while it is FIFO under the capacity scheduler. True

30.

31. The 3 v's of big data are volume, variety and Velocity

32. HBase is a NoSQL distributed database, part of Hadoop Storage layer, that runs on top of HDFS. It was inspired by Google Big Table.

33. Although the default replication factor in HDFS is 3, You should change it to 1. If you run Hadoop in the pseudo distributed mode (the operation mode used in our VM).

34. The mappers and reducers write their output to HDFS False

35. In HDFS HA, a method known as fencing is necessary to prevent the previously active namenode from doing any damage or corruption to shared edit log.