**Stock Exchange Data Analysis**

DESCRIPTION

**Objective**: To use hive features for data engineering or analysis and sharing the actionable insights

**Problem Statement:**  
NewYork stock exchange data of seven years, between 2010 to 2016, is captured for 500+ listed companies. The data set comprises of intra-day prices and volume traded for each listed company. The data serves both for machine learning and exploratory analysis projects, to automate the trading process and to predict the next trading-day winners or losers.. The scope of this project is limited to exploratory data analysis.

**Domain**: BFSI

**Analysis to be done:** Exploratory analysis to understand how MoM or YoY companies from different sectors or industries and states have progressed in a period of 7 years

**Content:** This data set contains prices.csv and securities.csv files having the following features:

Prices.csv:

1. Date: Trading date
2. Symbol: Ticker code or listed company code on NY stock exchange
3. Open: Intra-day opening price for each listed company
4. Close: Intra-day closing price for each listed company
5. Low: Intra-day lowest price for each listed company
6. High: Intra-day highest price for each listed company
7. Volume: Number of shares traded per day per company

Securities.csv:

1. Ticker\_Symbol: Country to which the customer belongs
2. Security: Legal name of the listed company
3. Sector: Business vertical of the listed company
4. Sub\_Industry: Business domain of the listed company within a Sector.
5. Headquarter: Headquarters address

**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

Check the TABLE description: STOCK\_PRICES

|  |  |
| --- | --- |
| Column Name | Datatype |
| Trading\_date | Date |
| Symbol | String |
| Open | double |
| Close | double |
| Low | double |
| High | double |
| Volume | int |

TABLE: STOCK\_COMPANIES

|  |  |
| --- | --- |
| Column Name | Datatype |
| Symbol | String |
| Company\_name | String |
| Sector | String |
| Sub\_industry | String |
| Headquarter | String |

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).

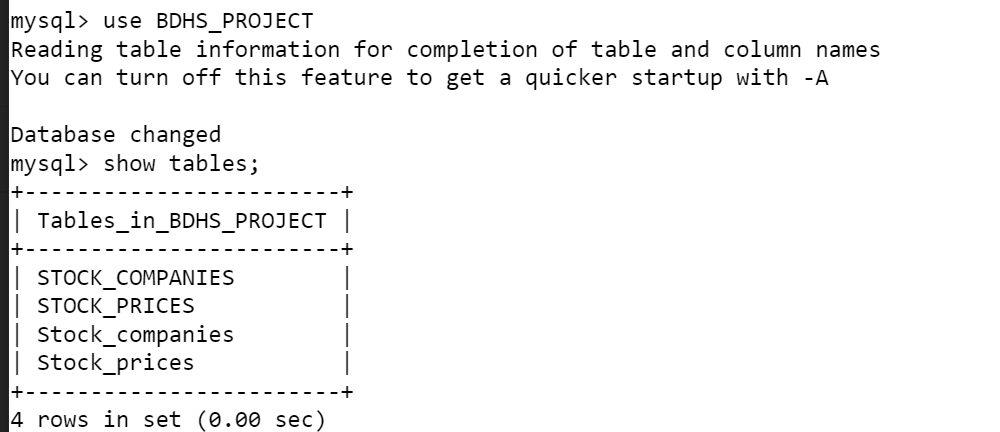
* Trading\_year: Should contain YYYY for each record
* Trading\_month: Should contain MM or MMM for each record
* Symbol: Ticker code
* CompanyName: Legal name of the listed company
* State: State to be extracted from headquarters value.
* Sector: Business vertical of the listed company
* Sub\_Industry: 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

**DATA ANALYSIS USING HIVE**

**First Step get data to hive**

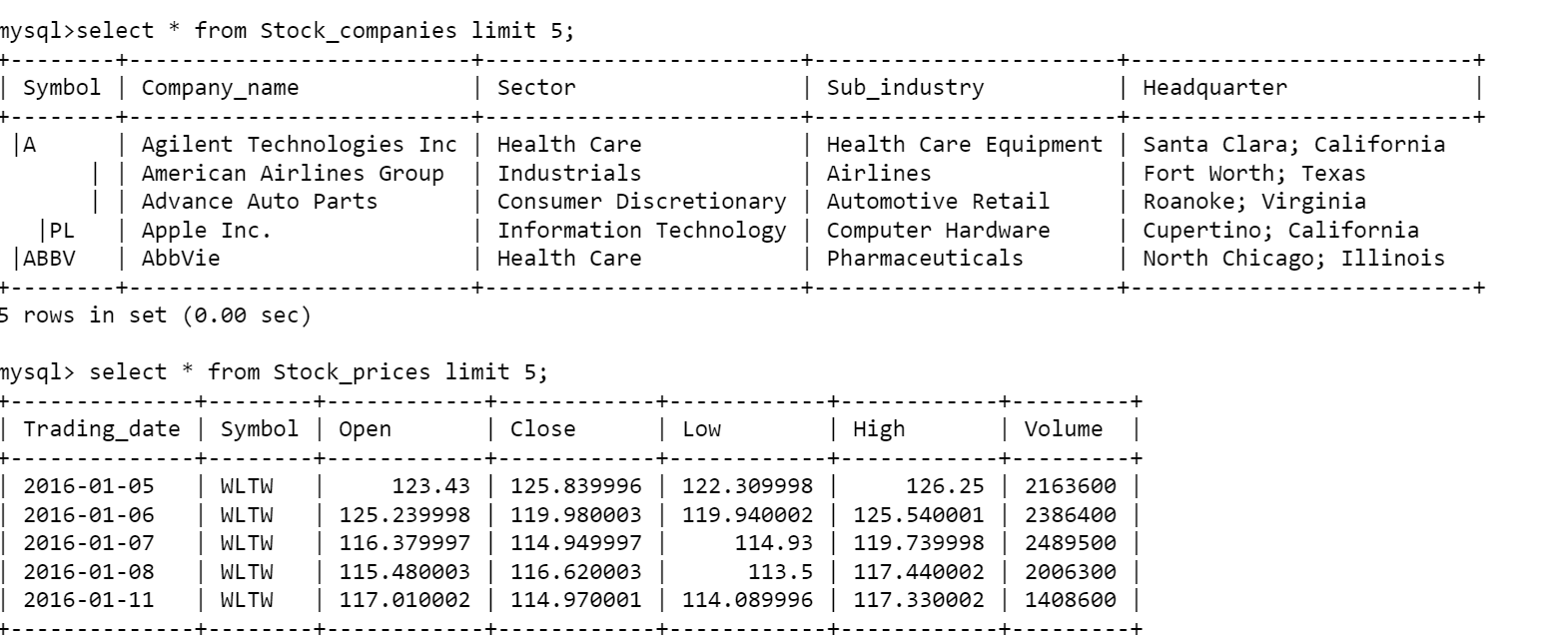
**Solution:**

**Use the mysql database to check if tables are present;**



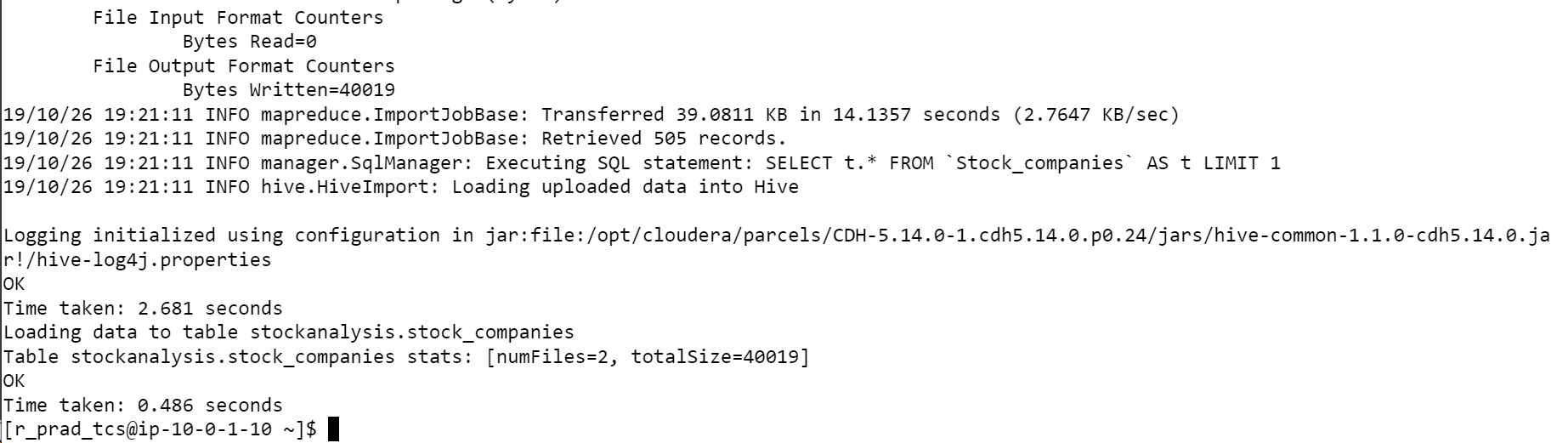
I was have an access issue so I create ticket to support team and they already get BDHS\_PROJECT database to hive so below steps just code to show how to get table to hive by using sqoop.

**Check table contents**

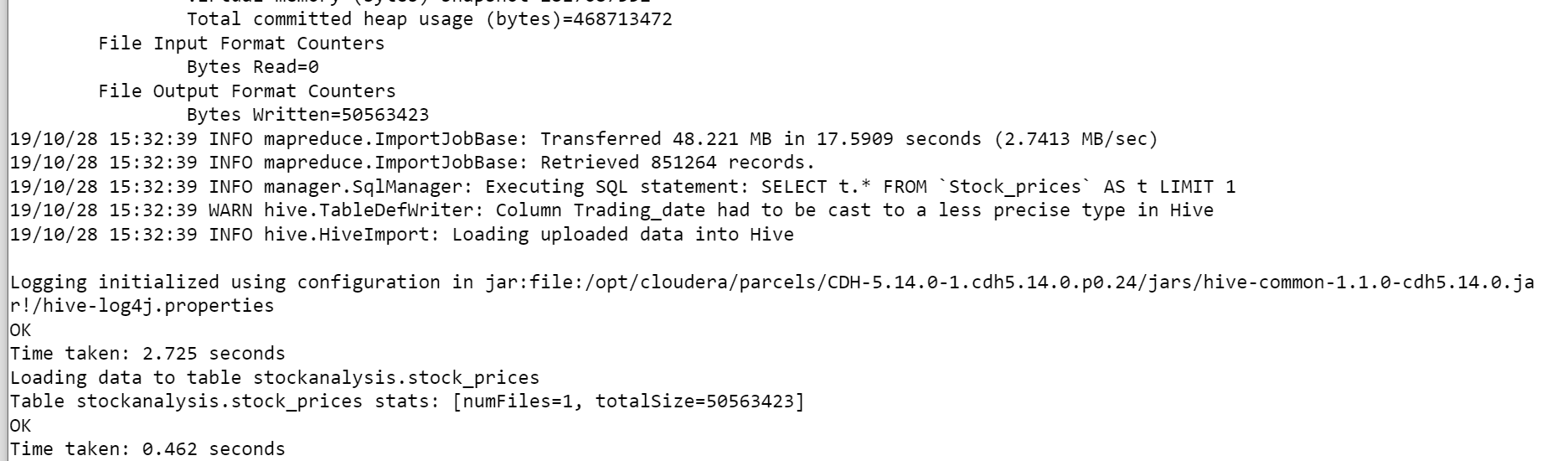


**Now import the data from mysql using sqoop to hive**

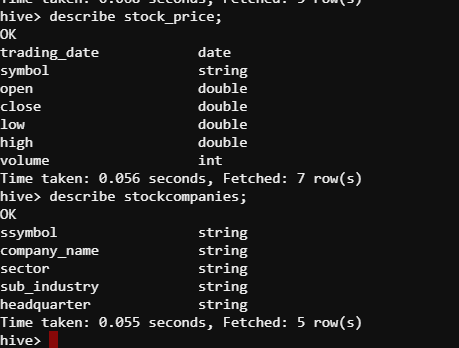
sqoop import --connect jdbc:mysql://localhost/BDHS\_PROJECT --username adhamfcihgmail --password \*\*\*\*\* --table Stock\_compaies --hive-import --hive-database stockanalysis -m 1



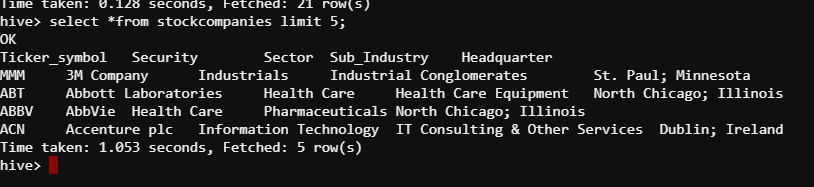
sqoop import --connect jdbc:mysql://ip-10-0-1-10.ec2.internal/BDHS\_PROJECT --username adhamfcihgmail --password \*\*\*\*\*\* --table Stock\_companies –hive-import -hive-database stockanalysis --m 1

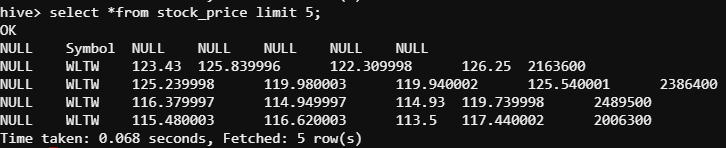


**Check structures of hive table:**



**Check contents of hive tables:**





**Create a new hive table by joining the above 2 hive tables by using appropriate hive built in functions for columns(a,b,e,h to l)**

**First step we create table with above requirement**

**create table stock\_data\_Adham\_P as select trading\_year,**

**trading\_month, sc.symbol, company\_name, trim(split(headquarter,"\;")[1]) state,**

**sector, sub\_industry, open, close, low, high, volume**

**from stock\_companies sc,**

**(select symbol, year(trading\_date) trading\_year, month(trading\_date) trading\_month,**

**round(avg(open),2) open, round(avg(close),2) close, round(avg(low),2) low,**

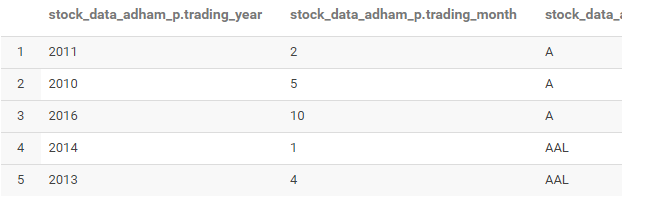
**round(avg(high),2) high, round(avg(volume),2) volume**

**from stock\_prices**

**group by symbol, month(trading\_date),year(trading\_date)) sp**

**where sc.symbol=sp.symbol;**

Now show content of table.

**select\*from stock\_data\_Adham\_P limit 5;** 

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

Step 1: Create ­table with required data

create table company\_horizon\_Adham\_P as select

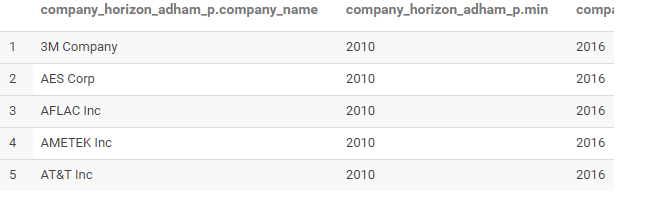
company\_name, min(trading\_year) min, max(trading\_year) max,

min(trading\_month) min\_month, max(trading\_month) max\_month

from stock\_data\_Adham\_P

group by company\_name;

select \*from company\_horizon\_Adham\_P limit 5;



Step 2: Get Top 5 growth based on the data in above table created

**select stock\_start.company\_name,**

**((close-open)/open)\*100 growth\_percent**

**from (select chv.company\_name, open from stock\_data\_Adham\_P sd, company\_horizon\_Adham\_P chv**

**where sd.trading\_year = chv.min**

**and sd.trading\_month = chv.min\_month**

**and sd.company\_name = chv.company\_name) stock\_start,**

**(select chv.company\_name,close from stock\_data\_Adham\_P sd, company\_horizon\_Adham\_P chv**

**where sd.trading\_year = chv.max**

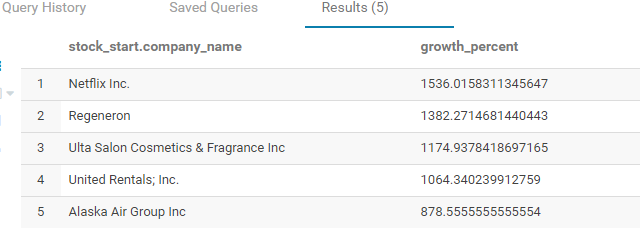
**and sd.trading\_month = chv.max\_month**

**and sd.company\_name = chv.company\_name) stock\_end**

**where stock\_start.company\_name = stock\_end.company\_name**

**sort by growth\_percent desc limit 5;**

**This shows the top 5 companies by growth**



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

Step 1:

**create table company\_growth\_Adham\_P as select**

**state, sub\_industry, stock\_start.company\_name, ((stock\_end.close-stock\_start.open)/stock\_start.open)\*100 growth\_percent**

**from (select chv.company\_name,open**

**from stock\_data\_Adham\_P sd, company\_horizon\_Adham\_P chv**

**where sd.trading\_year=chv.min and**

**sd.trading\_month=chv.min\_month and**

**sd.company\_name=chv.company\_name)stock\_start,**

**(select chv.company\_name, close**

**from stock\_data\_Adham\_P sd, company\_horizon\_Adham\_P chv**

**where sd.trading\_year=chv.max and**

**sd.trading\_month=chv.max\_month and**

**sd.company\_name=chv.company\_name)stock\_end,**

**(select company\_name, state, sub\_industry**

**from stock\_data\_Adham\_P**

**group by company\_name,state,sub\_industry)sd**

**where (stock\_end.close-stock\_start.open)>0 and**

**stock\_start.company\_name=stock\_end.company\_name and**

**sd.company\_name=stock\_start.company\_name;**

|  |  |  |  |
| --- | --- | --- | --- |
| **company\_growth\_adham\_p.state** | **company\_growth\_adham\_p.sub\_industry** | **company\_growth\_adham\_p.company\_name** | **company\_growth\_adham\_p.growth\_percent** |
| Minnesota | Industrial Conglomerates | 3M Company | 112.6081731 |
| Georgia | Life & Health Insurance | AFLAC Inc | 38.92215569 |
| Pennsylvania | Electrical Components & Equipment | AMETEK Inc | 192.1580189 |
| Texas | Integrated Telecommunications Services | AT&T Inc | 55.25028227 |
| Illinois | Pharmaceuticals | AbbVie | 73.49295775 |

Step 2:

**create table industry\_growth\_Adham\_P as**

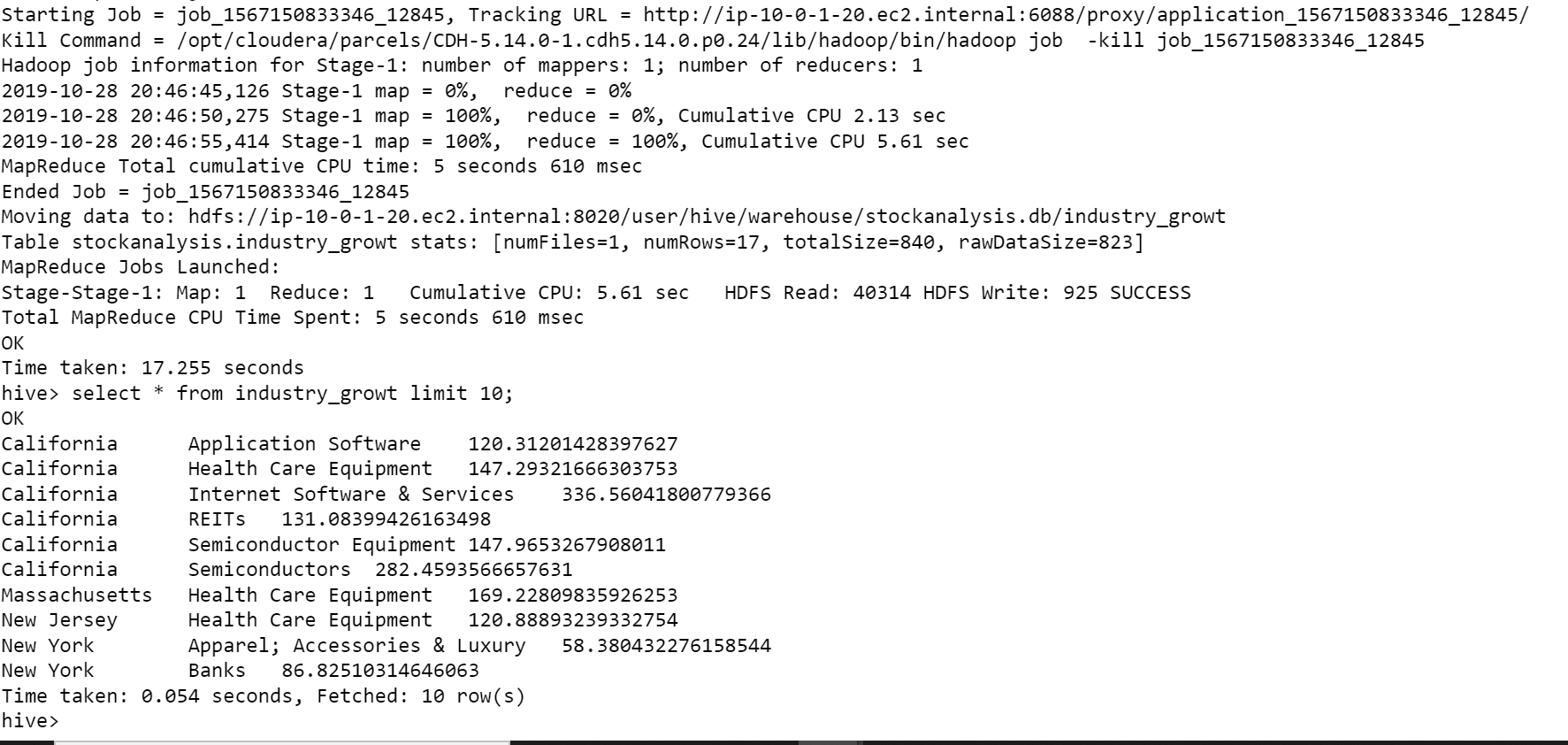
**select state,sub\_industry,**

**avg(growth\_percent) ind\_growth**

**from company\_growth\_Adham\_P**

**group by state, sub\_industry**

**having count(sub\_industry>2);**



Step 3:

**select ig.state, sub\_industry, ind\_growth**

**from industry\_growt ig,**

**(select state,max(ind\_growth) max\_growth**

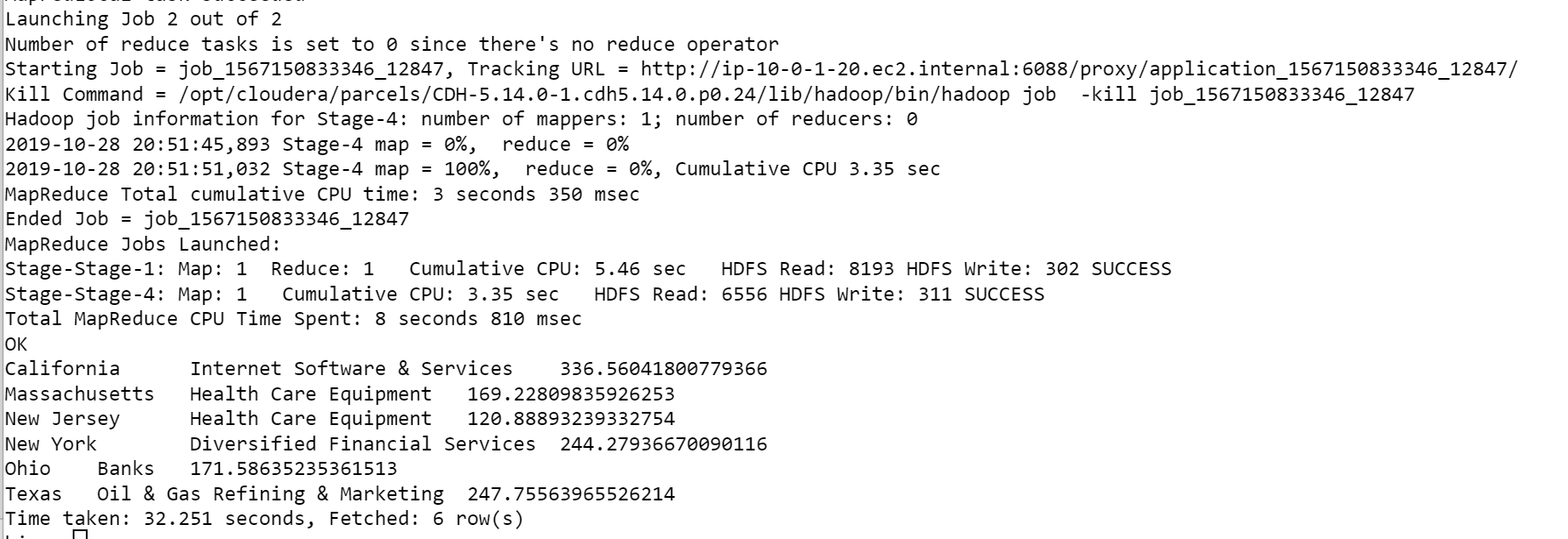
**from industry\_growth\_Adham\_P**

**group by state) inn\_ig**

**where inn\_ig.state = ig.state and**

**ig.ind\_growth = inn\_ig.max\_growth;**

This shows the best growing industry by each state.



3) For each sector find the following.

* Worst year
* Best year
* Stable year

Step 1: create table with sector growth

**create table sector\_growth\_Adham\_P as select open.sector, open.trading\_year,(close-open) growth**

**from (select sector,trading\_year,avg(open) open**

**from stock\_data\_Adham\_P**

**where trading\_month = 1**

**group by sector,trading\_year) open,**

**(select sector,trading\_year,avg(close) close**

**from stock\_data\_Adham\_P**

**where trading\_month=12**

**group by sector,trading\_year) close**

**where open.sector = close.sector and**

**open.trading\_year = close.trading\_year;**

**Select\*from sector\_growth\_Adham\_P limit 5;**



Step 2:

For the worst trading year by sector

**select x.sector,x.trading\_year,x.growth**

**from sector\_growth\_Adham\_P x,**

**(select sector,min(growth) growth**

**from sector\_growth\_Adham\_P**

**group by sector) y**

**where x.sector=y.sector and**

**x.growth=y.growth;**

As we see worest year for every sector.



For the best trading year by sector:

**select x.sector,x.trading\_year,x.growth**

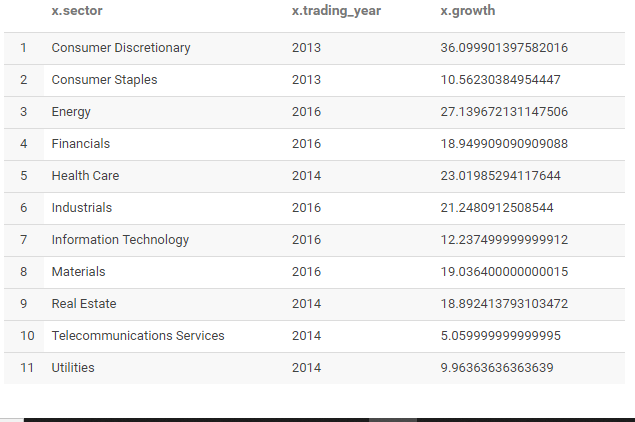
**from sector\_growth\_Adham\_P x,**

**(select sector,max(growth) growth**

**from sector\_growth\_Adham\_P**

**group by sector) y**

**where x.sector=y.sector and** x.growth=y.growth;

**As we see best year for every sector**

**For the stable year by sector**

For stable we get average growth

select x.sector,x.trading\_year,x.growth

from sector\_growth\_Adham\_P x,

(select sector,round(avg(growth),0) growth

from sector\_growth\_Adham\_P

group by sector) y

where x.sector=y.sector and

x.growth=y.growth

