Question – Project 1:

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
Trading_date
Symbol

Open
Close
Low
High
Volume

TABLE: STOCK_COMPANIES

Column Name
Symbol
Company_name
Sector
Sub_industry
Headquarter

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

3) Find the top five companies that are good for investment
4) Show the best-growing industry by each state, having at least two or more industries mapped.
5) For each sector find the following.

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

Write Up & Screenshots

Log in using Web Console:

MySQL Shell Command Structure

Username: nasehasgmail

Password: nasehasgmailonz2c

At prompt:

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

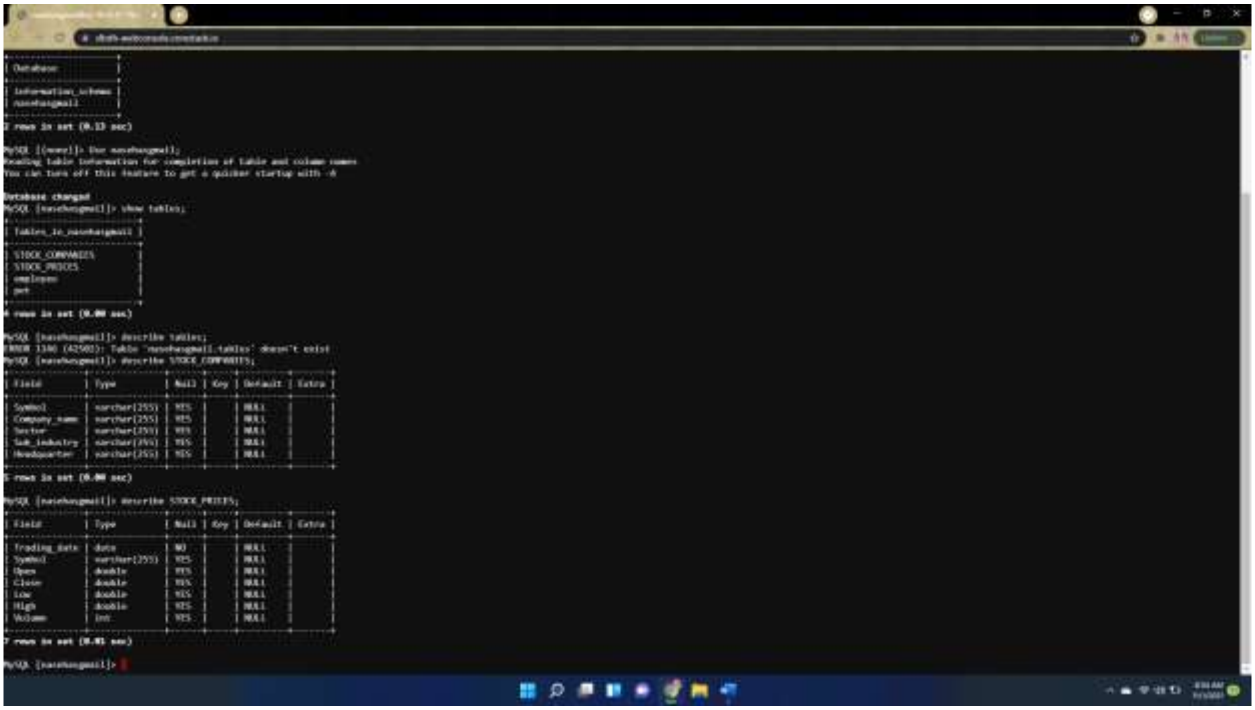
mysql -h sqoopdb.slbdh.cloudlabs.com -u nasehasgmail -pnasehasgmailonz2c

Show databases;

Use nasehasgmail;

Describe STOCK_COMPANIES;

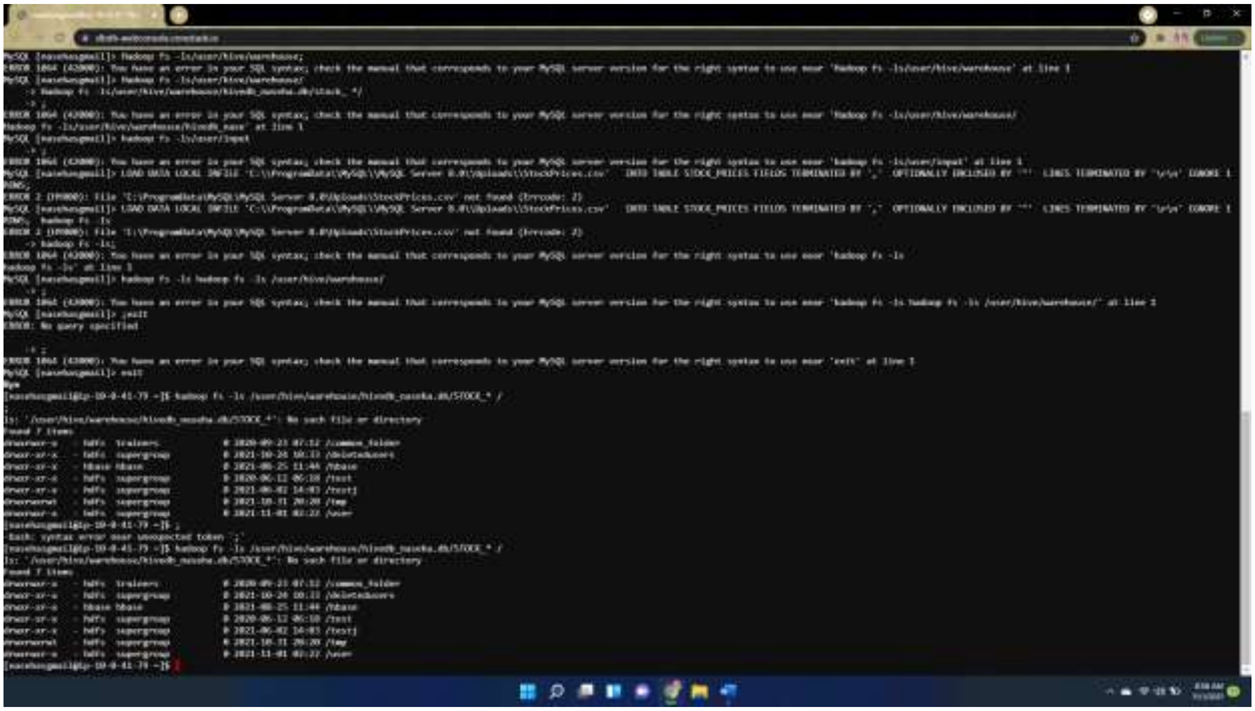
Describe STOCK_PRICES;



```
+--------------------+
| Database           |
+--------------------+
| information_schema |
| nasehasgmail       |
+--------------------+
2 rows in set (0.13 sec)

MySQL [(none)]> Use nasehasgmail;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
MySQL [nasehasgmail]> show tables;
+------------------------+
| Tables_in_nasehasgmail |
+------------------------+
| STOCK_COMPANIES        |
| STOCK_PRICES           |
| employee               |
| pet                    |
+------------------------+
4 rows in set (0.00 sec)

MySQL [nasehasgmail]> describe tables;
ERROR 1146 (42S02): Table 'nasehasgmail.tables' doesn't exist
MySQL [nasehasgmail]> describe STOCK_COMPANIES;
+--------------+--------------+------+-----+---------+-------+
| Field        | Type         | Null | Key | Default | Extra |
+--------------+--------------+------+-----+---------+-------+
| Symbol       | varchar(255) | YES  |     | NULL    |       |
| Company_name | varchar(255) | YES  |     | NULL    |       |
| Sector       | varchar(255) | YES  |     | NULL    |       |
| Sub_industry | varchar(255) | YES  |     | NULL    |       |
| Headquarter  | varchar(255) | YES  |     | NULL    |       |
+--------------+--------------+------+-----+---------+-------+
5 rows in set (0.00 sec)

MySQL [nasehasgmail]> describe STOCK_PRICES;
+--------------+--------------+------+-----+---------+-------+
| Field        | Type         | Null | Key | Default | Extra |
+--------------+--------------+------+-----+---------+-------+
| Trading_date | date         | NO   |     | NULL    |       |
| Symbol       | varchar(255) | YES  |     | NULL    |       |
| Open         | double       | YES  |     | NULL    |       |
| Close        | double       | YES  |     | NULL    |       |
| Low          | double       | YES  |     | NULL    |       |
| High         | double       | YES  |     | NULL    |       |
| Volume       | int          | YES  |     | NULL    |       |
+--------------+--------------+------+-----+---------+-------+
7 rows in set (0.01 sec)

MySQL [nasehasgmail]>
```

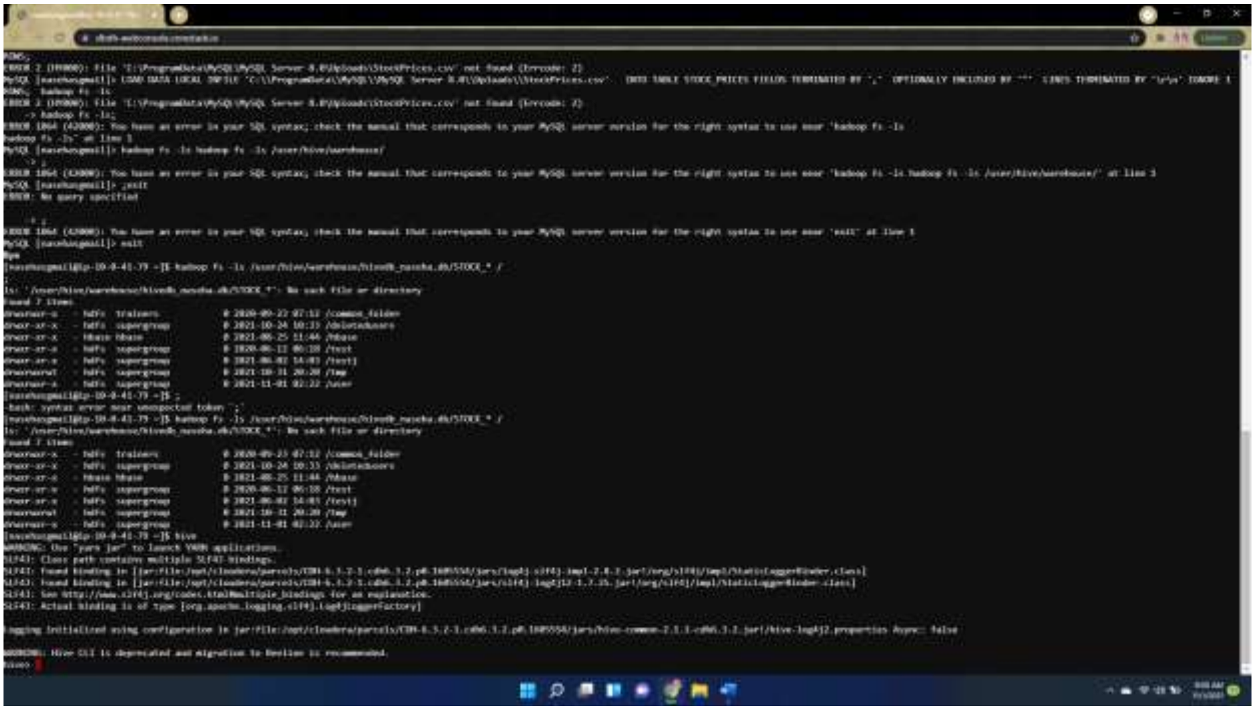
MYSQL> Exit

```
[nasehasgmail@ip-10-0-41-79 ~]$ hadoop fs -ls
/user/hive/warehouse/hivedb_naseha.db/STOCK_* /
```



```
MySQL [nasehasgmail]> Hadoop fs -ls/user/hive/warehouse;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near 'Hadoop fs -ls/user/hive/warehouse' at line 1
MySQL [nasehasgmail]> Hadoop fs -ls/user/hive/warehouse/
    -> hadoop fs -ls/user/hive/warehouse/hivedb_naseha.db/stock_ */
    -> ;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near 'Hadoop fs -ls/user/hive/warehouse/
hadoop fs -ls/user/hive/warehouse/hivedb_nase' at line 1
MySQL [nasehasgmail]> hadoop fs -ls/user/input
    -> ;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near 'hadoop fs -ls/user/input' at line 1
MySQL [nasehasgmail]> LOAD DATA LOCAL INFILE 'C:\\ProgramData\\MySQL\\MySQL Server 8.0\\Uploads\\StockPrices.csv'   INTO TABLE STOCK_PRICES FIELDS TERMINATED BY ','  OPTIONALLY ENCLOSED BY '"'  LINES TERMINATED BY '\r\n' IGNORE 1
ROWS;
ERROR 2 (HY000): File 'C:\ProgramData\MySQL\MySQL Server 8.0\Uploads\StockPrices.csv' not found (Errcode: 2)
MySQL [nasehasgmail]> LOAD DATA LOCAL INFILE 'C:\\ProgramData\\MySQL\\MySQL Server 8.0\\Uploads\\StockPrices.csv'   INTO TABLE STOCK_PRICES FIELDS TERMINATED BY ','  OPTIONALLY ENCLOSED BY '"'  LINES TERMINATED BY '\r\n' IGNORE 1
ROWS;  hadoop fs -ls
ERROR 2 (HY000): File 'C:\ProgramData\MySQL\MySQL Server 8.0\Uploads\StockPrices.csv' not found (Errcode: 2)
    -> hadoop fs -ls;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near 'hadoop fs -ls
hadoop fs -ls' at line 1
MySQL [nasehasgmail]> hadoop fs -ls hadoop fs -ls /user/hive/warehouse/
    -> ;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near 'hadoop fs -ls hadoop fs -ls /user/hive/warehouse/' at line 1
MySQL [nasehasgmail]> ;exit
ERROR: No query specified

    -> ;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near 'exit' at line 1
MySQL [nasehasgmail]> exit
Bye
[nasehasgmail@ip-10-0-41-79 ~]$ hadoop fs -ls /user/hive/warehouse/hivedb_naseha.db/STOCK_* /
;
ls: '/user/hive/warehouse/hivedb_naseha.db/STOCK_*': No such file or directory
Found 7 items
drwxrwxr-x   - hdfs  trainers          0 2020-09-23 07:12 /common_folder
drwxr-xr-x   - hdfs  supergroup        0 2021-10-24 10:33 /deletedusers
drwxr-xr-x   - hbase hbase             0 2021-08-25 11:44 /hbase
drwxr-xr-x   - hdfs  supergroup        0 2020-06-12 06:10 /test
drwxr-xr-x   - hdfs  supergroup        0 2021-06-02 14:03 /testj
drwxrwxrwt   - hdfs  supergroup        0 2021-10-31 20:20 /tmp
drwxrwxr-x   - hdfs  supergroup        0 2021-11-01 02:22 /user
[nasehasgmail@ip-10-0-41-79 ~]$ ;
-bash: syntax error near unexpected token `;'
[nasehasgmail@ip-10-0-41-79 ~]$ hadoop fs -ls /user/hive/warehouse/hivedb_naseha.db/STOCK_* /
ls: '/user/hive/warehouse/hivedb_naseha.db/STOCK_*': No such file or directory
Found 7 items
drwxrwxr-x   - hdfs  trainers          0 2020-09-23 07:12 /common_folder
drwxr-xr-x   - hdfs  supergroup        0 2021-10-24 10:33 /deletedusers
drwxr-xr-x   - hbase hbase             0 2021-08-25 11:44 /hbase
drwxr-xr-x   - hdfs  supergroup        0 2020-06-12 06:10 /test
drwxr-xr-x   - hdfs  supergroup        0 2021-06-02 14:03 /testj
drwxrwxrwt   - hdfs  supergroup        0 2021-10-31 20:20 /tmp
drwxrwxr-x   - hdfs  supergroup        0 2021-11-01 02:22 /user
[nasehasgmail@ip-10-0-41-79 ~]$
```

HIVE

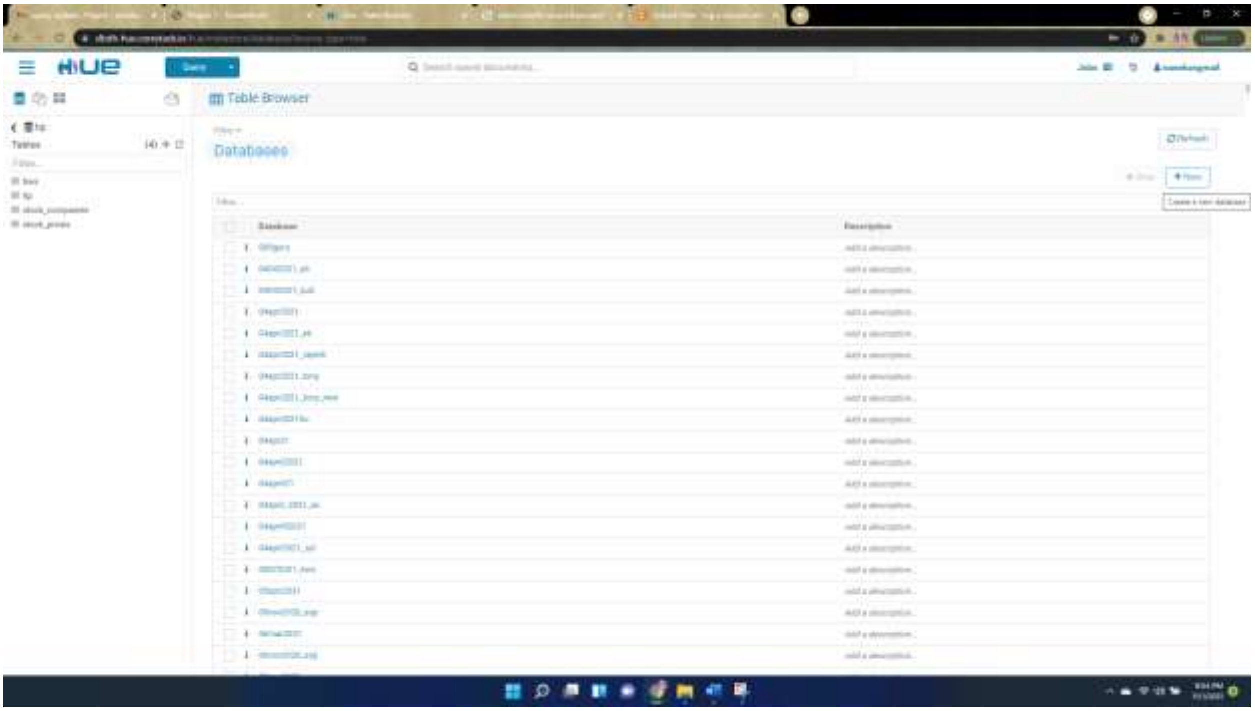


```
ROWS;
ERROR 2 (HY000): File 'C:\ProgramData\MySQL\MySQL Server 8.0\Uploads\StockPrices.csv' not found (Errcode: 2)
MySQL [naushongmail]> LOAD DATA LOCAL INFILE 'C:\\ProgramData\\MySQL\\MySQL Server 8.0\\Uploads\\StockPrices.csv'  INTO TABLE STOCK_PRICES FIELDS TERMINATED BY ',' OPTIONALLY ENCLOSED BY '"' LINES TERMINATED BY '\r\n' IGNORE 1
ROWS; hadoop fs -ls
ERROR 2 (HY000): File 'C:\ProgramData\MySQL\MySQL Server 8.0\Uploads\StockPrices.csv' not found (Errcode: 2)
    -> hadoop fs -ls;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near 'hadoop fs -ls
hadoop fs -ls' at line 1
MySQL [naushongmail]> hadoop fs -ls hadoop fs -ls /user/hive/warehouse/
    -> ;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near 'hadoop fs -ls hadoop fs -ls /user/hive/warehouse/' at line 1
MySQL [naushongmail]> ;exit
ERROR: No query specified

    -> ;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near 'exit' at line 1
MySQL [naushongmail]> exit
Bye
[naushongmail@ip-10-0-41-79 ~]$ hadoop fs -ls /user/hive/warehouse/hivedb_nausha.db/STOCK_* /
;
ls: `/user/hive/warehouse/hivedb_nausha.db/STOCK_*': No such file or directory
Found 7 items
drwxrwxr-x   - hdfs  trainers            0 2020-09-22 07:12 /common_folder
drwxr-xr-x   - hdfs  supergroup          0 2021-10-24 10:33 /deletedusers
drwxr-xr-x   - hbase hbase               0 2021-08-25 11:44 /hbase
drwxr-xr-x   - hdfs  supergroup          0 2020-06-12 06:18 /test
drwxr-xr-x   - hdfs  supergroup          0 2021-06-02 14:03 /test1
drwxrwxrwt   - hdfs  supergroup          0 2021-10-31 20:20 /tmp
drwxrwxr-x   - hdfs  supergroup          0 2021-11-01 02:22 /user
[naushongmail@ip-10-0-41-79 ~]$ ;
-bash: syntax error near unexpected token `;'
[naushongmail@ip-10-0-41-79 ~]$ hadoop fs -ls /user/hive/warehouse/hivedb_nausha.db/STOCK_* /
ls: `/user/hive/warehouse/hivedb_nausha.db/STOCK_*': No such file or directory
Found 7 items
drwxrwxr-x   - hdfs  trainers            0 2020-09-22 07:12 /common_folder
drwxr-xr-x   - hdfs  supergroup          0 2021-10-24 10:33 /deletedusers
drwxr-xr-x   - hbase hbase               0 2021-08-25 11:44 /hbase
drwxr-xr-x   - hdfs  supergroup          0 2020-06-12 06:18 /test
drwxr-xr-x   - hdfs  supergroup          0 2021-06-02 14:03 /test1
drwxrwxrwt   - hdfs  supergroup          0 2021-10-31 20:20 /tmp
drwxrwxr-x   - hdfs  supergroup          0 2021-11-01 02:22 /user
[naushongmail@ip-10-0-41-79 ~]$ hive
WARNING: Use "yarn jar" to launch YARN applications.
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/log4j-slf4j-impl-2.8.2.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/slf4j-log4j12-1.7.25.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.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/opt/cloudera/parcels/CDH-6.3.2-1.cdh6.3.2.p0.1605554/jars/hive-common-2.1.1-cdh6.3.2.jar!/hive-log4j2.properties Async: false

WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive>
```

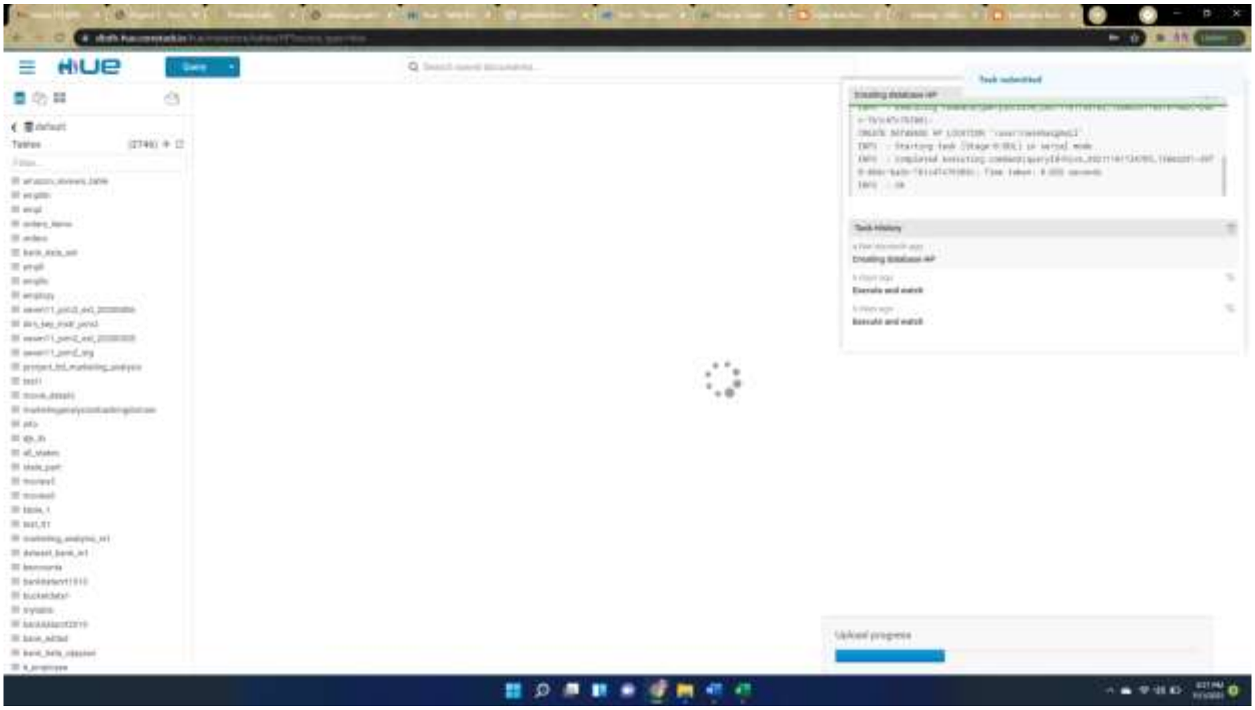
Used Hue Interface to do the rest of the questions

Creating a Database HP

File Section -> Tables -> Create Database



Database	Description
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...
[illegible]	Add a description...



Task submitted

Creating database HP

```
[illegible]
CREATE DATABASE HP LOCATION '/user/nasehasgmail'
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=[illegible]); Time taken: [illegible] seconds
INFO  : OK
```

Task History

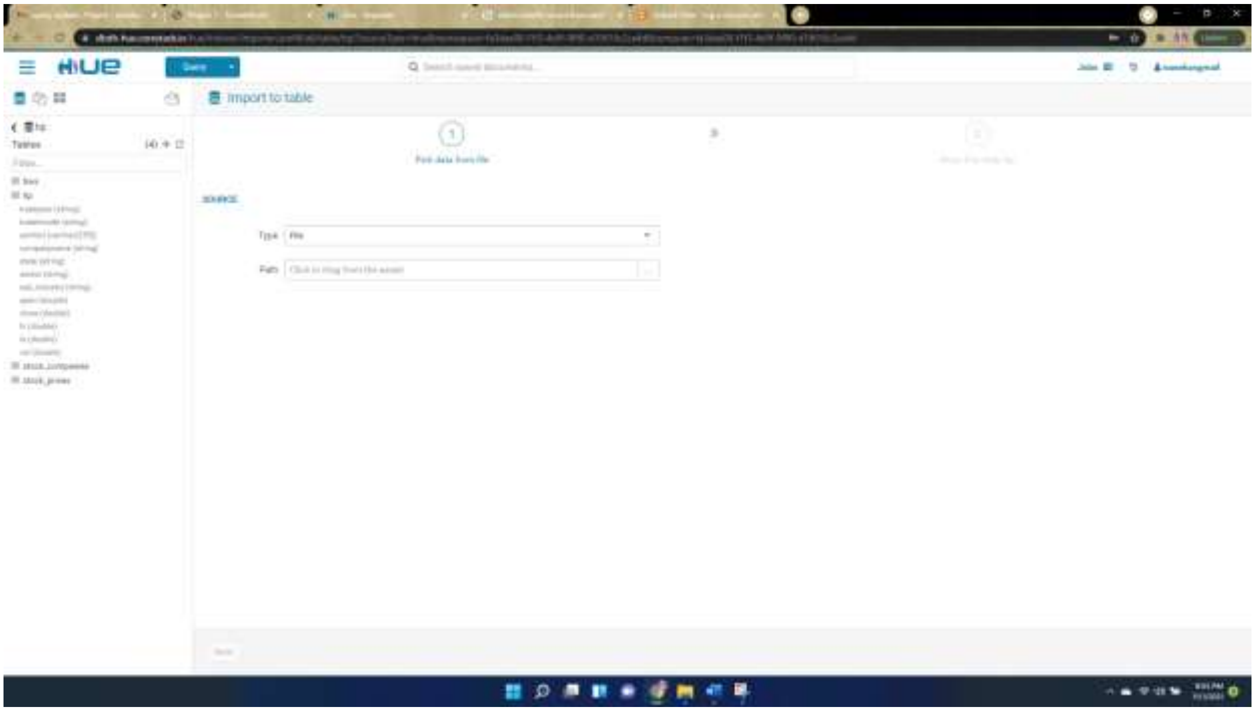
Creating database HP

Execute and watch

Execute and watch

Upload progress

Click on New -> Name the Database



Import to table

Pick data from file

SOURCE

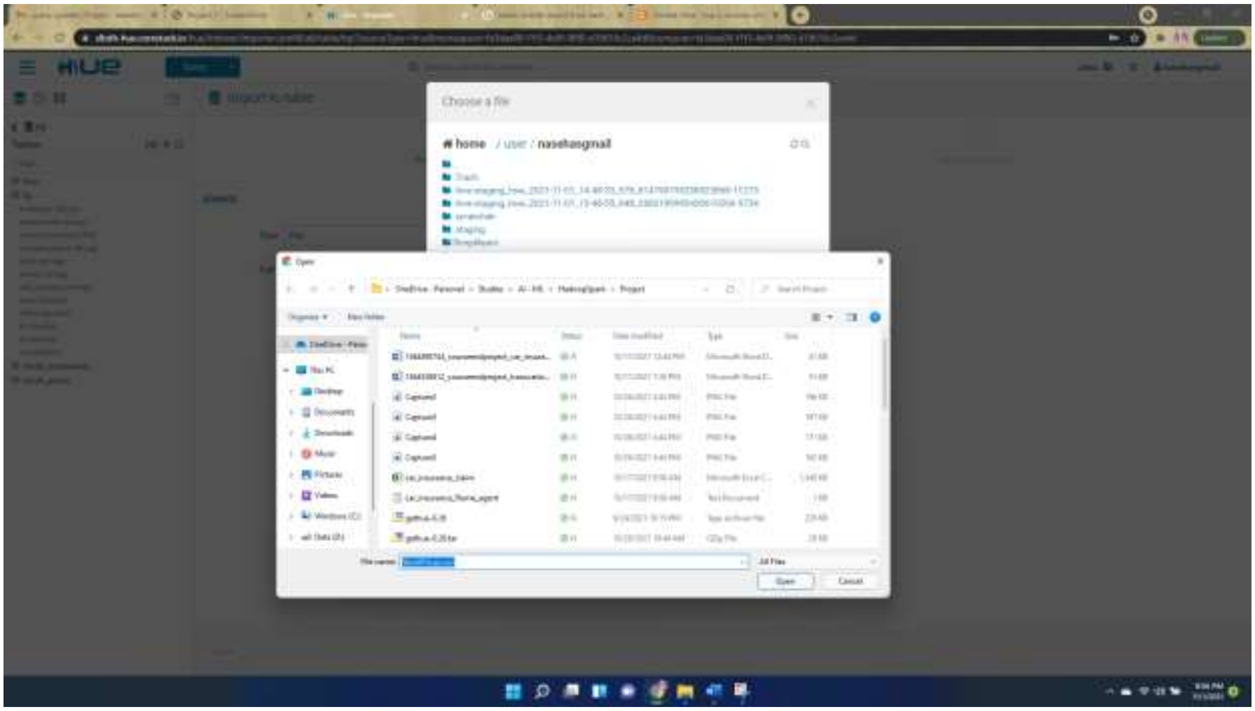
Type: File

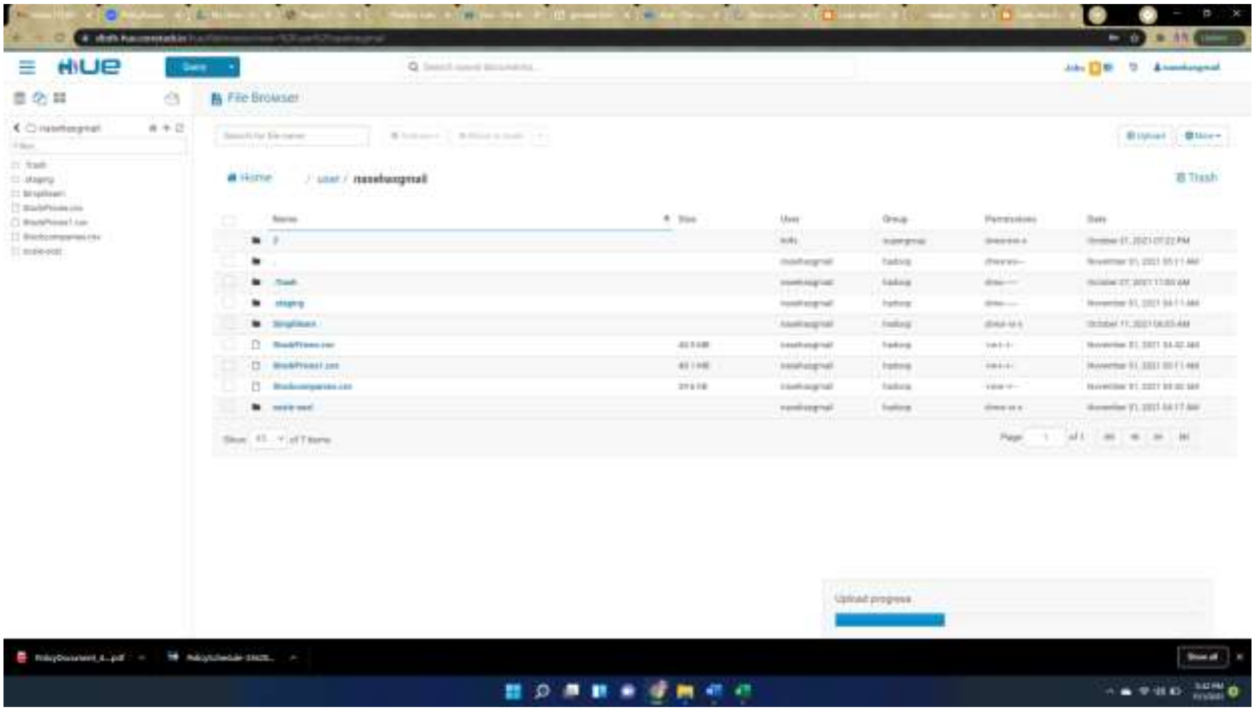
Path: Click to drag from the assist

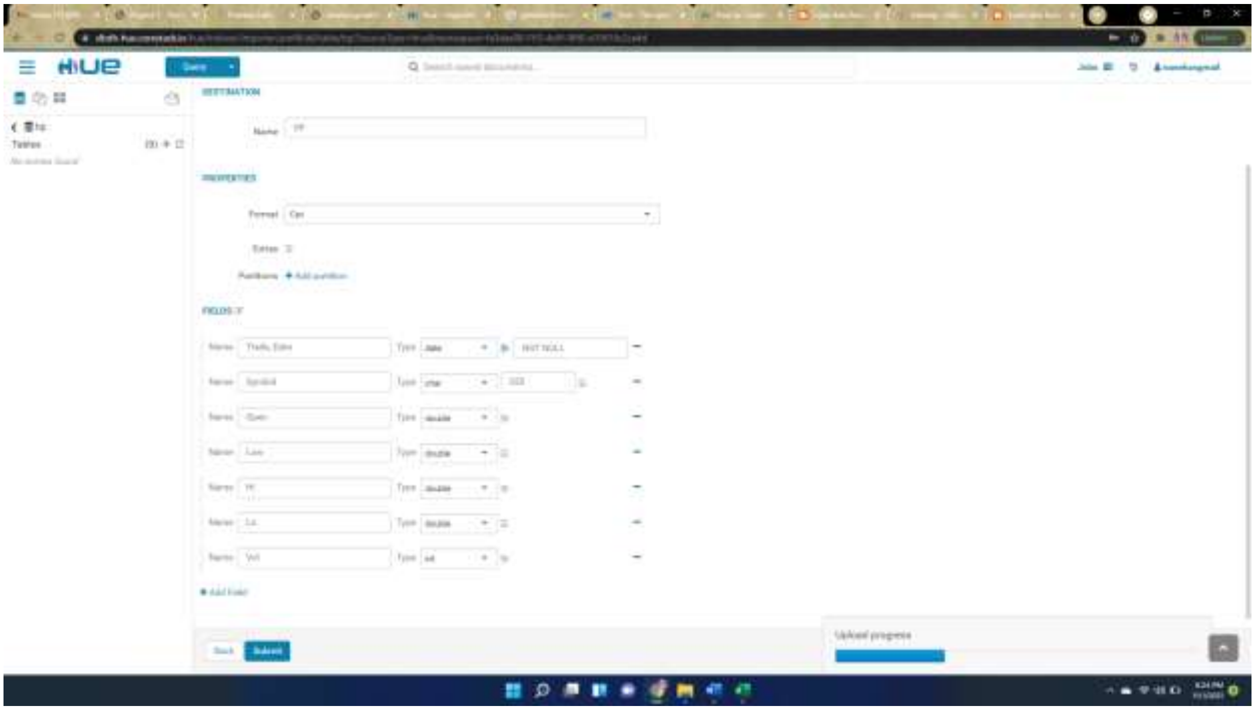
Here a database is created under my user id: Home/user/nasehasgmail

Then we upload the raw files (.csv) into the database

Upload both the .csv







DESTINATION

Name: FF

PROPERTIES

Format: Csv

Extras ☐

Partitions + Add partition

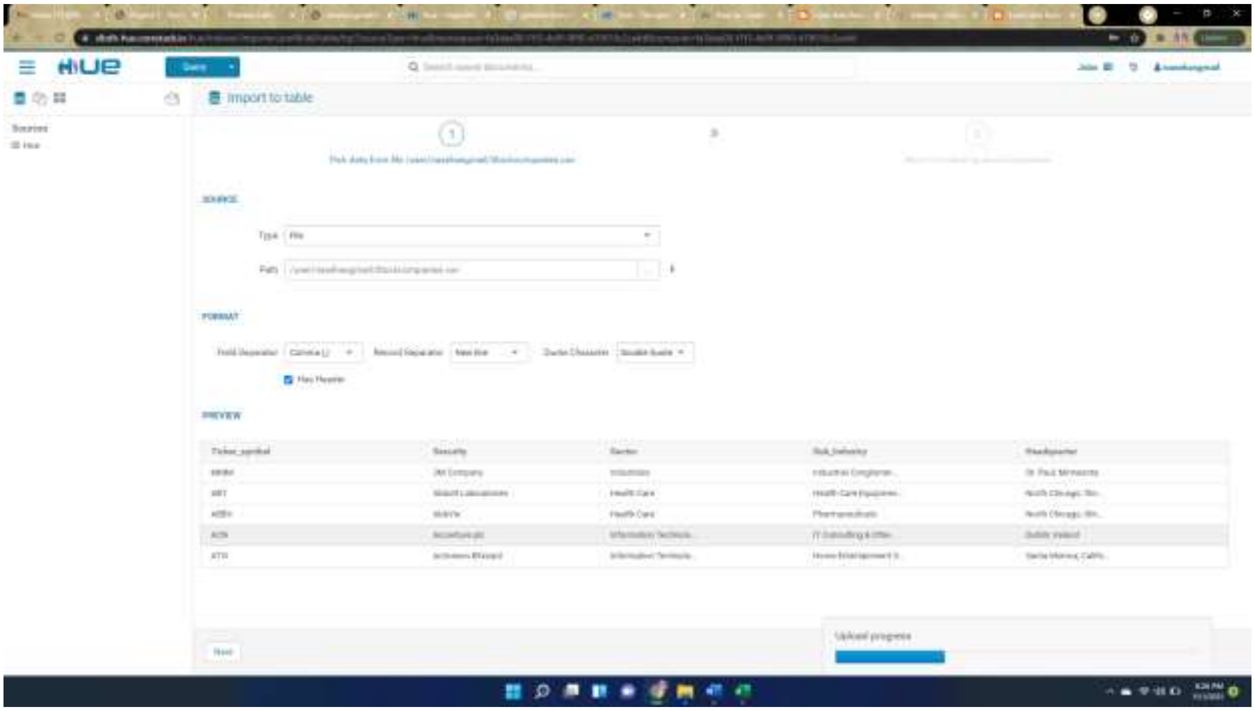
FIELDS 7

Name: Trade_Date	Type: date	☐	NOT NULL
Name: Symbol	Type: char	333	☐
Name: Open	Type: double	☐	
Name: Low	Type: double	☐	
Name: Hi	Type: double	☐	
Name: Lo	Type: double	☐	
Name: Vol	Type: int	☐	

+ Add Field

Back | Submit

Upload progress



Import to table

1 Pick data from file /user/.../Stockcompanies.csv

SOURCE

Type: File

Path: /user/.../Stockcompanies.csv

FORMAT

Field Separator: Comma (,) | Record Separator: New line | Quote Character: Double Quote

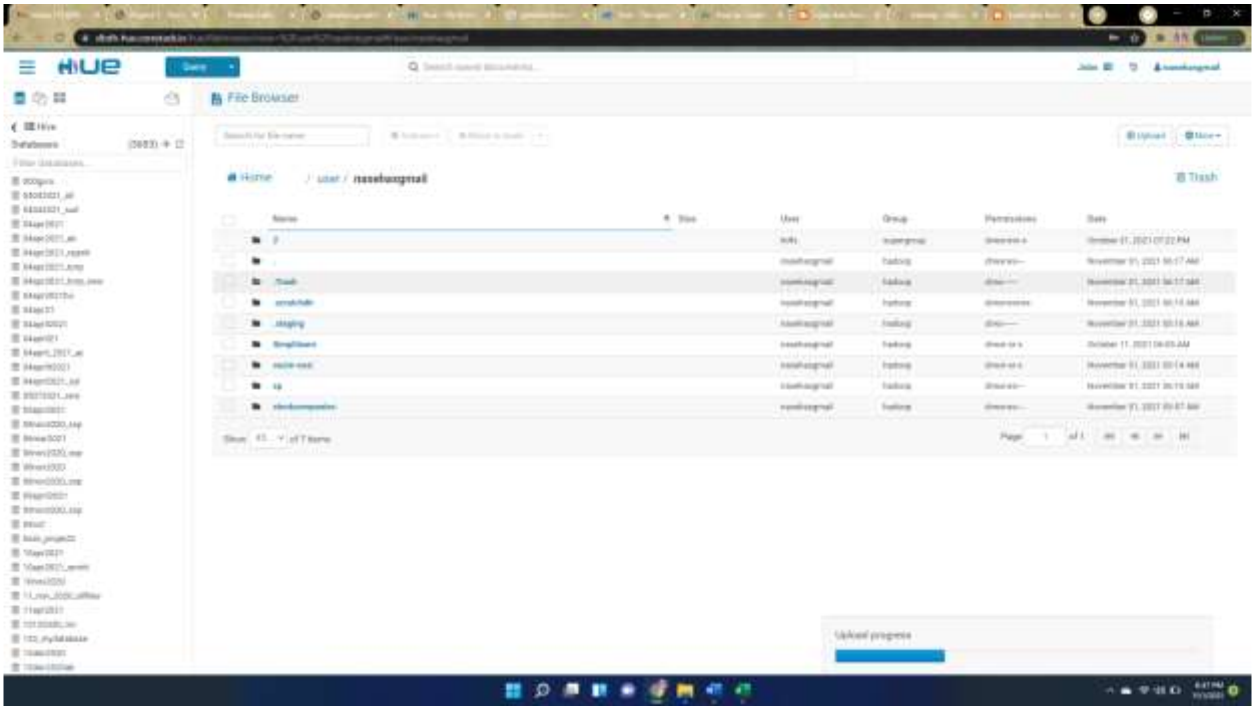
☑ Has Header

PREVIEW

Ticker_symbol	Security	Sector	Sub_Industry	Headquarter
MMM	3M Company	Industrials	Industrial Conglomer...	St. Paul, Minnesota
ABT	Abbott Laboratories	Health Care	Health Care Equipmen...	North Chicago, Illi...
ABBV	AbbVie	Health Care	Pharmaceuticals	North Chicago, Illi...
ACN	Accenture plc	Information Technolo...	IT Consulting & Other...	Dublin, Ireland
ATVI	Activision Blizzard	Information Technolo...	Home Entertainment S...	Santa Monica, Califo...

Next

Upload progress

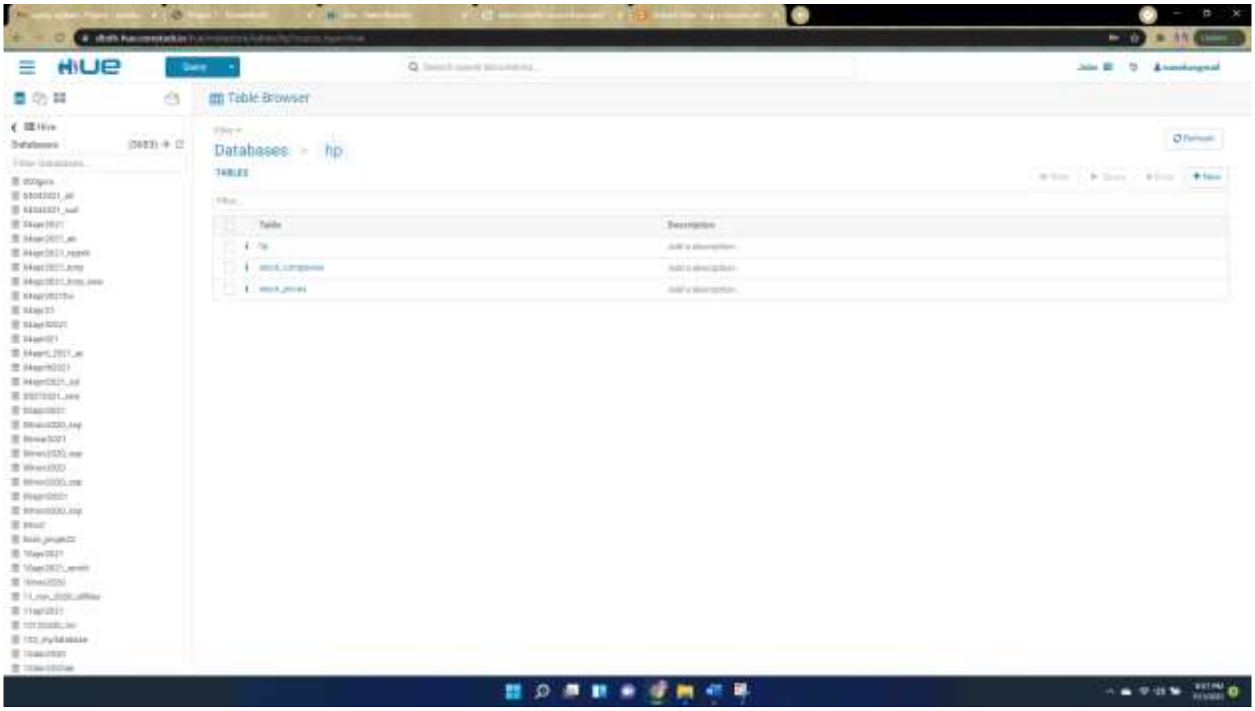


Home / user / nasebaxgmail

Name	Size	User	Group	Permissions	Date
..		hdfs	supergroup	drwxrwxrwx	October 31, 2021 07:22 PM
.		nasebaxgmail	hadoop	drwxr-xr--	November 01, 2021 06:17 AM
.Trash		nasebaxgmail	hadoop	drwx------	November 01, 2021 06:17 AM
[illegible]		nasebaxgmail	hadoop	[illegible]	November 01, 2021 [illegible] AM
.staging		nasebaxgmail	hadoop	drwx------	November 01, 2021 [illegible] AM
[illegible]		nasebaxgmail	hadoop	drwxr-xr-x	October 11, 2021 [illegible] AM
[illegible]		nasebaxgmail	hadoop	drwxr-xr-x	November 01, 2021 [illegible] AM
hp		nasebaxgmail	hadoop	drwxr-xr--	November 01, 2021 [illegible] AM
stockcompanies		nasebaxgmail	hadoop	drwxr-xr--	November 01, 2021 [illegible] AM

Show 15 of 7 items Page 1 of 1

Upload progress



Databases > hp

TABLES

Table	Description
hp	Add a description...
stock_companies	Add a description...
stock_prices	Add a description...

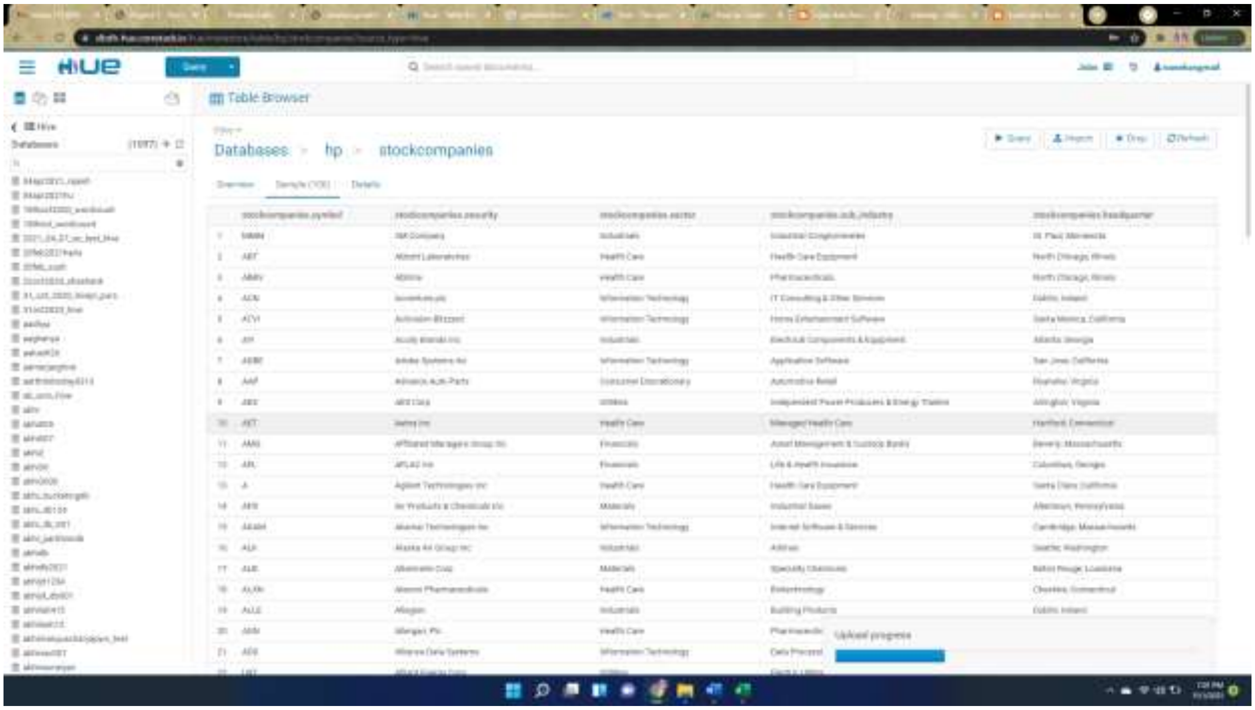


Table Browser

Databases > hp > stockcompanies

Overview | Sample (100) | Details

	stockcompanies.symbol	stockcompanies.security	stockcompanies.sector	stockcompanies.sub_industry	stockcompanies.headquarter
1	MMM	3M Company	Industrials	Industrial Conglomerates	St. Paul, Minnesota
2	ABT	Abbott Laboratories	Health Care	Health Care Equipment	North Chicago, Illinois
3	ABBV	AbbVie	Health Care	Pharmaceuticals	North Chicago, Illinois
4	ACN	Accenture plc	Information Technology	IT Consulting & Other Services	Dublin, Ireland
5	ATVI	Activision Blizzard	Information Technology	Home Entertainment Software	Santa Monica, California
6	AYI	Acuity Brands Inc	Industrials	Electrical Components & Equipment	Atlanta, Georgia
7	ADBE	Adobe Systems Inc	Information Technology	Application Software	San Jose, California
8	AAP	Advance Auto Parts	Consumer Discretionary	Automotive Retail	Roanoke, Virginia
9	AES	AES Corp	Utilities	Independent Power Producers & Energy Traders	Arlington, Virginia
10	AET	Aetna Inc	Health Care	Managed Health Care	Hartford, Connecticut
11	AMG	Affiliated Managers Group Inc	Financials	Asset Management & Custody Banks	Beverly, Massachusetts
12	AFL	AFLAC Inc	Financials	Life & Health Insurance	Columbus, Georgia
13	A	Agilent Technologies Inc	Health Care	Health Care Equipment	Santa Clara, California
14	APD	Air Products & Chemicals Inc	Materials	Industrial Gases	Allentown, Pennsylvania
15	AKAM	Akamai Technologies Inc	Information Technology	Internet Software & Services	Cambridge, Massachusetts
16	ALK	Alaska Air Group Inc	Industrials	Airlines	Seattle, Washington
17	ALB	Albemarle Corp	Materials	Specialty Chemicals	Baton Rouge, Louisiana
18	ALXN	Alexion Pharmaceuticals	Health Care	Biotechnology	Cheshire, Connecticut
19	ALLE	Allegion	Industrials	Building Products	Dublin, Ireland
20	AGN	Allergan, Plc	Health Care	Pharmaceutic	
21	ADS	Alliance Data Systems	Information Technology	Data Processi	

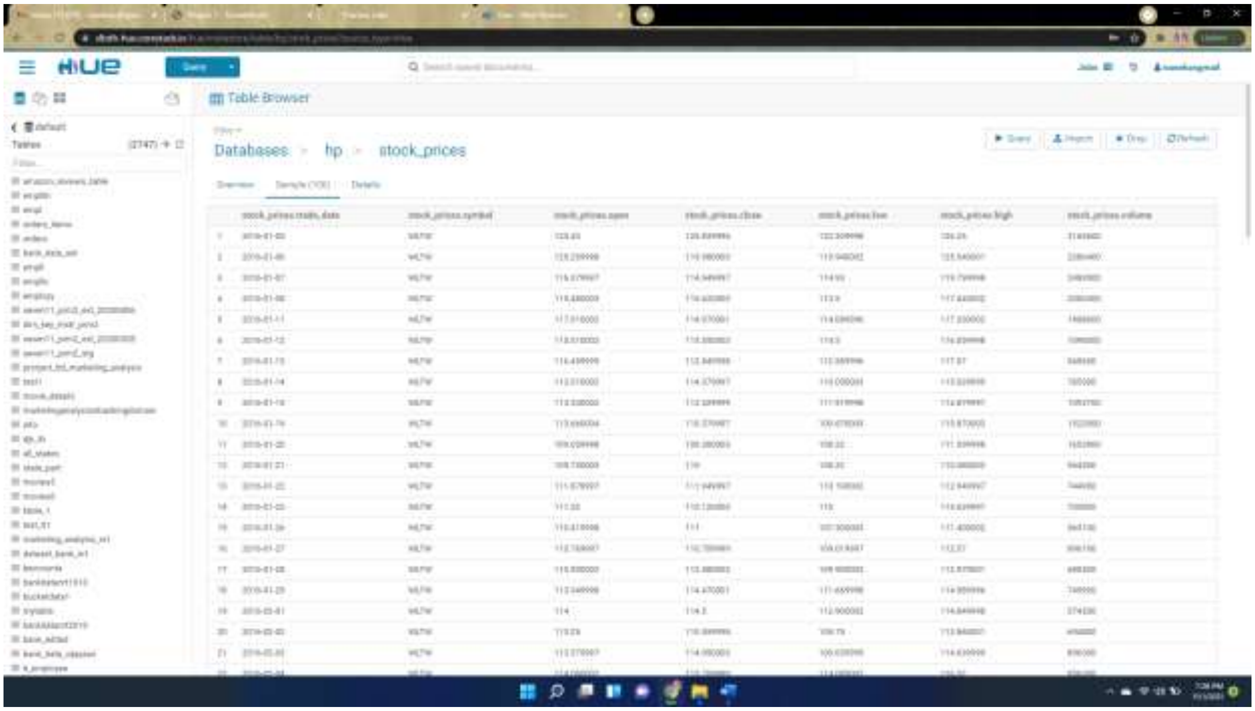
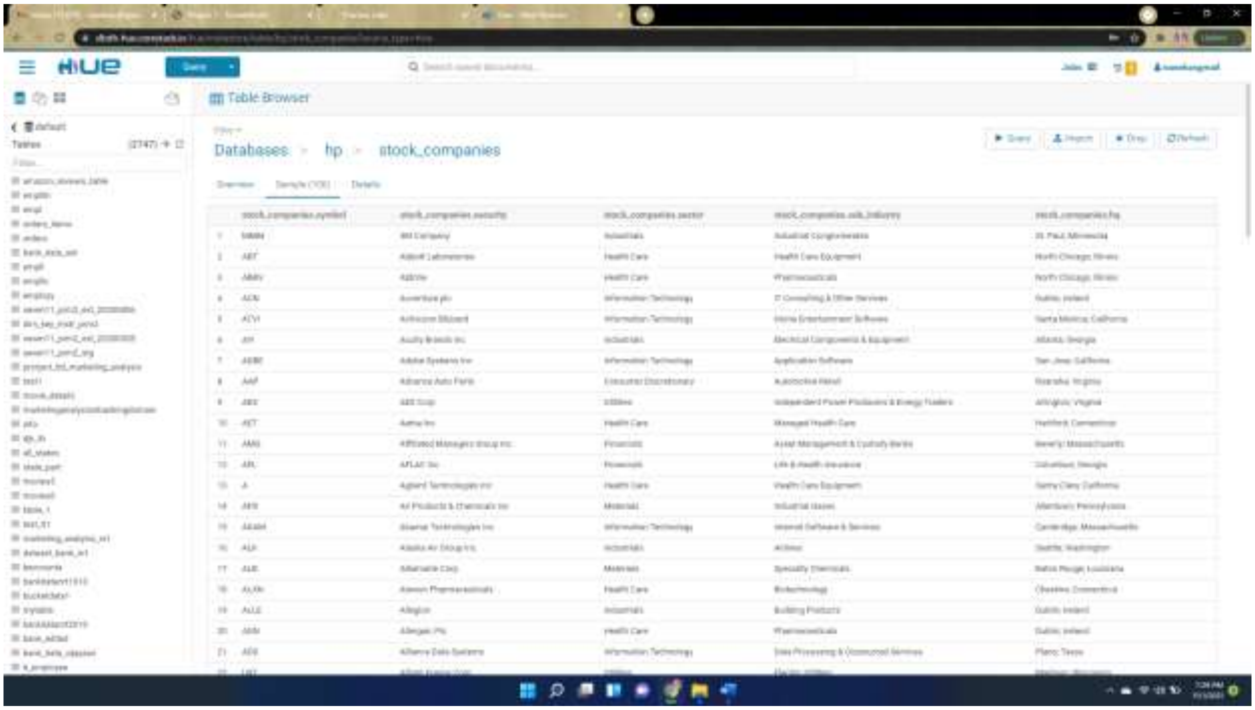


Table Browser

Databases > hp > stock_prices

Overview | Sample (100) | Details

	stock_prices.trade_date	stock_prices.symbol	stock_prices.open	stock_prices.close	stock_prices.low	stock_prices.high	stock_prices.volume
1	2016-01-05	WLTW	123.43	125.839996	122.309998	126.25	2163600
2	2016-01-06	WLTW	125.239998	119.980003	119.940002	125.540001	2386400
3	2016-01-07	WLTW	116.379997	114.949997	114.93	119.739998	2489500
4	2016-01-08	WLTW	115.480003	116.620003	113.5	117.440002	2006300
5	2016-01-11	WLTW	117.010002	114.970001	114.089996	117.330002	1408600
6	2016-01-12	WLTW	115.510002	115.550003	114.5	116.059998	1098000
7	2016-01-13	WLTW	116.459999	112.849998	112.589996	117.07	949600
8	2016-01-14	WLTW	113.510002	114.379997	110.050003	115.029999	785300
9	2016-01-15	WLTW	113.330002	112.529999	111.919998	114.879997	1093700
10	2016-01-19	WLTW	113.660004	110.379997	109.870003	115.870003	1523500
11	2016-01-20	WLTW	109.059998	109.300003	108.32	111.599998	1653900
12	2016-01-21	WLTW	109.730003	110	108.32	110.580002	944300
13	2016-01-22	WLTW	111.879997	111.949997	110.190002	112.949997	744900
14	2016-01-25	WLTW	111.32	110.120003	110	114.629997	703800
15	2016-01-26	WLTW	110.419998	111	107.300003	111.400002	563100
16	2016-01-27	WLTW	110.769997	110.709999	109.019997	112.57	896100
17	2016-01-28	WLTW	110.900002	113.089996	109.900002	112.970001	680400
18	2016-01-29	WLTW	113.349998	114.470001	111.669998	114.589996	749900
19	2016-02-01	WLTW	114	114.5	112.900002	114.849998	574200
20	2016-02-02	WLTW	113.25	110.559998	109.75	113.860001	694800
21	2016-02-03	WLTW	113.379997	114.050003	[illegible]	114.639999	896300



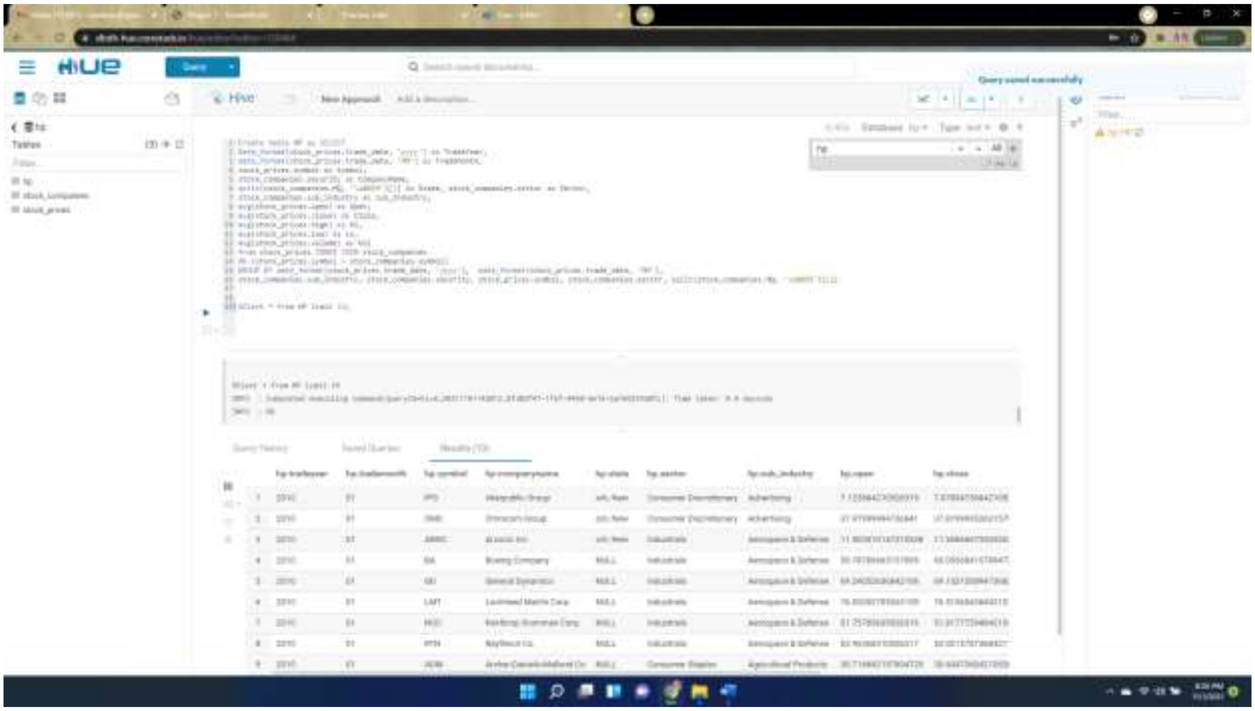
Databases > hp > stock_companies

	stock_companies.symbol	stock_companies.security	stock_companies.sector	stock_companies.sub_industry	stock_companies.hq
1	MMM	3M Company	Industrials	Industrial Conglomerates	St. Paul, Minnesota
2	ABT	Abbott Laboratories	Health Care	Health Care Equipment	North Chicago, Illinois
3	ABBV	AbbVie	Health Care	Pharmaceuticals	North Chicago, Illinois
4	ACN	Accenture plc	Information Technology	IT Consulting & Other Services	Dublin, Ireland
5	ATVI	Activision Blizzard	Information Technology	Home Entertainment Software	Santa Monica, California
6	AYI	Acuity Brands Inc	Industrials	Electrical Components & Equipment	Atlanta, Georgia
7	ADBE	Adobe Systems Inc	Information Technology	Application Software	San Jose, California
8	AAP	Advance Auto Parts	Consumer Discretionary	Automotive Retail	Roanoke, Virginia
9	AES	AES Corp	Utilities	Independent Power Producers & Energy Traders	Arlington, Virginia
10	AET	Aetna Inc	Health Care	Managed Health Care	Hartford, Connecticut
11	AMG	Affiliated Managers Group Inc	Financials	Asset Management & Custody Banks	Beverly, Massachusetts
12	AFL	AFLAC Inc	Financials	Life & Health Insurance	Columbus, Georgia
13	A	Agilent Technologies Inc	Health Care	Health Care Equipment	Santa Clara, California
14	APD	Air Products & Chemicals Inc	Materials	Industrial Gases	Allentown, Pennsylvania
15	AKAM	Akamai Technologies Inc	Information Technology	Internet Software & Services	Cambridge, Massachusetts
16	ALK	Alaska Air Group Inc	Industrials	Airlines	Seattle, Washington
17	ALB	Albemarle Corp	Materials	Specialty Chemicals	Baton Rouge, Louisiana
18	ALXN	Alexion Pharmaceuticals	Health Care	Biotechnology	Cheshire, Connecticut
19	ALLE	Allegion	Industrials	Building Products	Dublin, Ireland
20	AGN	Allergan, Plc	Health Care	Pharmaceuticals	Dublin, Ireland
21	ADS	Alliance Data Systems	Information Technology	Data Processing & Outsourced Services	Plano, Texas

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).

```
Create table HP as SELECT

Date_format(stock_prices.trade_date, 'yyyy') as TradeYear,

date_format(stock_prices.trade_date, 'MM') as TradeMonth,

stock_prices.symbol as Symbol,

stock_companies.security as CompanyName,

split(stock_companies.HQ, '\u0059')[1] as State, stock_companies.sector as Sector,

stock_companies.sub_industry as sub_Industry,

avg(stock_prices.open) as Open,

avg(stock_prices.close) as Close,

avg(stock_prices.high) as Hi,

avg(stock_prices.low) as Lo,

avg(stock_prices.volume) as Vol

from stock_prices INNER JOIN stock_companies

ON (stock_prices.symbol = stock_companies.symbol)

GROUP BY date_format(stock_prices.trade_date, 'yyyy'),  date_format(stock_prices.trade_date, 'MM'),
```

stock_companies.sub_Industry, stock_companies.security, stock_prices.symbol, stock_companies.sector, split(stock_companies.HQ, '\u0059')[1];



```
Create table HP as SELECT
date_format(stock_prices.trade_date, 'yyyy') as TradeYear,
date_format(stock_prices.trade_date, 'MM') as TradeMonth,
stock_prices.symbol as Symbol,
stock_companies.security as CompanyName,
split(stock_companies.HQ, '\u0059')[1] as State, stock_companies.sector as Sector,
stock_companies.sub_industry as Sub_Industry,
avg(stock_prices.open) as Open,
avg(stock_prices.close) as Close,
avg(stock_prices.high) as High,
avg(stock_prices.low) as Low,
avg(stock_prices.volume) as Vol
from stock_prices INNER JOIN stock_companies
ON (stock_prices.symbol = stock_companies.symbol)
GROUP BY date_format(stock_prices.trade_date, 'yyyy'), date_format(stock_prices.trade_date, 'MM'),
stock_companies.sub_industry, stock_companies.security, stock_prices.symbol, stock_companies.sector, split(stock_companies.HQ, '\u0059')[1];


Select * from HP limit 10;
```

```
Select * from HP limit 10
INFO  : Completed executing command(queryId=[illegible]); Time taken: [illegible] seconds
INFO  : OK
```

	hp.tradeyear	hp.trademonth	hp.symbol	hp.companyname	hp.state	hp.sector	hp.sub_industry	hp.open	hp.close
1	2010	01	IPG	Interpublic Group	[illegible]	Consumer Discretionary	Advertising	[illegible]	[illegible]
2	2010	01	OMC	Omnicom Group	[illegible]	Consumer Discretionary	Advertising	[illegible]	[illegible]
3	2010	01	ARNC	Arconic Inc	[illegible]	Industrials	Aerospace & Defense	[illegible]	[illegible]
4	2010	01	BA	Boeing Company	NULL	Industrials	Aerospace & Defense	[illegible]	[illegible]
5	2010	01	GD	General Dynamics	NULL	Industrials	Aerospace & Defense	[illegible]	[illegible]
6	2010	01	LMT	Lockheed Martin Corp.	NULL	Industrials	Aerospace & Defense	[illegible]	[illegible]
7	2010	01	NOC	Northrop Grumman Corp.	NULL	Industrials	Aerospace & Defense	[illegible]	[illegible]
8	2010	01	RTN	Raytheon Co.	NULL	Industrials	Aerospace & Defense	[illegible]	[illegible]
9	2010	01	ADM	Archer-Daniels-Midland Co	NULL	Consumer Staples	Agricultural Products	[illegible]	[illegible]

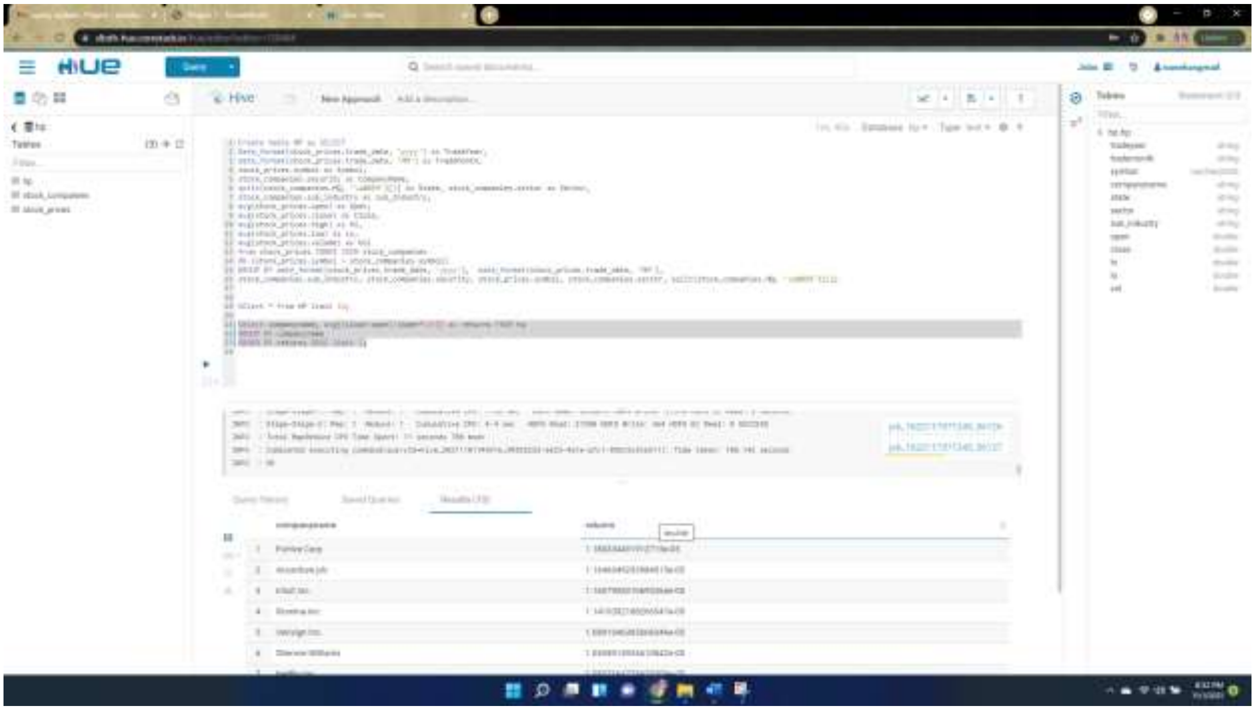
3) Find the top five companies that are good for investment

Select companyname, avg((close-open)/(open*100)) as returns FROM hp

GROUP BY companyname

ORDER BY returns DESC limit 5;

Describe formatted STOCK_PRICES;



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

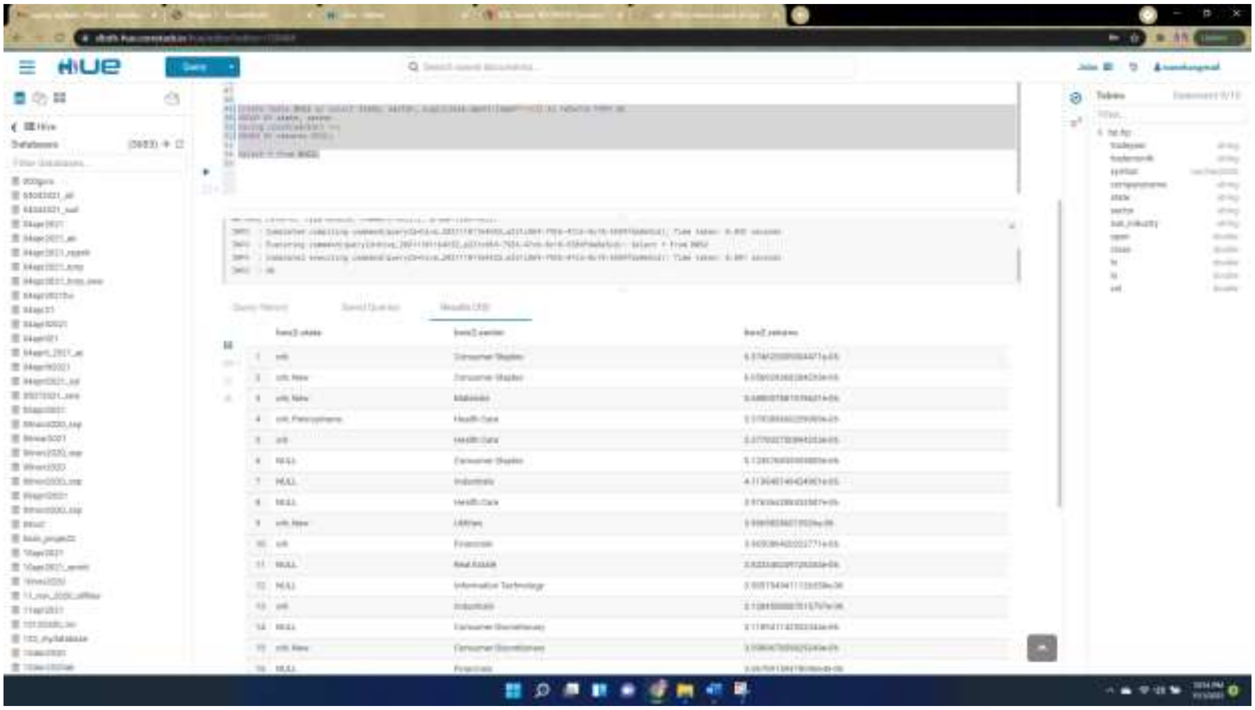
create table BWS2 as select state, sector, avg((close-open)/(open*100)) as returns FROM hp

GROUP BY state, sector

having count(sector) >=2

ORDER BY returns DESC;

Select * from BWS2;



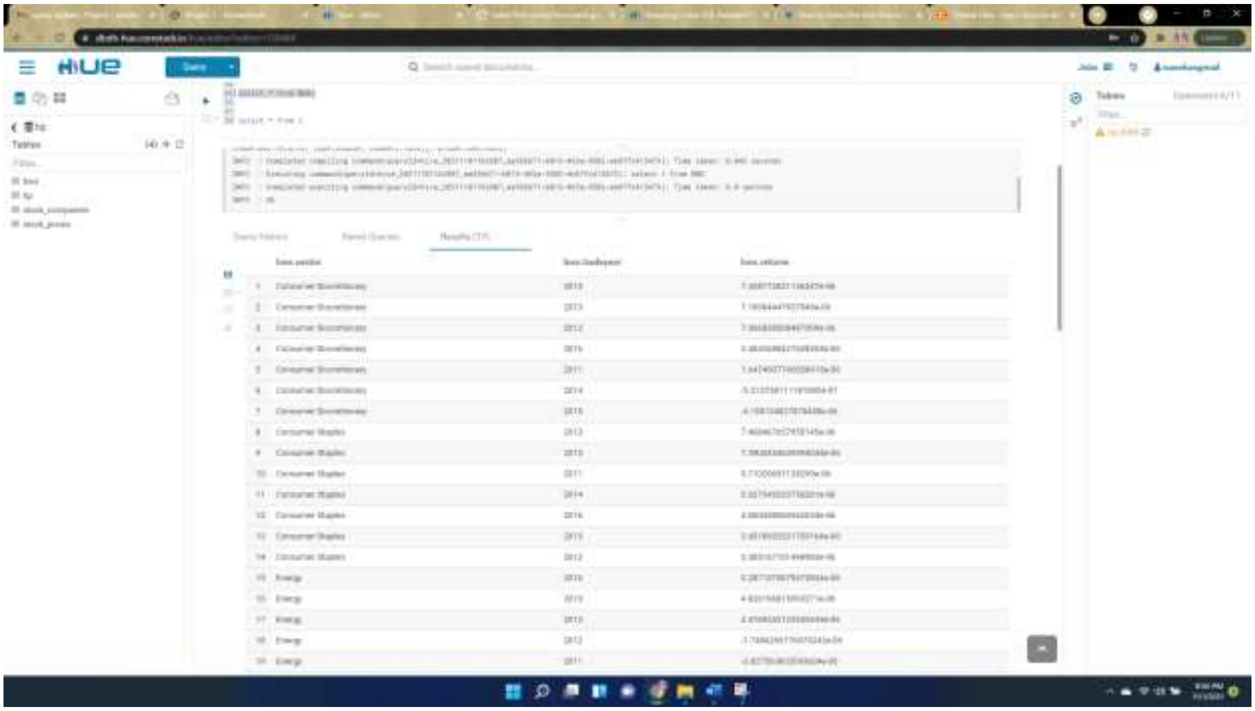
	[illegible]	[illegible]	[illegible]
1	[illegible]	Consumer Staples	[illegible]
2	[illegible]	Consumer Staples	[illegible]
3	[illegible]	Materials	[illegible]
4	[illegible]	Health Care	[illegible]
5	[illegible]	Health Care	[illegible]
6	NULL	Consumer Staples	[illegible]
7	NULL	Industrials	[illegible]
8	NULL	Health Care	[illegible]
9	[illegible]	Utilities	[illegible]
10	[illegible]	Financials	[illegible]
11	NULL	Real Estate	[illegible]
12	NULL	Information Technology	[illegible]
13	[illegible]	Industrials	[illegible]
14	NULL	Consumer Discretionary	[illegible]
15	[illegible]	Consumer Discretionary	[illegible]
16	NULL	Financials	[illegible]

5) For each sector find the following – worst, best and stable year;

To find the best and the worst years, I am creating another table with only sector, tradeyear and avg of the returns as columns.

```
create table BWS as select sector, tradeyear, avg((close-open)/(open*100)) as returns FROM hp

GROUP BY sector, tradeyear

ORDER BY sector, returns DESC;
```

select * from BWS;



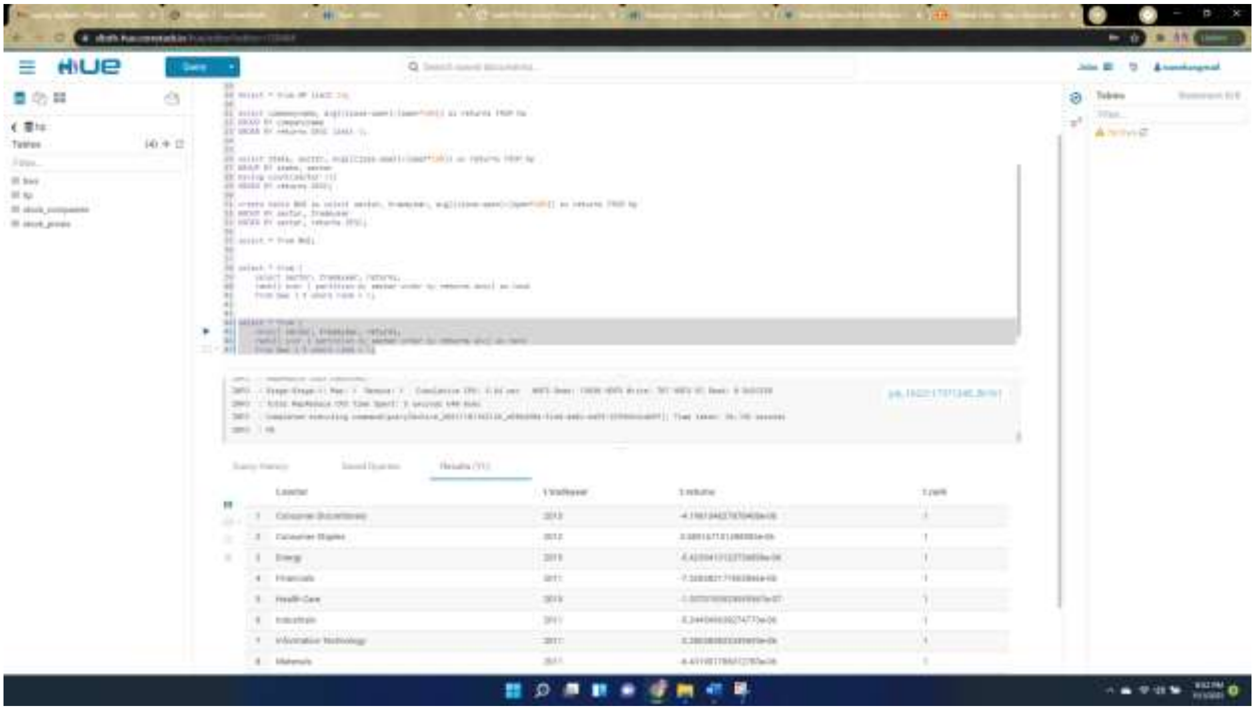
```
select * from BWS
```

	bws.sector	bws.tradeyear	bws.returns
1	Consumer Discretionary	2010	[illegible]
2	Consumer Discretionary	2013	[illegible]
3	Consumer Discretionary	2012	[illegible]
4	Consumer Discretionary	2015	[illegible]
5	Consumer Discretionary	2011	[illegible]
6	Consumer Discretionary	2014	[illegible]
7	Consumer Discretionary	2016	[illegible]
8	Consumer Staples	2013	[illegible]
9	Consumer Staples	2015	[illegible]
10	Consumer Staples	2011	[illegible]
11	Consumer Staples	2014	[illegible]
12	Consumer Staples	2016	[illegible]
13	Consumer Staples	2010	[illegible]
14	Consumer Staples	2012	[illegible]
15	Energy	2016	[illegible]
16	Energy	2010	[illegible]
17	Energy	2013	[illegible]
18	Energy	2012	[illegible]
19	Energy	2011	[illegible]

Once the table is ready, using the rank function to find the best and the worst year, and cume_dist() = .50 for the median. Median is assumed to be the most stable with min variance.

a. Worst Year

```
select * from (

        select sector, tradeyear, returns,

        rank() over ( partition by sector order by returns asc) as rank

        from bws ) t where rank = 1;
```



	t.sector	t.tradeyear	t.returns	t.rank
1	Consumer Discretionary	2013	-4.1[illegible]e-06	1
2	Consumer Staples	2012	[illegible]e-06	1
3	Energy	2015	-5.[illegible]e-06	1
4	Financials	2011	-7.[illegible]e-06	1
5	Health Care	2019	-1.[illegible]e-07	1
6	Industrials	2011	-5.[illegible]e-06	1
7	Information Technology	2011	-5.[illegible]e-06	1
8	Materials	2011	-6.[illegible]e-06	1

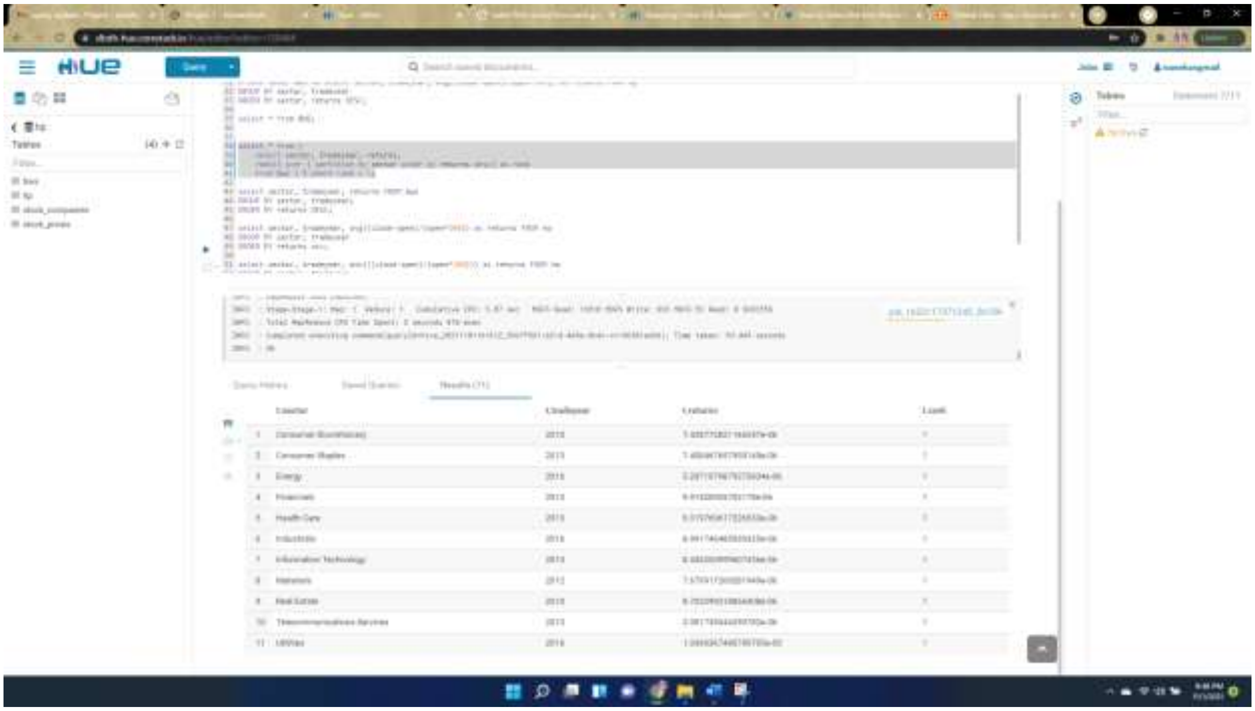
b. Best Year

select * from (

select sector, tradeyear, returns,

rank() over (partition by sector order by returns desc) as rank

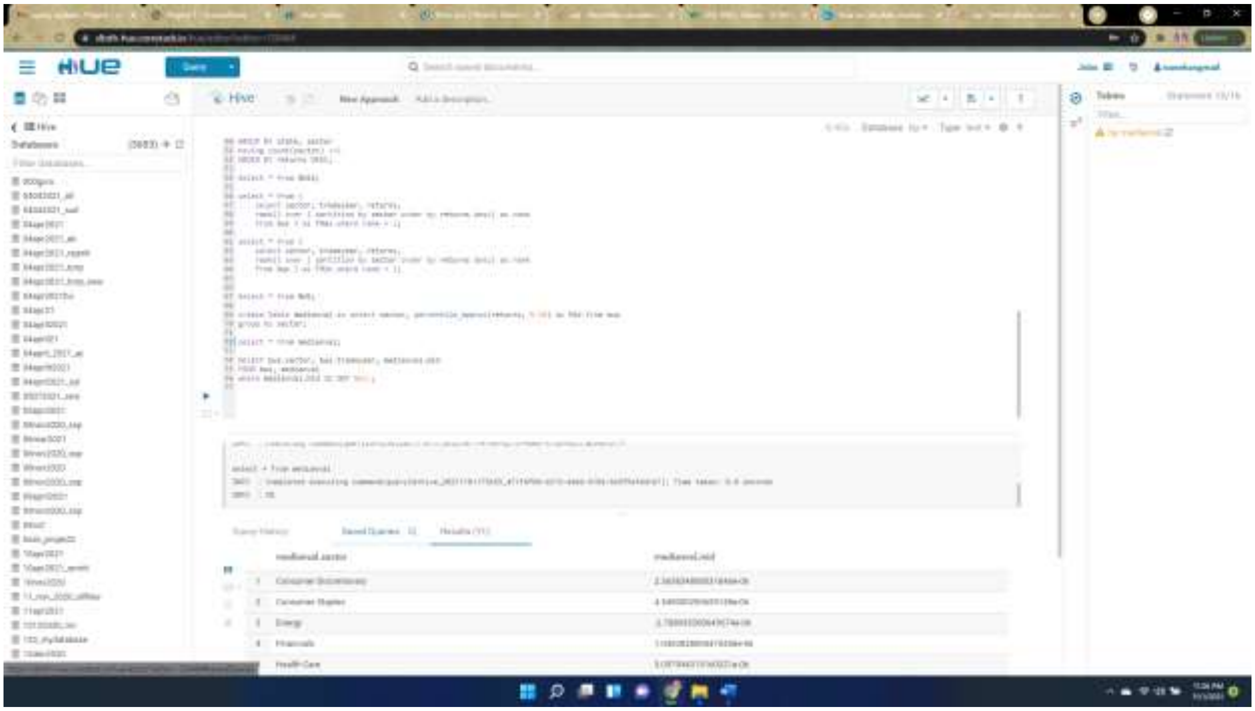
from bws) t where rank = 1;



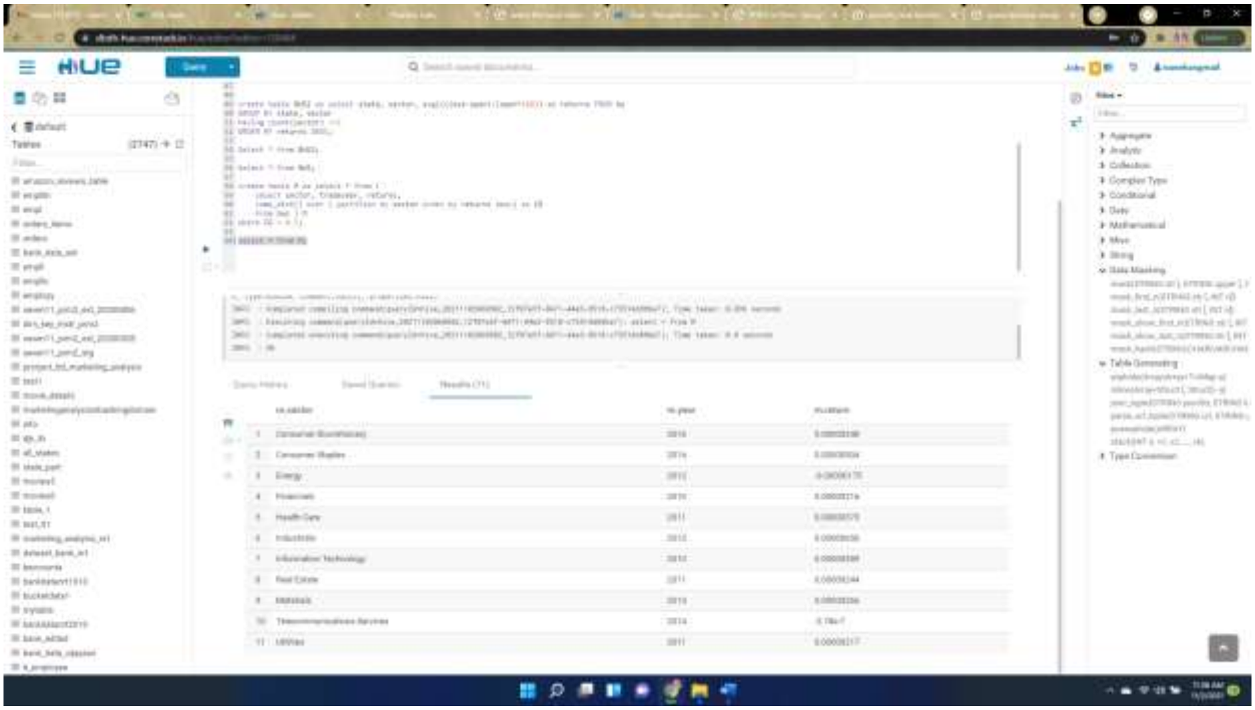
	t.sector	t.tradeyear	t.returns	t.rank
1	Consumer Discretionary	2013	1.[illegible]e-05	1
2	Consumer Staples	2013	1.[illegible]e-05	1
3	Energy	2016	5.[illegible]e-06	1
4	Financials	2013	9.[illegible]e-06	1
5	Health Care	2013	9.[illegible]e-06	1
6	Industrials	2016	8.[illegible]e-06	1
7	Information Technology	2013	8.[illegible]e-06	1
8	Materials	2012	7.[illegible]e-06	1
9	Real Estate	2016	8.[illegible]e-06	1
10	Telecommunications Services	2013	2.[illegible]e-06	1
11	Utilities	2016	1.[illegible]e-05	1

c. Stable year

```
create table M as select * from (

            select sector, tradeyear, returns,

            cume_dist() over ( partition by sector order by returns desc) as CD

            from bws ) M

where CD = 0.5;


select * from M;
```



	medianval.sector	medianval.[illegible]
1	Consumer Discretionary	[illegible]
2	Consumer Staples	[illegible]
3	Energy	[illegible]
4	Financials	[illegible]
	Health Care	[illegible]



	re.sector	re.year	re.return
1	Consumer Discretionary	2016	[illegible]
2	Consumer Staples	2016	[illegible]
3	Energy	2012	[illegible]
4	Financials	2010	[illegible]
5	Health Care	2011	[illegible]
6	Industrials	2013	[illegible]
7	Information Technology	2013	[illegible]
8	Real Estate	2011	[illegible]
9	Materials	2016	[illegible]
10	Telecommunications Services	2014	[illegible]
11	Utilities	2011	[illegible]

Code:

Create table HP as SELECT

Date_format(stock_prices.trade_date, 'yyyy') as TradeYear,

date_format(stock_prices.trade_date, 'MM') as TradeMonth,

stock_prices.symbol as Symbol,

stock_companies.security as CompanyName,

split(stock_companies.HQ, '\u0059')[1] as State, stock_companies.sector as Sector,

stock_companies.sub_industry as sub_Industry,

avg(stock_prices.open) as Open,

avg(stock_prices.close) as Close,

avg(stock_prices.high) as Hi,

avg(stock_prices.low) as Lo,

avg(stock_prices.volume) as Vol

from stock_prices INNER JOIN stock_companies

```
ON (stock_prices.symbol = stock_companies.symbol)
GROUP BY date_format(stock_prices.trade_date, 'yyyy'),  date_format(stock_prices.trade_date, 'MM'),
stock_companies.sub_Industry, stock_companies.security, stock_prices.symbol,
stock_companies.sector, split(stock_companies.HQ, '\u0059')[1];



SElect * from HP limit 10;


Select companyname, avg((close-open)/(open*100)) as returns FROM hp
GROUP BY companyname
ORDER BY returns DESC limit 5;


select state, sector, avg((close-open)/(open*100)) as returns FROM hp
GROUP BY state, sector
having count(sector >2)
ORDER BY returns DESC;


create table BWS as select sector, tradeyear, avg((close-open)/(open*100)) as returns FROM hp
GROUP BY sector, tradeyear
ORDER BY sector, returns DESC;


select * from BWS;


select * from (
        select sector, tradeyear, returns,
        rank() over ( partition by sector order by returns desc) as rank
        from bws ) t where rank = 1;



select * from (
        select sector, tradeyear, returns,
        rank() over ( partition by sector order by returns asc) as rank
        from bws ) t where rank = 1;



create table BWS2 as select state, sector, avg((close-open)/(open*100)) as returns FROM hp
GROUP BY state, sector
having count(sector) >=2
ORDER BY returns DESC;

Select * from BWS2;

Select * from BWS;

create table M as select * from (
        select sector, tradeyear, returns,
        cume_dist() over ( partition by sector order by returns desc) as CD
        from bws ) M
where CD = 0.5;

select * from M;
```