**Questions with Answers**

**1. As companies move past the experimental phase with Hadoop, many cite the need for additional capabilities, including:**   
  
d) Improved security, workload management and SQL support 

**2. Point out the correct statement:**   
  
b) Hadoop 2.0 allows live stream processing of real time data 

**3. According to analysts, for what can traditional IT systems provide a foundation when they’re integrated with big data technologies like Hadoop?**   
a) Big data management and data mining 

**4. Hadoop is a framework that works with a variety of related tools. Common cohorts include:**   
a) MapReduce, Hive and HBase 

**5. Point out the wrong statement:**   
  
c) The programming model, MapReduce, used by Hadoop is difficult to write and test 

**6. What was Hadoop named after?**   
  
c) The toy elephant of Cutting’s son 

**7. All of the following accurately describe Hadoop, EXCEPT:**   
  
b) Real-time 

**8. \_\_\_\_\_\_\_\_\_\_ can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data.**   
a) MapReduce 

**9. \_\_\_\_\_\_\_\_\_\_ has the world’s largest Hadoop cluster.**   
  
c) Facebook 

**10. Facebook Tackles Big Data With \_\_\_\_\_\_\_ based on Hadoop.**   
a) ‘Project Prism’ 

**11. Data locality feature in Hadoop means**

c) co-locate the data with the computing nodes.

**12. The main goal of HDFS High availability is**

b) To reduce the cycle time required to bring back a new primary name node after existing primary fails.

**13. When running on a pseudo distributed mode the replication factor is set to**

b) 1

**14. Which of the below property gets configured on hadoop-env.sh?**

d) Java Environment variables.

**15. The Hadoop tool used for uniformly spreading the data across the data nodes is named −**

b) Balancer

**16. Rack awareness in name node means**

b) It is aware of the mapping between the node and the rack

**17. The archive file created in Hadoop always has the extension of**

b) .har

**18. You can reserve the amount of disk usage in a data node by configuring the dfs.datanode.du.reserved in which of the following file**

a) Hdfs-site.xml

**19. What are the daemons that are required to start the HDFS?**

d) Data Node, Node Manager, Secondary Name Node

**20. XML file consists of**

d) All the above

**21. Identify the utility that allows you to create and run MapReduce jobs with any executable or script as the mapper and/or the reducer?**

d) Hadoop Streaming

**22.How are we overcoming Name node problem of Single point of failure in Gen2?**

d) All the above

**23.Which of the following is responsible for scheduling the job's component tasks on the slaves, monitoring them and re-executing the failed tasks?**

a) Slave

**24.How will the name node decide that which data node the data has to be written. Assume the replication factor is 3.**

c) It depends on the load on the data nodes

**25.Assume that there are 50 nodes in your Hadoop cluster with a total of 200 TB (4 TB per node) of raw disk space allocated HDFS storage. Assuming Hadoop's default configuration, how much data will you be able to store?**

d) Approximately 66 TB

**26.You need to move a file titled weblogs into HDFS. When you try to copy the file, you can’t. You know you have ample space on your Data Nodes. Which action should you take to relieve this situation and store more files in HDFS?**

c) Decrease the block size on your remaining files.

**27. You use Hadoop fs -put command to write a 300 MB file and HDFS block size of 64 MB. Just after this command has finished writing 200 MB of this file, what would another user see when trying to access this file?**

c) They would see the current state of the file through the last completed block

**28. You use the Hadoop fs -put command to add sales.txt to HDFS. This file is small enough that it fits into a single block, which is replicated to three nodes within your cluster. When and how will the cluster handle replication be following the failure of one of these nodes?**

d) The file will be re-replicated automatically after the Name Node determines it is under-replicated based on the block reports it receives from the Data Nodes.

**29. Which command does Hadoop offer to discover missing or corrupt HDFS data?**

a) fsck

**30.What metadata is stored on a Data Node when a block is written to it?**

b) Checksums for the data in the block, as a separate file.

**31.Identify which of the following cluster information will not be stored on disk on the Name Node?**

a) Names of the files in HDFS.

**32. Which describes how a client reads a file from HDFS?**

a) The client queries the NameNode for the block location(s). The NameNode returns the block location(s) to the client. The client reads the data directory off the DataNode(s).

**33. What is not part of the basic Hadoop Stack 'Zoo'?**

b) Horse

**34. What is part of the Apache Basic Hadoop Modules?**

a) HDFS

**35. What are the two major components of the MapReduce layer?**

c) NameNode

d) DataNode

**35. What does HDFS stand for?**

b) Hadoop Distributed File System

**36. What are the two majority types of nodes in HDFS?**

b) DataNode

d)  NameNode

**37. Could you run an existing MapReduce application using Yarn?**

b) Yes

**38. What are the two basic layers comprising the Hadoop Architecture?**

c) MapReduce and HDFS

**ASSIGNMENT**

**Let's assume that, you have 100 TB of data to store and process with Hadoop. The configuration of each available DataNode is as follows:**

* **8 GB RAM**
* **10 TB HDD**
* **100 MB/s read-write speed**

**You have a Hadoop Cluster with replication factor = 3 and block size = 64 MB.**

**In this case, the number of Data Nodes required to store would be:**

* **Total amount of Data \* Replication Factor / Disk Space available on each DataNode**
* **100 \* 3 / 10**
* **30 Data Nodes**

**Now, let's assume you need to process this 100 TB of data using MapReduce.**

**And, reading 100 TB data at a speed of 100 MB/s using only 1 node would take:**

* **Total data / Read-write speed**
* **100 \* 1024 \* 1024 / 100**
* **291.27 hours**
* **1048576 seconds**



**So, with 30 Data Nodes you would be able to finish this MapReduce job in:**

* **291.27 / 30**
* **9.70 hours**

**Problem Statement**

**How many such Data Nodes you would need to read 100TB data in 5 minutes in your**

**Hadoop Cluster?**

---1data node-( 100tb\*1024\*1024)/100=1048576 sec ~ 291.78 hrs.

DataNodes required to read the data in FIVE minutes ->

= (1048576 seconds/60)/5 minutes = 3495.253333 Data Nodes

SO, around 3495 data nodes required to read 100TB data in 5 mins.