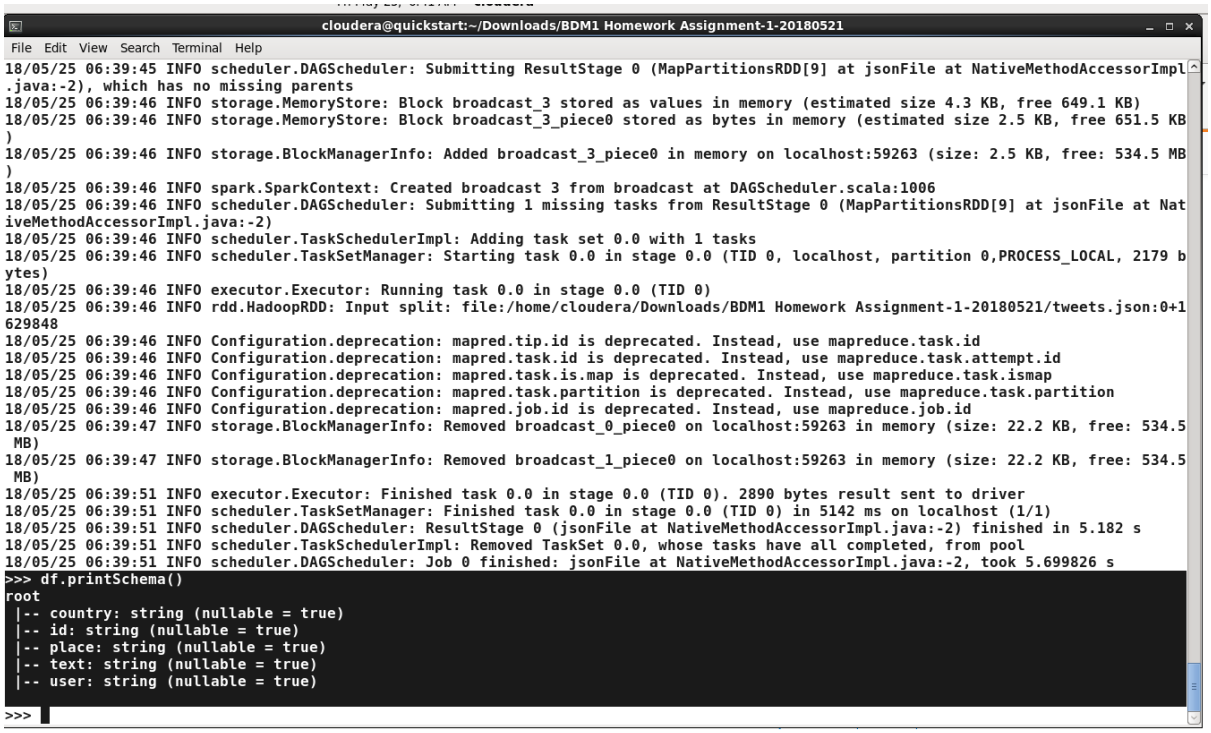
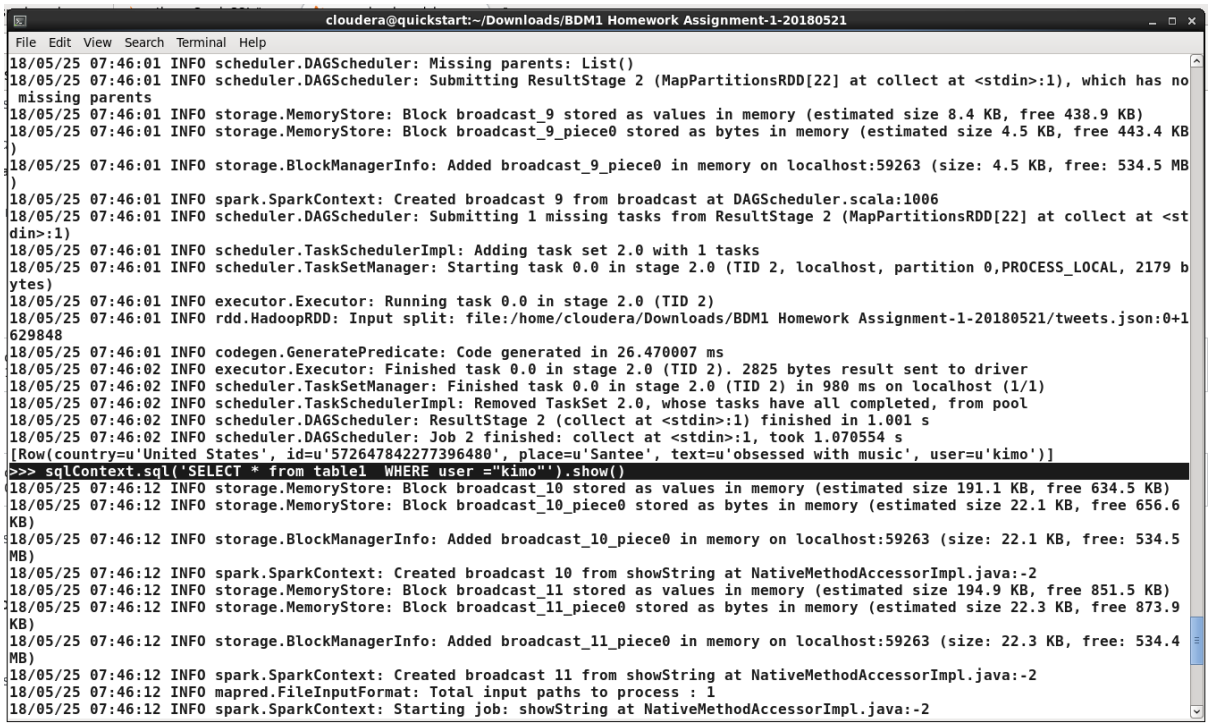
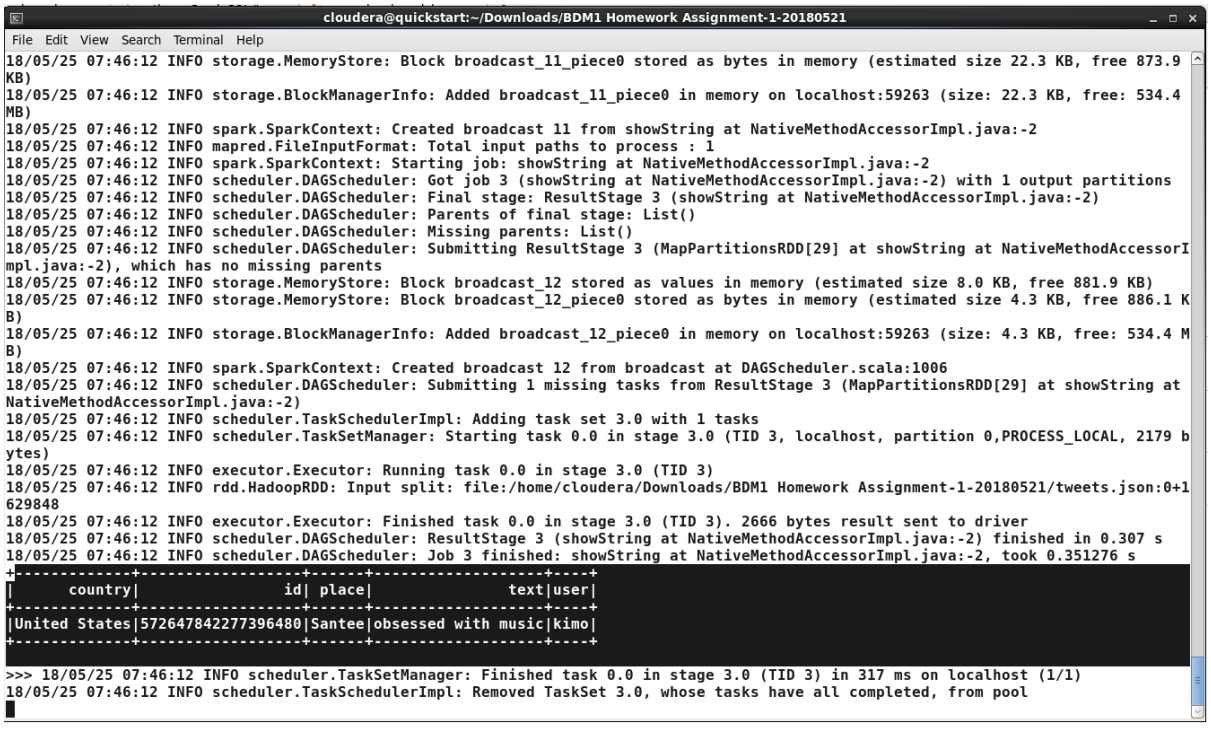
Answer 2

Save the dataset as a DataFrame, and print the schema.

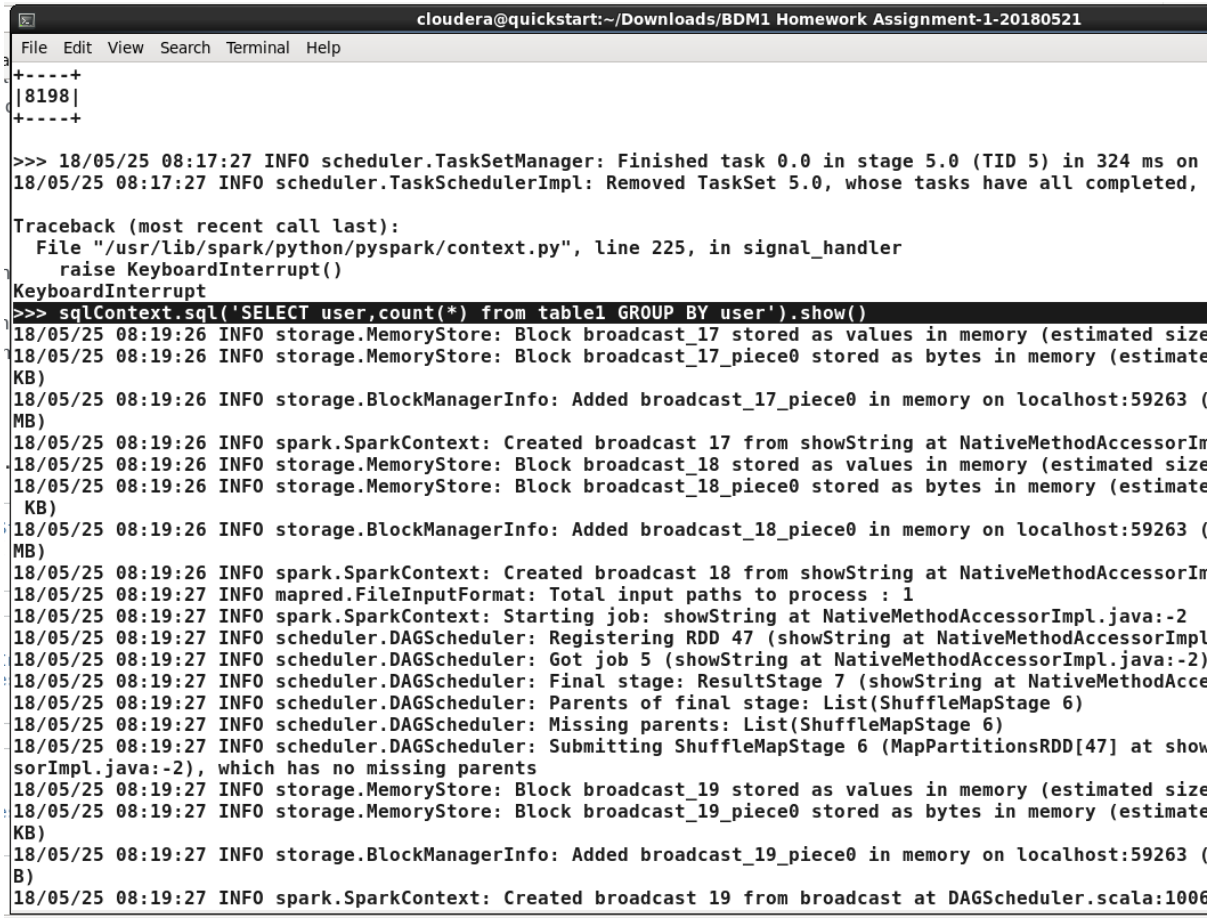


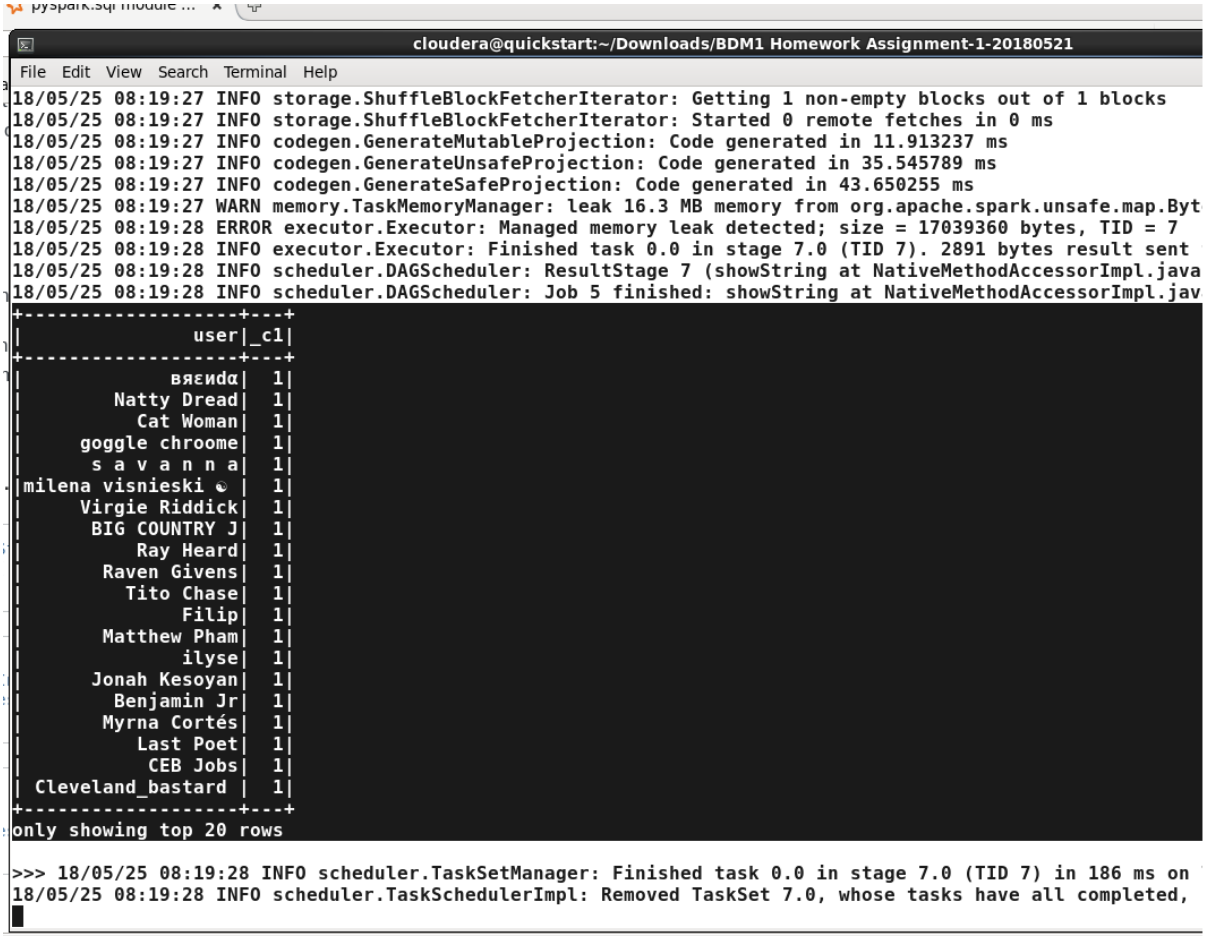
b)Get all of the tweets made by a user (any user would work. We should be able to replace user names to get tweets by that particular user).


cloudera@quickstart:~/Downloads/BDM1 Homework Assignment-1-20180521
File Edit View Search Terminal Help
18/05/25 07:46:01 INFO scheduler.DAGScheduler: Missing parents: List()
18/05/25 07:46:01 INFO scheduler.DAGScheduler: Submitting ResultStage 2 (MapPartitionsRDD[22] at collect at <stdin>:1), which has no missing parents
18/05/25 07:46:01 INFO storage.MemoryStore: Block broadcast_9 stored as values in memory (estimated size 8.4 KB, free 438.9 KB)
18/05/25 07:46:01 INFO storage.MemoryStore: Block broadcast_9_piece0 stored as bytes in memory (estimated size 4.5 KB, free 443.4 KB)
18/05/25 07:46:01 INFO storage.BlockManagerInfo: Added broadcast_9_piece0 in memory on localhost:59263 (size: 4.5 KB, free: 534.5 MB)
18/05/25 07:46:01 INFO spark.SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:1006
18/05/25 07:46:01 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 2 (MapPartitionsRDD[22] at collect at <stdin>:1)
18/05/25 07:46:01 INFO scheduler.TaskSchedulerImpl: Adding task set 2.0 with 1 tasks
18/05/25 07:46:01 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 2.0 (TID 2, localhost, partition 0,PROCESS_LOCAL, 2179 bytes)
18/05/25 07:46:01 INFO executor.Executor: Running task 0.0 in stage 2.0 (TID 2)
18/05/25 07:46:01 INFO rdd.HadoopRDD: Input split: file:/home/cloudera/Downloads/BDM1 Homework Assignment-1-20180521/tweets.json:0+1629848
18/05/25 07:46:01 INFO codegen.GeneratePredicate: Code generated in 26.470007 ms
18/05/25 07:46:02 INFO executor.Executor: Finished task 0.0 in stage 2.0 (TID 2). 2825 bytes result sent to driver
18/05/25 07:46:02 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 980 ms on localhost (1/1)
18/05/25 07:46:02 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
18/05/25 07:46:02 INFO scheduler.DAGScheduler: ResultStage 2 (collect at <stdin>:1) finished in 1.001 s
18/05/25 07:46:02 INFO scheduler.DAGScheduler: Job 2 finished: collect at <stdin>:1, took 1.070554 s
[Row(country=u'United States', id=u'572647842277396480', place=u'Santee', text=u'obsessed with music', user=u'kimo')]
>>> sqlContext.sql('SELECT * from table1 WHERE user ="kimo"').show()
18/05/25 07:46:12 INFO storage.MemoryStore: Block broadcast_10 stored as values in memory (estimated size 191.1 KB, free 634.5 KB)
18/05/25 07:46:12 INFO storage.MemoryStore: Block broadcast_10_piece0 stored as bytes in memory (estimated size 22.1 KB, free 656.6 KB)
18/05/25 07:46:12 INFO storage.BlockManagerInfo: Added broadcast_10_piece0 in memory on localhost:59263 (size: 22.1 KB, free: 534.5 MB)
18/05/25 07:46:12 INFO spark.SparkContext: Created broadcast 10 from showString at NativeMethodAccessorImpl.java:-2
18/05/25 07:46:12 INFO storage.MemoryStore: Block broadcast_11 stored as values in memory (estimated size 194.9 KB, free 851.5 KB)
18/05/25 07:46:12 INFO storage.MemoryStore: Block broadcast_11_piece0 stored as bytes in memory (estimated size 22.3 KB, free 873.9 KB)
18/05/25 07:46:12 INFO storage.BlockManagerInfo: Added broadcast_11_piece0 in memory on localhost:59263 (size: 22.3 KB, free: 534.4 MB)
18/05/25 07:46:12 INFO spark.SparkContext: Created broadcast 11 from showString at NativeMethodAccessorImpl.java:-2
18/05/25 07:46:12 INFO mapred.FileInputFormat: Total input paths to process : 1
18/05/25 07:46:12 INFO spark.SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:-2

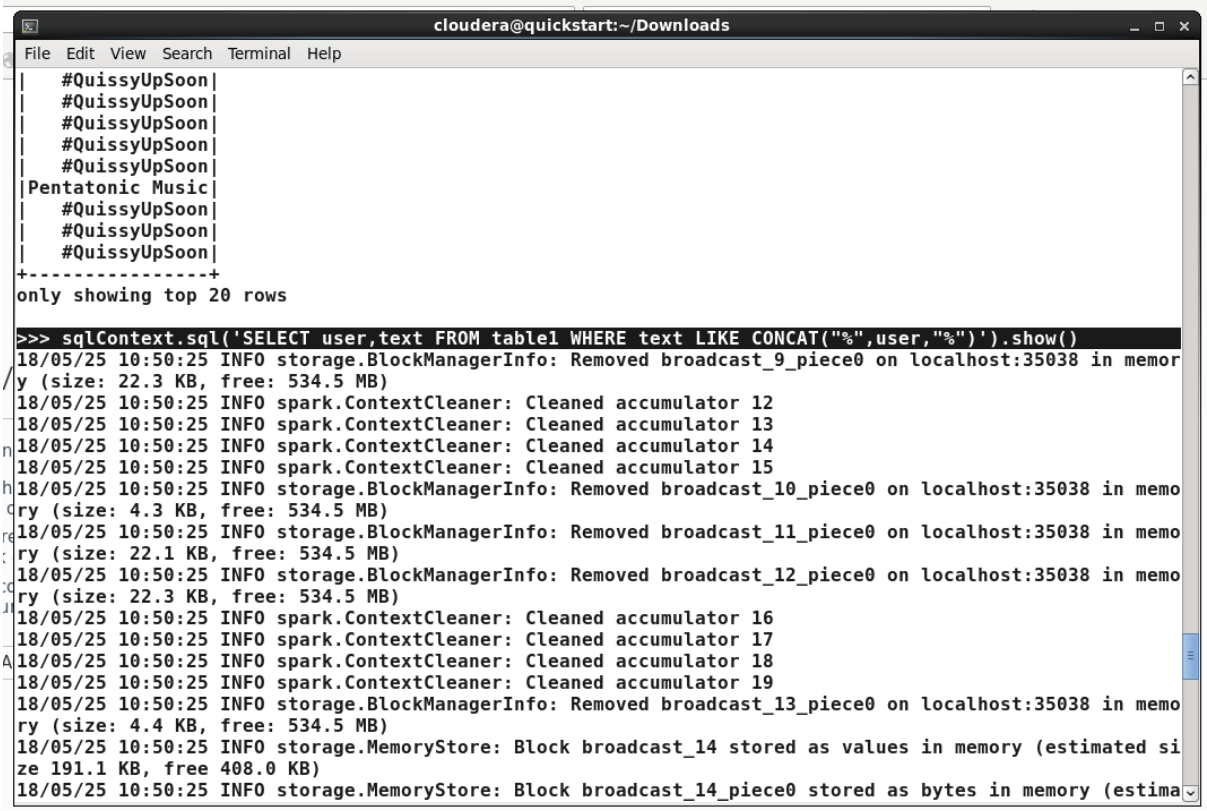

cloudera@quickstart:~/Downloads/BDM1 Homework Assignment-1-20180521
File Edit View Search Terminal Help
18/05/25 07:46:12 INFO storage.MemoryStore: Block broadcast_11_piece0 stored as bytes in memory (estimated size 22.3 KB, free 873.9 KB)
18/05/25 07:46:12 INFO storage.BlockManagerInfo: Added broadcast_11_piece0 in memory on localhost:59263 (size: 22.3 KB, free: 534.4 MB)
18/05/25 07:46:12 INFO spark.SparkContext: Created broadcast 11 from showString at NativeMethodAccessorImpl.java:-2
18/05/25 07:46:12 INFO mapred.FileInputFormat: Total input paths to process : 1
18/05/25 07:46:12 INFO spark.SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:-2
18/05/25 07:46:12 INFO scheduler.DAGScheduler: Got job 3 (showString at NativeMethodAccessorImpl.java:-2) with 1 output partitions
18/05/25 07:46:12 INFO scheduler.DAGScheduler: Final stage: ResultStage 3 (showString at NativeMethodAccessorImpl.java:-2)
18/05/25 07:46:12 INFO scheduler.DAGScheduler: Parents of final stage: List()
18/05/25 07:46:12 INFO scheduler.DAGScheduler: Missing parents: List()
18/05/25 07:46:12 INFO scheduler.DAGScheduler: Submitting ResultStage 3 (MapPartitionsRDD[29] at showString at NativeMethodAccessorImpl.java:-2), which has no missing parents
18/05/25 07:46:12 INFO storage.MemoryStore: Block broadcast_12 stored as values in memory (estimated size 8.0 KB, free 881.9 KB)
18/05/25 07:46:12 INFO storage.MemoryStore: Block broadcast_12_piece0 stored as bytes in memory (estimated size 4.3 KB, free 886.1 KB)
18/05/25 07:46:12 INFO storage.BlockManagerInfo: Added broadcast_12_piece0 in memory on localhost:59263 (size: 4.3 KB, free: 534.4 MB)
18/05/25 07:46:12 INFO spark.SparkContext: Created broadcast 12 from broadcast at DAGScheduler.scala:1006
18/05/25 07:46:12 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 3 (MapPartitionsRDD[29] at showString at NativeMethodAccessorImpl.java:-2)
18/05/25 07:46:12 INFO scheduler.TaskSchedulerImpl: Adding task set 3.0 with 1 tasks
18/05/25 07:46:12 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 3.0 (TID 3, localhost, partition 0,PROCESS_LOCAL, 2179 bytes)
18/05/25 07:46:12 INFO executor.Executor: Running task 0.0 in stage 3.0 (TID 3)
18/05/25 07:46:12 INFO rdd.HadoopRDD: Input split: file:/home/cloudera/Downloads/BDM1 Homework Assignment-1-20180521/tweets.json:0+1629848
18/05/25 07:46:12 INFO executor.Executor: Finished task 0.0 in stage 3.0 (TID 3). 2666 bytes result sent to driver
18/05/25 07:46:12 INFO scheduler.DAGScheduler: ResultStage 3 (showString at NativeMethodAccessorImpl.java:-2) finished in 0.307 s
18/05/25 07:46:12 INFO scheduler.DAGScheduler: Job 3 finished: showString at NativeMethodAccessorImpl.java:-2, took 0.351276 s
+-------------+------------------+------+------------------+----+
| country| id| place| text|user|
+-------------+------------------+------+------------------+----+
|United States|572647842277396480|Santee|obsessed with music|kimo|
+-------------+------------------+------+------------------+----+
>>> 18/05/25 07:46:12 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 317 ms on localhost (1/1)
18/05/25 07:46:12 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool

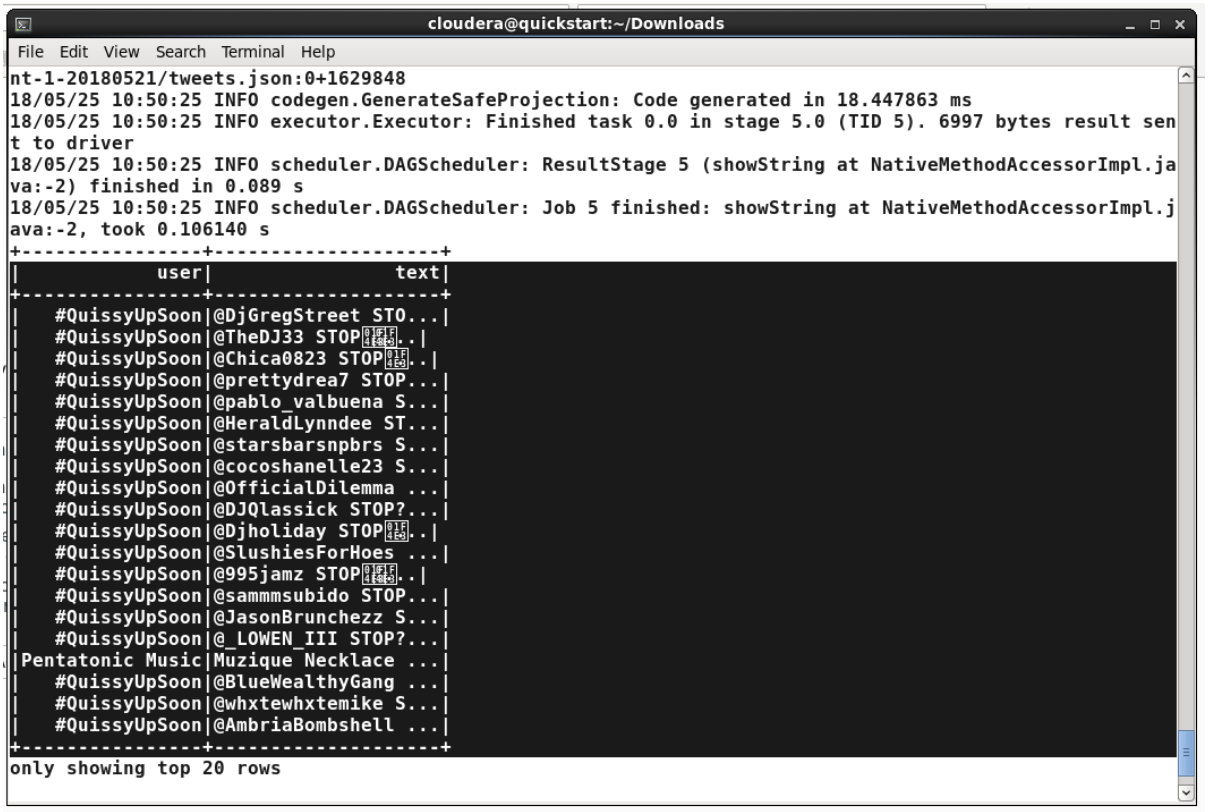
Find count of all tweets by each user user.



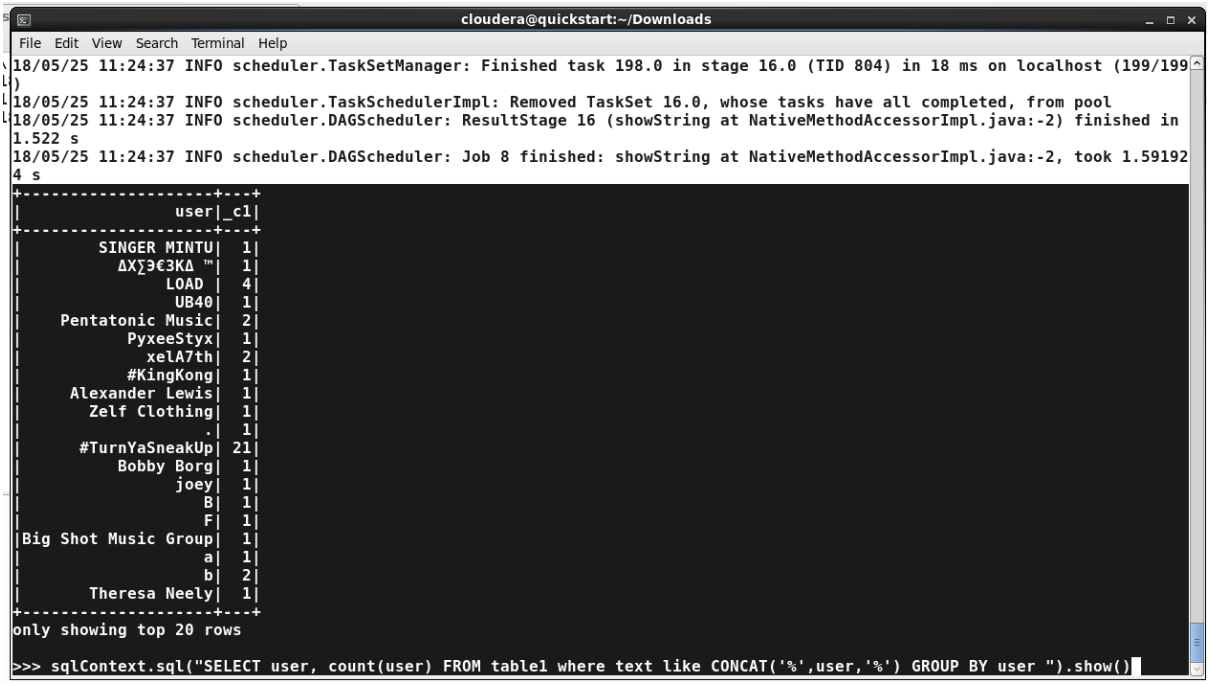


Get a list of all of the people who are mentioned in tweets.

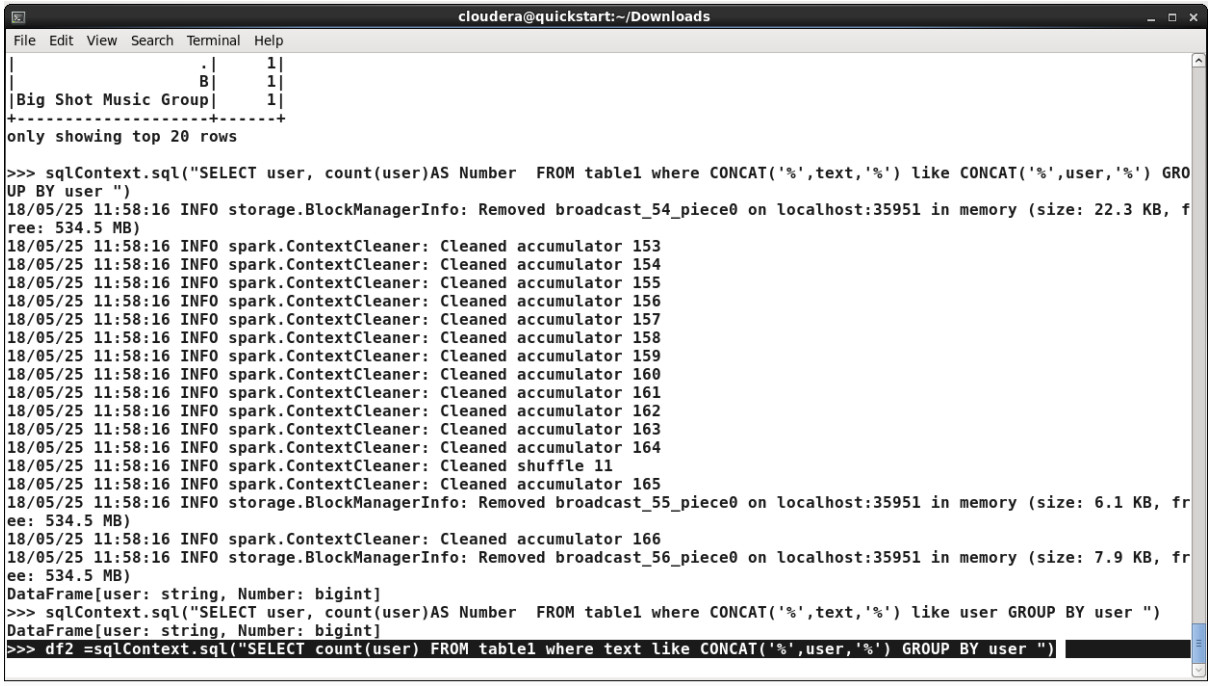


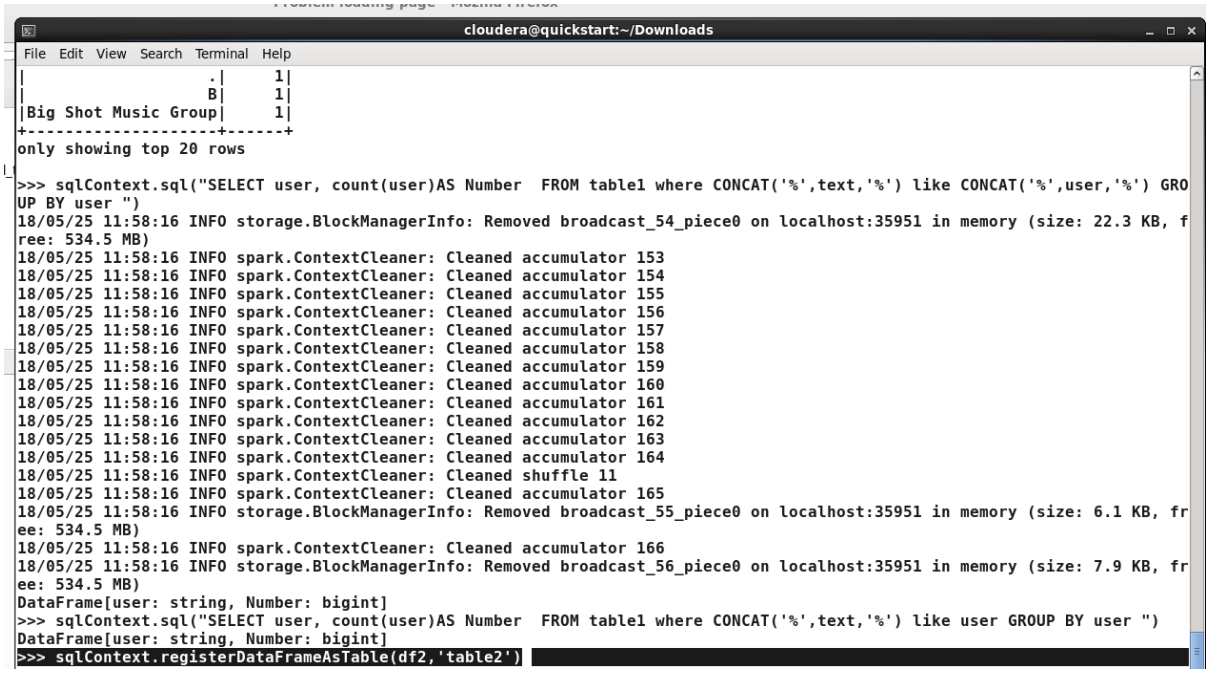


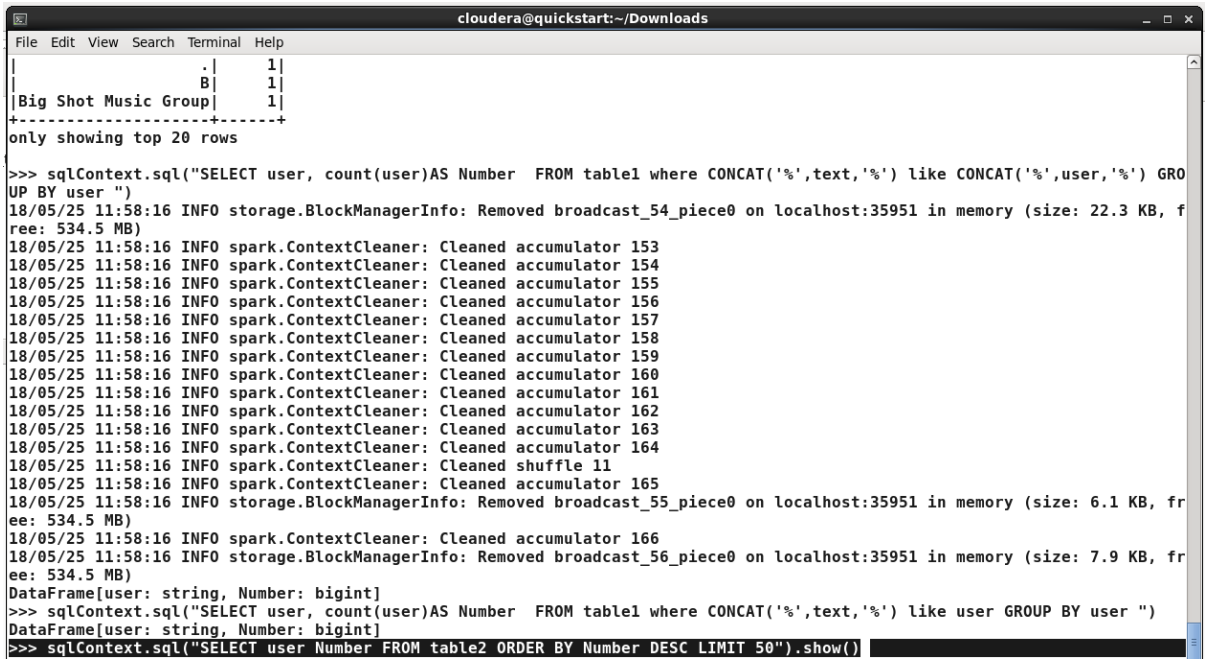
Count the number of time each person is mentioned in the entire dataset of tweets

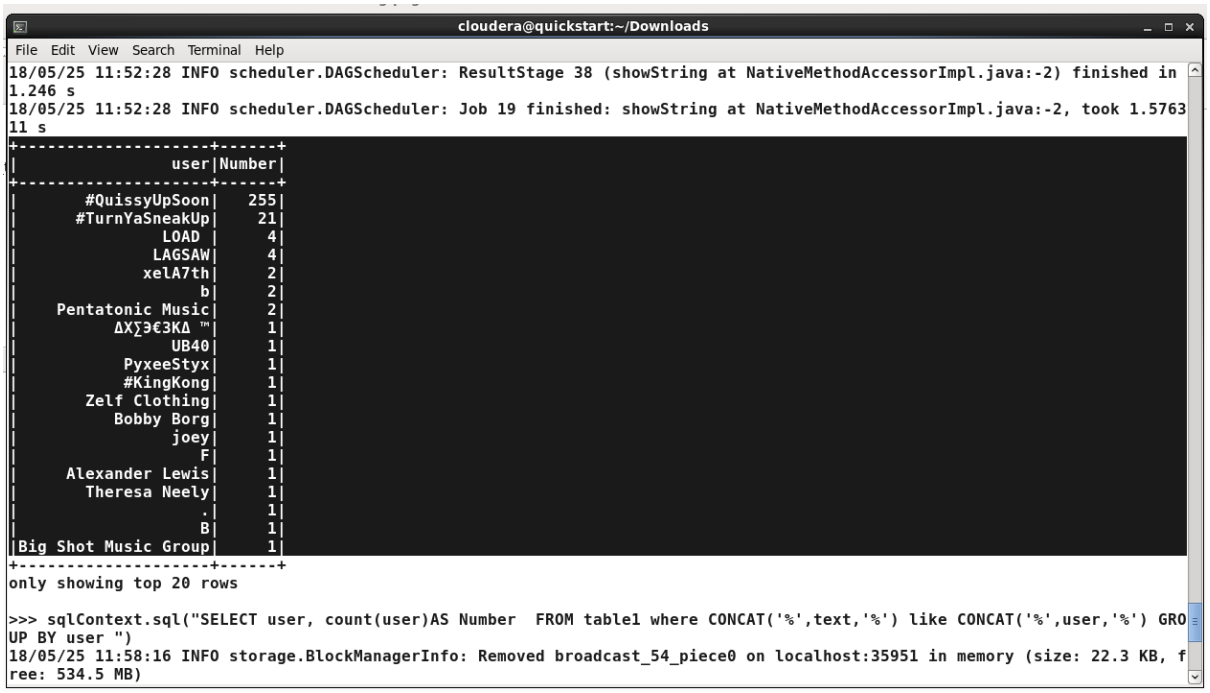


Give top 50 users who are mentioned the most.

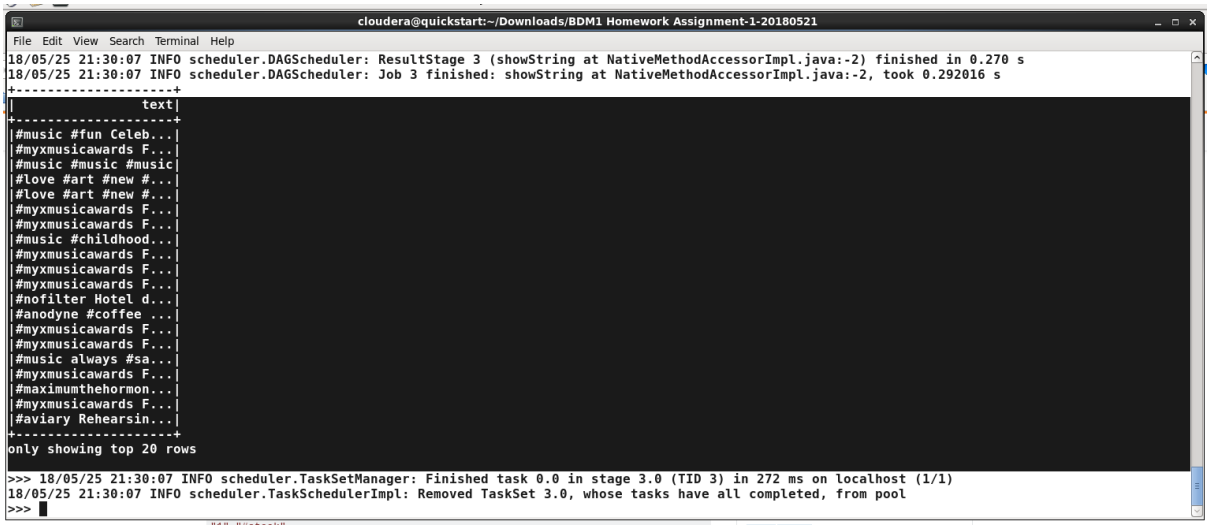




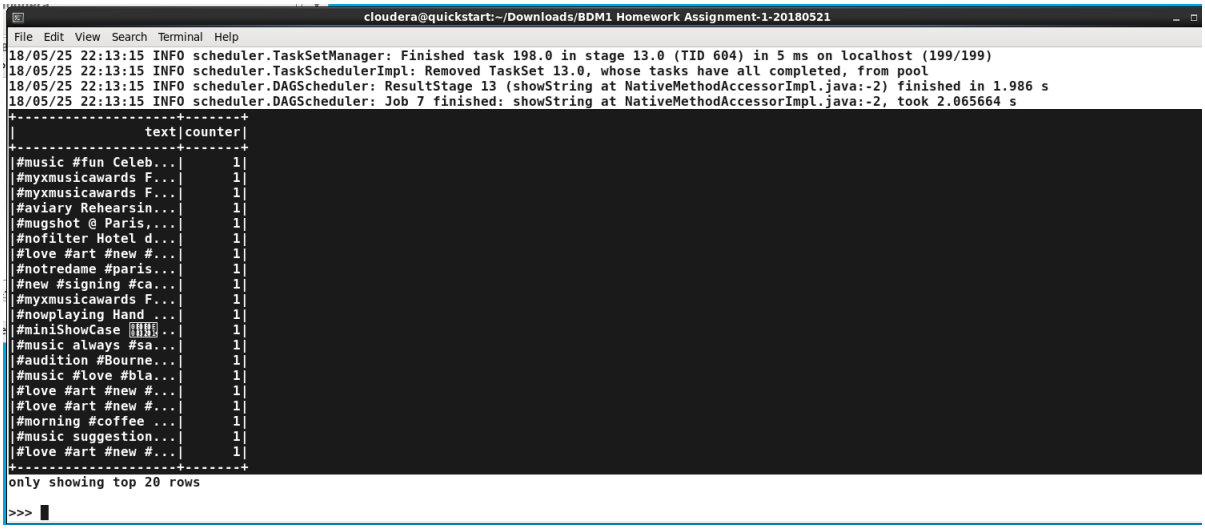

cloudera@quickstart:~/Downloads
File Edit View Search Terminal Help
| .| 1|
| B| 1|
|Big Shot Music Group| 1|
+--------------------+------+
only showing top 20 rows
>>> sqlContext.sql("SELECT user, count(user)AS Number FROM table1 where CONCAT('%',text,'%') like CONCAT('%',user,'%') GRO
UP BY user ")
18/05/25 11:58:16 INFO storage.BlockManagerInfo: Removed broadcast_54_piece0 on localhost:35951 in memory (size: 22.3 KB, f
ree: 534.5 MB)
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned accumulator 153
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned accumulator 154
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned accumulator 155
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned accumulator 156
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned accumulator 157
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned accumulator 158
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned accumulator 159
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned accumulator 160
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned accumulator 161
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned accumulator 162
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned accumulator 163
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned accumulator 164
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned shuffle 11
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned accumulator 165
18/05/25 11:58:16 INFO storage.BlockManagerInfo: Removed broadcast_55_piece0 on localhost:35951 in memory (size: 6.1 KB, fr
ee: 534.5 MB)
18/05/25 11:58:16 INFO spark.ContextCleaner: Cleaned accumulator 166
18/05/25 11:58:16 INFO storage.BlockManagerInfo: Removed broadcast_56_piece0 on localhost:35951 in memory (size: 7.9 KB, fr
ee: 534.5 MB)
DataFrame[user: string, Number: bigint]
>>> sqlContext.sql("SELECT user, count(user)AS Number FROM table1 where CONCAT('%',text,'%') like user GROUP BY user ")
DataFrame[user: string, Number: bigint]
>>> sqlContext.sql("SELECT user Number FROM table2 ORDER BY Number DESC LIMIT 50").show()


cloudera@quickstart:~/Downloads
File Edit View Search Terminal Help
18/05/25 11:52:28 INFO scheduler.DAGScheduler: ResultStage 38 (showString at NativeMethodAccessorImpl.java:-2) finished in
1.246 s
18/05/25 11:52:28 INFO scheduler.DAGScheduler: Job 19 finished: showString at NativeMethodAccessorImpl.java:-2, took 1.5763
11 s
+--------------------+------+
| user|Number|
+--------------------+------+
#QuissyUpSoon	255
#TurnYaSneakUp	21
LOAD	4
LAGSAW	4
xelA7th	2
b	2
Pentatonic Music	2
ΔXʃ3€3KΔ ™	1
UB40	1
PyxeeStyx	1
#KingKong	1
Zelf Clothing	1
Bobby Borg	1
joey	1
F	1
Alexander Lewis	1
Theresa Neely	1
.	1
B	1
Big Shot Music Group	1
+--------------------+------+
only showing top 20 rows
>>> sqlContext.sql("SELECT user, count(user)AS Number FROM table1 where CONCAT('%',text,'%') like CONCAT('%',user,'%') GRO
UP BY user ")
18/05/25 11:58:16 INFO storage.BlockManagerInfo: Removed broadcast_54_piece0 on localhost:35951 in memory (size: 22.3 KB, f
ree: 534.5 MB)

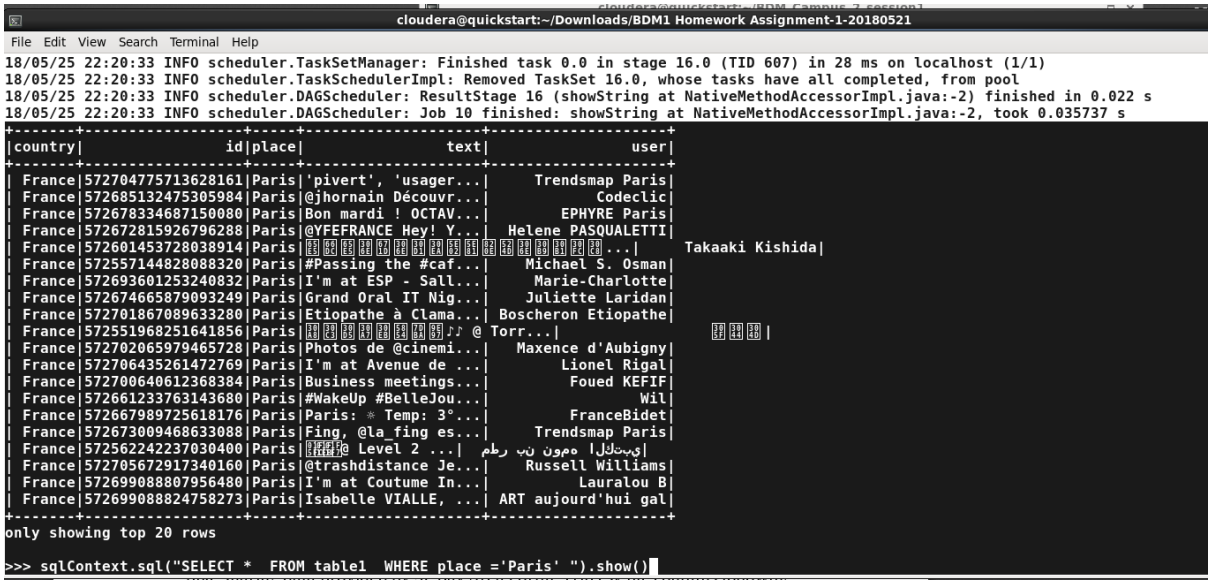
Get a list of all hashtags mentioned in the dataset.



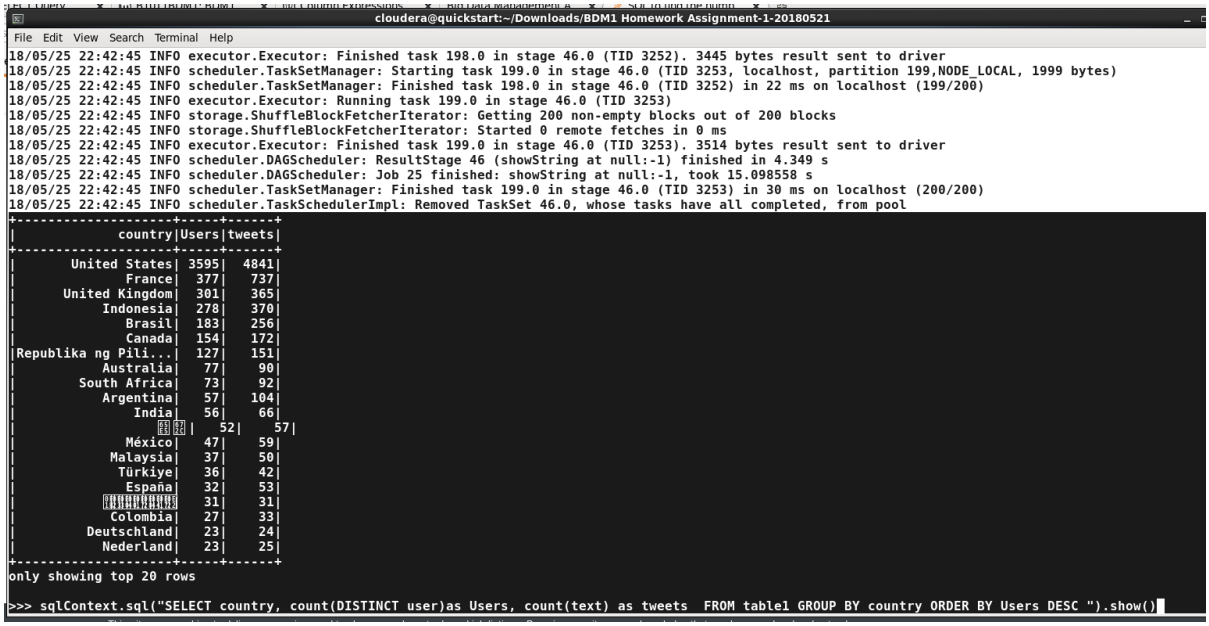
Find how many times each hashtag is mentioned in the dataset



Get a list of all of the people who are located in a particular city (e.g. Paris)



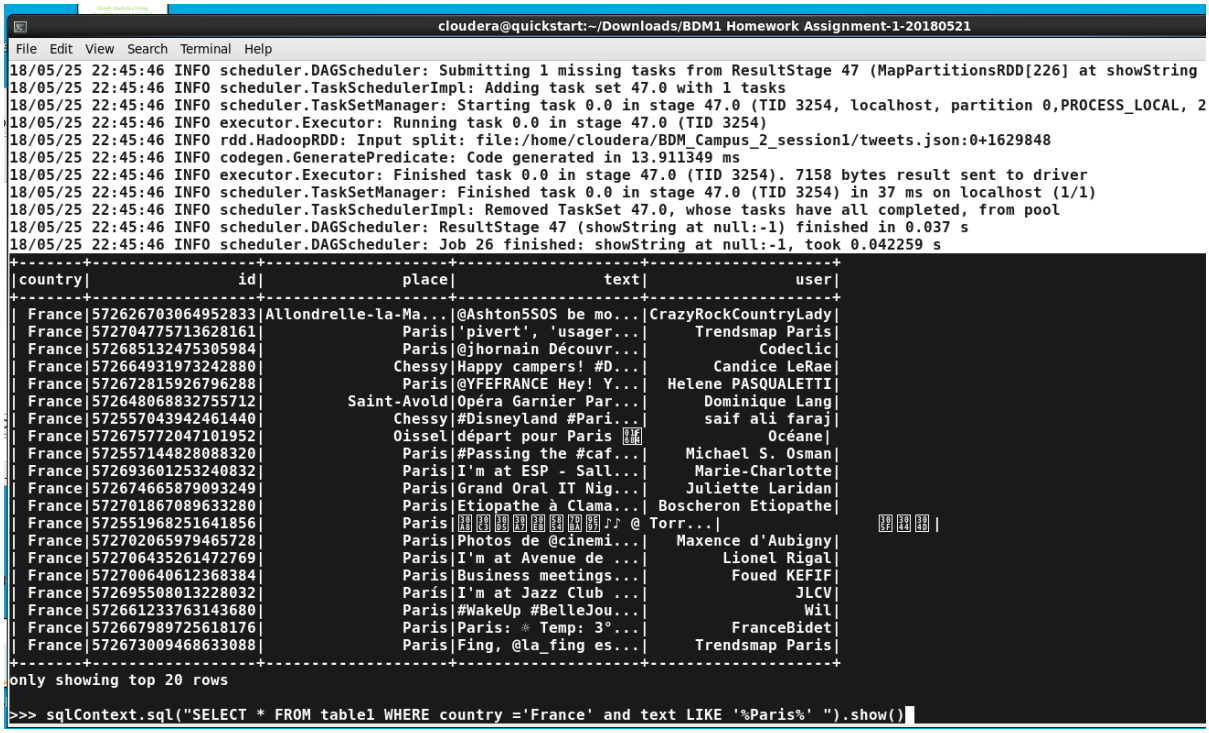
Get country wise distribution of users, and find out which country ranks highest in terms of number of tweets, and number of users.



```
cloudera@quickstart:~/Downloads/BDM1 Homework Assignment-1-20180521
File  Edit  View  Search  Terminal  Help
18/05/25 22:42:45 INFO executor.Executor: Finished task 198.0 in stage 46.0 (TID 3252). 3445 bytes result sent to driver
18/05/25 22:42:45 INFO scheduler.TaskSetManager: Starting task 199.0 in stage 46.0 (TID 3253, localhost, partition 199,NODE_LOCAL, 1999 bytes)
18/05/25 22:42:45 INFO scheduler.TaskSetManager: Finished task 198.0 in stage 46.0 (TID 3252) in 22 ms on localhost (199/200)
18/05/25 22:42:45 INFO executor.Executor: Running task 199.0 in stage 46.0 (TID 3253)
18/05/25 22:42:45 INFO storage.ShuffleBlockFetcherIterator: Getting 200 non-empty blocks out of 200 blocks
18/05/25 22:42:45 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/25 22:42:45 INFO executor.Executor: Finished task 199.0 in stage 46.0 (TID 3253). 3514 bytes result sent to driver
18/05/25 22:42:45 INFO scheduler.DAGScheduler: ResultStage 46 (showString at null:-1) finished in 4.349 s
18/05/25 22:42:45 INFO scheduler.DAGScheduler: Job 25 finished: showString at null:-1, took 15.098558 s
18/05/25 22:42:45 INFO scheduler.TaskSetManager: Finished task 199.0 in stage 46.0 (TID 3253) in 30 ms on localhost (200/200)
18/05/25 22:42:45 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 46.0, whose tasks have all completed, from pool
+--------------------+-----+------+
|             country|Users|tweets|
+--------------------+-----+------+
|       United States| 3595|  4841|
|              France|  377|   737|
|      United Kingdom|  301|   365|
|           Indonesia|  278|   370|
|              Brasil|  183|   256|
|              Canada|  154|   172|
|Republika ng Pili...|  127|   151|
|           Australia|   77|    90|
|        South Africa|   73|    92|
|           Argentina|   57|   104|
|               India|   56|    66|
|                 [illegible]|   52|    57|
|              México|   47|    59|
|            Malaysia|   37|    50|
|             Türkiye|   36|    42|
|              España|   32|    53|
|           [illegible]|   31|    31|
|            Colombia|   27|    33|
|         Deutschland|   23|    24|
|           Nederland|   23|    25|
+--------------------+-----+------+
only showing top 20 rows

>>> sqlContext.sql("SELECT country, count(DISTINCT user)as Users, count(text) as tweets  FROM table1 GROUP BY country ORDER BY Users DESC ").show()
```

Find out number of tweets where a user is from France and mentions Paris in their tweets.



```
cloudera@quickstart:~/Downloads/BDM1 Homework Assignment-1-20180521
File  Edit  View  Search  Terminal  Help
18/05/25 22:45:46 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 47 (MapPartitionsRDD[226] at showString
18/05/25 22:45:46 INFO scheduler.TaskSchedulerImpl: Adding task set 47.0 with 1 tasks
18/05/25 22:45:46 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 47.0 (TID 3254, localhost, partition 0,PROCESS_LOCAL, 2
18/05/25 22:45:46 INFO executor.Executor: Running task 0.0 in stage 47.0 (TID 3254)
18/05/25 22:45:46 INFO rdd.HadoopRDD: Input split: file:/home/cloudera/BDM_Campus_2_session1/tweets.json:0+1629848
18/05/25 22:45:46 INFO codegen.GeneratePredicate: Code generated in 13.911349 ms
18/05/25 22:45:46 INFO executor.Executor: Finished task 0.0 in stage 47.0 (TID 3254). 7158 bytes result sent to driver
18/05/25 22:45:46 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 47.0 (TID 3254) in 37 ms on localhost (1/1)
18/05/25 22:45:46 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 47.0, whose tasks have all completed, from pool
18/05/25 22:45:46 INFO scheduler.DAGScheduler: ResultStage 47 (showString at null:-1) finished in 0.037 s
18/05/25 22:45:46 INFO scheduler.DAGScheduler: Job 26 finished: showString at null:-1, took 0.042259 s
+-------+------------------+------------------+--------------------+--------------------+
|country|                id|             place|                text|                user|
+-------+------------------+------------------+--------------------+--------------------+
| France|572626703064952833|Allondrelle-la-Ma...|@Ashton5SOS be mo...|CrazyRockCountryLady|
| France|572704775713628161|             Paris|'pivert', 'usager...|     Trendsmap Paris|
| France|572685132475305984|             Paris|@jhornain Découvr...|            Codeclic|
| France|572664931973242880|            Chessy|Happy campers! #D...|       Candice LeRae|
| France|572672815926796288|             Paris|@YFEFRANCE Hey! Y...|  Helene PASQUALETTI|
| France|572648068832755712|       Saint-Avold|Opéra Garnier Par...|      Dominique Lang|
| France|572557043942461440|            Chessy|#Disneyland #Pari...|      saif ali faraj|
| France|572675772047101952|            Oissel|départ pour Paris [illegible]|              Océane|
| France|572557144828088320|             Paris|#Passing the #caf...|    Michael S. Osman|
| France|572693601253240832|             Paris|I'm at ESP - Sall...|     Marie-Charlotte|
| France|572674665879093249|             Paris|Grand Oral IT Nig...|    Juliette Laridan|
| France|572701867089633280|             Paris|Etiopathe à Clama...| Boscheron Etiopathe|
| France|572551968251641856|             Paris|[illegible]♪♪ @ Torr...|                [illegible]|
| France|572702065979465728|             Paris|Photos de @cinemi...|   Maxence d'Aubigny|
| France|572706435261472769|             Paris|I'm at Avenue de ...|        Lionel Rigal|
| France|572700640612368384|             Paris|Business meetings...|         Foued KEFIF|
| France|572695508013228032|             Paris|I'm at Jazz Club ...|                JLCV|
| France|572661233763143680|             Paris|#WakeUp #BelleJou...|                 Wil|
| France|572667989725618176|             Paris|Paris: ☼ Temp: 3°...|         FranceBidet|
| France|572673009468633088|             Paris|Fing, @la_fing es...|     Trendsmap Paris|
+-------+------------------+------------------+--------------------+--------------------+
only showing top 20 rows

>>> sqlContext.sql("SELECT * FROM table1 WHERE country ='France' and text LIKE '%Paris%' ").show()
```