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.

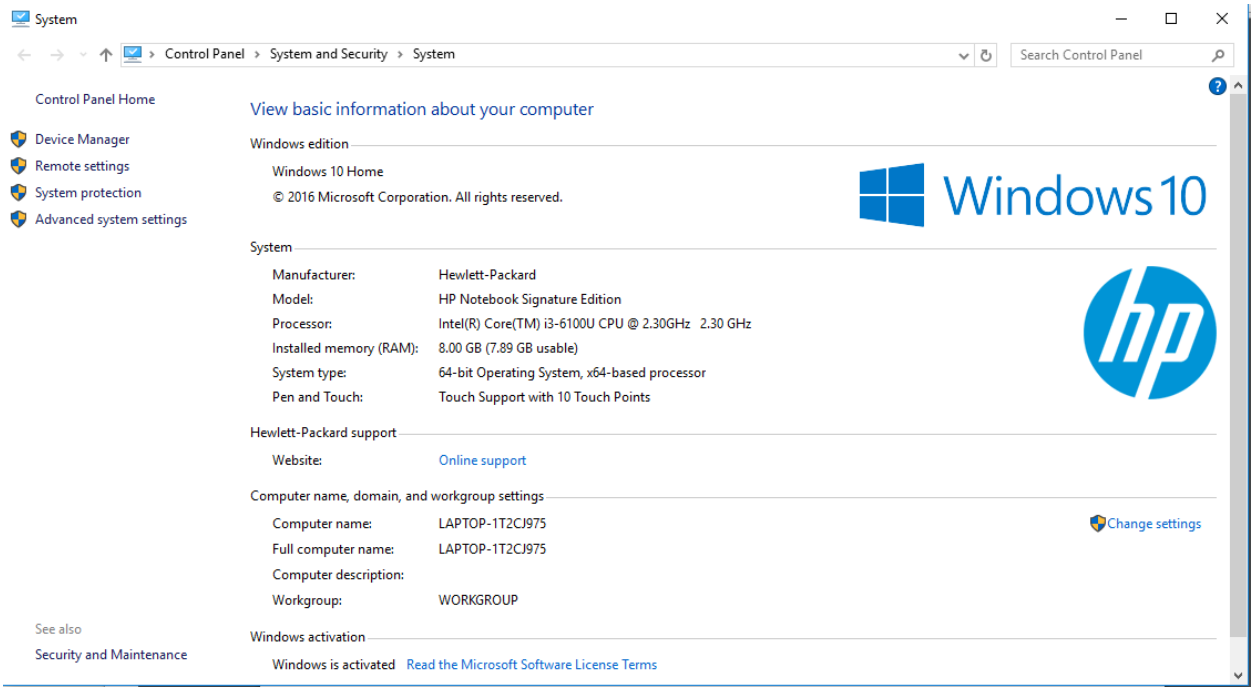
Software and programming languages Used:

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.

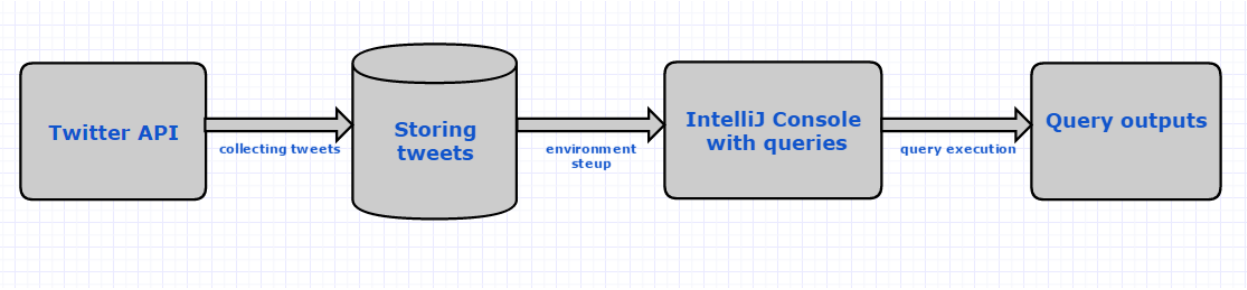
Architecture and Design:

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.

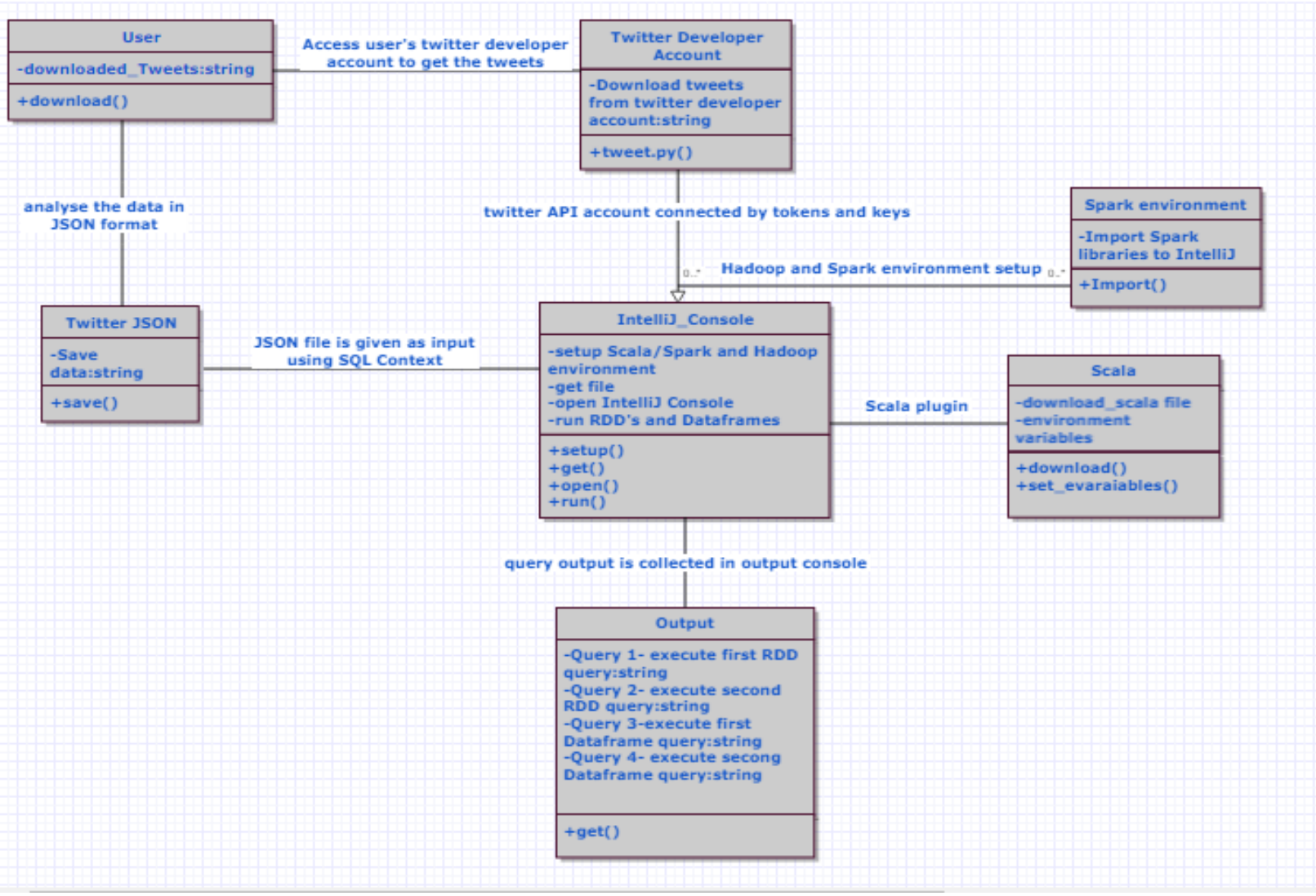
System Specifications:



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 newp1 {
  def main(args: Array[String]): Unit = {
    val conf = new
SparkConf().setAppName("query").setMaster("local[2]").set("spark.executor.memory",
```

```
"1g");

/*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 〉

scala1.scala × | newp1.scala × | newp2.scala × | newp3.scala ×

```scala
import ...

object newp1 {
  def main(args: Array[String]): Unit = {
    val conf = new SparkConf().setAppName("query").setMaster("local[2]").set("spark.executor.memory", "1g");
    val sc = new SparkContext("local[2]", "FbSpark")
    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(","))


    RDDQuery.take(100).foreach(println)

    val Data = RDD.filter(line => line.contains("protest"))
    Data.take(20).foreach(println)

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

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

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

  }
}
```

Run newp2 (1)

```
[793508881477562368,Last year on this day we got #Fan teaser2 .#memories #HappyBirthdaySRK]
[793508912188108801,RT @DelhiWalaSRK: The Man who taught a generation what true love is! #HappyBirthdaySRK https://t.co/iiZP7qx169]
[793508916508385280,@iamsrk Love Love Love Love Love Love Love Love Love Love Love Love Love Love Love Love Love Love Love Love #HappyBirthdaySRK]
[793508945297969153,RT @NextToSRK: RT if DDLJ is your fav movie  #HappyBirthdaySRK https://t.co/e0nvL9QEzW]
[793508947474907137,RT @BollywoodShine: #HappyBirthdaySRK is trending at 3rd Position in India!!! Comeon Shah Rukh Khan Fans Let's bring it to Number 1.
```

All files are up-to-date (6 minutes ago) 1:1 CRLF÷ 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

```
16/11/13 15:51:45 INFO DAGScheduler: Job 0 finished: take at newp1.scala:21, took 0.201890 s
[Ljava.lang.String;@d98a05
[Ljava.lang.String;@18ec001
[Ljava.lang.String;@95b653
[Ljava.lang.String;@16c6077
[Ljava.lang.String;@1c5563d
[Ljava.lang.String;@86090c
[Ljava.lang.String;@1120021
[Ljava.lang.String;@1f1a1ac
[Ljava.lang.String;@606c91
[Ljava.lang.String;@9cd0f8
[Ljava.lang.String;@1aed682
[Ljava.lang.String;@1e2f29a
[Ljava.lang.String;@3d1d93
[Ljava.lang.String;@2a027c
[Ljava.lang.String;@18dcbbd
[Ljava.lang.String;@1979d34
[Ljava.lang.String;@10ec690
[Ljava.lang.String;@1cb1baf
[Ljava.lang.String;@64f108
[Ljava.lang.String;@1dba4cd
[Ljava.lang.String;@a2d6e0
[Ljava.lang.String;@125a5b8
[Ljava.lang.String;@58eca5
[Ljava.lang.String;@fdb2f
[Ljava.lang.String;@380bd8
[Ljava.lang.String;@b87182
[Ljava.lang.String;@1768e74
[Ljava.lang.String;@140e316
[Ljava.lang.String;@15a4aed
[Ljava.lang.String;@991465
[Ljava.lang.String;@1ef2aac
[Ljava.lang.String;@1973a14
[Ljava.lang.String;@130a29
[Ljava.lang.String;@4d02b2
[Ljava.lang.String;@51f01d
```

All files are up-to-date (6 minutes ago) 17:1 CRLF÷ 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;@51f01d
[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;@4e8fdf
[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 up-to-date (9 minutes ago) 101:24 CRLF: UTF-8:

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

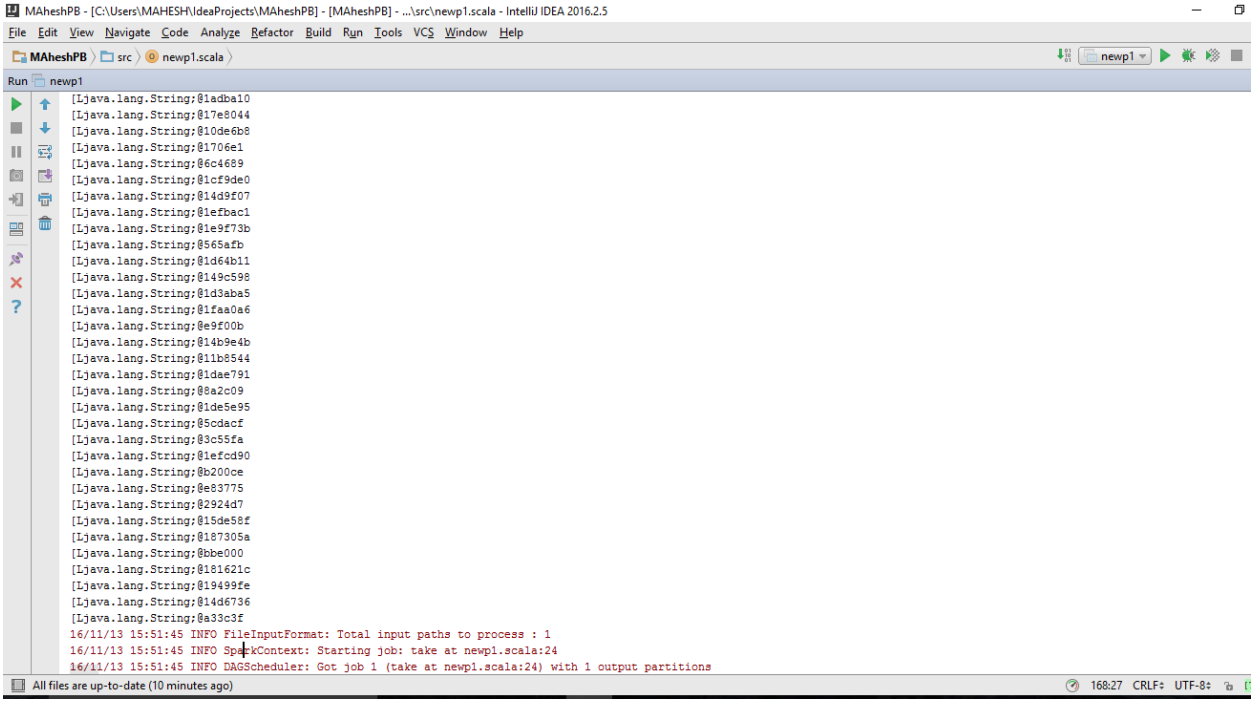
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help

MAheshPB 〉 src 〉 newp1.scala 〉 newp1

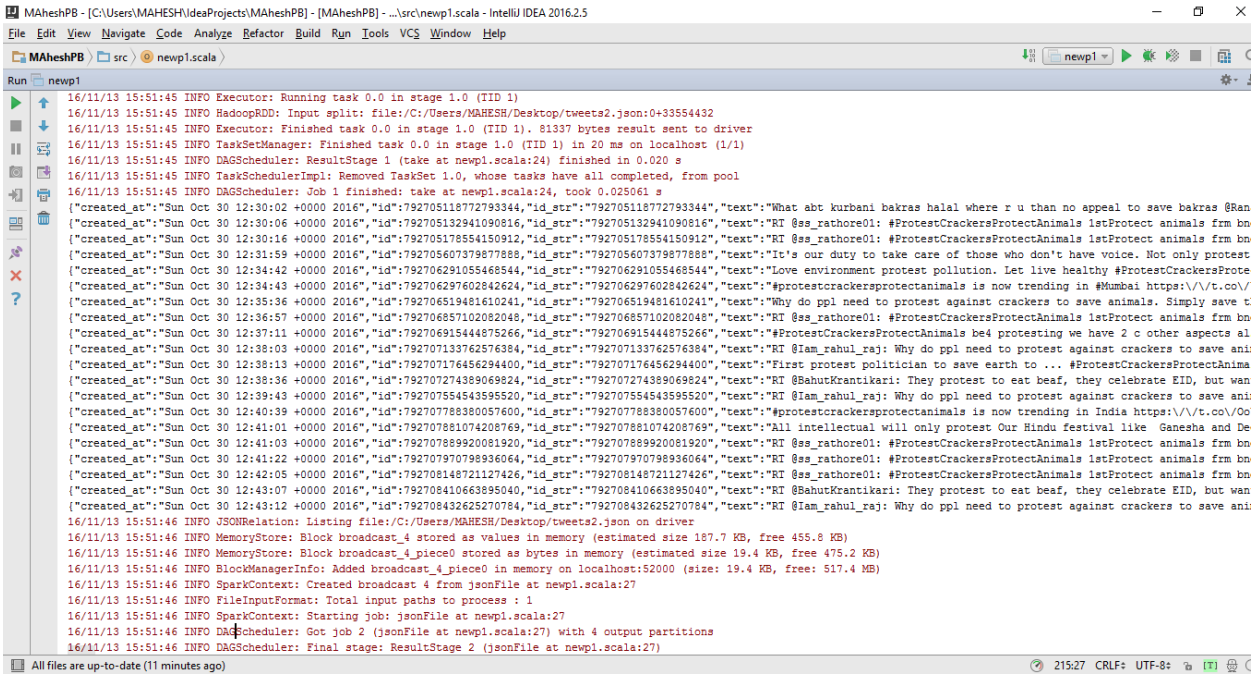
Run newp1

```
[Ljava.lang.String;@4d02b2
[Ljava.lang.String;@51f01d
[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;@4e8fdf
[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 up-to-date (9 minutes ago) 101:24 CRLF: UTF-8:



Query2 Output:



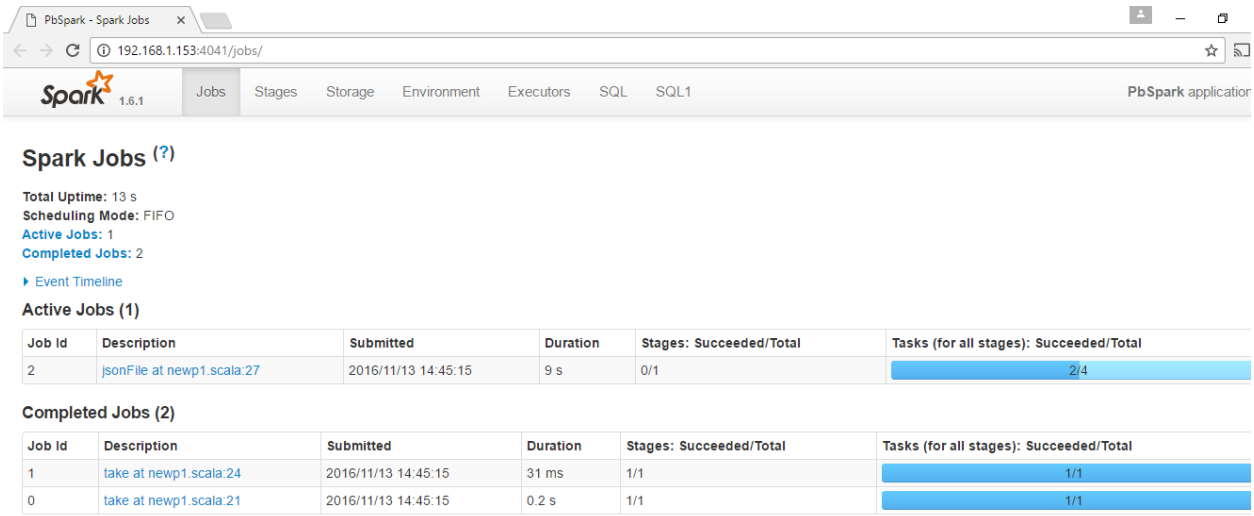
Query3 Output:

```
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 MB)
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|
|Kijiye 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:



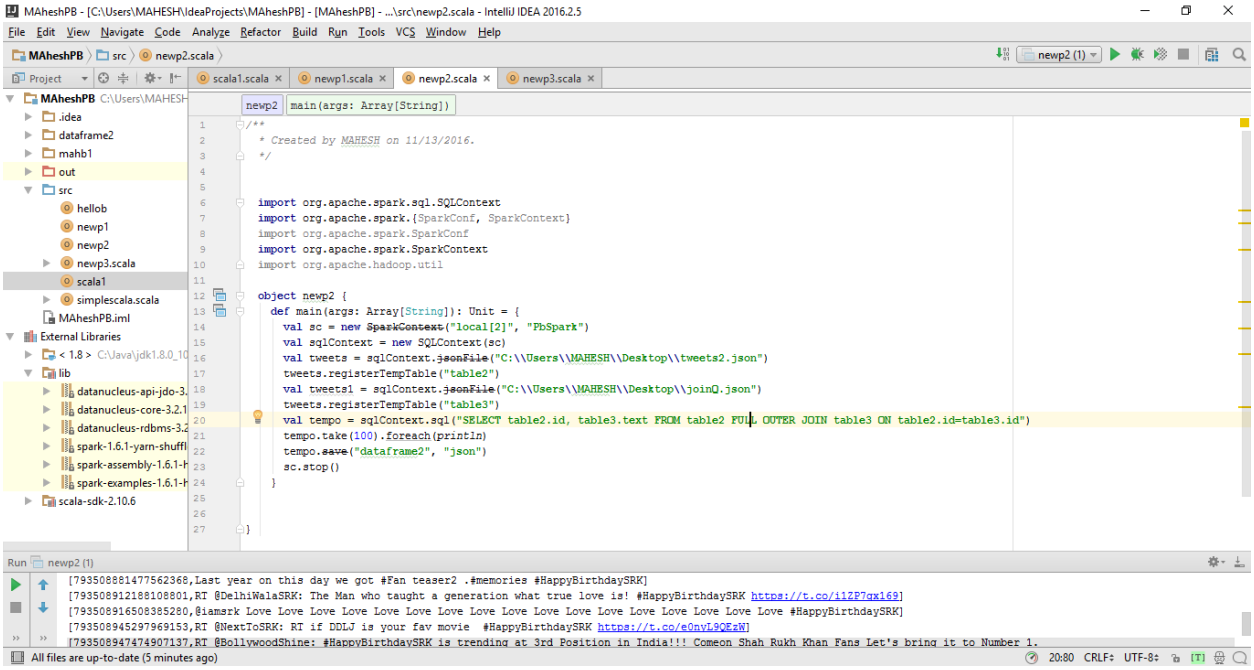
Join query

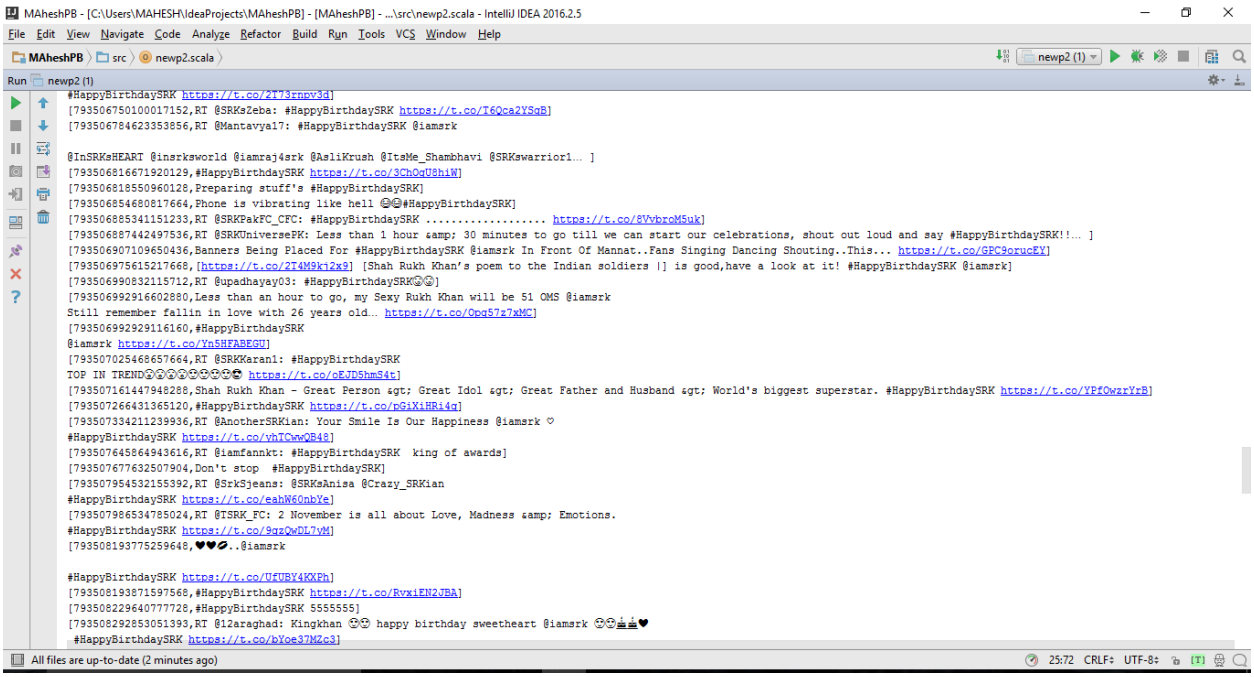
Joining the files Tweets2 and joinQ based on the column “id”. (Nothing but joining using “outer join” the two files data using the column “id”).

```
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")
      tweets.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:





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)

```
[793508292853051393,RT @12araghad: Kingkhan 😍😍 happy birthday sweetheart @iamsrk 😍😍🍰🍰❤
 #HappyBirthdaySRK https://t.co/bYoe37MZc3]
[793508296887984128,Wishing our dearest superhero @iamsrk the best birthday ever! We hope all your days are filled with love &amp; laughter... https://t.co/tCAZ5t9YHu]
[793508348377214980,RT @sunn_le: 1 year ho gaya shahrukh khan ko block kiye huye..
#HAppybirthdaySRK 😂]
[793508402035122176,RT @Shreyanjeet_SN: Today is Paro's birthday..
Tomorrow is Devdas's birthday..
#HappyBirthdayAishwarya
#HappyBirthdaySRK]
[793508491398897667,RT @I_Love_U_SRK: #HappyBirthdaySRK
The best husband
The best father
The best son
The best brother
The best motivator
The best Idol
T... ]
[793508548701454341,#HappyBirthdaySRK Thankful to have the honors of living during the same lifetime as Shahrukh Sir. What a connection. From Sweden to india!]
[793508570872487936,#HappyBirthdaySRK https://t.co/rtPr0a2ExL]
[793508572382502912,RT @AamirSahirKhan: Dil se happy birthday sir....#HappyBirthdaySRK https://t.co/05LlMWttHx]
[793508602635878401,Jio 4G ki Jai Ho

#HappyBirthdaySRK]
[793508638706978816,RT @BollywoodShine: #HappyBirthdaySRK is trending at 3rd Position in India!!! Comeon Shah Rukh Khan Fans Let's bring it to Number 1.
L... ]
[793508674073395200,RT @ahmedlovessrk: RT if u see this #HappyBirthdaySRK]
[793508744609067008,RT @SRKsChampion: Aaj jaan laga ke tweeting karunga BC #HappyBirthdaySRK]
[793508881477562368,Last year on this day we got #Fan teaser2 .#memories #HappyBirthdaySRK]
[793508912188108801,RT @DelhiWalaSRK: The Man who taught a generation what true love is! #HappyBirthdaySRK https://t.co/i1ZP7gx169]
[793508916508385280,@iamsrk Love Love Love Love Love Love Love Love Love Love Love Love Love Love Love Love #HappyBirthdaySRK]
[793508945297969153,RT @NextToSRK: RT if DDLJ is your fav movie  #HappyBirthdaySRK https://t.co/e0nyL9QEzW]
[793508947474907137,RT @BollywoodShine: #HappyBirthdaySRK is trending at 3rd Position in India!!! Comeon Shah Rukh Khan Fans Let's bring it to Number 1.
L... ]
[793509019608424448,Who brought Bollywood on Global platform
Representing INDIA
@iamsrk
```

All files are up-to-date (3 minutes ago) 25:72 CRLF÷ UTF-8÷

MAheshPB - [C:\Users\MAHESH\IdeaProjects\MAheshPB] - [MAheshPB] - ...\src\newp2.scala - IntelliJ IDEA 2016.2.5

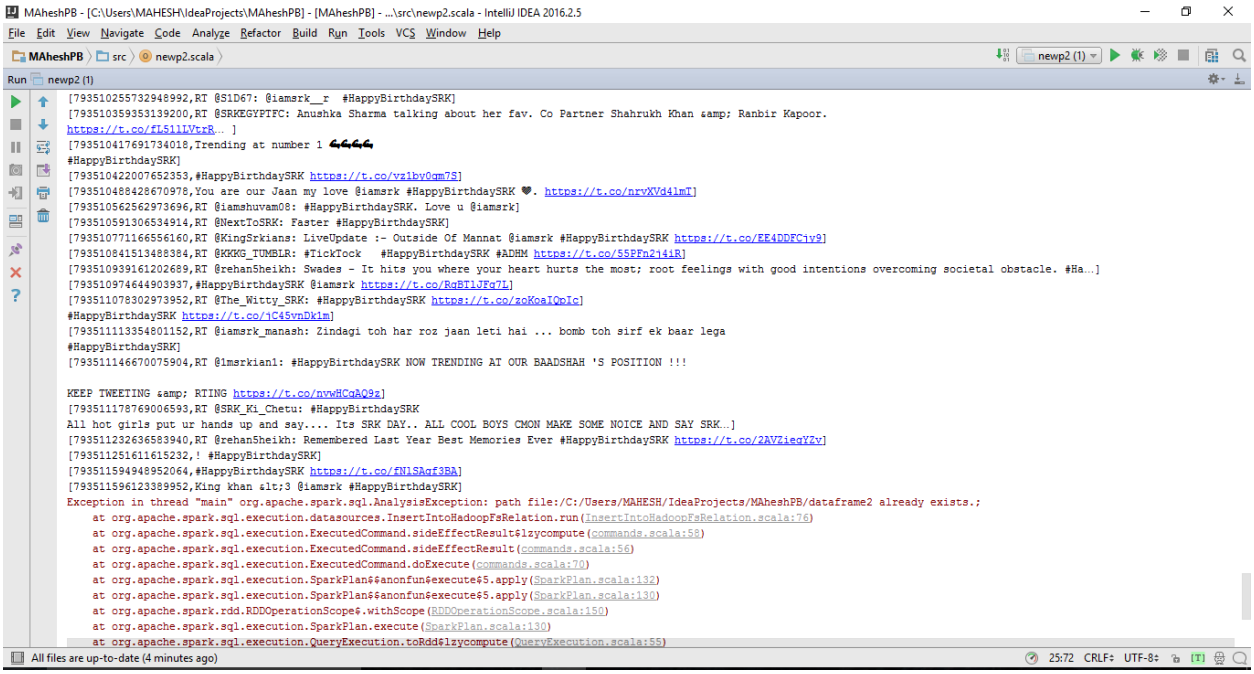
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help

MAheshPB > src > newp2.scala

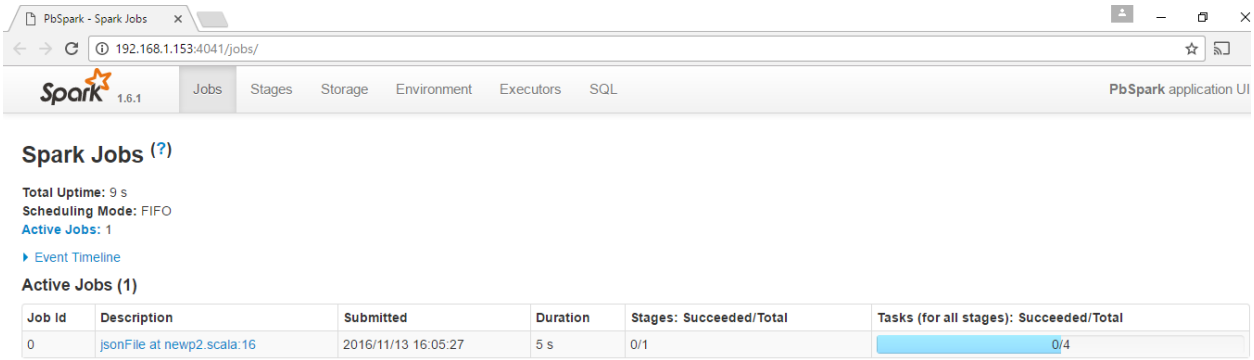
Run newp2 (1)

```
@iamsrk
#HappyBirthdaySRK ❤ https://t.co/gjxmL8Vlpo]
[793509051711811584,#HappyBirthdaySRK Ashoka ,Dil se under rated film of SRK ,realy luv these two,anothr underrated one hilariously funny duplicate luv it too]
[793509191734427648,RT @mad_for_SRK: #HappyBirthdaySRK https://t.co/R0LWH0nZ5J]
[793509260155944960,RT @MrsSamarAnand: Uffffff 💧 #HappyBirthdaySRK https://t.co/iU6M2hB2Lj]
#HappyBirthdaySRK]
[793509420764426242,RT @SRKian__: #HappyBirthdaySRK https://t.co/8JKRLjLSs1]
[793509534606106624,Veer Zaara
#HappyBirthdaySRK]
[793509670631636992,RT @nyk_pradeep22: #HappyBirthdaySRK  आतंकवादी मरते हे इस में भी ये कथाकथित सेक्युलर धर्म को बीच में ला रहे हे, हमारे हिन्दू धर्म को ऐसी गं...]
[793509703619870720,#HappyBirthdaySRK https://t.co/jdmiblVKtW]
[793509704613986304,RT @pramodsrkian: Hearts beating, clocks ticking, and it's down to only 3 hours 50 mint for the biggest day to come 😍... ]
[793509705318559744,RT @SRKCHENNAIFC: Pics - Grand celebrations started by Team SRKCHENNAIFC at Mannat.

#HappyBirthdaySRK - 2 https://t.co/evrqbKtjwp]
[793509771877949441,RT @SRKFanship: #HappyBirthdaySRK is now trending on KING KHAN's position. https://t.co/OcyoK9FvpX]
[793509773454954496,#HappyBirthdaySRK
Life😍]
[793509803255599105,Most loved superstar ever #HappyBirthdaySRK https://t.co/IQo4LruGTV]
 In 2009 @iamsrk Was Declared Amongst 500 Most Influential Muslims WW https://t.co/7dlI4zqqhm"]
[793509875913555968,Kal se classes open hai #HappyBirthdaySRK]
[793509910331871232,RT @SRKsChampion: I DONT WANT LIMIT SO RT MY TWEETS FROM @SRKCHAMP1 @srkchamp2 @srkchamp3 @srkchamp4 @srkchamp5 #HappyBirthdaySRK]
[793509911309058048,RT @SRKsCombat1: #HappyBirthdaySRK
World's Biggest movie Star https://t.co/FnhpaVYKBn]
[793509982025023488,RT @iamsrk_vraj: @iamsrk BAADSHAH &lt;3 #HappyBirthdaySRK https://t.co/bGper1EMTN]
[793510082818347008,@iamsrk #HappyBirthdaySRK https://t.co/f4Dj0VD8f7]
[793510085632942080,@iamsrk LoveLoveLoveLove Love Love Love Love Love Love Love Love Love Love Love Love Love Love Love Love Love Love Love Love #HappyBirthdaySRK]
[793510187588001792,RT @NextToSRK: Will surely do 500+ tweets tonight #HappyBirthdaySRK]
[793510222459351040,@IshqbaazIshaan dont copy me #HappyBirthdaySRK]
[793510255732948992,RT @S1D67: @iamsrk__r  #HappyBirthdaySRK]
[793510359353139200,RT @SRKEGYPTFC: Anushka Sharma talking about her fav. Co Partner Shahrukh Khan &amp; Ranbir Kapoor.
https://t.co/fL51lLVtrR... ]
[793510417691734018,Trending at number 1 🔥🔥🔥🔥
#HappyBirthdaySRK]
[793510422007652353,#HappyBirthdaySRK https://t.co/vzlbvOgm7S]
[793510488428670978,You are our Jaan my love @iamsrk #HappyBirthdaySRK ❤. https://t.co/nrvXVd4lmI]
```

All files are up-to-date (4 minutes ago) 25:72 CRLF÷ UTF-8÷



Runtime measurement for join query:



Public API:

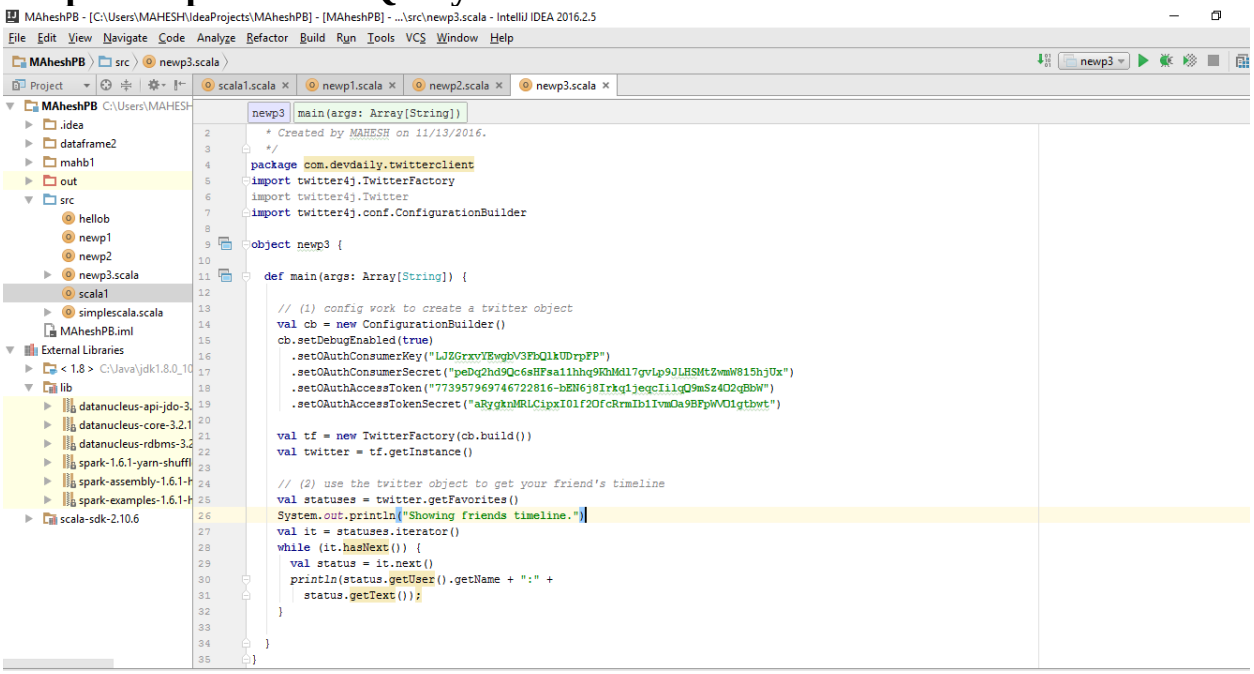
We have used Twitter Rest API

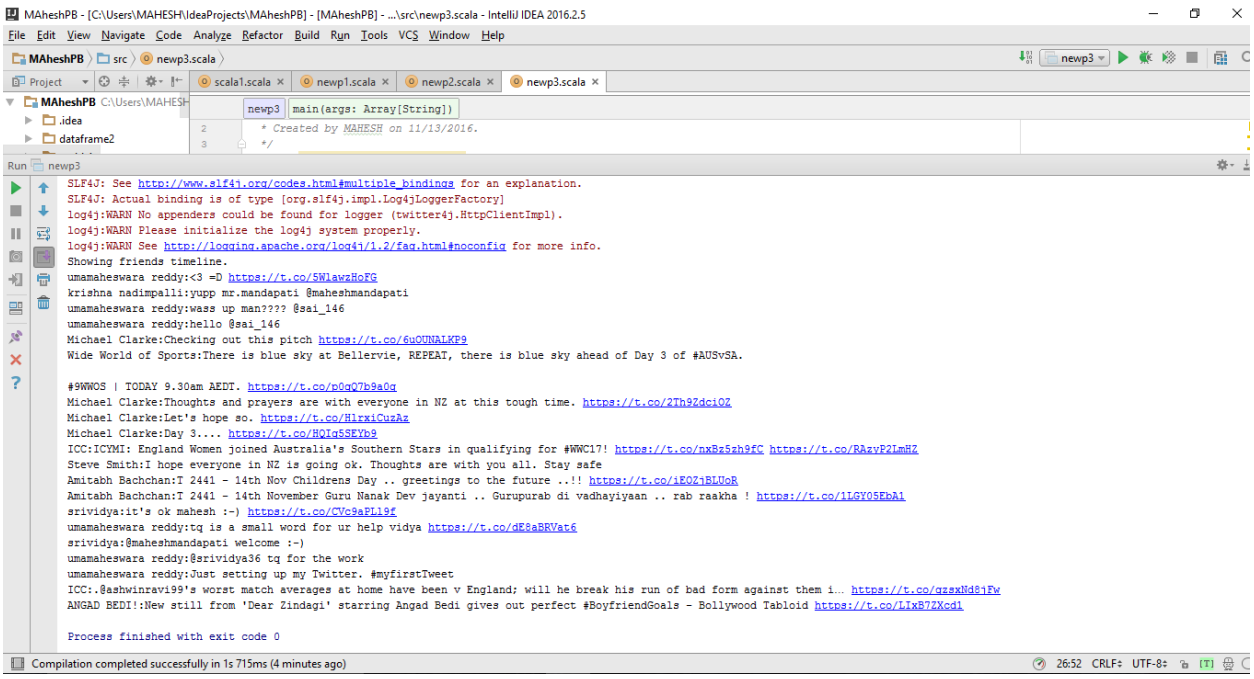
The REST APIs provide programmatic access to read and write Twitter data. Create a new Tweet, read user profile and follower data, and more. The REST API identifies Twitter applications and users using OAuth; responses are in JSON format.

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("LJZGrxvYEwgbV3FbQlkUDrpFP")
      .setOAuthConsumerSecret("peDq2hd9Qc6sHFsa11hhq9KhMdl7gvLp9JLHSMtZwmW815hjUx")
      .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:





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