Stock Market Prediction Using Twitter Data 🐦➔📶

Team members:

1.	Jayden Tran	16213471
2.	Kavin Kumar Arumugam	16262979
3.	Alper Erel	16254091

Motivation:

The stock market is considered a complicated and nonlinear system. Now stock market prediction is recognized as an attracting point for financial investors. The historical price is considered as the main factor to predict the stock market trend. Historical data may be unstructured and need special handling on storing and processing.

The purpose of this project is to analyze the stock market data and get general insight on this data through visualization to find stock behavior and value at risk for each stock.

Significance:

When it comes to investing in stocks, it is important that the investor is capable of conducting a thorough analysis. Technical analysis will allow us to do the process of forecasting future price movements based on past price movements within the stock data. It will be very helpful for the investors to make financial decisions of buying, holding, or selling stocks. Although it is impossible to make 100% accurate predictions, it can definitely help investors anticipate the future.

Objectives:

Stock Market Analysis (on Big Data Hadoop). This project is based on Big Data analysis of Stock Market. The daily commodity rates of various company shares are collected and are analyzed with the help of query method. One can easily have a market watch for any day he/she wants to look at falling in the year 2016. The user can find out his profit/loss for the share he/she owns with the help of current price rate of that share stored in our database. One can also compare different shares' highs and lows with respect to the market position. This project aims at providing simple and easy analysis of the Stock Market as per the user's requirement. The analysis result can be obtained in the form of tables, graphs and pie charts. The user gets a choice to choose the method of his analysis based on the script he selects. Relational structured data has been taken in order to complete this analysis task.

Features:

1. Collect the twitter data
2. Visualize the twitter data
3. Compare stock and twitter data

Technologies:

Name	Version
Python-3.7.5	gcloud==0.18.3 google==2.0.2 google-api-core==1.14.3 google-auth==1.6.3 google-auth-httplib2==0.0.3 google-cloud-core==1.0.3 google-cloud-language==1.3.0 nltk==3.4.5 notebook==6.0.1 oauth2client==4.1.2 oauthlib==3.1.0 psutil==5.6.3 py4j==0.10.7 pyspark==2.4.4 requests==2.22.0 requests-toolbelt==0.9.1
Java	1.8.0_221
Hadoop	3.1.2
Spark	2.4.4

Dataset:

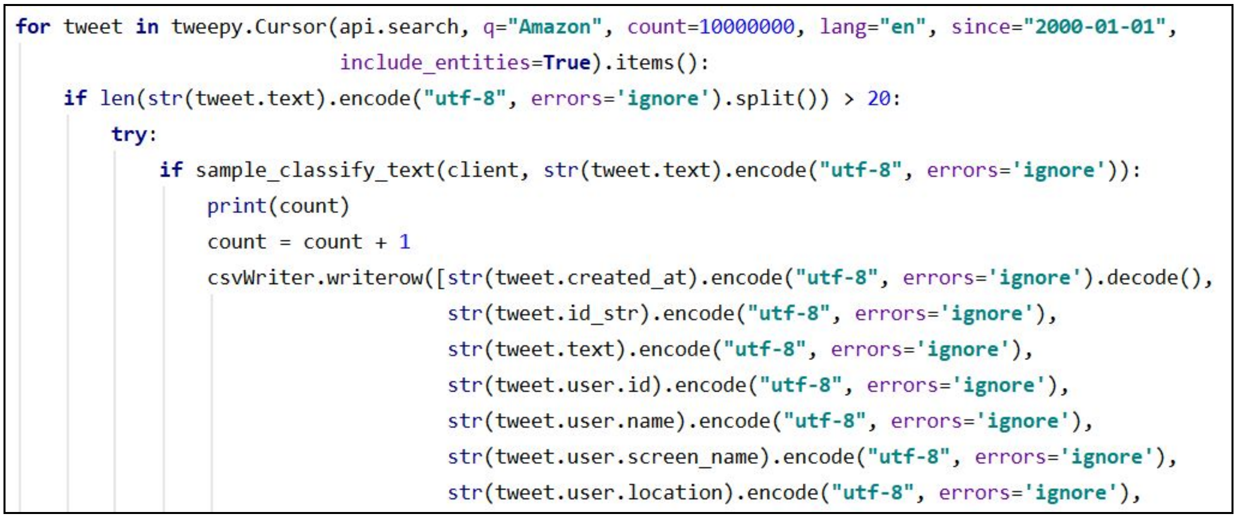
1. **Twitter data:**

a. Dataset Description:

We have collected dataset for companies

Company Name	Number of Tweets Collected	Dates
Amazon	10,996	10/27/2018 - 10/27/2019
Apple	11,869	10/27/2018 - 10/27/2019
Facebook	10,711	10/27/2018 - 10/27/2019
Intuitive Surgical	13,367	10/27/2018 - 10/27/2019
Netflix	14,546	10/27/2018 - 10/27/2019
Microsoft	13,343	10/27/2018 - 10/27/2019
iRobot	10,513	10/27/2018 - 10/27/2019
AT&T	10,575	10/27/2018 - 10/27/2019
Verizon Communications	11,093	10/27/2018 - 10/27/2019
Google	12,875	10/27/2018 - 10/27/2019

b. Code:

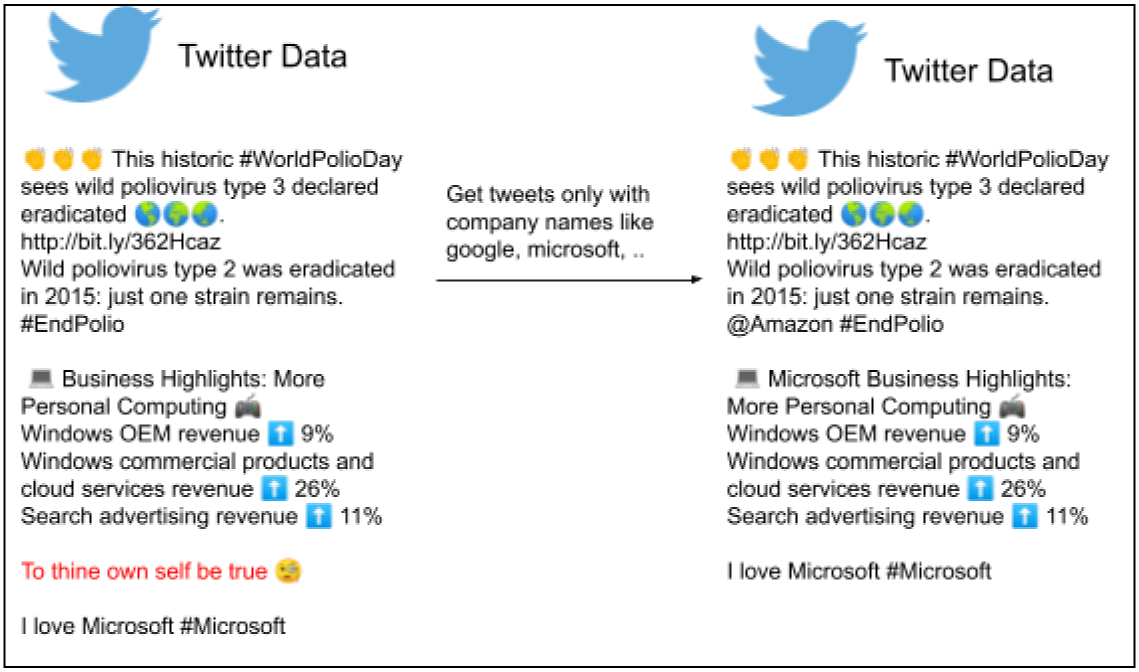


```
for tweet in tweepy.Cursor(api.search, q="Amazon", count=10000000, lang="en", since="2000-01-01",
                           include_entities=True).items():
    if len(str(tweet.text).encode("utf-8", errors='ignore').split()) > 20:
        try:
            if sample_classify_text(client, str(tweet.text).encode("utf-8", errors='ignore')):
                print(count)
                count = count + 1
                csvWriter.writerow([str(tweet.created_at).encode("utf-8", errors='ignore').decode(),
                                    str(tweet.id_str).encode("utf-8", errors='ignore'),
                                    str(tweet.text).encode("utf-8", errors='ignore'),
                                    str(tweet.user.id).encode("utf-8", errors='ignore'),
                                    str(tweet.user.name).encode("utf-8", errors='ignore'),
                                    str(tweet.user.screen_name).encode("utf-8", errors='ignore'),
                                    str(tweet.user.location).encode("utf-8", errors='ignore'),
```

c. Algorithm:

Step 1: Get tweets only with names like google and microsoft

Red ones are the removed ones



Twitter Data

👏 👏 👏 This historic #WorldPolioDay sees wild poliovirus type 3 declared eradicated 🌎🌍🌏. http://bit.ly/362Hcaz
Wild poliovirus type 2 was eradicated in 2015: just one strain remains. #EndPolio

💻 Business Highlights: More Personal Computing 🎮
Windows OEM revenue ⬆️ 9%
Windows commercial products and cloud services revenue ⬆️ 26%
Search advertising revenue ⬆️ 11%

To thine own self be true 🧐

I love Microsoft #Microsoft

Get tweets only with company names like google, microsoft, ..

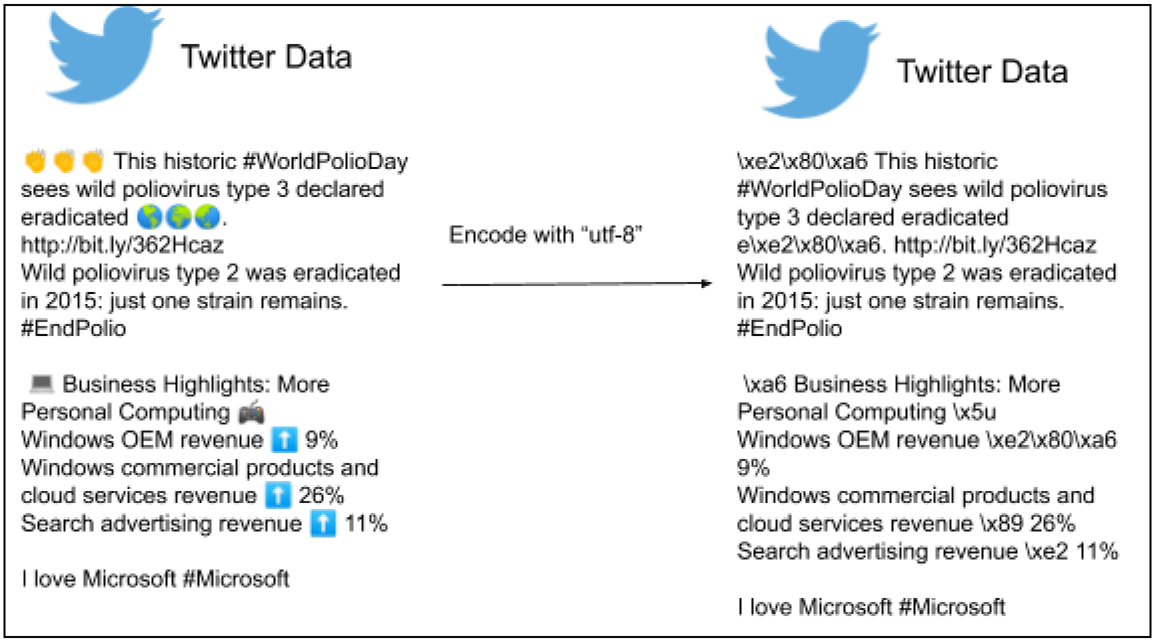
Twitter Data

👏 👏 👏 This historic #WorldPolioDay sees wild poliovirus type 3 declared eradicated 🌎🌍🌏. http://bit.ly/362Hcaz
Wild poliovirus type 2 was eradicated in 2015: just one strain remains. @Amazon #EndPolio

💻 Microsoft Business Highlights: More Personal Computing 🎮
Windows OEM revenue ⬆️ 9%
Windows commercial products and cloud services revenue ⬆️ 26%
Search advertising revenue ⬆️ 11%

I love Microsoft #Microsoft

Step 2: Encode the tweets with "utf-8



Twitter Data

👏 👏 👏 This historic #WorldPolioDay sees wild poliovirus type 3 declared eradicated 🌎🌍🌏. http://bit.ly/362Hcaz
Wild poliovirus type 2 was eradicated in 2015: just one strain remains. #EndPolio

💻 Business Highlights: More Personal Computing 🎮
Windows OEM revenue ⬆️ 9%
Windows commercial products and cloud services revenue ⬆️ 26%
Search advertising revenue ⬆️ 11%

I love Microsoft #Microsoft

Encode with "utf-8"

Twitter Data

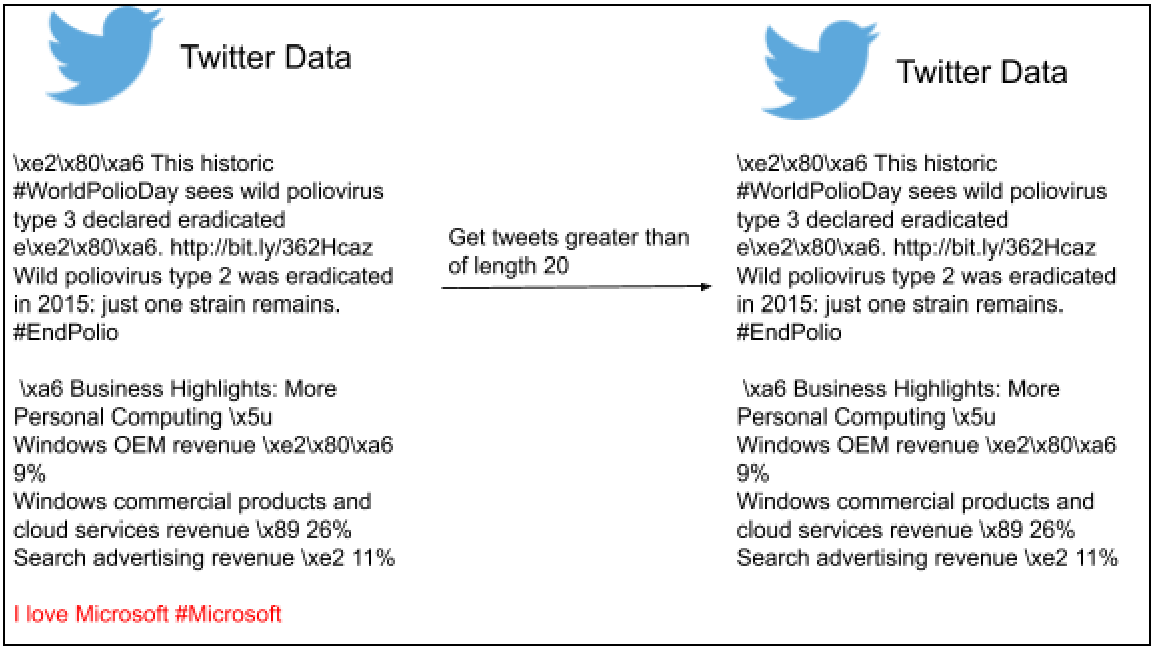
\xe2\x80\xa6 This historic #WorldPolioDay sees wild poliovirus type 3 declared eradicated e\xe2\x80\xa6. http://bit.ly/362Hcaz
Wild poliovirus type 2 was eradicated in 2015: just one strain remains. #EndPolio

\xa6 Business Highlights: More Personal Computing \x5u
Windows OEM revenue \xe2\x80\xa6 9%
Windows commercial products and cloud services revenue \x89 26%
Search advertising revenue \xe2 11%

I love Microsoft #Microsoft

Step 3: Remove tweets lesser than 20 words

Red ones are the removed ones



Twitter Data

\xe2\x80\xa6 This historic #WorldPolioDay sees wild poliovirus type 3 declared eradicated e\xe2\x80\xa6. http://bit.ly/362Hcaz Wild poliovirus type 2 was eradicated in 2015: just one strain remains. #EndPolio

\xa6 Business Highlights: More Personal Computing \x5u
Windows OEM revenue \xe2\x80\xa6 9%
Windows commercial products and cloud services revenue \x89 26%
Search advertising revenue \xe2 11%

I love Microsoft #Microsoft

Get tweets greater than of length 20

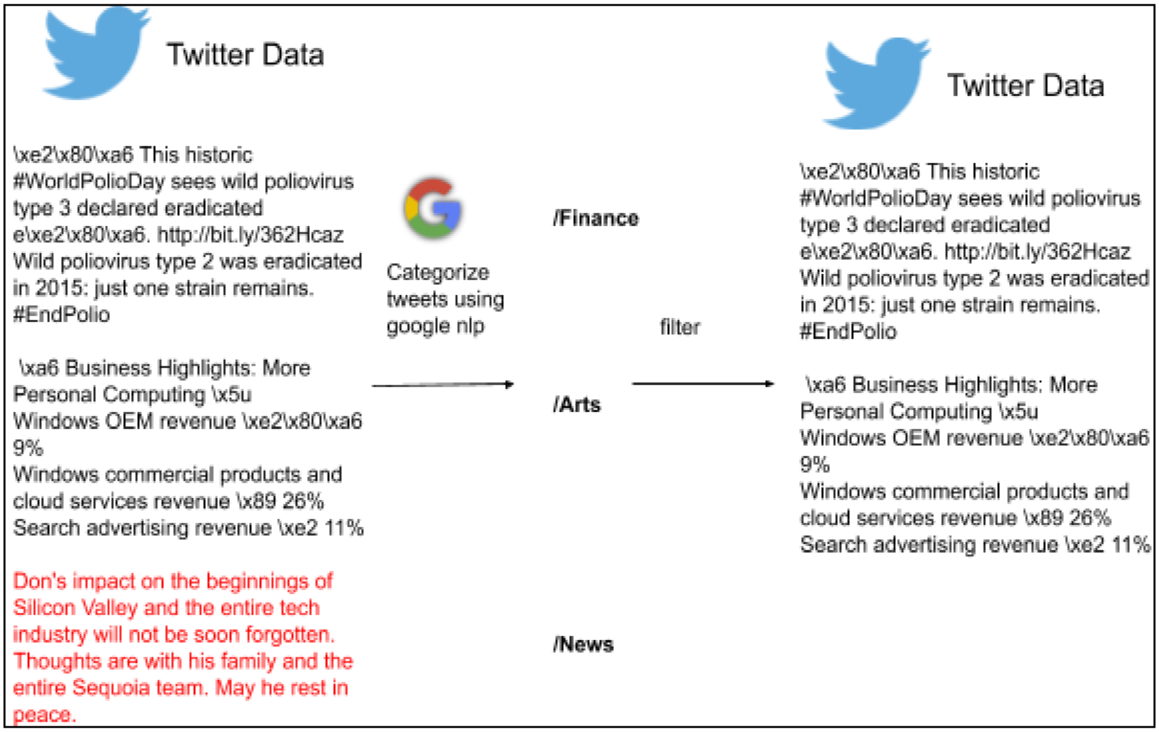
Twitter Data

\xe2\x80\xa6 This historic #WorldPolioDay sees wild poliovirus type 3 declared eradicated e\xe2\x80\xa6. http://bit.ly/362Hcaz Wild poliovirus type 2 was eradicated in 2015: just one strain remains. #EndPolio

\xa6 Business Highlights: More Personal Computing \x5u
Windows OEM revenue \xe2\x80\xa6 9%
Windows commercial products and cloud services revenue \x89 26%
Search advertising revenue \xe2 11%

Step 4: Classified the tweets and take only tweets related finance, business, news and science

Red ones are the removed ones



Twitter Data

\xe2\x80\xa6 This historic #WorldPolioDay sees wild poliovirus type 3 declared eradicated e\xe2\x80\xa6. http://bit.ly/362Hcaz Wild poliovirus type 2 was eradicated in 2015: just one strain remains. #EndPolio

\xa6 Business Highlights: More Personal Computing \x5u
Windows OEM revenue \xe2\x80\xa6 9%
Windows commercial products and cloud services revenue \x89 26%
Search advertising revenue \xe2 11%

Don's impact on the beginnings of Silicon Valley and the entire tech industry will not be soon forgotten. Thoughts are with his family and the entire Sequoia team. May he rest in peace.

Categorize tweets using google nlp

/Finance

/Arts

/News

filter

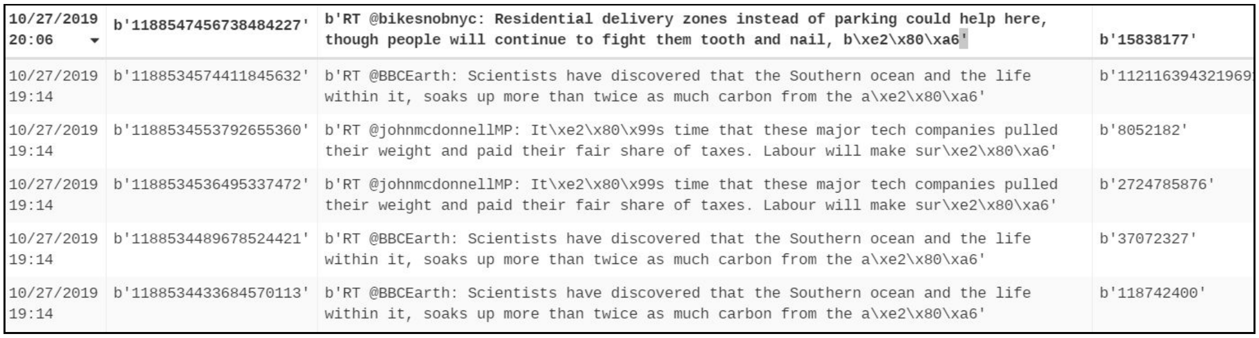
Twitter Data

\xe2\x80\xa6 This historic #WorldPolioDay sees wild poliovirus type 3 declared eradicated e\xe2\x80\xa6. http://bit.ly/362Hcaz Wild poliovirus type 2 was eradicated in 2015: just one strain remains. #EndPolio

\xa6 Business Highlights: More Personal Computing \x5u
Windows OEM revenue \xe2\x80\xa6 9%
Windows commercial products and cloud services revenue \x89 26%
Search advertising revenue \xe2 11%

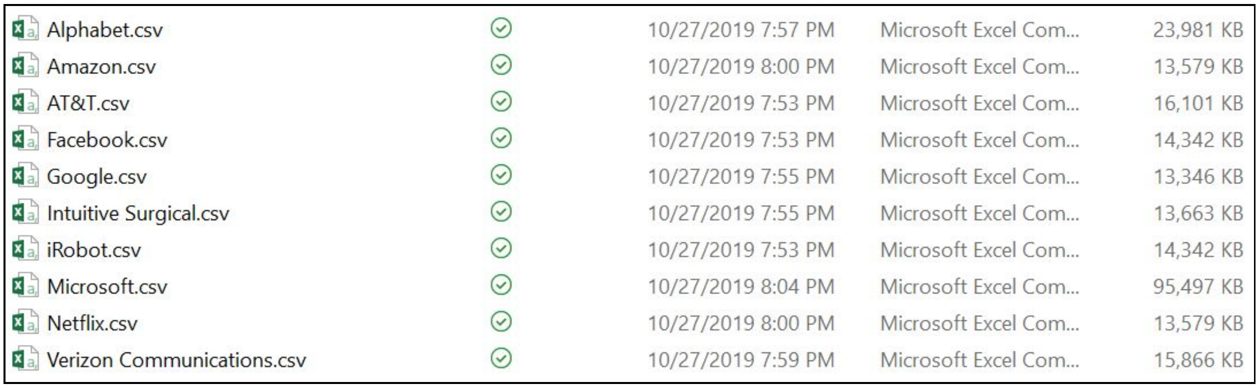
d. Result:

Final twitter data looks like this



10/27/2019 20:06	b'1188547456738484227'	b'RT @bikesnobnyc: Residential delivery zones instead of parking could help here, though people will continue to fight them tooth and nail, b\xe2\x80\xa6'	b'15838177'
10/27/2019 19:14	b'1188534574411845632'	b'RT @BBCEarth: Scientists have discovered that the Southern ocean and the life within it, soaks up more than twice as much carbon from the a\xe2\x80\xa6'	b'112116394321969
10/27/2019 19:14	b'1188534553792655360'	b'RT @johnmcdonnellMP: It\xe2\x80\x99s time that these major tech companies pulled their weight and paid their fair share of taxes. Labour will make sur\xe2\x80\xa6'	b'8052182'
10/27/2019 19:14	b'1188534536495337472'	b'RT @johnmcdonnellMP: It\xe2\x80\x99s time that these major tech companies pulled their weight and paid their fair share of taxes. Labour will make sur\xe2\x80\xa6'	b'2724785876'
10/27/2019 19:14	b'1188534489678524421'	b'RT @BBCEarth: Scientists have discovered that the Southern ocean and the life within it, soaks up more than twice as much carbon from the a\xe2\x80\xa6'	b'37072327'
10/27/2019 19:14	b'1188534433684570113'	b'RT @BBCEarth: Scientists have discovered that the Southern ocean and the life within it, soaks up more than twice as much carbon from the a\xe2\x80\xa6'	b'118742400'

We have created about 10 different csv files.

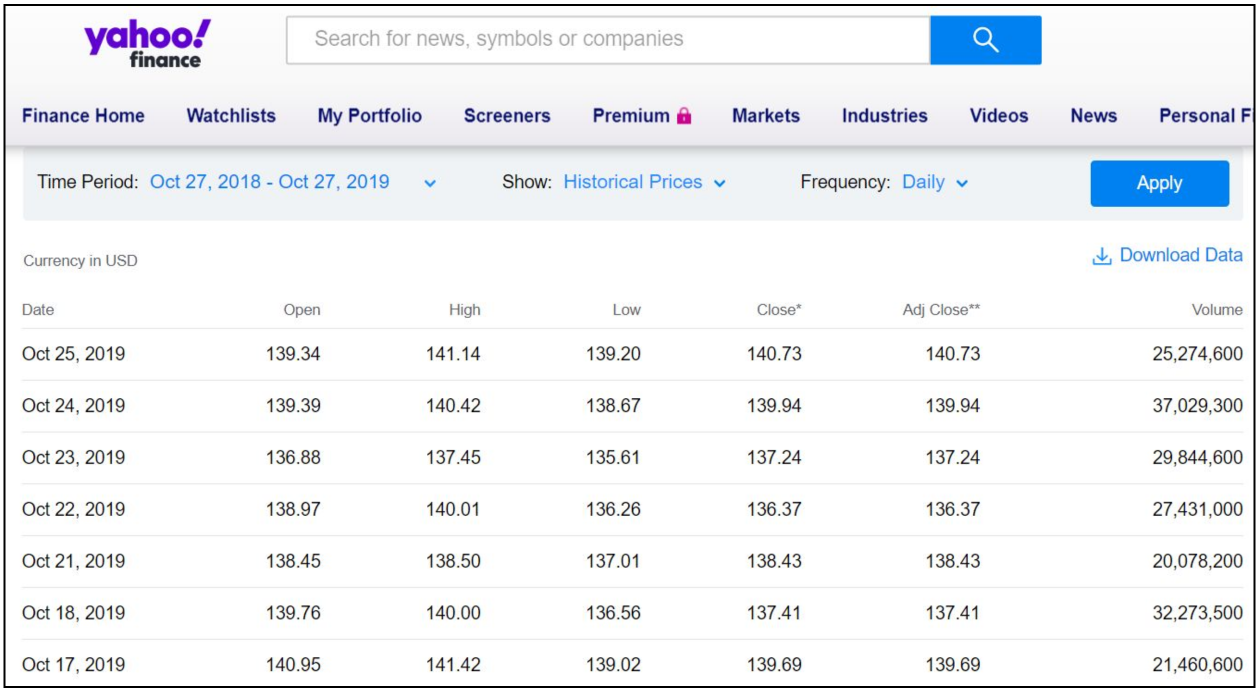


Alphabet.csv	✓	10/27/2019 7:57 PM	Microsoft Excel Com...	23,981 KB
Amazon.csv	✓	10/27/2019 8:00 PM	Microsoft Excel Com...	13,579 KB
AT&T.csv	✓	10/27/2019 7:53 PM	Microsoft Excel Com...	16,101 KB
Facebook.csv	✓	10/27/2019 7:53 PM	Microsoft Excel Com...	14,342 KB
Google.csv	✓	10/27/2019 7:55 PM	Microsoft Excel Com...	13,346 KB
Intuitive Surgical.csv	✓	10/27/2019 7:55 PM	Microsoft Excel Com...	13,663 KB
iRobot.csv	✓	10/27/2019 7:53 PM	Microsoft Excel Com...	14,342 KB
Microsoft.csv	✓	10/27/2019 8:04 PM	Microsoft Excel Com...	95,497 KB
Netflix.csv	✓	10/27/2019 8:00 PM	Microsoft Excel Com...	13,579 KB
Verizon Communications.csv	✓	10/27/2019 7:59 PM	Microsoft Excel Com...	15,866 KB

2. Stock data:

a. Dataset Description:

We collected stock data from 10/27/2018 to 10/27/2019 using finance.yahoo.com



Time Period: Oct 27, 2018 - Oct 27, 2019

Show: Historical Prices

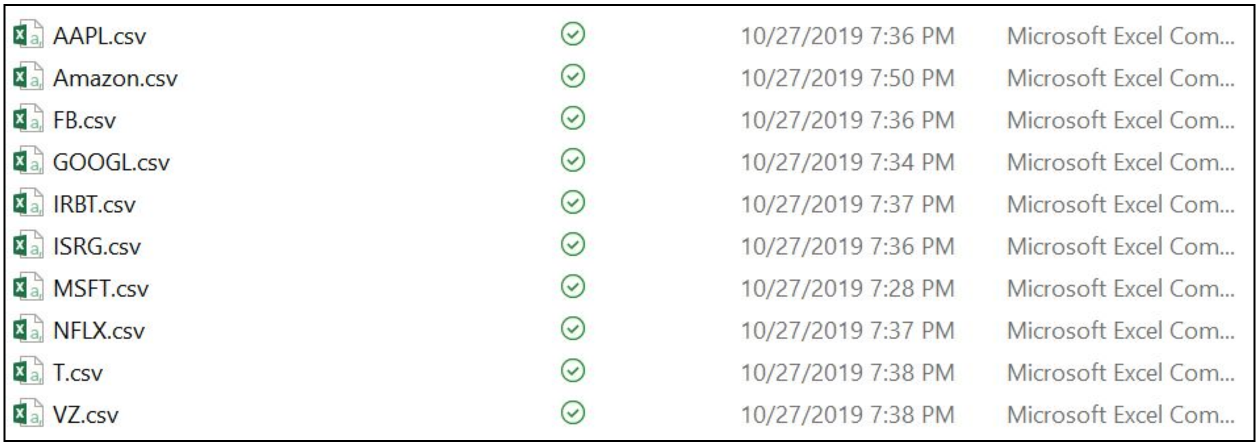
Frequency: Daily

Currency in USD

Date	Open	High	Low	Close*	Adj Close**	Volume
Oct 25, 2019	139.34	141.14	139.20	140.73	140.73	25,274,600
Oct 24, 2019	139.39	140.42	138.67	139.94	139.94	37,029,300
Oct 23, 2019	136.88	137.45	135.61	137.24	137.24	29,844,600
Oct 22, 2019	138.97	140.01	136.26	136.37	136.37	27,431,000
Oct 21, 2019	138.45	138.50	137.01	138.43	138.43	20,078,200
Oct 18, 2019	139.76	140.00	136.56	137.41	137.41	32,273,500
Oct 17, 2019	140.95	141.42	139.02	139.69	139.69	21,460,600

b. Result:

We have collected the stock market data for all of these companies

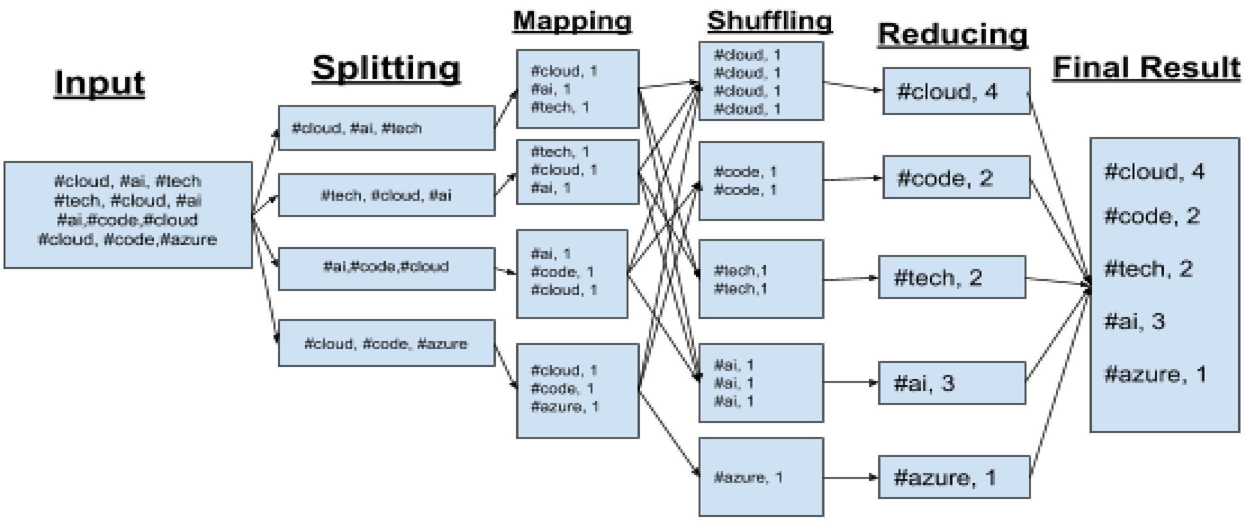


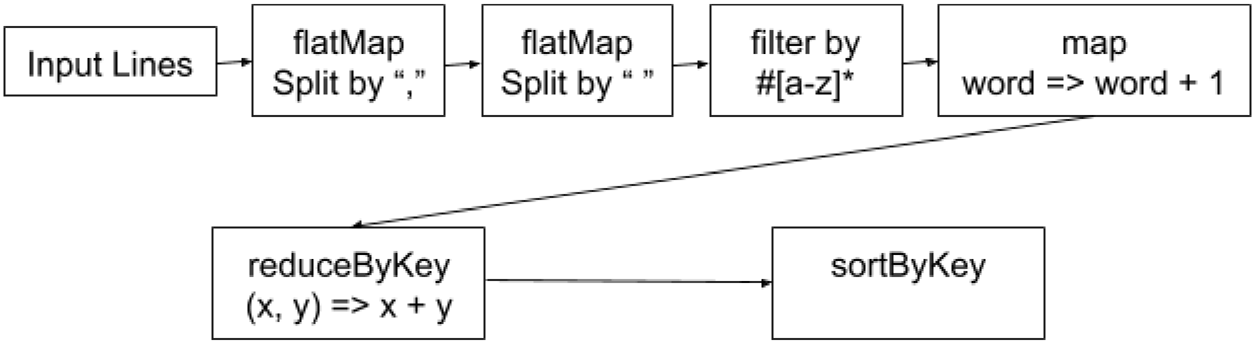
AAPL.csv	✓	10/27/2019 7:36 PM	Microsoft Excel Com...
Amazon.csv	✓	10/27/2019 7:50 PM	Microsoft Excel Com...
FB.csv	✓	10/27/2019 7:36 PM	Microsoft Excel Com...
GOOGL.csv	✓	10/27/2019 7:34 PM	Microsoft Excel Com...
IRBT.csv	✓	10/27/2019 7:37 PM	Microsoft Excel Com...
ISRG.csv	✓	10/27/2019 7:36 PM	Microsoft Excel Com...
MSFT.csv	✓	10/27/2019 7:28 PM	Microsoft Excel Com...
NFLX.csv	✓	10/27/2019 7:37 PM	Microsoft Excel Com...
T.csv	✓	10/27/2019 7:38 PM	Microsoft Excel Com...
VZ.csv	✓	10/27/2019 7:38 PM	Microsoft Excel Com...

Implementation:

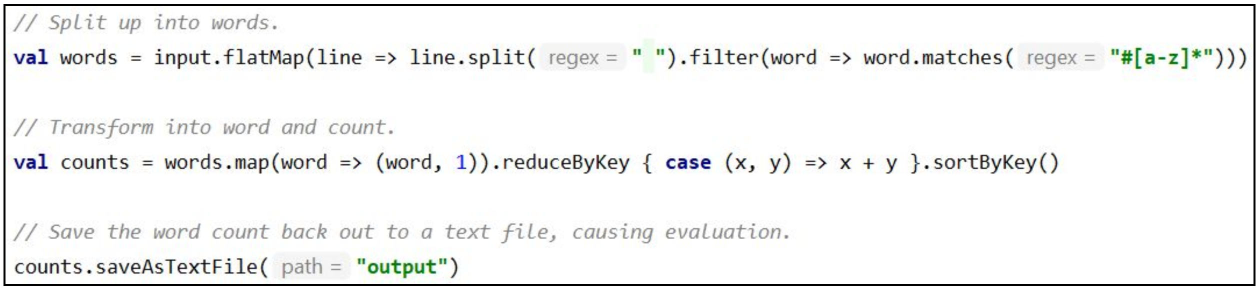
1. **Find most popular hashtags for each of the company.**

a. **Algorithm:**



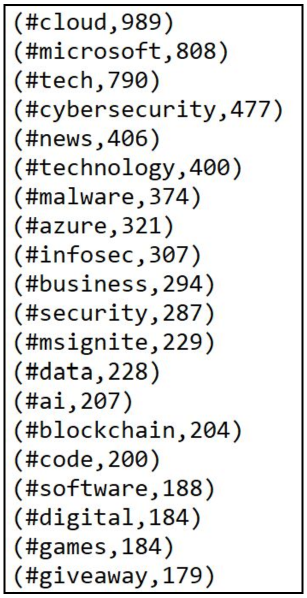


b. **Code:**

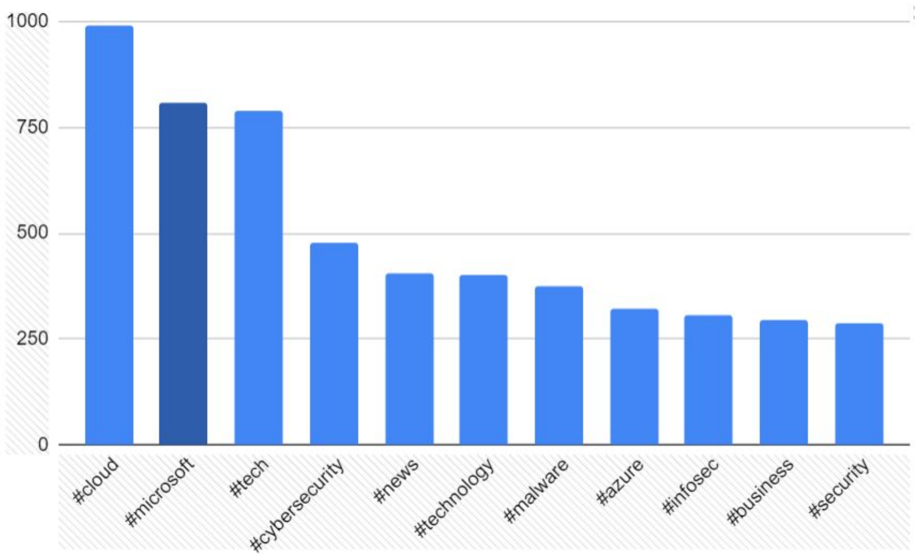


```
// Split up into words.
val words = input.flatMap(line => line.split( regex = " ").filter(word => word.matches( regex = "#[a-z]*")))

// Transform into word and count.
val counts = words.map(word => (word, 1)).reduceByKey { case (x, y) => x + y }.sortByKey()

// Save the word count back out to a text file, causing evaluation.
counts.saveAsTextFile( path = "output")
```

c. Result:

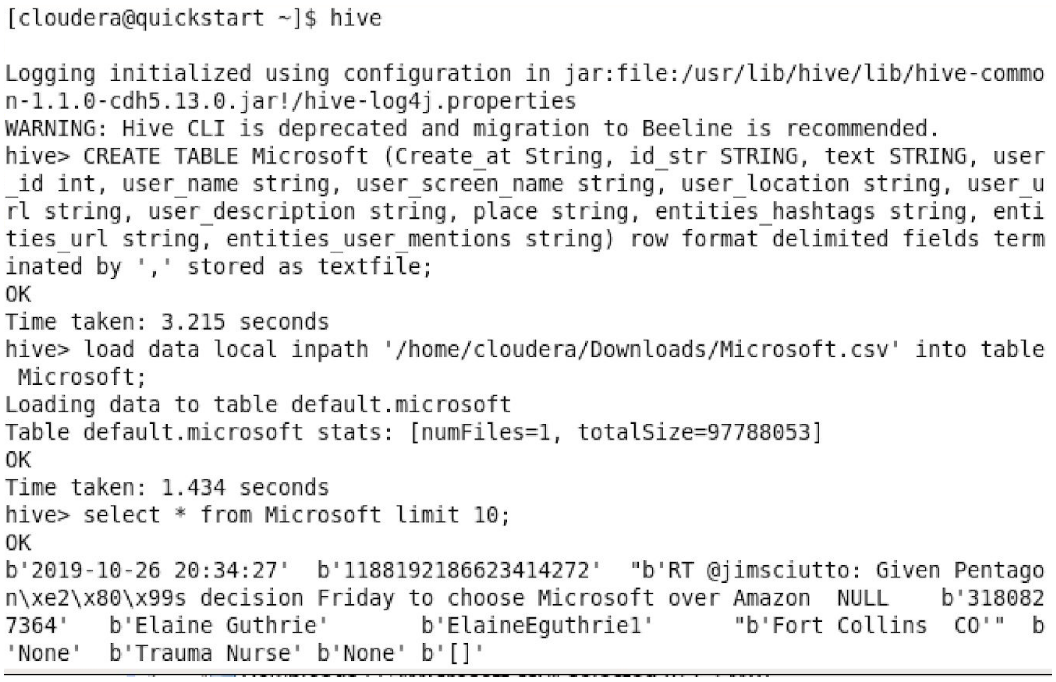


```
(#cloud,989)
(#microsoft,808)
(#tech,790)
(#cybersecurity,477)
(#news,406)
(#technology,400)
(#malware,374)
(#azure,321)
(#infosec,307)
(#business,294)
(#security,287)
(#msignite,229)
(#data,228)
(#ai,207)
(#blockchain,204)
(#code,200)
(#software,188)
(#digital,184)
(#games,184)
(#giveaway,179)
```



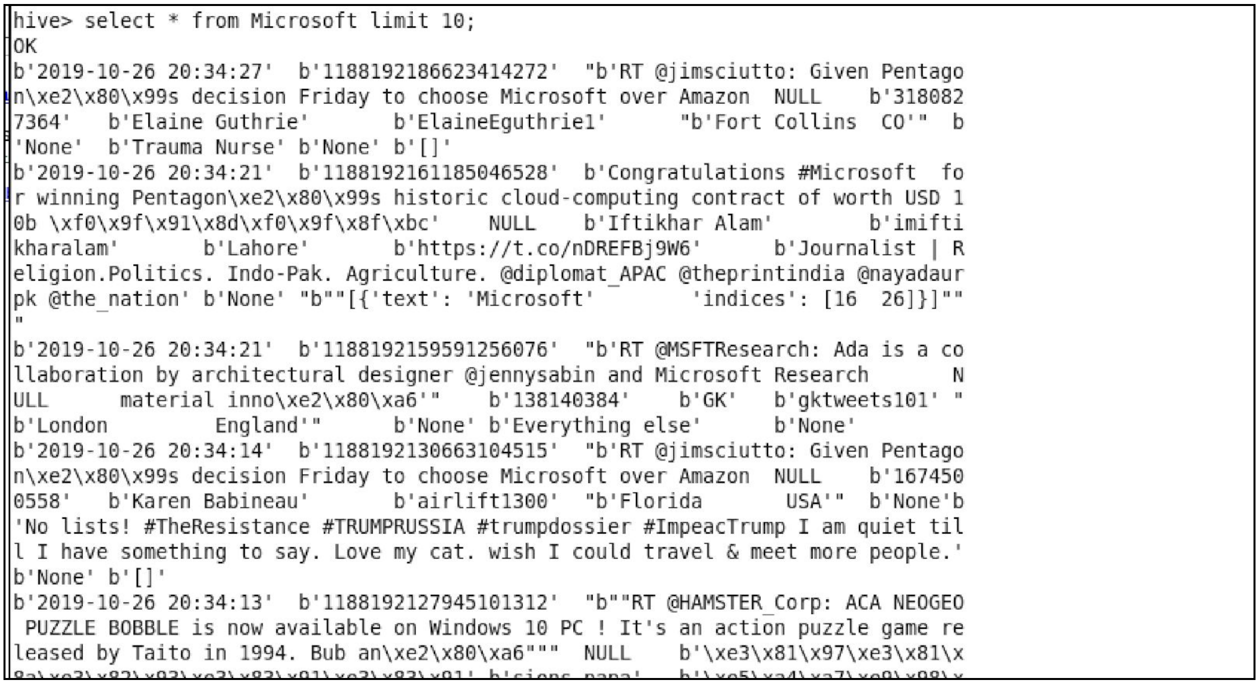
2. Create hive table and perform some queries:

Create Hive table for Microsoft file, and load the Microsoft.csv into Microsoft table.



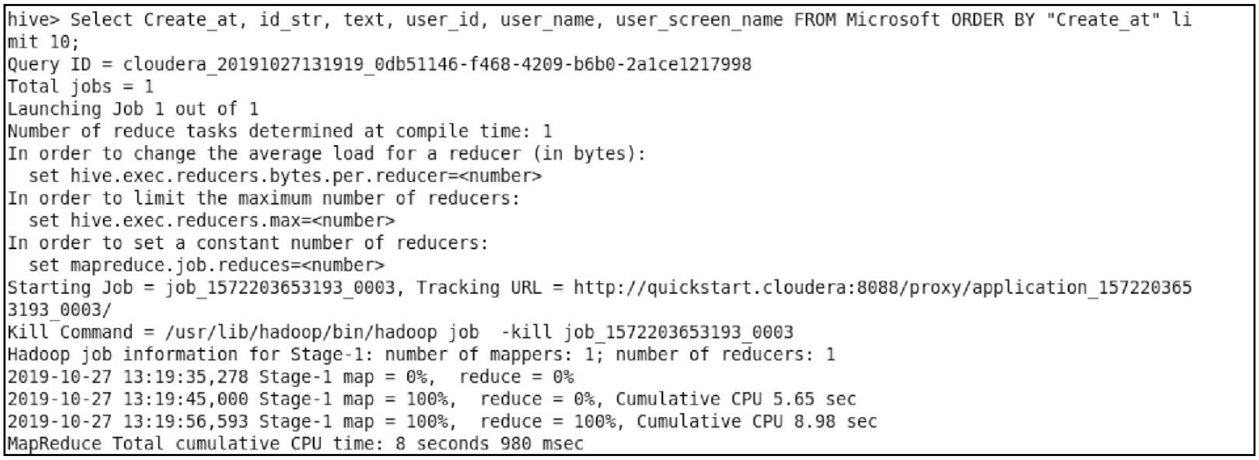
```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-commo
n-1.1.0-cdh5.13.0.jar!/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> CREATE TABLE Microsoft (Create_at String, id_str STRING, text STRING, user
_id int, user_name string, user_screen_name string, user_location string, user_u
rl string, user_description string, place string, entities_hashtags string, enti
ties_url string, entities_user_mentions string) row format delimited fields term
inated by ',' stored as textfile;
OK
Time taken: 3.215 seconds
hive> load data local inpath '/home/cloudera/Downloads/Microsoft.csv' into table
 Microsoft;
Loading data to table default.microsoft
Table default.microsoft stats: [numFiles=1, totalSize=97788053]
OK
Time taken: 1.434 seconds
hive> select * from Microsoft limit 10;
OK
b'2019-10-26 20:34:27'  b'1188192186623414272'  "b'RT @jimsciutto: Given Pentago
n\xe2\x80\x99s decision Friday to choose Microsoft over Amazon  NULL    b'318082
7364'   b'Elaine Guthrie'       b'ElaineEguthrie1'      "b'Fort Collins  CO'"  b
'None'  b'Trauma Nurse' b'None' b'[]'
```

Show first 10 row:

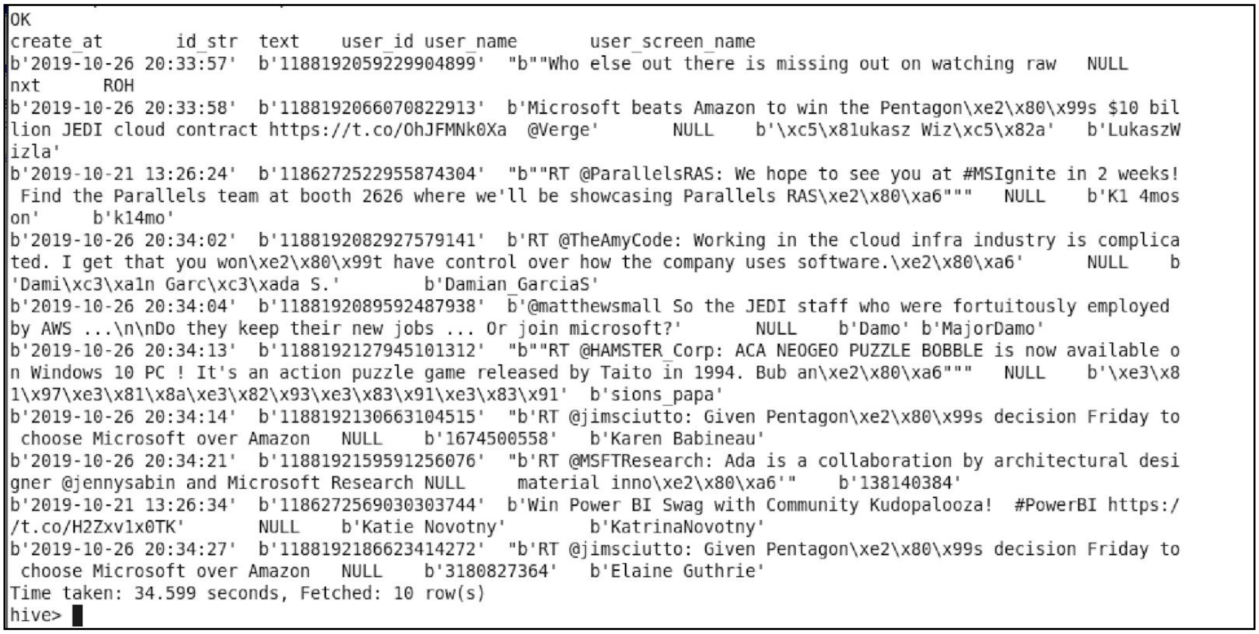


```
hive> select * from Microsoft limit 10;
OK
b'2019-10-26 20:34:27'  b'1188192186623414272'  "b'RT @jimsciutto: Given Pentago
n\xe2\x80\x99s decision Friday to choose Microsoft over Amazon  NULL    b'318082
7364'   b'Elaine Guthrie'       b'ElaineEguthrie1'      "b'Fort Collins  CO'"  b
'None'  b'Trauma Nurse' b'None' b'[]'
b'2019-10-26 20:34:21'  b'1188192161185046528'  b'Congratulations #Microsoft  fo
r winning Pentagon\xe2\x80\x99s historic cloud-computing contract of worth USD 1
0b \xf0\x9f\x91\x8d\xf0\x9f\x8f\xbc'    NULL    b'Iftikhar Alam'        b'imifti
kharalam'       b'Lahore'       b'https://t.co/nDREFBj9W6'      b'Journalist | R
eligion.Politics. Indo-Pak. Agriculture. @diplomat_APAC @theprintindia @nayadaur
pk @the_nation' b'None' "b""[{'text': 'Microsoft'       'indices': [16  26]}]""
"
b'2019-10-26 20:34:21'  b'1188192159591256076'  "b'RT @MSFTResearch: Ada is a co
llaboration by architectural designer @jennysabin and Microsoft Research       N
ULL     material inno\xe2\x80\xa6'"    b'138140384'    b'GK'   b'gktweets101' "
b'London         England'"      b'None' b'Everything else'      b'None'
b'2019-10-26 20:34:14'  b'1188192130663104515'  "b'RT @jimsciutto: Given Pentago
n\xe2\x80\x99s decision Friday to choose Microsoft over Amazon  NULL    b'167450
0558'   b'Karen Babineau'       b'airlift1300'  "b'Florida       USA'"  b'None'b
'No lists! #TheResistance #TRUMPRUSSIA #trumpdossier #ImpeacTrump I am quiet til
l I have something to say. Love my cat. wish I could travel & meet more people.'
b'None' b'[]'
b'2019-10-26 20:34:13'  b'1188192127945101312'  "b""RT @HAMSTER_Corp: ACA NEOGEO
 PUZZLE BOBBLE is now available on Windows 10 PC ! It's an action puzzle game re
leased by Taito in 1994. Bub an\xe2\x80\xa6"""  NULL    b'\xe3\x81\x97\xe3\x81\x
8a\xe3\x82\x93\xe3\x83\x91\xe3\x83\x91' b'sions_papa'   b'\xe5\xa4\xa7\xe9\x98\x
```

Show column into a table form with column name = true:



```
hive> Select Create_at, id_str, text, user_id, user_name, user_screen_name FROM Microsoft ORDER BY "Create_at" li
mit 10;
Query ID = cloudera_20191027131919_0db51146-f468-4209-b6b0-2a1ce1217998
Total jobs = 1
Launching Job 1 out of 1
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_1572203653193_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_157220365
3193_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1572203653193_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-10-27 13:19:35,278 Stage-1 map = 0%,  reduce = 0%
2019-10-27 13:19:45,000 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.65 sec
2019-10-27 13:19:56,593 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.98 sec
MapReduce Total cumulative CPU time: 8 seconds 980 msec
```



```
OK
create_at       id_str  text    user_id user_name       user_screen_name
b'2019-10-26 20:33:57'  b'1188192059229904899'  "b""Who else out there is missing out on watching raw   NULL
nxt      ROH
b'2019-10-26 20:33:58'  b'1188192066070822913'  b'Microsoft beats Amazon to win the Pentagon\xe2\x80\x99s $10 bil
lion JEDI cloud contract https://t.co/OhJFMNk0Xa  @Verge'       NULL    b'\xc5\x81ukasz Wiz\xc5\x82a'   b'LukaszW
izla'
b'2019-10-21 13:26:24'  b'1186272522955874304'  "b""RT @ParallelsRAS: We hope to see you at #MSIgnite in 2 weeks!
 Find the Parallels team at booth 2626 where we'll be showcasing Parallels RAS\xe2\x80\xa6"""   NULL    b'K1 4mos
on'     b'k14mo'
b'2019-10-26 20:34:02'  b'1188192082927579141'  b'RT @TheAmyCode: Working in the cloud infra industry is complica
ted. I get that you won\xe2\x80\x99t have control over how the company uses software.\xe2\x80\xa6'      NULL    b
'Dami\xc3\xa1n Garc\xc3\xada S.'        b'Damian_GarciaS'
b'2019-10-26 20:34:04'  b'1188192089592487938'  b'@matthewsmall So the JEDI staff who were fortuitously employed
by AWS ...\n\nDo they keep their new jobs ... Or join microsoft?'       NULL    b'Damo' b'MajorDamo'
b'2019-10-26 20:34:13'  b'1188192127945101312'  "b""RT @HAMSTER_Corp: ACA NEOGEO PUZZLE BOBBLE is now available o
n Windows 10 PC ! It's an action puzzle game released by Taito in 1994. Bub an\xe2\x80\xa6"""   NULL    b'\xe3\x8
1\x97\xe3\x81\x8a\xe3\x82\x93\xe3\x83\x91\xe3\x83\x91'  b'sions_papa'
b'2019-10-26 20:34:14'  b'1188192130663104515'  "b'RT @jimsciutto: Given Pentagon\xe2\x80\x99s decision Friday to
 choose Microsoft over Amazon   NULL    b'1674500558'   b'Karen Babineau'
b'2019-10-26 20:34:21'  b'1188192159591256076'  "b'RT @MSFTResearch: Ada is a collaboration by architectural desi
gner @jennysabin and Microsoft Research NULL     material inno\xe2\x80\xa6'"    b'138140384'
b'2019-10-21 13:26:34'  b'1186272569030303744'  b'Win Power BI Swag with Community Kudopalooza!  #PowerBI https:/
/t.co/H2Zxv1x0TK'       NULL    b'Katie Novotny'        b'KatrinaNovotny'
b'2019-10-26 20:34:27'  b'1188192186623414272'  "b'RT @jimsciutto: Given Pentagon\xe2\x80\x99s decision Friday to
 choose Microsoft over Amazon   NULL    b'3180827364'   b'Elaine Guthrie'
Time taken: 34.599 seconds, Fetched: 10 row(s)
hive> █
```

Implementation Status Report:

Work Completed			
Task	**Description**	**Contributor**	**Percentage**
1.	Dataset collection from twitter	Kavin Kumar Arumugam and Alpher Erel	33.33%
2.	Dataset preprocessing - using nlp techniques	Kavin Kumar Arumugam and Jayden Tran	
3.	Dataset preprocessing - using google nlp	Kavin Kumar Arumugam and Jayden Tran	
4.	MapReduce Algorithm on the preprocessed data using scala	Alper Erel	33.33%
5.	Visualization of data from the MapReduce Algorithm	Alper Erel	
6.	Analysis on MapReduce Algorithm	Alper Erel	
7.	Creating hive table using the schema from the downloaded csv	Jayden Tran	33.33%
8.	Loading the downloaded csv to the created table	Jayden Tran	
9.	Creating queries and do some analysis on the created table	Jayden Tran	

Work To Be Completed			
Task	**Description**	**Contributor**	**Percentage**
10.	Create some more queries on hive		
11.	Visualize the hive queries		
12.	Compare the predicted data and real stock data		

Preliminary Results:

- Upon evaluating the Microsoft dataset, we found out the most popular hashtags which define Microsoft including, but not limited to:
    - #cloud
    - #tech
    - #cybersecurity etc.
- These are the top words that people talk about that describes Microsoft.

References/Bibliography:

1) https://m.benzinga.com/article/9602734
2) https://www.investopedia.com/terms/s/stock-analysis.asp
3) https://cleartax.in/s/stock-market-analysis