Computer Sci Electrical Engr 5590 0004
Big Data Programming

Project Title

Stock Market Analysis using Hadoop Ecosystem

Team Members (Team-1)

- Jaisekhar Koya
- Sri Sai Nikhil Kantipudi
- Sai Rohith Guntupally
- Aarthi Nagireddy

Introduction

Study the stock market with graphs, news, important dates, and many more is always hard which leads to inaccuracy in analysis and investment in particular stock. This project helps to analyze the stock market using stock market data and twitter data using Hadoop ecosystem. It collects the twitter data associated with the stock symbols and analyze/predict the stock movement.

Background

There are many articles which tells about the handling of stock market using Hadoop Technologies. But there is no depth analysis of data. Only few articles talked about twitter data and did analysis only for couple of days and not used NLP Techniques. For example: https://www.3pillarglobal.com/insights/analyze-big-data-hadoop-technologies

Goals & Objectives

The main goal of this project is to analyze the stock market with the help of stock market data and also the social data(twitter data in this case) using Hadoop ecosystem. This system collects the twitter data associated with the stock symbols and calculate the sentiment of the stock based on the user tweets and extract the trend from the analysis.

- ### Motivation

We have tried hard to study the stock market with graphs, news, important dates, and many more, but we failed many times which leads to loss of profits and sometimes even the

investments. We always have an extremely hard time analyzing that huge amount of data and coming to a decision based on the analysis. And analyzing that huge data takes a lot of time and got to concentrate on various works too. So, to do this task for us we need some sort of Big data tools to analyze and give us filtered results which we can cross-check and invest in that stock confidently. Not only that as big data tools analyze data extremely fast, we can make decisions quicker than we do. We can also avoid human error using these tools.

- ### Significance
    - As more than half of the population are investing in stocks it is one of the most significant topics to be considered.
    - 30% of investors failing to analyze the stock and predicting the uptrend or downtrend of stock is one of the major reasons.
    - Big data tools help to analyze the huge data which helps to provide the efficient results.
    - It reduces the analysis time of investors and helps in making decisions faster and error free.

- ### Objectives

    To create a stock market analysis system, we will be collecting the twitter data using NLP techniques and processing the data. Different tools in Hadoop ecosystem will be used to analyze the data and help in predicting the movement of the stock.

- ### Features

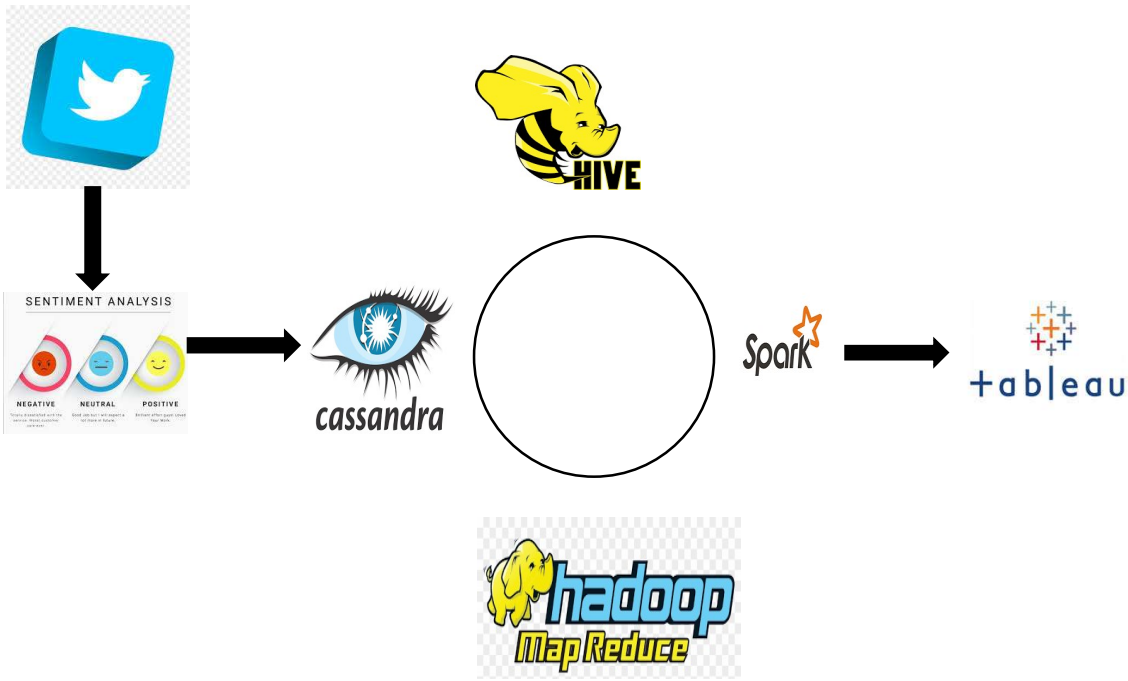
    The main features of this project is twitter data extraction, sentimental analysis on the extracted data. Using HDFS to store the data. Apache spark to process the data and Hive for querying the data. A visualization tool for showing out the results.

Dataset

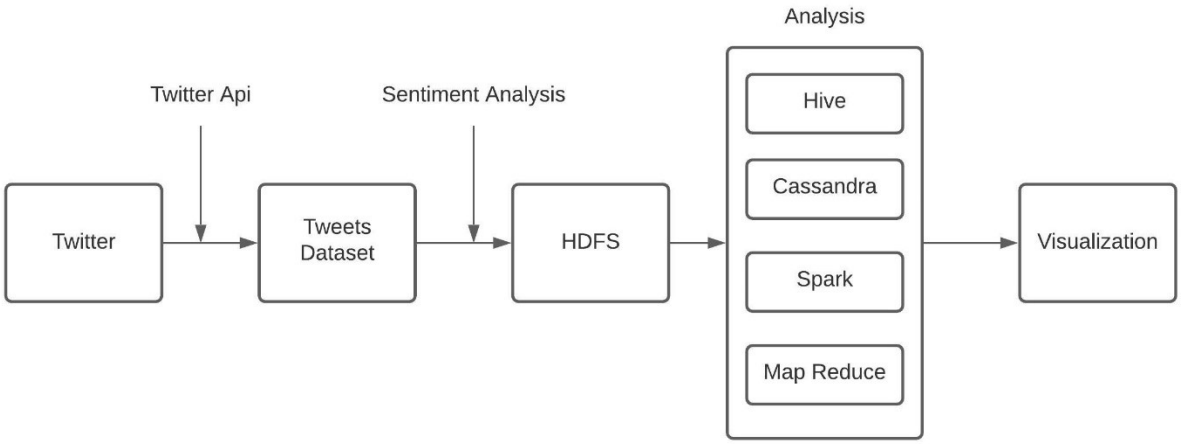
Twitter Dataset: (Only for project purpose. Do not share.)

https://mailmissouri-my.sharepoint.com/:u:/g/personal/skgyc_umsystem_edu/EXl5i-g3529CmwhdsDDSHoABtW7cOW4U_UtctybYaAsnyQ?e=WGriiy

Architecture



Design



Analysis: (Data)

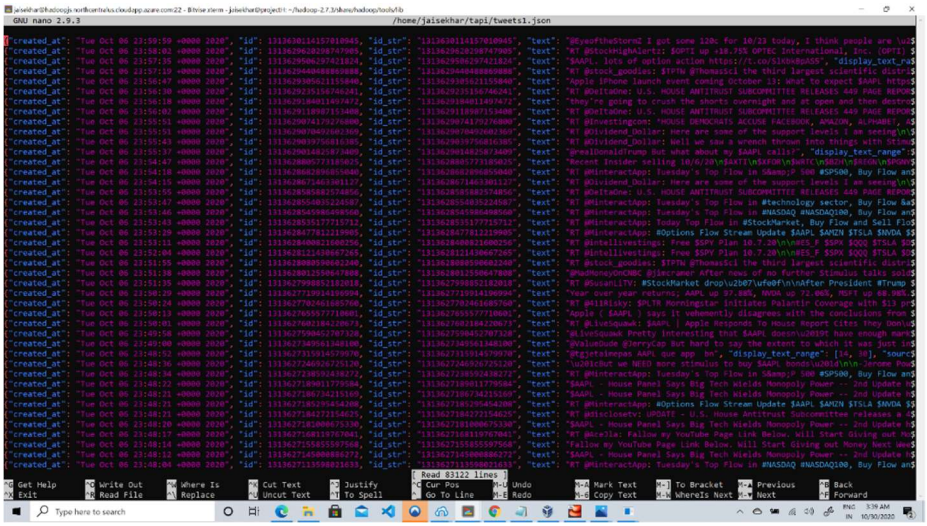
Twitter is one of the best data sources as it contains huge public opinions. Various number of users/ companies tweets the information regarding many things. This dataset is about AAPL stock information. Hence, we extracted all the tweets related to the AAPL stock. Since, our project is to analyze the previous data and predict/analyze the stock movement we are considering the tweets from October 1st to October 28th. We extracted more than 100K tweets and every tweet object contains complete information about the specific tweet like created_at, retweet_count, user details, entity details etc.

Keyword: AAPL

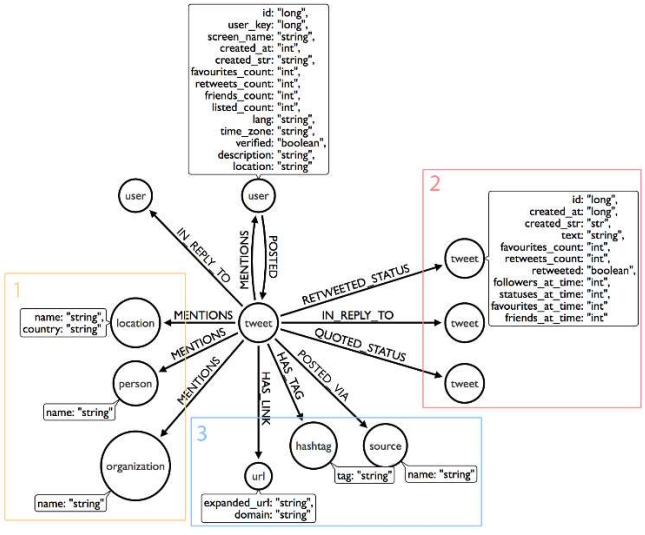
Date Range: 10/01/20 – 10/28/20

No. of tweets: 100k

The dataset will look like below



Features

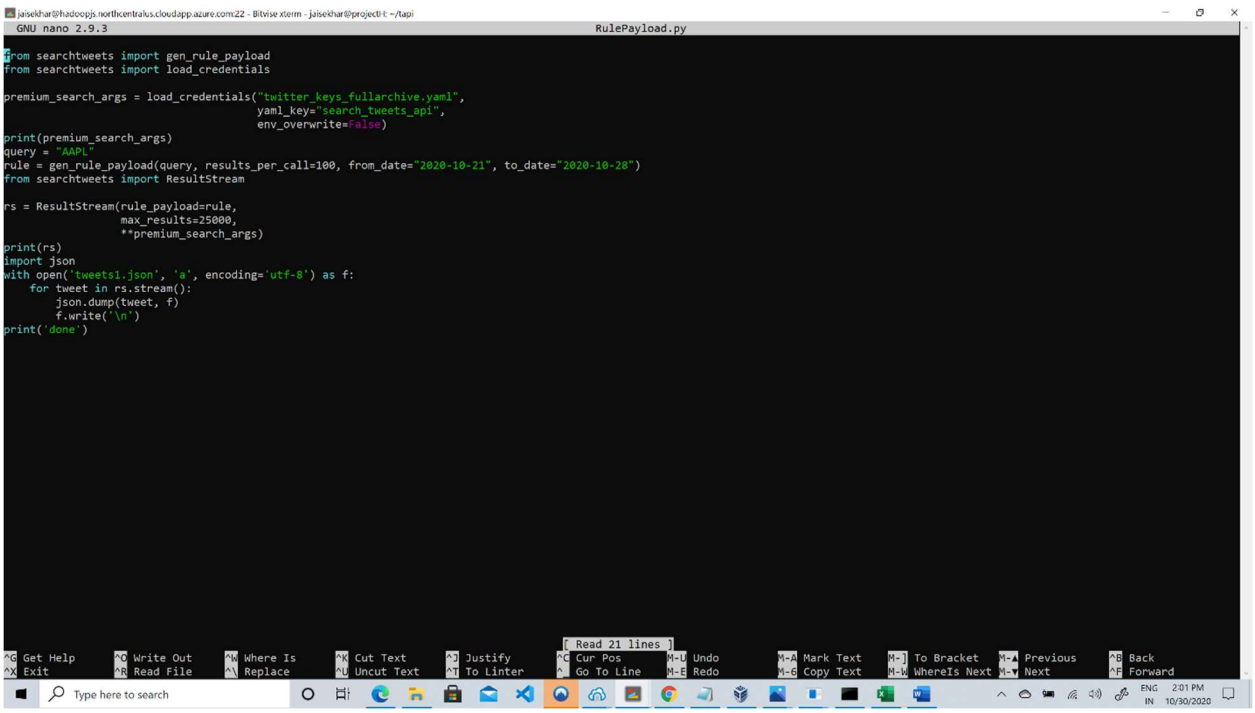


Implementation

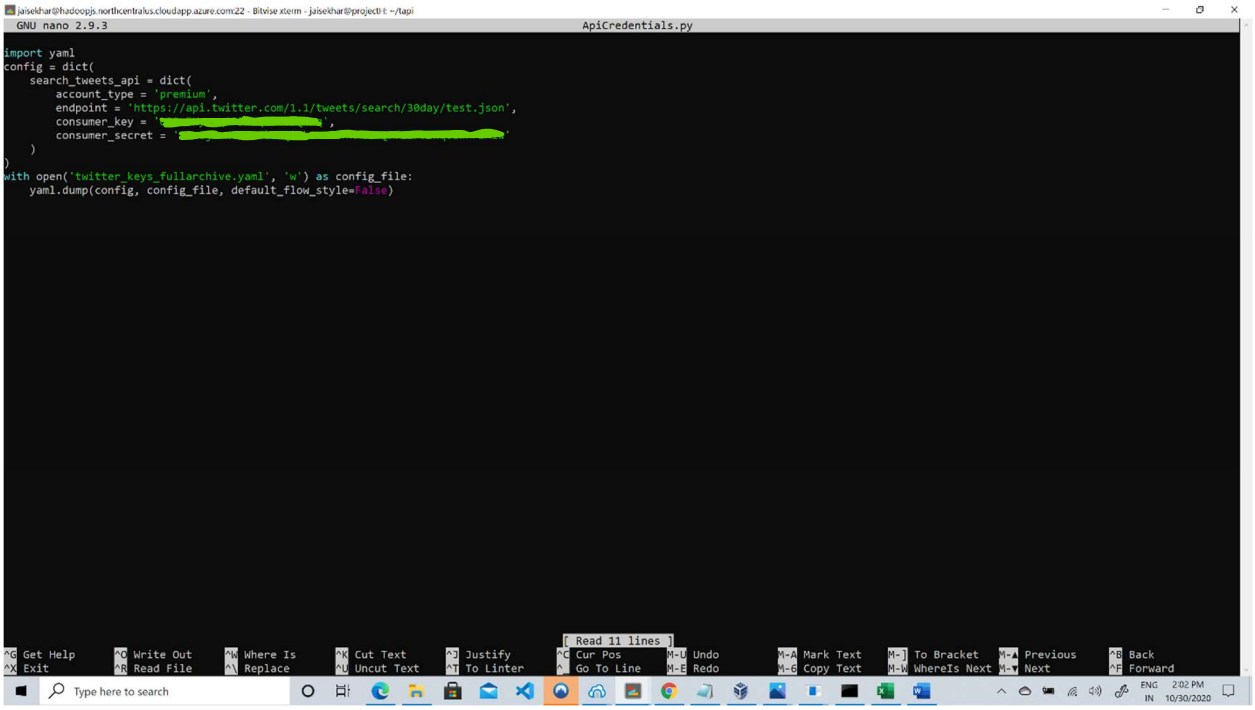
- ## Tweets Extraction:

To extract the tweets, we used different API's or library's in the python. Tweepy in the python is the widely used library to extract the tweets but it comes with limitations where we cannot extract tweets based on the data range. Even there were other libraries like twint or GetOldTweets3 but those were not working with the recent endpoint update at twitter api. Hence, we used searchtweets library which is nothing but a static library which makes direct API call to the twitter.

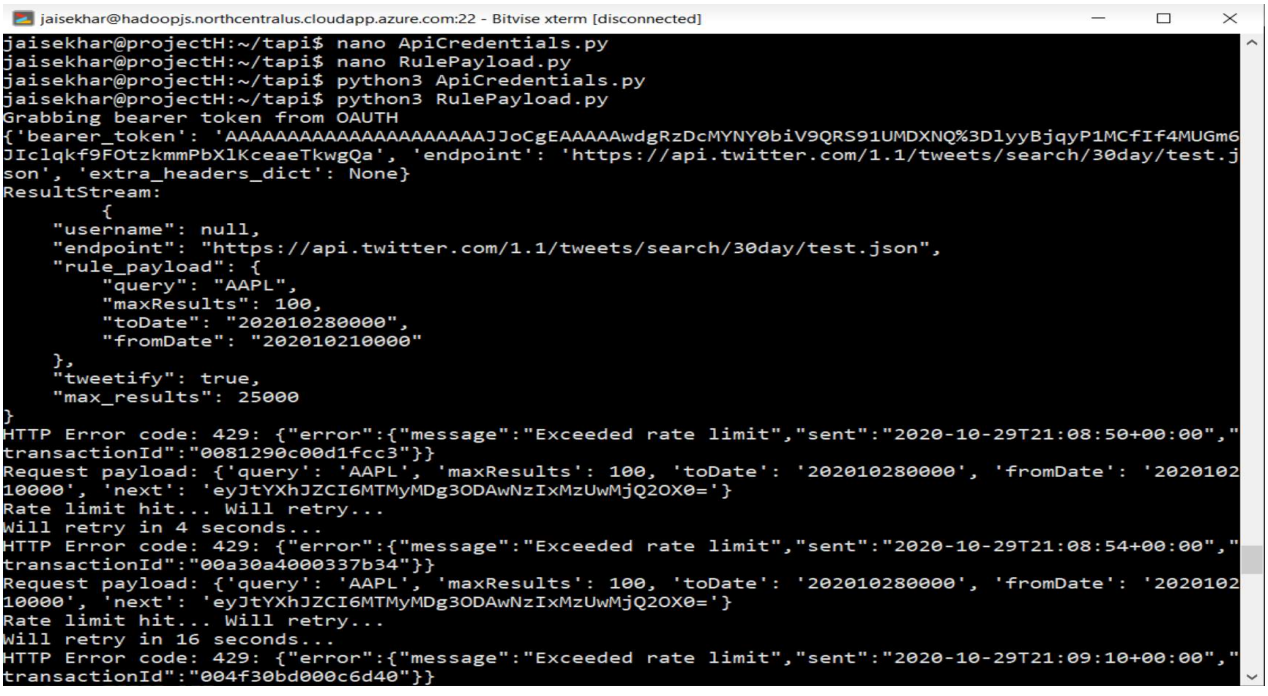
The below code is used to retrieve the tweets using searchtweets library. This code is importing the credentials from the other file and sending query(here we are querying AAPL), results_per_call, from_date, to_date (October month's) parameters to the api with max_results constraint which is set to 25000 here. As we used four keys we are able to retrieve approx 100k tweets by running four times of below code using four different twitter keys of ours.



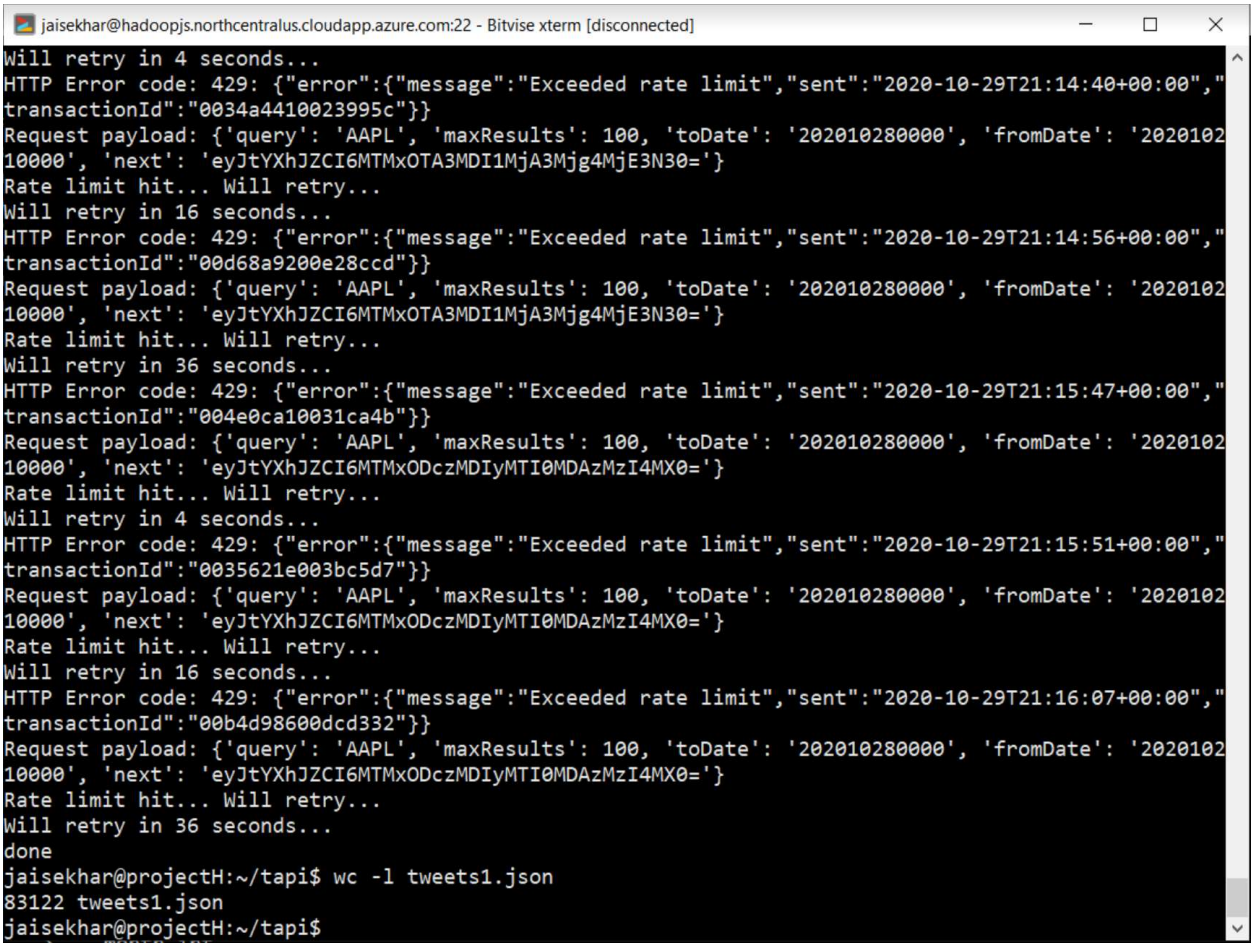
These are the credentials which are used to as the args to the searchtweets api .



Below you can see the execution of all the python files to extract the tweets.



The results are saved in tweets1.json and number lines in the below snippet is nothing but the number of tweets in the json file as the every line indicates one tweet object.

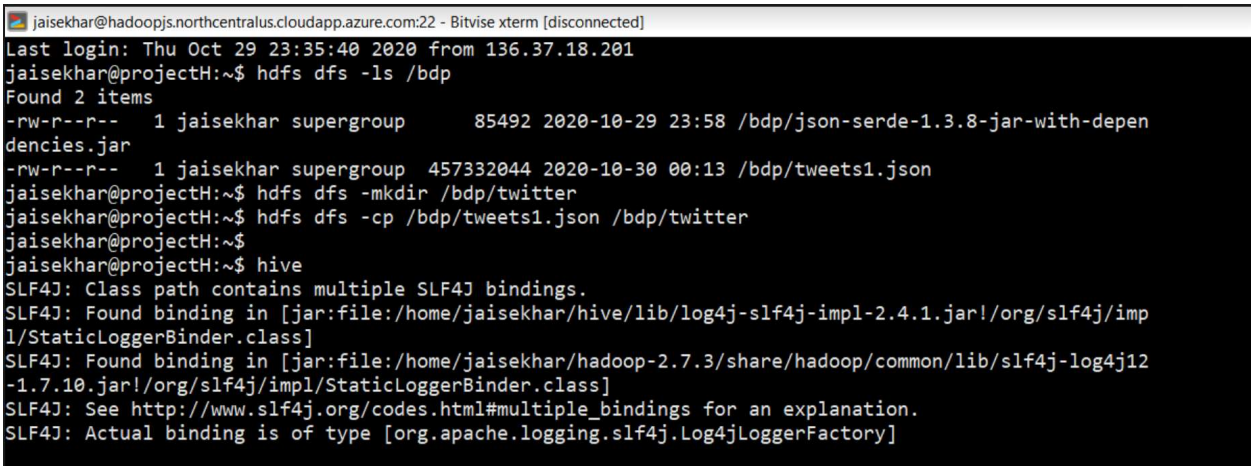


- ## Storing in HDFS:

The tweets file (tweets1.json) is copied to HDFS for further analysis using the the below commad

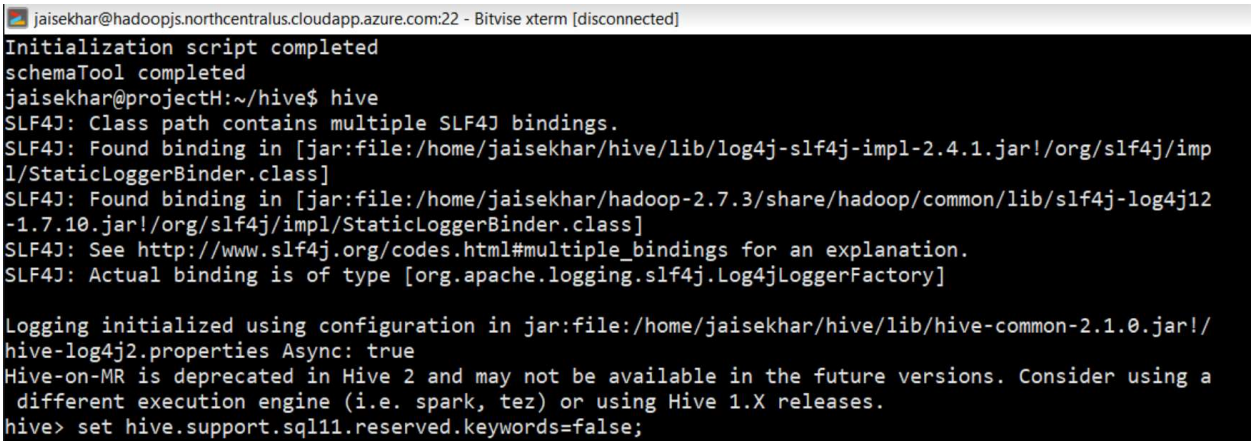
hdfs dfs -copyFromLocal ~/tapi/tweets1.json /bdp

Later, the file is moved to other directory which only contains that particular json file that helps while creating table in hive.

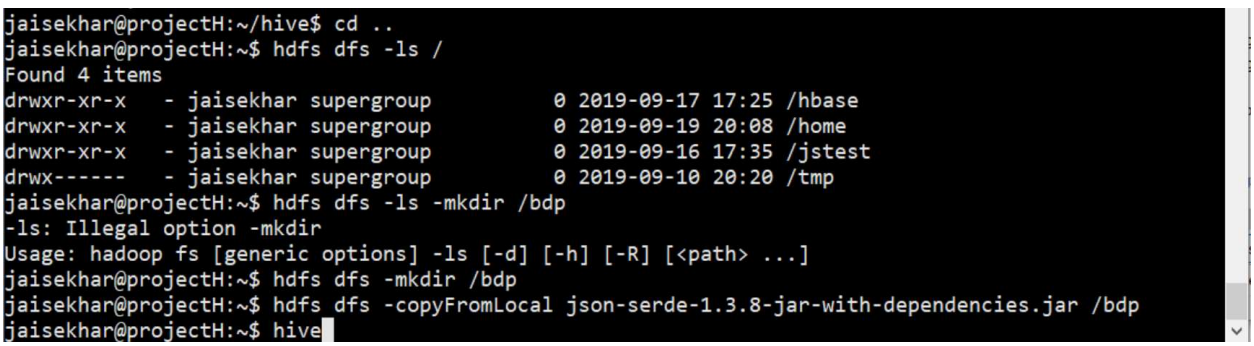


- Hive Operations: (initialization and Loading Data)

To start hive, we ran schematool to initialize the metastore_db and then we started the hive.



Later we added json serde dependency jar to the hive using following command. JSON SerDe is used to map the json object to the table schema. Thus, it helps in mapping the json data to the table.

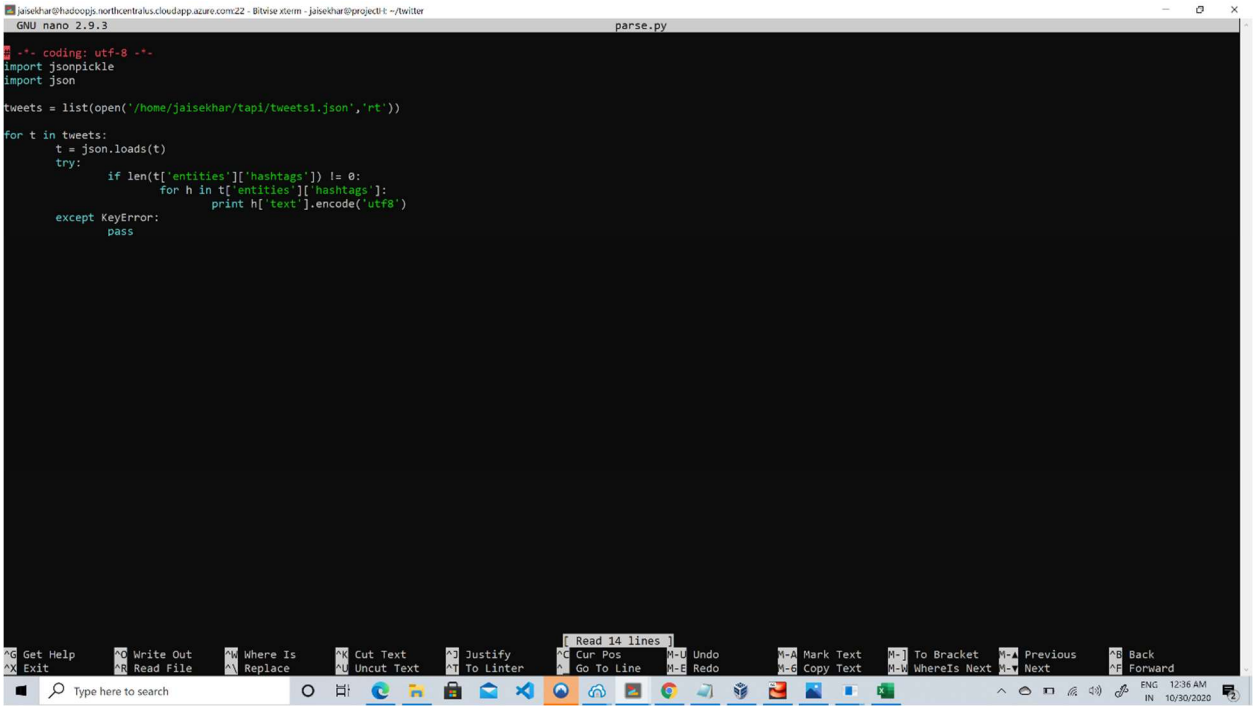


Next, we created the table in hive tweets_raw with the few required attributes from the tweet object which we extracted from twitter api. If you see here, the rows were delimited by serde which points to the added jar of json serde and location is pointing to the directory in hdfs which contains the tweets json file.

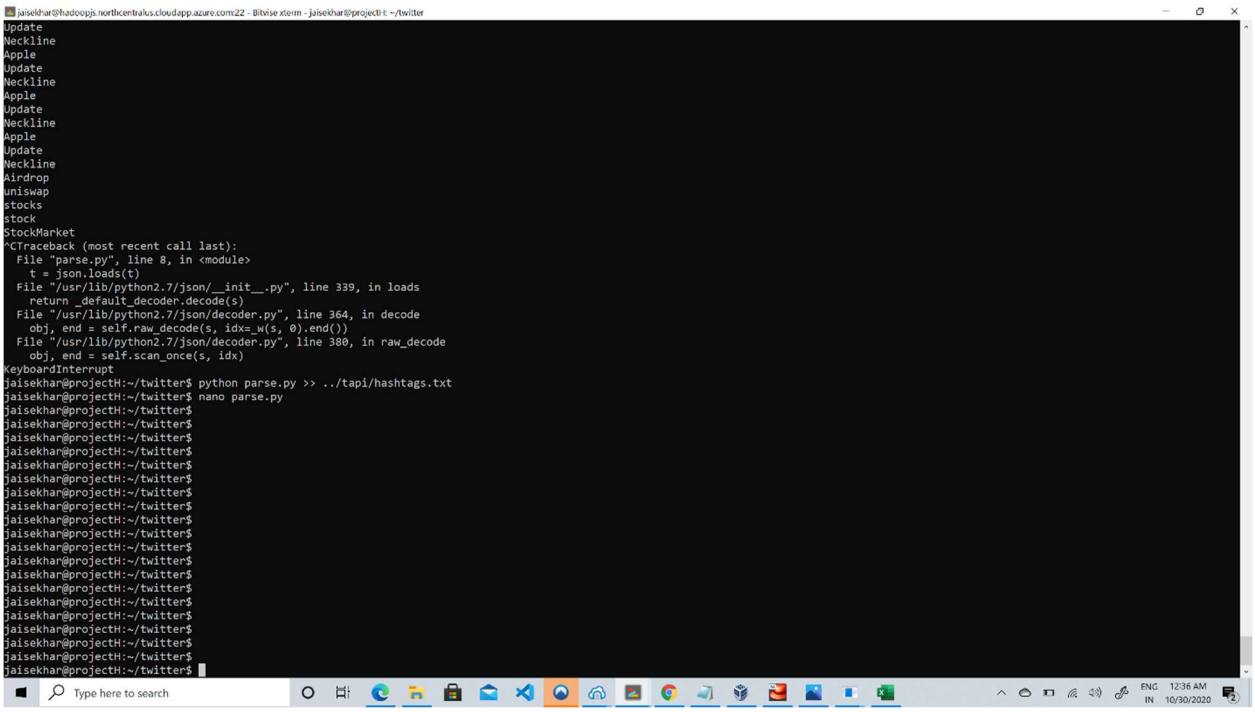
Now, the extracted tweets data is successfully imported to table in hive. Next, we will perform analysis on the data which we extracted into hive and hdfs.

- ## MapReduce – wordcount:

The below python code is used to parse the hashtags from the entities in tweets1.json



The python code is executed and saved the output in hashtags.txt using out operator.

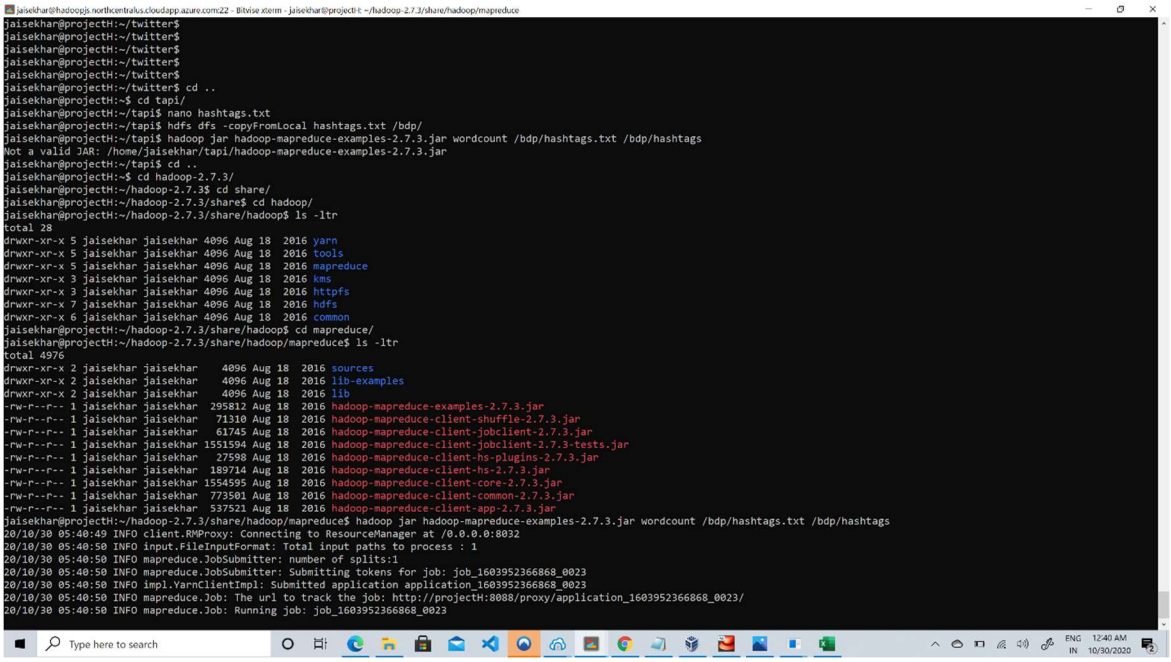


We successfully got the hashtags into the hashtags.txt. And as shown below the hashtags.txt is copied to hdfs using the below command

hdfs dfs -copyFromLocal hashtags.txt /bdp

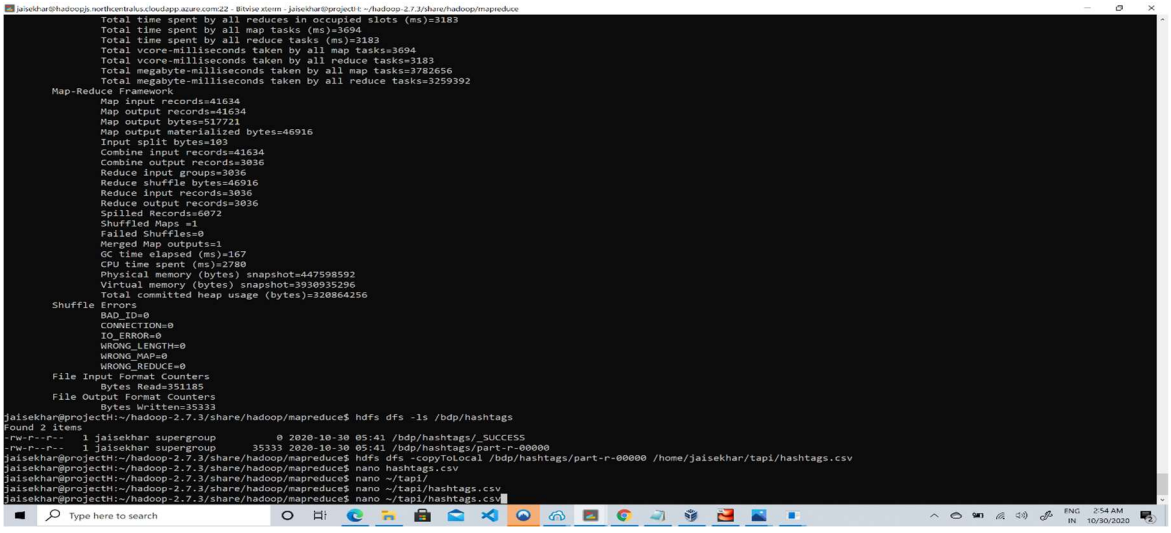
Now wordcount operation is performed using hadoop inbuilt example jar with hashtags.txt as input and saved the ouput in /bdp/hashtags directory in hdfs.

hadoop jar hadoop-mapreduce-examples-2.7.3.jar wordcount /bdp/hashtags.txt /bdp/hashtags

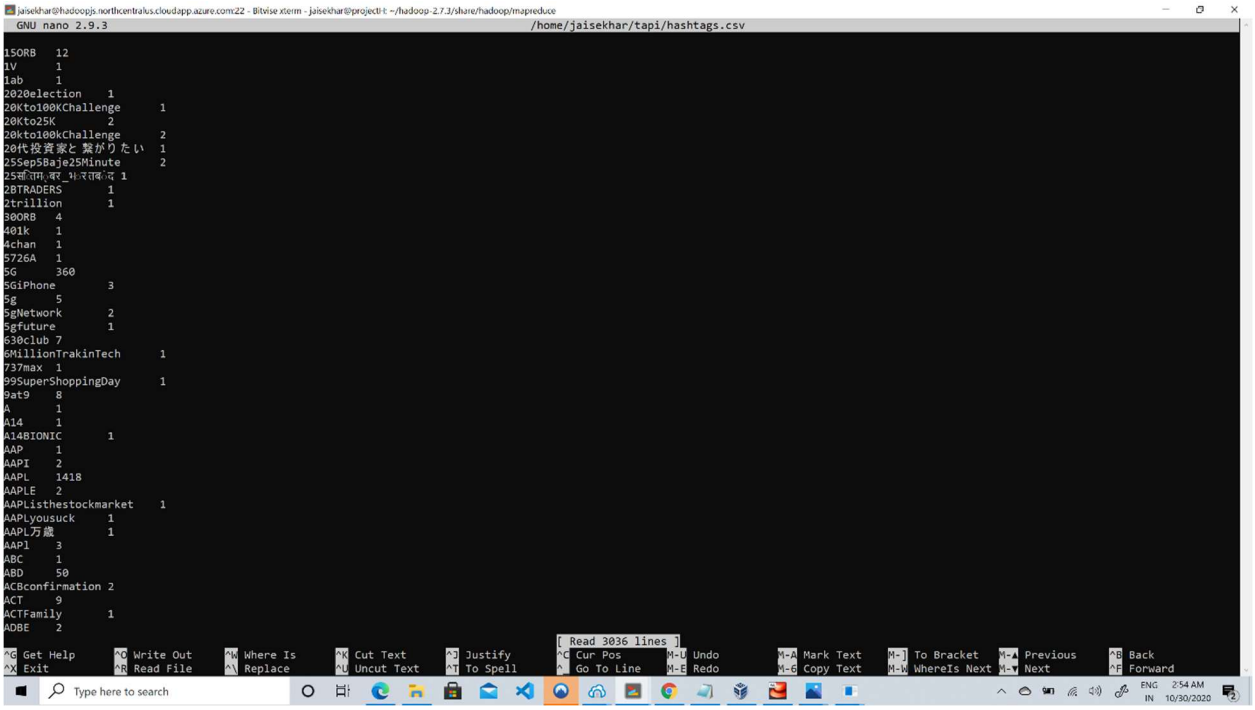


After the wordcount using the mapreduce, the saved output was sent to local for future analysis using the below commands.

hdfs dfs -copyToLocal /bdp/hashtags/part-r-00000 /home/jaisekhar/tapi/hashtags.csv



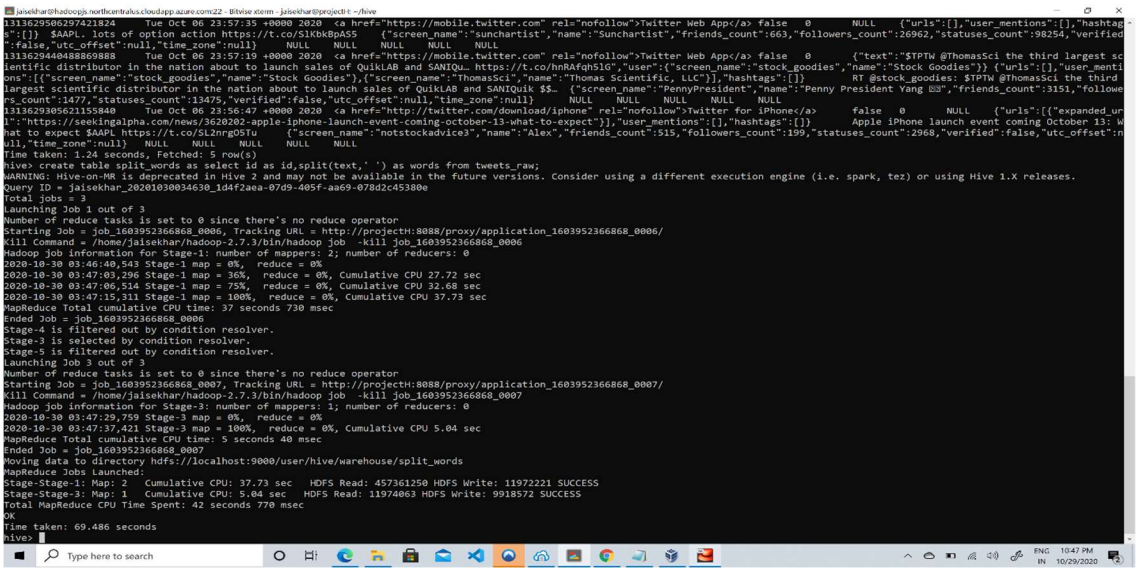
After performing the wordcount the output is like below where it shows the hashtag and their number of occurrences.



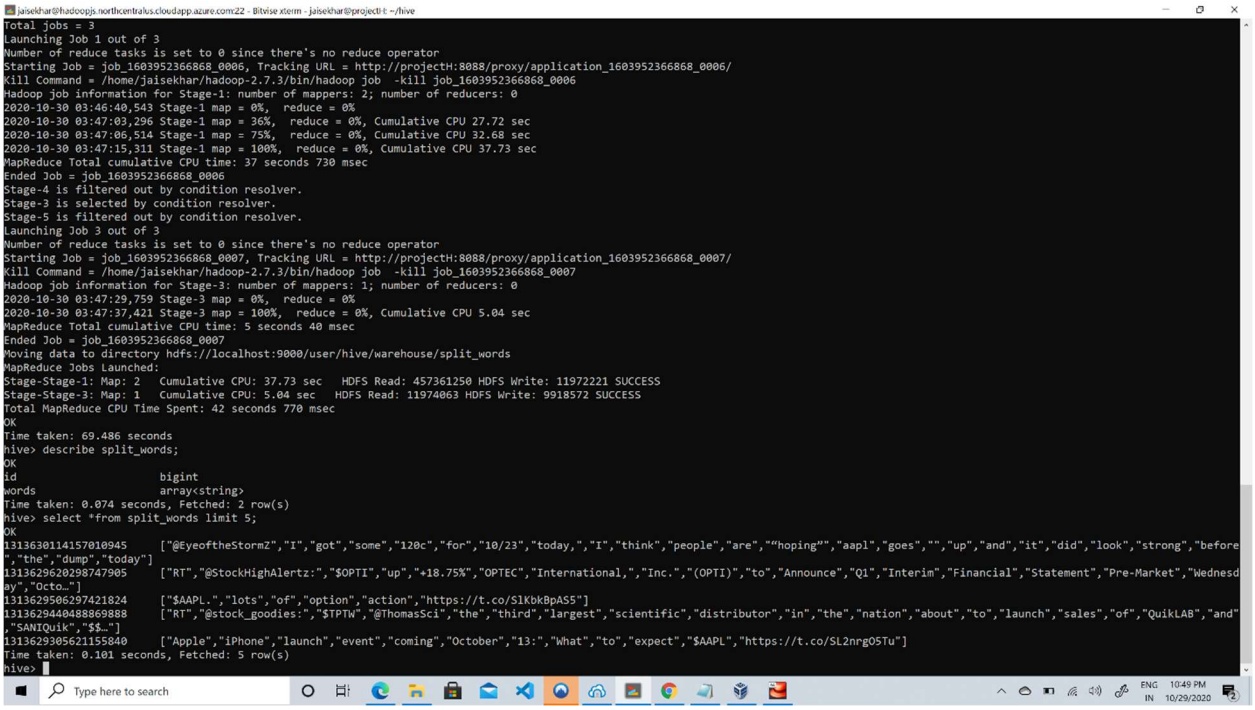
- ## Sentiment Analysis using Dictionary:

To perform sentiment analysis, we are using AFINN dictionary which contains more 2.5K words and values of polarity which ranges from -5.0 to +5.0 (Negative to Positive).

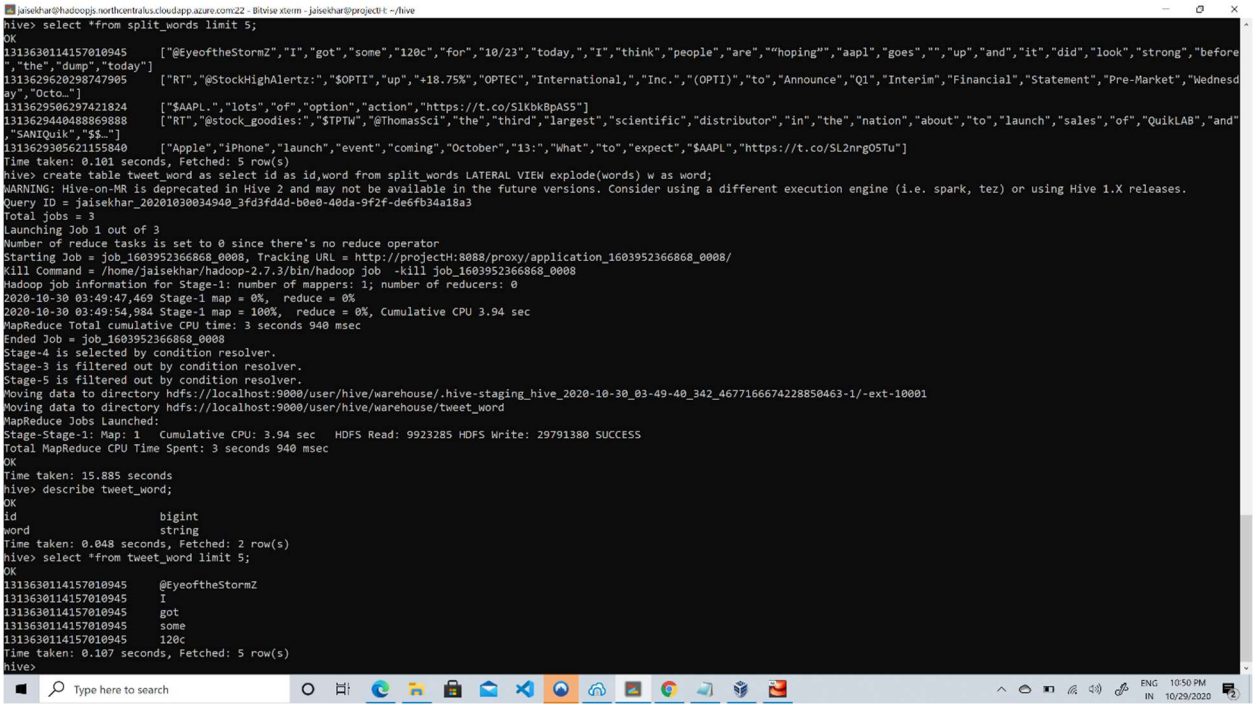
Initially, we separated the word using split function on text data in tweets _raw table and saved into split_words table using below command.



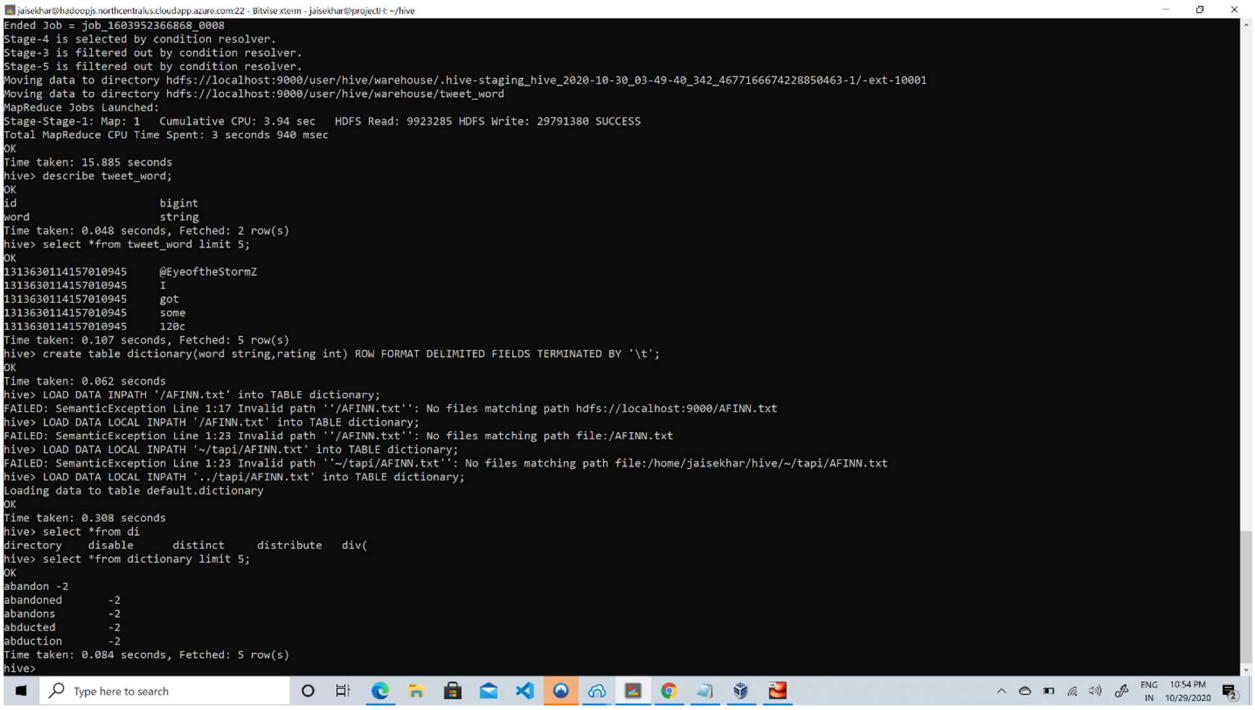
The split words in split_words table will look like below



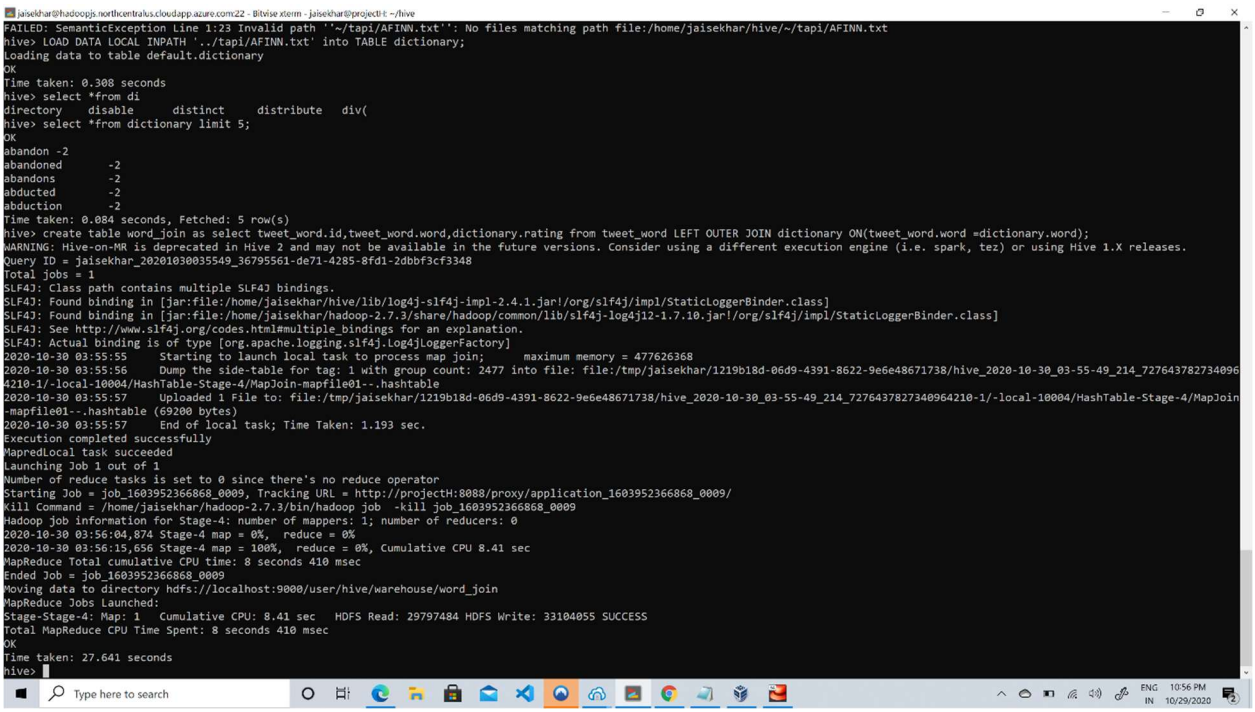
Now, we are splitting each word in the array as the new row. For this, we are using explode function which extracts the element from array into new row. Since it got some limitations we are using LATERAL VIEW. The output also shown below.



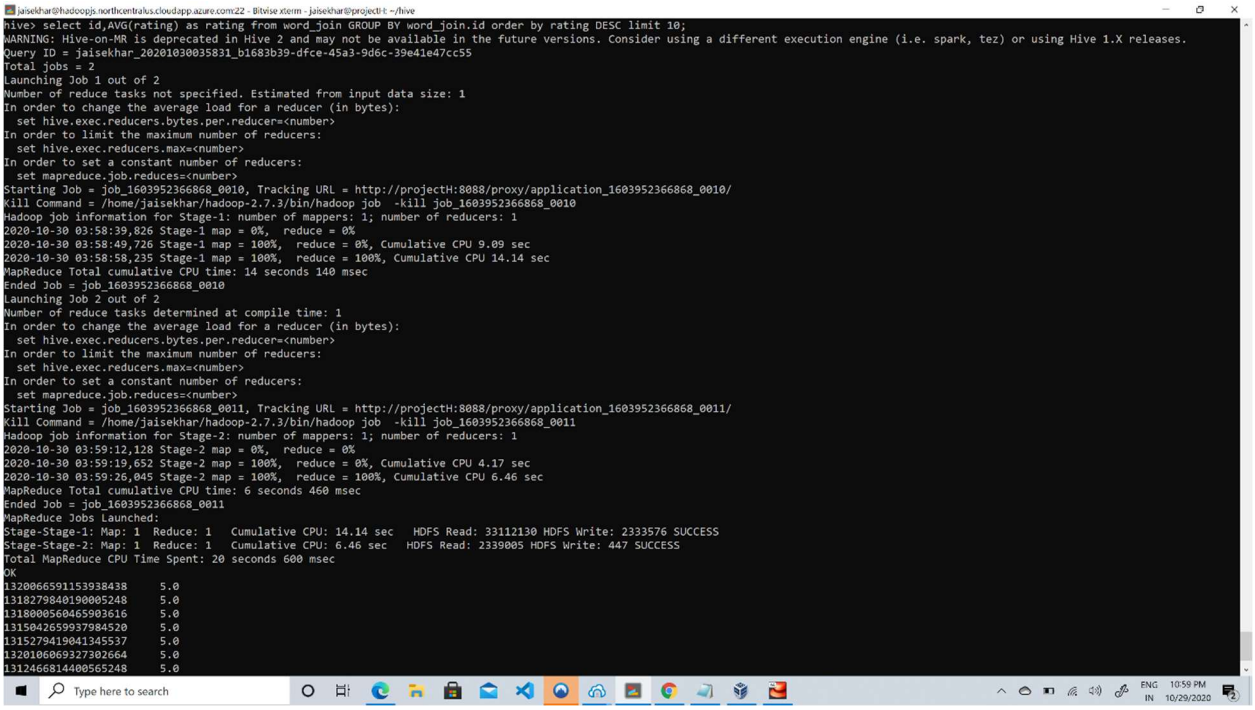
We got every word in new rows. Now, we will create new table named dictionary and will load the AFINN dictionary data into that table as as show below.



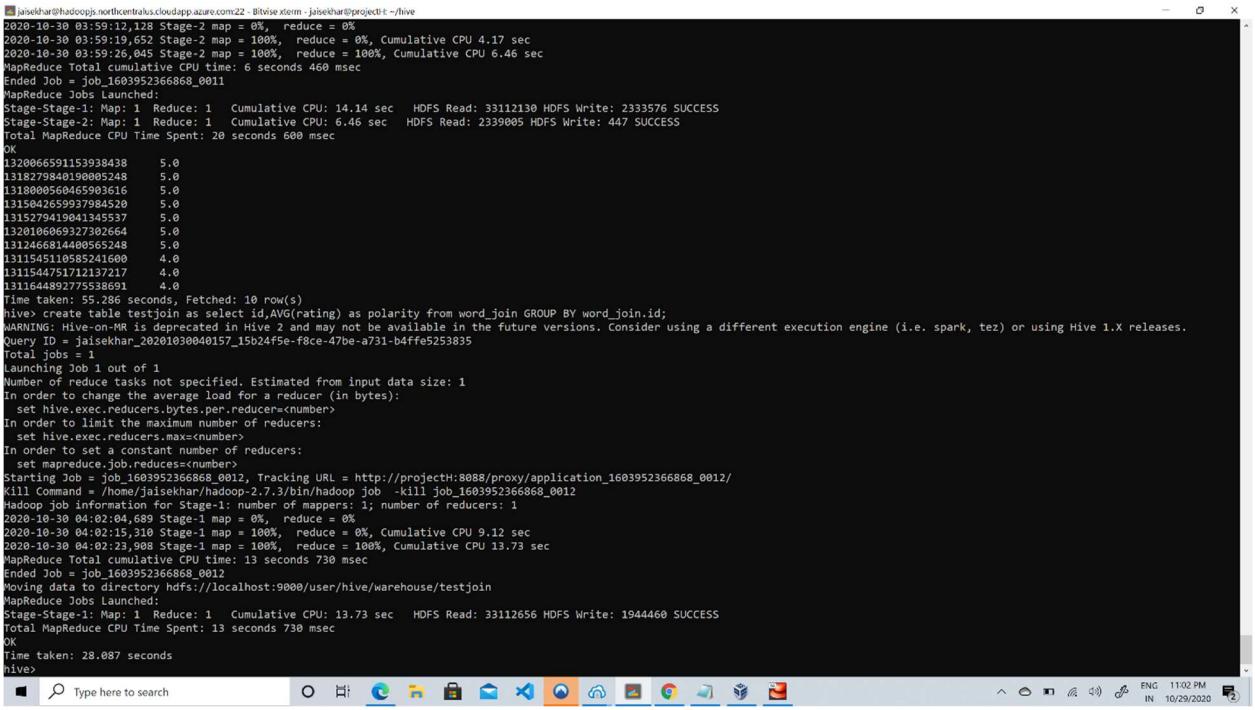
Create new table word_join by joining dictionary and tweet_word tables as shown below.



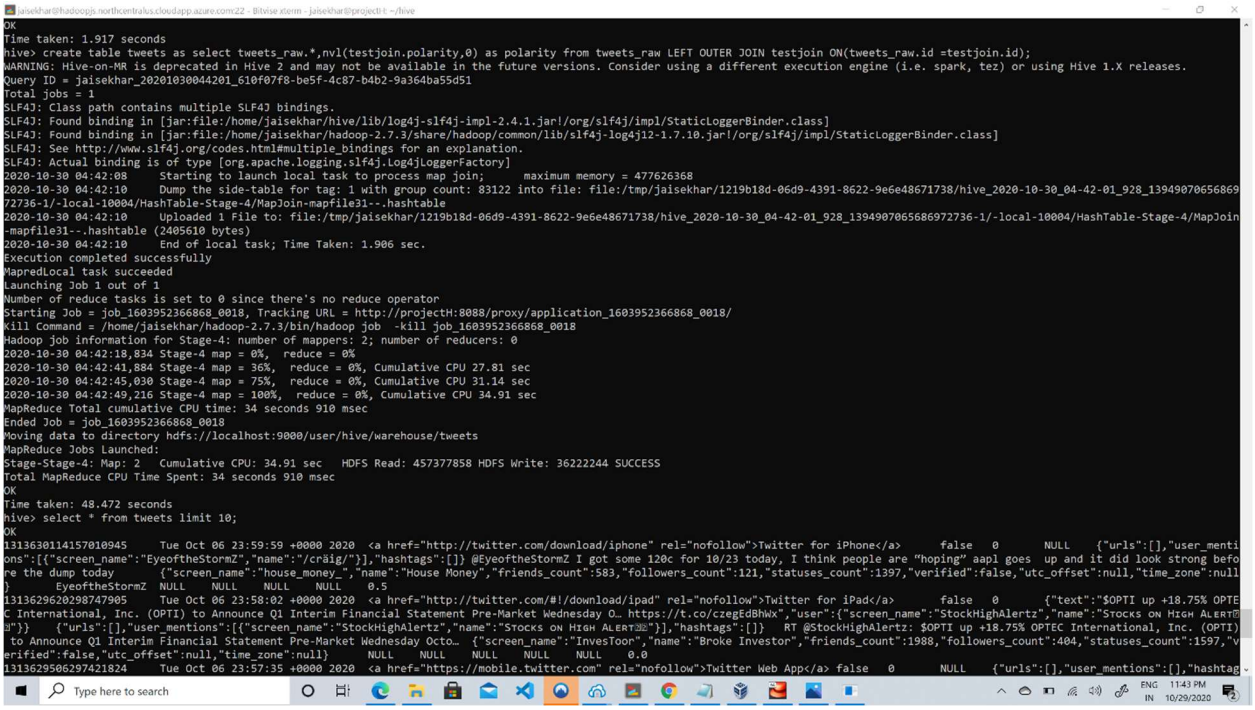
After joining now, we got the polarity value for each tweet in the word_join table which is shown below.



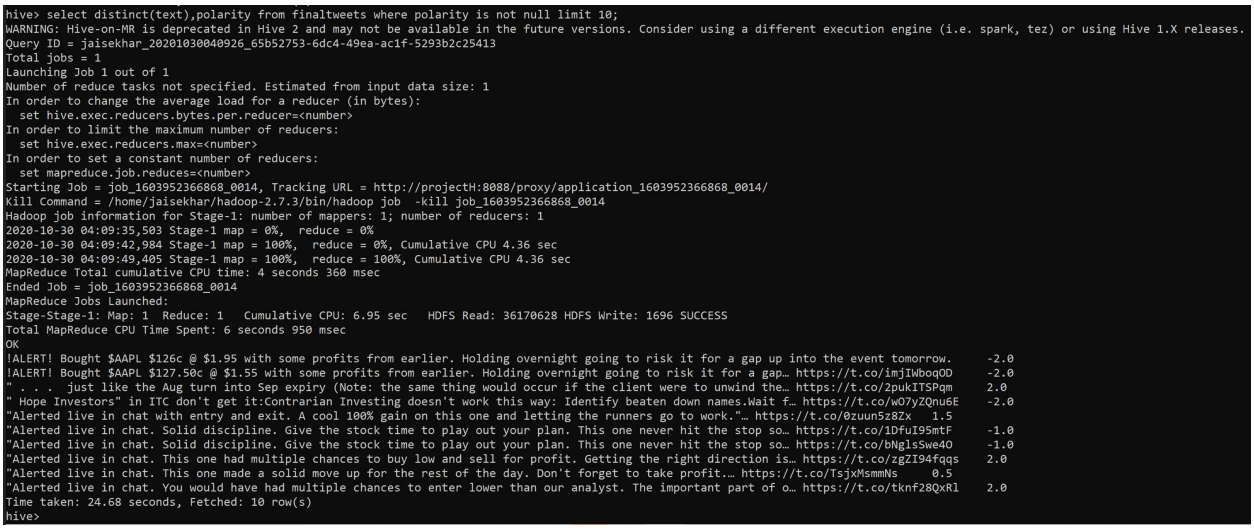
Now, we are saving the same id based grouped data into the testjoin table.



At last, we are joining the tweets_raw table and test join table based on id and created new table ‘tweets.



Below, we can the see sample data which the results show the tweet text and polarity value. This is one of the important part in this project. If you see the first tweets, it says holding the AAPL for overnight is risky hence it showed negative polarity value.

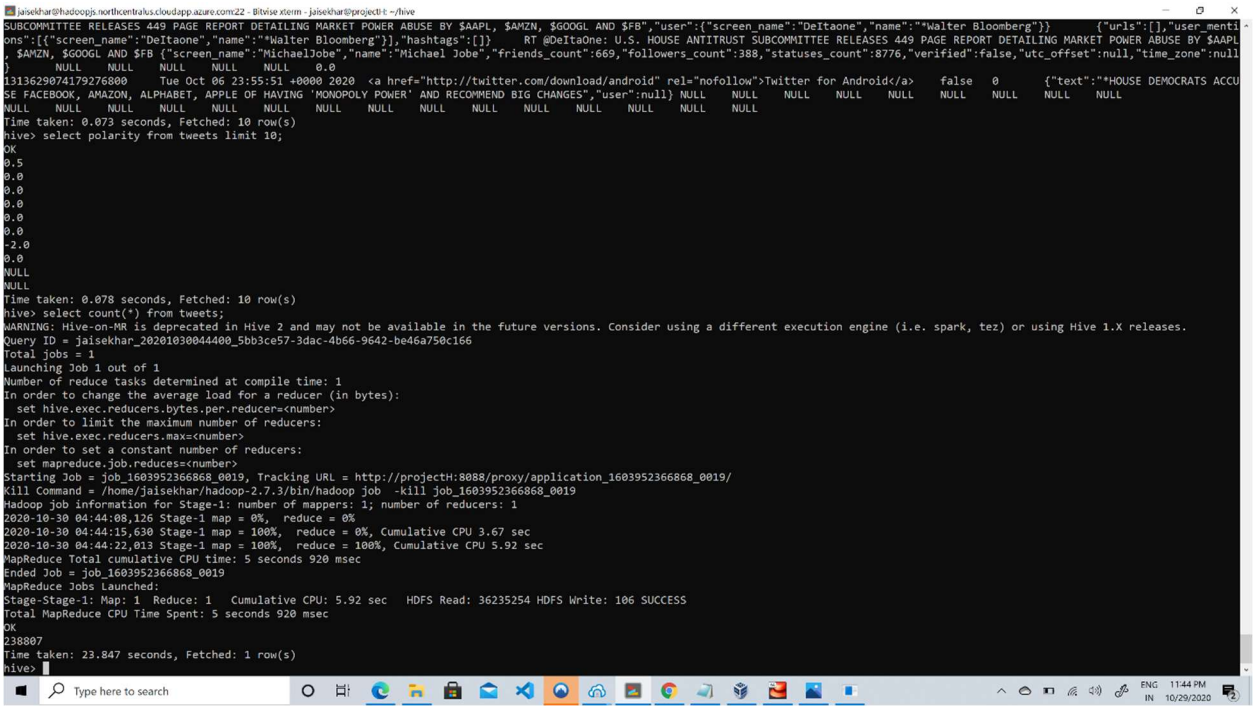


- ## Hive Analysis:

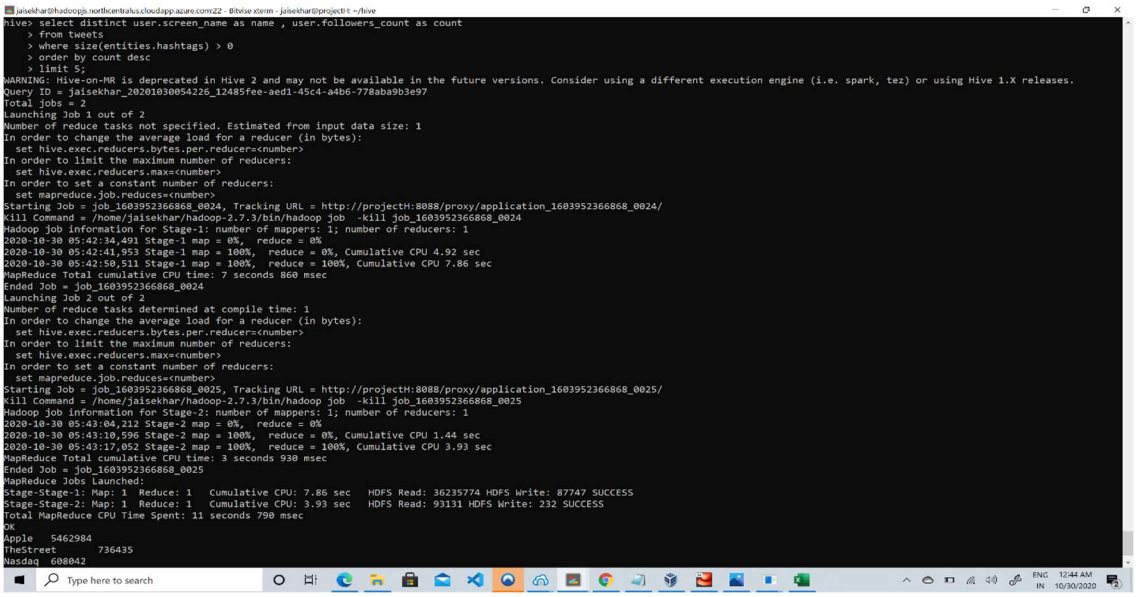
Below are the Queries for analysis using Hive.

- The below simple query shows number of records in the tweets table.

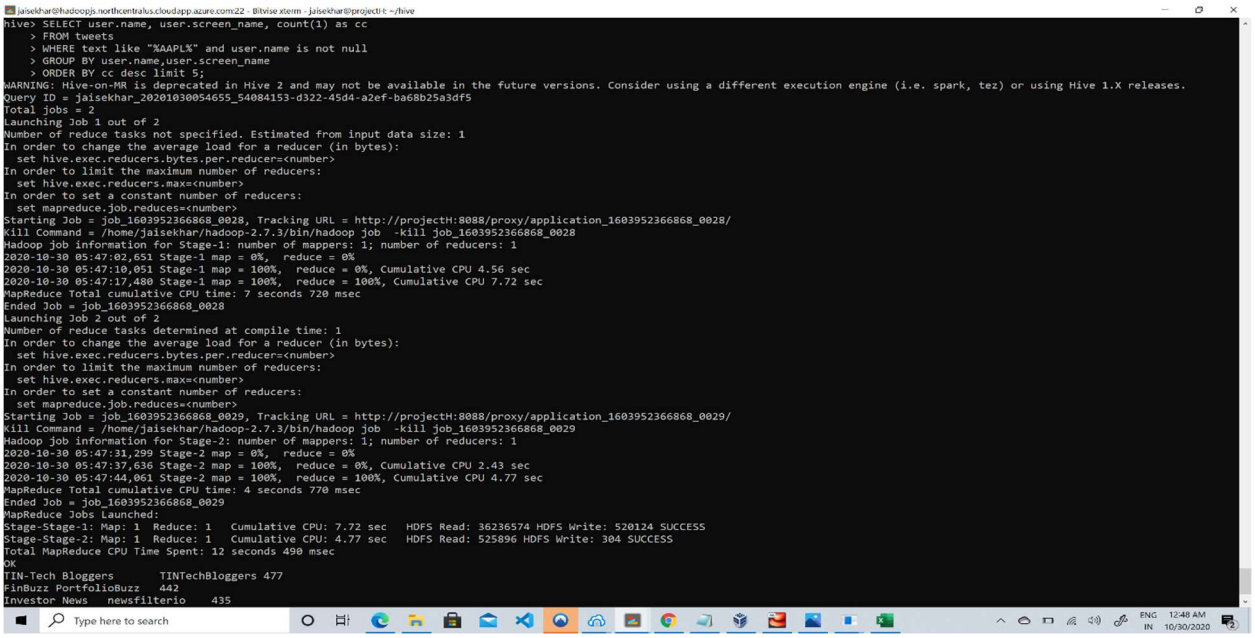
  select count (*) from tweets;



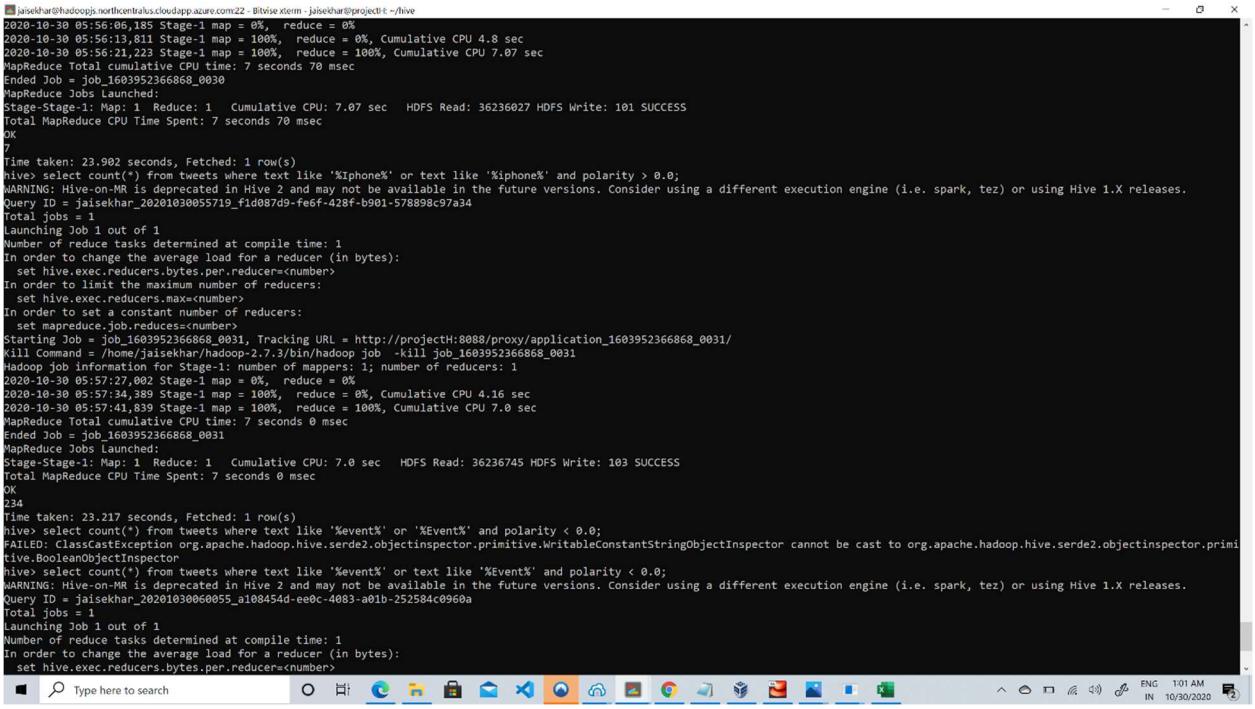
- The below query shows the screenname of the user and there followers count in the descending order in whole dataset. In this dataset, Apple user has the more number of followers.



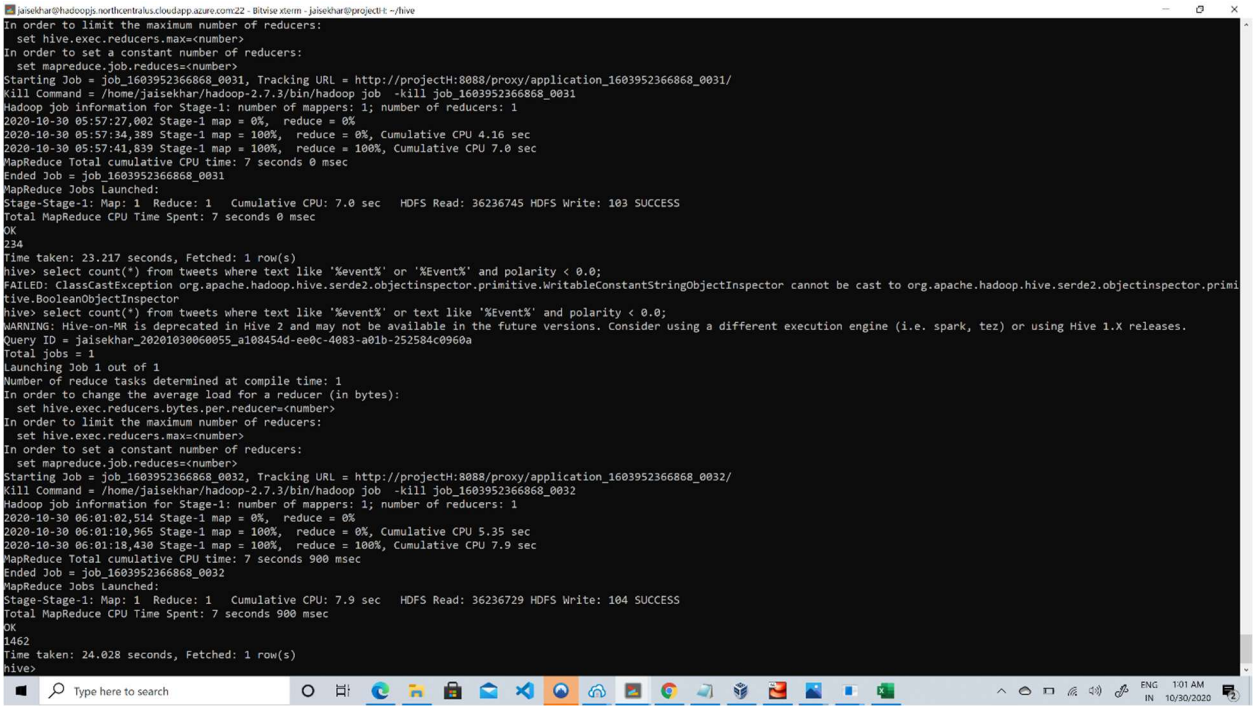
- Using the below query, we are fetching the user screennames and number of tweets tweeted by them in the descending order. In this dataset, TIN-Tech Bloggers has tweeted more about AAPL.



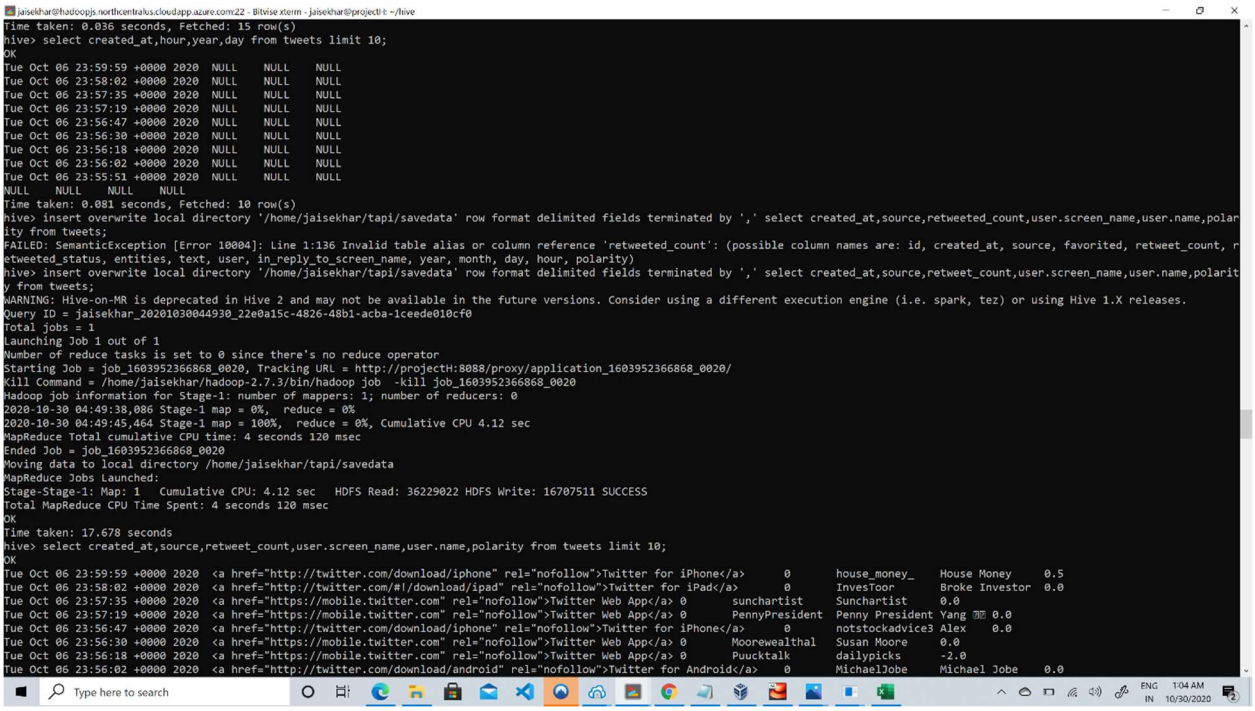
- In the below query, we are getting the number of positive tweets which contains the keyword ‘Iphone’. There are 234 positive tweets about Iphone.



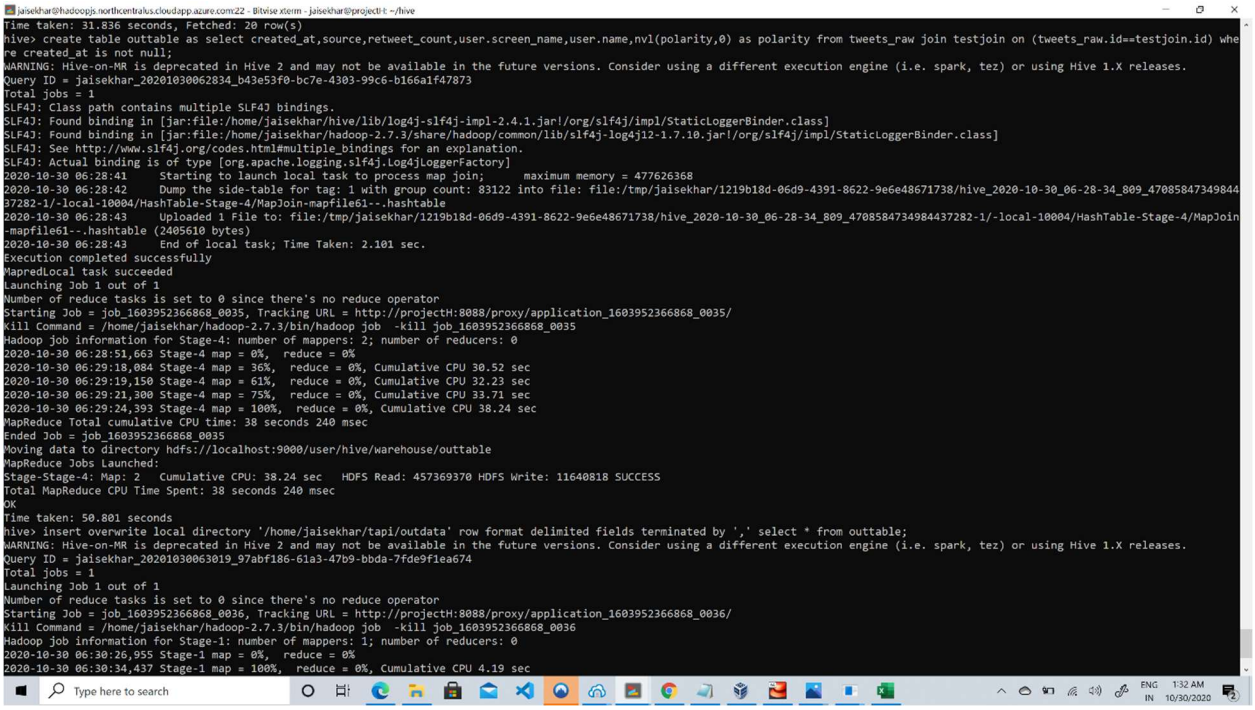
- Here, we are getting the number of negative tweets which contains the keyword ‘Event’ and there are 1462 tweets in this dataset.



- The below basic query returns the created time, source of the tweet, screenname, username and their polarities.

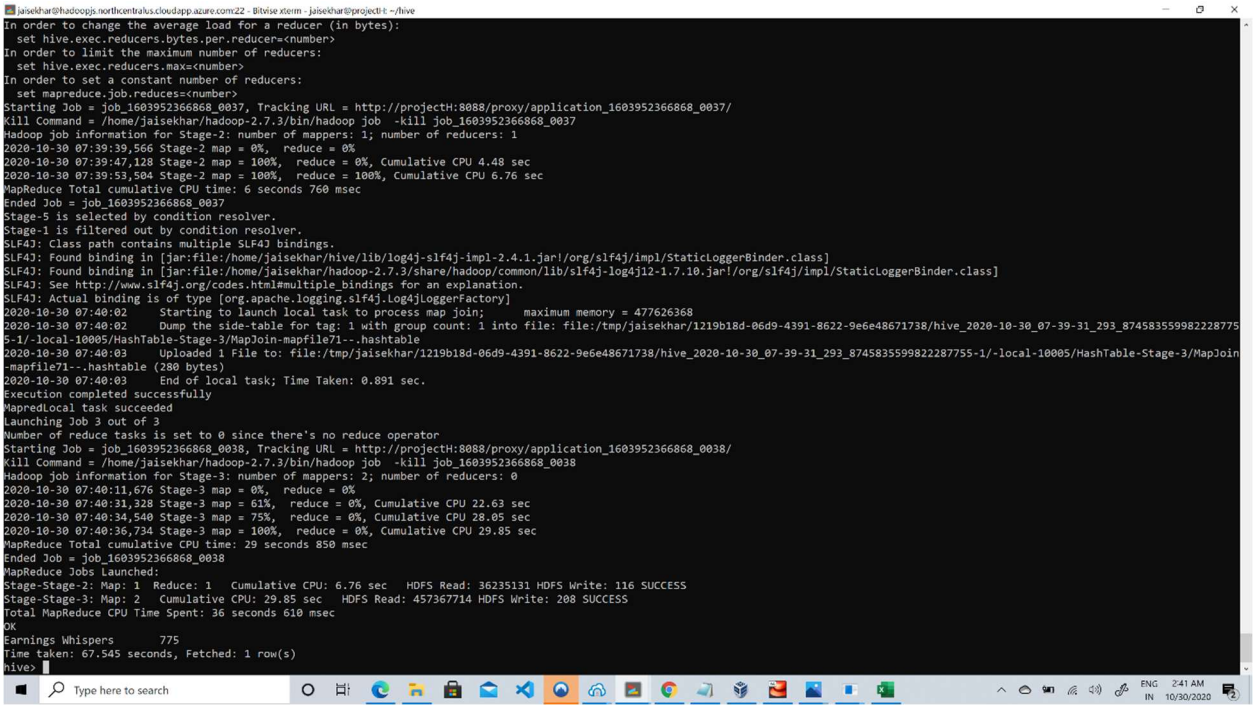


And the same data is stored in other csv file which helps in later part of analysis (Cassandra, visualization)

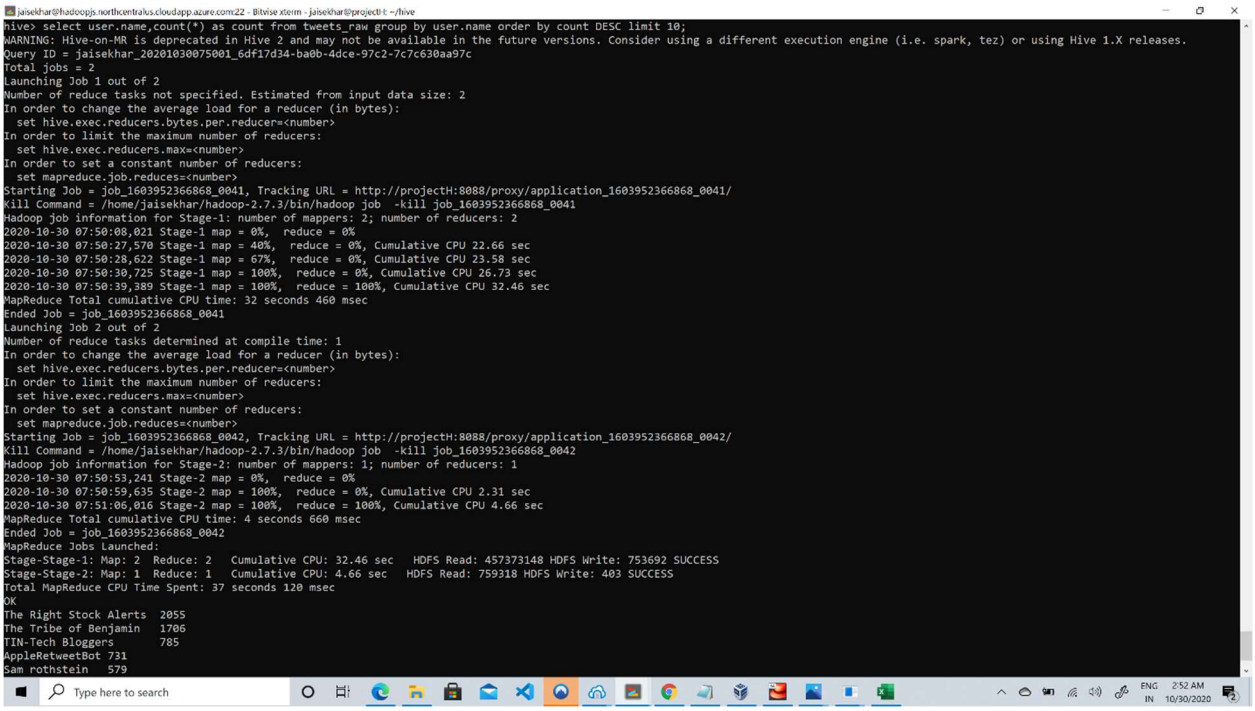


- The below query shows the user's name whose tweets has maximum number of retweet count. In this dataset, Earnings Whispers tweet has more tweet count.

```
hive> select user.name, retweet_count from tweets_raw where retweet_count in (select max(retweet_count) from tweets);
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution en
Query ID = jaisekhar_20201030073931_8dc5c22d-0da3-4921-8945-61633365e4e8
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1603952366868_0037, Tracking URL = http://projectH:8088/proxy/application_1603952366868_0037/
Kill Command = /home/jaisekhar/hadoop-2.7.3/bin/hadoop job  -kill job_1603952366868_0037
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2020-10-30 07:39:39,566 Stage-2 map = 0%,  reduce = 0%
2020-10-30 07:39:47,128 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 4.48 sec
2020-10-30 07:39:53,504 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 6.76 sec
MapReduce Total cumulative CPU time: 6 seconds 760 msec
Ended Job = job_1603952366868_0037
Stage-5 is selected by condition resolver.
Stage-1 is filtered out by condition resolver.
```

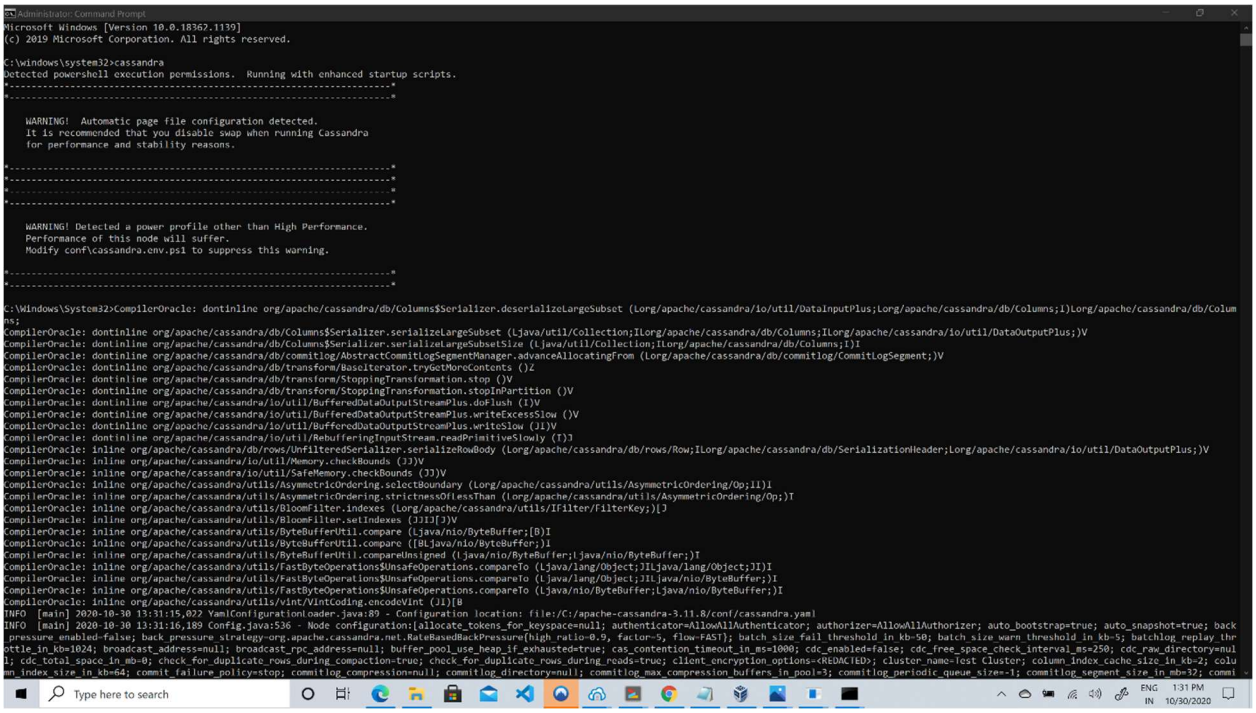


- Using this query, we are getting User's name and their total number of tweets in the descending order

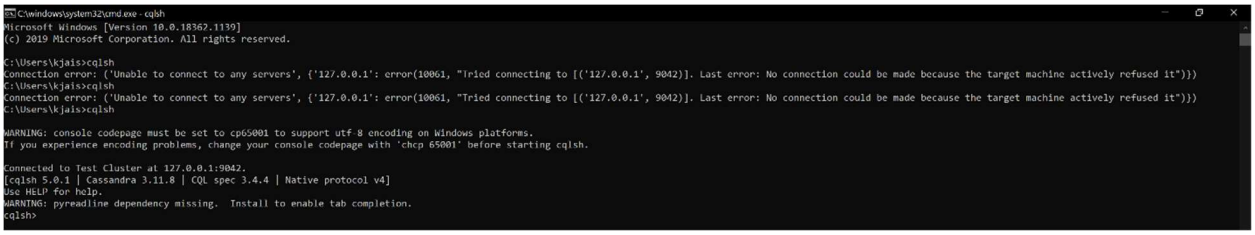


- ## Cassandra Analysis:

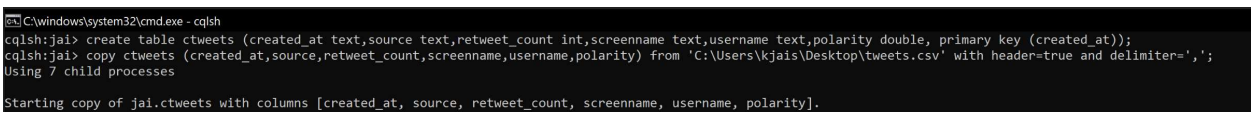
  Cassandra is started in command prompt administrator mode.



  Started cqlsh.



  Created table in Cassandra with required fields and copied the data from the tweets.csv into the ctweets table.



- The below query gives the count of number of most positive tweets which means the polarity values is more than 3.

```
cqlsh:jai> select count(*) as positive from ctweets where polarity > 3 allow filtering;

 positive
----------
      414

(1 rows)

Warnings :
Aggregation query used without partition key
```

- The below query gives the count of number of positive tweets whose polarity value is more than 0.

```
cqlsh:jai> select count(*) as positive from ctweets where polarity > 0 allow filtering;

 positive
----------
    12601

(1 rows)

Warnings :
Aggregation query used without partition key
```

- The below query gives the count of total negative tweets whose polarity value is less than 0.

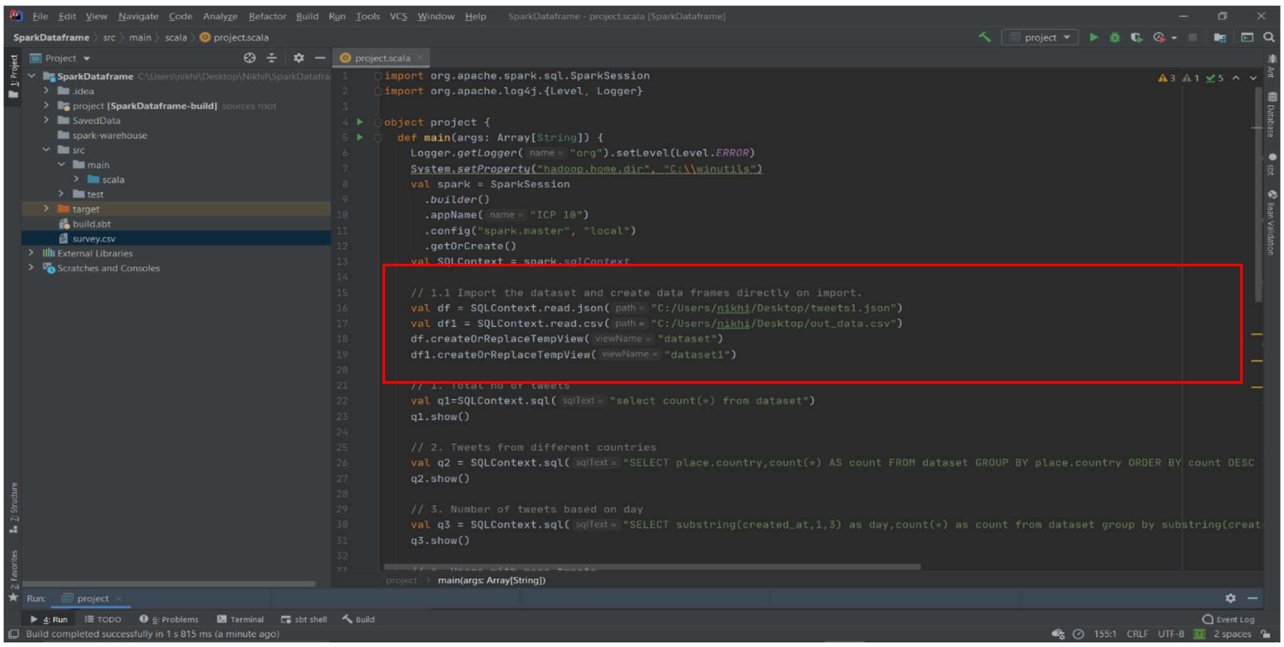
```
cqlsh:jai> select count(*) as negative from ctweets where polarity < 0 allow filtering;

 negative
----------
     4896

(1 rows)

Warnings :
Aggregation query used without partition key
```

- SPARK ANALYSIS

Below are the queries for analysis using spark.

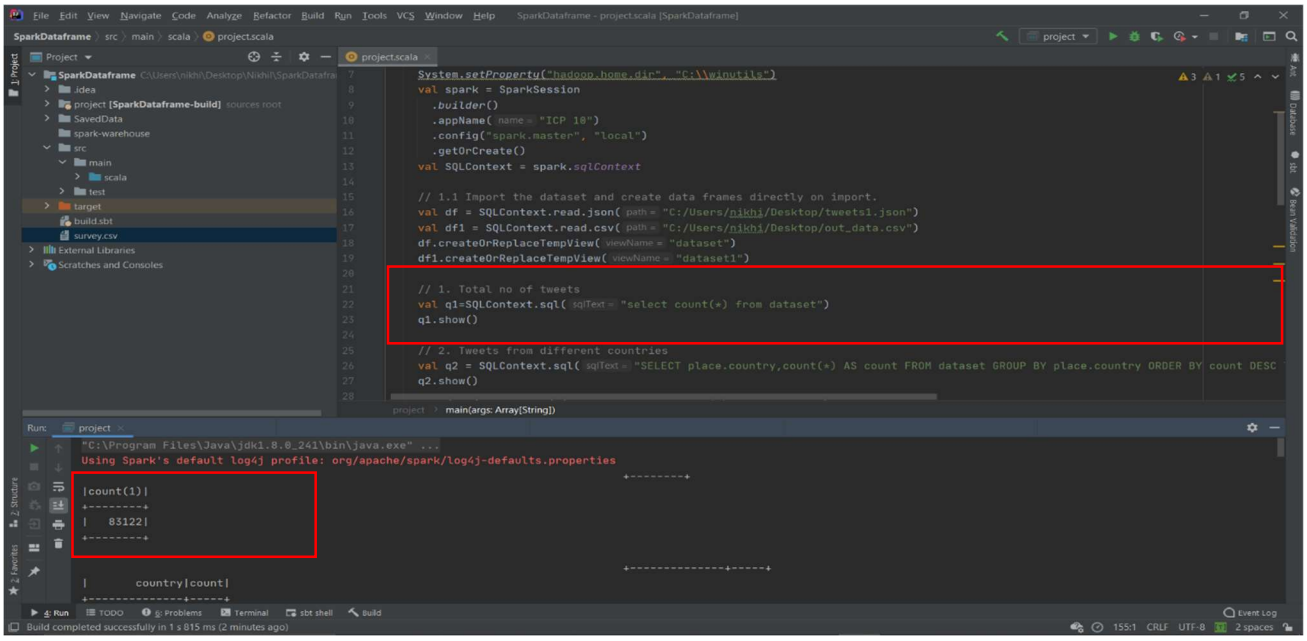
Loading the data

We created two data frames reading data from tweets1.json file which is the twitter data that is extracted and out_data.csv file is the data with polarity values.



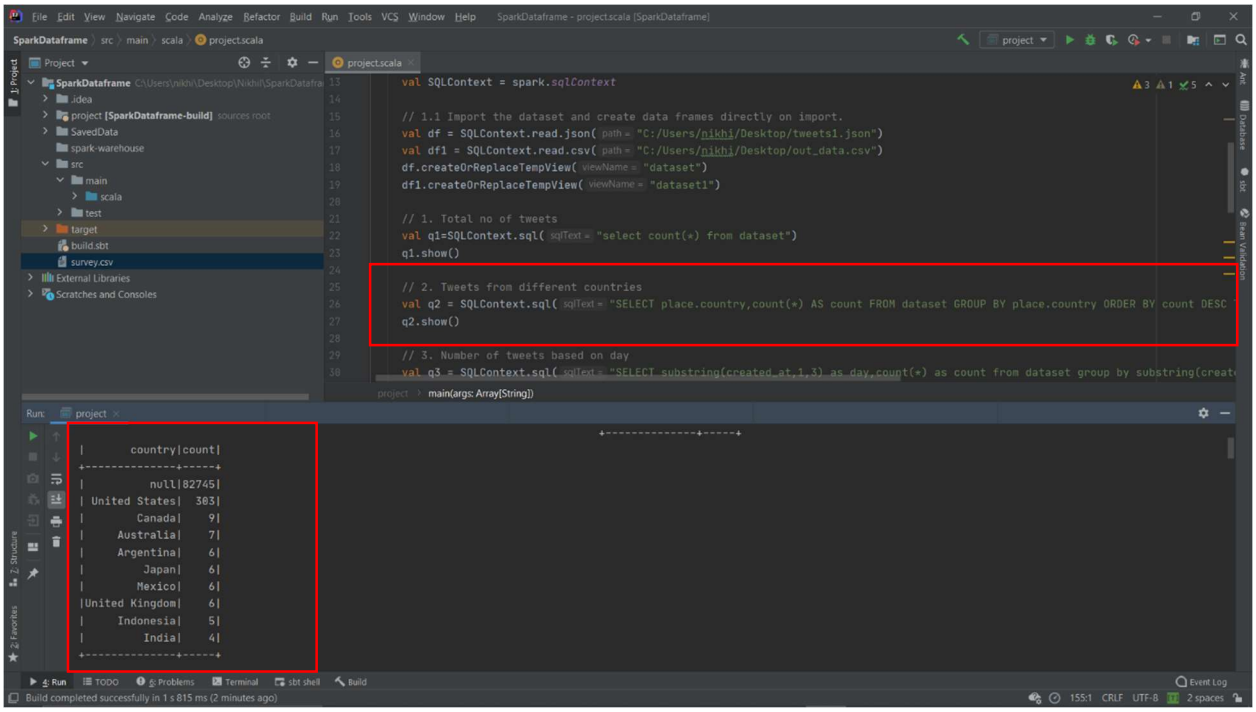
1. The below simple query shows number of records in the data frame dataset.

```
val q1=SQLContext.sql("select count(*) from dataset")
q1.show()
```



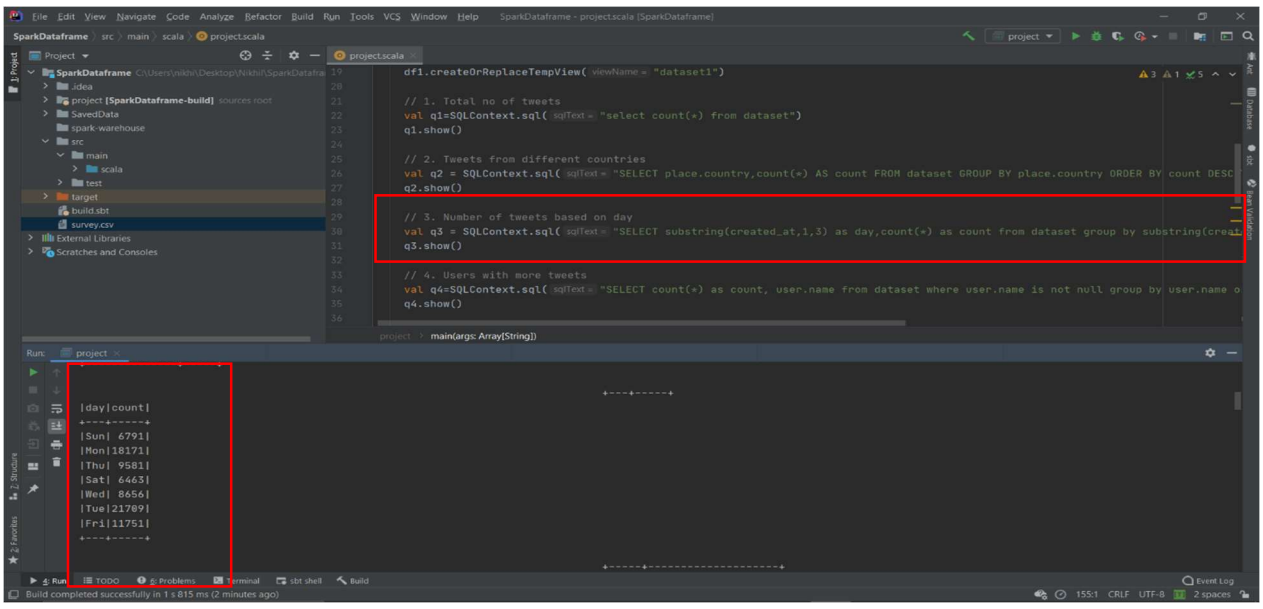
2. The query displays the tweets count from different countries

```
val q2 = SQLContext.sql("SELECT place.country,count(*) AS count FROM dataset GROUP BY place.country ORDER BY count DESC limit 10")
q2.show()
```



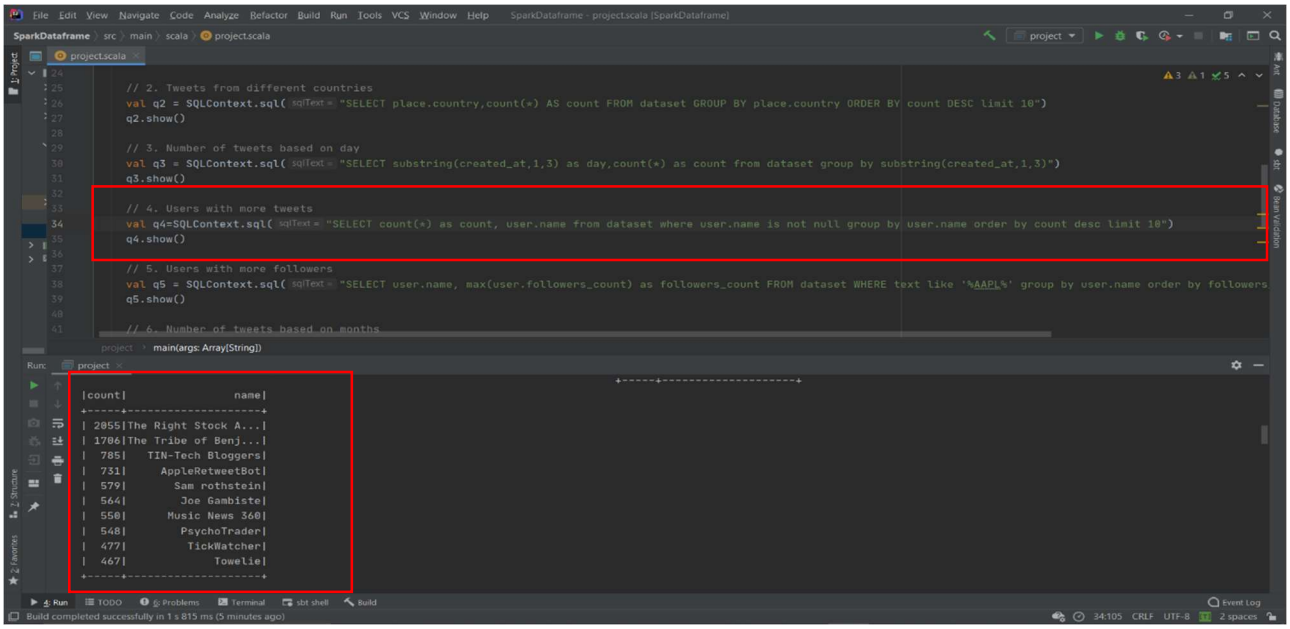
3. The query displays the count of tweets based on a day

```
val q3 = SQLContext.sql("SELECT substring(created_at,1,3) as day,count(*) as count from dataset group by substring(created_at,1,3)")
q3.show()
```



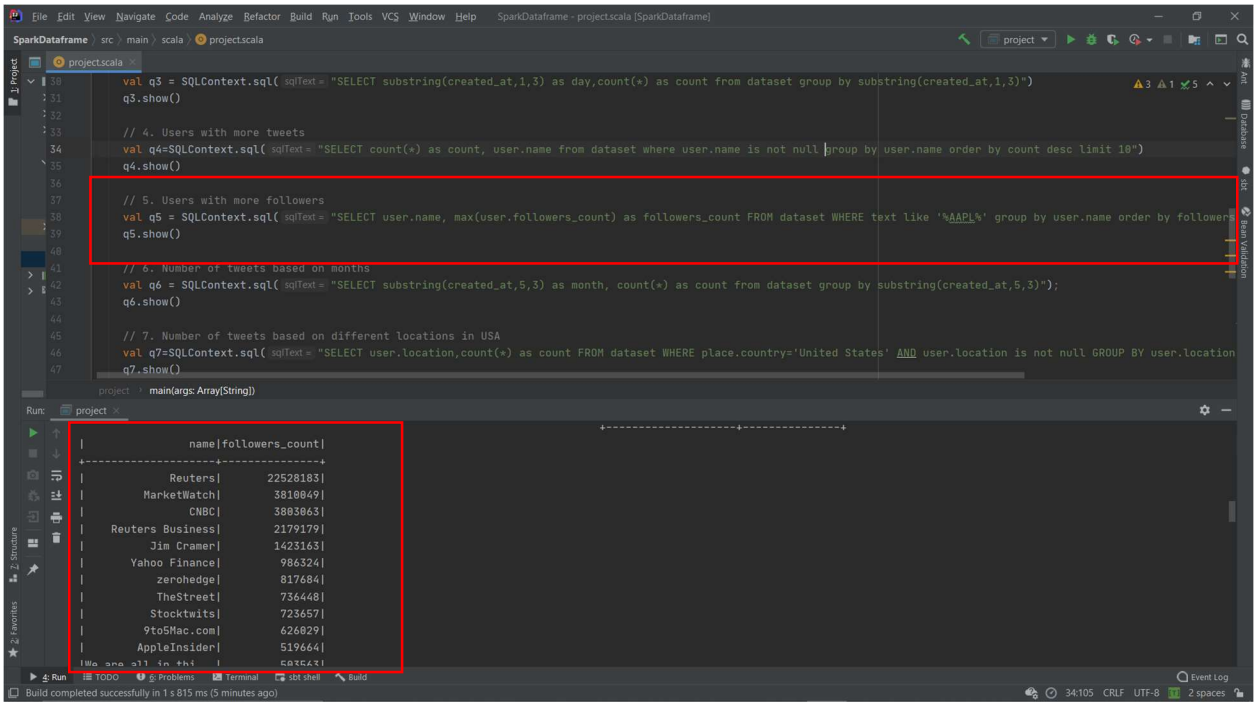
4. The query displays the user with more number of tweets

```
val q4=SQLContext.sql("SELECT count(*) as count, user.name from dataset where user.name is not null group by user.name order by count desc limit 10")
q4.show()
```



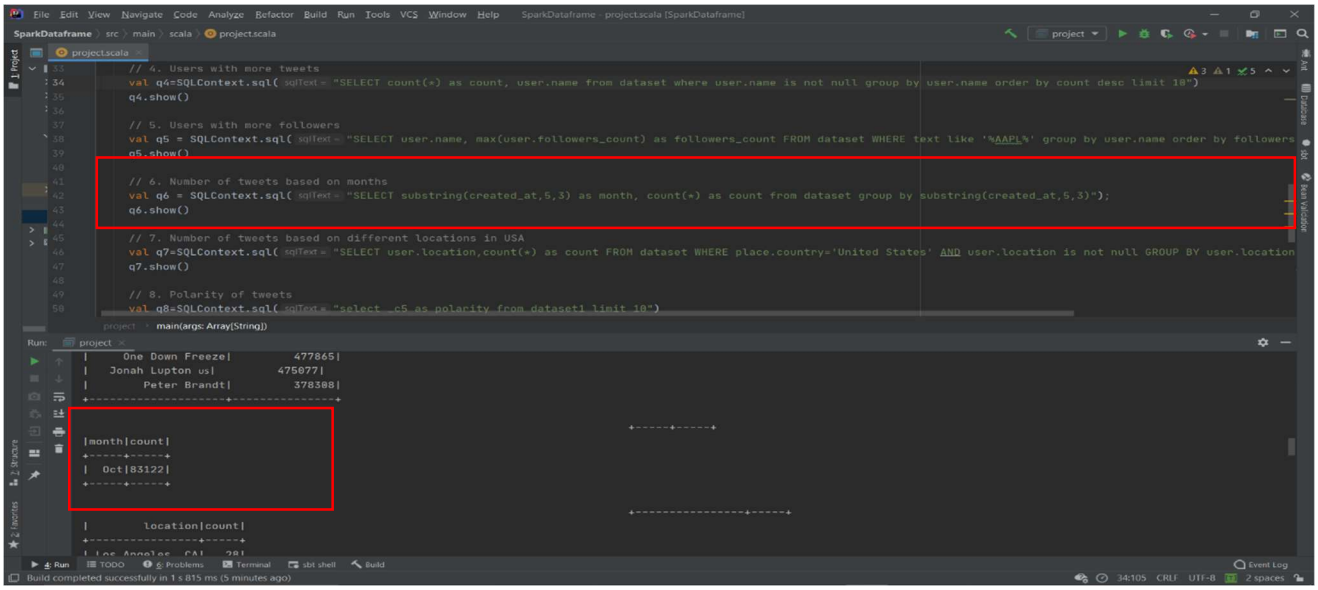
5. The query displays the users with more follower count

```
val q5 = SQLContext.sql("SELECT user.name, max(user.followers_count) as followers_count FROM dataset WHERE text like '%AAPL%' group by user.name order by followers_count desc limit 15");
q5.show()
```



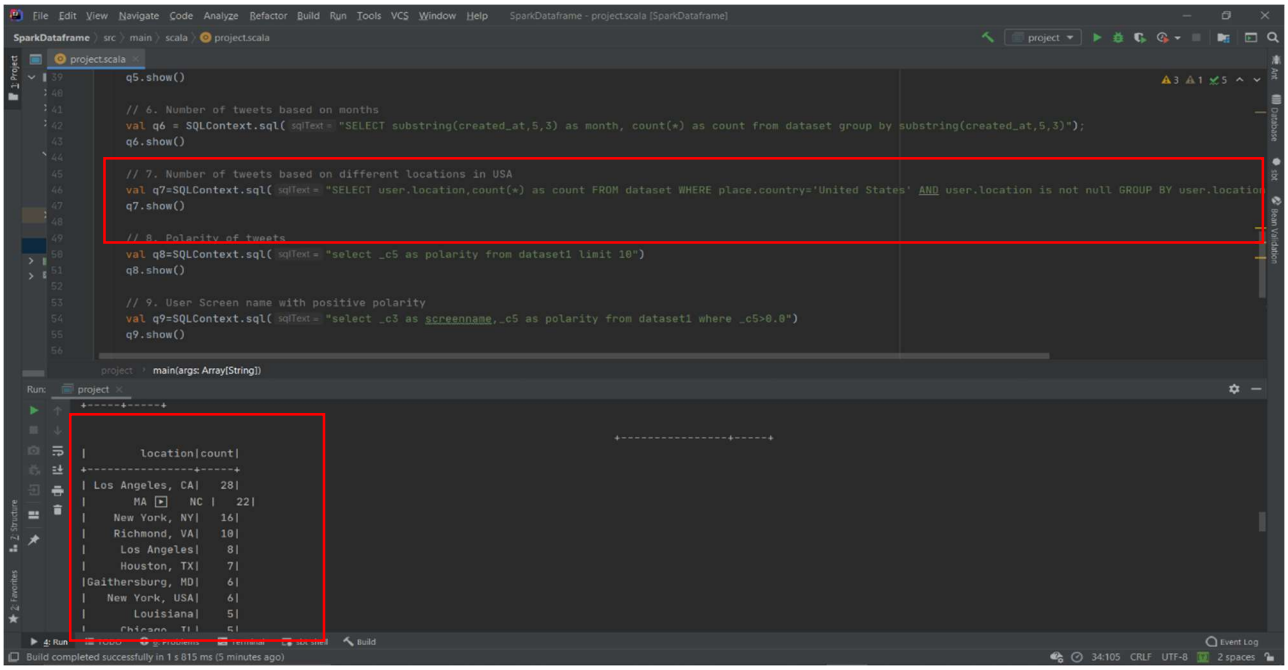
6. The query displays the tweet count based on months as we extracted the tweets only from October month only that month tweet count is shown.

```
val q6 = SQLContext.sql("SELECT substring(created_at,5,3) as month, count(*) as count from dataset group by substring(created_at,5,3)");
q6.show()
```



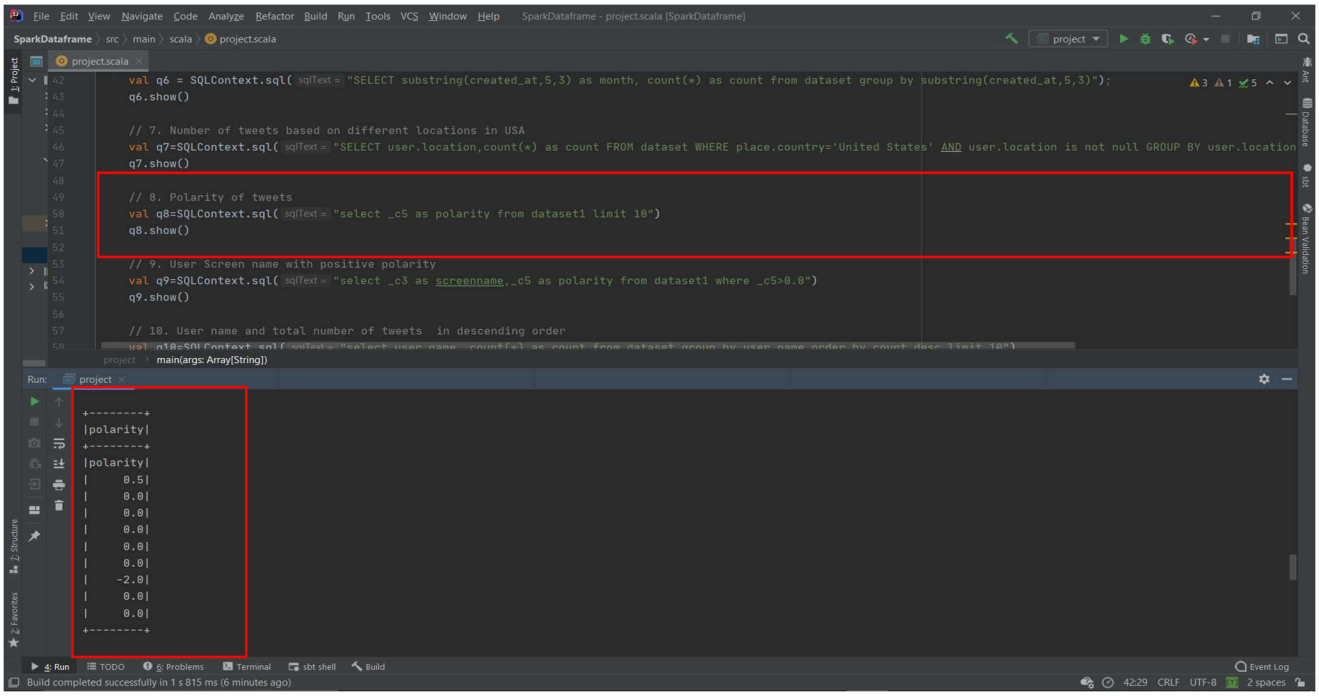
7. The query displays the tweets from different locations in USA

```
val q7=SQLContext.sql("SELECT user.location,count(*) as count FROM dataset WHERE place.country='United States' AND user.location is not null GROUP BY user.location ORDER BY count DESC LIMIT 15");
q7.show()
```



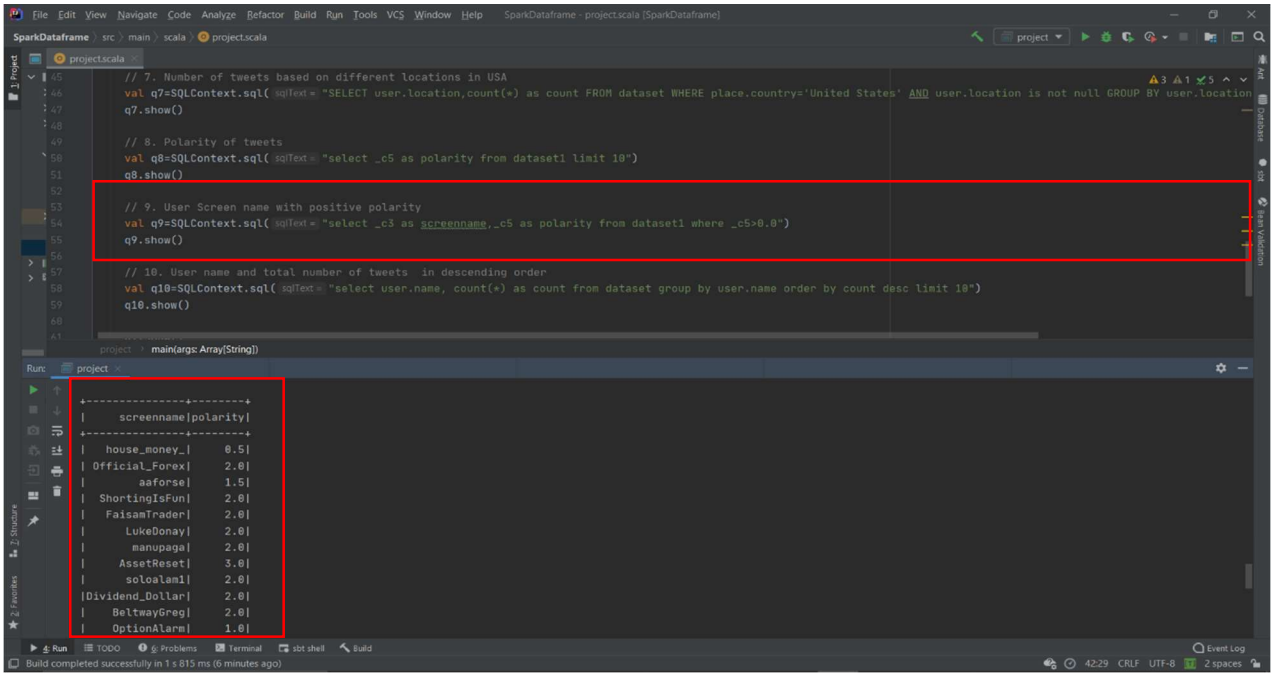
8. The query displays the polarity value of top 10 tweets

```
val q8=SQLContext.sql("select _c5 as polarity from dataset1 limit 10")
q8.show()
```



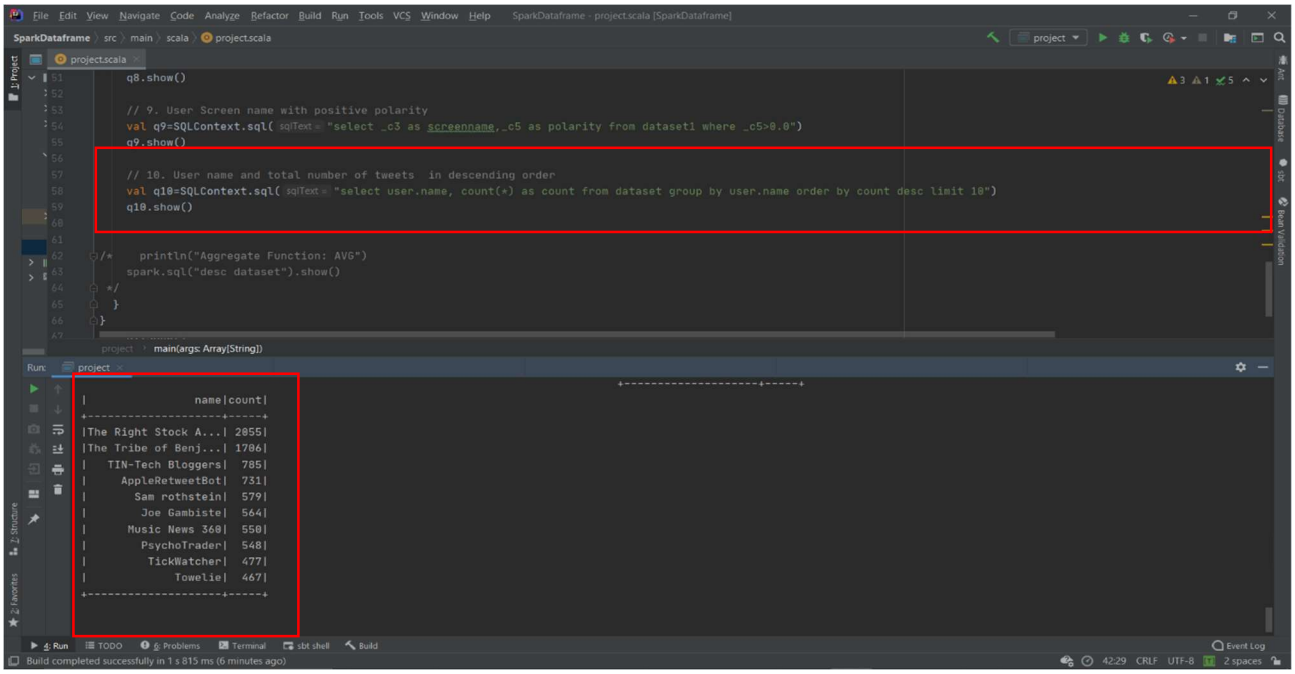
9. The query displays the tweets with users screen name having positive polarity

```
val q9=SQLContext.sql("select _c3 as screenname,_c5 as polarity from dataset1 where _c5>0.0")
q9.show()
```



10. The query displays the data with name and number of tweets they tweeted

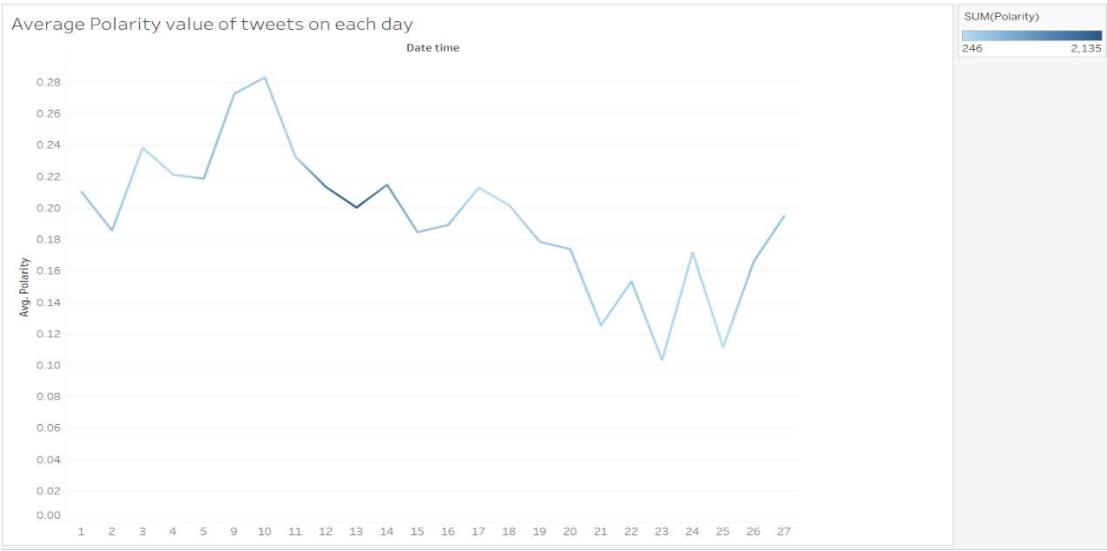
```
val q10=SQLContext.sql("select user.name, count(*) as count from dataset group by user.name order by count desc limit 10")
q10.show()
```



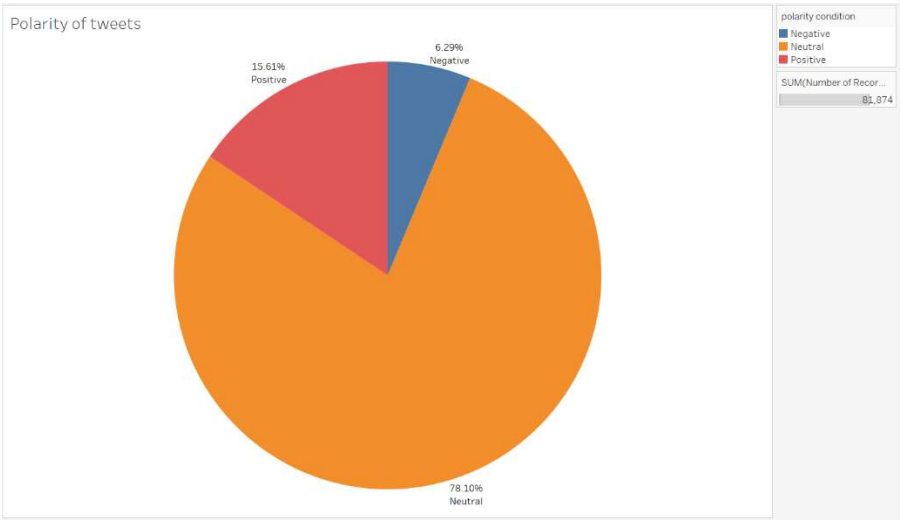
Visualization & Result Evaluation:

We used Tableau for the visualization

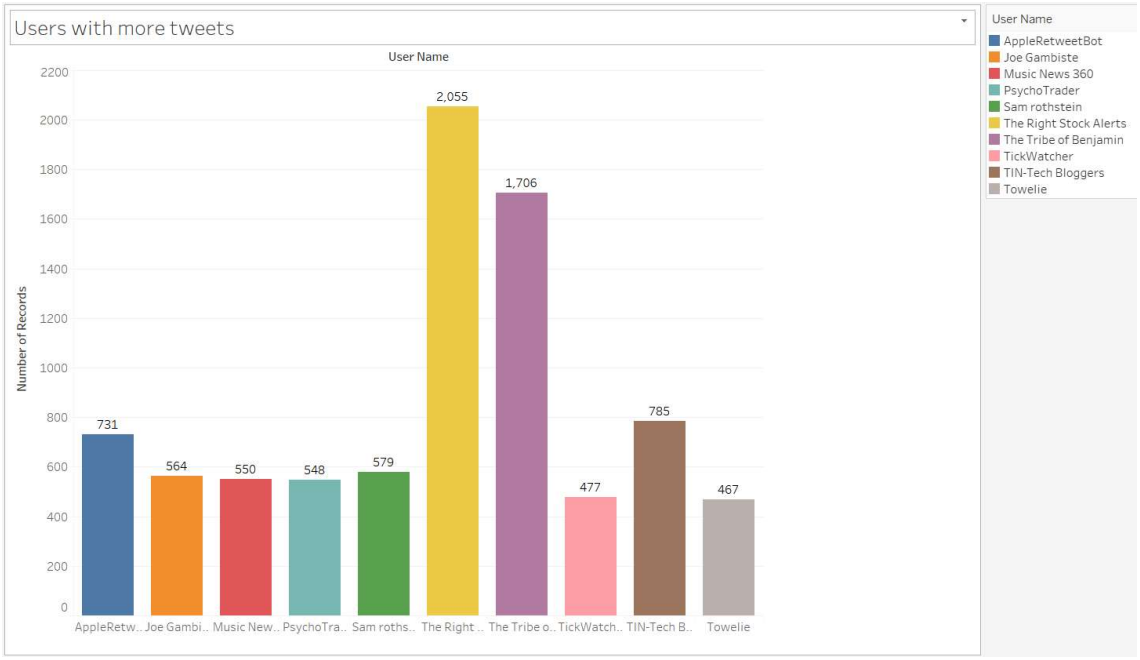
- The Average polarity of tweets based on keyword **AAPL** on each day of October month. From October 5th to 10th there is a gradual increase in the polarity value. The highest average polarity value is recorded on October 10th. From 10th of the month polarity value gradually decreases.



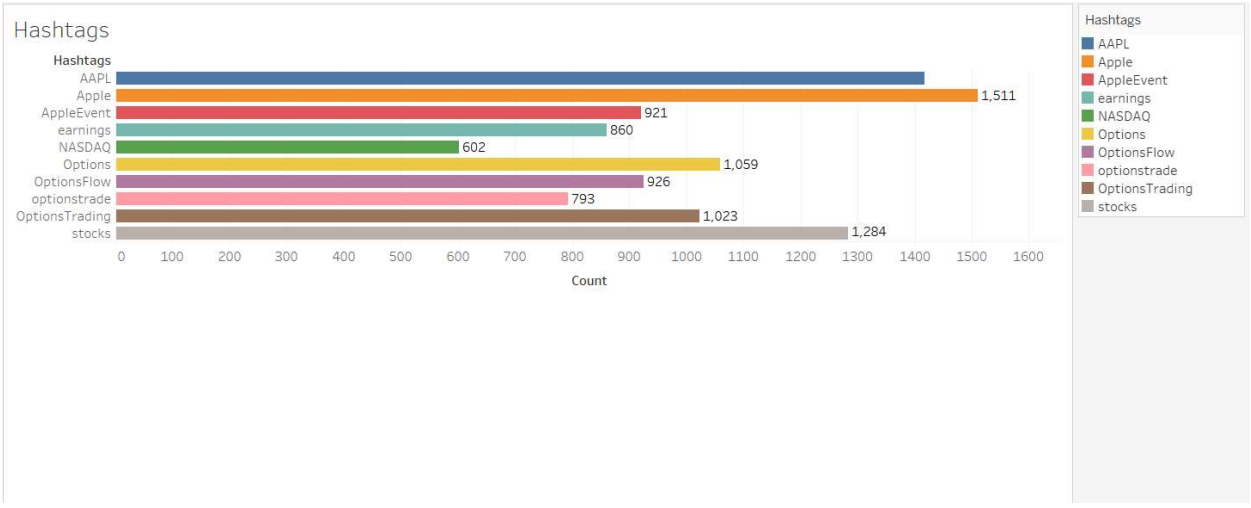
- Polarity value of each tweet is calculated by using sentiment analysis on each tweet. We get values in the rage between -5 to +5. If the polarity value lies between -5 and 0 it is Negative. If the polarity value lies between 0 and 5 it is Positive. If the polarity value is 0 and 1 it is Neutral.
  The data shows 78% tweets are neutral and remaining 22% of tweets are positive and negative.



- Data shows the top 10 users with a greater number of tweets in the month of October. User with username **The Right Stock Alerts** has more tweets on AAPL stock in the month of October.



- Data shows that Apple is the hashtag that has highest wordcount in the month of October

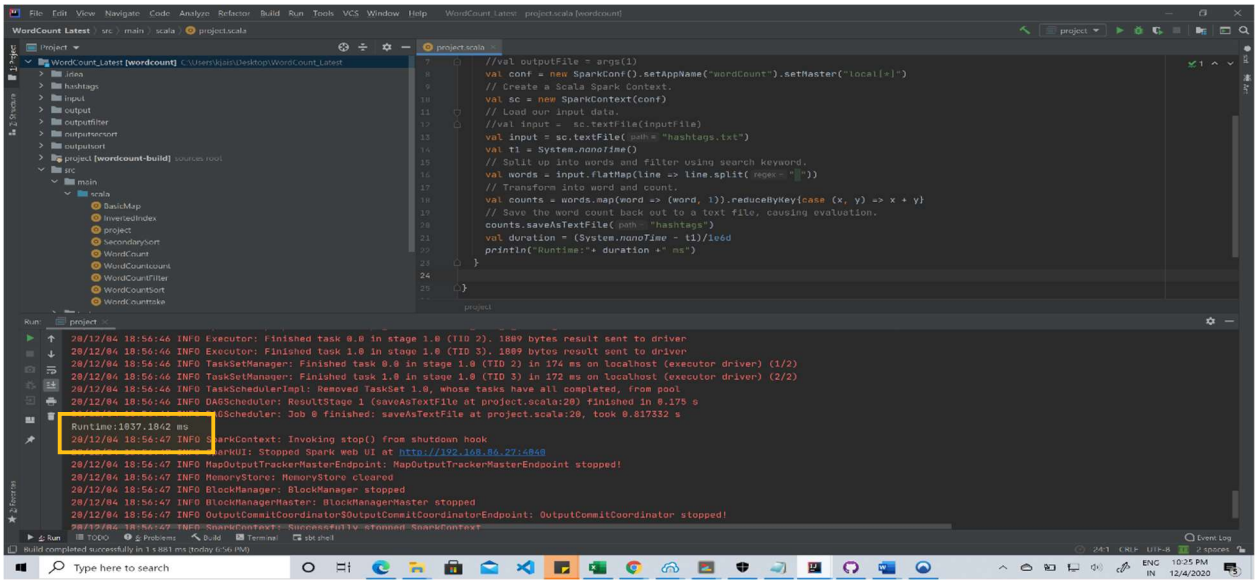


Technical Analysis

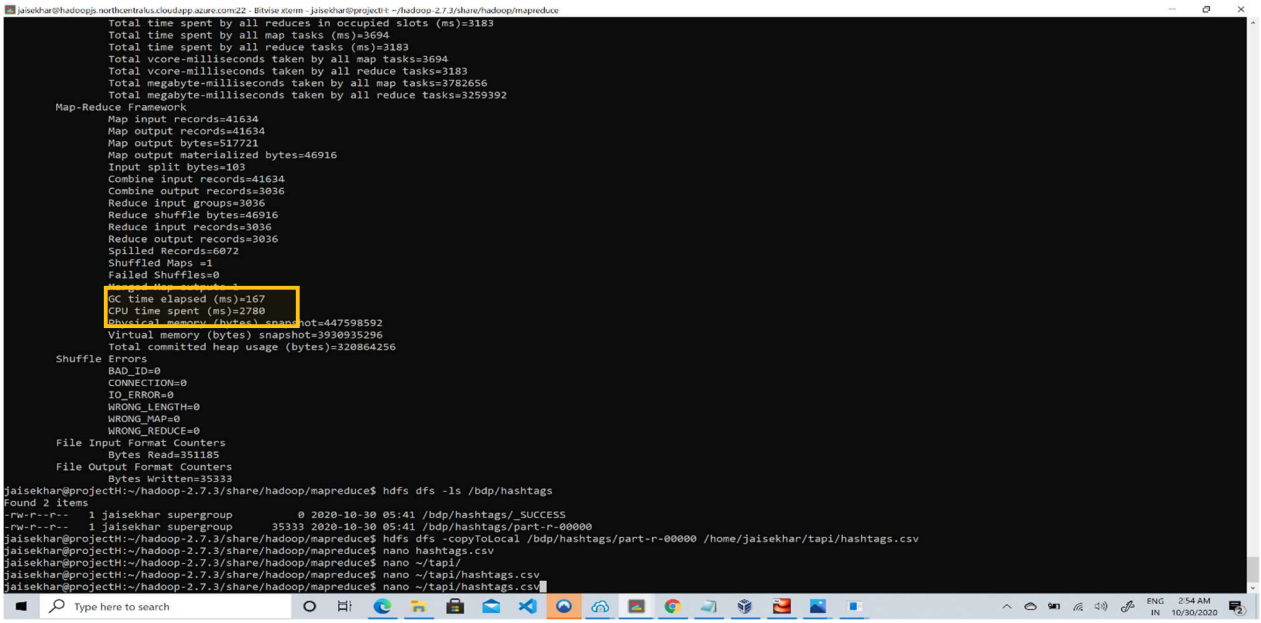
MapReduce vs Apache Spark

In the below we ran the wordcount program in both apache spark and mapreduce. The runtime for the wordcount in apache spark is 1037 ms and runtime in Hadoop mapreduce is 2790 ms (It can be more). By this we can say Apache Spark is way faster than Mapreduce in means of processing the data.

Apache Spark:



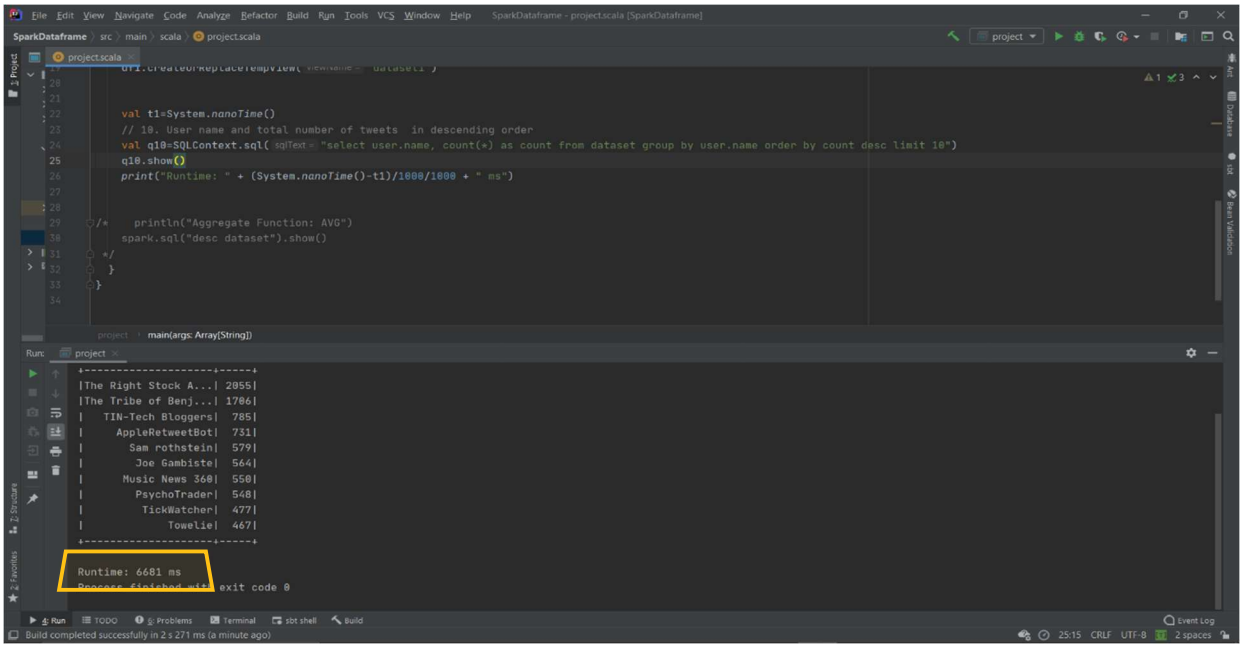
Hadoop Mapreduce:



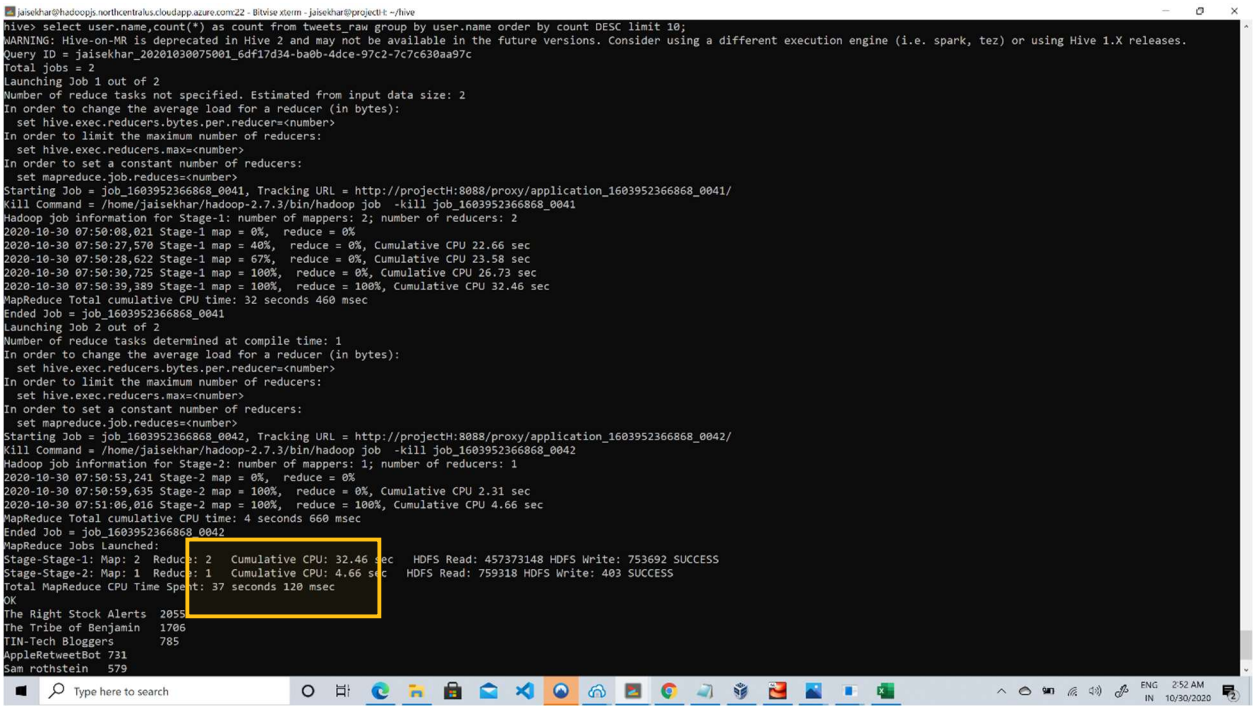
Hive vs Apache Spark

In the below snippet the comparison was drawn between processing time of Hive and Apache Spark using QL. Apache spark took only 6681 ms where as Hive took 37000 ms. By this we can say Apache spark is very faster than Hive.

Apache Spark:



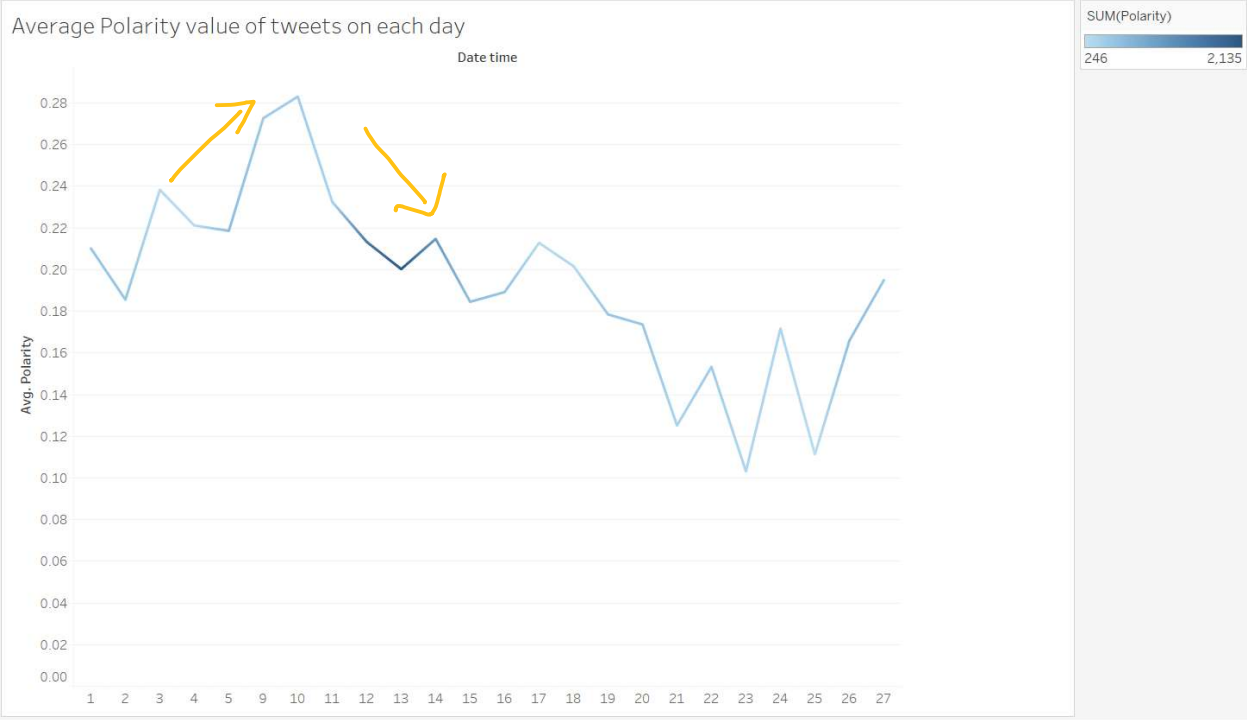
Hive:



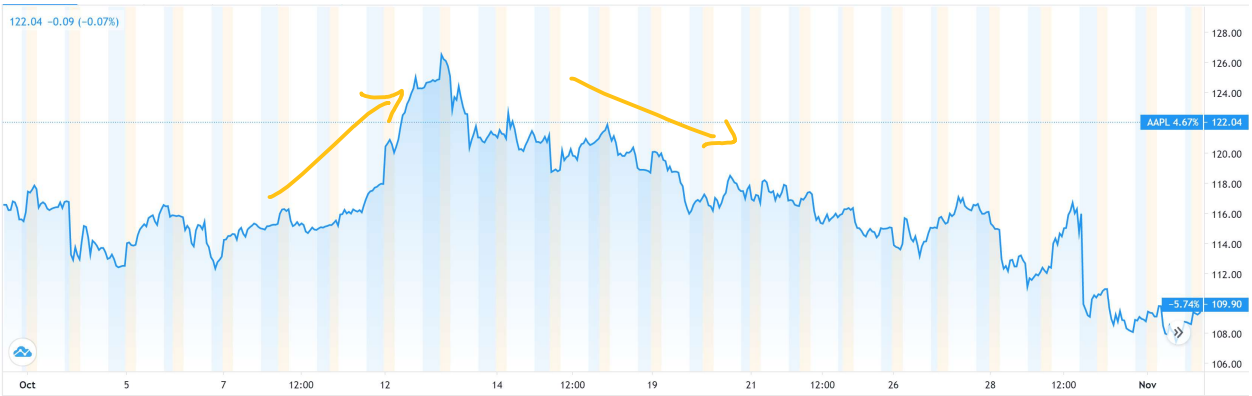
Stock Analysis: (Important)

The below two graphical representations are the trends obtained from analysis and other is the real stock market trend of the ticker AAPL stock in the month of October. Here, we can clearly see the identical trend between both the real and analytical trend which gives early signs for the investment. By this we can say that stock can be predictable with this system.

Our System: (Tableau Representation)



Real Stock: (Trading View)



Conclusion

Using this system, we successfully analyzed the data extracted from the twitter social opinions based on stock ticker with the help of hadoop ecosystem. In the above results, we able to get the trend after analysis which can be used for predictions of movement of stock. This trend might help for the investors in future investments. These trends are giving early signs for the investments which might help to gain the profits. Right now, this system might be only for one stock and for one month but it can be extended for future usge. Thus, we can say that we successfully achieved our goal using this system.

Future work

In this system, we analysed only AAPL stock ticker, which we extracted based on this keyword from the twitter for one month. In future, we can extend this range of extraction on keywords on time frames accordingly which might be dynamic in nature. Since our system is based on historical data but we can also try to analyze the live data with proper models.

Project Management

- ### Work completed:
    1. Extracted tweets using python and twitter API for the dataset.
    2. Loading the data into the HDFS.
    3. Sentiment Analysis done on every tweet in the dataset.
    4. Wordcount on hashtags using mapreduce.
    5. Creation of table and loading of data into Hive.
    6. Analysis of data using Hive
    7. Analysis of data using Cassandra
    8. Visualization using Tableau.
    9. Initialization and loading the data into Apache Spark.
    10. Analysis of data using SparkSQL
    11. Technical analysis by comparing Apache Spark, Mapreduce and Hive
    12. Final comparison between our system trend and real trend of stock (success)

- ### Contribution:

    Jaisekhar Koya : 25%

    Extracting tweets using Python. Done sentimental analysis using AFINN dictionary which gives polarity value. Extracted tweets were stored in HDFS. Created and loaded data to hive. Performed few queries in Hive and also Cassandra. Drawn the comparison between Apache Spark and Mapreduce. Contributed in drawing the comparison between trends of system and real.

Sri Sai Nikhil Kantipudi: 25%

Helped in doing sentimental analysis on every tweet in the dataset. Done wordcount on extracted hashtags using MapReduce. Contributed in loading and analysing the data in Hive. Performed few queries in Hive and also in Cassandra. Analyzed the data using Spark SQL. Also contributed to work for comparing system vs real trends.

Sai Rohith Guntupally: 25%

Tweets extracted and parsed for hashtags using python. Done hdfs operations on the extracted dataset. Performed few queries on the dataset using hive and Cassandra. Part of visualization in tableau. Wrote few queries in apache spark. Drawn the comparison between Apache Spark and Hive.

Aarthi Nagireddy: 25%

Sentiment Analysis using dictionary system. Hive initialization and loading of data of the dataset. Performed few queries in Hive and Cassandra for data analysis. Part of Visualization work in tableau. Initialization and loading the data into Apache Spark. Wrote few queries in apache spark.

Story Approach

Chapter – 1: Life

- **Who** are the people or communities in need of help?

  Investors who play a lot with stock market will get helped with this system which makes their stock selection and timing effective.

- **What** problem happened to them?

  Half of the investors are failing in stock selection with ineffective analyses.

- **When** did the problem take place?

  During investing in stock market on particular stocks by investors many people will loose their investment because of inaccurate approaches.

- **Where** means two things: 1) The environment and settings that the people or the community is living in, and 2) the place/location where the problem take place.

  Only in stock market this problem exists. There were different kinds of stock trading's. But, when it comes to the market we got only one stock market with different exchanges.

- **Why** means the possible causes and/or origin of the problem.

  People always wants to earn or save their money by buying shares in good company. But their selection of company or stock might be bad because of their process of analyses.

- **How:** If you would like, you can add a dimension of how. How did it happen? Sometimes, the answer to how can be covered by what, when and where.

  For example, if we take AAPL stock which belongs to apple. X likes the apple products so wants to invest in the company. So X bought the shares of AAPL stock. But, what X don't know is, at the same time Apple facing BendGate issue and stock just decreasing which bought loss to X. This system helps in analysing the user opinions and helps in selection.

Chapter -2: Data

- **Who** is the data set about? **Who** were sampled in this data set? **Who** were over sampled or under sampled? Are they representative of the main characters in Assignment 1? Is there any identifiable information or is there any risk of disclose identifiable information? This is fundamentally about the sampling issue, and anonymity.

  The dataset is about the users who tweeted with the keyword 'AAPL' (in this case) in the twitter. There are many numbers of users who tweets in twitter. Thus, it help us in analyzing more data. This dataset contains only public available information . It doesn't contain any undisclosable information. These users are the main representatives stated in the assignment 1.

- **What** events, activities, behaviors, and observations etc. are recorded by the data set? Does the data set record the targeted events, activities, behaviors, etc. in Assignment 1? This is fundamentally about the variables.

  The dataset contains the tweets extracted from the twitter which contains user information, tweet text, hashtags, created time and their entity information. This are the important variables used in the project. This covers all the required variables stated in the assignment 1.

- **When** did the event, activity, behavior, and observation, etc. take place? **When** were the data collected? Is it longitudinal or cross-sectional? Are they real time data? How old or fresh are the data? To what extent generalization can be made across time to inform Assignment 1? This is fundamentally about the temporal structure of the data set, and the external validity of the data set across time.

The data is generated when the user tweets about AAPL stock in the twitter. There will be a greater number of users who tweets same. The data in out dataset consists of all the tweets which contains AAPL keyword in Oct 1st to Oct 28th. Hence this is the old data not the real time data. This is permanent data which might need to extract more data in future for further analysis of other data range. It satisfied the abstract stated in assignment 1.

- **Where** did the event, activity, behavior, and observation, etc. take place? **Where** were the data collected if the information is available? What does the geographical coverage of the data set look like? Does the data set contain geographical information (GIS)? Is this a local, regional, national, or global data set? To what extent generalization can be made across settings to inform Assignment 1? This is fundamentally about geographic variables in the data set, and the external validity of the data set across settings.

The dataset consists of tweets which are extracted from twitter. Twitter consists of huge amount of data. It is an online social media platform. It has more number of users. This data set contains geographical information of the tweeted location. This dataset doesn't have geographical limitations. This contains the data as stated in Assignment 1.

- Why did the event, activity, behavior, or observation etc. take place? Why were the data collected?

Twitter is one of the important sources for the data. Many number of people puts their opinion in the twitter. In the same way, many companies or investors puts their news and opinions in twitter which effects in stock trend. Hence we are using their opinions(tweets) to analyze the stock movement.

- **How:** If you would like, you can add a dimension of how. How did it happen? Sometimes, the answer to how can be covered by what, when and where.

Most of this part was already covered above. But to be precise, we are collecting the tweets from twitter and doing sentiment analysis to find whether their opinion is positive or negative and do the analysis on the whole data using hadoop ecosystem.

- Calibration 1: Every aspect covered in Assignment 1 aligns with the Assignment 2. We have choose right dataset in Assignment 2 for analysis as mentioned in the Assignment 1. And, this is the relevant dataset to the project.
- Calibration 2: Experimented different models for our system using various NLP techniques. Aligned the data according to the process of using NLP techniques which is sentiment analysis. Applied accurately on the parsed data from twitter dataset.

Chapter -3: The Scientist and AI

- **Who:** In this story, there are three main characters: 1) the people/the community who needs help, 2) the data scientist (that is you), and 3) AI. How much does the data scientist understand Assignment 1 (domain) and Assignment 2 (data)?

The student/data scientist will understand the domain chapter as it will resemble the complete idea of the project which needs to done can also be considered as requirements technically. Domain knowledge is required to develop the project in later stages. Also, after getting the information related to the domain the data scientist needs to work on the data to process and analyze based on the requirements which stated in the chapter 2 Data.

- **What** models and analysis did the data scientist and AI apply to fulfill the need of the people or the community?
    - Can the data scientist estimate and select data for their goals from Assignment 1? Can they map data sets from Assignment 2 onto appropriate ML models?
    - Can the data scientist connect Story 1 with ML models/stories about what a ML model can do? To perform good ML research, what in-depth knowledge and experience with ML algorithms and ML stories does a data scientist need?

Data scientist selects the data according to the specified goals in the chapter 1. Using the same data, data scientist will gather the data as specified in the goals of chapter 2. Then works on the data using the domain knowledge. In this specific project, the data scientist needs the have knowledge about the hadoop ecosystem.

- **When** has to do with the iterations (Calibration 2). How much time did it take for experimentation? How efficient is the modeling/algorithm? Can the data scientist determine the acceptance level of the model (validation with accuracy and runtime performance) considering the targeted users?

Various technologies were used in this project. For every analysis done using one technology in hadoop ecosystem we will be having different run time performances and accurancies. In this project we dealt only NLP technique which is sentimental analysis and data analysis with different technologies in hadoop ecosystem.

- **Where** has to do with the learning environment. Where did this experiential learning process take place? For example, it was part of an online Deep Learning course.

It was part of the Big Data Programming course offered in University of Missouri-Kansas City and also few online courses for NLP Techniques.

- **How**: If you would like, you can add a dimension of how. How did it happen? Sometimes, the answer to how can be covered by what, when and where.

The data scientist can use data in the way it is required for the analysis or based on the tool. New data or data source can be added for further analysis as per required nothing static in this process.

Chapter 4: Users

- **Who**: the main character is the targeted user or audience?

The project is targeting the investors who is involved in the stock market. The analysis comes from this project will help them in picking the right stocks.

- **What** can the application do? What does the visualization show?

The analysis shows the one month analysis on the stock ticker's data collected from the twitter. It provides one month movement of stock using the social opinions by using different tools in hadoop ecosystem.

- **When** can the user use the application/visualization?

The investor is user here. The investor uses this analysis or visualization during the investment in stock market.

- **Where** will the visualization and applications be deployed, for example, mobile phones, the web, or IoT devices?

The visualization as of now is deployed in the tableau using the data retrieved after the analysis in the hadoop ecosystem

- **Why** is the visualization or application useful to the user?

This analysis helps the investor to pick the right stock to invest in the stock market and gain profits based on its historical movements and social opinions.

- **How** will the people/the community use this application or visualization to make changes?

The investor uses this analysis or visualization for any stock. They can change the stock ticker according to their requirement in the future.

Chapter 5: The society

- **Who** will be impacted? **Who** were sampled? **Who** were over sampled or under sampled? **Who** were the data scientists (yourself and your collaborators)?

The investor will get impacted using system. They get right stocks to pick after the results of the analysis. The data scientists here are the student who performs the analysis using the hadoop ecosystem. They just get benefited.

- **What** are the social and cultural impacts? **What** are the concerns about data privacy, security, and fairness?

The investors only get impacted as stated above. We are using the data which is publicly available only. We are not dealing with any impacting privacy or security or fairness of data.

- **When** will the social and cultural impacts take place? **When** should people be concerned about data privacy, security, and fairness?

During the investment in the stock market the investors might use this application which might impact in their investment. We are using the data which is publicly available only. We are not dealing with any impacting privacy or security or fairness of data.

- **Where** will the social and cultural impacts take place? **Where** will data privacy, security, and fairness issues, like data breach, and evaluative bias, likely to happen?

The impacts will take place during the investment in the stock market. It wont breach the data privacy not the security nor the fairness issues. Even, this system is not evaluative bias.

- **Why** are the social and cultural impacts important or consequential to the people and/or the community? **Why** should we be concerned about data privacy, security, and fairness issues?

This is very important in investors community. Because their daily routine is to pick the right stocks and do the investment. This system gives the analysis about the stock which helps them in the picking the right stock. It won't breach the data privacy not the security nor the fairness issues. Even, this system is not evaluative bias.

- **How** can we address these societal issues in ML using a community-in-the-loop approach?

Right now, this system is using NLP techniques to process the data which is sentimental analysis on the data collected from the twitter. This helps to retrieve the social opinions on large scale with out any biased in the results.

References

- ANALYSIS AND PREDICTION OF STOCK MARKET USING BIG DATA TECHNOLOGY https://www.academia.edu/36679481/ANALYSIS_AND_PREDICTION_OF_STOCK_MARKET_USING_BIG_DATA_TECHNOLOGY
- Stock Market Prediction on Bigdata Using Machine Learning Algorithm http://ijesc.org/upload/b91a9a994c4d79a72f5542393ca9469d.Stock%20Market%20Prediction%20on%20Bigdata%20Using%20Machine%20Learning%20Algorithm.pdf

GITHUB Link:
https://github.com/jaisekhar/Stock-Market-Analysis-using-Hadoop-Ecosystem