**Questions**

**1. As companies move past the experimental phase with Hadoop, many cite the need for additional capabilities, including:**  
a) Improved data storage and information retrieval  
b) Improved extract, transform and load features for data integration  
c) Improved data warehousing functionality  
d) Improved security, workload management and SQL support

**2. Point out the correct statement:**  
a) Hadoop do need specialized hardware to process the data  
b) Hadoop 2.0 allows live stream processing of real time data  
c) In Hadoop programming framework output files are divided in to lines or records  
d) None of the mentioned

**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  
b) Data warehousing and business intelligence  
c) Management of Hadoop clusters  
d) Collecting and storing unstructured data

**4. Hadoop is a framework that works with a variety of related tools. Common cohorts include:**  
a) MapReduce, Hive and HBase  
b) MapReduce, MySQL and Google Apps  
c) MapReduce, Hummer and Iguana  
d) MapReduce, Heron and Trumpet.

**5. Point out the wrong statement:**  
a) Hardtop’s processing capabilities are huge and its real advantage lies in the ability to process terabytes & petabytes of data  
b) Hadoop uses a programming model called “MapReduce”, all the programs should confirm to this model in order to work on Hadoop platform  
c) The programming model, MapReduce, used by Hadoop is difficult to write and test  
d) All of the mentioned

**6. What was Hadoop named after?**  
a) Creator Doug Cutting’s favorite circus act  
b) Cutting’s high school rock band  
c) The toy elephant of Cutting’s son  
d) A sound Cutting’s laptop made during Hadoop’s development

**7. All of the following accurately describe Hadoop, EXCEPT:**  
a) Open source  
b) Real-time  
c) Java-based  
d) Distributed computing approach

**8. \_\_\_\_\_\_\_\_\_\_ can best be described as a programming model used to develop Hadoop-based applications that can process massive amounts of data.**  
a) MapReduce  
b) Mahout  
c) Oozie  
d) All of the mentioned

**9. \_\_\_\_\_\_\_\_\_\_ has the world’s largest Hadoop cluster.**  
a) Apple  
b) Datamatics  
c) Facebook  
d) None of the mentioned

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

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

[a) store the same data across multiple nodes.](javascript:void(0);)

[b) relocate the data from one node to another.](javascript:void(0);)

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

[d) Distribute the data across multiple nodes.](javascript:void(0);)

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

[a) Faster creation of the replicas of primary name node.](javascript:void(0);)

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

[c)Prevent data loss due to failure of primary name node.](javascript:void(0);)

[d) Prevent the primary name node form becoming single point of failure.](javascript:void(0);)

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

[a) 2](javascript:void(0);)

b) 1

[c) 0](javascript:void(0);)

[d) 3](javascript:void(0);)

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

[a) Replication factor](javascript:void(0);)

[b) Directory names to store hdfs files](javascript:void(0);)

[c)Host and port where MapReduce task runs](javascript:void(0);)

d) Java Environment variables.

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

[a) Scheduler](javascript:void(0);)

b) Balancer

[c)Spreader](javascript:void(0);)

[d) Reporter](javascript:void(0);)

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

[a) It is aware how many racks are available in the cluster](javascript:void(0);)

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

[c) It is aware of the number of nodes in each of the rack](javascript:void(0);)

[d) It is aware which data nodes are unavailable in the cluster.](javascript:void(0);)

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

[a) .hrc](javascript:void(0);)

b) .har

[c) .hrh](javascript:void(0);)

[d) .hrar](javascript:void(0);)

**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

[b) Hdfs-defaukt.xml](javascript:void(0);)

[c) Core-site.xml](javascript:void(0);)

[d)Mapred-site.xml](javascript:void(0);)

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

a) Resource Manager, Name Node

b) Name Node, Secondary Name Node, Data Node

c) Data Node, Node Manager

d) Data Node, Node Manager, Secondary Name Node

**20. XML file consists of**

a) Structured data

b) Unstructured Data

c) Semi structured Data

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?**

a) Oozie

b) Flume

c) Sqoop

d) Hadoop Streaming

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

a) Active Name node

b) StandBy Name node

c) Secondary Name node

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

b) Master

c) Data node

d) Name node and Data node

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

a) It chooses randomly

b) It chooses the data nodes which are nearby in that cluster.

c) It depends on the load on the data nodes

d) Both 2 and 3

**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?**

a) Approximately 200TB

b) Approximately 100TB

c) Approximately 50TB

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?**

a) Increase the block size on all current files in HDFS.

b) Increase the block size on your remaining files

c) Decrease the block size on your remaining files.

d) Increase the amount of memory for the Name Node.

**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?**

a) They would see Hadoop throw an ConcurrentFileAccessException when they try to access this file.

b) They would see the current state of the file, up to the last bit written by the command.

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

d) They would see no content until the whole file written and closed.

#### **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?**

a) The cluster will make no attempt to re-replicate this block.

b) This block will be immediately re-replicated and all other HDFS operations on the cluster will halt while this is in progress.

c) The block will remain under-replicated until the administrator manually deletes and recreates the file.

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

b) du

c) dskchk

d) Hadoop does not provide any tools to discover missing or corrupt data; there is no need because three replicas are kept for each data block

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

a) None. Only the block itself is written.

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

c) Information on the files location in HDFS.

d) Node location of each block belonging to the same namespace.

#### **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.

b) The directory structure of the files in HDFS.

c) An edit log of changes that have been made since the last snapshot compaction by the Secondary Name Node

d) A catalog of Data Nodes and the blocks that are stored on them.

#### **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).

b) The client queries all Data Nodes in parallel. The DataNode that contains the requested data responds directly to the client. The client reads the data directly off the DataNode.

c) The client contacts the NameNode for the block location(s). The NameNode then queries the Data Nodes for block locations. The Data Nodes respond to the NameNode, and the NameNode redirects the client to the DataNode that holds the requested data block(s).

d) The client contacts the NameNode for the block location(s). The NameNode contacts the DataNode that holds the requested data block. Data is transferred from the DataNode to the NameNode, and then from the NameNode to the client.

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

a) Pig

b) Horse

c) Elephant

d) Hive

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

a) HDFS

b) Yarn

c) MapReduce

d) Impala

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

a) Task Manager

b) Job Tracker

c) NameNode

d) DataNode

#### **35. What does HDFS stand for?**

a) Hadoop Data File System

b) Hadoop Distributed File System

c) Hadoop Data File Scalability

d) Hadoop Data Node File Security

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

a) Rack Node

b) DataNode

c) Block Node

d) NameNode

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

a) No

b) Yes

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

a) Zoo Keeper and MapReduce

b) HDFS and Hive

c) MapReduce and HDFS

d) Impala 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?**

**Solution:**

**Time 1 Data Node takes = Total Data/Read –Write speed**

**Here to read 100TB Data time required , (100TB \* 1024\*1024) in MB / 100MB/s = 1048576 seconds or 291.77 hours.**

**Required time for reading 100TB data using 1 DataNode = 1048576 seconds or 291.77 hours.**

**No. of DataNodes required to read 100TB in 5 minutes**

**Time taken by 1 DataNode to read the 100TB data / Total time given to finish the read**

**= (1048576 seconds/60)/5 minutes**

**= 3495.253333 Data Nodes**

**DataNodes required to read the data in FIVE minutes = 3495.253333 Data Nodes.**

**So, we would need ~ 3495 such Data Nodes to read the 100TB data in 5 minutes in our Hadoop cluster.**