

Principles of Big Data

Project Phase-2

Submitted by,

Divya Gaddam

Mounika Prathipati

Swathi Jasthi

Uma Maheshwara Reddy Mandapati

Introduction:**Goal:**

The main goal of this project is to collect social media data and implement analytical queries using Apache Spark (Spark RDDs & Data frames).

Topic:

In this project, we have collected twitter data based on favorite hero criteria.

Hashtag – HappyBirthdaySRK

We have implemented five queries using Apache Spark

1. Two queries using Spark RDDs.
2. Two queries should use Spark Data Frames.
3. One query calling the public APIs.

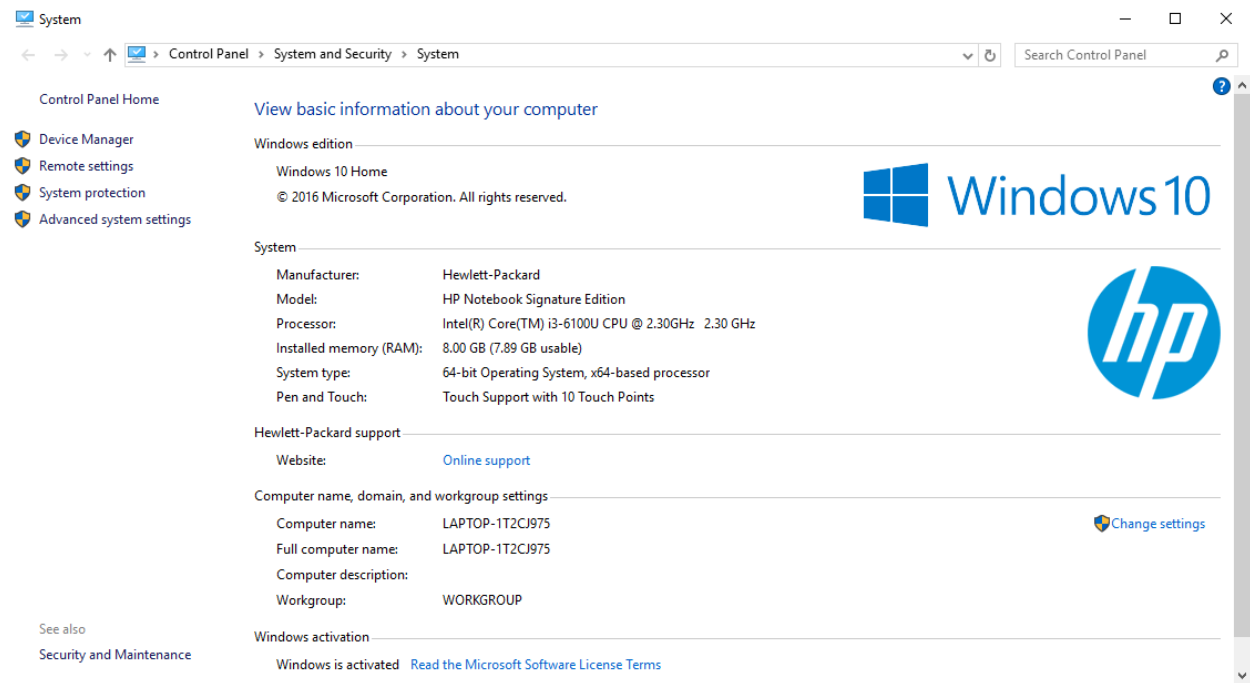
Software and programming languages Used:

1. Environment: Windows 10
2. Tweets collection: Python 2.10
3. Java 1.8.0
4. Scala 2.10.6
5. Spark 1.6.1
6. Hadoop 2.6
7. IntelliJ Idea Community V 16.

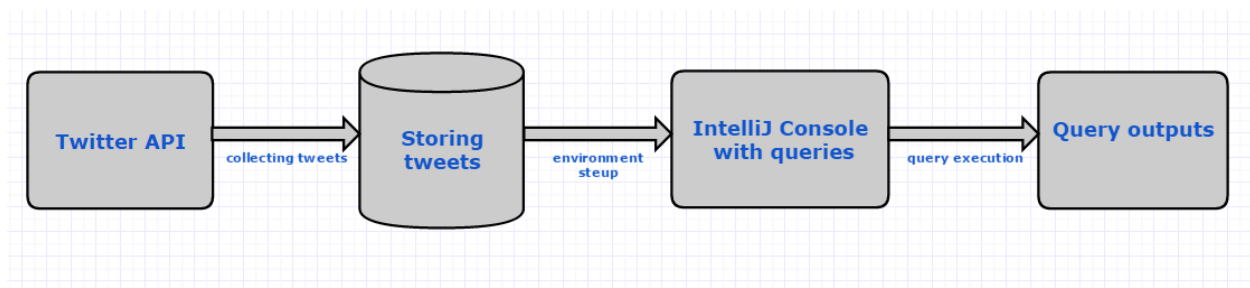
Architecture and Design:

1. Tweets are collected from twitter developer account using python program, based on twitter token.
2. The collected tweets are stores in JSON format.
3. Run the Spark RDD and Data frames analytical queries from IntelliJ using Scala plugins.
4. Outputs of executed queries are obtained in IntelliJ console.

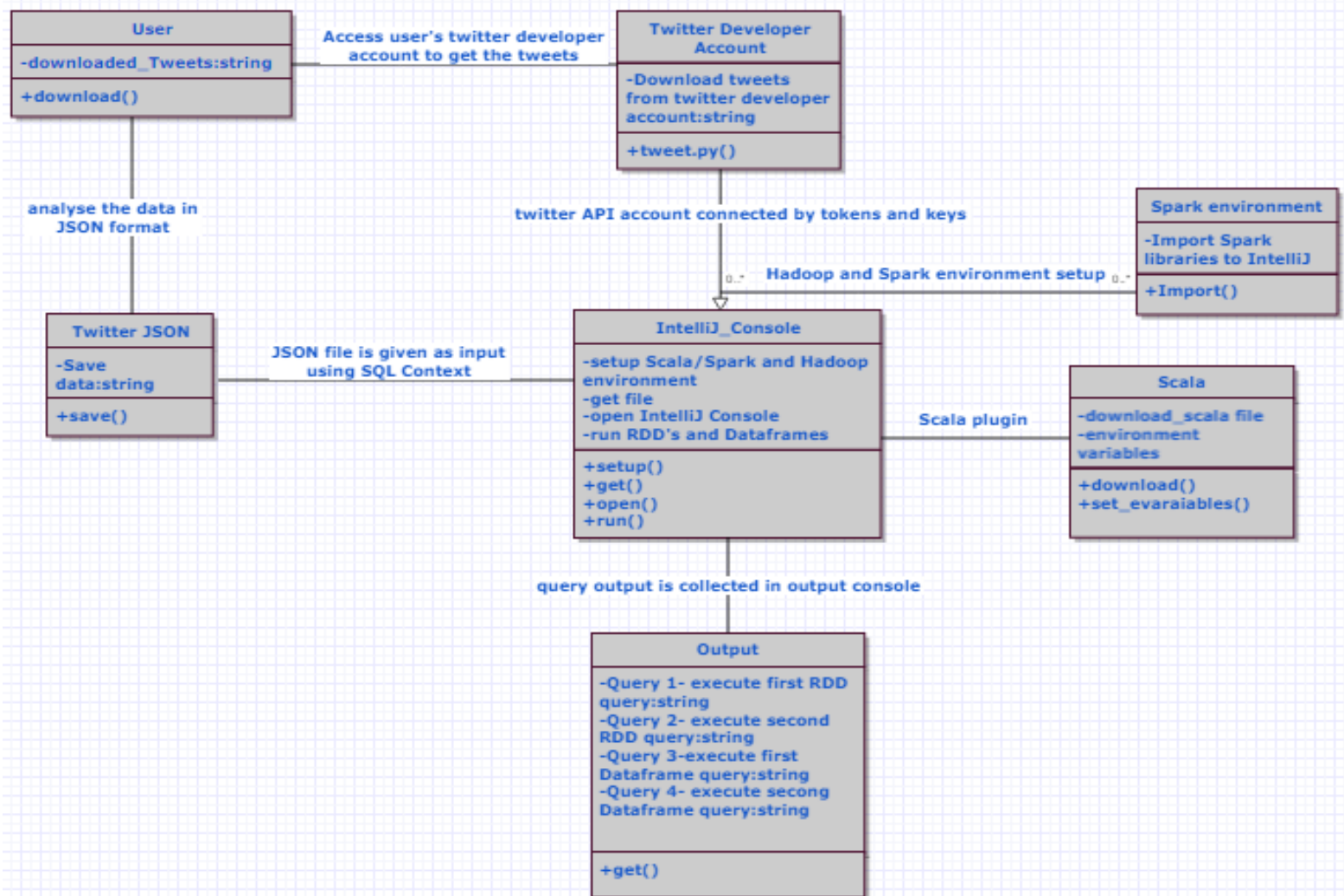
System Specifications:



Architecture diagram



UML diagram



DataFrame:

In Spark, a DataFrame is a distributed collection of data organized into named columns. It is conceptually equivalent to a table in a relational database or a data frame in R/Python, but with richer optimizations under the hood. DataFrames can be constructed from a wide array of sources such as: structured data files, tables in Hive, external databases, or existing RDDs.

RDD:

A Resilient Distributed Dataset (RDD), the basic abstraction in Spark. Represents an immutable, partitioned collection of elements that can be operated on in parallel. This class contains the basic operations available on all RDDs, such as map, filter,

and persist. Operations available on any RDD are imported from `org.apache.spark.SparkContext._`. (e.g. `RDD[(Int, Int)]`)

Internally, each RDD is characterized by five main properties:

- A list of partitions
- A function for computing each split
- A list of dependencies on other RDDs
- Optionally, a Partitioner for key-value RDDs (e.g. to say that the RDD is hash-partitioned)
- Optionally, a list of preferred locations to compute each split on (e.g. block locations for an HDFS file)

Queries:

Query1:

Firstly, reading the Json File “tweets2” and assigning it to the value RDD and mapping the values .

Query2:

Here we are filtering the line contains the word “protest” and displaying the first 20 results.

Query3:

Grouping the values in the column “filter_level” and getting the count of the values.

Query4:

Joining the files Tweets2 and joinQ based on the column “id”. (Nothing but joining using “outer join” the two files data using the column “id”).

Query5:

List out the data of the column “text”

Outer Join: Finds and returns the matching, dissimilar data from the tables tweets2 and joinQ

```
import org.apache.spark.sql.SQLContext
import org.apache.spark.{SparkConf, SparkContext}
import org.apache.spark.SparkConf
import org.apache.spark.SparkContext
import org.apache.hadoop.util

object newpl {
  def main(args: Array[String]): Unit = {
    val conf = new
    SparkConf().setAppName("query").setMaster("local[2]").set("spark.executor.memory",
```

```
"lg");
```

```
/*Query1*/
```

```
val sc = new SparkContext("local[2]", "PbSpark")
```

```
val sqlContext = new SQLContext(sc)
```

```
val RDD = sc.textFile("C:\\Users\\MAHESH\\Desktop\\tweets2.json/")
```

```
val RDDQuery =
```

```
sc.textFile("C:\\Users\\MAHESH\\Desktop\\tweets2.json").map(_.split(","))
```

```
/*Query2*/
```

```
RDDQuery.take(100).foreach(println)
```

```
val Data = RDD.filter(line => line.contains("protest"))
```

```
Data.take(20).foreach(println)
```

```
/*Query3*/
```

```
val stext = new SQLContext(sc)
```

```
val DataFrames = stext.jsonFile("C:\\Users\\MAHESH\\Desktop\\tweets2.json")
```

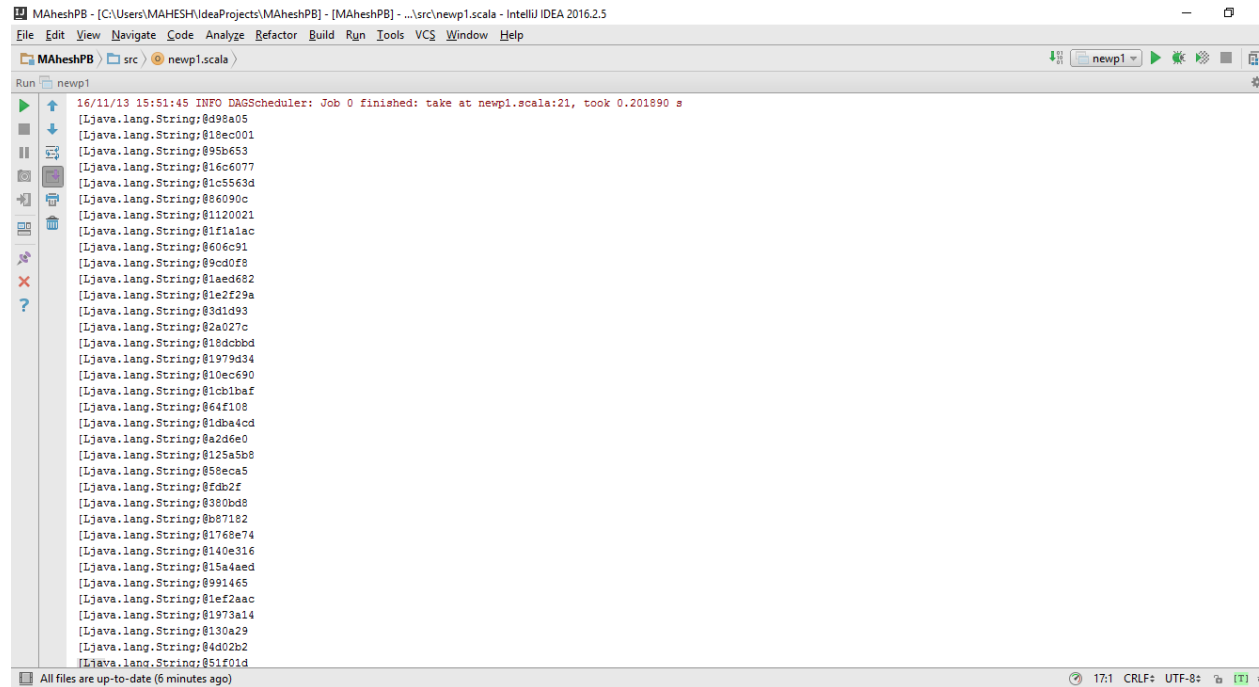
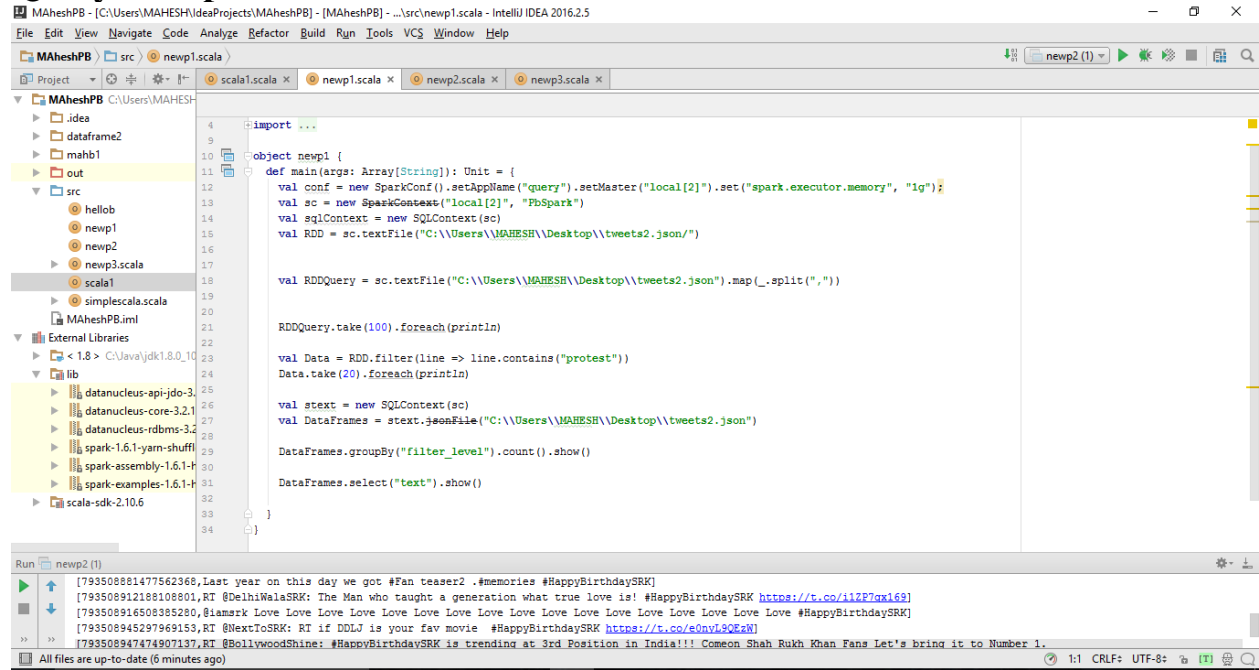
```
DataFrames.groupBy("filter_level").count().show()
```

```
/*Query4*/
```

```
DataFrames.select("text").show()
```

```
}  
}
```

Query1 Output:



MAheshPB - [C:\Users\MAHESH\IdeaProjects\MAheshPB] - [MAheshPB] - ...src\newp1.scala - IntelliJ IDEA 2016.2.5

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help

MAheshPB src newp1.scala newp1

Run newp1

```
[Ljava.lang.String;@4d02b2
[Ljava.lang.String;@51401d
[Ljava.lang.String;@503b19
[Ljava.lang.String;@fc2538
[Ljava.lang.String;@deec87
[Ljava.lang.String;@1342959
[Ljava.lang.String;@cf5915
[Ljava.lang.String;@1e76c11
[Ljava.lang.String;@cc183e
[Ljava.lang.String;@4ee8fd
[Ljava.lang.String;@1d6ef2
[Ljava.lang.String;@721044
[Ljava.lang.String;@5df5dc
[Ljava.lang.String;@a7b67f
[Ljava.lang.String;@1cc3f27
[Ljava.lang.String;@1aee22a
[Ljava.lang.String;@920bd6
[Ljava.lang.String;@1514735
[Ljava.lang.String;@1ee793f
[Ljava.lang.String;@15f69da
[Ljava.lang.String;@1b5021c
[Ljava.lang.String;@369db0
[Ljava.lang.String;@893e08
[Ljava.lang.String;@86807a
[Ljava.lang.String;@148d254
[Ljava.lang.String;@10f0c50
[Ljava.lang.String;@8f0007
[Ljava.lang.String;@90c13c
[Ljava.lang.String;@1c96e48
[Ljava.lang.String;@8598ad
[Ljava.lang.String;@d23c82
[Ljava.lang.String;@f78c85
[Ljava.lang.String;@be067d
[Ljava.lang.String;@4ab7f7
[Ljava.lang.String;@1adba10
[Ljava.lang.String;@17e8044
```

All files are un-to-date (9 minutes ago) 101:24 CRI F+ UTF-8

MAheshPB - [C:\Users\MAHESH\IdeaProjects\MAheshPB] - [MAheshPB] - ...src\newp1.scala - IntelliJ IDEA 2016.2.5

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help

MAheshPB src newp1.scala newp1

Run newp1

```
[Ljava.lang.String;@4d02b2
[Ljava.lang.String;@51401d
[Ljava.lang.String;@503b19
[Ljava.lang.String;@fc2538
[Ljava.lang.String;@deec87
[Ljava.lang.String;@1342959
[Ljava.lang.String;@cf5915
[Ljava.lang.String;@1e76c11
[Ljava.lang.String;@cc183e
[Ljava.lang.String;@4ee8fd
[Ljava.lang.String;@1d6ef2
[Ljava.lang.String;@721044
[Ljava.lang.String;@5df5dc
[Ljava.lang.String;@a7b67f
[Ljava.lang.String;@1cc3f27
[Ljava.lang.String;@1aee22a
[Ljava.lang.String;@920bd6
[Ljava.lang.String;@1514735
[Ljava.lang.String;@1ee793f
[Ljava.lang.String;@15f69da
[Ljava.lang.String;@1b5021c
[Ljava.lang.String;@369db0
[Ljava.lang.String;@893e08
[Ljava.lang.String;@86807a
[Ljava.lang.String;@148d254
[Ljava.lang.String;@10f0c50
[Ljava.lang.String;@8f0007
[Ljava.lang.String;@90c13c
[Ljava.lang.String;@1c96e48
[Ljava.lang.String;@8598ad
[Ljava.lang.String;@d23c82
[Ljava.lang.String;@f78c85
[Ljava.lang.String;@be067d
[Ljava.lang.String;@4ab7f7
[Ljava.lang.String;@1adba10
[Ljava.lang.String;@17e8044
```

All files are un-to-date (9 minutes ago) 101:24 CRI F+ UTF-8


```
MAheshPB - [C:\Users\MAHESH\IdeaProjects\MAheshPB] - [MAheshPB] - ...src\newp1.scala - IntelliJ IDEA 2016.2.5
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
MAheshPB src newp1.scala
Run newp1
[Ljava.lang.String;@1adba10
[Ljava.lang.String;@17e8044
[Ljava.lang.String;@10de6b8
[Ljava.lang.String;@1706e1
[Ljava.lang.String;@6c4689
[Ljava.lang.String;@1cf9de0
[Ljava.lang.String;@14d9f07
[Ljava.lang.String;@1efbac1
[Ljava.lang.String;@1e9273b
[Ljava.lang.String;@565afb
[Ljava.lang.String;@1d64b11
[Ljava.lang.String;@149c598
[Ljava.lang.String;@1d3aba5
[Ljava.lang.String;@1faa0a6
[Ljava.lang.String;@e9f00b
[Ljava.lang.String;@14b9e4b
[Ljava.lang.String;@11b8544
[Ljava.lang.String;@1dae791
[Ljava.lang.String;@8a2c09
[Ljava.lang.String;@1de5e95
[Ljava.lang.String;@5cdacf
[Ljava.lang.String;@3c55fa
[Ljava.lang.String;@1efcd90
[Ljava.lang.String;@b200ce
[Ljava.lang.String;@e83775
[Ljava.lang.String;@2924d7
[Ljava.lang.String;@15de58f
[Ljava.lang.String;@187305a
[Ljava.lang.String;@bbe000
[Ljava.lang.String;@181621c
[Ljava.lang.String;@19499fe
[Ljava.lang.String;@14d6736
[Ljava.lang.String;@a33c3f
16/11/13 15:51:45 INFO FileInputFormat: Total input paths to process : 1
16/11/13 15:51:45 INFO SparkContext: Starting job: take at newp1.scala:24
16/11/13 15:51:45 INFO DAGScheduler: Got job 1 (take at newp1.scala:24) with 1 output partitions
All files are up-to-date (10 minutes ago) 168-27 CRLF: UTF-8
```

Query2 Output:

```
MAheshPB - [C:\Users\MAHESH\IdeaProjects\MAheshPB] - [MAheshPB] - ...src\newp1.scala - IntelliJ IDEA 2016.2.5
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
MAheshPB src newp1.scala
Run newp1
16/11/13 15:51:45 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
16/11/13 15:51:45 INFO HadoopRDD: Input split: file:/C:/Users/MAHESH/Desktop/tweets2.json:0-33554432
16/11/13 15:51:45 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 81337 bytes result sent to driver
16/11/13 15:51:45 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 20 ms on localhost (1/1)
16/11/13 15:51:45 INFO DAGScheduler: ResultStage 1 (take at newp1.scala:24) finished in 0.020 s
16/11/13 15:51:45 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
16/11/13 15:51:45 INFO DAGScheduler: Job 1 finished: take at newp1.scala:24, took 0.025061 s
{"created_at":"Sun Oct 30 12:30:02 +0000 2016","id":"792705118772793344","id_str":"792705118772793344","text":"What abt kurbani bakras halal where r u than no appeal to save bakras @Ran
{"created_at":"Sun Oct 30 12:30:06 +0000 2016","id":"792705132941090816","id_str":"792705132941090816","text":"RT @s_rathore01: #ProtestCrackersProtectAnimals 1stProtect animals frm bn
{"created_at":"Sun Oct 30 12:30:16 +0000 2016","id":"792705178554150912","id_str":"792705178554150912","text":"RT @s_rathore01: #ProtestCrackersProtectAnimals 1stProtect animals frm bn
{"created_at":"Sun Oct 30 12:31:59 +0000 2016","id":"792705607379877888","id_str":"792705607379877888","text":"It's our duty to take care of those who don't have voice. Not only protest
{"created_at":"Sun Oct 30 12:34:42 +0000 2016","id":"792706291055468544","id_str":"792706291055468544","text":"Love environment protest pollution. Let live healthy #ProtestCrackersProte
{"created_at":"Sun Oct 30 12:34:43 +0000 2016","id":"792706291602842624","id_str":"792706291602842624","text":"#protestcrackersprotectanimals is now trending in #Mumbai https://t.co/v
{"created_at":"Sun Oct 30 12:35:36 +0000 2016","id":"792706519481610241","id_str":"792706519481610241","text":"Why do ppl need to protest against crackers to save animals. Simply save t
{"created_at":"Sun Oct 30 12:36:57 +0000 2016","id":"792706857102082048","id_str":"792706857102082048","text":"RT @s_rathore01: #ProtestCrackersProtectAnimals 1stProtect animals frm bn
{"created_at":"Sun Oct 30 12:37:11 +0000 2016","id":"792706915444875266","id_str":"792706915444875266","text":"#ProtestCrackersProtectAnimals be4 protesting we have 2 o other aspects al
{"created_at":"Sun Oct 30 12:38:03 +0000 2016","id":"792707133762576384","id_str":"792707133762576384","text":"RT @iam_rahul_raj: Why do ppl need to protest against crackers to save ani
{"created_at":"Sun Oct 30 12:38:13 +0000 2016","id":"792707176456294400","id_str":"792707176456294400","text":"First protest politician to save earth to ... #ProtestCrackersProtectAnima
{"created_at":"Sun Oct 30 12:38:36 +0000 2016","id":"7927072714389069824","id_str":"7927072714389069824","text":"RT @bahutKrantikari: They protest to eat beaf, they celebrate EID, but wan
{"created_at":"Sun Oct 30 12:39:43 +0000 2016","id":"792707554543595520","id_str":"792707554543595520","text":"RT @iam_rahul_raj: Why do ppl need to protest against crackers to save ani
{"created_at":"Sun Oct 30 12:40:39 +0000 2016","id":"792707788380057600","id_str":"792707788380057600","text":"#protestcrackersprotectanimals is now trending in India https://t.co/Oo
{"created_at":"Sun Oct 30 12:41:01 +0000 2016","id":"7927077881074208769","id_str":"7927077881074208769","text":"All intellectual will only protest Our Hindu festival like Ganesha and De
{"created_at":"Sun Oct 30 12:41:03 +0000 2016","id":"792707889920081920","id_str":"792707889920081920","text":"RT @s_rathore01: #ProtestCrackersProtectAnimals 1stProtect animals frm bn
{"created_at":"Sun Oct 30 12:41:22 +0000 2016","id":"792707970798936064","id_str":"792707970798936064","text":"RT @s_rathore01: #ProtestCrackersProtectAnimals 1stProtect animals frm bn
{"created_at":"Sun Oct 30 12:42:05 +0000 2016","id":"792708148721127426","id_str":"792708148721127426","text":"RT @s_rathore01: #ProtestCrackersProtectAnimals 1stProtect animals frm bn
{"created_at":"Sun Oct 30 12:43:07 +0000 2016","id":"792708410663895040","id_str":"792708410663895040","text":"RT @bahutKrantikari: They protest to eat beaf, they celebrate EID, but wan
{"created_at":"Sun Oct 30 12:43:12 +0000 2016","id":"792708432625270784","id_str":"792708432625270784","text":"RT @iam_rahul_raj: Why do ppl need to protest against crackers to save ani
16/11/13 15:51:46 INFO JSONRelation: Listing file:/C:/Users/MAHESH/Desktop/tweets2.json on driver
16/11/13 15:51:46 INFO MemoryStore: Block broadcast_4 stored as values in memory (estimated size 187.7 KB, free 455.8 KB)
16/11/13 15:51:46 INFO MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated size 19.4 KB, free 475.2 KB)
16/11/13 15:51:46 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory on localhost:52000 (size: 19.4 KB, free: 517.4 KB)
16/11/13 15:51:46 INFO SparkContext: Created broadcast 4 from jsonFile at newp1.scala:27
16/11/13 15:51:46 INFO FileInputFormat: Total input paths to process : 1
16/11/13 15:51:46 INFO SparkContext: Starting job: jsonFile at newp1.scala:27
16/11/13 15:51:46 INFO DAGScheduler: Got job 2 (jsonFile at newp1.scala:27) with 4 output partitions
16/11/13 15:51:46 INFO DAGScheduler: Final stage: ResultStage 2 (jsonFile at newp1.scala:27)
All files are up-to-date (11 minutes ago) 215-27 CRLF: UTF-8
```

Query3 Output:

```
MAheshPB - [C:\Users\MAHESH\IdeaProjects\MAheshPB] - [MAheshPB] - ...src\newp1.scala - IntelliJ IDEA 2016.2.5
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
MAheshPB src newp1.scala
Run newp1
16/11/13 15:52:01 INFO TaskSetManager: Finished task 194.0 in stage 6.0 (TID 205) in 4 ms on localhost (195/199)
16/11/13 15:52:01 INFO ShuffleBlockFetcherIterator: Getting 4 non-empty blocks out of 4 blocks
16/11/13 15:52:01 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/11/13 15:52:01 INFO ShuffleBlockFetcherIterator: Getting 4 non-empty blocks out of 4 blocks
16/11/13 15:52:01 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/11/13 15:52:01 INFO Executor: Finished task 196.0 in stage 6.0 (TID 207). 1609 bytes result sent to driver
16/11/13 15:52:01 INFO Executor: Finished task 195.0 in stage 6.0 (TID 206). 1609 bytes result sent to driver
16/11/13 15:52:01 INFO TaskSetManager: Starting task 197.0 in stage 6.0 (TID 208, localhost, partition 198, NODE_LOCAL, 1999 bytes)
16/11/13 15:52:01 INFO Executor: Running task 197.0 in stage 6.0 (TID 208)
16/11/13 15:52:01 INFO TaskSetManager: Finished task 196.0 in stage 6.0 (TID 207) in 4 ms on localhost (196/199)
16/11/13 15:52:01 INFO TaskSetManager: Finished task 195.0 in stage 6.0 (TID 206) in 4 ms on localhost (197/199)
16/11/13 15:52:01 INFO TaskSetManager: Starting task 198.0 in stage 6.0 (TID 209, localhost, partition 199, NODE_LOCAL, 1999 bytes)
16/11/13 15:52:01 INFO Executor: Running task 198.0 in stage 6.0 (TID 209)
16/11/13 15:52:01 INFO ShuffleBlockFetcherIterator: Getting 4 non-empty blocks out of 4 blocks
16/11/13 15:52:01 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/11/13 15:52:01 INFO ShuffleBlockFetcherIterator: Getting 4 non-empty blocks out of 4 blocks
16/11/13 15:52:01 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/11/13 15:52:01 INFO Executor: Finished task 197.0 in stage 6.0 (TID 208). 1609 bytes result sent to driver
16/11/13 15:52:01 INFO Executor: Finished task 198.0 in stage 6.0 (TID 209). 1609 bytes result sent to driver
16/11/13 15:52:01 INFO TaskSetManager: Finished task 197.0 in stage 6.0 (TID 208) in 4 ms on localhost (198/199)
16/11/13 15:52:01 INFO TaskSetManager: Finished task 198.0 in stage 6.0 (TID 209) in 4 ms on localhost (199/199)
16/11/13 15:52:01 INFO TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool
16/11/13 15:52:01 INFO DAGScheduler: ResultStage 6 (show at newp1.scala:29) finished in 0.540 s
16/11/13 15:52:01 INFO DAGScheduler: Job 4 finished: show at newp1.scala:29, took 0.579618 s

-----+-----+
|filter_level|count|
-----+-----+
|      low|22817|
|      null|22859|
-----+-----+

16/11/13 15:52:01 INFO MemoryStore: Block broadcast_11 stored as values in memory (estimated size 187.3 KB, free 688.3 KB)
16/11/13 15:52:01 INFO MemoryStore: Block broadcast_11_piece0 stored as bytes in memory (estimated size 19.3 KB, free 707.7 KB)
16/11/13 15:52:01 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on localhost:52000 (size: 19.3 KB, free: 517.3 KB)
16/11/13 15:52:01 INFO SparkContext: Created broadcast 11 from show at newp1.scala:31
16/11/13 15:52:01 INFO MemoryStore: Block broadcast_12 stored as values in memory (estimated size 187.7 KB, free 895.4 KB)
All files are up-to-date (12 minutes ago) 879:3 CRLF+ UTF-8+
```

Query4 Output:

```
MAheshPB - [C:\Users\MAHESH\IdeaProjects\MAheshPB] - [MAheshPB] - ...src\newp1.scala - IntelliJ IDEA 2016.2.5
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
MAheshPB src newp1.scala
Run newp1
16/11/13 15:52:01 INFO DAGScheduler: ResultStage 7 (show at newp1.scala:31) finished in 0.012 s
16/11/13 15:52:01 INFO TaskSchedulerImpl: Removed TaskSet 7.0, whose tasks have all completed, from pool
16/11/13 15:52:01 INFO DAGScheduler: Job 5 finished: show at newp1.scala:31, took 0.016065 s
16/11/13 15:52:01 INFO SparkContext: Invoking stop() from shutdown hook

-----+-----+
|      text|
-----+-----+
|RT @sunilsanjan: ...|
|      null|
|What abt kurbani ...|
|      null|
|Kisjiye boss ko im...|
|      null|
|RT @ss_rathore01:...|
|      null|
|RT @Bharat_Mantha...|
|      null|
|RT @KshatriyaVidh...|
|      null|
|Hindus scare poor...|
|      null|
|RT @ss_rathore01:...|
|      null|
|Stock Photo: Diwa...|
|      null|
|#ProtestCrackersP...|
|      null|
-----+-----+
only showing top 20 rows

16/11/13 15:52:01 INFO SparkUI: Stopped Spark web UI at http://192.168.1.153:4041
16/11/13 15:52:01 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
16/11/13 15:52:01 INFO MemoryStore: MemoryStore cleared
16/11/13 15:52:01 INFO BlockManager: BlockManager stopped
16/11/13 15:52:01 INFO BlockManagerMaster: BlockManagerMaster stopped
16/11/13 15:52:01 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
All files are up-to-date (12 minutes ago) 1609:58 CRLF+ UTF-8+
```

Runtime measurements for 4 Queries:

The screenshot shows the PbSpark application interface. At the top, there's a browser window with the address bar showing '192.168.1.153:4041/jobs/'. Below the browser window, the application has a navigation bar with tabs: 'Jobs', 'Stages', 'Storage', 'Environment', 'Executors', 'SQL', and 'SQL1'. The 'Jobs' tab is selected. The main content area is titled 'Spark Jobs (?)' and displays summary statistics: 'Total Uptime: 13 s', 'Scheduling Mode: FIFO', 'Active Jobs: 1', and 'Completed Jobs: 2'. There's a link for 'Event Timeline'. Below this, the 'Active Jobs (1)' section shows a table with one job. The 'Completed Jobs (2)' section shows a table with two jobs.

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
2	jsonFile at newp1.scala:27	2016/11/13 14:45:15	9 s	0/1	2/4

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
1	take at newp1.scala:24	2016/11/13 14:45:15	31 ms	1/1	1/1
0	take at newp1.scala:21	2016/11/13 14:45:15	0.2 s	1/1	1/1

Join query

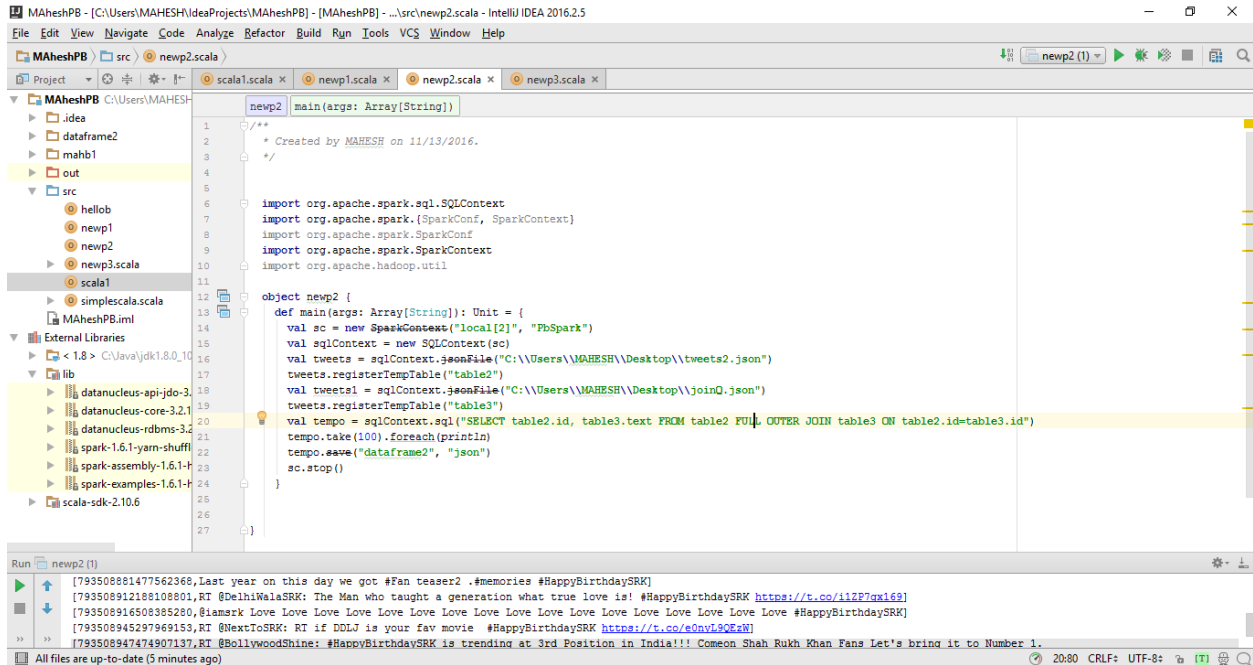
Joining the files Tweets2 and joinQ based on the column “id”. (Nothing but joining using “outer join” the two files data using the column “id”).

```
import org.apache.spark.sql.SQLContext
import org.apache.spark.{SparkConf, SparkContext}
import org.apache.spark.SparkConf
import org.apache.spark.SparkContext
import org.apache.hadoop.util

object newp2 {
  def main(args: Array[String]): Unit = {
    val sc = new SparkContext("local[2]", "PbSpark")
    val sqlContext = new SQLContext(sc)
    val tweets = sqlContext.jsonFile("C:\\Users\\MAHESH\\Desktop\\tweets2.json")
    tweets.registerTempTable("table2")
    val tweets1 = sqlContext.jsonFile("C:\\Users\\MAHESH\\Desktop\\joinQ.json")
    tweets1.registerTempTable("table3")
    val tempo = sqlContext.sql("SELECT table2.id, table3.text FROM table2 FULL OUTER
JOIN table3 ON table2.id=table3.id")
    tempo.take(100).foreach(println)
    tempo.save("dataframe2", "json")
    sc.stop()
  }
}
```

}

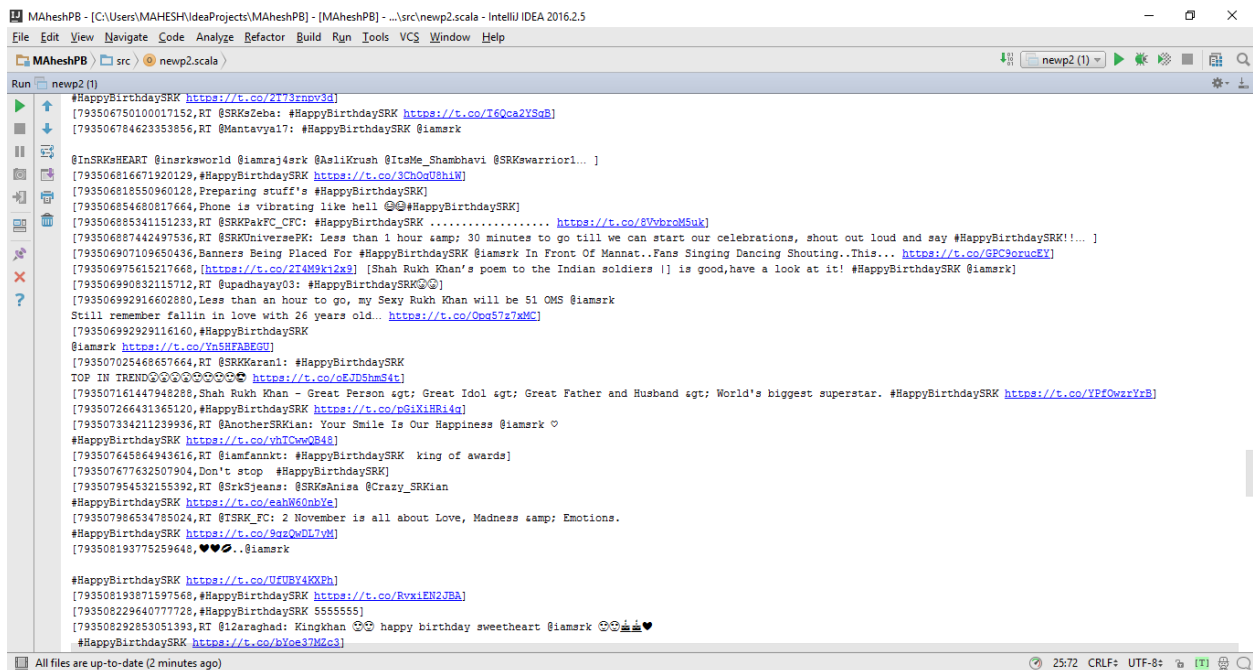
Join Query Output:



The screenshot shows the IntelliJ IDEA interface with a Scala file named `newp2.scala` open. The code defines a `newp2` object with a `main` function that uses Spark SQL to join two tables. The `main` function takes an array of strings as input, creates a SparkContext, and registers two temporary tables, `table2` and `table3`, from JSON files. It then performs a full outer join on these tables based on the `id` column and prints the results.

```
1  /**
2   * Created by MAHESH on 11/13/2016.
3   */
4
5
6  import org.apache.spark.sql.SQLContext
7  import org.apache.spark.{SparkConf, SparkContext}
8  import org.apache.spark.SparkConf
9  import org.apache.spark.SparkContext
10 import org.apache.hadoop.util
11
12 object newp2 {
13   def main(args: Array[String]): Unit = {
14     val sc = new SparkContext("local[2]", "PbSpark")
15     val sqlContext = new SQLContext(sc)
16     val tweets = sqlContext.jsonFile("C:\\Users\\MAHESH\\Desktop\\tweets2.json")
17     tweets.registerTempTable("table2")
18     val tweets1 = sqlContext.jsonFile("C:\\Users\\MAHESH\\Desktop\\joinQ.json")
19     tweets1.registerTempTable("table3")
20     val tempo = sqlContext.sql("SELECT table2.id, table3.text FROM table2 FULL OUTER JOIN table3 ON table2.id=table3.id")
21     tempo.take(100).foreach(println)
22     tempo.save("dataframe2", "json")
23     sc.stop()
24   }
25 }
26
27 }
```

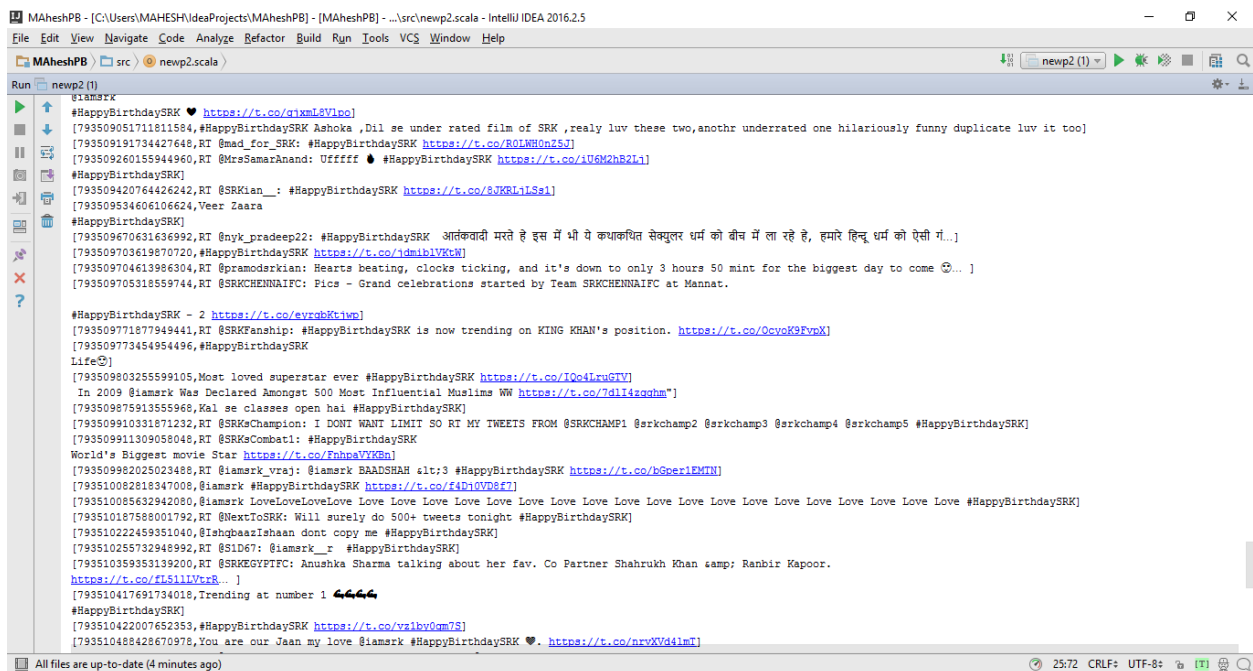
The Run console shows the output of the query, displaying a list of tweets with their IDs and text. The output is truncated with "...".



The screenshot shows the IntelliJ IDEA interface with a Scala file named `newp2.scala` open. The code defines a `newp2` object with a `main` function that uses Spark SQL to join two tables. The `main` function takes an array of strings as input, creates a SparkContext, and registers two temporary tables, `table2` and `table3`, from JSON files. It then performs a full outer join on these tables based on the `id` column and prints the results.

```
1  /**
2   * Created by MAHESH on 11/13/2016.
3   */
4
5
6  import org.apache.spark.sql.SQLContext
7  import org.apache.spark.{SparkConf, SparkContext}
8  import org.apache.spark.SparkConf
9  import org.apache.spark.SparkContext
10 import org.apache.hadoop.util
11
12 object newp2 {
13   def main(args: Array[String]): Unit = {
14     val sc = new SparkContext("local[2]", "PbSpark")
15     val sqlContext = new SQLContext(sc)
16     val tweets = sqlContext.jsonFile("C:\\Users\\MAHESH\\Desktop\\tweets2.json")
17     tweets.registerTempTable("table2")
18     val tweets1 = sqlContext.jsonFile("C:\\Users\\MAHESH\\Desktop\\joinQ.json")
19     tweets1.registerTempTable("table3")
20     val tempo = sqlContext.sql("SELECT table2.id, table3.text FROM table2 FULL OUTER JOIN table3 ON table2.id=table3.id")
21     tempo.take(100).foreach(println)
22     tempo.save("dataframe2", "json")
23     sc.stop()
24   }
25 }
26
27 }
```

The Run console shows the output of the query, displaying a list of tweets with their IDs and text. The output is truncated with "...".



```
MAHeshPB - [C:\Users\MAHESH\IdeaProjects\MAHeshPB] - [MAHeshPB] - ...src\newp2.scala - IntelliJ IDEA 2016.2.5
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
MAHeshPB src newp2.scala
Run newp2 (1)
[793510255732949992, RT @S1D67: @iamrk_r #HappyBirthdaySRK]
[793510359353139200, RT @SRKEGYPTFC: Anushka Sharma talking about her fav. Co Partner Shahrukh Khan &#amp; Ranbir Kapoor.
https://t.co/f1511Vtr8. ]
[793510417691734018, Trending at number 1 <<<<
#HappyBirthdaySRK]
[793510422007632353, #HappyBirthdaySRK https://t.co/vz1bv0gm7S]
[793510488428670978, You are our Jaan my love @iamrk #HappyBirthdaySRK ♡. https://t.co/nrvXVd4lml]
[793510562562973696, RT @iamshuvam08: #HappyBirthdaySRK. Love u @iamrk]
[793510591306534914, RT @NextToSRK: Faster #HappyBirthdaySRK]
[793510771166556160, RT @KingSrkians: LiveUpdate :- Outside Of Mannat @iamrk #HappyBirthdaySRK https://t.co/FE4DDFC1v9]
[793510841513488384, RT @KKKG_TUMBLR: #TickTock #HappyBirthdaySRK #ADHM https://t.co/5SPFn2141R]
[793510939161202689, RT @rehanSheikh: Swades - It hits you where your heart hurts the most; root feelings with good intentions overcoming societal obstacle. #Ha..]
[793510974644903937, #HappyBirthdaySRK @iamrk https://t.co/RgBT1vFg7L]
[793511078302973952, RT @The_Witty_SRK: #HappyBirthdaySRK https://t.co/soKoaIQnIc]
#HappyBirthdaySRK https://t.co/qC45vnDk1m
[793511113354801152, RT @iamrk_manash: Zindagi toh har roz jaan leti hai ... bomb toh sirf ek baar lega
#HappyBirthdaySRK]
[793511146670075904, RT @lmsrkian1: #HappyBirthdaySRK NOW TRENDING AT OUR BAADSHAH 'S POSITION !!!

KEEP TWEETING &#amp; RTING https://t.co/nvw8CoAQ9a]
[793511178769006593, RT @SRK_Ki_Chetu: #HappyBirthdaySRK
All hot girls put ur hands up and say.... Its SRK DAY.. ALL COOL BOYS CMON MAKE SOME NOICE AND SAY SRK..]
[793511232636583940, RT @rehanSheikh: Remembered Last Year Best Memories Ever #HappyBirthdaySRK https://t.co/2AVZ1eqYZv]
[793511251611615232, ! #HappyBirthdaySRK]
[793511594948952064, #HappyBirthdaySRK https://t.co/fN1SAgf38A]
[793511596123389952, King khan <lt;3 @iamrk #HappyBirthdaySRK]
Exception in thread "main" org.apache.spark.sql.AnalysisException: path file:/C:/Users/MAHESH/IdeaProjects/MAHeshPB/dataframe2 already exists.:
    at org.apache.spark.sql.execution.datasources.InsertIntoHadoopFSRelation.run(InsertIntoHadoopFSRelation.scala:76)
    at org.apache.spark.sql.execution.ExecutedCommand.sideEffectResult$lzycompute(commands.scala:58)
    at org.apache.spark.sql.execution.ExecutedCommand.sideEffectResult(commands.scala:56)
    at org.apache.spark.sql.execution.ExecutedCommand.doExecute(commands.scala:70)
    at org.apache.spark.sql.execution.SparkPlan.$anonfun$execute$5.apply(SparkPlan.scala:132)
    at org.apache.spark.sql.execution.SparkPlan.$anonfun$execute$5.apply(SparkPlan.scala:130)
    at org.apache.spark.rdd.RDDOperationScope$.withScope(RDDOperationScope.scala:150)
    at org.apache.spark.sql.execution.SparkPlan.execute(SparkPlan.scala:130)
    at org.apache.spark.sql.execution.QueryExecution.toRdd$lzycompute(QueryExecution.scala:55)
    at org.apache.spark.sql.execution.QueryExecution.toRdd(QueryExecution.scala:55)
All files are up-to-date (4 minutes ago) 25:72 CRLF+ UTF-8+ [?] [?]
```

Runtime measurement for join query:

PbSpark - Spark Jobs

192.168.1.153:4041/jobs/

Spark 1.6.1

Jobs

Stages

Storage

Environment

Executors

SQL

Pb Spark application UI

Spark Jobs (?)

Total Uptime: 9 s

Scheduling Mode: FIFO

Active Jobs: 1

Event Timeline

Active Jobs (1)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	jsonFile at newp2.scala:16	2016/11/13 16:05:27	5 s	0/1	0/4

Public API:

We have used Twitter Rest API

The REST APIs provide programmatic access to read and write Twitter data. Create a new Tweet, read user profile and follower data, and more. The REST API identifies Twitter applications and users using OAuth; responses are in JSON format.

Query:

```
package com.devdaily.twitterclient
import twitter4j.TwitterFactory
import twitter4j.Twitter
import twitter4j.conf.ConfigurationBuilder

object newp3 {

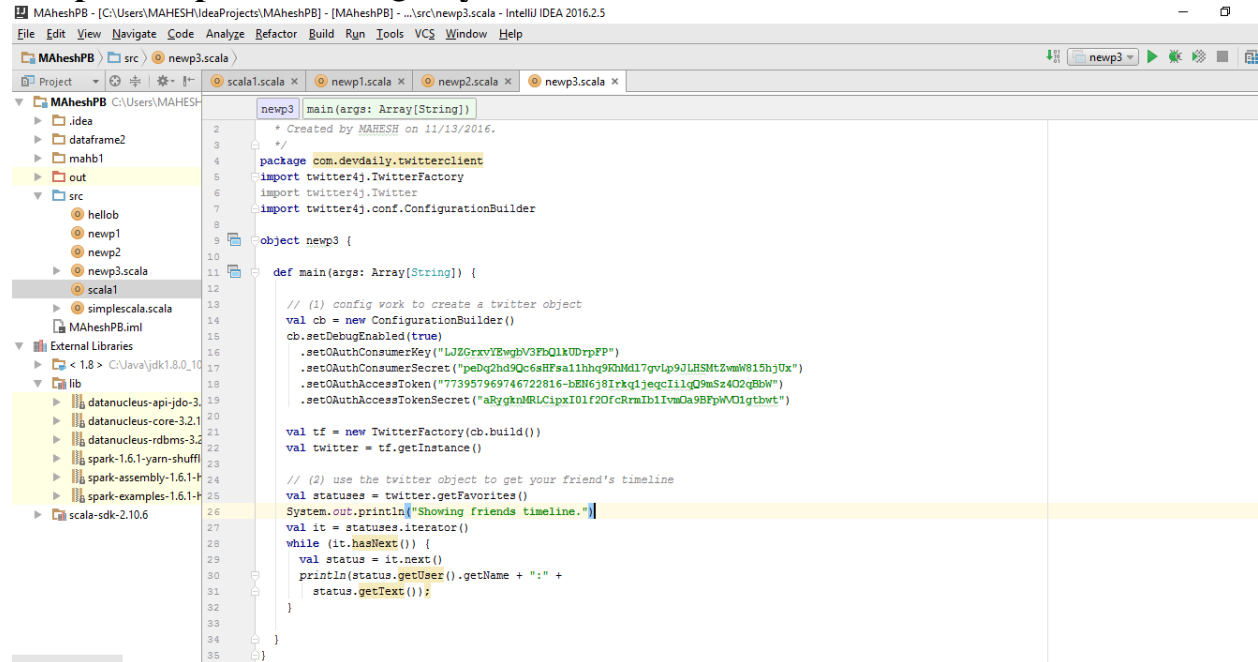
    def main(args: Array[String]) {

        // (1) config work to create a twitter object
        val cb = new ConfigurationBuilder()
        cb.setDebugEnabled(true)
            .setOAuthConsumerKey("LJZGrxvYEwgbV3FbQ1kUDrpFP")
            .setOAuthConsumerSecret("peDq2hd9Qc6sHFsa11hhq9KhMd17gvLp9JLHSMtZwmW815hjUx")
            .setOAuthAccessToken("773957969746722816-bEN6j8Irkq1jeqcIilqQ9mSz4O2qBbW")
            .setOAuthAccessTokenSecret("aRygknMRLCipxI0lf2OfcRrmIb1IvmOa9BFpWVO1gtbwt")

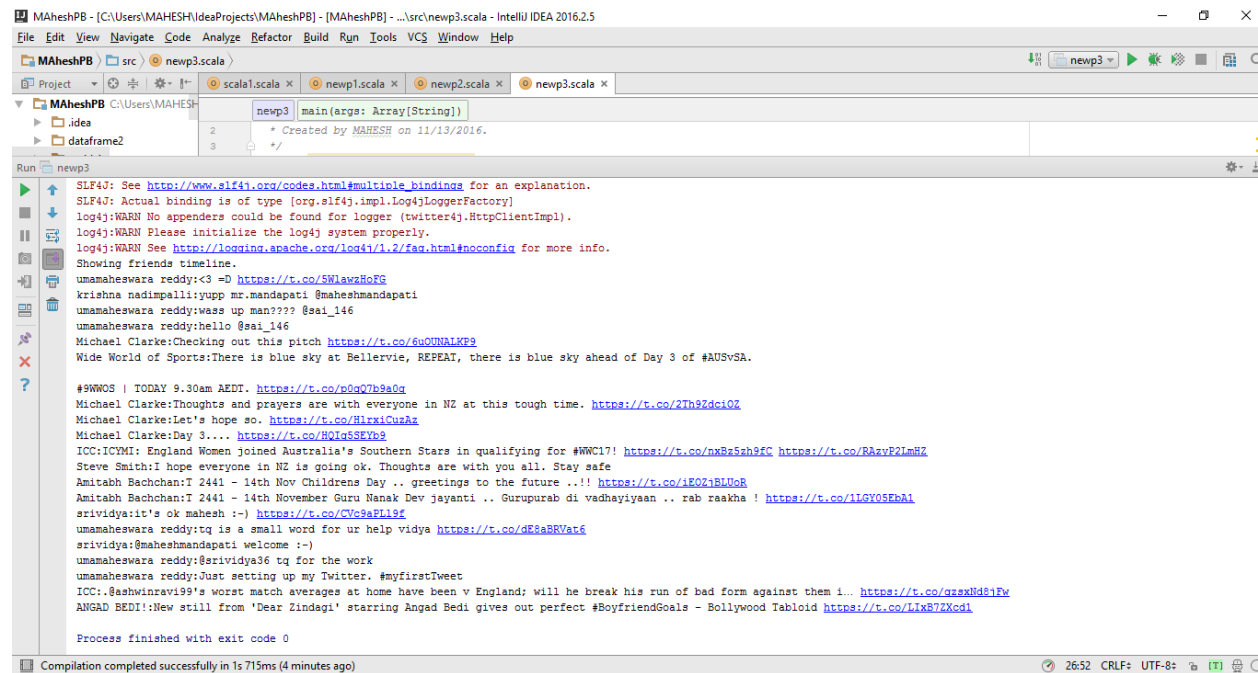
        val tf = new TwitterFactory(cb.build())
        val twitter = tf.getInstance()

        // (2) use the twitter object to get your friend's timeline
        val statuses = twitter.getFavorites()
        System.out.println("Showing friends timeline.")
        val it = statuses.iterator()
        while (it.hasNext()) {
            val status = it.next()
            println(status.getUser().getName + ":" +
                status.getText());
        }
    }
}
```


Output for public API Query:



```
newp3 main(args: Array[String])
2
3 * Created by MAHESH on 11/13/2016.
4
5 package com.devdaily.twitterclient
6 import twitter4j.TwitterFactory
7 import twitter4j.Twitter
8 import twitter4j.conf.ConfigurationBuilder
9
10 object newp3 {
11
12   def main(args: Array[String]) {
13
14     // (1) config work to create a twitter object
15     val cb = new ConfigurationBuilder()
16     cb.setDebugEnabled(true)
17     .setOAuthConsumerKey("lWZGrxvYwgbV3FbQlkUDrpFP")
18     .setOAuthConsumerSecret("peDq2hd9Qc6aHfSa11hhq9RmM17gvlp9JLHSMt2wmW615hjUx")
19     .setOAuthAccessToken("77395796974672816-bEN6j8Irkq1jeqI1q9mS402qBbW")
20     .setOAuthAccessTokenSecret("aRyqknRLCipxI01f2OfcRrmIb1IvmDa98FpW01gtbwt")
21
22     val tf = new TwitterFactory(cb.build())
23     val twitter = tf.getInstance()
24
25     // (2) use the twitter object to get your friend's timeline
26     val statuses = twitter.getFavorites()
27     val it = statuses.iterator()
28     while (it.hasNext()) {
29       val status = it.next()
30       println(status.getUser().getName + ":" +
31         status.getText())
32     }
33   }
34 }
35 }
```



```
Run newp3
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
log4j:WARN No appenders could be found for logger (twitter4j.HttpClientImpl).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Showing friends timeline.
umamaheswara_reddy: <3 =D https://t.co/5WJawzHoFG
kriehna nadimpalli:yupp mr.mandapati @maheshmandapati
umamaheswara_reddy:wass up man???? @sai_146
umamaheswara_reddy:hello @sai_146
Michael Clarke:Checking out this pitch https://t.co/6uQUNALKP9
Wide World of Sports:There is blue sky at Bellervie, REPEAT, there is blue sky ahead of Day 3 of #AUSvsSA.

#9WMO5 | TODAY 9.30am AEDT. https://t.co/p0gQ7b9a0q
Michael Clarke:Thoughts and prayers are with everyone in NZ at this tough time. https://t.co/2Th82dci0Z
Michael Clarke:Let's hope so. https://t.co/WlxwiCua2g
Michael Clarke:Day 3.... https://t.co/HQIGSEVB9
ICC:ICYMI: England Women joined Australia's Southern Stars in qualifying for #WNC17! https://t.co/nx8s5sh8fC https://t.co/RAzyF2lmWZ
Steve Smith:I hope everyone in NZ is going ok. Thoughts are with you all. Stay safe
Amitabh Bachchan:T 2441 - 14th Nov Childrens Day .. greetings to the future ...! https://t.co/iEQ2i8LUoR
Amitabh Bachchan:T 2441 - 14th November Guru Nanak Dev Jayanti .. Gurupurab di vadhaiyaan .. rab raakha ! https://t.co/1IGY0SEba1
srividya:it's ok mahesh :-)) https://t.co/CY9aPL19f
umamaheswara_reddy:tq is a small word for ur help vidya https://t.co/dE2aBRVar6
srividya:@maheshmandapati welcome :-))
umamaheswara_reddy:@srividya36 tq for the work
umamaheswara_reddy:Just setting up my Twitter. #myfirstTweet
ICC:.@ashwinravi99's worst match averages at home have been v England; will he break his run of bad form against them i... https://t.co/gzxwRd41Fv
ANGAD BEDI!:New still from 'Dear Zindagi' starring Angad Bedi gives out perfect #BoyfriendGoals - Bollywood Tabloid https://t.co/Lix87ZXcd1

Process finished with exit code 0
```

Reference:

<https://databricks.com/blog/2015/02/17/introducing-dataframes-in-spark-for-large-scale-data-science.html>

<https://spark.apache.org/docs/0.8.1/api/core/org/apache/spark/rdd/RDD.html>

<https://dev.twitter.com/rest/public>