

Principles of Big Data Management

Phase 1 Report

Team :

Murali krishna sai Chukka (16272588)

Sushma Manne (16271112)

Pujitha Koppanati (16273485)

Links :

https://github.com/chkrish9/PB_Project/tree/master/

Task 1

Collect tweets using twitters streaming API(tweepy)

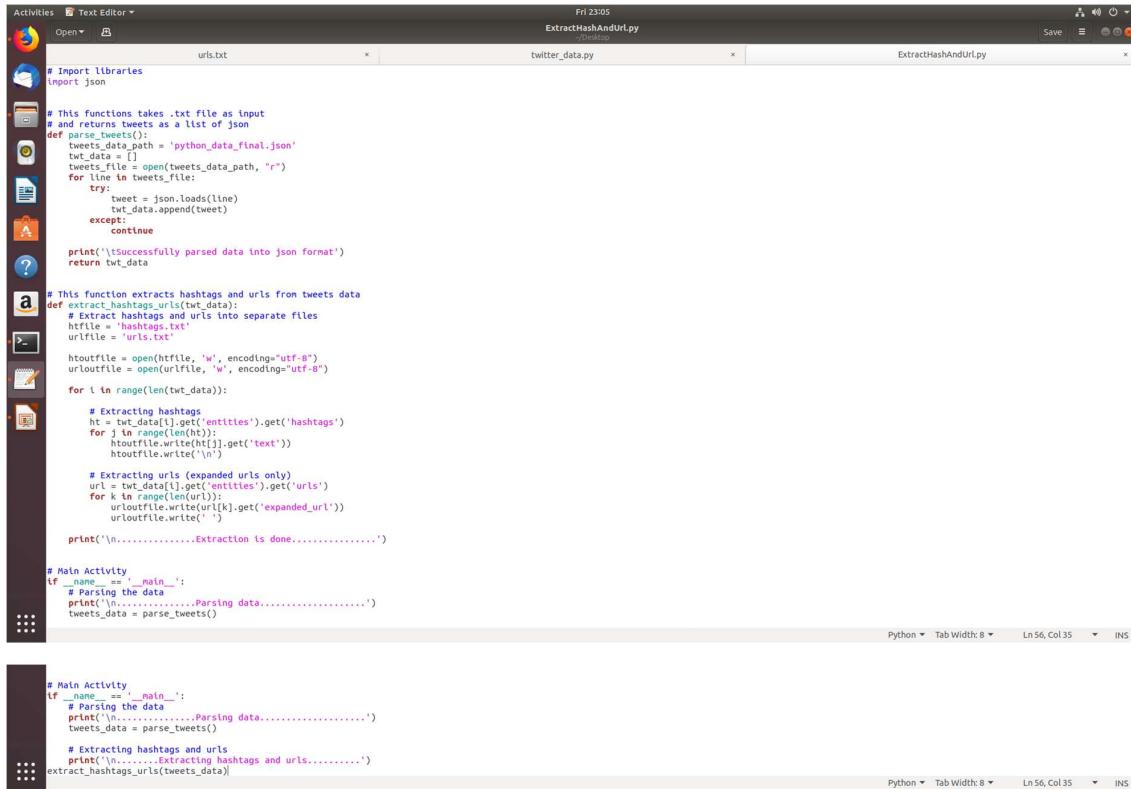
Python code for collecting tweets :

The above python code collects tweets into a text file by using Twitter's streaming API Tweepy. In order to connect to the Twitter's Streaming API, we need to authenticate using credentials from our twitter developer account. The program filters data by "Tesla" Keyword. The data is appended to file 'twitter_data.py'. We have collected around tweets.

Output file generated is of size 53MB

Task 2

Hashtags and URLs extraction from the twitter data



```
# Import libraries
import json

# This function takes .txt file as input
# and returns tweets as a list of json
def parse_tweets():
    tweets_data_path = 'python_data_final.json'
    twt_data = []
    tweets_file = open(tweets_data_path, "r")
    for line in tweets_file:
        try:
            tweet = json.loads(line)
            twt_data.append(tweet)
        except:
            continue
    print('Successfully parsed data into json format')
    return twt_data

# This function extracts hashtags and urls from tweets data
def extract_hashtags_urls(twt_data):
    # Extract hashtags and urls into separate files
    htfile = 'hashtags.txt'
    urlfile = 'urls.txt'

    htoutfile = open(htfile, 'w', encoding="utf-8")
    urloutfile = open(urloutfile, 'w', encoding="utf-8")

    for i in range(len(twt_data)):
        # Extracting hashtags
        ht = twt_data[i].get('entities').get('hashtags')
        for j in range(len(ht)):
            htoutfile.write(ht[j].get('text'))
            htoutfile.write('\n')

        # Extracting urls (expanded urls only)
        url = twt_data[i].get('entities').get('urls')
        for k in range(len(url)):
            urloutfile.write(url[k].get('expanded_url'))
            urloutfile.write('\n')

    print('\n.....Extraction is done.....')

# Main Activity
if __name__ == '__main__':
    # Parsing the data
    print('.....Parsing data.....')
    tweets_data = parse_tweets()

    # Extracting hashtags and urls
    print('.....Extracting hashtags and urls.....')
    extract_hashtags_urls(tweets_data)
```

The above program parses tweets file into json format and extracts required Hashtags and urls and writes to two different files.

Outputfiles :

urlsoutput

Hashtagsoutput



```
1 eduardofandinho
2 genzop
3 teslaFestival
4 teslaFestival2019
5 IsraeltoTheKon
6 tesla
7 tecnologia
8 science
9 technology
10 AUTOPILOT
11 autopilot
12 TeslaModel3
13 Tesla
14 klimaatrild
15 klassenstrid
16 Wunderwaffe
17 Kaufempfehlung
18 Tesla
19 ThirtyThursday
20 VeChain
21 latestComments
22 klimaatrild
23 klassenstrid
24 Tesla
25 teslaairbands
26 N
27 Tesla
28 China
29 Tesla
30 China
31 Tesla
32 ElonMusk
33 Tesla
34 BreakingNews
35 onmfc
36 CleanEnergyWillWin
37 ElonMusk
38 Tesla
39 Tesla
40 US
41 TIC
42 USA
43 US
44 USBiz
45 Tesla
46 CNBC
47 BusinessNews
48 ValorDelWeb
49 CNBC
50 BusinessNews
```

Task 3

WordCount Program Using Apache Hadoop.

Log generated by running above program on both hashtags file and url file

```
Activities Terminal Fri 23:22 hadoopusr@ubuntu:/home/kite/hadoop/share/hadoop/mapreduce
File Edit View Search Terminal Help
  at org.apache.hadoop.util.RunJar.main(RunJar.java:148)
hadoop@ubuntu:[/home/kite/hadoop/share/hadoop/mapreduce] hadoop jar ./hadoop-mapreduce-examples-2.8.1.jar wordcount /pb_project/hashtags.txt /pb_project/hashtags_result1.txt>>/home/kite/Desktop/hadoop.log
WARNING: An illegal reflective access operation has occurred
WARNING: illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.8.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings on other illegal reflective access operations
19/02/22 23:22:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable
19/02/22 23:22:31 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
19/02/22 23:22:31 INFO jvm.JvmMetrics: Initializing JVM Metrics with processname='jobTracker', sessionid='local1897427218_0001'
19/02/22 23:22:31 INFO mapred.JobClient: Total memory for this process : 1
19/02/22 23:22:31 INFO mapreduce.JobSubmitter: number of splits:1
19/02/22 23:22:31 INFO mapreduce.JobSubmitter: Submitting tokens for job _job_local1897427218_0001
19/02/22 23:22:32 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
19/02/22 23:22:32 INFO mapreduce.Job: number of splits:1
19/02/22 23:22:32 INFO mapred.LocalJobRunner: OutputCommitter set in config null
19/02/22 23:22:32 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
19/02/22 23:22:32 INFO mapred.LocalJobRunner: OutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
19/02/22 23:22:32 INFO mapred.LocalJobRunner: OutputCommitter class is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
19/02/22 23:22:32 INFO mapred.LocalJobRunner: Waiting for map tasks
19/02/22 23:22:32 INFO mapred.LocalJobRunner: Starting task: attempt_local1897427218_0001_m_000000_0
19/02/22 23:22:32 INFO mapred.FileOutputCommitter: File Output Committer Algorithm version is 1
19/02/22 23:22:32 INFO mapred.Task: TaskAttempted: attempt_local1897427218_0001_m_000000_0
19/02/22 23:22:32 INFO mapred.Task: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
19/02/22 23:22:32 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/pb_project/hashtags.txt+0+27780
19/02/22 23:22:32 INFO mapred.MapTask: (EQUATOR) o kvl: 26214396(104857584)
19/02/22 23:22:32 INFO mapred.MapTask: mapreduce.task.o.sort.mb: 100
19/02/22 23:22:32 INFO mapred.MapTask: mapreduce.task.o.map: 1
19/02/22 23:22:32 INFO mapred.MapTask: buffer_size = 0; bufvld = 104857600
19/02/22 23:22:32 INFO mapred.MapTask: kvstart = 0x26214396; length = 6553600
19/02/22 23:22:32 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
19/02/22 23:22:33 INFO mapreduce.Job: map 0% reduce 0%
19/02/22 23:22:33 INFO mapred.LocalJobRunner: mapreduce.job.map
19/02/22 23:22:33 INFO mapred.MapTask: Starting flush of map output
19/02/22 23:22:33 INFO mapred.MapTask: mapattempt_local1897427218_0001_m_000000_0
19/02/22 23:22:33 INFO mapred.MapTask: buffer_size = 42754; bufvld = 104857600
19/02/22 23:22:33 INFO mapred.MapTask: kvstart = 0x26214396(104857584); kvend = 0x2621104(104804416); length = 13293/6553600
19/02/22 23:22:33 INFO mapred.Task: TaskAttempted: attempt_local1897427218_0001_m_000000_0 is done. And is in the process of committing
19/02/22 23:22:33 INFO mapred.Task: TaskAttempted: attempt_local1897427218_0001_m_000000_0
19/02/22 23:22:33 INFO mapred.LocalJobRunner: map task executor complete.
19/02/22 23:22:33 INFO mapred.LocalJobRunner: Starting task: attempt_local1897427218_0001_r_000000_0
19/02/22 23:22:33 INFO mapred.LocalJobRunner: Starting task: attempt_local1897427218_0001_r_000000_0
19/02/22 23:22:33 INFO mapred.FileOutputCommitter: File Output Committer Algorithm version is 1
19/02/22 23:22:33 INFO mapred.FileOutputCommitter: File Output Committer skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
19/02/22 23:22:33 INFO mapred.Task: TaskAttempted: attempt_local1897427218_0001_r_000000_0
19/02/22 23:22:33 INFO mapred.ReduceManagerImpl: ReduceManager: memoryLimit=37589632, maxSingleShuffleLimit=2480343d8, mergeThreshold=2480343d8, ioSortFactor=10, memToMemMergeOutputsThreshold=10
19/02/22 23:22:33 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=37589632, maxSingleShuffleLimit=2480343d8, mergeThreshold=2480343d8, ioSortFactor=10, memToMemMergeOutputsThreshold=10
19/02/22 23:22:33 INFO reduce.ReduceLocality: localityFilter#1 about to shuffle task attempt_local1897427218_0001_r_000000_0
19/02/22 23:22:33 INFO reduce.ReduceLocality: localityFilter#1 about to attempt_local1897427218_0001_r_000000_0
Activities Terminal Fri 23:22 hadoopusr@ubuntu:/home/kite/hadoop/share/hadoop/mapreduce
File Edit View Search Terminal Help
hadoop@ubuntu:[/home/kite/hadoop/share/hadoop/mapreduce]
19/02/22 23:22:34 INFO mapreduce.Job: map 100% reduce 0%
19/02/22 23:22:34 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
19/02/22 23:22:34 INFO mapred.Task: TaskAttempted: attempt_local1897427218_0001_r_000000_0 is done. And is in the process of committing
19/02/22 23:22:34 INFO mapred.LocalJobRunner: 1 / 1 copied.
19/02/22 23:22:34 INFO mapred.Task: TaskAttempted: attempt_local1897427218_0001_r_000000_0 is allowed to commit now
19/02/22 23:22:34 INFO mapred.FileOutputCommitter: Saved output to hdfs://localhost:9000/pb_project/hashtags_result1.txt/_temporary/0/task_local1897427218_0001_r_000000_0
001_r_000000_0
19/02/22 23:22:34 INFO mapred.LocalJobRunner: reduce > reduce
19/02/22 23:22:34 INFO mapred.Task: TaskAttempted: attempt_local1897427218_0001_r_000000_0 done.
19/02/22 23:22:34 INFO mapred.LocalJobRunner: finishing task: attempt_local1897427218_0001_r_000000_0
19/02/22 23:22:34 INFO mapred.LocalJobRunner: reduce task executor complete.
19/02/22 23:22:35 INFO mapreduce.Job: map 100% reduce 100%
19/02/22 23:22:35 INFO mapreduce.Job: Job job_local1897427218_0001 completed successfully
19/02/22 23:22:35 INFO mapreduce.Job: Job counters: 35
  File System Counters
    FILE: Number of bytes read=634270
    FILE: Number of bytes written=1380280
    FILE: Number of bytes read=1380280
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=6553600
    HDFS: Number of bytes written=1322
    HDFS: Number of read operations=13
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=4
  Map-Reduce Framework
    Map output materialized bytes=15018
    Input File blocks=116
    Combine Input records=3324
    Combine output records=958
    Reduce Input groups=958
    Reduce Input file blocks=1518
    Reduce Input records=958
    Reduce output records=958
    Spilled Records=1916
    spilled Maps =0
    Failed Maps =0
    Merged Map outputs=1
    GC time elapsed (ms)=14
    total committed heap usage (bytes)=429916160
  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=12780
  File Output Format Counters
    Bytes Written=11222
hadoop@ubuntu:[/home/kite/hadoop/share/hadoop/mapreduce]
```


Output file showing hashtags, urls and their counts

File	Edit	View	Help
127881pacesyndicate	127881illion	5369Model3	47journeys7couvertures
AmericanRuckSimulator	1Android	1Antonopoulos	1Anymore
Autobahn	1Autobahn	1Autobahn	1Automotive
BabuBabu	1BabuBabu	1BabuBabu	1BabuBabu
Bloomberg	5BodyGuard	8Boeing4FriggleNuggets	48Busted
Calmbase	1Calmding	1CarChargingCompany	1Cartech
DriverlessVehicles	3Dumbus	1Dumbus	1Dumbus
ElectricVehicle	1ElectrictVehicles	6Electricity	1Elektroauto
Executive	3ExtremeMaster	2F	1FCC
GerryLynnWichmann	1Gigafactory	7Gems1me	461in1STEN
Hermes	1Heroes	3HitterHoldings	1HittinBeats1JillithHoldings
Invisible	3Ivys	1Ivys	1Ivys
Luxembourg	1LUXEM	1LUXEM	1LUXEM
Mobility	1Mobile1	1Mobile1	1Mobile1
Numbers	1Numerics	2011	1P3851PREVENT
Pixoxiamatics	1Python	1QuestQuicksilvers	1Quotes1QutesPortion
RoyalGold	45	15SPride	15SSSA
Sandiego	1Sandiego	1Sandiego	1Sandiego
3THREEREPID	1SOLARIS	1Sorry1SoundCloud	1Spotify1Spotify
SteslaChina	1SteslaFestival2019	1SteslaModel3	1SteslaModel3
XPKardot	1US	2US12	2VIXC
Xpeng	1XpengG3	1YU	2Youtubeh
Iaudinovius	1audinovius	1autodesk	1autodesk
Autodesk	1autodesk	2autodesk	1autodesk
Bullish	2Buseconnects	2Business	4Business1Slews
2Chness	3Chness	3ClassicalMetalradio	4Climatchange
Icostarican	1couple1Icostarican	1createor	1crypt0
Electric	1electrician	1electricians	1electricity
elektromodiano	1elonmusk	8emprende	1energy1EnergyStorage
faraday	1efeff2	2ff	2finanzen
for1founders	1finances1fractales	1finances1fractales	1finances1fractales
german	1gettingready	1gigafactory	1gigahbz
hotelsforinstgram	1hotelsHydrogen	1hyperloop	31heatAwards
Injustice_in_ua	1Injustice	9Innovate	21Innovate
Instastory	1instastory	1jimenez1jimenez	1instastory1jimenez
2late	11layered	1like1likelikes	1likes
manufacturingK	1marketing	1metal1metals	3minchies
Investor	1offer1investor	1imphorus1investor	1investor1investor
IPhone	1ipublico	1ipumping	1ipundtorecharge1ipuntrate
retweet4reuters	1irewels	1irewels1iroadtrip	1iroadtrip1iroadtrip
Isawthatnthon	1isaws1isawthatnthon	1isawthatnthon	1isawthatnthon1isawthatnthon
2seaspace	2seaspace	2seaspace1seaspace	2seaspace1seaspace
Teknikers	1tekniken	1tech1techcrunch	1technology1technology
festivals	2teslaFestival2019	2teslahas	3teslamodels
Trade1Itradearm	8trading	1transportation	1trucks1truck1trucks1tsqsl
vehicle	1vehiclesElectrics	1vehiculoselectrics	1veskappa1veskappa

Here Map function gets input as byte offset as key and each line as value. StringTokenizer breaks each line into words splitting with space character. Each word in all documents is combined with an integer value one and sent to combiner and reducer as value with the key as word. Each Reducer takes Similar key values from the mapper, and counts all the values of unique key and send it as output value. The result from the reducer is word and the number of its occurrences from all the documents.

WordCount Program Using Apache Spark

```
Activities Terminal Fri 23:31 hadoopuser@ubuntu: /home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/bin
File Edit View Search Terminal Help
19/02/02 23:29:44 INFO BlockManagerMaster: BlockManager stopped
19/02/02 23:29:44 INFO BlockManagerMaster: BlockManagerMaster stopped
19/02/02 23:29:44 INFO OutputCommitCoordinator$OutputCommitterEndpoint: OutputCommitCoordinator stopped!
19/02/02 23:29:44 INFO ShutdownHookManager: Successfully stopped SparkContext
19/02/02 23:29:44 INFO ShutdownHookManager: Shutdown hook called
19/02/02 23:29:44 INFO ShutdownHookManager: Deleting directory /tmp/.spark-submit/run-example_JavadocCount /home/kite/Desktop/pb/hashtags.txt>>/home/kite/Desktop/hashtags_output.txt
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
19/02/02 23:30:32 INFO SparkContext: Running Spark version 2.2.0
19/02/02 23:30:34 WARN NativeCodeLoader: Unable to load native library for your platform... using builtin-Java classes where applicable
19/02/02 23:30:34 WARN NativeCodeLoader: Native code for your platform was not found, so falling back to Java classes
19/02/02 23:30:34 WARN Util: Set SPARK_LOCAL_IP if you need to bind to another address
19/02/02 23:30:34 INFO SparkContext: Submitted application: JavadocCount
19/02/02 23:30:34 INFO SecurityManager: Changing view acls to 'hadoopuser'
19/02/02 23:30:34 INFO SecurityManager: Setting default view acls to 'hadoopuser'
19/02/02 23:30:34 INFO SecurityManager: Changing view acls groups to ''
19/02/02 23:30:34 INFO SecurityManager: Changing modify acls groups to ''
19/02/02 23:30:35 INFO SecurityManager: SecurityManager: authentication disabled; user acls disabled; users with view permissions: Set(hadoopuser); groups with view permissions: Set(); users with modify permissions: Set(); groups with modify permissions: Set()
19/02/02 23:30:35 INFO Util: Successfully started service 'sparkDriver' on port 38481.
19/02/02 23:30:34 INFO SparkEnv: Registering MapOutputTracker
19/02/02 23:30:34 INFO SparkEnv: Registering BlockManagerMaster
19/02/02 23:30:34 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
19/02/02 23:30:34 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-efd273b-c233-4eb3-a5b-2eebf84540875
19/02/02 23:30:35 INFO MemoryStore: MemoryStore started with capacity 366.3 MB
19/02/02 23:30:35 INFO SparkContext: Bound sparkURL to 0.0.0.0, and started at http://192.168.190.130:4040
19/02/02 23:30:35 INFO SparkContext: Added JAR file:/home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/examples/jars/scopt_2_11-3.3.0.jar at spark://192.168.190.130:38481/jars/scopt_2_11-3.3.0.jar with timestamp 15559353787
19/02/02 23:30:35 INFO SparkContext: Added JAR file:/home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/examples/jars/spark-examples_2_11-2.0.jar at spark://192.168.190.130:38481/jars/spark-examples_2_11-2.0.jar with timestamp 15559353788
19/02/02 23:30:35 INFO Executor: Starting executor ID driver on host localhost
19/02/02 23:30:36 INFO Util: Successfully started service org.apache.spark.blockTransfer.netty.NettyBlockTransferService' on port 39463.
19/02/02 23:30:36 INFO BlockManager: Registering block manager driver, 192.168.190.130, 39463, None
19/02/02 23:30:36 INFO BlockManagerMasterEndpoint: Registering block manager 192.168.190.130:39463 with 366.3 MB RAM, BlockManagerId(driver, 192.168.190.130, 39463, None)
19/02/02 23:30:36 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.190.130, 39463, None)
19/02/02 23:30:36 INFO SharedState: Hive metastore warehouse dir ('null') to the value of spark.sql.warehouse.dir ('file:/home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/bin/spark-warehouse').
19/02/02 23:30:36 INFO SharedState: Warehouse path is 'file:/home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/bin/spark-warehouse'.
19/02/02 23:30:43 INFO FileSourceStrategy: Pre-load directories with:
19/02/02 23:30:43 INFO FileSourceStrategy: Post-Scan Filters:
19/02/02 23:30:43 INFO FileSourceStrategy: Output Data Schema: struct<value: string>
19/02/02 23:30:45 INFO CodeGenerator: Code generated: 66,141,536 ns
19/02/02 23:30:45 INFO MemoryStore: Block broadcast_0 stored as bytes in memory (estimated size 277.3 KB, free 366.0 MB)
19/02/02 23:30:45 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 23.4 KB, free 366.0 MB)
19/02/02 23:30:45 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on 192.168.190.130:39463 (size: 23.4 KB, free: 366.3 MB)
19/02/02 23:30:45 INFO BlockManager: Registering block broadcast_0_piece0 at /tmp/blockmgr-efd273b-c233-4eb3-a5b-2eebf84540875
19/02/02 23:30:45 INFO FileSourceScanner: Planning scan with bin packing, max size: 4194304 bytes, open cost is considered as scanning 4194304 bytes.
19/02/02 23:30:46 INFO SparkContext: Starting job: collect at JavadocCount.java:53
```

Fri 23:30:46 INFO DAGScheduler: Submitting ShuffleMapStage 0 (MapPartitionsRDD[5] at mapToPair at JavaWordCount.java:49), which has no missing parents

19/02/22 23:30:46 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 12.5 KB, free 366.0 MB)

19/02/22 23:30:46 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on 192.168.190.130:44193 (size: 12.5 KB, free 366.0 MB)

19/02/22 23:30:46 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1008

19/02/22 23:30:46 INFO TaskSetManager: Partition 0 in stage 0 (ID 0, localhost, executor driver, partition 0, PROCESS_LOCAL, 5267 bytes)

19/02/22 23:30:47 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)

19/02/22 23:30:47 INFO Executor: Fetching spark://192.168.190.130:38481/jars/scopt_2.11-3.3.0.jar with timestamp 1550899835787

19/02/22 23:30:47 INFO Utils: Fetching spark://192.168.190.130:38481/jars/spark-examples_2.11-2.2.0.jar to /tmp/spark-53a0cae-d3a5-4772-83de-1d7674e53798/userFiles-01bcd44e-2678-4251-a799-6aacb0c5f8d2/fetchFileTemp3126435986883381219.tmp

19/02/22 23:30:47 INFO Executor: Adding file:/tmp/spark-53a0cae-d3a5-4772-83de-1d7674e53798/userFiles-01bcd44e-2678-4251-a799-6aacb0c5f8d2/scopt_2.11-3.3.0.jar to class loader

19/02/22 23:30:47 INFO Utils: Fetching spark://192.168.190.130:38481/jars/spark-examples_2.11-2.2.0.jar with timestamp 1550899835788

19/02/22 23:30:47 INFO Utils: Fetching spark://192.168.190.130:38481/jars/spark-examples_2.11-2.2.0.jar to /tmp/spark-53a0cae-d3a5-4772-83de-1d7674e53798/userFiles-01bcd44e-2678-4251-a799-6aacb0c5f8d2/feTchfileTemp01316490766688361.tmp

19/02/22 23:30:47 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 0 (MapPartitionsRDD[5] at mapToPair at JavaWordCount.java:49) (first 15 tasks are for partitions Vector(0))

19/02/22 23:30:47 INFO TaskSetManager: Partition 0 in stage 0 (ID 0, localhost, executor driver, partition 0, PROCESS_LOCAL, 5267 bytes)

19/02/22 23:30:47 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0)

19/02/22 23:30:47 INFO FileCacheDDB: Added broadcast 1 to FileCacheDDB: File path: file:///home/kite/Desktop/pb/tags.txt, range: 0-32780, partition values: [empty row]

19/02/22 23:30:47 INFO DAGScheduler: CodeGenerator: Code generated in 55.722835 ms

19/02/22 23:30:48 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0) - 1756 bytes result sent to driver

19/02/22 23:30:48 INFO TaskSetManager: Task 0.0 finished; collecting results from 1 task(s) in 1030 ms on localhost (executor driver) (1/1)

19/02/22 23:30:48 INFO DAGScheduler: Removed TaskSet 0.0, where tasks have all completed, from pool

19/02/22 23:30:48 INFO DAGScheduler: ShufflingMapstage 0 (mapToPair at JavaWordCount.java:49) finished in 1.721 s

19/02/22 23:30:48 INFO DAGScheduler: looking for newly runnable stages

19/02/22 23:30:48 INFO DAGScheduler: setting new root stage: Stage 1 (ResultStage1)

19/02/22 23:30:48 INFO DAGScheduler: failed: Set()

19/02/22 23:30:48 INFO DAGScheduler: Submitting ResultStage 1 (ShuffledRDD[6] at reduceByKey at JavaWordCount.java:51), which has no missing parents

19/02/22 23:30:48 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 3.7 KB, free 366.0 MB)

19/02/22 23:30:48 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on 192.168.190.130:39463 (size: 2.1 KB, free: 366.0 MB)

19/02/22 23:30:48 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1008

19/02/22 23:30:48 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (ShuffledRDD[6] at reduceByKey at JavaWordCount.java:51) (first 15 tasks are for partitions Vector(0))

19/02/22 23:30:48 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, executor driver, partition 0, ANY, 4621 bytes)

19/02/22 23:30:48 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)

19/02/22 23:30:48 INFO ShuffleFetcherIterator: Getting non-empty blocks out of 1 blocks

19/02/22 23:30:48 INFO Executor: Received 0 blocks from fetchers in 41 ms

19/02/22 23:30:49 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1), 23933 bytes result sent to driver

19/02/22 23:30:49 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 300 ms on localhost (executor driver) (1/1)

19/02/22 23:30:49 INFO DAGScheduler: 0.0 finished; collect at JavaWordCount.java:53 took 2.608590 s

19/02/22 23:30:49 INFO SparkUI: Stopped Spark web UI at http://192.168.190.130:4040

19/02/22 23:30:49 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!

19/02/22 23:30:49 INFO BlockManager: BlockManager stopped

19/02/22 23:30:49 INFO BlockManagerMaster: BlockManagerMaster stopped

19/02/22 23:30:49 INFO OutputCommitCoordinatorOutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!

19/02/22 23:30:49 INFO ShutdownHookManager: successfully registered hook

19/02/22 23:30:49 INFO ShutdownHookManager: Shutdown hook called

19/02/22 23:30:49 INFO ShutdownHookManager: Deleting directory /tmp/spark-53a0cae-d3a5-4772-83de-1d7674e53798

19/02/22 23:30:49 INFO hadoopuser@ubuntu:~/home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/bin\$]

Fri 23:32:04 INFO BlockManager: BlockManager stopped

19/02/22 23:30:49 INFO BlockManagerMaster: BlockManagerMaster stopped

19/02/22 23:30:49 INFO OutputCommitCoordinator: OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!

19/02/22 23:30:49 INFO SparkContext: Successfully stopped SparkContext

19/02/22 23:30:49 INFO ShutdownHookManager: Shutdown hook called

19/02/22 23:30:49 INFO ShutdownHookManager: Deleting directory /tmp/spark-53a0cae-d3a5-4772-83de-1d7674e53798

19/02/22 23:31:31 INFO hadoopuser@ubuntu:~/home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/bin\$ spark-submit run-example JavaWordCount /home/kite/Desktop/pb/urls.txt>/home/kite/Desktop/urls_output.txt

Using Spark's log4j profile: org/apache/spark/log4j.properties

19/02/22 23:31:31 INFO SparkContext: Running Spark version 2.2.0

19/02/22 23:31:32 WARN NativeCodeLoader: Unable to load native library. Using builtin-java classes where applicable

19/02/22 23:31:32 INFO SecurityManager: Setting spark.hadoop.fs.defaultFS to a local path to a loopback address: /tmp/192.168.190.130:1 using 192.168.190.130 instead (on interface ens33)

19/02/22 23:31:32 WARN Util: Set SPARK_LOCAL_IP if you need to bind to another address

19/02/22 23:31:32 INFO SparkContext: Submitted application: JavaWordCount

19/02/22 23:31:32 INFO SecurityManager: Changing view acls to hadoopuser

19/02/22 23:31:32 INFO SecurityManager: Changing view acls groups to hadoopuser

19/02/22 23:31:32 INFO SecurityManager: Changing view acls groups to:

19/02/22 23:31:32 INFO SecurityManager: SecurityManager: disabled; users with view permissions: Set(hadoopuser); groups with view permissions: Set(); users with modify permissions: Set(); groups with modify permissions: Set()

19/02/22 23:31:33 INFO Util: Successfully started service 'sparkDriver' on port 33385.

19/02/22 23:31:33 INFO SparkEnv: Registering MapoutputTracker

19/02/22 23:31:33 INFO SparkEnv: Registering BlockManagerMaster

19/02/22 23:31:33 INFO BlockManagerMasterEndpoint: blockManagerMasterEndpoint up

19/02/22 23:31:33 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-3591fsd2-6c8b-48db-8709-fc57cc215e84

19/02/22 23:31:33 INFO MemoryStore: MemoryStore started with capacity 366.3 MB

19/02/22 23:31:34 INFO Util: Registered BlockManager

19/02/22 23:31:34 INFO Util: Successfully started service 'SparkUI' on port 4040.

19/02/22 23:31:34 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://192.168.190.130:4040

19/02/22 23:31:34 INFO SparkContext: Added JAR file:/home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/examples/jars/scopt_2.11-3.3.0.jar at spark://192.168.190.130:33385/jars/scopt_2.11-3.3.0.jar with timestamp 155089984595

19/02/22 23:31:34 INFO Executors: Starting executor ID driver on host localhost

19/02/22 23:31:34 INFO Executor: Registered executor ID driver on host localhost with NetworkInterface{name='eth0', ip='192.168.190.130/24', broadcast='192.168.190.130', netmask='255.255.255.0', mac='00:0C:29:4E:4A:95', device='eth0'}

19/02/22 23:31:34 INFO NettyBlockTransferService: Server created on 192.168.190.130:44195

19/02/22 23:31:34 INFO NettyBlockTransferService: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy

19/02/22 23:31:34 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 192.168.190.130, 44195, None)

19/02/22 23:31:34 INFO BlockManagerMasterEndpoint: Registered BlockManager BlockManagerId(driver, 192.168.190.130, 44195, None)

19/02/22 23:31:34 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.190.130, 44195, None)

19/02/22 23:31:35 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('file:/home/kite/Desktop/spark-2.2.0-bin-hadoop2.7/bin/spark-warehouse').

19/02/22 23:31:35 INFO FileSourceStrategy: Hadoop path provider registered to coordinator endpoint

19/02/22 23:31:41 INFO FileSourceStrategy: Pruning directories with:

19/02/22 23:31:34 INFO FileSourceStrategy: Post-Scan Filters:

19/02/22 23:31:34 INFO FileSourceStrategy: OutputFormat: struct-value: string

19/02/22 23:31:34 INFO FileSourceStrategy: Skipped Filters:

19/02/22 23:31:42 INFO CodeGenerator: Code generated in 446.115771 ns

19/02/22 23:31:42 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 277.3 KB, free 366.0 MB)

19/02/22 23:31:42 INFO MemoryStore: Block broadcast_0_piece0 stored as values in memory (estimated size 23.4 KB, free 366.0 MB)

19/02/22 23:31:42 INFO BlockManager: Added broadcast_0_piece0 in memory on 192.168.190.130:44195 (size: 23.4 KB, free: 366.3 MB)

19/02/22 23:31:42 INFO SparkContext: Created broadcast 0 from JavaRDD at JavaWordCount.java:45

19/02/22 23:31:42 INFO FileSourceScanVec: Planning scan with bin packing, max size: 4194304 bytes, open cost is considered as scanning 4194304 bytes.

19/02/22 23:31:43 INFO SparkContext: Starting job: collect at JavaWordCount.java:53

19/02/22 23:31:43 INFO FileSourceScanVec: `FileSourceScanVec: max size: 4194304 bytes, open cost is considered as scanning 4194304 bytes.`

```
Activities Terminal Fri 23/02/2018 22:15:43 INFO DAGScheduler: Submitting ShuffledPartStage 0 (MapPartitionsRDD[5] at mapToPair at JavaWordCount.java:49), which has no missing parents
Fri 23/02/2018 22:15:43 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 12.5 KB, free 366.0 MB)
Fri 23/02/23 23:13:43 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 6.5 KB, free 366.0 MB)
Fri 23/02/23 23:13:43 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on 192.168.190.130:44195 (size: 6.5 KB, free: 366.3 MB)
Fri 23/02/23 23:13:43 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1006
Fri 23/02/23 23:13:43 INFO TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
Fri 23/02/23 23:13:43 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, executor driver, partition 0, PROCESS_LOCAL, 5263 bytes)
Fri 23/02/23 23:13:43 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
Fri 23/02/23 23:13:43 INFO TaskSetManager: Stage 0: 0 tasks left
Fri 23/02/23 23:13:43 INFO TaskSchedulerImpl: 0 tasks left in stage 0 (0 total)
Fri 23/02/23 23:13:43 INFO TransportLineFactory: Successfully created connection to /192.168.190.130:33385 after 53 ms (0 ms spent in bootstraps)
Fri 23/02/23 23:13:43 INFO Utils: Fetching spark://192.168.190.130:33385/jars/scott_2.11-2.2.0.jar to /tmp/spark-0a9d3c8b-508a-4193-a53a-dba347c8ef2e/userFiles-c0864d39-40c3-4eds-b3a1-704f4f56a29e/fetchfile
npo462965981422849908w.tgz
Fri 23/02/23 23:13:43 INFO Executor: Adding file:/tmp/spark-0a9d3c8b-508a-4193-a53a-dba347c8ef2e/userFiles-c0864d39-40c3-4eds-b3a1-704f4f56a29e/spark-examples_2.11-2.2.0.jar to class loader
Fri 23/02/23 23:13:43 INFO Executor: Fetching spark://192.168.190.130:33385/jars/spark-examples_2.11-2.2.0.jar with timestamp 155098984595
Fri 23/02/23 23:13:43 INFO Utils: Fetching spark://192.168.190.130:33385/jars/spark-examples_2.11-2.2.0.jar to /tmp/spark-0a9d3c8b-508a-4193-a53a-dba347c8ef2e/userFiles-c0864d39-40c3-4eds-b3a1-704f4f56a29e/fetchfile
tchFileTemp44429791867557762.s
Fri 23/02/23 23:13:43 INFO DAGScheduler: Adding file:/tmp/spark-0a9d3c8b-508a-4193-a53a-dba347c8ef2e/userFiles-c0864d39-40c3-4eds-b3a1-704f4f56a29e/spark-examples_2.11-2.2.0.jar to class loader
Fri 23/02/23 23:13:44 INFO CodeGenerator: Code generated in 40.464472 ms
Fri 23/02/23 23:13:44 INFO FileScanDesc: Reading File path: file:///Home/kite/Desktop/spark/ub.txt, range: 0-38162, partition values: [empty row]
Fri 23/02/23 23:13:44 INFO CodeGenerator: Code generated in 33.667355 ms
Fri 23/02/23 23:13:44 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
Fri 23/02/23 23:13:45 INFO TaskSchedulerImpl: Finished task 0.0 in stage 0.0 (TID 0) in 1768 ms on localhost (executor driver) (1/1)
Fri 23/02/23 23:13:45 INFO TaskSchedulerImpl: Removed Taskset 0.0, whose tasks have all completed, from pool
Fri 23/02/23 23:13:45 INFO DAGScheduler: ShuffleMapStage 0 (mapToPair at JavaWordCount.java:49) finished in 1.578 s
Fri 23/02/23 23:13:45 INFO DAGScheduler: Looking for newly runnable stages
Fri 23/02/23 23:13:45 INFO DAGScheduler: Found 1 stage(s) to run
Fri 23/02/23 23:13:45 INFO DAGScheduler: waiting: Set{ShuffledPartStage 0}
Fri 23/02/23 23:13:45 INFO DAGScheduler: failed: Set{ }
Fri 23/02/23 23:13:45 INFO DAGScheduler: Submitting ShuffledPartStage 1 (shuffledRDD[0] at reduceByKey at JavaWordCount.java:51), which has no missing parents
Fri 23/02/23 23:13:45 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 3.7 KB, free 366.0 MB)
Fri 23/02/23 23:13:45 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 2.1 KB, free 366.0 MB)
Fri 23/02/23 23:13:45 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on 192.168.190.130:44195 (size: 2.1 KB, free: 366.3 MB)
Fri 23/02/23 23:13:45 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1006
Fri 23/02/23 23:13:45 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
Fri 23/02/23 23:13:45 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 0, localhost, executor driver, partition 0, ANY, 4621 bytes)
Fri 23/02/23 23:13:45 INFO Executor: Running task 0.0 in stage 1.0 (TID 0)
Fri 23/02/23 23:13:45 INFO ShuffleBlockFetcherIterator: Getting 1000 blocks out of 1 blocks
Fri 23/02/23 23:13:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 5 ms
Fri 23/02/23 23:13:45 INFO Executor: Finished task 0.0 in stage 1.0 (TID 0) - 27071 bytes sent to driver
Fri 23/02/23 23:13:45 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 0) in 19 ms on localhost (executor driver) (1/1)
Fri 23/02/23 23:13:45 INFO TaskSchedulerImpl: Removed Taskset 1.0, whose tasks have all completed, from pool
Fri 23/02/23 23:13:45 INFO DAGScheduler: Job 0 finished: collect at JavaWordCount.java:53, took 2.26472 s
Fri 23/02/23 23:13:45 INFO SparkUI: Stopped Spark Web UI at http://192.168.190.130:4040
Fri 23/02/23 23:13:45 INFO MapOutputTrackerMasterEndpoint: mapOutputTrackerEndpoint stopped!
Fri 23/02/23 23:13:45 INFO BlockManager: BlockManager stopped
Fri 23/02/23 23:13:46 INFO BlockManagerMaster: BlockManagerMaster stopped
Fri 23/02/23 23:13:46 INFO OutputCommitCoordinator: OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
Fri 23/02/23 23:13:46 INFO ShutdownHookManager: Successfully stopped SparkContext
Fri 23/02/23 23:13:46 INFO ShutdownHookManager: Shutdown hook called
Fri 23/02/23 23:13:46 INFO ShutdownHookManager: Deleting directory /tmp/spark-0a9d3c8b-508a-4193-a53a-dba347c8ef2e
hadoopuser@ubuntu:~/kite/Desktopspark-2.2.0-bin-hadoop2.7$ bin/
```

In above spark code, first rdd is created by using textFile function of the sparkcontext from the hashtags and urls textfile. By using flatMap function, we generated words from each line by taking space as delimiter and appended one as value with the word as key. Using reduceByKey function we are adding values by the key and the output generated is the words and their sum is saved to a textfile.

```
spark_hashtag_output.txt
1 SebastianInthru: 3
2 weekend: 5
3 Blithub: 1
4 engenharia: 1
5 engenharia: 1
6 Gold: 4
7 MyStarShipPewPew: 1
8 socialdemokraterna: 1
9 mobilely: 1
10 electriques: 1
11 SmartTechnology: 1
12 autos: 2
13 stemzeug: 1
14 Mobilis: 1
15 sustainableenergy: 2
16 busconnects: 2
17 engineering: 1
18 google: 1
19 electriques: 1
20 technologiax: 1
21 ThePaper: 1
22 paymentsinnovation: 1
23 mobilely: 1
24 tecno: 5
25 models: 5
26 typography: 1
27 Fisker: 1
28 globelblz: 1
29 Autofahrer: 1
30 typography: 11
31 SahelNews: 1
32 book: 2
33 classictmetalradio: 1
34 europe: 1
35 NBC: 7
36 luke: 1
37 SolarEdge: 1
38 Kaufempfehlung: 1
39 automobile: 2
40 technologiax: 3
41 car: 1
42 Jaguar: 7
43 amsterdam: 1
44 shf: 1
45 bodyguard: 3
46 selfdriving: 3
47 ausland: 2
48 Leonesp: 2
49 metal: 1
50 topspeed: 2
```

The screenshot shows the Visual Studio Code interface with the following details:

- File Explorer (Left):** Shows the project structure with several files and folders:
 - OPEN EDITORS: spark.urls_output.txt, Spark_Logs
 - PHASE_1
 - Code
 - Hadoop_Logs
 - Spark_Logs
 - spark.urls_output.txt
 - spark.urls_output.txt
 - spark.urls_output.txt
 - python.urls_final.json
 - spark.urls_output.txt
 - Twitter_Data
- Editor (Right):** Displays the content of the file "spark.urls_output.txt". The file contains a list of approximately 50 URLs, each followed by a colon and the number "1". The URLs include various Twitter links, such as <https://twitter.com/i/web/status/10988483714094529665>, <https://twitter.com/i/web/status/109892495817936768>, and <https://www.flairmagazine.com/cryptocurrency/news/elon-musk-praises-bitcoin-refutes-rumors-of-tesla-crypto-plans/>.