# Dataset link to download

https://www.dropbox.com/s/0jftfm548qepmsv/h1b\_kaggle.csv?dl=0

# Few steps automation done using bash shell:

hive table creation, sqoop import, mysql table creation, hadoop service startup, dataset download etc



# Intially loading the cleaned dataset to Mysql

# Create table in mysql :

CREATE TABLE h1b(

s\_no integer,

case\_status LONGTEXT,

employer\_name LONGTEXT,

soc\_name LONGTEXT,

job\_title LONGTEXT,

full\_time\_position LONGTEXT,

prevailing\_wage integer,

year LONGTEXT,

worksite LONGTEXT,

longitute Double,

latitute Double);

# Initially loading data in created table from local to mysql

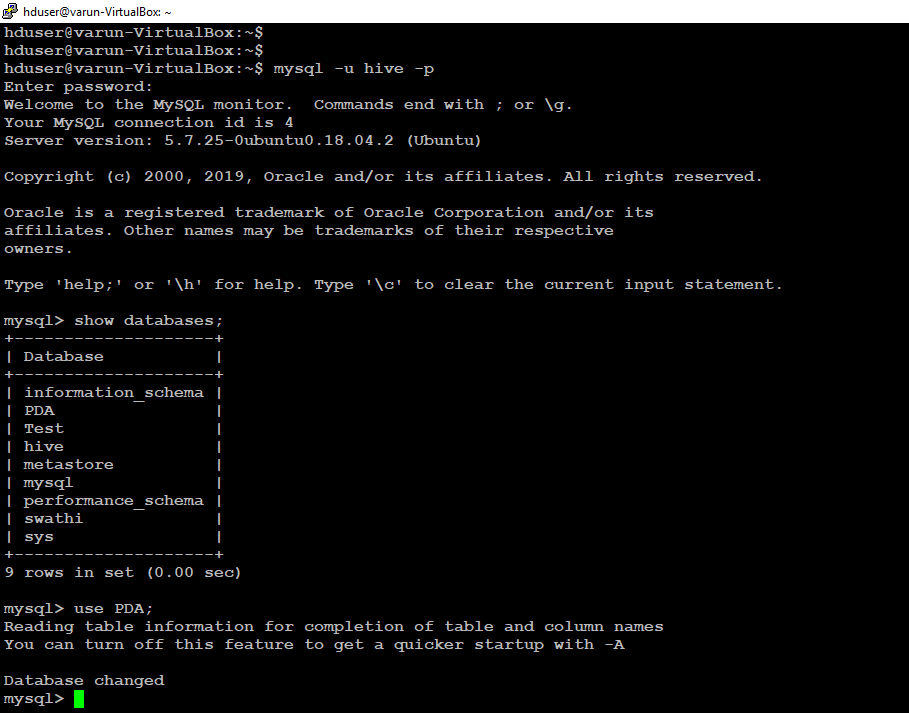
#mysql -u hive -p --local\_infile=1 PDA -e "LOAD DATA LOCAL INFILE '/home/hduser/h1b/dataset/h1b.csv' INTO TABLE h1b FIELDS TERMINATED BY ','"

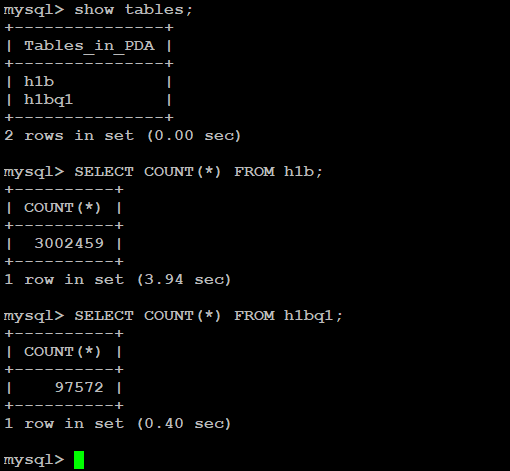
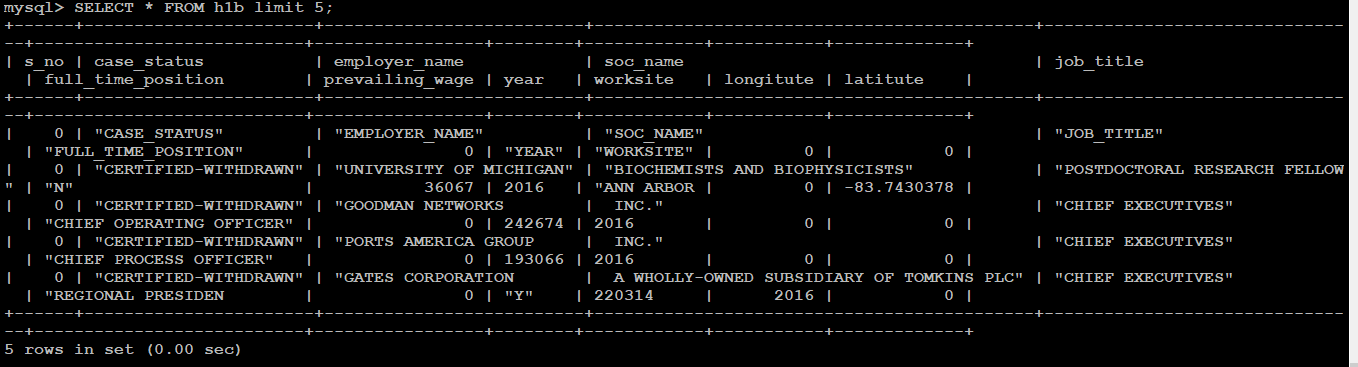
# Validating the loaded dataset

#SELECT COUNT(\*) FROM h1b;

#SELECT \* FROM h1b limit 5;

**Note :** Hive – one of the user in mysql database having full access to all databses.





# Data ingestion – Sqoop

Importing the data from mysql to hdfs using sqoop.

# sqoop import --connect jdbc:mysql://127.0.0.1/PDA --username hive --password admin --table h1b --target-dir **/h1b** -m 1

(note: mysql db - PDA , mysql tbl – H1b, destination dir in hdfs – h1b)

Changing the sqoop processed file into csv to load inside hive.

# hdfs dfs -cp /h1b/part-m-00000 /h1b/h1b.csv

# Creating table in hive

Hive:

# CREATE TABLE varun.h1b(s\_no int,case\_status string,employer\_name string, soc\_name string, job\_title string, full\_time\_position string,prevailing\_wage int,year string, worksite string, longitute double, latitute double )

ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'

WITH SERDEPROPERTIES (

"separatorChar" = ",",

"quoteChar" = "\""

)

stored as textfile

tblproperties ("skip.header.line.count"="1")

;

load data inpath '/h1b/h1b.csv' overwrite into table h1b;

# Hive Query – 1:

Top organization in United States who filed most number of h1b visa petition for the year 2016 & 2015.

***Note****: output of hive query redirected to local file system and shown top 5 employers for the visualization purpose.*

### ###

INSERT OVERWRITE LOCAL DIRECTORY '/tmp/hive\_output1' **#Output local path**

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE

select h1b.employer\_name,count(h1b.employer\_name) as no\_of\_applications,h1b.year from h1b where h1b.year = '2015' group by h1b.employer\_name,h1b.year order by no\_of\_applications desc limit 20;

### ###

INSERT OVERWRITE LOCAL DIRECTORY '/tmp/hive\_output2'

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE

select h1b.employer\_name,count(h1b.employer\_name) as no\_of\_applications,h1b.year from h1b where h1b.year = '2016' group by h1b.employer\_name,h1b.year order by no\_of\_applications desc limit 20;

### ###

# Hive Query – 2:

*Which is the most popular job in United States for the year 2016 and 2015?*

***Note****: output of hive query redirected to local file system and shown top 5 job for the visualization purpose.*

### ###

INSERT OVERWRITE LOCAL DIRECTORY '/tmp/hive\_output3'

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE

select h1b.job\_title,count(h1b.job\_title) as no\_of\_applications,h1b.year from h1b where h1b.year='2016' group by h1b.job\_title,h1b.year order by no\_of\_applications desc limit 10;

### ###

INSERT OVERWRITE LOCAL DIRECTORY '/tmp/hive\_output4'

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE

select h1b.job\_title,count(h1b.job\_title) as no\_of\_applications,h1b.year from h1b where h1b.year='2015' group by h1b.job\_title,h1b.year order by no\_of\_applications desc limit 10;

### ###

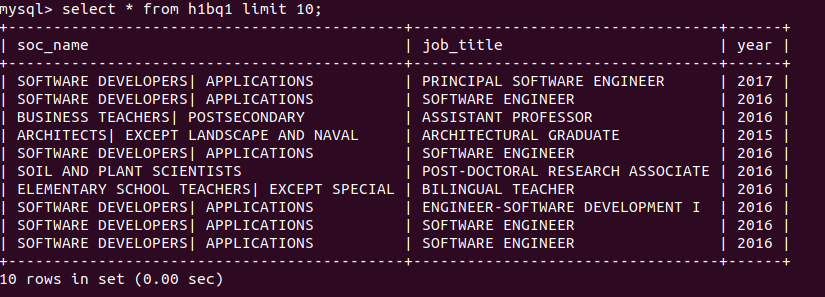
Hive outputs:



# MapReduce: Query - 1

Is the number of petitions increasing over year for data engineers?

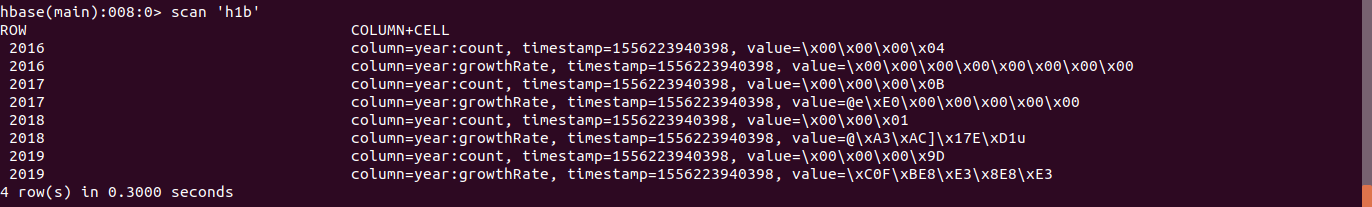
## Input: **Programmatically** accessing the source **Mysql**



## **MapReduce**:

Driver, mapper , reducer java files.



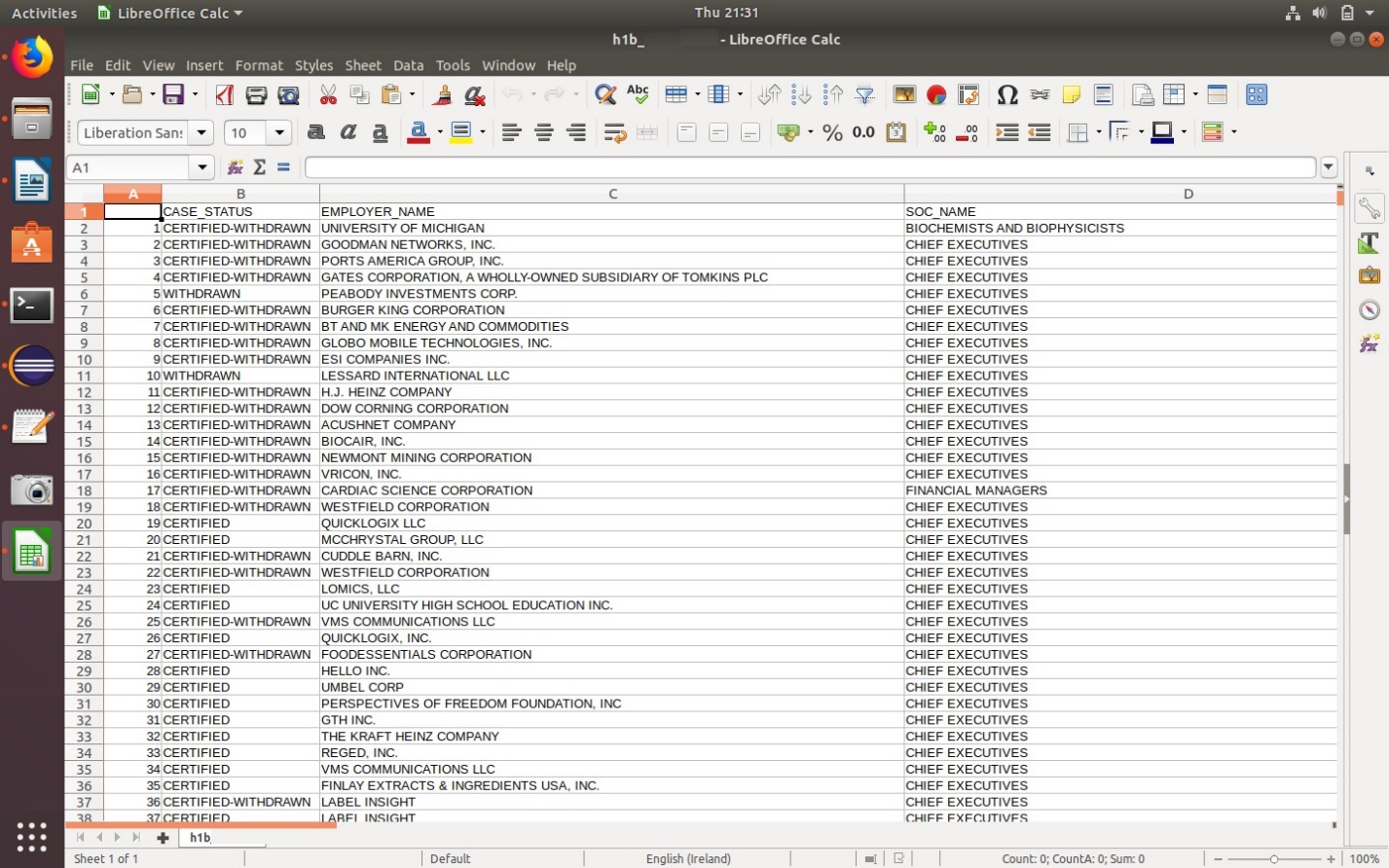
**Output: programmatically processed output to Hbase**



# MapReduce: Query – 2

Which Industry in the United States has most number of Data Scientist?

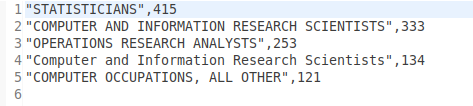
**Input: text file taken from Mysql**



## **MapReduce processing**:



**Output: eclipse screenshot**



****

**Design pattern used:** For this top 10 design pattern has been developed to process this in the MapReduce framework.

**Eclipse IDE :**

Since the Eclipse project files seems to have more data it has been attached in external drive**.**

[**https://www.dropbox.com/s/dc679cq2g80ocax/H1b.zip?dl=0**](https://www.dropbox.com/s/dc679cq2g80ocax/H1b.zip?dl=0)

All the MapReduce jobs can be validated by downloading the file which consist of actual / sample datasets as well.