1) Create a data pipeline using sqoop to pull the data from the table below from MYSQL server into Hive.

Transferring CSV files into local server.

```
[gyankrishnaagmail@ip-10-0-41-79 ~]$ ls
StockCompanies.csv  StockPrices.csv  T1.txt
[gyankrishnaagmail@ip-10-0-41-79 ~]$
```

Logging into MySQL .

```
[gyankrishnaagmail@ip-10-0-41-79 ~]$ ls
StockCompanies.csv  StockPrices.csv  T1.txt
[gyankrishnaagmail@ip-10-0-41-79 ~]$ mysql -h sqoopdb.slbdh.cloudlabs.com -u gyankrishnaagmail -pgyankrishnaagmailtltyi
Welcome to the MariaDB monitor.  Commands end with ; or \g.
Your MySQL connection id is 81515
Server version: 8.0.20 MySQL Community Server - GPL

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

MySQL [(none)]>
```

creating table stock_companies and loading data

```
MySQL [gyankrishnaagmail]> CREATE TABLE stock_companies(
    ->              symbol VARCHAR(5),
    ->              company_name VARCHAR(38),
    ->              sector VARCHAR(27),
    ->              sub_industry VARCHAR(47),
    ->              headquarter VARCHAR(47),
    ->              Primary Key (symbol));
Query OK, 0 rows affected (0.02 sec)

MySQL [gyankrishnaagmail]>
MySQL [gyankrishnaagmail]> LOAD DATA LOCAL INFILE '/mnt/home/gyankrishnaagmail/StockCompanies.csv'
    -> INTO TABLE stock_companies
    -> FIELDS TERMINATED BY ','
    -> OPTIONALLY ENCLOSED BY '"'
    -> LINES TERMINATED BY '\r\n'
    -> IGNORE 1 ROWS;
Query OK, 505 rows affected (0.03 sec)
Records: 505  Deleted: 0  Skipped: 0  Warnings: 0
```

```
CREATE TABLE stock_companies(
        symbol VARCHAR(5),
        company_name VARCHAR(38),
        sector VARCHAR(27),
        sub_industry VARCHAR(47),
        headquarter VARCHAR(47),
        Primary Key (symbol));
LOAD DATA LOCAL INFILE '/mnt/home/gyankrishnaagmail/StockCompanies.csv'
INTO TABLE stock_companies
FIELDS TERMINATED BY ','
OPTIONALLY ENCLOSED BY '"'
ESCAPED BY '\;'
LINES TERMINATED BY '\r\n'
```

IGNORE 1 ROWS;

creating table stock_prices and loading data

```
MySQL [gyankrishnaagmail]> CREATE TABLE stock_prices (
    ->              trading_date Date,
    ->              symbol VARCHAR(5),
    ->              open DOUBLE,
    ->              close DOUBLE,
    ->              low DOUBLE,
    ->              high DOUBLE,
    ->              volume INT
    ->              );
Query OK, 0 rows affected (0.05 sec)

MySQL [gyankrishnaagmail]>
MySQL [gyankrishnaagmail]> LOAD DATA LOCAL INFILE '/mnt/home/gyankrishnaagmail/StockPrices.csv'
    -> INTO TABLE stock_prices
    -> FIELDS TERMINATED BY ','
    -> ENCLOSED BY '"'
    -> LINES TERMINATED BY '\n'
    -> IGNORE 1 ROWS;
Query OK, 851264 rows affected (6.04 sec)
Records: 851264  Deleted: 0  Skipped: 0  Warnings: 0
```

CREATE TABLE stock_prices (

trading_date Date,

symbol VARCHAR(5),

open DOUBLE,

close DOUBLE,

low DOUBLE,

high DOUBLE,

volume INT

);

LOAD DATA LOCAL INFILE '/mnt/home/gyankrishnaagmail/StockPrices.csv'

INTO TABLE stock_prices

FIELDS TERMINATED BY ','

ENCLOSED BY '"'

LINES TERMINATED BY '\n'

IGNORE 1 ROWS;

Viewing loaded data

Stock_Companies

```
MySQL [gyankrishnaagmail]> select * from stock_companies;
+--------+----------------------------------------+----------------------------+-----------------------------------------
----+----------------------------------------+
| symbol | company_name                           | sector                     | sub_industry
    | headquarter                            |
+--------+----------------------------------------+----------------------------+-----------------------------------------
----+----------------------------------------+
| A      | Agilent Technologies Inc               | Health Care                | Health Care Equipment
    | Santa Clara; California                |
| AAL    | American Airlines Group                | Industrials                | Airlines
    | Fort Worth; Texas                      |
| AAP    | Advance Auto Parts                     | Consumer Discretionary     | Automotive Retail
    | Roanoke; Virginia                      |
| AAPL   | Apple Inc.                             | Information Technology     | Computer Hardware
    | Cupertino; California                  |
| ABBV   | AbbVie                                 | Health Care                | Pharmaceuticals
    | North Chicago; Illinois                |
| ABC    | AmerisourceBergen Corp                 | Health Care                | Health Care Distributors
    | Chesterbrook; Pennsylvania             |
| ABT    | Abbott Laboratories                    | Health Care                | Health Care Equipment
    | North Chicago; Illinois                |
| ACN    | Accenture plc                          | Information Technology     | IT Consulting & Other Services
    | Dublin; Ireland                        |
| ADBE   | Adobe Systems Inc                      | Information Technology     | Application Software
    | San Jose; California                   |
| ADI    | Analog Devices; Inc.                   | Information Technology     | Semiconductors
    | Norwood; Massachusetts                 |
| ADM    | Archer-Daniels-Midland Co              | Consumer Staples           | Agricultural Products
```

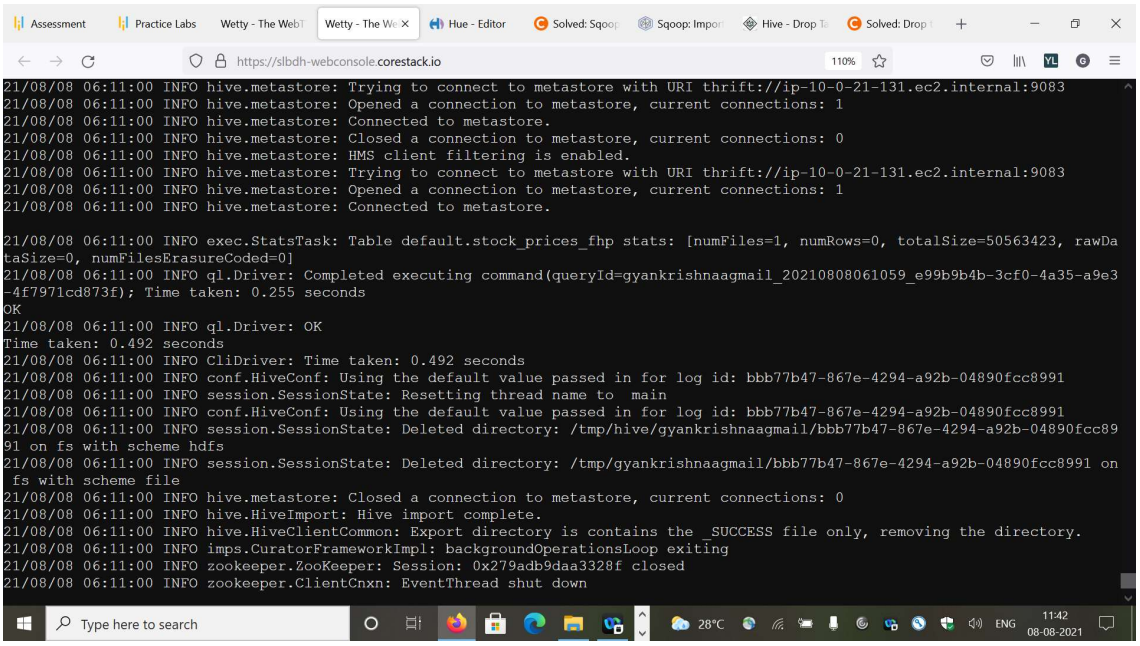
Stock_ Prices

```
MySQL [gyankrishnaagmail]> Select * from stock_prices;
+--------------+--------+-------------+-------------+-------------+-------------+-----------+
| trading_date | symbol | open        | close       | low         | high        | volume    |
+--------------+--------+-------------+-------------+-------------+-------------+-----------+
| 2016-01-05   | WLTW   |      123.43 |  125.839996 |  122.309998 |      126.25 |   2163600 |
| 2016-01-06   | WLTW   |  125.239998 |  119.980003 |  119.940002 |  125.540001 |   2386400 |
| 2016-01-07   | WLTW   |  116.379997 |  114.949997 |      114.93 |  119.739998 |   2489500 |
| 2016-01-08   | WLTW   |  115.480003 |  116.620003 |       113.5 |  117.440002 |   2006300 |
| 2016-01-11   | WLTW   |  117.010002 |  114.970001 |  114.089996 |  117.330002 |   1408600 |
| 2016-01-12   | WLTW   |  115.510002 |  115.550003 |       114.5 |  116.059998 |   1098000 |
| 2016-01-13   | WLTW   |  116.459999 |  112.849998 |  112.589996 |      117.07 |    949600 |
| 2016-01-14   | WLTW   |  113.510002 |  114.379997 |  110.050003 |  115.029999 |    785300 |
| 2016-01-15   | WLTW   |  113.330002 |  112.529999 |  111.919998 |  114.879997 |   1093700 |
| 2016-01-19   | WLTW   |  113.660004 |  110.379997 |  109.870003 |  115.870003 |   1523500 |
| 2016-01-20   | WLTW   |  109.059998 |  109.300003 |      108.32 |  111.599998 |   1653900 |
| 2016-01-21   | WLTW   |  109.730003 |         110 |      108.32 |  110.580002 |    944300 |
| 2016-01-22   | WLTW   |  111.879997 |  111.949997 |  110.190002 |  112.949997 |    744900 |
| 2016-01-25   | WLTW   |      111.32 |  110.120003 |         110 |  114.629997 |    703800 |
| 2016-01-26   | WLTW   |  110.419998 |         111 |  107.300003 |  111.400002 |    563100 |
| 2016-01-27   | WLTW   |  110.769997 |  110.709999 |  109.019997 |      112.57 |    896100 |
| 2016-01-28   | WLTW   |  110.900002 |  112.580002 |  109.900002 |  112.970001 |    680400 |
| 2016-01-29   | WLTW   |  113.349998 |  114.470001 |  111.669998 |  114.589996 |    749900 |
| 2016-02-01   | WLTW   |         114 |       114.5 |  112.900002 |  114.849998 |    574200 |
| 2016-02-02   | WLTW   |      113.25 |  110.559998 |       109.75 |  113.860001 |    694800 |
| 2016-02-03   | WLTW   |  113.379997 |  114.050003 |  109.639999 |  114.639999 |    896300 |
| 2016-02-04   | WLTW   |  114.080002 |  115.709999 |  114.080002 |      116.32 |    956300 |
| 2016-02-05   | WLTW   |  115.120003 |  114.019997 |  109.709999 |  116.489998 |    997100 |
| 2016-02-08   | WLTW   |  113.300003 |  111.160004 |  110.459999 |  113.300003 |   1200500 |
| 2016-02-09   | WLTW   |  111.169998 |  110.650002 |  109.639999 |  112.110001 |   1725200 |
```

Pulling data from RDBMS via sqoop into hive

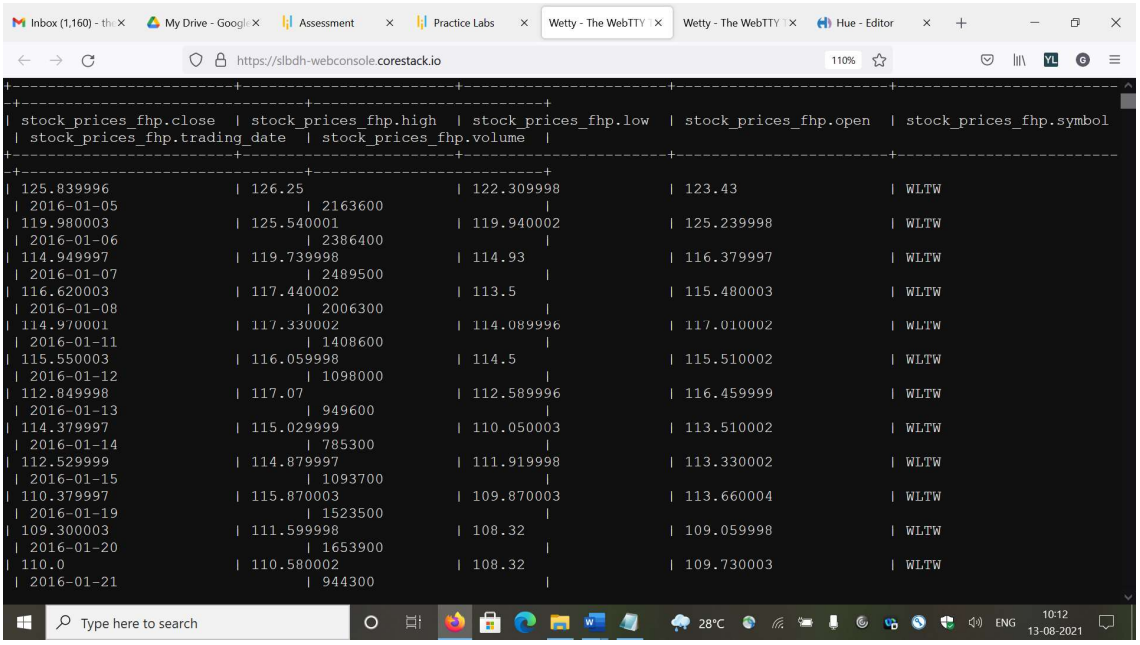
Stock_prices import

```
21/08/08 06:02:57 INFO hive.metastore: Connected to metastore.
[gyankrishnaagmail@ip-10-0-31-64 ~]$ sqoop import\
>  --connect jdbc:mysql://sqoopdb.slbdh.cloudlabs.com/gyankrishnaagmail\
>  --username gyankrishnaagmail\
>  --password 'gyankrishnaagmailtltyi'\
>  --split-by symbol\
>  --table stock_prices\
>  --fields-terminated-by ','\
>  --hive-import\
>  --hive-table stock_prices_fhp -m 1;
Warning: /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports w
ill fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/jars/slf4j-log4j12-1.7.25.jar!/org/slf
4j/impl/StaticLoggerBinder.class]
```




```
sqoop import\
--connect jdbc:mysql://sqoopdb.slbdh.cloudlabs.com/gyankrishnaagmail\
--username gyankrishnaagmail\
--password 'gyankrishnaagmailtltyi'\
--split-by symbol\
--table stock_prices\
--fields-terminated-by ','\
--hive-import\
--hive-table stock_prices_fhp -m 1;
```

Data after loading stock_prices

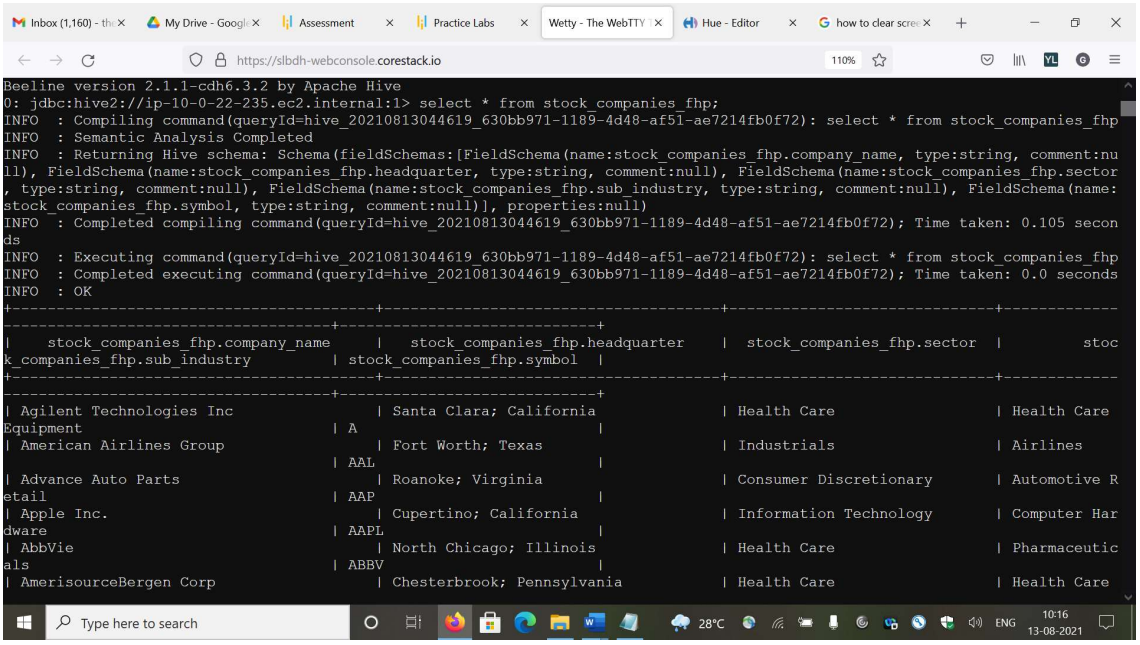


stock_prices_fhp.close	stock_prices_fhp.high	stock_prices_fhp.low	stock_prices_fhp.open	stock_prices_fhp.symbol	stock_prices_fhp.trading_date	stock_prices_fhp.volume
125.839996	126.25	122.309998	123.43	WLTW	2016-01-05	2163600
119.980003	125.540001	119.940002	125.239998	WLTW	2016-01-06	2386400
114.949997	119.739998	114.93	116.379997	WLTW	2016-01-07	2489500
116.620003	117.440002	113.5	115.480003	WLTW	2016-01-08	2006300
114.970001	117.330002	114.089996	117.010002	WLTW	2016-01-11	1408600
115.550003	116.059998	114.5	115.510002	WLTW	2016-01-12	1098000
112.849998	117.07	112.589996	116.459999	WLTW	2016-01-13	949600
114.379997	115.029999	110.050003	113.510002	WLTW	2016-01-14	785300
112.529999	114.879997	111.919998	113.330002	WLTW	2016-01-15	1093700
110.379997	115.870003	109.870003	113.660004	WLTW	2016-01-19	1523500
109.300003	111.599998	108.32	109.059998	WLTW	2016-01-20	1653900
110.0	110.580002	108.32	109.730003	WLTW	2016-01-21	944300

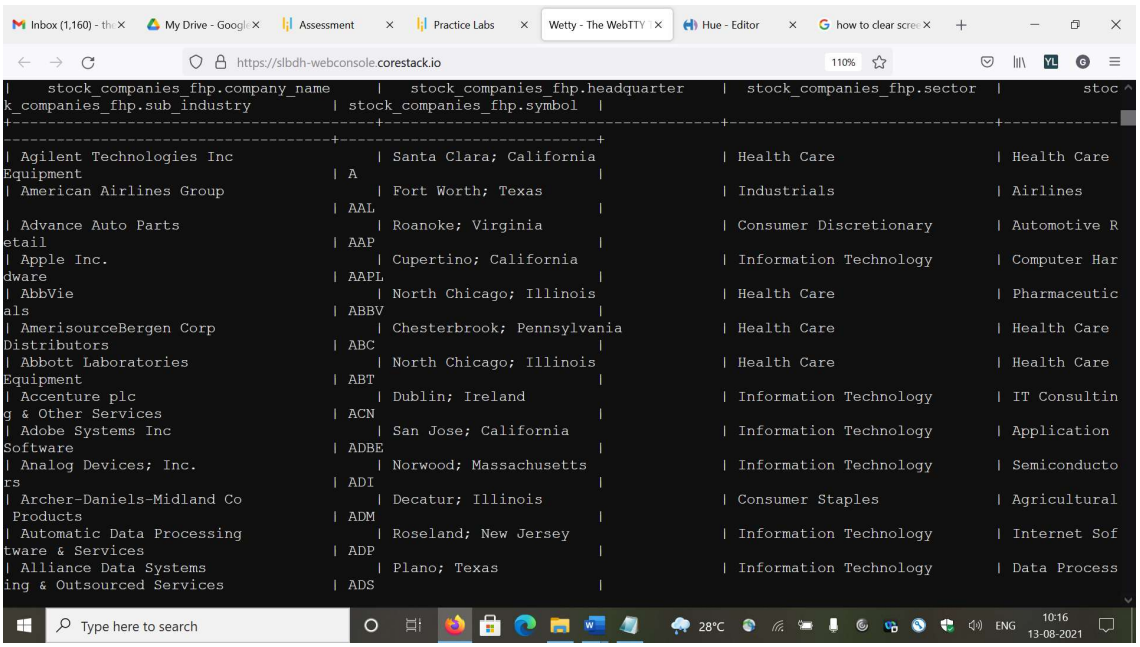
Similarly doing for other tables

```
sqoop import\
 --connect jdbc:mysql://sqoopdb.slbdh.cloudlabs.com/gyankrishnaagmail\
 --username gyankrishnaagmail\
 --password 'gyankrishnaagmailtltyi'\
 --split-by symbol\
 --table stock_companies\
 --fields-terminated-by ','\
 --hive-import\
 --hive-table stock_companies_fhp -m 1;
```

Data after loading stock_companies



```
Beeline version 2.1.1-cdh6.3.2 by Apache Hive
0: jdbc:hive2://ip-10-0-22-235.ec2.internal:1> select * from stock_companies_fhp;
INFO  : Compiling command(queryId=hive_20210813044619_630bb971-1189-4d48-af51-ae7214fb0f72): select * from stock_companies_fhp
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:stock_companies_fhp.company_name, type:string, comment:nu
ll), FieldSchema(name:stock_companies_fhp.headquarter, type:string, comment:null), FieldSchema(name:stock_companies_fhp.sector
, type:string, comment:null), FieldSchema(name:stock_companies_fhp.sub_industry, type:string, comment:null), FieldSchema(name:
stock_companies_fhp.symbol, type:string, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20210813044619_630bb971-1189-4d48-af51-ae7214fb0f72); Time taken: 0.105 secon
ds
INFO  : Executing command(queryId=hive_20210813044619_630bb971-1189-4d48-af51-ae7214fb0f72): select * from stock_companies_fhp
INFO  : Completed executing command(queryId=hive_20210813044619_630bb971-1189-4d48-af51-ae7214fb0f72); Time taken: 0.0 seconds
INFO  : OK
+-----------------------------------------+-----------------------------------------+-----------------------------+-------------
-----------------------------------+-----------------------------+
|    stock_companies_fhp.company_name     |   stock_companies_fhp.headquarter       |  stock_companies_fhp.sector |         stoc
k_companies_fhp.sub_industry         | stock_companies_fhp.symbol  |
+-----------------------------------------+-----------------------------------------+-----------------------------+-------------
-----------------------------------+-----------------------------+
| Agilent Technologies Inc                | Santa Clara; California                 | Health Care                 | Health Care
Equipment                            | A                           |
| American Airlines Group                 | Fort Worth; Texas                       | Industrials                 | Airlines
                                     | AAL                         |
| Advance Auto Parts                      | Roanoke; Virginia                       | Consumer Discretionary      | Automotive R
etail                                | AAP                         |
| Apple Inc.                              | Cupertino; California                   | Information Technology      | Computer Har
dware                                | AAPL                        |
| AbbVie                                  | North Chicago; Illinois                 | Health Care                 | Pharmaceutic
als                                  | ABBV                        |
| AmerisourceBergen Corp                  | Chesterbrook; Pennsylvania              | Health Care                 | Health Care
```

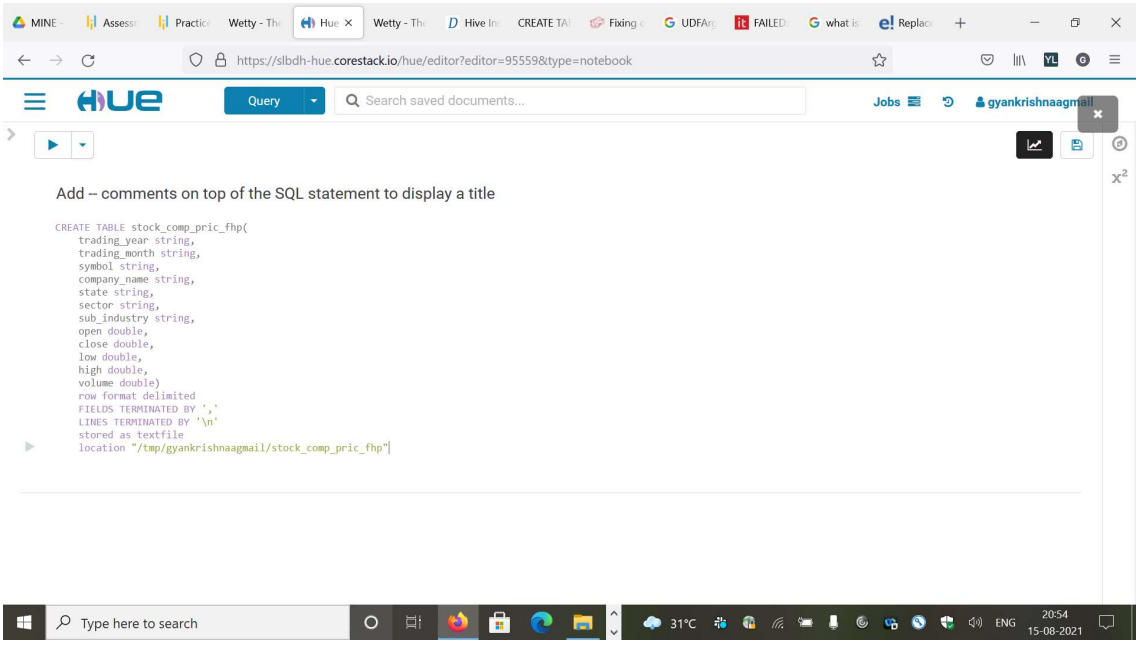


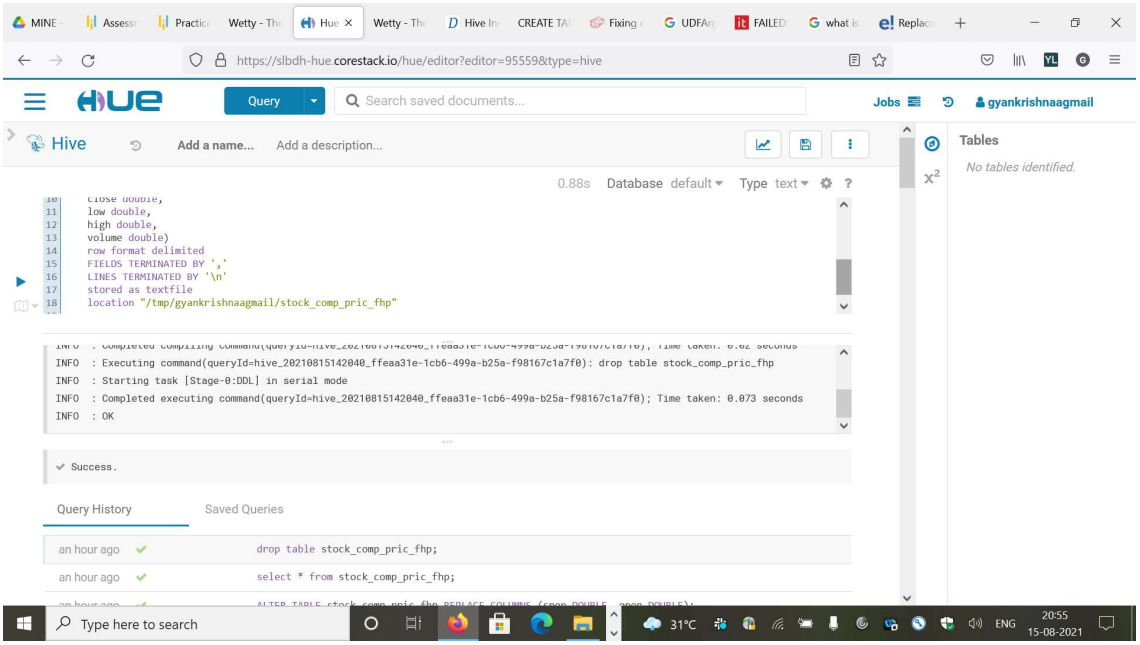
```
|    stock_companies_fhp.company_name     |   stock_companies_fhp.headquarter       |  stock_companies_fhp.sector |         stoc
k_companies_fhp.sub_industry         | stock_companies_fhp.symbol  |
+-----------------------------------------+-----------------------------------------+-----------------------------+-------------
-----------------------------------+-----------------------------+
| Agilent Technologies Inc                | Santa Clara; California                 | Health Care                 | Health Care
Equipment                            | A                           |
| American Airlines Group                 | Fort Worth; Texas                       | Industrials                 | Airlines
                                     | AAL                         |
| Advance Auto Parts                      | Roanoke; Virginia                       | Consumer Discretionary      | Automotive R
etail                                | AAP                         |
| Apple Inc.                              | Cupertino; California                   | Information Technology      | Computer Har
dware                                | AAPL                        |
| AbbVie                                  | North Chicago; Illinois                 | Health Care                 | Pharmaceutic
als                                  | ABBV                        |
| AmerisourceBergen Corp                  | Chesterbrook; Pennsylvania              | Health Care                 | Health Care
Distributors                         | ABC                         |
| Abbott Laboratories                     | North Chicago; Illinois                 | Health Care                 | Health Care
Equipment                            | ABT                         |
| Accenture plc                           | Dublin; Ireland                         | Information Technology      | IT Consultin
g & Other Services                   | ACN                         |
| Adobe Systems Inc                       | San Jose; California                    | Information Technology      | Application
Software                             | ADBE                        |
| Analog Devices; Inc.                    | Norwood; Massachusetts                  | Information Technology      | Semiconducto
rs                                   | ADI                         |
| Archer-Daniels-Midland Co               | Decatur; Illinois                       | Consumer Staples            | Agricultural
 Products                            | ADM                         |
| Automatic Data Processing               | Roseland; New Jersey                    | Information Technology      | Internet Sof
tware & Services                     | ADP                         |
| Alliance Data Systems                   | Plano; Texas                            | Information Technology      | Data Process
ing & Outsourced Services            | ADS                         |
```

2) Create a new hive table with the following fields by joining the above two hive tables. Please use appropriate Hive built-in functions for columns (a,b,e and h to l).

Creating table having all the fields from other two hive tables

```
CREATE TABLE stock_comp_pric_fhp(
   trading_year string,
   trading_month string,
   symbol string,
   company_name string,
   state string,
   sector string,
   sub_industry string,
   open double,
   close double,
   low double,
   high double,
   volume double)
   row format delimited
   FIELDS TERMINATED BY ','
   LINES TERMINATED BY '\n'
   stored as textfile
   location "/tmp/gyankrishnaagmail/stock_comp_pric_fhp"
```





Inserting data into the table using appropriate hive functions

INSERT INTO stock_comp_pric_fhp

SELECT year(P.trading_date) ,

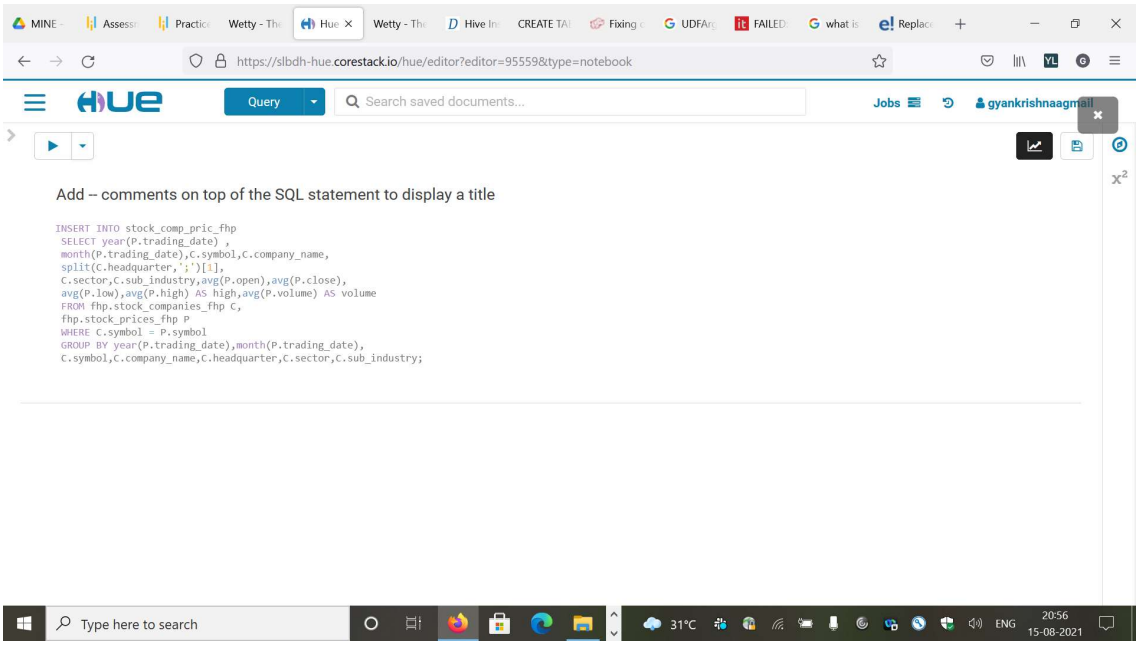
month(P.trading_date),C.symbol,C.company_name,

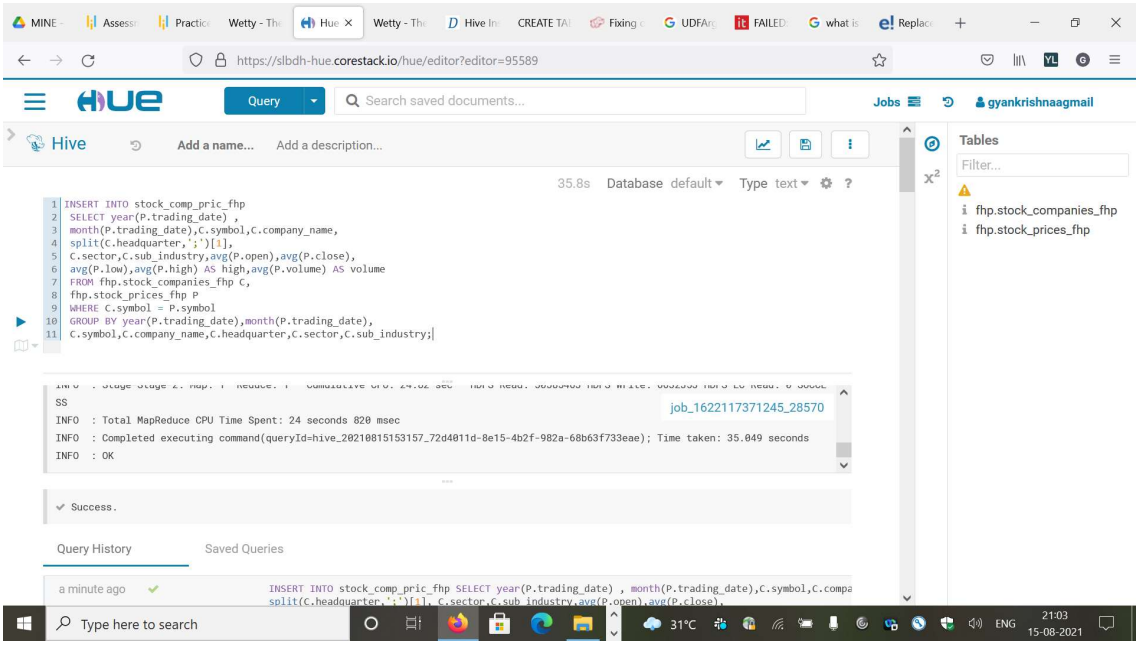
split(C.headquarter,';')[1],

C.sector,C.sub_industry,avg(P.open),avg(P.close),

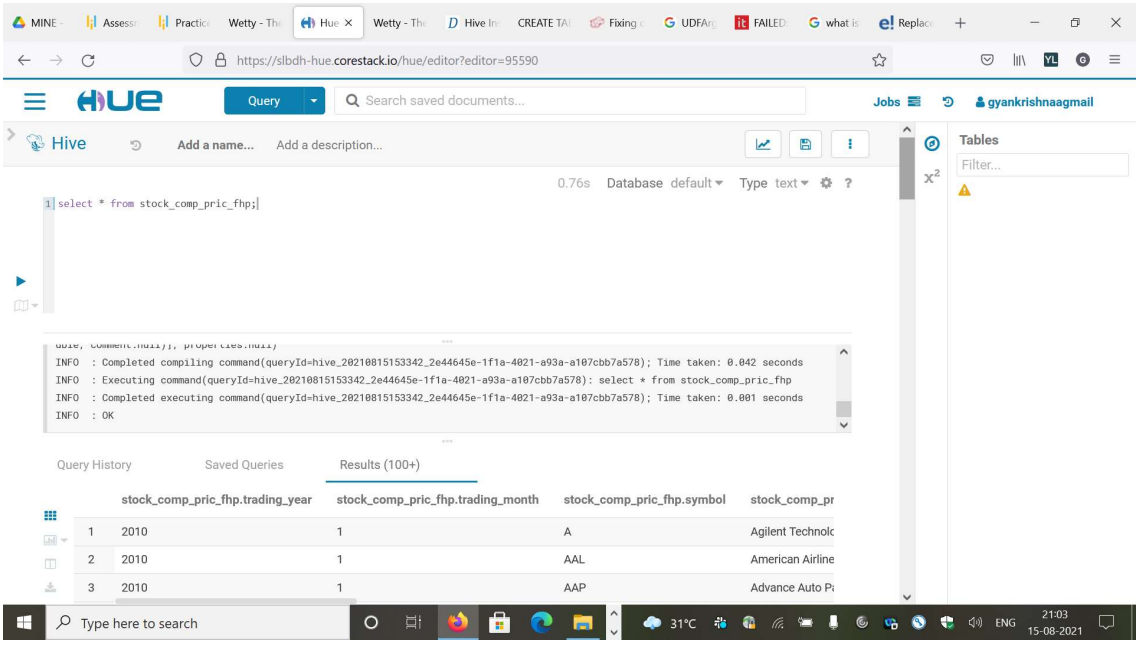
avg(P.low),avg(P.high) AS high,avg(P.volume) AS volume

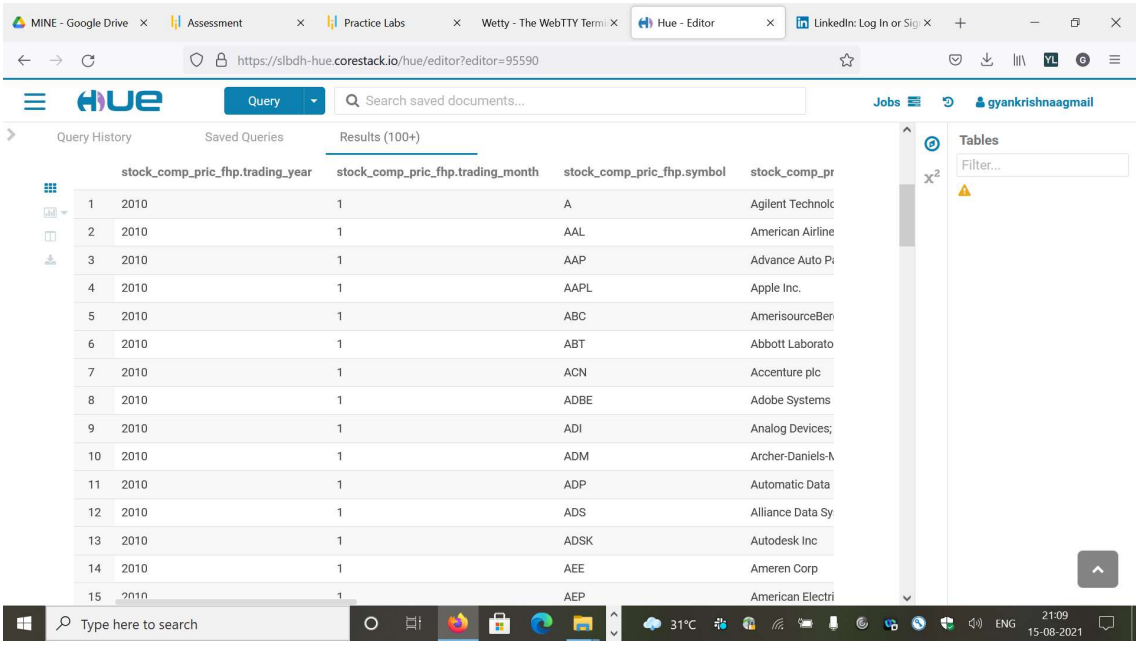
```
FROM fhp.stock_companies_fhp C,
fhp.stock_prices_fhp P
WHERE C.symbol = P.symbol
GROUP BY year(P.trading_date),month(P.trading_date),
C.symbol,C.company_name,C.headquarter,C.sector,C.sub_industry;
```





Selecting all data after loading the data





All the data are also present in the excel given below

query-hive-95604.csv

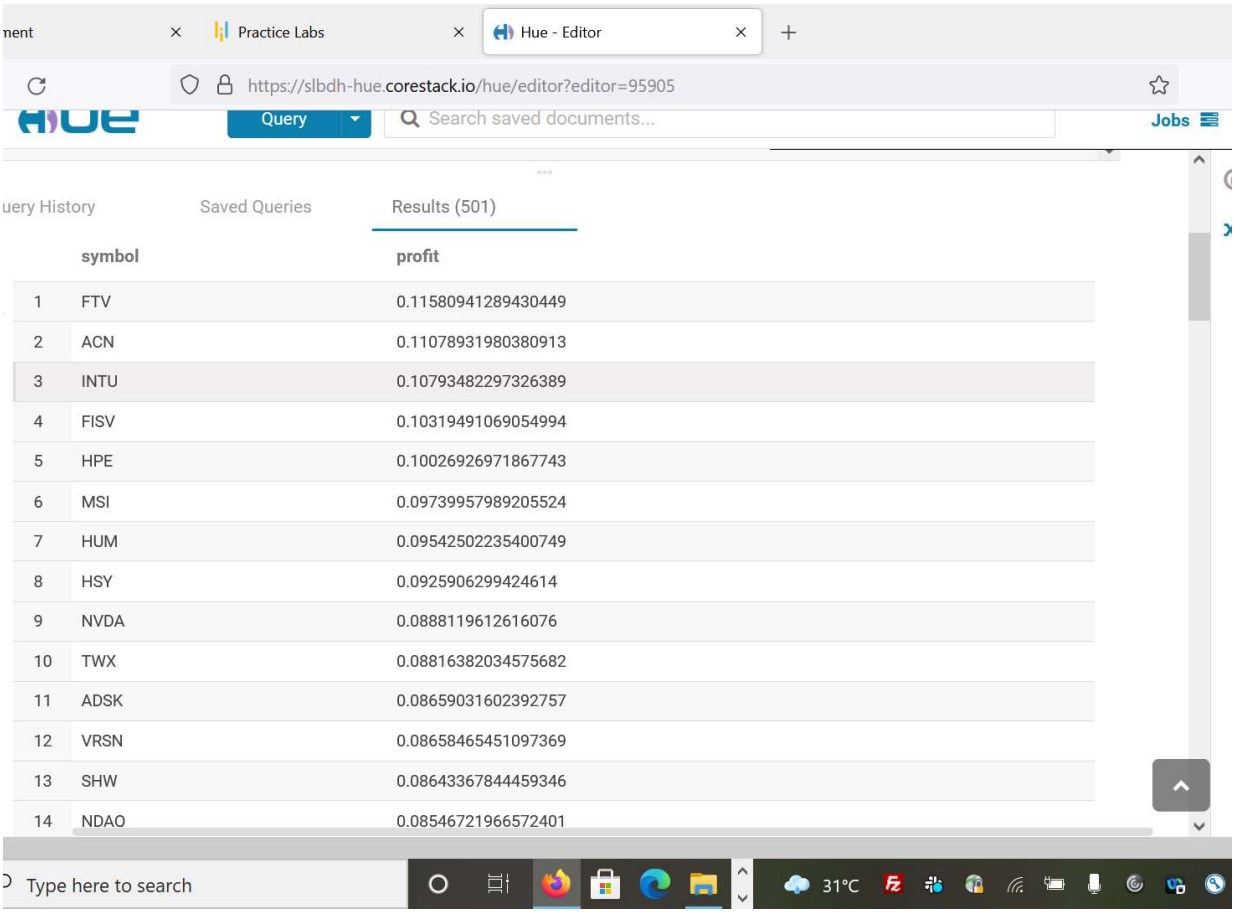
3) Find the top five companies that are good for investment

Finding the profit over the entire year , by using the below formula

SELECT symbol,(((sum(close)-sum(open))/sum(open))*100) AS profit FROM stock_comp_pric_fhp

GROUP BY symbol

SORT BY profit DESC;



	symbol	profit
1	FTV	0.11580941289430449
2	ACN	0.11078931980380913
3	INTU	0.10793482297326389
4	FISV	0.10319491069054994
5	HPE	0.10026926971867743
6	MSI	0.09739957989205524
7	HUM	0.09542502235400749
8	HSY	0.0925906299424614
9	NVDA	0.0888119612616076
10	TWX	0.08816382034575682
11	ADSK	0.08659031602392757
12	VRSN	0.08658465451097369
13	SHW	0.08643367844459346
14	NDAQ	0.08546721966572401

The above are the top companies which are profitable as there closing prices are higher than opening prices.

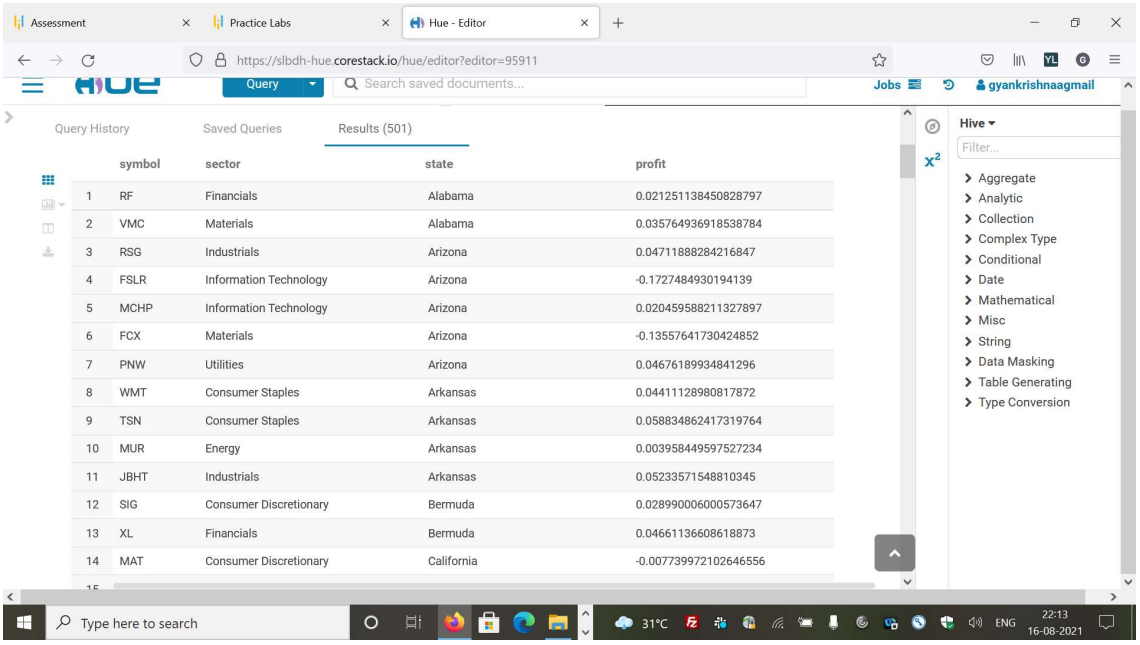
4) Show the best-growing industry by each state, having at least two or more industries mapped.

The best industry would be with highest profit in a state and sector. So, Below queries will extract the same

SELECT symbol,sector,state,(((sum(close)-sum(open))/sum(open))*100) AS profit FROM stock_comp_pric_fhp

GROUP BY symbol,sector,state

SORT BY state,sector,profit;



	symbol	sector	state	profit
1	RF	Financials	Alabama	0.021251138450828797
2	VMC	Materials	Alabama	0.035764936918538784
3	RSG	Industrials	Arizona	0.04711888284216847
4	FSLR	Information Technology	Arizona	-0.1727484930194139
5	MCHP	Information Technology	Arizona	0.020459588211327897
6	FCX	Materials	Arizona	-0.13557641730424852
7	PNW	Utilities	Arizona	0.04676189934841296
8	WMT	Consumer Staples	Arkansas	0.04411128980817872
9	TSN	Consumer Staples	Arkansas	0.058834862417319764
10	MUR	Energy	Arkansas	0.003958449597527234
11	JBHT	Industrials	Arkansas	0.05233571548810345
12	SIG	Consumer Discretionary	Bermuda	0.028990006000573647
13	XL	Financials	Bermuda	0.04661136608618873
14	MAT	Consumer Discretionary	California	-0.007739972102646556

The excel below contains all the data as shown in excel

query-hive-95911.csv

5) For each sector find the following.

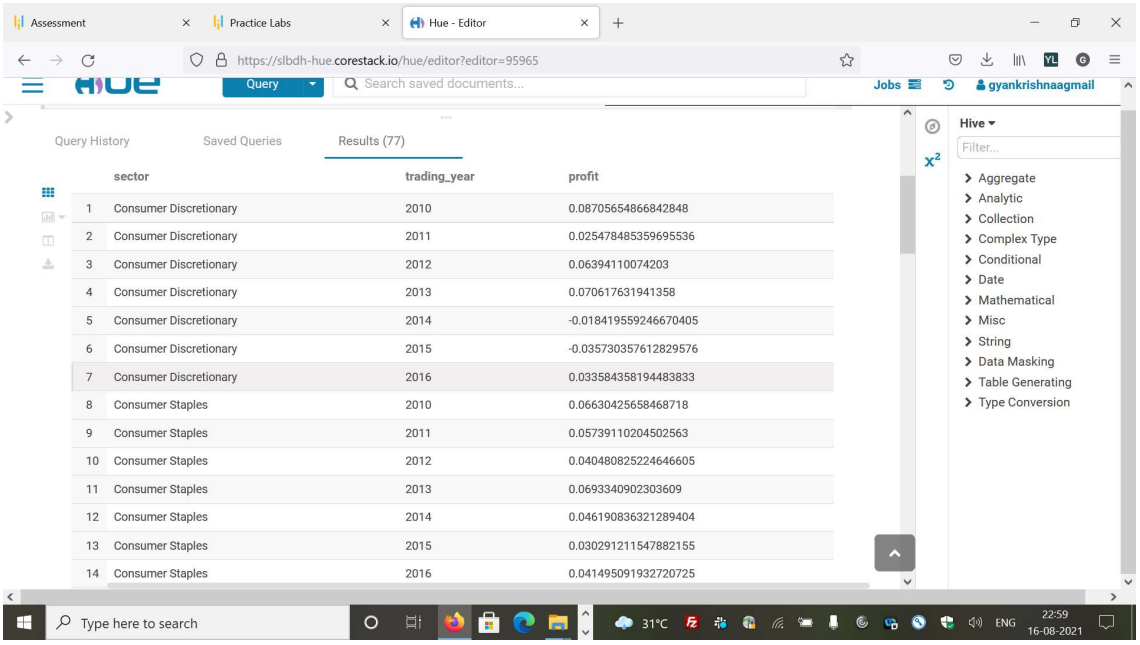
- Worst year
- b. Best year
- c. Stable year

For best , worst year, stable year , using the below queries.

SELECT sector,trading_year,(((sum(close)-sum(open))/sum(open))*100) AS profit FROM stock_comp_pric_fhp

GROUP BY sector,trading_year

ORDER BY sector, trading_year;



	sector	trading_year	profit
1	Consumer Discretionary	2010	0.08705654866842848
2	Consumer Discretionary	2011	0.025478485359695536
3	Consumer Discretionary	2012	0.06394110074203
4	Consumer Discretionary	2013	0.070617631941358
5	Consumer Discretionary	2014	-0.018419559246670405
6	Consumer Discretionary	2015	-0.035730357612829576
7	Consumer Discretionary	2016	0.033584358194483833
8	Consumer Staples	2010	0.06630425658468718
9	Consumer Staples	2011	0.05739110204502563
10	Consumer Staples	2012	0.040480825224646605
11	Consumer Staples	2013	0.0693340902303609
12	Consumer Staples	2014	0.046190836321289404
13	Consumer Staples	2015	0.030291211547882155
14	Consumer Staples	2016	0.041495091932720725

The below extract contains all the data from the query

query-hive-95965.csv