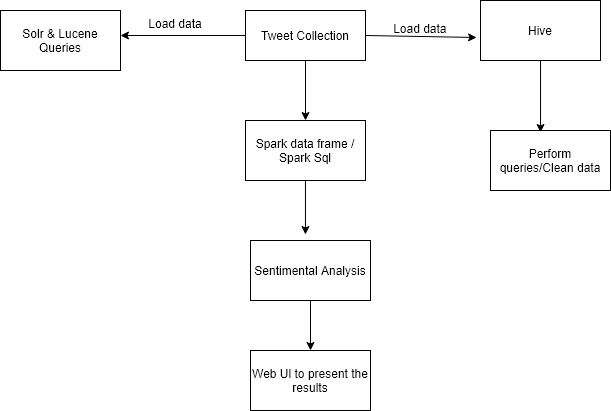
**Big Data Programming Project Increment 1**

**Dataset:**

We are collecting twitter data using Twitter batch API using twitter developer account credentials with Republicans and Democrats as keywords. All tweets are in CSV format. We collected about 3 million tweets which is about 450 Megabytes. It has various information about a tweet like user\_description, user\_location, coordinates, user\_name, user\_created, user\_followers(follower\_id, created), retweet\_count, polarity, subjectivity. We are collecting using batch API and able to download 4kb per second in which we are filtering for keywords like Democrat supporters, Republican supporters.

**Detail design of features:**



We have created a account in Twitter Developers API, then we got API tokens and credentials with which we downloaded the tweets using tweetpy and twitter streaming API in python and stored tweets into sqlite db.Imported the queries into csv file using python. Collected around 4 million tweets. Imported the csv file into HDFS. Visualized the csv data file in Hue. Processed the dataset using Hue by creating database& database table in hue and implemented various queries to get specific information on tweets. For next phases we are going to Implement solr& lucene on the data, create spark dataframes on CSV file and implement various transformations and actions on dataframes. We are also going to implement Spark sql, and query the data from dataframe. we will implement sentiment analysis on the data.

**Analysis of Increment-1:**

We have created a account in Twitter Developers API, then we got API tokens and credentials with which we downloaded the tweets using tweetpy and twitter streaming API in python and stored tweets into sqlite db.Imported the queries into csv file using python. Collected around 4 million tweets. Imported the csv file into HDFS. Visualized the csv data file in Hue. Processed the dataset using Hue by creating database& database table in hue and implemented various queries to get specific information on tweets.

**Implementation:**

**Collected tweets:**

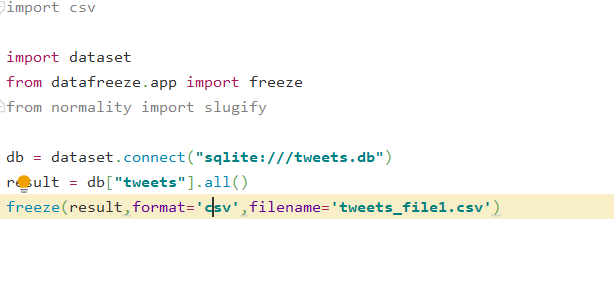
* Got developer access from twitter, then created an app to use their APIs.
* Using tweepy and twitter streaming API collected tweets in python and below is the code.



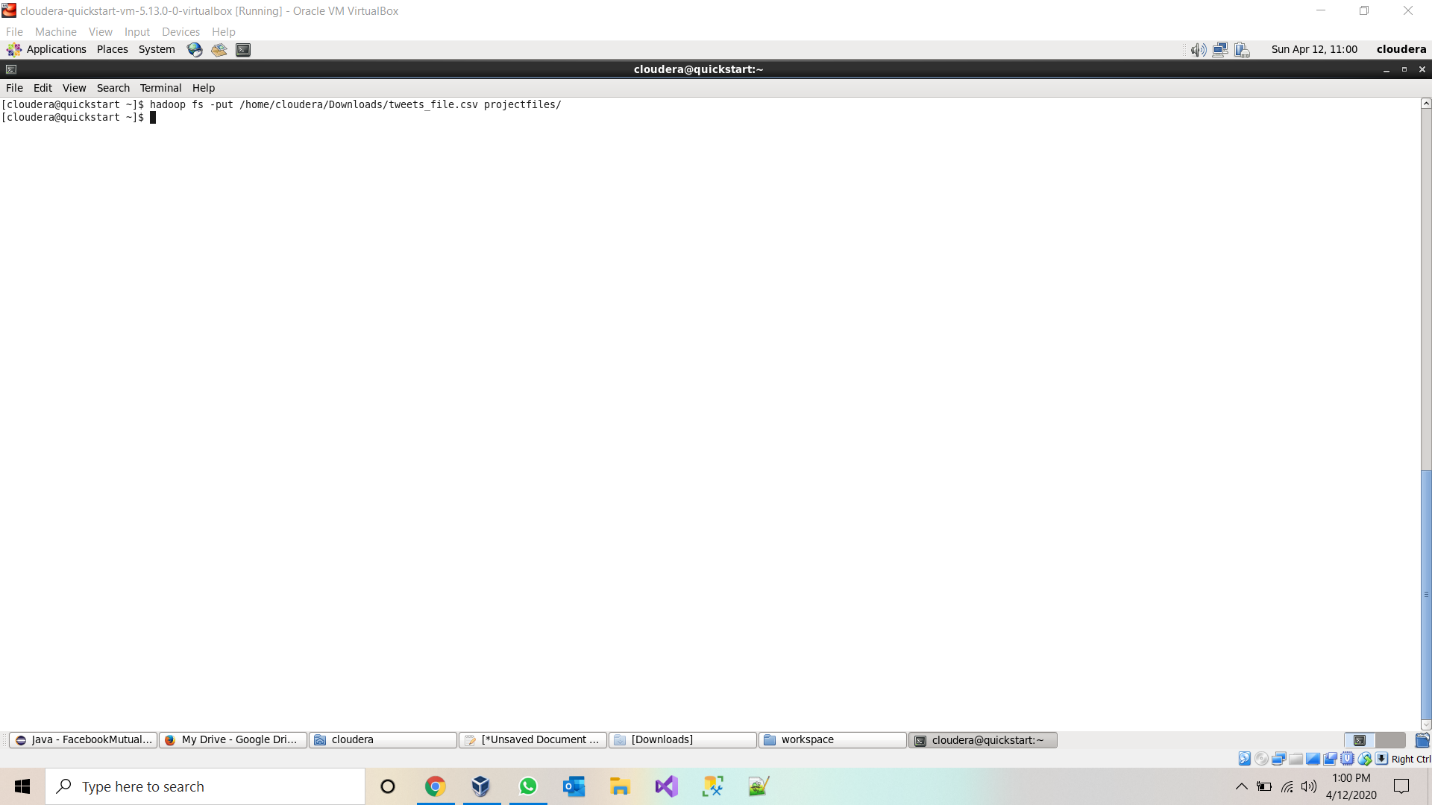
* Storing the tweets into sqlite db

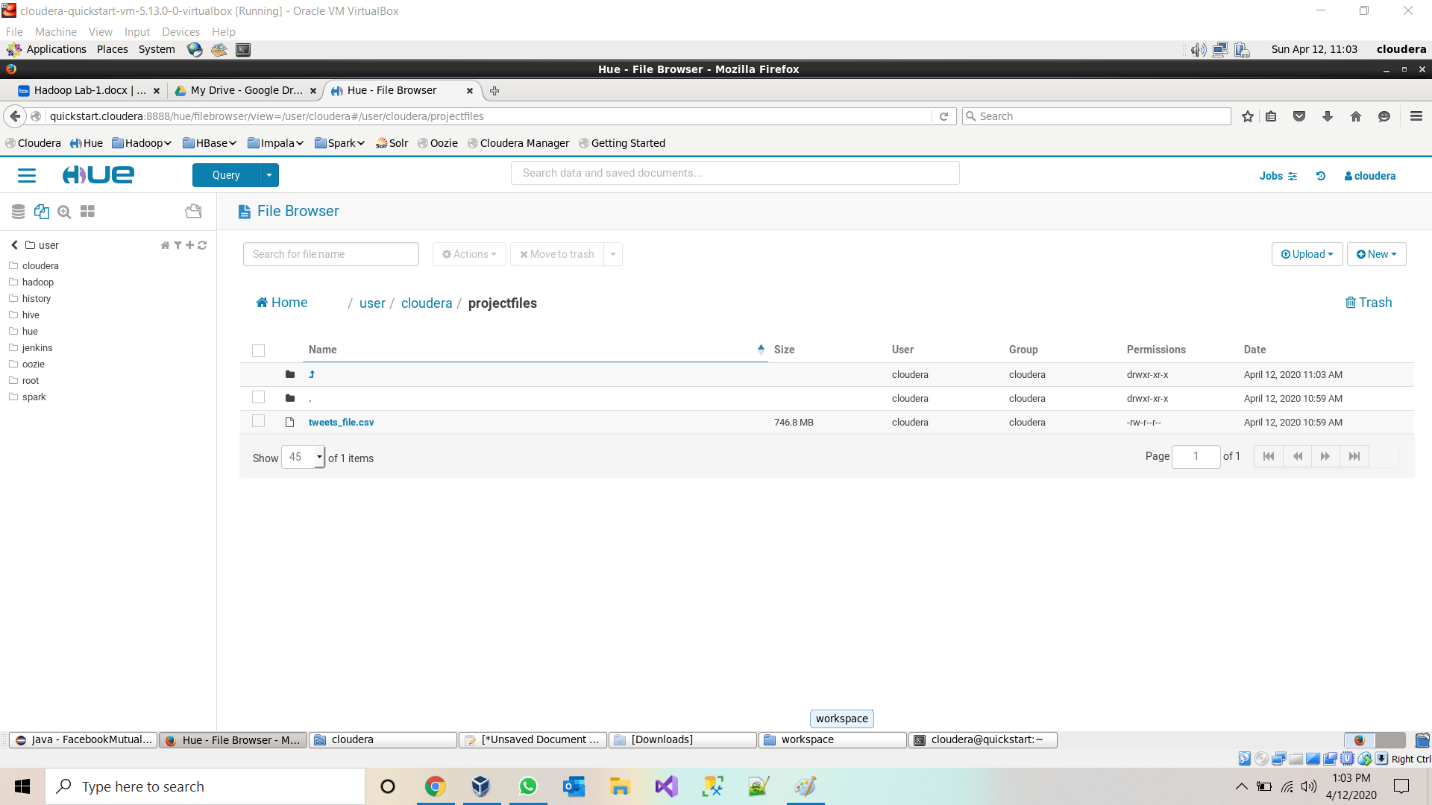


* After storing the tweets into sqlite db importing them into csv file using below code.

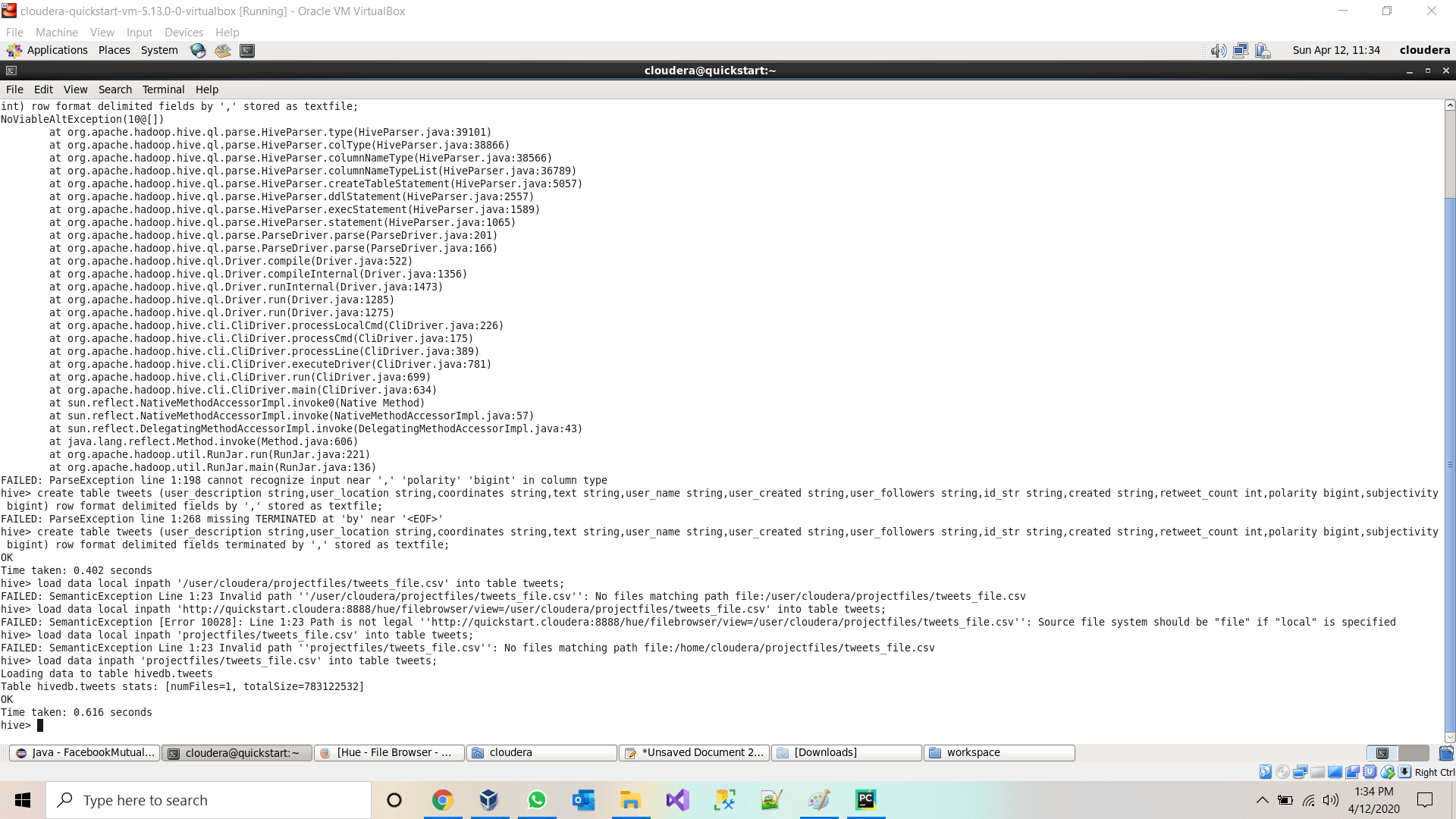


* Collected around 40 million tweets.
* Imported the file into hdfs.

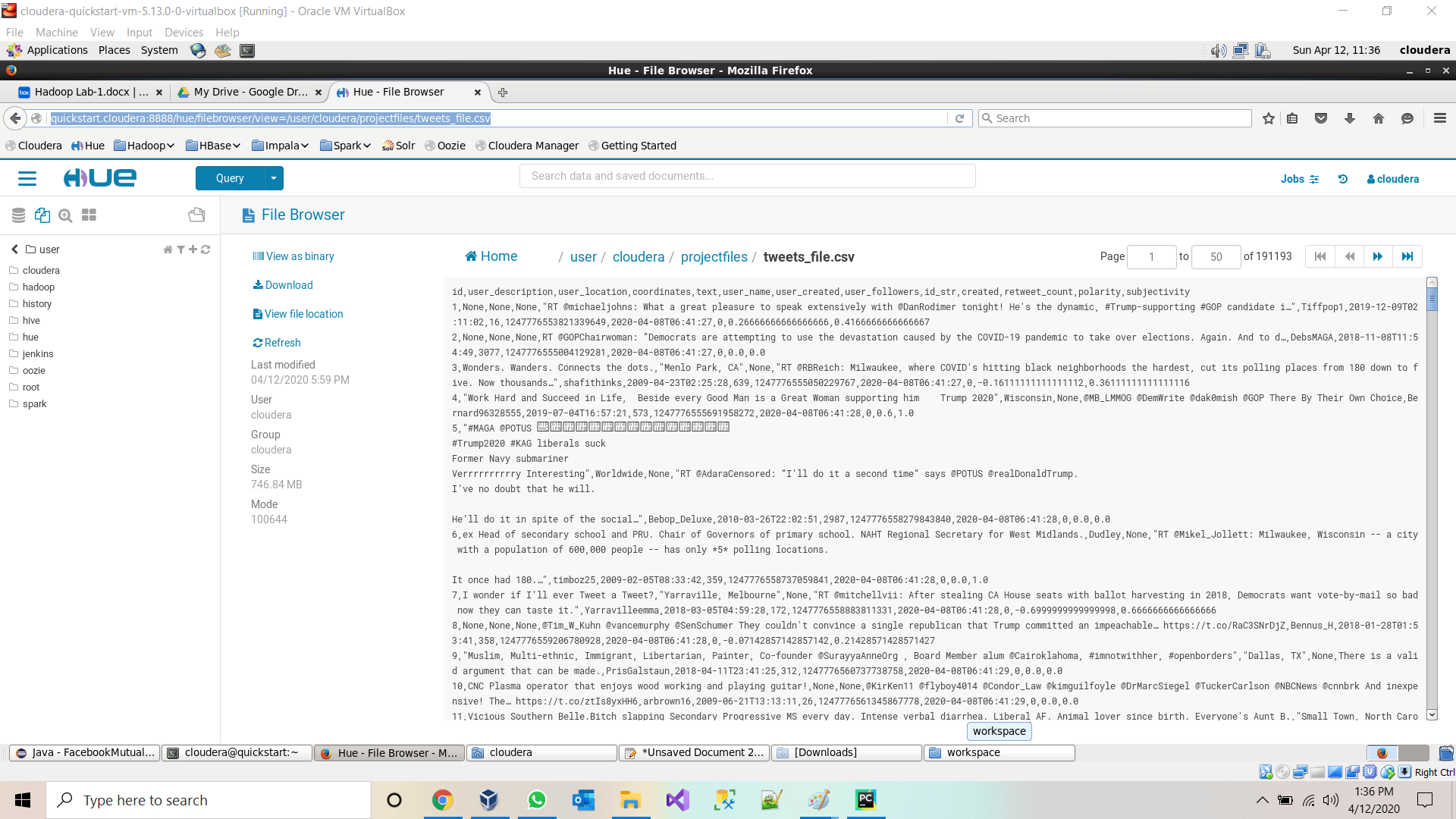




* Created a database and database table in hive and loaded the data into hive table.



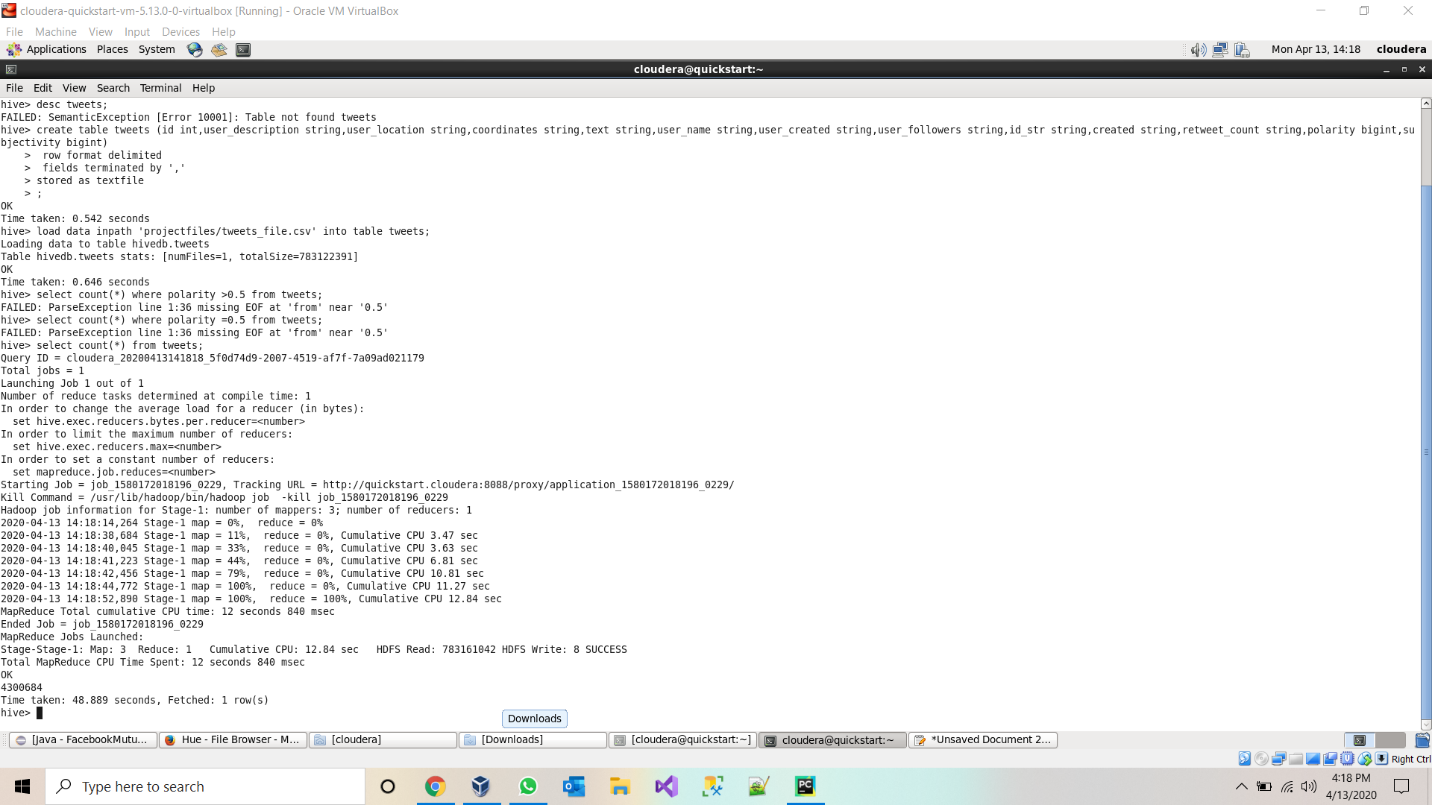
* Viewing data table in Hive.



* Querying data from Hive Table:

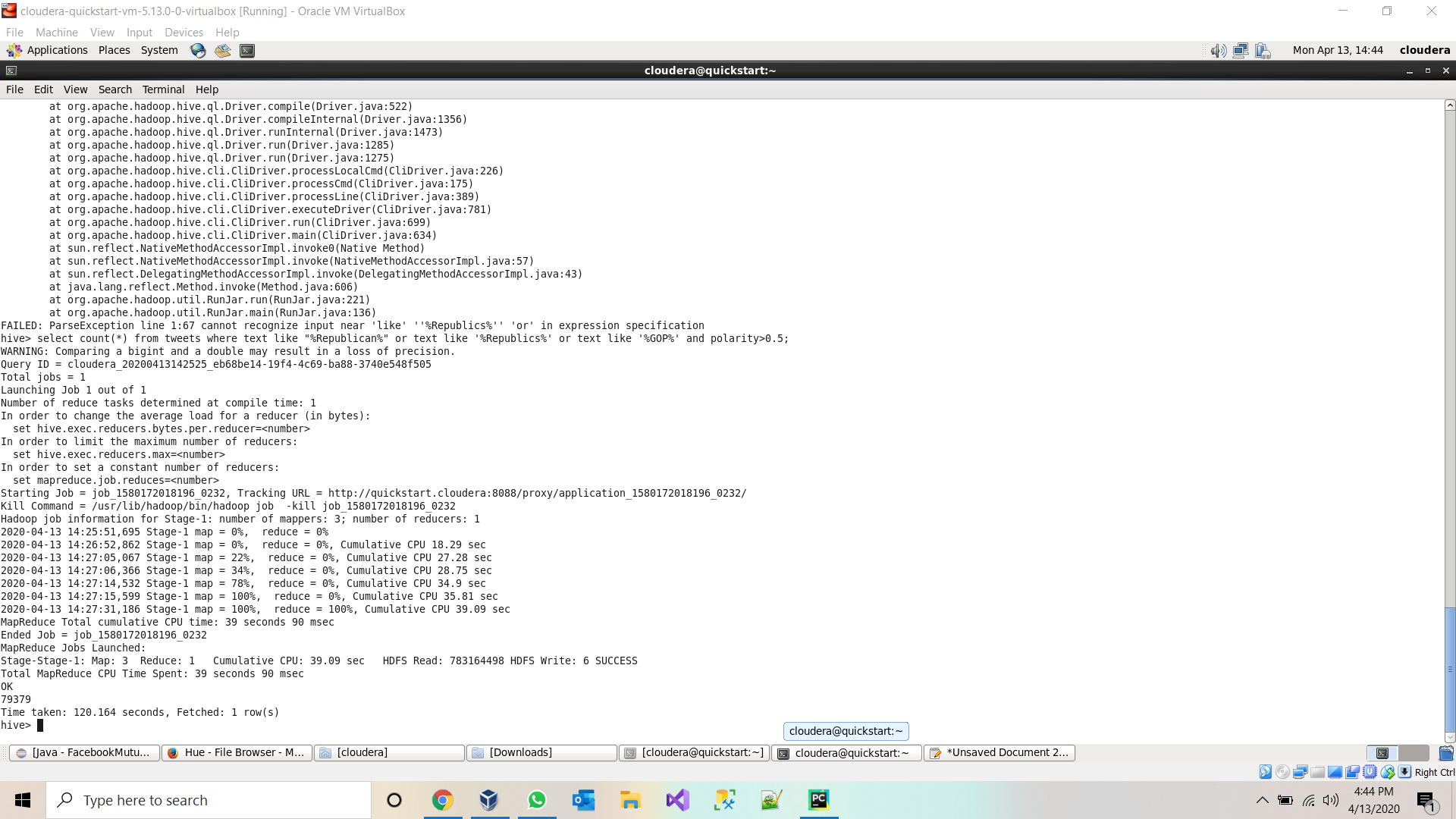
**Query 1:** 1.Basic query: total number of tweets using aggregate function

* Select count(\*) from tweets

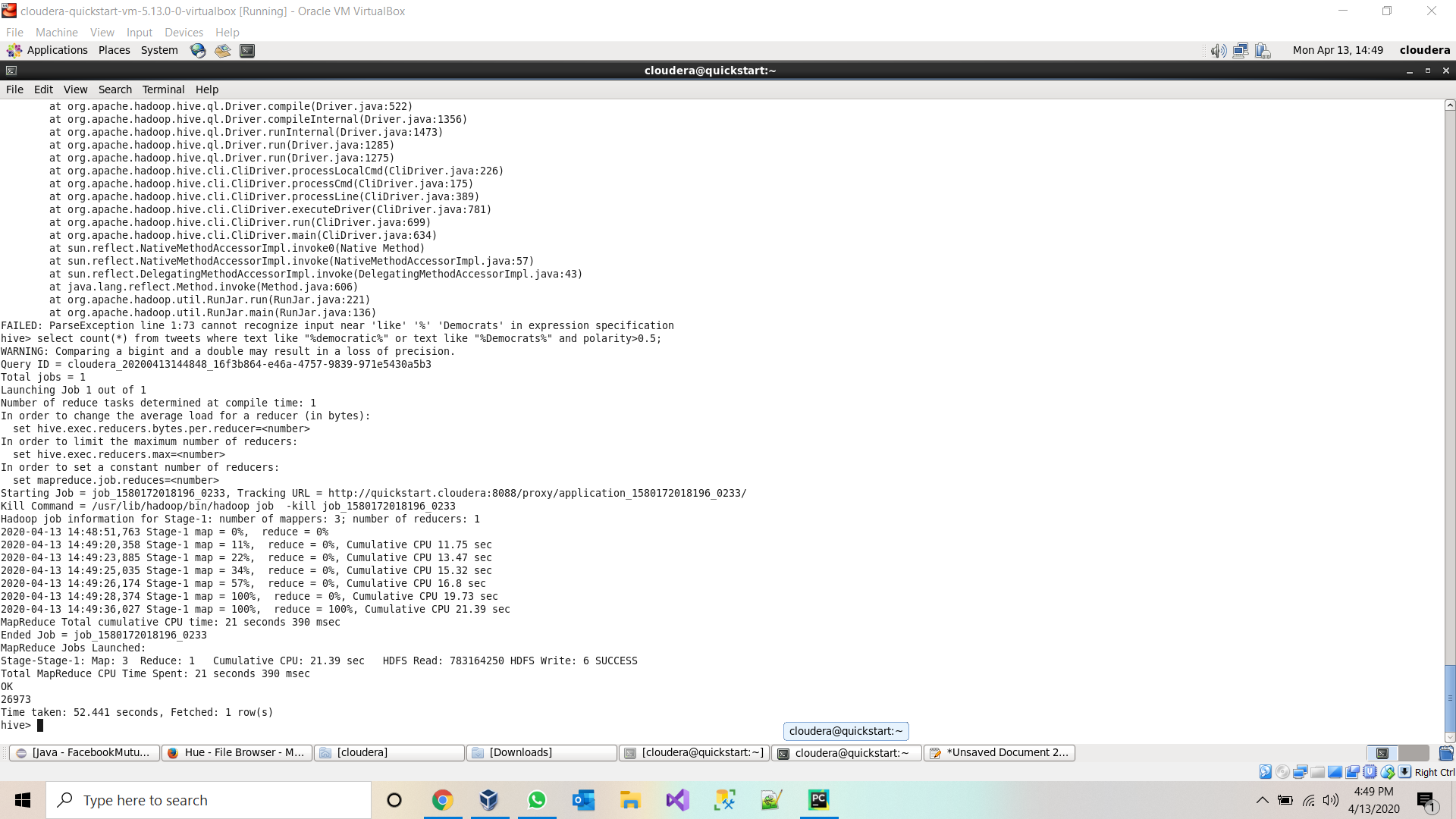


2.Number of positive tweets for republic party using polarity of sentimental analysis

select count(\*) from tweets where text like "%Republican%" or text like '%Republics%' or text like '%GOP%' and polarity>0.5;

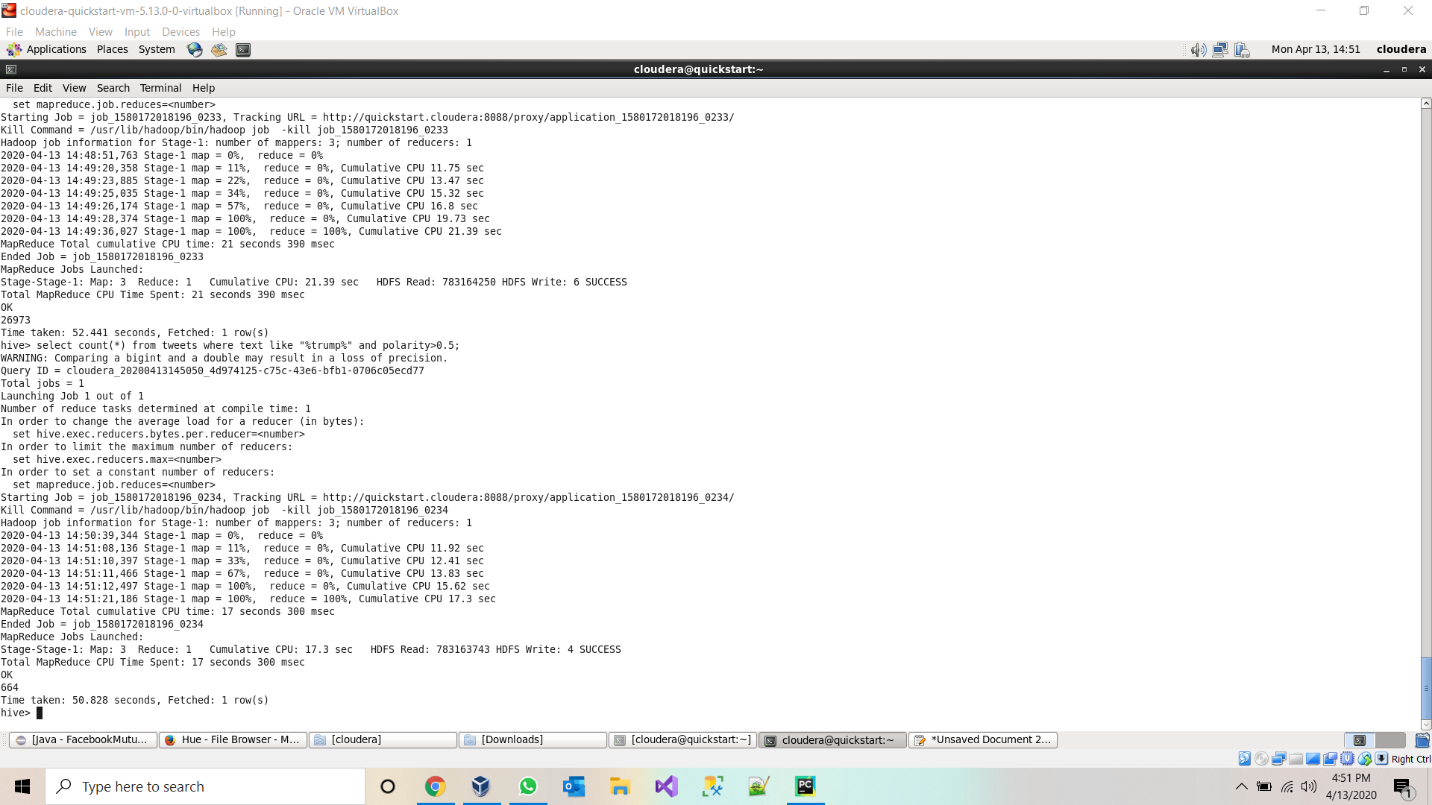


3.Count of positive tweets on democrats using polarity of sentimental analysis object. select count(\*) from tweets where text like "%democratic%" or text like "%Democrats%" and polarity>0.5;



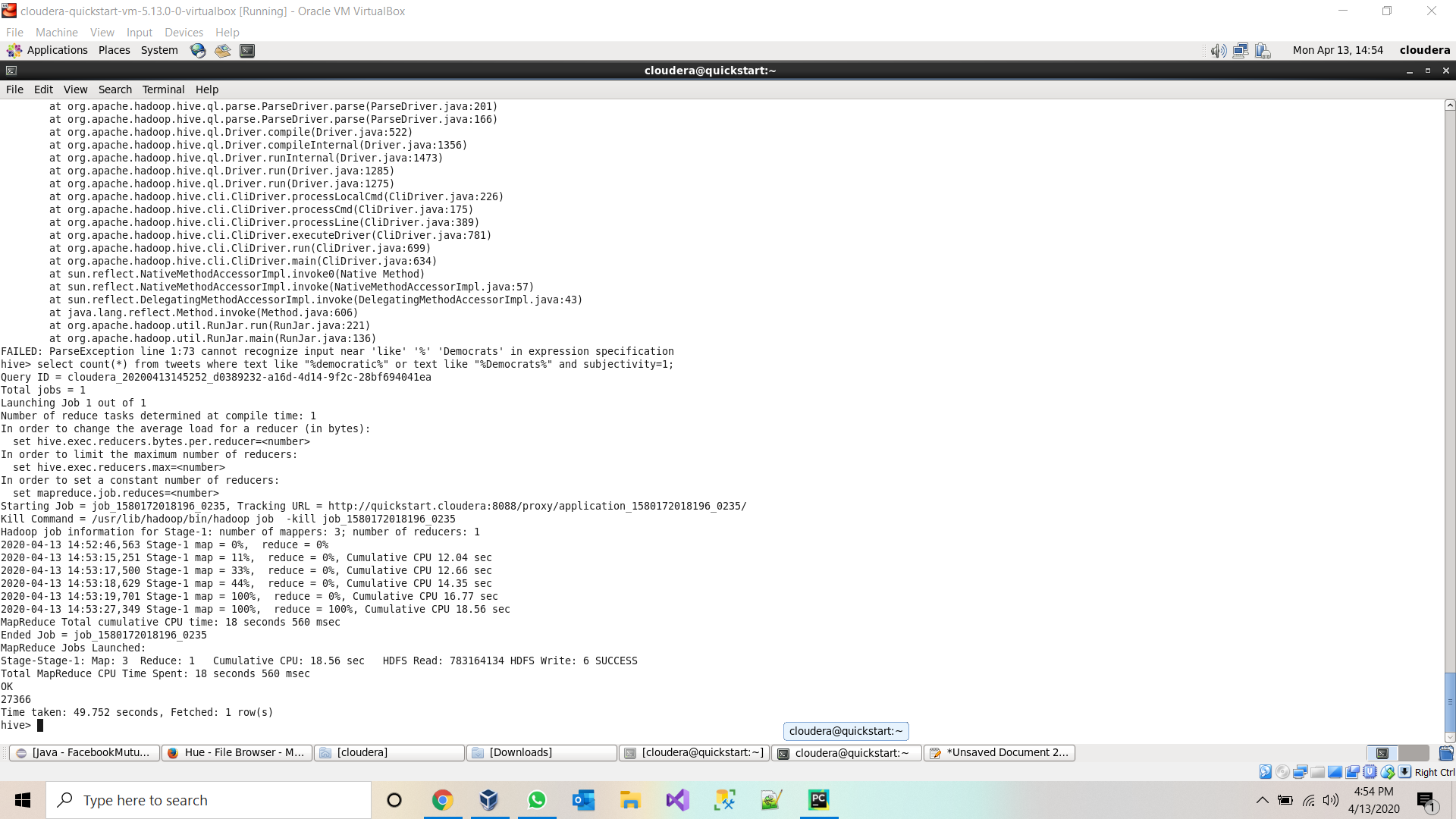
4.Tweets which contains text of US president trump and with positive.

select count(\*) from tweets where text like "%trump%" and polarity>0.5;

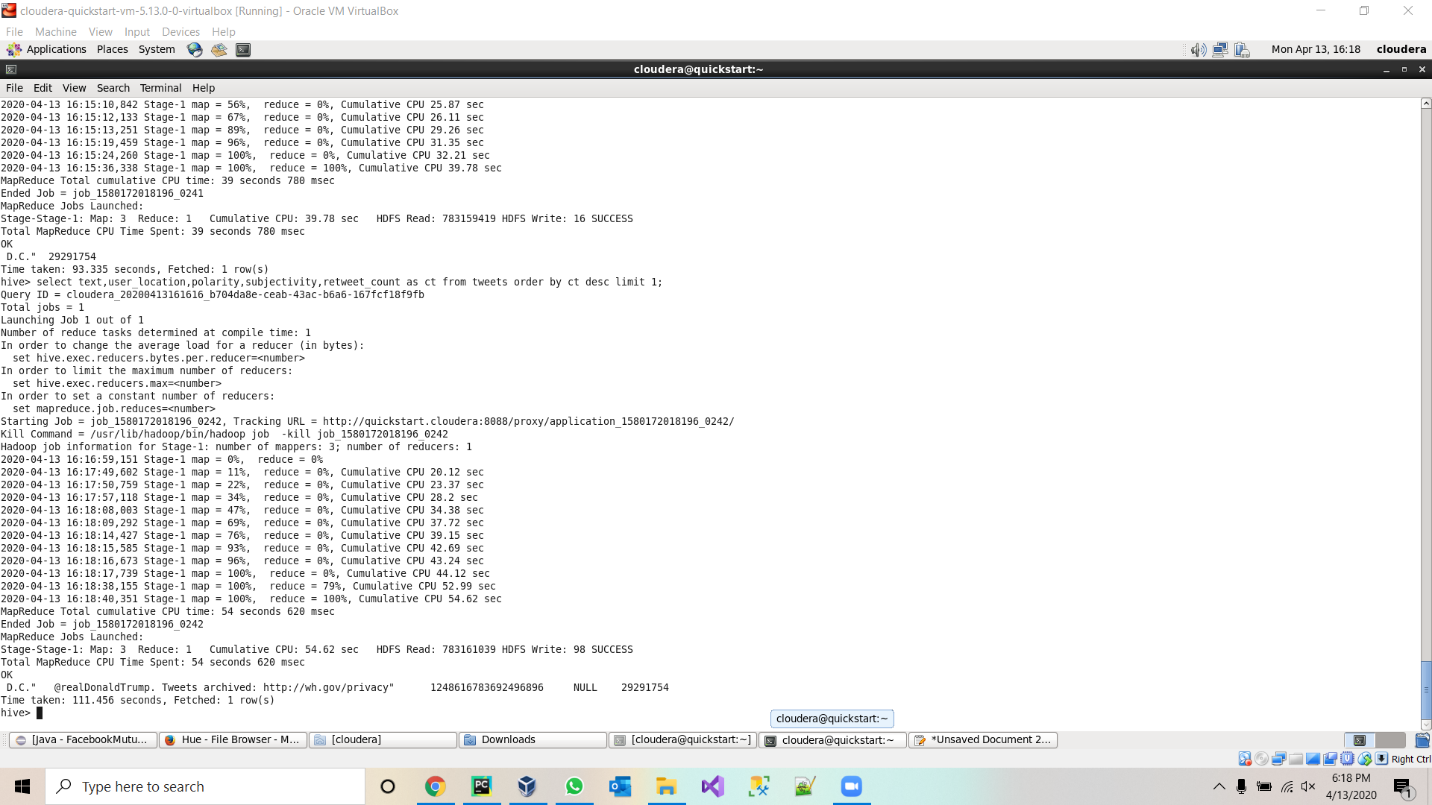


5.Count of tweets which are subjective towards democrats.

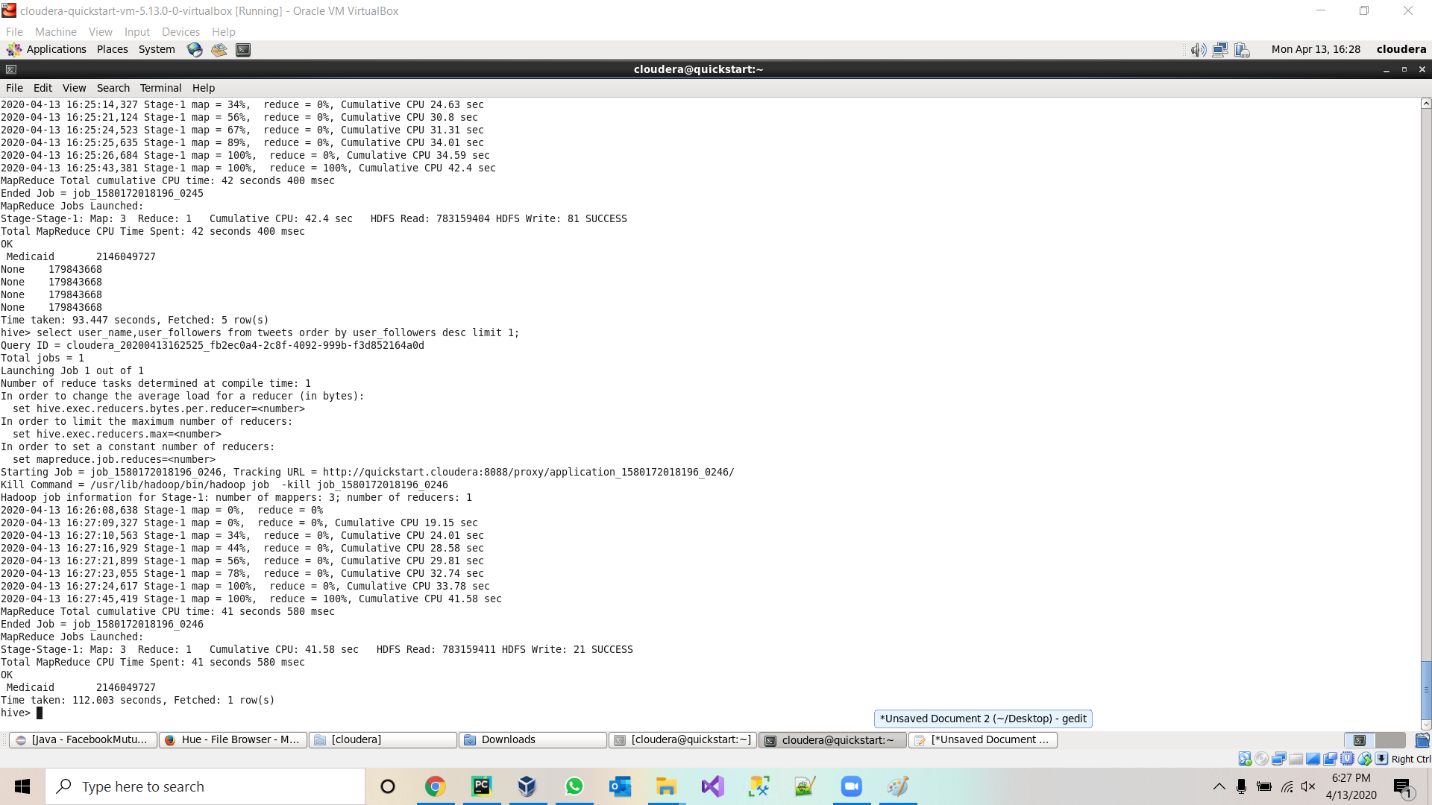
select count(\*) from tweets where text like "%democratic%" or text like "%Democrats%" and subjectivity=1;



6.Details of tweet which has highest retweet count select text,user\_location,polarity,subjectivity,retweet\_count as ct from tweets order by ct desc limit 1;

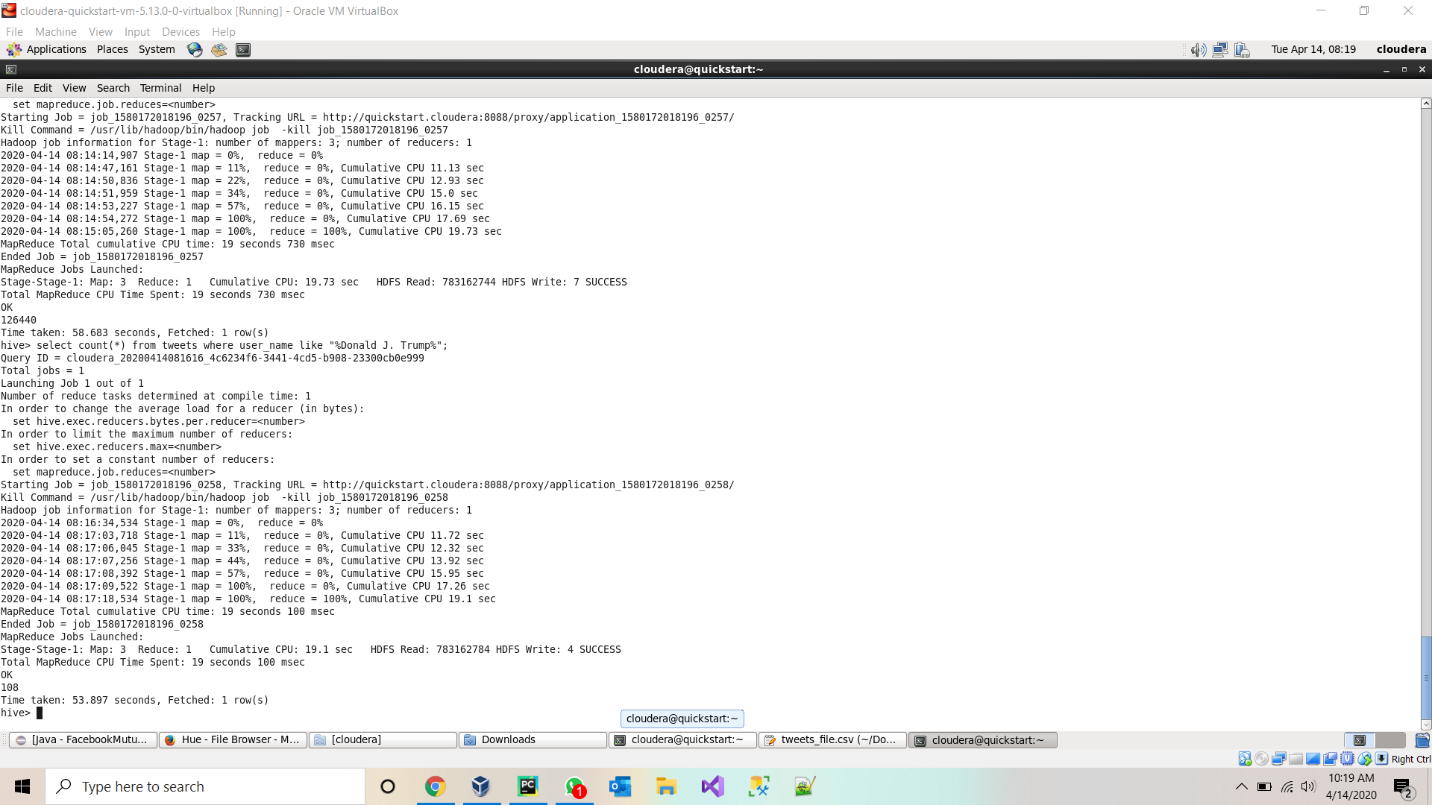


7.User which has highest follower select user\_name,user\_followers from tweets order by user\_followers desc limit 1;



8.Tweet by US president Donald Trump

Select count(\*) from tweets where user\_name like “%Donald J. Trump%”



**Project Management:**

**Implementation status report:**

**Work Completed:**

* Twitter data batch download & documented his part-Jagadeesh Maroju
* Data Cleanup and project implementation plan & documented his part- Paul Gomes
* HDFS import and Hue Visualization and Hive import & documented his part- Hari Raju
* Hive Queries and report- Praveen Poluri

**Work to be completed:**

* Loading data into solr and implement queries on the data in solr.
* create spark dataframes on CSV file and implement various transformations and actions on dataframes.
* We are also going to implement Spark sql, and query the data from dataframe.
* we will implement sentiment analysis on the data.

**References:**

* <https://towardsdatascience.com/twitter-data-collection-tutorial-using-python-3267d7cfa93e>
* <https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HDFSCommands.html>
* <https://docs.cloudera.com/HDPDocuments/HDP3/HDP-3.1.0/migrating-data/content/hive_moving_data_from_hdfs_to_hive.html>