**Steps to perform:**

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

**a. MYSQL DATABASE NAME: BDHS\_PROJECT**

**i. Stock\_prices  
ii. Stock\_companies**

Below is the command we use to sqoop the database From MYSQL.

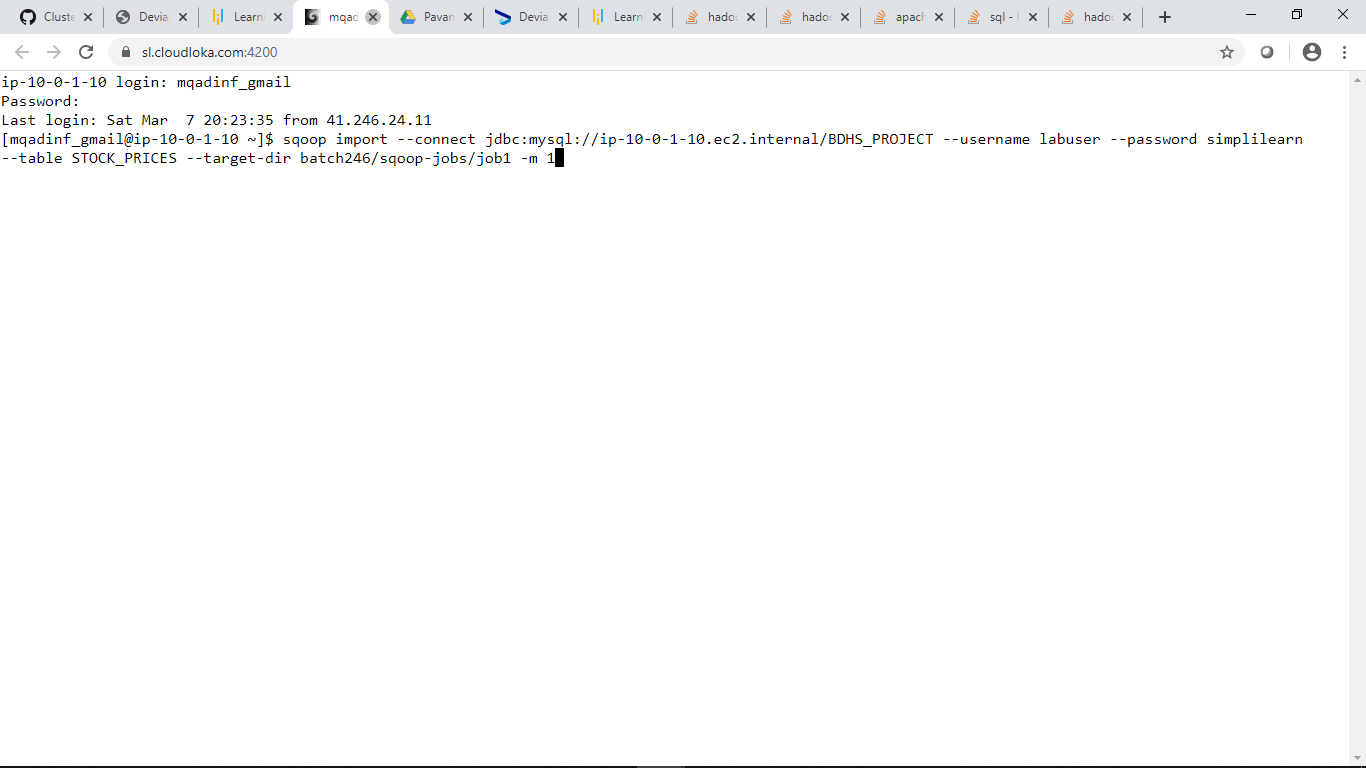
The database we use is **BDHS\_PROJECT** ,the table we use is **STOCK\_PRICES**

The target directory is **batch246/sqoop-jobs/job1**

**Below is the code to pull the data from the table**

**Code: sqoop import - -connect jdbc:mysql: //ip-10-0-1-10.ec2.internal/BDHS\_PROJECT - -username labuser - -password simplilearn - - table STOCK\_PRICES - - target-dir batch246/sqoop-jobs/job1 -m 1**

**Screenshot**

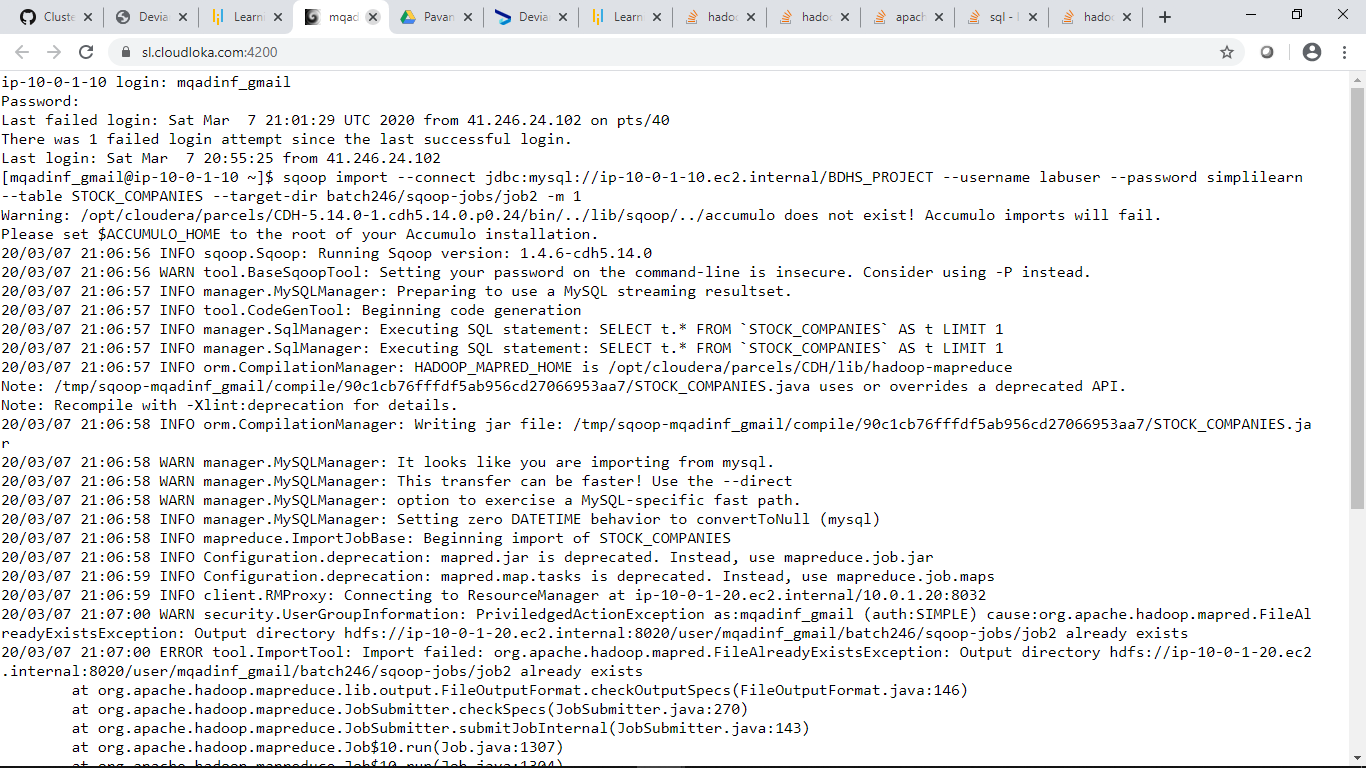


We do the same to the **STOCK\_COMPANIES** table

The target directory will be **batch246/sqoop-jobs/job2**

**Code: sqoop import - -connect jdbc:mysql: //ip-10-0-1-10.ec2.internal/BDHS\_PROJECT - -username labuser - -password simplilearn - - table STOCK\_COMPANIES - - target-dir batch246/sqoop-jobs/job2 -m 1**

**Screenshot**



**Check the TABLE description: STOCK\_PRICES**

**TABLE: STOCK\_COMPANIES**

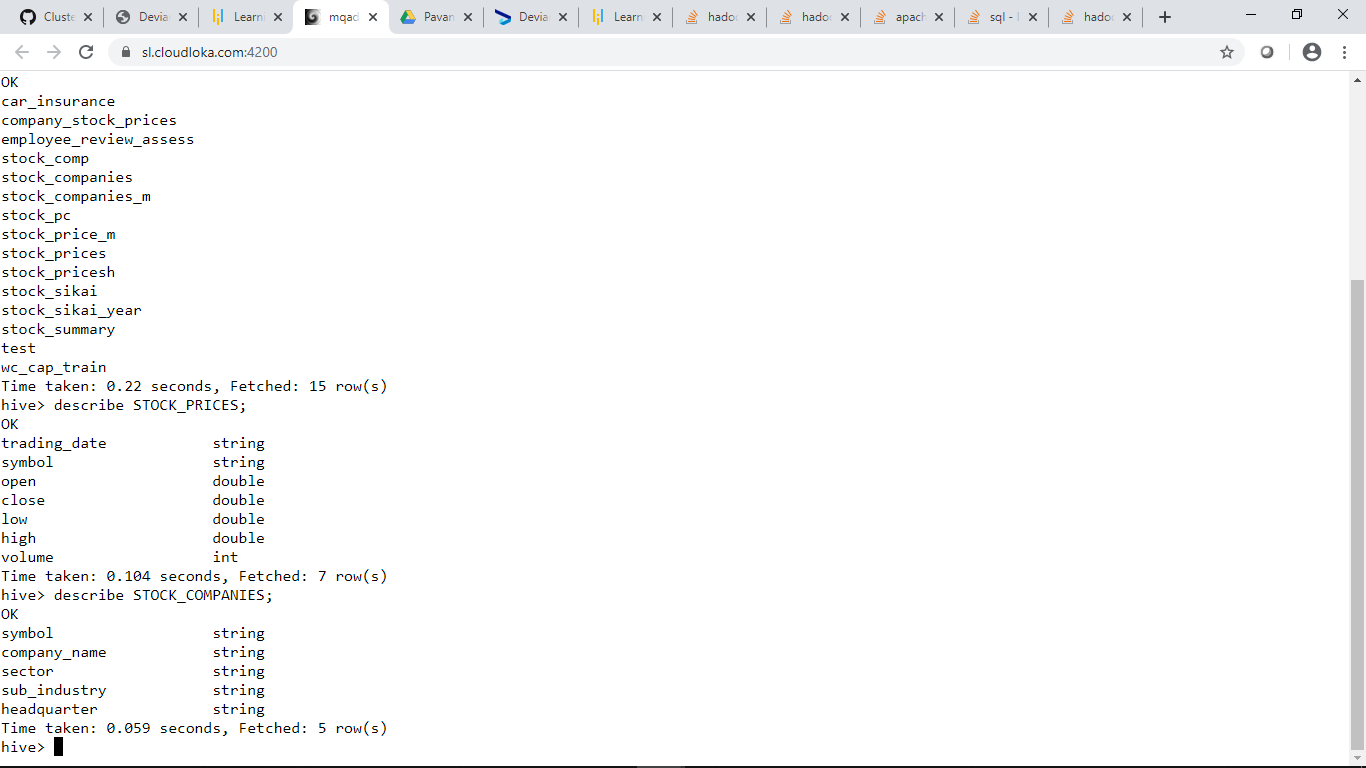
We can check the table descriptions on the tables STOCK\_PRICES and STOCK\_COMPANIES using hive describe function

**describe STOCK\_PRICES;**

**describe STOCK\_COMPANIES;**

All the Datatypes are appropriate besides the **trading\_date** datatype,it’s supposed to be a **Date** datatype,so we have to convert it to a **Date** datatype.

**Screenshot**



Now the trading\_date column is converted to an appropriate datatype which is a DATE

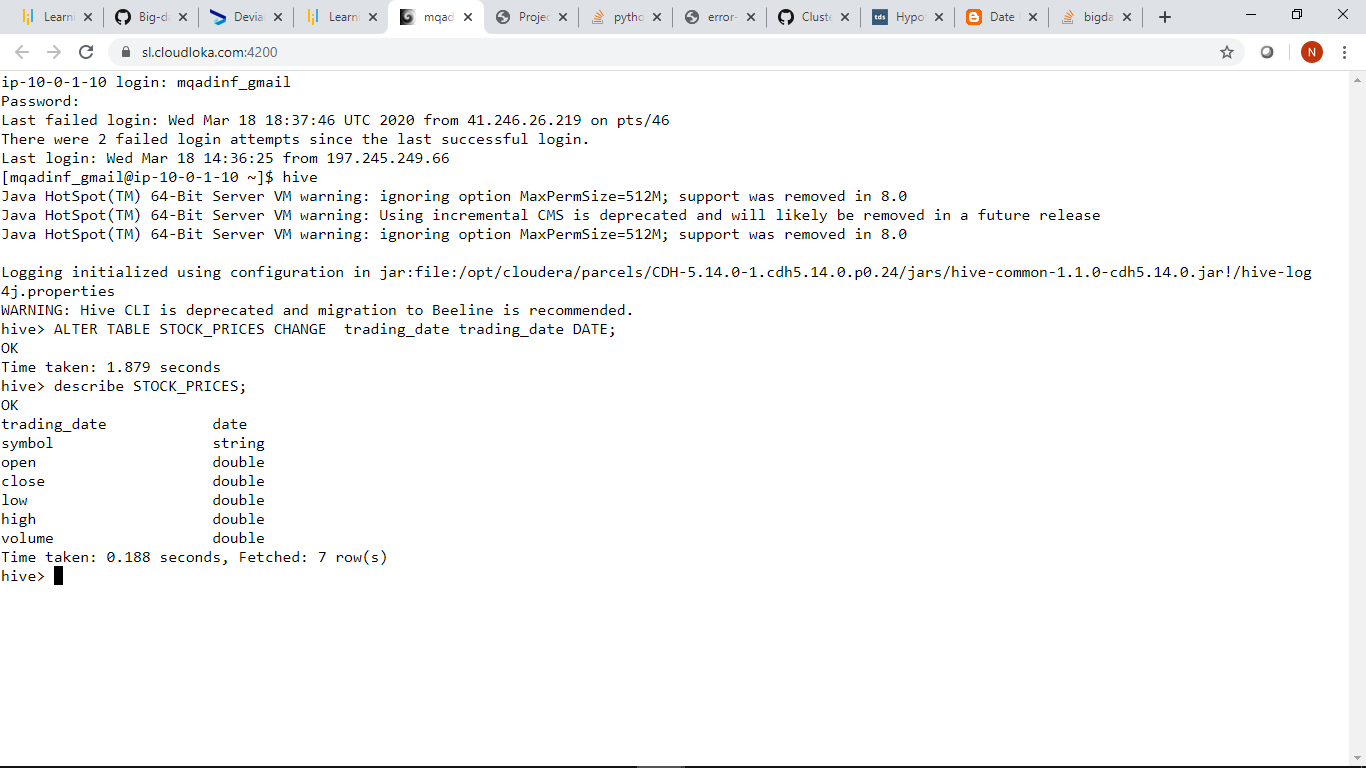
To convert to a DATE datatype we use the following code:

**ALTER TABLE STOCK\_PRICES CHANGE trading\_date trading\_date DATE;**

If we look below we can see that the trading\_date is converted after we run the command :

**describe STOCK\_PRICES;**

**Screen shot**



**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).**

* **Trading\_year: Should contain YYYY for each record**
* **Trading\_month: Should contain MM or MMM for each record**
* **Symbol: Ticker code**
* **CompanyName: Legal name of the listed company**
* **State: State to be extracted from headquarters value.**
* **Sector: Business vertical of the listed company**
* **Sub\_Industry: Business domain of the listed company within a sector**
* **Open: Average of intra-day opening price by month and year for each listed company**
* **Close: Average of intra-day closing price by month and year for each listed company**
* **Low: Average of intra-day lowest price by month and year for each listed company**
* **High: Average of intra-day highest price by month and year for each listed company**
* **Volume: Average of number of shares traded by month and year for each listed company**

Here we must first create a hive table named stockexchange\_hivedb2

In columns a,b,e and h to l appropriate Hive built-in functions are used.

-In column a,which is Trading\_year we used the function **date\_format(trading\_date,’YYYY’)**

- In column b,which is Trading\_month we used the function **date\_format(trading\_date,’MMM’)**

In these columns we used the column trading\_date from the STOCK\_PRICES table to be converted to Year and Month.

There are other alternatives however to accomplish the tasks of converting To Year and Month.

We could have used the function **Year(trading\_date)** for column a and **MONTH(trading\_date)** for column b as an alternative.

-for column e we used the function **split(headquarter,’[\;]’)[1]** to extract state from the headquarters value.Because the headquarter value is a string where the values are semi-colon separated,we will extract the second value from the headquarter with index of [1].

For columns h to l we used the aggregate function which is avg which is average:

**avg(open) as Open,**

**avg(close) as Close,**

**avg(low) as Low,**

**avg(high) as High,**

**avg(volume) as Volume**

Because we need the average of each of these by month and year we used a group by function.

We need to perform an inner join also inorder to create a new hive table.

We can accomplish this task by creating an alias for the STOCK\_PRICES table as s,and for the STOCK\_COMPANIES table as c.

Because the column symbol is appearing on both tables it is a primary key.Therefore the column symbol is a primary key.

Below is the code of creating a new hive table which stockexchange\_hivedb2

**Code:**

**create table stockexchange\_hivedb2**

**as**

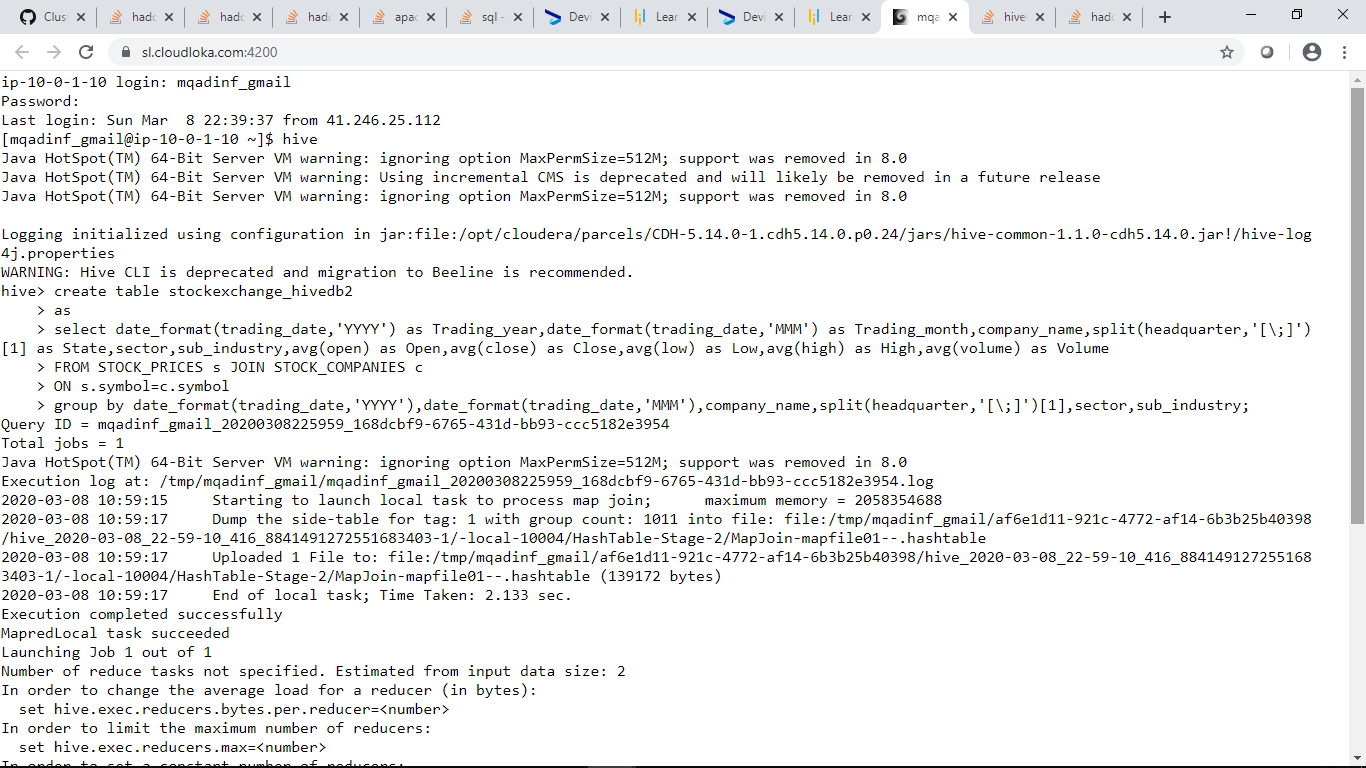
**select date\_format(trading\_date,’YYYY’) as Trading\_year,date\_format(trading\_date) as Trading\_month,company\_name,split(headquarter, ‘[\;]’)[1] as State,sector,sub\_industry,avg(open) as Open,avg(close) as Close,avg(low) as Low,avg(high) as High,avg(volume) as Volume**

**FROM STOCK\_PRICES s JOIN STOCK\_COMPANIES c**

**ON s.symbol=c.symbol**

**Group by date\_format(trading\_date,’YYYY’),date\_format(trading\_date,’MMM’),company\_name,split(headquarter,’[\;]’)[1],sector,sub\_industry;**

**Screen shot of the code**



**DATA ANALYSIS USING HIVE**

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

Here we will need to select the top 5 companies where the Volume (Average of number of shares traded by month and year for each listed company) is higher.

We will sort the Volume in Descending Order: Here’s the code below

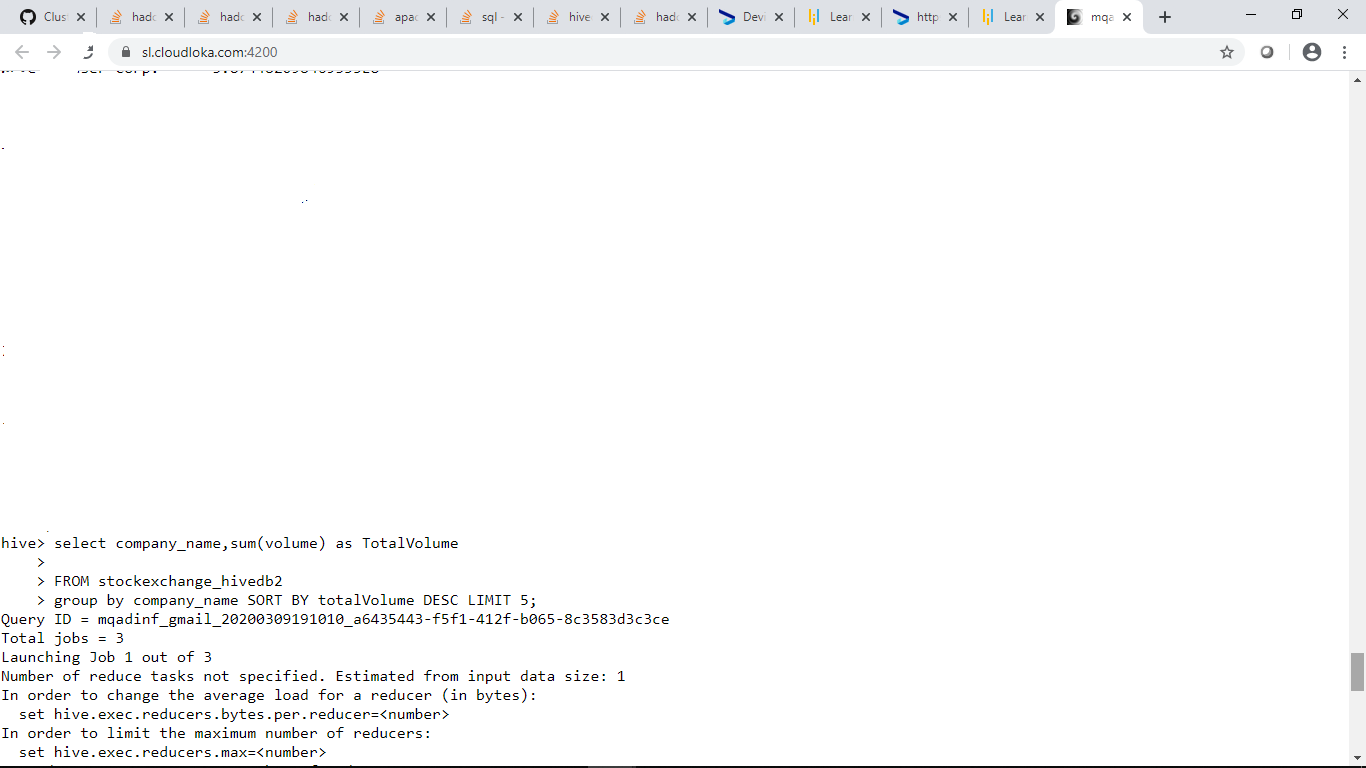
**Code:**

**select company\_name, sum(volume) as TotalVolume**

**FROM stockexchange\_hivedb2**

**Group by company\_name Sort by TotalVolume DESC LIMIT 5**

**Screen shot of the code**



**The Output:**

If we can see below we will get the first 5 companies with the TotalVolume which are higher.Therefore these are

the top 5 companies that are good for investment :Therefore the top 5 companies that are good for investment are:

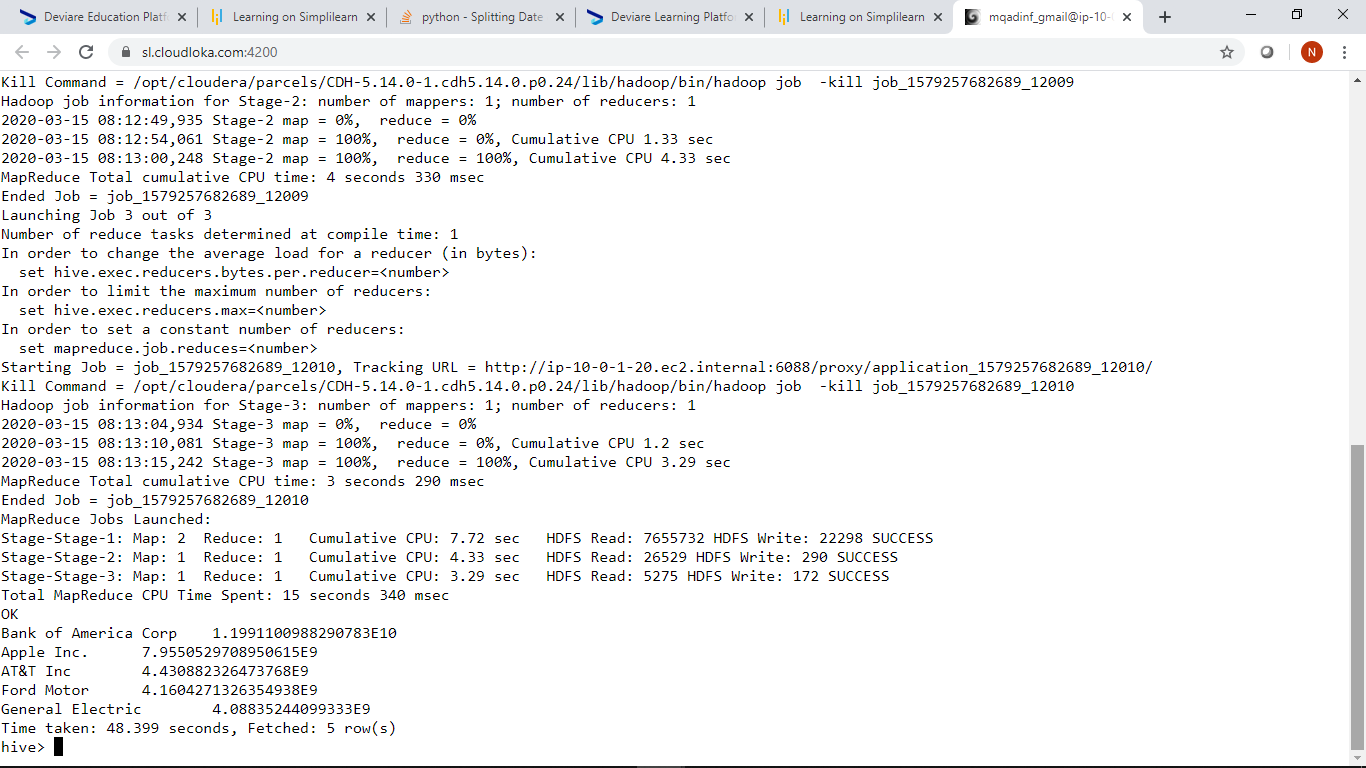
**Bank of America Corp**

**Apple Inc**

**AT&T Inc**

**Ford Motor**

**General Electric**



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

Here we will need to calculate the growth inorder to growing industry.With the help of the **LAG function** we can calculate the growth difference

with the current year value with that of the previous value,so the value we will use here is the volume value.

We will first create a new table called stockexchange\_hivedb3

We will select the columns state,trading\_year from the stockexchange\_hivedb2 table and then we add another column called Value\_Difference

The Value\_Difference column will calculate the volume value difference by subtracting the current volume value with that of the previous year volume value.

We accomplish this with the use of the LAG function.

Because we are required to show best growing industry having at least two or more industries mapped we require industries which are

separated by the symbol & For example :Trading & Inventory,Education & Maintenance etc.These are two or more industries mapped.

So to accomplish this we use a WHERE LIKE function.

**Code:**

**create table stockexchange\_hivedb3**

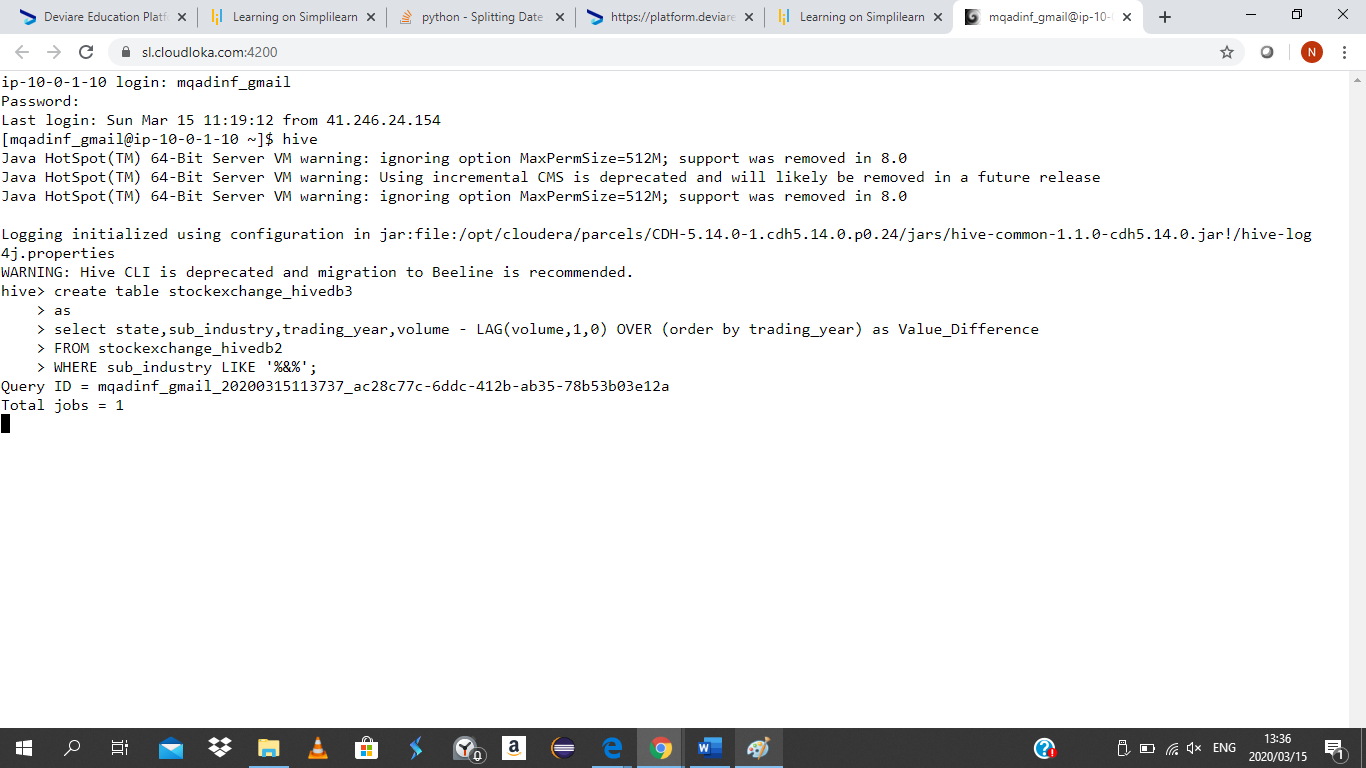
**as**

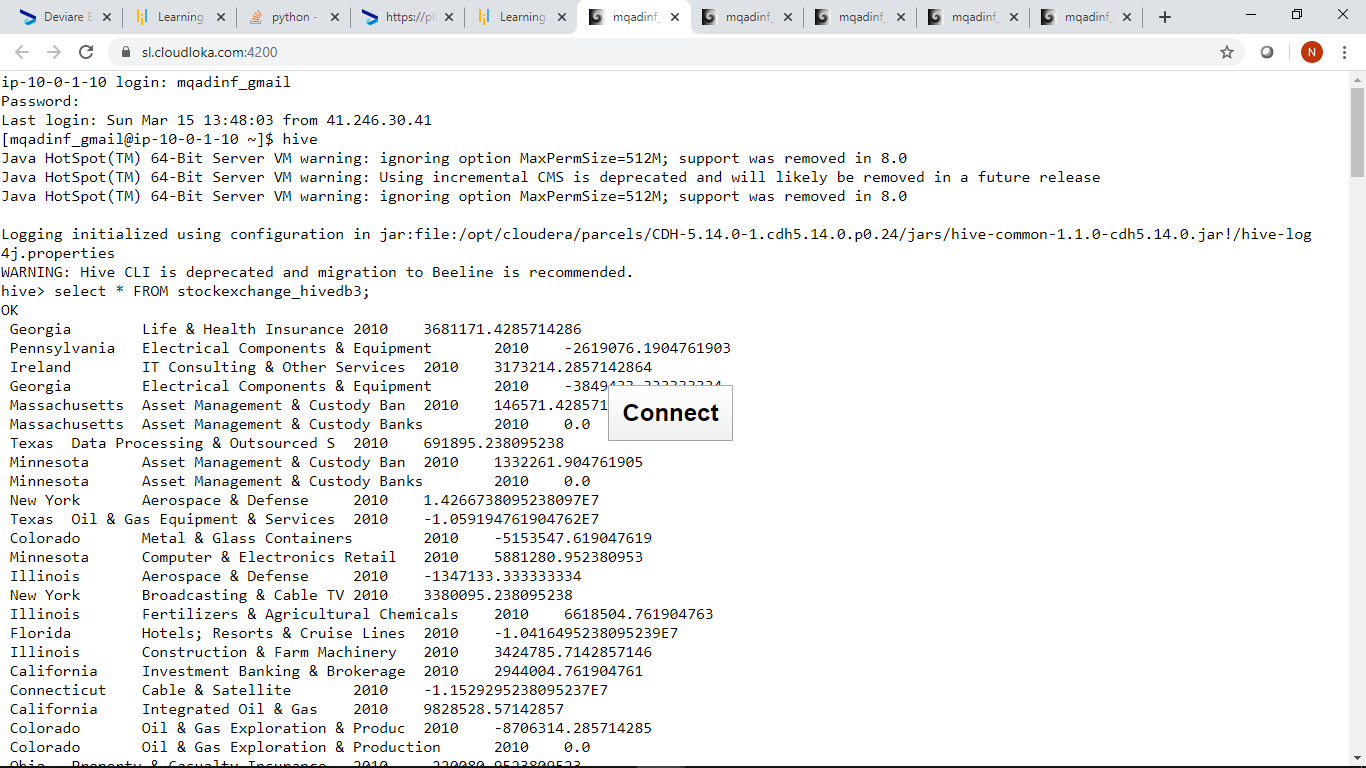
**select state, sub\_industry, trading\_year, volume - LAG(volume,1,0) OVER (order by trading\_year) as Value\_Difference**

**FROM stockexchange\_hivedb2**

**WHERE sub\_industry LIKE ‘%&%’;**

**Screen shot of the code**





We are not done yet.We need to show the best growing industry

To accomplish this we use a select statement where we select the maximum value difference,and we need to order by Max\_Diff in Descending order from highest to lowest.

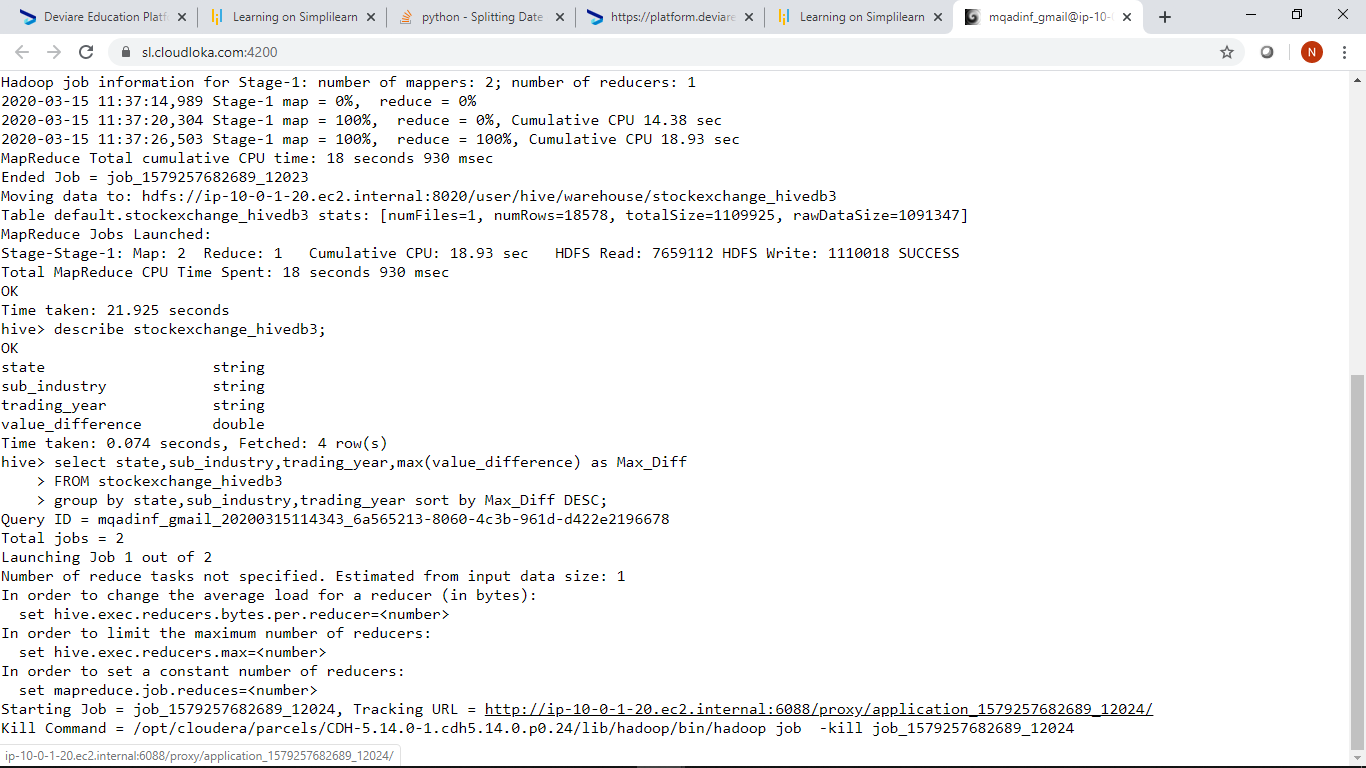
**Code:**

**select state,sub\_industry,trading\_year,max(value\_difference) as Max\_Diff**

**FROM stockexchange\_hivedb3**

**group by state,sub\_industry,trading\_year sort by Max\_Diff DESC;**

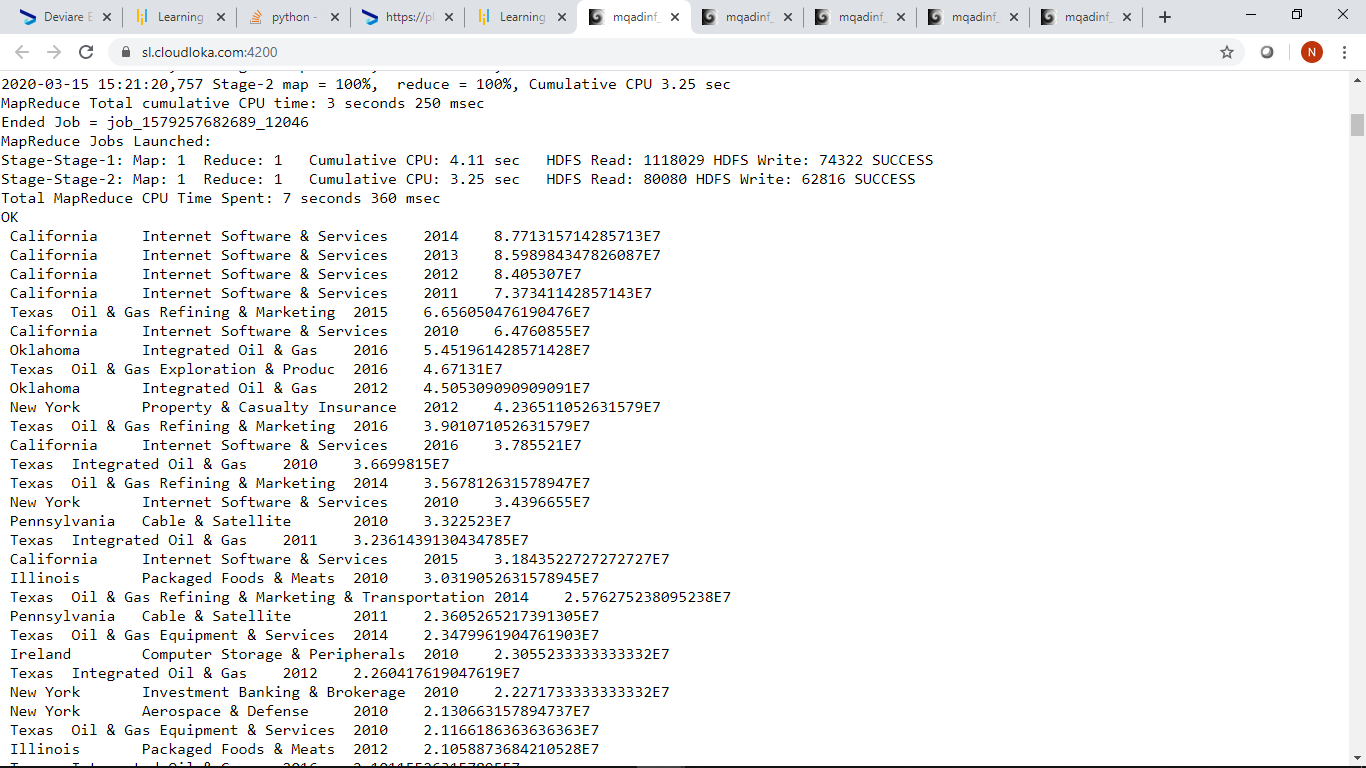
**Screen shot of the code**



**Screen shot of the output**

Here we can see that the Industry **Internet Software & Services** with the State of **California** is the first one appearing in Descending order.

Here we the all Industries are selected from Highest to lowest,but we need only one best growing industry.



To get one best growing industry here’s the code below:

We use LIMIT 1 to get one best growing industry.

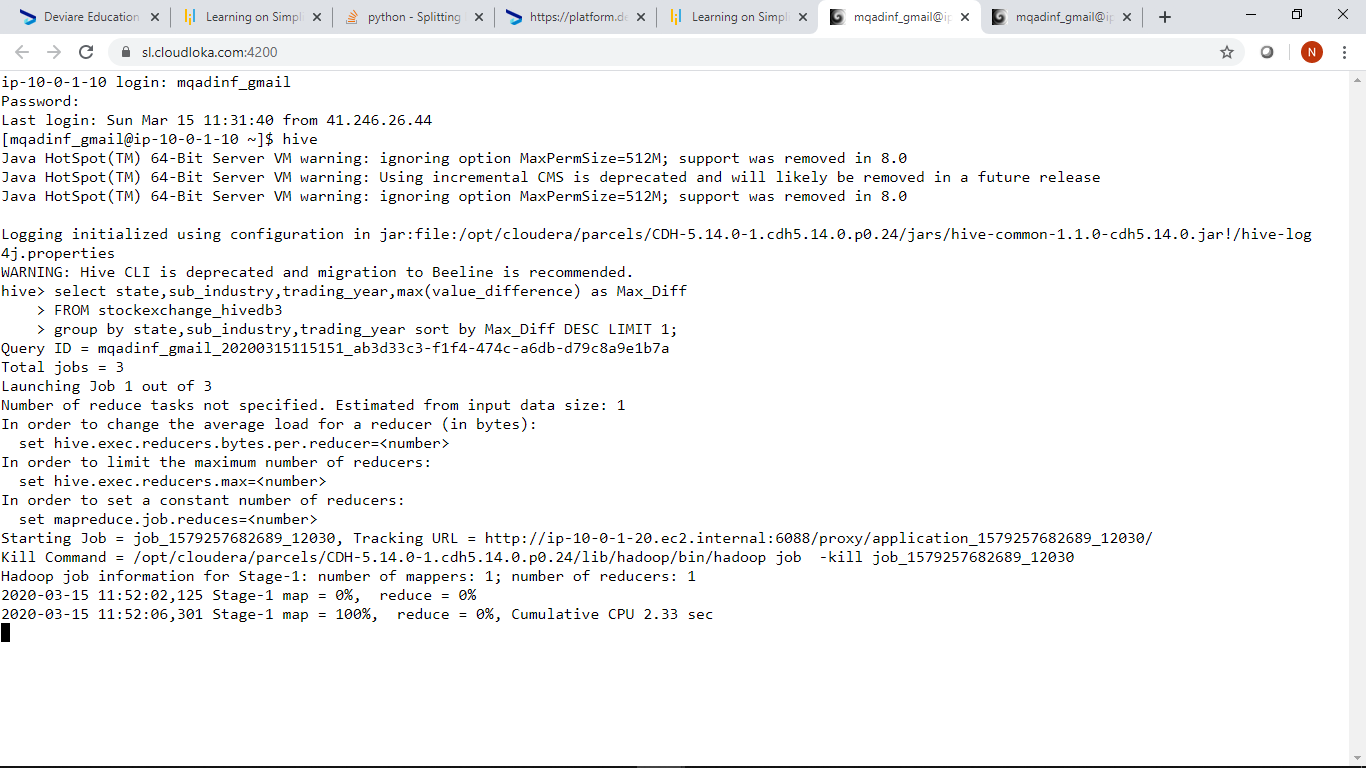
**Code:**

**Select state, sub\_industry, trading\_year, max(value\_difference) as Max\_Diff**

**FROM stockexchange\_hivedb3**

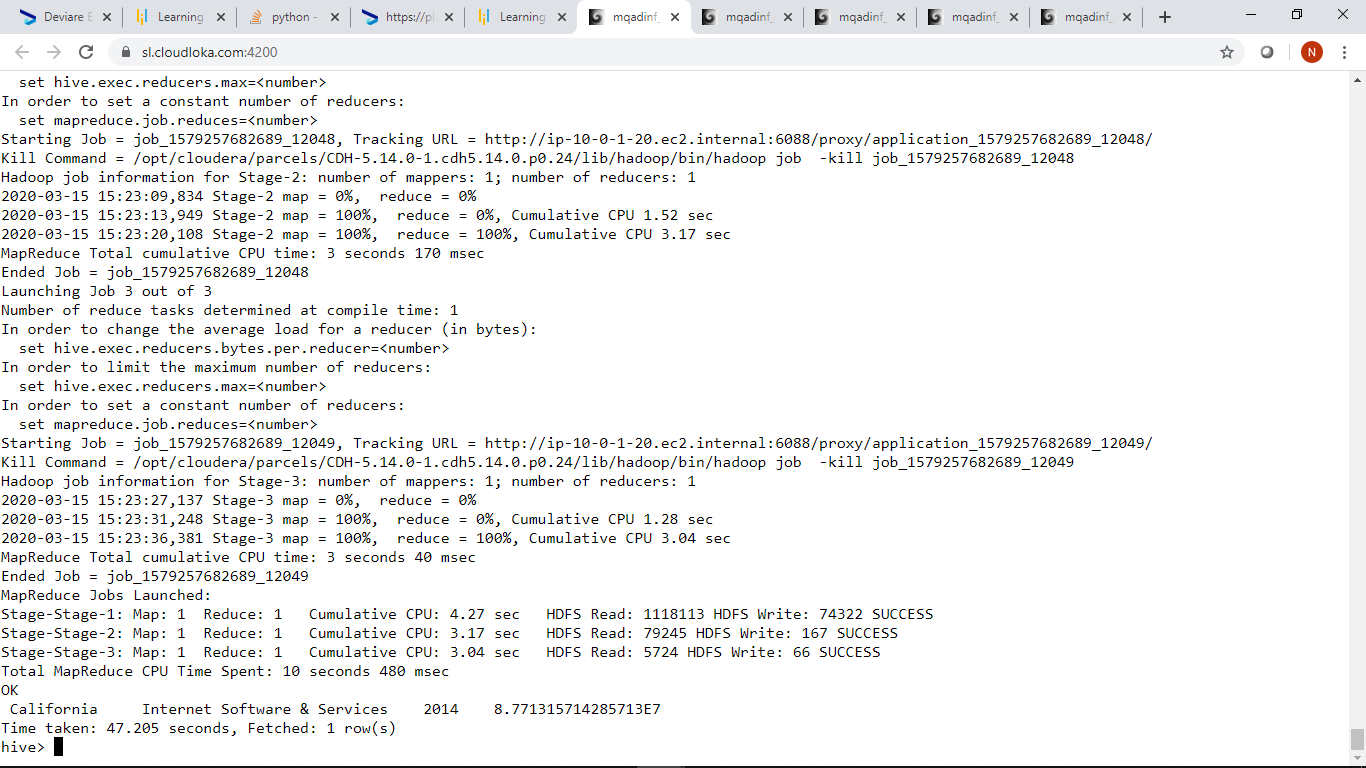
**group by state, sub\_industry, trading\_year sort by Max\_Diff DESC LIMIT 1;**

**Screen shot of the code**



**Screen shot of the output**

**Here the best growing industry is Internet Software & Services which is the state of California.**



**5) For each sector find the following.**

* **Worst year**
* **b. Best year**
* **c. Stable year**

Here we will need to create a new table named table stockexchange\_hivedb4

Here’s the code:

Here we will select the columns sector,trading\_year from the stockexchange\_hivedb2 table and add a new column TotalVolume with the use of sum function. We added a column TotalVolume because we need a TotalVolume based on each sector we need the Worst,Best and Stable year.The TotalVolume will give an indication on how the sector is performing

By sector and trading\_year.

**create table stockexchange\_hivedb4**

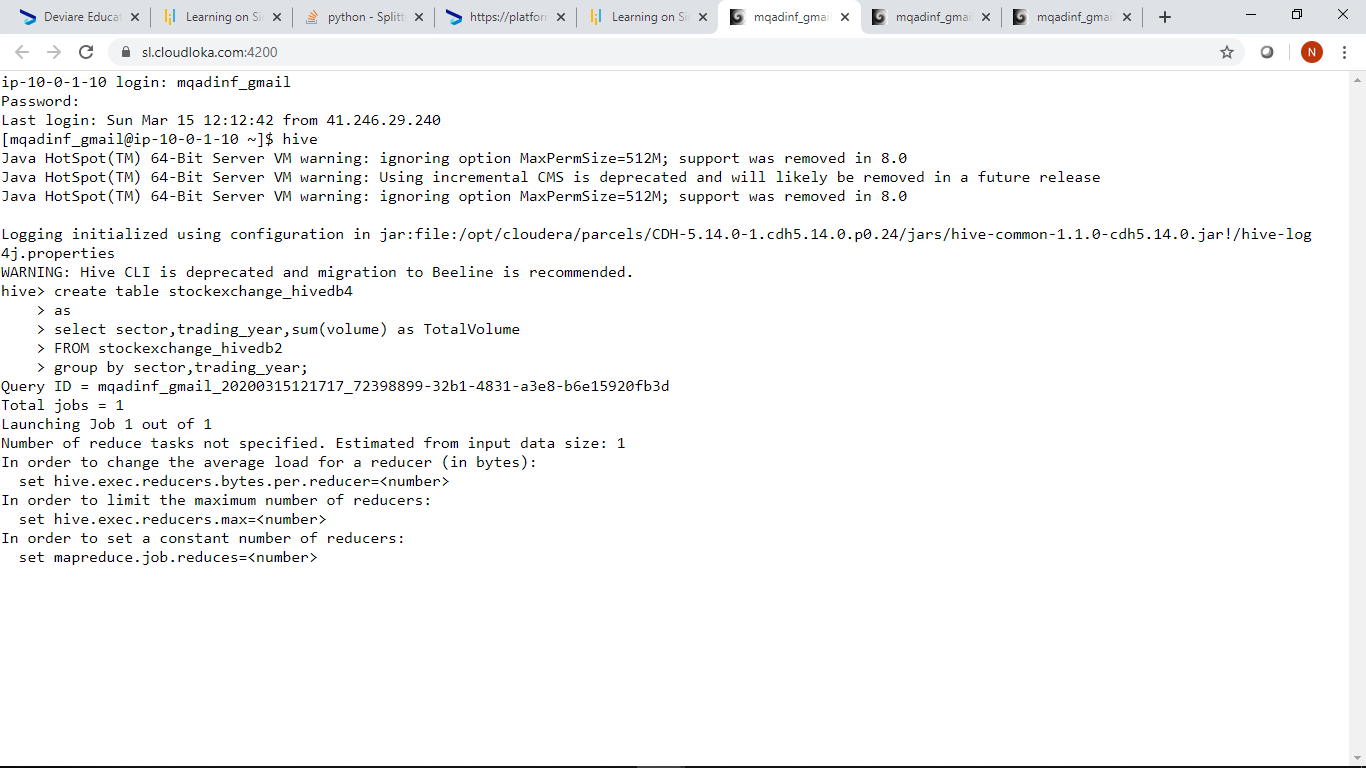
**as**

**select sector,trading\_year, sum(volume) as TotalVolume**

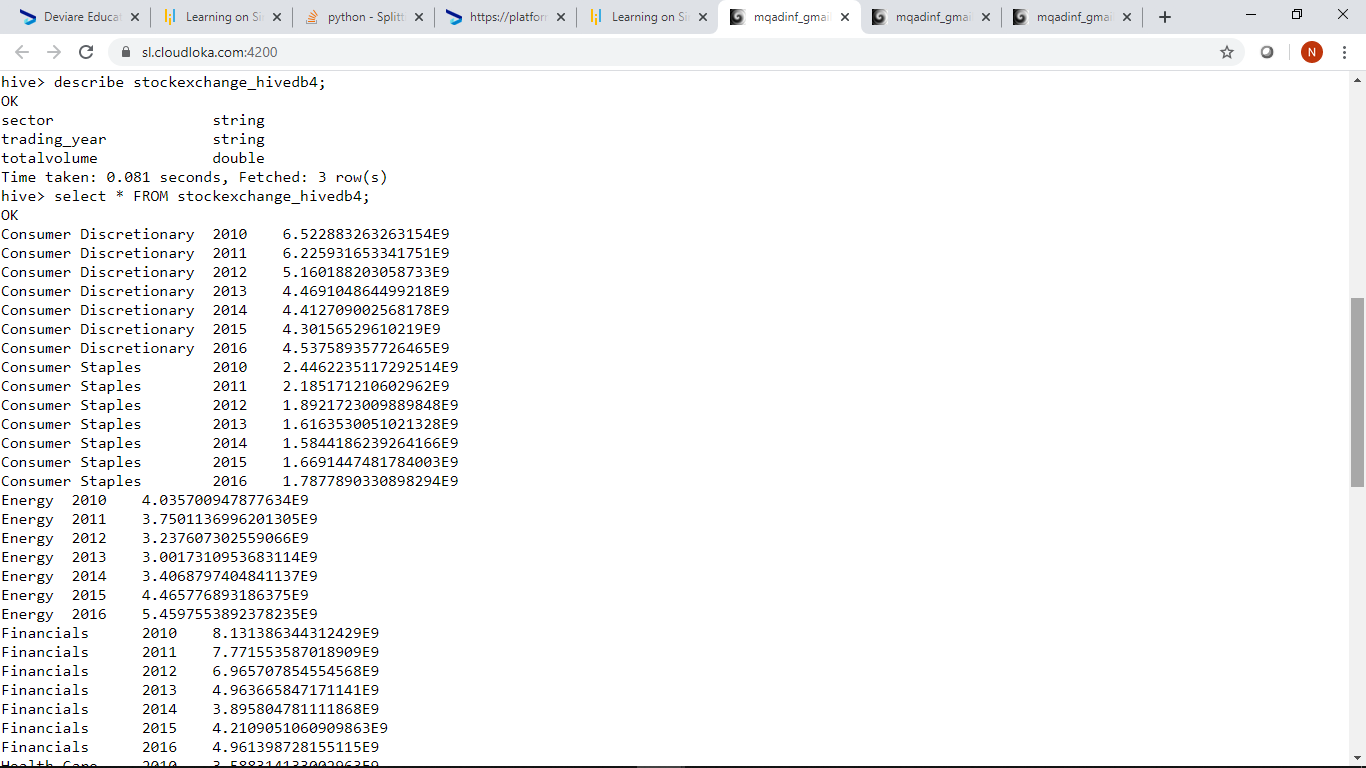
**FROM stockexchange\_hivedb2**

**group by sector,trading\_year;**

**Screen shot of the code**



With the use of the select statement this is how a stockexchange\_hivedb4 table looks like.



We are not done yet we need to find the Best,Worst and Stable Year for each sector

To find the Best Year:

To find the Best Year we select the maximum totalvolume based on each sector and trading\_year.

After that we group by sector and trading\_year .

Because we need to find the Best Year, we need to find the maximum totalvolume based on each sector and trading\_year.

We need to sort the Max\_Vol in Descending Order so that it will be easier for us to see the Max\_Vol based on each trading\_year.

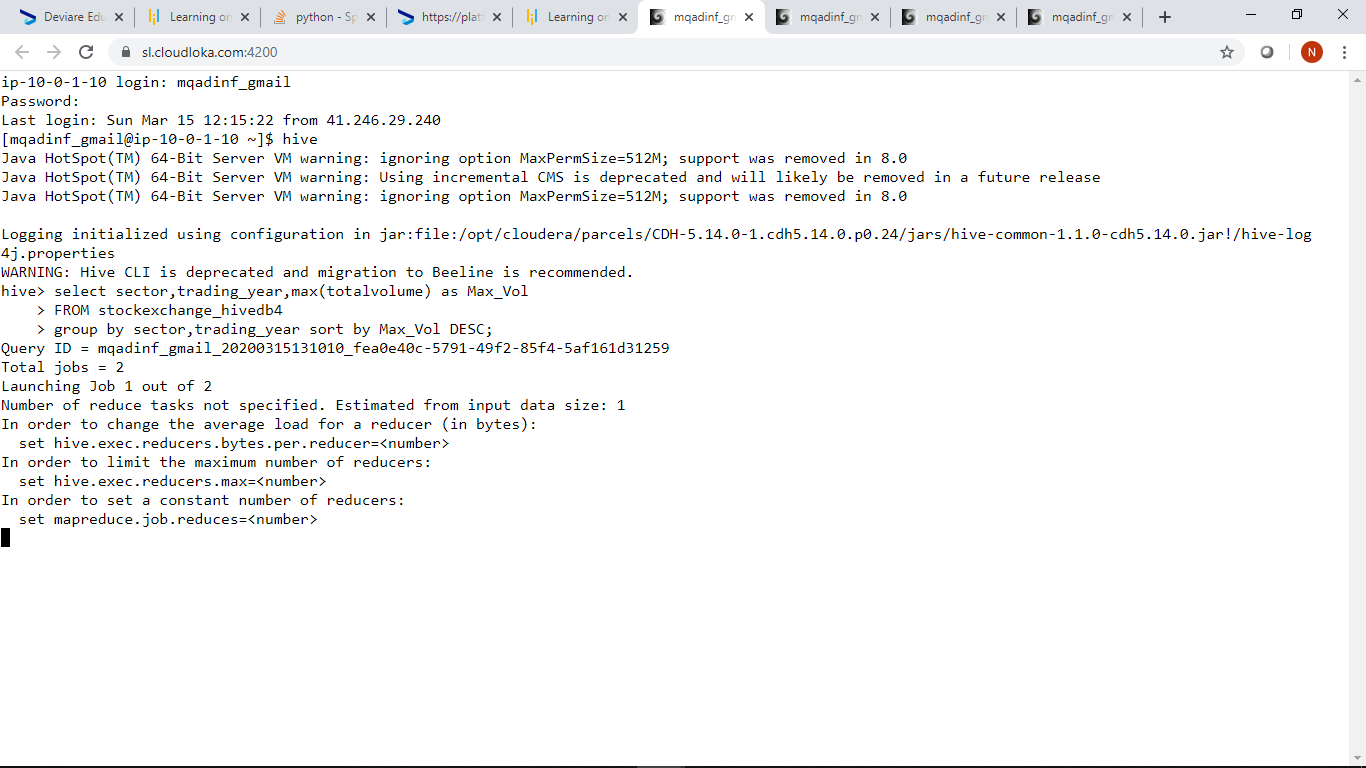
**Here’s the code:**

**Select sector,trading\_year,max(totalvolume) as Max\_Vol**

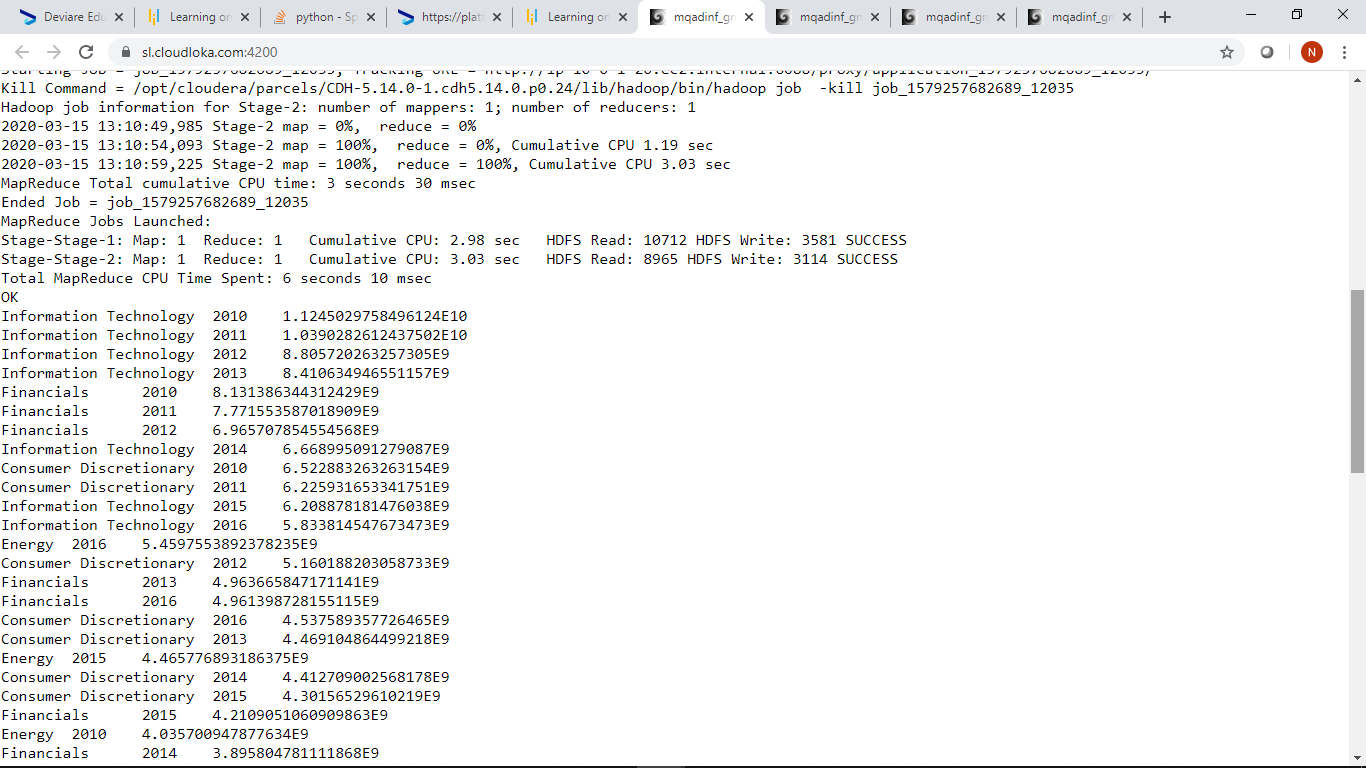
**FROM stockexchange\_hivedb4**

**group by sector,trading\_year sort by Max\_Vol DESC;**

**Screen shot of the code**



**Screen shot of output**



To get an answer for the Best Year where we find the Best year for each sector,we need to look at the topmost sector and trading\_year

**Answer For Best Year:**

**For Information Technology it is 2010**

**For Financials it is 2010**

**For Consumer Discretionary it is 2010**

**For Energy it is 2016**

**For Industrials it is 2010**

**For HealthCare it is 2010**

**For Consumer Staples it is 2010**

**For Telecommunication Services it is 2011**

**For Materials it 2010**

**For Utilities it is 2015**

**For Real Estate it is 2010**

To find the Worst Year we select minimum totalvolume based on each sector and trading\_year and after that we group by sector and trading\_year.

Because we need to find the Worst year,we need to find the minimum totalvolume based on each sector and trading\_year.

Here we will sort the Min\_Vol in ascending order so that it will be easier for us to see the Min\_Vol based on each trading\_year

**Here’s the code**

**select sector,trading\_year,min(totalvolume) as Min\_Vol**

**FROM stockexchange\_hivedb4**

**group by sector,trading\_year sort by Min\_Vol;**

To get an answer for the Worst year where we find the worst year for each sector,we need to look at the topmost and trading\_year.

**Answer for Worst Year:**

**For Real Estate it is 2014**

**For Utilities it is 2013**

**For Materials it is 2014**

**For Telecommunications Service it is 2013**

**For Consumer Staples it is 2014**

**For Health Care it is 2014**

**For Industrials it is 2014**

**For Energy it is 2013**

**For Financials it is 2014**

**For Consumer Discretionary it is 2015**

**For Information Technology it is 2016**

For Stable Year we need look at the middle position for each sector to get a Stable Year.As in each sector there a 7 duplicate sector values we need to

look at the middle position of each sector where we will find a stable year.

**For Real Estate there are 7 duplicate sectors values which has 7 distinct trading\_year values 2010,2011,2012,2013,2014,2015 and 2016**

**2014**

**2012**

**2013**

**2015**

**2016**

**2011**

**2010**

**So the Stable year for Real Estate is 2015**

**For Utilities there are also 7 duplicate sector values**

**2013**

**2012**

**2014**

**2016**

**2011**

**2010**

**2015**

**The Stable year for Utilities is 2016**

**For Materials there are also 7 duplicate sector values**

**2014**

**2013**

**2012**

**2015**

**2016**

**2011**

**2010**

**The Stable year for Materials is 2015**

**For Telecommunications Services there are also 7 duplicate sector values**

**2013**

**2012**

**2014**

**2016**

**2015**

**2010**

**2011**

**The Stable year for Telecommunications Services is 2016**

**For Consumer Staples there are also 7 duplicate sector values**

**2014**

**2013**

**2015**

**2016**

**2012**

**2011**

**2010**

**The Stable year for Consumer Staples is 2016**

**For Health Care there are also 7 duplicate sector values**

**2014**

**2015**

**2016**

**2013**

**2012**

**2011**

**2010**

**The Stable year for Health Care is 2013**

**For Industrials there are also 7 duplicate sector values**

**2014**

**2013**

**2016**

**2015**

**2012**

**2011**

**2010**

**The Stable year for Industrials is 2015**

**For Energy there are also 7 duplicate sector values**

**2013**

**2012**

**2014**

**2011**

**2010**

**2015**

**2016**

**The Stable year for Energy is 2011**

**For Financials there are also 7 duplicates values**

**2014**

**2015**

**2016**

**2013**

**2012**

**2011**

**2010**

**The Stable year for Financials is 2013**

**For Consumer Discretionary there are also 7 duplicate values**

**2015**

**2014**

**2013**

**2016**

**2012**

**2011**

**2010**

**The Stable year for Consumer Discretionary is 2016**

**And for Information Technology there are 7 duplicate values**

**2016**

**2015**

**2014**

**2013**

**2012**

**2011**

**2010**

**The Stable year for Information Technology is 2013.**