***This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on “History of Hadoop”.***

1. IBM and \_\_\_\_\_\_\_\_ have announced a major initiative to use Hadoop to support university courses in distributed computer programming.

a) Google Latitude

b) Android (operating system)

c) Google Variations

d) Google

2. Point out the correct statement :

a) Hadoop is an ideal environment for extracting and transforming small volumes of data

b) Hadoop stores data in HDFS and supports data compression/decompression

c) The Giraph framework is less useful than a MapReduce job to solve graph and machine learning

d) None of the mentioned

3. What license is Hadoop distributed under ?

a) Apache License 2.0

b) Mozilla Public License

c) Shareware

d) Commercial

4. Sun also has the Hadoop Live CD \_\_\_\_\_\_\_\_ project, which allows running a fully functional Hadoop cluster using a live CD.

a) OpenOffice.org

b) OpenSolaris

c) GNU

d) Linux

5. Which of the following genres does Hadoop produce ?

a) Distributed file system

b) JAX-RS

c) Java Message Service

d) Relational Database Management System

6. What was Hadoop written in ?

a) Java (software platform)

b) Perl

c) Java (programming language)

d) Lua (programming language)

7. Which of the following platforms does Hadoop run on ?

a) Bare metal

b) Debian

c) Cross-platform

d) Unix-like

8. Hadoop achieves reliability by replicating the data across multiple hosts, and hence does not require \_\_\_\_\_\_\_\_ storage on hosts.

a) RAID

b) Standard RAID levels

c) ZFS

d) Operating system

9. Above the file systems comes the \_\_\_\_\_\_\_\_ engine, which consists of one Job Tracker, to which client applications submit MapReduce jobs.

a) MapReduce

b) Google

c) Functional programming

d) Facebook

10. The Hadoop list includes the HBase database, the Apache Mahout \_\_\_\_\_\_\_\_ system, and matrix operations.

a) Machine learning

b) Pattern recognition

c) Statistical classification

d) Artificial intelligence

ANSWERS:

1.Answer: d

Explanation: Google and IBM Announce University Initiative to Address Internet-Scale.

2.Answer: b

Explanation: Data compression can be achieved using compression algorithms like bzip2, gzip, LZO, etc. Different algorithms can be used in different scenarios based on their capabilities.

3.Answer: a

Explanation: Hadoop is Open Source, released under Apache 2 license.

4.Answer: b

Explanation: The OpenSolaris Hadoop LiveCD project built a bootable CD-ROM image.

5.Answer: a

Explanation: The Hadoop Distributed File System (HDFS) is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user.

6.Answer: c

Explanation: The Hadoop framework itself is mostly written in the Java programming language, with some native code in C and command line utilities written as shell-scripts.

7.Answer: c

Explanation: Hadoop has support for cross platform operating system.

8.Answer: a

Explanation: With the default replication value, 3, data is stored on three nodes: two on the same rack, and one on a different rack.

9.Answer: a

Explanation: MapReduce engine uses to distribute work around a cluster.

10.Answer: a

Explanation: The Apache Mahout project’s goal is to build a scalable machine learning tool.

***This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on “Hadoop Ecosystem”.***

1. \_\_\_\_\_\_\_\_ is a platform for constructing data flows for extract, transform, and load (ETL) processing and analysis of large datasets.

a) Pig Latin

b) Oozie

c) Pig

d) Hive

2. Point out the correct statement :

a) Hive is not a relational database, but a query engine that supports the parts of SQL specific to querying data

b) Hive is a relational database with SQL support

c) Pig is a relational database with SQL support

d) All of the mentioned

3. \_\_\_\_\_\_\_\_\_ hides the limitations of Java behind a powerful and concise Clojure API for Cascading.

a) Scalding

b) HCatalog

c) Cascalog

d) All of the mentioned

4. Hive also support custom extensions written in :

a) C#

b) Java

c) C

d) C++

5. Point out the wrong statement :

a) Elastic MapReduce (EMR) is Facebook’s packaged Hadoop offering

b) Amazon Web Service Elastic MapReduce (EMR) is Amazon’s packaged Hadoop offering

c) Scalding is a Scala API on top of Cascading that removes most Java boilerplate

d) All of the mentioned

6. \_\_\_\_\_\_\_\_ is the most popular high-level Java API in Hadoop Ecosystem

a) Scalding

b) HCatalog

c) Cascalog

d) Cascading

7. \_\_\_\_\_\_\_\_\_\_\_ is general-purpose computing model and runtime system for distributed data analytics.

a) Mapreduce

b) Drill

c) Oozie

d) None of the mentioned

8. The Pig Latin scripting language is not only a higher-level data flow language but also has operators similar to :

a) SQL

b) JSON

c) XML

d) All of the mentioned

9. \_\_\_\_\_\_\_ jobs are optimized for scalability but not latency.

a) Mapreduce

b) Drill

c) Oozie

d) Hive

10. \_\_\_\_\_\_ is a framework for performing remote procedure calls and data serialization.

a) Drill

b) BigTop

c) Avro

d) Chukwa

ANSWERS:

1.Answer: c

Explanation: Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs.

2.Answer: a

Explanation: Hive is a SQL-based data warehouse system for Hadoop that facilitates data summarization, ad hoc queries, and the analysis of large datasets stored in Hadoop-compatible file systems.

3.Answer: c

Explanation: Cascalog also adds Logic Programming concepts inspired by Datalog. Hence the name “Cascalog” is a contraction of Cascading and Datalog.

4.Answer: b

Explanation: Hive also support custom extensions written in Java, including user-defined functions (UDFs) and serializer-deserializers for reading and optionally writing custom formats.

5.Answer: a

Explanation: Rather than building Hadoop deployments manually on EC2 (Elastic Compute Cloud) clusters, users can spin up fully configured Hadoop installations using simple invocation commands, either through the AWS Web Console or through command-line tools.

6.Answer: d

Explanation: Cascading hides many of the complexities of MapReduce programming behind more intuitive pipes and data flow abstractions.

7.Answer: a

Explanation: Mapreduce provides a flexible and scalable foundation for analytics, from traditional reporting to leading-edge machine learning algorithms.

8.Answer: a

Explanation: Pig Latin, in essence, is designed to fill the gap between the declarative style of SQL and the low-level procedural style of MapReduce.

9.Answer: d

Explanation: Hive Queries are translated to MapReduce jobs to exploit the scalability of MapReduce.

10.Answer: c

Explanation: In the context of Hadoop, Avro can be used to pass data from one program or language to another.

***This set of Multiple Choice Questions & Answers (MCQs) focuses on “MapReduce”.***

1. A \_\_\_\_\_\_\_\_ node acts as the Slave and is responsible for executing a Task assigned to it by the JobTracker.

a) MapReduce

b) Mapper

c) TaskTracker

d) JobTracker

2. Point out the correct statement :

a) MapReduce tries to place the data and the compute as close as possible

b) Map Task in MapReduce is performed using the Mapper() function

c) Reduce Task in MapReduce is performed using the Map() function

d) All of the mentioned

3. \_\_\_\_\_\_\_\_\_\_\_ part of the MapReduce is responsible for processing one or more chunks of data and producing the output results.

a) Maptask

b) Mapper

c) Task execution

d) All of the mentioned

4. \_\_\_\_\_\_\_\_\_ function is responsible for consolidating the results produced by each of the Map() functions/tasks.

a) Reduce

b) Map

c) Reducer

d) All of the mentioned

5. Point out the wrong statement :

a) A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner

b) The MapReduce framework operates exclusively on pairs

c) Applications typically implement the Mapper and Reducer interfaces to provide the map and reduce methods

6. Although the Hadoop framework is implemented in Java , MapReduce applications need not be written in :

a) Java

b) C

c) C#

d) None of the mentioned

7. \_\_\_\_\_\_\_\_ is a utility which allows users to create and run jobs with any executables as the mapper and/or the reducer.

a) Hadoop Strdata

b) Hadoop Streaming

c) Hadoop Stream

d) None of the mentioned

8. \_\_\_\_\_\_\_\_\_\_ maps input key/value pairs to a set of intermediate key/value pairs.

a) Mapper

b) Reducer

c) Both Mapper and Reducer

d) None of the mentioned

9. The number of maps is usually driven by the total size of :

a) inputs

b) outputs

c) tasks

d) None of the mentioned

10. \_\_\_\_\_\_\_\_\_ is the default Partitioner for partitioning key space.

a) HashPar

b) Partitioner

c) HashPartitioner

d) None of the mentioned

ANSWERS:

1.Answer: c

Explanation: TaskTracker receives the information necessary for execution of a Task from JobTracker, Executes the Task, and Sends the Results back to JobTracker.

2.Answer: a

Explanation: This feature of MapReduce is “Data Locality”.

3.Answer: a

Explanation: Map Task in MapReduce is performed using the Map() function.

4.Answer: a

Explanation: Reduce function collates the work and resolves the results.

5.Answer: d

Explanation: The MapReduce framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

6.Answer: a

Explanation: Hadoop Pipes is a SWIG- compatible C++ API to implement MapReduce applications (non JNITM based).

7.Answer: b

Explanation: Hadoop streaming is one of the most important utilities in the Apache Hadoop distribution.

8.Answer: a

Explanation: Maps are the individual tasks that transform input records into intermediate records.

9.Answer: a

Explanation: Total size of inputs means total number of blocks of the input files.

10.Answer: c

Explanation: The default partitioner in Hadoop is the HashPartitioner which has a method called getPartition to partition.

***This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on “Analyzing Data with Hadoop”.***

1. Mapper implementations are passed the JobConf for the job via the \_\_\_\_\_\_\_\_ method

a) JobConfigure.configure

b) JobConfigurable.configure

c) JobConfigurable.configureable

d) None of the mentioned

2. Point out the correct statement :

a) Applications can use the Reporter to report progress

b) The Hadoop MapReduce framework spawns one map task for each InputSplit generated by the InputFormat for the job

c) The intermediate, sorted outputs are always stored in a simple (key-len, key, value-len, value) format

d) All of the mentioned

3. Input to the \_\_\_\_\_\_\_ is the sorted output of the mappers.

a) Reducer

b) Mapper

c) Shuffle

d) All of the mentioned

4. The right number of reduces seems to be :

a) 0.90

b) 0.80

c) 0.36

d) 0.95

5. Point out the wrong statement :

a) Reducer has 2 primary phases

b) Increasing the number of reduces increases the framework overhead, but increases load balancing and lowers the cost of failures

c) It is legal to set the number of reduce-tasks to zero if no reduction is desired

d) The framework groups Reducer inputs by keys (since different mappers may have output the same key) in sort stage

6. The output of the \_\_\_\_\_\_\_ is not sorted in the Mapreduce framework for Hadoop.

a) Mapper

b) Cascader

c) Scalding

d) None of the mentioned

7. Which of the following phases occur simultaneously ?

a) Shuffle and Sort

b) Reduce and Sort

c) Shuffle and Map

d) All of the mentioned

8. Mapper and Reducer implementations can use the \_\_\_\_\_\_\_\_ to report progress or just indicate that they are alive.

a) Partitioner

b) OutputCollector

c) Reporter

d) All of the mentioned

9. \_\_\_\_\_\_\_\_\_\_ is a generalization of the facility provided by the MapReduce framework to collect data output by the Mapper or the Reducer

a) Partitioner

b) OutputCollector

c) Reporter

d) All of the mentioned

10. \_\_\_\_\_\_\_\_\_ is the primary interface for a user to describe a MapReduce job to the Hadoop framework for execution.

a) Map Parameters

b) JobConf

c) MemoryConf

d) None of the mentioned

ANSWERS:

1.Answer: b

Explanation: JobConfigurable.configure method is overridden to initialize themselves.

2.Answer: d

Explanation: Reporters can be used to set application-level status messages and update Counters.

3.Answer: a

Explanation: In Shuffle phase the framework fetches the relevant partition of the output of all the mappers, via HTTP.

4.Answer: d

Explanation: The right number of reduces seems to be 0.95 or 1.75.

5.Answer: a

Explanation: Reducer has 3 primary phases: shuffle, sort and reduce.

6.Answer: d

Explanation: The output of the reduce task is typically written to the FileSystem. The output of the Reducer is not sorted.

7.Answer: a

Explanation: The shuffle and sort phases occur simultaneously; while map-outputs are being fetched they are merged.

8.Answer: c

Explanation: Reporter is a facility for MapReduce applications to report progress, set application-level status messages and update Counters.

9.Answer: b

Explanation: Hadoop MapReduce comes bundled with a library of generally useful mappers, reducers, and partitioners.

10.Answer: b

Explanation: JobConf represents a MapReduce job configuration.

***This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on “Scaling out in Hadoop”.***

1. \_\_\_\_\_\_\_\_ systems are scale-out file-based (HDD) systems moving to more uses of memory in the nodes.

a) NoSQL

b) NewSQL

c) SQL

d) All of the mentioned

2. Point out the correct statement :

a) Hadoop is ideal for the analytical, post-operational, data-warehouse-ish type of workload

b) HDFS runs on a small cluster of commodity-class nodes

c) NEWSQL is frequently the collection point for big data

d) None of the mentioned

3. Hadoop data is not sequenced and is in 64MB to 256 MB block sizes of delimited record values with schema applied on read based on:

a) HCatalog

b) Hive

c) Hbase

d) All of the mentioned

4. \_\_\_\_\_\_\_\_\_\_ are highly resilient and eliminate the single-point-of-failure risk with traditional Hadoop deployments

a) EMR

b) Isilon solutions

c) AWS

d) None of the mentioned

5. Point out the wrong statement :

a) EMC Isilon Scale-out Storage Solutions for Hadoop combine a powerful yet simple and highly efficient storage platform

b) Isilon’s native HDFS integration means you can avoid the need to invest in a separate Hadoop infrastructure

c) NoSQL systems do provide high latency access and accommodate less concurrent users

d) None of the mentioned

6. HDFS and NoSQL file systems focus almost exclusively on adding nodes to :

a) Scale out

b) Scale up

c) Both Scale out and up

d) None of the mentioned

7. Which is the most popular NoSQL database for scalable big data store with Hadoop ?

a) Hbase

b) MongoDB

c) Cassandra

d) None of the mentioned

8. The \_\_\_\_\_\_\_\_\_\_\_ can also be used to distribute both jars and native libraries for use in the map and/or reduce tasks.

a) DataCache

b) DistributedData

c) DistributedCache

d) All of the mentioned

9. HBase provides \_\_\_\_\_\_\_\_\_\_\_ like capabilities on top of Hadoop and HDFS.

a) TopTable

b) BigTop

c) Bigtable

d) None of the mentioned

10. \_\_\_\_\_\_\_ refers to incremental costs with no major impact on solution design, performance and complexity.

a) Scale-out

b) Scale-down

c) Scale-up

d) None of the mentioned

ANSWERS:

1.Answer: a

Explanation: NoSQL systems make the most sense whenever the application is based on data with varying data types and the data can be stored in key-value notation.

2.Answer: a

Explanation: Hadoop together with a relational data warehouse, they can form very effective data warehouse infrastructure.

3.Answer: a

Explanation: Other means of tagging the values also can be used.

4.Answer: b

Explanation: enterprise data protection and security options including file system auditing and data-at-rest encryption to address compliance requirements is also provided by Isilon solution.

5.Answer: c

Explanation: NoSQL systems do provide low latency access and accommodate many concurrent users.

6.Answer: c

Explanation: NoSQL systems do provide low latency access and accommodate many concurrent users.

7.Answer: a

Explanation: HBase is the Hadoop database: a distributed, scalable Big Data store that lets you host very large tables — billions of rows multiplied by millions of columns — on clusters built with commodity hardware.

8.Answer: c

Explanation: The child-jvm always has its current working directory added to the java.library.path and LD\_LIBRARY\_PATH.

9.Answer: c

Explanation: Google Bigtable leverages the distributed data storage provided by the Google File System.

10.Answer: c

Explanation: Adding more CPU/RAM/Disk capacity to Hadoop DataNode that is already part of a cluster does not require additional network switches.

***This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on “Hadoop Streaming”.***

1. Streaming supports streaming command options as well as \_\_\_\_\_\_\_\_\_ command options.

a) generic

b) tool

c) library

d) task

2. Point out the correct statement :

a) You can specify any executable as the mapper and/or the reducer

b) You cannot supply a Java class as the mapper and/or the reducer

c) The class you supply for the output format should return key/value pairs of Text class

d) All of the mentioned

3. Which of the following Hadoop streaming command option parameter is required ?

a) output directoryname

b) mapper executable

c) input directoryname

d) all of the mentioned

4. To set an environment variable in a streaming command use:

a) -cmden EXAMPLE\_DIR=/home/example/dictionaries/

b) -cmdev EXAMPLE\_DIR=/home/example/dictionaries/

c) -cmdenv EXAMPLE\_DIR=/home/example/dictionaries/

d) -cmenv EXAMPLE\_DIR=/home/example/dictionaries/

5. Point out the wrong statement :

a) Hadoop has a library package called Aggregate

b) Aggregate allows you to define a mapper plugin class that is expected to generate “aggregatable items” for each input key/value pair of the mappers

c) To use Aggregate, simply specify “-mapper aggregate”

d) None of the mentioned

6. The \_\_\_\_\_\_\_\_ option allows you to copy jars locally to the current working directory of tasks and automatically unjar the files.

a) archives

b) files

c) task

d) none of the mentioned

7. \_\_\_\_\_\_\_\_\_\_\_\_\_\_ class allows the Map/Reduce framework to partition the map outputs based on certain key fields, not the whole keys.

a) KeyFieldPartitioner

b) KeyFieldBasedPartitioner

c) KeyFieldBased

d) None of the mentioned

8. Which of the following class provides a subset of features provided by the Unix/GNU Sort ?

a) KeyFieldBased

b) KeyFieldComparator

c) KeyFieldBasedComparator

d) All of the mentioned

9. Which of the following class is provided by Aggregate package ?

a) Map

b) Reducer

c) Reduce

d) None of the mentioned

10.Hadoop has a library class, org.apache.hadoop.mapred.lib.FieldSelectionMapReduce, that effectively allows you to process text data like the unix \_\_\_\_\_\_ utility.

a) Copy

b) Cut

c) Paste

d) Move

ANSWERS:

1.Answer: a

Explanation: Place the generic options before the streaming options, otherwise the command will fail.

2.Answer: a

Explanation: If you do not specify an input format class, the TextInputFormat is used as the default.

3.3. Which of the following Hadoop streaming command option parameter is required ?

a) output directoryname

b) mapper executable

c) input directoryname

d) all of the mentioned

4.Answer: c

Explanation: Environment Variable is set using cmdenv command.

5.Answer: c

Explanation: To use Aggregate, simply specify “-reducer aggregate”:

6.Answer: a

Explanation: Archives options is also a generic option.

7.Answer: b

Explanation: The primary key is used for partitioning, and the combination of the primary and secondary keys is used for sorting.

8.Answer: c

Explanation: Hadoop has a library class, KeyFieldBasedComparator, that is useful for many applications.

9.Answer: b

Explanation: Aggregate provides a special reducer class and a special combiner class, and a list of simple aggregators that perform aggregations such as “sum”, “max”, “min” and so on over a sequence of values.

10.Answer: b

Explanation: The map function defined in the class treats each input key/value pair as a list of fields.

***This set of Multiple Choice Questions & Answers (MCQs) focuses on “Hadoop Filesystem – HDFS”.***

1. A \_\_\_\_\_\_\_\_ serves as the master and there is only one NameNode per cluster.

a) Data Node

b) NameNode

c) Data block

d) Replication

2. Point out the correct statement :

a) DataNode is the slave/worker node and holds the user data in the form of Data Blocks

b) Each incoming file is broken into 32 MB by default

c) Data blocks are replicated across different nodes in the cluster to ensure a low degree of fault tolerance

d) None of the mentioned

3. HDFS works in a \_\_\_\_\_\_\_\_\_\_ fashion.

a) master-worker

b) master-slave

c) worker/slave

d) all of the mentioned

4. \_\_\_\_\_\_\_\_ NameNode is used when the Primary NameNode goes down.

a) Rack

b) Data

c) Secondary

d) None of the mentioned

5. Point out the wrong statement :

a) Replication Factor can be configured at a cluster level (Default is set to 3) and also at a file level

b) Block Report from each DataNode contains a list of all the blocks that are stored on that DataNode

c) User data is stored on the local file system of DataNodes

d) DataNode is aware of the files to which the blocks stored on it belong to

6. Which of the following scenario may not be a good fit for HDFS ?

a) HDFS is not suitable for scenarios requiring multiple/simultaneous writes to the same file

b) HDFS is suitable for storing data related to applications requiring low latency data access

c) HDFS is suitable for storing data related to applications requiring low latency data access

d) None of the mentioned

7. The need for data replication can arise in various scenarios like :

a) Replication Factor is changed

b) DataNode goes down

c) Data Blocks get corrupted

d) All of the mentioned

8. \_\_\_\_\_\_\_\_ is the slave/worker node and holds the user data in the form of Data Blocks.

a) DataNode

b) NameNode

c) Data block

d) Replication

9. HDFS provides a command line interface called \_\_\_\_\_\_\_\_\_\_ used to interact with HDFS.

a) “HDFS Shell”

b) “FS Shell”

c) “DFS Shell”

d) None of the mentioned

10. HDFS is implemented in \_\_\_\_\_\_\_\_\_\_\_\_\_ programming language.

a) C++

b) Java

c) Scala

d) None of the mentioned

ANSWERS:

1.Answer: b

Explanation: All the metadata related to HDFS including the information about data nodes, files stored on HDFS, and Replication, etc. are stored and maintained on the NameNode.

2.Answer: a

Explanation: There can be any number of DataNodes in a Hadoop Cluster.

3.Answer: a

Explanation: NameNode servers as the master and each DataNode servers as a worker/slave

4.Answer: c

Explanation: Secondary namenode is used for all time availability and reliability.

5.Answer: d

Explanation: NameNode is aware of the files to which the blocks stored on it belong to.

6.Answer: a

Explanation: HDFS can be used for storing archive data since it is cheaper as HDFS allows storing the data on low cost commodity hardware while ensuring a high degree of fault-tolerance.

7.Answer: d

Explanation: Data is replicated across different DataNodes to ensure a high degree of fault-tolerance.

8.Answer: a

Explanation: A DataNode stores data in the [HadoopFileSystem]. A functional filesystem has more than one DataNode, with data replicated across them.

9.Answer: b

Explanation: The File System (FS) shell includes various shell-like commands that directly interact with the Hadoop Distributed File System (HDFS).

10.Answer: b

Explanation: HDFS is implemented in Java and any computer which can run Java can host a NameNode/DataNode on it.

***This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on “Java Interface”.***

1. In order to read any file in HDFS, instance of \_\_\_\_\_\_\_\_\_\_ is required.

a) filesystem

b) datastream

c) outstream

d) inputstream

2. Point out the correct statement :

a) The framework groups Reducer inputs by keys

b) The shuffle and sort phases occur simultaneously i.e. while outputs are being fetched they are merged

c) Since JobConf.setOutputKeyComparatorClass(Class) can be used to control how intermediate keys are grouped, these can be used in conjunction to simulate secondary sort on values

d) All of the mentioned

3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_ is method to copy byte from input stream to any other stream in Hadoop.

a) IOUtils

b) Utils

c) IUtils

d) All of the mentioned

4. \_\_\_\_\_\_\_\_\_\_\_\_\_ is used to read data from bytes buffers .

a) write()

b) read()

c) readwrite()

d) all of the mentioned

5. Point out the wrong statement :

a) The framework calls reduce method for each pair in the grouped inputs

b) The output of the Reducer is re-sorted

c) reduce method reduces values for a given key

d) None of the mentioned

6. Interface \_\_\_\_\_\_\_\_\_\_\_\_ reduces a set of intermediate values which share a key to a smaller set of values.

a) Mapper

b) Reducer

c) Writable

d) Readable

7. Reducer is input the grouped output of a :

a) Mapper

b) Reducer

c) Writable

d) Readable

8. The output of the reduce task is typically written to the FileSystem via :

a) OutputCollector

b) InputCollector

c) OutputCollect

d) All of the mentioned

9. Applications can use the \_\_\_\_\_\_\_\_\_ provided to report progress or just indicate that they are alive.

a) Collector

b) Reporter

c) Dashboard

d) None of the mentioned

10. Which of the following parameter is to collect keys and combined values ?

a) key

b) values

c) reporter

d) output

ANSWERS:

1.Answer: a

Explanation: InputDataStream is used to read data from file.

2.Answer: d

Explanation: If equivalence rules for keys while grouping the intermediates are different from those for grouping keys before reduction, then one may specify a Comparator.

3.Answer: a

Explanation: IOUtils class is static method in Java interface.

4.Answer: a

Explanation: readfully method can also be used instead of read method.

5.Answer: b

Explanation: The output of the Reducer is not re-sorted.

6.Answer: b

Explanation: Reducer implementations can access the JobConf for the job.

7.Answer: a

Explanation: In the phase the framework, for each Reducer, fetches the relevant partition of the output of all the Mappers, via HTTP.

8.Answer: a

Explanation: In reduce phase the reduce(Object, Iterator, OutputCollector, Reporter) method is called for each pair in the grouped inputs.

9.Answer: b

Explanation: In scenarios where the application takes a significant amount of time to process individual key/value pairs, this is crucial since the framework might assume that the task has timed-out and kill that task.

10.Answer: d

Explanation: reporter parameter is for facility to report progress.

***This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on “Data Flow”.***

1. \_\_\_\_\_\_\_\_ is a programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks.

a) Hive

b) MapReduce

c) Pig

d) Lucene

2. Point out the correct statement :

a) Data locality means movement of algorithm to the data instead of data to algorithm

b) When the processing is done on the data algorithm is moved across the Action Nodes rather than data to the algorithm

c) Moving Computation is expensive than Moving Data

d) None of the mentioned

3. The daemons associated with the MapReduce phase are \_\_\_\_\_\_\_\_ and task-trackers.

a) job-tracker

b) map-tracker

c) reduce-tracker

d) all of the mentioned

4. The JobTracker pushes work out to available \_\_\_\_\_\_\_ nodes in the cluster, striving to keep the work as close to the data as possible

a) DataNodes

b) TaskTracker

c) ActionNodes

d) All of the mentioned

5. Point out the wrong statement :

a) The map function in Hadoop MapReduce have the following general form:map:(K1, V1) → list(K2, V2)

b) The reduce function in Hadoop MapReduce have the following general form: reduce: (K2, list(V2)) → list(K3, V3)

c) MapReduce has a complex model of data processing: inputs and outputs for the map and reduce functions are key-value pairs

d) None of the mentioned

6. InputFormat class calls the \_\_\_\_\_\_\_\_ function and computes splits for each file and then sends them to the jobtracker.

a) puts

b) gets

c) getSplits

d) all of the mentioned

7. On a tasktracker, the map task passes the split to the createRecordReader() method on InputFormat to obtain a \_\_\_\_\_\_\_\_\_ for that split.

a) InputReader

b) RecordReader

c) OutputReader

d) None of the mentioned

8. The default InputFormat is \_\_\_\_\_\_\_\_\_\_ which treats each value of input a new value and the associated key is byte offset.

a) TextFormat

b) TextInputFormat

c) InputFormat

d) All of the mentioned

9. \_\_\_\_\_\_\_\_\_\_ controls the partitioning of the keys of the intermediate map-outputs.

a) Collector

b) Partitioner

c) InputFormat

d) None of the mentioned

10. Output of the mapper is first written on the local disk for sorting and \_\_\_\_\_\_\_\_\_ process.

a) shuffling

b) secondary sorting

c) forking

d) reducing

ANSWERS:

1.Answer: b

Explanation: MapReduce is the heart of hadoop.

2.Answer: a

Explanation: Data flow framework possesses the feature of data locality.

3.Answer: a

Explanation: Map-Reduce jobs are submitted on job-tracker.

4.Answer: b

Explanation: A heartbeat is sent from the TaskTracker to the JobTracker every few minutes to check its status whether the node is dead or alive.

5.Answer: c

Explanation: MapReduce is relatively simple model to implement in Hadoop.

6.Answer: c

Explanation: InputFormat uses their storage locations to schedule map tasks to process them on the tasktrackers.

7.Answer: b

Explanation: The RecordReader loads data from its source and converts into key-value pairs suitable for reading by mapper.

8.Answer: b

Explanation: A RecordReader is little more than an iterator over records, and the map task uses one to generate record key-value pairs.

9.Answer: b

Explanation: The output of the mapper is sent to the partitioner.

10.Answer: a

Explanation: All values corresponding to the same key will go the same reducer.

***This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on “Hadoop Archives”.***

1. \_\_\_\_\_\_\_\_\_ is the name of the archive you would like to create.

a) archive

b) archiveName

c) name

d) none of the mentioned

2. Point out the correct statement :

a) A Hadoop archive maps to a file system directory

b) Hadoop archives are special format archives

c) A Hadoop archive always has a \*.har extension

d) All of the mentioned

3. Using Hadoop Archives in \_\_\_\_\_\_\_\_\_\_ is as easy as specifying a different input filesystem than the default file system.

a) Hive

b) Pig

c) MapReduce

d) All of the mentioned

4. The \_\_\_\_\_\_\_\_\_\_ guarantees that excess resources taken from a queue will be restored to it within N minutes of its need for them.

a) capacitor

b) scheduler

c) datanode

d) none of the mentioned

5. Point out the wrong statement :

a) The Hadoop archive exposes itself as a file system layer

b) Hadoop archives are immutable

c) Archive rename’s, deletes and creates return an error

d) None of the mentioned

6. \_\_\_\_\_\_\_\_\_ is a pluggable Map/Reduce scheduler for Hadoop which provides a way to share large clusters.

a) Flow Scheduler

b) Data Scheduler

c) Capacity Scheduler

d) None of the mentioned

7. Which of the following parameter describes destination directory which would contain the archive ?

a) -archiveName

b) <source>

c) <destination>

d) none of the mentioned

8. \_\_\_\_\_\_\_\_\_ identifies filesystem pathnames which work as usual with regular expressions.

a) -archiveName <name>

b) <source>

c) <destination>

d) none of the mentioned

9. \_\_\_\_\_\_\_\_\_\_ is the parent argument used to specify the relative path to which the files should be archived to

a) -archiveName <name>

b) -p <parent\_path>

c) <destination>

d) <source>

10. Which of the following is a valid syntax for hadoop archive ?

a)

hadooparchive [ Generic Options ] archive  
 -archiveName <name>  
 [-p <parent>]  
 <source>  
 <destination>

b)

hadooparch [ Generic Options ] archive  
 -archiveName <name>  
 [-p <parent>]  
 <source>  
 <destination>

c)

hadoop [ Generic Options ] archive  
 -archiveName <name>  
 [-p <parent>]  
 <source>  
 <destination>

d) None of the mentioned

ANSWERS:

1.Answer: b

Explanation: The name should have a \*.har extension.

2.Answer: d

Explanation: A Hadoop archive directory contains metadata (in the form of \_index and \_masterindex) and data (part-\*) files.

3.

Answer: c

Explanation: Hadoop Archives is exposed as a file system MapReduce will be able to use all the logical input files in Hadoop Archives as input.

4.Answer: b

Explanation: Free resources can be allocated to any queue beyond its guaranteed capacity.

5.Answer: d

Explanation: All the fs shell commands in the archives work but with a different URI.

6.Answer: c

Explanation: The Capacity Scheduler supports for multiple queues, where a job is submitted to a queue.

7.Answer: c

Explanation: -archiveName is the name of the archive to be created.

8.Answer: d

Explanation: identifies destination directory which would contain the archive.

9.Answer: b

Explanation: The hadoop archive command creates a Hadoop archive, a file that contains other files.

10.Answer: c

Explanation: The Hadoop archiving tool can be invoked using the following command format: hadoop archive -archiveName name -p \* .

***This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on “Hadoop I/O”.***

1. Hadoop I/O Hadoop comes with a set of \_\_\_\_\_\_\_\_ for data I/O.

a) methods

b) commands

c) classes

d) none of the mentioned

2. Point out the correct statement :

a) The sequence file also can contain a “secondary” key-value list that can be used as file Metadata

b) SequenceFile formats share a header that contains some information which allows the reader to recognize is format

c) There’re Key and Value Class Name’s that allow the reader to instantiate those classes, via reflection, for reading

d) All of the mentioned

3. Apache Hadoop’s \_\_\_\_\_\_\_\_\_\_\_ provides a persistent data structure for binary key-value pairs.

a) GetFile

b) SequenceFile

c) Putfile

d) All of the mentioned

4. How many formats of SequenceFile are present in Hadoop I/O ?

a) 2

b) 3

c) 4

d) 5

5. Point out the wrong statement :

a) The data file contains all the key, value records but key N + 1 must be greater then or equal to the key N

b) Sequence file is a kind of hadoop file based data structure

c) Map file type is splittable as it contains a sync point after several records

d) None of the mentioned

6. Which of the following format is more compression-aggressive ?

a) Partition Compressed

b) Record Compressed

c) Block-Compressed

d) Uncompressed

7. The \_\_\_\_\_\_\_\_\_\_ is a directory that contains two SequenceFile.

a) ReduceFile

b) MapperFile

c) MapFile

d) None of the mentioned

8. The \_\_\_\_\_\_ file is populated with the key and a LongWritable that contains the starting byte position of the record.

a) Array

b) Index

c) Immutable

d) All of the mentioned

9. The \_\_\_\_\_\_\_\_\_ as just the value field append(value) and the key is a LongWritable that contains the record number, count + 1.

a) SetFile

b) ArrayFile

c) BloomMapFile

d) None of the mentioned

10. \_\_\_\_\_\_\_\_\_\_\_\_ data file takes is based on avro serialization framework which was primarily created for hadoop.

a) Oozie

b) Avro

c) cTakes

d) Lucene

ANSWERS:

1.Answer: d

Explanation: Hadoop I/O consist of primitives for serialization and deserialization.

2.Answer: d

Explanation: In contrast with other persistent key-value data structures like B-Trees, you can’t seek to a specified key editing, adding or removing it.

3.Answer: b

Explanation: SequenceFile is append-only.

4.Answer: b

Explanation: SequenceFile has 3 available formats: An “Uncompressed” format, A “Record Compressed” format and a “Block-Compressed”.

5.Answer: c

Explanation: Map file is again a kind of hadoop file based data structure and it differs from a sequence file in a matter of the order.

6.Answer: c

Explanation: SequenceFile key-value list can be just a Text/Text pair, and is written to the file during the initialization that happens in the SequenceFile.

7.Answer: c

Explanation: Sequence files are data file (“/data”) and the index file (“/index”).

8.Answer: b

Explanation: Index does’t contains all the keys but just a fraction of the keys.

9.Answer: b

Explanation: The SetFile instead of append(key, value) as just the key field append(key) and the value is always the NullWritable instance.

10.Answer: b

Explanation: Avro is a splittable data format with a metadata section at the beginning and then a sequence of avro serialized objects.

***This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on “Compression”.***

1. The \_\_\_\_\_\_\_\_\_ codec from Google provides modest compression ratios.

a) Snapcheck

b) Snappy

c) FileCompress

d) None of the mentioned

2. Point out the correct statement :

a) Snappy is licensed under the GNU Public License (GPL)

b) BgCIK needs to create an index when it compresses a file

c) The Snappy codec is integrated into Hadoop Common, a set of common utilities that supports other Hadoop subprojects

d) None of the mentioned

3. Which of the following compression is similar to Snappy compression ?

a) LZO

b) Bzip2

c) Gzip

d) All of the mentioned

4. Which of the following supports splittable compression ?

a) LZO

b) Bzip2

c) Gzip

d) All of the mentioned

5. Point out the wrong statement :

a) From a usability standpoint, LZO and Gzip are similar.

b) Bzip2 generates a better compression ratio than does Gzip, but it’s much slower

c) Gzip is a compression utility that was adopted by the GNU project

d) None of the mentioned

6. Which of the following is the slowest compression technique ?

a) LZO

b) Bzip2

c) Gzip

d) All of the mentioned

7. Gzip (short for GNU zip) generates compressed files that have a \_\_\_\_\_\_\_\_\_ extension.

a) .gzip

b) .gz

c) .gzp

d) .g

8. Which of the following is based on the DEFLATE algorithm ?

a) LZO

b) Bzip2

c) Gzip

d) All of the mentioned

9. \_\_\_\_\_\_\_\_\_\_ typically compresses files to within 10% to 15% of the best available techniques.

a) LZO

b) Bzip2

c) Gzip

d) All of the mentioned

10. The LZO compression format is composed of approximately \_\_\_\_\_\_\_\_\_\_ blocks of compressed data.

a) 128k

b) 256k

c) 24k

d) 36k

ANSWERS:

1.Answer: b

Explanation: Snappy has fast compression and decompression speeds.

2.Answer: c

Explanation: You can use Snappy as an add-on for more recent versions of Hadoop that do not yet provide Snappy codec support.

3.Answer: a

Explanation: LZO is only really desirable if you need to compress text files.

4.Answer: a

Explanation: LZO enables the parallel processing of compressed text file splits by your MapReduce jobs.

5.Answer: a

Explanation: From a usability standpoint, Bzip2 and Gzip are similar.

6.Answer: b

Explanation: Of all the available compression codecs in Hadoop, Bzip2 is by far the slowest.

7.Answer: b

Explanation: You can use the gunzip command to decompress files that were created by a number of compression utilities, including Gzip.

8.Answer: c

Explanation: gzip is based on the DEFLATE algorithm, which is a combination of LZ77 and Huffman Coding.

9.Answer: b

Explanation: bzip2 is a freely available, patent free (see below), high-quality data compressor.

10.Answer: b

Explanation: LZO was designed with speed in mind: it decompresses about twice as fast as gzip, meaning it’s fast enough to keep up with hard drive read speeds.

***This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on “Data Integrity”.***

1. The HDFS client software implements \_\_\_\_\_\_\_\_\_\_ checking on the contents of HDFS files.

a) metastore

b) parity

c) checksum

d) none of the mentioned

2. Point out the correct statement :

a) The HDFS architecture is compatible with data rebalancing schemes

b) Datablocks support storing a copy of data at a particular instant of time

c) HDFS currently support snapshots

d) None of the mentioned

3. The \_\_\_\_\_\_\_\_\_\_\_ machine is a single point of failure for an HDFS cluster.

a) DataNode

b) NameNode

c) ActionNode

d) All of the mentioned

4. The \_\_\_\_\_\_\_\_\_\_\_\_ and the EditLog are central data structures of HDFS.

a) DsImage

b) FsImage

c) FsImages

d) All of the mentioned

5. Point out the wrong statement :

a) HDFS is designed to support small files only

b) Any update to either the FsImage or EditLog causes each of the FsImages and EditLogs to get updated synchronously

c) NameNode can be configured to support maintaining multiple copies of the FsImage and EditLog

d) None of the mentioned

6. \_\_\_\_\_\_\_\_\_\_ support storing a copy of data at a particular instant of time.

a) Data Image

b) Datanots

c) Snapshots

d) All of the mentioned

7. Automatic restart and \_\_\_\_\_\_\_\_\_\_\_\_ of the NameNode software to another machine is not supported.

a) failover

b) end

c) scalability

d) all of the mentioned

8. HDFS, by default, replicates each data block \_\_\_\_\_ times on different nodes and on at least \_\_\_\_ racks.

a) 3,2

b) 1,2

c) 2,3

d) All of the mentioned

9. \_\_\_\_\_\_\_\_\_ stores its metadata on multiple disks that typically include a non-local file server.

a) DataNode

b) NameNode

c) ActionNode

d) None of the mentioned

10. The HDFS file system is temporarily unavailable whenever the HDFS \_\_\_\_\_\_\_\_ is down.

a) DataNode

b) NameNode

c) ActionNode

d) None of the mentioned

ANSWERS:

1.Answer: c

Explanation: When a client creates an HDFS file, it computes a checksum of each block of the file and stores these checksums in a separate hidden file in the same HDFS namespace.

2.Answer: a

Explanation: A scheme might automatically move data from one DataNode to another if the free space on a DataNode falls below a certain threshold.

3.Answer: b

Explanation: If the NameNode machine fails, manual intervention is necessary. Currently, automatic restart and failover of the NameNode software to another machine is not supported.

4.Answer: b

Explanation: A corruption of these files can cause the HDFS instance to be non-functional.

5.Answer: a

Explanation: HDFS is designed to support very large files.

6.Answer: c

Explanation: One usage of the snapshot feature may be to roll back a corrupted HDFS instance to a previously known good point in time.

7.Answer: a

Explanation: If the NameNode machine fails, manual intervention is necessary.

8.Answer: a

Explanation: HDFS has a simple yet robust architecture that was explicitly designed for data reliability in the face of faults and failures in disks, nodes and networks.

9.Answer: b

Explanation: HDFS tolerates failures of storage servers (called DataNodes) and its disks.

10.Answer: b

Explanation: When the HDFS NameNode is restarted it recovers its metadata.

***This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on “Serialization”.***

1. Apache \_\_\_\_\_\_\_ is a serialization framework that produces data in a compact binary format.

a) Oozie

b) Impala

c) kafka

d) Avro

2. Point out the correct statement :

a) Apache Avro is a framework that allows you to serialize data in a format that has a schema built in

b) The serialized data is in a compact binary format that doesn’t require proxy objects or code generation

c) Including schemas with the Avro messages allows any application to deserialize the data

d) All of the mentioned

3. Avro schemas describe the format of the message and are defined using :

a) JSON

b) XML

c) JS

d) All of the mentioned

4. The \_\_\_\_\_\_\_\_\_\_\_\_ is an iterator which reads through the file and returns objects using the next() method.

a) DatReader

b) DatumReader

c) DatumRead

d) None of the mentioned

5. Point out the wrong statement :

a) Java code is used to deserialize the contents of the file into objects

b) Avro allows you to use complex data structures within Hadoop MapReduce jobs

c) The m2e plug-in automatically downloads the newly added JAR files and their dependencies

d) None of the mentioned

6. The \_\_\_\_\_\_\_\_\_\_\_\_ class extends and implements several Hadoop-supplied interfaces.

a) AvroReducer

b) Mapper

c) AvroMapper

d) None of the mentioned

7. \_\_\_\_\_\_\_\_\_\_\_\_ class accepts the values that the ModelCountMapper object has collected.

a) AvroReducer

b) Mapper

c) AvroMapper

d) None of the mentioned

8. The \_\_\_\_\_\_\_\_ method in the ModelCountReducer class “reduces” the values the mapper collects into a derived value

a) count

b) add

c) reduce

d) all of the mentioned

9. Which of the following works well with Avro ?

a) Lucene

b) kafka

c) MapReduce

d) None of the mentioned

10. \_\_\_\_\_\_\_\_\_\_ tools is used to generate proxy objects in Java to easily work with the objects.

a) Lucene

b) kafka

c) MapReduce

d) Avro

ANSWERS:

1.Answer: d

Explanation: Apache Avro doesn’t require proxy objects or code generation.

2.Answer: d

Explanation: Instead of using generated proxy libraries and strong typing, Avro relies heavily on the schemas that are sent along with the serialized data.

3.Answer: a

Explanation: The JSON schema content is put into a file.

4.Answer: b

Explanation: DatumReader reads the content through the DataFileReader implementation.

5.Answer: d

Explanation: A unit test is useful because you can make assertions to verify that the values of the deserialized object are the same as the original values.

6.Answer: c

Explanation: AvroMapper is used to provide the ability to collect or map data.

7.Answer: a

Explanation: AvroReducer summarizes them by looping through the values.

8.Answer: c

Explanation: In some case, it can be simple sum of the values.

9.Answer: c

Explanation: You can use Avro and MapReduce together to process many items serialized with Avro’s small binary format.

10.Answer: d

Explanation: Avro serialization includes the schema with it — in JSON format — which allows you to have different versions of objects.

***This set of Hadoop Multiple Choice Questions & Answers (MCQs) focuses on “Avro-1”.***

1. Avro schemas are defined with \_\_\_\_\_

a) JSON

b) XML

c) JAVA

d) All of the mentioned

2. Point out the correct statement :

a) Avro provides functionality similar to systems such as Thrift

b) When Avro is used in RPC, the client and server exchange data in the connection handshake

c) Apache Avro, Avro, Apache, and the Avro and Apache logos are trademarks of The Java Foundation

d) None of the mentioned

3. \_\_\_\_\_\_\_\_\_\_ facilitates construction of generic data-processing systems and languages.

a) Untagged data

b) Dynamic typing

c) No manually-assigned field IDs

d) All of the mentioned

4. With \_\_\_\_\_\_ we can store data and read it easily with various programming languages

a) Thrift

b) Protocol Buffers

c) Avro

d) None of the mentioned

5. Point out the wrong statement :

a) Apache Avro™ is a data serialization system

b) Avro provides simple integration with dynamic languages

c) Avro provides rich data structures

d) All of the mentioned

6. \_\_\_\_\_\_\_\_ are a way of encoding structured data in an efficient yet extensible format.

a) Thrift

b) Protocol Buffers

c) Avro

d) None of the mentioned

7. Thrift resolves possible conflicts through \_\_\_\_\_\_\_\_\_ of the field.

a) Name

b) Static number

c) UID

d) None of the mentioned

8. Avro is said to be the future \_\_\_\_\_\_\_ layer of Hadoop.

a) RMC

b) RPC

c) RDC

d) All of the mentioned

9. When using reflection to automatically build our schemas without code generation, we need to configure Avro using :

a) AvroJob.Reflect(jConf);

b) AvroJob.setReflect(jConf);

c) Job.setReflect(jConf);

d) None of the mentioned

10. We can declare the schema of our data either in a \_\_\_\_\_\_ file.

a) JSON

b) XML

c) SQL

d) R

ANSWERS:

1.Answer: a

Explanation: JSON implementation facilitates implementation in languages that already have JSON libraries.

2.Answer: a

Explanation: Avro differs from these systems in the fundamental aspects like untagged data.

3.Answer: b

Explanation: Avro does not require that code be generated.

4.Answer: c

Explanation: Avro is optimized to minimize the disk space needed by our data and it is flexible.

5.Answer: d

Explanation: Code generation is not required to read or write data files nor to use or implement RPC protocols in Avro.

6.Answer: b

Explanation: Google uses Protocol Buffers for almost all of its internal RPC protocols and file formats.

7.Answer: b

Explanation: Avro resolves possible conflicts through the name of the field.

8.Answer: b

Explanation: When Avro is used in RPC, the client and server exchange schemas in the connection handshake.

9.Answer: c

Explanation: For strongly typed languages like Java, it also provides a generation code layer, including RPC services code generation.

10.Answer: c

Explanation: Schema can be declared using an IDL or simply through Java beans by using reflection-based schema building.

***This set of Interview Questions and Answers focuses on “Avro-2”***

1. Which of the following is a primitive data type in Avro ?

a) null

b) boolean

c) float

d) all of the mentioned

2. Point out the correct statement :

a) Records use the type name “record” and support three attributes

b) Enum are represented using JSON arrays

c) Avro data is always serialized with its schema

d) All of the mentioned

3. Avro supports \_\_\_\_\_\_ kinds of complex types.

a) 3

b) 4

c) 6

d) 7

4.\_\_\_\_\_\_\_\_ are encoded as a series of blocks.

a) Arrays

b) Enum

c) Unions

d) Maps

5. Point out the wrong statement :

a) Record, enums and fixed are named types

b) Unions may immediately contain other unions

c) A namespace is a dot-separated sequence of such names

d) All of the mentioned

6. \_\_\_\_\_\_\_\_ instances are encoded using the number of bytes declared in the schema.

a) Fixed

b) Enum

c) Unions

d) Maps

7. \_\_\_\_\_\_\_\_ permits data written by one system to be efficiently sorted by another system.

a) Complex Data type

b) Order

c) Sort Order

d) All of the mentioned

8. \_\_\_\_\_\_\_\_\_\_\_\_\_ are used between blocks to permit efficient splitting of files for MapReduce processing.

a) Codec

b) Data Marker

c) Syncronization markers

d) All of the mentioned

9. The \_\_\_\_\_\_\_\_\_\_ codec uses Google’s Snappy compression library.

a) null

b) snappy

c) deflate

d) none of the mentioned

10. Avro messages are framed as a list of \_\_\_\_\_\_\_\_\_

a) buffers

b) frames

c) rows

d) none of the mentioned

ANSWERS:

1.Answer: d

Explanation: Primitive type names are also defined type names.

2.Answer: a

Explanation: A record is encoded by encoding the values of its fields in the order that they are declared.

3.Answer: d

Explanation: Avro supports six kinds of complex types: records, enums, arrays, maps, unions and fixed.

4.Answer: a

Explanation: Each block of array consists of a long count value, followed by that many array items. A block with count zero indicates the end of the array. Each item is encoded per the array’s item schema.

5.Answer: b

Explanation: Unions may not immediately contain other unions.

6.Answer: a

Explanation: Except for unions, the JSON encoding is the same as is used to encode field default values.

7.Answer: c

Explanation: Avro binary-encoded data can be efficiently ordered without deserializing it to objects.

8.Answer: c

Explanation: Avro includes a simple object container file format.

9.Answer: b

Explanation: Snappy is a compression library developed at Google, and, like many technologies that come from Google, Snappy was designed to be fast.

10.Answer: b

Explanation: Framing is a layer between messages and the transport. It exists to optimize certain operations.

***This set of Hadoop Interview Questions & Answers focuses on “MapReduce Development”.***

1. \_\_\_\_\_\_\_\_ job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner.

a) Tasker

b) MapReduce

c) Tasktrack

d) None of the mentioned

2. Point out the correct statement :

a) Another limitation of the Hadoop MapReduce framework is its pull-based scheduling model

b) The MapReduce framework sorts the outputs of the maps, which are then input to the reduce tasks

c) The MapReduce framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks

d) All of the mentioned

3. Hadoop \_\_\_\_\_\_\_\_\_\_ is a utility which allows users to create and run jobs with any executables

a) Streaming

b) Pipes

c) Orchestration

d) All of the mentioned

4. Hadoop \_\_\_\_\_\_\_\_\_ is a SWIG- compatible C++ API to implement MapReduce applications

a) Streaming

b) Pipes

c) Orchestration

d) All of the mentioned

5. Point out the wrong statement :

a) MapReduce configuration allows the framework to effectively schedule tasks on the nodes where data is already present

b) Typically the compute nodes and the storage nodes are different

c) The MapReduce framework consists of a single master JobTracker and one slave TaskTracker per cluster-node

d) None of the mentioned

6. The key and value classes have to be \_\_\_\_\_\_\_\_\_ by the Mapreduce framework.

a) collected

b) serializable

c) compacted

d) none of the mentioned

7. Key classes have to implement the \_\_\_\_\_\_\_\_\_\_ interface to facilitate sorting by the framework.

a) Writable

b) Comparable

c) WritableComparable

d) None of the mentioned

8. The \_\_\_\_\_\_\_\_ option allows applications to add jars to the classpaths of the maps and reduces.

a) optionname

b) -libjars

c) -archives

d) all of the mentioned

9. The option \_\_\_\_\_\_\_\_\_\_\_ allows to pass comma separated list of archives as arguments.

a) optionname

b) -libjars

c) -archives

d) none of the mentioned

10. Users can specify a different symbolic name for files and archives passed through -files and -archives option, using :

a) $

b) @

c) #

d) $

ANSWERS:

1.Answer: b

Explanation: Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data.

2.Answer: d

Explanation: Typically both the input and the output of the job are stored in a file-system.

3.Answer: a

Explanation: Applications specify the input/output locations and supply map and reduce functions.

4.Answer: b

Explanation: The MapReduce framework operates exclusively on pairs.

5.Answer: b

Explanation: MapReduce framework and the Hadoop Distributed File System are running on the same set of nodes.

6.Answer: b

Explanation: Writable interface need to be implemented for key classes.

7.Answer: c

Explanation: Input and Output types of a MapReduce job:(input) -> map -> -> combine -> -> reduce -> (output).

8.Answer: b

Explanation: Applications can specify a comma separated list of paths which would be present in the current working directory of the task using the option -files.

9.Answer: c

Explanation: These archives are unarchived and a link with name of the archive is created in the current working directory of tasks.

10.Answer: c

Explanation: MapReduce is the primary method for non-primary-key-based querying.

***This set of Questions & Answers focuses on “Hadoop MapReduce”.***

1. The Mapper implementation processes one line at a time via \_\_\_\_\_\_\_\_\_ method.

a) map

b) reduce

c) mapper

d) reducer

2. Point out the correct statement :

a) Mapper maps input key/value pairs to a set of intermediate key/value pairs

b) Applications typically implement the Mapper and Reducer interfaces to provide the map and reduce methods

c) Mapper and Reducer interfaces form the core of the job

d) None of the mentioned

3. The Hadoop MapReduce framework spawns one map task for each \_\_\_\_\_\_\_\_\_\_ generated by the InputFormat for the job.

a) OutputSplit

b) InputSplit

c) InputSplitStream

d) All of the mentioned

4. Users can control which keys (and hence records) go to which Reducer by implementing a custom

a) Partitioner

b) OutputSplit

c) Reporter

d) All of the mentioned

5. Point out the wrong statement :

a) The Mapper outputs are sorted and then partitioned per Reducer

b) The total number of partitions is the same as the number of reduce tasks for the job

c) The intermediate, sorted outputs are always stored in a simple (key-len, key, value-len, value) format

d) None of the mentioned

6. Applications can use the \_\_\_\_\_\_\_\_\_\_\_\_ to report progress and set application-level status messages

a) Partitioner

b) OutputSplit

c) Reporter

d) All of the mentioned

7. The right level of parallelism for maps seems to be around \_\_\_\_\_\_\_\_\_ maps per-node

a) 1-10

b) 10-100

c) 100-150

d) 150-200

8. The number of reduces for the job is set by the user via :

a) JobConf.setNumTasks(int)

b) JobConf.setNumReduceTasks(int)

c) JobConf.setNumMapTasks(int)

d) All of the mentioned

9. The framework groups Reducer inputs by key in \_\_\_\_\_\_\_\_\_ stage.

a) sort

b) shuffle

c) reduce

d) none of the mentioned

10. The output of the reduce task is typically written to the FileSystem via \_\_\_\_\_\_\_\_\_\_\_\_\_

a) OutputCollector.collect

b) OutputCollector.get

c) OutputCollector.receive

d) OutputCollector.put

ANSWERS:

1.Answer: a

Explanation: The Mapper outputs are sorted and then partitioned per Reducer.

2.Answer: d

Explanation: The transformed intermediate records do not need to be of the same type as the input records.

3.Answer: b

Explanation: Mapper implementations are passed the JobConf for the job via the JobConfigurable.configure(JobConf) method and override it to initialize themselves.

4.Answer: a

Explanation: Users can control the grouping by specifying a Comparator via JobConf.setOutputKeyComparatorClass(Class).

5.Answer: d

Explanation: All intermediate values associated with a given output key are subsequently grouped by the framework, and passed to the Reducer(s) to determine the final output.

6.Answer: c

Explanation: Reporter is also used to update Counters, or just indicate that they are alive.

7.Answer: b

Explanation: Task setup takes a while, so it is best if the maps take at least a minute to execute.

8.Answer: b

Explanation: Reducer has 3 primary phases: shuffle, sort and reduce.

9.Answer: a

Explanation: The shuffle and sort phases occur simultaneously; while map-outputs are being fetched they are merged.

10.Answer: a

Explanation: The output of the Reducer is not sorted.

***This set of Hadoop Questions & Answers for freshers focuses on “MapReduce Features”.***

1. Which of the following is the default Partitioner for Mapreduce ?

a) MergePartitioner

b) HashedPartitioner

c) HashPartitioner

d) None of the mentioned

2. Point out the correct statement :

a) The right number of reduces seems to be 0.95 or 1.75

b) Increasing the number of reduces increases the framework overhead

c) With 0.95 all of the reduces can launch immediately and start transferring map outputs as the maps finish

d) All of the mentioned

3. Which of the following partitions the key space ?

a) Partitioner

b) Compactor

c) Collector

d) All of the mentioned

4. \_\_\_\_\_\_\_\_\_\_\_\_ is a generalization of the facility provided by the MapReduce framework to collect data output by the Mapper or the Reducer

a) OutputCompactor

b) OutputCollector

c) InputCollector

d) All of the mentioned

5. Point out the wrong statement :

a) It is legal to set the number of reduce-tasks to zero if no reduction is desired

b) The outputs of the map-tasks go directly to the FileSystem

c) The Mapreduce framework does not sort the map-outputs before writing them out to the FileSystem

d) None of the mentioned

6. \_\_\_\_\_\_\_\_\_\_ is the primary interface for a user to describe a MapReduce job to the Hadoop framework for execution.

a) JobConfig

b) JobConf

c) JobConfiguration

d) All of the mentioned

7. The \_\_\_\_\_\_\_\_\_\_\_ executes the Mapper/ Reducer task as a child process in a separate jvm.

a) JobTracker

b) TaskTracker

c) TaskScheduler

d) None of the mentioned

8. Maximum virtual memory of the launched child-task is specified using :

a) mapv

b) mapred

c) mapvim

d) All of the mentioned

9. Which of the following parameter is the threshold for the accounting and serialization buffers ?

a) io.sort.spill.percent

b) io.sort.record.percent

c) io.sort.mb

d) None of the mentioned

10. \_\_\_\_\_\_\_\_\_\_\_\_\_\_ is percentage of memory relative to the maximum heapsize in which map outputs may be retained during the reduce.

a) mapred.job.shuffle.merge.percent

b) mapred.job.reduce.input.buffer.percen

c) mapred.inmem.merge.threshold

d) io.sort.factor

ANSWERS:

1.Answer: c

Explanation: The total number of partitions is the same as the number of reduce tasks for the job.