**Tools Used:**

**Oracle VM Virtualbox Manager** **->** used to create multiple nodes to communicate each other in the cluster. (For 8 nodes, created 8 new instances (m1,m2…m8) with 1 core, 1gb ram, 8gb harddisk)

**OS ->** Ubuntu 18.04

**HDFS ->** To load data in hdfs file system.

**Hive ->** To create relation tables for the given data file.

**Mysql db ->** Hive metastore

**Spark ->** To query the hive tables.

**Assignment Hands-on steps overview:**

**Step 1:** Using Oracle VM Virtualbox Manager, created namenodes and multiple datanodes with same configurations.

**Step2:** Installed Hadoop in all the nodes and update the datanode, namenode details in the configuration files. (core-site.xml, mapred-site.xml, hdfs-site.xml, yarn-site.xml)

**Step 3:** Updated the replication factor in hdfs-site xml file and load the data in to hdfs and verify the data loaded as per the replication factor value.

**Step 4:** To avoid the redundancy, created three new Hive tables to populate the normalized data of the given file (Coursera-1.csv). Tables are created using hive DDLs.

Hive tables: University\_Dim, Difficulty\_level\_Dim, Coursera\_Dim

**Step 5:** Verified the metadata of the table stored in metastore. (Mysql is our metastore of our Hive database)

**Step 6:** Loaded the normalized data into all the three new Hive tables using hive DMLs. Removed duplicates using ‘distinct’ keyword when inserting records from file to coursera\_dim hive table.

**Step 7:** Verified no duplicate records exists in the coursera\_dim hive table and dimension ids are properly populated.

**Step 8:** Used the ORC file format for the Hive tables. It heavily compresses the file size and also the performance of the column level manipulation/calculation is faster since the data stored in the columnar level.

**Step 9:** Verified the size of the hive data file with actual data file and found hive data file is five time lesser than the actual data size.

**Step 10:** Used Spark Sql to query the hive tables. Queried a table to retrieve data using an id column. Tried retrieving data from both hive CLI and spark sql.