

Answer 2

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

```
cloudera@quickstart:~/Downloads/BDM1 Homework Assignment-1-20180521
File Edit View Search Terminal Help

at org.apache.spark.sql.sources.HadoopFsRelation.schema$lzycompute(interfaces.scala:636)
at org.apache.spark.sql.sources.HadoopFsRelation.schema(interfaces.scala:635)
at org.apache.spark.sql.execution.datasources.LogicalRelation.<init>(LogicalRelation.scala:37)
at org.apache.spark.sql.DataFrameReader.load(DataFrameReader.scala:125)
at org.apache.spark.sql.DataFrameReader.load(DataFrameReader.scala:109)
at org.apache.spark.sql.DataFrameReader.json(DataFrameReader.scala:244)
at org.apache.spark.sql.SQLContext.jsonFile(SQLContext.scala:1032)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:606)
at py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:231)
at py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.java:381)
at py4j.Gateway.invoke(Gateway.java:259)
at py4j.commands.AbstractCommand.invokeMethod(AbstractCommand.java:133)
at py4j.commands.CallCommand.execute(CallCommand.java:79)
at py4j.GatewayConnection.run(Thread.java:209)
at java.lang.Thread.run(Thread.java:745)

>>> df.show()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'df' is not defined
>>> df = sqlContext.jsonFile('/home/cloudera/Downloads/BDM1 Homework Assignment-1-20180521/tweets.json')
Traceback (most recent call last):
  File "/usr/lib/spark/python/pyspark/context.py", line 225, in signal_handler
    raise KeyboardInterrupt()
KeyboardInterrupt

>>> sqlContext = SQLContext(sc)
>>> df = sqlContext.jsonFile('file:///home/cloudera/Downloads/BDM1 Homework Assignment-1-20180521/tweets.json')
18/05/25 06:39:42 INFO json.JSONRelation: Listing file:/home/cloudera/Downloads/BDM1 Homework Assignment-1-20180521/tweets.json on d
river
18/05/25 06:39:43 INFO storage.MemoryStore: Block broadcast_2 stored as values in memory (estimated size 194.9 KB, free 622.4 KB)
18/05/25 06:39:43 INFO storage.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 22.3 KB, free 644.7 K
B)
18/05/25 06:39:43 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on localhost:59263 (size: 22.3 KB, free: 534.5 M
B)
18/05/25 06:39:43 INFO spark.SparkContext: Created broadcast 2 from jsonFile at NativeMethodAccessorImpl.java:-2
18/05/25 06:39:44 INFO mapped.FileInputFormat: Total input paths to process : 1
```

```
cloudera@quickstart:~/Downloads/BDM1 Homework Assignment-1-20180521
File Edit View Search Terminal Help

18/05/25 06:39:45 INFO scheduler.DAGScheduler: Submitting ResultStage 0 (MapPartitionsRDD[9] at jsonFile at NativeMethodAccessorImpl
.java:-2), which has no missing parents
18/05/25 06:39:46 INFO storage.MemoryStore: Block broadcast_3 stored as values in memory (estimated size 4.3 KB, free 649.1 KB)
18/05/25 06:39:46 INFO storage.MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 2.5 KB, free 651.5 KB
)
18/05/25 06:39:46 INFO storage.BlockManagerInfo: Added broadcast_3_piece0 in memory on localhost:59263 (size: 2.5 KB, free: 534.5 MB
)
18/05/25 06:39:46 INFO spark.SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:1006
18/05/25 06:39:46 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 0 (MapPartitionsRDD[9] at jsonFile at Nat
iveMethodAccessorImpl.java:-2)
18/05/25 06:39:46 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
18/05/25 06:39:46 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, partition 0,PROCESS_LOCAL, 2179 b
ytes)
18/05/25 06:39:46 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
18/05/25 06:39:46 INFO rdd.HadoopRDD: Input split: file:/home/cloudera/Downloads/BDM1 Homework Assignment-1-20180521/tweets.json:0+1
629848
18/05/25 06:39:46 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
18/05/25 06:39:46 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
18/05/25 06:39:46 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
18/05/25 06:39:46 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
18/05/25 06:39:46 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
18/05/25 06:39:47 INFO storage.BlockManagerInfo: Removed broadcast_0_piece0 on localhost:59263 in memory (size: 22.2 KB, free: 534.5
MB)
18/05/25 06:39:47 INFO storage.BlockManagerInfo: Removed broadcast_1_piece0 on localhost:59263 in memory (size: 22.2 KB, free: 534.5
MB)
18/05/25 06:39:51 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 2890 bytes result sent to driver
18/05/25 06:39:51 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 5142 ms on localhost (1/1)
18/05/25 06:39:51 INFO scheduler.DAGScheduler: ResultStage 0 (jsonFile at NativeMethodAccessorImpl.java:-2) finished in 5.182 s
18/05/25 06:39:51 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
18/05/25 06:39:51 INFO scheduler.DAGScheduler: Job 0 finished: jsonFile at NativeMethodAccessorImpl.java:-2, took 5.699826 s
>>> df.printSchema()
root
|-- country: string (nullable = true)
|-- id: string (nullable = true)
|-- place: string (nullable = true)
|-- text: string (nullable = true)
|-- user: string (nullable = true)
>>>
```

```

cloudera@quickstart:~/Downloads/BDM1 Homework Assignment-1-20180521
File Edit View Search Terminal Help
18/05/25 06:43:37 INFO storage.BlockManagerInfo: Removed broadcast_3_piece0 on localhost:59263 in memory (size: 2.5 KB, free: 534.5 MB)
18/05/25 06:43:37 INFO spark.ContextCleaner: Cleaned accumulator 2
18/05/25 06:43:37 INFO storage.BlockManagerInfo: Removed broadcast_2_piece0 on localhost:59263 in memory (size: 22.3 KB, free: 534.5 MB)
18/05/25 06:43:37 INFO codegen.GenerateUnsafeProjection: Code generated in 709.36464 ms
18/05/25 06:43:37 INFO codegen.GenerateSafeProjection: Code generated in 69.312475 ms
18/05/25 06:43:37 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 1). 7110 bytes result sent to driver
18/05/25 06:43:37 INFO scheduler.DAGScheduler: ResultStage 1 (showString at NativeMethodAccessorImpl.java:-2) finished in 1.039 s
18/05/25 06:43:37 INFO scheduler.DAGScheduler: Job 1 finished: showString at NativeMethodAccessorImpl.java:-2, took 1.099288 s
18/05/25 06:43:37 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 1043 ms on localhost (1/1)
18/05/25 06:43:37 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool

+-----+-----+-----+-----+-----+
| country | id | place | text | user |
+-----+-----+-----+-----+-----+
| India | 572692378957430785 | Orissa | @always_nidhi @Yo... | Srkian nishu :) |
| United States | 572575240615796737 | Manhattan | @OnlyDancers Bell... | TagineDiningGlobal |
| United States | 572575243883036672 | Claremont | I/ "Without the a... | Daniel Beer |
| United States | 572575252020109313 | Vienna | idk why people ha... | someone actually |
| United States | 572575274539356160 | Boston | Taste of Iceland!... | BostonAttitude |
| United States | 572647819401670656 | Suwanee | Know what you don... | Collin A. Zimmerman |
| Indonesia | 572647831053312000 | Mario Riawa | Serasi ade haha @... | Rinie Syamsuddin |
| Indonesia | 572647839521767425 | Bogor Selatan | Akhirnya bisa jug... | Vinny Sylvia |
| United States | 572647841220337664 | Norwalk | @BeezyDH it's li... | Cas |
| United States | 572647842277396480 | Santee | obsessed with music | kimo |
| United States | 572631750163234816 | Tennessee | @blakeshelton You... | Jeff Morton |
| Indonesia | 572631763115249664 | Gambir | Happy Birthday Ps... | Rensus Paul |
| United States | 572606799712428033 | North Carolina | One night I'm ext... | KC |
| United States | 572606799649640449 | Baltimore | @DjGregStreet ST0... | #QuissyUpSoon |
| United States | 572606809216663552 | Cypress | always getting in... | lo |
| Negara Brunei Dar... | 572606812081410048 | Brunei | nigga in paris ht... | hafizzul |
| United States | 572616136963055616 | Portland | Boutta fall asLee... | Princess |
| United States | 572616139987144704 | Kentucky | Canadians are tak... | Nene Kiameso |
| United States | 572616165786185728 | Wahpeton | Chicago takin ove... | Chase |
| United Kingdom | 572667477949353984 | Ashton-under-Lyne | @traceyb65 I'm up... | stefan |
+-----+-----+-----+-----+-----+
only showing top 20 rows
>>>

```

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
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:606)
at py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:231)
at py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.java:381)
at py4j.Gateway.invoke(Gateway.java:259)
at py4j.commands.AbstractCommand.invokeMethod(AbstractCommand.java:133)
at py4j.commands.CallCommand.execute(CallCommand.java:79)
at py4j.GatewayConnection.run(GatewayConnection.java:209)
at java.lang.Thread.run(Thread.java:745)

>>> sqlContext.registerDataFrameAsTable(df, "table1")
>>> sqlContext.sql('SELECT * from table1 WHERE user = kim').collect()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/usr/lib/spark/python/pyspark/sql/context.py", line 580, in sql
    return DataFrame(self._ssql_ctx.sql(sqlQuery), self)
  File "/usr/lib/spark/python/lib/py4j-0.9-src.zip/py4j/java_gateway.py", line 813, in __call__
  File "/usr/lib/spark/python/pyspark/sql/utils.py", line 51, in deco
    raise AnalysisException(s.split(':', 1)[1], stackTrace)
pyspark.sql.utils.AnalysisException: u"cannot resolve 'kim' given input columns: [id, place, user, text, country];"
>>> sqlContext.sql('SELECT * from table1 WHERE user = kimo').collect()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/usr/lib/spark/python/pyspark/sql/context.py", line 580, in sql
    return DataFrame(self._ssql_ctx.sql(sqlQuery), self)
  File "/usr/lib/spark/python/lib/py4j-0.9-src.zip/py4j/java_gateway.py", line 813, in __call__
  File "/usr/lib/spark/python/pyspark/sql/utils.py", line 51, in deco
    raise AnalysisException(s.split(':', 1)[1], stackTrace)
pyspark.sql.utils.AnalysisException: u"cannot resolve 'kimo' given input columns: [id, place, user, text, country];"
>>> sqlContext.sql('SELECT * from table1 WHERE user = "kimo"').collect()
18/05/25 07:46:00 INFO storage.MemoryStore: Block broadcast_7 stored as values in memory (estimated size 191.1 KB, free
18/05/25 07:46:00 INFO storage.MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 22.1 KB, f
B)
18/05/25 07:46:00 INFO storage.BlockManagerInfo: Added broadcast_7_piece0 in memory on localhost:59263 (size: 22.1 KB, f
B)
18/05/25 07:46:00 INFO spark.SparkContext: Created broadcast 7 from collect at <stdin>:1
18/05/25 07:46:00 INFO storage.MemoryStore: Block broadcast_8 stored as values in memory (estimated size 194.9 KB, free
18/05/25 07:46:00 INFO storage.MemoryStore: Block broadcast_8_piece0 stored as bytes in memory (estimated size 22.3 KB,

```

```
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 <st
din>: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 b
ytes)
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+1
629848
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='United States', id='572647842277396480', place='Santee', text='obsessed with music', user='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
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.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 K
B)
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 M
B)
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 b
ytes)
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+1
629848
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.

```
cloudera@quickstart:~/Downloads/BDM1 Homework Assignment-1-20180521
File Edit View Search Terminal Help
+-----+
|8198|
+-----+

>>> 18/05/25 08:17:27 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 5.0 (TID 5) in 324 ms on
18/05/25 08:17:27 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed,

Traceback (most recent call last):
  File "/usr/lib/spark/python/pyspark/context.py", line 225, in signal_handler
    raise KeyboardInterrupt()
KeyboardInterrupt
>>> sqlContext.sql('SELECT user,count(*) from table1 GROUP BY user').show()
18/05/25 08:19:26 INFO storage.MemoryStore: Block broadcast_17 stored as values in memory (estimated size
18/05/25 08:19:26 INFO storage.MemoryStore: Block broadcast_17_piece0 stored as bytes in memory (estimate
KB)
18/05/25 08:19:26 INFO storage.BlockManagerInfo: Added broadcast_17_piece0 in memory on localhost:59263 (
MB)
18/05/25 08:19:26 INFO spark.SparkContext: Created broadcast 17 from showString at NativeMethodAccessorIn
18/05/25 08:19:26 INFO storage.MemoryStore: Block broadcast_18 stored as values in memory (estimated size
18/05/25 08:19:26 INFO storage.MemoryStore: Block broadcast_18_piece0 stored as bytes in memory (estimate
KB)
18/05/25 08:19:26 INFO storage.BlockManagerInfo: Added broadcast_18_piece0 in memory on localhost:59263 (
MB)
18/05/25 08:19:26 INFO spark.SparkContext: Created broadcast 18 from showString at NativeMethodAccessorIn
18/05/25 08:19:27 INFO mapred.FileInputFormat: Total input paths to process : 1
18/05/25 08:19:27 INFO spark.SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:-2
18/05/25 08:19:27 INFO scheduler.DAGScheduler: Registering RDD 47 (showString at NativeMethodAccessorImpl
18/05/25 08:19:27 INFO scheduler.DAGScheduler: Got job 5 (showString at NativeMethodAccessorImpl.java:-2)
18/05/25 08:19:27 INFO scheduler.DAGScheduler: Final stage: ResultStage 7 (showString at NativeMethodAcce
18/05/25 08:19:27 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 6)
18/05/25 08:19:27 INFO scheduler.DAGScheduler: Missing parents: List(ShuffleMapStage 6)
18/05/25 08:19:27 INFO scheduler.DAGScheduler: Submitting ShuffleMapStage 6 (MapPartitionsRDD[47] at show
sorImpl.java:-2), which has no missing parents
18/05/25 08:19:27 INFO storage.MemoryStore: Block broadcast_19 stored as values in memory (estimated size
18/05/25 08:19:27 INFO storage.MemoryStore: Block broadcast_19_piece0 stored as bytes in memory (estimate
KB)
18/05/25 08:19:27 INFO storage.BlockManagerInfo: Added broadcast_19_piece0 in memory on localhost:59263 (
B)
18/05/25 08:19:27 INFO spark.SparkContext: Created broadcast 19 from broadcast at DAGScheduler.scala:1006
```



```
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Help
nt-1-20180521/tweets.json:0+1629848
18/05/25 10:50:25 INFO codegen.GenerateSafeProjection: Code generated in 18.447863 ms
18/05/25 10:50:25 INFO executor.Executor: Finished task 0.0 in stage 5.0 (TID 5). 6997 bytes result sent to driver
18/05/25 10:50:25 INFO scheduler.DAGScheduler: ResultStage 5 (showString at NativeMethodAccessorImpl.java:-2) finished in 0.089 s
18/05/25 10:50:25 INFO scheduler.DAGScheduler: Job 5 finished: showString at NativeMethodAccessorImpl.java:-2, took 0.106140 s
+-----+
| user | text |
+-----+
| #QuissyUpSoon | @DjGregStreet ST0... |
| #QuissyUpSoon | @TheDJ33 STOP... |
| #QuissyUpSoon | @Chica0823 STOP... |
| #QuissyUpSoon | @prettydrea7 STOP... |
| #QuissyUpSoon | @pablo_valbuena S... |
| #QuissyUpSoon | @HeraldLynnndee ST... |
| #QuissyUpSoon | @starsbarsnpbrs S... |
| #QuissyUpSoon | @cocoshanelle23 S... |
| #QuissyUpSoon | @OfficialDilemma ... |
| #QuissyUpSoon | @DJQlassick STOP?... |
| #QuissyUpSoon | @Djholiday STOP... |
| #QuissyUpSoon | @SlushiesForHoes ... |
| #QuissyUpSoon | @995jamz STOP... |
| #QuissyUpSoon | @sammsubido STOP... |
| #QuissyUpSoon | @JasonBrunchezz S... |
| #QuissyUpSoon | @_LOWEN_III STOP?... |
| Pentatonic Music | Muzique Necklace ... |
| #QuissyUpSoon | @BlueWealthyGang ... |
| #QuissyUpSoon | @whxtewhxtemike S... |
| #QuissyUpSoon | @AmbriaBombshell ... |
+-----+
only showing top 20 rows
```

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

```
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Help
18/05/25 11:24:37 INFO scheduler.TaskSetManager: Finished task 198.0 in stage 16.0 (TID 804) in 18 ms on localhost (199/199)
18/05/25 11:24:37 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 16.0, whose tasks have all completed, from pool
18/05/25 11:24:37 INFO scheduler.DAGScheduler: ResultStage 16 (showString at NativeMethodAccessorImpl.java:-2) finished in 1.522 s
18/05/25 11:24:37 INFO scheduler.DAGScheduler: Job 8 finished: showString at NativeMethodAccessorImpl.java:-2, took 1.591924 s
+-----+
| user | _c1 |
+-----+
| SINGER MINTU | 1 |
| ΔXᵀ₃₆3KΔ ᵀᵀ | 1 |
| LOAD | 4 |
| UB40 | 1 |
| Pentatonic Music | 2 |
| PyxeeStyx | 1 |
| xelA7th | 2 |
| #KingKong | 1 |
| Alexander Lewis | 1 |
| Zelf Clothing | 1 |
| . | 1 |
| #TurnYaSneakUp | 21 |
| Bobby Borg | 1 |
| joey | 1 |
| B | 1 |
| F | 1 |
| Big Shot Music Group | 1 |
| a | 1 |
| b | 2 |
| Theresa Neely | 1 |
+-----+
only showing top 20 rows
>>> sqlContext.sql("SELECT user, count(user) FROM table1 where text like CONCAT('%',user,'%') GROUP BY user ").show()
```

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]
>>> df2 =sqlContext.sql("SELECT count(user) FROM table1 where text like CONCAT('%',user,'%') GROUP BY user ")
```

```
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.registerDataFrameAsTable(df2, 'table2')
```

```

cloudera@quickstart:~/Downloads
File Edit View Search Terminal Help
+-----+
|      | | 1 |
|      | | 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, fr
ee: 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 1.576311 s

+-----+
| user | Number |
+-----+
| #QuissyUpSoon | 255 |
| #TurnYaSneakUp | 21 |
| LOAD | 4 |
| LAGSAW | 4 |
| xeIA7th | 2 |
| b | 2 |
| Pentatonic Music | 2 |
| ΔXj9€3KA | 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,'%') GROUP BY user ")
18/05/25 11:58:16 INFO storage.BlockManagerInfo: Removed broadcast_54_piece0 on localhost:35951 in memory (size: 22.3 KB, free: 534.5 MB)

```


Get a list of all hashtags mentioned in the dataset.

```
cloudera@quickstart:~/Downloads/BDM1 Homework Assignment-1-20180521
File Edit View Search Terminal Help
18/05/25 21:30:07 INFO scheduler.DAGScheduler: ResultStage 3 (showString at NativeMethodAccessorImpl.java:-2) finished in 0.270 s
18/05/25 21:30:07 INFO scheduler.DAGScheduler: Job 3 finished: showString at NativeMethodAccessorImpl.java:-2, took 0.292016 s
+-----+
|      text      |
+-----+
|#music #fun Celeb...|
|#myxmusicawards F...|
|#music #music #music|
|#love #art #new #...|
|#love #art #new #...|
|#myxmusicawards F...|
|#myxmusicawards F...|
|#music #childhood...|
|#myxmusicawards F...|
|#myxmusicawards F...|
|#myxmusicawards F...|
|#nofilter Hotel d...|
|#anodyne #coffee ...|
|#myxmusicawards F...|
|#myxmusicawards F...|
|#music always #sa...|
|#myxmusicawards F...|
|#maximumthehormon...|
|#myxmusicawards F...|
|#aviary Rehearsin...|
+-----+
only showing top 20 rows

>>> 18/05/25 21:30:07 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 272 ms on localhost (1/1)
18/05/25 21:30:07 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
>>>
```

```
cloudera@quickstart:~/Downloads/BDM1 Homework Assignment-1-20180521
File Edit View Search Terminal Help
18/05/25 21:30:07 INFO scheduler.DAGScheduler: ResultStage 3 (showString at NativeMethodAccessorImpl.java:-2) finished in 0.270 s
18/05/25 21:30:07 INFO scheduler.DAGScheduler: Job 3 finished: showString at NativeMethodAccessorImpl.java:-2, took 0.292016 s
+-----+
|      text      |
+-----+
|#music #fun Celeb...|
|#myxmusicawards F...|
|#music #music #music|
|#love #art #new #...|
|#love #art #new #...|
|#myxmusicawards F...|
|#myxmusicawards F...|
|#music #childhood...|
|#myxmusicawards F...|
|#myxmusicawards F...|
|#myxmusicawards F...|
|#nofilter Hotel d...|
|#anodyne #coffee ...|
|#myxmusicawards F...|
|#myxmusicawards F...|
|#music always #sa...|
|#myxmusicawards F...|
|#maximumthehormon...|
|#myxmusicawards F...|
|#aviary Rehearsin...|
+-----+
only showing top 20 rows

>>> 18/05/25 21:30:07 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 272 ms on localhost (1/1)
18/05/25 21:30:07 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
>>> sqlContext.sql("SELECT text FROM Table1 WHERE text RLIKE '^#[[:alnum:]]+'.show()")
```

Find how many times each hashtag is mentioned in the dataset

```
cloudera@quickstart:~/Downloads/BDM1 Homework Assignment-1-20180521
File Edit View Search Terminal Help
18/05/25 22:13:15 INFO scheduler.TaskSetManager: Finished task 198.0 in stage 13.0 (TID 604) in 5 ms on localhost (199/199)
18/05/25 22:13:15 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 13.0, whose tasks have all completed, from pool
18/05/25 22:13:15 INFO scheduler.DAGScheduler: ResultStage 13 (showString at NativeMethodAccessorImpl.java:-2) finished in 1.986 s
18/05/25 22:13:15 INFO scheduler.DAGScheduler: Job 7 finished: showString at NativeMethodAccessorImpl.java:-2, took 2.065664 s
+-----+-----+
|      text      |counter|
+-----+-----+
|#music #fun Celeb...|      1|
|#myxmusicawards F...|      1|
|#myxmusicawards F...|      1|
|#aviary Rehearsin...|      1|
|#mugshot @ Paris,...|      1|
|#nofilter Hotel d...|      1|
|#love #art #new #...|      1|
|#notredame #paris...|      1|
|#new #signing #ca...|      1|
|#myxmusicawards F...|      1|
|#nowplaying Hand ...|      1|
|#miniShowCase [H...]|      1|
|#music always #sa...|      1|
|#audition #Bourne...|      1|
|#music #love #bla...|      1|
|#love #art #new #...|      1|
|#love #art #new #...|      1|
|#morning #coffee ...|      1|
|#music suggestion...|      1|
|#love #art #new #...|      1|
+-----+-----+
only showing top 20 rows

>>>
```

```

cloudera@quickstart:~/Downloads/BDM1 Homework Assignment-1-20180521
File Edit View Search Terminal Help
18/05/25 22:13:15 INFO scheduler.TaskSetManager: Finished task 198.0 in stage 13.0 (TID 604) in 5 ms on localhost (199/199)
18/05/25 22:13:15 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 13.0, whose tasks have all completed, from pool
18/05/25 22:13:15 INFO scheduler.DAGScheduler: ResultStage 13 (showString at NativeMethodAccessorImpl.java:-2) finished in 1.986 s
18/05/25 22:13:15 INFO scheduler.DAGScheduler: Job 7 finished: showString at NativeMethodAccessorImpl.java:-2, took 2.065664 s
+-----+-----+
|      text|counter|
+-----+-----+
|#music #fun Celeb...|1|
|#myxmusicawards F...|1|
|#myxmusicawards F...|1|
|#aviary Relhearsin...|1|
|#mugshot @ Paris...|1|
|#nofilter Hotel d...|1|
|#Love #art #new #...|1|
|#notredame #paris...|1|
|#new #signing #ca...|1|
|#myxmusicawards F...|1|
|#nowplaying Hand ...|1|
|#miniShowCase [REDACTED]|1|
|#music always #sa...|1|
|#audition #Bourne...|1|
|#music #love #bla...|1|
|#Love #art #new #...|1|
|#Love #art #new #...|1|
|#morning #coffee ...|1|
|#music suggestion...|1|
|#Love #art #new #...|1|
+-----+-----+
only showing top 20 rows

>>> sqlContext.sql("SELECT text ,count(1) AS counter FROM table1 WHERE text RLIKE '^#[[:alnum:]]' GROUP BY text ").show()

```

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

```

cloudera@quickstart:~/Downloads/BDM1 Homework Assignment-1-20180521
File Edit View Search Terminal Help
18/05/25 22:20:33 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 16.0 (TID 607) in 28 ms on localhost (1/1)
18/05/25 22:20:33 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 16.0, whose tasks have all completed, from pool
18/05/25 22:20:33 INFO scheduler.DAGScheduler: ResultStage 16 (showString at NativeMethodAccessorImpl.java:-2) finished in 0.022 s
18/05/25 22:20:33 INFO scheduler.DAGScheduler: Job 10 finished: showString at NativeMethodAccessorImpl.java:-2, took 0.035737 s
+-----+-----+-----+-----+
|country|id|place|text|user|
+-----+-----+-----+-----+
|France|572704775713628161|Paris|'pivert', 'usager...|Trendsmap Paris|
|France|572685132475305984|Paris|@jhornain Découvr...|Codeclac|
|France|572678334687150080|Paris|Bon mardi ! OCTAV...|EPHYRE Paris|
|France|572672815926796288|Paris|@YFEFRANCE Hey! Y...|Helene PASQUALETTI|
|France|572601453728038914|Paris|[REDACTED]|Takaaki Kishida|
|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|[REDACTED] @ Torr...|[REDACTED]|
|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|572661233763143680|Paris|#WakeUp #BelleJou...|Wil|
|France|572667989725618176|Paris|Paris: * Temp: 3°...|FranceBidet|
|France|572673009468633088|Paris|fing, @la fing es...|Trendsmap Paris|
|France|572562242237030400|Paris|[REDACTED] Level 2 ...|ابنكلل مومون ربم|
|France|572705672917340160|Paris|@trashdistance Je...|Russell Williams|
|France|572699088807956480|Paris|I'm at Coutume In...|Lauralou B|
|France|572699088824758273|Paris|Isabelle VIALLE, ...|ART aujourd'hui gal|
+-----+-----+-----+-----+
only showing top 20 rows

>>> sqlContext.sql("SELECT * FROM table1 WHERE place ='Paris' ").show()

```

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|
|Mexico|52|57|
|México|47|59|
|Malaysia|37|50|
|Türkiye|36|42|
|España|32|53|
|Colombia|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...|@Ashton5S0S be mo...|CrazyRockCountryLady|
|France|572704775713628161|Paris|'pivert', 'usager...|Trendsmap Paris|
|France|572685132475305984|Paris|@jhornain Découvr...|Codecllic|
|France|572664931973242880|Chessy|Happy campers! #D...|Candice LeRae|
|France|572672815926796288|Paris|@YFEFRANCE Hey! Y...|Helene PASQUALETTI|
|France|572648068832755712|Saint-Avoid|Opéra Garnier Par...|Dominique Lang|
|France|572557043942461440|Chessy|#Disneyland #Pari...|saif ali faraj|
|France|572675772047101952|Oissel|départ pour Paris |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|@Torr...|Torr...|
|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()

```

