**Practice questions on Cloudera Hadoop**

**SETUP:** Install Virtual Machine in your local PC and download CDH 5.x zip file 🡪 Create Cloudera quickstart VM and run it

<Right click and open terminal in order to run commands>

**Ques 1:** How to connect mySQL DB from the terminal

**Solution:** mysql -u <user\_name> -p

[Enter password]

**Ques 2:** Create and execute data ingestion job using sqoop import for retail\_db.categories table

**Solution:**

sqoop-job - - create categories\_sqoop\_import - - import - - connect jdbc:mysql://quickstart:3306/retail\_db - - table categories - - target-dir cat\_target - - fields-terminated-by ‘,’ - - lines-terminated-by ‘\n’;

Sqoop-job - - exec categories\_sqoop\_import -- -username retail\_dba -p

**Ques 3:** List all sqoop jobs

**Solution:** sqoop-job --list

**Ques 4:** Delete job created in ques 2

**Solution:** sqoop-job - -delete categories\_sqoop\_import

**Ques 5:** Command to clear the terminal

**Solution:** clear

**Ques 6:** List all files and directories in HDFS

**Solution:** hdfs dfs -ls

**Ques 7:** List content of files containing categories table data in HDFS

**Solution:** hdfs dfs -cat cat\_target/\*

**Ques 8:** Delete all files inside cat\_target directory created earlier

**Solution:** hdfs dfs -rmdir cat\_target/\*

**Ques 9:** Delete empty directory cat\_target

**Solution:** hdfs dfs -rm cat\_target

**Ques 10:** Using pyspark create an RDD from two collections and perform intersection operation

**Solution:**

pyspark - - master yarn

rdd1=sc.parallelize(range(1,10));

rdd2=sc.parallelize(range(5,20));

res=rdd1.intersection(rdd2);

res.collect();

**Ques 11:** Command to exit from pyspark shell

**Solution**: exit();

**Ques 12:** Using sqoop-import command import retail\_db.orders and retail\_db.order\_items tables into HDFS directory

**Solution:**

sqoop-import - -connect jdbc:mysql://quickstart:3306/retail\_db - -username retail\_dba - -password=’cloudera’ - - table orders - -target-dir orders

sqoop-import - -connect jdbc:mysql://quickstart:3306/retail\_db - -username retail\_dba - - table order\_items - -target-dir orders

**Ques 13:** Using pyspark find the number of orders on each date using imported data of ques 12

**Solution:**

pyspark

Orders\_rdd=sc.textFile(‘orders’);

Order\_items\_rdd=sc.textFIle(‘order\_items’);

Res=Orders\_rdd.map(lambda x:(x.split(‘,’)[0],x)).join(order\_items\_rdd.map(lambda x:(x.split(‘,’)[1],x))).map(lambda x:(x[1][0].split(‘,’)[1],x[0])).distinct().map(lambda x:(x[0],1)).reduceByKey(lambda a,b:a+b).sortByKey();