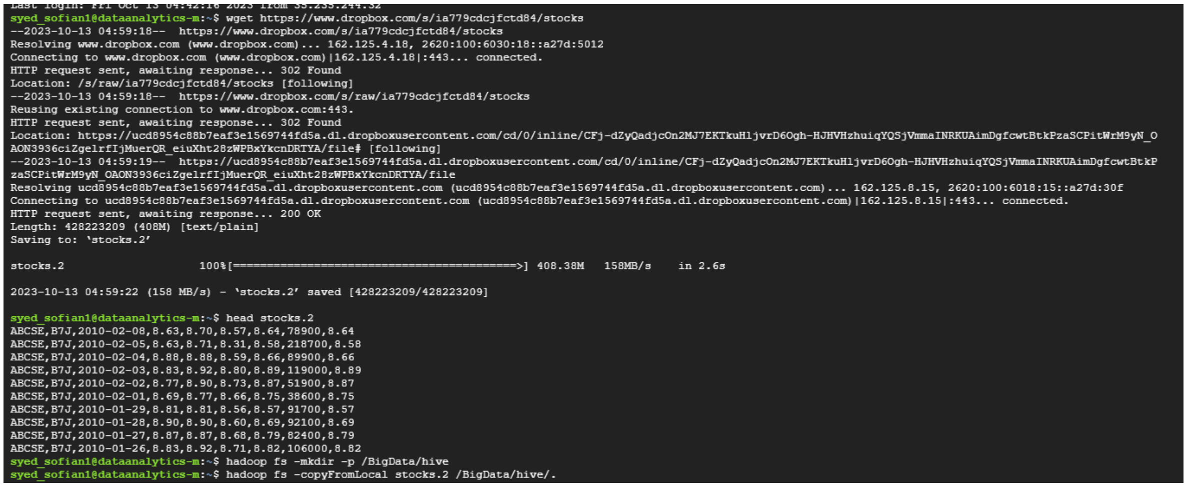
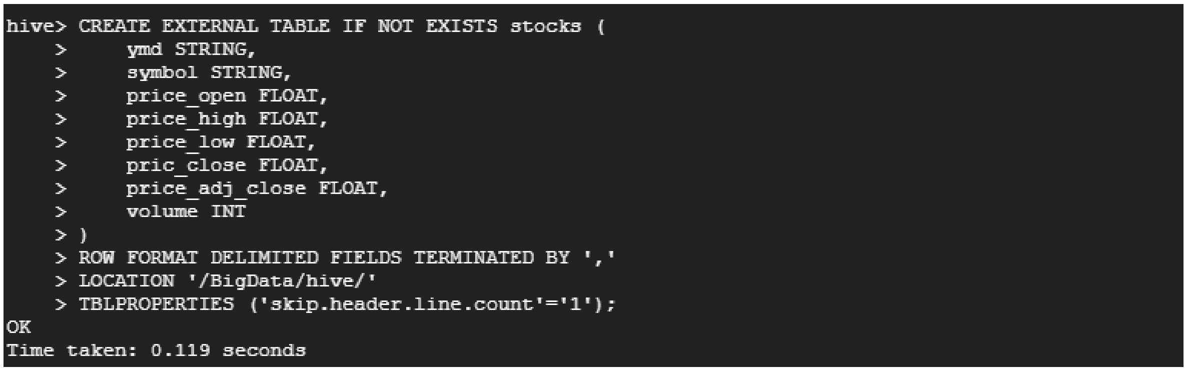
HiveQL Queries

1). Loading the SP 500 stocks dataset into HDFS and using the dataset to create a Hive table and using the LOCATION attribute.

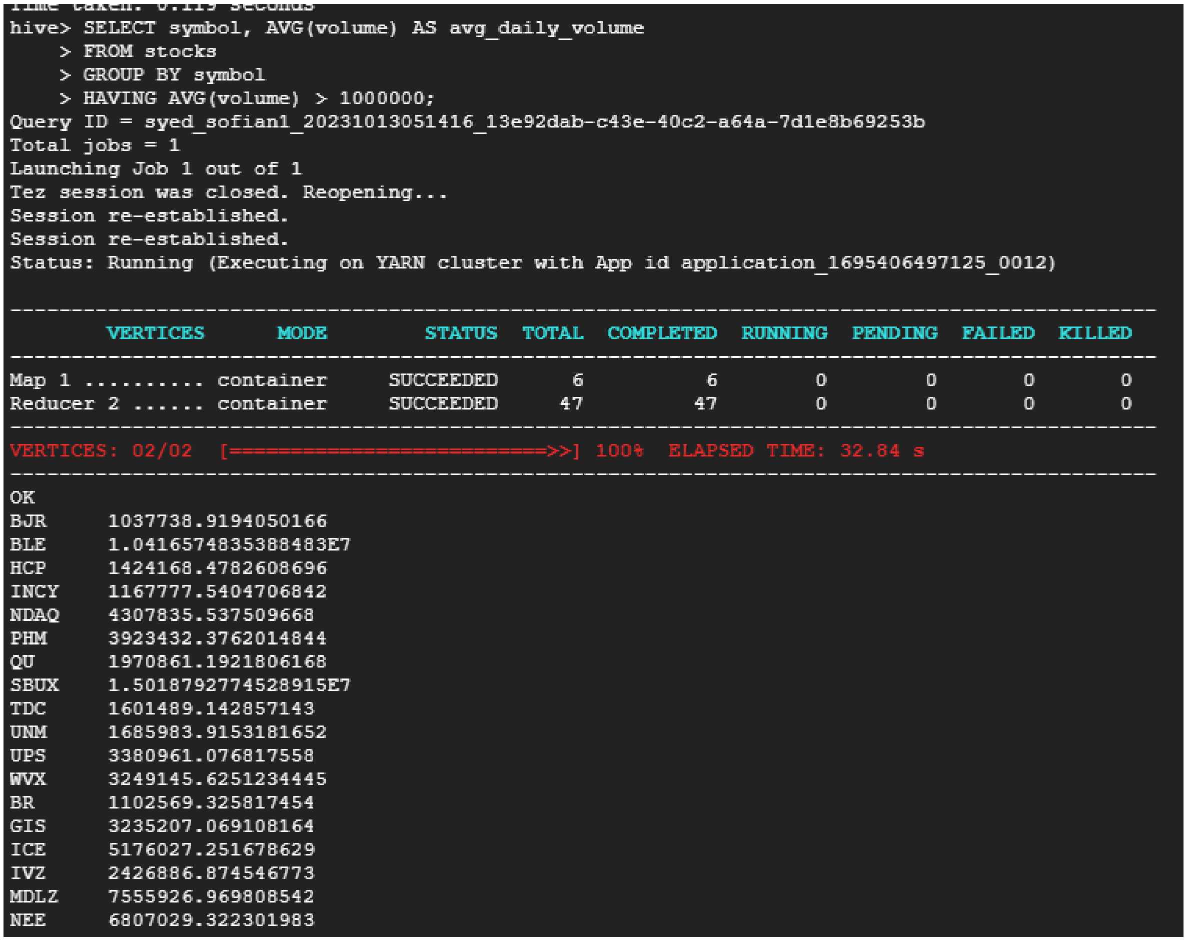


```
Last login: Fri Oct 13 04:42:16 2023 from 35.235.244.32
syed_sofian1@dataanalytics-m:~$ wget https://www.dropbox.com/s/ia779cdcjfctd84/stocks
--2023-10-13 04:59:18--  https://www.dropbox.com/s/ia779cdcjfctd84/stocks
Resolving www.dropbox.com (www.dropbox.com)... 162.125.4.18, 2620:100:6030:18::a27d:5012
Connecting to www.dropbox.com (www.dropbox.com)|162.125.4.18|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: /s/raw/ia779cdcjfctd84/stocks [following]
--2023-10-13 04:59:18--  https://www.dropbox.com/s/raw/ia779cdcjfctd84/stocks
Reusing existing connection to www.dropbox.com:443.
HTTP request sent, awaiting response... 302 Found
Location: https://ucd8954c88b7eaf3e1569744fd5a.dl.dropboxusercontent.com/cd/0/inline/CFj-dZyQadjcOn2MJ7EKTkuHljvrD6Ogh-HJHVHzhuiqYQSjVmmaINRKUAimDgfcwtBtkPzaSCPitWrM9yN_O
AON3936ciZgelrfIjMuerQR_eiuXht28zWPBxYkcnDRTYA/file# [following]
--2023-10-13 04:59:19--  https://ucd8954c88b7eaf3e1569744fd5a.dl.dropboxusercontent.com/cd/0/inline/CFj-dZyQadjcOn2MJ7EKTkuHljvrD6Ogh-HJHVHzhuiqYQSjVmmaINRKUAimDgfcwtBtkP
zaSCPitWrM9yN_OAON3936ciZgelrfIjMuerQR_eiuXht28zWPBxYkcnDRTYA/file
Resolving ucd8954c88b7eaf3e1569744fd5a.dl.dropboxusercontent.com (ucd8954c88b7eaf3e1569744fd5a.dl.dropboxusercontent.com)... 162.125.8.15, 2620:100:6018:15::a27d:30f
Connecting to ucd8954c88b7eaf3e1569744fd5a.dl.dropboxusercontent.com (ucd8954c88b7eaf3e1569744fd5a.dl.dropboxusercontent.com)|162.125.8.15|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 428223209 (408M) [text/plain]
Saving to: ‘stocks.2’

stocks.2                  100%[=======================================>] 408.38M   158MB/s    in 2.6s

2023-10-13 04:59:22 (158 MB/s) - ‘stocks.2’ saved [428223209/428223209]

syed_sofian1@dataanalytics-m:~$ head stocks.2
ABCSE,B7J,2010-02-08,8.63,8.70,8.57,8.64,78900,8.64
ABCSE,B7J,2010-02-05,8.63,8.71,8.31,8.58,218700,8.58
ABCSE,B7J,2010-02-04,8.88,8.88,8.59,8.66,89900,8.66
ABCSE,B7J,2010-02-03,8.83,8.92,8.80,8.89,119000,8.89
ABCSE,B7J,2010-02-02,8.77,8.90,8.73,8.87,51900,8.87
ABCSE,B7J,2010-02-01,8.69,8.77,8.66,8.75,38600,8.75
ABCSE,B7J,2010-01-29,8.81,8.81,8.56,8.57,91700,8.57
ABCSE,B7J,2010-01-28,8.90,8.90,8.60,8.69,92100,8.69
ABCSE,B7J,2010-01-27,8.87,8.87,8.68,8.79,82400,8.79
ABCSE,B7J,2010-01-26,8.83,8.92,8.71,8.82,106000,8.82
syed_sofian1@dataanalytics-m:~$ hadoop fs -mkdir -p /BigData/hive
syed_sofian1@dataanalytics-m:~$ hadoop fs -copyFromLocal stocks.2 /BigData/hive/.
```



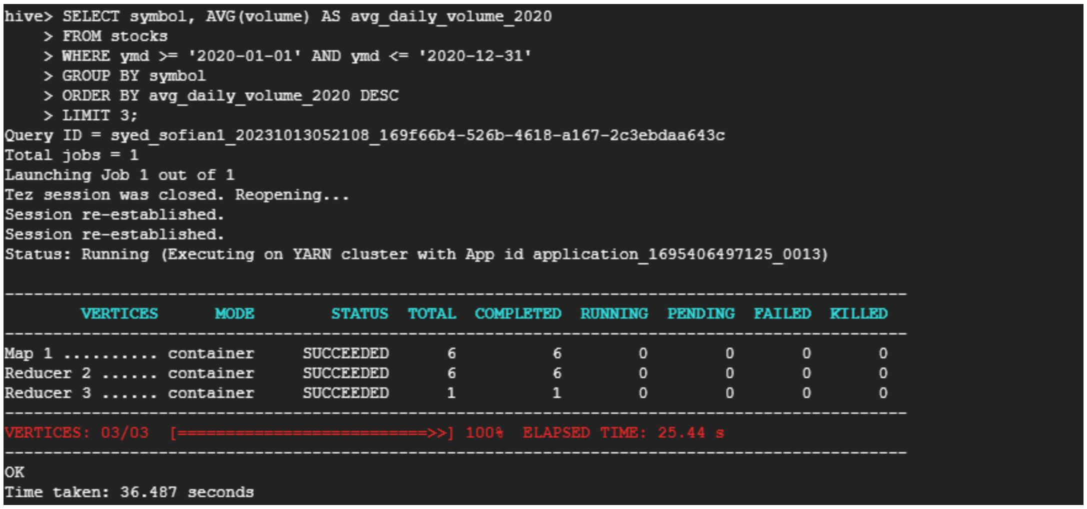
```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS stocks (
    >     ymd STRING,
    >     symbol STRING,
    >     price_open FLOAT,
    >     price_high FLOAT,
    >     price_low FLOAT,
    >     pric_close FLOAT,
    >     price_adj_close FLOAT,
    >     volume INT
    > )
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > LOCATION '/BigData/hive/'
    > TBLPROPERTIES ('skip.header.line.count'='1');
OK
Time taken: 0.119 seconds
```

2). Using Hive query to find the stocks with average daily volume larger than1 million shares.



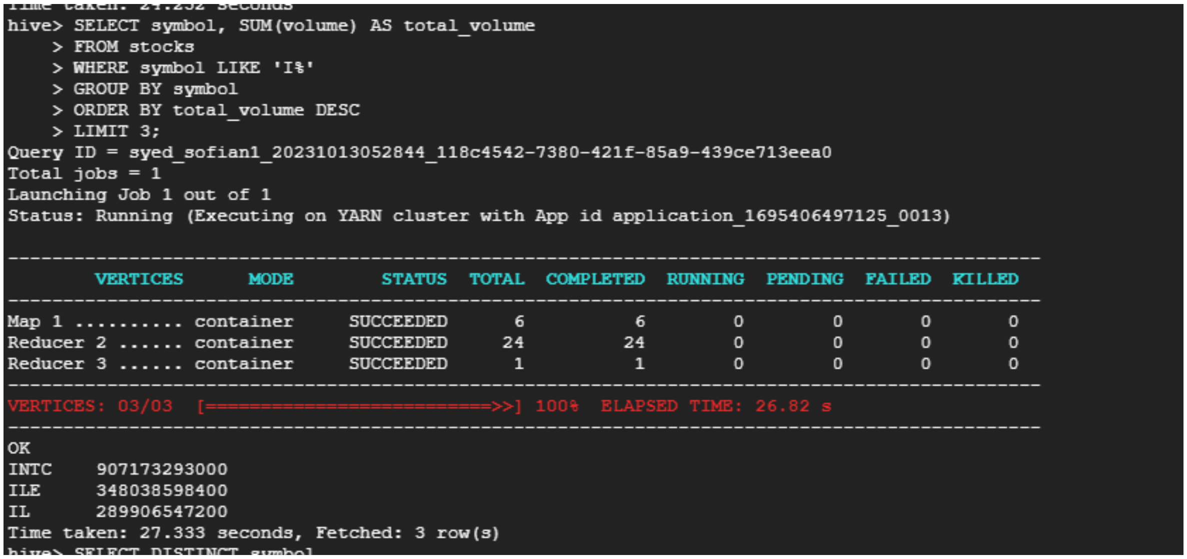
```
Time taken: 0.119 seconds
hive> SELECT symbol, AVG(volume) AS avg_daily_volume
    > FROM stocks
    > GROUP BY symbol
    > HAVING AVG(volume) > 1000000;
Query ID = syed_sofian1_20231013051416_13e92dab-c43e-40c2-a64a-7d1e8b69253b
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1695406497125_0012)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     47         47        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 32.84 s
----------------------------------------------------------------------------------------------
OK
BJR     1037738.9194050166
BLE     1.0416574835388483E7
HCP     1424168.4782608696
INCY    1167777.5404706842
NDAQ    4307835.537509668
PHM     3923432.3762014844
QU      1970861.1921806168
SBUX    1.5018792774528915E7
TDC     1601489.142857143
UNM     1685983.9153181652
UPS     3380961.076817558
WVX     3249145.6251234445
BR      1102569.325817454
GIS     3235207.069108164
ICE     5176027.251678629
IVZ     2426886.874546773
MDLZ    7555926.969808542
NEE     6807029.322301983
```

3). Using Hive query to find the top 3 stocks by average volume for the year 2020.



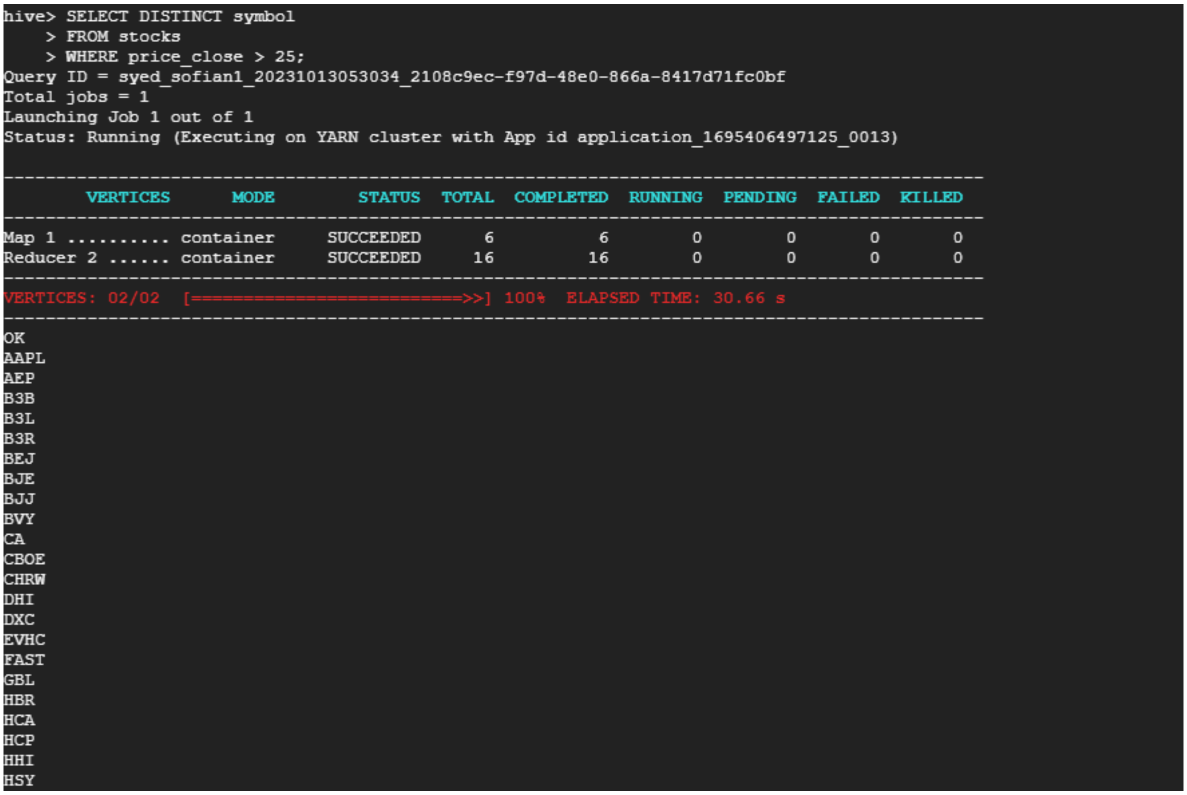
```
hive> SELECT symbol, AVG(volume) AS avg_daily_volume_2020
    > FROM stocks
    > WHERE ymd >= '2020-01-01' AND ymd <= '2020-12-31'
    > GROUP BY symbol
    > ORDER BY avg_daily_volume_2020 DESC
    > LIMIT 3;
Query ID = syed_sofian1_20231013052108_169f66b4-526b-4618-a167-2c3ebdaa643c
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1695406497125_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      6          6        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 25.44 s
----------------------------------------------------------------------------------------------
OK
Time taken: 36.487 seconds
```

4). Using Hive query to find the top 3 stocks by volume and whose symbol start with the letter I.

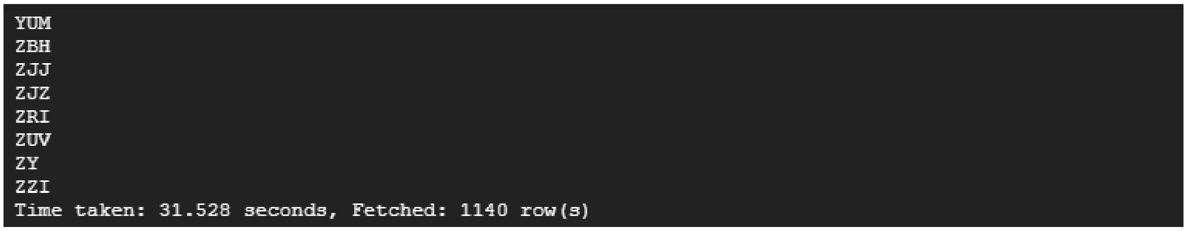


```
Time taken: 24.252 seconds
hive> SELECT symbol, SUM(volume) AS total_volume
    > FROM stocks
    > WHERE symbol LIKE 'I%'
    > GROUP BY symbol
    > ORDER BY total_volume DESC
    > LIMIT 3;
Query ID = syed_sofian1_20231013052844_118c4542-7380-421f-85a9-439ce713eea0
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1695406497125_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     24         24        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 26.82 s
----------------------------------------------------------------------------------------------
OK
INTC    907173293000
ILE     348038598400
IL      289906547200
Time taken: 27.333 seconds, Fetched: 3 row(s)
hive> SELECT DISTINCT symbol
```

5). Using Hive query to find all the stocks symbols whose closing price is larger than your age.

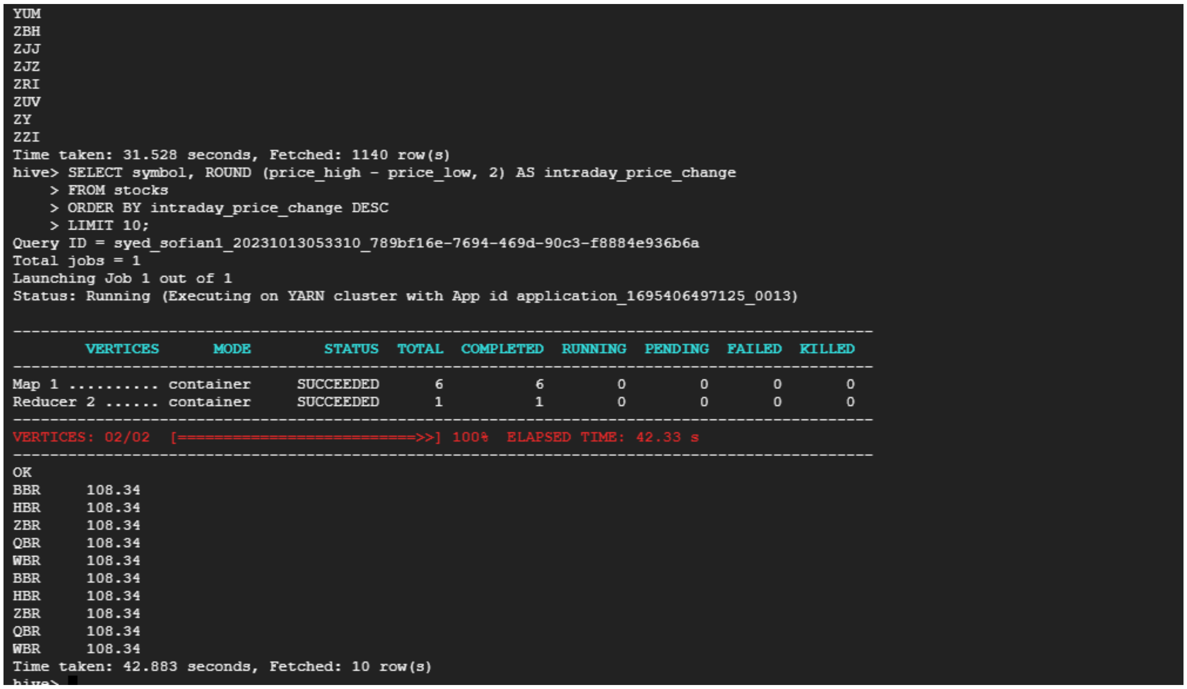


```
hive> SELECT DISTINCT symbol
    > FROM stocks
    > WHERE price_close > 25;
Query ID = syed_sofian1_20231013053034_2108c9ec-f97d-48e0-866a-8417d71fc0bf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1695406497125_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED     16         16        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 30.66 s
----------------------------------------------------------------------------------------------
OK
AAPL
AEP
B3B
B3L
B3R
BEJ
BJE
BJJ
BVY
CA
CBOE
CHRW
DHI
DXC
EVHC
FAST
GBL
HBR
HCA
HCP
HHI
HSY
```



```
YUM
ZBH
ZJJ
ZJZ
ZRI
ZUV
ZY
ZZI
Time taken: 31.528 seconds, Fetched: 1140 row(s)
```

6). Using Hive query to find the top 10 stocks with largest intraday price change and display the amount of the change.



```
YUM
ZBH
ZJJ
ZJZ
ZRI
ZUV
ZY
ZZI
Time taken: 31.528 seconds, Fetched: 1140 row(s)
hive> SELECT symbol, ROUND (price_high - price_low, 2) AS intraday_price_change
    > FROM stocks
    > ORDER BY intraday_price_change DESC
    > LIMIT 10;
Query ID = syed_sofian1_20231013053310_789bf16e-7694-469d-90c3-f8884e936b6a
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1695406497125_0013)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      6          6        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 42.33 s
----------------------------------------------------------------------------------------------
OK
BBR     108.34
HBR     108.34
ZBR     108.34
QBR     108.34
WBR     108.34
BBR     108.34
HBR     108.34
ZBR     108.34
QBR     108.34
WBR     108.34
Time taken: 42.883 seconds, Fetched: 10 row(s)
hive>
```