Big Data and Hadoop – 01

1. What is the default replication factor of Hadoop cluster?

a. 3

b. 2

c. 4

d. 1

Ans- a.3

1. Which component in Hadoop Cluster is responsible for serving read and write requests from the file system’s clients?

a. Name Node

b. Data Node

c. Broad a & b

d. None of the above

Ans- Data Node

1. Which component of Hadoop cluster manages the file system namespace and regulates access to files by clients?

a. Name Node

b. Data Node

c. Both a & b

d. None of the above

Ans- Data Node

1. If a file of size 100 MB is stored on HDFS, what would be the split size?

a. 64 MB & 64 MB

b. 64 MB and 36 MB

c. 100 MB

d. None of the above

Ans- d. None of the above

1. State true or false: MR2 support various MPP modes for data processing?
2. False
3. True

Ans-

1. Which command of HDFS helps copy files from HDFS to Local file system?
2. copyFromLocal
3. copyToLocal
4. put
5. mv

Ans- copyToLocal

7. Which Eco system component of Hadoop is good for non sql programmers?

a. Hive

b. Hbase

c. Flume

d. Pig

Ans- Pig

1. Block size of a Hadoop cluster is configurable by Administrator?

a. True

b. False

Ans- True

1. The functions performed by DataNodes in Hadoop Cluster is/are?

a. Data Block Creation

b. Data Block Deletion

c. Data Block Replication

d. All above

Ans- d.All of the above

1. Find error in below command:

hdfs dfs –put /home/user1/abc.txt

a. Target name missing

b. Source name should include hdfs://

c. No error

Ans- a.Target name missing

1. Hadoop block size should be multiple of which unit?

a. 32 MB

b. 50 MB

c. 64 MB

d. 70 MB

Ans-

1. Which component of the hadoop cluster manages data on slave nodes?

a. Name node

b. Data node

c. Task Tracker

d. Job Tracker

Ans- Name Node

1. MR1 and MR2 are two modes of processing in Hadoop?

a. True

b. False

Ans- a.True

1. What is Hadoop?

a. Open source software for reliable, scalable, distributed computing.

b. A framework that allows for the distributed processing of large data sets across clusters of

computers using simple programming models.

c. Both a & b

d. None of the above

Ans- c. Both a & b

1. Hadoop provides
2. A reliable distributed storage and processing system
3. Only distributed storage
4. Only processing system
5. None of the above

Ans- a. A reliable distributed storage and processing system

15.What is default replication factor and how will you change it at file level?

The default replication factor is 3. To change the replication at file level:

1. [anu@localhost ~]$ hadoop fs –setrep –w 3 /my/file

16.Why do we need replication >1 in production Hadoop cluster?

The default replication factor is 3. In a production cluster if we plan a maintenance activity on one node or suddenly a node stops working we still have a node available and that makes Hadoop fault tolerant. HDFS stores the data in such a way that if one datanode fails, the data is available in another node and if the entire rack fails, copies are available in different nodes of different rack. This makes Hadoop cluster fault tolerant and ensures high availability of data.

17. How will you combine the 4 part-r files of a mapreduce job?

hadoop fs -getmerge /output/dir/on/hdfs/ /desired/local/output/file.txt

18. What are the compression techniques in HDFS and which is the best one and why?

The four most widely used compression formats in Hadoop are as follows:

**1) GZIP**

Provides High compression ratio.

Uses high CPU resources to compress and decompress data.

Good choice for Cold data which is infrequently accessed.

Compressed data is not splittable and hence not suitable for MapReduce jobs.

**2) BZIP2**

Provides High compression ratio (even higher than GZIP).

Takes long time to compress and decompress data.

Good choice for Cold data which is infrequently accessed.

Compressed data is splittable.

Even though the compressed data is splittable, it is generally not suited for MR jobs because of high compression/decompression time.

**3) LZO**

Provides Low compression ratio.

Very fast in compressing and decompressing data.

Compressed data is splittable if an appropriate indexing algorithm is used.

Best suited for MR jobs

**4) SNAPPY**

Provides average compression ratio.

Aimed at very fast compression and decompression time.

Compressed data is not splittable if used with normal file like .txt

Generally used to compress Container file formats like Avro and SequenceFile because the files inside a Compressed Container file can be split.

19. How will you view the compressed files via HDFS command?

20. What is Secondary Namenode and its functionalities? Why do we need it?

The main function of the Secondary namenode is to store the latest copy of the FsImage and the Edits Log files. Secondary namenode keeps the checkpoint on the namenode, It reads the edit logs from the namenode continuously after a specific interval and applies it to the fsimage copy of secondary namenode. In this way the fsimage file will have the most recent state of HDFS.  
The secondary namenode copies new fsimage to primary, so fsimage is updated.

Since fsimage is updated, there will be no overhead of copying of edit logs at the moment of restarting the cluster. Secondary namenode is a helper node and can’t replace the namenode.

21. What is backup node and how is it different from secondary namenode?

Backup Node in hadoop is an extended checkpoint node that performs checkpointing and also supports online streaming of file system edits. It’s main memory is always in sync with primary namenode file system namespace. Since it maintains an in-memory, up-to-date copy of file system namespace and accepts a real time online stream of file system edits and applies these edits on its own copy of namespace in its main memory. Thus, at any point of time, it maintains a latest backup of current file system namespace.

The Backup node does not need to download fsimage and edits files from the active NameNode to create a checkpoint, as it already has an up-to-date state of the namespace in it’s own main memory. As the Backup node keeps a copy of the namespace in main memory similar to NameNode, its hardware specifications should be same as the NameNode.

Unlike Checkpoint nodes, there is only one Backup node is allowed to be registered with namenode at any time but multiple checkpoint nodes registration is possible. if a Backup node is in use, then there might not be need for checkpoint nodes and these may not be required to register with namenode. Backup Node in hadoop can be started with below command on the dedicated node configured in the cluster.

**$ hdfs namenode –backup**

22.What is FSimage and editlogs and how are they related?

FsImage is a file stored on the OS filesystem that contains the complete directory structure of the HDFS with details about the location of the data on the Data Blocks and which blocks are stored on which node. This file is used by the NameNode when it is started.

EditLogs is a transaction log that records the changes in the HDFS file system or any action performed on the HDFS cluster such as addition of a new block, replication, deletion etc. In short, it records the changes since the last FsImage was created.

Every time the NameNode restarts, EditLogs are applied to FsImage to get the latest snapshot of the file system. But NameNode restarts are rare in production clusters.

23.What is default block size in HDFS and why is it so large?

The default size of a block in HDFS is 128 MB (Hadoop 2. x). The reason of having this huge block size is to minimize the cost of seek and reduce the meta data information generated per block.

24.How will you copy a large file of 50GB into HDFS in parallel?

We can do it using put command.

**hdfs dfs –put <localpath> <destpath>**

25.What is Balancing in HDFS?

HDFS provides a balancer utility. This utility analyzes block placement and balances data across the DataNodes. It keeps on moving blocks until the cluster is deemed to be balanced, which means that the utilization of every DataNode is uniform.

26.What is expunge in HDFS?

This command is used to empty the trash in HDFS system.