**Hadoop Questions and Answers – The ZooKeeper Service-2**

This set of Hadoop Multiple Choice Questions and Answers focuses on “ZooKeeper Service”.

1. You need to have \_\_\_\_\_\_\_\_\_ installed before running ZooKeeper.  
**a) Java**  
b) C  
c) C++  
d) SQLGUI  
**View Answer**

**Answer: a  
Explanation: Client bindings are available in several other languages.**

2. Point out the correct statement :  
**a) All znodes are ephemeral, which means they are describing a “temporary” state**  
b) /hbase/replication/state contains the list of RegionServers in the main cluster  
c) Offline snapshots are coordinated by the Master using ZooKeeper to communicate with the RegionServers using a two-phase-commit-like transaction  
d) None of the mentioned  
View Answer

Answer: a  
Explanation: Although the Replication znodes do not describe a temporary state, they are meant to be the source of truth for the replication state, describing the replication state of each machine.

3. How many types of special znodes are present in Zookeeper ?  
a) 1  
**b) 2**  
c) 3  
d) All of the mentioned  
**View Answer**

**Answer: b  
Explanation: There are two special types of znode: sequential and ephemeral.**

4. To register a “watch” on a znode’s data, you need to use the \_\_\_\_\_\_\_ commands to access the current content or metadata  
**a) stat**  
b) put  
c) receive  
d) gets  
View Answer

Answer: a  
Explanation: ZooKeeper can also notify you of changes in a znode’s content or changes in a znode’s children.

5. Point out the wrong statement :  
a) All the znodes are prefixed using the default /hbase location  
b) ZooKeeper provides an interactive shell that allows you to explore the ZooKeeper state  
c) The znodes that you’ll most often see are the ones that coordinate operations like Region Assignment  
**d) All of the mentioned**  
View Answer

**Answer: d  
Explanation: The HBase root znode path is configurable using hbase-site.xml, and by default the location is “/hbase”.**

6. \_\_\_\_\_\_\_ has a design policy of using ZooKeeper only for transient data  
a) Hive  
b) Imphala  
**c) Hbase**  
d) Oozie  
View Answer

7. Zookeeper keep track of the cluster state such as the \_\_\_\_\_\_ table location.  
a) DOMAIN  
b) NODE  
**c) ROOT**  
d) All of the mentioned  
View Answer

Answer: c  
Explanation: Zookeeper keeps track of list of online RegionServers,unassigned Regions.

8. The \_\_\_\_\_\_\_\_ master will register its own address in this znode at startup, making this znode the source of truth for identifying which server is the Master.  
**a) active**  
b) passive  
c) region  
d) all of the mentioned  
View Answer

Answer: a **Explanation: Each inactive Master will register itself as backup Master by creating a sub-znode.**

9. \_\_\_\_\_\_\_\_\_\_\_ is used to decommission more than one RegionServer at a time by creating sub-znodes.  
a) /hbase/master  
**b) /hbase/draining**  
c) /hbase/passive  
d) None of the mentioned  
**View Answer**

**Answer: b  
Explanation: /hbase/draining lets you decommission multiple RegionServers without having the risk of regions temporarily moved to a RegionServer that will be decommissioned later.**

10. The \_\_\_\_\_\_ znode is used for synchronizing the changes made to the \_acl\_ table by the grant/revoke commands.  
a) zcl  
**b) acl**  
c) scl  
d) bnl  
View Answer

Answer: b  
Explanation: Each table will have a sub-znode (/hbase/acl/tableName) containing the ACLs of the

Tabl

**Hadoop Questions and Answers – The ZooKeeper Service-1**

**This set of Interview Questions & Answers focuses on “Zookeeper”.**

1. Zookeeper essentially mirrors the \_\_\_\_\_\_\_ functionality exposed in the Linux kernel.  
a) iread  
**b) inotify**  
c) iwrite  
d) icount  
View Answer

Answer: b  
Explanation: A client can request for Zookeeper to generate the node name to avoid collisions.

2. Point out the correct statement :  
**a) If we are to deploy Zookeeper in a distributed environment, we have to think about both the availability and scalability of the service**  
b) Chubby and Hbase are both much more than a distributed lock service  
c) Given a cluster of Zookeeper servers, many can act as leader  
d) None of the mentioned  
View Answer

Answer: a  
Explanation: At its core, Zookeeper is modeled after a straightforward, tree based, file system API.

3. A \_\_\_\_\_\_\_\_\_\_\_ server is a machine that keeps a copy of the state of the entire system and persists this information in local log files.  
a) Master  
b) Region  
**c) Zookeeper**  
d) All of the mentioned  
**View Answer**

**Answer: c  
Explanation: A very large Hadoop cluster can be supported by multiple ZooKeeper servers.**

4. ZooKeeper’s architecture supports high \_\_\_\_\_\_\_\_\_\_\_\_ through redundant services.  
a) flexibility  
b) scalability  
**c) availability**  
d) interactivity  
**View Answer**

**Answer: c  
Explanation: The clients can thus ask another ZooKeeper master if the first fails to answer.**

5. Point out the wrong statement :  
a) Cluster-wide status centralization service is essential for management and serialization tasks across a large distributed set of servers  
b) Within ZooKeeper, an application can create what is called a znode  
c) The znode can be updated by any node in the cluster, and any node in the cluster can register to be informed of changes to that znode  
**d) None of the mentioned**  
View Answer

Answer: d  
Explanation: ZooKeeper provides an infrastructure for cross-node synchronization and can be used by applications to ensure that tasks across the cluster are serialized or synchronized.

6. \_\_\_\_\_\_\_\_\_ serves distributed Lucene indexes in a grid environment.  
**a) Katta**  
b) Helprace  
c) Neo4j  
d) 101tec  
**View Answer**

**Answer: a  
Explanation: Zookeeper is used for node, master and index management in the grid.**

7. The Email & Apps team of \_\_\_\_\_\_\_\_\_\_\_ uses ZooKeeper to coordinate sharding and responsibility changes in a distributed e-mail client  
a) Katta  
b) Helprace  
**c) Rackspace**  
d) None of the mentioned  
View Answer

**Answer: c  
Explanation: ZooKeeper also provides distributed locking for connections to prevent a cluster from overwhelming servers.**

8. ZooKeeper is used for configuration, leader election in Cloud edition of :  
**a) Solr**  
b) Solur  
c) Solar101  
d) All of the mentioned  
View Answer

**Answer: a  
Explanation: ZooKeeper is used for internal application development with Solr and Hadoop with Hbase.**

9. Helprace is using ZooKeeper on a \_\_\_\_\_\_\_ cluster in conjunction with Hadoop and HBase.  
**a) 3-node**  
b) 4-node  
c) 5-node  
d) 6-node  
View Answer

Answer: a  
**Explanation: Zookeeper is used to manage a system build out of hadoop, katta, oracle batch jobs and a web component.**

10. The ZooKeeper Data Directory contains files which are \_\_\_\_\_\_\_\_\_ copy of the znodes stored by a particular serving ensemble.  
a) transient  
b) read only  
**c) persistent**  
d) none of the mentioned  
View Answer

**Answer: c  
Explanation: As changes are made to the znodes these changes are appended to a transaction log, occasionally, when a log grows large, a snapshot of the current state of all znodes will be written to the filesystem**.

e.

**Hadoop Questions and Answers – Introduction to Zookeeper**

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

1. ZooKeeper itself is intended to be replicated over a sets of hosts called :  
a) chunks  
**b) ensemble**  
c) subdomains  
d) none of the mentioned  
View Answer

**Answer: b  
Explanation: As long as a majority of the servers are available, the ZooKeeper service will be available**.

2. Point out the correct statement :  
a) ZooKeeper can achieve high throughput and high latency numbers.  
b) The fault tolerant ordering means that sophisticated synchronization primitives can be implemented at the client  
**c) The ZooKeeper implementation puts a premium on high performance, highly available, strictly ordered access**  
d) All of the mentioned  
View Answer

Answer: c  
Explanation: The performance aspects of ZooKeeper means it can be used in large, distributed systems.

3. Which of the guarantee is provided by Zookeeper ?  
a) Interactivity  
b) Flexibility  
c) Scalability  
**d) Reliability**  
View Answer

**Answer: d  
Explanation: Once an update has been applied, it will persist from that time forward until a client overwrites the update.**

4. ZooKeeper is especially fast in \_\_\_\_\_\_\_\_\_\_\_ workloads  
a) write  
**b) read-dominant**  
c) read-write  
d) none of the mentioned  
**View Answer**

**Answer: b  
Explanation: ZooKeeper applications run on thousands of machines, and it performs best where reads are more common than writes, at ratios of around 10:1.**

5. Point out the wrong statement :  
a) Distributed applications use SQL to store important configuration information  
b) The service maintains a record of all transactions, which can be used for higher-level abstractions, like synchronization primitives  
c) ZooKeeper maintain a standard hierarchical name space, similar to files and directories  
**d) ZooKeeper provides superior reliability through redundant services  
View Answer**

**Answer: d  
Explanation: Distributed applications use Zookeeper to store and mediate updates to important configuration information.**

6. When a \_\_\_\_\_\_\_ is triggered the client receives a packet saying that the znode has changed.  
a) event  
**b) watch**  
c) row  
d) value  
View Answer

Answer: b **Explanation: ZooKeeper supports the concept of watches. Clients can set a watch on a znodes.**

7. The underlying client-server protocol has changed in version \_\_\_\_\_\_\_ of ZooKeeper.  
a) 2.0.0  
**b) 3.0.0**  
c) 4.0.0  
d) 6.0.0  
View Answer

Answer: b  
Explanation: Old pre-3.0.0 clients are not guaranteed to operate against upgraded 3.0.0 servers and vice-versa.

8. The java package structure has changed from com.yahoo.zookeeper\* to :  
a) apache.zookeeper  
**b) org.apache.zookeeper**  
c) org.apache.zookeeper.package  
d) all of the mentioned  
**View Answer**

**Answer: b  
Explanation: A number of constants used in the client ZooKeeper API were re-specified using enums (rather than ints).**

9. A number of constants used in the client ZooKeeper API were renamed in order to reduce \_\_\_\_\_\_\_\_ collision  
a) value  
**b) namespace**  
c) counter  
d) none of the mentioned  
View Answer

Answer: b  
Explanation: ZOOKEEPER-18 removed KeeperStateChanged, use KeeperStateDisconnected instead.

10. ZooKeeper allows distributed processes to coordinate with each other through registers, known as :  
**a) znodes**b) hnodes  
c) vnodes  
d) rnodes  
View Answer

Answer: a  
Explanation: Every znode is identified by a path, with path elements separated by a slash.

**Hadoop Questions and Answers – Compression**

**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  
View Answer

Answer: b  
Explanation: Snappy has fast compression and decompression speeds.

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  
View Answer

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. Which of the following compression is similar to Snappy compression ?  
**a) LZO**  
b) Bzip2  
c) Gzip  
d) All of the mentioned  
View Answer

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

4. Which of the following supports splittable compression ?  
**a) LZO(**Lempel–Ziv–Oberhumer)  
b) Bzip2  
c) Gzip  
d) All of the mentioned  
View Answer

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

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  
View Answer

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

6. Which of the following is the slowest compression technique ?  
a) LZO  
**b) Bzip2**  
c) Gzip  
d) All of the mentioned  
View Answer

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

7. Gzip (short for GNU zip) generates compressed files that have a \_\_\_\_\_\_\_\_\_ extension.  
a) .gzip  
**b) .gz**  
c) .gzp  
d) .g  
View Answer

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

8. Which of the following is based on the DEFLATE algorithm ?  
a) LZO  
b) Bzip2  
**c) Gzip**  
d) All of the mentioned  
View Answer

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

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  
View Answer

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

10. The LZO compression format is composed of approximately \_\_\_\_\_\_\_\_\_\_ blocks of compressed data.  
a) 128k  
b) 256k  
c) 24k  
d) 36k  
View Answer

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.

**Hadoop Questions and Answers – Serialization**

**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**  
View Answer

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

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**  
View Answer

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. Avro schemas describe the format of the message and are defined using :  
**a) JSON**  
b) XML  
c) JS  
d) All of the mentioned  
View Answer

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

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  
View Answer

Answer: b  
Explanation: DatumReader reads the content through the DataFileReader implementation.

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**  
View Answer

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. The \_\_\_\_\_\_\_\_\_\_\_\_ class extends and implements several Hadoop-supplied interfaces.  
a) AvroReducer  
b) Mapper  
**c) AvroMapper**  
d) None of the mentioned  
View Answer

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

7. \_\_\_\_\_\_\_\_\_\_\_\_ class accepts the values that the ModelCountMapper object has collected.  
**a) AvroReducer**  
b) Mapper  
c) AvroMapper  
d) None of the mentioned  
View Answer

Answer: a  
Explanation: AvroReducer summarizes them by looping through the values.

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  
View Answer

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

9. Which of the following works well with Avro ?  
a) Lucene  
b) kafka  
**c) MapReduce**  
d) None of the mentioned  
View Answer

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

10. \_\_\_\_\_\_\_\_\_\_ tools is used to generate proxy objects in Java to easily work with the objects.  
a) Lucene  
b) kafka  
c) MapReduce  
**d) Avro**  
**View Answer**

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

**Hadoop Questions and Answers – Data Integrity**

**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  
View Answer

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. 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  
View Answer

**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. The \_\_\_\_\_\_\_\_\_\_\_ machine is a single point of failure for an HDFS cluster.  
a) DataNode  
**b) NameNode**  
c) ActionNode  
d) All of the mentioned  
View Answer

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. The \_\_\_\_\_\_\_\_\_\_\_\_ and the EditLog are central data structures of HDFS.  
a) DsImage  
**b) FsImage**  
c) FsImages  
d) All of the mentioned  
View Answer

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

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  
View Answer

Answer: a  
Explanation: HDFS is designed to support very large files.

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  
View Answer

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. Automatic restart and \_\_\_\_\_\_\_\_\_\_\_\_ of the NameNode software to another machine is not supported.  
**a) failover**  
b) end  
c) scalability  
d) all of the mentioned  
**View Answer**

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

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  
View Answer

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. \_\_\_\_\_\_\_\_\_ 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  
View Answer

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

10. The HDFS file system is temporarily unavailable whenever the HDFS \_\_\_\_\_\_\_\_ is down.  
a) DataNode  
**b) NameNode**  
c) ActionNode  
d) None of the mentioned  
View Answer

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

**Hadoop Questions and Answers – Avro-1**

**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  
View Answer

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

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  
View Answer

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

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  
View Answer

Answer: b  
Explanation: Avro does not require that code be generated.

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  
**View Answer**

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

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**  
**View Answer**

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

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  
View Answer

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

7. Thrift resolves possible conflicts through \_\_\_\_\_\_\_\_\_ of the field.  
a) Name  
**b) Static number**  
c) UID  
d) None of the mentioned  
**View Answer**

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

8. Avro is said to be the future \_\_\_\_\_\_\_ layer of Hadoop.  
a) RMC  
**b) RPC**(**Remote Procedural Call)**  
c) RDC  
d) All of the mentioned  
View Answer

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

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  
**View Answer**

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

10. We can declare the schema of our data either in a \_\_\_\_\_\_ file.  
a) JSON  
b) XML  
**c) SQL**  
d) R  
View Answer

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

**Hadoop Questions and Answers – Avro-2**

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

1. Which of the following is a primitive data type in Avro ?  
a) null  
b) boolean  
c) float  
d) all of the mentioned  
View Answer

Answer: d  
Explanation: Primitive type names are also defined type names.

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  
View Answer

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

3. Avro supports \_\_\_\_\_\_ kinds of complex types.  
a) 3  
b) 4  
c) 6  
**d) 7**  
View Answer

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

4.\_\_\_\_\_\_\_\_ are encoded as a series of blocks.  
**a) Arrays**  
b) Enum  
c) Unions  
d) Maps  
View Answer

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. 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  
View Answer

Answer: b  
Explanation: Unions may not immediately contain other unions.

6. \_\_\_\_\_\_\_\_ instances are encoded using the number of bytes declared in the schema.  
**a) Fixed**  
b) Enum  
c) Unions  
d) Maps  
View Answer

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

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  
View Answer

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

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  
View Answer

Answer: c  
Explanation: Avro includes a simple object container file format.

9. The \_\_\_\_\_\_\_\_\_\_ codec uses Google’s Snappy compression library.  
a) null  
**b) snappy**  
c) deflate  
d) none of the mentioned  
View Answer

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. Avro messages are framed as a list of \_\_\_\_\_\_\_\_\_  
a) buffers  
**b) frames**  
c) rows  
d) none of the mentioned  
**View Answer**

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

**Hadoop Questions and Answers – Sqoop with Hadoop-1**

**This set of Hadoop Questions & Answers for online test focuses on “Sqoop with Hadoop”.**

1. Which of the following interface is implemented by Sqoop for recording ?  
a) SqoopWrite  
**b) SqoopRecord**  
c) SqoopRead  
d) None of the mentioned  
View Answer

Answer: b  
Explanation: Class SqoopRecord is interface implemented by the classes generated by sqoop’s orm.ClassWriter.

2. Point out the correct statement :  
a) Interface FieldMapping is used for mapping of field  
**b) Interface FieldMappable is used for mapping of field**  
c) Sqoop is nothing but NoSQL to Hadoop  
d) Sqoop internally uses ODBC interface so it should work with any JDBC compatible database  
View Answer

Answer: b  
Explanation: FieldMappable Interface describes a class capable of returning a map of the fields of the object to their values.

3. Sqoop is an open source tool written at \_\_\_\_\_\_\_\_  
a) Cloudera  
b) IBM  
**c) Microsoft**  
d) All of the mentioned  
View Answer

Answer: c  
Explanation: Sqoop allows users to import data from their relational databases into HDFS and vice versa.

4. Sqoop uses \_\_\_\_\_\_\_\_\_ to fetch data from RDBMS and stores that on HDFS.  
a) Hive  
**b) Map reduce**  
c) Imphala  
d) BigTOP  
View Answer

Answer: b  
Explanation: While fetching , it throttles the number of mappers accessing data on RDBMS to avoid DDoS.

5. Point out the wrong statement :  
a) Sqoop is used to import complete database  
b) Sqoop is used to import selected columns from a particular table  
c) Sqoop is used to import selected tables  
**d) All of the mentioned**  
View Answer

Answer: d  
Explanation: Apache Sqoop is tool which allows users to import data from relational databases to HDFS and export data from HDFS to relational database.

6. \_\_\_\_\_\_\_\_\_ allows users to specify the target location inside of Hadoop.  
a) Imphala  
b) Oozie  
**c) Sqoop**  
d) Hive  
View Answer

Answer: c  
Explanation: Sqoop is a connectivity tool for moving data from non-Hadoop data stores – such as relational databases and data warehouses – into Hadoop.

7. Microsoft uses a Sqoop-based connector to help transfer data from \_\_\_\_\_\_\_\_\_ databases to Hadoop  
a) PostreSQL  
**b) SQL Server**  
c) Oracle  
d) MySQL  
View Answer

Answer: b  
Explanation: Sqoop is a command-line interface application for transferring data between relational databases and Hadoop.

8. \_\_\_\_\_\_\_\_\_\_ provides a Couchbase Server-Hadoop connector by means of Sqoop.  
**a) MemCache**  
b) Couchbase  
c) Hbase  
d) All of the mentioned  
View Answer

Answer: a  
Explanation: Exports can be used to put data from Hadoop into a relational database.

9. Sqoop’s direct mode does not support imports of \_\_\_\_\_\_ columns  
a) BLOB  
b) LONGVARBINARY  
c) CLOB  
**d) All of the mentioned**  
View Answer

**Answer: d  
Explanation: Use JDBC-based imports for these columns; do not supply the –direct argument to the import tool.**

10. Sqoop has been tested with Oracle \_\_\_\_\_\_ Express Edition.  
a) 11.2.0  
**b) 10.2.0**  
c) 12.2.0  
d) 10.3.0  
View Answer

Answer: b  
Explanation: Oracle is notable in its different approach to SQL from the ANSI standard, and its non-standard JDBC driver. Therefore, several features work differently.

**Hadoop Questions and Answers – Sqoop with Hadoop-2**

**This set of Interview Questions & Answers focuses on “Sqoop”.**

1. \_\_\_\_\_\_\_\_\_ tool can list all the available database schemas.  
a) sqoop-list-tables  
**b) sqoop-list-databases**c) sqoop-list-schema  
d) sqoop-list-columns  
**View Answer**

**Answer: b  
Explanation: Sqoop also includes a primitive SQL execution shell (the sqoop-eval tool).**

2. Point out the correct statement :  
**a) The sqoop command-line program is a wrapper which runs the bin/hadoop script shipped with Hadoop**  
b) If $HADOOP\_HOME is set, Sqoop will use the default installation location for Cloudera’s Distribution for Hadoop  
c) The active Hadoop configuration is loaded from $HADOOP\_HOME/conf/, unless the $HADOOP\_CONF\_DIR environment variable is unset  
d) None of the mentioned  
View Answer

Answer: a  
Explanation: If you have multiple installations of Hadoop present on your machine, you can select the Hadoop installation by setting the $HADOOP\_HOME environment variable.

3. Data can be imported in maximum \_\_\_\_\_\_ file formats.  
a) 1  
**b) 2**  
c) 3  
d) All of the mentioned  
View Answer

Answer: b  
Explanation: You can import data in one of two file formats: delimited text or SequenceFiles.

4. \_\_\_\_\_\_\_\_ text is appropriate for most non-binary data types.  
a) Character  
b) Binary  
**c) Delimited**  
d) None of the mentioned  
View Answer

Answer: c  
Explanation: Delimited text is the default import format.

5. Point out the wrong statement :  
a) Avro data files are a compact, efficient binary format that provides interoperability with applications written in other programming languages  
**b) By default, data is is compressed while importing**  
c) Delimited text also readily supports further manipulation by other tools, such as Hive  
d) None of the mentioned  
View Answer

**Answer: b  
Explanation: You can compress your data by using the deflate (gzip) algorithm with the -z or –compress argument, or specify any Hadoop compression codec using the –compression-codec argument.**

6. If you set the inline LOB limit to \_\_\_\_\_\_\_\_ all large objects will be placed in external storage.  
**a) 0**  
b) 1  
c) 2  
d) 3  
View Answer

Answer: a  
Explanation: The size at which lobs spill into separate files is controlled by the –inline-lob-limit argument, which takes a parameter specifying the largest lob size to keep inline, in bytes.

7. \_\_\_\_\_\_\_\_ does not support the notion of enclosing characters that may include field delimiters in the enclosed string.  
a) Imphala  
b) Oozie  
c) Sqoop  
**d) Hive**  
View Answer

Answer: d  
Explanation: Even though Hive supports escaping characters, it does not handle escaping of new-line character.

8. Sqoop can also import the data into Hive by generating and executing a \_\_\_\_\_\_\_\_\_\_\_\_ statement to define the data’s layout in Hive.  
a) SET TABLE  
**b) CREATE TABLE**  
c) INSERT TABLE  
d) All of the mentioned  
View Answer

Answer: b  
Explanation: Importing data into Hive is as simple as adding the –hive-import option to your Sqoop command line.

9. The \_\_\_\_\_\_\_\_\_\_ tool imports a set of tables from an RDBMS to HDFS.  
a) export-all-tables  
b) import-all-tables  
**c) import-tables**  
d) none of the mentioned  
View Answer

Answer: c  
Explanation: Data from each table is stored in a separate directory in HDFS.

10. Which of the following argument is not supported by import-all-tables tool ?  
**a) –class-name**  
b) –package-name  
c) –database-name  
d) –table-name  
**View Answer**

**Answer: a  
Explanation: You may, however, specify a package with –package-name in which all generated classes will be placed.**

**Hadoop Questions and Answers – Introduction to Sqoop**

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

1. BigDecimal is comprised of a \_\_\_\_\_\_\_\_ with an integer ‘scale’ field.  
a) BigInt  
**b) BigInteger**  
c) MediumInt  
d) SmallInt  
View Answer

Answer: b  
Explanation: The BigDecimal/BigInteger can also return itself as a ‘long’ value.

2. Point out the correct statement :  
a) BooleanSerializer is used to parse string representations of boolean values into boolean scalar types  
**b) BlobRef is a wrapper that holds a BLOB either directly**  
c) BooleanParse is used to parse string representations of boolean values into boolean scalar types  
d) All of the mentioned  
View Answer

Answer: b  
Explanation: BlobRef is used for reference to a file that holds the BLOB data.

3. ClobRef is a wrapper that holds a CLOB either directly, or a reference to a file that holds the \_\_\_\_\_\_ data.  
**a) CLOB**  
b) BLOB  
c) MLOB  
d) All of the mentioned  
View Answer

Answer: a  
Explanation: Create a ClobRef based on parsed data from a line of text.

4. \_\_\_\_\_\_\_\_\_\_ encapsulates a set of delimiters used to encode a record.  
a) LargeObjectLoader  
b) FieldMapProcessor  
**c) DelimiterSet**  
d) LobSerializer  
**View Answer**

**Answer: c  
Explanation: Delimiter set is created with the specified delimiters.**

5. Point out the wrong statement :  
a) Abstract base class that holds a reference to a Blob or a Clob  
b) ACCESSORTYPE is the type used to access this data in a streaming fashion  
c) CONTAINERTYPE is the type used to hold this data (e.g., BytesWritable)  
**d) None of the mentioned**  
**View Answer**

**Answer: d  
Explanation: DATATYPE is the type being held (e.g., a byte array).**

6. \_\_\_\_\_\_\_\_\_ supports null values for all types.  
a) SmallObjectLoader  
b) FieldMapProcessor  
c) DelimiterSet  
**d) JdbcWritableBridge**  
View Answer

Answer: d  
Explanation: JdbcWritableBridge class contains a set of methods which can read db columns from a ResultSet into Java types.

7. Which of the following is a singleton instance class ?  
**a) LargeObjectLoader**  
b) FieldMapProcessor  
c) DelimiterSet  
d) LobSerializer  
**View Answer**

**Answer: a  
Explanation: Lifetime is limited to the current TaskInputOutputContext’s life.**

8. Which of the following class is used for general processing of error ?  
a) LargeObjectLoader  
**b) ProcessingException**  
c) DelimiterSet  
d) LobSerializer  
View Answer

Answer: b  
Explanation: General error occurs during processing of a SqoopRecord.

9. Records are terminated by a \_\_\_\_\_\_\_\_\_\_ character.  
**a) RECORD\_DELIMITER**  
b) FIELD\_DELIMITER  
c) FIELD\_LIMITER  
d) None of the mentioned  
**View Answer**

**Answer: a  
Explanation: Class RecordParser parses a record containing one or more fields**.

10.The fields parsed by \_\_\_\_\_\_\_\_\_\_\_\_ are backed by an internal buffer.  
a) LargeObjectLoader  
b) ProcessingException  
**c) RecordParser**  
d) None of the Mentioned  
View Answer

Answer: c  
Explanation: Multiple threads must use separate instances of RecordParser.