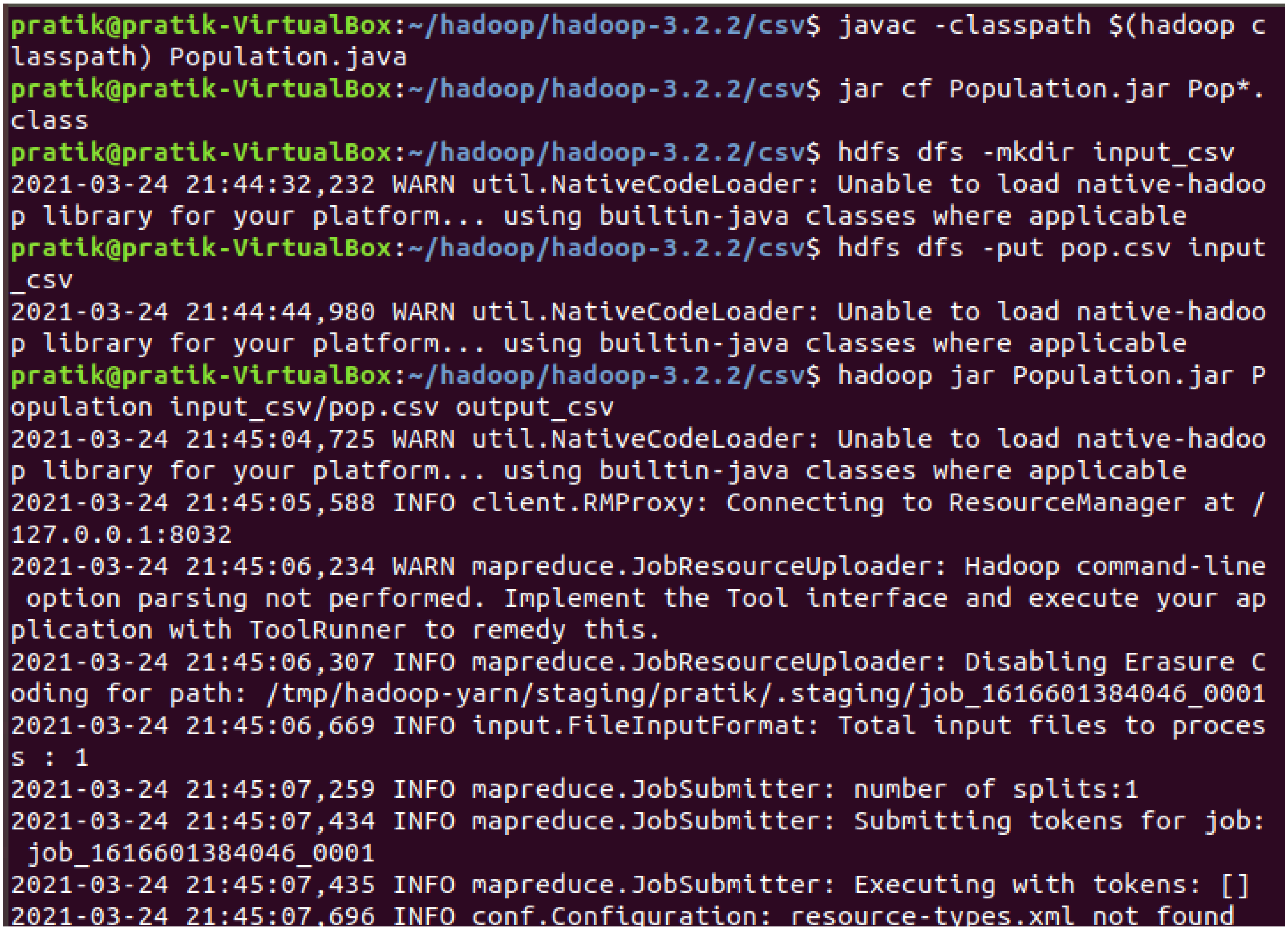
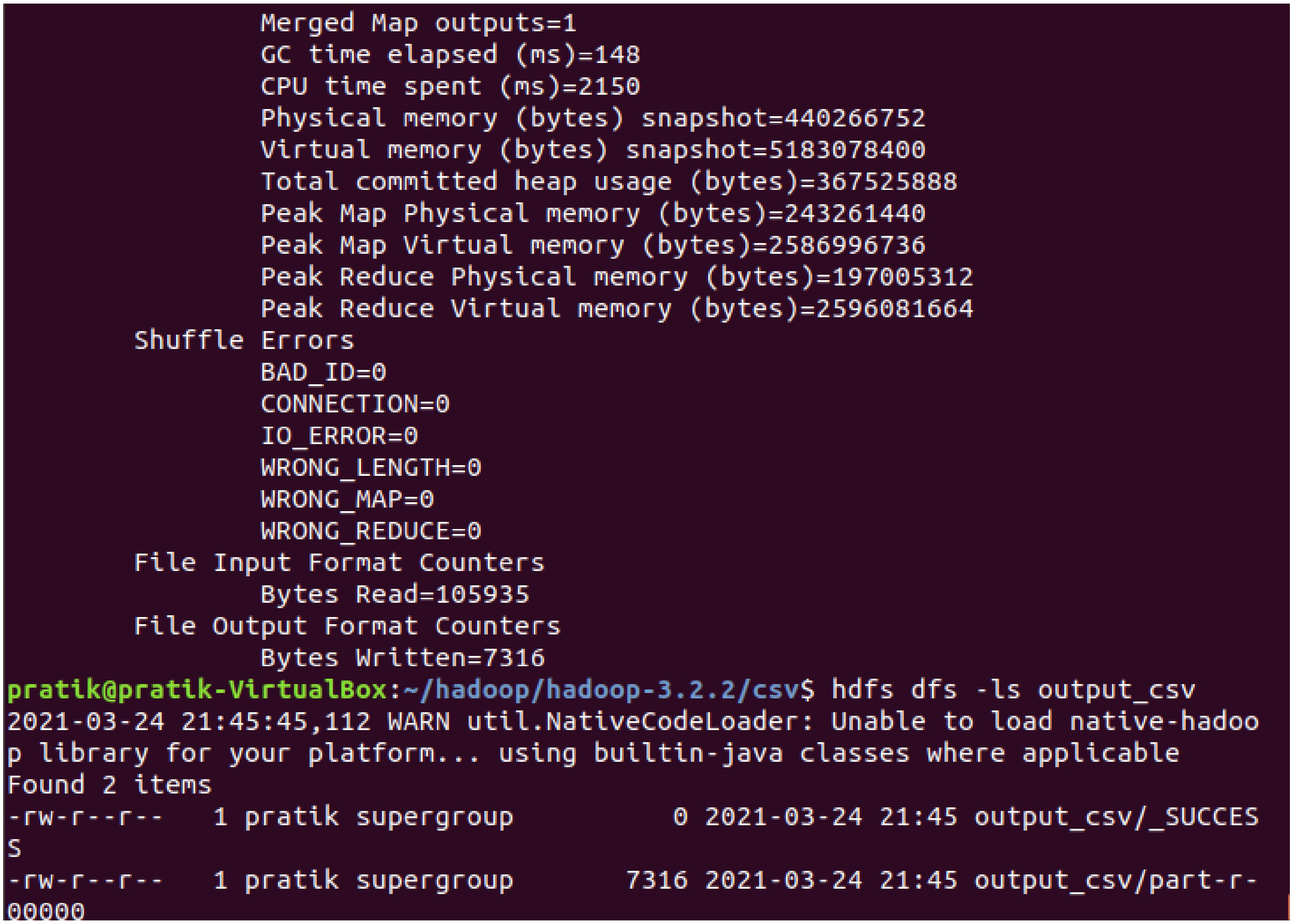
Big Data Week-7 workshop

Pratik Sarkar 2039301

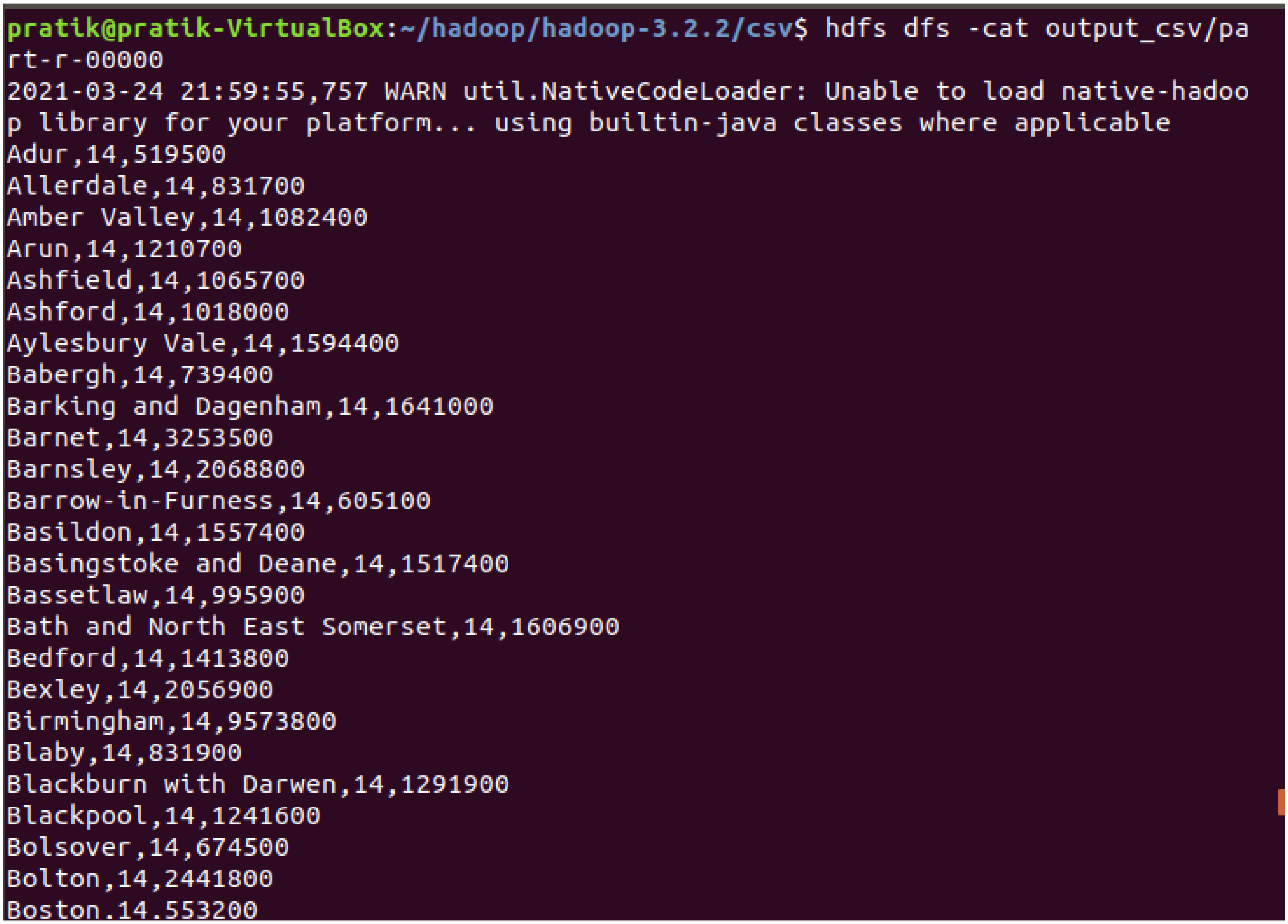
1. Running Population.java



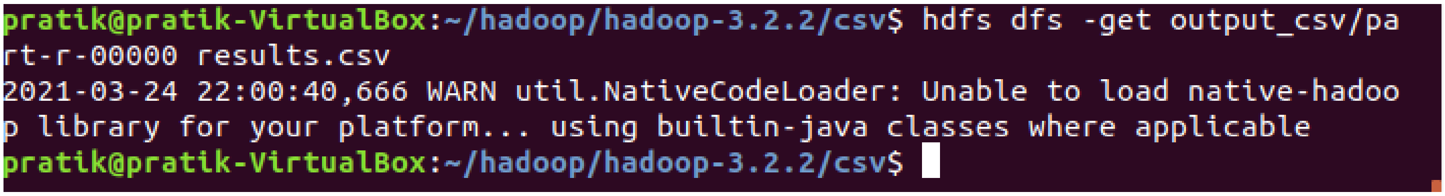
```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/csv$ javac -classpath $(hadoop c
lasspath) Population.java
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/csv$ jar cf Population.jar Pop*.
class
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/csv$ hdfs dfs -mkdir input_csv
2021-03-24 21:44:32,232 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/csv$ hdfs dfs -put pop.csv input
_csv
2021-03-24 21:44:44,980 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/csv$ hadoop jar Population.jar P
opulation input_csv/pop.csv output_csv
2021-03-24 21:45:04,725 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
2021-03-24 21:45:05,588 INFO client.RMProxy: Connecting to ResourceManager at /
127.0.0.1:8032
2021-03-24 21:45:06,234 WARN mapreduce.JobResourceUploader: Hadoop command-line
 option parsing not performed. Implement the Tool interface and execute your ap
plication with ToolRunner to remedy this.
2021-03-24 21:45:06,307 INFO mapreduce.JobResourceUploader: Disabling Erasure C
oding for path: /tmp/hadoop-yarn/staging/pratik/.staging/job_1616601384046_0001
2021-03-24 21:45:06,669 INFO input.FileInputFormat: Total input files to proces
s : 1
2021-03-24 21:45:07,259 INFO mapreduce.JobSubmitter: number of splits:1
2021-03-24 21:45:07,434 INFO mapreduce.JobSubmitter: Submitting tokens for job:
 job_1616601384046_0001
2021-03-24 21:45:07,435 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-03-24 21:45:07,696 INFO conf.Configuration: resource-types.xml not found
```



```
                Merged Map outputs=1
                GC time elapsed (ms)=148
                CPU time spent (ms)=2150
                Physical memory (bytes) snapshot=440266752
                Virtual memory (bytes) snapshot=5183078400
                Total committed heap usage (bytes)=367525888
                Peak Map Physical memory (bytes)=243261440
                Peak Map Virtual memory (bytes)=2586996736
                Peak Reduce Physical memory (bytes)=197005312
                Peak Reduce Virtual memory (bytes)=2596081664
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=105935
        File Output Format Counters
                Bytes Written=7316
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/csv$ hdfs dfs -ls output_csv
2021-03-24 21:45:45,112 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 pratik supergroup          0 2021-03-24 21:45 output_csv/_SUCCES
S
-rw-r--r--   1 pratik supergroup       7316 2021-03-24 21:45 output_csv/part-r-
00000
```

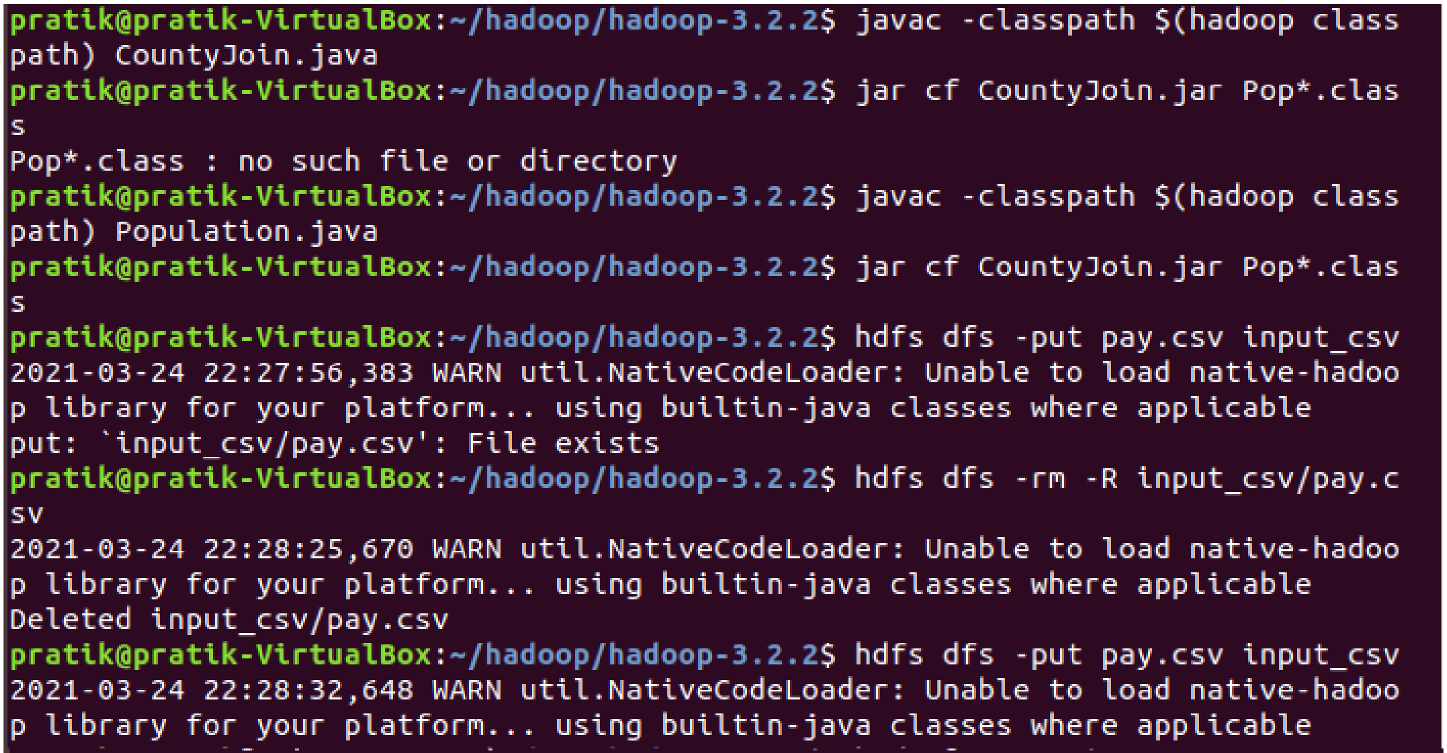


```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/csv$ hdfs dfs -cat output_csv/pa
rt-r-00000
2021-03-24 21:59:55,757 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Adur,14,519500
Allerdale,14,831700
Amber Valley,14,1082400
Arun,14,1210700
Ashfield,14,1065700
Ashford,14,1018000
Aylesbury Vale,14,1594400
Babergh,14,739400
Barking and Dagenham,14,1641000
Barnet,14,3253500
Barnsley,14,2068800
Barrow-in-Furness,14,605100
Basildon,14,1557400
Basingstoke and Deane,14,1517400
Bassetlaw,14,995900
Bath and North East Somerset,14,1606900
Bedford,14,1413800
Bexley,14,2056900
Birmingham,14,9573800
Blaby,14,831900
Blackburn with Darwen,14,1291900
Blackpool,14,1241600
Bolsover,14,674500
Bolton,14,2441800
Boston,14,553200
```

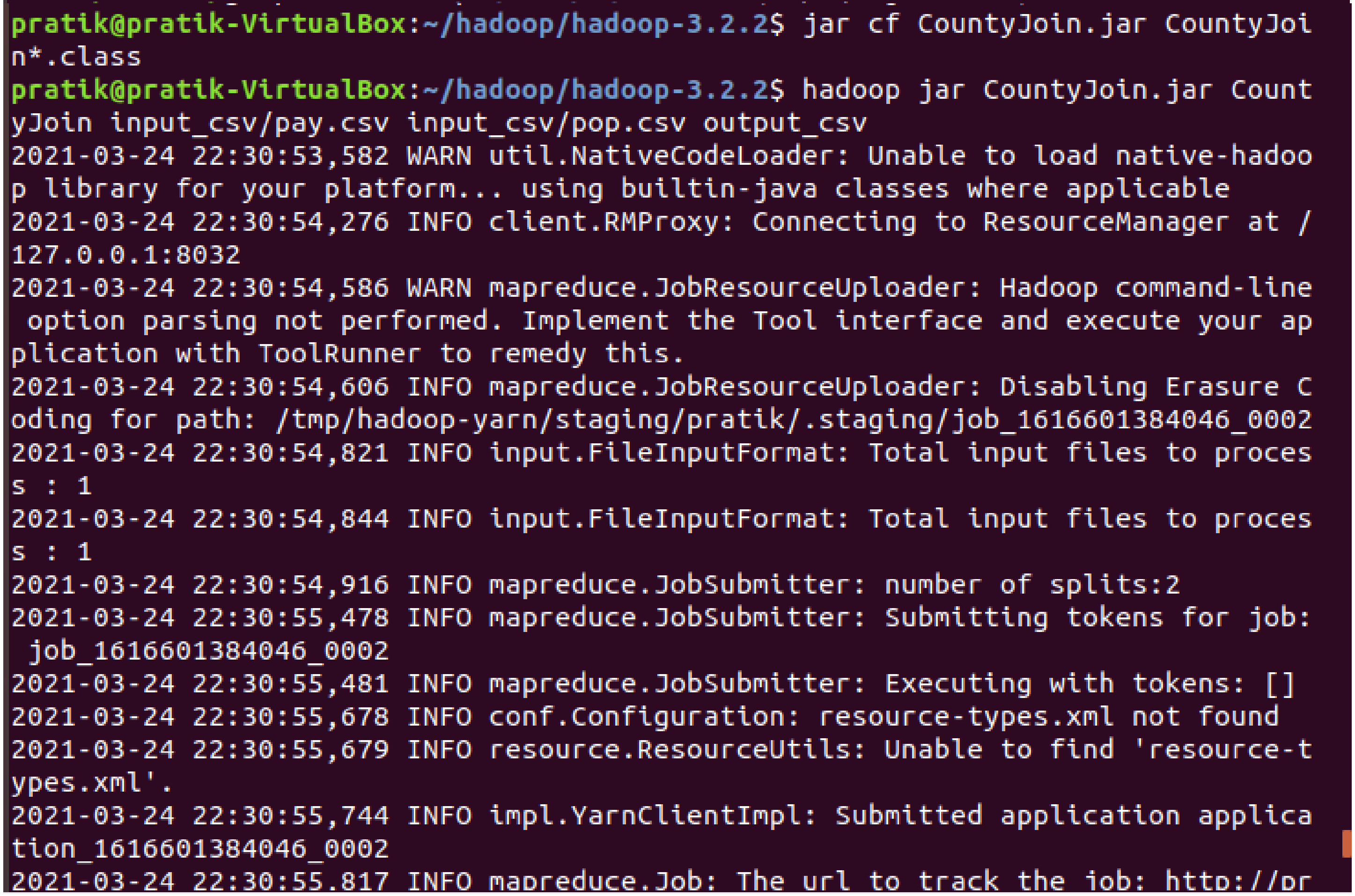


```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/csv$ hdfs dfs -get output_csv/pa
rt-r-00000 results.csv
2021-03-24 22:00:40,666 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2/csv$ 
```

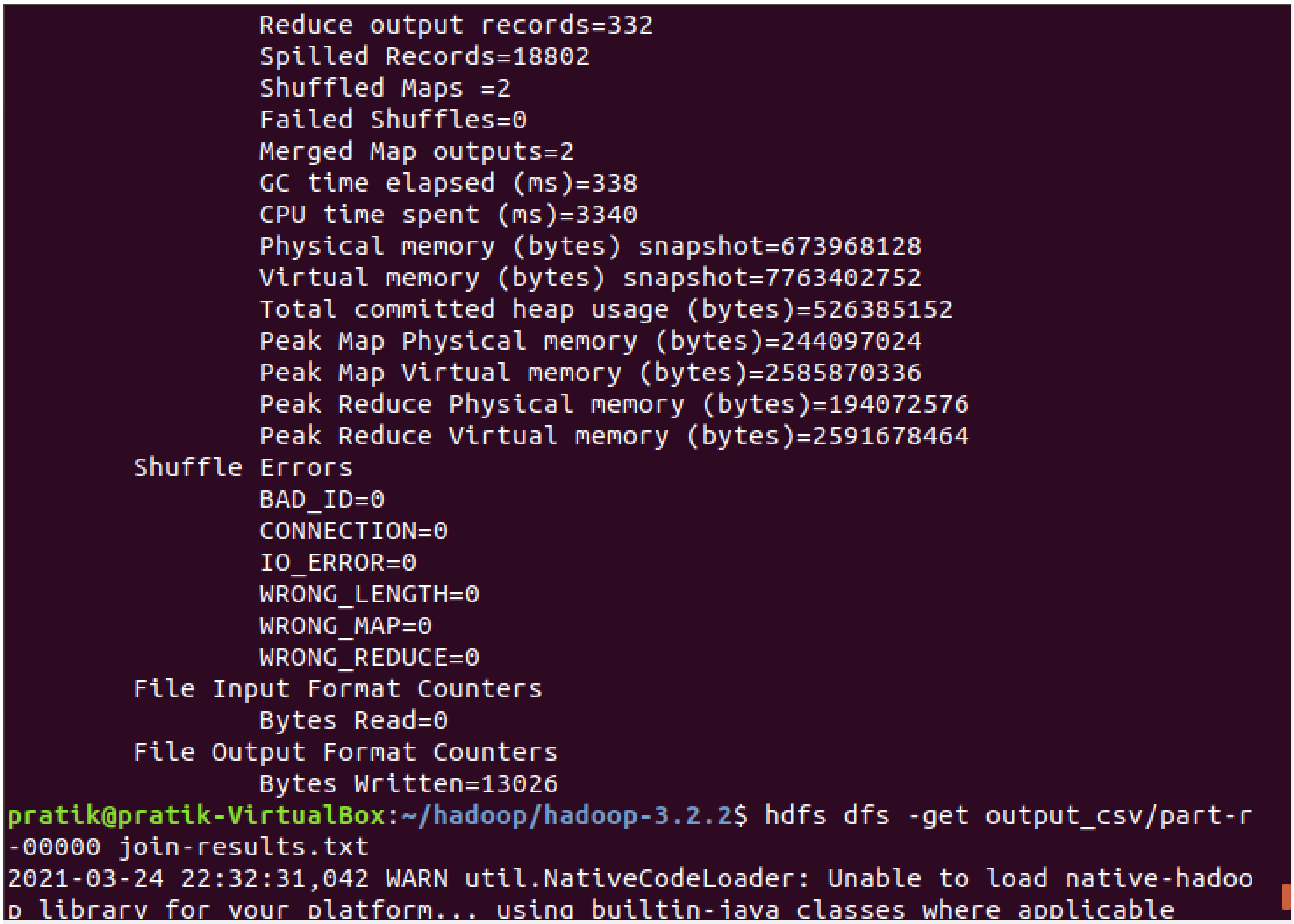
2. Joining the Datasets – CountyJoin.java



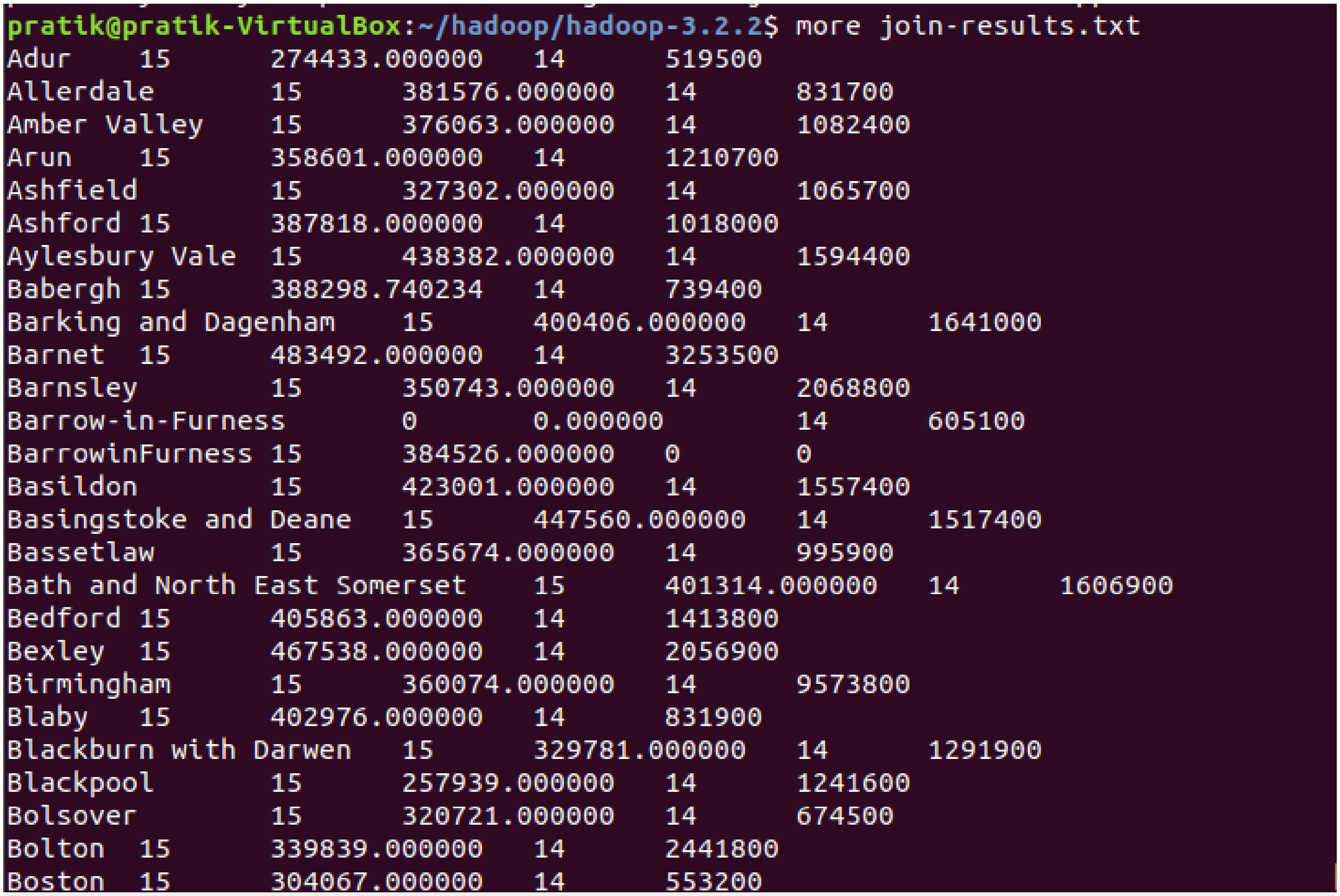
```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ javac -classpath $(hadoop class
path) CountyJoin.java
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ jar cf CountyJoin.jar Pop*.clas
s
Pop*.class : no such file or directory
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ javac -classpath $(hadoop class
path) Population.java
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ jar cf CountyJoin.jar Pop*.clas
s
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ hdfs dfs -put pay.csv input_csv
2021-03-24 22:27:56,383 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
put: `input_csv/pay.csv': File exists
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ hdfs dfs -rm -R input_csv/pay.c
sv
2021-03-24 22:28:25,670 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
Deleted input_csv/pay.csv
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ hdfs dfs -put pay.csv input_csv
2021-03-24 22:28:32,648 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
```



```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ jar cf CountyJoin.jar CountyJoi
n*.class
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ hadoop jar CountyJoin.jar Count
yJoin input_csv/pay.csv input_csv/pop.csv output_csv
2021-03-24 22:30:53,582 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
2021-03-24 22:30:54,276 INFO client.RMProxy: Connecting to ResourceManager at /
127.0.0.1:8032
2021-03-24 22:30:54,586 WARN mapreduce.JobResourceUploader: Hadoop command-line
 option parsing not performed. Implement the Tool interface and execute your ap
plication with ToolRunner to remedy this.
2021-03-24 22:30:54,606 INFO mapreduce.JobResourceUploader: Disabling Erasure C
oding for path: /tmp/hadoop-yarn/staging/pratik/.staging/job_1616601384046_0002
2021-03-24 22:30:54,821 INFO input.FileInputFormat: Total input files to proces
s : 1
2021-03-24 22:30:54,844 INFO input.FileInputFormat: Total input files to proces
s : 1
2021-03-24 22:30:54,916 INFO mapreduce.JobSubmitter: number of splits:2
2021-03-24 22:30:55,478 INFO mapreduce.JobSubmitter: Submitting tokens for job:
 job_1616601384046_0002
2021-03-24 22:30:55,481 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-03-24 22:30:55,678 INFO conf.Configuration: resource-types.xml not found
2021-03-24 22:30:55,679 INFO resource.ResourceUtils: Unable to find 'resource-t
ypes.xml'.
2021-03-24 22:30:55,744 INFO impl.YarnClientImpl: Submitted application applica
tion_1616601384046_0002
2021-03-24 22:30:55.817 INFO mapreduce.Job: The url to track the job: http://pr
```



```
                Reduce output records=332
                Spilled Records=18802
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=338
                CPU time spent (ms)=3340
                Physical memory (bytes) snapshot=673968128
                Virtual memory (bytes) snapshot=7763402752
                Total committed heap usage (bytes)=526385152
                Peak Map Physical memory (bytes)=244097024
                Peak Map Virtual memory (bytes)=2585870336
                Peak Reduce Physical memory (bytes)=194072576
                Peak Reduce Virtual memory (bytes)=2591678464
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=13026
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ hdfs dfs -get output_csv/part-r
-00000 join-results.txt
2021-03-24 22:32:31,042 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
```

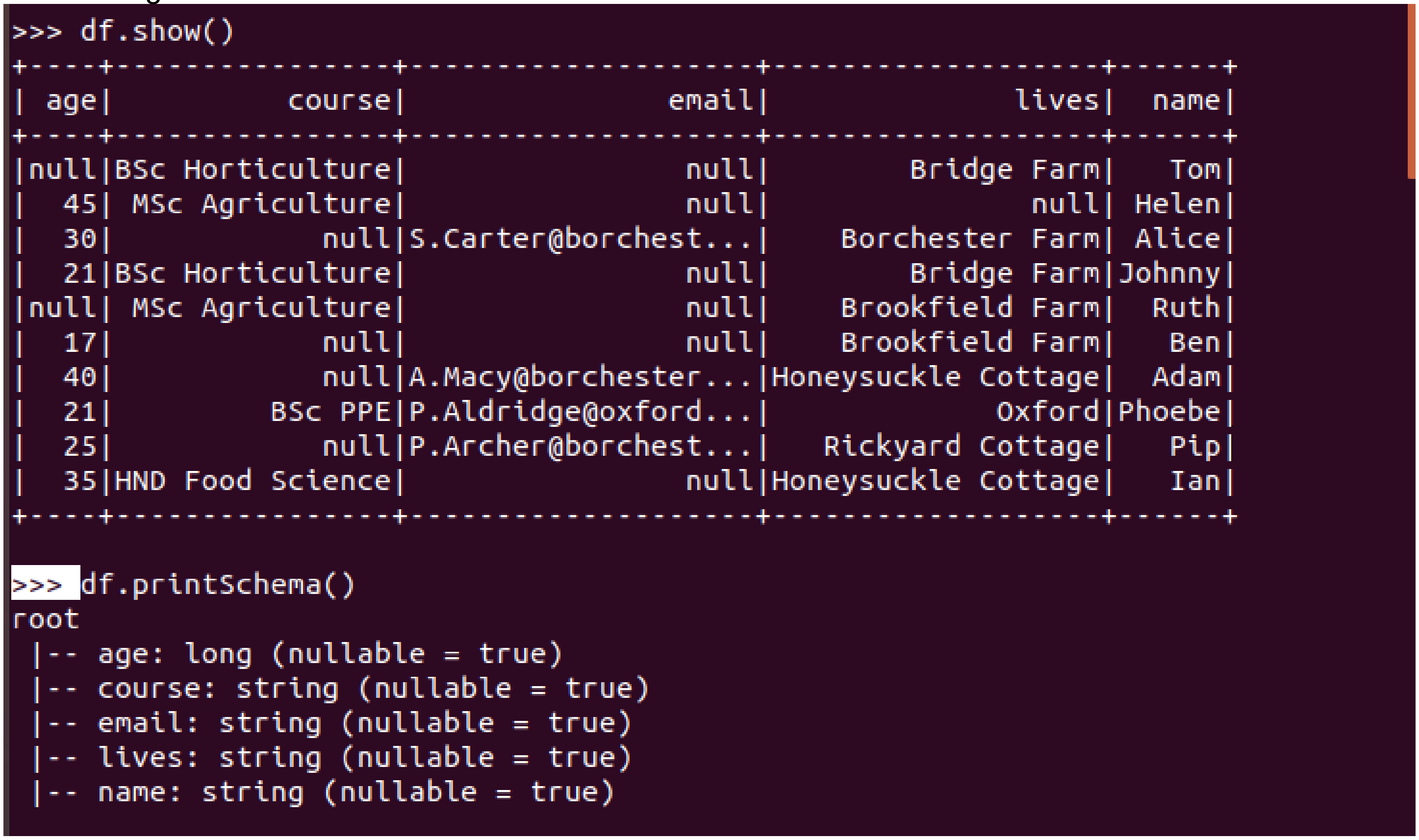


```
pratik@pratik-VirtualBox:~/hadoop/hadoop-3.2.2$ more join-results.txt
Adur    15      274433.000000   14      519500
Allerdale       15      381576.000000   14      831700
Amber Valley    15      376063.000000   14      1082400
Arun    15      358601.000000   14      1210700
Ashfield        15      327302.000000   14      1065700
Ashford 15      387818.000000   14      1018000
Aylesbury Vale  15      438382.000000   14      1594400
Babergh 15      388298.740234   14      739400
Barking and Dagenham    15      400406.000000   14      1641000
Barnet  15      483492.000000   14      3253500
Barnsley        15      350743.000000   14      2068800
Barrow-in-Furness       0       0.000000        14      605100
BarrowinFurness 15      384526.000000   0       0
Basildon        15      423001.000000   14      1557400
Basingstoke and Deane   15      447560.000000   14      1517400
Bassetlaw       15      365674.000000   14      995900
Bath and North East Somerset    15      401314.000000   14      1606900
Bedford 15      405863.000000   14      1413800
Bexley  15      467538.000000   14      2056900
Birmingham      15      360074.000000   14      9573800
Blaby   15      402976.000000   14      831900
Blackburn with Darwen   15      329781.000000   14      1291900
Blackpool       15      257939.000000   14      1241600
Bolsover        15      320721.000000   14      674500
Bolton  15      339839.000000   14      2441800
Boston  15      304067.000000   14      553200
```

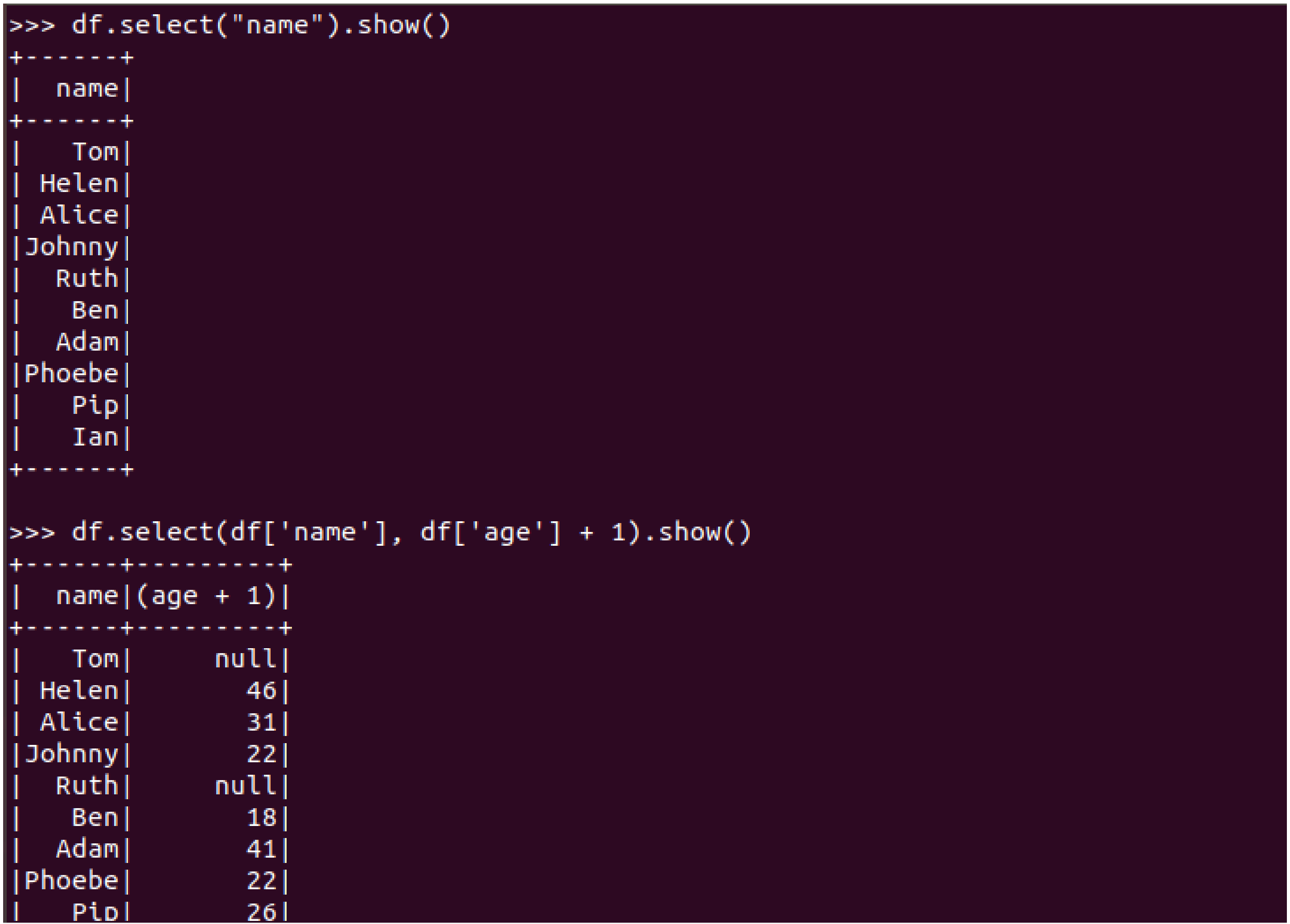
3. Apache Spark

3.1 Student.json

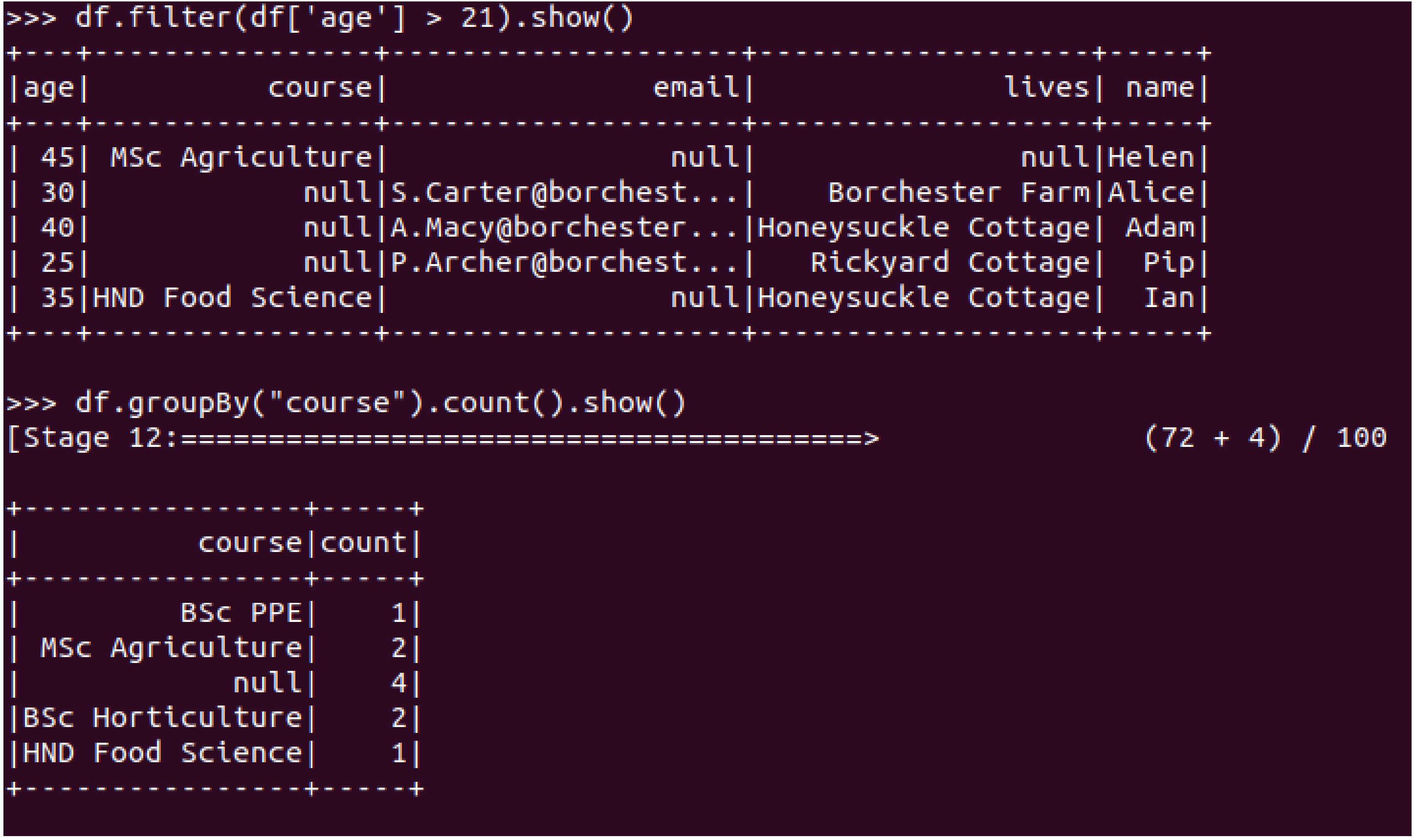
3.1.1 Using Data Frames



```
>>> df.show()
+----+----------------+--------------------+-------------------+------+
| age|          course|               email|              lives|  name|
+----+----------------+--------------------+-------------------+------+
|null|BSc Horticulture|                null|        Bridge Farm|   Tom|
|  45| MSc Agriculture|                null|               null| Helen|
|  30|            null|S.Carter@borchest...|    Borchester Farm| Alice|
|  21|BSc Horticulture|                null|        Bridge Farm|Johnny|
|null| MSc Agriculture|                null|    Brookfield Farm|  Ruth|
|  17|            null|                null|    Brookfield Farm|   Ben|
|  40|            null|A.Macy@borchester...|Honeysuckle Cottage|  Adam|
|  21|         BSc PPE|P.Aldridge@oxford...|             Oxford|Phoebe|
|  25|            null|P.Archer@borchest...|   Rickyard Cottage|   Pip|
|  35|HND Food Science|                null|Honeysuckle Cottage|   Ian|
+----+----------------+--------------------+-------------------+------+

>>> df.printSchema()
root
 |-- age: long (nullable = true)
 |-- course: string (nullable = true)
 |-- email: string (nullable = true)
 |-- lives: string (nullable = true)
 |-- name: string (nullable = true)
```

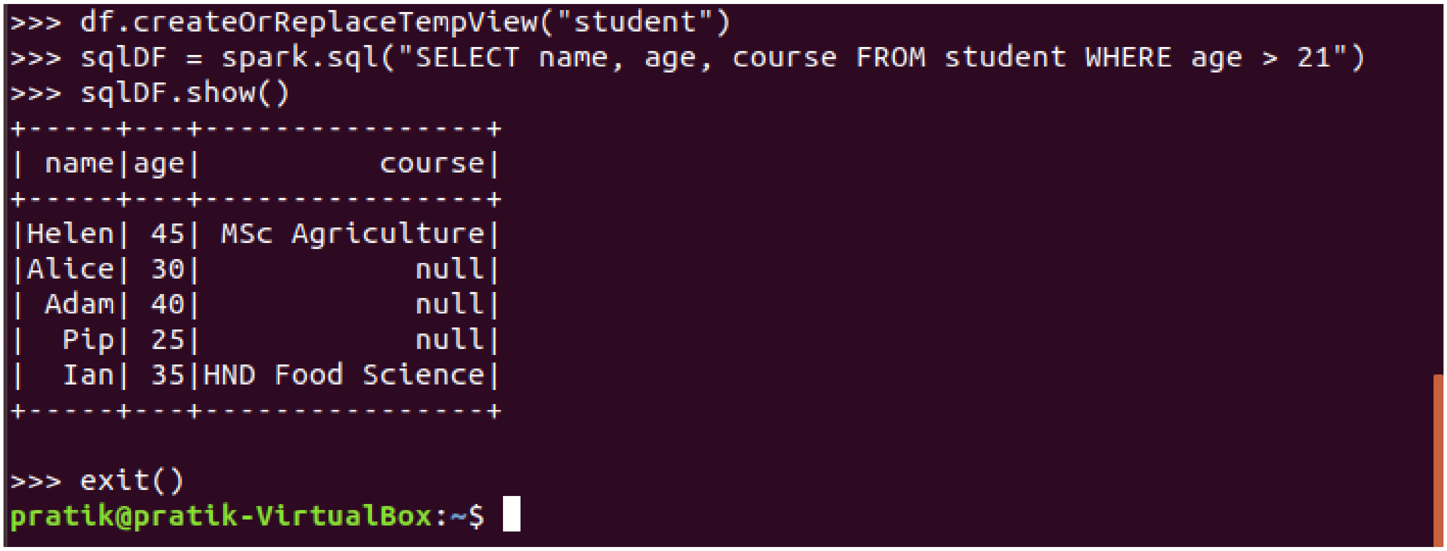


```
>>> df.select("name").show()
+------+
|  name|
+------+
|   Tom|
| Helen|
| Alice|
|Johnny|
|  Ruth|
|   Ben|
|  Adam|
|Phoebe|
|   Pip|
|   Ian|
+------+

>>> df.select(df['name'], df['age'] + 1).show()
+------+---------+
|  name|(age + 1)|
+------+---------+
|   Tom|     null|
| Helen|       46|
| Alice|       31|
|Johnny|       22|
|  Ruth|     null|
|   Ben|       18|
|  Adam|       41|
|Phoebe|       22|
|   Pip|       26|
```



```
>>> df.filter(df['age'] > 21).show()
+---+----------------+--------------------+-------------------+-----+
|age|          course|               email|              lives| name|
+---+----------------+--------------------+-------------------+-----+
| 45| MSc Agriculture|                null|               null|Helen|
| 30|            null|S.Carter@borchest...|    Borchester Farm|Alice|
| 40|            null|A.Macy@borchester...|Honeysuckle Cottage| Adam|
| 25|            null|P.Archer@borchest...|   Rickyard Cottage|  Pip|
| 35|HND Food Science|                null|Honeysuckle Cottage|  Ian|
+---+----------------+--------------------+-------------------+-----+

>>> df.groupBy("course").count().show()
[Stage 12:=======================================>               (72 + 4) / 100

+----------------+-----+
|          course|count|
+----------------+-----+
|         BSc PPE|    1|
| MSc Agriculture|    2|
|            null|    4|
|BSc Horticulture|    2|
|HND Food Science|    1|
+----------------+-----+
```

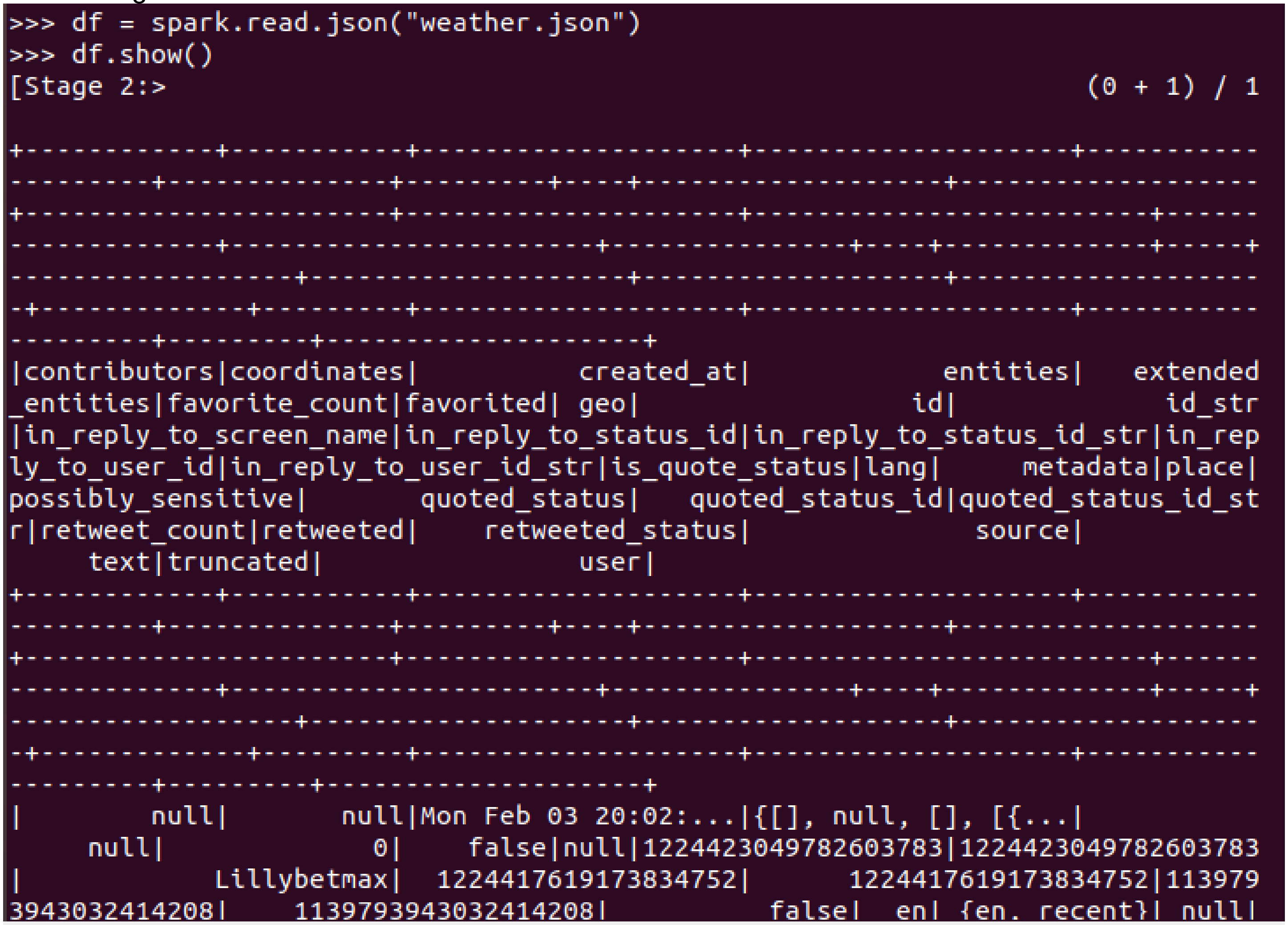
3.1.2 Using SQL



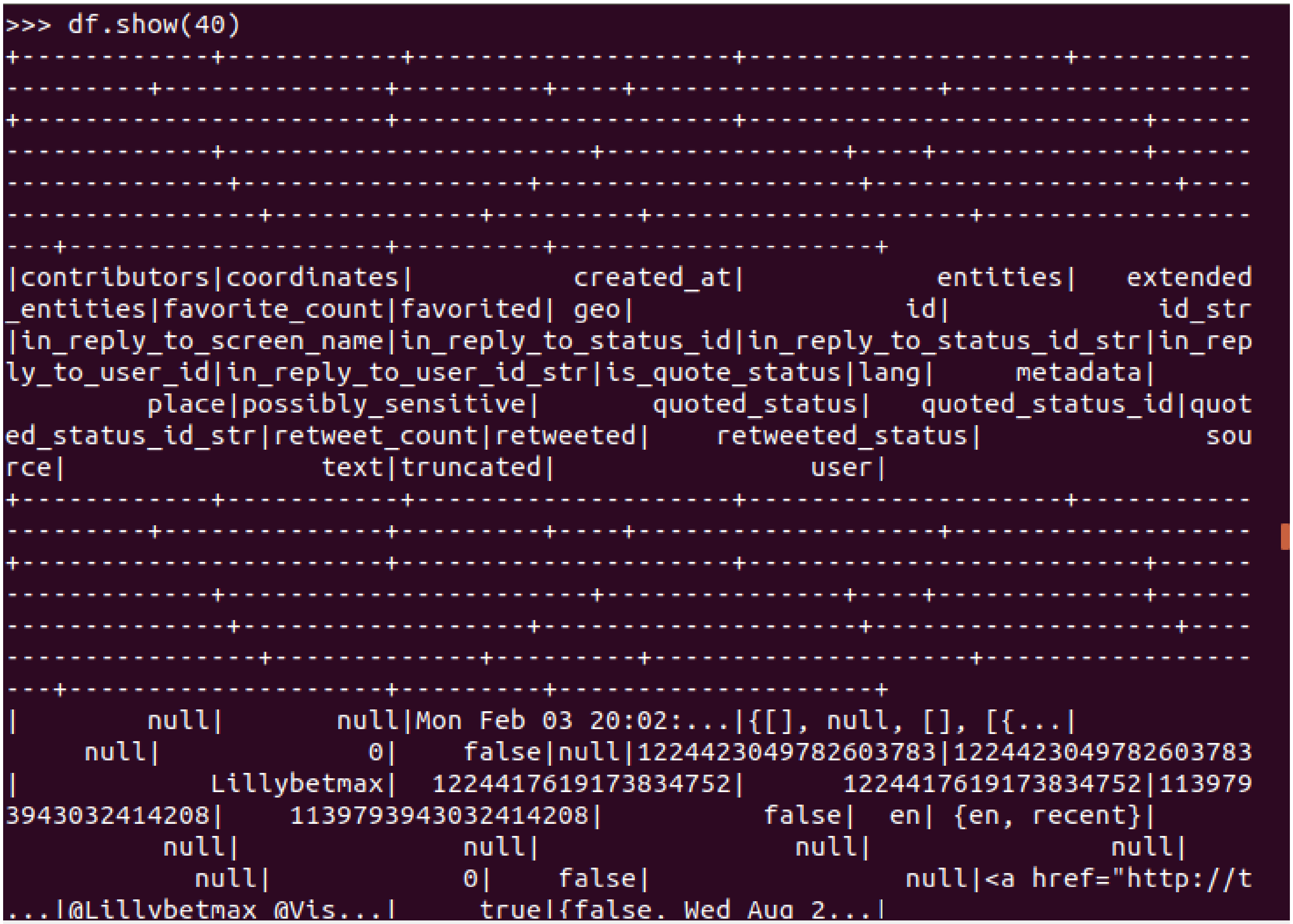
```
>>> df.createOrReplaceTempView("student")
>>> sqlDF = spark.sql("SELECT name, age, course FROM student WHERE age > 21")
>>> sqlDF.show()
+-----+---+----------------+
| name|age|          course|
+-----+---+----------------+
|Helen| 45| MSc Agriculture|
|Alice| 30|            null|
| Adam| 40|            null|
|  Pip| 25|            null|
|  Ian| 35|HND Food Science|
+-----+---+----------------+

>>> exit()
pratik@pratik-VirtualBox:~$ 
```

3.2 Weather.json

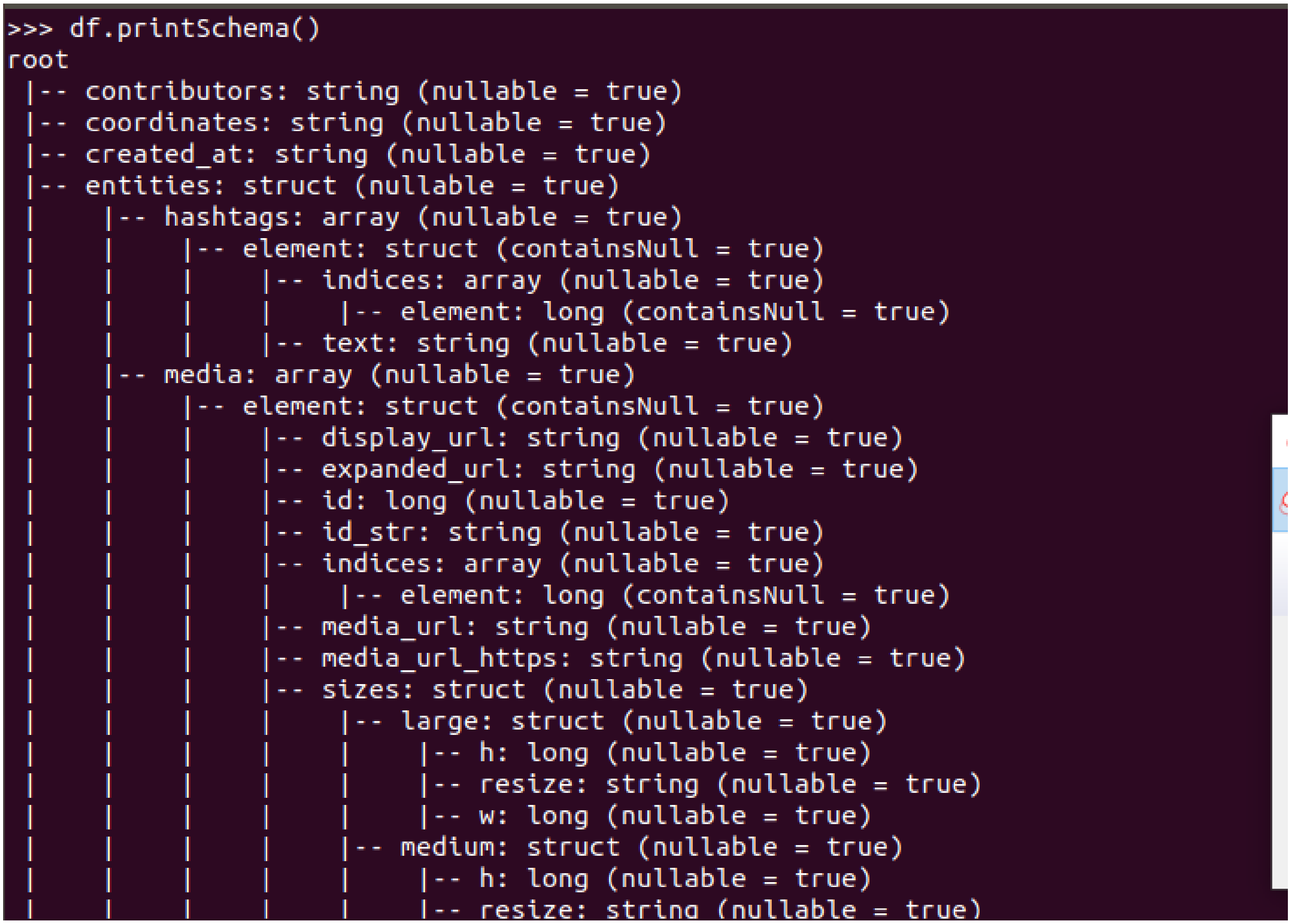
3.2.1 Using Data Frames



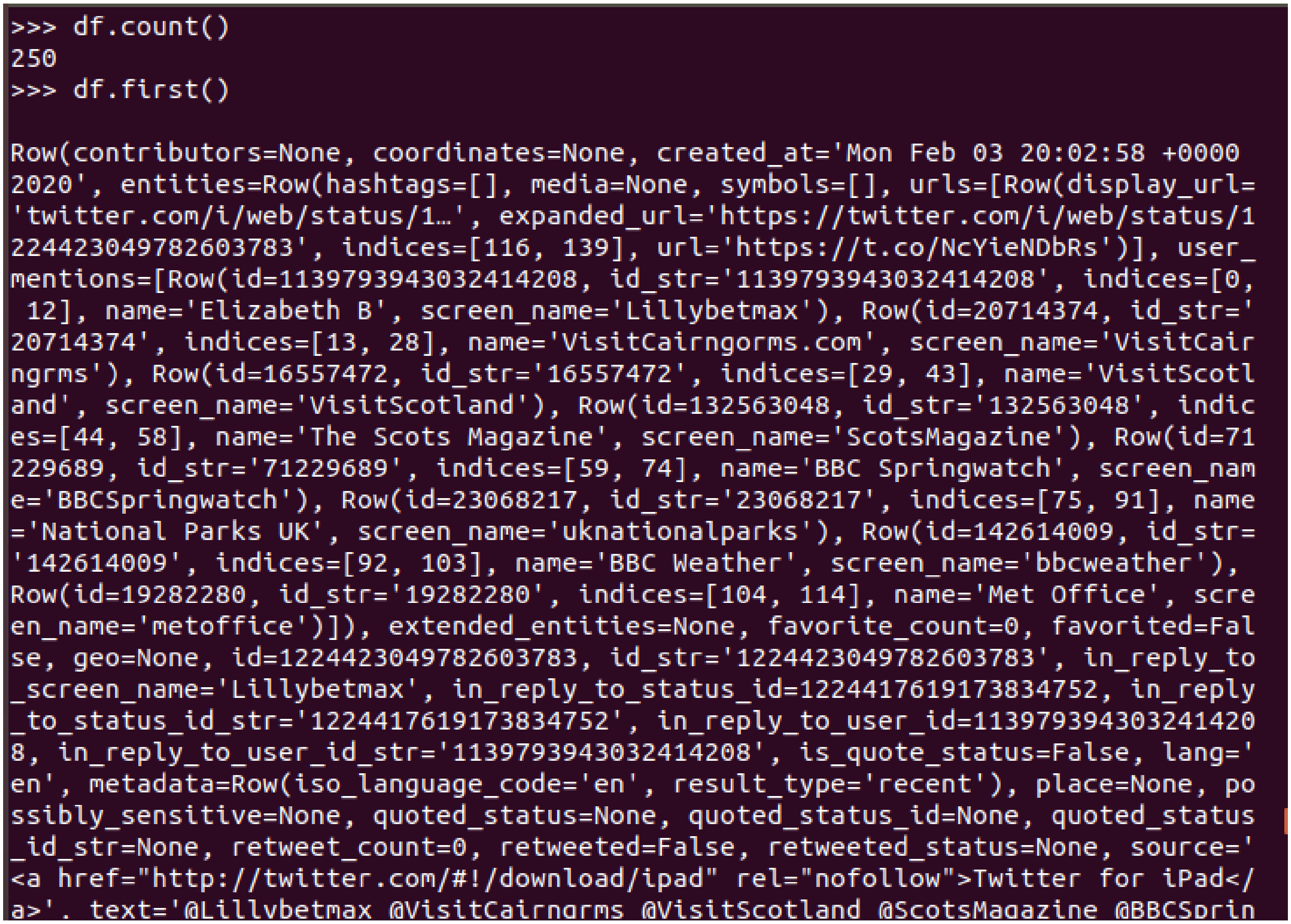
```
>>> df = spark.read.json("weather.json")
>>> df.show()
[Stage 2:>                                                          (0 + 1) / 1

+------------+-----------+--------------------+--------------------+-----------
---------+--------------+---------+----+-------------------+-------------------
+-----------------------+---------------------+-------------------------+------
-------------+-----------------------+---------------+----+-------------+-----+
------------------+--------------------+-------------------+-------------------
-+-------------+---------+--------------------+--------------------+-----------
---------+---------+--------------------+
|contributors|coordinates|          created_at|            entities|   extended
_entities|favorite_count|favorited| geo|                 id|             id_str
|in_reply_to_screen_name|in_reply_to_status_id|in_reply_to_status_id_str|in_rep
ly_to_user_id|in_reply_to_user_id_str|is_quote_status|lang|     metadata|place|
possibly_sensitive|       quoted_status|   quoted_status_id|quoted_status_id_st
r|retweet_count|retweeted|    retweeted_status|              source|
     text|truncated|                user|
+------------+-----------+--------------------+--------------------+-----------
---------+--------------+---------+----+-------------------+-------------------
+-----------------------+---------------------+-------------------------+------
-------------+-----------------------+---------------+----+-------------+-----+
------------------+--------------------+-------------------+-------------------
-+-------------+---------+--------------------+--------------------+-----------
---------+---------+--------------------+
|        null|       null|Mon Feb 03 20:02:...|{[], null, [], [{...|
     null|             0|    false|null|1224423049782603783|1224423049782603783
|            Lillybetmax|  1224417619173834752|      1224417619173834752|113979
3943032414208|    1139793943032414208|          false|  en| {en, recent}| null|
```



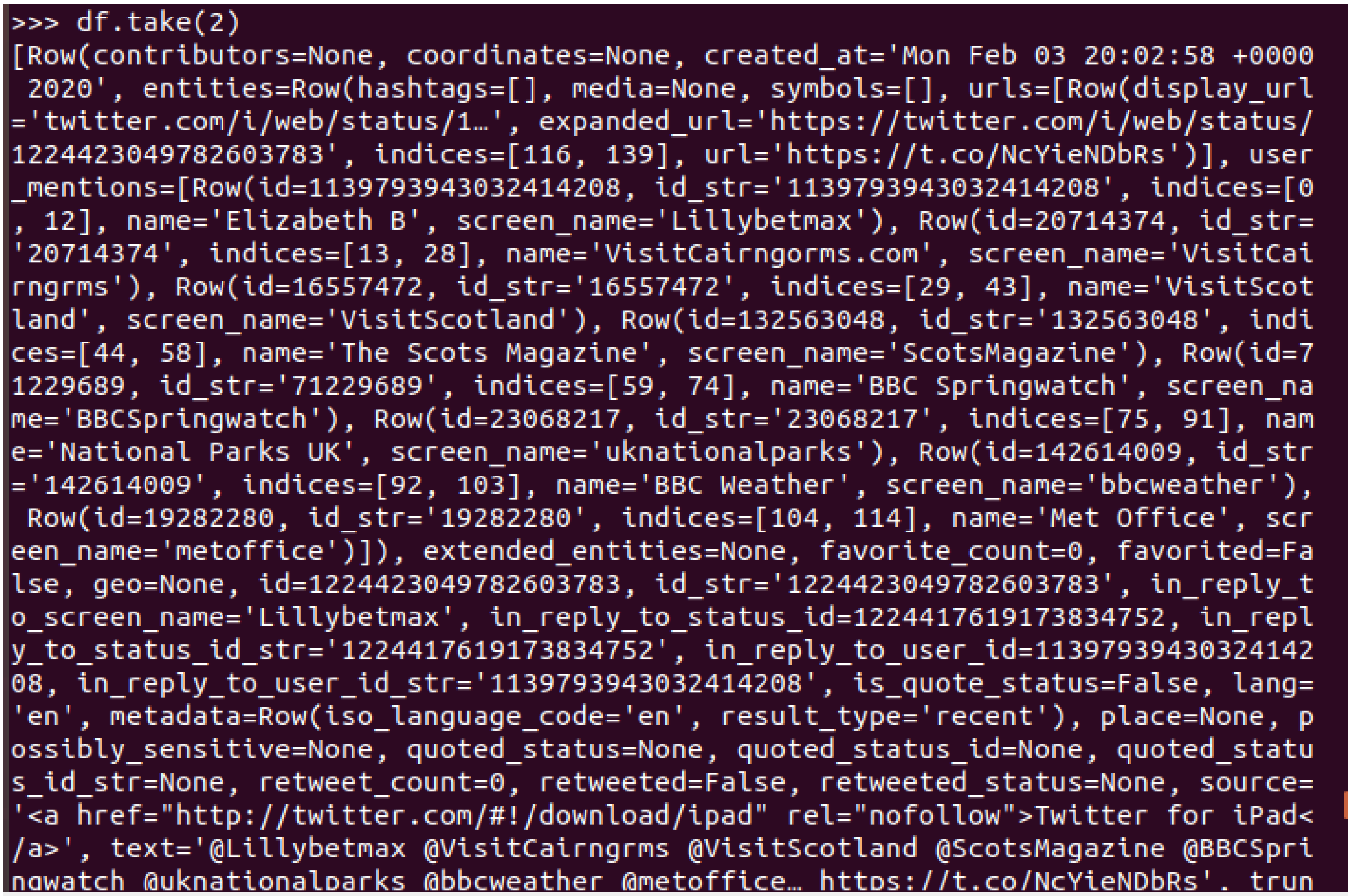
```
>>> df.show(40)
+------------+-----------+--------------------+--------------------+-----------
---------+--------------+---------+----+-------------------+-------------------
+-----------------------+---------------------+-------------------------+------
-------------+-----------------------+---------------+----+-------------+------
--------------+------------------+--------------------+-------------------+----
----------------+-------------+---------+--------------------+-----------------
---+--------------------+---------+--------------------+
|contributors|coordinates|          created_at|            entities|   extended
_entities|favorite_count|favorited| geo|                 id|             id_str
|in_reply_to_screen_name|in_reply_to_status_id|in_reply_to_status_id_str|in_rep
ly_to_user_id|in_reply_to_user_id_str|is_quote_status|lang|     metadata|
         place|possibly_sensitive|       quoted_status|   quoted_status_id|quot
ed_status_id_str|retweet_count|retweeted|    retweeted_status|              sou
rce|                text|truncated|                user|
+------------+-----------+--------------------+--------------------+-----------
---------+--------------+---------+----+-------------------+-------------------
+-----------------------+---------------------+-------------------------+------
-------------+-----------------------+---------------+----+-------------+------
--------------+------------------+--------------------+-------------------+----
----------------+-------------+---------+--------------------+-----------------
---+--------------------+---------+--------------------+
|        null|       null|Mon Feb 03 20:02:...|{[], null, [], [{...|
     null|             0|    false|null|1224423049782603783|1224423049782603783
|           Lillybetmax|  1224417619173834752|      1224417619173834752|113979
3943032414208|    1139793943032414208|          false|  en| {en, recent}|
          null|                null|               null|              null|
            null|            0|    false|                null|<a href="http://t
...|@Lillybetmax @Vis...|     true|{false, Wed Aug 2...|
```



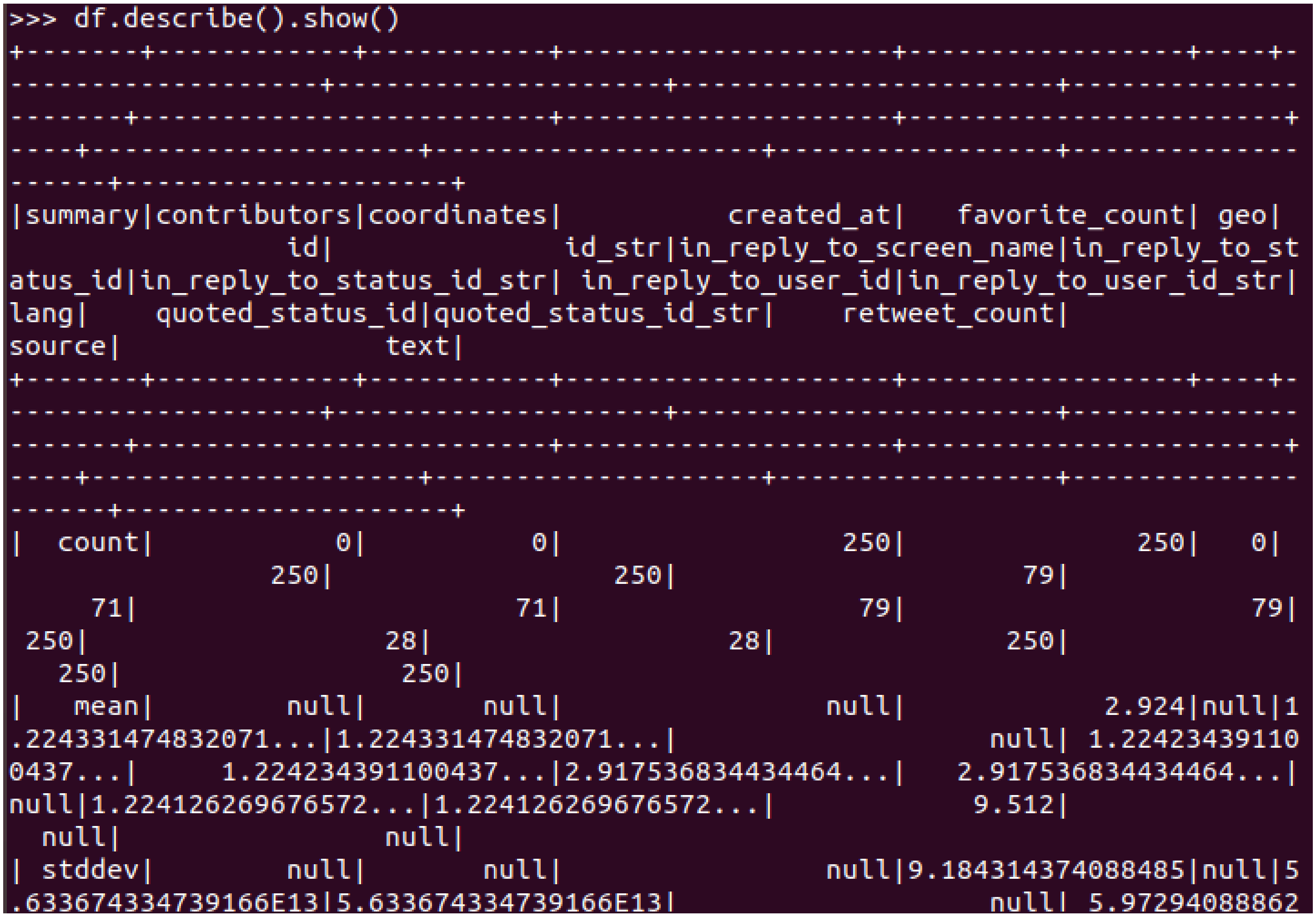
```
>>> df.printSchema()
root
 |-- contributors: string (nullable = true)
 |-- coordinates: string (nullable = true)
 |-- created_at: string (nullable = true)
 |-- entities: struct (nullable = true)
 |    |-- hashtags: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |-- text: string (nullable = true)
 |    |-- media: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |-- id: long (nullable = true)
 |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |-- media_url: string (nullable = true)
 |    |    |    |-- media_url_https: string (nullable = true)
 |    |    |    |-- sizes: struct (nullable = true)
 |    |    |    |    |-- large: struct (nullable = true)
 |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |-- medium: struct (nullable = true)
 |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |-- resize: string (nullable = true)
```



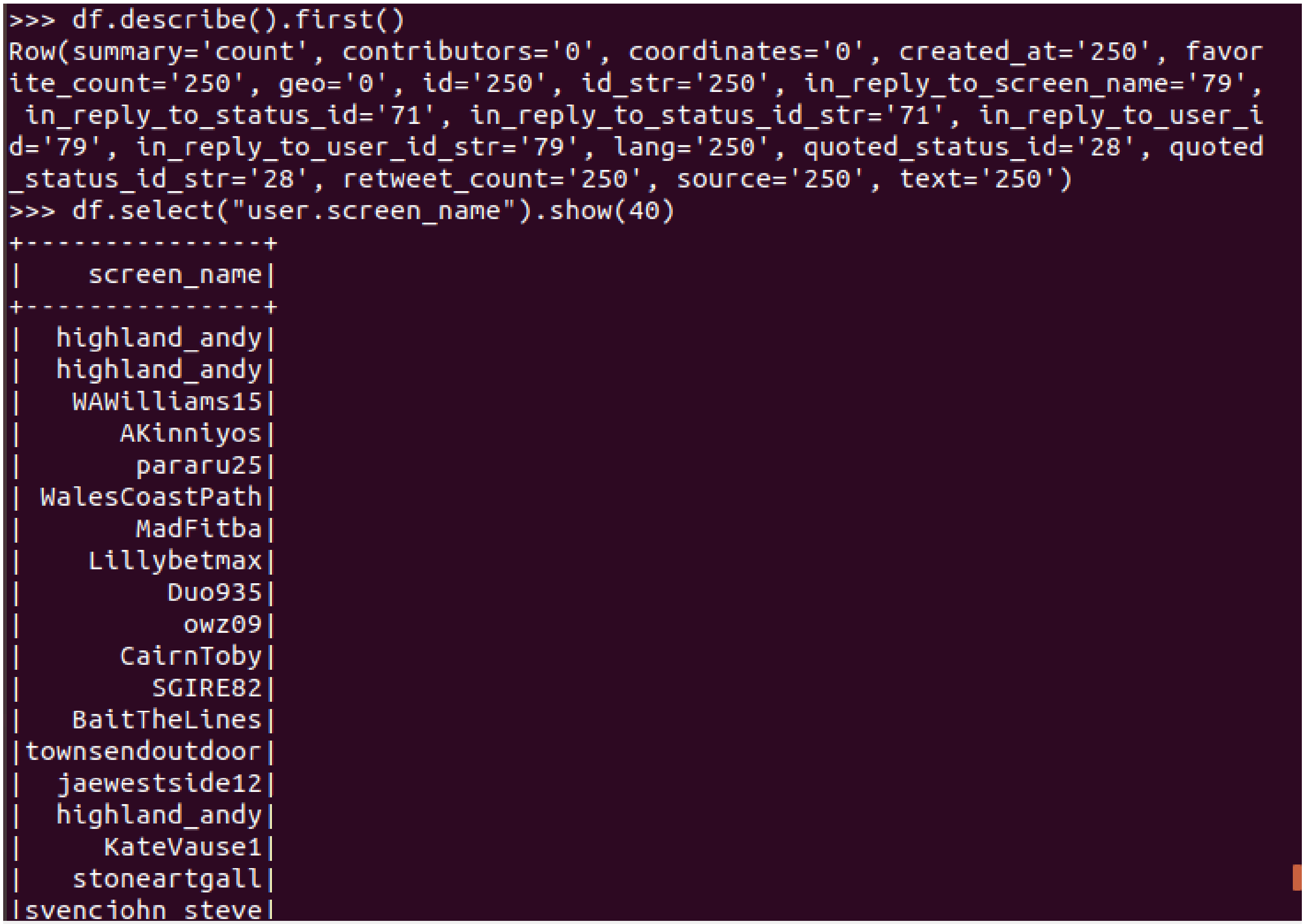
```
>>> df.count()
250
>>> df.first()

Row(contributors=None, coordinates=None, created_at='Mon Feb 03 20:02:58 +0000
2020', entities=Row(hashtags=[], media=None, symbols=[], urls=[Row(display_url=
'twitter.com/i/web/status/1…', expanded_url='https://twitter.com/i/web/status/1
224423049782603783', indices=[116, 139], url='https://t.co/NcYieNDbRs')], user_
mentions=[Row(id=1139793943032414208, id_str='1139793943032414208', indices=[0,
 12], name='Elizabeth B', screen_name='Lillybetmax'), Row(id=20714374, id_str='
20714374', indices=[13, 28], name='VisitCairngorms.com', screen_name='VisitCair
ngrms'), Row(id=16557472, id_str='16557472', indices=[29, 43], name='VisitScotl
and', screen_name='VisitScotland'), Row(id=132563048, id_str='132563048', indic
es=[44, 58], name='The Scots Magazine', screen_name='ScotsMagazine'), Row(id=71
229689, id_str='71229689', indices=[59, 74], name='BBC Springwatch', screen_nam
e='BBCSpringwatch'), Row(id=23068217, id_str='23068217', indices=[75, 91], name
='National Parks UK', screen_name='uknationalparks'), Row(id=142614009, id_str=
'142614009', indices=[92, 103], name='BBC Weather', screen_name='bbcweather'),
Row(id=19282280, id_str='19282280', indices=[104, 114], name='Met Office', scre
en_name='metoffice')]), extended_entities=None, favorite_count=0, favorited=Fal
se, geo=None, id=1224423049782603783, id_str='1224423049782603783', in_reply_to
_screen_name='Lillybetmax', in_reply_to_status_id=1224417619173834752, in_reply
_to_status_id_str='1224417619173834752', in_reply_to_user_id=113979394303241420
8, in_reply_to_user_id_str='1139793943032414208', is_quote_status=False, lang='
en', metadata=Row(iso_language_code='en', result_type='recent'), place=None, po
ssibly_sensitive=None, quoted_status=None, quoted_status_id=None, quoted_status
_id_str=None, retweet_count=0, retweeted=False, retweeted_status=None, source='
<a href="http://twitter.com/#!/download/ipad" rel="nofollow">Twitter for iPad</
a>', text='@Lillybetmax @VisitCairngrms @VisitScotland @ScotsMagazine @BBCSprin
```



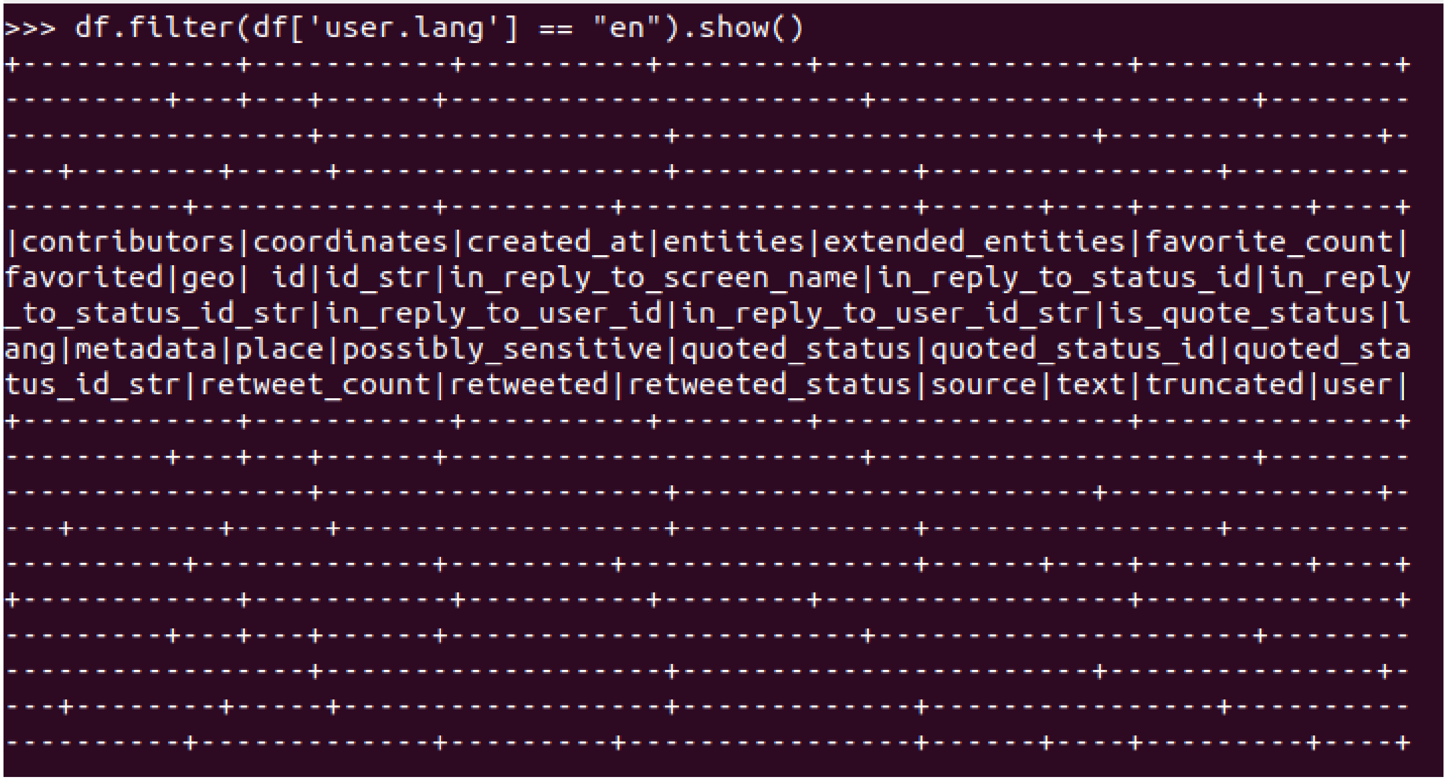
```
>>> df.take(2)
[Row(contributors=None, coordinates=None, created_at='Mon Feb 03 20:02:58 +0000
 2020', entities=Row(hashtags=[], media=None, symbols=[], urls=[Row(display_url
='twitter.com/i/web/status/1…', expanded_url='https://twitter.com/i/web/status/
1224423049782603783', indices=[116, 139], url='https://t.co/NcYieNDbRs')], user
_mentions=[Row(id=1139793943032414208, id_str='1139793943032414208', indices=[0
, 12], name='Elizabeth B', screen_name='Lillybetmax'), Row(id=20714374, id_str=
'20714374', indices=[13, 28], name='VisitCairngorms.com', screen_name='VisitCai
rngrms'), Row(id=16557472, id_str='16557472', indices=[29, 43], name='VisitScot
land', screen_name='VisitScotland'), Row(id=132563048, id_str='132563048', indi
ces=[44, 58], name='The Scots Magazine', screen_name='ScotsMagazine'), Row(id=7
1229689, id_str='71229689', indices=[59, 74], name='BBC Springwatch', screen_na
me='BBCSpringwatch'), Row(id=23068217, id_str='23068217', indices=[75, 91], nam
e='National Parks UK', screen_name='uknationalparks'), Row(id=142614009, id_str
='142614009', indices=[92, 103], name='BBC Weather', screen_name='bbcweather'),
 Row(id=19282280, id_str='19282280', indices=[104, 114], name='Met Office', scr
een_name='metoffice')]), extended_entities=None, favorite_count=0, favorited=Fa
lse, geo=None, id=1224423049782603783, id_str='1224423049782603783', in_reply_t
o_screen_name='Lillybetmax', in_reply_to_status_id=1224417619173834752, in_repl
y_to_status_id_str='1224417619173834752', in_reply_to_user_id=11397939430324142
08, in_reply_to_user_id_str='1139793943032414208', is_quote_status=False, lang=
'en', metadata=Row(iso_language_code='en', result_type='recent'), place=None, p
ossibly_sensitive=None, quoted_status=None, quoted_status_id=None, quoted_statu
s_id_str=None, retweet_count=0, retweeted=False, retweeted_status=None, source=
'<a href="http://twitter.com/#!/download/ipad" rel="nofollow">Twitter for iPad<
/a>', text='@Lillybetmax @VisitCairngrms @VisitScotland @ScotsMagazine @BBCSpri
ngwatch @uknationalparks @bbcweather @metoffice… https://t.co/NcYieNDbRs', trun
```



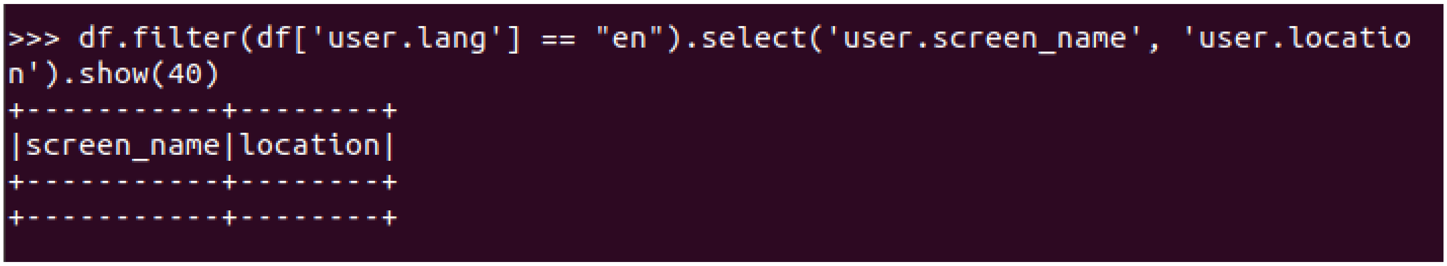
```
>>> df.describe().show()
+-------+------------+-----------+--------------------+-----------------+----+-
-------------------+--------------------+-----------------------+--------------
-------+-------------------------+--------------------+-----------------------+
----+--------------------+--------------------+------------------+-------------
------+--------------------+
|summary|contributors|coordinates|          created_at|   favorite_count| geo| 
                 id|              id_str|in_reply_to_screen_name|in_reply_to_st
atus_id|in_reply_to_status_id_str| in_reply_to_user_id|in_reply_to_user_id_str|
lang|    quoted_status_id|quoted_status_id_str|     retweet_count|             
source|                text|
+-------+------------+-----------+--------------------+-----------------+----+-
-------------------+--------------------+-----------------------+--------------
-------+-------------------------+--------------------+-----------------------+
----+--------------------+--------------------+------------------+-------------
------+--------------------+
|  count|           0|          0|                 250|              250|   0| 
                250|                 250|                     79|              
     71|                       71|                  79|                     79|
 250|                  28|                  28|               250|             
   250|                 250|
|   mean|        null|       null|                null|            2.924|null|1
.224331474832071...|1.224331474832071...|                   null| 1.22423439110
0437...|     1.224234391100437...|2.917536834434464...|   2.917536834434464...|
null|1.224126269676572...|1.224126269676572...|             9.512|             
  null|                null|
| stddev|        null|       null|                null|9.184314374088485|null|5
.633674334739166E13|5.633674334739166E13|                   null| 5.97294088862
```



```
>>> df.describe().first()
Row(summary='count', contributors='0', coordinates='0', created_at='250', favor
ite_count='250', geo='0', id='250', id_str='250', in_reply_to_screen_name='79',
 in_reply_to_status_id='71', in_reply_to_status_id_str='71', in_reply_to_user_i
d='79', in_reply_to_user_id_str='79', lang='250', quoted_status_id='28', quoted
_status_id_str='28', retweet_count='250', source='250', text='250')
>>> df.select("user.screen_name").show(40)
+---------------+
|    screen_name|
+---------------+
|  highland_andy|
|  highland_andy|
|   WAWilliams15|
|      AKinniyos|
|       pararu25|
| WalesCoastPath|
|       MadFitba|
|    Lillybetmax|
|         Duo935|
|          owz09|
|      CairnToby|
|        SGIRE82|
|   BaitTheLines|
|townsendoutdoor|
|  jaewestside12|
|  highland_andy|
|     KateVause1|
|   stoneartgall|
|svenciohn_steve|
```

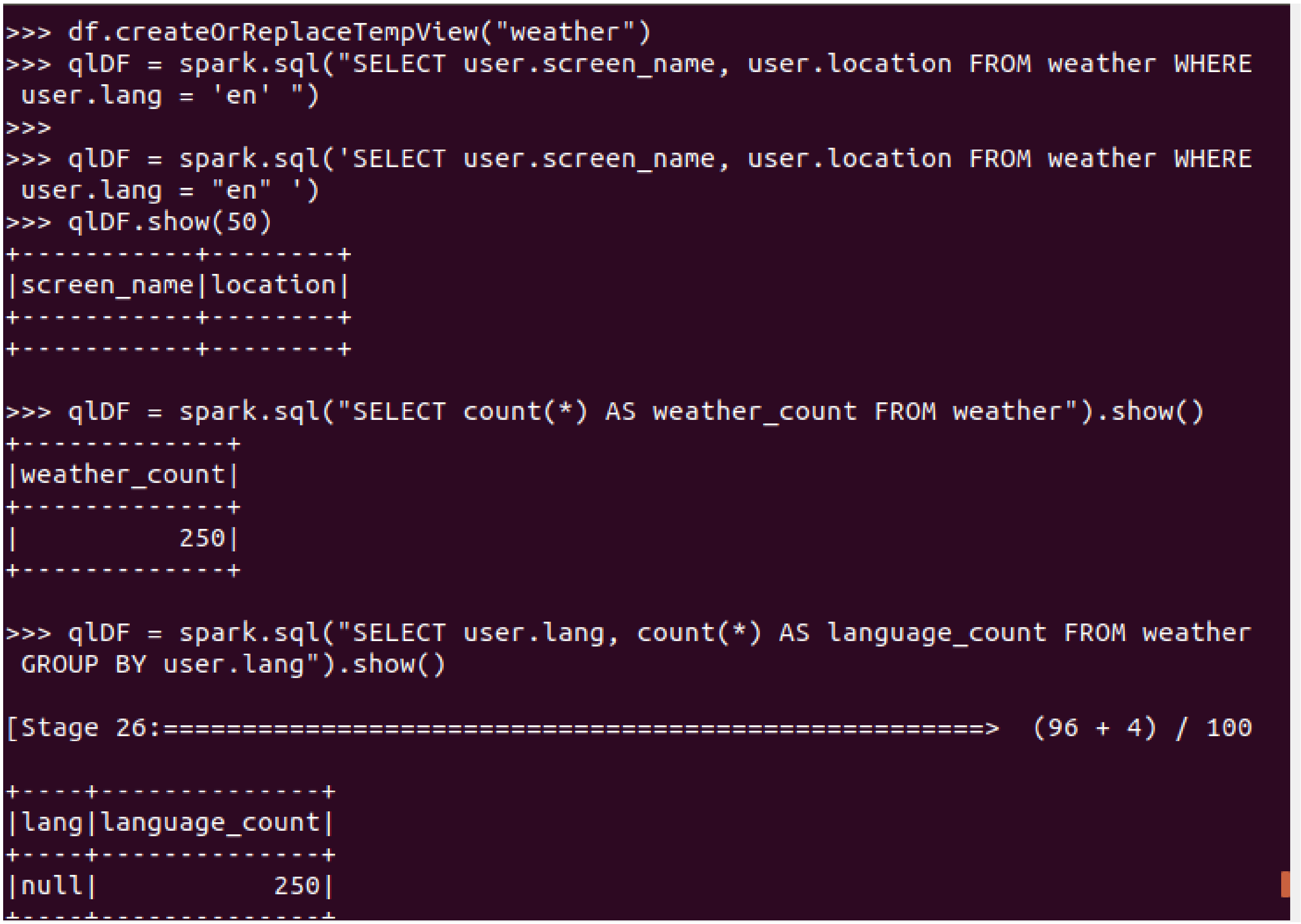


```
>>> df.filter(df['user.lang'] == "en").show()
+------------+-----------+----------+--------+-----------------+--------------+
---------+---+---+------+-----------------------+---------------------+--------
-----------------+-------------------+-----------------------+---------------+-
---+--------+-----+------------------+-------------+----------------+----------
----------+-------------+---------+----------------+------+----+---------+----+
|contributors|coordinates|created_at|entities|extended_entities|favorite_count|
favorited|geo| id|id_str|in_reply_to_screen_name|in_reply_to_status_id|in_reply
_to_status_id_str|in_reply_to_user_id|in_reply_to_user_id_str|is_quote_status|l
ang|metadata|place|possibly_sensitive|quoted_status|quoted_status_id|quoted_sta
tus_id_str|retweet_count|retweeted|retweeted_status|source|text|truncated|user|
+------------+-----------+----------+--------+-----------------+--------------+
---------+---+---+------+-----------------------+---------------------+--------
-----------------+-------------------+-----------------------+---------------+-
---+--------+-----+------------------+-------------+----------------+----------
----------+-------------+---------+----------------+------+----+---------+----+
+------------+-----------+----------+--------+-----------------+--------------+
---------+---+---+------+-----------------------+---------------------+--------
-----------------+-------------------+-----------------------+---------------+-
---+--------+-----+------------------+-------------+----------------+----------
----------+-------------+---------+----------------+------+----+---------+----+
```

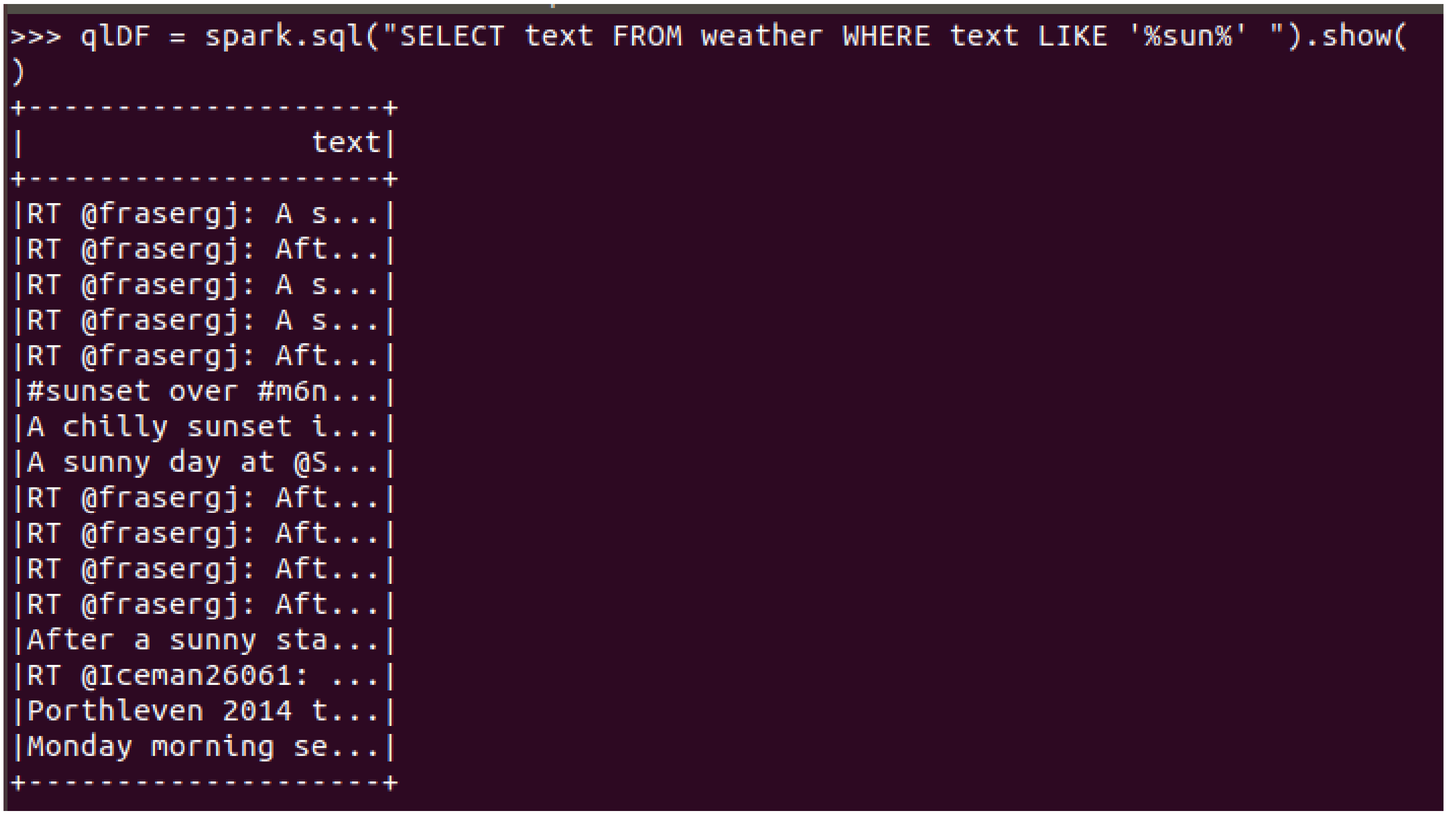


```
>>> df.filter(df['user.lang'] == "en").select('user.screen_name', 'user.locatio
n').show(40)
+-----------+--------+
|screen_name|location|
+-----------+--------+
+-----------+--------+
```

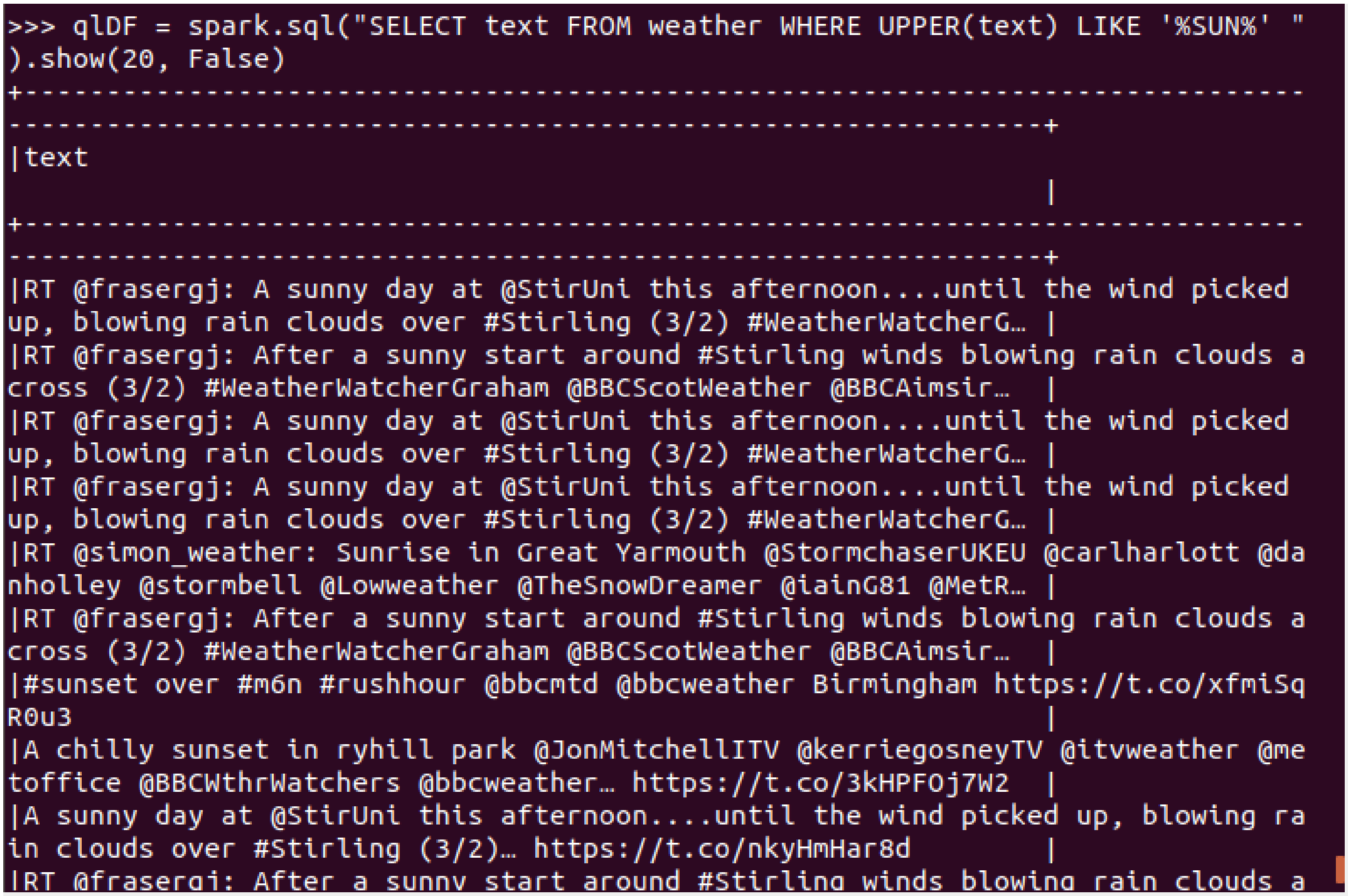
3.2.2 Using SQL



```
>>> df.createOrReplaceTempView("weather")
>>> qlDF = spark.sql("SELECT user.screen_name, user.location FROM weather WHERE
 user.lang = 'en' ")
>>>
>>> qlDF = spark.sql('SELECT user.screen_name, user.location FROM weather WHERE
 user.lang = "en" ')
>>> qlDF.show(50)
+-----------+--------+
|screen_name|location|
+-----------+--------+
+-----------+--------+

>>> qlDF = spark.sql("SELECT count(*) AS weather_count FROM weather").show()
+-------------+
|weather_count|
+-------------+
|          250|
+-------------+

>>> qlDF = spark.sql("SELECT user.lang, count(*) AS language_count FROM weather
 GROUP BY user.lang").show()

[Stage 26:====================================================>  (96 + 4) / 100

+----+--------------+
|lang|language_count|
+----+--------------+
|null|           250|
+----+--------------+
```

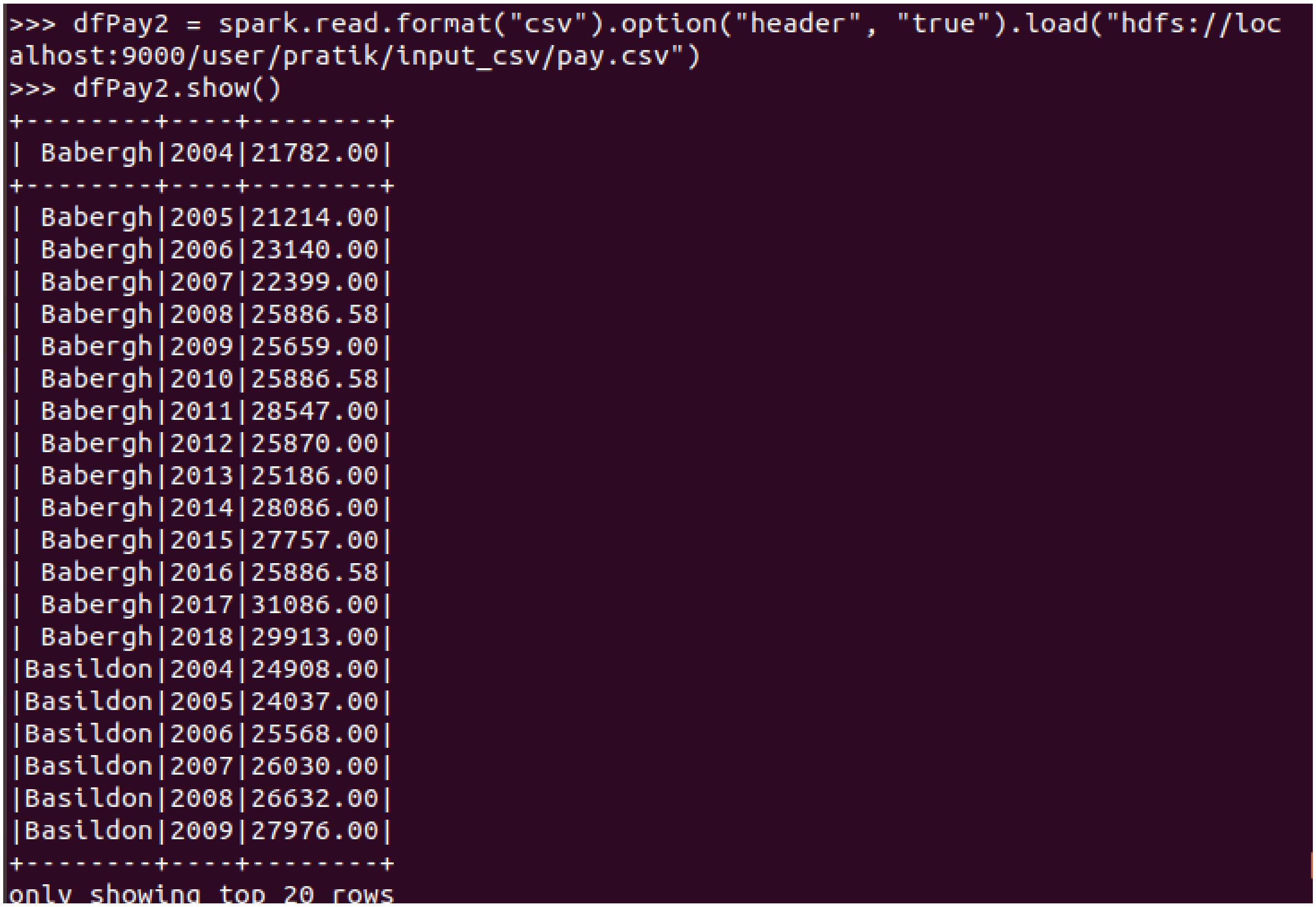


```
>>> qlDF = spark.sql("SELECT text FROM weather WHERE text LIKE '%sun%' ").show(
)
+--------------------+
|                text|
+--------------------+
|RT @frasergj: A s...|
|RT @frasergj: Aft...|
|RT @frasergj: A s...|
|RT @frasergj: A s...|
|RT @frasergj: Aft...|
|#sunset over #m6n...|
|A chilly sunset i...|
|A sunny day at @S...|
|RT @frasergj: Aft...|
|RT @frasergj: Aft...|
|RT @frasergj: Aft...|
|RT @frasergj: Aft...|
|After a sunny sta...|
|RT @Iceman26061: ...|
|Porthleven 2014 t...|
|Monday morning se...|
+--------------------+
```

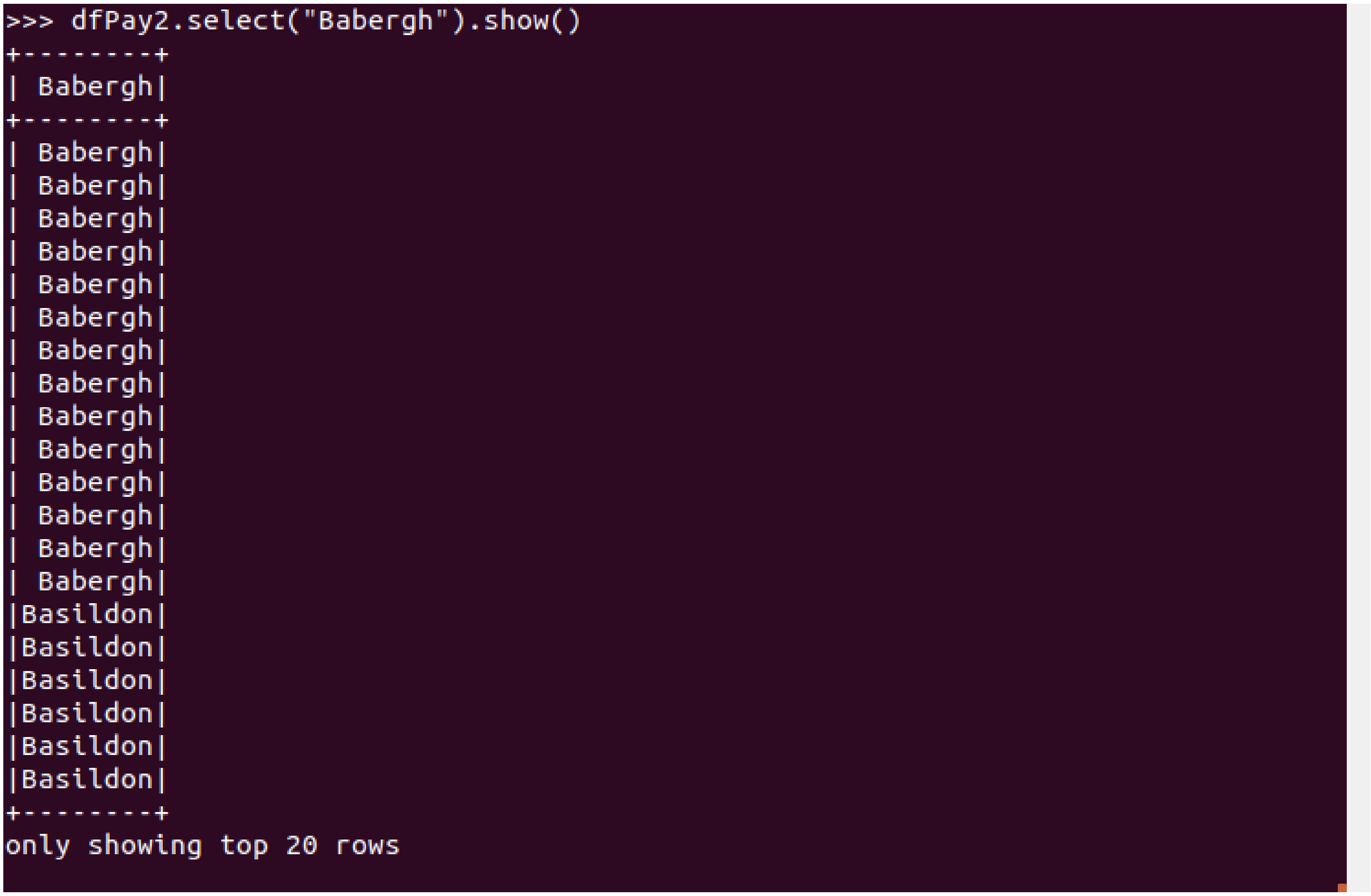


```
>>> qlDF = spark.sql("SELECT text FROM weather WHERE UPPER(text) LIKE '%SUN%' "
).show(20, False)
+-------------------------------------------------------------------------------
---------------------------------------------------------------+
|text
                                                               |
+-------------------------------------------------------------------------------
---------------------------------------------------------------+
|RT @frasergj: A sunny day at @StirUni this afternoon....until the wind picked
up, blowing rain clouds over #Stirling (3/2) #WeatherWatcherG… |
|RT @frasergj: After a sunny start around #Stirling winds blowing rain clouds a
cross (3/2) #WeatherWatcherGraham @BBCScotWeather @BBCAimsir…  |
|RT @frasergj: A sunny day at @StirUni this afternoon....until the wind picked
up, blowing rain clouds over #Stirling (3/2) #WeatherWatcherG… |
|RT @frasergj: A sunny day at @StirUni this afternoon....until the wind picked
up, blowing rain clouds over #Stirling (3/2) #WeatherWatcherG… |
|RT @simon_weather: Sunrise in Great Yarmouth @StormchaserUKEU @carlharlott @da
nholley @stormbell @Lowweather @TheSnowDreamer @iainG81 @MetR… |
|RT @frasergj: After a sunny start around #Stirling winds blowing rain clouds a
cross (3/2) #WeatherWatcherGraham @BBCScotWeather @BBCAimsir…  |
|#sunset over #m6n #rushhour @bbcmtd @bbcweather Birmingham https://t.co/xfmiSq
R0u3                                                           |
|A chilly sunset in ryhill park @JonMitchellITV @kerriegosneyTV @itvweather @me
toffice @BBCWthrWatchers @bbcweather… https://t.co/3kHPFOj7W2  |
|A sunny day at @StirUni this afternoon....until the wind picked up, blowing ra
in clouds over #Stirling (3/2)… https://t.co/nkyHmHar8d        |
|RT @frasergj: After a sunny start around #Stirling winds blowing rain clouds a
```

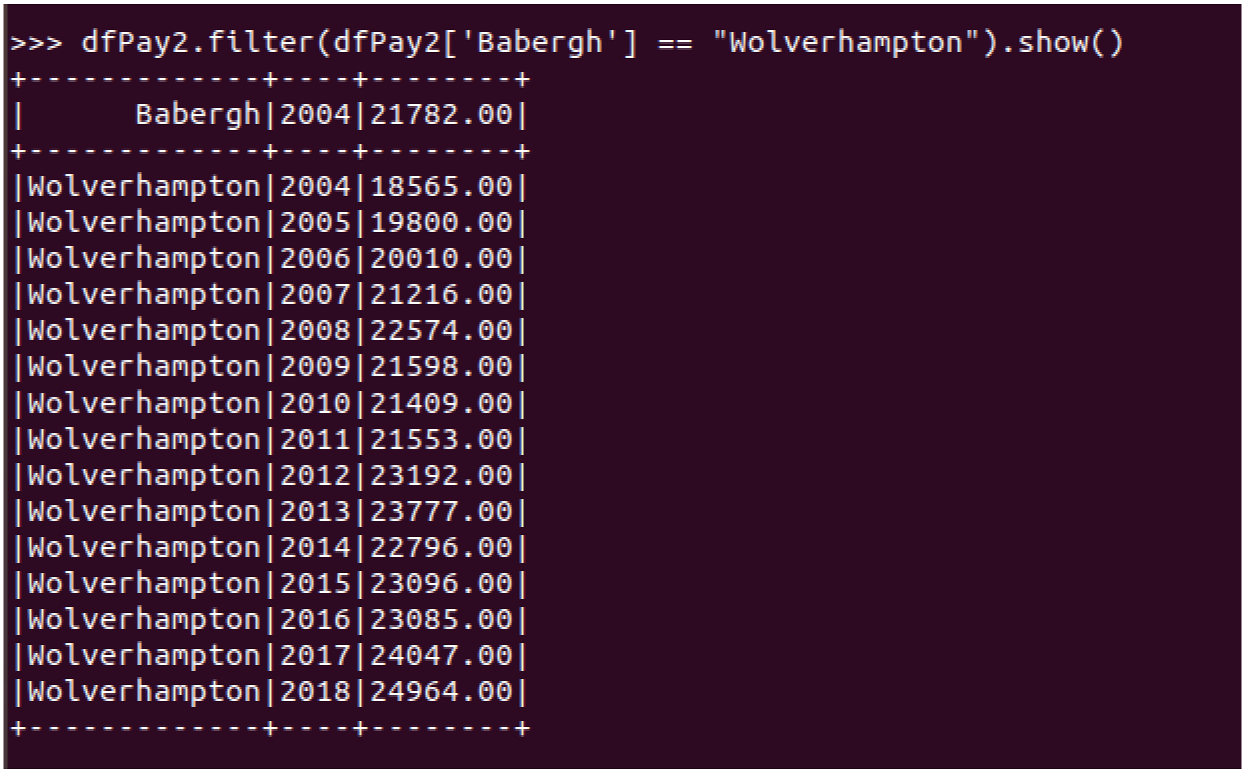
4. HDFS and Apache Spark



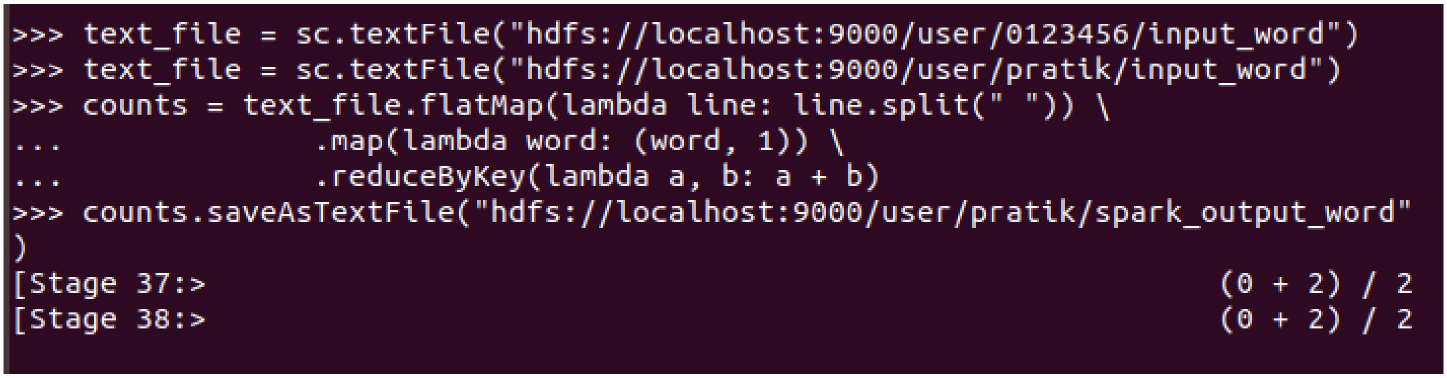
```
>>> dfPay2 = spark.read.format("csv").option("header", "true").load("hdfs://loc
alhost:9000/user/pratik/input_csv/pay.csv")
>>> dfPay2.show()
+--------+----+--------+
| Babergh|2004|21782.00|
+--------+----+--------+
| Babergh|2005|21214.00|
| Babergh|2006|23140.00|
| Babergh|2007|22399.00|
| Babergh|2008|25886.58|
| Babergh|2009|25659.00|
| Babergh|2010|25886.58|
| Babergh|2011|28547.00|
| Babergh|2012|25870.00|
| Babergh|2013|25186.00|
| Babergh|2014|28086.00|
| Babergh|2015|27757.00|
| Babergh|2016|25886.58|
| Babergh|2017|31086.00|
| Babergh|2018|29913.00|
|Basildon|2004|24908.00|
|Basildon|2005|24037.00|
|Basildon|2006|25568.00|
|Basildon|2007|26030.00|
|Basildon|2008|26632.00|
|Basildon|2009|27976.00|
+--------+----+--------+
only showing top 20 rows
```



```
>>> dfPay2.select("Babergh").show()
+--------+
| Babergh|
+--------+
| Babergh|
| Babergh|
| Babergh|
| Babergh|
| Babergh|
| Babergh|
| Babergh|
| Babergh|
| Babergh|
| Babergh|
| Babergh|
| Babergh|
| Babergh|
| Babergh|
|Basildon|
|Basildon|
|Basildon|
|Basildon|
|Basildon|
|Basildon|
+--------+
only showing top 20 rows
```



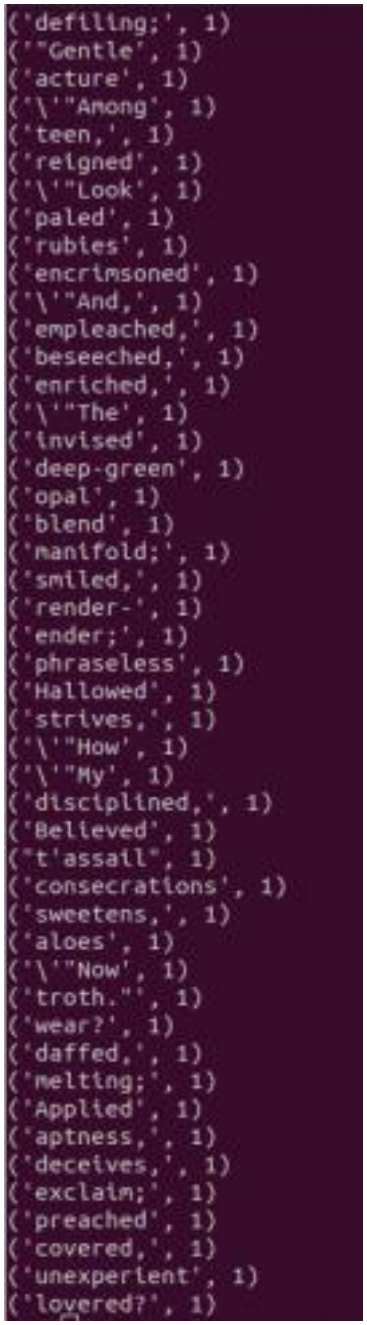
```
>>> dfPay2.filter(dfPay2['Babergh'] == "Wolverhampton").show()
+-------------+----+--------+
|      Babergh|2004|21782.00|
+-------------+----+--------+
|Wolverhampton|2004|18565.00|
|Wolverhampton|2005|19800.00|
|Wolverhampton|2006|20010.00|
|Wolverhampton|2007|21216.00|
|Wolverhampton|2008|22574.00|
|Wolverhampton|2009|21598.00|
|Wolverhampton|2010|21409.00|
|Wolverhampton|2011|21553.00|
|Wolverhampton|2012|23192.00|
|Wolverhampton|2013|23777.00|
|Wolverhampton|2014|22796.00|
|Wolverhampton|2015|23096.00|
|Wolverhampton|2016|23085.00|
|Wolverhampton|2017|24047.00|
|Wolverhampton|2018|24964.00|
+-------------+----+--------+
```



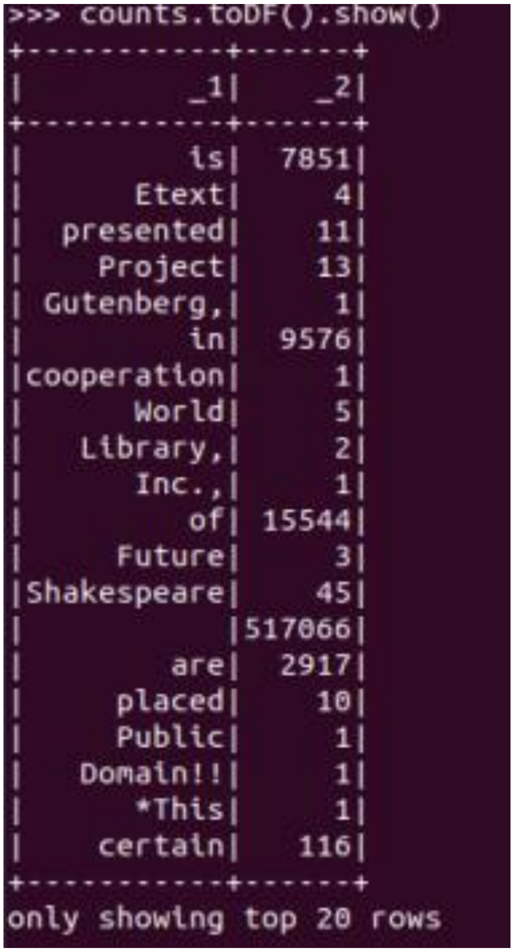
```
>>> text_file = sc.textFile("hdfs://localhost:9000/user/0123456/input_word")
>>> text_file = sc.textFile("hdfs://localhost:9000/user/pratik/input_word")
>>> counts = text_file.flatMap(lambda line: line.split(" ")) \
...              .map(lambda word: (word, 1)) \
...              .reduceByKey(lambda a, b: a + b)
>>> counts.saveAsTextFile("hdfs://localhost:9000/user/pratik/spark_output_word"
)
[Stage 37:>                                                          (0 + 2) / 2
[Stage 38:>                                                          (0 + 2) / 2
```

For x in counts.collect:

Print(x)

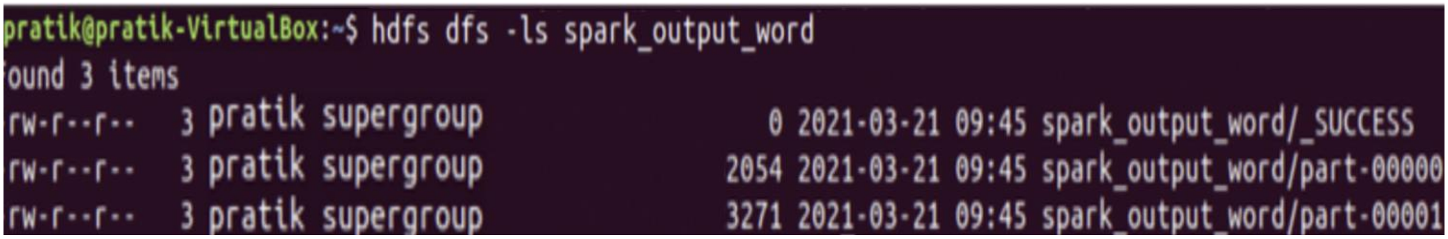


```
('defiling;', 1)
('"Gentle', 1)
('acture', 1)
('\'"Among', 1)
('teen,', 1)
('reigned', 1)
('\'"Look', 1)
('paled', 1)
('rubies', 1)
('encrimsoned', 1)
('\'"And,', 1)
('empleached,', 1)
('beseeched,', 1)
('enriched,', 1)
('\'"The', 1)
('invised', 1)
('deep-green', 1)
('opal', 1)
('blend', 1)
('manifold;', 1)
('smiled,', 1)
('render-', 1)
('ender;', 1)
('phraseless', 1)
('Hallowed', 1)
('strives,', 1)
('\'"How', 1)
('\'"My', 1)
('disciplined,', 1)
('Believed', 1)
("t'assail", 1)
('consecrations', 1)
('sweetens,', 1)
('aloes', 1)
('\'"Now', 1)
('troth."', 1)
('wear?', 1)
('daffed,', 1)
('melting;', 1)
('Applied', 1)
('aptness,', 1)
('deceives,', 1)
('exclaim;', 1)
('preached', 1)
('covered,', 1)
('unexperient', 1)
('lovered?', 1)
```



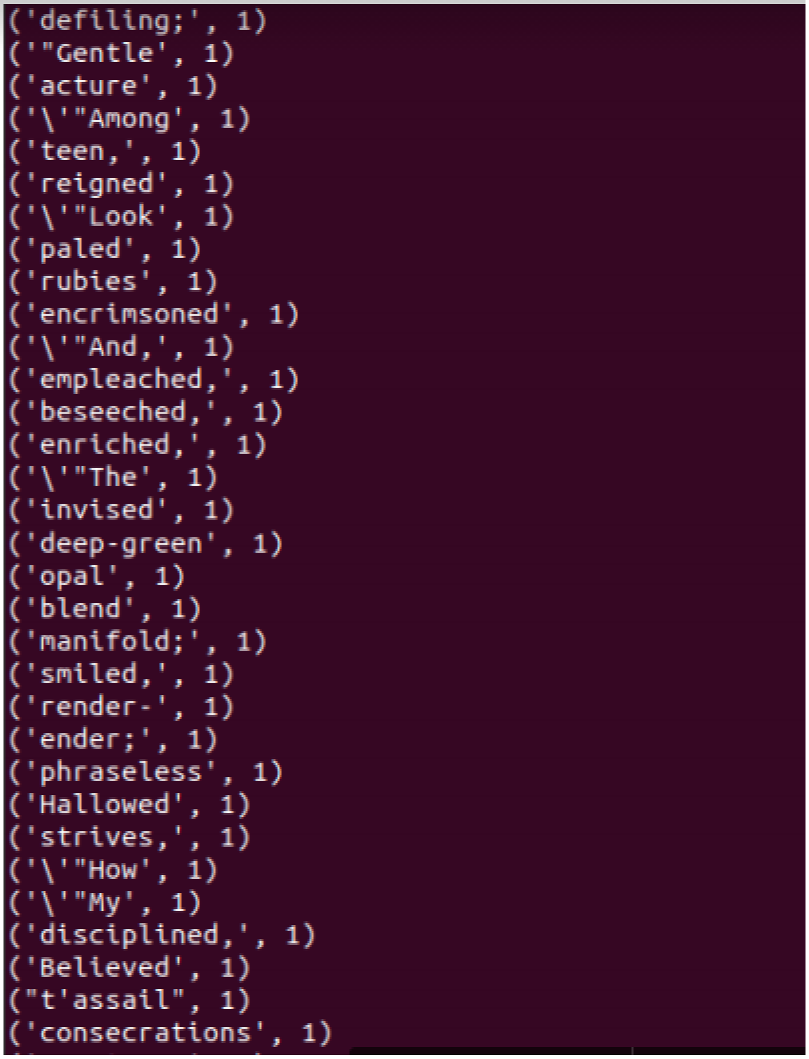
```
>>> counts.toDF().show()
+-----------+------+
|         _1|    _2|
+-----------+------+
|         is|  7851|
|      Etext|     4|
|  presented|    11|
|    Project|    13|
| Gutenberg,|     1|
|         in|  9576|
|cooperation|     1|
|      World|     5|
|   Library,|     2|
|      Inc.,|     1|
|         of| 15544|
|     Future|     3|
|Shakespeare|    45|
|           |517066|
|        are|  2917|
|     placed|    10|
|     Public|     1|
|   Domain!!|     1|
|      *This|     1|
|    certain|   116|
+-----------+------+
only showing top 20 rows
```

In hdfs dfs:



```
pratik@pratik-VirtualBox:~$ hdfs dfs -ls spark_output_word
ound 3 items
rw-r--r--   3 pratik supergroup          0 2021-03-21 09:45 spark_output_word/_SUCCESS
rw-r--r--   3 pratik supergroup       2054 2021-03-21 09:45 spark_output_word/part-00000
rw-r--r--   3 pratik supergroup       3271 2021-03-21 09:45 spark_output_word/part-00001
```

hdfs dfs -cat spark_output_word/part-00001



```
('defiling;', 1)
('"Gentle', 1)
('acture', 1)
('\'"Among', 1)
('teen,', 1)
('reigned', 1)
('\'"Look', 1)
('paled', 1)
('rubies', 1)
('encrimsoned', 1)
('\'"And,', 1)
('empleached,', 1)
('beseeched,', 1)
('enriched,', 1)
('\'"The', 1)
('invised', 1)
('deep-green', 1)
('opal', 1)
('blend', 1)
('manifold;', 1)
('smiled,', 1)
('render-', 1)
('ender;', 1)
('phraseless', 1)
('Hallowed', 1)
('strives,', 1)
('\'"How', 1)
('\'"My', 1)
('disciplined,', 1)
('Believed', 1)
("t'assail", 1)
('consecrations', 1)
```