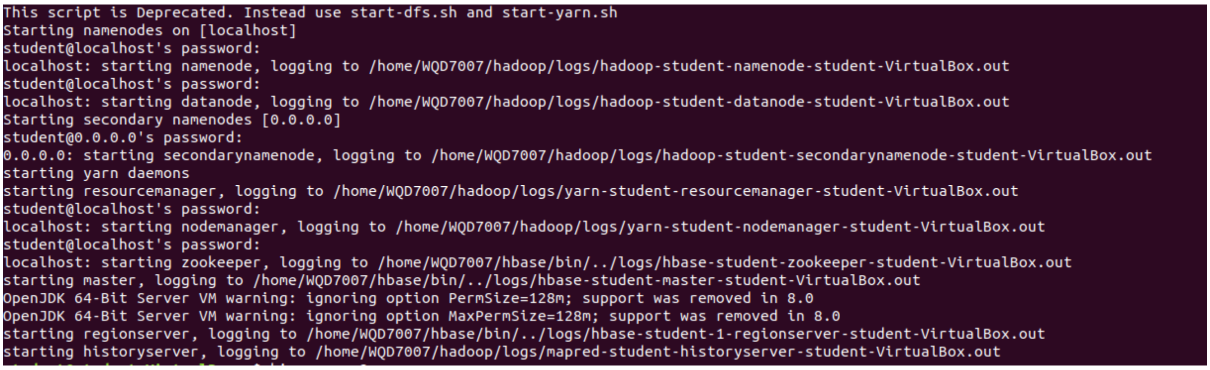
WQD7005 Data Mining

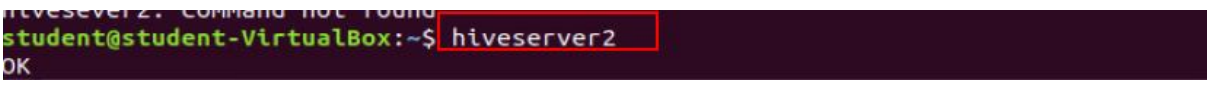
Video 3: https://www.loom.com/share/b0d7878ccc4341c482e3aa4d1d09ad94

Milestone 3: Accessing Hive Data Warehouse by using Python

In terminal, start Hadoop and hiveserver2

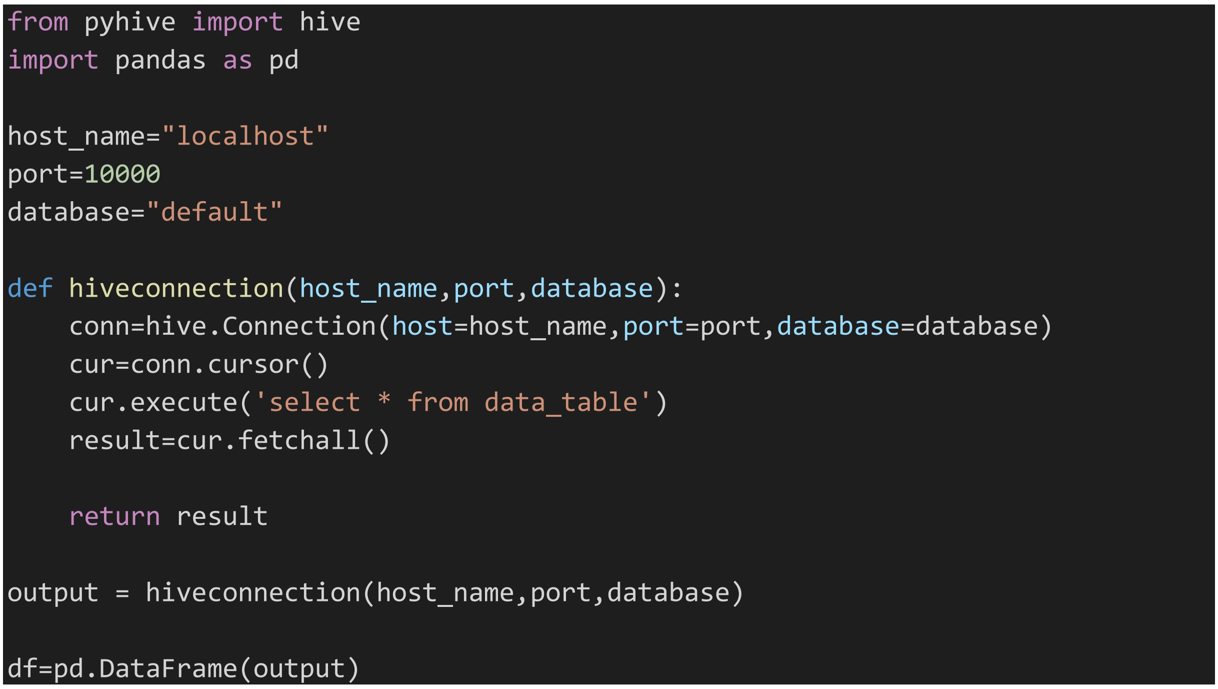


```
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
student@localhost's password:
localhost: starting namenode, logging to /home/WQD7007/hadoop/logs/hadoop-student-namenode-student-VirtualBox.out
student@localhost's password:
localhost: starting datanode, logging to /home/WQD7007/hadoop/logs/hadoop-student-datanode-student-VirtualBox.out
Starting secondary namenodes [0.0.0.0]
student@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /home/WQD7007/hadoop/logs/hadoop-student-secondarynamenode-student-VirtualBox.out
starting yarn daemons
starting resourcemanager, logging to /home/WQD7007/hadoop/logs/yarn-student-resourcemanager-student-VirtualBox.out
student@localhost's password:
localhost: starting nodemanager, logging to /home/WQD7007/hadoop/logs/yarn-student-nodemanager-student-VirtualBox.out
student@localhost's password:
localhost: starting zookeeper, logging to /home/WQD7007/hbase/bin/../logs/hbase-student-zookeeper-student-VirtualBox.out
starting master, logging to /home/WQD7007/hbase/bin/../logs/hbase-student-master-student-VirtualBox.out
OpenJDK 64-Bit Server VM warning: ignoring option PermSize=128m; support was removed in 8.0
OpenJDK 64-Bit Server VM warning: ignoring option MaxPermSize=128m; support was removed in 8.0
starting regionserver, logging to /home/WQD7007/hbase/bin/../logs/hbase-student-1-regionserver-student-VirtualBox.out
starting historyserver, logging to /home/WQD7007/hadoop/logs/mapred-student-historyserver-student-VirtualBox.out
```



```
hivesever2: command not found
student@student-VirtualBox:~$ hiveserver2
OK
```

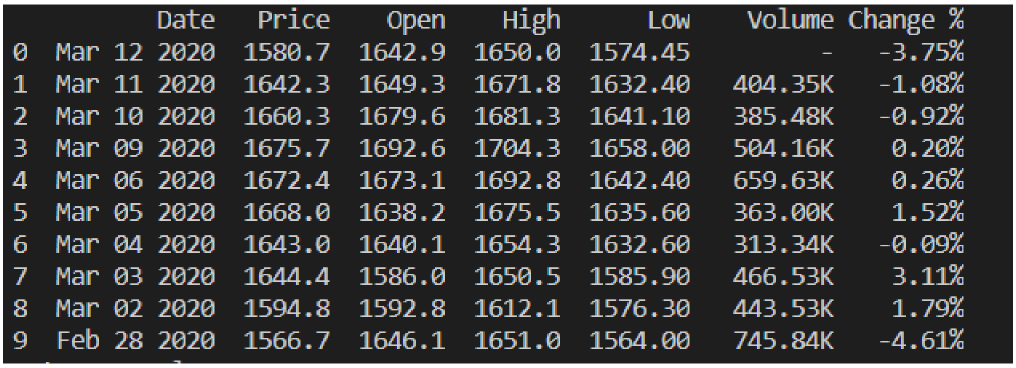
To access data from hive, we import pyhive library in python script to connect to hive server. host_name=localhost, port=port & database=database



```python
from pyhive import hive
import pandas as pd

host_name="localhost"
port=10000
database="default"

def hiveconnection(host_name,port,database):
    conn=hive.Connection(host=host_name,port=port,database=database)
    cur=conn.cursor()
    cur.execute('select * from data_table')
    result=cur.fetchall()

    return result

output = hiveconnection(host_name,port,database)

df=pd.DataFrame(output)
```

Output:

The data are loaded and top 10 row of data is read.



```
         Date   Price    Open    High      Low    Volume Change %
0  Mar 12 2020  1580.7  1642.9  1650.0  1574.45         -   -3.75%
1  Mar 11 2020  1642.3  1649.3  1671.8  1632.40   404.35K   -1.08%
2  Mar 10 2020  1660.3  1679.6  1681.3  1641.10   385.48K   -0.92%
3  Mar 09 2020  1675.7  1692.6  1704.3  1658.00   504.16K    0.20%
4  Mar 06 2020  1672.4  1673.1  1692.8  1642.40   659.63K    0.26%
5  Mar 05 2020  1668.0  1638.2  1675.5  1635.60   363.00K    1.52%
6  Mar 04 2020  1643.0  1640.1  1654.3  1632.60   313.34K   -0.09%
7  Mar 03 2020  1644.4  1586.0  1650.5  1585.90   466.53K    3.11%
8  Mar 02 2020  1594.8  1592.8  1612.1  1576.30   443.53K    1.79%
9  Feb 28 2020  1566.7  1646.1  1651.0  1564.00   745.84K   -4.61%
```