Shop Revenue Analysis

Introduction

Authors

Thomas Charuel, Kyota Lannelongue

Subject

The goal is to write a small Spark script that allows to display different statistics on the shop's performance:

- Average monthly income of the shop in France (on a 1 year data)
- Total revenue per city per year
- Average monthly income of the shop in each city (on a 1 year data)
- Total revenue per store per year
- The store that achieves the best performance in each month

We also wrote the following query (not required):

- The average income of the shop in France (on a 1 year data) per month

Script

```
import sys
from pyspark import SparkContext

# Instanciate the spark context
sc = SparkContext()

# Load the the input files
# The RDD is the key/value list with key the filename and value the file
content
files = sc.wholeTextFiles(sys.argv[1])

# Generate an object for each line on each file like ('store_name', 'city',
'month', 'income')
f1 = files.map(lambda file: (file[0].split("/")[-1], file[1]))
f2 = f1.map(lambda file: (file[0].split(".txt")[0], file[1]))
f3 = f2.flatMapValues(lambda v: v.split("\r\n"))
f4 = f3.map(lambda kv: (kv[0], kv[0].split("_")[0], kv[1].split(" ")[0],
kv[1].split(" ")[1]))
cityAsKey = f4.map(lambda scmr: (scmr[1], scmr[0], scmr[2], int(scmr[3])))

# persist the cityAsKey rdd, because this rdd is used by multiple queries
cityAsKey.cache()

## Query 1
# Get all income values
```

```
income_list = cityAsKey.map(lambda income_line: income_line[3])

# Compute the total income for the shop on 1 year
total_income = income_list.reduce(lambda income1, income2: income1 + income2)

# Compute the monthly average over 1 year (12 months in a year)
average_monthly_income_on_1_year = total_income / 12


## Query 2
# Prepare the data, by transforming the cityAsKey rdd as a tuple list with key:
city, value: income value
income_list_with_city_as_key = cityAsKey.map(lambda income_line:
(income_line[0], income_line[3]))

# Compute the revenue on 1 year for each city
total_revenue_per_city = income_list_with_city_as_key.reduceByKey(lambda
income1, income2: income1 + income2)

# persist the total_revenue_per_city rdd cause we will use it again with the
query 3
total_revenue_per_city.persist()


## Query 3
# Using the revenue on 1 year for each city result from the previous query,
# we can easily compute the average monthly income per city by dividing the
previous results by 12
average_monthly_income_on_1_year_per_city =
total_revenue_per_city.mapValues(lambda income_on_1_year: income_on_1_year/12)


## Query 4
# Prepare the data, by transforming the cityAsKey rdd as a tuple list with key:
store, value: income value
income_list_with_store_as_key = cityAsKey.map(lambda income_line:
(income_line[1], income_line[3]))

# Compute th revenue on 1 year for each store
total_revenue_per_store = income_list_with_store_as_key.reduceByKey(lambda
income1, income2: income1 + income2)


## Query 5
# Prepare the data, by transforming the cityAsKey rdd as a tuple list with key:
month, value: (store, income)
income_with_store_list_with_month_as_key = cityAsKey.map(lambda income_line:
(income_line[2], (income_line[1], income_line[3])))

# Find the best income for each month
best_income_per_month =
income_with_store_list_with_month_as_key.reduceByKey(lambda income_line1,
```

```
income_line2: income_line1 if income_line1[1] > income_line2[1] else
income_line2)

# Get the store name for each best income line of the month
best_performance_store_per_month = best_income_per_month.mapValues(lambda
income_line: income_line[0])


# Query 6 (not required)
# This query displays the average income of the shop in France (on a 1 year
data) per month
# Prepare the data, by transforming the cityAsKey rdd as a tuple list with key:
month, value: income
income_list_with_month_as_key = cityAsKey.map(lambda income_line:
(income_line[2], income_line[3]))

# Regroup the incomes per month
incomes_per_month = income_list_with_month_as_key.groupByKey()

# Compute the average income per month
average_income_per_month = incomes_per_month.mapValues(lambda incomes:
sum(incomes)/len(incomes))


# Display the results
# Query 1:
print("\nAverage monthly income of the shop in France (on a 1 year data):")
print(average_monthly_income_on_1_year)

# Query 2:
print("\nTotal revenue per city per year:")
for city, income in total_revenue_per_city.collect():
  print(city, ": ", income)

# Query 3:
print("\nAverage monthly income of the shop in each city (on a 1 year data):")
for city, income in average_monthly_income_on_1_year_per_city.collect():
  print(city, ": ", income)

# Query 4:
print("\nTotal revenue per store per year:")
for store, income in total_revenue_per_store.collect():
  print(store, ": ", income)

# Query 5:
print("\nThe store that achieves the best performance in each month:")
for month, store in best_performance_store_per_month.collect():
  print(month, ": ", store)

# Query 6:
print("\nAverage income of the shop in France (on a 1 year data) per month:")
for month, average_income in average_income_per_month.collect():
```

```
  print(month, ": ", average_income)
```

Results

Results without logs

```
Average monthly income of the shop in France (on a 1 year data):
301.5833333333333

Total revenue per city per year:
anger :  166
lyon :  193
nice :  203
paris :  1568
troyes :  214
marseilles :  515
nantes :  207
orlean :  196
rennes :  180
toulouse :  177

Average monthly income of the shop in each city (on a 1 year data):
anger :  13.833333333333334
lyon :  16.083333333333332
nice :  16.916666666666668
paris :  130.66666666666666
troyes :  17.833333333333332
marseilles :  42.916666666666664
nantes :  17.25
orlean :  16.333333333333332
rennes :  15.0
toulouse :  14.75

Total revenue per store per year:
anger :  166
lyon :  193
marseilles_1 :  284
nice :  203
paris_2 :  642
paris_3 :  330
troyes :  214
marseilles_2 :  231
nantes :  207
orlean :  196
paris_1 :  596
rennes :  180
toulouse :  177

The store that achieves the best performance in each month:
APR :  paris_1
MAY :  paris_2
```

```
AUG :  paris_2
JAN :  paris_1
FEB :  paris_2
MAR :  paris_2
JUN :  paris_2
JUL :  paris_1
SEP :  paris_2
OCT :  paris_1
NOV :  paris_2
DEC :  paris_1

Average income of the shop in France (on a 1 year data) per month:
APR :  20.23076923076923
MAY :  22.46153846153846
AUG :  23.076923076923077
JAN :  20.76923076923077
FEB :  19.153846153846153
MAR :  17.53846153846154
JUN :  27.846153846153847
JUL :  21.692307692307693
SEP :  25.53846153846154
OCT :  26.53846153846154
NOV :  24.53846153846154
DEC :  29.0
```

Results with logs

```
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
17/10/26 15:47:49 INFO SparkContext: Running Spark version 2.2.0
17/10/26 15:47:50 WARN NativeCodeLoader: Unable to load native-hadoop library
for your platform... using builtin-java classes where applicable
17/10/26 15:47:50 INFO SparkContext: Submitted application: main.py
17/10/26 15:47:50 INFO SecurityManager: Changing view acls to: thomas
17/10/26 15:47:50 INFO SecurityManager: Changing modify acls to: thomas
17/10/26 15:47:50 INFO SecurityManager: Changing view acls groups to:
17/10/26 15:47:50 INFO SecurityManager: Changing modify acls groups to:
17/10/26 15:47:50 INFO SecurityManager: SecurityManager: authentication
disabled; ui acls disabled; users  with view permissions: Set(thomas); groups
with view permissions: Set(); users  with modify permissions: Set(thomas);
groups with modify permissions: Set()
17/10/26 15:47:50 INFO Utils: Successfully started service 'sparkDriver' on
port 59434.
17/10/26 15:47:50 INFO SparkEnv: Registering MapOutputTracker
17/10/26 15:47:50 INFO SparkEnv: Registering BlockManagerMaster
17/10/26 15:47:50 INFO BlockManagerMasterEndpoint: Using
org.apache.spark.storage.DefaultTopologyMapper for getting topology information
17/10/26 15:47:50 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint
up
17/10/26 15:47:50 INFO DiskBlockManager: Created local directory at
/private/var/folders/6_/n45kv5s50_j7lphv4rz2795r0000gn/T/blockmgr-436a7a15-
535a-4998-9da8-4e29bd4a94f3
17/10/26 15:47:50 INFO MemoryStore: MemoryStore started with capacity 366.3 MB
```

```
17/10/26 15:47:51 INFO SparkEnv: Registering OutputCommitCoordinator
17/10/26 15:47:51 INFO Utils: Successfully started service 'SparkUI' on port
4040.
17/10/26 15:47:51 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at
http://10.4.186.168:4040
17/10/26 15:47:51 INFO SparkContext: Added file
file:/Users/thomas/Desktop/ING5/Spark/lab1/main.py at
file:/Users/thomas/Desktop/ING5/Spark/lab1/main.py with timestamp 1509025671788
17/10/26 15:47:51 INFO Utils: Copying
/Users/thomas/Desktop/ING5/Spark/lab1/main.py to
/private/var/folders/6_/n45kv5s50_j7lphv4rz2795r0000gn/T/spark-d0cd7b11-34bc-
48e9-b114-2ef4d64da860/userFiles-a1c676ef-7b8e-4975-8a91-6d287e1f38ab/main.py
17/10/26 15:47:51 INFO Executor: Starting executor ID driver on host localhost
17/10/26 15:47:52 INFO Utils: Successfully started service
'org.apache.spark.network.netty.NettyBlockTransferService' on port 59440.
17/10/26 15:47:52 INFO NettyBlockTransferService: Server created on
10.4.186.168:59440
17/10/26 15:47:52 INFO BlockManager: Using
org.apache.spark.storage.RandomBlockReplicationPolicy for block replication
policy
17/10/26 15:47:52 INFO BlockManagerMaster: Registering BlockManager
BlockManagerId(driver, 10.4.186.168, 59440, None)
17/10/26 15:47:52 INFO BlockManagerMasterEndpoint: Registering block manager
10.4.186.168:59440 with 366.3 MB RAM, BlockManagerId(driver, 10.4.186.168,
59440, None)
17/10/26 15:47:52 INFO BlockManagerMaster: Registered BlockManager
BlockManagerId(driver, 10.4.186.168, 59440, None)
17/10/26 15:47:52 INFO BlockManager: Initialized BlockManager:
BlockManagerId(driver, 10.4.186.168, 59440, None)
17/10/26 15:47:52 INFO MemoryStore: Block broadcast_0 stored as values in
memory (estimated size 275.4 KB, free 366.0 MB)
17/10/26 15:47:53 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in
memory (estimated size 23.0 KB, free 366.0 MB)
17/10/26 15:47:53 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on
10.4.186.168:59440 (size: 23.0 KB, free: 366.3 MB)
17/10/26 15:47:53 INFO SparkContext: Created broadcast 0 from wholeTextFiles at
NativeMethodAccessorImpl.java:0
17/10/26 15:47:53 INFO FileInputFormat: Total input paths to process : 13
17/10/26 15:47:53 INFO FileInputFormat: Total input paths to process : 13
17/10/26 15:47:53 INFO CombineFileInputFormat: DEBUG: Terminated node
allocation with : CompletedNodes: 1, size left: 563
17/10/26 15:47:53 INFO SparkContext: Starting job: reduce at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:26
17/10/26 15:47:53 INFO DAGScheduler: Got job 0 (reduce at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:26) with 2 output partitions
17/10/26 15:47:53 INFO DAGScheduler: Final stage: ResultStage 0 (reduce at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:26)
17/10/26 15:47:53 INFO DAGScheduler: Parents of final stage: List()
17/10/26 15:47:53 INFO DAGScheduler: Missing parents: List()
17/10/26 15:47:53 INFO DAGScheduler: Submitting ResultStage 0 (PythonRDD[3] at
reduce at /Users/thomas/Desktop/ING5/Spark/lab1/main.py:26), which has no
missing parents
```

```
17/10/26 15:47:53 INFO MemoryStore: Block broadcast_1 stored as values in
memory (estimated size 8.4 KB, free 366.0 MB)
17/10/26 15:47:53 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in
memory (estimated size 5.1 KB, free 366.0 MB)
17/10/26 15:47:53 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on
10.4.186.168:59440 (size: 5.1 KB, free: 366.3 MB)
17/10/26 15:47:53 INFO SparkContext: Created broadcast 1 from broadcast at
DAGScheduler.scala:1006
17/10/26 15:47:53 INFO DAGScheduler: Submitting 2 missing tasks from
ResultStage 0 (PythonRDD[3] at reduce at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:26) (first 15 tasks are for
partitions Vector(0, 1))
17/10/26 15:47:53 INFO TaskSchedulerImpl: Adding task set 0.0 with 2 tasks
17/10/26 15:47:53 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0,
localhost, executor driver, partition 0, PROCESS_LOCAL, 5427 bytes)
17/10/26 15:47:53 INFO TaskSetManager: Starting task 1.0 in stage 0.0 (TID 1,
localhost, executor driver, partition 1, PROCESS_LOCAL, 5348 bytes)
17/10/26 15:47:53 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
17/10/26 15:47:53 INFO Executor: Running task 1.0 in stage 0.0 (TID 1)
17/10/26 15:47:53 INFO Executor: Fetching
file:/Users/thomas/Desktop/ING5/Spark/lab1/main.py with timestamp 1509025671788
17/10/26 15:47:53 INFO Utils: /Users/thomas/Desktop/ING5/Spark/lab1/main.py has
been previously copied to
/private/var/folders/6_/n45kv5s50_j7lphv4rz2795r0000gn/T/spark-d0cd7b11-34bc-
48e9-b114-2ef4d64da860/userFiles-a1c676ef-7b8e-4975-8a91-6d287e1f38ab/main.py
17/10/26 15:47:53 INFO WholeTextFileRDD: Input split:
Paths:/Users/thomas/Desktop/ING5/Spark/lab1/input1/paris_1.txt:0+94,/Users/thom
as/Desktop/ING5/Spark/lab1/input1/paris_2.txt:0+94,/Users/thomas/Desktop/ING5/S
park/lab1/input1/paris_3.txt:0+94,/Users/thomas/Desktop/ING5/Spark/lab1/input1/
rennes.txt:0+93,/Users/thomas/Desktop/ING5/Spark/lab1/input1/toulouse.txt:0+94,
/Users/thomas/Desktop/ING5/Spark/lab1/input1/troyes.txt:0+94
17/10/26 15:47:53 INFO WholeTextFileRDD: Input split:
Paths:/Users/thomas/Desktop/ING5/Spark/lab1/input1/anger.txt:0+93,/Users/thomas
/Desktop/ING5/Spark/lab1/input1/lyon.txt:0+94,/Users/thomas/Desktop/ING5/Spark/
lab1/input1/marseilles_1.txt:0+94,/Users/thomas/Desktop/ING5/Spark/lab1/input1/
marseilles_2.txt:0+94,/Users/thomas/Desktop/ING5/Spark/lab1/input1/nantes.txt:0
+94,/Users/thomas/Desktop/ING5/Spark/lab1/input1/nice.txt:0+93,/Users/thomas/De
sktop/ING5/Spark/lab1/input1/orlean.txt:0+93
17/10/26 15:47:54 INFO PythonRunner: Times: total = 629, boot = 407, init =
221, finish = 1
17/10/26 15:47:54 INFO PythonRunner: Times: total = 670, boot = 412, init =
257, finish = 1
17/10/26 15:47:54 INFO MemoryStore: Block rdd_2_0 stored as bytes in memory
(estimated size 1162.0 B, free 366.0 MB)
17/10/26 15:47:54 INFO MemoryStore: Block rdd_2_1 stored as bytes in memory
(estimated size 1056.0 B, free 366.0 MB)
17/10/26 15:47:54 INFO BlockManagerInfo: Added rdd_2_0 in memory on
10.4.186.168:59440 (size: 1162.0 B, free: 366.3 MB)
17/10/26 15:47:54 INFO BlockManagerInfo: Added rdd_2_1 in memory on
10.4.186.168:59440 (size: 1056.0 B, free: 366.3 MB)
17/10/26 15:47:54 INFO PythonRunner: Times: total = 3, boot = -220, init = 222,
finish = 1
```

```
17/10/26 15:47:54 INFO PythonRunner: Times: total = 3, boot = -259, init = 262,
finish = 0
17/10/26 15:47:54 INFO Executor: Finished task 1.0 in stage 0.0 (TID 1). 1837
bytes result sent to driver
17/10/26 15:47:54 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1837
bytes result sent to driver
17/10/26 15:47:54 INFO TaskSetManager: Finished task 1.0 in stage 0.0 (TID 1)
in 1131 ms on localhost (executor driver) (1/2)
17/10/26 15:47:54 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0)
in 1159 ms on localhost (executor driver) (2/2)
17/10/26 15:47:54 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have
all completed, from pool
17/10/26 15:47:54 INFO DAGScheduler: ResultStage 0 (reduce at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:26) finished in 1,197 s
17/10/26 15:47:54 INFO DAGScheduler: Job 0 finished: reduce at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:26, took 1,323679 s

Average monthly income of the shop in France (on a 1 year data):
301.5833333333333

Total revenue per city per year:
17/10/26 15:47:54 INFO SparkContext: Starting job: collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:87
17/10/26 15:47:54 INFO DAGScheduler: Registering RDD 5 (reduceByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:37)
17/10/26 15:47:54 INFO DAGScheduler: Got job 1 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:87) with 2 output partitions
17/10/26 15:47:54 INFO DAGScheduler: Final stage: ResultStage 2 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:87)
17/10/26 15:47:54 INFO DAGScheduler: Parents of final stage:
List(ShuffleMapStage 1)
17/10/26 15:47:54 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 1)
17/10/26 15:47:54 INFO DAGScheduler: Submitting ShuffleMapStage 1
(PairwiseRDD[5] at reduceByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:37), which has no missing parents
17/10/26 15:47:54 INFO MemoryStore: Block broadcast_2 stored as values in
memory (estimated size 11.1 KB, free 366.0 MB)
17/10/26 15:47:54 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in
memory (estimated size 6.9 KB, free 366.0 MB)
17/10/26 15:47:54 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on
10.4.186.168:59440 (size: 6.9 KB, free: 366.3 MB)
17/10/26 15:47:54 INFO SparkContext: Created broadcast 2 from broadcast at
DAGScheduler.scala:1006
17/10/26 15:47:54 INFO DAGScheduler: Submitting 2 missing tasks from
ShuffleMapStage 1 (PairwiseRDD[5] at reduceByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:37) (first 15 tasks are for
partitions Vector(0, 1))
17/10/26 15:47:54 INFO TaskSchedulerImpl: Adding task set 1.0 with 2 tasks
17/10/26 15:47:54 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 2,
localhost, executor driver, partition 0, PROCESS_LOCAL, 5416 bytes)
17/10/26 15:47:54 INFO TaskSetManager: Starting task 1.0 in stage 1.0 (TID 3,
localhost, executor driver, partition 1, PROCESS_LOCAL, 5337 bytes)
```

```
17/10/26 15:47:54 INFO Executor: Running task 0.0 in stage 1.0 (TID 2)
17/10/26 15:47:54 INFO Executor: Running task 1.0 in stage 1.0 (TID 3)
17/10/26 15:47:54 INFO BlockManager: Found block rdd_2_0 locally
17/10/26 15:47:54 INFO BlockManager: Found block rdd_2_1 locally
17/10/26 15:47:54 INFO PythonRunner: Times: total = 4, boot = -246, init = 249,
finish = 1
17/10/26 15:47:54 INFO PythonRunner: Times: total = 5, boot = -248, init = 252,
finish = 1
17/10/26 15:47:54 INFO Executor: Finished task 1.0 in stage 1.0 (TID 3). 1568
bytes result sent to driver
17/10/26 15:47:54 INFO Executor: Finished task 0.0 in stage 1.0 (TID 2). 1568
bytes result sent to driver
17/10/26 15:47:54 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (TID 3)
in 73 ms on localhost (executor driver) (1/2)
17/10/26 15:47:54 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2)
in 76 ms on localhost (executor driver) (2/2)
17/10/26 15:47:54 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have
all completed, from pool
17/10/26 15:47:54 INFO DAGScheduler: ShuffleMapStage 1 (reduceByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:37) finished in 0,079 s
17/10/26 15:47:54 INFO DAGScheduler: looking for newly runnable stages
17/10/26 15:47:54 INFO DAGScheduler: running: Set()
17/10/26 15:47:54 INFO DAGScheduler: waiting: Set(ResultStage 2)
17/10/26 15:47:54 INFO DAGScheduler: failed: Set()
17/10/26 15:47:54 INFO DAGScheduler: Submitting ResultStage 2 (PythonRDD[8] at
RDD at PythonRDD.scala:48), which has no missing parents
17/10/26 15:47:54 INFO MemoryStore: Block broadcast_3 stored as values in
memory (estimated size 6.3 KB, free 366.0 MB)
17/10/26 15:47:54 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in
memory (estimated size 4.0 KB, free 366.0 MB)
17/10/26 15:47:54 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on
10.4.186.168:59440 (size: 4.0 KB, free: 366.3 MB)
17/10/26 15:47:54 INFO SparkContext: Created broadcast 3 from broadcast at
DAGScheduler.scala:1006
17/10/26 15:47:54 INFO DAGScheduler: Submitting 2 missing tasks from
ResultStage 2 (PythonRDD[8] at RDD at PythonRDD.scala:48) (first 15 tasks are
for partitions Vector(0, 1))
17/10/26 15:47:54 INFO TaskSchedulerImpl: Adding task set 2.0 with 2 tasks
17/10/26 15:47:54 INFO TaskSetManager: Starting task 0.0 in stage 2.0 (TID 4,
localhost, executor driver, partition 0, ANY, 4621 bytes)
17/10/26 15:47:54 INFO TaskSetManager: Starting task 1.0 in stage 2.0 (TID 5,
localhost, executor driver, partition 1, ANY, 4621 bytes)
17/10/26 15:47:54 INFO Executor: Running task 0.0 in stage 2.0 (TID 4)
17/10/26 15:47:54 INFO Executor: Running task 1.0 in stage 2.0 (TID 5)
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks
out of 2 blocks
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks
out of 2 blocks
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in
5 ms
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in
5 ms
```

```
17/10/26 15:47:55 INFO PythonRunner: Times: total = 6, boot = -86, init = 92,
finish = 0
17/10/26 15:47:55 INFO MemoryStore: Block rdd_8_0 stored as bytes in memory
(estimated size 154.0 B, free 366.0 MB)
17/10/26 15:47:55 INFO BlockManagerInfo: Added rdd_8_0 in memory on
10.4.186.168:59440 (size: 154.0 B, free: 366.3 MB)
17/10/26 15:47:55 INFO PythonRunner: Times: total = 6, boot = -88, init = 94,
finish = 0
17/10/26 15:47:55 INFO MemoryStore: Block rdd_8_1 stored as bytes in memory
(estimated size 159.0 B, free 366.0 MB)
17/10/26 15:47:55 INFO BlockManagerInfo: Added rdd_8_1 in memory on
10.4.186.168:59440 (size: 159.0 B, free: 366.3 MB)
17/10/26 15:47:55 INFO Executor: Finished task 0.0 in stage 2.0 (TID 4). 2041
bytes result sent to driver
17/10/26 15:47:55 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 4)
in 53 ms on localhost (executor driver) (1/2)
17/10/26 15:47:55 INFO Executor: Finished task 1.0 in stage 2.0 (TID 5). 2053
bytes result sent to driver
17/10/26 15:47:55 INFO TaskSetManager: Finished task 1.0 in stage 2.0 (TID 5)
in 57 ms on localhost (executor driver) (2/2)
17/10/26 15:47:55 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have
all completed, from pool
17/10/26 15:47:55 INFO DAGScheduler: ResultStage 2 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:87) finished in 0,060 s
17/10/26 15:47:55 INFO DAGScheduler: Job 1 finished: collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:87, took 0,192788 s
anger :  166
lyon :  193
nice :  203
paris :  1568
troyes :  214
marseilles :  515
nantes :  207
orlean :  196
rennes :  180
toulouse :  177

Average monthly income of the shop in each city (on a 1 year data):
17/10/26 15:47:55 INFO SparkContext: Starting job: collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:92
17/10/26 15:47:55 INFO MapOutputTrackerMaster: Size of output statuses for
shuffle 0 is 160 bytes
17/10/26 15:47:55 INFO DAGScheduler: Got job 2 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:92) with 2 output partitions
17/10/26 15:47:55 INFO DAGScheduler: Final stage: ResultStage 4 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:92)
17/10/26 15:47:55 INFO DAGScheduler: Parents of final stage:
List(ShuffleMapStage 3)
17/10/26 15:47:55 INFO DAGScheduler: Missing parents: List()
17/10/26 15:47:55 INFO DAGScheduler: Submitting ResultStage 4 (PythonRDD[21] at
collect at /Users/thomas/Desktop/ING5/Spark/lab1/main.py:92), which has no
missing parents
```

```
17/10/26 15:47:55 INFO MemoryStore: Block broadcast_4 stored as values in
memory (estimated size 7.4 KB, free 366.0 MB)
17/10/26 15:47:55 INFO MemoryStore: Block broadcast_4_piece0 stored as bytes in
memory (estimated size 4.4 KB, free 366.0 MB)
17/10/26 15:47:55 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory on
10.4.186.168:59440 (size: 4.4 KB, free: 366.3 MB)
17/10/26 15:47:55 INFO SparkContext: Created broadcast 4 from broadcast at
DAGScheduler.scala:1006
17/10/26 15:47:55 INFO DAGScheduler: Submitting 2 missing tasks from
ResultStage 4 (PythonRDD[21] at collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:92) (first 15 tasks are for
partitions Vector(0, 1))
17/10/26 15:47:55 INFO TaskSchedulerImpl: Adding task set 4.0 with 2 tasks
17/10/26 15:47:55 INFO TaskSetManager: Starting task 0.0 in stage 4.0 (TID 6,
localhost, executor driver, partition 0, PROCESS_LOCAL, 4621 bytes)
17/10/26 15:47:55 INFO TaskSetManager: Starting task 1.0 in stage 4.0 (TID 7,
localhost, executor driver, partition 1, PROCESS_LOCAL, 4621 bytes)
17/10/26 15:47:55 INFO Executor: Running task 0.0 in stage 4.0 (TID 6)
17/10/26 15:47:55 INFO Executor: Running task 1.0 in stage 4.0 (TID 7)
17/10/26 15:47:55 INFO BlockManager: Found block rdd_8_1 locally
17/10/26 15:47:55 INFO BlockManager: Found block rdd_8_0 locally
17/10/26 15:47:55 INFO PythonRunner: Times: total = 1, boot = -55, init = 56,
finish = 0
17/10/26 15:47:55 INFO Executor: Finished task 0.0 in stage 4.0 (TID 6). 1506
bytes result sent to driver
17/10/26 15:47:55 INFO PythonRunner: Times: total = 1, boot = -55, init = 56,
finish = 0
17/10/26 15:47:55 INFO Executor: Finished task 1.0 in stage 4.0 (TID 7). 1518
bytes result sent to driver
17/10/26 15:47:55 INFO TaskSetManager: Finished task 0.0 in stage 4.0 (TID 6)
in 21 ms on localhost (executor driver) (1/2)
17/10/26 15:47:55 INFO TaskSetManager: Finished task 1.0 in stage 4.0 (TID 7)
in 21 ms on localhost (executor driver) (2/2)
17/10/26 15:47:55 INFO TaskSchedulerImpl: Removed TaskSet 4.0, whose tasks have
all completed, from pool
17/10/26 15:47:55 INFO DAGScheduler: ResultStage 4 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:92) finished in 0,024 s
17/10/26 15:47:55 INFO DAGScheduler: Job 2 finished: collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:92, took 0,037678 s
anger :  13.833333333333334
lyon :  16.083333333333332
nice :  16.916666666666668
paris :  130.66666666666666
troyes :  17.833333333333332
marseilles :  42.916666666666664
nantes :  17.25
orlean :  16.333333333333332
rennes :  15.0
toulouse :  14.75

Total revenue per store per year:
17/10/26 15:47:55 INFO SparkContext: Starting job: collect at
```

```
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:97
17/10/26 15:47:55 INFO DAGScheduler: Registering RDD 10 (reduceByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:54)
17/10/26 15:47:55 INFO DAGScheduler: Got job 3 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:97) with 2 output partitions
17/10/26 15:47:55 INFO DAGScheduler: Final stage: ResultStage 6 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:97)
17/10/26 15:47:55 INFO DAGScheduler: Parents of final stage:
List(ShuffleMapStage 5)
17/10/26 15:47:55 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 5)
17/10/26 15:47:55 INFO DAGScheduler: Submitting ShuffleMapStage 5
(PairwiseRDD[10] at reduceByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:54), which has no missing parents
17/10/26 15:47:55 INFO MemoryStore: Block broadcast_5 stored as values in
memory (estimated size 11.1 KB, free 365.9 MB)
17/10/26 15:47:55 INFO MemoryStore: Block broadcast_5_piece0 stored as bytes in
memory (estimated size 6.9 KB, free 365.9 MB)
17/10/26 15:47:55 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on
10.4.186.168:59440 (size: 6.9 KB, free: 366.2 MB)
17/10/26 15:47:55 INFO SparkContext: Created broadcast 5 from broadcast at
DAGScheduler.scala:1006
17/10/26 15:47:55 INFO DAGScheduler: Submitting 2 missing tasks from
ShuffleMapStage 5 (PairwiseRDD[10] at reduceByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:54) (first 15 tasks are for
partitions Vector(0, 1))
17/10/26 15:47:55 INFO TaskSchedulerImpl: Adding task set 5.0 with 2 tasks
17/10/26 15:47:55 INFO TaskSetManager: Starting task 0.0 in stage 5.0 (TID 8,
localhost, executor driver, partition 0, PROCESS_LOCAL, 5416 bytes)
17/10/26 15:47:55 INFO TaskSetManager: Starting task 1.0 in stage 5.0 (TID 9,
localhost, executor driver, partition 1, PROCESS_LOCAL, 5337 bytes)
17/10/26 15:47:55 INFO Executor: Running task 1.0 in stage 5.0 (TID 9)
17/10/26 15:47:55 INFO Executor: Running task 0.0 in stage 5.0 (TID 8)
17/10/26 15:47:55 INFO BlockManager: Found block rdd_2_1 locally
17/10/26 15:47:55 INFO BlockManager: Found block rdd_2_0 locally
17/10/26 15:47:55 INFO PythonRunner: Times: total = 1, boot = -34, init = 35,
finish = 0
17/10/26 15:47:55 INFO PythonRunner: Times: total = 2, boot = -37, init = 39,
finish = 0
17/10/26 15:47:55 INFO Executor: Finished task 1.0 in stage 5.0 (TID 9). 1525
bytes result sent to driver
17/10/26 15:47:55 INFO TaskSetManager: Finished task 1.0 in stage 5.0 (TID 9)
in 24 ms on localhost (executor driver) (1/2)
17/10/26 15:47:55 INFO Executor: Finished task 0.0 in stage 5.0 (TID 8). 1525
bytes result sent to driver
17/10/26 15:47:55 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 8)
in 32 ms on localhost (executor driver) (2/2)
17/10/26 15:47:55 INFO TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have
all completed, from pool
17/10/26 15:47:55 INFO DAGScheduler: ShuffleMapStage 5 (reduceByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:54) finished in 0,033 s
17/10/26 15:47:55 INFO DAGScheduler: looking for newly runnable stages
17/10/26 15:47:55 INFO DAGScheduler: running: Set()
```

```
17/10/26 15:47:55 INFO DAGScheduler: waiting: Set(ResultStage 6)
17/10/26 15:47:55 INFO DAGScheduler: failed: Set()
17/10/26 15:47:55 INFO DAGScheduler: Submitting ResultStage 6 (PythonRDD[22] at
collect at /Users/thomas/Desktop/ING5/Spark/lab1/main.py:97), which has no
missing parents
17/10/26 15:47:55 INFO MemoryStore: Block broadcast_6 stored as values in
memory (estimated size 6.3 KB, free 365.9 MB)
17/10/26 15:47:55 INFO MemoryStore: Block broadcast_6_piece0 stored as bytes in
memory (estimated size 4.0 KB, free 365.9 MB)
17/10/26 15:47:55 INFO BlockManagerInfo: Added broadcast_6_piece0 in memory on
10.4.186.168:59440 (size: 4.0 KB, free: 366.2 MB)
17/10/26 15:47:55 INFO SparkContext: Created broadcast 6 from broadcast at
DAGScheduler.scala:1006
17/10/26 15:47:55 INFO DAGScheduler: Submitting 2 missing tasks from
ResultStage 6 (PythonRDD[22] at collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:97) (first 15 tasks are for
partitions Vector(0, 1))
17/10/26 15:47:55 INFO TaskSchedulerImpl: Adding task set 6.0 with 2 tasks
17/10/26 15:47:55 INFO TaskSetManager: Starting task 0.0 in stage 6.0 (TID 10,
localhost, executor driver, partition 0, ANY, 4621 bytes)
17/10/26 15:47:55 INFO TaskSetManager: Starting task 1.0 in stage 6.0 (TID 11,
localhost, executor driver, partition 1, ANY, 4621 bytes)
17/10/26 15:47:55 INFO Executor: Running task 0.0 in stage 6.0 (TID 10)
17/10/26 15:47:55 INFO Executor: Running task 1.0 in stage 6.0 (TID 11)
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks
out of 2 blocks
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in
0 ms
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks
out of 2 blocks
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in
1 ms
17/10/26 15:47:55 INFO PythonRunner: Times: total = 4, boot = -28, init = 32,
finish = 0
17/10/26 15:47:55 INFO PythonRunner: Times: total = 2, boot = -28, init = 30,
finish = 0
17/10/26 15:47:55 INFO Executor: Finished task 1.0 in stage 6.0 (TID 11). 1677
bytes result sent to driver
17/10/26 15:47:55 INFO TaskSetManager: Finished task 1.0 in stage 6.0 (TID 11)
in 13 ms on localhost (executor driver) (1/2)
17/10/26 15:47:55 INFO Executor: Finished task 0.0 in stage 6.0 (TID 10). 1691
bytes result sent to driver
17/10/26 15:47:55 INFO TaskSetManager: Finished task 0.0 in stage 6.0 (TID 10)
in 16 ms on localhost (executor driver) (2/2)
17/10/26 15:47:55 INFO TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have
all completed, from pool
17/10/26 15:47:55 INFO DAGScheduler: ResultStage 6 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:97) finished in 0,017 s
17/10/26 15:47:55 INFO DAGScheduler: Job 3 finished: collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:97, took 0,067008 s
anger :  166
lyon :  193
```

```
marseilles_1 :  284
nice :  203
paris_2 :  642
paris_3 :  330
troyes :  214
marseilles_2 :  231
nantes :  207
orlean :  196
paris_1 :  596
rennes :  180
toulouse :  177

The store that achieves the best performance in each month:
17/10/26 15:47:55 INFO SparkContext: Starting job: collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:102
17/10/26 15:47:55 INFO DAGScheduler: Registering RDD 14 (reduceByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:62)
17/10/26 15:47:55 INFO DAGScheduler: Got job 4 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:102) with 2 output partitions
17/10/26 15:47:55 INFO DAGScheduler: Final stage: ResultStage 8 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:102)
17/10/26 15:47:55 INFO DAGScheduler: Parents of final stage:
List(ShuffleMapStage 7)
17/10/26 15:47:55 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 7)
17/10/26 15:47:55 INFO DAGScheduler: Submitting ShuffleMapStage 7
(PairwiseRDD[14] at reduceByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:62), which has no missing parents
17/10/26 15:47:55 INFO MemoryStore: Block broadcast_7 stored as values in
memory (estimated size 11.2 KB, free 365.9 MB)
17/10/26 15:47:55 INFO MemoryStore: Block broadcast_7_piece0 stored as bytes in
memory (estimated size 6.9 KB, free 365.9 MB)
17/10/26 15:47:55 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on
10.4.186.168:59440 (size: 6.9 KB, free: 366.2 MB)
17/10/26 15:47:55 INFO SparkContext: Created broadcast 7 from broadcast at
DAGScheduler.scala:1006
17/10/26 15:47:55 INFO DAGScheduler: Submitting 2 missing tasks from
ShuffleMapStage 7 (PairwiseRDD[14] at reduceByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:62) (first 15 tasks are for
partitions Vector(0, 1))
17/10/26 15:47:55 INFO TaskSchedulerImpl: Adding task set 7.0 with 2 tasks
17/10/26 15:47:55 INFO TaskSetManager: Starting task 0.0 in stage 7.0 (TID 12,
localhost, executor driver, partition 0, PROCESS_LOCAL, 5416 bytes)
17/10/26 15:47:55 INFO TaskSetManager: Starting task 1.0 in stage 7.0 (TID 13,
localhost, executor driver, partition 1, PROCESS_LOCAL, 5337 bytes)
17/10/26 15:47:55 INFO Executor: Running task 0.0 in stage 7.0 (TID 12)
17/10/26 15:47:55 INFO Executor: Running task 1.0 in stage 7.0 (TID 13)
17/10/26 15:47:55 INFO BlockManager: Found block rdd_2_1 locally
17/10/26 15:47:55 INFO BlockManager: Found block rdd_2_0 locally
17/10/26 15:47:55 INFO PythonRunner: Times: total = 3, boot = -31, init = 33,
finish = 1
17/10/26 15:47:55 INFO PythonRunner: Times: total = 4, boot = -35, init = 39,
finish = 0
```

```
17/10/26 15:47:55 INFO Executor: Finished task 1.0 in stage 7.0 (TID 13). 1525
bytes result sent to driver
17/10/26 15:47:55 INFO TaskSetManager: Finished task 1.0 in stage 7.0 (TID 13)
in 27 ms on localhost (executor driver) (1/2)
17/10/26 15:47:55 INFO Executor: Finished task 0.0 in stage 7.0 (TID 12). 1525
bytes result sent to driver
17/10/26 15:47:55 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID 12)
in 31 ms on localhost (executor driver) (2/2)
17/10/26 15:47:55 INFO TaskSchedulerImpl: Removed TaskSet 7.0, whose tasks have
all completed, from pool
17/10/26 15:47:55 INFO DAGScheduler: ShuffleMapStage 7 (reduceByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:62) finished in 0,032 s
17/10/26 15:47:55 INFO DAGScheduler: looking for newly runnable stages
17/10/26 15:47:55 INFO DAGScheduler: running: Set()
17/10/26 15:47:55 INFO DAGScheduler: waiting: Set(ResultStage 8)
17/10/26 15:47:55 INFO DAGScheduler: failed: Set()
17/10/26 15:47:55 INFO DAGScheduler: Submitting ResultStage 8 (PythonRDD[23] at
collect at /Users/thomas/Desktop/ING5/Spark/lab1/main.py:102), which has no
missing parents
17/10/26 15:47:55 INFO MemoryStore: Block broadcast_8 stored as values in
memory (estimated size 6.8 KB, free 365.9 MB)
17/10/26 15:47:55 INFO MemoryStore: Block broadcast_8_piece0 stored as bytes in
memory (estimated size 4.4 KB, free 365.9 MB)
17/10/26 15:47:55 INFO BlockManagerInfo: Added broadcast_8_piece0 in memory on
10.4.186.168:59440 (size: 4.4 KB, free: 366.2 MB)
17/10/26 15:47:55 INFO SparkContext: Created broadcast 8 from broadcast at
DAGScheduler.scala:1006
17/10/26 15:47:55 INFO DAGScheduler: Submitting 2 missing tasks from
ResultStage 8 (PythonRDD[23] at collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:102) (first 15 tasks are for
partitions Vector(0, 1))
17/10/26 15:47:55 INFO TaskSchedulerImpl: Adding task set 8.0 with 2 tasks
17/10/26 15:47:55 INFO TaskSetManager: Starting task 0.0 in stage 8.0 (TID 14,
localhost, executor driver, partition 0, ANY, 4621 bytes)
17/10/26 15:47:55 INFO TaskSetManager: Starting task 1.0 in stage 8.0 (TID 15,
localhost, executor driver, partition 1, ANY, 4621 bytes)
17/10/26 15:47:55 INFO Executor: Running task 1.0 in stage 8.0 (TID 15)
17/10/26 15:47:55 INFO Executor: Running task 0.0 in stage 8.0 (TID 14)
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks
out of 2 blocks
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in
1 ms
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks
out of 2 blocks
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in
0 ms
17/10/26 15:47:55 INFO PythonRunner: Times: total = 4, boot = -19, init = 22,
finish = 1
17/10/26 15:47:55 INFO PythonRunner: Times: total = 3, boot = -21, init = 24,
finish = 0
17/10/26 15:47:55 INFO Executor: Finished task 0.0 in stage 8.0 (TID 14). 1610
bytes result sent to driver
```

```
17/10/26 15:47:55 INFO TaskSetManager: Finished task 0.0 in stage 8.0 (TID 14)
in 13 ms on localhost (executor driver) (1/2)
17/10/26 15:47:55 INFO Executor: Finished task 1.0 in stage 8.0 (TID 15). 1799
bytes result sent to driver
17/10/26 15:47:55 INFO TaskSetManager: Finished task 1.0 in stage 8.0 (TID 15)
in 13 ms on localhost (executor driver) (2/2)
17/10/26 15:47:55 INFO TaskSchedulerImpl: Removed TaskSet 8.0, whose tasks have
all completed, from pool
17/10/26 15:47:55 INFO DAGScheduler: ResultStage 8 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:102) finished in 0,015 s
17/10/26 15:47:55 INFO DAGScheduler: Job 4 finished: collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:102, took 0,063869 s
APR :  paris_1
MAY :  paris_2
AUG :  paris_2
JAN :  paris_1
FEB :  paris_2
MAR :  paris_2
JUN :  paris_2
JUL :  paris_1
SEP :  paris_2
OCT :  paris_1
NOV :  paris_2
DEC :  paris_1

Average income of the shop in France (on a 1 year data) per month:
17/10/26 15:47:55 INFO SparkContext: Starting job: collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:107
17/10/26 15:47:55 INFO DAGScheduler: Registering RDD 18 (groupByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:74)
17/10/26 15:47:55 INFO DAGScheduler: Got job 5 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:107) with 2 output partitions
17/10/26 15:47:55 INFO DAGScheduler: Final stage: ResultStage 10 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:107)
17/10/26 15:47:55 INFO DAGScheduler: Parents of final stage:
List(ShuffleMapStage 9)
17/10/26 15:47:55 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 9)
17/10/26 15:47:55 INFO DAGScheduler: Submitting ShuffleMapStage 9
(PairwiseRDD[18] at groupByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:74), which has no missing parents
17/10/26 15:47:55 INFO MemoryStore: Block broadcast_9 stored as values in
memory (estimated size 11.3 KB, free 365.9 MB)
17/10/26 15:47:55 INFO MemoryStore: Block broadcast_9_piece0 stored as bytes in
memory (estimated size 7.0 KB, free 365.9 MB)
17/10/26 15:47:55 INFO BlockManagerInfo: Added broadcast_9_piece0 in memory on
10.4.186.168:59440 (size: 7.0 KB, free: 366.2 MB)
17/10/26 15:47:55 INFO SparkContext: Created broadcast 9 from broadcast at
DAGScheduler.scala:1006
17/10/26 15:47:55 INFO DAGScheduler: Submitting 2 missing tasks from
ShuffleMapStage 9 (PairwiseRDD[18] at groupByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:74) (first 15 tasks are for
partitions Vector(0, 1))
```

```
17/10/26 15:47:55 INFO TaskSchedulerImpl: Adding task set 9.0 with 2 tasks
17/10/26 15:47:55 INFO TaskSetManager: Starting task 0.0 in stage 9.0 (TID 16,
localhost, executor driver, partition 0, PROCESS_LOCAL, 5416 bytes)
17/10/26 15:47:55 INFO TaskSetManager: Starting task 1.0 in stage 9.0 (TID 17,
localhost, executor driver, partition 1, PROCESS_LOCAL, 5337 bytes)
17/10/26 15:47:55 INFO Executor: Running task 0.0 in stage 9.0 (TID 16)
17/10/26 15:47:55 INFO Executor: Running task 1.0 in stage 9.0 (TID 17)
17/10/26 15:47:55 INFO BlockManager: Found block rdd_2_1 locally
17/10/26 15:47:55 INFO BlockManager: Found block rdd_2_0 locally
17/10/26 15:47:55 INFO PythonRunner: Times: total = 2, boot = -30, init = 32,
finish = 0
17/10/26 15:47:55 INFO PythonRunner: Times: total = 2, boot = -27, init = 29,
finish = 0
17/10/26 15:47:55 INFO Executor: Finished task 0.0 in stage 9.0 (TID 16). 1525
bytes result sent to driver
17/10/26 15:47:55 INFO TaskSetManager: Finished task 0.0 in stage 9.0 (TID 16)
in 23 ms on localhost (executor driver) (1/2)
17/10/26 15:47:55 INFO Executor: Finished task 1.0 in stage 9.0 (TID 17). 1525
bytes result sent to driver
17/10/26 15:47:55 INFO TaskSetManager: Finished task 1.0 in stage 9.0 (TID 17)
in 29 ms on localhost (executor driver) (2/2)
17/10/26 15:47:55 INFO TaskSchedulerImpl: Removed TaskSet 9.0, whose tasks have
all completed, from pool
17/10/26 15:47:55 INFO DAGScheduler: ShuffleMapStage 9 (groupByKey at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:74) finished in 0,031 s
17/10/26 15:47:55 INFO DAGScheduler: looking for newly runnable stages
17/10/26 15:47:55 INFO DAGScheduler: running: Set()
17/10/26 15:47:55 INFO DAGScheduler: waiting: Set(ResultStage 10)
17/10/26 15:47:55 INFO DAGScheduler: failed: Set()
17/10/26 15:47:55 INFO DAGScheduler: Submitting ResultStage 10 (PythonRDD[24]
at collect at /Users/thomas/Desktop/ING5/Spark/lab1/main.py:107), which has no
missing parents
17/10/26 15:47:55 INFO MemoryStore: Block broadcast_10 stored as values in
memory (estimated size 7.1 KB, free 365.9 MB)
17/10/26 15:47:55 INFO MemoryStore: Block broadcast_10_piece0 stored as bytes
in memory (estimated size 4.6 KB, free 365.9 MB)
17/10/26 15:47:55 INFO BlockManagerInfo: Added broadcast_10_piece0 in memory on
10.4.186.168:59440 (size: 4.6 KB, free: 366.2 MB)
17/10/26 15:47:55 INFO SparkContext: Created broadcast 10 from broadcast at
DAGScheduler.scala:1006
17/10/26 15:47:55 INFO DAGScheduler: Submitting 2 missing tasks from
ResultStage 10 (PythonRDD[24] at collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:107) (first 15 tasks are for
partitions Vector(0, 1))
17/10/26 15:47:55 INFO TaskSchedulerImpl: Adding task set 10.0 with 2 tasks
17/10/26 15:47:55 INFO TaskSetManager: Starting task 0.0 in stage 10.0 (TID 18,
localhost, executor driver, partition 0, ANY, 4621 bytes)
17/10/26 15:47:55 INFO TaskSetManager: Starting task 1.0 in stage 10.0 (TID 19,
localhost, executor driver, partition 1, ANY, 4621 bytes)
17/10/26 15:47:55 INFO Executor: Running task 0.0 in stage 10.0 (TID 18)
17/10/26 15:47:55 INFO Executor: Running task 1.0 in stage 10.0 (TID 19)
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks
```

```
out of 2 blocks
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in
1 ms
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Getting 2 non-empty blocks
out of 2 blocks
17/10/26 15:47:55 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in
0 ms
17/10/26 15:47:55 INFO PythonRunner: Times: total = 2, boot = -23, init = 25,
finish = 0
17/10/26 15:47:55 INFO Executor: Finished task 1.0 in stage 10.0 (TID 19). 1821
bytes result sent to driver
17/10/26 15:47:55 INFO PythonRunner: Times: total = 3, boot = -28, init = 30,
finish = 1
17/10/26 15:47:55 INFO TaskSetManager: Finished task 1.0 in stage 10.0 (TID 19)
in 13 ms on localhost (executor driver) (1/2)
17/10/26 15:47:55 INFO Executor: Finished task 0.0 in stage 10.0 (TID 18). 1607
bytes result sent to driver
17/10/26 15:47:55 INFO TaskSetManager: Finished task 0.0 in stage 10.0 (TID 18)
in 19 ms on localhost (executor driver) (2/2)
17/10/26 15:47:55 INFO TaskSchedulerImpl: Removed TaskSet 10.0, whose tasks
have all completed, from pool
17/10/26 15:47:55 INFO DAGScheduler: ResultStage 10 (collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:107) finished in 0,021 s
17/10/26 15:47:55 INFO DAGScheduler: Job 5 finished: collect at
/Users/thomas/Desktop/ING5/Spark/lab1/main.py:107, took 0,069113 s
APR :  20.23076923076923
MAY :  22.46153846153846
AUG :  23.076923076923077
JAN :  20.76923076923077
FEB :  19.153846153846153
MAR :  17.53846153846154
JUN :  27.846153846153847
JUL :  21.692307692307693
SEP :  25.53846153846154
OCT :  26.53846153846154
NOV :  24.53846153846154
DEC :  29.0
17/10/26 15:47:55 INFO SparkContext: Invoking stop() from shutdown hook
17/10/26 15:47:55 INFO SparkUI: Stopped Spark web UI at
http://10.4.186.168:4040
17/10/26 15:47:55 INFO MapOutputTrackerMasterEndpoint:
MapOutputTrackerMasterEndpoint stopped!
17/10/26 15:47:55 INFO MemoryStore: MemoryStore cleared
17/10/26 15:47:55 INFO BlockManager: BlockManager stopped
17/10/26 15:47:55 INFO BlockManagerMaster: BlockManagerMaster stopped
17/10/26 15:47:55 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint:
OutputCommitCoordinator stopped!
17/10/26 15:47:55 INFO SparkContext: Successfully stopped SparkContext
17/10/26 15:47:55 INFO ShutdownHookManager: Shutdown hook called
17/10/26 15:47:55 INFO ShutdownHookManager: Deleting directory
/private/var/folders/6_/n45kv5s50_j7lphv4rz2795r0000gn/T/spark-d0cd7b11-34bc-
48e9-b114-2ef4d64da860/pyspark-337b6096-dd58-42ab-af6a-4effa265e4e2
```

```
17/10/26 15:47:55 INFO ShutdownHookManager: Deleting directory
/private/var/folders/6_/n45kv5s50_j7lphv4rz2795r0000gn/T/spark-d0cd7b11-34bc-
48e9-b114-2ef4d64da860
```