PRINCIPLES OF BIG DATA MANAGEMENT

PROJECT PHASE 1

Submitted by,

Sneha Lagandula

Abhiram Ampabathina

Dinesh Reddy Paduru

(Team-5)

Code for Tweet Collection:

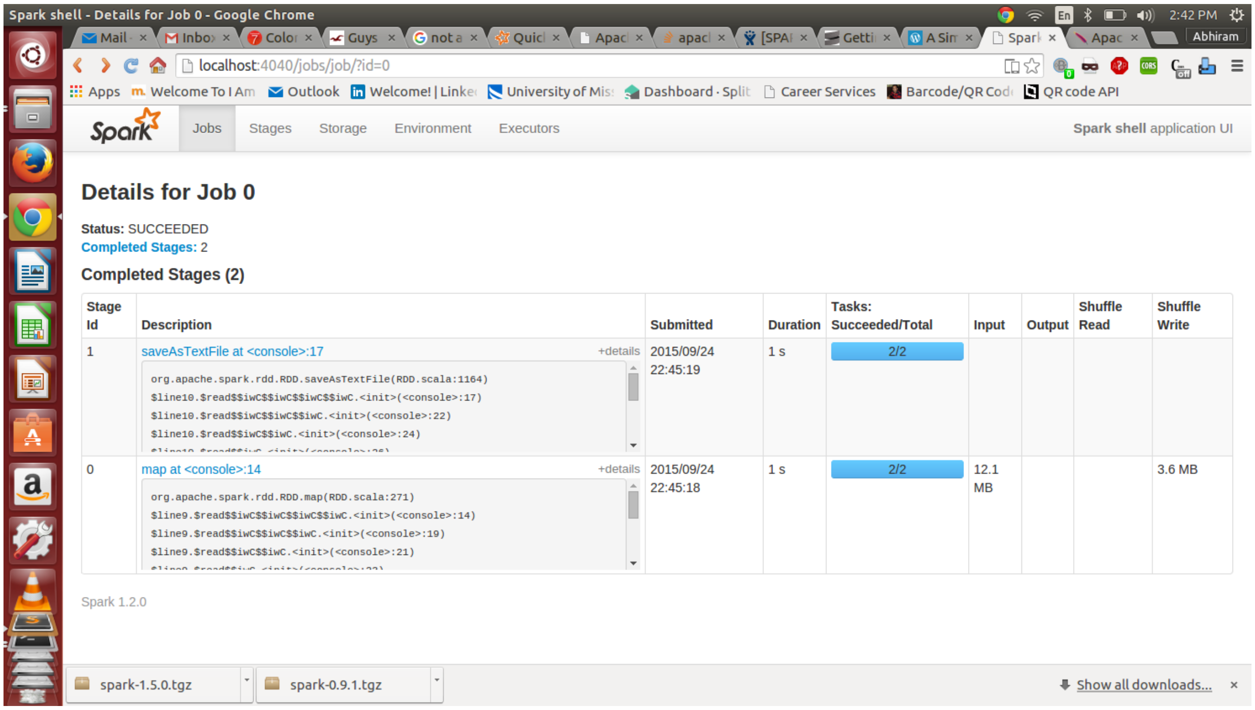
```
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream
import json, time, sys , tweepy
import re
import unicodedata
import codecs
#customer keys
consumer_key = 'yGRxlXyDG39iHr1akrN0zRU9F'
consumer_secret = '4LiY6VeolUtPPaV9dmEsh6jJvvXCAcGpHXIrbQZaxHpcc0ANMt'
access_key = '3700831395-31MrA3aG8vPRWlKzA2eu8dONZ7GhGfso2AuUe9g'
access_secret = 'ft3LSqsjk8hEe7AABYG5uga9320mLoOTZIA0yPGI6iwY1'
#auth keys
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_key, access_secret)
api = tweepy.API(auth)
#stream listener
class StdOutListener(StreamListener):
    def on_status(self, status):
        twit_id = status.id
        text = status.text
        created = status.created_at
        time_zone = status.user.time_zone
        location = status.user.location
        language = status.lang
        hashtags = re.findall(r"#(\w+)" , text)

  for tag in hashtags:
```

```
                          record = unicode(twit_id) + '\t' +  unicode(created) + '\t' + unicode(language) + '\t' +
                          unicode(location)  + '\t' + unicode(time_zone) + '\t' +  unicode(tag) + '\t' + unicode(text) +
                          '\t'

                          tweetfile = codecs.open('sampletweets4.txt' , 'a' , 'utf8')

                           tweetfile.write(unicode(record))

                          #new line for every tweet

                            tweetfile.write('\n')

                             #close tweet file

                             tweetfile.close()


                            return True


    def on_error(self, status):

        print 'Error on status', status


    def on_limit(self, status):

        print 'Limit threshold exceeded', status


    def on_timeout(self, status):

        print 'Stream disconnected; continuing...'


stream = Stream(auth, StdOutListener())

stream.filter(track = ['#NFL', '#football', '#baseball', '#royals', '#seattlemariners', '@tamba', '#Orioles',
'#Yankees', '#marlins', '#redsox', '#MLB', '@StarSports', '#FMRedHawks', '@USABaseball',
'@ForrestWhitley', '@MickeyMoniak', '@colestobbe', '#USAbaseball', '@baseball', 'Bryce Harper', 'Albert
Pujols', 'David Ortiz', 'Alex Rodriguez', 'Derek Jeter', 'Clayton Kershaw', 'Andrew McCutchen', 'Yasiel
Puig', 'Miguel Cabrera', 'Mike Trout', 'Johnny Bench', '#sports', '#kcroyals', 'pirates', 'rockies', 'sports',
'#indians', '#twins', '#gaints', '#padres', '#marlins', '#phillies', '#royals', '#mariners', '#cardinals', '#brewers',
'#mets', '#reds', '#pirates', '#rockies', '#rangers', '#athletics', '#whitesox', '#yankees', '#orioles', '#nationals',
'#dodgers', '#diamondbacks', '#rays', '#redsox', '#mariners', '#angels', '#LaMarcus Aldridge', '#Carmelo
Anthony', '#Harrison Barnes', '#Boston Celtics', '#brooklyn nets', '#houston rockets', '#new york knicks'])
```

Code for Spark Word Count:

```
val textFile = spark.textFile("sample_tweets.txt")
val counts = textFile.flatMap(line => line.split(" "))
                .map(word => (word, 1))
                .reduceByKey(_ + _)
counts.saveAsTextFile("tweets_output")
```

Spark Job:



Details for Job 0

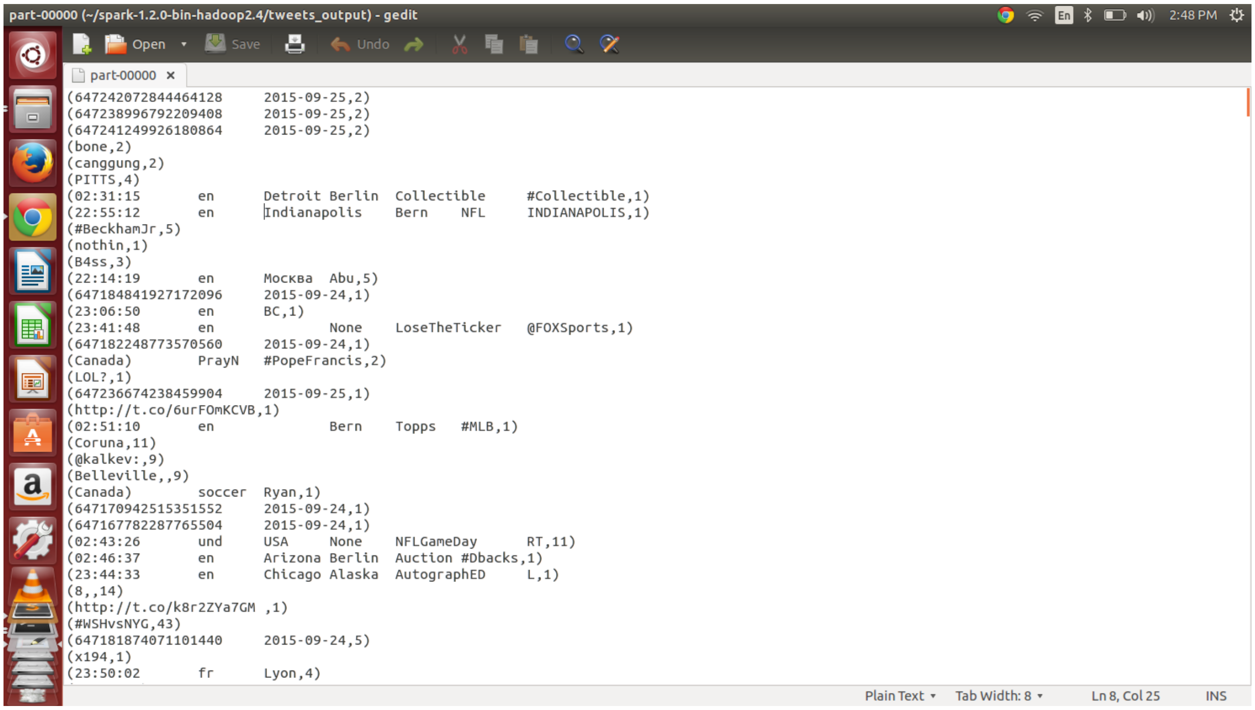
Status: SUCCEEDED
Completed Stages: 2

Completed Stages (2)

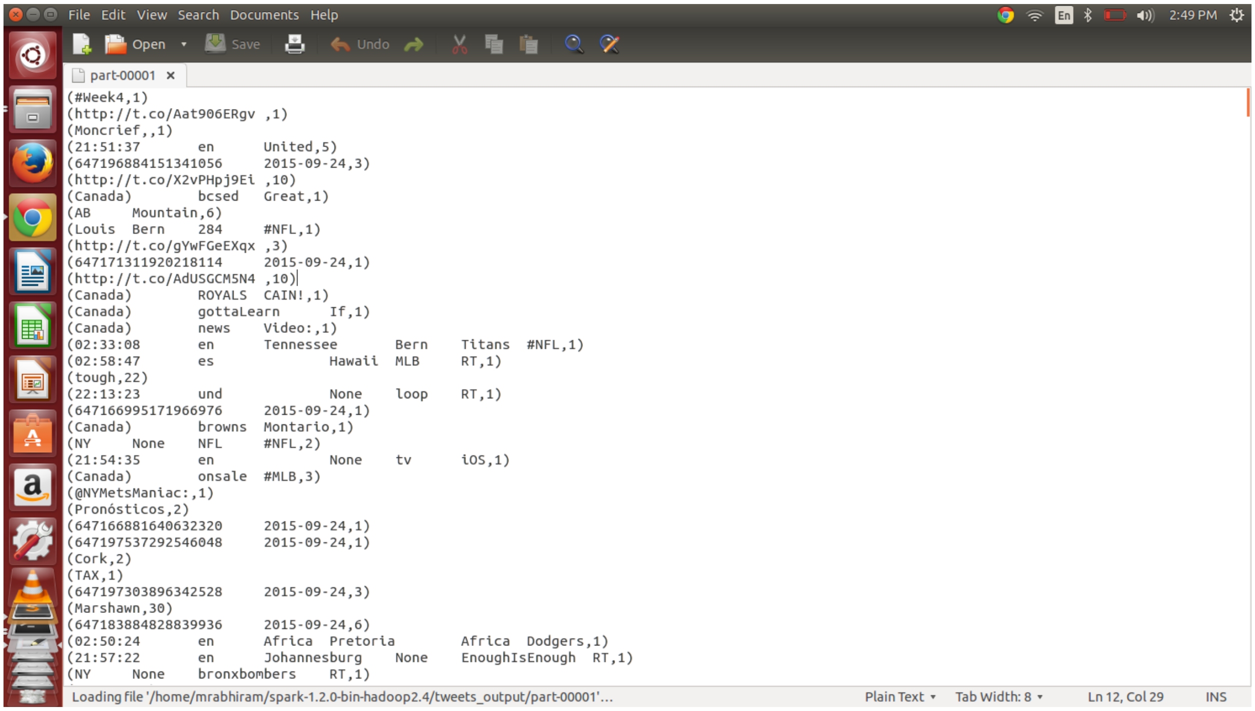
Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
1	saveAsTextFile at <console>:17 +details org.apache.spark.rdd.RDD.saveAsTextFile(RDD.scala:1164) $line10.$read$$iwC$$iwC$$iwC$$iwC.<init>(<console>:17) $line10.$read$$iwC$$iwC$$iwC.<init>(<console>:22) $line10.$read$$iwC$$iwC.<init>(<console>:24) $line10.$read$$iwC.<init>(<console>:26)	2015/09/24 22:45:19	1 s	2/2				
0	map at <console>:14 +details org.apache.spark.rdd.RDD.map(RDD.scala:271) $line9.$read$$iwC$$iwC$$iwC$$iwC.<init>(<console>:14) $line9.$read$$iwC$$iwC$$iwC.<init>(<console>:19) $line9.$read$$iwC$$iwC.<init>(<console>:21) $line9.$read$$iwC.<init>(<console>:23)	2015/09/24 22:45:18	1 s	2/2	12.1 MB			3.6 MB

Spark 1.2.0

Word Count Output:

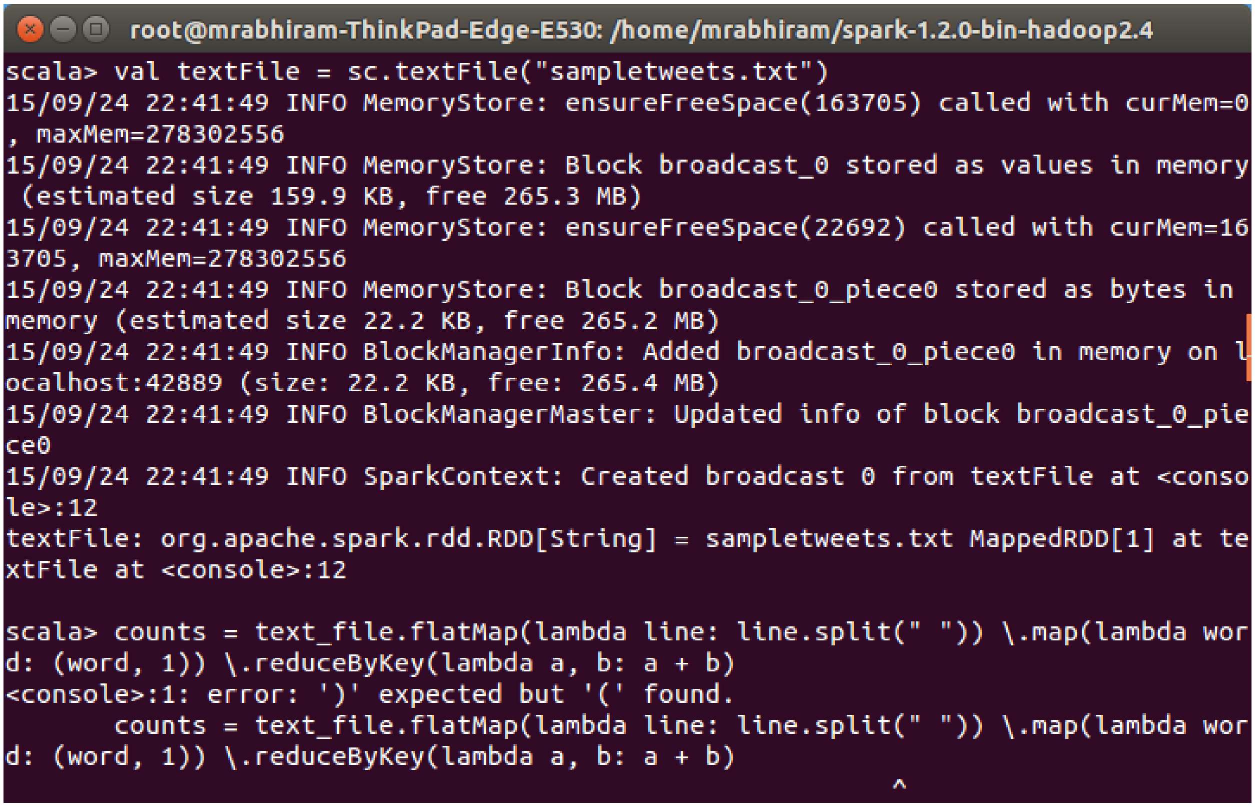


```
(647242072844464128     2015-09-25,2)
(647238996792209408     2015-09-25,2)
(647241249926180864     2015-09-25,2)
(bone,2)
(canggung,2)
(PITTS,4)
(02:31:15       en      Detroit Berlin  Collectible     #Collectible,1)
(22:55:12       en      Indianapolis    Bern    NFL     INDIANAPOLIS,1)
(#BeckhamJr,5)
(nothin,1)
(B4ss,3)
(22:14:19       en      Москва  Abu,5)
(647184841927172096     2015-09-24,1)
(23:06:50       en      BC,1)
(23:41:48       en              None    LoseTheTicker   @FOXSports,1)
(647182248773570560     2015-09-24,1)
(Canada)        PrayN   #PopeFrancis,2)
(LOL?,1)
(647236674238459904     2015-09-25,1)
(http://t.co/6urFOmKCVB,1)
(02:51:10       en              Bern    Topps   #MLB,1)
(Coruna,11)
(@kalkev:,9)
(Belleville,,9)
(Canada)        soccer  Ryan,1)
(647170942515351552     2015-09-24,1)
(647167782287765504     2015-09-24,1)
(02:43:26       und     USA     None    NFLGameDay      RT,11)
(02:46:37       en      Arizona Berlin  Auction #Dbacks,1)
(23:44:33       en      Chicago Alaska  AutographED     L,1)
(8,,14)
(http://t.co/k8r2ZYa7GM ,1)
(#WSHvsNYG,43)
(647181874071101440     2015-09-24,5)
(x194,1)
(23:50:02       fr      Lyon,4)
```

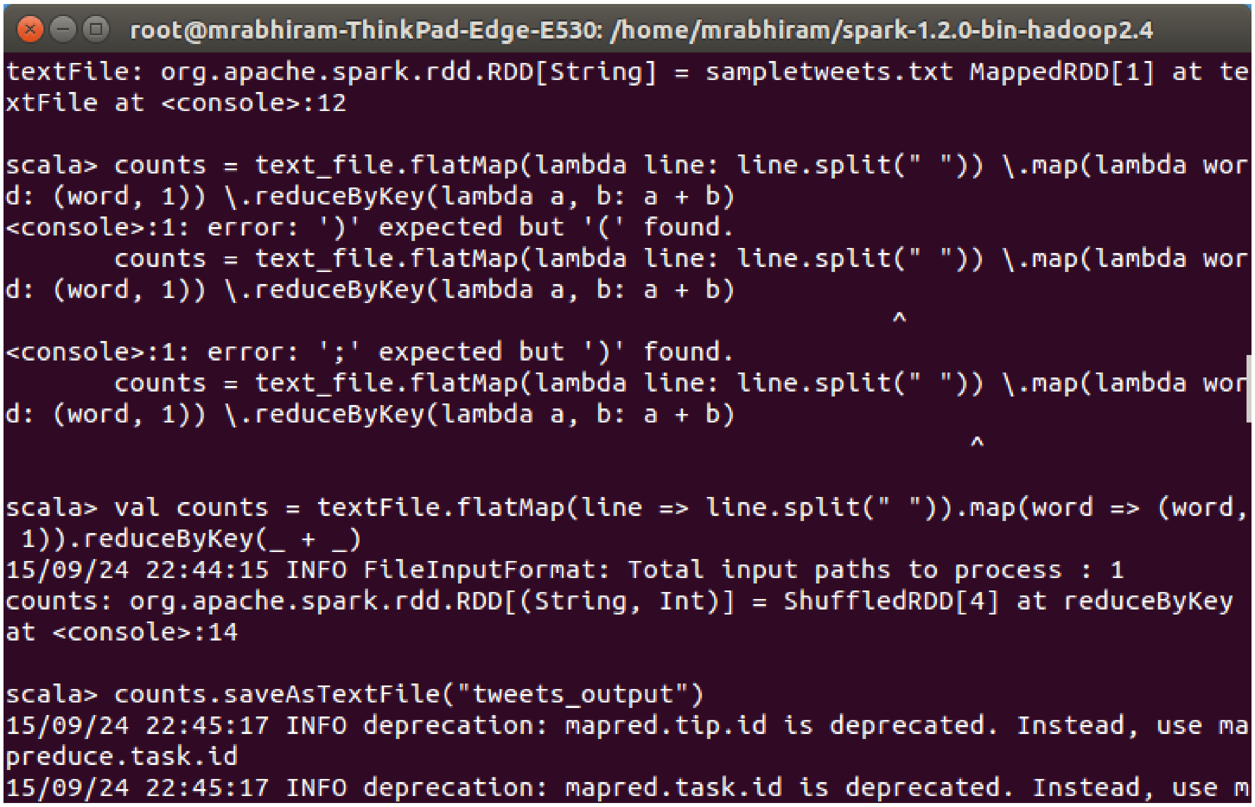


```
(#Week4,1)
(http://t.co/Aat906ERgv ,1)
(Moncrief,,1)
(21:51:37       en      United,5)
(647196884151341056     2015-09-24,3)
(http://t.co/X2vPHpj9Ei ,10)
(Canada)        bcsed   Great,1)
(AB     Mountain,6)
(Louis  Bern    284     #NFL,1)
(http://t.co/gYwFGeEXqx ,3)
(647171311920218114     2015-09-24,1)
(http://t.co/AdUSGCM5N4 ,10)
(Canada)        ROYALS  CAIN!,1)
(Canada)        gottaLearn      If,1)
(Canada)        news    Video:,1)
(02:33:08       en      Tennessee       Bern    Titans  #NFL,1)
(02:58:47       es              Hawaii  MLB     RT,1)
(tough,22)
(22:13:23       und             None    loop    RT,1)
(647166995171966976     2015-09-24,1)
(Canada)        browns  Montario,1)
(NY     None    NFL     #NFL,2)
(21:54:35       en              None    tv      iOS,1)
(Canada)        onsale  #MLB,3)
(@NYMetsManiac:,1)
(Pronósticos,2)
(647166881640632320     2015-09-24,1)
(647197537292546048     2015-09-24,1)
(Cork,2)
(TAX,1)
(647197303896342528     2015-09-24,3)
(Marshawn,30)
(647183884828839936     2015-09-24,6)
(02:50:24       en      Africa  Pretoria        Africa  Dodgers,1)
(21:57:22       en      Johannesburg    None    EnoughIsEnough  RT,1)
(NY     None    bronxbombers    RT,1)
```

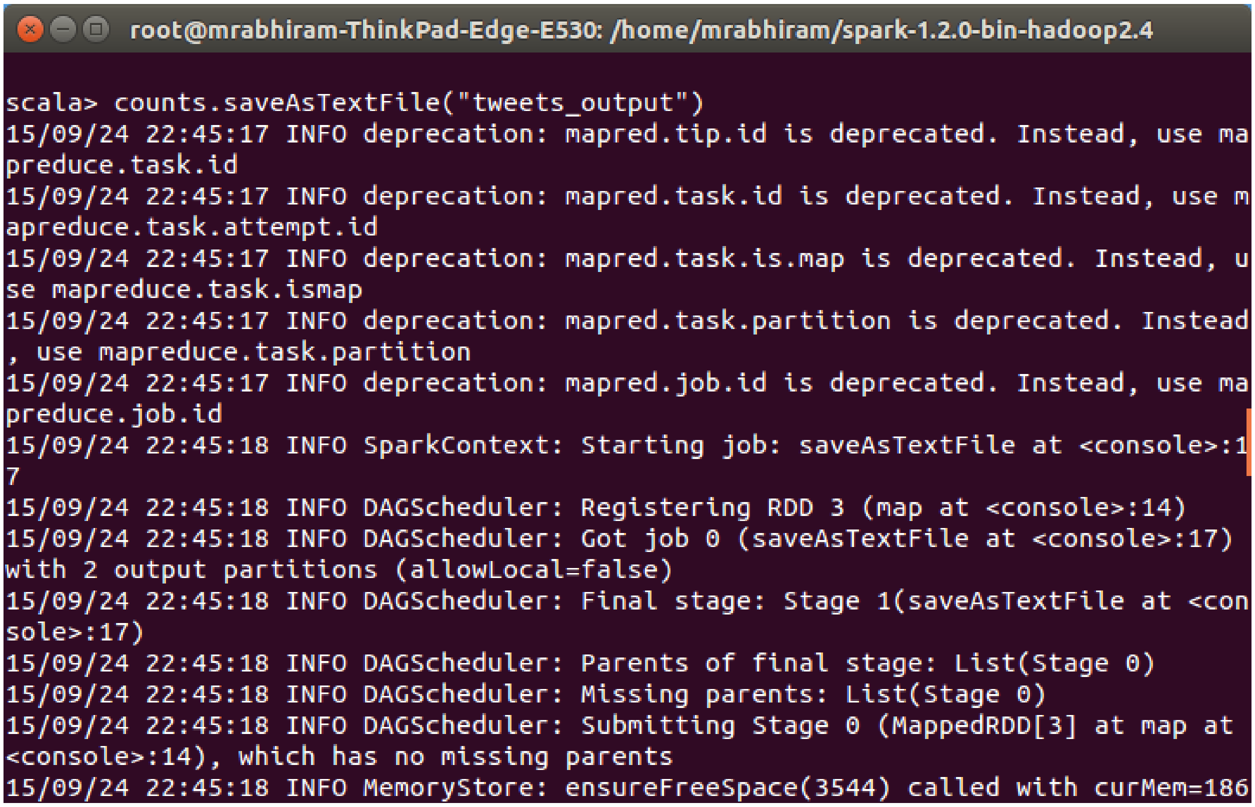
Word Count Execution:


root@mrabhiram-ThinkPad-Edge-E530: /home/mrabhiram/spark-1.2.0-bin-hadoop2.4

```
scala> val textFile = sc.textFile("sampletweets.txt")
15/09/24 22:41:49 INFO MemoryStore: ensureFreeSpace(163705) called with curMem=0
, maxMem=278302556
15/09/24 22:41:49 INFO MemoryStore: Block broadcast_0 stored as values in memory
 (estimated size 159.9 KB, free 265.3 MB)
15/09/24 22:41:49 INFO MemoryStore: ensureFreeSpace(22692) called with curMem=16
3705, maxMem=278302556
15/09/24 22:41:49 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in
memory (estimated size 22.2 KB, free 265.2 MB)
15/09/24 22:41:49 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on l
ocalhost:42889 (size: 22.2 KB, free: 265.4 MB)
15/09/24 22:41:49 INFO BlockManagerMaster: Updated info of block broadcast_0_pie
ce0
15/09/24 22:41:49 INFO SparkContext: Created broadcast 0 from textFile at <conso
le>:12
textFile: org.apache.spark.rdd.RDD[String] = sampletweets.txt MappedRDD[1] at te
xtFile at <console>:12

scala> counts = text_file.flatMap(lambda line: line.split(" ")) \.map(lambda wor
d: (word, 1)) \.reduceByKey(lambda a, b: a + b)
<console>:1: error: ')' expected but '(' found.
       counts = text_file.flatMap(lambda line: line.split(" ")) \.map(lambda wor
d: (word, 1)) \.reduceByKey(lambda a, b: a + b)
                                                         ^
```


root@mrabhiram-ThinkPad-Edge-E530: /home/mrabhiram/spark-1.2.0-bin-hadoop2.4

```
textFile: org.apache.spark.rdd.RDD[String] = sampletweets.txt MappedRDD[1] at te
xtFile at <console>:12

scala> counts = text_file.flatMap(lambda line: line.split(" ")) \.map(lambda wor
d: (word, 1)) \.reduceByKey(lambda a, b: a + b)
<console>:1: error: ')' expected but '(' found.
       counts = text_file.flatMap(lambda line: line.split(" ")) \.map(lambda wor
d: (word, 1)) \.reduceByKey(lambda a, b: a + b)
                                                         ^

<console>:1: error: ';' expected but ')' found.
       counts = text_file.flatMap(lambda line: line.split(" ")) \.map(lambda wor
d: (word, 1)) \.reduceByKey(lambda a, b: a + b)
                                                              ^


scala> val counts = textFile.flatMap(line => line.split(" ")).map(word => (word,
 1)).reduceByKey(_ + _)
15/09/24 22:44:15 INFO FileInputFormat: Total input paths to process : 1
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[4] at reduceByKey
at <console>:14

scala> counts.saveAsTextFile("tweets_output")
15/09/24 22:45:17 INFO deprecation: mapred.tip.id is deprecated. Instead, use ma
preduce.task.id
15/09/24 22:45:17 INFO deprecation: mapred.task.id is deprecated. Instead, use m
```


root@mrabhiram-ThinkPad-Edge-E530: /home/mrabhiram/spark-1.2.0-bin-hadoop2.4

```
scala> counts.saveAsTextFile("tweets_output")
15/09/24 22:45:17 INFO deprecation: mapred.tip.id is deprecated. Instead, use ma
preduce.task.id
15/09/24 22:45:17 INFO deprecation: mapred.task.id is deprecated. Instead, use m
apreduce.task.attempt.id
15/09/24 22:45:17 INFO deprecation: mapred.task.is.map is deprecated. Instead, u
se mapreduce.task.ismap
15/09/24 22:45:17 INFO deprecation: mapred.task.partition is deprecated. Instead
, use mapreduce.task.partition
15/09/24 22:45:17 INFO deprecation: mapred.job.id is deprecated. Instead, use ma
preduce.job.id
15/09/24 22:45:18 INFO SparkContext: Starting job: saveAsTextFile at <console>:1
7
15/09/24 22:45:18 INFO DAGScheduler: Registering RDD 3 (map at <console>:14)
15/09/24 22:45:18 INFO DAGScheduler: Got job 0 (saveAsTextFile at <console>:17)
with 2 output partitions (allowLocal=false)
15/09/24 22:45:18 INFO DAGScheduler: Final stage: Stage 1(saveAsTextFile at <con
sole>:17)
15/09/24 22:45:18 INFO DAGScheduler: Parents of final stage: List(Stage 0)
15/09/24 22:45:18 INFO DAGScheduler: Missing parents: List(Stage 0)
15/09/24 22:45:18 INFO DAGScheduler: Submitting Stage 0 (MappedRDD[3] at map at
<console>:14), which has no missing parents
15/09/24 22:45:18 INFO MemoryStore: ensureFreeSpace(3544) called with curMem=186
```