This exam serves as the assessment for those students who cannot utilize the Hadoop system and/or Ambari GUI.

1. Describe the process of MapReduce.

Answer :

1. -

**Map-Reduce** is a programming model that is mainly divided into two phases i.e. Map Phase and Reduce Phase. It is designed for processing the data in parallel which is divided on various machines(nodes). The Hadoop [Java](https://www.geeksforgeeks.org/java/)programs consist of Mapper class and Reducer class along with the driver class. Reducer is the second part of the Map-Reduce programming model. The Mapper produces the output in the form of key-value pairs which works as input for the Reducer.

But before sending this intermediate key-value pairs directly to the Reducer some process will be done which shuffle and sort the key-value pairs according to its key values, which means the value of the key is the main decisive factor for sorting. The output generated by the Reducer will be the final output which is then stored on [HDFS(Hadoop Distributed File System)](https://www.geeksforgeeks.org/introduction-to-hadoop-distributed-file-systemhdfs/). Reducer mainly performs some computation operation like addition, filtration, and aggregation. By default, the number of reducers utilized for process the output of the Mapper is 1 which is configurable and can be changed by the user according to the requirement.

1. -

**MapReduce** is a programming model or pattern within the [Hadoop](http://hadoop.apache.org/) framework that is used to access big data stored in the Hadoop File System (HDFS). It is a core component, integral to the functioning of the Hadoop framework.

MapReduce facilitates concurrent processing by splitting petabytes of data into smaller chunks, and processing them in parallel on Hadoop commodity servers. In the end, it aggregates all the data from multiple servers to return a consolidated output back to the application.

**How MapReduce Works**

At the crux of MapReduce are two functions: Map and Reduce. They are sequenced one after the other.

* The **Map** function takes input from the disk as <key,value> pairs, processes them, and produces another set of intermediate <key,value> pairs as output.
* The **Reduce** function also takes inputs as <key,value> pairs, and produces <key,value> pairs as output.

### **Map**

The input data is first split into smaller blocks. Each block is then assigned to a mapper for processing.

For example, if a file has 100 records to be processed, 100 mappers can run together to process one record each. Or maybe 50 mappers can run together to process two records each. The Hadoop framework decides how many mappers to use, based on the size of the data to be processed and the memory block available on each mapper server.

### **Reduce**

After all the mappers complete processing, the framework shuffles and sorts the results before passing them on to the reducers. A reducer cannot start while a mapper is still in progress. All the map output values that have the same key are assigned to a single reducer, which then aggregates the values for that key.

1. What is the difference between schema on read and schema on write?
   1. On read applies structure when the data is being processed, on write doesn’t
   2. On read applies structure all the time, on write doesn’t
2. True or False? “You can use SQL in Hive.”
   1. True
   2. False
3. Why do you think sqoop would be a useful tool?

Answer :

1. -

**Apache Sqoop** is designed to efficiently transfer enormous volumes of data between Apache Hadoop and structured datastores such as relational databases(RDBMS). It helps to offload certain tasks, such asETL processing, from an enterprise data warehouse to Hadoop, for efficient execution at a much lower cost. Sqoop also makes it easy to extract data from Hadoop and export it to external structured datastores.

1. -

The traditional application management system, that is, the interaction of applications with relational database using RDBMS, is one of the sources that generate Big Data. Such Big Data, generated by RDBMS, is stored in Relational **Database Servers** in the relational database structure.

When Big Data storages and analyzers such as MapReduce, Hive, HBase, Cassandra, Pig, etc. of the Hadoop ecosystem came into picture, they required a tool to interact with the relational database servers for importing and exporting the Big Data residing in them. Here, Sqoop occupies a place in the Hadoop ecosystem to provide feasible interaction between relational database server and Hadoop’s HDFS.

**Sqoop** − “SQL to Hadoop and Hadoop to SQL”

Sqoop is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases. It is provided by the Apache Software Foundation.