Principles Of Bigdata Management

Phase-2 Report

By

Team-23

Hemanth Kumar Reddy Dantu(16233525)

Mahesh Chowdary Jamallamudi(16234558)

Theyab Alharbi(16194252)

Arun Kumar Anthati(16232818)

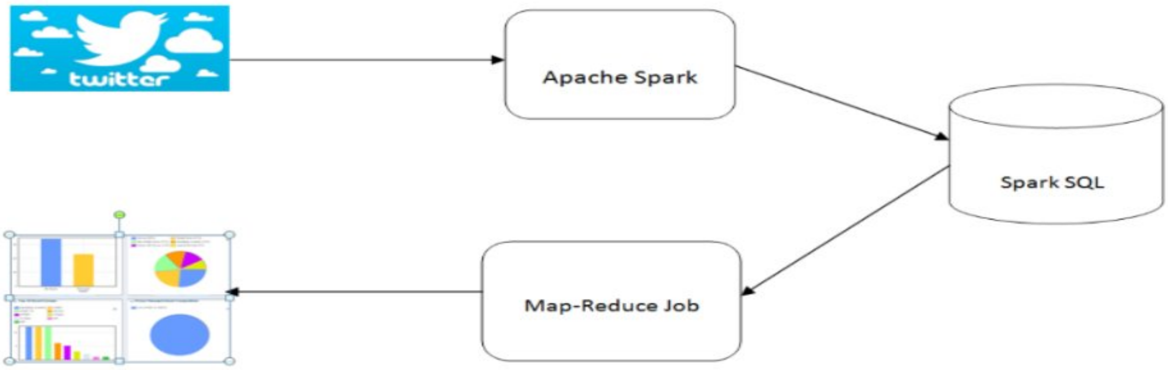
Project Description:

For the phase-2 we have collected to more than 1 GB tweets on Banking systems. We have used hashtags for banks and collected the tweets on Banking systems.

After the tweets collections, we have analyzed the banking tweets and visualized using "d3.js" and "high charts". Total we have implemented 6 queries, 3 queries using "Java RDD" and 3 queries using "DataFrames".

We have implemented the dynamic web project it first process the data based on the query and the result stored on .csv file and immediately it calls html page to visualize the collected results.

Architecture Diagram:



Software Technologies and Tools:

User Interface: HTML,CSS. D3.js and High Charts.

Tools: Eclipse

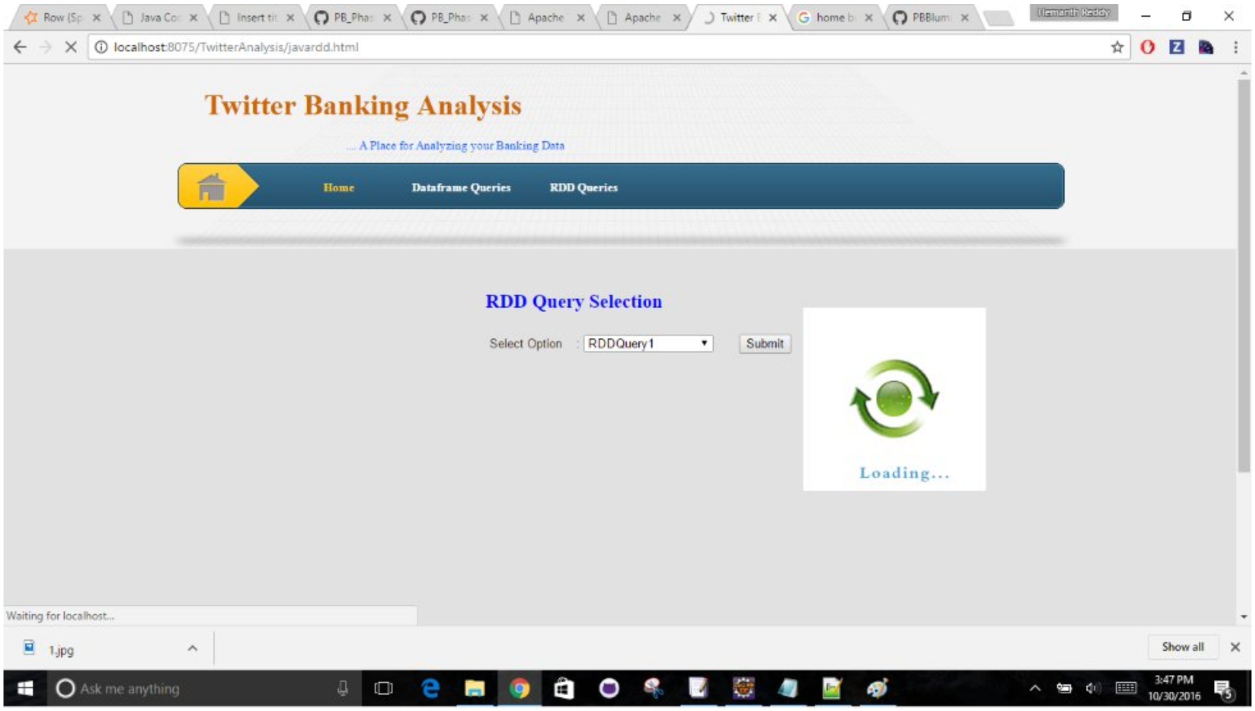
Environment: Apache Spark

No SQL Database: Spark SQL.

Backend: Java Spark and Servlets.

RDD Queries:

Home Page Visualization:



localhost:8075/TwitterAnalysis/javardd.html

Twitter Banking Analysis

... A Place for Analyzing your Banking Data

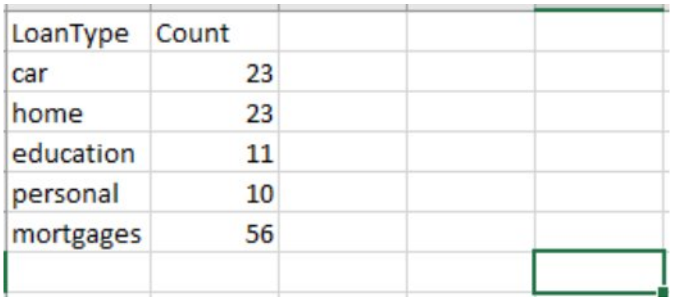
RDD Query Selection

Select Option : RDDQuery1

Waiting for localhost...

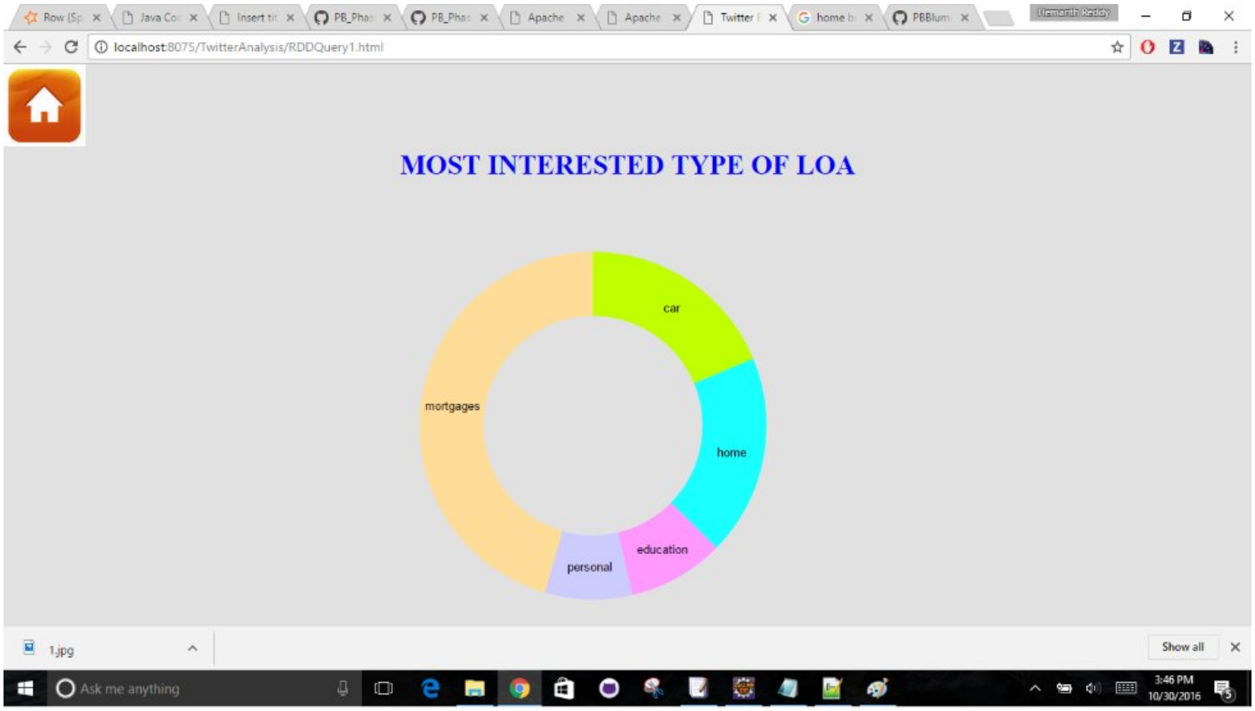
RDD Query-1: The most interested loan types that the users discussing on twitter.

.CSV file Results:



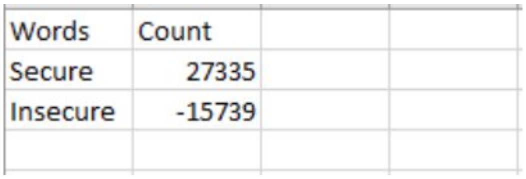
LoanType	Count
car	23
home	23
education	11
personal	10
mortgages	56

Visualization Results:



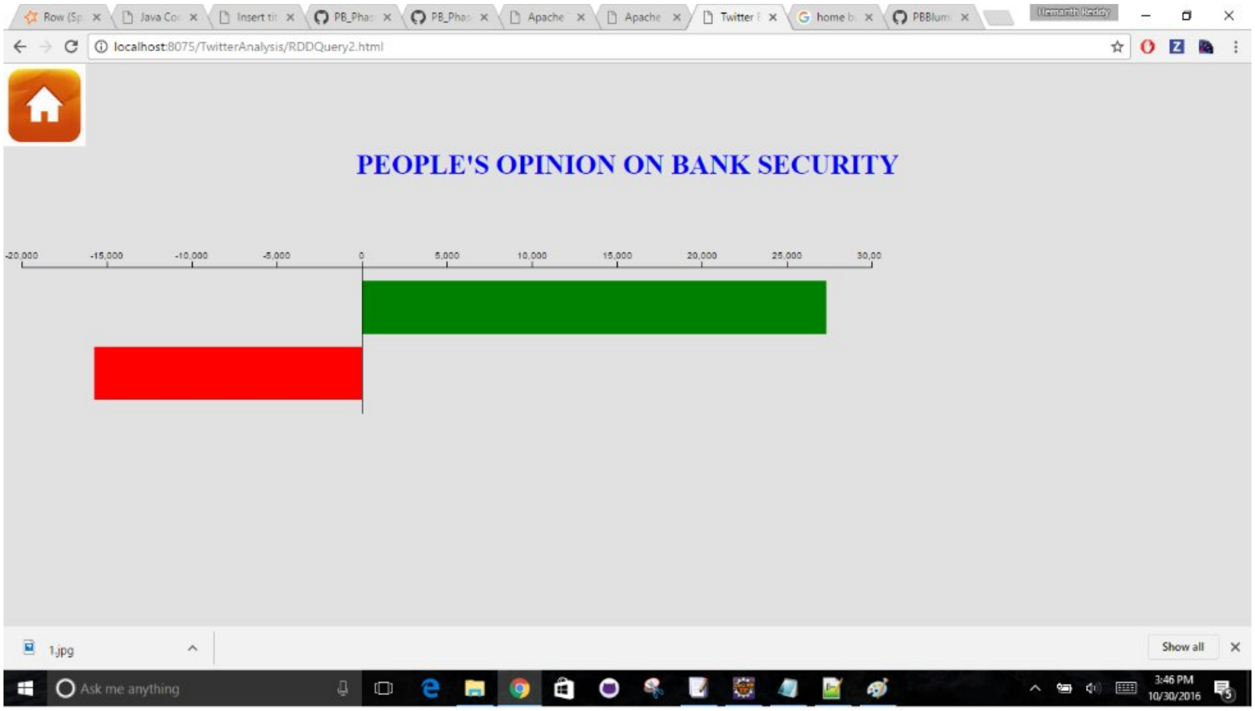
RDD Query-2: This query analyzes people's opinion on bank security. We have analyzed the positive(Secure) and negative (insecure) tweets.

.CSV file Results:



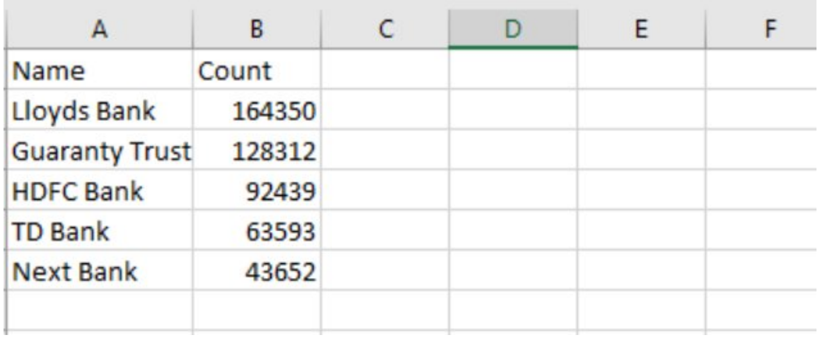
Words	Count
Secure	27335
Insecure	-15739

Visualization Results:



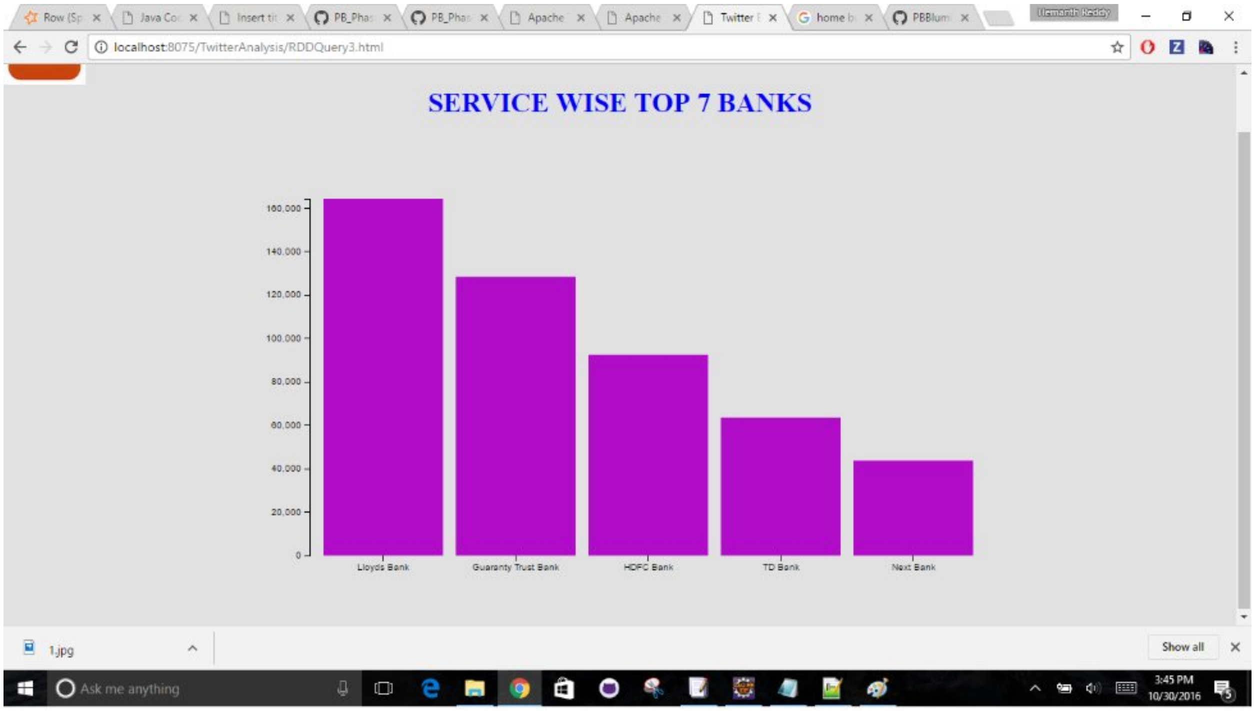
RDD Query-3: This query analyses the top banks on service wise.

.CSV file Results:



A	B	C	D	E	F
Name	Count				
Lloyds Bank	164350				
Guaranty Trust	128312				
HDFC Bank	92439				
TD Bank	63593				
Next Bank	43652				

Visualization Results:

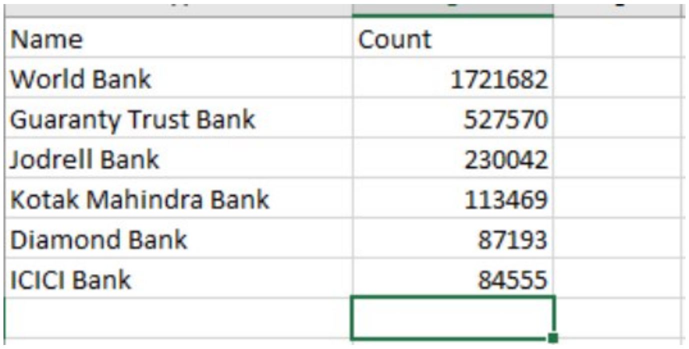


localhost:8075/TwitterAnalysis/RDDQuery3.html

DataFrame Queries:

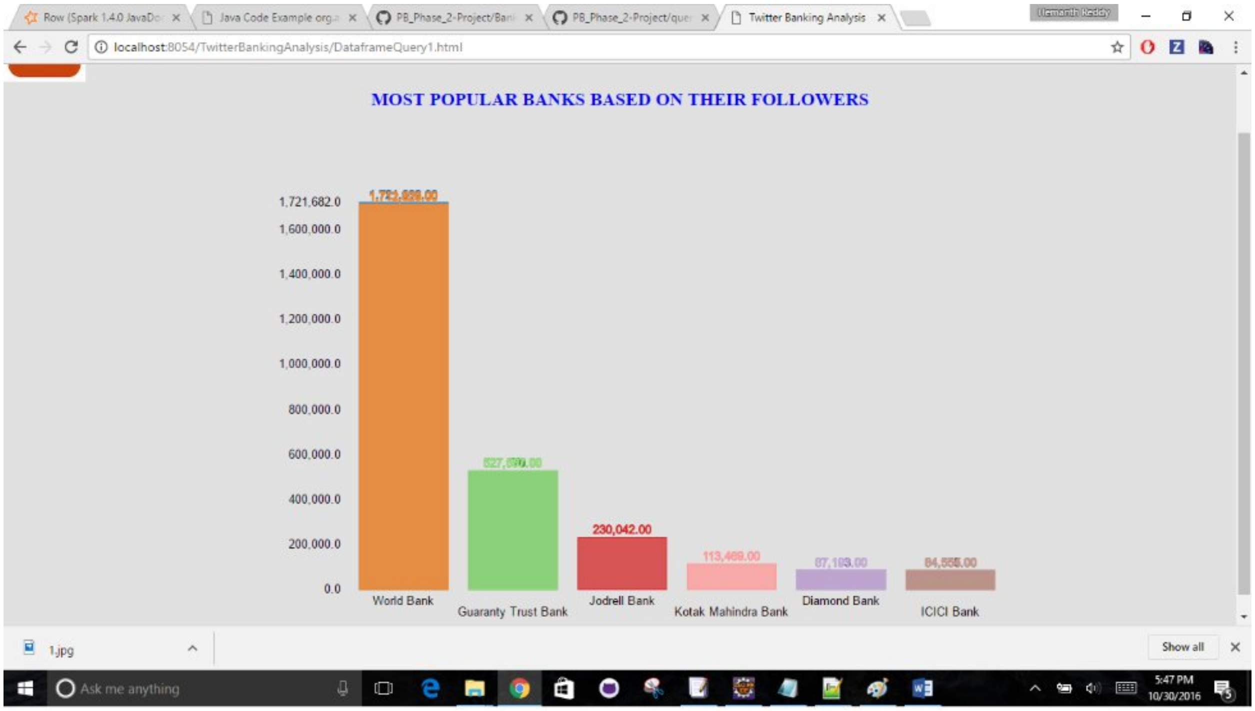
DataFrame Query-1: The most popular banks on twitter. We have analyzed the most followers of the each banks and visualized the top 6 banks.

.CSV file Results:



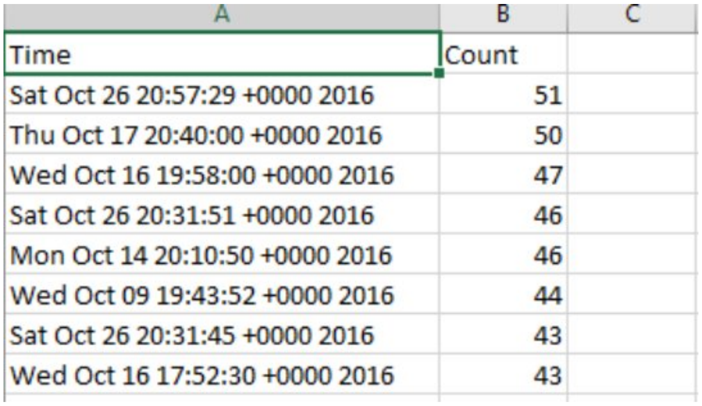
Name	Count
World Bank	1721682
Guaranty Trust Bank	527570
Jodrell Bank	230042
Kotak Mahindra Bank	113469
Diamond Bank	87193
ICICI Bank	84555

Visualization Results:



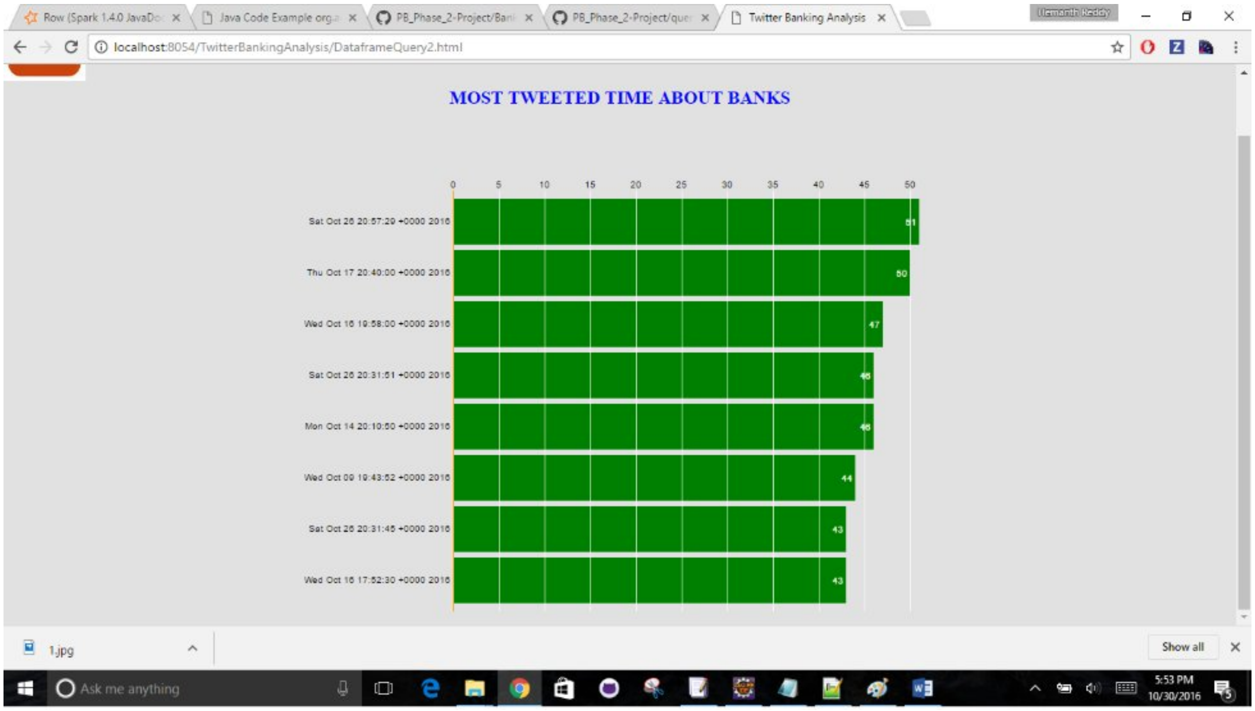
DataFrame Query-2: The most tweeted timings on banks. We have analyzed the top 8 twitting times on banks.

.CSV file Results:



Time	Count
Sat Oct 26 20:57:29 +0000 2016	51
Thu Oct 17 20:40:00 +0000 2016	50
Wed Oct 16 19:58:00 +0000 2016	47
Sat Oct 26 20:31:51 +0000 2016	46
Mon Oct 14 20:10:50 +0000 2016	46
Wed Oct 09 19:43:52 +0000 2016	44
Sat Oct 26 20:31:45 +0000 2016	43
Wed Oct 16 17:52:30 +0000 2016	43

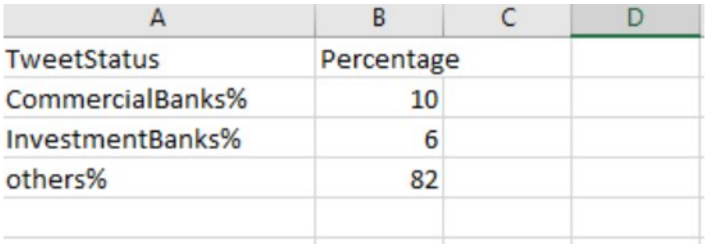
Visualization Results:


Row (Spark 1.4.0 JavaDo
Java Code Example org.a
PB_Phase_2-Project/Ban
PB_Phase_2-Project/que
Twitter Banking Analysis
Hemanth Reddy
localhost:8054/TwitterBankingAnalysis/DataframeQuery2.html
0
5
10
15
20
25
30
35
40
45
50
Sat Oct 26 20:57:29 +0000 2016
51
Thu Oct 17 20:40:00 +0000 2016
50
Wed Oct 16 19:58:00 +0000 2016
47
Sat Oct 26 20:31:51 +0000 2016
46
Mon Oct 14 20:10:50 +0000 2016
46
Wed Oct 09 19:43:52 +0000 2016
44
Sat Oct 26 20:31:45 +0000 2016
43
Wed Oct 16 17:52:30 +0000 2016
43
1.jpg
Show all
Ask me anything
5:53 PM
10/30/2016

MOST TWEETED TIME ABOUT BANKS

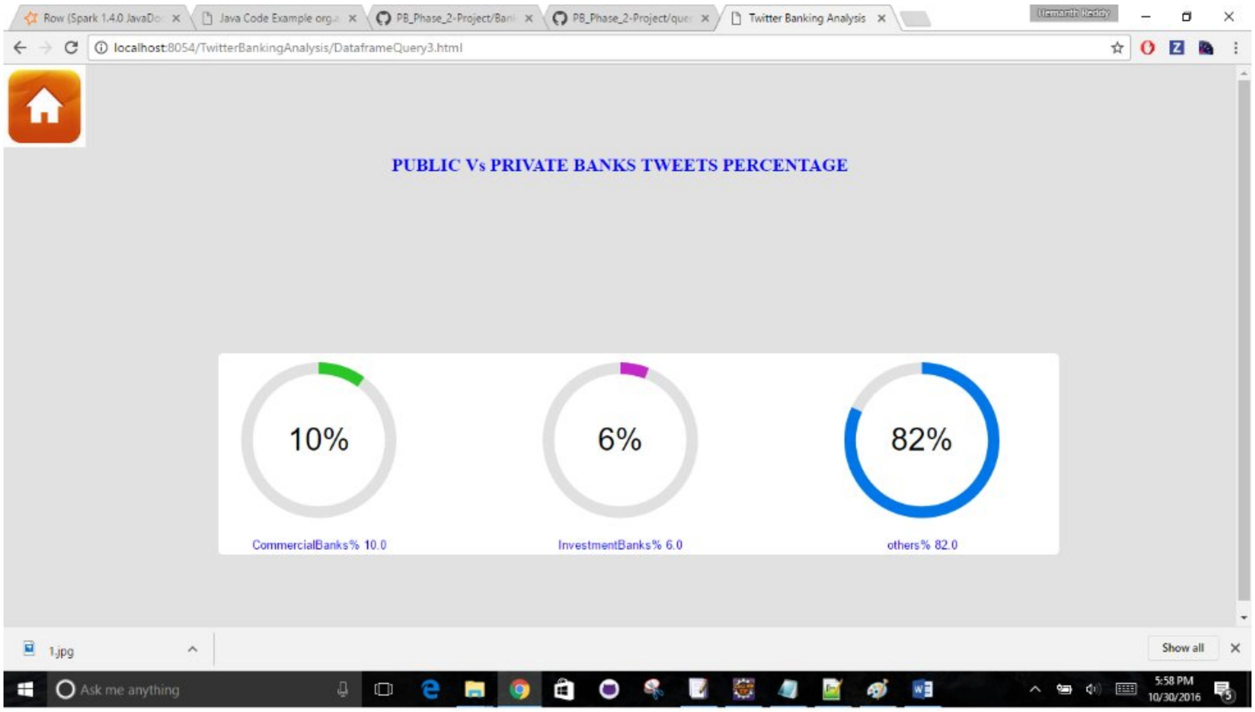
DataFrame Query-3: This query analyzes the Public and Private and other banks tweets percentage.

.CSV file Results:

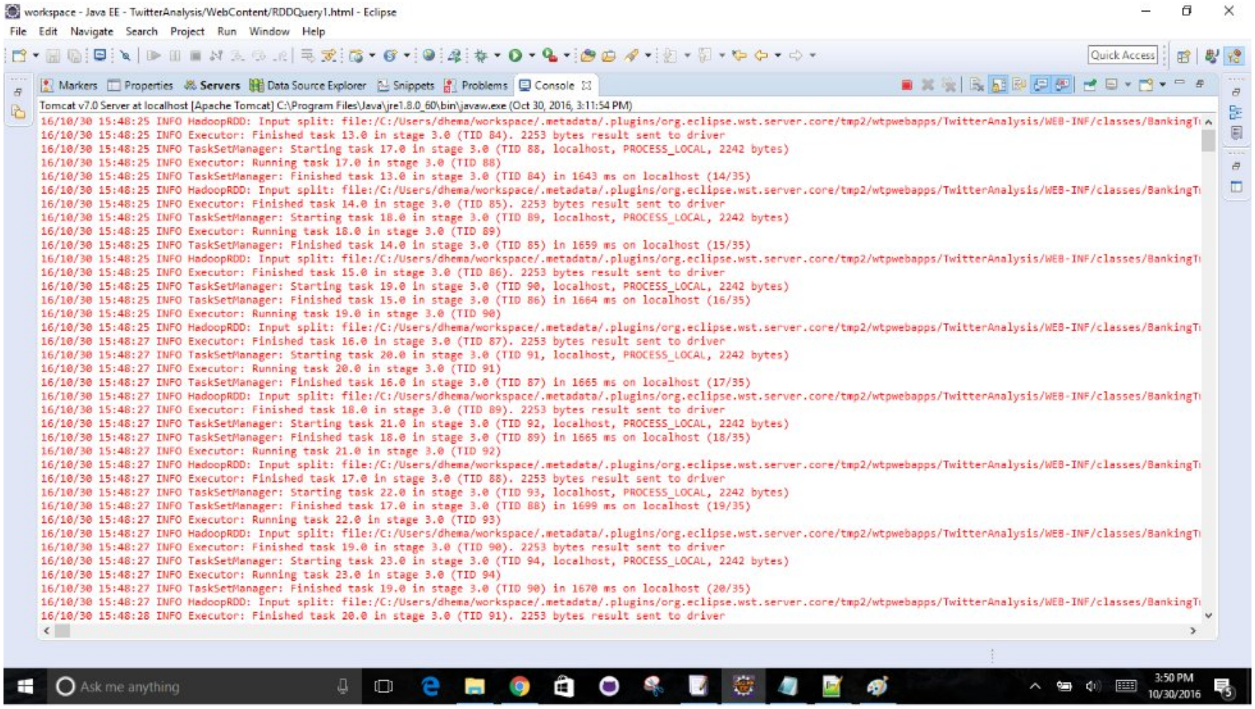


A	B	C	D
TweetStatus	Percentage		
CommercialBanks%	10		
InvestmentBanks%	6		
others%	82		

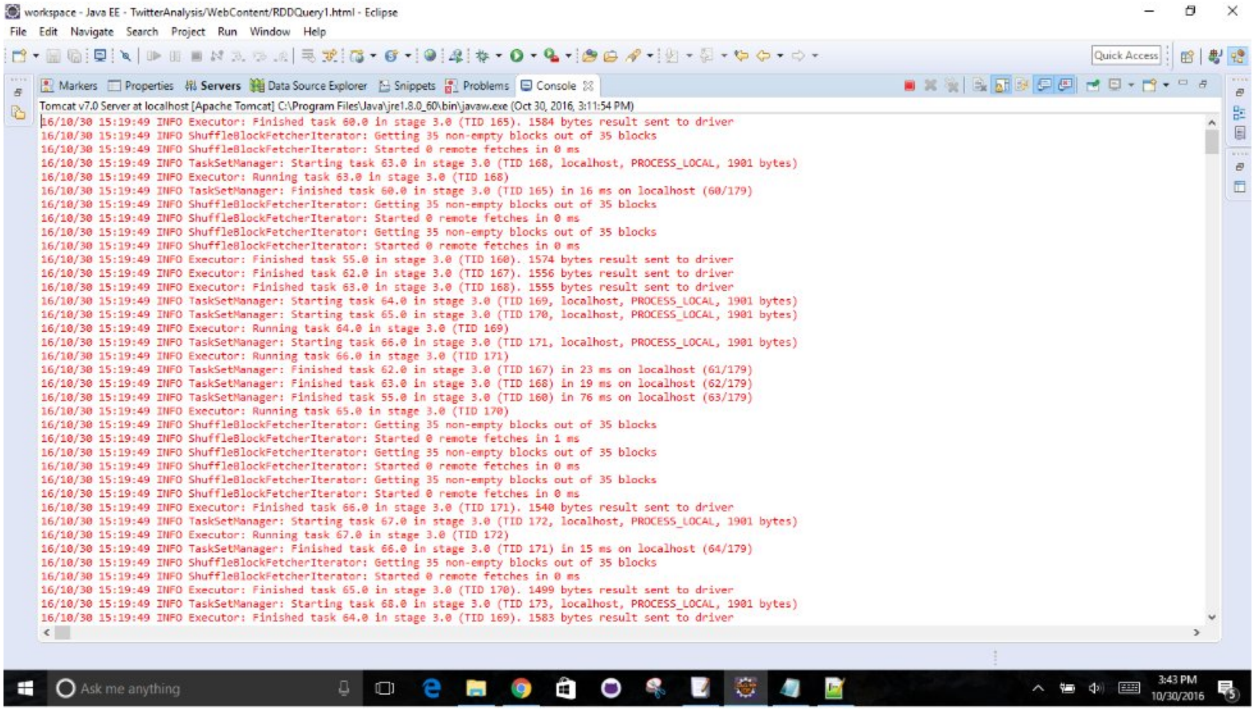
Visualization Results:



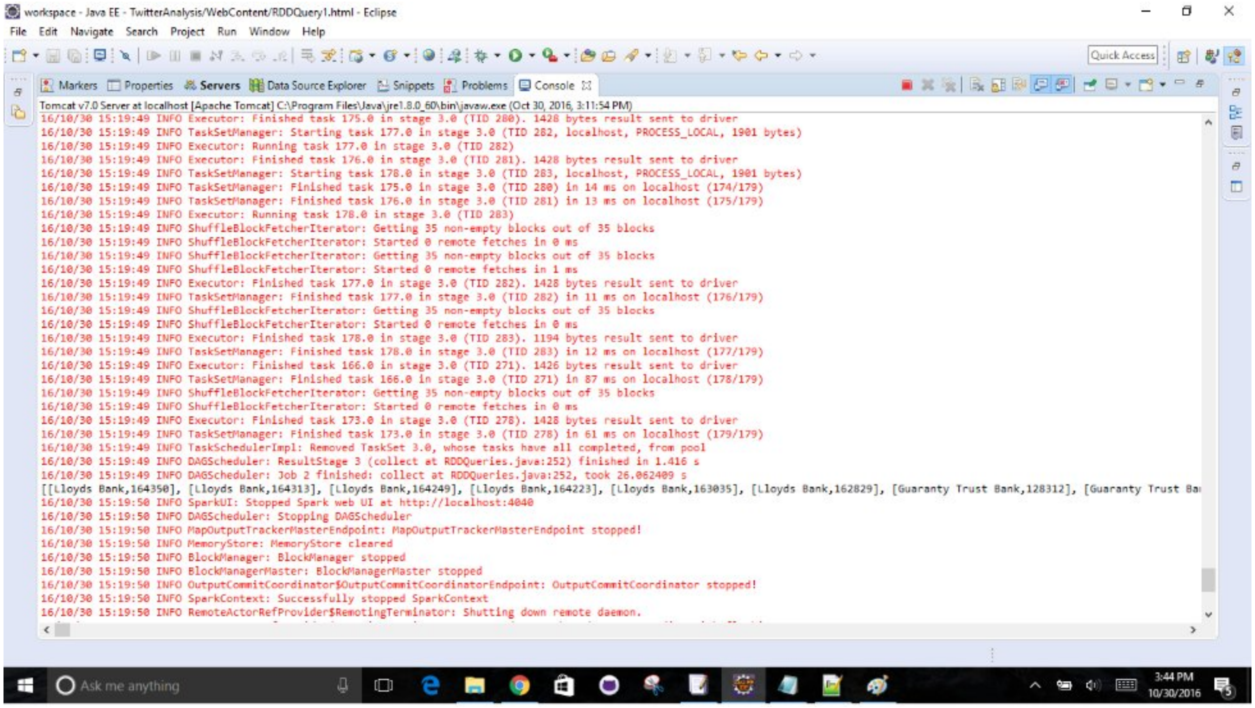
Log files:


workspace - Java EE - TwitterAnalysis/WebContent/RDDQuery1.html - Eclipse
File Edit Navigate Search Project Run Window Help
Quick Access
Markers Properties Servers Data Source Explorer Snippets Problems Console
Ask me anything
3:50 PM
10/30/2016

```
Tomcat v7.0 Server at localhost [Apache Tomcat] C:\Program Files\Java\jre1.8.0_60\bin\javaw.exe (Oct 30, 2016, 3:11:54 PM)
16/10/30 15:48:25 INFO HadoopRDD: Input split: file:/C:/Users/dhema/workspace/.metadata/.plugins/org.eclipse.wst.server.core/tmp2/wtpwebapps/TwitterAnalysis/WEB-INF/classes/BankingT
16/10/30 15:48:25 INFO Executor: Finished task 13.0 in stage 3.0 (TID 84). 2253 bytes result sent to driver
16/10/30 15:48:25 INFO TaskSetManager: Starting task 17.0 in stage 3.0 (TID 88, localhost, PROCESS_LOCAL, 2242 bytes)
16/10/30 15:48:25 INFO Executor: Running task 17.0 in stage 3.0 (TID 88)
16/10/30 15:48:25 INFO TaskSetManager: Finished task 13.0 in stage 3.0 (TID 84) in 1643 ms on localhost (14/35)
16/10/30 15:48:25 INFO HadoopRDD: Input split: file:/C:/Users/dhema/workspace/.metadata/.plugins/org.eclipse.wst.server.core/tmp2/wtpwebapps/TwitterAnalysis/WEB-INF/classes/BankingT
16/10/30 15:48:25 INFO Executor: Finished task 14.0 in stage 3.0 (TID 85). 2253 bytes result sent to driver
16/10/30 15:48:25 INFO TaskSetManager: Starting task 18.0 in stage 3.0 (TID 89, localhost, PROCESS_LOCAL, 2242 bytes)
16/10/30 15:48:25 INFO Executor: Running task 18.0 in stage 3.0 (TID 89)
16/10/30 15:48:25 INFO TaskSetManager: Finished task 14.0 in stage 3.0 (TID 85) in 1659 ms on localhost (15/35)
16/10/30 15:48:25 INFO HadoopRDD: Input split: file:/C:/Users/dhema/workspace/.metadata/.plugins/org.eclipse.wst.server.core/tmp2/wtpwebapps/TwitterAnalysis/WEB-INF/classes/BankingT
16/10/30 15:48:25 INFO Executor: Finished task 15.0 in stage 3.0 (TID 86). 2253 bytes result sent to driver
16/10/30 15:48:25 INFO TaskSetManager: Starting task 19.0 in stage 3.0 (TID 90, localhost, PROCESS_LOCAL, 2242 bytes)
16/10/30 15:48:25 INFO TaskSetManager: Finished task 15.0 in stage 3.0 (TID 86) in 1664 ms on localhost (16/35)
16/10/30 15:48:25 INFO Executor: Running task 19.0 in stage 3.0 (TID 90)
16/10/30 15:48:25 INFO HadoopRDD: Input split: file:/C:/Users/dhema/workspace/.metadata/.plugins/org.eclipse.wst.server.core/tmp2/wtpwebapps/TwitterAnalysis/WEB-INF/classes/BankingT
16/10/30 15:48:27 INFO Executor: Finished task 16.0 in stage 3.0 (TID 87). 2253 bytes result sent to driver
16/10/30 15:48:27 INFO TaskSetManager: Starting task 20.0 in stage 3.0 (TID 91, localhost, PROCESS_LOCAL, 2242 bytes)
16/10/30 15:48:27 INFO Executor: Running task 20.0 in stage 3.0 (TID 91)
16/10/30 15:48:27 INFO TaskSetManager: Finished task 16.0 in stage 3.0 (TID 87) in 1665 ms on localhost (17/35)
16/10/30 15:48:27 INFO HadoopRDD: Input split: file:/C:/Users/dhema/workspace/.metadata/.plugins/org.eclipse.wst.server.core/tmp2/wtpwebapps/TwitterAnalysis/WEB-INF/classes/BankingT
16/10/30 15:48:27 INFO Executor: Finished task 18.0 in stage 3.0 (TID 89). 2253 bytes result sent to driver
16/10/30 15:48:27 INFO TaskSetManager: Starting task 21.0 in stage 3.0 (TID 92, localhost, PROCESS_LOCAL, 2242 bytes)
16/10/30 15:48:27 INFO TaskSetManager: Finished task 18.0 in stage 3.0 (TID 89) in 1665 ms on localhost (18/35)
16/10/30 15:48:27 INFO Executor: Running task 21.0 in stage 3.0 (TID 92)
16/10/30 15:48:27 INFO HadoopRDD: Input split: file:/C:/Users/dhema/workspace/.metadata/.plugins/org.eclipse.wst.server.core/tmp2/wtpwebapps/TwitterAnalysis/WEB-INF/classes/BankingT
16/10/30 15:48:27 INFO Executor: Finished task 17.0 in stage 3.0 (TID 88). 2253 bytes result sent to driver
16/10/30 15:48:27 INFO TaskSetManager: Starting task 22.0 in stage 3.0 (TID 93, localhost, PROCESS_LOCAL, 2242 bytes)
16/10/30 15:48:27 INFO TaskSetManager: Finished task 17.0 in stage 3.0 (TID 88) in 1699 ms on localhost (19/35)
16/10/30 15:48:27 INFO Executor: Running task 22.0 in stage 3.0 (TID 93)
16/10/30 15:48:27 INFO HadoopRDD: Input split: file:/C:/Users/dhema/workspace/.metadata/.plugins/org.eclipse.wst.server.core/tmp2/wtpwebapps/TwitterAnalysis/WEB-INF/classes/BankingT
16/10/30 15:48:27 INFO Executor: Finished task 19.0 in stage 3.0 (TID 90). 2253 bytes result sent to driver
16/10/30 15:48:27 INFO TaskSetManager: Starting task 23.0 in stage 3.0 (TID 94, localhost, PROCESS_LOCAL, 2242 bytes)
16/10/30 15:48:27 INFO Executor: Running task 23.0 in stage 3.0 (TID 94)
16/10/30 15:48:27 INFO TaskSetManager: Finished task 19.0 in stage 3.0 (TID 90) in 1670 ms on localhost (20/35)
16/10/30 15:48:27 INFO HadoopRDD: Input split: file:/C:/Users/dhema/workspace/.metadata/.plugins/org.eclipse.wst.server.core/tmp2/wtpwebapps/TwitterAnalysis/WEB-INF/classes/BankingT
16/10/30 15:48:28 INFO Executor: Finished task 20.0 in stage 3.0 (TID 91). 2253 bytes result sent to driver
```


workspace - Java EE - TwitterAnalysis/WebContent/RDDQuery1.html - Eclipse
File Edit Navigate Search Project Run Window Help
Quick Access
Markers Properties Servers Data Source Explorer Snippets Problems Console
Ask me anything
3:43 PM
10/30/2016

```
Tomcat v7.0 Server at localhost [Apache Tomcat] C:\Program Files\Java\jre1.8.0_60\bin\javaw.exe (Oct 30, 2016, 3:11:54 PM)
16/10/30 15:19:49 INFO Executor: Finished task 60.0 in stage 3.0 (TID 165). 1584 bytes result sent to driver
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Getting 35 non-empty blocks out of 35 blocks
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/10/30 15:19:49 INFO TaskSetManager: Starting task 63.0 in stage 3.0 (TID 168, localhost, PROCESS_LOCAL, 1901 bytes)
16/10/30 15:19:49 INFO Executor: Running task 63.0 in stage 3.0 (TID 168)
16/10/30 15:19:49 INFO TaskSetManager: Finished task 60.0 in stage 3.0 (TID 165) in 16 ms on localhost (60/179)
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Getting 35 non-empty blocks out of 35 blocks
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Getting 35 non-empty blocks out of 35 blocks
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/10/30 15:19:49 INFO Executor: Finished task 55.0 in stage 3.0 (TID 160). 1574 bytes result sent to driver
16/10/30 15:19:49 INFO Executor: Finished task 62.0 in stage 3.0 (TID 167). 1556 bytes result sent to driver
16/10/30 15:19:49 INFO Executor: Finished task 63.0 in stage 3.0 (TID 168). 1555 bytes result sent to driver
16/10/30 15:19:49 INFO TaskSetManager: Starting task 64.0 in stage 3.0 (TID 169, localhost, PROCESS_LOCAL, 1901 bytes)
16/10/30 15:19:49 INFO TaskSetManager: Starting task 65.0 in stage 3.0 (TID 170, localhost, PROCESS_LOCAL, 1901 bytes)
16/10/30 15:19:49 INFO Executor: Running task 64.0 in stage 3.0 (TID 169)
16/10/30 15:19:49 INFO TaskSetManager: Starting task 66.0 in stage 3.0 (TID 171, localhost, PROCESS_LOCAL, 1901 bytes)
16/10/30 15:19:49 INFO Executor: Running task 66.0 in stage 3.0 (TID 171)
16/10/30 15:19:49 INFO TaskSetManager: Finished task 62.0 in stage 3.0 (TID 167) in 23 ms on localhost (61/179)
16/10/30 15:19:49 INFO TaskSetManager: Finished task 63.0 in stage 3.0 (TID 168) in 19 ms on localhost (62/179)
16/10/30 15:19:49 INFO TaskSetManager: Finished task 55.0 in stage 3.0 (TID 160) in 76 ms on localhost (63/179)
16/10/30 15:19:49 INFO Executor: Running task 65.0 in stage 3.0 (TID 170)
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Getting 35 non-empty blocks out of 35 blocks
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Getting 35 non-empty blocks out of 35 blocks
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Getting 35 non-empty blocks out of 35 blocks
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/10/30 15:19:49 INFO Executor: Finished task 66.0 in stage 3.0 (TID 171). 1540 bytes result sent to driver
16/10/30 15:19:49 INFO TaskSetManager: Starting task 67.0 in stage 3.0 (TID 172, localhost, PROCESS_LOCAL, 1901 bytes)
16/10/30 15:19:49 INFO Executor: Running task 67.0 in stage 3.0 (TID 172)
16/10/30 15:19:49 INFO TaskSetManager: Finished task 66.0 in stage 3.0 (TID 171) in 15 ms on localhost (64/179)
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Getting 35 non-empty blocks out of 35 blocks
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/10/30 15:19:49 INFO Executor: Finished task 65.0 in stage 3.0 (TID 170). 1499 bytes result sent to driver
16/10/30 15:19:49 INFO TaskSetManager: Starting task 68.0 in stage 3.0 (TID 173, localhost, PROCESS_LOCAL, 1901 bytes)
16/10/30 15:19:49 INFO Executor: Finished task 64.0 in stage 3.0 (TID 169). 1583 bytes result sent to driver
```



workspace - Java EE - TwitterAnalysis/WebContent/RDDQuery1.html - Eclipse
File Edit Navigate Search Project Run Window Help
Quick Access
Markers Properties Servers Data Source Explorer Snippets Problems Console
Ask me anything
3:44 PM
10/30/2016

```
Tomcat v7.0 Server at localhost [Apache Tomcat] C:\Program Files\Java\jre1.8.0_60\bin\javaw.exe (Oct 30, 2016, 3:11:54 PM)
16/10/30 15:19:49 INFO Executor: Finished task 175.0 in stage 3.0 (TID 280). 1428 bytes result sent to driver
16/10/30 15:19:49 INFO TaskSetManager: Starting task 177.0 in stage 3.0 (TID 282, localhost, PROCESS_LOCAL, 1901 bytes)
16/10/30 15:19:49 INFO Executor: Running task 177.0 in stage 3.0 (TID 282)
16/10/30 15:19:49 INFO Executor: Finished task 176.0 in stage 3.0 (TID 281). 1428 bytes result sent to driver
16/10/30 15:19:49 INFO TaskSetManager: Starting task 178.0 in stage 3.0 (TID 283, localhost, PROCESS_LOCAL, 1901 bytes)
16/10/30 15:19:49 INFO TaskSetManager: Finished task 175.0 in stage 3.0 (TID 280) in 14 ms on localhost (174/179)
16/10/30 15:19:49 INFO TaskSetManager: Finished task 176.0 in stage 3.0 (TID 281) in 13 ms on localhost (175/179)
16/10/30 15:19:49 INFO Executor: Running task 178.0 in stage 3.0 (TID 283)
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Getting 35 non-empty blocks out of 35 blocks
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Getting 35 non-empty blocks out of 35 blocks
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
16/10/30 15:19:49 INFO Executor: Finished task 177.0 in stage 3.0 (TID 282). 1428 bytes result sent to driver
16/10/30 15:19:49 INFO TaskSetManager: Finished task 177.0 in stage 3.0 (TID 282) in 11 ms on localhost (176/179)
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Getting 35 non-empty blocks out of 35 blocks
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/10/30 15:19:49 INFO Executor: Finished task 178.0 in stage 3.0 (TID 283). 1194 bytes result sent to driver
16/10/30 15:19:49 INFO TaskSetManager: Finished task 178.0 in stage 3.0 (TID 283) in 12 ms on localhost (177/179)
16/10/30 15:19:49 INFO Executor: Finished task 166.0 in stage 3.0 (TID 271). 1426 bytes result sent to driver
16/10/30 15:19:49 INFO TaskSetManager: Finished task 166.0 in stage 3.0 (TID 271) in 87 ms on localhost (178/179)
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Getting 35 non-empty blocks out of 35 blocks
16/10/30 15:19:49 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/10/30 15:19:49 INFO Executor: Finished task 173.0 in stage 3.0 (TID 278). 1428 bytes result sent to driver
16/10/30 15:19:49 INFO TaskSetManager: Finished task 173.0 in stage 3.0 (TID 278) in 61 ms on localhost (179/179)
16/10/30 15:19:49 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
16/10/30 15:19:49 INFO DAGScheduler: ResultStage 3 (collect at RDDQueries.java:252) finished in 1.416 s
16/10/30 15:19:49 INFO DAGScheduler: Job 2 finished: collect at RDDQueries.java:252, took 26.062409 s
[[Lloyds Bank,164350], [Lloyds Bank,164313], [Lloyds Bank,164249], [Lloyds Bank,164223], [Lloyds Bank,163035], [Lloyds Bank,162829], [Guaranty Trust Bank,128312], [Guaranty Trust Ba
16/10/30 15:19:50 INFO SparkUI: Stopped Spark web UI at http://localhost:4040
16/10/30 15:19:50 INFO DAGScheduler: Stopping DAGScheduler
16/10/30 15:19:50 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
16/10/30 15:19:50 INFO MemoryStore: MemoryStore cleared
16/10/30 15:19:50 INFO BlockManager: BlockManager stopped
16/10/30 15:19:50 INFO BlockManagerMaster: BlockManagerMaster stopped
16/10/30 15:19:50 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
16/10/30 15:19:50 INFO SparkContext: Successfully stopped SparkContext
16/10/30 15:19:50 INFO RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
```

Source Code:

Source code is available on GitHub. Please find the URL below.

https://github.com/HemanthDantu/PB_Phase-2_Project/tree/master/Source%20Code

References:

https://github.com/nivdul/spark-in-practice/blob/master/src/main/java/com/handson/spark/dataframe/DataFrameOnTweets.java

https://github.com/AgilData/spark-rdd-dataframe-dataset/blob/master/src/main/java/example_java/dataframe/JavaDataFrameExample.java

https://spark.apache.org/docs/1.6.0/sql-programming-guide.html

http://www.programcreek.com/java-api-examples/index.php?api=org.apache.spark.sql.Row

https://github.com/stefani75/workspace/blob/b90a63f2f3028fb358b28a77ac416b223d37a52a/projet1/Hands-On-Spark-java-solution/src/main/java/com/duchessfr/spark/part3/sparksql/FunWithSparkSQL.java

https://github.com/AshokYaganti/PB_Phase2_TwitterAnalysis