**Problem-1**

This Assignment will test your knowledge on Hadoop filesystem and Apache Hive(HQL).

Input file **bhav040416\_080416.csv** has trading data on NSE from 4th Apr 2016 to 8th Apr 2016.

This input file has trading data from 5 days in comma separated format with following columns as header...

SYMBOL,SERIES,OPEN,HIGH,LOW,CLOSE,LAST,PREVCLOSE,TOTTRDQTY,TOTTRDVAL,TIMESTAMP,TOTALTRADES,ISIN.

This input file is part of download.

Perform the following tasks:

1. Copy input file from local filesystem directory/home/osgdev/Downloads to hadoop filesystem directory /user/hadoopTraining/Assignment1/input/.

2. Using Apache Hive create a database, an external table 'stockTrdData' in database and load this data in Hive table. Verify the data loaded by selecting first 100 rows from table.

3. Using Apache Hive queries, list symbol on each series with highest TOTTRDVAL. Copy results to local filesystem.

4. Using Apache Hive queries, list symbol on each series with highest TOTTRQTY. Check the results on Hadoop filesystem using cat command.

5. Using Apache Hive queries, list symbol on only EQ series with highest TOTTRQTY.

**Problem-2**

This assignment will test your knowledge on Hive advance concepts like Partitioning and more.

Input file **u.user** has users data who are rating movies in a movies database. This data is '|' delimited and each row has five fields.

Fields have following values on each row in order : userid, age(in years), gender(M/F), occupation, zipcode.

ex: 902|45|F|artist|97203.

Tasks to perform on this data:

1. In Hive, Create a database and in this DB create an internal (managed) table to load this data.

2. Load this data from local filesystem (not from HDFS) to internal (managed) table created above.

3. Create a partitioned table with partition on gender column (i.e. M and F partitions) and load data into this table from table loaded on task 2.

4. List the HDFS filesystem and list the HDFS file created for above partitioned table.

5. Run query to get the below on non-partitioned internal table and partitioned internal table created above...

Avg age of female audience?

Avg age of male audience?

Can you calculate note the time taken for each and calculate improved performance in case of partitioned tables.

**Problem-3**

This assignment will test your knowledge on Hive advance concept like Bucketing, Sampling etc.

Source file **u.data** has movie rating data in tab delimited format with columns:

user id | item id | rating | timestamp(time stamps are unix seconds since 1/1/1970 UTC).

Perform the following task on this data:

1. Load source data from native OS filesystem in Hive tables: Non Bucketing

2. Load this data from Hive table to bucketed hive table with 3 buckets on userid column.

3. Count the records loaded in each bucket. List the Hadoop files created for this table.

Do you see the difference from partitioned table?

4. Fetch average rating over 5% on sample data on bucketted rating table.

5. Fetch average rating on 3rd bucket of bucketted rating table.

Fetch average rating on full rating bucketted table and compare results.