



PRINCIPLES OF BIG DATA **MANAGEMENT**

PROJECT PHASE-1



TEAM: 24

TEAM MEMBERS:

Barath Naravula Loganathan

Aishvarya Natarajan Iyer

Gayathree Natarajan Iyer

Bhavani Prasad Meena

TASK-1

PYTHON CODE FOR TWEET COLLECTION:

Tweets_Collection.py

```
#Import the necessary methods from tweepy library
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream

#Variables that contains the user credentials to access Twitter API
access_token = "2935099867-8tNtrkyR0Rggg5AzV5ebmibmsnZKQxRw90nrDfa"
access_token_secret = "bduYTPwQwXQi5suSLI9AMmwAxFAj0u7M4N4uX5TgXy9FF"
consumer_key = "OmCUg40OHHyqvTUb24hx3F1mQ"
consumer_secret = "BMnGdkIJAlWMyp18ff9rzcFZlkdWJjUUcQc3WExyHvc27rnhcQ"

#This is a basic listener that just prints received tweets to stdout.
class StdOutListener(StreamListener):

    def on_data(self, data):
        print(data)
        import json
        with open('tweet_data.txt','a') as outfile:
            json.dump(data,outfile)
        return True

    def on_error(self, status):
        print(status)

if __name__ == '__main__':

    #This handles Twitter authentication and the connection to Twitter Streaming API
    l = StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, l)

    #This line filter Twitter Streams to capture data by the keywords: 'python', 'javascript', 'ruby'
    stream.filter(track=['python', 'javascript', 'ruby'])
```

TWEETS COLLECTED IN JSON FORMAT:*tweet_data.txt*

```

{"created_at":"Sat Sep 10 20:47:36 +0000
2016","id":"774710943372705794","id_str":"774710943372705794","text":"Taking Baby Steps with
NodeJS.\n 100% Gluton FREE! :) https://t.co/gn6jQJV5S0","source":{"u003ca
href=\\\"http://heroku.com\\\"
rel=\\\"nofollow\\\"","u003eadstweetbot","u003c\\a\\\"u003e"},"truncated":false,"in_reply_to_status_id":null,"in_re
ply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name"
:null,"user":{"id":"3021155140","id_str":"3021155140","name":"AdsTweetBot","screen_name":"adstweetb
ot","location":null,"url":null,"description":"Advertise your products on Us and we are happy to publish your
Ads on
Twitter.","protected":false,"verified":false,"followers_count":223,"friends_count":1361,"listed_count":34,"f
avourites_count":94,"statuses_count":40370,"created_at":"Fri Feb 06 06:50:57 +0000
2015","utc_offset":null,"time_zone":null,"geo_enabled":false,"lang":"en","contributors_enabled":false,"is
_translator":false,"profile_background_color":"CODEED","profile_background_image_url":"http://abs.twi
mg.com/images/themes/theme1/bg.png","profile_background_image_url_https":"https://abs.twimg.com/
images/themes/theme1/bg.png","profile_background_tile":false,"profile_link_color":"0084B4","profile_si
debar_border_color":"CODEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","pr
ofile_use_background_image":true,"profile_image_url":"http://abs.twimg.com/sticky/default_profile_image
s/default_profile_3_normal.png","profile_image_url_https":"https://abs.twimg.com/sticky/default_profile_
images/default_profile_3_normal.png","profile_banner_url":"https://pbs.twimg.com/profile_banners/3021
155140/1424245486","default_profile":true,"default_profile_image":true,"following":null,"follow_request_s
ent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_status
":false,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"urls":[{"url":"https://t.co/gn6jQ
JV5S0","expanded_url":"http://www.amazon.com/BEANS-Bootstrap-ExpressJS-Socket-IO-How-
JavaScript/dp/1502541149/ref=pd_ybh_1?1473540456851","display_url":"amazon.com/BEANS-
Bootstrap/2026","indices":[53,76]}],"user_mentions":[],"symbols":[],"favorited":false,"retweeted":false,"
possibly_sensitive":false,"filter_level":"low","lang":"en","timestamp_ms":"1473540456990"}r\n{"creat
ed_at":"Sat Sep 10 20:47:42 +0000
2016","id":"774710967276101637","id_str":"774710967276101637","text":"@_vanita5 Python 2 oder
3?","source":{"u003ca href=\\\"https://about.twitter.com/products/tweetdeck\\\"
rel=\\\"nofollow\\\"","u003eTweetDeck","u003c\\a\\\"u003e"},"truncated":false,"in_reply_to_status_id":7747105697
48328448,"in_reply_to_status_id_str":"774710569748328448","in_reply_to_user_id":335920521,"in_reply_to
_user_id_str":"335920521","in_reply_to_screen_name":"_vanita5","user":{"id":"970676712","id_str":"9706
76712","name":"Random Guy
\\ud83e\\udd3b\\u20e0","screen_name":"Random_Guy_32","location":"Marderheim an der
M\\u00f6lle","url":"http://randomguy32.de","description":"Ich mag Flaggen, Exklaven,
U\\u0336n\\u0336i\\u0336c\\u0336o\\u0336d\\u0336e\\u0336 und wei\\u00dfe Schokolade. Manchmal finde ich
tolle Sachen auf Wikipedia.
\\u2640?","protected":false,"verified":false,"followers_count":261,"friends_count":400,"listed_count":12,"f
avourites_count":1759,"statuses_count":51335,"created_at":"Sun Nov 25 20:01:36 +0000
2012","utc_offset":7200,"time_zone":"Berlin","geo_enabled":false,"lang":"de","contributors_enabled":fal
se,"is_translator":false,"profile_background_color":"CODEED","profile_background_image_url":"http://pb
s.twimg.com/profile_background_images/537290199202414592\\uVUjgR3A.png","profile_background_image
_url_https":"https://pbs.twimg.com/profile_background_images/537290199202414592\\uVUjgR3A.png","
profile_background_tile":true,"profile_link_color":"009B77","profile_sidebar_border_color":"FFFFFF","pro
file_sidebar_fill_color":"EFEFEF","profile_text_color":"333333","profile_use_background_image":true,"pro
file_image_url":"http://pbs.twimg.com/profile_images/755117890735595522\\FXmOsTxp_normal.jpg","pr
ofile_image_url_https":"https://pbs.twimg.com/profile_images/755117890735595522\\FXmOsTxp_normal.j
pg","profile_banner_url":"https://pbs.twimg.com/profile_banners/970676712/1459548556"}

```

JAVA CODE FOR TEXT EXTRACTION:*Tweet_Parse.java*

```
import java.io.*;
import java.io.BufferedReader;
import java.io.FileReader;
import java.util.*;
import java.util.regex.*;

public class Tweet_Parse{
    public static void main(String args[]) throws IOException {
        try (BufferedReader br = new BufferedReader(new FileReader("./tweet_data.txt"))) {
            String linef,line,result;
            //File operation for output file
            File fout = new File("./tweet_text_only.txt");
            FileOutputStream fos = new FileOutputStream(fout);
            BufferedWriter bw = new BufferedWriter(new OutputStreamWriter(fos));
            while ((linef = br.readLine()) != null) {
                //Using Regular Expression to extract text only
                Pattern p = Pattern.compile("(?=\btext\b).*?(?=\bsource\b)");
                Matcher m = p.matcher(linef);
                //List<String> matches = new ArrayList<String>();
                while (m.find()) {
                    line=m.group();
                    //System.out.println("----->" +line);
                    result = line.substring(line.indexOf("text") + 9, line.indexOf(",")-2);
                    System.out.println("<***Extracting Text only***> " +result);
                    bw.write(result);
                    bw.newLine();
                    //matches.add(result);
                }
                //System.out.println(matches);
            }
            bw.close();
        }
        catch(IOException e){
            System.out.println(e);
        }
    }
}
```

TEXT EXTRACTED FROM JSON FORMAT:*tweet_text_only.txt*

Taking Baby Steps with NodeJS.\\n 100% Gluton FREE! :) https:\\\\t.co\\gn6jQJV sS0
 @_vanita5 Python 2 oder 3?
 'The European Union Experience' Book Illuminates Controversial Director's Career #fakeheadlinebot
 #learntocode #makeatwitterbot #javascript
 fakeheadlinebot
 ruby rose \\\\/ march
 #ebay #USA #Deals #8834 0.82 Ct Round Red Created Ruby 925 Sterling Silver Ring
 https:\\\\t.co\\miGWsljntj https:\\\\t.co\\BJI3GG2gec
 ebay
 python
 javascript
 Senior Software Engineer Python @redhat Virtual USA https:\\\\t.co\\WHT0WDDri2 #Debian #Jenkins
 #Linux
 Debian
 Senior Software Engineer Python @redhat Virtual USA https:\\\\t.co\\spWQid3q1F #Debian #Jenkins
 #Linux
 Debian
 Why Ruby is an acceptable Lisp by via Hacker News https:\\\\t.co\\VP1g8JJylU
 STEPHON MARBURY 2003-04 UPPER DECK REFLECTIONS RUBY PARALLEL #55 KNICKS
 #D\\u2026 https:\\\\t.co\\DY0Y0zpnys KNICKS #D \\/500 https:\\\\t.co\\g81ewgAA7Z
 Enter to win a pyth
 Enter to win a pyth
 STEPHON MARBURY 2003-04 UPPER DECK REFLECTIONS RUBY PARALLEL #55 KNICKS #D
 \\/500 https:\\\\t.co\\KlxVcSSe9 https:\\\\t.co\\XLUdahHyr9
 RT @STYLESCONFUSA: 61 - ruby rose https:\\\\t.co\\MQPb1j306j
 61 - ruby rose https:\\\\t.co\\MQPb1j306j
 RT @RangerKarl: Congrats @bombaJB
 Congrats @bombaJB
 Teaching Python. https:\\\\t.co\\THBLL7f5bi
 Teaching Python. https:\\\\t.co\\THBLL7f5bi
 #ProgrammingLanguages\\n\\nBlack Hat Python: Python Programming for Hackers and Pentesters
 https:\\\\t.co\\id85eesGUQ #python #programming #hacki\\u2026
 ProgrammingLanguages
 Teaching Python. https:\\\\t.co\\THBLL7f5bi
 Hire Top JavaScript Talent https:\\\\t.co\\IPAVTWWMW5
 10 Python Machine Learning Projects on GitHub https:\\\\t.co\\8rJU1dMI3c
 https:\\\\t.co\\JkoUNugKPs
 THE YOUCHIKA SONG IS NOW MY FAVOURITE AQOURS SO
 You know Ruby got them cherries \\u26b0

TASK-2

MOVING TO HDFS:

Stored the text content (e.g. tweet's text) from the data into a file in HDFS.

```

C:\Barath\Software\hadoop-2.6.2>hadoop fs -ls /wordcount/input
Found 1 items
-rw-r--r-- 1 barath supergroup      1867 2016-09-11 20:45 /wordcount/input/hdfs_tweet.txt

C:\Barath\Software\hadoop-2.6.2>hadoop fs -cat /wordcount/input/hdfs_tweet.txt
Taking Baby Steps with NodeJS.\n 100% Gluten FREE! :) https://t.co/gn6jQJVs0
@_vanita5 Python 2 oder 3?
'The European Union Experience' Book Illuminates Controversial Director's Career #fakeheadlinebot #learntocode #makeatwitterbot #javascript
fakeheadlinebot
ruby rose \\\\/ march
#ebay #USA #Deals #8834 0.82 Ct Round Red Created Ruby 925 Sterling Silver Ring https://t.co/miGwsljntj https://t.co/BJI3G62gec
ebay
python
javascript
Senior Software Engineer Python @redhat Virtual USA https://t.co/WHT0WDDri2 #Debian #Jenkins #Linux
Debian
Senior Software Engineer Python @redhat Virtual USA https://t.co/spWQid3q1F #Debian #Jenkins #Linux
Debian
Why Ruby is an acceptable Lisp by via Hacker News https://t.co/NP1g8JJyIU
STEPHON MARBURY 2003-04 UPPER DECK REFLECTIONS RUBY PARALLEL #55 KNICKS #'D \u2026 https://t.co/DY0Y0zpnys KNICKS #'D \u2026 https://t.co/g81ewgAA7Z
Enter to win a pyth
Enter to win a pyth
STEPHON MARBURY 2003-04 UPPER DECK REFLECTIONS RUBY PARALLEL #55 KNICKS #'D \u2026 https://t.co/KIXVcCSse9 https://t.co/XLUdahHyr9
RT @STYLESCONFUSA: 61 - ruby rose https://t.co/MQPb1j306j
61 - ruby rose https://t.co/MQPb1j306j
RT @RangerKarl: Congrats @bombaJB
Congrats @bombaJB
Teaching Python. https://t.co/THBL7f5bi
Teaching Python. https://t.co/THBL7f5bi
#ProgrammingLanguages\n\nBlack Hat Python: Python Programming for Hackers and Pentesters https://t.co/id85eesGUQ #python #programming #hacki\u2026
ProgrammingLanguages
Teaching Python. https://t.co/THBL7f5bi
Hire Top JavaScript Talent https://t.co/IPAVTMMW5
10 Python Machine Learning Projects on GitHub https://t.co/8rJU1dM13c https://t.co/JkoUNugKPs
THE YOUNG CHICKA SONG IS NOW MY FAVOURITE AOURS SO
You know Ruby got them cherries \u26b0
C:\Barath\Software\hadoop-2.6.2>

```

TASK-3

WORD COUNT PROGRAM IN APACHE SPARK:

```
scala> val textFile = sc.textFile("hdfs://localhost:50071/wordcount/input/hdfs_tweet.txt")
textFile: org.apache.spark.rdd.RDD[String] = hdfs://localhost:50071/wordcount/input/hdfs_tweet.txt MapPartitionsRDD[17] at textFile at <console>:24

scala> val counts = textFile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(_ + _)
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[20] at reduceByKey at <console>:26

scala> counts.saveAsTextFile("hdfs://localhost:50071/wordcount/output")

scala>
```

OUTPUT:

```
cmd Command Prompt
Microsoft Windows [Version 10.0.14393]
(c) 2016 Microsoft Corporation. All rights reserved.

C:\Users\barat>cd C:\Barath\Software\hadoop-2.6.2

C:\Barath\Software\hadoop-2.6.2>hadoop fs -ls /wordcount
Found 2 items
drwxr-xr-x   - barat supergroup          0 2016-09-11 20:45 /wordcount/input
drwxr-xr-x   - barat supergroup          0 2016-09-11 21:06 /wordcount/output


C:\Barath\Software\hadoop-2.6.2>hadoop fs -ls /wordcount/output
Found 3 items
-rw-r--r--   3 barat supergroup          0 2016-09-11 21:06 /wordcount/output/_SUCCESS
-rw-r--r--   3 barat supergroup        990 2016-09-11 21:06 /wordcount/output/part-00000
-rw-r--r--   3 barat supergroup       1047 2016-09-11 21:06 /wordcount/output/part-00001

C:\Barath\Software\hadoop-2.6.2>
```

OUTPUT – PART 00000:

cmd Command Prompt

```
C:\Barath\Software\hadoop-2.6.2>
C:\Barath\Software\hadoop-2.6.2>
C:\Barath\Software\hadoop-2.6.2>
C:\Barath\Software\hadoop-2.6.2>
C:\Barath\Software\hadoop-2.6.2>hadoop fs -cat /wordcount/output/part-00000
(Created,1)
(\\/\,1)
(Python:,1)
(Python,5)
(is,1)
(Round,1)
(MARBURY,2)
(https:\\/\t.co\\DY0Y0zpnys,1)
(acceptable,1)
(#programming,1)
(0.82,1)
(Sterling,1)
(SO,1)
(https:\\/\t.co\\g81ewgAA7Z,1)
(with,1)
(2,1)
(Hire,1)
(Senior,2)
(#makeatwitterbot,1)
(#Debian,2)
(Talent,1)
(NOW,1)
('The,1)
(Ring,1)
(100%,1)
(Python.,3)
(UPPER,2)
(#8834,1)
(fakeheadlinebot,1)
(ruby,3)
(Hacker,1)
(3?,1)
(#ProgrammingLanguages\\n\\nBlack,1)
(#javascript,1)
(Taking,1)
(https:\\/\t.co\\MQPb1j306j,2)
(Learning,1)
(got,1)
(#USA,1)
(RT,2)
(via,1)
(925,1)
(IS,1)
(#ebay,1)
```


OUTPUT – PART 00001: Command Prompt

```
C:\Barath\Software\hadoop-2.6.2>hadoop fs -cat /wordcount/output/part-00001
(@_vanita5,1)
(Silver,1)
(Software,2)
(Glutton,1)
(News,1)
(rose,3)
(@STYLESCONFUSA:,1)
(FREE!,1)
(https:\\\\t.co\\miGws1jntj,1)
(Book,1)
(THE,1)
(#learntocode,1)
(Pentesters,1)
(Controversial,1)
(on,1)
(You,1)
(march,1)
(https:\\\\t.co\\THBLL7f5bi,3)
(USA,2)
(Experience',1)
(#python,1)
(Hat,1)
(https:\\\\t.co\\WHT0WDDri2,1)
(pyth,2)
(https:\\\\t.co\\spwQid3q1F,1)
(Congrats,2)
(Illuminates,1)
(ebay,1)
(-,2)
(61,2)
(for,1)
(@RangerKarl:,1)
(Engineer,2)
(#55,2)
(GitHub,1)
(javascript,1)
(Programming,1)
(SONG,1)
(Teaching,3)
(10,1)
(https:\\\\t.co\\JkoUNugKPs,1)
(@bombaJB,2)
(https:\\\\t.co\\BJI3GG2gec,1)
(#fakeheadlinebot,1)
(FAVOURITE,1)
(\\u26b0,1)
(cherries,1)
(#'D\\u2026,1)
```

OUTPUT ON SPARK:

The screenshot shows the Spark web interface for a Spark shell application. The top navigation bar includes links for Jobs, Stages, Storage, Environment, Execution, and SQL. The main content area displays 'Spark Jobs (7)' with summary statistics: 'User: root', 'Total Uptime: 26 min', 'Scheduling Mode: FIFO', and 'Completed Jobs: 1'. A 'Completed Jobs (1)' section contains a table with the following data:

Job id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	saveAsTextFile all records to /tmp	2016/09/11 21:00:25	2 s	2/2	100/100