**BIG DATA -Assignment 2:**

**Submission Date: 21st June 2019**

1. Download any CSV dataset
2. Run query with 2 Group by clauses, 1 nested select and 1 sum operation
3. Run query with 2 sort clauses, 1 grouping and 1 average operation

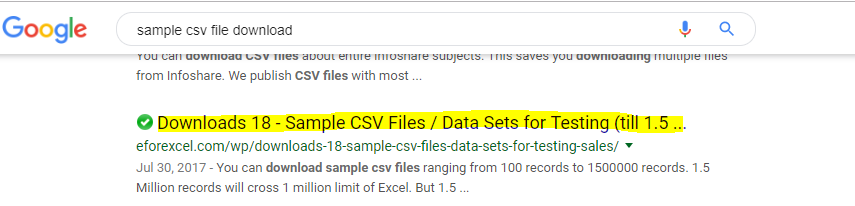
Implement it with

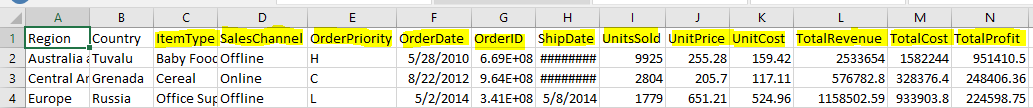
1. HUE- Impala - > \* .sql file
2. Python - sqlAlclemy -> \*.py file
3. SQL online editor.

**BIG DATA -Assignment 3:**

**Submission Date: 21st June 2019**

1. Download Source Tree
2. Set up a github account
3. Upload all assignments on github
4. **HUE-Impala – Solution**
5. Download csv file

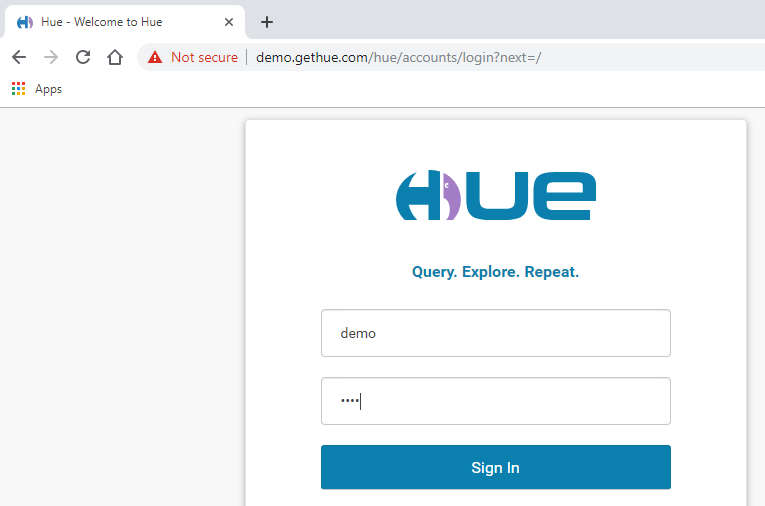


1. Edit csv file to remove blank space from titles of column 
2. Login in to Hue (Use chrome browser)

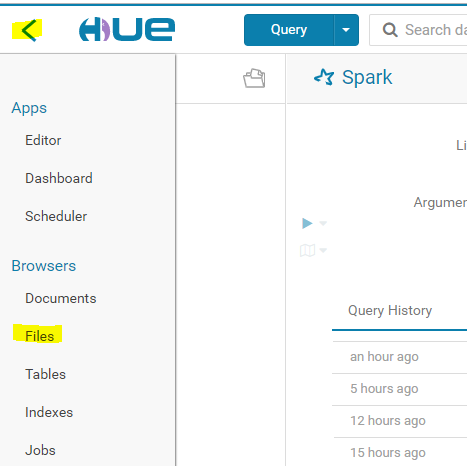
url: <http://demo.gethue.com>

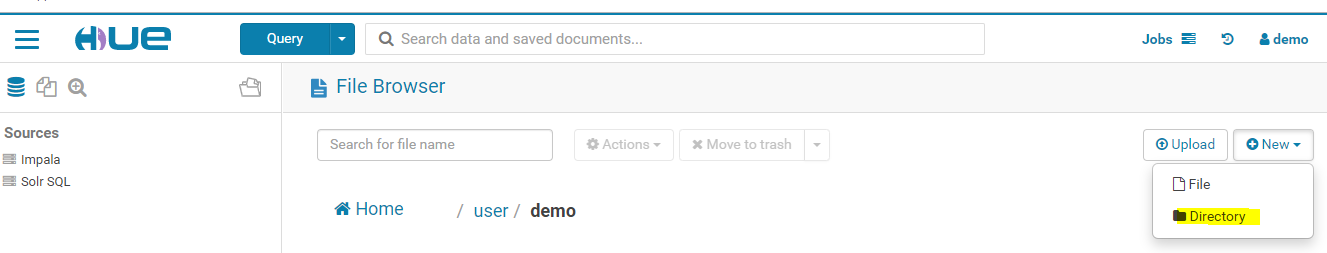
user: demo

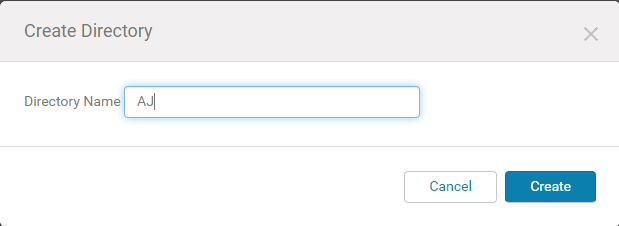
password: demo

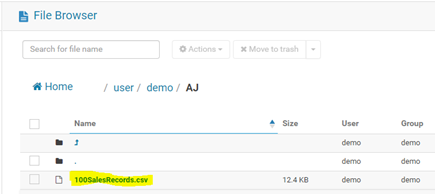


1. Upload CSV file:

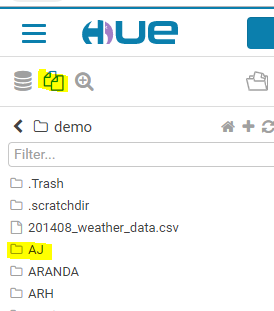


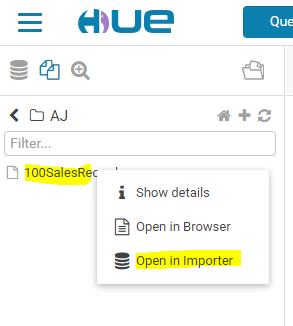


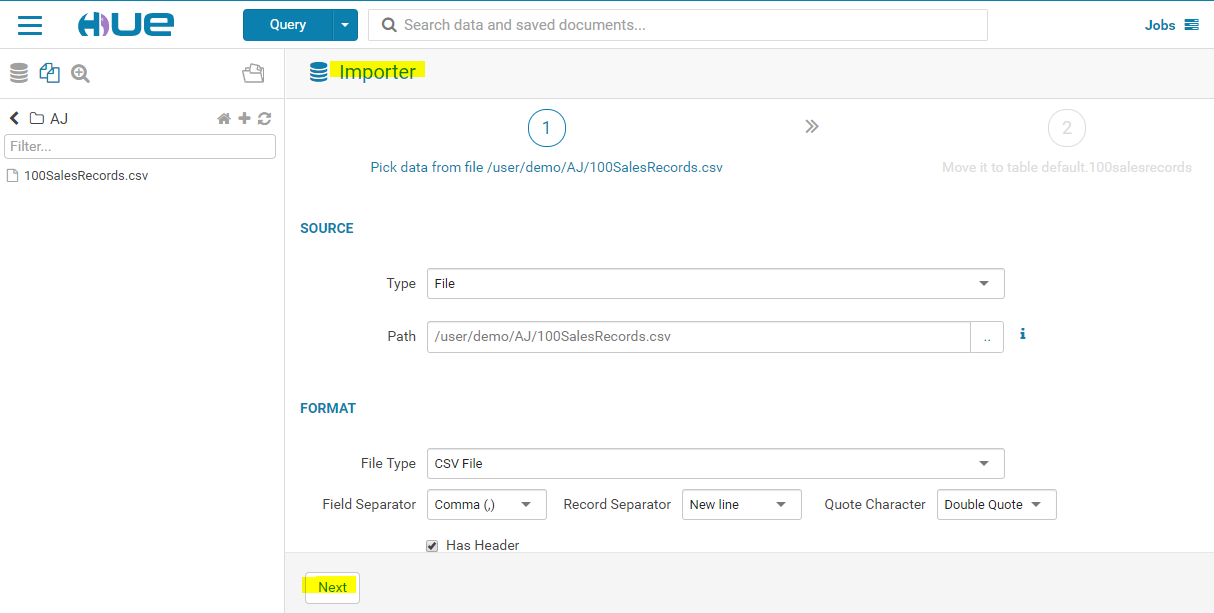


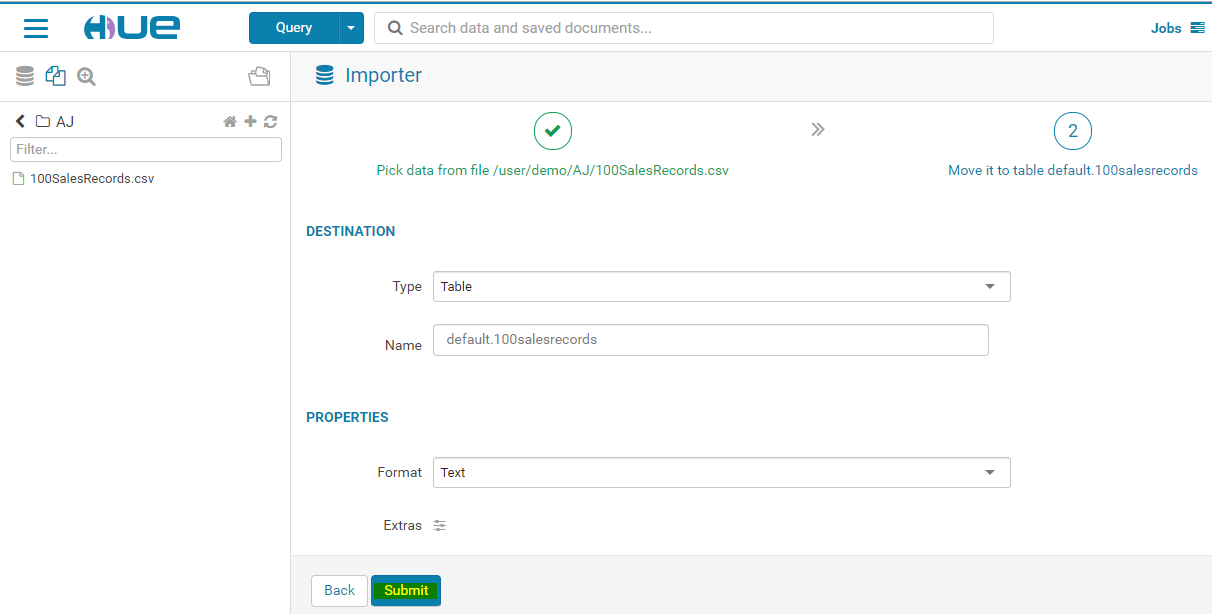


1. Move data from CSV to SQL table

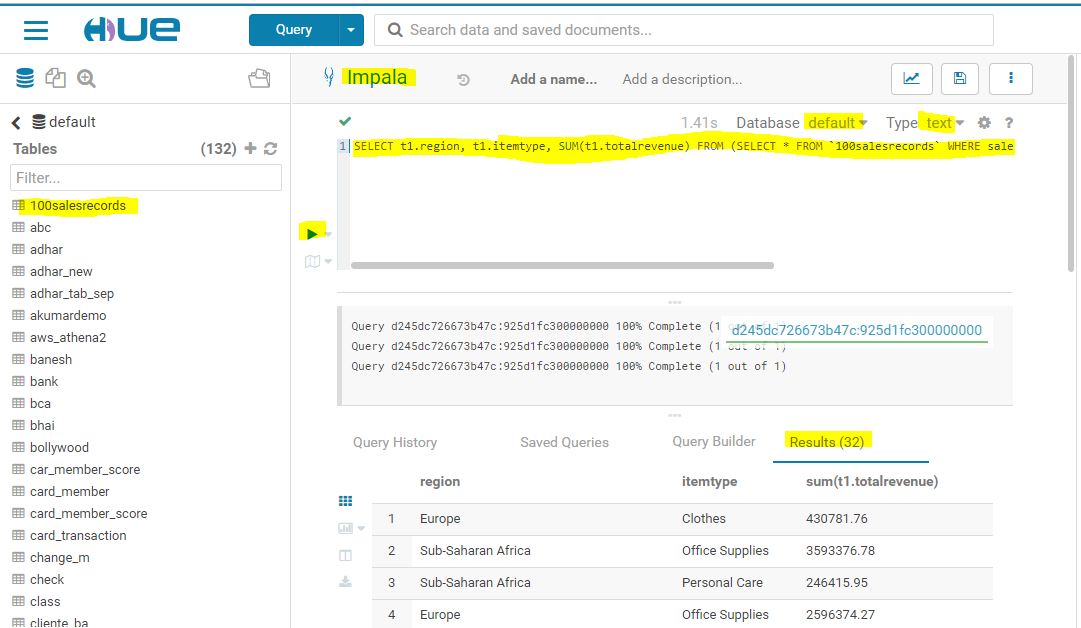








1. Write SQL query and verify it in Impala editor.



/\* Query1 -> 1 nested (inner) query, 2 Group By clause and 1 SUM operation \*/

/\* Query to get total revenue for online saleschannel, regionwise and itemtypewise\*/

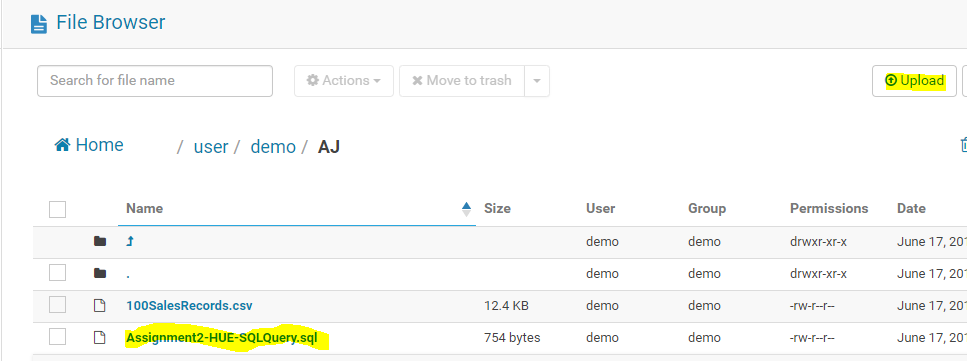
SELECT t1.region, t1.itemtype, SUM(t1.totalrevenue) FROM (SELECT \* FROM `100salesrecords` WHERE saleschannel = 'Online')t1 GROUP BY t1.region, t1.itemtype

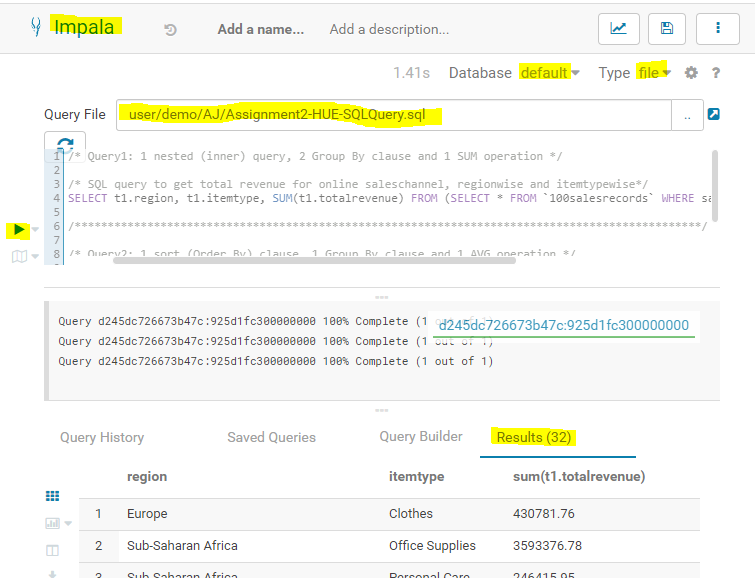
/\* Query2 -> 1 sort (Order By) clause, 1 Group By clause and 1 AVG operation \*/

/\* Query to get average of total profit, regionwise, order by region and total profit\*/

SELECT region, AVG(totalprofit) FROM `100salesrecords` GROUP BY region ORDER BY region, AVG(totalprofit)

1. Create sql file, upload and test it in Impala editor.





1. **Python – Solution**

Python code:

from sqlalchemy import create\_engine

import pandas as pd

# Note: Due to memory issue reduced no.of records to 50 rows

data = pd.read\_csv('C:\\PGDDS\\BigData\\50SalesRecords.csv')

# Create the db engine

engine = create\_engine('sqlite:///:memory:')

# Store the dataframe as a table

data.to\_sql('data\_table', engine)

# Query 1 to get total revenue for online saleschannel, regionwise and itemtypewise\*/

print('Query 1 to get total revenue for online saleschannel, regionwise and itemtypewise')

res1 = pd.read\_sql\_query('SELECT t1.region, t1.itemtype, SUM(t1.totalrevenue) FROM (SELECT \* FROM data\_table WHERE saleschannel = \'Online\')t1 GROUP BY t1.region, t1.itemtype', engine)

print('Result 1')

print(res1)

print('\n----------------------------------------------------------------------------------------\n')

# Query 2 to get average of total profit, regionwise, order by region and total profit

print('Query 2 to get average of total profit, regionwise, order by region and total profit')

res2 = pd.read\_sql\_query('SELECT region, AVG(totalprofit) FROM data\_table GROUP BY region ORDER BY region, AVG(totalprofit)', engine)

print('Result 2')

print(res2)

print('\n----------------------------------------------------------------------------------------\n')