

Big data programming Assignment -5

1. Explanation of the source code

Answer:

```
"""  
A Python program with spark implementation for counting tweets by state.  
Run with:  
spark-submit /home/vbhamidipati1/spark/workspace/TweetsCount.py  
"""  
from __future__ import print_function  
  
# $example on:init_session$  
from pyspark.sql import SparkSession  
# $example off:init_session$  
  
# $example on:schema_inferring$  
from pyspark.sql import Row  
# $example off:schema_inferring$  
  
# $example on:programmatic_schema$  
# Import data types  
from pyspark.sql.types import *  
# $example off:programmatic_schema$  
  
def count_tweets_state(spark):  
    # spark is an existing SparkSession  
    tweetsDF = spark.read.json("/home/vbhamidipati1/spark/workspace/data/tweets.j  
son")  
    # Displays the content of the DataFrame to stdout  
    tweetsDF.show()  
    # +-----+-----+-----+  
    # |          geo|          tweet|   user|  
    # +-----+-----+-----+  
    # |    Atlanta| It is a sunny day!|   Bob|  
    # |    Athens|We have a footbal...| Susan|  
    # |    Atlanta|   Today is cold.| David|  
    # |    Auburn|I love Auburn Uni...|  Lisa|  
    # | Birmingham|I will go to Atla...|   Ben|  
    # |San Francisco|We watch a movie ...|  Paul|  
    # |   San Diego|It is hot today. ...| Smith|  
    # |Log Angeles|Oscar ceremony is...| Ethan|  
    # |Log Angeles|I love Oscar cere...| Emma|  
    # |   Orlando|I will go to the ...|Rolando|  
    # |    Miami|          Sunny Day!|   Mia|  
    # +-----+-----+-----+  
    # Dataframe describing the tweets json
```

Big data programming Assignment -5

```
cityStateMapDF = spark.read.json("/home/vbhamidipati1/spark/workspace/data/ci
tyStateMap.json")
# Displays the content of the DataFrame to stdout
cityStateMapDF.show()
# +-----+-----+
# |      city|    state|
# +-----+-----+
# |    Atlanta|  Georgia|
# |    Athens|  Georgia|
# |    Miami|   Florida|
# |   Orlando|  Florida|
# |Birmingham| Alabama|
# |    Auburn|  Alabama|
# |Log Angeles|California|
# |San Francisco|California|
# |   San Diego|California|
# +-----+-----+
# Dataframe describing the cityStateMap json

tweetsMapJoin = tweetsDF.join(cityStateMapDF, tweetsDF.geo == cityStateMapDF.
city).drop('city')
tweetsMapJoin.show()

# +-----+-----+-----+-----+
# |      geo|      tweet|    user|    state|
# +-----+-----+-----+-----+
# |    Atlanta| It is a sunny day!|    Bob|  Georgia|
# |    Athens|We have a footbal...|   Susan|  Georgia|
# |    Atlanta|    Today is cold.|   David|  Georgia|
# |    Auburn|I love Auburn Uni...|    Lisa| Alabama|
# |Birmingham|I will go to Atla...|    Ben| Alabama|
# |San Francisco|We watch a movie ...|   Paul|California|
# |   San Diego|It is hot today. ...|  Smith|California|
# |Log Angeles|Oscar ceremony is...|   Ethan|California|
# |Log Angeles|I love Oscar cere...|   Emma|California|
# |    Orlando|I will go to the ...|Rolando|  Florida|
# |    Miami|    Sunny Day!|    Mia|  Florida|
# +-----+-----+-----+-----+
# Dataframe describing the join of tweets and map jsons

# We calculate the count of tweets for state using GROUP BY
tweetsMapJoin.createGlobalTempView("tweetsMapJoin")
spark.sql("SELECT state, count(tweet) as count FROM global_temp.tweetsMapJoin
GROUP BY state").show()
```

Big data programming Assignment -5

```
# +-----+-----+
# |      state|count|
# +-----+-----+
# |   Georgia|    3|
# |  Alabama|    2|
# |California|    4|
# |   Florida|    2|
# +-----+-----+
```

```
if __name__ == "__main__":
    # $example on:init_session$
    spark = SparkSession \
        .builder \
        .appName("Count Number of tweets based on state") \
        .getOrCreate()
    # $example off:init_session$

    count_tweets_state(spark)

    spark.stop()
```

2. Screenshots of the output. Since we plan to use Dataframe in Spark, it is easy to type in "DF.show()" to visualize the table in the terminal. Please do so and take a screenshot of the output in the terminal. The screenshot "output.PNG" of the output in my VM is given. You can use it to verify your outputs.

Answer:

Big data programming Assignment -5

The image shows a Visual Studio Code editor window with a Python script named `TweetsCount.py` open. The script is designed to read a JSON file of tweets and count them by state. The terminal output shows the successful execution of the script, including the registration of Spark components and the final output of the tweet counts.

```
15
16 # $example on:programmatic_schema$
17 # Import data types
18 from pyspark.sql.types import *
19 # $example off:programmatic_schema$
20
21
22 def count_tweets_state(spark):
23     # spark is an existing SparkSession
24     tweetsDF = spark.read.json("/home/vbhamidipati1/spark/workspace/data/tweets.json")
25     # Displays the content of the DataFrame to stdout
26     tweetsDF.show()
27     # +-----+-----+-----+
28     # | geo | tweet | user |
29     # +-----+-----+-----+
30     # | Atlanta | It is a sunny day! | Bob |
31     # | Athens | We have a footbal... | Susan |
32     # | Atlanta | Today is cold. | David |
33     # | Auburn | I love Auburn Uni... | Lisa |
34     # | Birmingham | I will go to Atla... | Ben |
35     # | San Francisco | We watch a movie ... | Paul |
36     # | San Diego | It is hot today. ... | Smith |
37     # | Log Angeles | Oscar ceremony is... | Ethan |
38     # | Log Angeles | I love Oscar cere... | Emma |
39     # | Orlando | I will go to the ... | Rolando |
40     # | Miami | Sunny Day! | Mia |
```

The terminal output shows the following logs:

```
20/04/08 13:17:44 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
20/04/08 13:17:45 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-eed7ac8c-1174-4bb8-b599-b3d17af3703
e
20/04/08 13:17:45 INFO MemoryStore: MemoryStore started with capacity 413.9 MiB
20/04/08 13:17:45 INFO SparkEnv: Registering OutputCommitCoordinator
20/04/08 13:17:45 INFO Utils: Successfully started service 'SparkUI' on port 4040.
20/04/08 13:17:45 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://192.168.122.108:4040
```

The terminal output also shows the execution of the script:

```
driver) (1/1)
20/04/08 13:17:54 INFO DAGScheduler: ResultStage 1 (showString at NativeMethodAccessorImpl.java:0) finished in 0.231
s
20/04/08 13:17:54 INFO DAGScheduler: Job 1 is finished. Cancelling potential speculative or zombie tasks for this job
20/04/08 13:17:54 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
20/04/08 13:17:54 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
20/04/08 13:17:54 INFO DAGScheduler: Job 1 finished: showString at NativeMethodAccessorImpl.java:0, took 0.248912 s
20/04/08 13:17:55 INFO CodeGenerator: Code generated in 47.432273 ms
```

The final output of the script is a table of tweet counts by state:

geo	tweet	user
Atlanta	It is a sunny day!	Bob
Athens	We have a footbal...	Susan
Atlanta	Today is cold.	David
Auburn	I love Auburn Uni...	Lisa
Birmingham	I will go to Atla...	Ben
San Francisco	We watch a movie ...	Paul
San Diego	It is hot today. ...	Smith
Log Angeles	Oscar ceremony is...	Ethan
Log Angeles	I love Oscar cere...	Emma
Orlando	I will go to the ...	Rolando
Miami	Sunny Day!	Mia

The terminal output also shows the following logs:

```
20/04/08 13:17:55 INFO FileSourceStrategy: Pruning directories with:
20/04/08 13:17:55 INFO FileSourceStrategy: Pushed Filters:
20/04/08 13:17:55 INFO FileSourceStrategy: Post-Scan Filters:
20/04/08 13:17:55 INFO FileSourceStrategy: Output Data Schema: struct<value: string>
20/04/08 13:17:55 INFO MemoryStore: Block broadcast_5 stored as values in memory (estimated size 174.2 KiB, free 413.
```

Big data programming Assignment -5

Visual Studio Code interface showing the execution of a PySpark program. The Explorer pane on the left shows the project structure with files like PageRank.py, TweetsCount.py, and WordCount.py. The Editor pane shows the code for TweetsCount.py, which includes comments for example on/off and a function to count tweets by state. The Terminal pane shows the execution output, including logs from TaskSchedulerImpl, DAGScheduler, and CodeGenerator, followed by a table of tweet counts by state.

```
15
16 # $example on:programmatic_schema$
17 # Import data types
18 from pyspark.sql.types import *
19 # $example off:programmatic_schema$
20
21
22 def count_tweets_state(spark):
```

20/04/08 13:17:55 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
20/04/08 13:17:55 INFO TaskSchedulerImpl: Killing all running tasks in stage 3: Stage finished
20/04/08 13:17:55 INFO DAGScheduler: Job 3 finished: showString at NativeMethodAccessorImpl.java:0, took 0.125097 s
20/04/08 13:17:55 INFO CodeGenerator: Code generated in 32.332183 ms

city	state
Atlanta	Georgia
Athens	Georgia
Miami	Florida
Orlando	Florida
Birmingham	Alabama
Auburn	Alabama
Log Angeles	California
San Francisco	California
San Diego	California

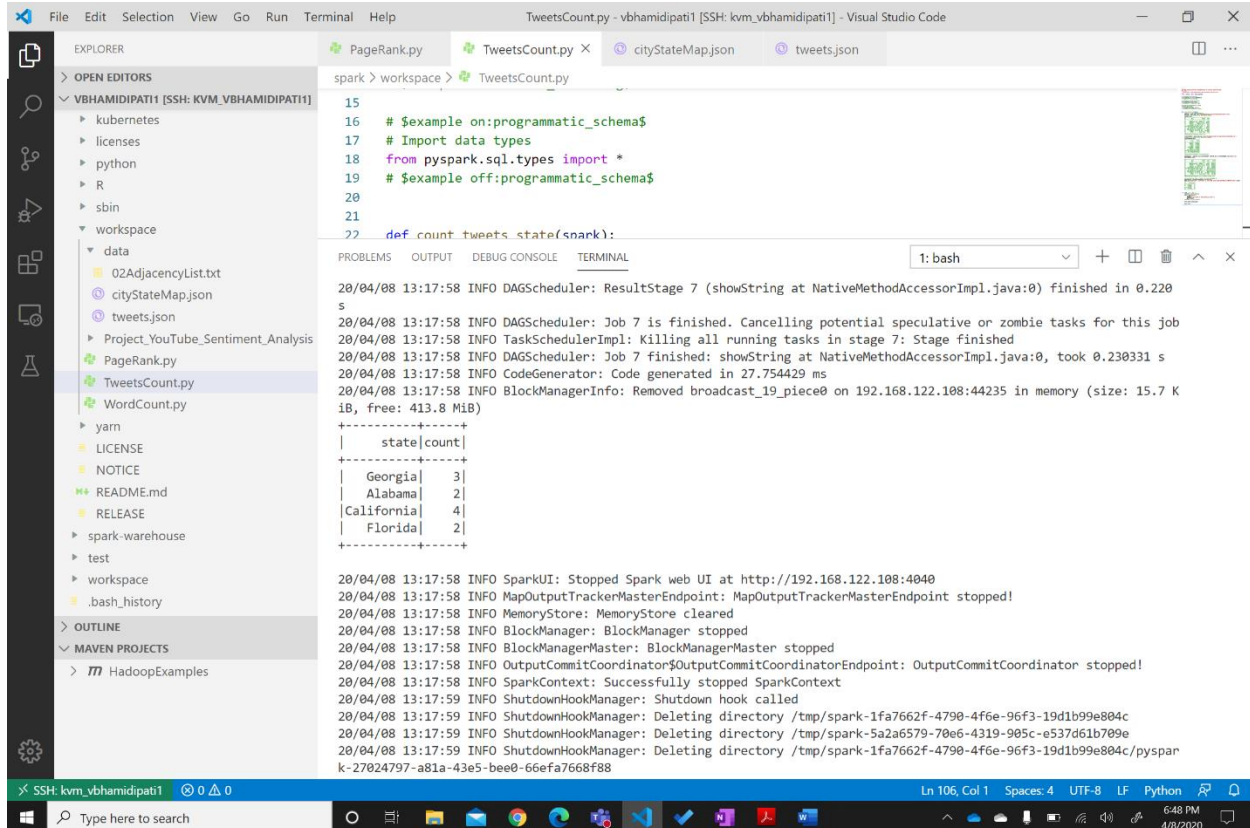
20/04/08 13:17:56 INFO BlockManagerInfo: Removed broadcast_9_piece0 on 192.168.122.108:44235 in memory (size: 5.3 KiB, free: 413.8 MiB)
20/04/08 13:17:56 INFO V2ScanRelationPushDown: Pushing operators to json file:/home/vbhamidipati1/spark/workspace/data/tweets.json
Pushed Filters:
Post-Scan Filters: isNotNull(geo#7)
Output: geo#7, tweet#8, user#9
20/04/08 13:17:56 INFO V2ScanRelationPushDown: Pushing operators to json file:/home/vbhamidipati1/spark/workspace/data/cityStateMap.json
Pushed Filters:

```
$
20/04/08 13:17:57 INFO DAGScheduler: Job 5 is finished. Cancelling potential speculative or zombie tasks for this job
20/04/08 13:17:57 INFO TaskSchedulerImpl: Killing all running tasks in stage 5: Stage finished
20/04/08 13:17:57 INFO DAGScheduler: Job 5 finished: showString at NativeMethodAccessorImpl.java:0, took 0.063541 s
20/04/08 13:17:57 INFO CodeGenerator: Code generated in 30.776623 ms
```

geo	tweet	user	state
Atlanta	It is a sunny day!	Bob	Georgia
Athens	We have a footbal...	Susan	Georgia
Atlanta	Today is cold.	David	Georgia
Auburn	I love Auburn Uni...	Lisa	Alabama
Birmingham	I will go to Atla...	Ben	Alabama
San Francisco	We watch a movie ...	Paul	California
San Diego	It is hot today. ...	Smith	California
Log Angeles	Oscar ceremony is...	Ethan	California
Log Angeles	I love Oscar cere...	Emma	California
Orlando	I will go to the ...	Rolando	Florida
Miami	Sunny Day!	Mia	Florida

20/04/08 13:17:57 INFO BlockManagerInfo: Removed broadcast_14_piece0 on 192.168.122.108:44235 in memory (size: 6.2 KiB, free: 413.9 MiB)
20/04/08 13:17:57 INFO V2ScanRelationPushDown: Pushing operators to json file:/home/vbhamidipati1/spark/workspace/data/tweets.json
Pushed Filters:
Post-Scan Filters: isNotNull(geo#7)
Output: geo#7, tweet#8

Big data programming Assignment -5



```
15
16 # $example on:programmatic_schema$
17 # Import data types
18 from pyspark.sql.types import *
19 # $example off:programmatic_schema$
20
21
22 def count_tweets(state(spark):
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
```

```
20/04/08 13:17:58 INFO DAGScheduler: ResultStage 7 (showString at NativeMethodAccessorImpl.java:0) finished in 0.220 s
20/04/08 13:17:58 INFO DAGScheduler: Job 7 is finished. Cancelling potential speculative or zombie tasks for this job
20/04/08 13:17:58 INFO TaskSchedulerImpl: Killing all running tasks in stage 7: Stage finished
20/04/08 13:17:58 INFO DAGScheduler: Job 7 finished: showString at NativeMethodAccessorImpl.java:0, took 0.230331 s
20/04/08 13:17:58 INFO CodeGenerator: Code generated in 27.754429 ms
20/04/08 13:17:58 INFO BlockManagerInfo: Removed broadcast_19_piece0 on 192.168.122.108:44235 in memory (size: 15.7 K iB, free: 413.8 MiB)
+-----+-----+
| state|count|
+-----+-----+
| Georgia| 3|
| Alabama| 2|
| California| 4|
| Florida| 2|
+-----+-----+
20/04/08 13:17:58 INFO SparkUI: Stopped Spark web UI at http://192.168.122.108:4040
20/04/08 13:17:58 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
20/04/08 13:17:58 INFO MemoryStore: MemoryStore cleared
20/04/08 13:17:58 INFO BlockManager: BlockManager stopped
20/04/08 13:17:58 INFO BlockManagerMaster: BlockManagerMaster stopped
20/04/08 13:17:58 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
20/04/08 13:17:58 INFO SparkContext: Successfully stopped SparkContext
20/04/08 13:17:59 INFO ShutdownHookManager: Shutdown hook called
20/04/08 13:17:59 INFO ShutdownHookManager: Deleting directory /tmp/spark-1fa7662f-4790-4f6e-96f3-19d1b99e804c
20/04/08 13:17:59 INFO ShutdownHookManager: Deleting directory /tmp/spark-5a2a6579-70e6-4319-905c-e537d61b709e
20/04/08 13:17:59 INFO ShutdownHookManager: Deleting directory /tmp/spark-1fa7662f-4790-4f6e-96f3-19d1b99e804c/pyspar
k-27024797-a81a-43e5-bee0-66efa7668f88
```

3. Explain your results. Does your implementation give the right answer?

Answer: Since I have performed an inner join My implementation gave exact right answers as expected.

```
vbhamidipati1@kvm_vbhamidipati1:~$ spark-submit
/home/vbhamidipati1/spark/workspace/TweetsCount.py
20/04/08 13:17:42 WARN Utils: Your hostname, kvm_vbhamidipati1 resolves to a loopback address:
127.0.1.1; using 192.168.122.108 instead (on interface ens3)
20/04/08 13:17:42 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform
(file:/home/vbhamidipati1/spark/jars/spark-unsafe_2.12-3.0.0-preview2.jar) to constructor
java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/04/08 13:17:43 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform...
using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
20/04/08 13:17:44 INFO SparkContext: Running Spark version 3.0.0-preview2
20/04/08 13:17:44 INFO ResourceUtils:
=====
20/04/08 13:17:44 INFO ResourceUtils: Resources for spark.driver:
```

Big data programming Assignment -5

20/04/08 13:17:44 INFO ResourceUtils:

=====

20/04/08 13:17:44 INFO SparkContext: Submitted application: Count Number of tweets based on state

20/04/08 13:17:44 INFO SecurityManager: Changing view acls to: vbhamidipati1

20/04/08 13:17:44 INFO SecurityManager: Changing modify acls to: vbhamidipati1

20/04/08 13:17:44 INFO SecurityManager: Changing view acls groups to:

20/04/08 13:17:44 INFO SecurityManager: Changing modify acls groups to:

20/04/08 13:17:44 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled;
users with view permissions: Set(vbhamidipati1); groups with view permissions: Set(); users with
modify permissions: Set(vbhamidipati1); groups with modify permissions: Set()

20/04/08 13:17:44 INFO Utils: Successfully started service 'sparkDriver' on port 38353.

20/04/08 13:17:44 INFO SparkEnv: Registering MapOutputTracker

20/04/08 13:17:44 INFO SparkEnv: Registering BlockManagerMaster

20/04/08 13:17:44 INFO BlockManagerMasterEndpoint: Using
org.apache.spark.storage.DefaultTopologyMapper for getting topology information

20/04/08 13:17:44 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up

20/04/08 13:17:44 INFO SparkEnv: Registering BlockManagerMasterHeartbeat

20/04/08 13:17:45 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-eed7ac8c-1174-
4bb8-b599-b3d17af3703e

20/04/08 13:17:45 INFO MemoryStore: MemoryStore started with capacity 413.9 MiB

20/04/08 13:17:45 INFO SparkEnv: Registering OutputCommitCoordinator

20/04/08 13:17:45 INFO Utils: Successfully started service 'SparkUI' on port 4040.

20/04/08 13:17:45 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://192.168.122.108:4040

20/04/08 13:17:45 INFO Executor: Starting executor ID driver on host 192.168.122.108

20/04/08 13:17:46 INFO Utils: Successfully started service

'org.apache.spark.network.netty.NettyBlockTransferService' on port 44235.

20/04/08 13:17:46 INFO NettyBlockTransferService: Server created on 192.168.122.108:44235

20/04/08 13:17:46 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy
for block replication policy

20/04/08 13:17:46 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver,
192.168.122.108, 44235, None)

20/04/08 13:17:46 INFO BlockManagerMasterEndpoint: Registering block manager
192.168.122.108:44235 with 413.9 MiB RAM, BlockManagerId(driver, 192.168.122.108, 44235, None)

20/04/08 13:17:46 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver,
192.168.122.108, 44235, None)

20/04/08 13:17:46 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver,
192.168.122.108, 44235, None)

20/04/08 13:17:46 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of
spark.sql.warehouse.dir ('file:/home/vbhamidipati1/spark-warehouse/').

20/04/08 13:17:46 INFO SharedState: Warehouse path is 'file:/home/vbhamidipati1/spark-warehouse/'.

20/04/08 13:17:51 INFO FileSourceStrategy: Pruning directories with:

20/04/08 13:17:51 INFO FileSourceStrategy: Pushed Filters:

20/04/08 13:17:51 INFO FileSourceStrategy: Post-Scan Filters:

20/04/08 13:17:51 INFO FileSourceStrategy: Output Data Schema: struct<value: string>

20/04/08 13:17:51 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size
174.2 KiB, free 413.8 MiB)

20/04/08 13:17:52 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated
size 27.5 KiB, free 413.7 MiB)

Big data programming Assignment -5

20/04/08 13:17:52 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on 192.168.122.108:44235 (size: 27.5 KiB, free: 413.9 MiB)

20/04/08 13:17:52 INFO SparkContext: Created broadcast 0 from json at NativeMethodAccessorImpl.java:0

20/04/08 13:17:52 INFO FileSourceScanExec: Planning scan with bin packing, max size: 4195117 bytes, open cost is considered as scanning 4194304 bytes.

20/04/08 13:17:52 INFO SparkContext: Starting job: json at NativeMethodAccessorImpl.java:0

20/04/08 13:17:52 INFO DAGScheduler: Got job 0 (json at NativeMethodAccessorImpl.java:0) with 1 output partitions

20/04/08 13:17:52 INFO DAGScheduler: Final stage: ResultStage 0 (json at NativeMethodAccessorImpl.java:0)

20/04/08 13:17:52 INFO DAGScheduler: Parents of final stage: List()

20/04/08 13:17:52 INFO DAGScheduler: Missing parents: List()

20/04/08 13:17:52 INFO DAGScheduler: Submitting ResultStage 0 (MapPartitionsRDD[3] at json at NativeMethodAccessorImpl.java:0), which has no missing parents

20/04/08 13:17:52 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 11.0 KiB, free 413.7 MiB)

20/04/08 13:17:52 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 5.7 KiB, free 413.7 MiB)

20/04/08 13:17:52 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on 192.168.122.108:44235 (size: 5.7 KiB, free: 413.9 MiB)

20/04/08 13:17:52 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1206

20/04/08 13:17:52 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 0 (MapPartitionsRDD[3] at json at NativeMethodAccessorImpl.java:0) (first 15 tasks are for partitions Vector(0))

20/04/08 13:17:52 INFO TaskSchedulerImpl: Adding task set 0.0 with 1 tasks

20/04/08 13:17:52 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, 192.168.122.108, executor driver, partition 0, PROCESS_LOCAL, 7756 bytes)

20/04/08 13:17:52 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)

20/04/08 13:17:53 INFO FileScanRDD: Reading File path: file:///home/vbhamidipati1/spark/workspace/data/tweets.json, range: 0-813, partition values: [empty row]

20/04/08 13:17:53 INFO CodeGenerator: Code generated in 510.615494 ms

20/04/08 13:17:54 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1989 bytes result sent to driver

20/04/08 13:17:54 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 1166 ms on 192.168.122.108 (executor driver) (1/1)

20/04/08 13:17:54 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool

20/04/08 13:17:54 INFO DAGScheduler: ResultStage 0 (json at NativeMethodAccessorImpl.java:0) finished in 1.476 s

20/04/08 13:17:54 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job

20/04/08 13:17:54 INFO TaskSchedulerImpl: Killing all running tasks in stage 0: Stage finished

20/04/08 13:17:54 INFO DAGScheduler: Job 0 finished: json at NativeMethodAccessorImpl.java:0, took 1.600429 s

20/04/08 13:17:54 INFO V2ScanRelationPushDown: Pushing operators to json file:/home/vbhamidipati1/spark/workspace/data/tweets.json

Pushed Filters:

Big data programming Assignment -5

Post-Scan Filters:

Output: geo#7, tweet#8, user#9

```
20/04/08 13:17:54 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size
174.2 KiB, free 413.5 MiB)
20/04/08 13:17:54 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated
size 27.5 KiB, free 413.5 MiB)
20/04/08 13:17:54 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on
192.168.122.108:44235 (size: 27.5 KiB, free: 413.9 MiB)
20/04/08 13:17:54 INFO SparkContext: Created broadcast 2 from showString at
NativeMethodAccessorImpl.java:0
20/04/08 13:17:54 INFO MemoryStore: Block broadcast_3 stored as values in memory (estimated size
174.2 KiB, free 413.3 MiB)
20/04/08 13:17:54 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated
size 27.5 KiB, free 413.3 MiB)
20/04/08 13:17:54 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on
192.168.122.108:44235 (size: 27.5 KiB, free: 413.8 MiB)
20/04/08 13:17:54 INFO SparkContext: Created broadcast 3 from showString at
NativeMethodAccessorImpl.java:0
20/04/08 13:17:54 INFO CodeGenerator: Code generated in 48.35184 ms
20/04/08 13:17:54 INFO SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0
20/04/08 13:17:54 INFO DAGScheduler: Got job 1 (showString at NativeMethodAccessorImpl.java:0)
with 1 output partitions
20/04/08 13:17:54 INFO DAGScheduler: Final stage: ResultStage 1 (showString at
NativeMethodAccessorImpl.java:0)
20/04/08 13:17:54 INFO DAGScheduler: Parents of final stage: List()
20/04/08 13:17:54 INFO DAGScheduler: Missing parents: List()
20/04/08 13:17:54 INFO DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[7] at showString
at NativeMethodAccessorImpl.java:0), which has no missing parents
20/04/08 13:17:54 INFO MemoryStore: Block broadcast_4 stored as values in memory (estimated size
10.7 KiB, free 413.3 MiB)
20/04/08 13:17:54 INFO MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated
size 5.4 KiB, free 413.3 MiB)
20/04/08 13:17:54 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory on
192.168.122.108:44235 (size: 5.4 KiB, free: 413.8 MiB)
20/04/08 13:17:54 INFO SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:1206
20/04/08 13:17:54 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1
(MapPartitionsRDD[7] at showString at NativeMethodAccessorImpl.java:0) (first 15 tasks are for
partitions Vector(0))
20/04/08 13:17:54 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
20/04/08 13:17:54 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, 192.168.122.108,
executor driver, partition 0, PROCESS_LOCAL, 7925 bytes)
20/04/08 13:17:54 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)
20/04/08 13:17:54 INFO CodeGenerator: Code generated in 38.480591 ms
20/04/08 13:17:54 INFO FilePartitionReader: Reading file path:
file:///home/vbhamidipati1/spark/workspace/data/tweets.json, range: 0-813, partition values: [empty
row]
20/04/08 13:17:54 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 2105 bytes result sent to driver
```

Big data programming Assignment -5

20/04/08 13:17:54 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 177 ms on 192.168.122.108 (executor driver) (1/1)
20/04/08 13:17:54 INFO DAGScheduler: ResultStage 1 (showString at NativeMethodAccessorImpl.java:0) finished in 0.231 s
20/04/08 13:17:54 INFO DAGScheduler: Job 1 is finished. Cancelling potential speculative or zombie tasks for this job
20/04/08 13:17:54 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
20/04/08 13:17:54 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
20/04/08 13:17:54 INFO DAGScheduler: Job 1 finished: showString at NativeMethodAccessorImpl.java:0, took 0.248912 s
20/04/08 13:17:55 INFO CodeGenerator: Code generated in 47.432273 ms

```
+-----+-----+-----+
|   geo|   tweet|  user|
+-----+-----+-----+
| Atlanta| It is a sunny day!| Bob|
| Athens| We have a footbal...| Susan|
| Atlanta| Today is cold.| David|
| Auburn| I love Auburn Uni...| Lisa|
| Birmingham| I will go to Atla...| Ben|
| San Francisco| We watch a movie ...| Paul|
| San Diego| It is hot today. ...| Smith|
| Log Angeles| Oscar ceremony is...| Ethan|
| Log Angeles| I love Oscar cere...| Emma|
| Orlando| I will go to the ...| Rolando|
| Miami| Sunny Day!| Mia|
+-----+-----+-----+
```

20/04/08 13:17:55 INFO FileSourceStrategy: Pruning directories with:
20/04/08 13:17:55 INFO FileSourceStrategy: Pushed Filters:
20/04/08 13:17:55 INFO FileSourceStrategy: Post-Scan Filters:
20/04/08 13:17:55 INFO FileSourceStrategy: Output Data Schema: struct<value: string>
20/04/08 13:17:55 INFO MemoryStore: Block broadcast_5 stored as values in memory (estimated size 174.2 KiB, free 413.1 MiB)
20/04/08 13:17:55 INFO MemoryStore: Block broadcast_5_piece0 stored as bytes in memory (estimated size 27.5 KiB, free 413.1 MiB)
20/04/08 13:17:55 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on 192.168.122.108:44235 (size: 27.5 KiB, free: 413.8 MiB)
20/04/08 13:17:55 INFO SparkContext: Created broadcast 5 from json at NativeMethodAccessorImpl.java:0
20/04/08 13:17:55 INFO FileSourceScanExec: Planning scan with bin packing, max size: 4194694 bytes, open cost is considered as scanning 4194304 bytes.
20/04/08 13:17:55 INFO SparkContext: Starting job: json at NativeMethodAccessorImpl.java:0
20/04/08 13:17:55 INFO DAGScheduler: Got job 2 (json at NativeMethodAccessorImpl.java:0) with 1 output partitions
20/04/08 13:17:55 INFO DAGScheduler: Final stage: ResultStage 2 (json at NativeMethodAccessorImpl.java:0)
20/04/08 13:17:55 INFO DAGScheduler: Parents of final stage: List()

Big data programming Assignment -5

20/04/08 13:17:55 INFO DAGScheduler: Missing parents: List()
20/04/08 13:17:55 INFO DAGScheduler: Submitting ResultStage 2 (MapPartitionsRDD[11] at json at NativeMethodAccessorImpl.java:0), which has no missing parents
20/04/08 13:17:55 INFO MemoryStore: Block broadcast_6 stored as values in memory (estimated size 11.0 KiB, free 413.1 MiB)
20/04/08 13:17:55 INFO MemoryStore: Block broadcast_6_piece0 stored as bytes in memory (estimated size 5.7 KiB, free 413.1 MiB)
20/04/08 13:17:55 INFO BlockManagerInfo: Added broadcast_6_piece0 in memory on 192.168.122.108:44235 (size: 5.7 KiB, free: 413.8 MiB)
20/04/08 13:17:55 INFO SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:1206
20/04/08 13:17:55 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 2 (MapPartitionsRDD[11] at json at NativeMethodAccessorImpl.java:0) (first 15 tasks are for partitions Vector(0))
20/04/08 13:17:55 INFO TaskSchedulerImpl: Adding task set 2.0 with 1 tasks
20/04/08 13:17:55 INFO BlockManagerInfo: Removed broadcast_4_piece0 on 192.168.122.108:44235 in memory (size: 5.4 KiB, free: 413.8 MiB)
20/04/08 13:17:55 INFO TaskSetManager: Starting task 0.0 in stage 2.0 (TID 2, 192.168.122.108, executor driver, partition 0, PROCESS_LOCAL, 7762 bytes)
20/04/08 13:17:55 INFO Executor: Running task 0.0 in stage 2.0 (TID 2)
20/04/08 13:17:55 INFO FileScanRDD: Reading File path: file:///home/vbhamidipati1/spark/workspace/data/cityStateMap.json, range: 0-390, partition values: [empty row]
20/04/08 13:17:55 INFO Executor: Finished task 0.0 in stage 2.0 (TID 2). 1923 bytes result sent to driver
20/04/08 13:17:55 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 44 ms on 192.168.122.108 (executor driver) (1/1)
20/04/08 13:17:55 INFO DAGScheduler: ResultStage 2 (json at NativeMethodAccessorImpl.java:0) finished in 0.079 s
20/04/08 13:17:55 INFO DAGScheduler: Job 2 is finished. Cancelling potential speculative or zombie tasks for this job
20/04/08 13:17:55 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
20/04/08 13:17:55 INFO TaskSchedulerImpl: Killing all running tasks in stage 2: Stage finished
20/04/08 13:17:55 INFO DAGScheduler: Job 2 finished: json at NativeMethodAccessorImpl.java:0, took 0.088071 s
20/04/08 13:17:55 INFO V2ScanRelationPushDown: Pushing operators to json file:/home/vbhamidipati1/spark/workspace/data/cityStateMap.json
Pushed Filters:
Post-Scan Filters:
Output: city#42, state#43

20/04/08 13:17:55 INFO MemoryStore: Block broadcast_7 stored as values in memory (estimated size 174.2 KiB, free 412.9 MiB)
20/04/08 13:17:55 INFO BlockManagerInfo: Removed broadcast_6_piece0 on 192.168.122.108:44235 in memory (size: 5.7 KiB, free: 413.8 MiB)
20/04/08 13:17:55 INFO MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 27.5 KiB, free 412.9 MiB)
20/04/08 13:17:55 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on 192.168.122.108:44235 (size: 27.5 KiB, free: 413.8 MiB)

Big data programming Assignment -5

20/04/08 13:17:55 INFO SparkContext: Created broadcast 7 from showString at NativeMethodAccessorImpl.java:0

20/04/08 13:17:55 INFO MemoryStore: Block broadcast_8 stored as values in memory (estimated size 174.2 KiB, free 412.8 MiB)

20/04/08 13:17:55 INFO BlockManagerInfo: Removed broadcast_5_piece0 on 192.168.122.108:44235 in memory (size: 27.5 KiB, free: 413.8 MiB)

20/04/08 13:17:55 INFO MemoryStore: Block broadcast_8_piece0 stored as bytes in memory (estimated size 27.5 KiB, free 412.9 MiB)

20/04/08 13:17:55 INFO BlockManagerInfo: Added broadcast_8_piece0 in memory on 192.168.122.108:44235 (size: 27.5 KiB, free: 413.8 MiB)

20/04/08 13:17:55 INFO SparkContext: Created broadcast 8 from showString at NativeMethodAccessorImpl.java:0

20/04/08 13:17:55 INFO CodeGenerator: Code generated in 28.216505 ms

20/04/08 13:17:55 INFO SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0

20/04/08 13:17:55 INFO DAGScheduler: Got job 3 (showString at NativeMethodAccessorImpl.java:0) with 1 output partitions

20/04/08 13:17:55 INFO DAGScheduler: Final stage: ResultStage 3 (showString at NativeMethodAccessorImpl.java:0)

20/04/08 13:17:55 INFO DAGScheduler: Parents of final stage: List()

20/04/08 13:17:55 INFO DAGScheduler: Missing parents: List()

20/04/08 13:17:55 INFO DAGScheduler: Submitting ResultStage 3 (MapPartitionsRDD[15] at showString at NativeMethodAccessorImpl.java:0), which has no missing parents

20/04/08 13:17:55 INFO MemoryStore: Block broadcast_9 stored as values in memory (estimated size 10.3 KiB, free 412.9 MiB)

20/04/08 13:17:55 INFO MemoryStore: Block broadcast_9_piece0 stored as bytes in memory (estimated size 5.3 KiB, free 412.9 MiB)

20/04/08 13:17:55 INFO BlockManagerInfo: Added broadcast_9_piece0 in memory on 192.168.122.108:44235 (size: 5.3 KiB, free: 413.8 MiB)

20/04/08 13:17:55 INFO SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:1206

20/04/08 13:17:55 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 3 (MapPartitionsRDD[15] at showString at NativeMethodAccessorImpl.java:0) (first 15 tasks are for partitions Vector(0))

20/04/08 13:17:55 INFO TaskSchedulerImpl: Adding task set 3.0 with 1 tasks

20/04/08 13:17:55 INFO TaskSetManager: Starting task 0.0 in stage 3.0 (TID 3, 192.168.122.108, executor driver, partition 0, PROCESS_LOCAL, 7931 bytes)

20/04/08 13:17:55 INFO Executor: Running task 0.0 in stage 3.0 (TID 3)

20/04/08 13:17:55 INFO CodeGenerator: Code generated in 25.062007 ms

20/04/08 13:17:55 INFO FilePartitionReader: Reading file path: file:///home/vbhamidipati1/spark/workspace/data/cityStateMap.json, range: 0-390, partition values: [empty row]

20/04/08 13:17:55 INFO Executor: Finished task 0.0 in stage 3.0 (TID 3). 1630 bytes result sent to driver

20/04/08 13:17:55 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 84 ms on 192.168.122.108 (executor driver) (1/1)

20/04/08 13:17:55 INFO DAGScheduler: ResultStage 3 (showString at NativeMethodAccessorImpl.java:0) finished in 0.113 s

20/04/08 13:17:55 INFO DAGScheduler: Job 3 is finished. Cancelling potential speculative or zombie tasks for this job

Big data programming Assignment -5

20/04/08 13:17:55 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
20/04/08 13:17:55 INFO TaskSchedulerImpl: Killing all running tasks in stage 3: Stage finished
20/04/08 13:17:55 INFO DAGScheduler: Job 3 finished: showString at NativeMethodAccessorImpl.java:0, took 0.125097 s
20/04/08 13:17:55 INFO CodeGenerator: Code generated in 32.332183 ms

```
+-----+-----+
|   city|   state|
+-----+-----+
| Atlanta| Georgia|
| Athens| Georgia|
| Miami| Florida|
| Orlando| Florida|
| Birmingham| Alabama|
| Auburn| Alabama|
| Log Angeles| California|
| San Francisco| California|
| San Diego| California|
+-----+-----+
```

20/04/08 13:17:56 INFO BlockManagerInfo: Removed broadcast_9_piece0 on 192.168.122.108:44235 in memory (size: 5.3 KiB, free: 413.8 MiB)
20/04/08 13:17:56 INFO V2ScanRelationPushDown:
Pushing operators to json file:/home/vbhamidipati1/spark/workspace/data/tweets.json
Pushed Filters:
Post-Scan Filters: isnotnull(geo#7)
Output: geo#7, tweet#8, user#9

20/04/08 13:17:56 INFO V2ScanRelationPushDown:
Pushing operators to json file:/home/vbhamidipati1/spark/workspace/data/cityStateMap.json
Pushed Filters:
Post-Scan Filters: isnotnull(city#42)
Output: city#42, state#43

20/04/08 13:17:56 INFO MemoryStore: Block broadcast_10 stored as values in memory (estimated size 174.2 KiB, free 412.8 MiB)
20/04/08 13:17:56 INFO MemoryStore: Block broadcast_10_piece0 stored as bytes in memory (estimated size 27.5 KiB, free 412.7 MiB)
20/04/08 13:17:56 INFO BlockManagerInfo: Added broadcast_10_piece0 in memory on 192.168.122.108:44235 (size: 27.5 KiB, free: 413.8 MiB)
20/04/08 13:17:56 INFO SparkContext: Created broadcast 10 from showString at NativeMethodAccessorImpl.java:0
20/04/08 13:17:56 INFO MemoryStore: Block broadcast_11 stored as values in memory (estimated size 174.2 KiB, free 412.6 MiB)
20/04/08 13:17:56 INFO BlockManagerInfo: Removed broadcast_7_piece0 on 192.168.122.108:44235 in memory (size: 27.5 KiB, free: 413.8 MiB)
20/04/08 13:17:56 INFO MemoryStore: Block broadcast_11_piece0 stored as bytes in memory (estimated size 27.5 KiB, free 412.7 MiB)

Big data programming Assignment -5

20/04/08 13:17:56 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on 192.168.122.108:44235 (size: 27.5 KiB, free: 413.8 MiB)

20/04/08 13:17:56 INFO SparkContext: Created broadcast 11 from showString at NativeMethodAccessorImpl.java:0

20/04/08 13:17:56 INFO BlockManagerInfo: Removed broadcast_8_piece0 on 192.168.122.108:44235 in memory (size: 27.5 KiB, free: 413.8 MiB)

20/04/08 13:17:56 INFO CodeGenerator: Code generated in 28.036973 ms

20/04/08 13:17:56 INFO SparkContext: Starting job: call at FutureTask.java:264

20/04/08 13:17:56 INFO DAGScheduler: Got job 4 (call at FutureTask.java:264) with 1 output partitions

20/04/08 13:17:56 INFO DAGScheduler: Final stage: ResultStage 4 (call at FutureTask.java:264)

20/04/08 13:17:56 INFO DAGScheduler: Parents of final stage: List()

20/04/08 13:17:56 INFO DAGScheduler: Missing parents: List()

20/04/08 13:17:56 INFO DAGScheduler: Submitting ResultStage 4 (MapPartitionsRDD[19] at call at FutureTask.java:264), which has no missing parents

20/04/08 13:17:56 INFO MemoryStore: Block broadcast_12 stored as values in memory (estimated size 10.7 KiB, free 412.9 MiB)

20/04/08 13:17:56 INFO MemoryStore: Block broadcast_12_piece0 stored as bytes in memory (estimated size 5.4 KiB, free 412.9 MiB)

20/04/08 13:17:56 INFO BlockManagerInfo: Added broadcast_12_piece0 in memory on 192.168.122.108:44235 (size: 5.4 KiB, free: 413.8 MiB)

20/04/08 13:17:56 INFO SparkContext: Created broadcast 12 from broadcast at DAGScheduler.scala:1206

20/04/08 13:17:56 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 4 (MapPartitionsRDD[19] at call at FutureTask.java:264) (first 15 tasks are for partitions Vector(0))

20/04/08 13:17:56 INFO TaskSchedulerImpl: Adding task set 4.0 with 1 tasks

20/04/08 13:17:56 INFO TaskSetManager: Starting task 0.0 in stage 4.0 (TID 4, 192.168.122.108, executor driver, partition 0, PROCESS_LOCAL, 7931 bytes)

20/04/08 13:17:56 INFO Executor: Running task 0.0 in stage 4.0 (TID 4)

20/04/08 13:17:56 INFO FilePartitionReader: Reading file path: file:///home/vbhamidipati1/spark/workspace/data/cityStateMap.json, range: 0-390, partition values: [empty row]

20/04/08 13:17:56 INFO Executor: Finished task 0.0 in stage 4.0 (TID 4). 1728 bytes result sent to driver

20/04/08 13:17:56 INFO TaskSetManager: Finished task 0.0 in stage 4.0 (TID 4) in 34 ms on 192.168.122.108 (executor driver) (1/1)

20/04/08 13:17:56 INFO TaskSchedulerImpl: Removed TaskSet 4.0, whose tasks have all completed, from pool

20/04/08 13:17:56 INFO DAGScheduler: ResultStage 4 (call at FutureTask.java:264) finished in 0.048 s

20/04/08 13:17:56 INFO DAGScheduler: Job 4 is finished. Cancelling potential speculative or zombie tasks for this job

20/04/08 13:17:56 INFO TaskSchedulerImpl: Killing all running tasks in stage 4: Stage finished

20/04/08 13:17:56 INFO DAGScheduler: Job 4 finished: call at FutureTask.java:264, took 0.054570 s

20/04/08 13:17:56 INFO BlockManagerInfo: Removed broadcast_12_piece0 on 192.168.122.108:44235 in memory (size: 5.4 KiB, free: 413.8 MiB)

20/04/08 13:17:56 INFO MemoryStore: Block broadcast_13 stored as values in memory (estimated size 16.0 MiB, free 396.9 MiB)

20/04/08 13:17:56 INFO MemoryStore: Block broadcast_13_piece0 stored as bytes in memory (estimated size 491.0 B, free 396.9 MiB)

Big data programming Assignment -5

20/04/08 13:17:56 INFO BlockManagerInfo: Added broadcast_13_piece0 in memory on 192.168.122.108:44235 (size: 491.0 B, free: 413.8 MiB)

20/04/08 13:17:56 INFO SparkContext: Created broadcast 13 from call at FutureTask.java:264

20/04/08 13:17:56 INFO BlockManagerInfo: Removed broadcast_3_piece0 on 192.168.122.108:44235 in memory (size: 27.5 KiB, free: 413.8 MiB)

20/04/08 13:17:56 INFO BlockManagerInfo: Removed broadcast_2_piece0 on 192.168.122.108:44235 in memory (size: 27.5 KiB, free: 413.8 MiB)

20/04/08 13:17:56 INFO CodeGenerator: Code generated in 38.773107 ms

20/04/08 13:17:56 INFO BlockManagerInfo: Removed broadcast_1_piece0 on 192.168.122.108:44235 in memory (size: 5.7 KiB, free: 413.8 MiB)

20/04/08 13:17:56 INFO BlockManagerInfo: Removed broadcast_0_piece0 on 192.168.122.108:44235 in memory (size: 27.5 KiB, free: 413.9 MiB)

20/04/08 13:17:56 INFO SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0

20/04/08 13:17:56 INFO DAGScheduler: Got job 5 (showString at NativeMethodAccessorImpl.java:0) with 1 output partitions

20/04/08 13:17:56 INFO DAGScheduler: Final stage: ResultStage 5 (showString at NativeMethodAccessorImpl.java:0)

20/04/08 13:17:56 INFO DAGScheduler: Parents of final stage: List()

20/04/08 13:17:56 INFO DAGScheduler: Missing parents: List()

20/04/08 13:17:56 INFO DAGScheduler: Submitting ResultStage 5 (MapPartitionsRDD[23] at showString at NativeMethodAccessorImpl.java:0), which has no missing parents

20/04/08 13:17:57 INFO MemoryStore: Block broadcast_14 stored as values in memory (estimated size 12.9 KiB, free 397.5 MiB)

20/04/08 13:17:57 INFO MemoryStore: Block broadcast_14_piece0 stored as bytes in memory (estimated size 6.2 KiB, free 397.5 MiB)

20/04/08 13:17:57 INFO BlockManagerInfo: Added broadcast_14_piece0 in memory on 192.168.122.108:44235 (size: 6.2 KiB, free: 413.9 MiB)

20/04/08 13:17:57 INFO SparkContext: Created broadcast 14 from broadcast at DAGScheduler.scala:1206

20/04/08 13:17:57 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 5 (MapPartitionsRDD[23] at showString at NativeMethodAccessorImpl.java:0) (first 15 tasks are for partitions Vector(0))

20/04/08 13:17:57 INFO TaskSchedulerImpl: Adding task set 5.0 with 1 tasks

20/04/08 13:17:57 INFO TaskSetManager: Starting task 0.0 in stage 5.0 (TID 5, 192.168.122.108, executor driver, partition 0, PROCESS_LOCAL, 7925 bytes)

20/04/08 13:17:57 INFO Executor: Running task 0.0 in stage 5.0 (TID 5)

20/04/08 13:17:57 INFO FilePartitionReader: Reading file path: file:///home/vbhamidipati1/spark/workspace/data/tweets.json, range: 0-813, partition values: [empty row]

20/04/08 13:17:57 INFO Executor: Finished task 0.0 in stage 5.0 (TID 5). 2324 bytes result sent to driver

20/04/08 13:17:57 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 5) in 39 ms on 192.168.122.108 (executor driver) (1/1)

20/04/08 13:17:57 INFO TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool

20/04/08 13:17:57 INFO DAGScheduler: ResultStage 5 (showString at NativeMethodAccessorImpl.java:0) finished in 0.053 s

20/04/08 13:17:57 INFO DAGScheduler: Job 5 is finished. Cancelling potential speculative or zombie tasks for this job

Big data programming Assignment -5

20/04/08 13:17:57 INFO TaskSchedulerImpl: Killing all running tasks in stage 5: Stage finished
20/04/08 13:17:57 INFO DAGScheduler: Job 5 finished: showString at NativeMethodAccessorImpl.java:0, took 0.063541 s
20/04/08 13:17:57 INFO CodeGenerator: Code generated in 30.776623 ms

```
+-----+-----+-----+-----+
|   geo|   tweet| user|   state|
+-----+-----+-----+-----+
| Atlanta| It is a sunny day!| Bob| Georgia|
| Athens| We have a footbal...| Susan| Georgia|
| Atlanta| Today is cold.| David| Georgia|
| Auburn| I love Auburn Uni...| Lisa| Alabama|
| Birmingham| I will go to Atla...| Ben| Alabama|
| San Francisco| We watch a movie ...| Paul| California|
| San Diego| It is hot today. ...| Smith| California|
| Log Angeles| Oscar ceremony is...| Ethan| California|
| Log Angeles| I love Oscar cere...| Emma| California|
| Orlando| I will go to the ...| Rolando| Florida|
| Miami| Sunny Day!| Mia| Florida|
+-----+-----+-----+-----+
```

20/04/08 13:17:57 INFO BlockManagerInfo: Removed broadcast_14_piece0 on 192.168.122.108:44235 in memory (size: 6.2 KiB, free: 413.9 MiB)

20/04/08 13:17:57 INFO V2ScanRelationPushDown:

Pushing operators to json file:/home/vbhamidipati1/spark/workspace/data/tweets.json

Pushed Filters:

Post-Scan Filters: isnotnull(geo#7)

Output: geo#7, tweet#8

20/04/08 13:17:57 INFO V2ScanRelationPushDown:

Pushing operators to json file:/home/vbhamidipati1/spark/workspace/data/cityStateMap.json

Pushed Filters:

Post-Scan Filters: isnotnull(city#42)

Output: city#42, state#43

20/04/08 13:17:57 INFO MemoryStore: Block broadcast_15 stored as values in memory (estimated size 174.2 KiB, free 397.4 MiB)

20/04/08 13:17:57 INFO MemoryStore: Block broadcast_15_piece0 stored as bytes in memory (estimated size 27.5 KiB, free 397.3 MiB)

20/04/08 13:17:57 INFO BlockManagerInfo: Added broadcast_15_piece0 in memory on 192.168.122.108:44235 (size: 27.5 KiB, free: 413.8 MiB)

20/04/08 13:17:57 INFO SparkContext: Created broadcast 15 from showString at NativeMethodAccessorImpl.java:0

20/04/08 13:17:57 INFO MemoryStore: Block broadcast_16 stored as values in memory (estimated size 174.2 KiB, free 397.2 MiB)

20/04/08 13:17:58 INFO MemoryStore: Block broadcast_16_piece0 stored as bytes in memory (estimated size 27.5 KiB, free 397.1 MiB)

20/04/08 13:17:58 INFO BlockManagerInfo: Added broadcast_16_piece0 in memory on 192.168.122.108:44235 (size: 27.5 KiB, free: 413.8 MiB)

Big data programming Assignment -5

20/04/08 13:17:58 INFO SparkContext: Created broadcast 16 from showString at NativeMethodAccessorImpl.java:0

20/04/08 13:17:58 INFO SparkContext: Starting job: call at FutureTask.java:264

20/04/08 13:17:58 INFO DAGScheduler: Got job 6 (call at FutureTask.java:264) with 1 output partitions

20/04/08 13:17:58 INFO DAGScheduler: Final stage: ResultStage 6 (call at FutureTask.java:264)

20/04/08 13:17:58 INFO DAGScheduler: Parents of final stage: List()

20/04/08 13:17:58 INFO DAGScheduler: Missing parents: List()

20/04/08 13:17:58 INFO DAGScheduler: Submitting ResultStage 6 (MapPartitionsRDD[27] at call at FutureTask.java:264), which has no missing parents

20/04/08 13:17:58 INFO MemoryStore: Block broadcast_17 stored as values in memory (estimated size 10.7 KiB, free 397.1 MiB)

20/04/08 13:17:58 INFO MemoryStore: Block broadcast_17_piece0 stored as bytes in memory (estimated size 5.4 KiB, free 397.1 MiB)

20/04/08 13:17:58 INFO BlockManagerInfo: Added broadcast_17_piece0 in memory on 192.168.122.108:44235 (size: 5.4 KiB, free: 413.8 MiB)

20/04/08 13:17:58 INFO SparkContext: Created broadcast 17 from broadcast at DAGScheduler.scala:1206

20/04/08 13:17:58 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 6 (MapPartitionsRDD[27] at call at FutureTask.java:264) (first 15 tasks are for partitions Vector(0))

20/04/08 13:17:58 INFO TaskSchedulerImpl: Adding task set 6.0 with 1 tasks

20/04/08 13:17:58 INFO TaskSetManager: Starting task 0.0 in stage 6.0 (TID 6, 192.168.122.108, executor driver, partition 0, PROCESS_LOCAL, 7931 bytes)

20/04/08 13:17:58 INFO Executor: Running task 0.0 in stage 6.0 (TID 6)

20/04/08 13:17:58 INFO FilePartitionReader: Reading file path: file:///home/vbhamidipati1/spark/workspace/data/cityStateMap.json, range: 0-390, partition values: [empty row]

20/04/08 13:17:58 INFO Executor: Finished task 0.0 in stage 6.0 (TID 6). 1728 bytes result sent to driver

20/04/08 13:17:58 INFO TaskSetManager: Finished task 0.0 in stage 6.0 (TID 6) in 24 ms on 192.168.122.108 (executor driver) (1/1)

20/04/08 13:17:58 INFO TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool

20/04/08 13:17:58 INFO DAGScheduler: ResultStage 6 (call at FutureTask.java:264) finished in 0.037 s

20/04/08 13:17:58 INFO DAGScheduler: Job 6 is finished. Cancelling potential speculative or zombie tasks for this job

20/04/08 13:17:58 INFO TaskSchedulerImpl: Killing all running tasks in stage 6: Stage finished

20/04/08 13:17:58 INFO DAGScheduler: Job 6 finished: call at FutureTask.java:264, took 0.043113 s

20/04/08 13:17:58 INFO MemoryStore: Block broadcast_18 stored as values in memory (estimated size 16.0 MiB, free 381.1 MiB)

20/04/08 13:17:58 INFO MemoryStore: Block broadcast_18_piece0 stored as bytes in memory (estimated size 491.0 B, free 381.1 MiB)

20/04/08 13:17:58 INFO BlockManagerInfo: Added broadcast_18_piece0 in memory on 192.168.122.108:44235 (size: 491.0 B, free: 413.8 MiB)

20/04/08 13:17:58 INFO SparkContext: Created broadcast 18 from call at FutureTask.java:264

20/04/08 13:17:58 INFO BlockManagerInfo: Removed broadcast_17_piece0 on 192.168.122.108:44235 in memory (size: 5.4 KiB, free: 413.8 MiB)

20/04/08 13:17:58 INFO CodeGenerator: Code generated in 136.147226 ms

20/04/08 13:17:58 INFO SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0

Big data programming Assignment -5

```
20/04/08 13:17:58 INFO DAGScheduler: Got job 7 (showString at NativeMethodAccessorImpl.java:0)
with 1 output partitions
20/04/08 13:17:58 INFO DAGScheduler: Final stage: ResultStage 7 (showString at
NativeMethodAccessorImpl.java:0)
20/04/08 13:17:58 INFO DAGScheduler: Parents of final stage: List()
20/04/08 13:17:58 INFO DAGScheduler: Missing parents: List()
20/04/08 13:17:58 INFO DAGScheduler: Submitting ResultStage 7 (MapPartitionsRDD[31] at showString
at NativeMethodAccessorImpl.java:0), which has no missing parents
20/04/08 13:17:58 INFO MemoryStore: Block broadcast_19 stored as values in memory (estimated size
36.8 KiB, free 381.1 MiB)
20/04/08 13:17:58 INFO MemoryStore: Block broadcast_19_piece0 stored as bytes in memory
(estimated size 15.7 KiB, free 381.1 MiB)
20/04/08 13:17:58 INFO BlockManagerInfo: Added broadcast_19_piece0 in memory on
192.168.122.108:44235 (size: 15.7 KiB, free: 413.8 MiB)
20/04/08 13:17:58 INFO SparkContext: Created broadcast 19 from broadcast at
DAGScheduler.scala:1206
20/04/08 13:17:58 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 7
(MapPartitionsRDD[31] at showString at NativeMethodAccessorImpl.java:0) (first 15 tasks are for
partitions Vector(0))
20/04/08 13:17:58 INFO TaskSchedulerImpl: Adding task set 7.0 with 1 tasks
20/04/08 13:17:58 INFO TaskSetManager: Starting task 0.0 in stage 7.0 (TID 7, 192.168.122.108,
executor driver, partition 0, PROCESS_LOCAL, 7925 bytes)
20/04/08 13:17:58 INFO Executor: Running task 0.0 in stage 7.0 (TID 7)
20/04/08 13:17:58 INFO CodeGenerator: Code generated in 18.157749 ms
20/04/08 13:17:58 INFO CodeGenerator: Code generated in 14.027567 ms
20/04/08 13:17:58 INFO FilePartitionReader: Reading file path:
file:///home/vbhamidipati1/spark/workspace/data/tweets.json, range: 0-813, partition values: [empty
row]
20/04/08 13:17:58 INFO Executor: Finished task 0.0 in stage 7.0 (TID 7). 3640 bytes result sent to driver
20/04/08 13:17:58 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID 7) in 202 ms on
192.168.122.108 (executor driver) (1/1)
20/04/08 13:17:58 INFO TaskSchedulerImpl: Removed TaskSet 7.0, whose tasks have all completed,
from pool
20/04/08 13:17:58 INFO DAGScheduler: ResultStage 7 (showString at NativeMethodAccessorImpl.java:0)
finished in 0.220 s
20/04/08 13:17:58 INFO DAGScheduler: Job 7 is finished. Cancelling potential speculative or zombie
tasks for this job
20/04/08 13:17:58 INFO TaskSchedulerImpl: Killing all running tasks in stage 7: Stage finished
20/04/08 13:17:58 INFO DAGScheduler: Job 7 finished: showString at NativeMethodAccessorImpl.java:0,
took 0.230331 s
20/04/08 13:17:58 INFO CodeGenerator: Code generated in 27.754429 ms
20/04/08 13:17:58 INFO BlockManagerInfo: Removed broadcast_19_piece0 on 192.168.122.108:44235
in memory (size: 15.7 KiB, free: 413.8 MiB)
+-----+-----+
| state|count|
+-----+-----+
| Georgia| 3|
| Alabama| 2|
```

Big data programming Assignment -5

|California| 4|

| Florida| 2|

+-----+-----+

20/04/08 13:17:58 INFO SparkUI: Stopped Spark web UI at http://192.168.122.108:4040

20/04/08 13:17:58 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!

20/04/08 13:17:58 INFO MemoryStore: MemoryStore cleared

20/04/08 13:17:58 INFO BlockManager: BlockManager stopped

20/04/08 13:17:58 INFO BlockManagerMaster: BlockManagerMaster stopped

20/04/08 13:17:58 INFO OutputCommitCoordinator\$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!

20/04/08 13:17:58 INFO SparkContext: Successfully stopped SparkContext

20/04/08 13:17:59 INFO ShutdownHookManager: Shutdown hook called

20/04/08 13:17:59 INFO ShutdownHookManager: Deleting directory /tmp/spark-1fa7662f-4790-4f6e-96f3-19d1b99e804c

20/04/08 13:17:59 INFO ShutdownHookManager: Deleting directory /tmp/spark-5a2a6579-70e6-4319-905c-e537d61b709e

20/04/08 13:17:59 INFO ShutdownHookManager: Deleting directory /tmp/spark-1fa7662f-4790-4f6e-96f3-19d1b99e804c/pyspark-27024797-a81a-43e5-bee0-66efa7668f88