`***Big Data Hadoop and Spark Developer: **Project_Presentation*****`

In partial fulfillment of Simplilearn Master Data Science Certification course.

Due Date: Feb 17 2021

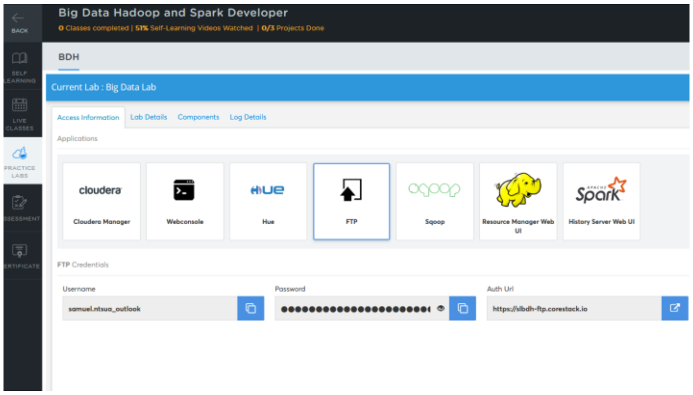
Project name: Stock Exchange Data Analysis

Modeler and presenter : ***Samuel_Y._Ntsua***

Trainer and Mentor: ***Ajaykuma***

Creating the data pipeline:

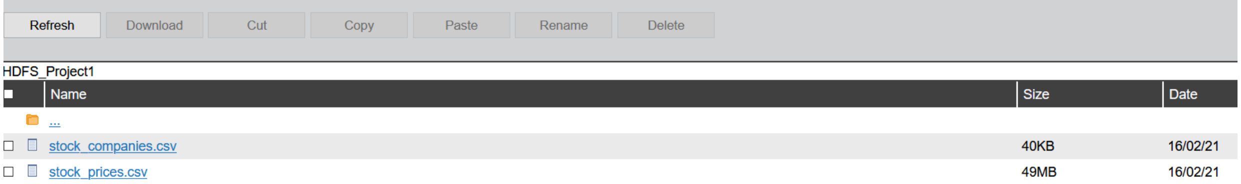
Use FTP to upload csv data from local desktop to LMS:



Username: samuel.ntsua_outlook

Password: ●●●●●●●●●●●●●●●●●●●●●

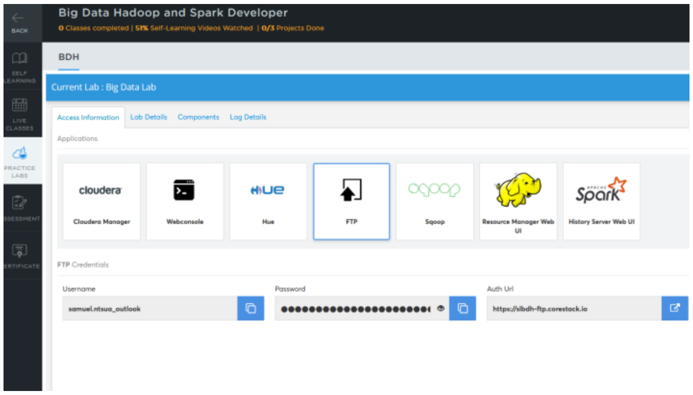
Auth Url: https://slbdh-ftp.corestack.io



HDFS_Project1

■	Name	Size	Date
	...		
□	stock_companies.csv	40KB	16/02/21
□	stock_prices.csv	49MB	16/02/21

Since the plan is to use Sqoop to move the data into Hive, I logon to the server hosting MySQL and Sqoop:

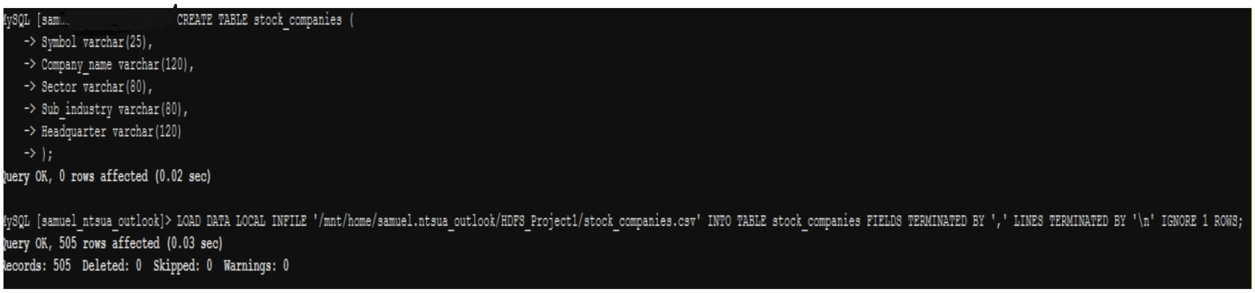


Username: samuel.ntsua_outlook

Password: ●●●●●●●●●●●●●●●●●●●●●

Auth Url: https://slbdh-ftp.corestack.io

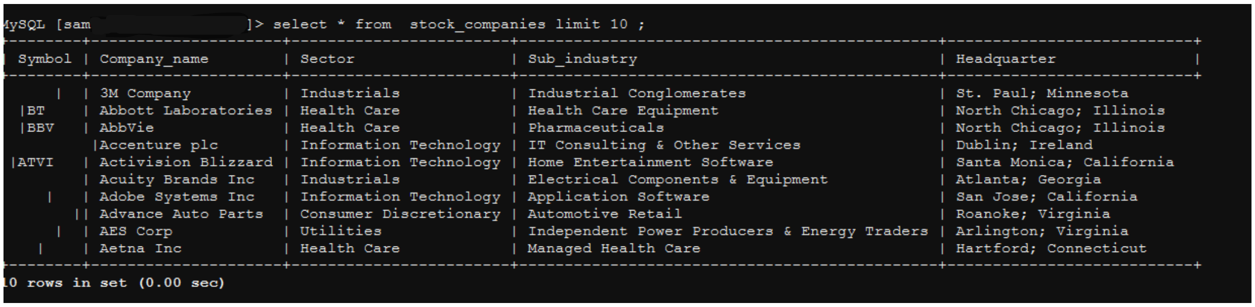
Create a table in MySQL, then load the STOCK_PRICES.csv and STOCK_COMPANIES.csv:



```
MySQL [sam[redacted] CREATE TABLE stock_companies (
    -> Symbol varchar(25),
    -> Company_name varchar(120),
    -> Sector varchar(80),
    -> Sub_industry varchar(80),
    -> Headquarter varchar(120)
    -> );
Query OK, 0 rows affected (0.02 sec)

MySQL [samuel_ntsua_outlook]> LOAD DATA LOCAL INFILE '/mnt/home/samuel.ntsua_outlook/HDFS_Project1/stock_companies.csv' INTO TABLE stock_companies FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' IGNORE 1 ROWS;
Query OK, 505 rows affected (0.03 sec)
Records: 505  Deleted: 0  Skipped: 0  Warnings: 0
```

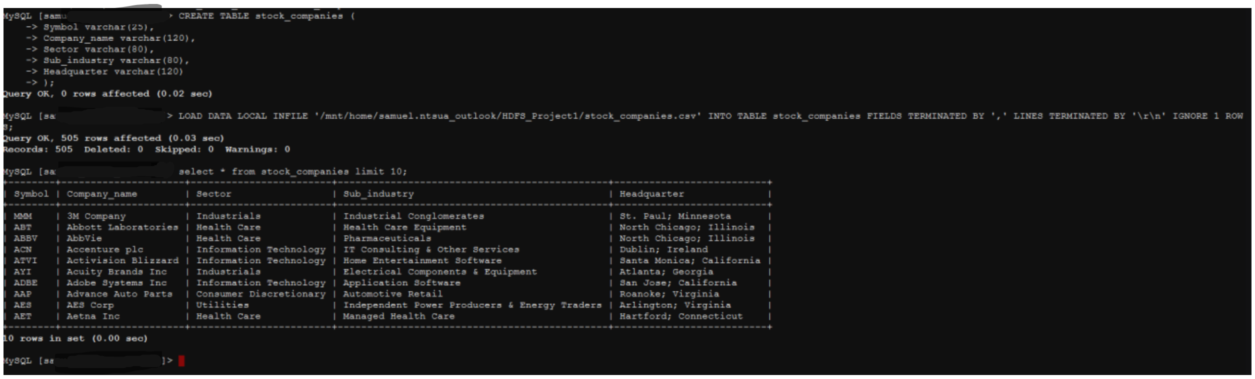
Below is the STOCK_COMPANIES table in MySQL.

A closer look at the first column shows the first letter is cut off. I suspect it is end-of-line issue. To fix that, I dropped the table then re-load the table by changing the option line-terminated from '\n' to '\r\n.



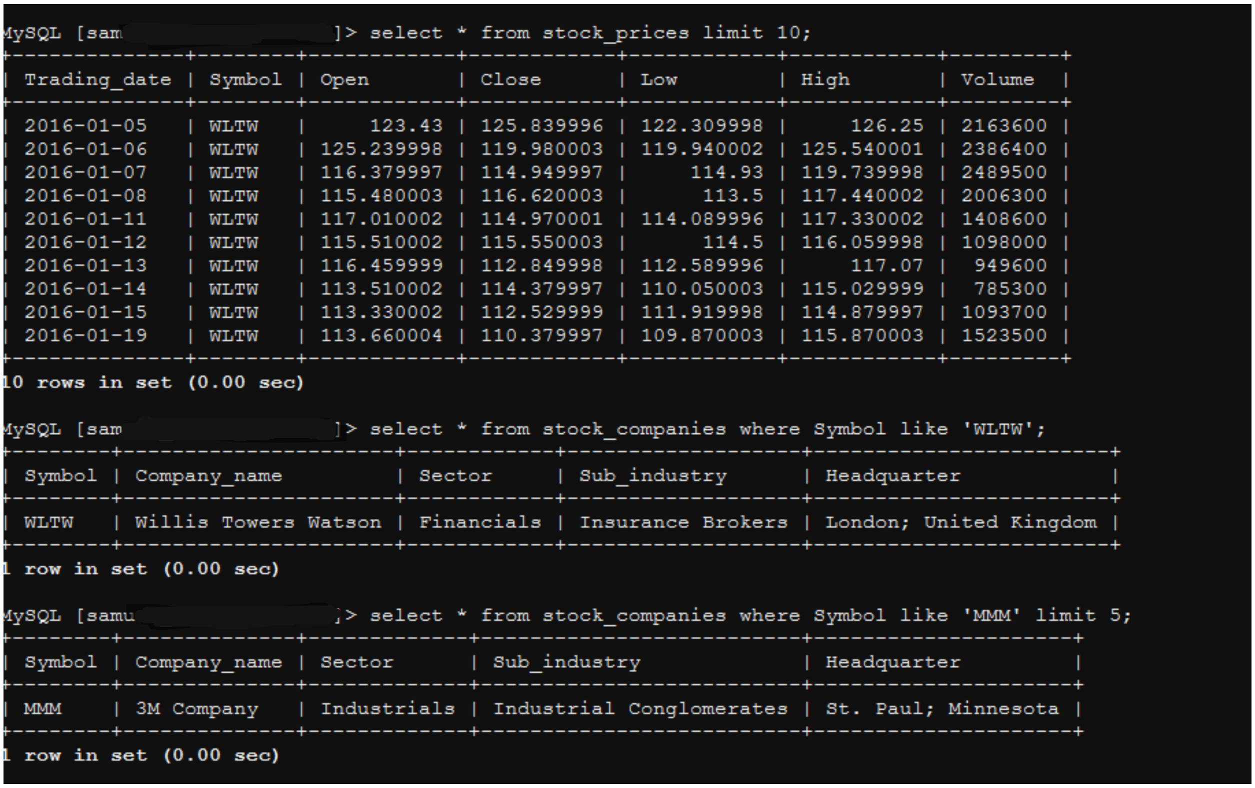
```
MySQL [sam[redacted]]> select * from  stock_companies limit 10 ;
+--------+----------------------+------------------------+------------------------------------------------+--------------------------+
| Symbol | Company_name         | Sector                 | Sub_industry                                   | Headquarter              |
+--------+----------------------+------------------------+------------------------------------------------+--------------------------+
     |  | 3M Company           | Industrials            | Industrial Conglomerates                       | St. Paul; Minnesota
 |BT    | Abbott Laboratories  | Health Care            | Health Care Equipment                          | North Chicago; Illinois
 |BBV   | AbbVie               | Health Care            | Pharmaceuticals                                | North Chicago; Illinois
        |Accenture plc         | Information Technology | IT Consulting & Other Services                 | Dublin; Ireland
|ATVI   | Activision Blizzard  | Information Technology | Home Entertainment Software                    | Santa Monica; California
        | Acuity Brands Inc    | Industrials            | Electrical Components & Equipment              | Atlanta; Georgia
    |   | Adobe Systems Inc    | Information Technology | Application Software                           | San Jose; California
       || Advance Auto Parts   | Consumer Discretionary | Automotive Retail                              | Roanoke; Virginia
     |  | AES Corp             | Utilities              | Independent Power Producers & Energy Traders   | Arlington; Virginia
   |    | Aetna Inc            | Health Care            | Managed Health Care                            | Hartford; Connecticut
+--------+----------------------+------------------------+------------------------------------------------+--------------------------+
10 rows in set (0.00 sec)
```

It worked! The Symbol are now well aligned in the table below.



```
MySQL [samu[redacted]> CREATE TABLE stock_companies (
    -> Symbol varchar(25),
    -> Company_name varchar(120),
    -> Sector varchar(80),
    -> Sub_industry varchar(80),
    -> Headquarter varchar(120)
    -> );
Query OK, 0 rows affected (0.02 sec)

MySQL [sa[redacted]> LOAD DATA LOCAL INFILE '/mnt/home/samuel.ntsua_outlook/HDFS_Project1/stock_companies.csv' INTO TABLE stock_companies FIELDS TERMINATED BY ',' LINES TERMINATED BY '\r\n' IGNORE 1 ROW
S;
Query OK, 505 rows affected (0.03 sec)
Records: 505  Deleted: 0  Skipped: 0  Warnings: 0

MySQL [sa[redacted] select * from stock_companies limit 10;
+--------+----------------------+------------------------+----------------------------------------------+--------------------------+
| Symbol | Company_name         | Sector                 | Sub_industry                                 | Headquarter              |
+--------+----------------------+------------------------+----------------------------------------------+--------------------------+
| MMM    | 3M Company           | Industrials            | Industrial Conglomerates                     | St. Paul; Minnesota      |
| ABT    | Abbott Laboratories  | Health Care            | Health Care Equipment                        | North Chicago; Illinois  |
| ABBV   | AbbVie               | Health Care            | Pharmaceuticals                              | North Chicago; Illinois  |
| ACN    | Accenture plc        | Information Technology | IT Consulting & Other Services               | Dublin; Ireland          |
| ATVI   | Activision Blizzard  | Information Technology | Home Entertainment Software                  | Santa Monica; California |
| AYI    | Acuity Brands Inc    | Industrials            | Electrical Components & Equipment            | Atlanta; Georgia         |
| ADBE   | Adobe Systems Inc    | Information Technology | Application Software                         | San Jose; California     |
| AAP    | Advance Auto Parts   | Consumer Discretionary | Automotive Retail                            | Roanoke; Virginia        |
| AES    | AES Corp             | Utilities              | Independent Power Producers & Energy Traders | Arlington; Virginia      |
| AET    | Aetna Inc            | Health Care            | Managed Health Care                          | Hartford; Connecticut    |
+--------+----------------------+------------------------+----------------------------------------------+--------------------------+
10 rows in set (0.00 sec)

MySQL [sa[redacted]]>
```

Below, I just checked on some random rows in the tables to make sure the files was properly loaded into the tables.



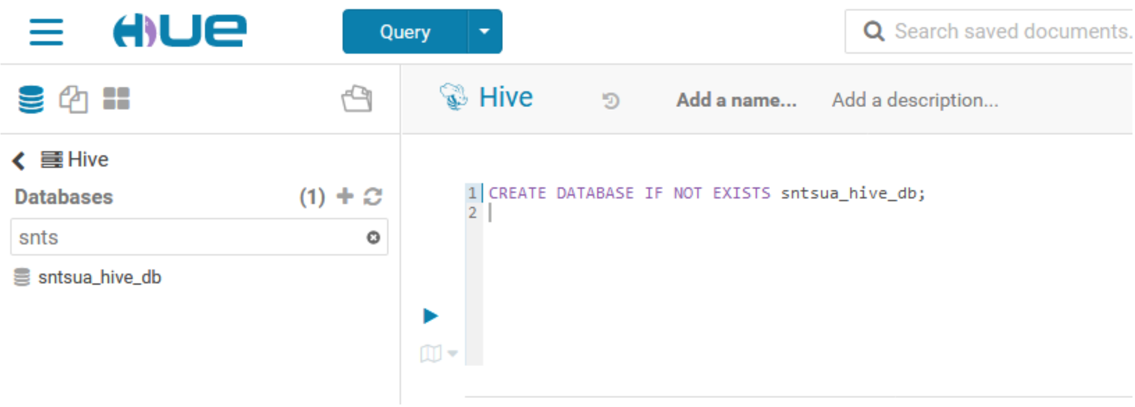
```
MySQL [sam[redacted]]> select * from stock_prices limit 10;
+--------------+--------+------------+------------+------------+------------+---------+
| Trading_date | Symbol | Open       | Close      | Low        | High       | Volume  |
+--------------+--------+------------+------------+------------+------------+---------+
| 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 |
+--------------+--------+------------+------------+------------+------------+---------+
10 rows in set (0.00 sec)

MySQL [sam[redacted]]> select * from stock_companies where Symbol like 'WLTW';
+--------+----------------------+------------+-------------------+-------------------------+
| Symbol | Company_name         | Sector     | Sub_industry      | Headquarter             |
+--------+----------------------+------------+-------------------+-------------------------+
| WLTW   | Willis Towers Watson | Financials | Insurance Brokers | London; United Kingdom |
+--------+----------------------+------------+-------------------+-------------------------+
1 row in set (0.00 sec)

MySQL [samu[redacted]]> select * from stock_companies where Symbol like 'MMM' limit 5;
+--------+--------------+-------------+---------------------------+---------------------+
| Symbol | Company_name | Sector      | Sub_industry              | Headquarter         |
+--------+--------------+-------------+---------------------------+---------------------+
| MMM    | 3M Company   | Industrials | Industrial Conglomerates | St. Paul; Minnesota |
+--------+--------------+-------------+---------------------------+---------------------+
1 row in set (0.00 sec)
```

Now that we have the two tables straighten up, and can be queried by Sqoop, we can now Sqoop them to Hive.

Since we did not set PRIMARY KEY in the tables, Sqoop will complain because Sqoop uses the key to “split” the file to load. We can tell Sqoop to load the file without splitting it by   passing the argument –autoreset-to-one-mapper.

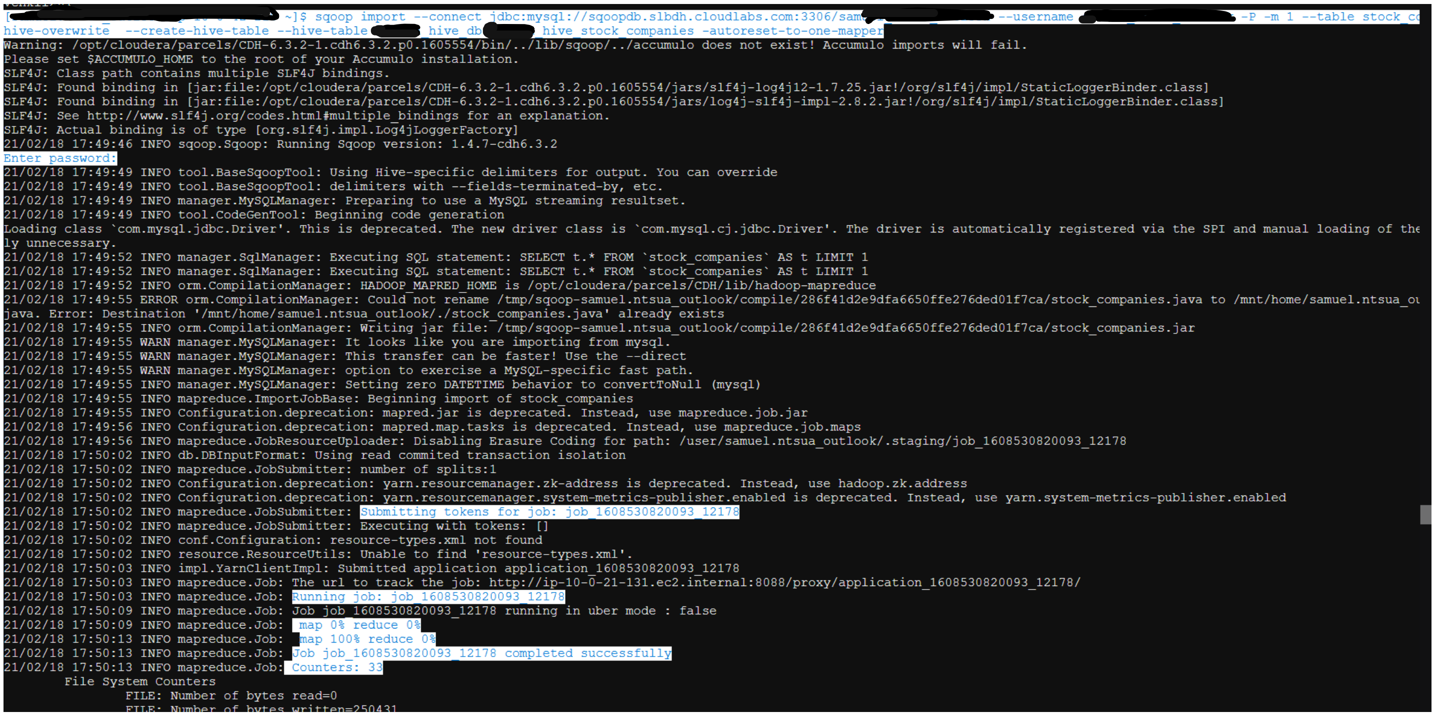
But before moving the tables to Hive, I create a database in Hive where I will store my tables.

I opted for this because I do not want Sqoop to put my tables in the default database. So I will specify a directory where Sqoop will put the tables.

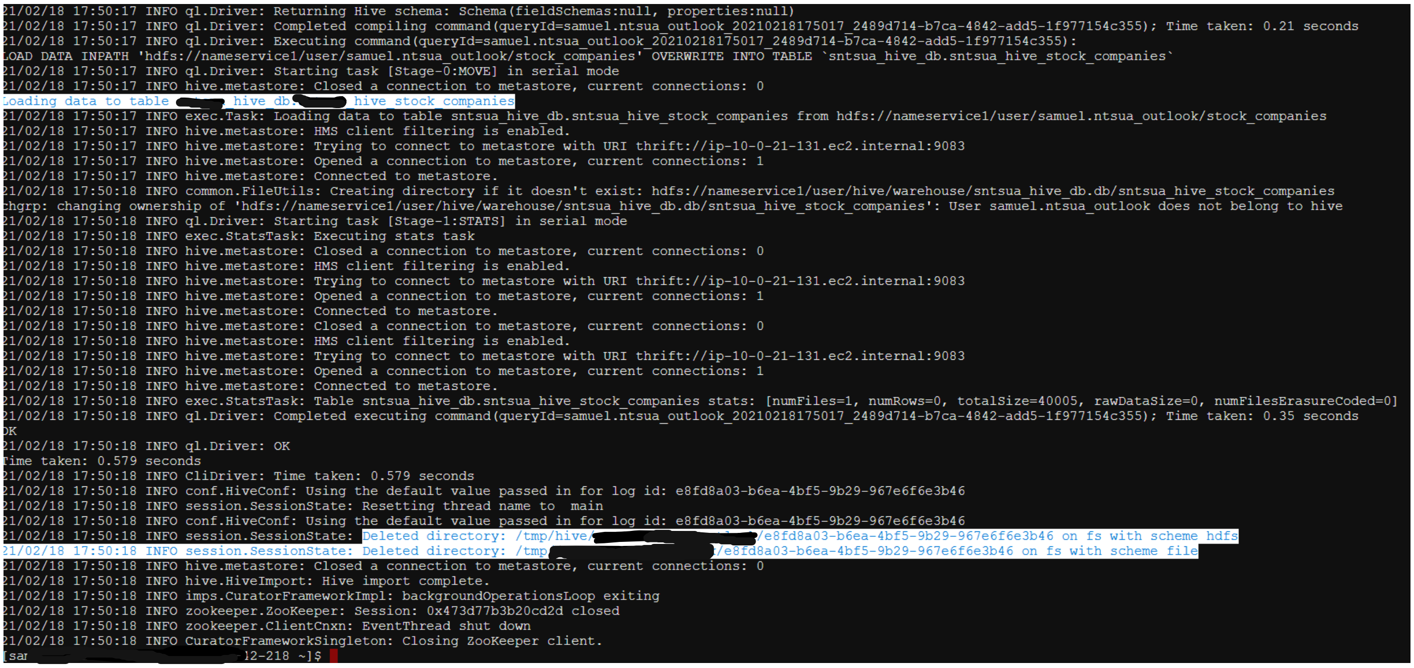


```
CREATE DATABASE IF NOT EXISTS sntsua_hive_db;
```

Now let’s Sqoop the tables to Hive. In the next two screenshot I have highlighted the Sqoop parameters used to transfer STOCK_COMPANIES , as well as some key output that show the transfer has been successful.



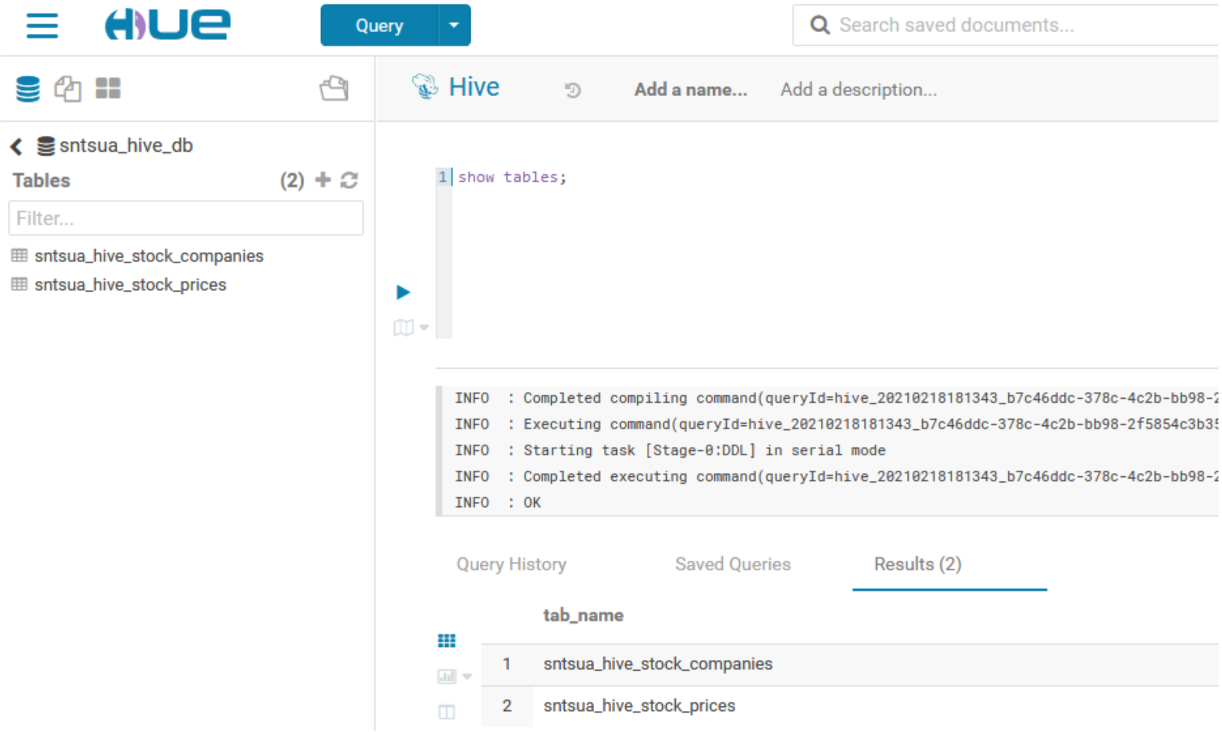
```
[[redacted] ~]$ sqoop import --connect jdbc:mysql://sqoopdb.slbdh.cloudlabs.com:3306/sam[redacted] --username [redacted] -P -m 1 --table stock_co
hive-overwrite  --create-hive-table --hive-table [redacted] hive_db[redacted] hive_stock_companies -autoreset-to-one-mapper
Warning: /opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports will 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/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/jars/log4j-slf4j-impl-2.8.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
21/02/18 17:49:46 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7-cdh6.3.2
Enter password:
21/02/18 17:49:49 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
21/02/18 17:49:49 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
21/02/18 17:49:49 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
21/02/18 17:49:49 INFO tool.CodeGenTool: Beginning code generation
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the
ly unnecessary.
21/02/18 17:49:52 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `stock_companies` AS t LIMIT 1
21/02/18 17:49:52 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `stock_companies` AS t LIMIT 1
21/02/18 17:49:52 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce
21/02/18 17:49:55 ERROR orm.CompilationManager: Could not rename /tmp/sqoop-samuel.ntsua_outlook/compile/286f41d2e9dfa6650ffe276ded01f7ca/stock_companies.java to /mnt/home/samuel.ntsua_ou
java. Error: Destination '/mnt/home/samuel.ntsua_outlook/./stock_companies.java' already exists
21/02/18 17:49:55 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-samuel.ntsua_outlook/compile/286f41d2e9dfa6650ffe276ded01f7ca/stock_companies.jar
21/02/18 17:49:55 WARN manager.MySQLManager: It looks like you are importing from mysql.
21/02/18 17:49:55 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
21/02/18 17:49:55 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
21/02/18 17:49:55 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
21/02/18 17:49:55 INFO mapreduce.ImportJobBase: Beginning import of stock_companies
21/02/18 17:49:55 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
21/02/18 17:49:56 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
21/02/18 17:49:56 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/samuel.ntsua_outlook/.staging/job_1608530820093_12178
21/02/18 17:50:02 INFO db.DBInputFormat: Using read commited transaction isolation
21/02/18 17:50:02 INFO mapreduce.JobSubmitter: number of splits:1
21/02/18 17:50:02 INFO Configuration.deprecation: yarn.resourcemanager.zk-address is deprecated. Instead, use hadoop.zk.address
21/02/18 17:50:02 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
21/02/18 17:50:02 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1608530820093_12178
21/02/18 17:50:02 INFO mapreduce.JobSubmitter: Executing with tokens: []
21/02/18 17:50:02 INFO conf.Configuration: resource-types.xml not found
21/02/18 17:50:02 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/02/18 17:50:03 INFO impl.YarnClientImpl: Submitted application application_1608530820093_12178
21/02/18 17:50:03 INFO mapreduce.Job: The url to track the job: http://ip-10-0-21-131.ec2.internal:8088/proxy/application_1608530820093_12178/
21/02/18 17:50:03 INFO mapreduce.Job: Running job: job_1608530820093_12178
21/02/18 17:50:09 INFO mapreduce.Job: Job job_1608530820093_12178 running in uber mode : false
21/02/18 17:50:09 INFO mapreduce.Job:  map 0% reduce 0%
21/02/18 17:50:13 INFO mapreduce.Job:  map 100% reduce 0%
21/02/18 17:50:13 INFO mapreduce.Job: Job job_1608530820093_12178 completed successfully
21/02/18 17:50:13 INFO mapreduce.Job: Counters: 33
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=250431
```



```
21/02/18 17:50:17 INFO ql.Driver: Returning Hive schema: Schema(fieldSchemas:null, properties:null)
21/02/18 17:50:17 INFO ql.Driver: Completed compiling command(queryId=samuel.ntsua_outlook_20210218175017_2489d714-b7ca-4842-add5-1f977154c355); Time taken: 0.21 seconds
21/02/18 17:50:17 INFO ql.Driver: Executing command(queryId=samuel.ntsua_outlook_20210218175017_2489d714-b7ca-4842-add5-1f977154c355):
LOAD DATA INPATH 'hdfs://nameservice1/user/samuel.ntsua_outlook/stock_companies' OVERWRITE INTO TABLE `sntsua_hive_db.sntsua_hive_stock_companies`
21/02/18 17:50:17 INFO ql.Driver: Starting task [Stage-0:MOVE] in serial mode
21/02/18 17:50:17 INFO hive.metastore: Closed a connection to metastore, current connections: 0
Loading data to table [redacted] hive_db.[redacted] hive_stock_companies
21/02/18 17:50:17 INFO exec.Task: Loading data to table sntsua_hive_db.sntsua_hive_stock_companies from hdfs://nameservice1/user/samuel.ntsua_outlook/stock_companies
21/02/18 17:50:17 INFO hive.metastore: HMS client filtering is enabled.
21/02/18 17:50:17 INFO hive.metastore: Trying to connect to metastore with URI thrift://ip-10-0-21-131.ec2.internal:9083
21/02/18 17:50:17 INFO hive.metastore: Opened a connection to metastore, current connections: 1
21/02/18 17:50:17 INFO hive.metastore: Connected to metastore.
21/02/18 17:50:18 INFO common.FileUtils: Creating directory if it doesn't exist: hdfs://nameservice1/user/hive/warehouse/sntsua_hive_db.db/sntsua_hive_stock_companies
chgrp: changing ownership of 'hdfs://nameservice1/user/hive/warehouse/sntsua_hive_db.db/sntsua_hive_stock_companies': User samuel.ntsua_outlook does not belong to hive
21/02/18 17:50:18 INFO ql.Driver: Starting task [Stage-1:STATS] in serial mode
21/02/18 17:50:18 INFO exec.StatsTask: Executing stats task
21/02/18 17:50:18 INFO hive.metastore: Closed a connection to metastore, current connections: 0
21/02/18 17:50:18 INFO hive.metastore: HMS client filtering is enabled.
21/02/18 17:50:18 INFO hive.metastore: Trying to connect to metastore with URI thrift://ip-10-0-21-131.ec2.internal:9083
21/02/18 17:50:18 INFO hive.metastore: Opened a connection to metastore, current connections: 1
21/02/18 17:50:18 INFO hive.metastore: Connected to metastore.
21/02/18 17:50:18 INFO hive.metastore: Closed a connection to metastore, current connections: 0
21/02/18 17:50:18 INFO hive.metastore: HMS client filtering is enabled.
21/02/18 17:50:18 INFO hive.metastore: Trying to connect to metastore with URI thrift://ip-10-0-21-131.ec2.internal:9083
21/02/18 17:50:18 INFO hive.metastore: Opened a connection to metastore, current connections: 1
21/02/18 17:50:18 INFO hive.metastore: Connected to metastore.
21/02/18 17:50:18 INFO exec.StatsTask: Table sntsua_hive_db.sntsua_hive_stock_companies stats: [numFiles=1, numRows=0, totalSize=40005, rawDataSize=0, numFilesErasureCoded=0]
21/02/18 17:50:18 INFO ql.Driver: Completed executing command(queryId=samuel.ntsua_outlook_20210218175017_2489d714-b7ca-4842-add5-1f977154c355); Time taken: 0.35 seconds
OK
21/02/18 17:50:18 INFO ql.Driver: OK
Time taken: 0.579 seconds
21/02/18 17:50:18 INFO CliDriver: Time taken: 0.579 seconds
21/02/18 17:50:18 INFO conf.HiveConf: Using the default value passed in for log id: e8fd8a03-b6ea-4bf5-9b29-967e6f6e3b46
21/02/18 17:50:18 INFO session.SessionState: Resetting thread name to  main
21/02/18 17:50:18 INFO conf.HiveConf: Using the default value passed in for log id: e8fd8a03-b6ea-4bf5-9b29-967e6f6e3b46
21/02/18 17:50:18 INFO session.SessionState: Deleted directory: /tmp/hive/[redacted]/e8fd8a03-b6ea-4bf5-9b29-967e6f6e3b46 on fs with scheme hdfs
21/02/18 17:50:18 INFO session.SessionState: Deleted directory: /tmp[redacted]/e8fd8a03-b6ea-4bf5-9b29-967e6f6e3b46 on fs with scheme file
21/02/18 17:50:18 INFO hive.metastore: Closed a connection to metastore, current connections: 0
21/02/18 17:50:18 INFO hive.HiveImport: Hive import complete.
21/02/18 17:50:18 INFO imps.CuratorFrameworkImpl: backgroundOperationsLoop exiting
21/02/18 17:50:18 INFO zookeeper.ZooKeeper: Session: 0x473d77b3b20cd2d closed
21/02/18 17:50:18 INFO zookeeper.ClientCnxn: EventThread shut down
21/02/18 17:50:18 INFO CuratorFrameworkSingleton: Closing ZooKeeper client.
[sa[redacted]12-218 ~]$
```

I do the same for STOCK_PRICES table, and then I check in HUE interface.

Checking in HUE.



```
show tables;
```

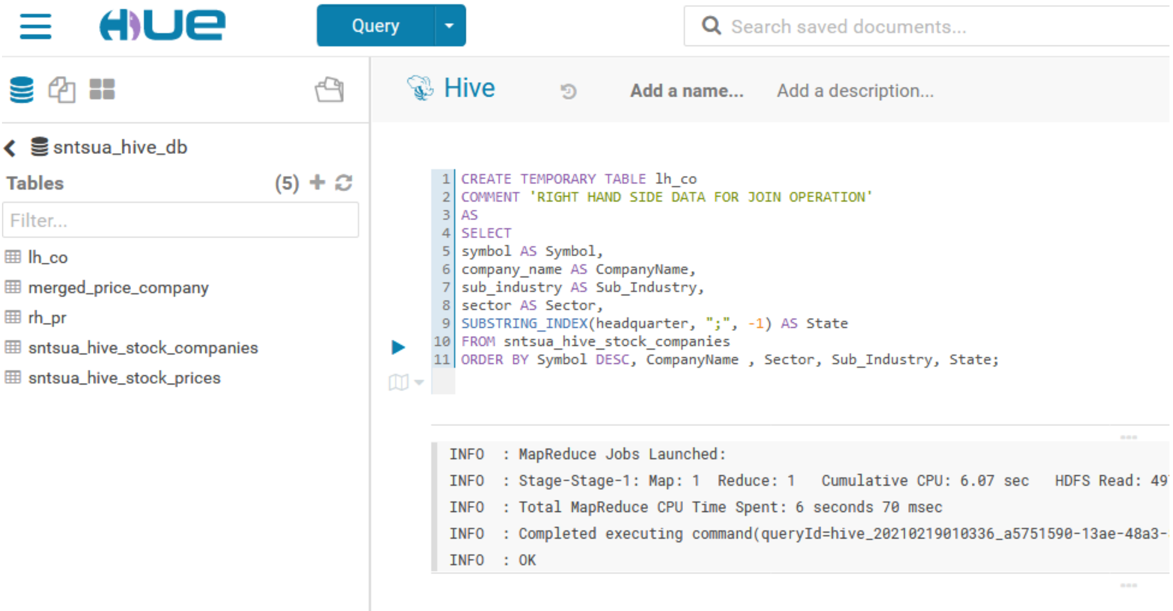
```
INFO  : Completed compiling command(queryId=hive_20210218181343_b7c46ddc-378c-4c2b-bb98-2
INFO  : Executing command(queryId=hive_20210218181343_b7c46ddc-378c-4c2b-bb98-2f5854c3b35
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20210218181343_b7c46ddc-378c-4c2b-bb98-2
INFO  : OK
```

	tab_name
1	sntsua_hive_stock_companies
2	sntsua_hive_stock_prices

Now that I have the two tables in Hive, I can JOIN them, then answer the business questions.

To JOIN the tables, I created a TEMPORARY table to hold intermediary aggregates. This way, a more complex JOIN that could lead to error is avoided.

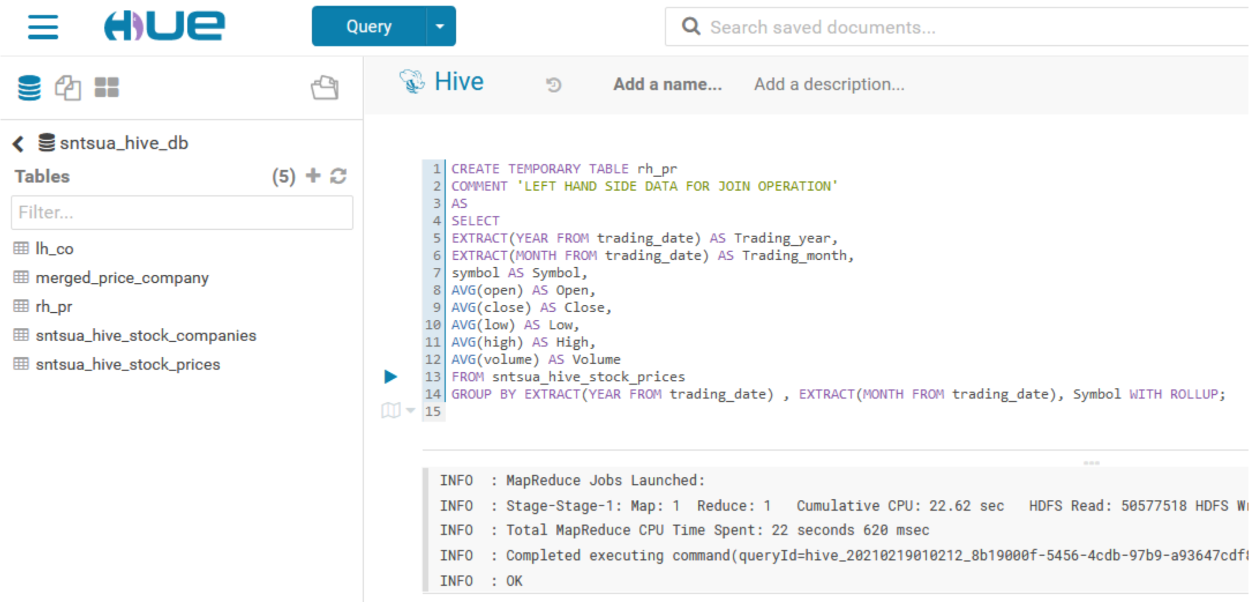
Temp table for stock_companies: lh_co



```
CREATE TEMPORARY TABLE lh_co
COMMENT 'RIGHT HAND SIDE DATA FOR JOIN OPERATION'
AS
SELECT
symbol AS Symbol,
company_name AS CompanyName,
sub_industry AS Sub_Industry,
sector AS Sector,
SUBSTRING_INDEX(headquarter, ";", -1) AS State
FROM sntsua_hive_stock_companies
ORDER BY Symbol DESC, CompanyName , Sector, Sub_Industry, State;
```

```
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.07 sec   HDFS Read: 49
INFO  : Total MapReduce CPU Time Spent: 6 seconds 70 msec
INFO  : Completed executing command(queryId=hive_20210219010336_a5751590-13ae-48a3-
INFO  : OK
```

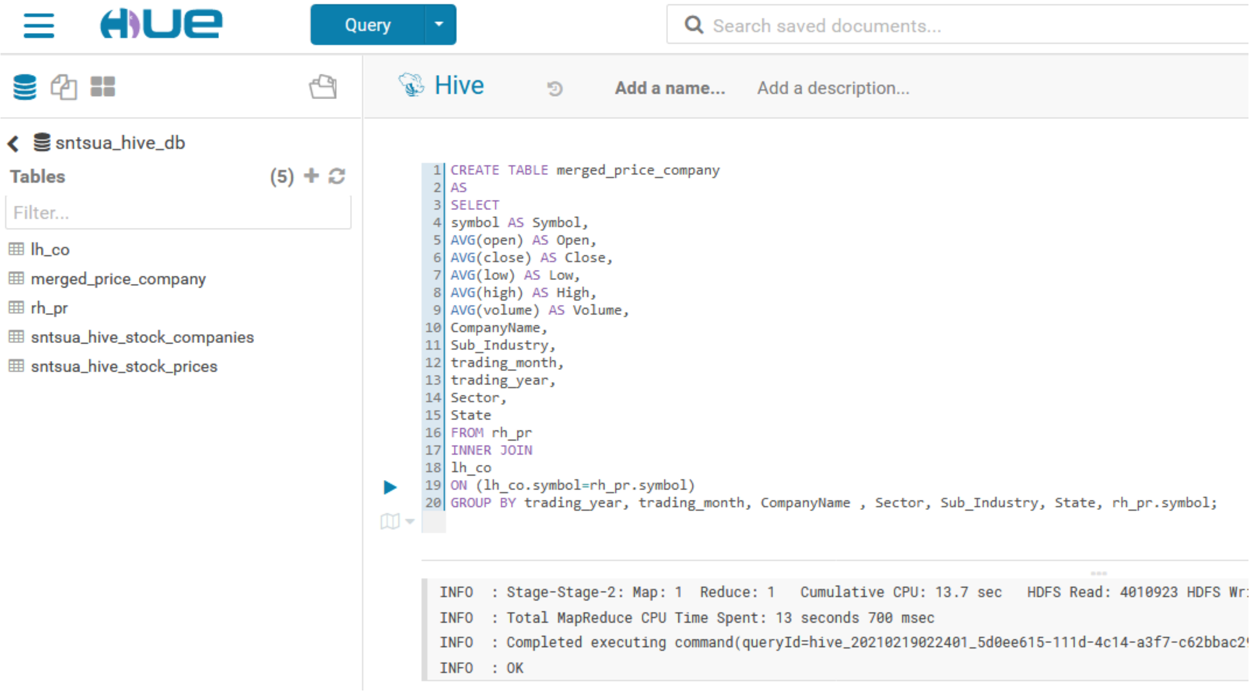
Temp table for stock_price: rh_pr



```
CREATE TEMPORARY TABLE rh_pr
COMMENT 'LEFT HAND SIDE DATA FOR JOIN OPERATION'
AS
SELECT
EXTRACT(YEAR FROM trading_date) AS Trading_year,
EXTRACT(MONTH FROM trading_date) AS Trading_month,
symbol AS Symbol,
AVG(open) AS Open,
AVG(close) AS Close,
AVG(low) AS Low,
AVG(high) AS High,
AVG(volume) AS Volume
FROM sntsua_hive_stock_prices
GROUP BY EXTRACT(YEAR FROM trading_date) , EXTRACT(MONTH FROM trading_date), Symbol WITH ROLLUP;
```

```
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 22.62 sec   HDFS Read: 50577518 HDFS W
INFO  : Total MapReduce CPU Time Spent: 22 seconds 620 msec
INFO  : Completed executing command(queryId=hive_20210219010212_8b19000f-5456-4cdb-97b9-a93647cdf
INFO  : OK
```

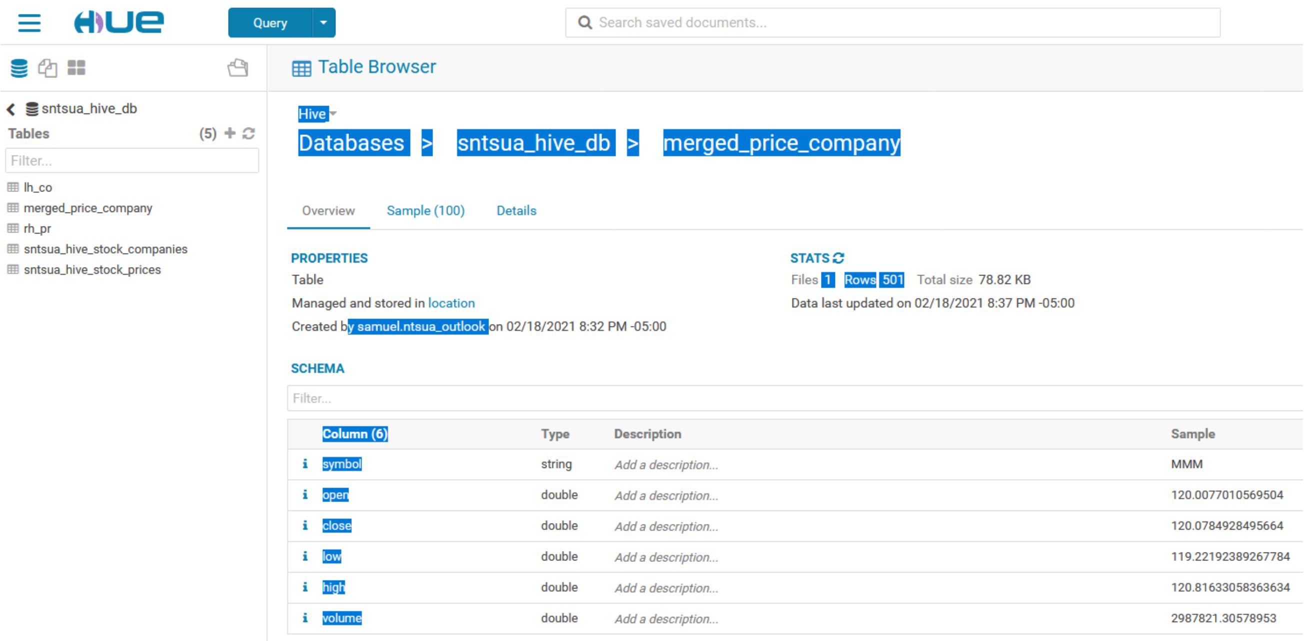
JOIN lh_co and rh_pr: merged_price_company



```
CREATE TABLE merged_price_company
AS
SELECT
symbol AS Symbol,
AVG(open) AS Open,
AVG(close) AS Close,
AVG(low) AS Low,
AVG(high) AS High,
AVG(volume) AS Volume,
CompanyName,
Sub_Industry,
trading_month,
trading_year,
Sector,
State
FROM rh_pr
INNER JOIN
lh_co
ON (lh_co.symbol=rh_pr.symbol)
GROUP BY trading_year, trading_month, CompanyName , Sector, Sub_Industry, State, rh_pr.symbol;
```

```
INFO  : Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 13.7 sec   HDFS Read: 4010923 HDFS Wr
INFO  : Total MapReduce CPU Time Spent: 13 seconds 700 msec
INFO  : Completed executing command(queryId=hive_20210219022401_5d0ee615-111d-4c14-a3f7-c62bbac2
INFO  : OK
```

A quick check on the merged table: merged_price_company. Key features are highlighted.



Hive

Databases > sntsua_hive_db > merged_price_company

PROPERTIES

Table

Managed and stored in location

Created by samuel.ntsua_outlook on 02/18/2021 8:32 PM -05:00

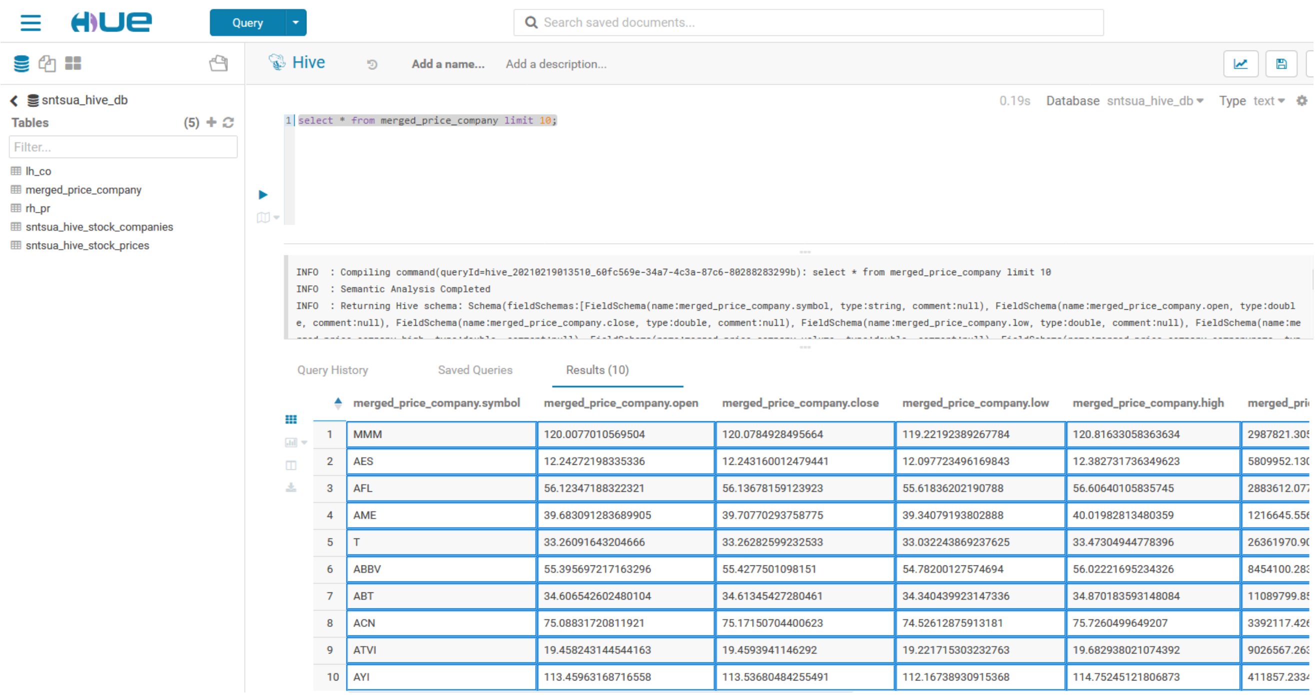
STATS

Files 1 Rows 501 Total size 78.82 KB

Data last updated on 02/18/2021 8:37 PM -05:00

SCHEMA

Column (6)	Type	Description	Sample
symbol	string	Add a description...	MMM
open	double	Add a description...	120.0077010569504
close	double	Add a description...	120.0784928495664
low	double	Add a description...	119.22192389267784
high	double	Add a description...	120.81633058363634
volume	double	Add a description...	2987821.30578953



0.19s Database sntsua_hive_db Type text

```
select * from merged_price_company limit 10;
```

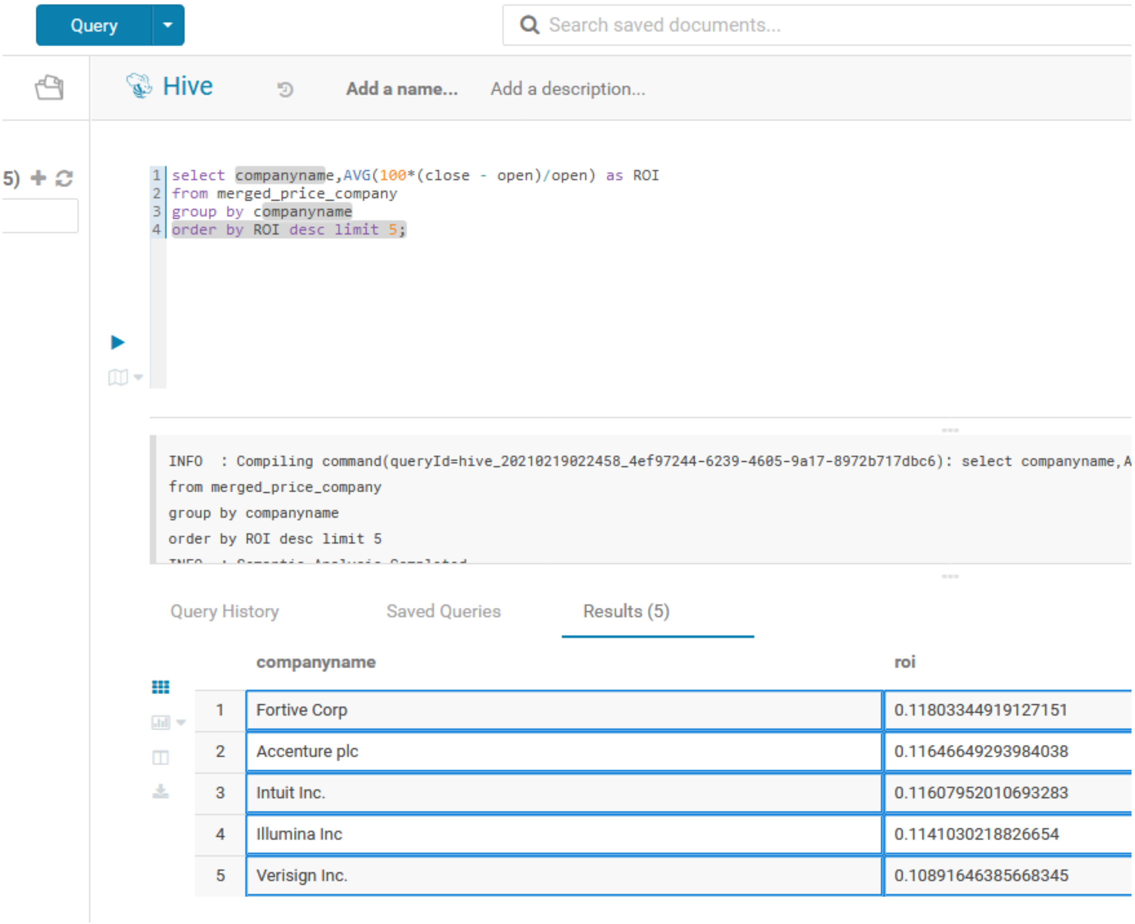
```
INFO  : Compiling command(queryId=hive_20210219013510_60fc569e-34a7-4c3a-87c6-80288283299b): select * from merged_price_company limit 10
INFO  : Semantic Analysis Completed
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:merged_price_company.symbol, type:string, comment:null), FieldSchema(name:merged_price_company.open, type:doubl
e, comment:null), FieldSchema(name:merged_price_company.close, type:double, comment:null), FieldSchema(name:merged_price_company.low, type:double, comment:null), FieldSchema(name:me
```

	merged_price_company.symbol	merged_price_company.open	merged_price_company.close	merged_price_company.low	merged_price_company.high	merged_pri
1	MMM	120.0077010569504	120.0784928495664	119.22192389267784	120.81633058363634	2987821.305
2	AES	12.24272198335336	12.243160012479441	12.097723496169843	12.382731736349623	5809952.130
3	AFL	56.12347188322321	56.13678159123923	55.61836202190788	56.60640105835745	2883612.077
4	AME	39.683091283689905	39.70770293758775	39.34079193802888	40.01982813480359	1216645.556
5	T	33.26091643204666	33.26282599232533	33.032243869237625	33.47304944778396	26361970.90
6	ABBV	55.395697217163296	55.4277501098151	54.78200127574694	56.02221695234326	8454100.283
7	ABT	34.606542602480104	34.61345427280461	34.340439923147336	34.870183593148084	11089799.85
8	ACN	75.08831720811921	75.17150704400623	74.52612875913181	75.7260499649207	3392117.426
9	ATVI	19.458243144544163	19.4593941146292	19.221715303232763	19.682938021074392	9026567.263
10	AYI	113.45963168716558	113.53680484255491	112.16738930915368	114.75245121806873	411857.2334

Answer to business questions:

3) Top 5 Return on investment :

```
Rate of return=100*(Current_value-Initial_value)/Initial_value
```

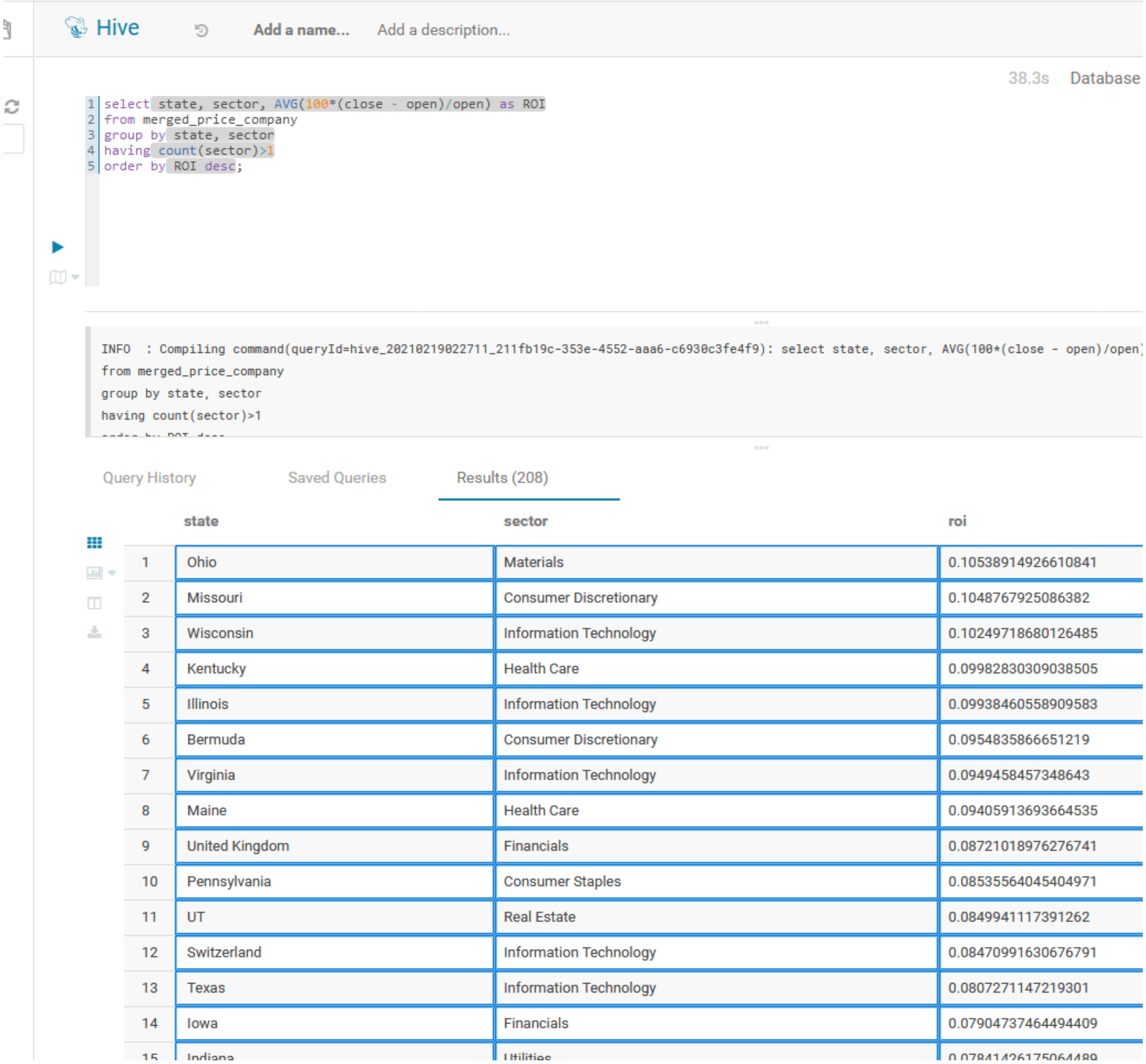


```
select companyname,AVG(100*(close - open)/open) as ROI
from merged_price_company
group by companyname
order by ROI desc limit 5;
```

```
INFO  : Compiling command(queryId=hive_20210219022458_4ef97244-6239-4605-9a17-8972b717dbc6): select companyname,A
from merged_price_company
group by companyname
order by ROI desc limit 5
INFO  : Semantic Analysis Completed
```

	companyname	roi
1	Fortive Corp	0.11803344919127151
2	Accenture plc	0.11646649293984038
3	Intuit Inc.	0.11607952010693283
4	Illumina Inc	0.1141030218826654
5	Verisign Inc.	0.10891646385668345

4) Show the best growing INDUSTRY by each STATE, having at least two or more INDUSTRIES mapped.



```
select state, sector, AVG(100*(close - open)/open) as ROI
from merged_price_company
group by state, sector
having count(sector)>1
order by ROI desc;
```

```
INFO  : Compiling command(queryId=hive_20210219022711_211fb19c-353e-4552-aaa6-c6930c3fe4f9): select state, sector, AVG(100*(close - open)/open)
from merged_price_company
group by state, sector
having count(sector)>1
order by ROI desc
```

	state	sector	roi
1	Ohio	Materials	0.10538914926610841
2	Missouri	Consumer Discretionary	0.1048767925086382
3	Wisconsin	Information Technology	0.10249718680126485
4	Kentucky	Health Care	0.09982830309038505
5	Illinois	Information Technology	0.09938460558909583
6	Bermuda	Consumer Discretionary	0.0954835866651219
7	Virginia	Information Technology	0.0949458457348643
8	Maine	Health Care	0.09405913693664535
9	United Kingdom	Financials	0.08721018976276741
10	Pennsylvania	Consumer Staples	0.08535564045404971
11	UT	Real Estate	0.0849941117391262
12	Switzerland	Information Technology	0.08470991630676791
13	Texas	Information Technology	0.0807271147219301
14	Iowa	Financials	0.07904737464494409
15	Indiana	Utilities	0.07841426175064489

5) For each SECTOR find the following: a. Worst YEAR b. Best YEAR c. Stable YEAR.

Strategy to arrive at the correct answer:

I will answer this question in two stages: I compute the MIN(ROI) for worst year, MAX(ROI) for best year, as well as the AVG(ROI) for each sector and for each year.

The worst year will be determined by their AVG(ROI) > MIN(ROI)

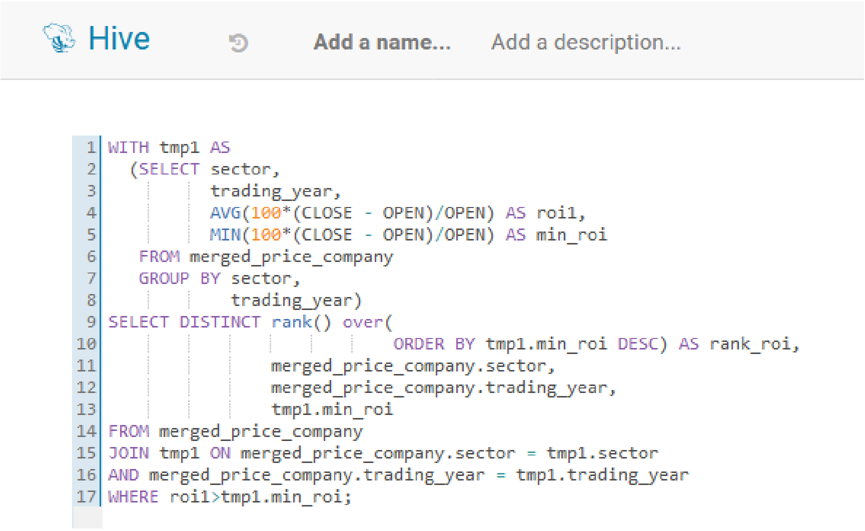
In each stage of computation, the values that will be close to zero will correspond to the stable years.

How the code works: after computing the MIN/MAX(ROI) by sector and by year, distinct rows for sector, year and ROI were retained (similar to dropping duplicates).

Determining Worst YEAR

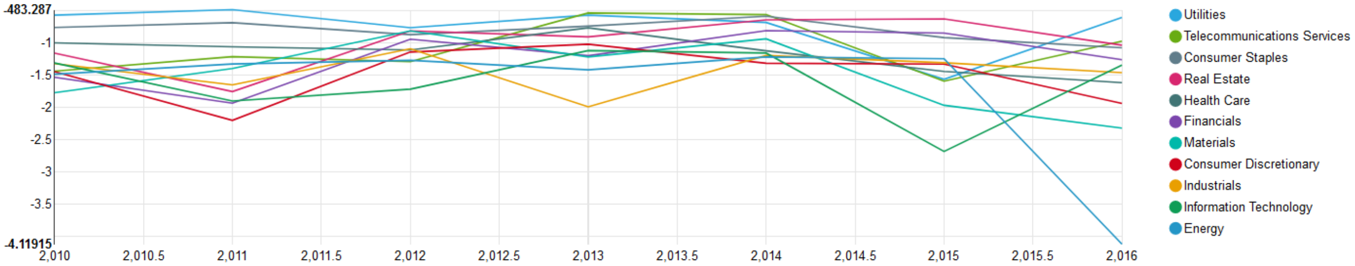
The worst year will be determined by their AVG(ROI) > MIN(ROI)

The HiveQL for MIN(ROI) = Worst Year for each sector:



```
WITH tmp1 AS
  (SELECT sector,
          trading_year,
          AVG(100*(CLOSE - OPEN)/OPEN) AS roi1,
          MIN(100*(CLOSE - OPEN)/OPEN) AS min_roi
   FROM merged_price_company
   GROUP BY sector,
            trading_year)
SELECT DISTINCT rank() over(
                            ORDER BY tmp1.min_roi DESC) AS rank_roi,
                merged_price_company.sector,
                merged_price_company.trading_year,
                tmp1.min_roi
FROM merged_price_company
JOIN tmp1 ON merged_price_company.sector = tmp1.sector
AND merged_price_company.trading_year = tmp1.trading_year
WHERE roi1>tmp1.min_roi;
```

The Hive Graph for MIN(ROI) = Worst Year for each sector:



Reading the output Graph:

Note the color legend showing the various sectors.

I place MIN(ROI) on the Y-axis, and “Trading-Year” on X-axis. Each line shows the worst ROI for a sector(Legend is color-coded for sector).

We can see that Energy sector had its worst year in 2016.

The Information Technology had its worst year in 2015, Consumer Discretionary in 2011

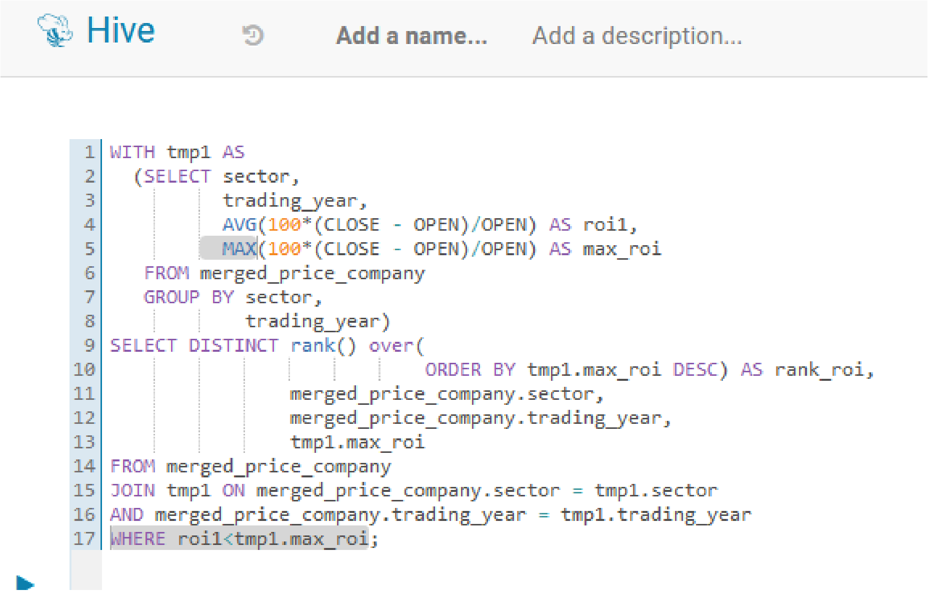
Utility sector seems to be fairly flat, just a little below 0, from 2010 to 2016, except for 2015 where it deeps to its lowest. We can say that Utility sector had very stable years in general from 2010 to 2016.

Determining Best YEAR

The worst year will be determined by their AVG(ROI) <MAX(ROI)

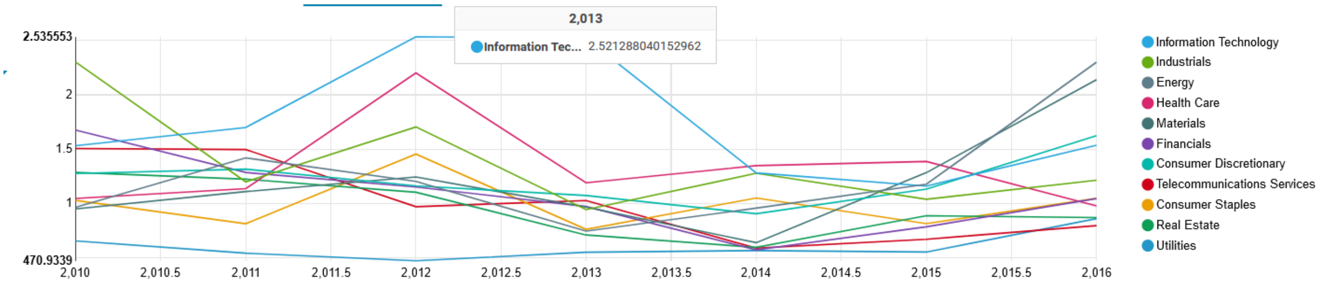
The same code structure is used, except that MAX (ROI) is computed and the AVG(ROI) is lower than the MAX(ROI).

The HiveQL for MAX(ROI) = Best Year for each sector:



```
WITH tmp1 AS
  (SELECT sector,
          trading_year,
          AVG(100*(CLOSE - OPEN)/OPEN) AS roi1,
          MAX(100*(CLOSE - OPEN)/OPEN) AS max_roi
   FROM merged_price_company
   GROUP BY sector,
            trading_year)
SELECT DISTINCT rank() over(
                            ORDER BY tmp1.max_roi DESC) AS rank_roi,
                merged_price_company.sector,
                merged_price_company.trading_year,
                tmp1.max_roi
FROM merged_price_company
JOIN tmp1 ON merged_price_company.sector = tmp1.sector
AND merged_price_company.trading_year = tmp1.trading_year
WHERE roi1<tmp1.max_roi;
```

The Hive Graph for MIN(ROI) = Worst Year for each sector:



Reading the output Graph:

With MAX(ROI) on the Y-axis, and “Trading-Year” on X-axis, each line shows the best ROI for a sector(Legend is color-coded for sector).

Here, Information Technology sector had its best years in 2012 and 2013.

Best year for industrials is 2010. The year 2012 is best for many: Consumer staples, Health Care and IT.

Consumer stables has been mostly flat from 2010 to 2016, hovering around the X-axis, which is an indication of a stable ROI for that sector.

Energy sector had its best years in 2016, still with ROI below zero, but for most part of 2010 to 2016 had shown stability in ROI.

Telecommunication Services and Real Estate have shown less fluctuation in ROI, but continuing decrease in ROI from 2010 to 2016, with 2010 being their best year.

FIN!