# \*\*\*Big Data Hadoop and Spark Developer: Project\_Presentation\*\*\*

In partial fulfillment of Simplilearn Master Data Science Certification course.

Due Date: Feb 17 2021

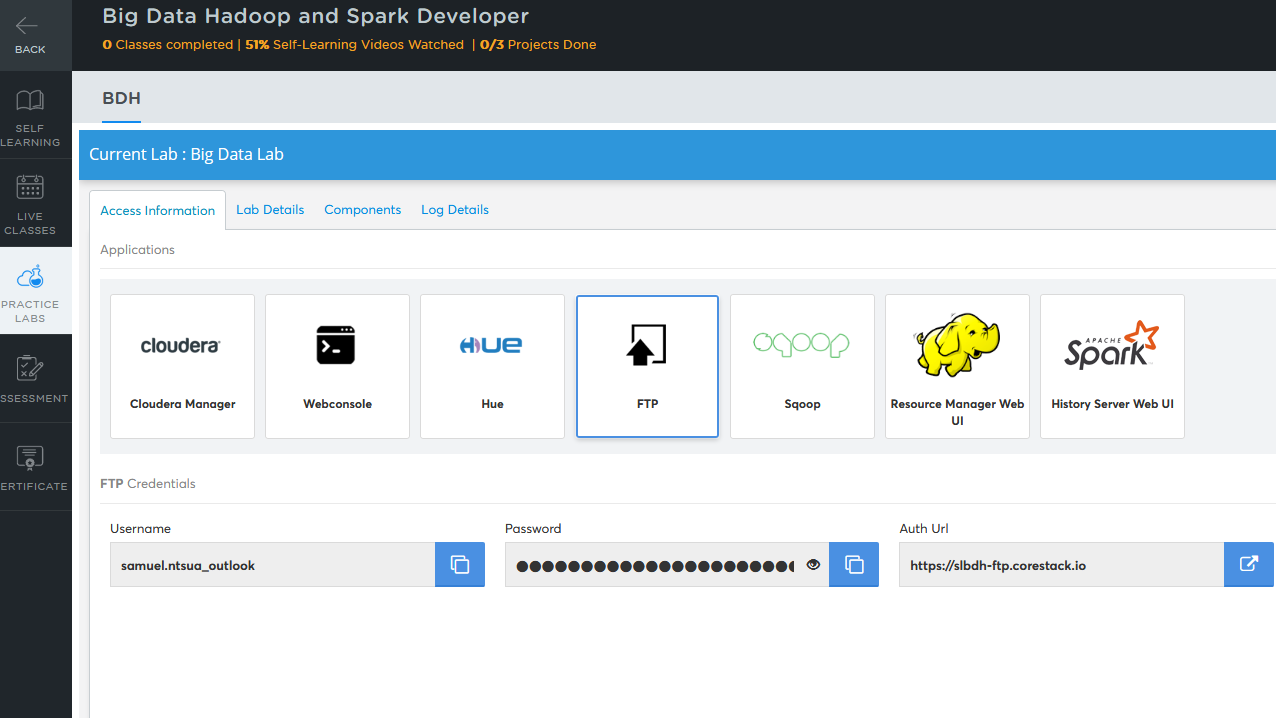
Project name: Stock Exchange Data Analysis

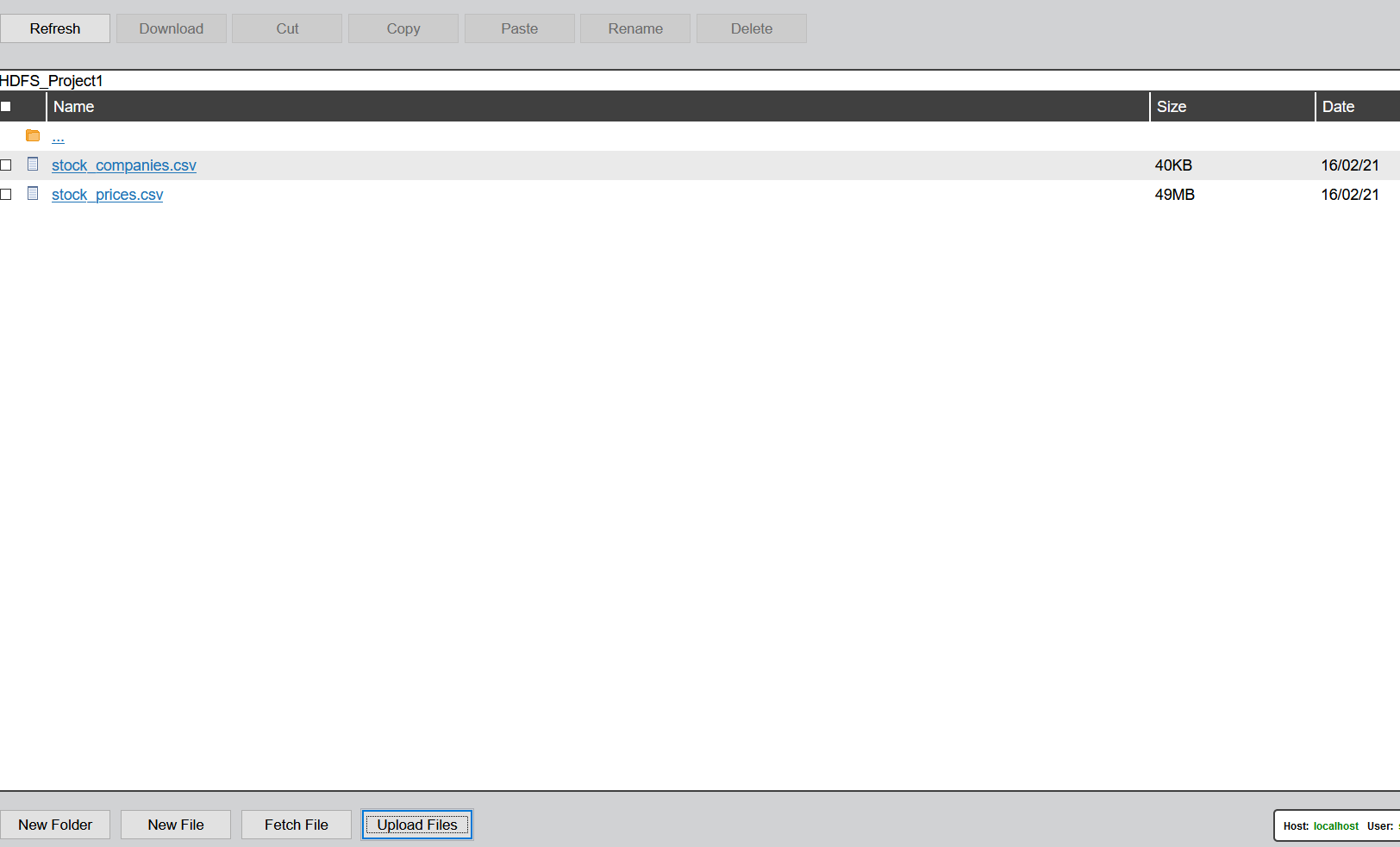
Modeler and presenter : \*\*\*Samuel\_Y.\_Ntsua\*\*\*

Trainer and Mentor: \*\*\*Ajaykuma\*\*\*

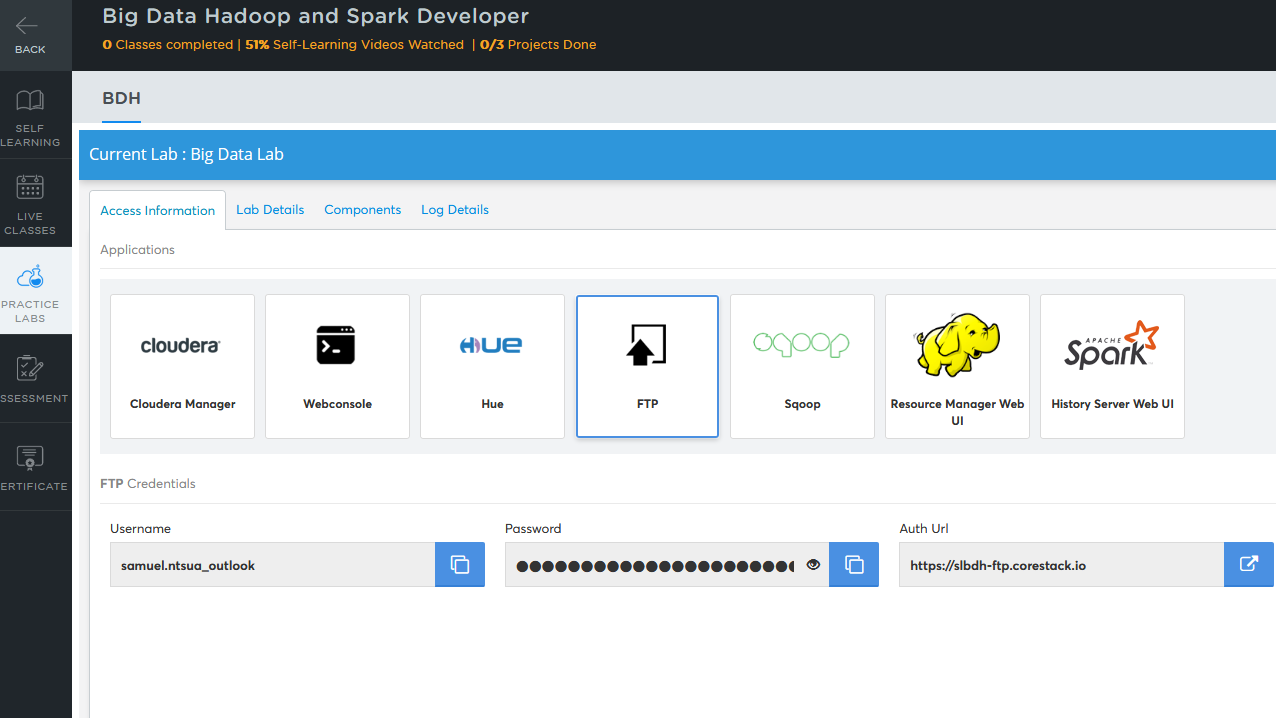
Creating the data pipeline:

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

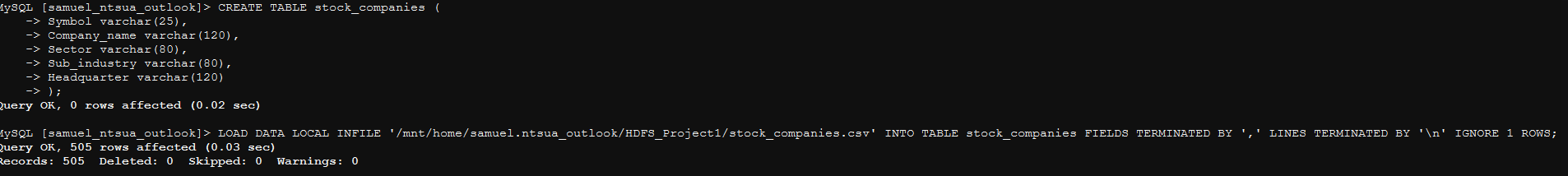




Since the plan is to use Sqoop to move the data into Hive, I logon to the server hosting MySQL and Sqoop:



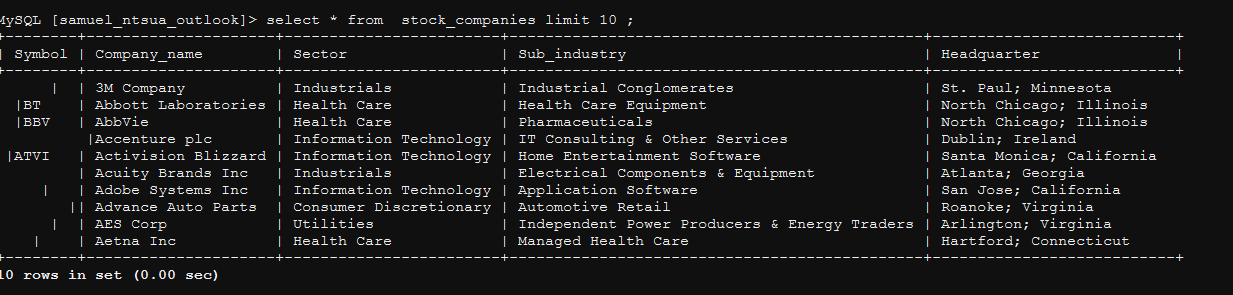
Create a table in MySQL, then load the STOCK\_PRICES.csv and STOCK\_COMPANIES.csv:





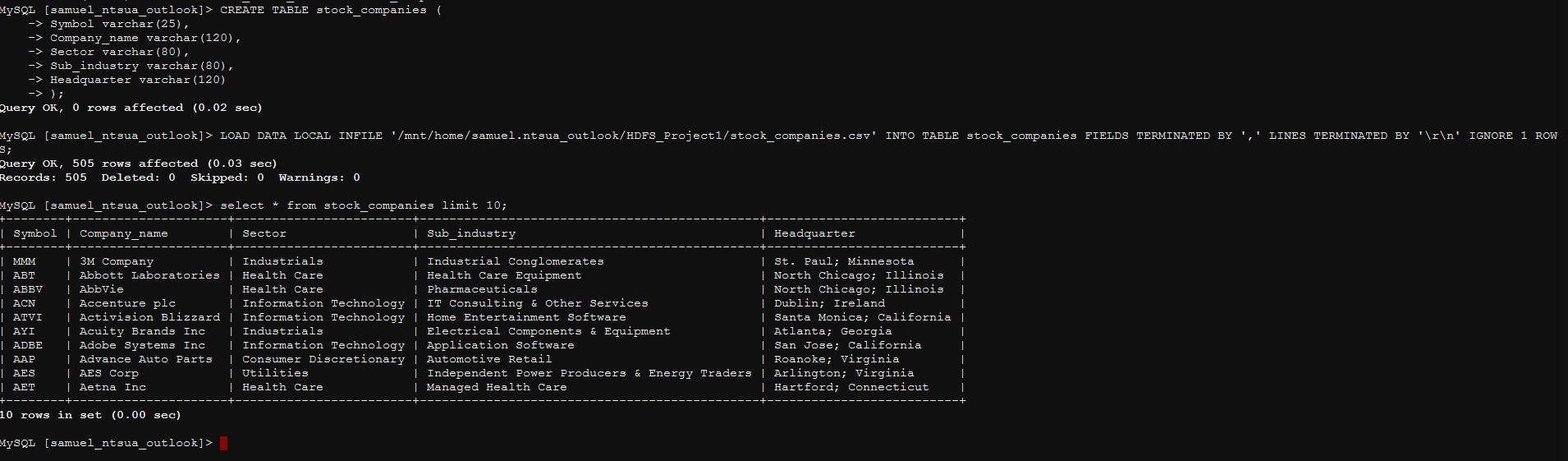
Below is the STOCK\_COMPANIES table in MySQL.

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.



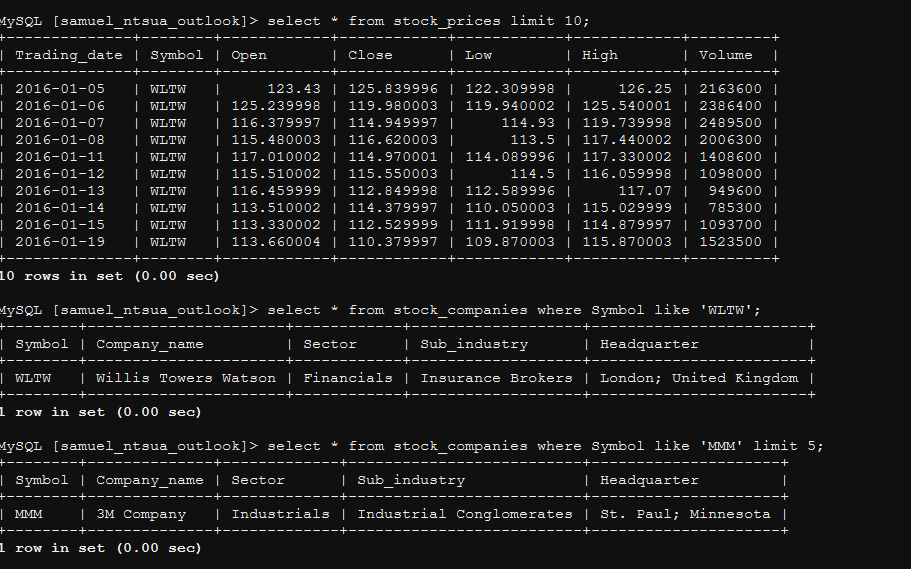


It worked! The Symbol are now well aligned in the table below.





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



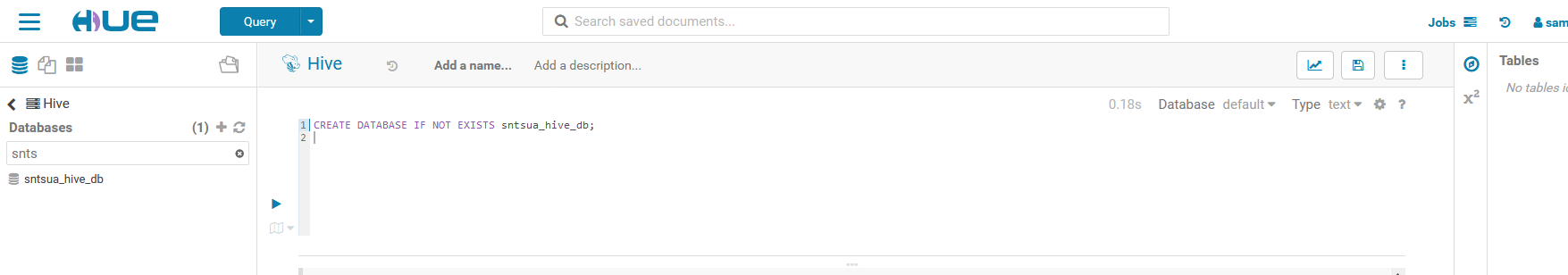


Now that we have the two tables straighten up, and can be queried by Sqoop, we can now Sqoop them to Hive.

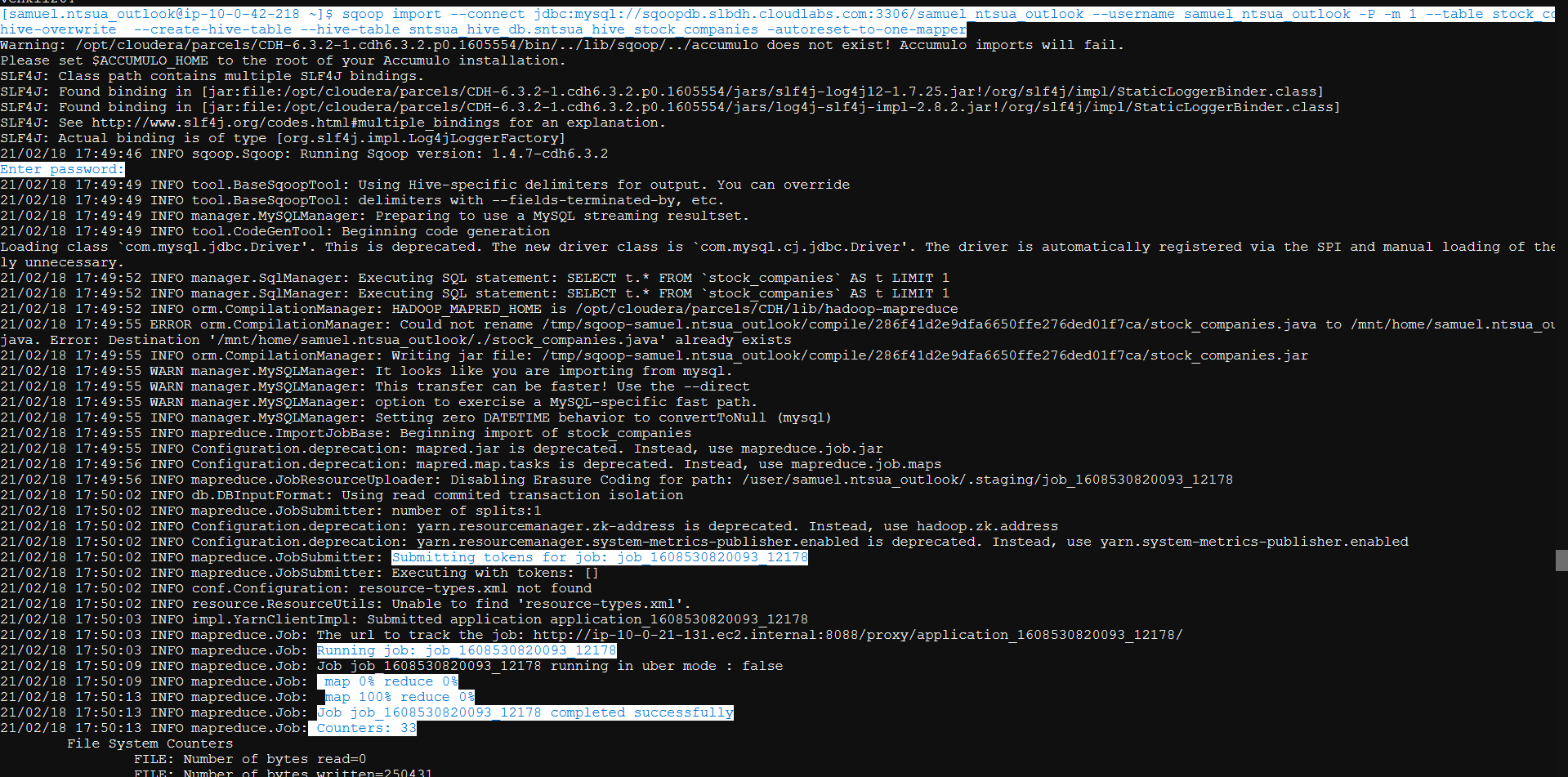
Since we did not set PRIMARY KEY in the tables, Sqoop will complain because Sqoop uses the key to “split” the file to load. We can tell Sqoop to load the file without splitting it by passing the argument –autoreset-to-one-mapper.

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

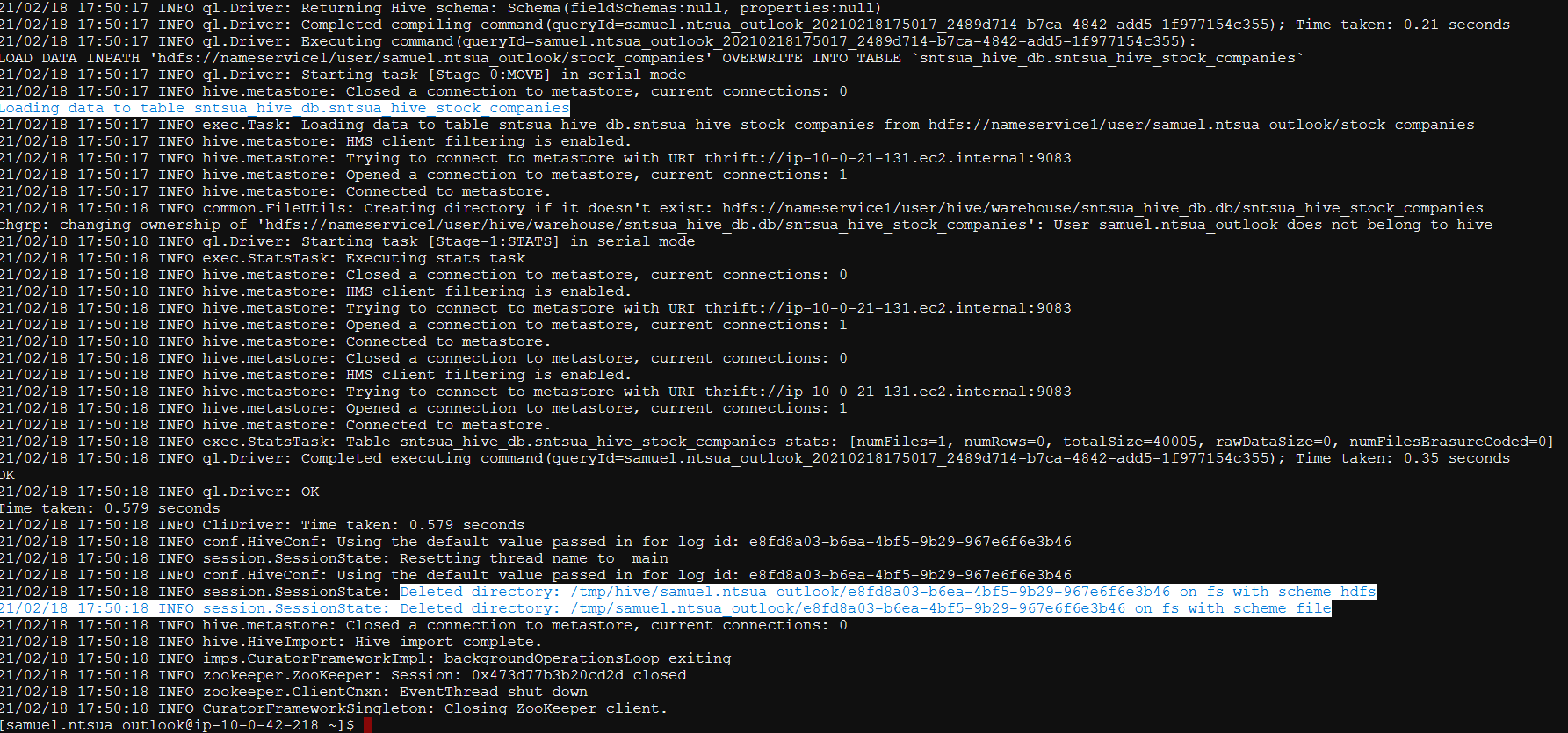
I opted for this because I do not want Sqoop to put my tables in the default database. So I will specify a directory where Sqoop will put the tables.



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.



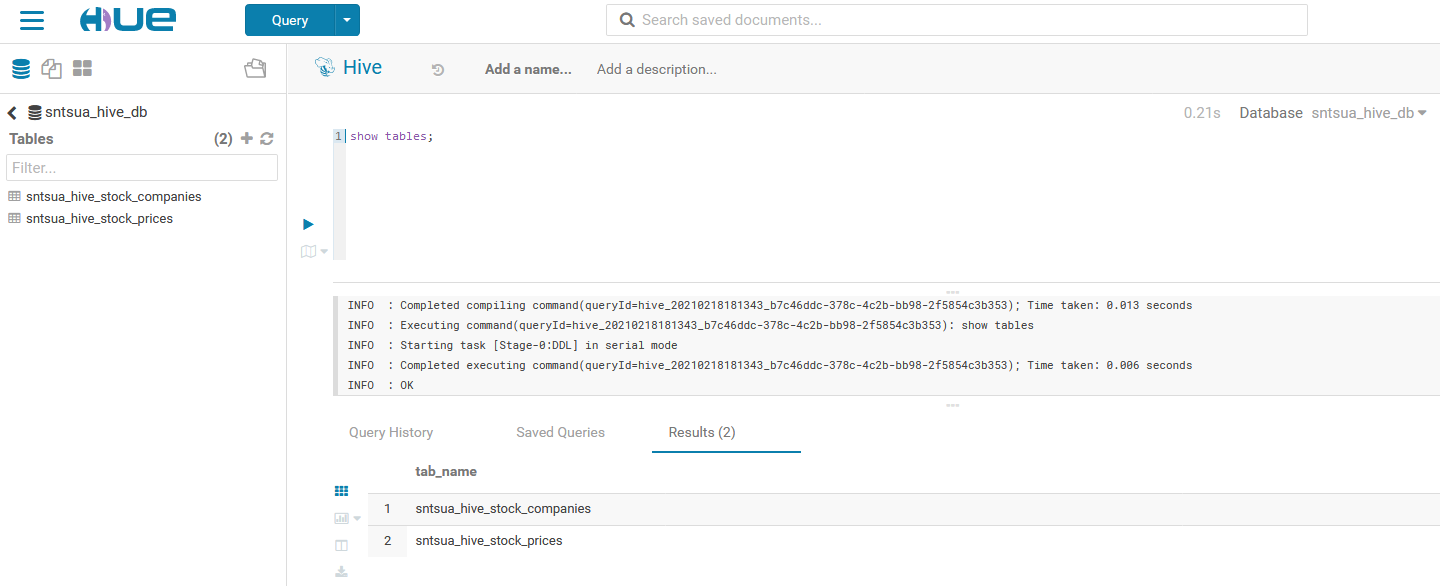






I do the same for STOCK\_PRICES table, and then I check in HUE interface.

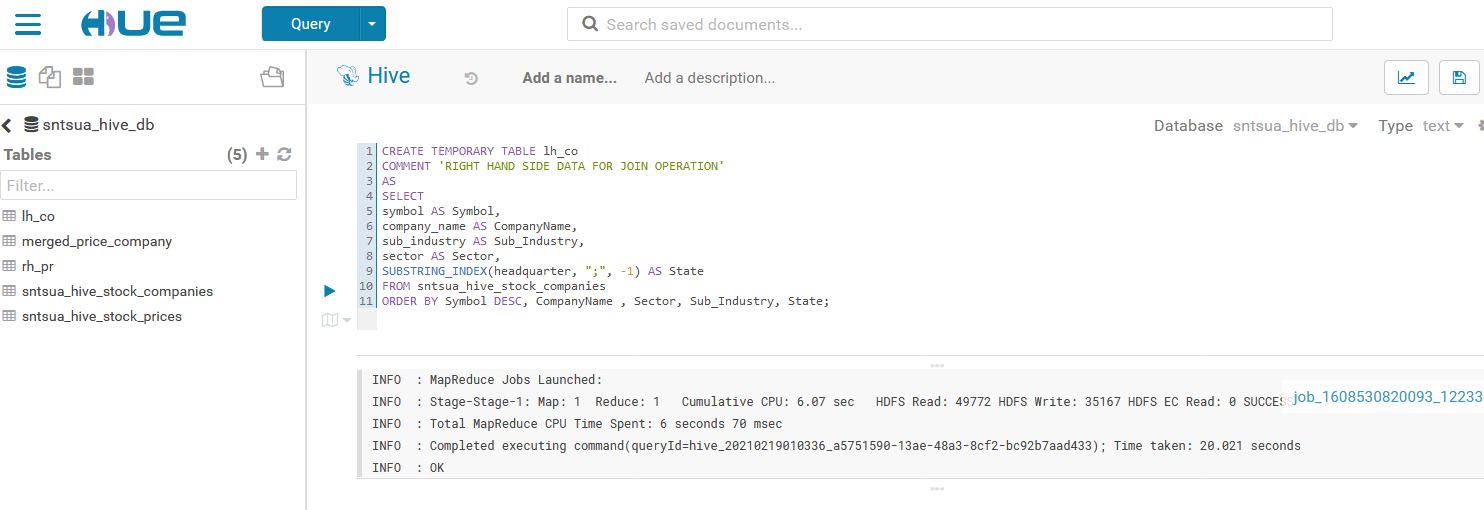
Checking in HUE.



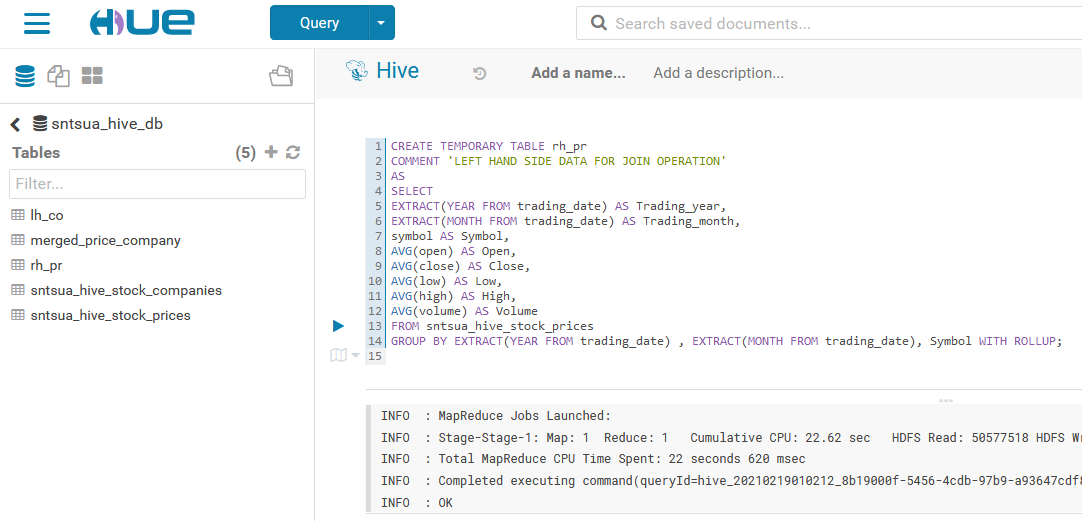
Now that I have the two tables in Hive, I can JOIN them, then answer the business questions.

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.

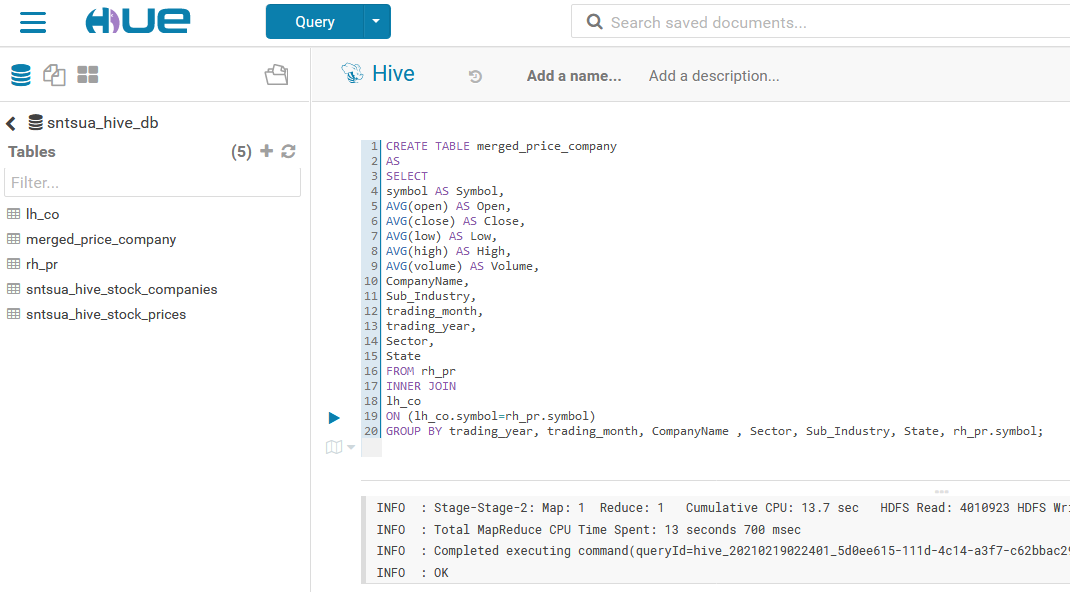
Temp table for stock\_companies: lh\_co



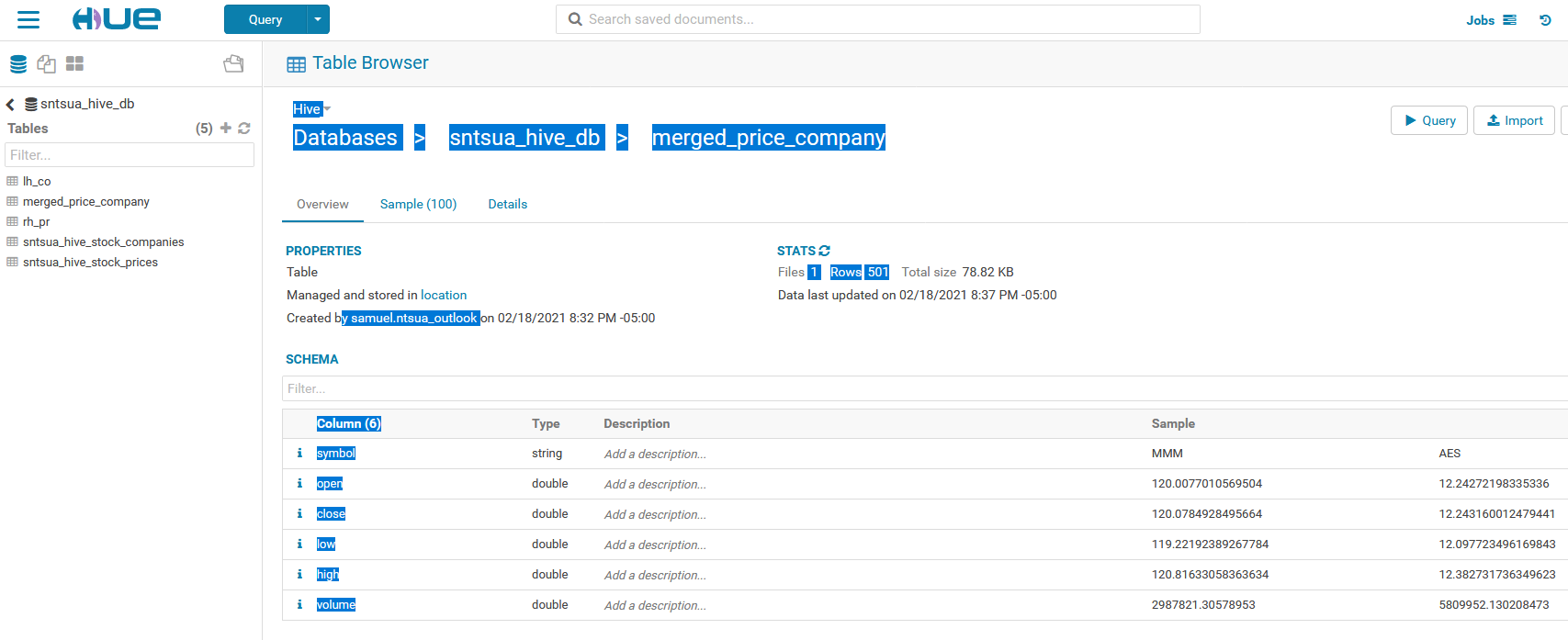
Temp table for stock\_price: rh\_pr

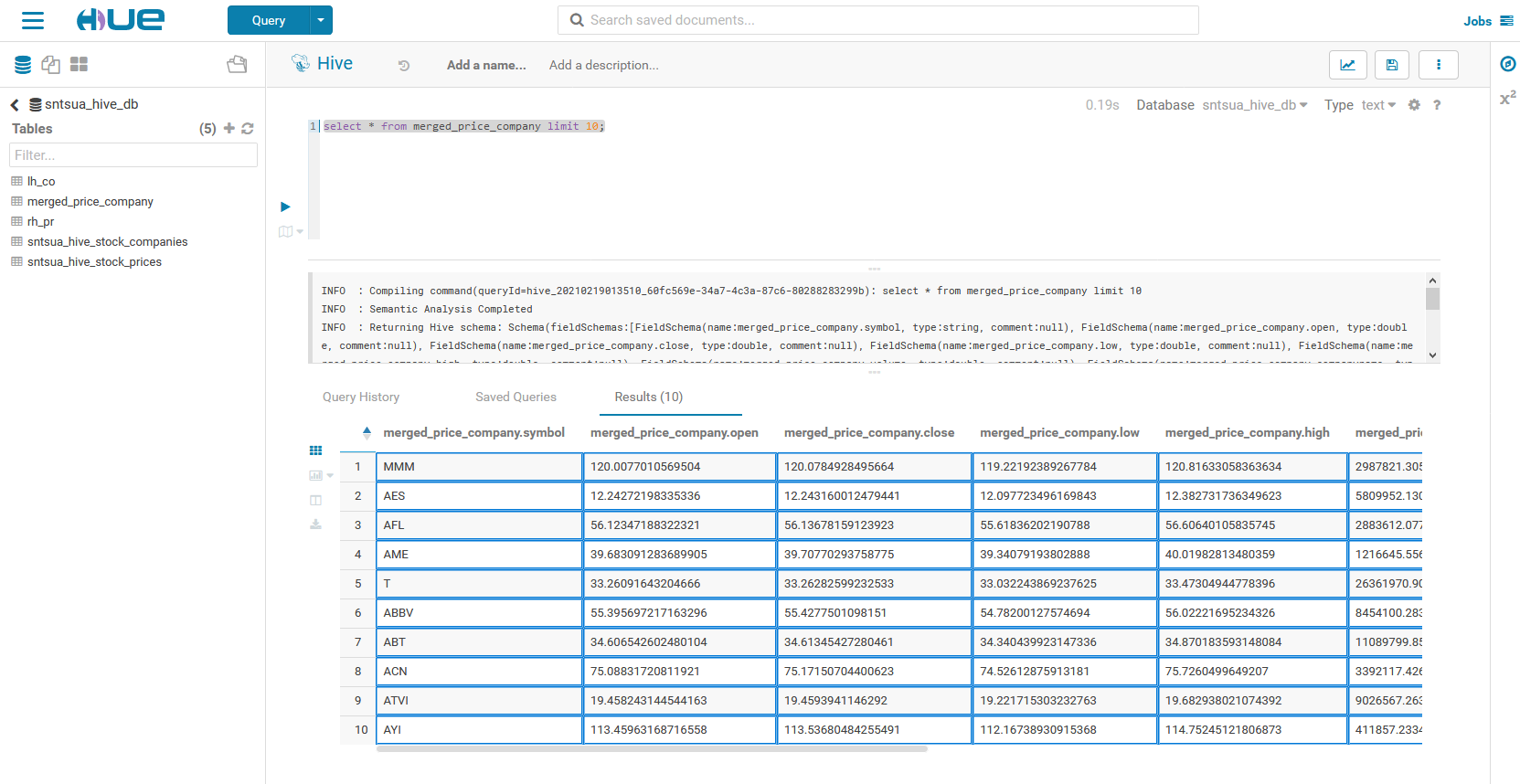


JOIN lh\_co and rh\_pr: merged\_price\_company



A quick check on the merged table: merged\_price\_company. Key features are highlighted.

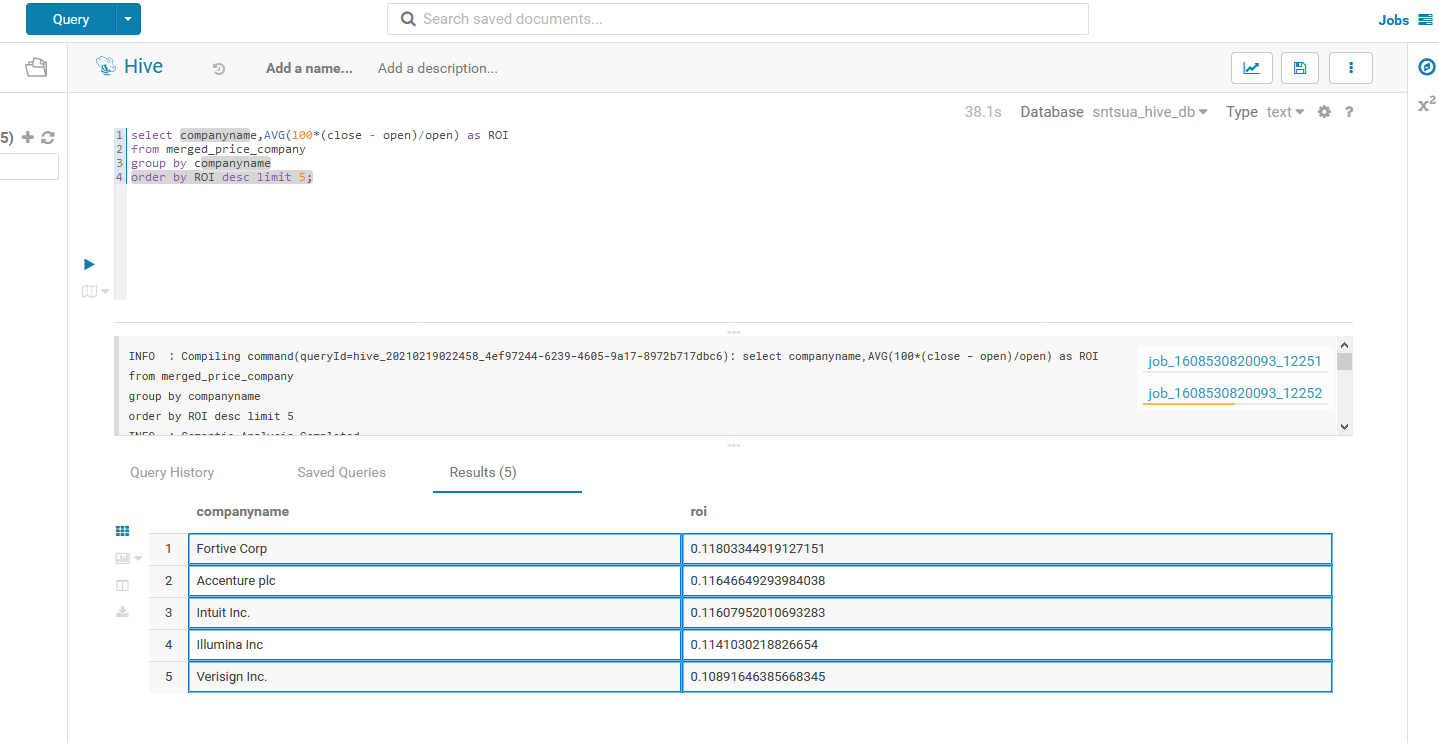




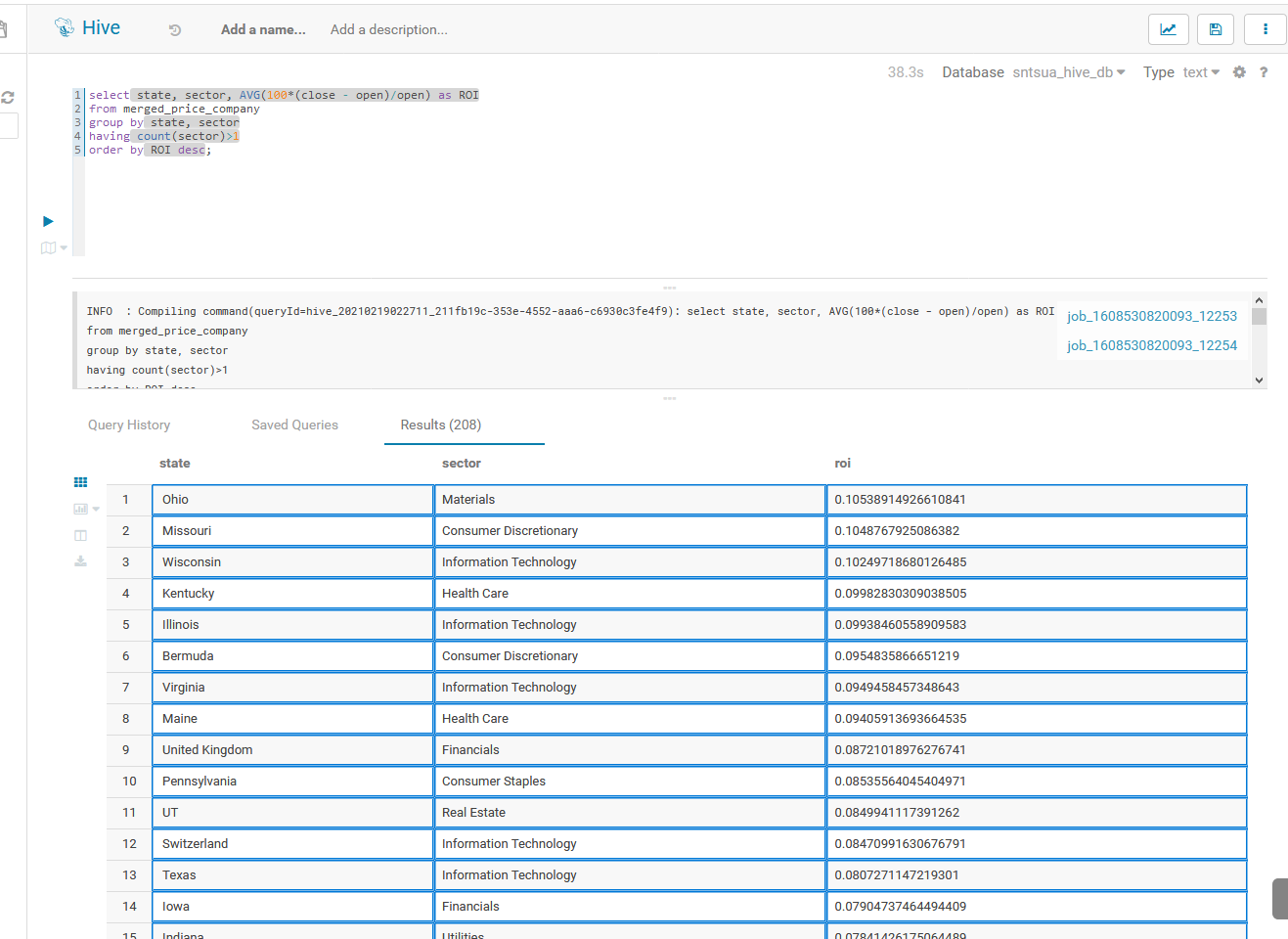
***Answer to business questions:***

**3) Top 5 Return on investment :**

Rate of return=100\*(Current\_value−Initial\_value)/Initial\_value



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



**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)

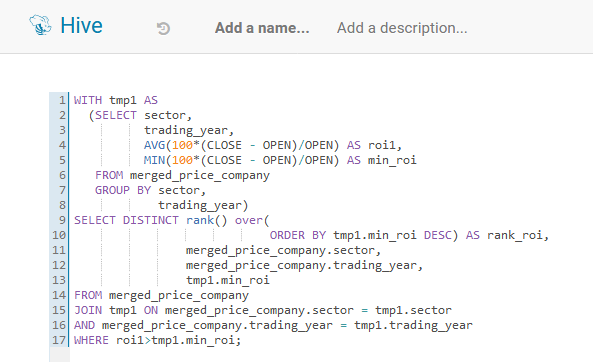
In each stage of computation, the values that will be close to zero will correspond to the stable years.

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).

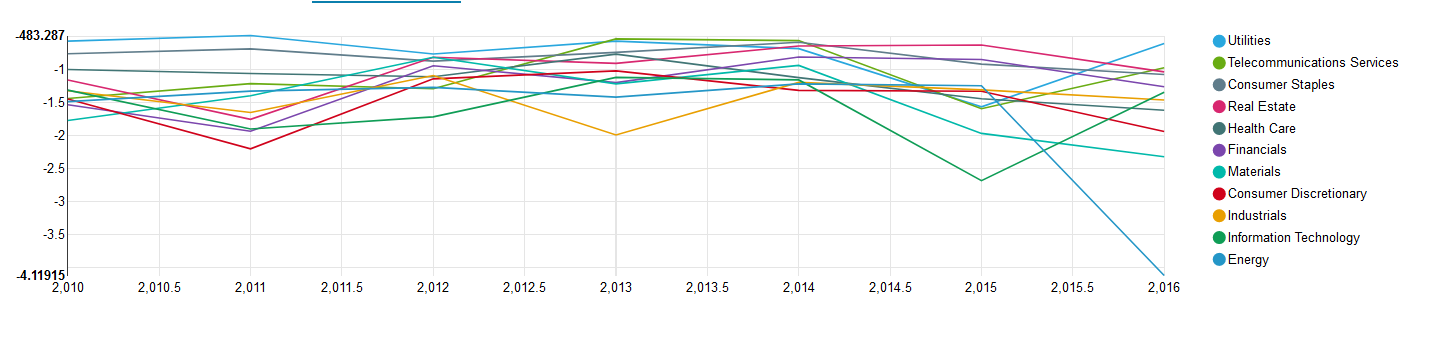
Determining Worst YEAR

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

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



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.

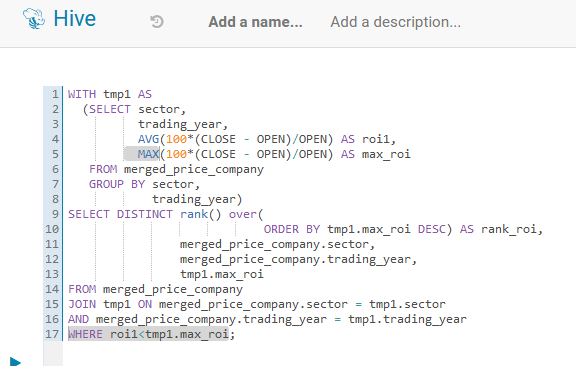
The Information Technology had its worst year in 2015, Consumer Discretionary in 2011

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.

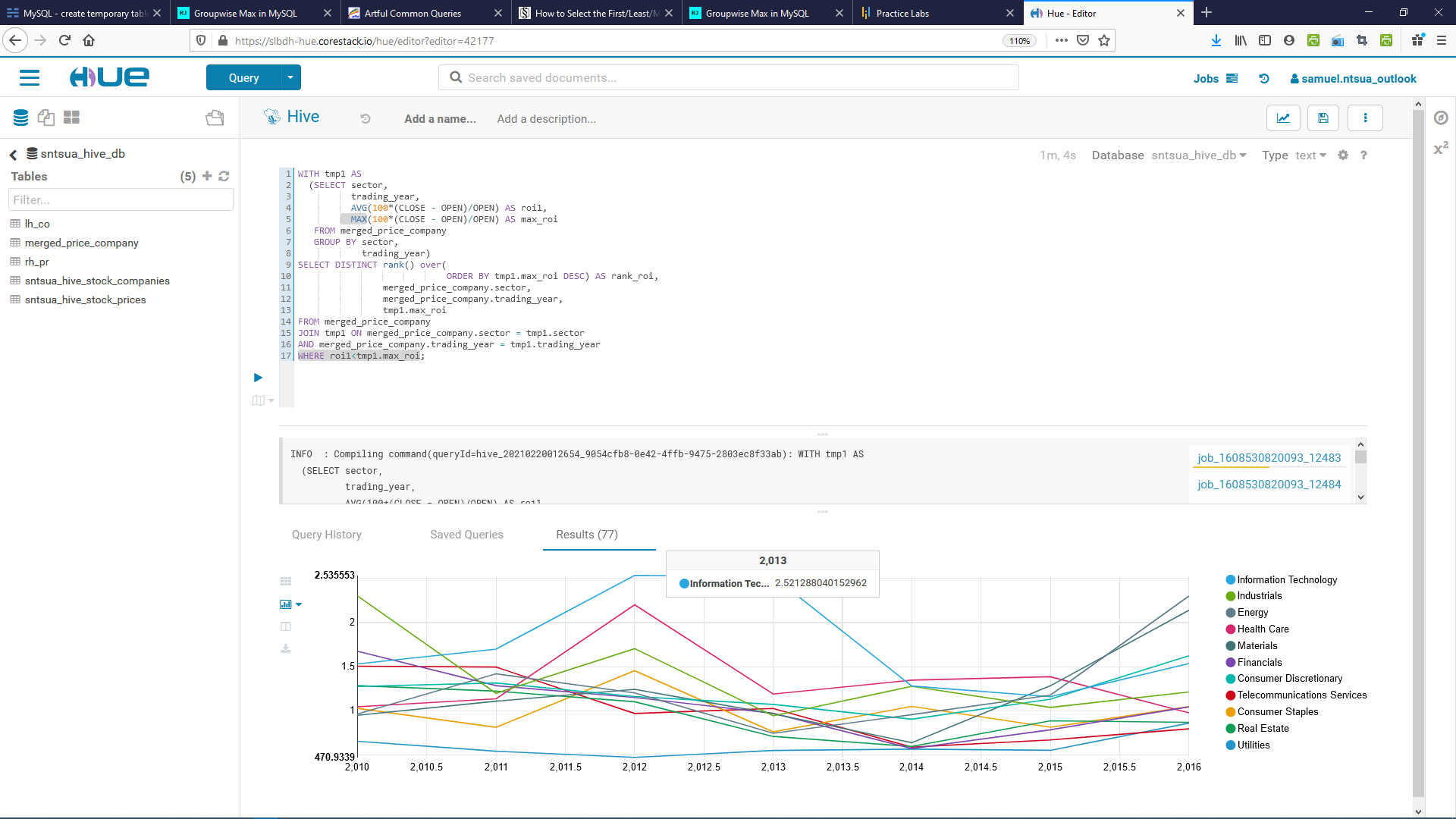
Determining Best YEAR

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:

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!**