WQD7005 Data Mining

Video 2: https://www.loom.com/share/5ac650128d234834801bb6d28df271b4

Milestone 2: Store data into hive data warehouse

In terminal, create a directory called 'datamining' in hadoop folder. Then, put the *goldprice.csv* file into the folder created.

```
student@student-VirtualBox:~$ hadoop fs -mkdir /user/hdfs/datamining
```

```
student@student-VirtualBox:~$ hadoop fs -put /home/student/Downloads/goldprice.csv /user/hdfs/datamining
student@student-VirtualBox:~$ hadoop fs -ls /user/hdfs/datamining
Found 1 items
-rw-r--r--   1 student supergroup      17489 2020-03-22 21:21 /user/hdfs/datamining/goldprice.csv
```

In hive terminal, create a table called 'data_table' and stored the data into hive warehouse. Then, the header of the table is removed and the first 5 rows of data is selected and print out in the result.

```
student@student-VirtualBox:~$ hive
ls: cannot access '/home/WQD7007/spark/lib/spark-assembly-*.jar': No such file or directory

Logging initialized using configuration in jar:file:/home/WQD7007/hive/lib/hive-common-1.2.2.jar!/hive-l
og4j.properties
hive> DROP TABLE data_table;
OK
Time taken: 1.659 seconds
hive> CREATE TABLE data_table
    > (Date_s STRING, Price DOUBLE, Open DOUBLE, High DOUBLE, Low DOUBLE, Volume STRING, Change STRING)
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.399 seconds
hive> LOAD DATA INPATH '/user/hdfs/datamining/goldprice.csv' into table data_table;
Loading data to table default.data_table
Table default.data_table stats: [numFiles=1, totalSize=17489]
OK
Time taken: 0.417 seconds
hive> alter table data_table set tblproperties ("skip.header.line.count"="1");
OK
Time taken: 0.092 seconds
hive> select * from data_table limit 5;
OK
Mar 20 2020     1501.15 1473.5  1518.9  1473.5   -       1.85%
Mar 19 2020     1473.95 1500.25 1500.3  1457.9   -       -0.27%
Mar 18 2020     1477.9  1527.6  1547.0  1473.3   415.49K         -3.14%
Mar 17 2020     1525.8  1512.8  1554.3  1465.6   434.51K         2.64%
Mar 16 2020     1486.5  1563.8  1574.8  1450.9   565.98K         -1.99%
Time taken: 0.45 seconds, Fetched: 5 row(s)
```