COVID-19 TWITTER ANALYSIS

COURSE: BIG DATA PROGRAMMING

PROJECT INCREMENT-1

Vidyullatha Lakshmi Kaza- 8

Aparna Manda- 11

Lohitha Yenugu- 19

Extracted COVID-19 related Live Stream Tweets using Python, performed analysis on data using Map-Reduce and Hive in Increment-I.

Python Code to extract live stream twitter data:

Open

TwitterDataExtraction.py
~/Desktop/BigDataProgramming

```
    def on_status(self, status):
        print(status.id_str)
        # if "retweeted_status" attribute exists, flag this tweet as a retweet.
        is_retweet = hasattr(status, "retweeted_status")

        # check if text has been truncated
        if hasattr(status,"extended_tweet"):
            text = status.extended_tweet["full_text"]
        else:
            text = status.text

        # check if this is a quote tweet.
        is_quote = hasattr(status, "quoted_status")
        quoted_text = ""
        if is_quote:
            # check if quoted tweet's text has been truncated before recording it
            if hasattr(status.quoted_status,"extended_tweet"):
                quoted_text = status.quoted_status.extended_tweet["full_text"]
            else:
                quoted_text = status.quoted_status.text

        # remove characters that might cause problems with csv encoding
        remove_characters = [",","\n"]
        for c in remove_characters:
            text.replace(c," ")
            quoted_text.replace(c, " ")

        with open("/home/lohitha/Desktop/BigDataProgramming/COVID19_Data.json", "a", encoding='utf-8') as f:
            f.write("%s,%s,%s,%s,%s,%s\n" % (status.created_at,status.user.screen_name,is_retweet,is_quote,text,quoted_text))

    def on_error(self, status_code):
        print("Encountered streaming error (", status_code, ")")
        sys.exit()

if __name__ == "__main__":
    # complete authorization and initialize API endpoint
    auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    api = tweepy.API(auth)

    # initialize stream
    streamListener = StreamListener()
    stream = tweepy.Stream(auth=api.auth, listener=streamListener,tweet_mode='extended')
    with open("/home/lohitha/Desktop/BigDataProgramming/COVID19_Data.json", "w", encoding='utf-8') as f:
        f.write("date,user,is_retweet,is_quote,text,quoted_text\n")
    tags = ["covid19","covid-19","COVID-19","COVID19","corona","CORONA"]
    stream.filter(track=tags)
```

Collected Tweets:

```
date,user,is_retweet,is_quote,text,quoted_text
2020-04-15 02:06:43,kimpetty64,False,True,😂😂😂🤦 and just like Corona beer,Meanwhile, Trump supporters burn their 8-tracks cursing out Keith Moon, Pete Townshend and Roger Daltrey.

Blaming WHO
2020-04-15 02:06:43,2021Diary,True,False,RT @PinkeshOfficial: While PM Modi having serious discussion with states on Covid19, Aditya Thackeray is busy with his mobile.

Anjana was…,
2020-04-15 02:06:43,rdc_south,True,False,RT @chennaicorp: Here's the Graphical Representation of total COVID-19 positive cases in Chennai as on 14-04-2020.

#Covid19Chennai
#GCC #…,
2020-04-15 02:06:42,openletterbot,False,False,🖉 Support Patrick by signing “Support the USPS!” and I’ll deliver a copy to your officials too: https://t.co/GjO2cQ6Sqq

📫 Last delivered to @RepStephenLynch, @SenMarkey and @SenWarren #MA08 #MApoli #MApols #COVID19 https://t.co/JSsgHhZjoI,
2020-04-15 02:06:43,ahernandez85b,True,True,RT @KelemenCari: It's been 35 years and there is still no vaccine for AIDS. https://t.co/X8yKxTSt3G,As we reopen #Ohio, people will have to be ve
2020-04-15 02:06:43,lucas_bhmg,True,True,RT @AbdelDeuxFois: Sans promo ni rien. Le monde chico 🌍,FR ALERTE INFO - 36,7 millions de téléspectateurs ont suivi l'allocution d'Emmanuel #Macron
2020-04-15 02:06:43,Celinve,True,False,RT @MonsieurQFB: Aquí los análisis de casos de COVID—19 en  México MX y países de América, Europa US IT ES KR BR DE JP FR PE AR EC CL CO y…,
2020-04-15 02:06:43,MrSol54252546,True,False,RT @islamramahdotco: Kiai Anwar Zahid: Patuhi Protokoler Covid-19

"Kita ajak seluruh umat agar taat dan patuh instruksi pemerintah dan pat…,
2020-04-15 02:06:43,drjaswantpatil,False,False,@SrBachchan # Homeopathy can beat Corona.                                Want to see the results allow homoeopaths to treat and see the change.
2020-04-15 02:06:43,xznxxz,True,False,RT @gunawan_anas: Jujurnya pemerintah masih setengah2.. selain data saat ini, pmrintah harus terbuka mengenai forecasting versi mereka akan…,
2020-04-15 02:06:43,WUNNA_1,True,False,RT @21savage: Bang outside I hang outside
don’t come out da house cuz corona outside,
2020-04-15 02:06:43,abnesdad,True,False,RT @TrumpNoodles: 🏠Michigan kid videotaping his dad who hates the Michigan governor

🏠Funny as heck 😂😂🤣🤣

🏠He sounds like my mom when it co…,
2020-04-15 02:06:43,respirovondy,True,False,RT @crushdobbb20: Flávia Pavanelli fez foto, fez preenchimento, arrumou cabelo, unhas, encontrou amigos em casa e agora vem falar q acha q…,
2020-04-15 02:06:43,carterpillar82,True,False,RT @franjuero: Un reportaje de la Televisión Italiana del 2015 donde habla de que China experimenta con un virus SARS insertandole proteína…,
2020-04-15 02:06:43,Kak77742001,True,False,RT @BastienParisot: 🔴 THREAD : Une proche m'envoie ce soir ces photos. Elle travaille dans un hôpital public de la région parisienne. Des b…,
2020-04-15 02:06:43,FloreCsGo,True,True,RT @AbdelDeuxFois: Sans promo ni rien. Le monde chico 🌍,FR ALERTE INFO - 36,7 millions de téléspectateurs ont suivi l'allocution d'Emmanuel #Macron h
2020-04-15 02:06:43,QAlwaysWins,True,False,RT @21savage: Bang outside I hang outside
don’t come out da house cuz corona outside,
2020-04-15 02:06:43,poppy_yifan,True,True,RT @McKe3z: อีสัด ประชาชนไทยทั้งสู้โควิดกันเอง บริจาคเงินกันเอง แบ่งของกินของใช้กันเอง ดับไฟป่าเอง มาถึงขั้นสร้างความสัมพันธ์ระหว่างประเทศ…,Thank you our friends from #Thailand. Thailand has long been a popu
ได้ทั้งแผ่นดินไทย สู้ไปด้วยกันนะครับTW TH
#nnevvy
2020-04-15 02:06:43,ikanbadutz,False,False,Makin tanbah hari, akhir dr wabah corona ini makin dekat 😁

Dan semua kembali normal, yes 😁😁😁,
2020-04-15 02:06:43,babyyis_,True,False,RT @21savage: Bang outside I hang outside
```

Map Reduce program to count the tweets by each user:

UserTweetsCount.java

```
13
14    public static class Map extends Mapper<LongWritable, Text, Text,
15 IntWritable> {
16
17        private Text user = new Text();
18        private final static IntWritable one = new IntWritable(1);
19        public void map(LongWritable key, Text value, Context context )
20 throws IOException, InterruptedException {
21            String line = value.toString();
22            String str[]=line.split(",");
23            if(str.length>1){
24                 user.set(str[1]);
25            }
26       context.write(user, one);
27       }
28
29     }
30
31    public static class Reduce extends Reducer<Text, IntWritable,
32 Text, IntWritable> {
33
34        public void reduce(Text key, Iterable<IntWritable> values,
35 Context context)
36          throws IOException, InterruptedException {
37            int sum = 0;
38            int l=0;
39            for (IntWritable val : values) {
40                    l+=1;
41                sum += val.get();
42            }
43            sum=sum/l;
44            context.write(key, new IntWritable(sum));
45        }
46     }
47
48    public static void main(String[] args) throws Exception {
49        Configuration conf = new Configuration();
50
51            @SuppressWarnings("deprecation")
52                 Job job = new Job(conf, "UserCount");
53            job.setJarByClass(UserTweetsCount.class);
54
55            job.setMapOutputKeyClass(Text.class);
56            job.setMapOutputValueClass(IntWritable.class);
57
58        job.setOutputKeyClass(Text.class);
59        job.setOutputValueClass(IntWritable.class);
```

Starting HDFS namenode and datanodes:

```
lohitha@ubuntu:~/hadoop/hadoop-2.8.1/sbin$ start-dfs.sh
20/04/14 20:23:28 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/lohitha/hadoop/hadoop-2.8.1/logs/hadoop-lohitha-namenode-ubuntu.out
localhost: starting datanode, logging to /home/lohitha/hadoop/hadoop-2.8.1/logs/hadoop-lohitha-datanode-ubuntu.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/lohitha/hadoop/hadoop-2.8.1/logs/hadoop-lohitha-secondarynamenode-ubuntu.out
20/04/14 20:23:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
lohitha@ubuntu:~/hadoop/hadoop-2.8.1/sbin$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /home/lohitha/hadoop/hadoop-2.8.1/logs/yarn-lohitha-resourcemanager-ubuntu.out
localhost: starting nodemanager, logging to /home/lohitha/hadoop/hadoop-2.8.1/logs/yarn-lohitha-nodemanager-ubuntu.out
```

Loading input data from local to HDFS:

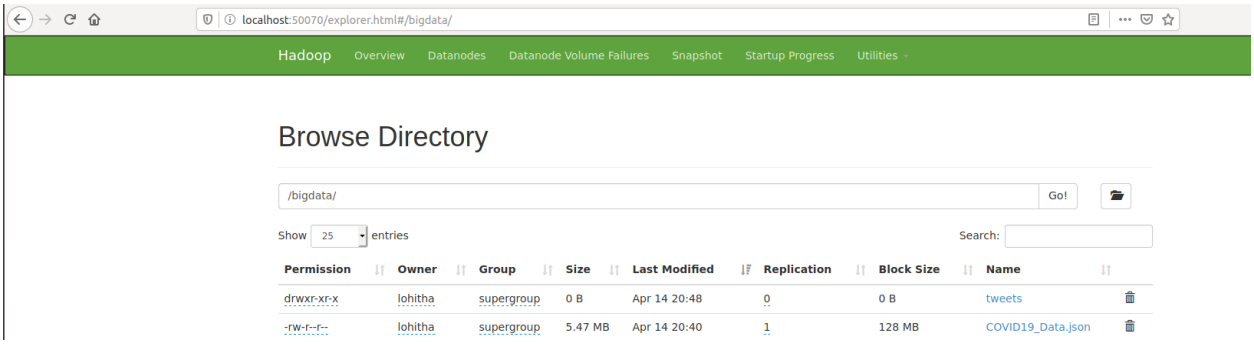
```
lohitha@ubuntu:~/Desktop/BigDataProgramming$ hdfs dfs -put COVID19_Data.json /bigdata/
20/04/14 20:40:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

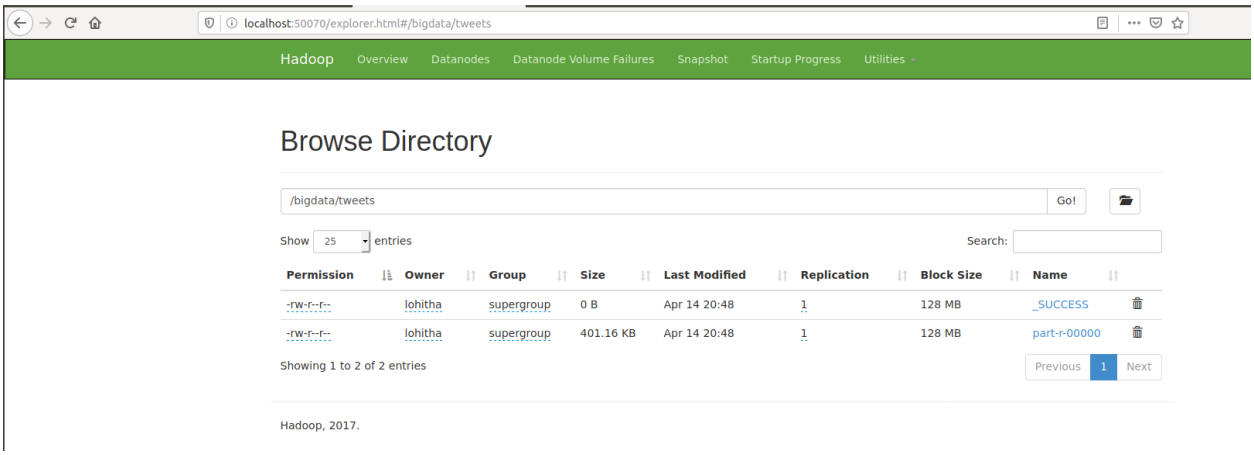
Running MapReduce:

```
lohitha@ubuntu:~/Downloads$ hadoop jar UserCount.jar UserTweetsCount /bigdata/COVID19_Data.json /bigdata/tweets/
20/04/14 20:47:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
20/04/14 20:47:58 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
20/04/14 20:47:59 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your ap
20/04/14 20:48:00 INFO input.FileInputFormat: Total input files to process : 1
20/04/14 20:48:00 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:927)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:578)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:755)
20/04/14 20:48:00 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:927)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:578)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:755)
20/04/14 20:48:00 INFO mapreduce.JobSubmitter: number of splits:1
20/04/14 20:48:00 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1586921057152_0003
20/04/14 20:48:01 INFO impl.YarnClientImpl: Submitted application application_1586921057152_0003
20/04/14 20:48:02 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1586921057152_0003/
20/04/14 20:48:02 INFO mapreduce.Job: Running job: job_1586921057152_0003
20/04/14 20:48:17 INFO mapreduce.Job: Job job_1586921057152_0003 running in uber mode : false
20/04/14 20:48:17 INFO mapreduce.Job:  map 0% reduce 0%
20/04/14 20:48:32 INFO mapreduce.Job:  map 100% reduce 0%
20/04/14 20:48:47 INFO mapreduce.Job:  map 100% reduce 100%
20/04/14 20:48:50 INFO mapreduce.Job: Job job_1586921057152_0003 completed successfully
20/04/14 20:48:50 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=1172059
                FILE: Number of bytes written=2617277
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5741057
                HDFS: Number of bytes written=410786
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
```

```
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=11966
                Total time spent by all reduces in occupied slots (ms)=12023
                Total time spent by all map tasks (ms)=11966
                Total time spent by all reduce tasks (ms)=12023
                Total vcore-milliseconds taken by all map tasks=11966
                Total vcore-milliseconds taken by all reduce tasks=12023
                Total megabyte-milliseconds taken by all map tasks=12253184
                Total megabyte-milliseconds taken by all reduce tasks=12311552
        Map-Reduce Framework
                Map input records=52473
                Map output records=52473
                Map output bytes=1066546
                Map output materialized bytes=1172059
                Input split bytes=112
                Combine input records=0
                Combine output records=0
                Reduce input groups=24116
                Reduce shuffle bytes=1172059
                Reduce input records=52473
                Reduce output records=24116
                Spilled Records=104946
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=216
                CPU time spent (ms)=3330
                Physical memory (bytes) snapshot=302473216
                Virtual memory (bytes) snapshot=3893735424
                Total committed heap usage (bytes)=170004480
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=5740945
        File Output Format Counters
                Bytes Written=410786
```

HDFS File Structure:





Output File:

```
lohitha@ubuntu:~/Downloads$ hdfs dfs -cat /bigdata/tweets/part-r-00000 | head
20/04/14 22:29:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
        1
        1
  111 se han recuper…   1
  2 NAVOLATO    1
  Gue kirimi buat Makan selama pedemi Virus Corona      1
  aseguran que hay varias muertes clasificadas como “neumonía atípica” con todas las características de ser #coronavirus. https://t.co/vpe0pZYfLM       1
  very inspirational. This drug is a game changer.      1
  where we are measured by our humanity not valued in silver or other coin.     1
  देश के मजदूर को LAN   1
  "not responsible" and put all the responsibility on the States. The States then accepted the challenge and sourced their own supplies. NOW THIS???     1
cat: Unable to write to output stream.
lohitha@ubuntu:~/Downloads$
```

Hive Queries:

Create Tweets Table:

```
hive> create table tweets (tweet_date STRING,user STRING,is_retweet STRING,is_qu
ote STRING,text STRING,qouted_text STRING)
    > [cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> create table tweets (tweet_date STRING,user STRING,is_retweet STRING,is_qu
ote STRING,text STRING,qouted_text STRING)
    > row format delimited fields terminated by "," stored as textfile;
OK
Time taken: 1.493 seconds
```

Load Twitter Data into Tweets table:

```
hive> load data local inpath "/home/cloudera/Downloads/COVID19_Data.csv" into ta
ble tweets;
Loading data to table default.tweets
Table default.tweets stats: [numFiles=1, totalSize=5740945]
OK
Time taken: 1.58 seconds
```

Query 1: Fetch top users with more number of tweets

```
hive> select user,count(1) as tweet_count from tweets where text is not null gro
up by user order by tweet_count desc limit 10;
Query ID = cloudera_20200414212424_9bed624d-dab4-488a-891d-0dcc5d58ce86
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 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_1586916751037_0004, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1586916751037_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1586916751037_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-04-14 21:24:29,303 Stage-1 map = 0%,  reduce = 0%
2020-04-14 21:24:49,156 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.51 sec
2020-04-14 21:25:10,351 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 11.04 sec
MapReduce Total cumulative CPU time: 11 seconds 40 msec
Ended Job = job_1586916751037_0004
Launching Job 2 out of 2
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_1586916751037_0005, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1586916751037_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1586916751037_0005
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2020-04-14 21:25:30,720 Stage-2 map = 0%,  reduce = 0%
2020-04-14 21:25:49,266 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 5.59 sec
2020-04-14 21:26:06,936 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 9.05 sec
MapReduce Total cumulative CPU time: 9 seconds 50 msec
Ended Job = job_1586916751037_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 11.04 sec   HDFS Read: 5748913 HDFS Write: 690474 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 9.05 sec   HDFS Read: 695482 HDFS Write: 201 SUCCESS
Total MapReduce CPU Time Spent: 20 seconds 90 msec
OK
openletterbot   55
 Dy CM  26
HollyDeinert    21
syedsalu11      20
paragpa10920537 17
Em coletiva de imprensa realizada no @Planalto  16
JPNicholasDabot 15
```

```
OK
openletterbot   55
 Dy CM  26
HollyDeinert    21
syedsalu11      20
paragpa10920537 17
Em coletiva de imprensa realizada no @Planalto  16
JPNicholasDabot 15
RDSharm91052874 13
rogerperadelles 13
VinayaKantRai2  12
Time taken: 118.462 seconds, Fetched: 10 row(s)
hive>
```

Query 2: Retweet count

```
hive> select count(1) retweet_count from tweets where trim(lower(is_retweet)) = "true";
Query ID = cloudera_20200414213232_bfebce0c-70ca-48e4-89ed-2f15680b54ef
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_1586916751037_0008, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1586916751037_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1586916751037_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-04-14 21:32:37,284 Stage-1 map = 0%,  reduce = 0%
2020-04-14 21:32:57,614 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.85 sec
2020-04-14 21:33:13,541 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.44 sec
MapReduce Total cumulative CPU time: 7 seconds 440 msec
Ended Job = job_1586916751037_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.44 sec   HDFS Read: 5749863 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 440 msec
OK
19731
Time taken: 54.324 seconds, Fetched: 1 row(s)
hive>
```

Query 3: Tweets per minute

```
hive> ( select count(1) tweet_count from tweets where trim(lower(tweet_date)) like '%2020-04-15 02:07%';
NoViableAltException(290@[])
        at org.apache.hadoop.hive.ql.parse.HiveParser.statement(HiveParser.java:1028)
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:201)
        at org.apache.hadoop.hive.ql.parse.ParseDriver.parse(ParseDriver.java:166)
        at org.apache.hadoop.hive.ql.Driver.compile(Driver.java:522)
        at org.apache.hadoop.hive.ql.Driver.compileInternal(Driver.java:1356)
        at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:1473)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1285)
        at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1275)
        at org.apache.hadoop.hive.cli.CliDriver.processLocalCmd(CliDriver.java:226)
        at org.apache.hadoop.hive.cli.CliDriver.processCmd(CliDriver.java:175)
        at org.apache.hadoop.hive.cli.CliDriver.processLine(CliDriver.java:389)
        at org.apache.hadoop.hive.cli.CliDriver.executeDriver(CliDriver.java:781)
        at org.apache.hadoop.hive.cli.CliDriver.run(CliDriver.java:699)
        at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:634)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:606)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 1:0 cannot recognize input near '(' 'select' 'count'
hive> select count(1) tweet_count from tweets where trim(lower(tweet_date)) like '%2020-04-15 02:07%';
Query ID = cloudera_20200414213939_1c2210e7-c840-431f-a32e-b2eb35171d20
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_1586916751037_0009, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1586916751037_0009/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1586916751037_0009
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-04-14 21:39:45,329 Stage-1 map = 0%,  reduce = 0%
2020-04-14 21:40:04,322 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.46 sec
2020-04-14 21:40:21,539 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.14 sec
MapReduce Total cumulative CPU time: 7 seconds 140 msec
Ended Job = job_1586916751037_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.14 sec   HDFS Read: 5749880 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 140 msec
OK
3017
Time taken: 57.374 seconds, Fetched: 1 row(s)
hive>
```

Query 4: Tweets on Layoffs

```
hive> select count(1) tweet_count from tweets where trim(lower(text)) like '%layoff%';
Query ID = cloudera_20200414214141_9c3cbfd4-11d9-416d-b135-e51dbd2875f3
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_1586916751037_0010, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1586916751037_0010/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1586916751037_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-04-14 21:41:58,116 Stage-1 map = 0%,  reduce = 0%
2020-04-14 21:42:15,545 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.7 sec
2020-04-14 21:42:36,210 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.93 sec
MapReduce Total cumulative CPU time: 7 seconds 930 msec
Ended Job = job_1586916751037_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.93 sec   HDFS Read: 5749863 HDFS Write: 2 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 930 msec
OK
2
Time taken: 59.565 seconds, Fetched: 1 row(s)
hive>
```

Query 5: Tweets on USA

```
hive> select count(1) tweet_count from tweets where trim(lower(text)) like '%usa%' or trim(lower(text)) like '%united states%';
Query ID = cloudera_20200414214545_ba40e4d2-1995-4b17-8f46-d647f76cdf95
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_1586916751037_0011, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1586916751037_0011/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1586916751037_0011
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-04-14 21:45:46,527 Stage-1 map = 0%,  reduce = 0%
2020-04-14 21:46:02,673 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.32 sec
2020-04-14 21:46:19,715 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.03 sec
MapReduce Total cumulative CPU time: 7 seconds 30 msec
Ended Job = job_1586916751037_0011
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.03 sec   HDFS Read: 5750249 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 30 msec
OK
417
Time taken: 52.837 seconds, Fetched: 1 row(s)
hive>
```

Work Completed:

- Collected Data
- Analyzed data using map reduce and hive

Responsibility:

Vidyullatha Lakshmi Kaza- Analyzed twitter data using hive queries.

Aparna Manda- Collected live streaming twitter data

Lohitha Yenugu- Map Reduce job to analyze twitter data

Work to be Completed:

- Analyzing data using Spark SQL