

**Test on Hadoop and Ecosystems**

**BIGDATA & HADOOP**

1. Which one of the following is true about Bigdata.
   1. Bigdata is a methodology that deals with huge volume of relational and structured data.
   2. Bigdata is a software framework that defines the aquire, cure, process and store of huge volume of variety of data that comes in different velocity.
   3. Bigdata is a phrase or methodology that defines the aquire, cure, process and store of huge volume of variety of data that comes in different velocity.
   4. Using bigdata solution we can handle data volume in a maximum of Exa bytes.
   5. Bigdata does not support veracity of data.
2. Pick the right size hierarchy arranged in chronological order

*(Acronyms : MB – Megabyte, GB –Gigabyte, ZB – Zetabyte, YB – Yotabyte, TB – Terabyte, EB – Exabyte, PB - Petabyte).*

* 1. MB 🡪 GB 🡪 TB 🡪 EB 🡪 PB 🡪ZB 🡪 YB
  2. MB 🡪 GB 🡪 TB 🡪 PB 🡪 ZB 🡪EB 🡪 YB
  3. MB 🡪 GB 🡪 TB 🡪 PB 🡪 EB 🡪YB 🡪 ZB
  4. MB 🡪 GB 🡪 TB 🡪 PB 🡪 EB 🡪ZB 🡪 YB

1. Which of the following is correct about HADOOP (choose 3 correct answers)?
   1. Hadoop is a open source Apache software framework/tool that provides distributed storage and processing of large data sets.
   2. Hadoop is a open source Apache software framework/tool that runs on cluster of commodity hardware that provides distributed storage and processing of large data sets.
   3. Features – High Throughput, handles huge volume of data, easily scalable, data locality, fault tolerant, cost effective, highly reliable and flexible.
   4. Features – High Throughput, Low latency, handles huge volume of data, easily scalable, data locality, fault tolerant, cost effective, highly reliable and flexible.
   5. Good for – Batch processing, Write many read many, High volume of small files.
   6. Good for – Batch processing, Write once read many, Low volume of huge files.
2. Which describes how a client reads a file from HDFS?
   1. The client queries all DataNodes in parallel. The DataNode that contains the requested data responds directly to the client. The client reads the data directly off the DataNode.
   2. The client contacts the NameNode for the block location(s). The NameNode then queries the DataNodes for block locations. The DataNodes respond to the NameNode, and the NameNode redirects the client to the DataNode that holds the requested data block(s). The client then reads the data directly off the DataNode.
   3. 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).
   4. 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.
3. In a HDFS system with block size 64MB we store a file which is less than 64MB. Which of the following is true?
   1. [The file will consume 64MB](javascript:void(0);)
   2. [The file will consume more than 64MB](javascript:void(0);)
   3. [The file will consume less than 64MB.](javascript:void(0);)
   4. [Cannot be predicted.](javascript:void(0);)
4. Which of the below property gets configured on hdfs-site.xml choose all that applies
   1. [Replication factor](javascript:void(0);)
   2. [Directory names to store hdfs files.](javascript:void(0);)
   3. [Host and port where MapReduce task runs.](javascript:void(0);)
   4. [Java Environment variables.](javascript:void(0);)
5. Under Hadoop High Availability, Fencing means
   1. [Preventing a previously active namenode from start running again.](javascript:void(0);)
   2. [Preventing the start of a failover in the event of network failure with the active namenode.](javascript:void(0);)
   3. [Preventing the power down to the previously active namenode.](javascript:void(0);)
   4. [Preventing a  previously active namenode from writing to the edit log.](javascript:void(0);)
6. Which of this is not a scheduler options available with YARN? (Choose all that applies)
   1. [Optimal Scheduler](javascript:void(0);)
   2. [FIFO scheduler](javascript:void(0);)
   3. [Capacity scheduler](javascript:void(0);)
   4. [Fair scheduler](javascript:void(0);)
   5. Queue scheduler
7. When a name node is restarted gracefully
   1. Namenode will be in safemode until checkpoint completes.
   2. [Namenode](javascript:void(0);) will not respond any client request.
   3. After a configurable percentage of safely replicated data blocks checks in with the NameNode, the NameNode exits the Safemode state.
   4. Fsimage and editlog will be copied from secondary name node by the administrator for checkpointing and then the Namenode will be started.
   5. All the above
   6. Options a,b and c.
8. Which one of the following is true about secondary name node SNN.
   1. SNN acts as backup node for storing fsimage and editlog of up to date changes happened in the cluster
   2. SNN acts as backup node for storing fsimage only of up to date changes happened in the cluster
   3. SNN acts as standby name node for storing fsimage of the recent changes happened in the cluster for about maximum an hour by default.
   4. SNN acts as backup node for storing fsimage of the recent changes happened in the cluster for about maximum an hour by default.
   5. SNN acts as a failover node to provide high availability.
   6. All the above.
9. Which of the following is true about block replication (choose all that applies).
   1. Default replication factor is 5.
   2. Replication factor can be defined in dfs.replication in hdfs-site.xml.
   3. Replication factor can be defined in dfs.replication in core-site.xml
   4. Replication factor can be defined at file copy time using “hadoop fs –setrep’ command.
   5. Place the first replica in the local node if HDFS Client is internal in the cluster, place the second replica in a different rack, place the third replica in a different rack other than first and second replica placed.
   6. Place the first replica in the local node if HDFS Client is internal in the cluster, place the second replica in a different rack, place the third replica in the same rack as the second replica. Replicas more than 3 will be placed randomly anywhere in the cluster.
   7. Same block of different replica will not be placed in the same node.
10. How to create zero byte file directly in HDFS.
    1. hadoop fs –vi zerobytefile.txt (open in vi mode and save the file using :wq)
    2. hadoop fs –touch /hdfspath/zerobytefile.txt
    3. hadoop fs –touchz /hdfspath/zerobytefile.txt
    4. hadoop fs –cat /linuxpath/zerobytefile.txt /hdfspath/zerobytefile.txt
    5. hadoop fs –cp /linuxpath/zerobytefile.txt /hdfspath/zerobytefile.txt
11. Where data block metadata informations are stored.
    1. FSImage in disk
    2. Editlog
    3. Namenode’s FSImage in memory.
    4. Secondary namenodes memory.
    5. Namenodes disk
    6. FSImage and editlog
12. Which of the following is true about HDFS (Choose all that applies)
    1. High throughput.
    2. Low latency
    3. Write once read many
    4. Write many read many
    5. Random read and Sequencial write.
    6. Random read and random write.
    7. Good for huge volume of small files.
13. Which of the following is true about HDFS High availability (Choose all that applies)
    1. Standby name node act as active name node in case of name node failure.
    2. Zookeeper failover controller monitors the health and manages the high availability state.
    3. Active and standby nodes communicates directly to stream edit logs.
    4. Active and standby nodes communicates through journal nodes to stream edit logs.
    5. Data Nodes are configured with the location of both the Name Nodes and send block location information and heartbeats to both Name Node machines.
    6. Standby name node act as secondary name node for checkpointing.
    7. Data Nodes are configured with the location of both the Name Nodes and send block location information and heartbeats to only Name Node machine and not to the standby name node.

**Hadoop Processing**

1. Your cluster’s HDFS block size is 64MB, you have directory containing 100 plain text files, each of which is 100MB in size. Determine how many mappers will run?
   1. 64
   2. 100
   3. 200
   4. 640
   5. 0
2. The input split used in MapReduce indicates
   1. [The average size of the data blocks used as input for the program](javascript:void(0);)
   2. [The location details of where the first whole record in a block begins and the last whole record in the block ends.](javascript:void(0);)
   3. [Splitting the input data to a MapReduce program into a size already configured in the mapred-site.xml](javascript:void(0);)
   4. [None of these](javascript:void(0);)
3. MapReduce v2 (MRv2/YARN) is designed to address which the three issues?
   1. Single point of failure in the NameNode.
   2. Reduce pressure on the JobTracker.
   3. HDFS latency.
   4. Ability to run frameworks other than MapReduce.
   5. Fixed number of map and reduce slots.
4. Identify the MapReduce v2 (MRv2 / YARN) daemon responsible for launching application containers and monitoring application CPU and RAM resource usage?
   1. ResourceManager
   2. NodeManager
   3. ApplicationMaster
   4. ApplicationMasterService
   5. TaskTracker
   6. JobTracker
5. [Which](http://www.fromdev.com/2010/12/interview-questions-hadoop-mapreduce.html#When-reducer-started) of the following is true about YARN (choose all that applies)
   1. YARN means Yet another resource navigator.
   2. YARN is designed to address supporting multiple frameworks including map reduce and act as a data operating system.
   3. YARN Resource manager is responsible for scheduling, monitoring and managing the resources.
   4. YARN Application master is responsible for negotiating resources with the Resource-Manager and for working with the Node-Managers to start the containers.
   5. YARN Resource manager is responsible for scheduling and managing the applications.
6. Which of the following are true about compression techniques (choose all that applies)
   1. Gzip compression uses deflate algorithm.
   2. Bzip is splittable only with index.
   3. Bzip2 is splittable.
   4. Snappy is splittable with index.
   5. LZO is not splittable.
   6. LZO is splitable with index.
7. When is the earliest point at which the reduce method of a given Reducer can be called?
   1. As soon as atleast any one of the mapper has finished processing its input split.
   2. As soon as a mapper has emitted at least one record.
   3. Not until all mappers finished and have finished processing all records.
   4. After the very first mapper finished processing all records.
8. Which of the following are true about failure handling in YARN (Choose all that applies)
   1. When a container fails, node-manager detects the failure event and launches a new container to replace the failing container and restart the task execution in the new container.
   2. When a container fails, application master detects the failure event and launches a new container to replace the failing container and restart the task execution in the new container.
   3. When a application master fails, In this Resource manager will start a new Application Master under node manager.
   4. When a application master fails, In this Node manager will start a new Application Master.
   5. Resource manager failure is single point of failure, but can be addressed using resource manager high availability.

**SQOOP**

1. Which of the following is true about sqoop import (choose all that applies)
2. [default](javascript:void(0);) number of mapper is 3
3. We can define number of mappers with - -splitby columns or m 1 or if we have primary key in the table.
4. Sqoop does not support composite key columns.
5. Always uses JDBC driver to extract the table metadata even when use direct mode for import.
6. Split by queries can be used instead of boundary queries to improve the performance.
7. [All of the above](javascript:void(0);).
8. In the import involving join of two tables if there are two columns with matching name between two tables then this conflict can be resolved by
   1. [Using table aliases](javascript:void(0);)
   2. [Column aliases](javascript:void(0);)
   3. [First creating temporary tables form each table with different column names](javascript:void(0);)
   4. [Rename the columns in the source system and then import](javascript:void(0);)
9. To import data to Hadoop cluster from relational database sqoop create a mapreduce job. In this job
10. [All the data is transferred in one go.](javascript:void(0);)
11. [Each mapper transfers a slice of Table's data](javascript:void(0);)
12. [Each mapper transfers tables' data along with table's metadata (name of the columns etc)](javascript:void(0);)
13. [Only the schema of relational table is validated without fetching data](javascript:void(0);)
14. What is achieved by the command – sqoop job –exec myjob
15. [Sqoop job named myjob is saved to sqoop metastore](javascript:void(0);)
16. [Sqoop job named myjob starts running](javascript:void(0);)
17. [Sqoop job named myjob is scheduled](javascript:void(0);)
18. [Sqoop job named myjob gets created](javascript:void(0);)
19. If the table to which data is being exported has more columns than the data present in the hdfs file then
20. [The load definitely fails](javascript:void(0);)
21. [The load will be done only for the relevant columns present in HDFS file](javascript:void(0);) if - -columns option is not used.
22. [The load will be done only for the relevant columns present in HDFS file](javascript:void(0);) if - -columns option is used.
23. [The data will populate values into the wrong columns](javascript:void(0);)
24. [The load does not start](javascript:void(0);)
25. When does sqoop gather the metadata of the relational table into which it exports/imports the data?
26. [Gathers the metadata of all tables only once during establishing the connection to the database](javascript:void(0);)
27. [Never as it relies on the user to ensure the exported data matches the table’s structure](javascript:void(0);)
28. Gathers the metadata of a particular table, every time the sqoop export command it submitted and just before the data transfer starts
29. [Only if the export fails, Sqoop accesses the metadata of the table](javascript:void(0);)
30. Which of the following is a disadvantage of using the –staging-table parameter?
31. [Data is stored twice and consumes more storage](javascript:void(0);)
32. [The overall export time is more than direct export to final table](javascript:void(0);)
33. [User should ensure the structure of staging table and final tables are in Sync.](javascript:void(0);)
34. [All of the above](javascript:void(0);)
35. Which of the following is true about sqoop export (choose all that applies)
36. Sqoop performs export row by row if we don’t leverage batch mode option.
37. Enabling batch mode will export more than one row at a time as batch of rows.
38. Sqoop export does not generates a Java record container class based on the target table definition.
39. Increasing the number of mappers in sqoop will degrade the data transfer rate (throughput).
40. Increasing the number of mappers in sqoop will improve the data transfer rate (throughput) and degrade the source DB performance due to increased parallel sessions.

**Flume**

1. Definition of Flume says ‘Flume collects data efficiently, aggregates and moves large amount of log data from many sources to a centralized data store for real time steaming analytics’, which is a true/false statement?
2. True.
3. False.
4. Which of the following are true about Flume (choose all that applies)
5. Source is a passive component which receives the event and places it in the Channel
6. Source is an active component which receives the event and places it in the Channel
7. Channel is a passive component which pulls the event from source and send it to Sink.
8. Sink consumes the data (events) from the channels and delivers it to the destination.
9. Agent in flume shares a JVM with other flume agents.
10. Number of flume Agents can run on its own JVM.
11. Which of the following are true about Flume (choose all that applies)
12. Flume supports contextual routing.
13. Fan out flow is connecting multiple sources to a channel.
14. Fan in flow is connecting multiple sources to a channel.
15. Channel can use memory or disk or jdbc for the buffering of events.
16. Sink processor is used for failover and load balancing of sinks.
17. Which of the following are the features of flume (choose all that applies)
18. Flume provides data reliability with event acknowledgment.
19. Flume supports interceptors to filter or enrich events.
20. Sink processor provides fail over and load balancing.
21. Channel supports both transient and persistent storage.
22. All the above.

**HIVE**

36. Hive supports row-level Insert/update and Delete using the ACID features only on which file format?

1. Sequence File.
2. Text File.
3. **ORC file.**
4. RC file.

37. On dropping an external table

1. [**The schema gets dropped without dropping the data**](javascript:void(0);)
2. [The data gets dropped without dropping the schema](javascript:void(0);)
3. [An error is thrown](javascript:void(0);)
4. [Both the schema and the data is dropped](javascript:void(0);)

38. The main advantage of creating table partition is

1. [Effective storage memory utilization](javascript:void(0);)
2. [**faster query performance**](javascript:void(0);)
3. [Less RAM required by namenode](javascript:void(0);)
4. [Simpler query syntax](javascript:void(0);)

39. If the schema of the table does not match with the  data types present in the file containing the table then Hive

1. [Automatically drops the file](javascript:void(0);)
2. [Automatically corrects the data](javascript:void(0);)
3. **Reports Null values for mismatched data**
4. [Does not allow any query to run on the table](javascript:void(0);)

40. Which of the following hint is used to optimize the join queries

1. [/\* joinlast(table\_name) \*/](javascript:void(0);)
2. [/\* joinfirst(table\_name) \*/](javascript:void(0);)
3. [**/\* streamtable(table\_name) \*/**](javascript:void(0);)
4. [/\* cacheable(table\_name) \*/](javascript:void(0);)

41. When importing data to using SerDe, if a row is found to have more columns than expected then

1. [The extra columns are replaced with NULL](javascript:void(0);)
2. [The row is skipped](javascript:void(0);)
3. [The import halts with error](javascript:void(0);)
4. **The Columns are ignored.**

42. If the directory for a partition does not exist, and a query is executed for this partition then

1. [Error is thrown](javascript:void(0);)
2. [mapreduce job is not triggered](javascript:void(0);)
3. [Result from a random partition is returned.](javascript:void(0);)
4. **No result are returned**

43. Which of the following is not a complex  data type in Hive?

1. [**Matrix**](javascript:void(0);)
2. [Array](javascript:void(0);)
3. [Map](javascript:void(0);)
4. [STRUCT](javascript:void(0);)
5. UNION
6. **COMPLEX**

44. On dropping an external table

1. [**The schema gets dropped without dropping the data**](javascript:void(0);)
2. [The data gets dropped without dropping the schema](javascript:void(0);)
3. [An error is thrown](javascript:void(0);)
4. [Both the schema and the data is dropped](javascript:void(0);)

**45. If you run hive as a server, what are the available mechanism for connecting it from application?**

1. Thrift Client: Using thrift you can call hive commands from a various programming languages e.g. C++, Java, PHP, Python and Ruby.
2. JDBC Driver : It supports the Type 4 (pure Java) JDBC Driver
3. ODBC Driver: It supports ODBC protocol.
4. **All the above**

46. Which classes are used by the Hive to Read and Write HDFS Files

1. **TextInputFormat/HiveIgnoreKeyTextOutputFormat: These 2 classes read/write data in plain text file format.**
2. **SequenceFileInputFormat/SequenceFileOutputFormat: These 2 classes read/write data in hadoop SequenceFile format.**
3. KeyValueTextInputformat/ColumnarTextInputFormat : These 2 classes read/write data in columnar text file format.
4. All the above
5. Hive does not use any input/output format

47. What is SerDe in Apache Hive ?

1. A SerDe is a short name for a Serializer Deserializer.
2. Hive uses SerDe (and FileFormat) to read and write data from tables.
3. Users are able to write files to HDFS with whatever tools/mechanism, Hive uses serde to parse the data stored.
4. **All the above**

48. What are the processing engines supported by hive

1. **Tez.**
2. Hue.
3. **Mapreduce.**
4. **Spark.**
5. Scala.
6. Hive native engine.

49. Which of the following is true about partition and bucketing

1. **Using static partition we can reduce the number of partitions and mappers.**
2. Bucket can be created only in the partitioned tables.
3. **Bucket will improve the performance if we join two tables that are bucketed on the same columns.**
4. **Bucket a table is to make sampling more efficient**
5. All the above.

50. Which of the following is true about tuning in Hive

1. **Sort by with distribute by is better than using order by.**
2. Order by with distribute by is better than using sort by.
3. **Compression will improve performance when set mapred.output.compress=true.**
4. **Hive uses local mode to perform all the tasks for the job on a single machine and sometimes in the same process.**
5. All the above