MID 1

1. Once compiled a map reduce job can be submitted to Hadoop running in standalone, pseudo distributed, or fully distributed mode, without any changes or reconfiguration but necessary T or F ? Ans: **T**

2. Fencing is used in namemode federation to prevent namenodes from changing each other namespace metadata T or F ? Ans : **T**

3. In MapReduce 2 ,Progress tracking is handled by application, Masters , rather than the daemon , the resource manager . T or F ? Ans : **T**

4. FIFO scheduling is good for shared clusters T or F ? Ans : **F**

5. Hadoop is meant for batch processing rather than real time processing T or F ? Ans : **T**

6. To use HCFS services we have to start the…… **Name node**……and the ……**Datanode** ……..daemons on the master and worker machines respectively .

7. Among YARN available schedulers the FIFO scheduler is the one recommended for shared clusters . T or F ? Ans : **F**

8. Memory becomes a limiting factor for scaling if we keep using a single name mode. HDFS ……**Federation** ……. Allows the cluster to scale by adding name modes, each of which manages a portion of the file system name space.

9. The only programming language that can be used in MapReduce application is Java. T or F ? Ans : **F**

10. Data security in MapReduce applies only to the Mappers Rather than the mappers and Reducers . T or F ? Ans : **F**

11. In Contract 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. T or F ? Ans: **T**

12. In MapReduce !(before YARN) we used containers to manage and allocate resources among jobs ? T or F ? Ans : **T**

13. Hadoop distcp is much more efficient than HDFS cp command for copying data to and from Hadoop filesystems in parallel. T or F? Ans : **T**

14. In Hadoop the idea of running the computation where the data resides is called …**Data Locality**……

15. In HDFS HA graceful is a failover that is initiated by the administrator usually for routine maintenance ? T or F ? Ans: **F**

16. The Command to start YARN in the VM is …**start-yarn.sh**……..The command will start the Resource Manager and …**Node Manager**……..demons .

17. When we submit an application to YARN. The application will use the first allocated container to launch and run its…**Application Master Container**……..Process.

18. In java code of the driver of map Reduce application, 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 be submit the job without blocking the driver is ……**job.submit()……**

19. Deleting the Mapper and Reducer classes for the MapReduce application in the driver code if optional. T or F? Ans: **F (not sure)**

20. The MapReduce application will use Zero reduces. If you skip setting the reducer clases in the Driver. T or F ? Ans : **F**

21. The …**Quorum Journal Manager**…… 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.

22. HDFS is an efficient distributed filesystem for reading and writing large amount of data in parallel . T or F ? Ans : **T**

23. Consider the wordcount MapReduce application we covered in class let

- The Man class be WordCount
- The jar file be wordcount.jar
- The input be stored in a directory called wc-in under the user home directory in HDFS
- The output directory be wc-out. To be created under the user 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/***

24. The HDFS command to display the content of the root directory in HDFS is…… **hdfs dfs -ls /**……

25. Before YARN was added to Hadoop, we had to start the……**Job Tracker**……. And the Task Tracker daemons to allow jobs to run in the cluster.

26. In HDFS high availability (HA) , in the event of failure of the active namemode the secondary namemode takes over its duties. T or F ? Ans : **T**

27. One of the common Hadoop Processing workloads is ETL. ETL stands for… **Extract** ……, … **Transform** …….. and …… **Load** …… Respectively .

28. The HDFS command to upload stocks.csv file , stored locally in the current working directory , to the user home directory in HDFS is ……**hdfs dfs -Put stocks.csv ~/**……

Note : your command should work independent of the user name in Hadoop.

29. The default Scheduling policy within each queue is fair under the Fair Scheduler , While it is fifo under the capacity Scheduler . T or F ? Ans : **T**

30. The three Vs of Bigdata are Volume , Variety and … **velocity** …….

31. …… **Apache HBase** …….. is a NoSQL distributed database . Part of Hadoop storage Layer , That runs on top of HDFS. It was inspired by Google Bigdata table .

32. 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).

33. The Mappers and reducers write their Output to HDFS ? T or F ? Ans : **T**

34. In HDFS HA , a method Known as …… **Fencing** ……. Is necessary to prevent the previously active name mode from doing any damage or corruption to the shared edit log.