Estudiante: Junior Efraín Franco Pérez.

Taller Práctico Realizado En Módulo V: Implementación De Proyectos Bigdata.

Importar las librerías, crear la sesión

In [187…

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, sum, desc, format_number
import pandas as pd

# Crear sesión de Spark
spark = SparkSession.builder.appName("Taller PySpark RDD UGB").getOrCreate()
```

```
23/07/08 18:10:32 WARN SparkSession: Using an existing Spark session; only runtime S
QL configurations will take effect.
```

Obtener los archivos de Hadoop.

Importante: Se cargan los archivos como RDD utilizando el contexto de Spark (sc).

In [188…

```
from pyspark.sql import SparkSession

# Crear una instancia de SparkSession
spark = SparkSession.builder.getOrCreate()

# Obtener el SparkContext
sc = spark.sparkContext

# Cargar archivos que tengo en hadoop como RDD
rdd_canal = sc.textFile("/datos/CanalDeVenta.csv")
rdd_cliente = sc.textFile("/datos/Cliente.csv")
rdd_empleado = sc.textFile("/datos/Empleado.csv")
rdd_producto = sc.textFile("/datos/Producto.csv")
rdd_sucursal = sc.textFile("/datos/Sucursal.csv")
rdd_venta = sc.textFile("/datos/Ventas.csv")
rdd_DIMDATE_DATAONLY = sc.textFile("/datos/dim/DIMDATE-DATAONLY.csv")
rdd_DIMDATE = sc.textFile("/datos/dim/DIMDATE.csv")
rdd_DIMTIME_DATAONLY = sc.textFile("/datos/dim/DIMTIME-DATAONLY.csv")
rdd_DIMTIME = sc.textFile("/datos/dim/DIMTIME.csv")
```

```
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1515 stored as values in memory
(estimated size 499.7 KiB, free 348.0 MiB)
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1515_piece0 stored as bytes in m
emory (estimated size 52.7 KiB, free 348.0 MiB)
23/07/08 18:10:45 INFO BlockManagerInfo: Added broadcast_1515_piece0 in memory on 17
2.30.115.138:43839 (size: 52.7 KiB, free: 364.5 MiB)
23/07/08 18:10:45 INFO SparkContext: Created broadcast 1515 from textFile at <unknow
n>:0
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1516 stored as values in memory
(estimated size 499.7 KiB, free 347.5 MiB)
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1516_piece0 stored as bytes in m
emory (estimated size 52.7 KiB, free 347.4 MiB)
23/07/08 18:10:45 INFO BlockManagerInfo: Added broadcast_1516_piece0 in memory on 17
2.30.115.138:43839 (size: 52.7 KiB, free: 364.5 MiB)
23/07/08 18:10:45 INFO SparkContext: Created broadcast 1516 from textFile at <unknow
n>:0
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1517 stored as values in memory
(estimated size 499.7 KiB, free 346.9 MiB)
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1517_piece0 stored as bytes in m
emory (estimated size 52.7 KiB, free 346.9 MiB)
23/07/08 18:10:45 INFO BlockManagerInfo: Added broadcast_1517_piece0 in memory on 17
2.30.115.138:43839 (size: 52.7 KiB, free: 364.4 MiB)
23/07/08 18:10:45 INFO SparkContext: Created broadcast 1517 from textFile at <unknow
n>:0
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1518 stored as values in memory
(estimated size 499.7 KiB, free 346.4 MiB)
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1518_piece0 stored as bytes in m
emory (estimated size 52.7 KiB, free 346.3 MiB)
23/07/08 18:10:45 INFO BlockManagerInfo: Added broadcast_1518_piece0 in memory on 17
2.30.115.138:43839 (size: 52.7 KiB, free: 364.4 MiB)
23/07/08 18:10:45 INFO SparkContext: Created broadcast 1518 from textFile at <unknow
n>:0
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1519 stored as values in memory
(estimated size 499.7 KiB, free 345.9 MiB)
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1519_piece0 stored as bytes in m
emory (estimated size 52.7 KiB, free 345.8 MiB)
23/07/08 18:10:45 INFO BlockManagerInfo: Added broadcast_1519_piece0 in memory on 17
2.30.115.138:43839 (size: 52.7 KiB, free: 364.3 MiB)
23/07/08 18:10:45 INFO SparkContext: Created broadcast 1519 from textFile at <unknow
n>:0
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1520 stored as values in memory
(estimated size 499.7 KiB, free 345.3 MiB)
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1520_piece0 stored as bytes in m
emory (estimated size 52.7 KiB, free 345.3 MiB)
23/07/08 18:10:45 INFO BlockManagerInfo: Added broadcast_1520_piece0 in memory on 17
2.30.115.138:43839 (size: 52.7 KiB, free: 364.3 MiB)
23/07/08 18:10:45 INFO SparkContext: Created broadcast 1520 from textFile at <unknow
n>:0
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1521 stored as values in memory
(estimated size 499.7 KiB, free 344.8 MiB)
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1521_piece0 stored as bytes in m
emory (estimated size 52.7 KiB, free 344.7 MiB)
23/07/08 18:10:45 INFO BlockManagerInfo: Added broadcast_1521_piece0 in memory on 17
2.30.115.138:43839 (size: 52.7 KiB, free: 364.2 MiB)
23/07/08 18:10:45 INFO SparkContext: Created broadcast 1521 from textFile at <unknow
n>:0
```

```
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1522 stored as values in memory
(estimated size 499.7 KiB, free 344.2 MiB)
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1522_piece0 stored as bytes in m
emory (estimated size 52.7 KiB, free 344.2 MiB)
23/07/08 18:10:45 INFO BlockManagerInfo: Added broadcast_1522_piece0 in memory on 17
2.30.115.138:43839 (size: 52.7 KiB, free: 364.2 MiB)
23/07/08 18:10:45 INFO SparkContext: Created broadcast 1522 from textFile at <unknow
n>:0
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1523 stored as values in memory
(estimated size 499.7 KiB, free 343.7 MiB)
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1523_piece0 stored as bytes in m
emory (estimated size 52.7 KiB, free 343.7 MiB)
23/07/08 18:10:45 INFO BlockManagerInfo: Added broadcast_1523_piece0 in memory on 17
2.30.115.138:43839 (size: 52.7 KiB, free: 364.1 MiB)
23/07/08 18:10:45 INFO SparkContext: Created broadcast 1523 from textFile at <unknow
n>:0
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1524 stored as values in memory
(estimated size 499.7 KiB, free 343.2 MiB)
23/07/08 18:10:45 INFO MemoryStore: Block broadcast_1524_piece0 stored as bytes in m
emory (estimated size 52.7 KiB, free 343.1 MiB)
23/07/08 18:10:45 INFO BlockManagerInfo: Added broadcast_1524_piece0 in memory on 17
2.30.115.138:43839 (size: 52.7 KiB, free: 364.1 MiB)
23/07/08 18:10:45 INFO SparkContext: Created broadcast 1524 from textFile at <unknow
n>:0
```

Obtener la primera fila de cada RDD como encabezado

In [189…

```
header_canal = rdd_canal.first()
header_cliente = rdd_cliente.first()
header_empleado = rdd_empleado.first()
header_producto = rdd_producto.first()
header_sucursal = rdd_sucursal.first()
header_venta = rdd_venta.first()
header_DIMDATE_DATAONLY = rdd_DIMDATE_DATAONLY.first()
header_DIMDATE = rdd_DIMDATE.first()
header_DIMTIME_DATAONLY = rdd_DIMTIME_DATAONLY.first()
header_DIMTIME = rdd_DIMTIME.first()
```

```
23/07/08 18:10:49 INFO FileInputFormat: Total input files to process : 1
23/07/08 18:10:49 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:49 INFO DAGScheduler: Got job 1239 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:49 INFO DAGScheduler: Final stage: ResultStage 2680 (runJob at Python
RDD.scala:179)
23/07/08 18:10:49 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:49 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:49 INFO DAGScheduler: Submitting ResultStage 2680 (PythonRDD[3846] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:49 INFO MemoryStore: Block broadcast_1525 stored as values in memory
(estimated size 7.8 KiB, free 343.1 MiB)
23/07/08 18:10:49 INFO MemoryStore: Block broadcast_1525_piece0 stored as bytes in m
emory (estimated size 4.8 KiB, free 343.1 MiB)
23/07/08 18:10:49 INFO BlockManagerInfo: Added broadcast_1525_piece0 in memory on 17
2.30.115.138:43839 (size: 4.8 KiB, free: 364.1 MiB)
23/07/08 18:10:49 INFO SparkContext: Created broadcast 1525 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:49 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 268
0 (PythonRDD[3846] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:49 INFO TaskSchedulerImpl: Adding task set 2680.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:49 INFO TaskSetManager: Starting task 0.0 in stage 2680.0 (TID 1705)
(172.30.115.138, executor driver, partition 0, ANY, 7423 bytes)
23/07/08 18:10:49 INFO Executor: Running task 0.0 in stage 2680.0 (TID 1705)
23/07/08 18:10:49 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/CanalDeVe
nta.csv:0+29
23/07/08 18:10:49 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:49 INFO PythonRunner: Times: total = 149, boot = 11, init = 137, fini
sh = 1
23/07/08 18:10:49 INFO Executor: Finished task 0.0 in stage 2680.0 (TID 1705). 1428
bytes result sent to driver
23/07/08 18:10:49 INFO TaskSetManager: Finished task 0.0 in stage 2680.0 (TID 1705)
in 164 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:49 INFO TaskSchedulerImpl: Removed TaskSet 2680.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:49 INFO DAGScheduler: ResultStage 2680 (runJob at PythonRDD.scala:17
9) finished in 0.176 s
23/07/08 18:10:49 INFO DAGScheduler: Job 1239 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:49 INFO TaskSchedulerImpl: Killing all running tasks in stage 2680: S
tage finished
23/07/08 18:10:49 INFO DAGScheduler: Job 1239 finished: runJob at PythonRDD.scala:17
9, took 0.179837 s
23/07/08 18:10:49 INFO FileInputFormat: Total input files to process : 1
23/07/08 18:10:49 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:49 INFO DAGScheduler: Got job 1240 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:49 INFO DAGScheduler: Final stage: ResultStage 2681 (runJob at Python
RDD.scala:179)
23/07/08 18:10:49 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:49 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:49 INFO DAGScheduler: Submitting ResultStage 2681 (PythonRDD[3847] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:49 INFO MemoryStore: Block broadcast_1526 stored as values in memory
```

```
(estimated size 7.8 KiB, free 343.1 MiB)
23/07/08 18:10:49 INFO MemoryStore: Block broadcast_1526_piece0 stored as bytes in m
emory (estimated size 4.8 KiB, free 343.1 MiB)
23/07/08 18:10:49 INFO BlockManagerInfo: Added broadcast_1526_piece0 in memory on 17
2.30.115.138:43839 (size: 4.8 KiB, free: 364.1 MiB)
23/07/08 18:10:49 INFO SparkContext: Created broadcast 1526 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:49 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 268
1 (PythonRDD[3847] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:49 INFO TaskSchedulerImpl: Adding task set 2681.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:49 INFO TaskSetManager: Starting task 0.0 in stage 2681.0 (TID 1706)
(172.30.115.138, executor driver, partition 0, ANY, 7418 bytes)
23/07/08 18:10:49 INFO Executor: Running task 0.0 in stage 2681.0 (TID 1706)
23/07/08 18:10:49 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Cliente.c
sv:0+214936
23/07/08 18:10:49 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:49 INFO PythonRunner: Times: total = 120, boot = 4, init = 115, finis
h = 1
23/07/08 18:10:49 INFO Executor: Finished task 0.0 in stage 2681.0 (TID 1706). 1484
bytes result sent to driver
23/07/08 18:10:49 INFO TaskSetManager: Finished task 0.0 in stage 2681.0 (TID 1706)
in 128 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:49 INFO TaskSchedulerImpl: Removed TaskSet 2681.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:49 INFO DAGScheduler: ResultStage 2681 (runJob at PythonRDD.scala:17
9) finished in 0.136 s
23/07/08 18:10:49 INFO DAGScheduler: Job 1240 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:49 INFO TaskSchedulerImpl: Killing all running tasks in stage 2681: S
tage finished
23/07/08 18:10:49 INFO DAGScheduler: Job 1240 finished: runJob at PythonRDD.scala:17
9, took 0.137573 s
23/07/08 18:10:49 INFO FileInputFormat: Total input files to process : 1
23/07/08 18:10:49 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:49 INFO DAGScheduler: Got job 1241 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:49 INFO DAGScheduler: Final stage: ResultStage 2682 (runJob at Python
RDD.scala:179)
23/07/08 18:10:49 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:49 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:49 INFO DAGScheduler: Submitting ResultStage 2682 (PythonRDD[3848] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:49 INFO MemoryStore: Block broadcast_1527 stored as values in memory
(estimated size 7.8 KiB, free 343.1 MiB)
23/07/08 18:10:49 INFO MemoryStore: Block broadcast_1527_piece0 stored as bytes in m
emory (estimated size 4.8 KiB, free 343.1 MiB)
23/07/08 18:10:49 INFO BlockManagerInfo: Added broadcast_1527_piece0 in memory on 17
2.30.115.138:43839 (size: 4.8 KiB, free: 364.1 MiB)
23/07/08 18:10:49 INFO SparkContext: Created broadcast 1527 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:49 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 268
2 (PythonRDD[3848] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:49 INFO TaskSchedulerImpl: Adding task set 2682.0 with 1 tasks resour
```

```
ce profile 0
23/07/08 18:10:49 INFO TaskSetManager: Starting task 0.0 in stage 2682.0 (TID 1707)
(172.30.115.138, executor driver, partition 0, ANY, 7419 bytes)
23/07/08 18:10:49 INFO Executor: Running task 0.0 in stage 2682.0 (TID 1707)
23/07/08 18:10:49 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Empleado.
csv:0+8119
23/07/08 18:10:49 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:49 INFO PythonRunner: Times: total = 115, boot = 4, init = 111, finis
h = 0
23/07/08 18:10:49 INFO Executor: Finished task 0.0 in stage 2682.0 (TID 1707). 1467
bytes result sent to driver
23/07/08 18:10:49 INFO TaskSetManager: Finished task 0.0 in stage 2682.0 (TID 1707)
in 121 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:49 INFO TaskSchedulerImpl: Removed TaskSet 2682.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:49 INFO DAGScheduler: ResultStage 2682 (runJob at PythonRDD.scala:17
9) finished in 0.126 s
23/07/08 18:10:49 INFO DAGScheduler: Job 1241 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:49 INFO TaskSchedulerImpl: Killing all running tasks in stage 2682: S
tage finished
23/07/08 18:10:49 INFO DAGScheduler: Job 1241 finished: runJob at PythonRDD.scala:17
9, took 0.127470 s
23/07/08 18:10:49 INFO FileInputFormat: Total input files to process : 1
23/07/08 18:10:49 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:49 INFO DAGScheduler: Got job 1242 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:49 INFO DAGScheduler: Final stage: ResultStage 2683 (runJob at Python
RDD.scala:179)
23/07/08 18:10:49 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:49 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:49 INFO DAGScheduler: Submitting ResultStage 2683 (PythonRDD[3849] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:49 INFO MemoryStore: Block broadcast_1528 stored as values in memory
(estimated size 7.8 KiB, free 343.1 MiB)
23/07/08 18:10:49 INFO MemoryStore: Block broadcast_1528_piece0 stored as bytes in m
emory (estimated size 4.8 KiB, free 343.1 MiB)
23/07/08 18:10:49 INFO BlockManagerInfo: Added broadcast_1528_piece0 in memory on 17
2.30.115.138:43839 (size: 4.8 KiB, free: 364.1 MiB)
23/07/08 18:10:49 INFO SparkContext: Created broadcast 1528 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:49 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 268
3 (PythonRDD[3849] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:49 INFO TaskSchedulerImpl: Adding task set 2683.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:49 INFO TaskSetManager: Starting task 0.0 in stage 2683.0 (TID 1708)
(172.30.115.138, executor driver, partition 0, ANY, 7419 bytes)
23/07/08 18:10:49 INFO Executor: Running task 0.0 in stage 2683.0 (TID 1708)
23/07/08 18:10:49 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Producto.
csv:0+8440
23/07/08 18:10:49 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:49 INFO PythonRunner: Times: total = 115, boot = 4, init = 111, finis
h = 0
23/07/08 18:10:49 INFO Executor: Finished task 0.0 in stage 2683.0 (TID 1708). 1444
bytes result sent to driver
```

```
23/07/08 18:10:49 INFO TaskSetManager: Finished task 0.0 in stage 2683.0 (TID 1708)
in 122 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:49 INFO TaskSchedulerImpl: Removed TaskSet 2683.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:49 INFO DAGScheduler: ResultStage 2683 (runJob at PythonRDD.scala:17
9) finished in 0.127 s
23/07/08 18:10:49 INFO DAGScheduler: Job 1242 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:49 INFO TaskSchedulerImpl: Killing all running tasks in stage 2683: S
tage finished
23/07/08 18:10:49 INFO DAGScheduler: Job 1242 finished: runJob at PythonRDD.scala:17
9, took 0.128620 s
23/07/08 18:10:49 INFO FileInputFormat: Total input files to process : 1
23/07/08 18:10:49 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:49 INFO DAGScheduler: Got job 1243 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:49 INFO DAGScheduler: Final stage: ResultStage 2684 (runJob at Python
RDD.scala:179)
23/07/08 18:10:49 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:49 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:49 INFO DAGScheduler: Submitting ResultStage 2684 (PythonRDD[3850] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:49 INFO MemoryStore: Block broadcast_1529 stored as values in memory
(estimated size 7.8 KiB, free 343.1 MiB)
23/07/08 18:10:49 INFO MemoryStore: Block broadcast_1529_piece0 stored as bytes in m
emory (estimated size 4.8 KiB, free 343.1 MiB)
23/07/08 18:10:49 INFO BlockManagerInfo: Added broadcast_1529_piece0 in memory on 17
2.30.115.138:43839 (size: 4.8 KiB, free: 364.1 MiB)
23/07/08 18:10:49 INFO SparkContext: Created broadcast 1529 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:49 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 268
4 (PythonRDD[3850] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:49 INFO TaskSchedulerImpl: Adding task set 2684.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:49 INFO TaskSetManager: Starting task 0.0 in stage 2684.0 (TID 1709)
(172.30.115.138, executor driver, partition 0, ANY, 7419 bytes)
23/07/08 18:10:49 INFO Executor: Running task 0.0 in stage 2684.0 (TID 1709)
23/07/08 18:10:49 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Sucursal.
csv:0+1266
23/07/08 18:10:49 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:50 INFO PythonRunner: Times: total = 118, boot = 5, init = 113, finis
h = 0
23/07/08 18:10:50 INFO Executor: Finished task 0.0 in stage 2684.0 (TID 1709). 1468
bytes result sent to driver
23/07/08 18:10:50 INFO TaskSetManager: Finished task 0.0 in stage 2684.0 (TID 1709)
in 124 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:50 INFO TaskSchedulerImpl: Removed TaskSet 2684.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:50 INFO DAGScheduler: ResultStage 2684 (runJob at PythonRDD.scala:17
9) finished in 0.129 s
23/07/08 18:10:50 INFO DAGScheduler: Job 1243 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:50 INFO TaskSchedulerImpl: Killing all running tasks in stage 2684: S
tage finished
23/07/08 18:10:50 INFO DAGScheduler: Job 1243 finished: runJob at PythonRDD.scala:17
```

```
9, took 0.130207 s
23/07/08 18:10:50 INFO FileInputFormat: Total input files to process : 1
23/07/08 18:10:50 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:50 INFO DAGScheduler: Got job 1244 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:50 INFO DAGScheduler: Final stage: ResultStage 2685 (runJob at Python
RDD.scala:179)
23/07/08 18:10:50 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:50 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:50 INFO DAGScheduler: Submitting ResultStage 2685 (PythonRDD[3851] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:50 INFO MemoryStore: Block broadcast_1530 stored as values in memory
(estimated size 7.8 KiB, free 343.0 MiB)
23/07/08 18:10:50 INFO MemoryStore: Block broadcast_1530_piece0 stored as bytes in m
emory (estimated size 4.8 KiB, free 343.0 MiB)
23/07/08 18:10:50 INFO BlockManagerInfo: Added broadcast_1530_piece0 in memory on 17
2.30.115.138:43839 (size: 4.8 KiB, free: 364.1 MiB)
23/07/08 18:10:50 INFO SparkContext: Created broadcast 1530 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:50 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 268
5 (PythonRDD[3851] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:50 INFO TaskSchedulerImpl: Adding task set 2685.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:50 INFO TaskSetManager: Starting task 0.0 in stage 2685.0 (TID 1710)
(172.30.115.138, executor driver, partition 0, ANY, 7417 bytes)
23/07/08 18:10:50 INFO Executor: Running task 0.0 in stage 2685.0 (TID 1710)
23/07/08 18:10:50 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:0+1310749
23/07/08 18:10:50 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:50 INFO PythonRunner: Times: total = 120, boot = 3, init = 117, finis
h = 0
23/07/08 18:10:50 INFO Executor: Finished task 0.0 in stage 2685.0 (TID 1710). 1504
bytes result sent to driver
23/07/08 18:10:50 INFO TaskSetManager: Finished task 0.0 in stage 2685.0 (TID 1710)
in 126 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:50 INFO TaskSchedulerImpl: Removed TaskSet 2685.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:50 INFO DAGScheduler: ResultStage 2685 (runJob at PythonRDD.scala:17
9) finished in 0.132 s
23/07/08 18:10:50 INFO DAGScheduler: Job 1244 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:50 INFO TaskSchedulerImpl: Killing all running tasks in stage 2685: S
tage finished
23/07/08 18:10:50 INFO DAGScheduler: Job 1244 finished: runJob at PythonRDD.scala:17
9, took 0.133508 s
23/07/08 18:10:50 INFO FileInputFormat: Total input files to process : 1
23/07/08 18:10:50 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:50 INFO DAGScheduler: Got job 1245 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:50 INFO DAGScheduler: Final stage: ResultStage 2686 (runJob at Python
RDD.scala:179)
23/07/08 18:10:50 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:50 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:50 INFO DAGScheduler: Submitting ResultStage 2686 (PythonRDD[3852] at
RDD at PythonRDD.scala:53), which has no missing parents
```

```
23/07/08 18:10:50 INFO MemoryStore: Block broadcast_1531 stored as values in memory
(estimated size 7.8 KiB, free 343.0 MiB)
23/07/08 18:10:50 INFO MemoryStore: Block broadcast_1531_piece0 stored as bytes in m
emory (estimated size 4.8 KiB, free 343.0 MiB)
23/07/08 18:10:50 INFO BlockManagerInfo: Added broadcast_1531_piece0 in memory on 17
2.30.115.138:43839 (size: 4.8 KiB, free: 364.1 MiB)
23/07/08 18:10:50 INFO SparkContext: Created broadcast 1531 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:50 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 268
6 (PythonRDD[3852] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:50 INFO TaskSchedulerImpl: Adding task set 2686.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:50 INFO TaskSetManager: Starting task 0.0 in stage 2686.0 (TID 1711)
(172.30.115.138, executor driver, partition 0, ANY, 7431 bytes)
23/07/08 18:10:50 INFO Executor: Running task 0.0 in stage 2686.0 (TID 1711)
23/07/08 18:10:50 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE-DATAONLY.csv:0+365617
23/07/08 18:10:50 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:50 INFO PythonRunner: Times: total = 120, boot = 3, init = 117, finis
h = 0
23/07/08 18:10:50 INFO Executor: Finished task 0.0 in stage 2686.0 (TID 1711). 1517
bytes result sent to driver
23/07/08 18:10:50 INFO TaskSetManager: Finished task 0.0 in stage 2686.0 (TID 1711)
in 127 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:50 INFO TaskSchedulerImpl: Removed TaskSet 2686.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:50 INFO DAGScheduler: ResultStage 2686 (runJob at PythonRDD.scala:17
9) finished in 0.132 s
23/07/08 18:10:50 INFO DAGScheduler: Job 1245 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:50 INFO TaskSchedulerImpl: Killing all running tasks in stage 2686: S
tage finished
23/07/08 18:10:50 INFO DAGScheduler: Job 1245 finished: runJob at PythonRDD.scala:17
9, took 0.133236 s
23/07/08 18:10:50 INFO FileInputFormat: Total input files to process : 1
23/07/08 18:10:50 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:50 INFO DAGScheduler: Got job 1246 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:50 INFO DAGScheduler: Final stage: ResultStage 2687 (runJob at Python
RDD.scala:179)
23/07/08 18:10:50 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:50 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:50 INFO DAGScheduler: Submitting ResultStage 2687 (PythonRDD[3853] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:50 INFO MemoryStore: Block broadcast_1532 stored as values in memory
(estimated size 7.8 KiB, free 343.0 MiB)
23/07/08 18:10:50 INFO MemoryStore: Block broadcast_1532_piece0 stored as bytes in m
emory (estimated size 4.8 KiB, free 343.0 MiB)
23/07/08 18:10:50 INFO BlockManagerInfo: Added broadcast_1532_piece0 in memory on 17
2.30.115.138:43839 (size: 4.8 KiB, free: 364.0 MiB)
23/07/08 18:10:50 INFO SparkContext: Created broadcast 1532 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:50 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 268
7 (PythonRDD[3853] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
```

```
23/07/08 18:10:50 INFO TaskSchedulerImpl: Adding task set 2687.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:50 INFO TaskSetManager: Starting task 0.0 in stage 2687.0 (TID 1712)
(172.30.115.138, executor driver, partition 0, ANY, 7422 bytes)
23/07/08 18:10:50 INFO Executor: Running task 0.0 in stage 2687.0 (TID 1712)
23/07/08 18:10:50 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:0+450275
23/07/08 18:10:50 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:50 INFO PythonRunner: Times: total = 115, boot = 4, init = 111, finis
h = 0
23/07/08 18:10:50 INFO Executor: Finished task 0.0 in stage 2687.0 (TID 1712). 1533
bytes result sent to driver
23/07/08 18:10:50 INFO TaskSetManager: Finished task 0.0 in stage 2687.0 (TID 1712)
in 123 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:50 INFO TaskSchedulerImpl: Removed TaskSet 2687.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:50 INFO DAGScheduler: ResultStage 2687 (runJob at PythonRDD.scala:17
9) finished in 0.128 s
23/07/08 18:10:50 INFO DAGScheduler: Job 1246 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:50 INFO TaskSchedulerImpl: Killing all running tasks in stage 2687: S
tage finished
23/07/08 18:10:50 INFO DAGScheduler: Job 1246 finished: runJob at PythonRDD.scala:17
9, took 0.129091 s
23/07/08 18:10:50 INFO FileInputFormat: Total input files to process : 1
23/07/08 18:10:50 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:50 INFO DAGScheduler: Got job 1247 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:50 INFO DAGScheduler: Final stage: ResultStage 2688 (runJob at Python
RDD.scala:179)
23/07/08 18:10:50 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:50 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:50 INFO DAGScheduler: Submitting ResultStage 2688 (PythonRDD[3854] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:50 INFO MemoryStore: Block broadcast_1533 stored as values in memory
(estimated size 7.8 KiB, free 343.0 MiB)
23/07/08 18:10:50 INFO MemoryStore: Block broadcast_1533_piece0 stored as bytes in m
emory (estimated size 4.8 KiB, free 343.0 MiB)
23/07/08 18:10:50 INFO BlockManagerInfo: Added broadcast_1533_piece0 in memory on 17
2.30.115.138:43839 (size: 4.8 KiB, free: 364.0 MiB)
23/07/08 18:10:50 INFO SparkContext: Created broadcast 1533 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:50 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 268
8 (PythonRDD[3854] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:50 INFO TaskSchedulerImpl: Adding task set 2688.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:50 INFO TaskSetManager: Starting task 0.0 in stage 2688.0 (TID 1713)
(172.30.115.138, executor driver, partition 0, ANY, 7431 bytes)
23/07/08 18:10:50 INFO Executor: Running task 0.0 in stage 2688.0 (TID 1713)
23/07/08 18:10:50 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMTI
ME-DATAONLY.csv:0+75866
23/07/08 18:10:50 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:50 INFO PythonRunner: Times: total = 115, boot = 4, init = 110, finis
h = 1
23/07/08 18:10:50 INFO Executor: Finished task 0.0 in stage 2688.0 (TID 1713). 1508
```

```
bytes result sent to driver
23/07/08 18:10:50 INFO TaskSetManager: Finished task 0.0 in stage 2688.0 (TID 1713)
in 124 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:50 INFO TaskSchedulerImpl: Removed TaskSet 2688.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:50 INFO DAGScheduler: ResultStage 2688 (runJob at PythonRDD.scala:17
9) finished in 0.129 s
23/07/08 18:10:50 INFO DAGScheduler: Job 1247 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:50 INFO TaskSchedulerImpl: Killing all running tasks in stage 2688: S
tage finished
23/07/08 18:10:50 INFO DAGScheduler: Job 1247 finished: runJob at PythonRDD.scala:17
9, took 0.130099 s
23/07/08 18:10:50 INFO FileInputFormat: Total input files to process : 1
23/07/08 18:10:50 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:50 INFO DAGScheduler: Got job 1248 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:50 INFO DAGScheduler: Final stage: ResultStage 2689 (runJob at Python
RDD.scala:179)
23/07/08 18:10:50 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:50 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:50 INFO DAGScheduler: Submitting ResultStage 2689 (PythonRDD[3855] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:50 INFO MemoryStore: Block broadcast_1534 stored as values in memory
(estimated size 7.8 KiB, free 343.0 MiB)
23/07/08 18:10:50 INFO MemoryStore: Block broadcast_1534_piece0 stored as bytes in m
emory (estimated size 4.8 KiB, free 343.0 MiB)
23/07/08 18:10:50 INFO BlockManagerInfo: Added broadcast_1534_piece0 in memory on 17
2.30.115.138:43839 (size: 4.8 KiB, free: 364.0 MiB)
23/07/08 18:10:50 INFO SparkContext: Created broadcast 1534 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:50 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 268
9 (PythonRDD[3855] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:50 INFO TaskSchedulerImpl: Adding task set 2689.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:50 INFO TaskSetManager: Starting task 0.0 in stage 2689.0 (TID 1714)
(172.30.115.138, executor driver, partition 0, ANY, 7422 bytes)
23/07/08 18:10:50 INFO Executor: Running task 0.0 in stage 2689.0 (TID 1714)
23/07/08 18:10:50 INFO BlockManagerInfo: Removed broadcast_1525_piece0 on 172.30.11
5.138:43839 in memory (size: 4.8 KiB, free: 364.0 MiB)
23/07/08 18:10:50 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMTI
ME.csv:0+22350
23/07/08 18:10:50 INFO BlockManagerInfo: Removed broadcast_1528_piece0 on 172.30.11
5.138:43839 in memory (size: 4.8 KiB, free: 364.0 MiB)
23/07/08 18:10:50 INFO BlockManagerInfo: Removed broadcast_1526_piece0 on 172.30.11
5.138:43839 in memory (size: 4.8 KiB, free: 364.1 MiB)
23/07/08 18:10:50 INFO BlockManagerInfo: Removed broadcast_1527_piece0 on 172.30.11
5.138:43839 in memory (size: 4.8 KiB, free: 364.1 MiB)
23/07/08 18:10:50 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:50 INFO BlockManagerInfo: Removed broadcast_1532_piece0 on 172.30.11
5.138:43839 in memory (size: 4.8 KiB, free: 364.1 MiB)
23/07/08 18:10:50 INFO BlockManagerInfo: Removed broadcast_1531_piece0 on 172.30.11
5.138:43839 in memory (size: 4.8 KiB, free: 364.1 MiB)
23/07/08 18:10:50 INFO BlockManagerInfo: Removed broadcast_1533_piece0 on 172.30.11
5.138:43839 in memory (size: 4.8 KiB, free: 364.1 MiB)
```

```
23/07/08 18:10:50 INFO BlockManagerInfo: Removed broadcast_1530_piece0 on 172.30.11
5.138:43839 in memory (size: 4.8 KiB, free: 364.1 MiB)
23/07/08 18:10:50 INFO BlockManagerInfo: Removed broadcast_1529_piece0 on 172.30.11
5.138:43839 in memory (size: 4.8 KiB, free: 364.1 MiB)
23/07/08 18:10:50 INFO PythonRunner: Times: total = 122, boot = 4, init = 118, finis
h = 0
23/07/08 18:10:50 INFO Executor: Finished task 0.0 in stage 2689.0 (TID 1714). 1465
bytes result sent to driver
23/07/08 18:10:50 INFO TaskSetManager: Finished task 0.0 in stage 2689.0 (TID 1714)
in 129 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:50 INFO TaskSchedulerImpl: Removed TaskSet 2689.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:50 INFO DAGScheduler: ResultStage 2689 (runJob at PythonRDD.scala:17
9) finished in 0.136 s
23/07/08 18:10:50 INFO DAGScheduler: Job 1248 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:50 INFO TaskSchedulerImpl: Killing all running tasks in stage 2689: S
tage finished
23/07/08 18:10:50 INFO DAGScheduler: Job 1248 finished: runJob at PythonRDD.scala:17
9, took 0.137616 s
```

Crear DataFrames a partir de RDD con encabezados

In [190…

```
df_canal = rdd_canal.filter(lambda line: line != header_canal).map(lambda line: lin
df_cliente = rdd_cliente.filter(lambda line: line != header_cliente).map(lambda lin
df_empleado = rdd_empleado.filter(lambda line: line != header_empleado).map(lambda
df_producto = rdd_producto.filter(lambda line: line != header_producto).map(lambda
df_sucursal = rdd_sucursal.filter(lambda line: line != header_sucursal).map(lambda
df_venta = rdd_venta.filter(lambda line: line != header_venta).map(lambda line: lin
##Genera columna de venta a partir del producto de cantidad * precio
df_venta = df_venta.withColumn("total_venta", (col("Cantidad") * col("Precio")))

df_DIMDATE_DATAONLY = rdd_DIMDATE_DATAONLY.filter(lambda line: line != header_DIMDA
df_DIMDATE = rdd_DIMDATE.filter(lambda line: line != header_DIMDATE).map(lambda lin
df_DIMTIME_DATAONLY = rdd_DIMTIME_DATAONLY.filter(lambda line: line != header_DIMTI
df_DIMTIME = rdd_DIMTIME.filter(lambda line: line != header_DIMTIME).map(lambda lin
```

```
23/07/08 18:10:57 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:57 INFO DAGScheduler: Got job 1249 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:57 INFO DAGScheduler: Final stage: ResultStage 2690 (runJob at Python
RDD.scala:179)
23/07/08 18:10:57 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:57 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:57 INFO DAGScheduler: Submitting ResultStage 2690 (PythonRDD[3856] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:57 INFO MemoryStore: Block broadcast_1535 stored as values in memory
(estimated size 9.5 KiB, free 343.1 MiB)
23/07/08 18:10:57 INFO MemoryStore: Block broadcast_1535_piece0 stored as bytes in m
emory (estimated size 5.6 KiB, free 343.1 MiB)
23/07/08 18:10:57 INFO BlockManagerInfo: Added broadcast_1535_piece0 in memory on 17
2.30.115.138:43839 (size: 5.6 KiB, free: 364.1 MiB)
23/07/08 18:10:57 INFO SparkContext: Created broadcast 1535 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:57 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 269
0 (PythonRDD[3856] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:57 INFO TaskSchedulerImpl: Adding task set 2690.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:57 INFO TaskSetManager: Starting task 0.0 in stage 2690.0 (TID 1715)
(172.30.115.138, executor driver, partition 0, ANY, 7423 bytes)
23/07/08 18:10:57 INFO Executor: Running task 0.0 in stage 2690.0 (TID 1715)
23/07/08 18:10:57 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/CanalDeVe
nta.csv:0+29
23/07/08 18:10:57 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:57 INFO PythonRunner: Times: total = 123, boot = 4, init = 119, finis
h = 0
23/07/08 18:10:57 INFO Executor: Finished task 0.0 in stage 2690.0 (TID 1715). 1429
bytes result sent to driver
23/07/08 18:10:57 INFO TaskSetManager: Finished task 0.0 in stage 2690.0 (TID 1715)
in 129 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:57 INFO TaskSchedulerImpl: Removed TaskSet 2690.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:57 INFO DAGScheduler: ResultStage 2690 (runJob at PythonRDD.scala:17
9) finished in 0.134 s
23/07/08 18:10:57 INFO DAGScheduler: Job 1249 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:57 INFO TaskSchedulerImpl: Killing all running tasks in stage 2690: S
tage finished
23/07/08 18:10:57 INFO DAGScheduler: Job 1249 finished: runJob at PythonRDD.scala:17
9, took 0.135666 s
23/07/08 18:10:57 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:57 INFO DAGScheduler: Got job 1250 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:57 INFO DAGScheduler: Final stage: ResultStage 2691 (runJob at Python
RDD.scala:179)
23/07/08 18:10:57 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:57 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:57 INFO DAGScheduler: Submitting ResultStage 2691 (PythonRDD[3861] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:57 INFO MemoryStore: Block broadcast_1536 stored as values in memory
(estimated size 9.6 KiB, free 343.1 MiB)
23/07/08 18:10:57 INFO MemoryStore: Block broadcast_1536_piece0 stored as bytes in m
```

```
emory (estimated size 5.6 KiB, free 343.1 MiB)
23/07/08 18:10:57 INFO BlockManagerInfo: Added broadcast_1536_piece0 in memory on 17
2.30.115.138:43839 (size: 5.6 KiB, free: 364.1 MiB)
23/07/08 18:10:57 INFO SparkContext: Created broadcast 1536 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:57 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 269
1 (PythonRDD[3861] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:57 INFO TaskSchedulerImpl: Adding task set 2691.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:57 INFO TaskSetManager: Starting task 0.0 in stage 2691.0 (TID 1716)
(172.30.115.138, executor driver, partition 0, ANY, 7418 bytes)
23/07/08 18:10:57 INFO Executor: Running task 0.0 in stage 2691.0 (TID 1716)
23/07/08 18:10:57 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Cliente.c
sv:0+214936
23/07/08 18:10:57 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:58 INFO PythonRunner: Times: total = 121, boot = 4, init = 117, finis
h = 0
23/07/08 18:10:58 INFO Executor: Finished task 0.0 in stage 2691.0 (TID 1716). 1559
bytes result sent to driver
23/07/08 18:10:58 INFO TaskSetManager: Finished task 0.0 in stage 2691.0 (TID 1716)
in 128 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:58 INFO TaskSchedulerImpl: Removed TaskSet 2691.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:58 INFO DAGScheduler: ResultStage 2691 (runJob at PythonRDD.scala:17
9) finished in 0.133 s
23/07/08 18:10:58 INFO DAGScheduler: Job 1250 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:58 INFO TaskSchedulerImpl: Killing all running tasks in stage 2691: S
tage finished
23/07/08 18:10:58 INFO DAGScheduler: Job 1250 finished: runJob at PythonRDD.scala:17
9, took 0.134959 s
23/07/08 18:10:58 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:58 INFO DAGScheduler: Got job 1251 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:58 INFO DAGScheduler: Final stage: ResultStage 2692 (runJob at Python
RDD.scala:179)
23/07/08 18:10:58 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:58 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:58 INFO DAGScheduler: Submitting ResultStage 2692 (PythonRDD[3866] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:58 INFO MemoryStore: Block broadcast_1537 stored as values in memory
(estimated size 9.6 KiB, free 343.1 MiB)
23/07/08 18:10:58 INFO MemoryStore: Block broadcast_1537_piece0 stored as bytes in m
emory (estimated size 5.6 KiB, free 343.1 MiB)
23/07/08 18:10:58 INFO BlockManagerInfo: Added broadcast_1537_piece0 in memory on 17
2.30.115.138:43839 (size: 5.6 KiB, free: 364.1 MiB)
23/07/08 18:10:58 INFO SparkContext: Created broadcast 1537 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:58 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 269
2 (PythonRDD[3866] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:58 INFO TaskSchedulerImpl: Adding task set 2692.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:58 INFO TaskSetManager: Starting task 0.0 in stage 2692.0 (TID 1717)
(172.30.115.138, executor driver, partition 0, ANY, 7419 bytes)
```

```
23/07/08 18:10:58 INFO Executor: Running task 0.0 in stage 2692.0 (TID 1717)
23/07/08 18:10:58 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Empleado.
csv:0+8119
23/07/08 18:10:58 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:58 INFO PythonRunner: Times: total = 120, boot = 4, init = 115, finis
h = 1
23/07/08 18:10:58 INFO Executor: Finished task 0.0 in stage 2692.0 (TID 1717). 1494
bytes result sent to driver
23/07/08 18:10:58 INFO TaskSetManager: Finished task 0.0 in stage 2692.0 (TID 1717)
in 126 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:58 INFO TaskSchedulerImpl: Removed TaskSet 2692.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:58 INFO DAGScheduler: ResultStage 2692 (runJob at PythonRDD.scala:17
9) finished in 0.130 s
23/07/08 18:10:58 INFO DAGScheduler: Job 1251 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:58 INFO TaskSchedulerImpl: Killing all running tasks in stage 2692: S
tage finished
23/07/08 18:10:58 INFO DAGScheduler: Job 1251 finished: runJob at PythonRDD.scala:17
9, took 0.131879 s
23/07/08 18:10:58 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:58 INFO DAGScheduler: Got job 1252 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:58 INFO DAGScheduler: Final stage: ResultStage 2693 (runJob at Python
RDD.scala:179)
23/07/08 18:10:58 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:58 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:58 INFO DAGScheduler: Submitting ResultStage 2693 (PythonRDD[3871] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:58 INFO MemoryStore: Block broadcast_1538 stored as values in memory
(estimated size 9.6 KiB, free 343.0 MiB)
23/07/08 18:10:58 INFO MemoryStore: Block broadcast_1538_piece0 stored as bytes in m
emory (estimated size 5.6 KiB, free 343.0 MiB)
23/07/08 18:10:58 INFO BlockManagerInfo: Added broadcast_1538_piece0 in memory on 17
2.30.115.138:43839 (size: 5.6 KiB, free: 364.1 MiB)
23/07/08 18:10:58 INFO SparkContext: Created broadcast 1538 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:58 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 269
3 (PythonRDD[3871] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:58 INFO TaskSchedulerImpl: Adding task set 2693.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:58 INFO TaskSetManager: Starting task 0.0 in stage 2693.0 (TID 1718)
(172.30.115.138, executor driver, partition 0, ANY, 7419 bytes)
23/07/08 18:10:58 INFO Executor: Running task 0.0 in stage 2693.0 (TID 1718)
23/07/08 18:10:58 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Producto.
csv:0+8440
23/07/08 18:10:58 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:58 INFO PythonRunner: Times: total = 143, boot = 4, init = 139, finis
h = 0
23/07/08 18:10:58 INFO Executor: Finished task 0.0 in stage 2693.0 (TID 1718). 1465
bytes result sent to driver
23/07/08 18:10:58 INFO TaskSetManager: Finished task 0.0 in stage 2693.0 (TID 1718)
in 149 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:58 INFO TaskSchedulerImpl: Removed TaskSet 2693.0, whose tasks have a
ll completed, from pool
```

```
23/07/08 18:10:58 INFO DAGScheduler: ResultStage 2693 (runJob at PythonRDD.scala:17
9) finished in 0.154 s
23/07/08 18:10:58 INFO DAGScheduler: Job 1252 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:58 INFO TaskSchedulerImpl: Killing all running tasks in stage 2693: S
tage finished
23/07/08 18:10:58 INFO DAGScheduler: Job 1252 finished: runJob at PythonRDD.scala:17
9, took 0.155233 s
23/07/08 18:10:58 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:58 INFO DAGScheduler: Got job 1253 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:58 INFO DAGScheduler: Final stage: ResultStage 2694 (runJob at Python
RDD.scala:179)
23/07/08 18:10:58 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:58 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:58 INFO DAGScheduler: Submitting ResultStage 2694 (PythonRDD[3876] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:58 INFO MemoryStore: Block broadcast_1539 stored as values in memory
(estimated size 9.6 KiB, free 343.0 MiB)
23/07/08 18:10:58 INFO MemoryStore: Block broadcast_1539_piece0 stored as bytes in m
emory (estimated size 5.6 KiB, free 343.0 MiB)
23/07/08 18:10:58 INFO BlockManagerInfo: Added broadcast_1539_piece0 in memory on 17
2.30.115.138:43839 (size: 5.6 KiB, free: 364.1 MiB)
23/07/08 18:10:58 INFO SparkContext: Created broadcast 1539 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:58 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 269
4 (PythonRDD[3876] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:58 INFO TaskSchedulerImpl: Adding task set 2694.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:58 INFO TaskSetManager: Starting task 0.0 in stage 2694.0 (TID 1719)
(172.30.115.138, executor driver, partition 0, ANY, 7419 bytes)
23/07/08 18:10:58 INFO Executor: Running task 0.0 in stage 2694.0 (TID 1719)
23/07/08 18:10:58 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Sucursal.
csv:0+1266
23/07/08 18:10:58 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:58 INFO PythonRunner: Times: total = 117, boot = 5, init = 111, finis
h = 1
23/07/08 18:10:58 INFO Executor: Finished task 0.0 in stage 2694.0 (TID 1719). 1522
bytes result sent to driver
23/07/08 18:10:58 INFO TaskSetManager: Finished task 0.0 in stage 2694.0 (TID 1719)
in 123 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:58 INFO TaskSchedulerImpl: Removed TaskSet 2694.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:58 INFO DAGScheduler: ResultStage 2694 (runJob at PythonRDD.scala:17
9) finished in 0.128 s
23/07/08 18:10:58 INFO DAGScheduler: Job 1253 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:58 INFO TaskSchedulerImpl: Killing all running tasks in stage 2694: S
tage finished
23/07/08 18:10:58 INFO DAGScheduler: Job 1253 finished: runJob at PythonRDD.scala:17
9, took 0.129219 s
23/07/08 18:10:58 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:58 INFO DAGScheduler: Got job 1254 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:58 INFO DAGScheduler: Final stage: ResultStage 2695 (runJob at Python
```

```
RDD.scala:179)
23/07/08 18:10:58 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:58 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:58 INFO DAGScheduler: Submitting ResultStage 2695 (PythonRDD[3881] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:58 INFO MemoryStore: Block broadcast_1540 stored as values in memory
(estimated size 9.6 KiB, free 343.0 MiB)
23/07/08 18:10:58 INFO MemoryStore: Block broadcast_1540_piece0 stored as bytes in m
emory (estimated size 5.7 KiB, free 343.0 MiB)
23/07/08 18:10:58 INFO BlockManagerInfo: Added broadcast_1540_piece0 in memory on 17
2.30.115.138:43839 (size: 5.7 KiB, free: 364.0 MiB)
23/07/08 18:10:58 INFO SparkContext: Created broadcast 1540 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:58 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 269
5 (PythonRDD[3881] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:58 INFO TaskSchedulerImpl: Adding task set 2695.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:58 INFO TaskSetManager: Starting task 0.0 in stage 2695.0 (TID 1720)
(172.30.115.138, executor driver, partition 0, ANY, 7417 bytes)
23/07/08 18:10:58 INFO Executor: Running task 0.0 in stage 2695.0 (TID 1720)
23/07/08 18:10:58 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:0+1310749
23/07/08 18:10:58 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:58 INFO PythonRunner: Times: total = 117, boot = 3, init = 114, finis
h = 0
23/07/08 18:10:58 INFO Executor: Finished task 0.0 in stage 2695.0 (TID 1720). 1484
bytes result sent to driver
23/07/08 18:10:58 INFO TaskSetManager: Finished task 0.0 in stage 2695.0 (TID 1720)
in 126 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:58 INFO TaskSchedulerImpl: Removed TaskSet 2695.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:58 INFO DAGScheduler: ResultStage 2695 (runJob at PythonRDD.scala:17
9) finished in 0.131 s
23/07/08 18:10:58 INFO DAGScheduler: Job 1254 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:58 INFO TaskSchedulerImpl: Killing all running tasks in stage 2695: S
tage finished
23/07/08 18:10:58 INFO DAGScheduler: Job 1254 finished: runJob at PythonRDD.scala:17
9, took 0.133375 s
23/07/08 18:10:59 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:59 INFO DAGScheduler: Got job 1255 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:59 INFO DAGScheduler: Final stage: ResultStage 2696 (runJob at Python
RDD.scala:179)
23/07/08 18:10:59 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:59 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:59 INFO DAGScheduler: Submitting ResultStage 2696 (PythonRDD[3886] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:59 INFO MemoryStore: Block broadcast_1541 stored as values in memory
(estimated size 9.7 KiB, free 343.0 MiB)
23/07/08 18:10:59 INFO MemoryStore: Block broadcast_1541_piece0 stored as bytes in m
emory (estimated size 5.7 KiB, free 343.0 MiB)
23/07/08 18:10:59 INFO BlockManagerInfo: Added broadcast_1541_piece0 in memory on 17
2.30.115.138:43839 (size: 5.7 KiB, free: 364.0 MiB)
23/07/08 18:10:59 INFO SparkContext: Created broadcast 1541 from broadcast at DAGSch
```

```
eduler.scala:1535
23/07/08 18:10:59 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 269
6 (PythonRDD[3886] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:59 INFO TaskSchedulerImpl: Adding task set 2696.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:59 INFO TaskSetManager: Starting task 0.0 in stage 2696.0 (TID 1721)
(172.30.115.138, executor driver, partition 0, ANY, 7431 bytes)
23/07/08 18:10:59 INFO Executor: Running task 0.0 in stage 2696.0 (TID 1721)
23/07/08 18:10:59 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE-DATAONLY.csv:0+365617
23/07/08 18:10:59 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:59 INFO PythonRunner: Times: total = 118, boot = 4, init = 113, finis
h = 1
23/07/08 18:10:59 INFO Executor: Finished task 0.0 in stage 2696.0 (TID 1721). 1470
bytes result sent to driver
23/07/08 18:10:59 INFO TaskSetManager: Finished task 0.0 in stage 2696.0 (TID 1721)
in 125 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:59 INFO TaskSchedulerImpl: Removed TaskSet 2696.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:59 INFO DAGScheduler: ResultStage 2696 (runJob at PythonRDD.scala:17
9) finished in 0.128 s
23/07/08 18:10:59 INFO DAGScheduler: Job 1255 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:59 INFO TaskSchedulerImpl: Killing all running tasks in stage 2696: S
tage finished
23/07/08 18:10:59 INFO DAGScheduler: Job 1255 finished: runJob at PythonRDD.scala:17
9, took 0.129685 s
23/07/08 18:10:59 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:59 INFO DAGScheduler: Got job 1256 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:59 INFO DAGScheduler: Final stage: ResultStage 2697 (runJob at Python
RDD.scala:179)
23/07/08 18:10:59 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:59 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:59 INFO DAGScheduler: Submitting ResultStage 2697 (PythonRDD[3891] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:59 INFO MemoryStore: Block broadcast_1542 stored as values in memory
(estimated size 9.7 KiB, free 343.0 MiB)
23/07/08 18:10:59 INFO MemoryStore: Block broadcast_1542_piece0 stored as bytes in m
emory (estimated size 5.7 KiB, free 343.0 MiB)
23/07/08 18:10:59 INFO BlockManagerInfo: Added broadcast_1542_piece0 in memory on 17
2.30.115.138:43839 (size: 5.7 KiB, free: 364.0 MiB)
23/07/08 18:10:59 INFO SparkContext: Created broadcast 1542 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:59 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 269
7 (PythonRDD[3891] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:59 INFO TaskSchedulerImpl: Adding task set 2697.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:59 INFO TaskSetManager: Starting task 0.0 in stage 2697.0 (TID 1722)
(172.30.115.138, executor driver, partition 0, ANY, 7422 bytes)
23/07/08 18:10:59 INFO Executor: Running task 0.0 in stage 2697.0 (TID 1722)
23/07/08 18:10:59 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:0+450275
23/07/08 18:10:59 INFO LineRecordReader: Found UTF-8 BOM and skipped it
```

```
23/07/08 18:10:59 INFO PythonRunner: Times: total = 116, boot = 4, init = 112, finis
h = 0
23/07/08 18:10:59 INFO Executor: Finished task 0.0 in stage 2697.0 (TID 1722). 1480
bytes result sent to driver
23/07/08 18:10:59 INFO TaskSetManager: Finished task 0.0 in stage 2697.0 (TID 1722)
in 123 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:59 INFO TaskSchedulerImpl: Removed TaskSet 2697.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:59 INFO DAGScheduler: ResultStage 2697 (runJob at PythonRDD.scala:17
9) finished in 0.127 s
23/07/08 18:10:59 INFO DAGScheduler: Job 1256 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:59 INFO TaskSchedulerImpl: Killing all running tasks in stage 2697: S
tage finished
23/07/08 18:10:59 INFO DAGScheduler: Job 1256 finished: runJob at PythonRDD.scala:17
9, took 0.128462 s
23/07/08 18:10:59 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:59 INFO DAGScheduler: Got job 1257 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:59 INFO DAGScheduler: Final stage: ResultStage 2698 (runJob at Python
RDD.scala:179)
23/07/08 18:10:59 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:59 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:59 INFO DAGScheduler: Submitting ResultStage 2698 (PythonRDD[3896] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:59 INFO MemoryStore: Block broadcast_1543 stored as values in memory
(estimated size 9.6 KiB, free 343.0 MiB)
23/07/08 18:10:59 INFO MemoryStore: Block broadcast_1543_piece0 stored as bytes in m
emory (estimated size 5.6 KiB, free 343.0 MiB)
23/07/08 18:10:59 INFO BlockManagerInfo: Added broadcast_1543_piece0 in memory on 17
2.30.115.138:43839 (size: 5.6 KiB, free: 364.0 MiB)
23/07/08 18:10:59 INFO SparkContext: Created broadcast 1543 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:59 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 269
8 (PythonRDD[3896] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:59 INFO TaskSchedulerImpl: Adding task set 2698.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:59 INFO TaskSetManager: Starting task 0.0 in stage 2698.0 (TID 1723)
(172.30.115.138, executor driver, partition 0, ANY, 7431 bytes)
23/07/08 18:10:59 INFO Executor: Running task 0.0 in stage 2698.0 (TID 1723)
23/07/08 18:10:59 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMTI
ME-DATAONLY.csv:0+75866
23/07/08 18:10:59 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:59 INFO PythonRunner: Times: total = 120, boot = 4, init = 116, finis
h = 0
23/07/08 18:10:59 INFO Executor: Finished task 0.0 in stage 2698.0 (TID 1723). 1451
bytes result sent to driver
23/07/08 18:10:59 INFO TaskSetManager: Finished task 0.0 in stage 2698.0 (TID 1723)
in 126 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:59 INFO TaskSchedulerImpl: Removed TaskSet 2698.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:59 INFO DAGScheduler: ResultStage 2698 (runJob at PythonRDD.scala:17
9) finished in 0.130 s
23/07/08 18:10:59 INFO DAGScheduler: Job 1257 is finished. Cancelling potential spec
ulative or zombie tasks for this job
```

```
23/07/08 18:10:59 INFO TaskSchedulerImpl: Killing all running tasks in stage 2698: S
tage finished
23/07/08 18:10:59 INFO DAGScheduler: Job 1257 finished: runJob at PythonRDD.scala:17
9, took 0.131050 s
23/07/08 18:10:59 INFO SparkContext: Starting job: runJob at PythonRDD.scala:179
23/07/08 18:10:59 INFO DAGScheduler: Got job 1258 (runJob at PythonRDD.scala:179) wi
th 1 output partitions
23/07/08 18:10:59 INFO DAGScheduler: Final stage: ResultStage 2699 (runJob at Python
RDD.scala:179)
23/07/08 18:10:59 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:10:59 INFO DAGScheduler: Missing parents: List()
23/07/08 18:10:59 INFO DAGScheduler: Submitting ResultStage 2699 (PythonRDD[3901] at
RDD at PythonRDD.scala:53), which has no missing parents
23/07/08 18:10:59 INFO MemoryStore: Block broadcast_1544 stored as values in memory
(estimated size 9.6 KiB, free 343.0 MiB)
23/07/08 18:10:59 INFO MemoryStore: Block broadcast_1544_piece0 stored as bytes in m
emory (estimated size 5.6 KiB, free 343.0 MiB)
23/07/08 18:10:59 INFO BlockManagerInfo: Added broadcast_1544_piece0 in memory on 17
2.30.115.138:43839 (size: 5.6 KiB, free: 364.0 MiB)
23/07/08 18:10:59 INFO SparkContext: Created broadcast 1544 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:10:59 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 269
9 (PythonRDD[3901] at RDD at PythonRDD.scala:53) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:10:59 INFO TaskSchedulerImpl: Adding task set 2699.0 with 1 tasks resour
ce profile 0
23/07/08 18:10:59 INFO TaskSetManager: Starting task 0.0 in stage 2699.0 (TID 1724)
(172.30.115.138, executor driver, partition 0, ANY, 7422 bytes)
23/07/08 18:10:59 INFO Executor: Running task 0.0 in stage 2699.0 (TID 1724)
23/07/08 18:10:59 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMTI
ME.csv:0+22350
23/07/08 18:10:59 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:10:59 INFO PythonRunner: Times: total = 125, boot = 4, init = 121, finis
h = 0
23/07/08 18:10:59 INFO Executor: Finished task 0.0 in stage 2699.0 (TID 1724). 1451
bytes result sent to driver
23/07/08 18:10:59 INFO TaskSetManager: Finished task 0.0 in stage 2699.0 (TID 1724)
in 131 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:10:59 INFO TaskSchedulerImpl: Removed TaskSet 2699.0, whose tasks have a
ll completed, from pool
23/07/08 18:10:59 INFO DAGScheduler: ResultStage 2699 (runJob at PythonRDD.scala:17
9) finished in 0.136 s
23/07/08 18:10:59 INFO DAGScheduler: Job 1258 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:10:59 INFO TaskSchedulerImpl: Killing all running tasks in stage 2699: S
tage finished
23/07/08 18:10:59 INFO DAGScheduler: Job 1258 finished: runJob at PythonRDD.scala:17
9, took 0.137522 s
```

Imprimir los primeros 5 registros de cada DataFrame Creado

In [191…

```
print("CanalDeVenta:")
df_canal.show(5)
```

```
print("Cliente:")
df_cliente.show(5)

print("Empleado:")
df_empleado.show(5)

print("Producto:")
df_producto.show(5)

print("Sucursal:")
df_sucursal.show(5)

print("Venta:")
df_venta.show(5)

## TABLAS DE FECHAS
print("DIMDATE-DATAONLY:")
df_DIMDATE_DATAONLY.show(5)

print("DIMDATE:")
df_DIMDATE.show(5)

print("DIMTIME-DATAONLY:")
df_DIMTIME_DATAONLY.show(5)

print("DIMTIME:")
df_DIMTIME.show(5)
```

```
CanalDeVenta:
```

```
23/07/08 18:11:04 INFO CodeGenerator: Code generated in 25.455055 ms
23/07/08 18:11:04 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:11:04 INFO DAGScheduler: Got job 1259 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:11:04 INFO DAGScheduler: Final stage: ResultStage 2700 (showString at <u
nknown>:0)
23/07/08 18:11:04 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:04 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:04 INFO DAGScheduler: Submitting ResultStage 2700 (MapPartitionsRDD[3
907] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:04 INFO MemoryStore: Block broadcast_1545 stored as values in memory
(estimated size 17.0 KiB, free 342.9 MiB)
23/07/08 18:11:04 INFO MemoryStore: Block broadcast_1545_piece0 stored as bytes in m
emory (estimated size 8.9 KiB, free 342.9 MiB)
23/07/08 18:11:04 INFO BlockManagerInfo: Added broadcast_1545_piece0 in memory on 17
2.30.115.138:43839 (size: 8.9 KiB, free: 364.0 MiB)
23/07/08 18:11:04 INFO SparkContext: Created broadcast 1545 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:04 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 270
0 (MapPartitionsRDD[3907] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:11:04 INFO TaskSchedulerImpl: Adding task set 2700.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:04 INFO TaskSetManager: Starting task 0.0 in stage 2700.0 (TID 1725)
(172.30.115.138, executor driver, partition 0, ANY, 7423 bytes)
23/07/08 18:11:04 INFO Executor: Running task 0.0 in stage 2700.0 (TID 1725)
23/07/08 18:11:04 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/CanalDeVe
nta.csv:0+29
23/07/08 18:11:04 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:04 INFO BlockManagerInfo: Removed broadcast_1542_piece0 on 172.30.11
5.138:43839 in memory (size: 5.7 KiB, free: 364.0 MiB)
23/07/08 18:11:04 INFO BlockManagerInfo: Removed broadcast_1541_piece0 on 172.30.11
5.138:43839 in memory (size: 5.7 KiB, free: 364.0 MiB)
23/07/08 18:11:04 INFO BlockManagerInfo: Removed broadcast_1540_piece0 on 172.30.11
5.138:43839 in memory (size: 5.7 KiB, free: 364.0 MiB)
23/07/08 18:11:04 INFO BlockManagerInfo: Removed broadcast_1543_piece0 on 172.30.11
5.138:43839 in memory (size: 5.6 KiB, free: 364.0 MiB)
23/07/08 18:11:04 INFO BlockManagerInfo: Removed broadcast_1536_piece0 on 172.30.11
5.138:43839 in memory (size: 5.6 KiB, free: 364.0 MiB)
23/07/08 18:11:04 INFO BlockManagerInfo: Removed broadcast_1539_piece0 on 172.30.11
5.138:43839 in memory (size: 5.6 KiB, free: 364.0 MiB)
23/07/08 18:11:04 INFO BlockManagerInfo: Removed broadcast_1538_piece0 on 172.30.11
5.138:43839 in memory (size: 5.6 KiB, free: 364.1 MiB)
23/07/08 18:11:04 INFO BlockManagerInfo: Removed broadcast_1535_piece0 on 172.30.11
5.138:43839 in memory (size: 5.6 KiB, free: 364.1 MiB)
23/07/08 18:11:04 INFO BlockManagerInfo: Removed broadcast_1544_piece0 on 172.30.11
5.138:43839 in memory (size: 5.6 KiB, free: 364.1 MiB)
23/07/08 18:11:04 INFO BlockManagerInfo: Removed broadcast_1534_piece0 on 172.30.11
5.138:43839 in memory (size: 4.8 KiB, free: 364.1 MiB)
23/07/08 18:11:04 INFO BlockManagerInfo: Removed broadcast_1537_piece0 on 172.30.11
5.138:43839 in memory (size: 5.6 KiB, free: 364.1 MiB)
23/07/08 18:11:04 INFO PythonRunner: Times: total = 136, boot = 5, init = 131, finis
h = 0
23/07/08 18:11:04 INFO Executor: Finished task 0.0 in stage 2700.0 (TID 1725). 2062
bytes result sent to driver
23/07/08 18:11:04 INFO TaskSetManager: Finished task 0.0 in stage 2700.0 (TID 1725)
```

```
in 147 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:04 INFO TaskSchedulerImpl: Removed TaskSet 2700.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:04 INFO DAGScheduler: ResultStage 2700 (showString at <unknown>:0) fi
nished in 0.160 s
23/07/08 18:11:04 INFO DAGScheduler: Job 1259 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:04 INFO TaskSchedulerImpl: Killing all running tasks in stage 2700: S
tage finished
23/07/08 18:11:04 INFO DAGScheduler: Job 1259 finished: showString at <unknown>:0, t
ook 0.161450 s
23/07/08 18:11:04 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:11:04 INFO DAGScheduler: Got job 1260 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:11:04 INFO DAGScheduler: Final stage: ResultStage 2701 (showString at <u
nknown>:0)
23/07/08 18:11:04 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:04 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:04 INFO DAGScheduler: Submitting ResultStage 2701 (MapPartitionsRDD[3
907] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:04 INFO MemoryStore: Block broadcast_1546 stored as values in memory
(estimated size 17.0 KiB, free 343.1 MiB)
23/07/08 18:11:04 INFO MemoryStore: Block broadcast_1546_piece0 stored as bytes in m
emory (estimated size 8.9 KiB, free 343.1 MiB)
23/07/08 18:11:04 INFO BlockManagerInfo: Added broadcast_1546_piece0 in memory on 17
2.30.115.138:43839 (size: 8.9 KiB, free: 364.1 MiB)
23/07/08 18:11:04 INFO SparkContext: Created broadcast 1546 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:04 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 270
1 (MapPartitionsRDD[3907] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(1))
23/07/08 18:11:04 INFO TaskSchedulerImpl: Adding task set 2701.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:04 INFO TaskSetManager: Starting task 0.0 in stage 2701.0 (TID 1726)
(172.30.115.138, executor driver, partition 1, ANY, 7423 bytes)
23/07/08 18:11:04 INFO Executor: Running task 0.0 in stage 2701.0 (TID 1726)
23/07/08 18:11:04 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/CanalDeVe
nta.csv:29+29
23/07/08 18:11:04 INFO PythonRunner: Times: total = 222, boot = 112, init = 110, fin
ish = 0
23/07/08 18:11:04 INFO Executor: Finished task 0.0 in stage 2701.0 (TID 1726). 2002
bytes result sent to driver
23/07/08 18:11:04 INFO TaskSetManager: Finished task 0.0 in stage 2701.0 (TID 1726)
in 230 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:04 INFO TaskSchedulerImpl: Removed TaskSet 2701.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:04 INFO DAGScheduler: ResultStage 2701 (showString at <unknown>:0) fi
nished in 0.234 s
23/07/08 18:11:04 INFO DAGScheduler: Job 1260 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:04 INFO TaskSchedulerImpl: Killing all running tasks in stage 2701: S
tage finished
23/07/08 18:11:04 INFO DAGScheduler: Job 1260 finished: showString at <unknown>:0, t
ook 0.235988 s
23/07/08 18:11:04 INFO CodeGenerator: Code generated in 11.972496 ms
23/07/08 18:11:05 INFO SparkContext: Starting job: showString at <unknown>:0
```

```
23/07/08 18:11:05 INFO DAGScheduler: Got job 1261 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:11:05 INFO DAGScheduler: Final stage: ResultStage 2702 (showString at <u
nknown>:0)
23/07/08 18:11:05 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:05 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:05 INFO DAGScheduler: Submitting ResultStage 2702 (MapPartitionsRDD[3
909] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:05 INFO MemoryStore: Block broadcast_1547 stored as values in memory
(estimated size 20.2 KiB, free 343.0 MiB)
23/07/08 18:11:05 INFO MemoryStore: Block broadcast_1547_piece0 stored as bytes in m
emory (estimated size 9.5 KiB, free 343.0 MiB)
23/07/08 18:11:05 INFO BlockManagerInfo: Added broadcast_1547_piece0 in memory on 17
2.30.115.138:43839 (size: 9.5 KiB, free: 364.1 MiB)
23/07/08 18:11:05 INFO SparkContext: Created broadcast 1547 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:05 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 270
2 (MapPartitionsRDD[3909] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:11:05 INFO TaskSchedulerImpl: Adding task set 2702.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:05 INFO TaskSetManager: Starting task 0.0 in stage 2702.0 (TID 1727)
(172.30.115.138, executor driver, partition 0, ANY, 7418 bytes)
23/07/08 18:11:05 INFO Executor: Running task 0.0 in stage 2702.0 (TID 1727)
23/07/08 18:11:05 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Cliente.c
sv:0+214936
23/07/08 18:11:05 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:05 INFO Executor: Finished task 0.0 in stage 2702.0 (TID 1727). 2834
bytes result sent to driver
23/07/08 18:11:05 INFO TaskSetManager: Finished task 0.0 in stage 2702.0 (TID 1727)
in 119 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:05 INFO TaskSchedulerImpl: Removed TaskSet 2702.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:05 INFO DAGScheduler: ResultStage 2702 (showString at <unknown>:0) fi
nished in 0.123 s
23/07/08 18:11:05 INFO DAGScheduler: Job 1261 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:05 INFO TaskSchedulerImpl: Killing all running tasks in stage 2702: S
tage finished
23/07/08 18:11:05 INFO DAGScheduler: Job 1261 finished: showString at <unknown>:0, t
ook 0.124263 s
23/07/08 18:11:05 INFO CodeGenerator: Code generated in 6.611027 ms
23/07/08 18:11:05 INFO CodeGenerator: Code generated in 6.630925 ms
23/07/08 18:11:05 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:11:05 INFO DAGScheduler: Got job 1262 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:11:05 INFO DAGScheduler: Final stage: ResultStage 2703 (showString at <u
nknown>:0)
23/07/08 18:11:05 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:05 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:05 INFO DAGScheduler: Submitting ResultStage 2703 (MapPartitionsRDD[3
911] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:05 INFO MemoryStore: Block broadcast_1548 stored as values in memory
(estimated size 19.0 KiB, free 343.0 MiB)
```

```
+------+-----------+
|CODIGO|DESCRIPCION|
+------+-----------+
|     1| Telefónica|
|     2|     OnLine|
|     3| Presencial|
+------+-----------+
```

Cliente:

```
+---+-------------+--------------------+--------------------+--------+----+---------
----------+------------+------------+-----+
| ID|    Provincia|   Nombre_y_Apellido|           Domicilio|Telefono|Edad|
Localidad|           X|           Y|col10|
+---+-------------+--------------------+--------------------+--------+----+---------
----------+------------+------------+-----+
|  1|             |  HEBER JONI SANTANA|LAS HERAS Y BAT. ...| 42-5161|  58|
LOMA VERDE|-58,81850307|-34,30997088|     |
|  2| Buenos Aires|         ANA SAPRIZA|PUEYRREDON Y DUPU...| 49-7578|  61|
SANTA ROSA|-58,73073751|-34,93908311|     |
|  3|Buenos Aires |FERNANDO LUIS SAR...|CALDERON DE LA BA...| 49-3435|  15|
TORRES|-59,12794068|-34,43082199|     |
|  4| Buenos Aires|    MANUELA SARASOLA|RUTA 36 KM 45,500...| 49-2883|  29|
RUTA SOL|-58,14393954|-34,92052706|     |
|  5| Buenos Aires|  MARIO RAÚL SARASUA|492 Y 186 S/N  CO...|491-4608|  34|JOSE MELC
HOR ROMERO|  -58,089381| -34,9444471|     |
+---+-------------+--------------------+--------------------+--------+----+---------
----------+------------+------------+-----+
only showing top 5 rows
```

Empleado:

```
23/07/08 18:11:05 INFO MemoryStore: Block broadcast_1548_piece0 stored as bytes in m
emory (estimated size 9.3 KiB, free 343.0 MiB)
23/07/08 18:11:05 INFO BlockManagerInfo: Added broadcast_1548_piece0 in memory on 17
2.30.115.138:43839 (size: 9.3 KiB, free: 364.0 MiB)
23/07/08 18:11:05 INFO SparkContext: Created broadcast 1548 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:05 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 270
3 (MapPartitionsRDD[3911] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:11:05 INFO TaskSchedulerImpl: Adding task set 2703.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:05 INFO TaskSetManager: Starting task 0.0 in stage 2703.0 (TID 1728)
(172.30.115.138, executor driver, partition 0, ANY, 7419 bytes)
23/07/08 18:11:05 INFO Executor: Running task 0.0 in stage 2703.0 (TID 1728)
23/07/08 18:11:05 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Empleado.
csv:0+8119
23/07/08 18:11:05 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:05 INFO Executor: Finished task 0.0 in stage 2703.0 (TID 1728). 2313
bytes result sent to driver
23/07/08 18:11:05 INFO TaskSetManager: Finished task 0.0 in stage 2703.0 (TID 1728)
in 131 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:05 INFO TaskSchedulerImpl: Removed TaskSet 2703.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:05 INFO DAGScheduler: ResultStage 2703 (showString at <unknown>:0) fi
nished in 0.137 s
23/07/08 18:11:05 INFO DAGScheduler: Job 1262 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:05 INFO TaskSchedulerImpl: Killing all running tasks in stage 2703: S
tage finished
23/07/08 18:11:05 INFO DAGScheduler: Job 1262 finished: showString at <unknown>:0, t
ook 0.138048 s
23/07/08 18:11:05 INFO CodeGenerator: Code generated in 7.321447 ms
23/07/08 18:11:05 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:11:05 INFO DAGScheduler: Got job 1263 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:11:05 INFO DAGScheduler: Final stage: ResultStage 2704 (showString at <u
nknown>:0)
23/07/08 18:11:05 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:05 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:05 INFO DAGScheduler: Submitting ResultStage 2704 (MapPartitionsRDD[3
913] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:05 INFO MemoryStore: Block broadcast_1549 stored as values in memory
(estimated size 17.8 KiB, free 343.0 MiB)
23/07/08 18:11:05 INFO MemoryStore: Block broadcast_1549_piece0 stored as bytes in m
emory (estimated size 9.0 KiB, free 343.0 MiB)
23/07/08 18:11:05 INFO BlockManagerInfo: Added broadcast_1549_piece0 in memory on 17
2.30.115.138:43839 (size: 9.0 KiB, free: 364.0 MiB)
23/07/08 18:11:05 INFO SparkContext: Created broadcast 1549 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:05 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 270
4 (MapPartitionsRDD[3913] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:11:05 INFO TaskSchedulerImpl: Adding task set 2704.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:05 INFO TaskSetManager: Starting task 0.0 in stage 2704.0 (TID 1729)
(172.30.115.138, executor driver, partition 0, ANY, 7419 bytes)
```

```
23/07/08 18:11:05 INFO Executor: Running task 0.0 in stage 2704.0 (TID 1729)
23/07/08 18:11:05 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Producto.
csv:0+8440
23/07/08 18:11:05 INFO LineRecordReader: Found UTF-8 BOM and skipped it
+-----------+---------+---------+--------+--------------+--------------+--------+
|ID_empleado| Apellido|   Nombre|Sucursal|        Sector|         Cargo| Salario|
+-----------+---------+---------+--------+--------------+--------------+--------+
|       1968|   Burgos| Jeronimo| Caseros|Administración|Administrativo|32000,00|
|       1674| Villegas|Estefania| Caseros|Administración|      Vendedor|32000,00|
|       1516|Fernandez|Guillermo| Caseros|Administración|      Vendedor|45000,00|
|       1330|  Ramirez|   Eliana| Caseros|Administración|      Vendedor|32000,00|
|       1657|  Carmona|     Jose| Caseros|Administración|      Vendedor|32000,00|
+-----------+---------+---------+--------+--------------+--------------+--------+
only showing top 5 rows

Producto:
+-----------+--------------------+-----------+--------+
|ID_PRODUCTO|            Concepto|       Tipo| Precio |
+-----------+--------------------+-----------+--------+
|      42737|EPSON COPYFAX 2000  |  IMPRESIÓN| 1658,00|
|      42754|  MOT ASROCK H110...|INFORMATICA| 1237,50|
|      42755|MOT ASROCK A58M-V...|INFORMATICA| 1079,32|
|      42756|  MOT ECS KAM1-I AM1|INFORMATICA|  638,66|
|      42757|MOT ASROCK B150M-...|INFORMATICA| 1784,42|
+-----------+--------------------+-----------+--------+
only showing top 5 rows

Sucursal:
```

```
23/07/08 18:11:05 INFO Executor: Finished task 0.0 in stage 2704.0 (TID 1729). 2233
bytes result sent to driver
23/07/08 18:11:05 INFO TaskSetManager: Finished task 0.0 in stage 2704.0 (TID 1729)
in 127 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:05 INFO TaskSchedulerImpl: Removed TaskSet 2704.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:05 INFO DAGScheduler: ResultStage 2704 (showString at <unknown>:0) fi
nished in 0.131 s
23/07/08 18:11:05 INFO DAGScheduler: Job 1263 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:05 INFO TaskSchedulerImpl: Killing all running tasks in stage 2704: S
tage finished
23/07/08 18:11:05 INFO DAGScheduler: Job 1263 finished: showString at <unknown>:0, t
ook 0.133341 s
23/07/08 18:11:05 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:11:05 INFO DAGScheduler: Got job 1264 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:11:05 INFO DAGScheduler: Final stage: ResultStage 2705 (showString at <u
nknown>:0)
23/07/08 18:11:05 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:05 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:05 INFO DAGScheduler: Submitting ResultStage 2705 (MapPartitionsRDD[3
915] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:05 INFO MemoryStore: Block broadcast_1550 stored as values in memory
(estimated size 19.0 KiB, free 343.0 MiB)
23/07/08 18:11:05 INFO MemoryStore: Block broadcast_1550_piece0 stored as bytes in m
emory (estimated size 9.2 KiB, free 343.0 MiB)
23/07/08 18:11:05 INFO BlockManagerInfo: Added broadcast_1550_piece0 in memory on 17
2.30.115.138:43839 (size: 9.2 KiB, free: 364.0 MiB)
23/07/08 18:11:05 INFO SparkContext: Created broadcast 1550 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:05 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 270
5 (MapPartitionsRDD[3915] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:11:05 INFO TaskSchedulerImpl: Adding task set 2705.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:05 INFO TaskSetManager: Starting task 0.0 in stage 2705.0 (TID 1730)
(172.30.115.138, executor driver, partition 0, ANY, 7419 bytes)
23/07/08 18:11:05 INFO Executor: Running task 0.0 in stage 2705.0 (TID 1730)
23/07/08 18:11:05 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Sucursal.
csv:0+1266
23/07/08 18:11:05 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:05 INFO Executor: Finished task 0.0 in stage 2705.0 (TID 1730). 2498
bytes result sent to driver
23/07/08 18:11:05 INFO TaskSetManager: Finished task 0.0 in stage 2705.0 (TID 1730)
in 127 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:05 INFO TaskSchedulerImpl: Removed TaskSet 2705.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:05 INFO DAGScheduler: ResultStage 2705 (showString at <unknown>:0) fi
nished in 0.131 s
23/07/08 18:11:05 INFO DAGScheduler: Job 1264 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:05 INFO TaskSchedulerImpl: Killing all running tasks in stage 2705: S
tage finished
23/07/08 18:11:05 INFO DAGScheduler: Job 1264 finished: showString at <unknown>:0, t
ook 0.132173 s
```

```
+---+----------+--------------------+--------------------+--------------------+-----
------+-----------+
| ID|  Sucursal|           Direccion|           Localidad|           Provincia|    L
atitud|   Longitud|
+---+----------+--------------------+--------------------+--------------------+-----
------+-----------+
|  1|   Cabildo|    Av. Cabildo 1342|Ciudad de Buenos ...|Ciudad de Buenos ...|-34,5
678060|-58,4495720|
|  2| Palermo 1|      Guatemala 5701|                CABA|                CABA|-34,5
790350|-58,4335660|
|  3| Palermo 2|Gral. Lucio Norbe...|                CABA|    C deBuenos Aires|-34,5
959660|-58,4051500|
|  4|Corrientes| Av. Corrientes 2352|Ciudad de Buenos ...|               Bs As|-34,6
046850|-58,3987640|
|  5|   Almagro|      Venezuela 3650|             Capital|             Bs.As. |-34,6
173080|-58,4161790|
+---+----------+--------------------+--------------------+--------------------+-----
------+-----------+
only showing top 5 rows

Venta:
+-------+----------+-------------+-------+---------+----------+----------+----------
+------+--------+-----------+
|IdVenta|     Fecha|Fecha_Entrega|IdCanal|IdCliente|IdSucursal|IdEmpleado|IdProducto
|Precio|Cantidad|total_venta|
+-------+----------+-------------+-------+---------+----------+----------+----------
+------+--------+-----------+
|      1|09/03/2018|   17/03/2018|      3|      969|        13|      1674|     42817
|813.12|       2|    1626.24|
|      2|28/12/2018|   29/12/2018|      2|      884|        13|      1674|     42795
|543.18|       3|    1629.54|
|      3|28/03/2016|   31/03/2016|      2|     1722|        13|      1674|     42837
|430.32|       1|     430.32|
|      4|23/10/2017|   24/10/2017|      3|     2876|        13|      1674|     42834
|818.84|       2|    1637.68|
|      5|22/11/2017|   25/11/2017|      2|      678|        13|      1674|     42825
|554.18|       3|    1662.54|
+-------+----------+-------------+-------+---------+----------+----------+----------
+------+--------+-----------+
only showing top 5 rows

DIMDATE-DATAONLY:
```

```
23/07/08 18:11:05 INFO CodeGenerator: Code generated in 16.667712 ms
23/07/08 18:11:05 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:11:05 INFO DAGScheduler: Got job 1265 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:11:05 INFO DAGScheduler: Final stage: ResultStage 2706 (showString at <u
nknown>:0)
23/07/08 18:11:05 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:05 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:05 INFO DAGScheduler: Submitting ResultStage 2706 (MapPartitionsRDD[3
917] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:05 INFO MemoryStore: Block broadcast_1551 stored as values in memory
(estimated size 21.8 KiB, free 342.9 MiB)
23/07/08 18:11:05 INFO MemoryStore: Block broadcast_1551_piece0 stored as bytes in m
emory (estimated size 10.1 KiB, free 342.9 MiB)
23/07/08 18:11:05 INFO BlockManagerInfo: Added broadcast_1551_piece0 in memory on 17
2.30.115.138:43839 (size: 10.1 KiB, free: 364.0 MiB)
23/07/08 18:11:05 INFO SparkContext: Created broadcast 1551 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:05 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 270
6 (MapPartitionsRDD[3917] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:11:05 INFO TaskSchedulerImpl: Adding task set 2706.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:05 INFO TaskSetManager: Starting task 0.0 in stage 2706.0 (TID 1731)
(172.30.115.138, executor driver, partition 0, ANY, 7417 bytes)
23/07/08 18:11:05 INFO Executor: Running task 0.0 in stage 2706.0 (TID 1731)
23/07/08 18:11:05 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:0+1310749
23/07/08 18:11:05 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:05 INFO Executor: Finished task 0.0 in stage 2706.0 (TID 1731). 2324
bytes result sent to driver
23/07/08 18:11:05 INFO TaskSetManager: Finished task 0.0 in stage 2706.0 (TID 1731)
in 131 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:05 INFO TaskSchedulerImpl: Removed TaskSet 2706.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:05 INFO DAGScheduler: ResultStage 2706 (showString at <unknown>:0) fi
nished in 0.135 s
23/07/08 18:11:05 INFO DAGScheduler: Job 1265 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:05 INFO TaskSchedulerImpl: Killing all running tasks in stage 2706: S
tage finished
23/07/08 18:11:05 INFO DAGScheduler: Job 1265 finished: showString at <unknown>:0, t
ook 0.136694 s
23/07/08 18:11:05 INFO CodeGenerator: Code generated in 6.342837 ms
23/07/08 18:11:05 INFO CodeGenerator: Code generated in 10.001542 ms
23/07/08 18:11:05 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:11:05 INFO DAGScheduler: Got job 1266 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:11:05 INFO DAGScheduler: Final stage: ResultStage 2707 (showString at <u
nknown>:0)
23/07/08 18:11:05 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:05 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:05 INFO DAGScheduler: Submitting ResultStage 2707 (MapPartitionsRDD[3
919] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:05 INFO MemoryStore: Block broadcast_1552 stored as values in memory
(estimated size 21.9 KiB, free 342.9 MiB)
```

```
23/07/08 18:11:05 INFO MemoryStore: Block broadcast_1552_piece0 stored as bytes in m
emory (estimated size 9.8 KiB, free 342.9 MiB)
23/07/08 18:11:05 INFO BlockManagerInfo: Added broadcast_1552_piece0 in memory on 17
2.30.115.138:43839 (size: 9.8 KiB, free: 364.0 MiB)
23/07/08 18:11:05 INFO SparkContext: Created broadcast 1552 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:05 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 270
7 (MapPartitionsRDD[3919] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:11:05 INFO TaskSchedulerImpl: Adding task set 2707.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:05 INFO TaskSetManager: Starting task 0.0 in stage 2707.0 (TID 1732)
(172.30.115.138, executor driver, partition 0, ANY, 7431 bytes)
23/07/08 18:11:05 INFO Executor: Running task 0.0 in stage 2707.0 (TID 1732)
23/07/08 18:11:05 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE-DATAONLY.csv:0+365617
23/07/08 18:11:05 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:05 INFO Executor: Finished task 0.0 in stage 2707.0 (TID 1732). 2120
bytes result sent to driver
23/07/08 18:11:05 INFO TaskSetManager: Finished task 0.0 in stage 2707.0 (TID 1732)
in 132 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:05 INFO TaskSchedulerImpl: Removed TaskSet 2707.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:05 INFO DAGScheduler: ResultStage 2707 (showString at <unknown>:0) fi
nished in 0.136 s
23/07/08 18:11:05 INFO DAGScheduler: Job 1266 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:05 INFO TaskSchedulerImpl: Killing all running tasks in stage 2707: S
tage finished
23/07/08 18:11:05 INFO DAGScheduler: Job 1266 finished: showString at <unknown>:0, t
ook 0.137325 s
23/07/08 18:11:06 INFO CodeGenerator: Code generated in 8.132989 ms
23/07/08 18:11:06 INFO CodeGenerator: Code generated in 8.721889 ms
23/07/08 18:11:06 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:11:06 INFO DAGScheduler: Got job 1267 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:11:06 INFO DAGScheduler: Final stage: ResultStage 2708 (showString at <u
nknown>:0)
23/07/08 18:11:06 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:06 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:06 INFO DAGScheduler: Submitting ResultStage 2708 (MapPartitionsRDD[3
921] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:06 INFO MemoryStore: Block broadcast_1553 stored as values in memory
(estimated size 22.3 KiB, free 342.9 MiB)
23/07/08 18:11:06 INFO MemoryStore: Block broadcast_1553_piece0 stored as bytes in m
emory (estimated size 9.8 KiB, free 342.9 MiB)
23/07/08 18:11:06 INFO BlockManagerInfo: Added broadcast_1553_piece0 in memory on 17
2.30.115.138:43839 (size: 9.8 KiB, free: 364.0 MiB)
23/07/08 18:11:06 INFO SparkContext: Created broadcast 1553 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:06 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 270
8 (MapPartitionsRDD[3921] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:11:06 INFO TaskSchedulerImpl: Adding task set 2708.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:06 INFO TaskSetManager: Starting task 0.0 in stage 2708.0 (TID 1733)
```

```
(172.30.115.138, executor driver, partition 0, ANY, 7422 bytes)
23/07/08 18:11:06 INFO Executor: Running task 0.0 in stage 2708.0 (TID 1733)
23/07/08 18:11:06 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:0+450275
23/07/08 18:11:06 INFO LineRecordReader: Found UTF-8 BOM and skipped it
+--------+----------+---+---+----+------+--------+---------+-----------+--------+--
---------+---------+---+---------+
|      ID|     FECHA|DIA|MES| AÑO|SEMANA|BIMESTRE|TRIMESTRE|CUATRIMESTRE|SEMESTRE|FI
NDESEMANA|DIADELAÑO|AÑO|DIASEMANA|
+--------+----------+---+---+----+------+--------+---------+-----------+--------+--
---------+---------+---+---------+
|20100101|01/01/2010|  1|  1|2010|     1|       1|        1|           1|       1|
0|        1|  1|        5|
|20100102|02/01/2010|  2|  1|2010|     1|       1|        1|           1|       1|
1|        2|  1|        6|
|20100103|03/01/2010|  3|  1|2010|     1|       1|        1|           1|       1|
1|        3|  1|        7|
|20100104|04/01/2010|  4|  1|2010|     1|       1|        1|           1|       1|
0|        4|  1|        1|
|20100105|05/01/2010|  5|  1|2010|     1|       1|        1|           1|       1|
0|        5|  1|        2|
+--------+----------+---+---+----+------+--------+---------+-----------+--------+--
---------+---------+---+---------+
only showing top 5 rows

DIMDATE:
+--------+----------+---+---+----+------+--------+---------+-----------+--------+--
---------+---------+----+---------+--------------+
|      ID|     FECHA|DIA|MES| AÑO|SEMANA|BIMESTRE|TRIMESTRE|CUATRIMESTRE|SEMESTRE|FI
NDESEMANA|DIADELAÑO|AÑO2|DIASEMANA|DIASEMANATEXTO|
+--------+----------+---+---+----+------+--------+---------+-----------+--------+--
---------+---------+----+---------+--------------+
|20100101|01/01/2010|  1|  1|2010|     1|       1|        1|           1|       1|
0|        1|   1|        5|       VIERNES|
|20100102|02/01/2010|  2|  1|2010|     1|       1|        1|           1|       1|
1|        2|   1|        6|        SABADO|
|20100103|03/01/2010|  3|  1|2010|     1|       1|        1|           1|       1|
1|        3|   1|        7|       DOMINGO|
|20100104|04/01/2010|  4|  1|2010|     1|       1|        1|           1|       1|
0|        4|   1|        1|         LUNES|
|20100105|05/01/2010|  5|  1|2010|     1|       1|        1|           1|       1|
0|        5|   1|        2|        MARTES|
+--------+----------+---+---+----+------+--------+---------+-----------+--------+--
---------+---------+----+---------+--------------+
only showing top 5 rows

DIMTIME-DATAONLY:
```

```
23/07/08 18:11:06 INFO Executor: Finished task 0.0 in stage 2708.0 (TID 1733). 2197
bytes result sent to driver
23/07/08 18:11:06 INFO TaskSetManager: Finished task 0.0 in stage 2708.0 (TID 1733)
in 128 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:06 INFO TaskSchedulerImpl: Removed TaskSet 2708.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:06 INFO DAGScheduler: ResultStage 2708 (showString at <unknown>:0) fi
nished in 0.133 s
23/07/08 18:11:06 INFO DAGScheduler: Job 1267 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:06 INFO TaskSchedulerImpl: Killing all running tasks in stage 2708: S
tage finished
23/07/08 18:11:06 INFO DAGScheduler: Job 1267 finished: showString at <unknown>:0, t
ook 0.134573 s
23/07/08 18:11:06 INFO CodeGenerator: Code generated in 10.630056 ms
23/07/08 18:11:06 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:11:06 INFO DAGScheduler: Got job 1268 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:11:06 INFO DAGScheduler: Final stage: ResultStage 2709 (showString at <u
nknown>:0)
23/07/08 18:11:06 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:06 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:06 INFO DAGScheduler: Submitting ResultStage 2709 (MapPartitionsRDD[3
923] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:06 INFO MemoryStore: Block broadcast_1554 stored as values in memory
(estimated size 19.0 KiB, free 342.8 MiB)
23/07/08 18:11:06 INFO MemoryStore: Block broadcast_1554_piece0 stored as bytes in m
emory (estimated size 9.3 KiB, free 342.8 MiB)
23/07/08 18:11:06 INFO BlockManagerInfo: Added broadcast_1554_piece0 in memory on 17
2.30.115.138:43839 (size: 9.3 KiB, free: 364.0 MiB)
23/07/08 18:11:06 INFO SparkContext: Created broadcast 1554 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:06 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 270
9 (MapPartitionsRDD[3923] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:11:06 INFO TaskSchedulerImpl: Adding task set 2709.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:06 INFO TaskSetManager: Starting task 0.0 in stage 2709.0 (TID 1734)
(172.30.115.138, executor driver, partition 0, ANY, 7431 bytes)
23/07/08 18:11:06 INFO Executor: Running task 0.0 in stage 2709.0 (TID 1734)
23/07/08 18:11:06 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMTI
ME-DATAONLY.csv:0+75866
23/07/08 18:11:06 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:06 INFO Executor: Finished task 0.0 in stage 2709.0 (TID 1734). 2071
bytes result sent to driver
23/07/08 18:11:06 INFO TaskSetManager: Finished task 0.0 in stage 2709.0 (TID 1734)
in 129 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:06 INFO TaskSchedulerImpl: Removed TaskSet 2709.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:06 INFO DAGScheduler: ResultStage 2709 (showString at <unknown>:0) fi
nished in 0.133 s
23/07/08 18:11:06 INFO DAGScheduler: Job 1268 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:06 INFO TaskSchedulerImpl: Killing all running tasks in stage 2709: S
tage finished
23/07/08 18:11:06 INFO DAGScheduler: Job 1268 finished: showString at <unknown>:0, t
```

```
ook 0.135029 s
23/07/08 18:11:06 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:11:06 INFO DAGScheduler: Got job 1269 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:11:06 INFO DAGScheduler: Final stage: ResultStage 2710 (showString at <u
nknown>:0)
23/07/08 18:11:06 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:06 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:06 INFO DAGScheduler: Submitting ResultStage 2710 (MapPartitionsRDD[3
925] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:06 INFO MemoryStore: Block broadcast_1555 stored as values in memory
(estimated size 19.0 KiB, free 342.8 MiB)
23/07/08 18:11:06 INFO MemoryStore: Block broadcast_1555_piece0 stored as bytes in m
emory (estimated size 9.2 KiB, free 342.8 MiB)
23/07/08 18:11:06 INFO BlockManagerInfo: Added broadcast_1555_piece0 in memory on 17
2.30.115.138:43839 (size: 9.2 KiB, free: 364.0 MiB)
23/07/08 18:11:06 INFO SparkContext: Created broadcast 1555 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:06 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 271
0 (MapPartitionsRDD[3925] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:11:06 INFO TaskSchedulerImpl: Adding task set 2710.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:06 INFO TaskSetManager: Starting task 0.0 in stage 2710.0 (TID 1735)
(172.30.115.138, executor driver, partition 0, ANY, 7422 bytes)
23/07/08 18:11:06 INFO Executor: Running task 0.0 in stage 2710.0 (TID 1735)
23/07/08 18:11:06 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMTI
ME.csv:0+22350
```

```
+----+------+----+------+---------------+-----------+-----+
|  ID|TIEMPO|HORA|MINUTO|BLOQUEMEDIAHORA|HORAOFICINA|AM-PM|
+----+------+----+------+---------------+-----------+-----+
|0000| 00:00|   0|     0|          0:00 |         NO|   AM|
|0001| 00:01|   0|     1|          0:00 |         NO|   AM|
|0002| 00:02|   0|     2|          0:00 |         NO|   AM|
|0003| 00:03|   0|     3|          0:00 |         NO|   AM|
|0004| 00:04|   0|     4|          0:00 |         NO|   AM|
+----+------+----+------+---------------+-----------+-----+
only showing top 5 rows

DIMTIME:
+----+------+----+------+---------------+-----------+-----+
|  ID|TIEMPO|HORA|MINUTO|BLOQUEMEDIAHORA|HORAOFICINA|AM-PM|
+----+------+----+------+---------------+-----------+-----+
|0000| 00:00|   0|     0|          0:00 |         NO|   AM|
|0001| 00:01|   0|     1|          0:00 |         NO|   AM|
|0002| 00:02|   0|     2|          0:00 |         NO|   AM|
|0003| 00:03|   0|     3|          0:00 |         NO|   AM|
|0004| 00:04|   0|     4|          0:00 |         NO|   AM|
+----+------+----+------+---------------+-----------+-----+
only showing top 5 rows
```

```
23/07/08 18:11:06 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:06 INFO Executor: Finished task 0.0 in stage 2710.0 (TID 1735). 2071
bytes result sent to driver
23/07/08 18:11:06 INFO TaskSetManager: Finished task 0.0 in stage 2710.0 (TID 1735)
in 124 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:06 INFO TaskSchedulerImpl: Removed TaskSet 2710.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:06 INFO DAGScheduler: ResultStage 2710 (showString at <unknown>:0) fi
nished in 0.130 s
23/07/08 18:11:06 INFO DAGScheduler: Job 1269 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:06 INFO TaskSchedulerImpl: Killing all running tasks in stage 2710: S
tage finished
23/07/08 18:11:06 INFO DAGScheduler: Job 1269 finished: showString at <unknown>:0, t
ook 0.131541 s
```

Relaciones del modelo proporcionado.

In [192…

```
r1 = df_canal.join(df_venta, df_canal["CODIGO"] == df_venta["IdCanal"])
r2 = df_cliente.join(df_venta, df_cliente["ID"] == df_venta["IdCliente"])
r3 = df_producto.join(df_venta, df_producto["ID_PRODUCTO"] == df_venta["IdProducto"
r4 = df_empleado.join(df_venta, df_empleado["ID_empleado"] == df_venta["IdEmpleado"
r5 = df_sucursal.join(df_venta, df_sucursal["ID"] == df_venta["IdSucursal"])
r6 = df_DIMDATE.join(df_venta, df_DIMDATE["FECHA"] == df_venta["Fecha"])

r1.show(5)
```

```
23/07/08 18:11:12 INFO DAGScheduler: Registering RDD 3927 (showString at <unknown>:
0) as input to shuffle 644
23/07/08 18:11:12 INFO DAGScheduler: Got map stage job 1270 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:11:12 INFO DAGScheduler: Final stage: ShuffleMapStage 2711 (showString a
t <unknown>:0)
23/07/08 18:11:12 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:12 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:12 INFO DAGScheduler: Submitting ShuffleMapStage 2711 (MapPartitionsR
DD[3927] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:12 INFO MemoryStore: Block broadcast_1556 stored as values in memory
(estimated size 20.1 KiB, free 342.8 MiB)
23/07/08 18:11:12 INFO MemoryStore: Block broadcast_1556_piece0 stored as bytes in m
emory (estimated size 10.4 KiB, free 342.8 MiB)
23/07/08 18:11:12 INFO BlockManagerInfo: Added broadcast_1556_piece0 in memory on 17
2.30.115.138:43839 (size: 10.4 KiB, free: 364.0 MiB)
23/07/08 18:11:12 INFO CodeGenerator: Code generated in 16.107678 ms
23/07/08 18:11:12 INFO SparkContext: Created broadcast 1556 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:12 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2711 (MapPartitionsRDD[3927] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:11:12 INFO TaskSchedulerImpl: Adding task set 2711.0 with 2 tasks resour
ce profile 0
23/07/08 18:11:12 INFO DAGScheduler: Registering RDD 3929 (showString at <unknown>:
0) as input to shuffle 645
23/07/08 18:11:12 INFO DAGScheduler: Got map stage job 1271 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:11:12 INFO DAGScheduler: Final stage: ShuffleMapStage 2712 (showString a
t <unknown>:0)
23/07/08 18:11:12 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:12 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:12 INFO DAGScheduler: Submitting ShuffleMapStage 2712 (MapPartitionsR
DD[3929] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:12 INFO TaskSetManager: Starting task 0.0 in stage 2711.0 (TID 1736)
(172.30.115.138, executor driver, partition 0, ANY, 7412 bytes)
23/07/08 18:11:12 INFO TaskSetManager: Starting task 1.0 in stage 2711.0 (TID 1737)
(172.30.115.138, executor driver, partition 1, ANY, 7412 bytes)
23/07/08 18:11:12 INFO Executor: Running task 0.0 in stage 2711.0 (TID 1736)
23/07/08 18:11:12 INFO Executor: Running task 1.0 in stage 2711.0 (TID 1737)
23/07/08 18:11:12 INFO MemoryStore: Block broadcast_1557 stored as values in memory
(estimated size 25.6 KiB, free 342.7 MiB)
23/07/08 18:11:12 INFO MemoryStore: Block broadcast_1557_piece0 stored as bytes in m
emory (estimated size 11.8 KiB, free 342.7 MiB)
23/07/08 18:11:12 INFO BlockManagerInfo: Added broadcast_1557_piece0 in memory on 17
2.30.115.138:43839 (size: 11.8 KiB, free: 364.0 MiB)
23/07/08 18:11:12 INFO SparkContext: Created broadcast 1557 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:12 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2712 (MapPartitionsRDD[3929] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:11:12 INFO TaskSchedulerImpl: Adding task set 2712.0 with 2 tasks resour
ce profile 0
23/07/08 18:11:12 INFO TaskSetManager: Starting task 0.0 in stage 2712.0 (TID 1738)
(172.30.115.138, executor driver, partition 0, ANY, 7406 bytes)
23/07/08 18:11:12 INFO TaskSetManager: Starting task 1.0 in stage 2712.0 (TID 1739)
```

```
(172.30.115.138, executor driver, partition 1, ANY, 7406 bytes)
23/07/08 18:11:12 INFO Executor: Running task 0.0 in stage 2712.0 (TID 1738)
23/07/08 18:11:12 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/CanalDeVe
nta.csv:29+29
23/07/08 18:11:12 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/CanalDeVe
nta.csv:0+29
23/07/08 18:11:12 INFO Executor: Running task 1.0 in stage 2712.0 (TID 1739)
23/07/08 18:11:12 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:1310749+1310750
23/07/08 18:11:12 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:0+1310749
23/07/08 18:11:12 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:12 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:12 INFO BlockManagerInfo: Removed broadcast_1554_piece0 on 172.30.11
5.138:43839 in memory (size: 9.3 KiB, free: 364.0 MiB)
23/07/08 18:11:12 INFO BlockManagerInfo: Removed broadcast_1548_piece0 on 172.30.11
5.138:43839 in memory (size: 9.3 KiB, free: 364.0 MiB)
23/07/08 18:11:12 INFO BlockManagerInfo: Removed broadcast_1546_piece0 on 172.30.11
5.138:43839 in memory (size: 8.9 KiB, free: 364.0 MiB)
23/07/08 18:11:12 INFO BlockManagerInfo: Removed broadcast_1550_piece0 on 172.30.11
5.138:43839 in memory (size: 9.2 KiB, free: 364.0 MiB)
23/07/08 18:11:12 INFO BlockManagerInfo: Removed broadcast_1555_piece0 on 172.30.11
5.138:43839 in memory (size: 9.2 KiB, free: 364.0 MiB)
23/07/08 18:11:12 INFO BlockManagerInfo: Removed broadcast_1553_piece0 on 172.30.11
5.138:43839 in memory (size: 9.8 KiB, free: 364.0 MiB)
23/07/08 18:11:12 INFO BlockManagerInfo: Removed broadcast_1552_piece0 on 172.30.11
5.138:43839 in memory (size: 9.8 KiB, free: 364.0 MiB)
23/07/08 18:11:12 INFO BlockManagerInfo: Removed broadcast_1547_piece0 on 172.30.11
5.138:43839 in memory (size: 9.5 KiB, free: 364.0 MiB)
23/07/08 18:11:12 INFO BlockManagerInfo: Removed broadcast_1549_piece0 on 172.30.11
5.138:43839 in memory (size: 9.0 KiB, free: 364.0 MiB)
23/07/08 18:11:12 INFO BlockManagerInfo: Removed broadcast_1551_piece0 on 172.30.11
5.138:43839 in memory (size: 10.1 KiB, free: 364.1 MiB)
23/07/08 18:11:12 INFO PythonRunner: Times: total = 158, boot = 10, init = 146, fini
sh = 2
23/07/08 18:11:12 INFO Executor: Finished task 0.0 in stage 2711.0 (TID 1736). 2524
bytes result sent to driver
23/07/08 18:11:12 INFO TaskSetManager: Finished task 0.0 in stage 2711.0 (TID 1736)
in 214 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:11:12 INFO PythonRunner: Times: total = 185, boot = 17, init = 162, fini
sh = 6
23/07/08 18:11:12 INFO Executor: Finished task 1.0 in stage 2711.0 (TID 1737). 2524
bytes result sent to driver
23/07/08 18:11:12 INFO TaskSetManager: Finished task 1.0 in stage 2711.0 (TID 1737)
in 219 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:11:12 INFO TaskSchedulerImpl: Removed TaskSet 2711.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:12 INFO DAGScheduler: ShuffleMapStage 2711 (showString at <unknown>:
0) finished in 0.242 s
23/07/08 18:11:12 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:11:12 INFO DAGScheduler: running: Set(ShuffleMapStage 2712)
23/07/08 18:11:12 INFO DAGScheduler: waiting: Set()
23/07/08 18:11:12 INFO DAGScheduler: failed: Set()
23/07/08 18:11:12 INFO PythonRunner: Times: total = 337, boot = 10, init = 175, fini
sh = 152
23/07/08 18:11:12 INFO Executor: Finished task 0.0 in stage 2712.0 (TID 1738). 2524
```

```
bytes result sent to driver
23/07/08 18:11:12 INFO TaskSetManager: Finished task 0.0 in stage 2712.0 (TID 1738)
in 414 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:11:12 INFO PythonRunner: Times: total = 324, boot = 13, init = 169, fini
sh = 142
23/07/08 18:11:13 INFO Executor: Finished task 1.0 in stage 2712.0 (TID 1739). 2524
bytes result sent to driver
23/07/08 18:11:13 INFO TaskSetManager: Finished task 1.0 in stage 2712.0 (TID 1739)
in 424 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:11:13 INFO TaskSchedulerImpl: Removed TaskSet 2712.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:13 INFO DAGScheduler: ShuffleMapStage 2712 (showString at <unknown>:
0) finished in 0.432 s
23/07/08 18:11:13 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:11:13 INFO DAGScheduler: running: Set()
23/07/08 18:11:13 INFO DAGScheduler: waiting: Set()
23/07/08 18:11:13 INFO DAGScheduler: failed: Set()
23/07/08 18:11:13 INFO ShufflePartitionsUtil: For shuffle(644, 645), advisory target
size: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:11:13 INFO CodeGenerator: Code generated in 16.146883 ms
23/07/08 18:11:13 INFO CodeGenerator: Code generated in 9.855484 ms
23/07/08 18:11:13 INFO CodeGenerator: Code generated in 11.507634 ms
23/07/08 18:11:13 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:11:13 INFO DAGScheduler: Got job 1272 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:11:13 INFO DAGScheduler: Final stage: ResultStage 2715 (showString at <u
nknown>:0)
23/07/08 18:11:13 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
13, ShuffleMapStage 2714)
23/07/08 18:11:13 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:13 INFO DAGScheduler: Submitting ResultStage 2715 (MapPartitionsRDD[3
936] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:13 INFO MemoryStore: Block broadcast_1558 stored as values in memory
(estimated size 60.5 KiB, free 343.0 MiB)
23/07/08 18:11:13 INFO MemoryStore: Block broadcast_1558_piece0 stored as bytes in m
emory (estimated size 27.8 KiB, free 342.9 MiB)
23/07/08 18:11:13 INFO BlockManagerInfo: Added broadcast_1558_piece0 in memory on 17
2.30.115.138:43839 (size: 27.8 KiB, free: 364.0 MiB)
23/07/08 18:11:13 INFO SparkContext: Created broadcast 1558 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:13 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 271
5 (MapPartitionsRDD[3936] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:11:13 INFO TaskSchedulerImpl: Adding task set 2715.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:13 INFO TaskSetManager: Starting task 0.0 in stage 2715.0 (TID 1740)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7645 bytes)
23/07/08 18:11:13 INFO Executor: Running task 0.0 in stage 2715.0 (TID 1740)
23/07/08 18:11:13 INFO ShuffleBlockFetcherIterator: Getting 2 (194.0 B) non-empty bl
ocks including 2 (194.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-
local and 0 (0.0 B) remote blocks
23/07/08 18:11:13 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:11:13 INFO ShuffleBlockFetcherIterator: Getting 2 (1419.3 KiB) non-empty
blocks including 2 (1419.3 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:11:13 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
```

```
+------+-----------+-------+----------+-------------+-------+---------+----------+--
--------+----------+------+--------+-----------+
|CODIGO|DESCRIPCION|IdVenta|     Fecha|Fecha_Entrega|IdCanal|IdCliente|IdSucursal|Id
Empleado|IdProducto|Precio|Cantidad|total_venta|
+------+-----------+-------+----------+-------------+-------+---------+----------+--
--------+----------+------+--------+-----------+
|     1| Telefónica|     10|16/03/2019|   17/03/2019|      1|     1003|        13|
1674|     42894|   515|       2|     1030.0|
|     1| Telefónica|     12|17/02/2015|   18/02/2015|      1|     2866|        13|
1674|     42969|  3839|       2|     7678.0|
|     1| Telefónica|     30|07/05/2019|   16/05/2019|      1|     2939|        13|
1674|     42861|   542|       1|      542.0|
|     1| Telefónica|     35|29/12/2015|   06/01/2016|      1|      895|        13|
1674|     43016|   454|       3|     1362.0|
|     1| Telefónica|     36|28/01/2016|   31/01/2016|      1|      572|        13|
1674|     42970|  3001|       2|     6002.0|
+------+-----------+-------+----------+-------------+-------+---------+----------+--
--------+----------+------+--------+-----------+
only showing top 5 rows
```

```
23/07/08 18:11:13 INFO Executor: Finished task 0.0 in stage 2715.0 (TID 1740). 6992
bytes result sent to driver
23/07/08 18:11:13 INFO TaskSetManager: Finished task 0.0 in stage 2715.0 (TID 1740)
in 108 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:13 INFO TaskSchedulerImpl: Removed TaskSet 2715.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:13 INFO DAGScheduler: ResultStage 2715 (showString at <unknown>:0) fi
nished in 0.115 s
23/07/08 18:11:13 INFO DAGScheduler: Job 1272 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:13 INFO TaskSchedulerImpl: Killing all running tasks in stage 2715: S
tage finished
23/07/08 18:11:13 INFO DAGScheduler: Job 1272 finished: showString at <unknown>:0, t
ook 0.118718 s
```

a) Calcular las ventas anuales de cada sucursal, crear un archivo de salida con el resultado

In [193…

```
from pyspark.sql.functions import desc, format_number, col

Sucursal = df_sucursal.alias("sucursal")

# Calcular las ventas anuales de cada sucursal
ventas_anuales_sucursal = r6.join(Sucursal, r5["IdSucursal"] == Sucursal["ID"]) \
    .groupBy(Sucursal["Sucursal"], r6["AÑO"]) \
    .agg(sum(r5["total_venta"]).alias("VentasAnuales"))

ventas_anuales_sucursal = ventas_anuales_sucursal.withColumn("Ventas_Anuales_format
format_number(col("VentasAnuales"), 2).alias("Ventas_Anuales_format"))

#Guardando el Archivo
ventas_anuales_sucursal.write.csv("/datos/salida/salida1.csv", header=True, mode="o
ventas_anuales_sucursal.show()
```

```
23/07/08 18:11:14 INFO DAGScheduler: Registering RDD 3938 (csv at <unknown>:0) as in
put to shuffle 646
23/07/08 18:11:14 INFO DAGScheduler: Got map stage job 1273 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:11:14 INFO DAGScheduler: Final stage: ShuffleMapStage 2716 (csv at <unkn
own>:0)
23/07/08 18:11:14 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:14 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:14 INFO DAGScheduler: Submitting ShuffleMapStage 2716 (MapPartitionsR
DD[3938] at csv at <unknown>:0), which has no missing parents
23/07/08 18:11:14 INFO MemoryStore: Block broadcast_1559 stored as values in memory
(estimated size 22.5 KiB, free 342.9 MiB)
23/07/08 18:11:14 INFO MemoryStore: Block broadcast_1559_piece0 stored as bytes in m
emory (estimated size 10.8 KiB, free 342.9 MiB)
23/07/08 18:11:14 INFO BlockManagerInfo: Added broadcast_1559_piece0 in memory on 17
2.30.115.138:43839 (size: 10.8 KiB, free: 364.0 MiB)
23/07/08 18:11:14 INFO SparkContext: Created broadcast 1559 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:14 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2716 (MapPartitionsRDD[3938] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:11:14 INFO TaskSchedulerImpl: Adding task set 2716.0 with 2 tasks resour
ce profile 0
23/07/08 18:11:14 INFO TaskSetManager: Starting task 0.0 in stage 2716.0 (TID 1741)
(172.30.115.138, executor driver, partition 0, ANY, 7411 bytes)
23/07/08 18:11:14 INFO TaskSetManager: Starting task 1.0 in stage 2716.0 (TID 1742)
(172.30.115.138, executor driver, partition 1, ANY, 7411 bytes)
23/07/08 18:11:14 INFO Executor: Running task 0.0 in stage 2716.0 (TID 1741)
23/07/08 18:11:14 INFO Executor: Running task 1.0 in stage 2716.0 (TID 1742)
23/07/08 18:11:14 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:450275+450275
23/07/08 18:11:14 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:0+450275
23/07/08 18:11:14 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:14 INFO CodeGenerator: Code generated in 23.482324 ms
23/07/08 18:11:14 INFO DAGScheduler: Registering RDD 3940 (csv at <unknown>:0) as in
put to shuffle 647
23/07/08 18:11:14 INFO DAGScheduler: Got map stage job 1274 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:11:14 INFO DAGScheduler: Final stage: ShuffleMapStage 2717 (csv at <unkn
own>:0)
23/07/08 18:11:14 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:14 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:14 INFO DAGScheduler: Submitting ShuffleMapStage 2717 (MapPartitionsR
DD[3940] at csv at <unknown>:0), which has no missing parents
23/07/08 18:11:14 INFO MemoryStore: Block broadcast_1560 stored as values in memory
(estimated size 23.3 KiB, free 342.9 MiB)
23/07/08 18:11:14 INFO MemoryStore: Block broadcast_1560_piece0 stored as bytes in m
emory (estimated size 11.3 KiB, free 342.9 MiB)
23/07/08 18:11:14 INFO BlockManagerInfo: Added broadcast_1560_piece0 in memory on 17
2.30.115.138:43839 (size: 11.3 KiB, free: 364.0 MiB)
23/07/08 18:11:14 INFO SparkContext: Created broadcast 1560 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:14 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2717 (MapPartitionsRDD[3940] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
```

```
23/07/08 18:11:14 INFO TaskSchedulerImpl: Adding task set 2717.0 with 2 tasks resour
ce profile 0
23/07/08 18:11:14 INFO DAGScheduler: Registering RDD 3942 (csv at <unknown>:0) as in
put to shuffle 648
23/07/08 18:11:14 INFO DAGScheduler: Got map stage job 1275 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:11:14 INFO DAGScheduler: Final stage: ShuffleMapStage 2718 (csv at <unkn
own>:0)
23/07/08 18:11:14 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:14 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:14 INFO TaskSetManager: Starting task 0.0 in stage 2717.0 (TID 1743)
(172.30.115.138, executor driver, partition 0, ANY, 7406 bytes)
23/07/08 18:11:14 INFO TaskSetManager: Starting task 1.0 in stage 2717.0 (TID 1744)
(172.30.115.138, executor driver, partition 1, ANY, 7406 bytes)
23/07/08 18:11:14 INFO DAGScheduler: Submitting ShuffleMapStage 2718 (MapPartitionsR
DD[3942] at csv at <unknown>:0), which has no missing parents
23/07/08 18:11:14 INFO Executor: Running task 0.0 in stage 2717.0 (TID 1743)
23/07/08 18:11:14 INFO Executor: Running task 1.0 in stage 2717.0 (TID 1744)
23/07/08 18:11:14 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:0+1310749
23/07/08 18:11:14 INFO MemoryStore: Block broadcast_1561 stored as values in memory
(estimated size 21.1 KiB, free 342.8 MiB)
23/07/08 18:11:14 INFO MemoryStore: Block broadcast_1561_piece0 stored as bytes in m
emory (estimated size 10.6 KiB, free 342.8 MiB)
23/07/08 18:11:14 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:1310749+1310750
23/07/08 18:11:14 INFO BlockManagerInfo: Added broadcast_1561_piece0 in memory on 17
2.30.115.138:43839 (size: 10.6 KiB, free: 364.0 MiB)
23/07/08 18:11:14 INFO SparkContext: Created broadcast 1561 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:14 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2718 (MapPartitionsRDD[3942] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:11:14 INFO TaskSchedulerImpl: Adding task set 2718.0 with 2 tasks resour
ce profile 0
23/07/08 18:11:14 INFO TaskSetManager: Starting task 0.0 in stage 2718.0 (TID 1745)
(172.30.115.138, executor driver, partition 0, ANY, 7408 bytes)
23/07/08 18:11:14 INFO TaskSetManager: Starting task 1.0 in stage 2718.0 (TID 1746)
(172.30.115.138, executor driver, partition 1, ANY, 7408 bytes)
23/07/08 18:11:14 INFO Executor: Running task 0.0 in stage 2718.0 (TID 1745)
23/07/08 18:11:14 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:14 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Sucursal.
csv:0+1266
23/07/08 18:11:15 INFO Executor: Running task 1.0 in stage 2718.0 (TID 1746)
23/07/08 18:11:15 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Sucursal.
csv:1266+1266
23/07/08 18:11:15 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:15 INFO PythonRunner: Times: total = 203, boot = -2119, init = 2259,
finish = 63
23/07/08 18:11:15 INFO PythonRunner: Times: total = 212, boot = -2095, init = 2235,
finish = 72
23/07/08 18:11:15 INFO Executor: Finished task 1.0 in stage 2716.0 (TID 1742). 2481
bytes result sent to driver
23/07/08 18:11:15 INFO TaskSetManager: Finished task 1.0 in stage 2716.0 (TID 1742)
in 297 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:11:15 INFO PythonRunner: Times: total = 198, boot = 10, init = 187, fini
```

```
sh = 1
23/07/08 18:11:15 INFO Executor: Finished task 0.0 in stage 2718.0 (TID 1745). 2567
bytes result sent to driver
23/07/08 18:11:15 INFO PythonRunner: Times: total = 182, boot = 10, init = 172, fini
sh = 0
23/07/08 18:11:15 INFO Executor: Finished task 1.0 in stage 2718.0 (TID 1746). 2524
bytes result sent to driver
23/07/08 18:11:15 INFO TaskSetManager: Finished task 1.0 in stage 2718.0 (TID 1746)
in 363 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:11:15 INFO Executor: Finished task 0.0 in stage 2716.0 (TID 1741). 2524
bytes result sent to driver
23/07/08 18:11:15 INFO TaskSetManager: Finished task 0.0 in stage 2718.0 (TID 1745)
in 363 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:11:15 INFO TaskSchedulerImpl: Removed TaskSet 2718.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:15 INFO TaskSetManager: Finished task 0.0 in stage 2716.0 (TID 1741)
in 409 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:11:15 INFO TaskSchedulerImpl: Removed TaskSet 2716.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:15 INFO DAGScheduler: ShuffleMapStage 2718 (csv at <unknown>:0) finis
hed in 0.375 s
23/07/08 18:11:15 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:11:15 INFO DAGScheduler: running: Set(ShuffleMapStage 2716, ShuffleMapSt
age 2717)
23/07/08 18:11:15 INFO DAGScheduler: waiting: Set()
23/07/08 18:11:15 INFO DAGScheduler: failed: Set()
23/07/08 18:11:15 INFO DAGScheduler: ShuffleMapStage 2716 (csv at <unknown>:0) finis
hed in 0.417 s
23/07/08 18:11:15 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:11:15 INFO DAGScheduler: running: Set(ShuffleMapStage 2717)
23/07/08 18:11:15 INFO DAGScheduler: waiting: Set()
23/07/08 18:11:15 INFO DAGScheduler: failed: Set()
23/07/08 18:11:15 INFO ShufflePartitionsUtil: For shuffle(646), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:11:15 INFO SparkContext: Starting job: $anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266
23/07/08 18:11:15 INFO DAGScheduler: Got job 1276 ($anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266) with 1 output partitions
23/07/08 18:11:15 INFO DAGScheduler: Final stage: ResultStage 2720 ($anonfun$withThr
eadLocalCaptured$1 at FutureTask.java:266)
23/07/08 18:11:15 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
19)
23/07/08 18:11:15 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:15 INFO DAGScheduler: Submitting ResultStage 2720 (MapPartitionsRDD[3
944] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:266), which has no mis
sing parents
23/07/08 18:11:15 INFO MemoryStore: Block broadcast_1562 stored as values in memory
(estimated size 8.2 KiB, free 342.8 MiB)
23/07/08 18:11:15 INFO MemoryStore: Block broadcast_1562_piece0 stored as bytes in m
emory (estimated size 4.2 KiB, free 342.8 MiB)
23/07/08 18:11:15 INFO BlockManagerInfo: Added broadcast_1562_piece0 in memory on 17
2.30.115.138:43839 (size: 4.2 KiB, free: 364.0 MiB)
23/07/08 18:11:15 INFO SparkContext: Created broadcast 1562 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:15 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 272
0 (MapPartitionsRDD[3944] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:2
```

```
66) (first 15 tasks are for partitions Vector(0))
23/07/08 18:11:15 INFO TaskSchedulerImpl: Adding task set 2720.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:15 INFO TaskSetManager: Starting task 0.0 in stage 2720.0 (TID 1747)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7379 bytes)
23/07/08 18:11:15 INFO Executor: Running task 0.0 in stage 2720.0 (TID 1747)
23/07/08 18:11:15 INFO ShuffleBlockFetcherIterator: Getting 2 (234.6 KiB) non-empty
blocks including 2 (234.6 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:11:15 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:11:15 INFO Executor: Finished task 0.0 in stage 2720.0 (TID 1747). 17174
1 bytes result sent to driver
23/07/08 18:11:15 INFO TaskSetManager: Finished task 0.0 in stage 2720.0 (TID 1747)
in 13 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:15 INFO TaskSchedulerImpl: Removed TaskSet 2720.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:15 INFO DAGScheduler: ResultStage 2720 ($anonfun$withThreadLocalCaptu
red$1 at FutureTask.java:266) finished in 0.023 s
23/07/08 18:11:15 INFO DAGScheduler: Job 1276 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:15 INFO TaskSchedulerImpl: Killing all running tasks in stage 2720: S
tage finished
23/07/08 18:11:15 INFO DAGScheduler: Job 1276 finished: $anonfun$withThreadLocalCapt
ured$1 at FutureTask.java:266, took 0.025084 s
23/07/08 18:11:15 INFO MemoryStore: Block broadcast_1563 stored as values in memory
(estimated size 2.5 MiB, free 340.3 MiB)
23/07/08 18:11:15 INFO MemoryStore: Block broadcast_1563_piece0 stored as bytes in m
emory (estimated size 299.6 KiB, free 340.0 MiB)
23/07/08 18:11:15 INFO BlockManagerInfo: Added broadcast_1563_piece0 in memory on 17
2.30.115.138:43839 (size: 299.6 KiB, free: 363.7 MiB)
23/07/08 18:11:15 INFO SparkContext: Created broadcast 1563 from $anonfun$withThread
LocalCaptured$1 at FutureTask.java:266
23/07/08 18:11:15 INFO PythonRunner: Times: total = 356, boot = -2000, init = 2148,
finish = 208
23/07/08 18:11:15 INFO PythonRunner: Times: total = 396, boot = -1980, init = 2158,
finish = 218
23/07/08 18:11:15 INFO Executor: Finished task 1.0 in stage 2717.0 (TID 1744). 2524
bytes result sent to driver
23/07/08 18:11:15 INFO TaskSetManager: Finished task 1.0 in stage 2717.0 (TID 1744)
in 513 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:11:15 INFO Executor: Finished task 0.0 in stage 2717.0 (TID 1743). 2524
bytes result sent to driver
23/07/08 18:11:15 INFO TaskSetManager: Finished task 0.0 in stage 2717.0 (TID 1743)
in 520 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:11:15 INFO TaskSchedulerImpl: Removed TaskSet 2717.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:15 INFO DAGScheduler: ShuffleMapStage 2717 (csv at <unknown>:0) finis
hed in 0.526 s
23/07/08 18:11:15 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:11:15 INFO DAGScheduler: running: Set()
23/07/08 18:11:15 INFO DAGScheduler: waiting: Set()
23/07/08 18:11:15 INFO DAGScheduler: failed: Set()
23/07/08 18:11:15 INFO ShufflePartitionsUtil: For shuffle(647), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:11:15 INFO CodeGenerator: Code generated in 6.888735 ms
23/07/08 18:11:15 INFO DAGScheduler: Registering RDD 3947 (csv at <unknown>:0) as in
```

```
put to shuffle 649
23/07/08 18:11:15 INFO DAGScheduler: Got map stage job 1277 (csv at <unknown>:0) wit
h 1 output partitions
23/07/08 18:11:15 INFO DAGScheduler: Final stage: ShuffleMapStage 2722 (csv at <unkn
own>:0)
23/07/08 18:11:15 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
21)
23/07/08 18:11:15 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:15 INFO DAGScheduler: Submitting ShuffleMapStage 2722 (MapPartitionsR
DD[3947] at csv at <unknown>:0), which has no missing parents
23/07/08 18:11:15 INFO MemoryStore: Block broadcast_1564 stored as values in memory
(estimated size 15.3 KiB, free 340.0 MiB)
23/07/08 18:11:15 INFO MemoryStore: Block broadcast_1564_piece0 stored as bytes in m
emory (estimated size 7.4 KiB, free 340.0 MiB)
23/07/08 18:11:15 INFO BlockManagerInfo: Added broadcast_1564_piece0 in memory on 17
2.30.115.138:43839 (size: 7.4 KiB, free: 363.7 MiB)
23/07/08 18:11:15 INFO SparkContext: Created broadcast 1564 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:15 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
2722 (MapPartitionsRDD[3947] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0))
23/07/08 18:11:15 INFO TaskSchedulerImpl: Adding task set 2722.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:15 INFO TaskSetManager: Starting task 0.0 in stage 2722.0 (TID 1748)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7368 bytes)
23/07/08 18:11:15 INFO Executor: Running task 0.0 in stage 2722.0 (TID 1748)
23/07/08 18:11:15 INFO ShuffleBlockFetcherIterator: Getting 2 (479.7 KiB) non-empty
blocks including 2 (479.7 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:11:15 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:11:15 INFO Executor: Finished task 0.0 in stage 2722.0 (TID 1748). 4301
bytes result sent to driver
23/07/08 18:11:15 INFO TaskSetManager: Finished task 0.0 in stage 2722.0 (TID 1748)
in 62 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:15 INFO TaskSchedulerImpl: Removed TaskSet 2722.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:15 INFO DAGScheduler: ShuffleMapStage 2722 (csv at <unknown>:0) finis
hed in 0.065 s
23/07/08 18:11:15 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:11:15 INFO DAGScheduler: running: Set()
23/07/08 18:11:15 INFO DAGScheduler: waiting: Set()
23/07/08 18:11:15 INFO DAGScheduler: failed: Set()
23/07/08 18:11:15 INFO ShufflePartitionsUtil: For shuffle(649, 648), advisory target
size: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:11:15 INFO CodeGenerator: Code generated in 32.941127 ms
23/07/08 18:11:15 INFO BlockManagerInfo: Removed broadcast_1564_piece0 on 172.30.11
5.138:43839 in memory (size: 7.4 KiB, free: 363.7 MiB)
23/07/08 18:11:15 INFO BlockManagerInfo: Removed broadcast_1562_piece0 on 172.30.11
5.138:43839 in memory (size: 4.2 KiB, free: 363.7 MiB)
23/07/08 18:11:15 INFO DAGScheduler: Registering RDD 3954 (csv at <unknown>:0) as in
put to shuffle 650
23/07/08 18:11:15 INFO DAGScheduler: Got map stage job 1278 (csv at <unknown>:0) wit
h 1 output partitions
23/07/08 18:11:15 INFO DAGScheduler: Final stage: ShuffleMapStage 2726 (csv at <unkn
own>:0)
23/07/08 18:11:15 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
```

```
24, ShuffleMapStage 2725)
23/07/08 18:11:15 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:15 INFO DAGScheduler: Submitting ShuffleMapStage 2726 (MapPartitionsR
DD[3954] at csv at <unknown>:0), which has no missing parents
23/07/08 18:11:15 INFO MemoryStore: Block broadcast_1565 stored as values in memory
(estimated size 94.4 KiB, free 340.0 MiB)
23/07/08 18:11:15 INFO MemoryStore: Block broadcast_1565_piece0 stored as bytes in m
emory (estimated size 40.7 KiB, free 339.9 MiB)
23/07/08 18:11:15 INFO BlockManagerInfo: Added broadcast_1565_piece0 in memory on 17
2.30.115.138:43839 (size: 40.7 KiB, free: 363.7 MiB)
23/07/08 18:11:15 INFO SparkContext: Created broadcast 1565 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:15 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
2726 (MapPartitionsRDD[3954] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0))
23/07/08 18:11:15 INFO TaskSchedulerImpl: Adding task set 2726.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:15 INFO TaskSetManager: Starting task 0.0 in stage 2726.0 (TID 1749)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7634 bytes)
23/07/08 18:11:15 INFO Executor: Running task 0.0 in stage 2726.0 (TID 1749)
23/07/08 18:11:15 INFO ShuffleBlockFetcherIterator: Getting 1 (384.5 KiB) non-empty
blocks including 1 (384.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:11:15 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:11:15 INFO CodeGenerator: Code generated in 5.761281 ms
23/07/08 18:11:15 INFO ShuffleBlockFetcherIterator: Getting 2 (2.7 KiB) non-empty bl
ocks including 2 (2.7 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-
local and 0 (0.0 B) remote blocks
23/07/08 18:11:15 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:11:15 INFO CodeGenerator: Code generated in 6.346044 ms
23/07/08 18:11:15 INFO Executor: Finished task 0.0 in stage 2726.0 (TID 1749). 10354
bytes result sent to driver
23/07/08 18:11:15 INFO TaskSetManager: Finished task 0.0 in stage 2726.0 (TID 1749)
in 126 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:15 INFO TaskSchedulerImpl: Removed TaskSet 2726.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:15 INFO DAGScheduler: ShuffleMapStage 2726 (csv at <unknown>:0) finis
hed in 0.132 s
23/07/08 18:11:15 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:11:15 INFO DAGScheduler: running: Set()
23/07/08 18:11:15 INFO DAGScheduler: waiting: Set()
23/07/08 18:11:15 INFO DAGScheduler: failed: Set()
23/07/08 18:11:15 INFO ShufflePartitionsUtil: For shuffle(650), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:11:15 INFO FileOutputCommitter: File Output Committer Algorithm version
is 1
23/07/08 18:11:15 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _tempor
ary folders under output directory:false, ignore cleanup failures: false
23/07/08 18:11:15 INFO SQLHadoopMapReduceCommitProtocol: Using output committer clas
s org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
23/07/08 18:11:15 INFO HashAggregateExec: spark.sql.codegen.aggregate.map.twolevel.e
nabled is set to true, but current version of codegened fast hashmap does not suppor
t this aggregate.
23/07/08 18:11:15 INFO CodeGenerator: Code generated in 14.098957 ms
23/07/08 18:11:15 INFO SparkContext: Starting job: csv at <unknown>:0
23/07/08 18:11:15 INFO DAGScheduler: Got job 1279 (csv at <unknown>:0) with 1 output
```

```
partitions
23/07/08 18:11:15 INFO DAGScheduler: Final stage: ResultStage 2731 (csv at <unknown
>:0)
23/07/08 18:11:15 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
30)
23/07/08 18:11:15 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:15 INFO DAGScheduler: Submitting ResultStage 2731 (MapPartitionsRDD[3
957] at csv at <unknown>:0), which has no missing parents
23/07/08 18:11:15 INFO MemoryStore: Block broadcast_1566 stored as values in memory
(estimated size 389.7 KiB, free 339.5 MiB)
23/07/08 18:11:15 INFO MemoryStore: Block broadcast_1566_piece0 stored as bytes in m
emory (estimated size 145.0 KiB, free 339.4 MiB)
23/07/08 18:11:15 INFO BlockManagerInfo: Added broadcast_1566_piece0 in memory on 17
2.30.115.138:43839 (size: 145.0 KiB, free: 363.5 MiB)
23/07/08 18:11:15 INFO SparkContext: Created broadcast 1566 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:15 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 273
1 (MapPartitionsRDD[3957] at csv at <unknown>:0) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:11:15 INFO TaskSchedulerImpl: Adding task set 2731.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:15 INFO TaskSetManager: Starting task 0.0 in stage 2731.0 (TID 1750)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7363 bytes)
23/07/08 18:11:15 INFO Executor: Running task 0.0 in stage 2731.0 (TID 1750)
23/07/08 18:11:15 INFO ShuffleBlockFetcherIterator: Getting 1 (13.6 KiB) non-empty b
locks including 1 (13.6 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:11:15 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:11:15 INFO FileOutputCommitter: File Output Committer Algorithm version
is 1
23/07/08 18:11:15 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _tempor
ary folders under output directory:false, ignore cleanup failures: false
23/07/08 18:11:15 INFO SQLHadoopMapReduceCommitProtocol: Using output committer clas
s org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
23/07/08 18:11:16 INFO FileOutputCommitter: Saved output of task 'attempt_2023070818
1115161690748021256 2057_2731_m_000000_1750' to hdfs://127.0.0.1:9000/datos/salida/sa
lida1.csv/_temporary/0/task_202307081811151616907480212562057_2731_m_000000
23/07/08 18:11:16 INFO SparkHadoopMapRedUtil: attempt_202307081811151616907480212562
057_2731_m_000000_1750: Committed. Elapsed time: 9 ms.
23/07/08 18:11:16 INFO Executor: Finished task 0.0 in stage 2731.0 (TID 1750). 12652
bytes result sent to driver
23/07/08 18:11:16 INFO TaskSetManager: Finished task 0.0 in stage 2731.0 (TID 1750)
in 491 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:16 INFO TaskSchedulerImpl: Removed TaskSet 2731.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:16 INFO DAGScheduler: ResultStage 2731 (csv at <unknown>:0) finished
in 0.526 s
23/07/08 18:11:16 INFO DAGScheduler: Job 1279 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:16 INFO TaskSchedulerImpl: Killing all running tasks in stage 2731: S
tage finished
23/07/08 18:11:16 INFO DAGScheduler: Job 1279 finished: csv at <unknown>:0, took 0.5
28820 s
23/07/08 18:11:16 INFO FileFormatWriter: Start to commit write Job 13521809-b716-40a
f-8114-26cfdf0402dc.
23/07/08 18:11:16 INFO FileFormatWriter: Write Job 13521809-b716-40af-8114-26cfdf040
```

```
2dc committed. Elapsed time: 34 ms.
23/07/08 18:11:16 INFO FileFormatWriter: Finished processing stats for write job 135
21809-b716-40af-8114-26cfdf0402dc.
23/07/08 18:11:16 INFO DAGScheduler: Registering RDD 3959 (showString at <unknown>:
0) as input to shuffle 651
23/07/08 18:11:16 INFO DAGScheduler: Got map stage job 1280 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:11:16 INFO DAGScheduler: Final stage: ShuffleMapStage 2732 (showString a
t <unknown>:0)
23/07/08 18:11:16 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:16 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:16 INFO DAGScheduler: Submitting ShuffleMapStage 2732 (MapPartitionsR
DD[3959] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:16 INFO MemoryStore: Block broadcast_1567 stored as values in memory
(estimated size 22.5 KiB, free 339.4 MiB)
23/07/08 18:11:16 INFO MemoryStore: Block broadcast_1567_piece0 stored as bytes in m
emory (estimated size 10.8 KiB, free 339.4 MiB)
23/07/08 18:11:16 INFO BlockManagerInfo: Added broadcast_1567_piece0 in memory on 17
2.30.115.138:43839 (size: 10.8 KiB, free: 363.5 MiB)
23/07/08 18:11:16 INFO SparkContext: Created broadcast 1567 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:16 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2732 (MapPartitionsRDD[3959] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:11:16 INFO TaskSchedulerImpl: Adding task set 2732.0 with 2 tasks resour
ce profile 0
23/07/08 18:11:16 INFO TaskSetManager: Starting task 0.0 in stage 2732.0 (TID 1751)
(172.30.115.138, executor driver, partition 0, ANY, 7411 bytes)
23/07/08 18:11:16 INFO TaskSetManager: Starting task 1.0 in stage 2732.0 (TID 1752)
(172.30.115.138, executor driver, partition 1, ANY, 7411 bytes)
23/07/08 18:11:16 INFO Executor: Running task 1.0 in stage 2732.0 (TID 1752)
23/07/08 18:11:16 INFO Executor: Running task 0.0 in stage 2732.0 (TID 1751)
23/07/08 18:11:16 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:0+450275
23/07/08 18:11:16 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:450275+450275
23/07/08 18:11:16 INFO DAGScheduler: Registering RDD 3961 (showString at <unknown>:
0) as input to shuffle 652
23/07/08 18:11:16 INFO DAGScheduler: Got map stage job 1281 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:11:16 INFO DAGScheduler: Final stage: ShuffleMapStage 2733 (showString a
t <unknown>:0)
23/07/08 18:11:16 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:16 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:16 INFO DAGScheduler: Submitting ShuffleMapStage 2733 (MapPartitionsR
DD[3961] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:16 INFO MemoryStore: Block broadcast_1568 stored as values in memory
(estimated size 23.3 KiB, free 339.3 MiB)
23/07/08 18:11:16 INFO MemoryStore: Block broadcast_1568_piece0 stored as bytes in m
emory (estimated size 11.3 KiB, free 339.3 MiB)
23/07/08 18:11:16 INFO BlockManagerInfo: Added broadcast_1568_piece0 in memory on 17
2.30.115.138:43839 (size: 11.3 KiB, free: 363.5 MiB)
23/07/08 18:11:16 INFO SparkContext: Created broadcast 1568 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:16 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2733 (MapPartitionsRDD[3961] at showString at <unknown>:0) (first 15 tasks are for p
```

```
artitions Vector(0, 1))
23/07/08 18:11:16 INFO TaskSchedulerImpl: Adding task set 2733.0 with 2 tasks resour
ce profile 0
23/07/08 18:11:16 INFO DAGScheduler: Registering RDD 3963 (showString at <unknown>:
0) as input to shuffle 653
23/07/08 18:11:16 INFO DAGScheduler: Got map stage job 1282 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:11:16 INFO TaskSetManager: Starting task 0.0 in stage 2733.0 (TID 1753)
(172.30.115.138, executor driver, partition 0, ANY, 7406 bytes)
23/07/08 18:11:16 INFO DAGScheduler: Final stage: ShuffleMapStage 2734 (showString a
t <unknown>:0)
23/07/08 18:11:16 INFO TaskSetManager: Starting task 1.0 in stage 2733.0 (TID 1754)
(172.30.115.138, executor driver, partition 1, ANY, 7406 bytes)
23/07/08 18:11:16 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:11:16 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:16 INFO Executor: Running task 1.0 in stage 2733.0 (TID 1754)
23/07/08 18:11:16 INFO Executor: Running task 0.0 in stage 2733.0 (TID 1753)
23/07/08 18:11:16 INFO DAGScheduler: Submitting ShuffleMapStage 2734 (MapPartitionsR
DD[3963] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:16 INFO MemoryStore: Block broadcast_1569 stored as values in memory
(estimated size 21.1 KiB, free 339.3 MiB)
23/07/08 18:11:16 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:1310749+1310750
23/07/08 18:11:16 INFO MemoryStore: Block broadcast_1569_piece0 stored as bytes in m
emory (estimated size 10.6 KiB, free 339.3 MiB)
23/07/08 18:11:16 INFO BlockManagerInfo: Added broadcast_1569_piece0 in memory on 17
2.30.115.138:43839 (size: 10.6 KiB, free: 363.5 MiB)
23/07/08 18:11:16 INFO SparkContext: Created broadcast 1569 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:16 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2734 (MapPartitionsRDD[3963] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:11:16 INFO TaskSchedulerImpl: Adding task set 2734.0 with 2 tasks resour
ce profile 0
23/07/08 18:11:16 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:0+1310749
23/07/08 18:11:16 INFO TaskSetManager: Starting task 0.0 in stage 2734.0 (TID 1755)
(172.30.115.138, executor driver, partition 0, ANY, 7408 bytes)
23/07/08 18:11:16 INFO TaskSetManager: Starting task 1.0 in stage 2734.0 (TID 1756)
(172.30.115.138, executor driver, partition 1, ANY, 7408 bytes)
23/07/08 18:11:16 INFO Executor: Running task 0.0 in stage 2734.0 (TID 1755)
23/07/08 18:11:16 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Sucursal.
csv:0+1266
23/07/08 18:11:16 INFO Executor: Running task 1.0 in stage 2734.0 (TID 1756)
23/07/08 18:11:16 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Sucursal.
csv:1266+1266
23/07/08 18:11:16 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:16 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:16 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:11:16 INFO PythonRunner: Times: total = 188, boot = -1139, init = 1326,
finish = 1
23/07/08 18:11:16 INFO PythonRunner: Times: total = 173, boot = -1124, init = 1296,
finish = 1
23/07/08 18:11:16 INFO PythonRunner: Times: total = 203, boot = -1289, init = 1433,
finish = 59
23/07/08 18:11:16 INFO Executor: Finished task 0.0 in stage 2732.0 (TID 1751). 2524
```

```
bytes result sent to driver
23/07/08 18:11:16 INFO TaskSetManager: Finished task 0.0 in stage 2732.0 (TID 1751)
in 337 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:11:16 INFO Executor: Finished task 1.0 in stage 2734.0 (TID 1756). 2524
bytes result sent to driver
23/07/08 18:11:16 INFO TaskSetManager: Finished task 1.0 in stage 2734.0 (TID 1756)
in 313 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:11:16 INFO Executor: Finished task 0.0 in stage 2734.0 (TID 1755). 2524
bytes result sent to driver
23/07/08 18:11:16 INFO TaskSetManager: Finished task 0.0 in stage 2734.0 (TID 1755)
in 326 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:11:16 INFO TaskSchedulerImpl: Removed TaskSet 2734.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:16 INFO DAGScheduler: ShuffleMapStage 2734 (showString at <unknown>:
0) finished in 0.330 s
23/07/08 18:11:16 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:11:16 INFO DAGScheduler: running: Set(ShuffleMapStage 2732, ShuffleMapSt
age 2733)
23/07/08 18:11:16 INFO DAGScheduler: waiting: Set()
23/07/08 18:11:16 INFO DAGScheduler: failed: Set()
23/07/08 18:11:16 INFO BlockManagerInfo: Removed broadcast_1565_piece0 on 172.30.11
5.138:43839 in memory (size: 40.7 KiB, free: 363.5 MiB)
23/07/08 18:11:16 INFO PythonRunner: Times: total = 243, boot = -1179, init = 1352,
finish = 70
23/07/08 18:11:16 INFO Executor: Finished task 1.0 in stage 2732.0 (TID 1752). 2524
bytes result sent to driver
23/07/08 18:11:16 INFO TaskSetManager: Finished task 1.0 in stage 2732.0 (TID 1752)
in 374 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:11:16 INFO TaskSchedulerImpl: Removed TaskSet 2732.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:16 INFO DAGScheduler: ShuffleMapStage 2732 (showString at <unknown>:
0) finished in 0.378 s
23/07/08 18:11:16 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:11:16 INFO DAGScheduler: running: Set(ShuffleMapStage 2733)
23/07/08 18:11:16 INFO DAGScheduler: waiting: Set()
23/07/08 18:11:16 INFO DAGScheduler: failed: Set()
23/07/08 18:11:16 INFO ShufflePartitionsUtil: For shuffle(651), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:11:16 INFO SparkContext: Starting job: $anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266
23/07/08 18:11:16 INFO DAGScheduler: Got job 1283 ($anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266) with 1 output partitions
23/07/08 18:11:16 INFO DAGScheduler: Final stage: ResultStage 2736 ($anonfun$withThr
eadLocalCaptured$1 at FutureTask.java:266)
23/07/08 18:11:16 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
35)
23/07/08 18:11:16 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:16 INFO DAGScheduler: Submitting ResultStage 2736 (MapPartitionsRDD[3
965] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:266), which has no mis
sing parents
23/07/08 18:11:16 INFO MemoryStore: Block broadcast_1570 stored as values in memory
(estimated size 8.2 KiB, free 339.4 MiB)
23/07/08 18:11:16 INFO MemoryStore: Block broadcast_1570_piece0 stored as bytes in m
emory (estimated size 4.2 KiB, free 339.4 MiB)
23/07/08 18:11:16 INFO BlockManagerInfo: Added broadcast_1570_piece0 in memory on 17
2.30.115.138:43839 (size: 4.2 KiB, free: 363.5 MiB)
```

```
23/07/08 18:11:16 INFO SparkContext: Created broadcast 1570 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:16 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 273
6 (MapPartitionsRDD[3965] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:2
66) (first 15 tasks are for partitions Vector(0))
23/07/08 18:11:16 INFO TaskSchedulerImpl: Adding task set 2736.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:16 INFO TaskSetManager: Starting task 0.0 in stage 2736.0 (TID 1757)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7379 bytes)
23/07/08 18:11:16 INFO Executor: Running task 0.0 in stage 2736.0 (TID 1757)
23/07/08 18:11:16 INFO ShuffleBlockFetcherIterator: Getting 2 (234.6 KiB) non-empty
blocks including 2 (234.6 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:11:16 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:11:16 INFO PythonRunner: Times: total = 299, boot = -1272, init = 1419,
finish = 152
23/07/08 18:11:16 INFO Executor: Finished task 0.0 in stage 2736.0 (TID 1757). 17174
1 bytes result sent to driver
23/07/08 18:11:16 INFO TaskSetManager: Finished task 0.0 in stage 2736.0 (TID 1757)
in 11 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:16 INFO TaskSchedulerImpl: Removed TaskSet 2736.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:16 INFO DAGScheduler: ResultStage 2736 ($anonfun$withThreadLocalCaptu
red$1 at FutureTask.java:266) finished in 0.017 s
23/07/08 18:11:16 INFO DAGScheduler: Job 1283 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:16 INFO TaskSchedulerImpl: Killing all running tasks in stage 2736: S
tage finished
23/07/08 18:11:16 INFO DAGScheduler: Job 1283 finished: $anonfun$withThreadLocalCapt
ured$1 at FutureTask.java:266, took 0.018483 s
23/07/08 18:11:16 INFO MemoryStore: Block broadcast_1571 stored as values in memory
(estimated size 2.5 MiB, free 336.9 MiB)
23/07/08 18:11:16 INFO MemoryStore: Block broadcast_1571_piece0 stored as bytes in m
emory (estimated size 299.6 KiB, free 336.6 MiB)
23/07/08 18:11:16 INFO BlockManagerInfo: Added broadcast_1571_piece0 in memory on 17
2.30.115.138:43839 (size: 299.6 KiB, free: 363.2 MiB)
23/07/08 18:11:16 INFO SparkContext: Created broadcast 1571 from $anonfun$withThread
LocalCaptured$1 at FutureTask.java:266
23/07/08 18:11:16 INFO Executor: Finished task 1.0 in stage 2733.0 (TID 1754). 2524
bytes result sent to driver
23/07/08 18:11:16 INFO TaskSetManager: Finished task 1.0 in stage 2733.0 (TID 1754)
in 421 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:11:16 INFO BlockManagerInfo: Removed broadcast_1570_piece0 on 172.30.11
5.138:43839 in memory (size: 4.2 KiB, free: 363.2 MiB)
23/07/08 18:11:16 INFO PythonRunner: Times: total = 369, boot = -1241, init = 1420,
finish = 190
23/07/08 18:11:16 INFO Executor: Finished task 0.0 in stage 2733.0 (TID 1753). 2524
bytes result sent to driver
23/07/08 18:11:16 INFO TaskSetManager: Finished task 0.0 in stage 2733.0 (TID 1753)
in 490 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:11:16 INFO TaskSchedulerImpl: Removed TaskSet 2733.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:16 INFO DAGScheduler: ShuffleMapStage 2733 (showString at <unknown>:
0) finished in 0.494 s
23/07/08 18:11:16 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:11:16 INFO DAGScheduler: running: Set()
```

```
23/07/08 18:11:16 INFO DAGScheduler: waiting: Set()
23/07/08 18:11:16 INFO DAGScheduler: failed: Set()
23/07/08 18:11:16 INFO ShufflePartitionsUtil: For shuffle(652), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:11:16 INFO DAGScheduler: Registering RDD 3968 (showString at <unknown>:
0) as input to shuffle 654
23/07/08 18:11:16 INFO DAGScheduler: Got map stage job 1284 (showString at <unknown
>:0) with 1 output partitions
23/07/08 18:11:16 INFO DAGScheduler: Final stage: ShuffleMapStage 2738 (showString a
t <unknown>:0)
23/07/08 18:11:16 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
37)
23/07/08 18:11:16 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:16 INFO DAGScheduler: Submitting ShuffleMapStage 2738 (MapPartitionsR
DD[3968] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:16 INFO MemoryStore: Block broadcast_1572 stored as values in memory
(estimated size 15.3 KiB, free 336.6 MiB)
23/07/08 18:11:16 INFO MemoryStore: Block broadcast_1572_piece0 stored as bytes in m
emory (estimated size 7.4 KiB, free 336.6 MiB)
23/07/08 18:11:16 INFO BlockManagerInfo: Added broadcast_1572_piece0 in memory on 17
2.30.115.138:43839 (size: 7.4 KiB, free: 363.2 MiB)
23/07/08 18:11:16 INFO SparkContext: Created broadcast 1572 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:16 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
2738 (MapPartitionsRDD[3968] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0))
23/07/08 18:11:16 INFO TaskSchedulerImpl: Adding task set 2738.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:16 INFO TaskSetManager: Starting task 0.0 in stage 2738.0 (TID 1758)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7368 bytes)
23/07/08 18:11:16 INFO Executor: Running task 0.0 in stage 2738.0 (TID 1758)
23/07/08 18:11:16 INFO ShuffleBlockFetcherIterator: Getting 2 (479.7 KiB) non-empty
blocks including 2 (479.7 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:11:16 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:11:17 INFO Executor: Finished task 0.0 in stage 2738.0 (TID 1758). 4301
bytes result sent to driver
23/07/08 18:11:17 INFO TaskSetManager: Finished task 0.0 in stage 2738.0 (TID 1758)
in 48 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:17 INFO TaskSchedulerImpl: Removed TaskSet 2738.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:17 INFO DAGScheduler: ShuffleMapStage 2738 (showString at <unknown>:
0) finished in 0.052 s
23/07/08 18:11:17 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:11:17 INFO DAGScheduler: running: Set()
23/07/08 18:11:17 INFO DAGScheduler: waiting: Set()
23/07/08 18:11:17 INFO DAGScheduler: failed: Set()
23/07/08 18:11:17 INFO ShufflePartitionsUtil: For shuffle(654, 653), advisory target
size: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:11:17 INFO DAGScheduler: Registering RDD 3975 (showString at <unknown>:
0) as input to shuffle 655
23/07/08 18:11:17 INFO DAGScheduler: Got map stage job 1285 (showString at <unknown
>:0) with 1 output partitions
23/07/08 18:11:17 INFO DAGScheduler: Final stage: ShuffleMapStage 2742 (showString a
t <unknown>:0)
23/07/08 18:11:17 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
```

```
40, ShuffleMapStage 2741)
23/07/08 18:11:17 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:17 INFO DAGScheduler: Submitting ShuffleMapStage 2742 (MapPartitionsR
DD[3975] at showString at <unknown>:0), which has no missing parents
23/07/08 18:11:17 INFO MemoryStore: Block broadcast_1573 stored as values in memory
(estimated size 94.5 KiB, free 336.5 MiB)
23/07/08 18:11:17 INFO MemoryStore: Block broadcast_1573_piece0 stored as bytes in m
emory (estimated size 40.8 KiB, free 336.5 MiB)
23/07/08 18:11:17 INFO BlockManagerInfo: Added broadcast_1573_piece0 in memory on 17
2.30.115.138:43839 (size: 40.8 KiB, free: 363.2 MiB)
23/07/08 18:11:17 INFO SparkContext: Created broadcast 1573 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:17 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
2742 (MapPartitionsRDD[3975] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0))
23/07/08 18:11:17 INFO TaskSchedulerImpl: Adding task set 2742.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:17 INFO TaskSetManager: Starting task 0.0 in stage 2742.0 (TID 1759)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7634 bytes)
23/07/08 18:11:17 INFO Executor: Running task 0.0 in stage 2742.0 (TID 1759)
23/07/08 18:11:17 INFO ShuffleBlockFetcherIterator: Getting 1 (384.5 KiB) non-empty
blocks including 1 (384.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:11:17 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:11:17 INFO ShuffleBlockFetcherIterator: Getting 2 (2.7 KiB) non-empty bl
ocks including 2 (2.7 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-
local and 0 (0.0 B) remote blocks
23/07/08 18:11:17 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:11:17 INFO Executor: Finished task 0.0 in stage 2742.0 (TID 1759). 10354
bytes result sent to driver
23/07/08 18:11:17 INFO TaskSetManager: Finished task 0.0 in stage 2742.0 (TID 1759)
in 69 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:17 INFO TaskSchedulerImpl: Removed TaskSet 2742.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:17 INFO DAGScheduler: ShuffleMapStage 2742 (showString at <unknown>:
0) finished in 0.076 s
23/07/08 18:11:17 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:11:17 INFO DAGScheduler: running: Set()
23/07/08 18:11:17 INFO DAGScheduler: waiting: Set()
23/07/08 18:11:17 INFO DAGScheduler: failed: Set()
23/07/08 18:11:17 INFO ShufflePartitionsUtil: For shuffle(655), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:11:17 INFO HashAggregateExec: spark.sql.codegen.aggregate.map.twolevel.e
nabled is set to true, but current version of codegened fast hashmap does not suppor
t this aggregate.
23/07/08 18:11:17 INFO CodeGenerator: Code generated in 13.70464 ms
23/07/08 18:11:17 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:11:17 INFO DAGScheduler: Got job 1286 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:11:17 INFO DAGScheduler: Final stage: ResultStage 2747 (showString at <u
nknown>:0)
23/07/08 18:11:17 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
46)
23/07/08 18:11:17 INFO DAGScheduler: Missing parents: List()
23/07/08 18:11:17 INFO DAGScheduler: Submitting ResultStage 2747 (MapPartitionsRDD[3
978] at showString at <unknown>:0), which has no missing parents
```

```
23/07/08 18:11:17 INFO MemoryStore: Block broadcast_1574 stored as values in memory
(estimated size 88.0 KiB, free 336.4 MiB)
23/07/08 18:11:17 INFO MemoryStore: Block broadcast_1574_piece0 stored as bytes in m
emory (estimated size 34.3 KiB, free 336.4 MiB)
23/07/08 18:11:17 INFO BlockManagerInfo: Added broadcast_1574_piece0 in memory on 17
2.30.115.138:43839 (size: 34.3 KiB, free: 363.2 MiB)
23/07/08 18:11:17 INFO SparkContext: Created broadcast 1574 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:11:17 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 274
7 (MapPartitionsRDD[3978] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:11:17 INFO TaskSchedulerImpl: Adding task set 2747.0 with 1 tasks resour
ce profile 0
23/07/08 18:11:17 INFO TaskSetManager: Starting task 0.0 in stage 2747.0 (TID 1760)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7363 bytes)
23/07/08 18:11:17 INFO Executor: Running task 0.0 in stage 2747.0 (TID 1760)
23/07/08 18:11:17 INFO ShuffleBlockFetcherIterator: Getting 1 (13.6 KiB) non-empty b
locks including 1 (13.6 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:11:17 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
```

Sucursal	AÑO	VentasAnuales	Ventas_Anuales_format
La Plata	2019	585806.24	585,806.24
San Justo	2018	5554329.759999999	5,554,329.76
Mendoza1	2017	917432.68	917,432.68
MDQ1	2018	1866711.9200000004	1,866,711.92
Palermo 2	2016	1202901.1800000004	1,202,901.18
San Justo	2015	3730566.5200000005	3,730,566.52
Mendoza1	2015	534599.0400000002	534,599.04
MDQ2	2015	463976.27999999997	463,976.28
Avellaneda	2018	1589915.5999999999	1,589,915.60
Palermo 2	2018	1930545.0	1,930,545.00
Caballito	2015	1358026.6600000004	1,358,026.66
Caballito	2020	2152175.4999999995	2,152,175.50
Córdoba Quiroz	2015	2458103.1599999983	2,458,103.16
Mendoza1	2018	1612505.0999999999	1,612,505.10
Deposito	2016	1001208.2600000001	1,001,208.26
Quilmes	2015	118648.95	118,648.95
Vicente Lopez	2017	660427.7799999999	660,427.78
Velez	2015	5.724828531E7	57,248,285.31
Bariloche	2018	7.47928616E7	74,792,861.60
Palermo 2	2020	4165637.750000001	4,165,637.75

only showing top 20 rows

```
23/07/08 18:11:17 INFO Executor: Finished task 0.0 in stage 2747.0 (TID 1760). 12772
bytes result sent to driver
23/07/08 18:11:17 INFO TaskSetManager: Finished task 0.0 in stage 2747.0 (TID 1760)
in 12 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:11:17 INFO TaskSchedulerImpl: Removed TaskSet 2747.0, whose tasks have a
ll completed, from pool
23/07/08 18:11:17 INFO DAGScheduler: ResultStage 2747 (showString at <unknown>:0) fi
nished in 0.018 s
23/07/08 18:11:17 INFO DAGScheduler: Job 1286 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:11:17 INFO TaskSchedulerImpl: Killing all running tasks in stage 2747: S
tage finished
23/07/08 18:11:17 INFO DAGScheduler: Job 1286 finished: showString at <unknown>:0, t
ook 0.020526 s
```

b) Calcular las ventas semestrales de los productos más vendidos en el periodo 2019, crear un archivo de salida con el resultado

In [195…

```python
from pyspark.sql.functions import sum
from pyspark.sql.window import Window
from pyspark.sql.functions import rank

# Filtrar el periodo 2019
ventas_2019 = r6.filter(r6["AÑO"] == 2019)

# Calcular las ventas semestrales de los productos
ventas_semestrales_productos = ventas_2019.groupBy("IdProducto", "SEMESTRE") \
    .agg(sum("total_venta").alias("VentasSemestrales"))

# Calcular el ranking de las ventas semestrales por producto
windowSpec = Window.partitionBy("SEMESTRE").orderBy(ventas_semestrales_productos["V
ventas_semestrales_productos_ranked = ventas_semestrales_productos.withColumn("Top"

# Filtrar los productos más vendidos en cada semestre (SOLO LOS PRIMEROS 5)
productos_mas_vendidos = ventas_semestrales_productos_ranked.filter(ventas_semestra

##eSTO para generar una columna con buen formato y definir 2 decimales.
productos_mas_vendidos = productos_mas_vendidos.withColumn("Ventas_Semestrales_form
format_number(col("VentasSemestrales"), 2).alias("Ventas_Semestrales_format"))

#Guardando el Archivo
productos_mas_vendidos.write.csv("/datos/salida/salida2", header=True, mode="overwr

# Mostrar el resultado
productos_mas_vendidos.show()
```

```
23/07/08 18:12:15 INFO BlockManagerInfo: Removed broadcast_1588_piece0 on 172.30.11
5.138:43839 in memory (size: 145.5 KiB, free: 363.1 MiB)
23/07/08 18:12:15 INFO BlockManagerInfo: Removed broadcast_1587_piece0 on 172.30.11
5.138:43839 in memory (size: 30.0 KiB, free: 363.2 MiB)
23/07/08 18:12:15 INFO BlockManagerInfo: Removed broadcast_1586_piece0 on 172.30.11
5.138:43839 in memory (size: 31.3 KiB, free: 363.2 MiB)
23/07/08 18:12:15 INFO DAGScheduler: Registering RDD 4014 (csv at <unknown>:0) as in
put to shuffle 664
23/07/08 18:12:15 INFO DAGScheduler: Got map stage job 1299 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:12:15 INFO DAGScheduler: Final stage: ShuffleMapStage 2774 (csv at <unkn
own>:0)
23/07/08 18:12:15 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:12:15 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:15 INFO DAGScheduler: Submitting ShuffleMapStage 2774 (MapPartitionsR
DD[4014] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:15 INFO MemoryStore: Block broadcast_1589 stored as values in memory
(estimated size 23.2 KiB, free 332.4 MiB)
23/07/08 18:12:15 INFO MemoryStore: Block broadcast_1589_piece0 stored as bytes in m
emory (estimated size 11.2 KiB, free 332.4 MiB)
23/07/08 18:12:15 INFO BlockManagerInfo: Added broadcast_1589_piece0 in memory on 17
2.30.115.138:43839 (size: 11.2 KiB, free: 363.2 MiB)
23/07/08 18:12:15 INFO SparkContext: Created broadcast 1589 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:15 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2774 (MapPartitionsRDD[4014] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:12:15 INFO TaskSchedulerImpl: Adding task set 2774.0 with 2 tasks resour
ce profile 0
23/07/08 18:12:15 INFO TaskSetManager: Starting task 0.0 in stage 2774.0 (TID 1777)
(172.30.115.138, executor driver, partition 0, ANY, 7411 bytes)
23/07/08 18:12:15 INFO TaskSetManager: Starting task 1.0 in stage 2774.0 (TID 1778)
(172.30.115.138, executor driver, partition 1, ANY, 7411 bytes)
23/07/08 18:12:15 INFO Executor: Running task 0.0 in stage 2774.0 (TID 1777)
23/07/08 18:12:15 INFO Executor: Running task 1.0 in stage 2774.0 (TID 1778)
23/07/08 18:12:15 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:0+450275
23/07/08 18:12:15 INFO DAGScheduler: Registering RDD 4016 (csv at <unknown>:0) as in
put to shuffle 665
23/07/08 18:12:15 INFO DAGScheduler: Got map stage job 1300 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:12:15 INFO DAGScheduler: Final stage: ShuffleMapStage 2775 (csv at <unkn
own>:0)
23/07/08 18:12:15 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:12:15 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:15 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:450275+450275
23/07/08 18:12:15 INFO DAGScheduler: Submitting ShuffleMapStage 2775 (MapPartitionsR
DD[4016] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:15 INFO MemoryStore: Block broadcast_1590 stored as values in memory
(estimated size 23.2 KiB, free 332.4 MiB)
23/07/08 18:12:15 INFO MemoryStore: Block broadcast_1590_piece0 stored as bytes in m
emory (estimated size 11.3 KiB, free 332.4 MiB)
23/07/08 18:12:15 INFO BlockManagerInfo: Added broadcast_1590_piece0 in memory on 17
2.30.115.138:43839 (size: 11.3 KiB, free: 363.2 MiB)
23/07/08 18:12:15 INFO SparkContext: Created broadcast 1590 from broadcast at DAGSch
```

```
eduler.scala:1535
23/07/08 18:12:15 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2775 (MapPartitionsRDD[4016] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:12:15 INFO TaskSchedulerImpl: Adding task set 2775.0 with 2 tasks resour
ce profile 0
23/07/08 18:12:15 INFO TaskSetManager: Starting task 0.0 in stage 2775.0 (TID 1779)
(172.30.115.138, executor driver, partition 0, ANY, 7406 bytes)
23/07/08 18:12:15 INFO TaskSetManager: Starting task 1.0 in stage 2775.0 (TID 1780)
(172.30.115.138, executor driver, partition 1, ANY, 7406 bytes)
23/07/08 18:12:15 INFO Executor: Running task 1.0 in stage 2775.0 (TID 1780)
23/07/08 18:12:15 INFO Executor: Running task 0.0 in stage 2775.0 (TID 1779)
23/07/08 18:12:15 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:1310749+1310750
23/07/08 18:12:15 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:0+1310749
23/07/08 18:12:15 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:12:15 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:12:16 INFO PythonRunner: Times: total = 184, boot = -33504, init = 3363
2, finish = 56
23/07/08 18:12:16 INFO Executor: Finished task 1.0 in stage 2774.0 (TID 1778). 2352
bytes result sent to driver
23/07/08 18:12:16 INFO TaskSetManager: Finished task 1.0 in stage 2774.0 (TID 1778)
in 224 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:12:16 INFO PythonRunner: Times: total = 203, boot = -33505, init = 3364
8, finish = 60
23/07/08 18:12:16 INFO Executor: Finished task 0.0 in stage 2774.0 (TID 1777). 2524
bytes result sent to driver
23/07/08 18:12:16 INFO TaskSetManager: Finished task 0.0 in stage 2774.0 (TID 1777)
in 312 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:12:16 INFO TaskSchedulerImpl: Removed TaskSet 2774.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:16 INFO DAGScheduler: ShuffleMapStage 2774 (csv at <unknown>:0) finis
hed in 0.318 s
23/07/08 18:12:16 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:16 INFO DAGScheduler: running: Set(ShuffleMapStage 2775)
23/07/08 18:12:16 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:16 INFO DAGScheduler: failed: Set()
23/07/08 18:12:16 INFO ShufflePartitionsUtil: For shuffle(664), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:16 INFO SparkContext: Starting job: $anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266
23/07/08 18:12:16 INFO DAGScheduler: Got job 1301 ($anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266) with 1 output partitions
23/07/08 18:12:16 INFO DAGScheduler: Final stage: ResultStage 2777 ($anonfun$withThr
eadLocalCaptured$1 at FutureTask.java:266)
23/07/08 18:12:16 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
76)
23/07/08 18:12:16 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:16 INFO DAGScheduler: Submitting ResultStage 2777 (MapPartitionsRDD[4
018] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:266), which has no mis
sing parents
23/07/08 18:12:16 INFO MemoryStore: Block broadcast_1591 stored as values in memory
(estimated size 8.2 KiB, free 332.4 MiB)
23/07/08 18:12:16 INFO MemoryStore: Block broadcast_1591_piece0 stored as bytes in m
emory (estimated size 4.2 KiB, free 332.4 MiB)
```

```
23/07/08 18:12:16 INFO BlockManagerInfo: Added broadcast_1591_piece0 in memory on 17
2.30.115.138:43839 (size: 4.2 KiB, free: 363.2 MiB)
23/07/08 18:12:16 INFO SparkContext: Created broadcast 1591 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:16 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 277
7 (MapPartitionsRDD[4018] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:2
66) (first 15 tasks are for partitions Vector(0))
23/07/08 18:12:16 INFO TaskSchedulerImpl: Adding task set 2777.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:16 INFO TaskSetManager: Starting task 0.0 in stage 2777.0 (TID 1781)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7379 bytes)
23/07/08 18:12:16 INFO Executor: Running task 0.0 in stage 2777.0 (TID 1781)
23/07/08 18:12:16 INFO ShuffleBlockFetcherIterator: Getting 1 (18.0 KiB) non-empty b
locks including 1 (18.0 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:12:16 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:16 INFO Executor: Finished task 0.0 in stage 2777.0 (TID 1781). 6186
bytes result sent to driver
23/07/08 18:12:16 INFO TaskSetManager: Finished task 0.0 in stage 2777.0 (TID 1781)
in 8 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:16 INFO TaskSchedulerImpl: Removed TaskSet 2777.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:16 INFO DAGScheduler: ResultStage 2777 ($anonfun$withThreadLocalCaptu
red$1 at FutureTask.java:266) finished in 0.018 s
23/07/08 18:12:16 INFO DAGScheduler: Job 1301 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:12:16 INFO TaskSchedulerImpl: Killing all running tasks in stage 2777: S
tage finished
23/07/08 18:12:16 INFO DAGScheduler: Job 1301 finished: $anonfun$withThreadLocalCapt
ured$1 at FutureTask.java:266, took 0.020179 s
23/07/08 18:12:16 INFO MemoryStore: Block broadcast_1592 stored as values in memory
(estimated size 2.0 MiB, free 330.4 MiB)
23/07/08 18:12:16 INFO MemoryStore: Block broadcast_1592_piece0 stored as bytes in m
emory (estimated size 5.7 KiB, free 330.4 MiB)
23/07/08 18:12:16 INFO BlockManagerInfo: Added broadcast_1592_piece0 in memory on 17
2.30.115.138:43839 (size: 5.7 KiB, free: 363.1 MiB)
23/07/08 18:12:16 INFO SparkContext: Created broadcast 1592 from $anonfun$withThread
LocalCaptured$1 at FutureTask.java:266
23/07/08 18:12:16 INFO PythonRunner: Times: total = 302, boot = -32838, init = 3297
6, finish = 164
23/07/08 18:12:16 INFO PythonRunner: Times: total = 330, boot = -32832, init = 3297
3, finish = 189
23/07/08 18:12:16 INFO Executor: Finished task 1.0 in stage 2775.0 (TID 1780). 2524
bytes result sent to driver
23/07/08 18:12:16 INFO TaskSetManager: Finished task 1.0 in stage 2775.0 (TID 1780)
in 383 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:12:16 INFO Executor: Finished task 0.0 in stage 2775.0 (TID 1779). 2524
bytes result sent to driver
23/07/08 18:12:16 INFO TaskSetManager: Finished task 0.0 in stage 2775.0 (TID 1779)
in 399 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:12:16 INFO TaskSchedulerImpl: Removed TaskSet 2775.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:16 INFO DAGScheduler: ShuffleMapStage 2775 (csv at <unknown>:0) finis
hed in 0.410 s
23/07/08 18:12:16 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:16 INFO DAGScheduler: running: Set()
```

```
23/07/08 18:12:16 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:16 INFO DAGScheduler: failed: Set()
23/07/08 18:12:16 INFO ShufflePartitionsUtil: For shuffle(665), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:16 INFO DAGScheduler: Registering RDD 4021 (csv at <unknown>:0) as in
put to shuffle 666
23/07/08 18:12:16 INFO DAGScheduler: Got map stage job 1302 (csv at <unknown>:0) wit
h 1 output partitions
23/07/08 18:12:16 INFO DAGScheduler: Final stage: ShuffleMapStage 2779 (csv at <unkn
own>:0)
23/07/08 18:12:16 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
78)
23/07/08 18:12:16 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:16 INFO DAGScheduler: Submitting ShuffleMapStage 2779 (MapPartitionsR
DD[4021] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:16 INFO MemoryStore: Block broadcast_1593 stored as values in memory
(estimated size 72.4 KiB, free 330.3 MiB)
23/07/08 18:12:16 INFO MemoryStore: Block broadcast_1593_piece0 stored as bytes in m
emory (estimated size 31.0 KiB, free 330.3 MiB)
23/07/08 18:12:16 INFO BlockManagerInfo: Added broadcast_1593_piece0 in memory on 17
2.30.115.138:43839 (size: 31.0 KiB, free: 363.1 MiB)
23/07/08 18:12:16 INFO SparkContext: Created broadcast 1593 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:16 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
2779 (MapPartitionsRDD[4021] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0))
23/07/08 18:12:16 INFO TaskSchedulerImpl: Adding task set 2779.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:16 INFO TaskSetManager: Starting task 0.0 in stage 2779.0 (TID 1782)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7368 bytes)
23/07/08 18:12:16 INFO Executor: Running task 0.0 in stage 2779.0 (TID 1782)
23/07/08 18:12:16 INFO ShuffleBlockFetcherIterator: Getting 2 (518.3 KiB) non-empty
blocks including 2 (518.3 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:12:16 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:16 INFO BlockManagerInfo: Removed broadcast_1591_piece0 on 172.30.11
5.138:43839 in memory (size: 4.2 KiB, free: 363.1 MiB)
23/07/08 18:12:16 INFO Executor: Finished task 0.0 in stage 2779.0 (TID 1782). 7136
bytes result sent to driver
23/07/08 18:12:16 INFO TaskSetManager: Finished task 0.0 in stage 2779.0 (TID 1782)
in 63 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:16 INFO TaskSchedulerImpl: Removed TaskSet 2779.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:16 INFO DAGScheduler: ShuffleMapStage 2779 (csv at <unknown>:0) finis
hed in 0.069 s
23/07/08 18:12:16 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:16 INFO DAGScheduler: running: Set()
23/07/08 18:12:16 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:16 INFO DAGScheduler: failed: Set()
23/07/08 18:12:16 INFO ShufflePartitionsUtil: For shuffle(666), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:16 INFO HashAggregateExec: spark.sql.codegen.aggregate.map.twolevel.e
nabled is set to true, but current version of codegened fast hashmap does not suppor
t this aggregate.
23/07/08 18:12:16 INFO DAGScheduler: Registering RDD 4024 (csv at <unknown>:0) as in
put to shuffle 667
```

```
23/07/08 18:12:16 INFO DAGScheduler: Got map stage job 1303 (csv at <unknown>:0) wit
h 1 output partitions
23/07/08 18:12:16 INFO DAGScheduler: Final stage: ShuffleMapStage 2782 (csv at <unkn
own>:0)
23/07/08 18:12:16 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
81)
23/07/08 18:12:16 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:16 INFO DAGScheduler: Submitting ShuffleMapStage 2782 (MapPartitionsR
DD[4024] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:16 INFO MemoryStore: Block broadcast_1594 stored as values in memory
(estimated size 71.0 KiB, free 330.2 MiB)
23/07/08 18:12:16 INFO MemoryStore: Block broadcast_1594_piece0 stored as bytes in m
emory (estimated size 30.0 KiB, free 330.2 MiB)
23/07/08 18:12:16 INFO BlockManagerInfo: Added broadcast_1594_piece0 in memory on 17
2.30.115.138:43839 (size: 30.0 KiB, free: 363.1 MiB)
23/07/08 18:12:16 INFO SparkContext: Created broadcast 1594 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:16 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
2782 (MapPartitionsRDD[4024] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0))
23/07/08 18:12:16 INFO TaskSchedulerImpl: Adding task set 2782.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:16 INFO TaskSetManager: Starting task 0.0 in stage 2782.0 (TID 1783)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7352 bytes)
23/07/08 18:12:16 INFO Executor: Running task 0.0 in stage 2782.0 (TID 1783)
23/07/08 18:12:16 INFO ShuffleBlockFetcherIterator: Getting 1 (24.5 KiB) non-empty b
locks including 1 (24.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:12:16 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:16 INFO Executor: Finished task 0.0 in stage 2782.0 (TID 1783). 8717
bytes result sent to driver
23/07/08 18:12:16 INFO TaskSetManager: Finished task 0.0 in stage 2782.0 (TID 1783)
in 17 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:16 INFO TaskSchedulerImpl: Removed TaskSet 2782.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:16 INFO DAGScheduler: ShuffleMapStage 2782 (csv at <unknown>:0) finis
hed in 0.023 s
23/07/08 18:12:16 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:16 INFO DAGScheduler: running: Set()
23/07/08 18:12:16 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:16 INFO DAGScheduler: failed: Set()
23/07/08 18:12:16 INFO ShufflePartitionsUtil: For shuffle(667), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:16 INFO FileOutputCommitter: File Output Committer Algorithm version
is 1
23/07/08 18:12:16 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _tempor
ary folders under output directory:false, ignore cleanup failures: false
23/07/08 18:12:16 INFO SQLHadoopMapReduceCommitProtocol: Using output committer clas
s org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
23/07/08 18:12:16 INFO SparkContext: Starting job: csv at <unknown>:0
23/07/08 18:12:16 INFO DAGScheduler: Got job 1304 (csv at <unknown>:0) with 1 output
partitions
23/07/08 18:12:16 INFO DAGScheduler: Final stage: ResultStage 2786 (csv at <unknown
>:0)
23/07/08 18:12:16 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
85)
```

```
23/07/08 18:12:16 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:16 INFO DAGScheduler: Submitting ResultStage 2786 (MapPartitionsRDD[4
029] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:16 INFO MemoryStore: Block broadcast_1595 stored as values in memory
(estimated size 379.7 KiB, free 329.8 MiB)
23/07/08 18:12:16 INFO MemoryStore: Block broadcast_1595_piece0 stored as bytes in m
emory (estimated size 145.5 KiB, free 329.7 MiB)
23/07/08 18:12:16 INFO BlockManagerInfo: Added broadcast_1595_piece0 in memory on 17
2.30.115.138:43839 (size: 145.5 KiB, free: 363.0 MiB)
23/07/08 18:12:16 INFO SparkContext: Created broadcast 1595 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:16 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 278
6 (MapPartitionsRDD[4029] at csv at <unknown>:0) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:12:16 INFO TaskSchedulerImpl: Adding task set 2786.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:16 INFO TaskSetManager: Starting task 0.0 in stage 2786.0 (TID 1784)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7363 bytes)
23/07/08 18:12:16 INFO Executor: Running task 0.0 in stage 2786.0 (TID 1784)
23/07/08 18:12:16 INFO ShuffleBlockFetcherIterator: Getting 1 (7.1 KiB) non-empty bl
ocks including 1 (7.1 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-
local and 0 (0.0 B) remote blocks
23/07/08 18:12:16 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:16 INFO FileOutputCommitter: File Output Committer Algorithm version
is 1
23/07/08 18:12:16 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _tempor
ary folders under output directory:false, ignore cleanup failures: false
23/07/08 18:12:16 INFO SQLHadoopMapReduceCommitProtocol: Using output committer clas
s org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
23/07/08 18:12:16 INFO FileOutputCommitter: Saved output of task 'attempt_2023070818
121691266670917920906 89_2786_m_000000_1784' to hdfs://127.0.0.1:9000/datos/salida/sa
lida2/_temporary/0/task_20230708181216912666709179209 0689_2786_m_000000
23/07/08 18:12:16 INFO SparkHadoopMapRedUtil: attempt_20230708181216912666709179209
0689_2786_m_000000_1784: Committed. Elapsed time: 7 ms.
23/07/08 18:12:16 INFO Executor: Finished task 0.0 in stage 2786.0 (TID 1784). 10985
bytes result sent to driver
23/07/08 18:12:16 INFO TaskSetManager: Finished task 0.0 in stage 2786.0 (TID 1784)
in 481 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:16 INFO TaskSchedulerImpl: Removed TaskSet 2786.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:16 INFO DAGScheduler: ResultStage 2786 (csv at <unknown>:0) finished
in 0.508 s
23/07/08 18:12:16 INFO DAGScheduler: Job 1304 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:12:16 INFO TaskSchedulerImpl: Killing all running tasks in stage 2786: S
tage finished
23/07/08 18:12:16 INFO DAGScheduler: Job 1304 finished: csv at <unknown>:0, took 0.5
09534 s
23/07/08 18:12:16 INFO FileFormatWriter: Start to commit write Job b75ab42a-5a88-473
c-9461-c16c281040d4.
23/07/08 18:12:17 INFO FileFormatWriter: Write Job b75ab42a-5a88-473c-9461-c16c28104
0d4 committed. Elapsed time: 53 ms.
23/07/08 18:12:17 INFO FileFormatWriter: Finished processing stats for write job b75
ab42a-5a88-473c-9461-c16c281040d4.
23/07/08 18:12:17 INFO DAGScheduler: Registering RDD 4031 (showString at <unknown>:
0) as input to shuffle 668
```

```
23/07/08 18:12:17 INFO DAGScheduler: Got map stage job 1305 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:12:17 INFO DAGScheduler: Final stage: ShuffleMapStage 2787 (showString a
t <unknown>:0)
23/07/08 18:12:17 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:12:17 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:17 INFO DAGScheduler: Submitting ShuffleMapStage 2787 (MapPartitionsR
DD[4031] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:17 INFO MemoryStore: Block broadcast_1596 stored as values in memory
(estimated size 23.2 KiB, free 329.6 MiB)
23/07/08 18:12:17 INFO MemoryStore: Block broadcast_1596_piece0 stored as bytes in m
emory (estimated size 11.2 KiB, free 329.6 MiB)
23/07/08 18:12:17 INFO BlockManagerInfo: Added broadcast_1596_piece0 in memory on 17
2.30.115.138:43839 (size: 11.2 KiB, free: 362.9 MiB)
23/07/08 18:12:17 INFO SparkContext: Created broadcast 1596 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:17 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2787 (MapPartitionsRDD[4031] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:12:17 INFO TaskSchedulerImpl: Adding task set 2787.0 with 2 tasks resour
ce profile 0
23/07/08 18:12:17 INFO DAGScheduler: Registering RDD 4033 (showString at <unknown>:
0) as input to shuffle 669
23/07/08 18:12:17 INFO DAGScheduler: Got map stage job 1306 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:12:17 INFO DAGScheduler: Final stage: ShuffleMapStage 2788 (showString a
t <unknown>:0)
23/07/08 18:12:17 INFO TaskSetManager: Starting task 0.0 in stage 2787.0 (TID 1785)
(172.30.115.138, executor driver, partition 0, ANY, 7411 bytes)
23/07/08 18:12:17 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:12:17 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:17 INFO TaskSetManager: Starting task 1.0 in stage 2787.0 (TID 1786)
(172.30.115.138, executor driver, partition 1, ANY, 7411 bytes)
23/07/08 18:12:17 INFO DAGScheduler: Submitting ShuffleMapStage 2788 (MapPartitionsR
DD[4033] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:17 INFO Executor: Running task 0.0 in stage 2787.0 (TID 1785)
23/07/08 18:12:17 INFO Executor: Running task 1.0 in stage 2787.0 (TID 1786)
23/07/08 18:12:17 INFO MemoryStore: Block broadcast_1597 stored as values in memory
(estimated size 23.2 KiB, free 329.6 MiB)
23/07/08 18:12:17 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:450275+450275
23/07/08 18:12:17 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:0+450275
23/07/08 18:12:17 INFO MemoryStore: Block broadcast_1597_piece0 stored as bytes in m
emory (estimated size 11.3 KiB, free 329.6 MiB)
23/07/08 18:12:17 INFO BlockManagerInfo: Added broadcast_1597_piece0 in memory on 17
2.30.115.138:43839 (size: 11.3 KiB, free: 362.9 MiB)
23/07/08 18:12:17 INFO SparkContext: Created broadcast 1597 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:17 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2788 (MapPartitionsRDD[4033] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:12:17 INFO TaskSchedulerImpl: Adding task set 2788.0 with 2 tasks resour
ce profile 0
23/07/08 18:12:17 INFO TaskSetManager: Starting task 0.0 in stage 2788.0 (TID 1787)
(172.30.115.138, executor driver, partition 0, ANY, 7406 bytes)
```

```
23/07/08 18:12:17 INFO TaskSetManager: Starting task 1.0 in stage 2788.0 (TID 1788)
(172.30.115.138, executor driver, partition 1, ANY, 7406 bytes)
23/07/08 18:12:17 INFO Executor: Running task 1.0 in stage 2788.0 (TID 1788)
23/07/08 18:12:17 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:1310749+1310750
23/07/08 18:12:17 INFO Executor: Running task 0.0 in stage 2788.0 (TID 1787)
23/07/08 18:12:17 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:0+1310749
23/07/08 18:12:17 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:12:17 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:12:17 INFO BlockManagerInfo: Removed broadcast_1594_piece0 on 172.30.11
5.138:43839 in memory (size: 30.0 KiB, free: 363.0 MiB)
23/07/08 18:12:17 INFO PythonRunner: Times: total = 175, boot = -33891, init = 3401
5, finish = 51
23/07/08 18:12:17 INFO PythonRunner: Times: total = 181, boot = -33916, init = 3404
1, finish = 56
23/07/08 18:12:17 INFO Executor: Finished task 1.0 in stage 2787.0 (TID 1786). 2395
bytes result sent to driver
23/07/08 18:12:17 INFO TaskSetManager: Finished task 1.0 in stage 2787.0 (TID 1786)
in 235 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:12:17 INFO BlockManagerInfo: Removed broadcast_1595_piece0 on 172.30.11
5.138:43839 in memory (size: 145.5 KiB, free: 363.1 MiB)
23/07/08 18:12:17 INFO Executor: Finished task 0.0 in stage 2787.0 (TID 1785). 2524
bytes result sent to driver
23/07/08 18:12:17 INFO TaskSetManager: Finished task 0.0 in stage 2787.0 (TID 1785)
in 248 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:12:17 INFO TaskSchedulerImpl: Removed TaskSet 2787.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:17 INFO DAGScheduler: ShuffleMapStage 2787 (showString at <unknown>:
0) finished in 0.253 s
23/07/08 18:12:17 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:17 INFO DAGScheduler: running: Set(ShuffleMapStage 2788)
23/07/08 18:12:17 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:17 INFO DAGScheduler: failed: Set()
23/07/08 18:12:17 INFO ShufflePartitionsUtil: For shuffle(668), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:17 INFO SparkContext: Starting job: $anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266
23/07/08 18:12:17 INFO DAGScheduler: Got job 1307 ($anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266) with 1 output partitions
23/07/08 18:12:17 INFO DAGScheduler: Final stage: ResultStage 2790 ($anonfun$withThr
eadLocalCaptured$1 at FutureTask.java:266)
23/07/08 18:12:17 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
89)
23/07/08 18:12:17 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:17 INFO DAGScheduler: Submitting ResultStage 2790 (MapPartitionsRDD[4
035] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:266), which has no mis
sing parents
23/07/08 18:12:17 INFO MemoryStore: Block broadcast_1598 stored as values in memory
(estimated size 8.2 KiB, free 330.2 MiB)
23/07/08 18:12:17 INFO MemoryStore: Block broadcast_1598_piece0 stored as bytes in m
emory (estimated size 4.2 KiB, free 330.2 MiB)
23/07/08 18:12:17 INFO BlockManagerInfo: Added broadcast_1598_piece0 in memory on 17
2.30.115.138:43839 (size: 4.2 KiB, free: 363.1 MiB)
23/07/08 18:12:17 INFO SparkContext: Created broadcast 1598 from broadcast at DAGSch
eduler.scala:1535
```

```
23/07/08 18:12:17 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 279
0 (MapPartitionsRDD[4035] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:2
66) (first 15 tasks are for partitions Vector(0))
23/07/08 18:12:17 INFO TaskSchedulerImpl: Adding task set 2790.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:17 INFO TaskSetManager: Starting task 0.0 in stage 2790.0 (TID 1789)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7379 bytes)
23/07/08 18:12:17 INFO Executor: Running task 0.0 in stage 2790.0 (TID 1789)
23/07/08 18:12:17 INFO ShuffleBlockFetcherIterator: Getting 1 (18.0 KiB) non-empty b
locks including 1 (18.0 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:12:17 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:17 INFO Executor: Finished task 0.0 in stage 2790.0 (TID 1789). 6186
bytes result sent to driver
23/07/08 18:12:17 INFO TaskSetManager: Finished task 0.0 in stage 2790.0 (TID 1789)
in 34 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:17 INFO TaskSchedulerImpl: Removed TaskSet 2790.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:17 INFO DAGScheduler: ResultStage 2790 ($anonfun$withThreadLocalCaptu
red$1 at FutureTask.java:266) finished in 0.039 s
23/07/08 18:12:17 INFO DAGScheduler: Job 1307 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:12:17 INFO TaskSchedulerImpl: Killing all running tasks in stage 2790: S
tage finished
23/07/08 18:12:17 INFO DAGScheduler: Job 1307 finished: $anonfun$withThreadLocalCapt
ured$1 at FutureTask.java:266, took 0.040852 s
23/07/08 18:12:17 INFO MemoryStore: Block broadcast_1599 stored as values in memory
(estimated size 2.0 MiB, free 328.2 MiB)
23/07/08 18:12:17 INFO MemoryStore: Block broadcast_1599_piece0 stored as bytes in m
emory (estimated size 5.7 KiB, free 328.2 MiB)
23/07/08 18:12:17 INFO BlockManagerInfo: Added broadcast_1599_piece0 in memory on 17
2.30.115.138:43839 (size: 5.7 KiB, free: 363.1 MiB)
23/07/08 18:12:17 INFO SparkContext: Created broadcast 1599 from $anonfun$withThread
LocalCaptured$1 at FutureTask.java:266
23/07/08 18:12:17 INFO PythonRunner: Times: total = 270, boot = -948, init = 1076, f
inish = 142
23/07/08 18:12:17 INFO PythonRunner: Times: total = 264, boot = -925, init = 1055, f
inish = 134
23/07/08 18:12:17 INFO Executor: Finished task 0.0 in stage 2788.0 (TID 1787). 2524
bytes result sent to driver
23/07/08 18:12:17 INFO TaskSetManager: Finished task 0.0 in stage 2788.0 (TID 1787)
in 356 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:12:17 INFO Executor: Finished task 1.0 in stage 2788.0 (TID 1788). 2524
bytes result sent to driver
23/07/08 18:12:17 INFO TaskSetManager: Finished task 1.0 in stage 2788.0 (TID 1788)
in 358 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:12:17 INFO TaskSchedulerImpl: Removed TaskSet 2788.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:17 INFO DAGScheduler: ShuffleMapStage 2788 (showString at <unknown>:
0) finished in 0.365 s
23/07/08 18:12:17 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:17 INFO DAGScheduler: running: Set()
23/07/08 18:12:17 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:17 INFO DAGScheduler: failed: Set()
23/07/08 18:12:17 INFO ShufflePartitionsUtil: For shuffle(669), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
```

```
23/07/08 18:12:17 INFO DAGScheduler: Registering RDD 4038 (showString at <unknown>:
0) as input to shuffle 670
23/07/08 18:12:17 INFO DAGScheduler: Got map stage job 1308 (showString at <unknown
>:0) with 1 output partitions
23/07/08 18:12:17 INFO DAGScheduler: Final stage: ShuffleMapStage 2792 (showString a
t <unknown>:0)
23/07/08 18:12:17 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
91)
23/07/08 18:12:17 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:17 INFO DAGScheduler: Submitting ShuffleMapStage 2792 (MapPartitionsR
DD[4038] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:17 INFO MemoryStore: Block broadcast_1600 stored as values in memory
(estimated size 72.6 KiB, free 328.1 MiB)
23/07/08 18:12:17 INFO MemoryStore: Block broadcast_1600_piece0 stored as bytes in m
emory (estimated size 31.2 KiB, free 328.1 MiB)
23/07/08 18:12:17 INFO BlockManagerInfo: Added broadcast_1600_piece0 in memory on 17
2.30.115.138:43839 (size: 31.2 KiB, free: 363.1 MiB)
23/07/08 18:12:17 INFO SparkContext: Created broadcast 1600 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:17 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
2792 (MapPartitionsRDD[4038] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0))
23/07/08 18:12:17 INFO TaskSchedulerImpl: Adding task set 2792.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:17 INFO TaskSetManager: Starting task 0.0 in stage 2792.0 (TID 1790)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7368 bytes)
23/07/08 18:12:17 INFO Executor: Running task 0.0 in stage 2792.0 (TID 1790)
23/07/08 18:12:17 INFO ShuffleBlockFetcherIterator: Getting 2 (518.3 KiB) non-empty
blocks including 2 (518.3 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:12:17 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
```

IdProducto	SEMESTRE	VentasSemestrales	Top	Ventas_Semestrales_format
42766	1	2306822.98	1	2,306,822.98
42967	1	1046515.0	2	1,046,515.00
42957	1	889691.0	3	889,691.00
42963	1	805410.0	4	805,410.00
42934	1	769270.0	5	769,270.00
42779	2	3449369.4400000013	1	3,449,369.44
42969	2	1286065.0	2	1,286,065.00
42951	2	1065050.0	3	1,065,050.00
42846	2	1033308.5399999998	4	1,033,308.54
42971	2	1016416.0	5	1,016,416.00

```
23/07/08 18:12:17 INFO Executor: Finished task 0.0 in stage 2792.0 (TID 1790). 7093
bytes result sent to driver
23/07/08 18:12:17 INFO TaskSetManager: Finished task 0.0 in stage 2792.0 (TID 1790)
in 53 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:17 INFO TaskSchedulerImpl: Removed TaskSet 2792.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:17 INFO DAGScheduler: ShuffleMapStage 2792 (showString at <unknown>:
0) finished in 0.059 s
23/07/08 18:12:17 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:17 INFO DAGScheduler: running: Set()
23/07/08 18:12:17 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:17 INFO DAGScheduler: failed: Set()
23/07/08 18:12:17 INFO ShufflePartitionsUtil: For shuffle(670), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:17 INFO HashAggregateExec: spark.sql.codegen.aggregate.map.twolevel.e
nabled is set to true, but current version of codegened fast hashmap does not suppor
t this aggregate.
23/07/08 18:12:17 INFO DAGScheduler: Registering RDD 4041 (showString at <unknown>:
0) as input to shuffle 671
23/07/08 18:12:17 INFO DAGScheduler: Got map stage job 1309 (showString at <unknown
>:0) with 1 output partitions
23/07/08 18:12:17 INFO DAGScheduler: Final stage: ShuffleMapStage 2795 (showString a
t <unknown>:0)
23/07/08 18:12:17 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
94)
23/07/08 18:12:17 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:17 INFO DAGScheduler: Submitting ShuffleMapStage 2795 (MapPartitionsR
DD[4041] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:17 INFO MemoryStore: Block broadcast_1601 stored as values in memory
(estimated size 71.0 KiB, free 328.0 MiB)
23/07/08 18:12:17 INFO MemoryStore: Block broadcast_1601_piece0 stored as bytes in m
emory (estimated size 30.0 KiB, free 328.0 MiB)
23/07/08 18:12:17 INFO BlockManagerInfo: Added broadcast_1601_piece0 in memory on 17
2.30.115.138:43839 (size: 30.0 KiB, free: 363.0 MiB)
23/07/08 18:12:17 INFO SparkContext: Created broadcast 1601 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:17 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
2795 (MapPartitionsRDD[4041] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0))
23/07/08 18:12:17 INFO TaskSchedulerImpl: Adding task set 2795.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:17 INFO TaskSetManager: Starting task 0.0 in stage 2795.0 (TID 1791)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7352 bytes)
23/07/08 18:12:17 INFO Executor: Running task 0.0 in stage 2795.0 (TID 1791)
23/07/08 18:12:17 INFO ShuffleBlockFetcherIterator: Getting 1 (24.5 KiB) non-empty b
locks including 1 (24.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:12:17 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:17 INFO Executor: Finished task 0.0 in stage 2795.0 (TID 1791). 8717
bytes result sent to driver
23/07/08 18:12:17 INFO TaskSetManager: Finished task 0.0 in stage 2795.0 (TID 1791)
in 19 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:17 INFO TaskSchedulerImpl: Removed TaskSet 2795.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:17 INFO DAGScheduler: ShuffleMapStage 2795 (showString at <unknown>:
0) finished in 0.025 s
```

```
23/07/08 18:12:17 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:17 INFO DAGScheduler: running: Set()
23/07/08 18:12:17 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:17 INFO DAGScheduler: failed: Set()
23/07/08 18:12:17 INFO ShufflePartitionsUtil: For shuffle(671), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:17 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:12:17 INFO DAGScheduler: Got job 1310 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:12:17 INFO DAGScheduler: Final stage: ResultStage 2799 (showString at <u
nknown>:0)
23/07/08 18:12:17 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 27
98)
23/07/08 18:12:17 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:17 INFO DAGScheduler: Submitting ResultStage 2799 (MapPartitionsRDD[4
046] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:17 INFO MemoryStore: Block broadcast_1602 stored as values in memory
(estimated size 77.9 KiB, free 327.9 MiB)
23/07/08 18:12:17 INFO MemoryStore: Block broadcast_1602_piece0 stored as bytes in m
emory (estimated size 34.5 KiB, free 327.9 MiB)
23/07/08 18:12:17 INFO BlockManagerInfo: Added broadcast_1602_piece0 in memory on 17
2.30.115.138:43839 (size: 34.5 KiB, free: 363.0 MiB)
23/07/08 18:12:17 INFO SparkContext: Created broadcast 1602 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:17 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 279
9 (MapPartitionsRDD[4046] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:12:17 INFO TaskSchedulerImpl: Adding task set 2799.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:17 INFO TaskSetManager: Starting task 0.0 in stage 2799.0 (TID 1792)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7363 bytes)
23/07/08 18:12:17 INFO Executor: Running task 0.0 in stage 2799.0 (TID 1792)
23/07/08 18:12:17 INFO ShuffleBlockFetcherIterator: Getting 1 (7.1 KiB) non-empty bl
ocks including 1 (7.1 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-
local and 0 (0.0 B) remote blocks
23/07/08 18:12:17 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:17 INFO Executor: Finished task 0.0 in stage 2799.0 (TID 1792). 10411
bytes result sent to driver
23/07/08 18:12:17 INFO TaskSetManager: Finished task 0.0 in stage 2799.0 (TID 1792)
in 13 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:17 INFO TaskSchedulerImpl: Removed TaskSet 2799.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:17 INFO DAGScheduler: ResultStage 2799 (showString at <unknown>:0) fi
nished in 0.020 s
23/07/08 18:12:17 INFO DAGScheduler: Job 1310 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:12:17 INFO TaskSchedulerImpl: Killing all running tasks in stage 2799: S
tage finished
23/07/08 18:12:17 INFO DAGScheduler: Job 1310 finished: showString at <unknown>:0, t
ook 0.021212 s
```

c) Encontrar el Top 10 ordenado de los mejores vendedores, crear un archivo de salida con el resultado

In [196…

```
from pyspark.sql.functions import desc, format_number, col
# Calcular la suma sin formatear
ventas_por_vendedor = r4.groupBy(
    r4["Id_empleado"], r4["Nombre"], r4["Apellido"]
).agg(
    sum(r4["total_venta"]).alias("VentasTotales")
)

# Ordenar usando la suma sin formatear
vendedores_ordenados = ventas_por_vendedor.orderBy(
    desc("VentasTotales")
)

# Tomar solo los 10 primeros
vendedores_top_10 = vendedores_ordenados.limit(10)

# Formatear la suma
vendedores_top_10 = vendedores_top_10.withColumn("total_ventas_format",
                    format_number(col("VentasTotales"), 2).alias("total_ventas_form
# Mostrar el resultado
vendedores_top_10.show()

#Guardando el Archivo
vendedores_top_10.write.csv("/datos/salida/salida3", header=True, mode="overwrite")
```

```
23/07/08 18:12:27 INFO CodeGenerator: Code generated in 6.98614 ms
23/07/08 18:12:27 INFO DAGScheduler: Registering RDD 4048 (showString at <unknown>:
0) as input to shuffle 672
23/07/08 18:12:27 INFO DAGScheduler: Got map stage job 1311 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:12:27 INFO DAGScheduler: Final stage: ShuffleMapStage 2800 (showString a
t <unknown>:0)
23/07/08 18:12:27 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:12:27 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:27 INFO DAGScheduler: Submitting ShuffleMapStage 2800 (MapPartitionsR
DD[4048] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1603 stored as values in memory
(estimated size 21.5 KiB, free 327.8 MiB)
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1603_piece0 stored as bytes in m
emory (estimated size 10.7 KiB, free 327.8 MiB)
23/07/08 18:12:27 INFO BlockManagerInfo: Added broadcast_1603_piece0 in memory on 17
2.30.115.138:43839 (size: 10.7 KiB, free: 363.0 MiB)
23/07/08 18:12:27 INFO SparkContext: Created broadcast 1603 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:27 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2800 (MapPartitionsRDD[4048] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:12:27 INFO TaskSchedulerImpl: Adding task set 2800.0 with 2 tasks resour
ce profile 0
23/07/08 18:12:27 INFO TaskSetManager: Starting task 0.0 in stage 2800.0 (TID 1793)
(172.30.115.138, executor driver, partition 0, ANY, 7408 bytes)
23/07/08 18:12:27 INFO TaskSetManager: Starting task 1.0 in stage 2800.0 (TID 1794)
(172.30.115.138, executor driver, partition 1, ANY, 7408 bytes)
23/07/08 18:12:27 INFO Executor: Running task 0.0 in stage 2800.0 (TID 1793)
23/07/08 18:12:27 INFO Executor: Running task 1.0 in stage 2800.0 (TID 1794)
23/07/08 18:12:27 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Empleado.
csv:0+8119
23/07/08 18:12:27 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Empleado.
csv:8119+8119
23/07/08 18:12:27 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:12:27 INFO CodeGenerator: Code generated in 18.249304 ms
23/07/08 18:12:27 INFO DAGScheduler: Registering RDD 4050 (showString at <unknown>:
0) as input to shuffle 673
23/07/08 18:12:27 INFO DAGScheduler: Got map stage job 1312 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:12:27 INFO DAGScheduler: Final stage: ShuffleMapStage 2801 (showString a
t <unknown>:0)
23/07/08 18:12:27 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:12:27 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:27 INFO DAGScheduler: Submitting ShuffleMapStage 2801 (MapPartitionsR
DD[4050] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1604 stored as values in memory
(estimated size 22.9 KiB, free 327.8 MiB)
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1604_piece0 stored as bytes in m
emory (estimated size 11.2 KiB, free 327.8 MiB)
23/07/08 18:12:27 INFO BlockManagerInfo: Added broadcast_1604_piece0 in memory on 17
2.30.115.138:43839 (size: 11.2 KiB, free: 363.0 MiB)
23/07/08 18:12:27 INFO SparkContext: Created broadcast 1604 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:27 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2801 (MapPartitionsRDD[4050] at showString at <unknown>:0) (first 15 tasks are for p
```

```
artitions Vector(0, 1))
23/07/08 18:12:27 INFO TaskSchedulerImpl: Adding task set 2801.0 with 2 tasks resour
ce profile 0
23/07/08 18:12:27 INFO TaskSetManager: Starting task 0.0 in stage 2801.0 (TID 1795)
(172.30.115.138, executor driver, partition 0, ANY, 7406 bytes)
23/07/08 18:12:27 INFO TaskSetManager: Starting task 1.0 in stage 2801.0 (TID 1796)
(172.30.115.138, executor driver, partition 1, ANY, 7406 bytes)
23/07/08 18:12:27 INFO Executor: Running task 0.0 in stage 2801.0 (TID 1795)
23/07/08 18:12:27 INFO Executor: Running task 1.0 in stage 2801.0 (TID 1796)
23/07/08 18:12:27 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:1310749+1310750
23/07/08 18:12:27 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:0+1310749
23/07/08 18:12:27 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:12:27 INFO BlockManagerInfo: Removed broadcast_1602_piece0 on 172.30.11
5.138:43839 in memory (size: 34.5 KiB, free: 363.0 MiB)
23/07/08 18:12:27 INFO BlockManagerInfo: Removed broadcast_1601_piece0 on 172.30.11
5.138:43839 in memory (size: 30.0 KiB, free: 363.0 MiB)
23/07/08 18:12:27 INFO BlockManagerInfo: Removed broadcast_1600_piece0 on 172.30.11
5.138:43839 in memory (size: 31.2 KiB, free: 363.1 MiB)
23/07/08 18:12:27 INFO PythonRunner: Times: total = 137, boot = -10750, init = 1088
6, finish = 1
23/07/08 18:12:27 INFO PythonRunner: Times: total = 142, boot = -10765, init = 1090
5, finish = 2
23/07/08 18:12:27 INFO Executor: Finished task 1.0 in stage 2800.0 (TID 1794). 2524
bytes result sent to driver
23/07/08 18:12:27 INFO TaskSetManager: Finished task 1.0 in stage 2800.0 (TID 1794)
in 182 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:12:27 INFO Executor: Finished task 0.0 in stage 2800.0 (TID 1793). 2524
bytes result sent to driver
23/07/08 18:12:27 INFO TaskSetManager: Finished task 0.0 in stage 2800.0 (TID 1793)
in 186 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:12:27 INFO TaskSchedulerImpl: Removed TaskSet 2800.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:27 INFO DAGScheduler: ShuffleMapStage 2800 (showString at <unknown>:
0) finished in 0.193 s
23/07/08 18:12:27 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:27 INFO DAGScheduler: running: Set(ShuffleMapStage 2801)
23/07/08 18:12:27 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:27 INFO DAGScheduler: failed: Set()
23/07/08 18:12:27 INFO ShufflePartitionsUtil: For shuffle(672), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:27 INFO SparkContext: Starting job: $anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266
23/07/08 18:12:27 INFO DAGScheduler: Got job 1313 ($anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266) with 1 output partitions
23/07/08 18:12:27 INFO DAGScheduler: Final stage: ResultStage 2803 ($anonfun$withThr
eadLocalCaptured$1 at FutureTask.java:266)
23/07/08 18:12:27 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 28
02)
23/07/08 18:12:27 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:27 INFO DAGScheduler: Submitting ResultStage 2803 (MapPartitionsRDD[4
052] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:266), which has no mis
sing parents
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1605 stored as values in memory
(estimated size 8.2 KiB, free 328.1 MiB)
```

```
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1605_piece0 stored as bytes in m
emory (estimated size 4.2 KiB, free 328.1 MiB)
23/07/08 18:12:27 INFO BlockManagerInfo: Added broadcast_1605_piece0 in memory on 17
2.30.115.138:43839 (size: 4.2 KiB, free: 363.1 MiB)
23/07/08 18:12:27 INFO SparkContext: Created broadcast 1605 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:27 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 280
3 (MapPartitionsRDD[4052] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:2
66) (first 15 tasks are for partitions Vector(0))
23/07/08 18:12:27 INFO TaskSchedulerImpl: Adding task set 2803.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:27 INFO TaskSetManager: Starting task 0.0 in stage 2803.0 (TID 1797)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7379 bytes)
23/07/08 18:12:27 INFO Executor: Running task 0.0 in stage 2803.0 (TID 1797)
23/07/08 18:12:27 INFO ShuffleBlockFetcherIterator: Getting 2 (22.5 KiB) non-empty b
locks including 2 (22.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:12:27 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:27 INFO Executor: Finished task 0.0 in stage 2803.0 (TID 1797). 10564
bytes result sent to driver
23/07/08 18:12:27 INFO TaskSetManager: Finished task 0.0 in stage 2803.0 (TID 1797)
in 8 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:27 INFO TaskSchedulerImpl: Removed TaskSet 2803.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:27 INFO DAGScheduler: ResultStage 2803 ($anonfun$withThreadLocalCaptu
red$1 at FutureTask.java:266) finished in 0.019 s
23/07/08 18:12:27 INFO DAGScheduler: Job 1313 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:12:27 INFO TaskSchedulerImpl: Killing all running tasks in stage 2803: S
tage finished
23/07/08 18:12:27 INFO DAGScheduler: Job 1313 finished: $anonfun$withThreadLocalCapt
ured$1 at FutureTask.java:266, took 0.026069 s
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1606 stored as values in memory
(estimated size 2.0 MiB, free 326.1 MiB)
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1606_piece0 stored as bytes in m
emory (estimated size 8.5 KiB, free 326.1 MiB)
23/07/08 18:12:27 INFO BlockManagerInfo: Added broadcast_1606_piece0 in memory on 17
2.30.115.138:43839 (size: 8.5 KiB, free: 363.1 MiB)
23/07/08 18:12:27 INFO SparkContext: Created broadcast 1606 from $anonfun$withThread
LocalCaptured$1 at FutureTask.java:266
23/07/08 18:12:27 INFO BlockManagerInfo: Removed broadcast_1605_piece0 on 172.30.11
5.138:43839 in memory (size: 4.2 KiB, free: 363.1 MiB)
23/07/08 18:12:27 INFO PythonRunner: Times: total = 304, boot = -9771, init = 9919,
finish = 156
23/07/08 18:12:27 INFO Executor: Finished task 0.0 in stage 2801.0 (TID 1795). 2524
bytes result sent to driver
23/07/08 18:12:27 INFO TaskSetManager: Finished task 0.0 in stage 2801.0 (TID 1795)
in 349 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:12:27 INFO PythonRunner: Times: total = 289, boot = -9764, init = 9909,
finish = 144
23/07/08 18:12:27 INFO Executor: Finished task 1.0 in stage 2801.0 (TID 1796). 2524
bytes result sent to driver
23/07/08 18:12:27 INFO TaskSetManager: Finished task 1.0 in stage 2801.0 (TID 1796)
in 365 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:12:27 INFO TaskSchedulerImpl: Removed TaskSet 2801.0, whose tasks have a
ll completed, from pool
```

```
23/07/08 18:12:27 INFO DAGScheduler: ShuffleMapStage 2801 (showString at <unknown>:
0) finished in 0.370 s
23/07/08 18:12:27 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:27 INFO DAGScheduler: running: Set()
23/07/08 18:12:27 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:27 INFO DAGScheduler: failed: Set()
23/07/08 18:12:27 INFO ShufflePartitionsUtil: For shuffle(673), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:27 INFO HashAggregateExec: spark.sql.codegen.aggregate.map.twolevel.e
nabled is set to true, but current version of codegened fast hashmap does not suppor
t this aggregate.
23/07/08 18:12:27 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:12:27 INFO DAGScheduler: Got job 1314 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:12:27 INFO DAGScheduler: Final stage: ResultStage 2805 (showString at <u
nknown>:0)
23/07/08 18:12:27 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 28
04)
23/07/08 18:12:27 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:27 INFO DAGScheduler: Submitting ResultStage 2805 (MapPartitionsRDD[4
056] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1607 stored as values in memory
(estimated size 79.9 KiB, free 326.0 MiB)
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1607_piece0 stored as bytes in m
emory (estimated size 34.3 KiB, free 326.0 MiB)
23/07/08 18:12:27 INFO BlockManagerInfo: Added broadcast_1607_piece0 in memory on 17
2.30.115.138:43839 (size: 34.3 KiB, free: 363.0 MiB)
23/07/08 18:12:27 INFO SparkContext: Created broadcast 1607 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:27 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 280
5 (MapPartitionsRDD[4056] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:12:27 INFO TaskSchedulerImpl: Adding task set 2805.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:27 INFO TaskSetManager: Starting task 0.0 in stage 2805.0 (TID 1798)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7363 bytes)
23/07/08 18:12:27 INFO Executor: Running task 0.0 in stage 2805.0 (TID 1798)
23/07/08 18:12:27 INFO ShuffleBlockFetcherIterator: Getting 2 (341.4 KiB) non-empty
blocks including 2 (341.4 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:12:27 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:27 INFO CodeGenerator: Code generated in 6.671922 ms
23/07/08 18:12:27 INFO Executor: Finished task 0.0 in stage 2805.0 (TID 1798). 7855
bytes result sent to driver
23/07/08 18:12:27 INFO TaskSetManager: Finished task 0.0 in stage 2805.0 (TID 1798)
in 60 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:27 INFO TaskSchedulerImpl: Removed TaskSet 2805.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:27 INFO DAGScheduler: ResultStage 2805 (showString at <unknown>:0) fi
nished in 0.067 s
23/07/08 18:12:27 INFO DAGScheduler: Job 1314 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:12:27 INFO TaskSchedulerImpl: Killing all running tasks in stage 2805: S
tage finished
23/07/08 18:12:27 INFO DAGScheduler: Job 1314 finished: showString at <unknown>:0, t
ook 0.069927 s
```

```
23/07/08 18:12:27 INFO CodeGenerator: Code generated in 9.112101 ms
23/07/08 18:12:27 INFO CodeGenerator: Code generated in 5.56049 ms
23/07/08 18:12:27 INFO DAGScheduler: Registering RDD 4058 (csv at <unknown>:0) as in
put to shuffle 674
23/07/08 18:12:27 INFO DAGScheduler: Got map stage job 1315 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:12:27 INFO DAGScheduler: Final stage: ShuffleMapStage 2806 (csv at <unkn
own>:0)
23/07/08 18:12:27 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:12:27 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:27 INFO DAGScheduler: Submitting ShuffleMapStage 2806 (MapPartitionsR
DD[4058] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1608 stored as values in memory
(estimated size 21.5 KiB, free 326.0 MiB)
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1608_piece0 stored as bytes in m
emory (estimated size 10.7 KiB, free 325.9 MiB)
23/07/08 18:12:27 INFO BlockManagerInfo: Added broadcast_1608_piece0 in memory on 17
2.30.115.138:43839 (size: 10.7 KiB, free: 363.0 MiB)
23/07/08 18:12:27 INFO SparkContext: Created broadcast 1608 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:27 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2806 (MapPartitionsRDD[4058] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:12:27 INFO TaskSchedulerImpl: Adding task set 2806.0 with 2 tasks resour
ce profile 0
23/07/08 18:12:27 INFO DAGScheduler: Registering RDD 4060 (csv at <unknown>:0) as in
put to shuffle 675
23/07/08 18:12:27 INFO DAGScheduler: Got map stage job 1316 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:12:27 INFO DAGScheduler: Final stage: ShuffleMapStage 2807 (csv at <unkn
own>:0)
23/07/08 18:12:27 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:12:27 INFO TaskSetManager: Starting task 0.0 in stage 2806.0 (TID 1799)
(172.30.115.138, executor driver, partition 0, ANY, 7408 bytes)
23/07/08 18:12:27 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:27 INFO TaskSetManager: Starting task 1.0 in stage 2806.0 (TID 1800)
(172.30.115.138, executor driver, partition 1, ANY, 7408 bytes)
23/07/08 18:12:27 INFO DAGScheduler: Submitting ShuffleMapStage 2807 (MapPartitionsR
DD[4060] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:27 INFO Executor: Running task 0.0 in stage 2806.0 (TID 1799)
23/07/08 18:12:27 INFO Executor: Running task 1.0 in stage 2806.0 (TID 1800)
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1609 stored as values in memory
(estimated size 22.9 KiB, free 325.9 MiB)
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1609_piece0 stored as bytes in m
emory (estimated size 11.2 KiB, free 325.9 MiB)
23/07/08 18:12:27 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Empleado.
csv:8119+8119
23/07/08 18:12:27 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Empleado.
csv:0+8119
23/07/08 18:12:27 INFO BlockManagerInfo: Added broadcast_1609_piece0 in memory on 17
2.30.115.138:43839 (size: 11.2 KiB, free: 363.0 MiB)
23/07/08 18:12:27 INFO SparkContext: Created broadcast 1609 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:27 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2807 (MapPartitionsRDD[4060] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
```

```
23/07/08 18:12:27 INFO TaskSchedulerImpl: Adding task set 2807.0 with 2 tasks resour
ce profile 0
23/07/08 18:12:27 INFO TaskSetManager: Starting task 0.0 in stage 2807.0 (TID 1801)
(172.30.115.138, executor driver, partition 0, ANY, 7406 bytes)
23/07/08 18:12:27 INFO TaskSetManager: Starting task 1.0 in stage 2807.0 (TID 1802)
(172.30.115.138, executor driver, partition 1, ANY, 7406 bytes)
23/07/08 18:12:27 INFO Executor: Running task 0.0 in stage 2807.0 (TID 1801)
23/07/08 18:12:27 INFO Executor: Running task 1.0 in stage 2807.0 (TID 1802)
23/07/08 18:12:27 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:1310749+1310750
23/07/08 18:12:27 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:0+1310749
23/07/08 18:12:27 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:12:27 INFO LineRecordReader: Found UTF-8 BOM and skipped it
```

Id_empleado	Nombre	Apellido	VentasTotales	total_ventas_format
3043	Jacobo	Higuita	7.520016199E7	75,200,161.99
2351	David	Gracía	5.4225736059999995E7	54,225,736.06
1966	Camila	Dominguez	4.059298618E7	40,592,986.18
1531	Angela	Alzate	2.018225373E7	20,182,253.73
1329	Julieth	Osorio	8563928.940000001	8,563,928.94
1675	Jorge	Zea	8173400.110000002	8,173,400.11
1675	Luis	Melano	8173400.110000002	8,173,400.11
1012	Julian	Duque	7285232.22	7,285,232.22
1426	Pablo	Rojas	7255398.9899999965	7,255,398.99
1673	Stepania	Zapata	7008054.420000001	7,008,054.42

```
23/07/08 18:12:27 INFO PythonRunner: Times: total = 112, boot = -10197, init = 1030
8, finish = 1
23/07/08 18:12:27 INFO PythonRunner: Times: total = 114, boot = -10204, init = 1031
7, finish = 1
23/07/08 18:12:27 INFO Executor: Finished task 0.0 in stage 2806.0 (TID 1799). 2567
bytes result sent to driver
23/07/08 18:12:27 INFO TaskSetManager: Finished task 0.0 in stage 2806.0 (TID 1799)
in 164 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:12:27 INFO BlockManagerInfo: Removed broadcast_1607_piece0 on 172.30.11
5.138:43839 in memory (size: 34.3 KiB, free: 363.0 MiB)
23/07/08 18:12:27 INFO Executor: Finished task 1.0 in stage 2806.0 (TID 1800). 2524
bytes result sent to driver
23/07/08 18:12:27 INFO TaskSetManager: Finished task 1.0 in stage 2806.0 (TID 1800)
in 169 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:12:27 INFO TaskSchedulerImpl: Removed TaskSet 2806.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:27 INFO DAGScheduler: ShuffleMapStage 2806 (csv at <unknown>:0) finis
hed in 0.173 s
23/07/08 18:12:27 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:27 INFO DAGScheduler: running: Set(ShuffleMapStage 2807)
23/07/08 18:12:27 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:27 INFO DAGScheduler: failed: Set()
23/07/08 18:12:27 INFO ShufflePartitionsUtil: For shuffle(674), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:27 INFO SparkContext: Starting job: $anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266
23/07/08 18:12:27 INFO DAGScheduler: Got job 1317 ($anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266) with 1 output partitions
23/07/08 18:12:27 INFO DAGScheduler: Final stage: ResultStage 2809 ($anonfun$withThr
eadLocalCaptured$1 at FutureTask.java:266)
23/07/08 18:12:27 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 28
08)
23/07/08 18:12:27 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:27 INFO DAGScheduler: Submitting ResultStage 2809 (MapPartitionsRDD[4
062] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:266), which has no mis
sing parents
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1610 stored as values in memory
(estimated size 8.2 KiB, free 326.0 MiB)
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1610_piece0 stored as bytes in m
emory (estimated size 4.2 KiB, free 326.0 MiB)
23/07/08 18:12:27 INFO BlockManagerInfo: Added broadcast_1610_piece0 in memory on 17
2.30.115.138:43839 (size: 4.2 KiB, free: 363.0 MiB)
23/07/08 18:12:27 INFO SparkContext: Created broadcast 1610 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:27 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 280
9 (MapPartitionsRDD[4062] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:2
66) (first 15 tasks are for partitions Vector(0))
23/07/08 18:12:27 INFO TaskSchedulerImpl: Adding task set 2809.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:27 INFO TaskSetManager: Starting task 0.0 in stage 2809.0 (TID 1803)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7379 bytes)
23/07/08 18:12:27 INFO Executor: Running task 0.0 in stage 2809.0 (TID 1803)
23/07/08 18:12:27 INFO ShuffleBlockFetcherIterator: Getting 2 (22.5 KiB) non-empty b
locks including 2 (22.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:12:27 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
```

```
23/07/08 18:12:27 INFO Executor: Finished task 0.0 in stage 2809.0 (TID 1803). 10564
bytes result sent to driver
23/07/08 18:12:27 INFO TaskSetManager: Finished task 0.0 in stage 2809.0 (TID 1803)
in 6 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:27 INFO TaskSchedulerImpl: Removed TaskSet 2809.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:27 INFO DAGScheduler: ResultStage 2809 ($anonfun$withThreadLocalCaptu
red$1 at FutureTask.java:266) finished in 0.011 s
23/07/08 18:12:27 INFO DAGScheduler: Job 1317 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:12:27 INFO TaskSchedulerImpl: Killing all running tasks in stage 2809: S
tage finished
23/07/08 18:12:27 INFO DAGScheduler: Job 1317 finished: $anonfun$withThreadLocalCapt
ured$1 at FutureTask.java:266, took 0.013582 s
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1611 stored as values in memory
(estimated size 2.0 MiB, free 324.0 MiB)
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1611_piece0 stored as bytes in m
emory (estimated size 8.5 KiB, free 324.0 MiB)
23/07/08 18:12:27 INFO BlockManagerInfo: Added broadcast_1611_piece0 in memory on 17
2.30.115.138:43839 (size: 8.5 KiB, free: 363.0 MiB)
23/07/08 18:12:27 INFO SparkContext: Created broadcast 1611 from $anonfun$withThread
LocalCaptured$1 at FutureTask.java:266
23/07/08 18:12:27 INFO PythonRunner: Times: total = 232, boot = -384, init = 492, fi
nish = 124
23/07/08 18:12:27 INFO BlockManagerInfo: Removed broadcast_1610_piece0 on 172.30.11
5.138:43839 in memory (size: 4.2 KiB, free: 363.0 MiB)
23/07/08 18:12:27 INFO Executor: Finished task 0.0 in stage 2807.0 (TID 1801). 2524
bytes result sent to driver
23/07/08 18:12:27 INFO TaskSetManager: Finished task 0.0 in stage 2807.0 (TID 1801)
in 286 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:12:27 INFO PythonRunner: Times: total = 244, boot = -389, init = 505, fi
nish = 128
23/07/08 18:12:27 INFO Executor: Finished task 1.0 in stage 2807.0 (TID 1802). 2524
bytes result sent to driver
23/07/08 18:12:27 INFO TaskSetManager: Finished task 1.0 in stage 2807.0 (TID 1802)
in 299 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:12:27 INFO TaskSchedulerImpl: Removed TaskSet 2807.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:27 INFO DAGScheduler: ShuffleMapStage 2807 (csv at <unknown>:0) finis
hed in 0.305 s
23/07/08 18:12:27 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:27 INFO DAGScheduler: running: Set()
23/07/08 18:12:27 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:27 INFO DAGScheduler: failed: Set()
23/07/08 18:12:27 INFO ShufflePartitionsUtil: For shuffle(675), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:27 INFO FileOutputCommitter: File Output Committer Algorithm version
is 1
23/07/08 18:12:27 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _tempor
ary folders under output directory:false, ignore cleanup failures: false
23/07/08 18:12:27 INFO SQLHadoopMapReduceCommitProtocol: Using output committer clas
s org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
23/07/08 18:12:27 INFO HashAggregateExec: spark.sql.codegen.aggregate.map.twolevel.e
nabled is set to true, but current version of codegened fast hashmap does not suppor
t this aggregate.
23/07/08 18:12:27 INFO SparkContext: Starting job: csv at <unknown>:0
```

```
23/07/08 18:12:27 INFO DAGScheduler: Got job 1318 (csv at <unknown>:0) with 1 output
partitions
23/07/08 18:12:27 INFO DAGScheduler: Final stage: ResultStage 2811 (csv at <unknown
>:0)
23/07/08 18:12:27 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 28
10)
23/07/08 18:12:27 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:27 INFO DAGScheduler: Submitting ResultStage 2811 (MapPartitionsRDD[4
066] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1612 stored as values in memory
(estimated size 384.0 KiB, free 323.6 MiB)
23/07/08 18:12:27 INFO MemoryStore: Block broadcast_1612_piece0 stored as bytes in m
emory (estimated size 146.1 KiB, free 323.5 MiB)
23/07/08 18:12:27 INFO BlockManagerInfo: Added broadcast_1612_piece0 in memory on 17
2.30.115.138:43839 (size: 146.1 KiB, free: 362.9 MiB)
23/07/08 18:12:27 INFO SparkContext: Created broadcast 1612 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:27 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 281
1 (MapPartitionsRDD[4066] at csv at <unknown>:0) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:12:27 INFO TaskSchedulerImpl: Adding task set 2811.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:27 INFO TaskSetManager: Starting task 0.0 in stage 2811.0 (TID 1804)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7363 bytes)
23/07/08 18:12:27 INFO Executor: Running task 0.0 in stage 2811.0 (TID 1804)
23/07/08 18:12:27 INFO ShuffleBlockFetcherIterator: Getting 2 (341.4 KiB) non-empty
blocks including 2 (341.4 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:12:27 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:28 INFO CodeGenerator: Code generated in 9.2261 ms
23/07/08 18:12:28 INFO FileOutputCommitter: File Output Committer Algorithm version
is 1
23/07/08 18:12:28 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _tempor
ary folders under output directory:false, ignore cleanup failures: false
23/07/08 18:12:28 INFO SQLHadoopMapReduceCommitProtocol: Using output committer clas
s org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
23/07/08 18:12:28 INFO FileOutputCommitter: Saved output of task 'attempt_2023070818
122790001413119887930 06_2811_m_000000_1804' to hdfs://127.0.0.1:9000/datos/salida/sa
lida3/_temporary/0/task_20230708181227900014131198879300 6_2811_m_000000
23/07/08 18:12:28 INFO SparkHadoopMapRedUtil: attempt_20230708181227900014131198879
3006_2811_m_000000_1804: Committed. Elapsed time: 6 ms.
23/07/08 18:12:28 INFO Executor: Finished task 0.0 in stage 2811.0 (TID 1804). 9180
bytes result sent to driver
23/07/08 18:12:28 INFO TaskSetManager: Finished task 0.0 in stage 2811.0 (TID 1804)
in 519 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:28 INFO TaskSchedulerImpl: Removed TaskSet 2811.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:28 INFO DAGScheduler: ResultStage 2811 (csv at <unknown>:0) finished
in 0.550 s
23/07/08 18:12:28 INFO DAGScheduler: Job 1318 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:12:28 INFO TaskSchedulerImpl: Killing all running tasks in stage 2811: S
tage finished
23/07/08 18:12:28 INFO DAGScheduler: Job 1318 finished: csv at <unknown>:0, took 0.5
52151 s
23/07/08 18:12:28 INFO FileFormatWriter: Start to commit write Job 42915dff-271c-467
```

```
8-8a29-7a71bddbec64.
23/07/08 18:12:28 INFO FileFormatWriter: Write Job 42915dff-271c-4678-8a29-7a71bddbe
c64 committed. Elapsed time: 27 ms.
23/07/08 18:12:28 INFO FileFormatWriter: Finished processing stats for write job 429
15dff-271c-4678-8a29-7a71bddbec64.
```

d) Buscar el canal de venta con el mejor resultado en ventas, agrupado por año, crear un archivo de salida con el resultado

In [197…

```python
from pyspark.sql.functions import desc, format_number, col, row_number
from pyspark.sql.window import Window

Canal = df_canal.alias("canal")

# Calcular las ventas anuales de cada canal
ventas_por_canal_y_anio = r6.join(Canal, r6["IdCanal"] == Canal["CODIGO"]) \
    .groupBy(Canal["DESCRIPCION"], r6["AÑO"]) \
    .agg(sum(r1["total_venta"]).alias("VentasAnuales"))

# Ordenar los resultados por año y ventas anuales en orden descendente
windowSpec = Window.partitionBy(r6["AÑO"]).orderBy(desc("VentasAnuales"))
ventas_por_canal_y_anio_ranked = ventas_por_canal_y_anio.withColumn("rank", row_num

# Filtrar solo el mejor canal por año
mejor_canal_por_anio = ventas_por_canal_y_anio_ranked.filter(col("rank") == 1)

# Formatear la suma
mejor_canal_por_anio = mejor_canal_por_anio.withColumn("total_ventas_format", forma

#Guardando el Archivo
mejor_canal_por_anio.write.csv("/datos/salida/salida4", header=True, mode="overwrit

# Mostrar el resultado
mejor_canal_por_anio.show()
```

```
23/07/08 18:12:43 INFO DAGScheduler: Registering RDD 4068 (csv at <unknown>:0) as in
put to shuffle 676
23/07/08 18:12:43 INFO DAGScheduler: Got map stage job 1319 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:12:43 INFO DAGScheduler: Final stage: ShuffleMapStage 2812 (csv at <unkn
own>:0)
23/07/08 18:12:43 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:12:43 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:43 INFO DAGScheduler: Submitting ShuffleMapStage 2812 (MapPartitionsR
DD[4068] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:43 INFO MemoryStore: Block broadcast_1613 stored as values in memory
(estimated size 22.5 KiB, free 323.5 MiB)
23/07/08 18:12:43 INFO MemoryStore: Block broadcast_1613_piece0 stored as bytes in m
emory (estimated size 10.8 KiB, free 323.5 MiB)
23/07/08 18:12:43 INFO BlockManagerInfo: Added broadcast_1613_piece0 in memory on 17
2.30.115.138:43839 (size: 10.8 KiB, free: 362.9 MiB)
23/07/08 18:12:43 INFO SparkContext: Created broadcast 1613 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:43 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2812 (MapPartitionsRDD[4068] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:12:43 INFO TaskSchedulerImpl: Adding task set 2812.0 with 2 tasks resour
ce profile 0
23/07/08 18:12:43 INFO TaskSetManager: Starting task 0.0 in stage 2812.0 (TID 1805)
(172.30.115.138, executor driver, partition 0, ANY, 7411 bytes)
23/07/08 18:12:43 INFO TaskSetManager: Starting task 1.0 in stage 2812.0 (TID 1806)
(172.30.115.138, executor driver, partition 1, ANY, 7411 bytes)
23/07/08 18:12:43 INFO Executor: Running task 0.0 in stage 2812.0 (TID 1805)
23/07/08 18:12:43 INFO Executor: Running task 1.0 in stage 2812.0 (TID 1806)
23/07/08 18:12:43 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:0+450275
23/07/08 18:12:43 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:450275+450275
23/07/08 18:12:43 INFO DAGScheduler: Registering RDD 4070 (csv at <unknown>:0) as in
put to shuffle 677
23/07/08 18:12:43 INFO DAGScheduler: Got map stage job 1320 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:12:43 INFO DAGScheduler: Final stage: ShuffleMapStage 2813 (csv at <unkn
own>:0)
23/07/08 18:12:43 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:12:43 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:43 INFO DAGScheduler: Submitting ShuffleMapStage 2813 (MapPartitionsR
DD[4070] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:43 INFO MemoryStore: Block broadcast_1614 stored as values in memory
(estimated size 23.3 KiB, free 323.4 MiB)
23/07/08 18:12:43 INFO MemoryStore: Block broadcast_1614_piece0 stored as bytes in m
emory (estimated size 11.3 KiB, free 323.4 MiB)
23/07/08 18:12:43 INFO BlockManagerInfo: Added broadcast_1614_piece0 in memory on 17
2.30.115.138:43839 (size: 11.3 KiB, free: 362.9 MiB)
23/07/08 18:12:43 INFO SparkContext: Created broadcast 1614 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:43 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2813 (MapPartitionsRDD[4070] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:12:43 INFO TaskSchedulerImpl: Adding task set 2813.0 with 2 tasks resour
ce profile 0
```

```
23/07/08 18:12:43 INFO DAGScheduler: Registering RDD 4072 (csv at <unknown>:0) as in
put to shuffle 678
23/07/08 18:12:43 INFO DAGScheduler: Got map stage job 1321 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:12:43 INFO DAGScheduler: Final stage: ShuffleMapStage 2814 (csv at <unkn
own>:0)
23/07/08 18:12:43 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:12:43 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:43 INFO DAGScheduler: Submitting ShuffleMapStage 2814 (MapPartitionsR
DD[4072] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:43 INFO MemoryStore: Block broadcast_1615 stored as values in memory
(estimated size 20.1 KiB, free 323.4 MiB)
23/07/08 18:12:43 INFO TaskSetManager: Starting task 0.0 in stage 2813.0 (TID 1807)
(172.30.115.138, executor driver, partition 0, ANY, 7406 bytes)
23/07/08 18:12:43 INFO TaskSetManager: Starting task 1.0 in stage 2813.0 (TID 1808)
(172.30.115.138, executor driver, partition 1, ANY, 7406 bytes)
23/07/08 18:12:43 INFO MemoryStore: Block broadcast_1615_piece0 stored as bytes in m
emory (estimated size 10.4 KiB, free 323.4 MiB)
23/07/08 18:12:43 INFO Executor: Running task 0.0 in stage 2813.0 (TID 1807)
23/07/08 18:12:43 INFO BlockManagerInfo: Added broadcast_1615_piece0 in memory on 17
2.30.115.138:43839 (size: 10.4 KiB, free: 362.9 MiB)
23/07/08 18:12:43 INFO SparkContext: Created broadcast 1615 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:43 INFO Executor: Running task 1.0 in stage 2813.0 (TID 1808)
23/07/08 18:12:43 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2814 (MapPartitionsRDD[4072] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:12:43 INFO TaskSchedulerImpl: Adding task set 2814.0 with 2 tasks resour
ce profile 0
23/07/08 18:12:43 INFO TaskSetManager: Starting task 0.0 in stage 2814.0 (TID 1809)
(172.30.115.138, executor driver, partition 0, ANY, 7412 bytes)
23/07/08 18:12:43 INFO TaskSetManager: Starting task 1.0 in stage 2814.0 (TID 1810)
(172.30.115.138, executor driver, partition 1, ANY, 7412 bytes)
23/07/08 18:12:43 INFO Executor: Running task 0.0 in stage 2814.0 (TID 1809)
23/07/08 18:12:43 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:1310749+1310750
23/07/08 18:12:43 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:0+1310749
23/07/08 18:12:43 INFO Executor: Running task 1.0 in stage 2814.0 (TID 1810)
23/07/08 18:12:43 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/CanalDeVe
nta.csv:0+29
23/07/08 18:12:43 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:12:43 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/CanalDeVe
nta.csv:29+29
23/07/08 18:12:43 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:12:43 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:12:43 INFO PythonRunner: Times: total = 141, boot = -15779, init = 1591
9, finish = 1
23/07/08 18:12:43 INFO Executor: Finished task 0.0 in stage 2814.0 (TID 1809). 2481
bytes result sent to driver
23/07/08 18:12:43 INFO TaskSetManager: Finished task 0.0 in stage 2814.0 (TID 1809)
in 171 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:12:43 INFO PythonRunner: Times: total = 134, boot = -15693, init = 1582
7, finish = 0
23/07/08 18:12:43 INFO BlockManagerInfo: Removed broadcast_1612_piece0 on 172.30.11
5.138:43839 in memory (size: 146.1 KiB, free: 363.0 MiB)
```

```
23/07/08 18:12:43 INFO Executor: Finished task 1.0 in stage 2814.0 (TID 1810). 2524
bytes result sent to driver
23/07/08 18:12:43 INFO TaskSetManager: Finished task 1.0 in stage 2814.0 (TID 1810)
in 263 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:12:43 INFO TaskSchedulerImpl: Removed TaskSet 2814.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:43 INFO DAGScheduler: ShuffleMapStage 2814 (csv at <unknown>:0) finis
hed in 0.269 s
23/07/08 18:12:43 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:43 INFO DAGScheduler: running: Set(ShuffleMapStage 2812, ShuffleMapSt
age 2813)
23/07/08 18:12:43 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:43 INFO DAGScheduler: failed: Set()
23/07/08 18:12:43 INFO PythonRunner: Times: total = 217, boot = -16118, init = 1626
0, finish = 75
23/07/08 18:12:43 INFO PythonRunner: Times: total = 196, boot = -16104, init = 1624
6, finish = 54
23/07/08 18:12:43 INFO Executor: Finished task 1.0 in stage 2812.0 (TID 1806). 2524
bytes result sent to driver
23/07/08 18:12:43 INFO TaskSetManager: Finished task 1.0 in stage 2812.0 (TID 1806)
in 308 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:12:43 INFO Executor: Finished task 0.0 in stage 2812.0 (TID 1805). 2524
bytes result sent to driver
23/07/08 18:12:43 INFO TaskSetManager: Finished task 0.0 in stage 2812.0 (TID 1805)
in 326 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:12:43 INFO TaskSchedulerImpl: Removed TaskSet 2812.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:43 INFO DAGScheduler: ShuffleMapStage 2812 (csv at <unknown>:0) finis
hed in 0.330 s
23/07/08 18:12:43 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:43 INFO DAGScheduler: running: Set(ShuffleMapStage 2813)
23/07/08 18:12:43 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:43 INFO DAGScheduler: failed: Set()
23/07/08 18:12:43 INFO ShufflePartitionsUtil: For shuffle(676), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:43 INFO SparkContext: Starting job: $anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266
23/07/08 18:12:43 INFO DAGScheduler: Got job 1322 ($anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266) with 1 output partitions
23/07/08 18:12:43 INFO DAGScheduler: Final stage: ResultStage 2816 ($anonfun$withThr
eadLocalCaptured$1 at FutureTask.java:266)
23/07/08 18:12:43 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 28
15)
23/07/08 18:12:43 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:43 INFO DAGScheduler: Submitting ResultStage 2816 (MapPartitionsRDD[4
074] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:266), which has no mis
sing parents
23/07/08 18:12:43 INFO MemoryStore: Block broadcast_1616 stored as values in memory
(estimated size 8.2 KiB, free 323.9 MiB)
23/07/08 18:12:43 INFO MemoryStore: Block broadcast_1616_piece0 stored as bytes in m
emory (estimated size 4.2 KiB, free 323.9 MiB)
23/07/08 18:12:43 INFO BlockManagerInfo: Added broadcast_1616_piece0 in memory on 17
2.30.115.138:43839 (size: 4.2 KiB, free: 363.0 MiB)
23/07/08 18:12:43 INFO SparkContext: Created broadcast 1616 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:43 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 281
```

```
6 (MapPartitionsRDD[4074] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:2
66) (first 15 tasks are for partitions Vector(0))
23/07/08 18:12:43 INFO TaskSchedulerImpl: Adding task set 2816.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:43 INFO TaskSetManager: Starting task 0.0 in stage 2816.0 (TID 1811)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7379 bytes)
23/07/08 18:12:43 INFO Executor: Running task 0.0 in stage 2816.0 (TID 1811)
23/07/08 18:12:43 INFO ShuffleBlockFetcherIterator: Getting 2 (234.6 KiB) non-empty
blocks including 2 (234.6 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:12:43 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:43 INFO PythonRunner: Times: total = 267, boot = -15780, init = 1591
7, finish = 130
23/07/08 18:12:43 INFO Executor: Finished task 0.0 in stage 2816.0 (TID 1811). 17174
1 bytes result sent to driver
23/07/08 18:12:43 INFO TaskSetManager: Finished task 0.0 in stage 2816.0 (TID 1811)
in 14 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:43 INFO TaskSchedulerImpl: Removed TaskSet 2816.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:43 INFO DAGScheduler: ResultStage 2816 ($anonfun$withThreadLocalCaptu
red$1 at FutureTask.java:266) finished in 0.020 s
23/07/08 18:12:43 INFO DAGScheduler: Job 1322 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:12:43 INFO TaskSchedulerImpl: Killing all running tasks in stage 2816: S
tage finished
23/07/08 18:12:43 INFO DAGScheduler: Job 1322 finished: $anonfun$withThreadLocalCapt
ured$1 at FutureTask.java:266, took 0.020769 s
23/07/08 18:12:43 INFO MemoryStore: Block broadcast_1617 stored as values in memory
(estimated size 2.5 MiB, free 321.4 MiB)
23/07/08 18:12:43 INFO MemoryStore: Block broadcast_1617_piece0 stored as bytes in m
emory (estimated size 299.6 KiB, free 321.1 MiB)
23/07/08 18:12:43 INFO BlockManagerInfo: Added broadcast_1617_piece0 in memory on 17
2.30.115.138:43839 (size: 299.6 KiB, free: 362.7 MiB)
23/07/08 18:12:43 INFO SparkContext: Created broadcast 1617 from $anonfun$withThread
LocalCaptured$1 at FutureTask.java:266
23/07/08 18:12:43 INFO Executor: Finished task 1.0 in stage 2813.0 (TID 1808). 2524
bytes result sent to driver
23/07/08 18:12:43 INFO TaskSetManager: Finished task 1.0 in stage 2813.0 (TID 1808)
in 387 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:12:43 INFO PythonRunner: Times: total = 305, boot = -15690, init = 1586
2, finish = 133
23/07/08 18:12:43 INFO Executor: Finished task 0.0 in stage 2813.0 (TID 1807). 2524
bytes result sent to driver
23/07/08 18:12:43 INFO TaskSetManager: Finished task 0.0 in stage 2813.0 (TID 1807)
in 409 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:12:43 INFO TaskSchedulerImpl: Removed TaskSet 2813.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:43 INFO DAGScheduler: ShuffleMapStage 2813 (csv at <unknown>:0) finis
hed in 0.420 s
23/07/08 18:12:43 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:43 INFO DAGScheduler: running: Set()
23/07/08 18:12:43 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:43 INFO DAGScheduler: failed: Set()
23/07/08 18:12:43 INFO ShufflePartitionsUtil: For shuffle(677), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:43 INFO DAGScheduler: Registering RDD 4077 (csv at <unknown>:0) as in
```

```
put to shuffle 679
23/07/08 18:12:43 INFO DAGScheduler: Got map stage job 1323 (csv at <unknown>:0) wit
h 1 output partitions
23/07/08 18:12:43 INFO DAGScheduler: Final stage: ShuffleMapStage 2818 (csv at <unkn
own>:0)
23/07/08 18:12:43 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 28
17)
23/07/08 18:12:43 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:43 INFO DAGScheduler: Submitting ShuffleMapStage 2818 (MapPartitionsR
DD[4077] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:43 INFO MemoryStore: Block broadcast_1618 stored as values in memory
(estimated size 15.3 KiB, free 321.1 MiB)
23/07/08 18:12:43 INFO MemoryStore: Block broadcast_1618_piece0 stored as bytes in m
emory (estimated size 7.4 KiB, free 321.1 MiB)
23/07/08 18:12:43 INFO BlockManagerInfo: Added broadcast_1618_piece0 in memory on 17
2.30.115.138:43839 (size: 7.4 KiB, free: 362.7 MiB)
23/07/08 18:12:43 INFO SparkContext: Created broadcast 1618 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:43 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
2818 (MapPartitionsRDD[4077] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0))
23/07/08 18:12:43 INFO TaskSchedulerImpl: Adding task set 2818.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:43 INFO TaskSetManager: Starting task 0.0 in stage 2818.0 (TID 1812)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7368 bytes)
23/07/08 18:12:43 INFO Executor: Running task 0.0 in stage 2818.0 (TID 1812)
23/07/08 18:12:43 INFO ShuffleBlockFetcherIterator: Getting 2 (484.3 KiB) non-empty
blocks including 2 (484.3 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:12:43 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:44 INFO Executor: Finished task 0.0 in stage 2818.0 (TID 1812). 4301
bytes result sent to driver
23/07/08 18:12:44 INFO TaskSetManager: Finished task 0.0 in stage 2818.0 (TID 1812)
in 74 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:44 INFO TaskSchedulerImpl: Removed TaskSet 2818.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:44 INFO DAGScheduler: ShuffleMapStage 2818 (csv at <unknown>:0) finis
hed in 0.078 s
23/07/08 18:12:44 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:44 INFO DAGScheduler: running: Set()
23/07/08 18:12:44 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:44 INFO DAGScheduler: failed: Set()
23/07/08 18:12:44 INFO ShufflePartitionsUtil: For shuffle(679, 678), advisory target
size: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:44 INFO DAGScheduler: Registering RDD 4084 (csv at <unknown>:0) as in
put to shuffle 680
23/07/08 18:12:44 INFO DAGScheduler: Got map stage job 1324 (csv at <unknown>:0) wit
h 1 output partitions
23/07/08 18:12:44 INFO DAGScheduler: Final stage: ShuffleMapStage 2822 (csv at <unkn
own>:0)
23/07/08 18:12:44 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 28
21, ShuffleMapStage 2820)
23/07/08 18:12:44 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:44 INFO DAGScheduler: Submitting ShuffleMapStage 2822 (MapPartitionsR
DD[4084] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:44 INFO MemoryStore: Block broadcast_1619 stored as values in memory
```

```
(estimated size 92.7 KiB, free 321.0 MiB)
23/07/08 18:12:44 INFO MemoryStore: Block broadcast_1619_piece0 stored as bytes in m
emory (estimated size 40.1 KiB, free 321.0 MiB)
23/07/08 18:12:44 INFO BlockManagerInfo: Added broadcast_1619_piece0 in memory on 17
2.30.115.138:43839 (size: 40.1 KiB, free: 362.7 MiB)
23/07/08 18:12:44 INFO SparkContext: Created broadcast 1619 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:44 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
2822 (MapPartitionsRDD[4084] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0))
23/07/08 18:12:44 INFO TaskSchedulerImpl: Adding task set 2822.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:44 INFO TaskSetManager: Starting task 0.0 in stage 2822.0 (TID 1813)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7634 bytes)
23/07/08 18:12:44 INFO Executor: Running task 0.0 in stage 2822.0 (TID 1813)
23/07/08 18:12:44 INFO ShuffleBlockFetcherIterator: Getting 1 (406.5 KiB) non-empty
blocks including 1 (406.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:12:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:44 INFO ShuffleBlockFetcherIterator: Getting 2 (274.0 B) non-empty bl
ocks including 2 (274.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-
local and 0 (0.0 B) remote blocks
23/07/08 18:12:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:44 INFO Executor: Finished task 0.0 in stage 2822.0 (TID 1813). 10354
bytes result sent to driver
23/07/08 18:12:44 INFO TaskSetManager: Finished task 0.0 in stage 2822.0 (TID 1813)
in 72 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:44 INFO TaskSchedulerImpl: Removed TaskSet 2822.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:44 INFO DAGScheduler: ShuffleMapStage 2822 (csv at <unknown>:0) finis
hed in 0.078 s
23/07/08 18:12:44 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:44 INFO DAGScheduler: running: Set()
23/07/08 18:12:44 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:44 INFO DAGScheduler: failed: Set()
23/07/08 18:12:44 INFO ShufflePartitionsUtil: For shuffle(680), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:44 INFO HashAggregateExec: spark.sql.codegen.aggregate.map.twolevel.e
nabled is set to true, but current version of codegened fast hashmap does not suppor
t this aggregate.
23/07/08 18:12:44 INFO CodeGenerator: Code generated in 10.370682 ms
23/07/08 18:12:44 INFO DAGScheduler: Registering RDD 4087 (csv at <unknown>:0) as in
put to shuffle 681
23/07/08 18:12:44 INFO DAGScheduler: Got map stage job 1325 (csv at <unknown>:0) wit
h 1 output partitions
23/07/08 18:12:44 INFO DAGScheduler: Final stage: ShuffleMapStage 2827 (csv at <unkn
own>:0)
23/07/08 18:12:44 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 28
26)
23/07/08 18:12:44 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:44 INFO DAGScheduler: Submitting ShuffleMapStage 2827 (MapPartitionsR
DD[4087] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:44 INFO MemoryStore: Block broadcast_1620 stored as values in memory
(estimated size 84.8 KiB, free 320.9 MiB)
23/07/08 18:12:44 INFO MemoryStore: Block broadcast_1620_piece0 stored as bytes in m
emory (estimated size 32.4 KiB, free 320.8 MiB)
```

```
23/07/08 18:12:44 INFO BlockManagerInfo: Added broadcast_1620_piece0 in memory on 17
2.30.115.138:43839 (size: 32.4 KiB, free: 362.6 MiB)
23/07/08 18:12:44 INFO SparkContext: Created broadcast 1620 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:44 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
2827 (MapPartitionsRDD[4087] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0))
23/07/08 18:12:44 INFO TaskSchedulerImpl: Adding task set 2827.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:44 INFO TaskSetManager: Starting task 0.0 in stage 2827.0 (TID 1814)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7352 bytes)
23/07/08 18:12:44 INFO Executor: Running task 0.0 in stage 2827.0 (TID 1814)
23/07/08 18:12:44 INFO ShuffleBlockFetcherIterator: Getting 1 (1739.0 B) non-empty b
locks including 1 (1739.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:12:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:44 INFO Executor: Finished task 0.0 in stage 2827.0 (TID 1814). 11978
bytes result sent to driver
23/07/08 18:12:44 INFO TaskSetManager: Finished task 0.0 in stage 2827.0 (TID 1814)
in 18 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:44 INFO TaskSchedulerImpl: Removed TaskSet 2827.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:44 INFO DAGScheduler: ShuffleMapStage 2827 (csv at <unknown>:0) finis
hed in 0.023 s
23/07/08 18:12:44 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:44 INFO DAGScheduler: running: Set()
23/07/08 18:12:44 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:44 INFO DAGScheduler: failed: Set()
23/07/08 18:12:44 INFO ShufflePartitionsUtil: For shuffle(681), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:44 INFO FileOutputCommitter: File Output Committer Algorithm version
is 1
23/07/08 18:12:44 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _tempor
ary folders under output directory:false, ignore cleanup failures: false
23/07/08 18:12:44 INFO SQLHadoopMapReduceCommitProtocol: Using output committer clas
s org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
23/07/08 18:12:44 INFO CodeGenerator: Code generated in 8.861084 ms
23/07/08 18:12:44 INFO CodeGenerator: Code generated in 11.104265 ms
23/07/08 18:12:44 INFO SparkContext: Starting job: csv at <unknown>:0
23/07/08 18:12:44 INFO DAGScheduler: Got job 1326 (csv at <unknown>:0) with 1 output
partitions
23/07/08 18:12:44 INFO DAGScheduler: Final stage: ResultStage 2833 (csv at <unknown
>:0)
23/07/08 18:12:44 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 28
32)
23/07/08 18:12:44 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:44 INFO DAGScheduler: Submitting ResultStage 2833 (MapPartitionsRDD[4
092] at csv at <unknown>:0), which has no missing parents
23/07/08 18:12:44 INFO MemoryStore: Block broadcast_1621 stored as values in memory
(estimated size 391.7 KiB, free 320.5 MiB)
23/07/08 18:12:44 INFO MemoryStore: Block broadcast_1621_piece0 stored as bytes in m
emory (estimated size 147.7 KiB, free 320.3 MiB)
23/07/08 18:12:44 INFO BlockManagerInfo: Added broadcast_1621_piece0 in memory on 17
2.30.115.138:43839 (size: 147.7 KiB, free: 362.5 MiB)
23/07/08 18:12:44 INFO BlockManagerInfo: Removed broadcast_1616_piece0 on 172.30.11
5.138:43839 in memory (size: 4.2 KiB, free: 362.5 MiB)
```

```
23/07/08 18:12:44 INFO SparkContext: Created broadcast 1621 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:44 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 283
3 (MapPartitionsRDD[4092] at csv at <unknown>:0) (first 15 tasks are for partitions
Vector(0))
23/07/08 18:12:44 INFO TaskSchedulerImpl: Adding task set 2833.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:44 INFO TaskSetManager: Starting task 0.0 in stage 2833.0 (TID 1815)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7363 bytes)
23/07/08 18:12:44 INFO Executor: Running task 0.0 in stage 2833.0 (TID 1815)
23/07/08 18:12:44 INFO BlockManagerInfo: Removed broadcast_1619_piece0 on 172.30.11
5.138:43839 in memory (size: 40.1 KiB, free: 362.5 MiB)
23/07/08 18:12:44 INFO BlockManagerInfo: Removed broadcast_1618_piece0 on 172.30.11
5.138:43839 in memory (size: 7.4 KiB, free: 362.5 MiB)
23/07/08 18:12:44 INFO BlockManagerInfo: Removed broadcast_1620_piece0 on 172.30.11
5.138:43839 in memory (size: 32.4 KiB, free: 362.6 MiB)
23/07/08 18:12:44 INFO ShuffleBlockFetcherIterator: Getting 1 (1026.0 B) non-empty b
locks including 1 (1026.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:12:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:44 INFO CodeGenerator: Code generated in 4.237945 ms
23/07/08 18:12:44 INFO FileOutputCommitter: File Output Committer Algorithm version
is 1
23/07/08 18:12:44 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _tempor
ary folders under output directory:false, ignore cleanup failures: false
23/07/08 18:12:44 INFO SQLHadoopMapReduceCommitProtocol: Using output committer clas
s org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
23/07/08 18:12:44 INFO FileOutputCommitter: Saved output of task 'attempt_2023070818
12447620861055560810963_2833_m_000000_1815' to hdfs://127.0.0.1:9000/datos/salida/sa
lida4/_temporary/0/task_20230708181244762086105556081 0963_2833_m_000000
23/07/08 18:12:44 INFO SparkHadoopMapRedUtil: attempt_20230708181244762086105556081
0963_2833_m_000000_1815: Committed. Elapsed time: 8 ms.
23/07/08 18:12:44 INFO Executor: Finished task 0.0 in stage 2833.0 (TID 1815). 14233
bytes result sent to driver
23/07/08 18:12:44 INFO TaskSetManager: Finished task 0.0 in stage 2833.0 (TID 1815)
in 483 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:44 INFO TaskSchedulerImpl: Removed TaskSet 2833.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:44 INFO DAGScheduler: ResultStage 2833 (csv at <unknown>:0) finished
in 0.518 s
23/07/08 18:12:44 INFO DAGScheduler: Job 1326 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:12:44 INFO TaskSchedulerImpl: Killing all running tasks in stage 2833: S
tage finished
23/07/08 18:12:44 INFO DAGScheduler: Job 1326 finished: csv at <unknown>:0, took 0.5
20093 s
23/07/08 18:12:44 INFO FileFormatWriter: Start to commit write Job f9659423-df0b-457
5-8b5b-aef9db6b5bb2.
23/07/08 18:12:44 INFO FileFormatWriter: Write Job f9659423-df0b-4575-8b5b-aef9db6b5
bb2 committed. Elapsed time: 29 ms.
23/07/08 18:12:44 INFO FileFormatWriter: Finished processing stats for write job f96
59423-df0b-4575-8b5b-aef9db6b5bb2.
23/07/08 18:12:44 INFO DAGScheduler: Registering RDD 4094 (showString at <unknown>:
0) as input to shuffle 682
23/07/08 18:12:44 INFO DAGScheduler: Got map stage job 1327 (showString at <unknown
>:0) with 2 output partitions
```

```
23/07/08 18:12:44 INFO DAGScheduler: Final stage: ShuffleMapStage 2834 (showString a
t <unknown>:0)
23/07/08 18:12:44 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:12:44 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:44 INFO DAGScheduler: Submitting ShuffleMapStage 2834 (MapPartitionsR
DD[4094] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:44 INFO MemoryStore: Block broadcast_1622 stored as values in memory
(estimated size 22.5 KiB, free 320.6 MiB)
23/07/08 18:12:44 INFO MemoryStore: Block broadcast_1622_piece0 stored as bytes in m
emory (estimated size 10.8 KiB, free 320.6 MiB)
23/07/08 18:12:44 INFO BlockManagerInfo: Added broadcast_1622_piece0 in memory on 17
2.30.115.138:43839 (size: 10.8 KiB, free: 362.6 MiB)
23/07/08 18:12:44 INFO SparkContext: Created broadcast 1622 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:44 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2834 (MapPartitionsRDD[4094] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:12:44 INFO TaskSchedulerImpl: Adding task set 2834.0 with 2 tasks resour
ce profile 0
23/07/08 18:12:44 INFO DAGScheduler: Registering RDD 4096 (showString at <unknown>:
0) as input to shuffle 683
23/07/08 18:12:44 INFO DAGScheduler: Got map stage job 1328 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:12:44 INFO TaskSetManager: Starting task 0.0 in stage 2834.0 (TID 1816)
(172.30.115.138, executor driver, partition 0, ANY, 7411 bytes)
23/07/08 18:12:44 INFO DAGScheduler: Final stage: ShuffleMapStage 2835 (showString a
t <unknown>:0)
23/07/08 18:12:44 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:12:44 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:44 INFO TaskSetManager: Starting task 1.0 in stage 2834.0 (TID 1817)
(172.30.115.138, executor driver, partition 1, ANY, 7411 bytes)
23/07/08 18:12:44 INFO DAGScheduler: Submitting ShuffleMapStage 2835 (MapPartitionsR
DD[4096] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:44 INFO Executor: Running task 0.0 in stage 2834.0 (TID 1816)
23/07/08 18:12:44 INFO Executor: Running task 1.0 in stage 2834.0 (TID 1817)
23/07/08 18:12:44 INFO MemoryStore: Block broadcast_1623 stored as values in memory
(estimated size 23.3 KiB, free 320.5 MiB)
23/07/08 18:12:44 INFO MemoryStore: Block broadcast_1623_piece0 stored as bytes in m
emory (estimated size 11.3 KiB, free 320.5 MiB)
23/07/08 18:12:44 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:450275+450275
23/07/08 18:12:44 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:0+450275
23/07/08 18:12:44 INFO BlockManagerInfo: Added broadcast_1623_piece0 in memory on 17
2.30.115.138:43839 (size: 11.3 KiB, free: 362.5 MiB)
23/07/08 18:12:44 INFO SparkContext: Created broadcast 1623 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:44 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2835 (MapPartitionsRDD[4096] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:12:44 INFO TaskSchedulerImpl: Adding task set 2835.0 with 2 tasks resour
ce profile 0
23/07/08 18:12:44 INFO TaskSetManager: Starting task 0.0 in stage 2835.0 (TID 1818)
(172.30.115.138, executor driver, partition 0, ANY, 7406 bytes)
23/07/08 18:12:44 INFO TaskSetManager: Starting task 1.0 in stage 2835.0 (TID 1819)
(172.30.115.138, executor driver, partition 1, ANY, 7406 bytes)
```

```
23/07/08 18:12:44 INFO Executor: Running task 1.0 in stage 2835.0 (TID 1819)
23/07/08 18:12:44 INFO Executor: Running task 0.0 in stage 2835.0 (TID 1818)
23/07/08 18:12:44 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:0+1310749
23/07/08 18:12:44 INFO DAGScheduler: Registering RDD 4098 (showString at <unknown>:
0) as input to shuffle 684
23/07/08 18:12:44 INFO DAGScheduler: Got map stage job 1329 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:12:44 INFO DAGScheduler: Final stage: ShuffleMapStage 2836 (showString a
t <unknown>:0)
23/07/08 18:12:44 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:12:44 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:44 INFO DAGScheduler: Submitting ShuffleMapStage 2836 (MapPartitionsR
DD[4098] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:44 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:1310749+1310750
23/07/08 18:12:44 INFO MemoryStore: Block broadcast_1624 stored as values in memory
(estimated size 20.1 KiB, free 320.5 MiB)
23/07/08 18:12:44 INFO MemoryStore: Block broadcast_1624_piece0 stored as bytes in m
emory (estimated size 10.4 KiB, free 320.5 MiB)
23/07/08 18:12:44 INFO BlockManagerInfo: Added broadcast_1624_piece0 in memory on 17
2.30.115.138:43839 (size: 10.4 KiB, free: 362.5 MiB)
23/07/08 18:12:44 INFO SparkContext: Created broadcast 1624 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:44 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
2836 (MapPartitionsRDD[4098] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:12:44 INFO TaskSchedulerImpl: Adding task set 2836.0 with 2 tasks resour
ce profile 0
23/07/08 18:12:44 INFO TaskSetManager: Starting task 0.0 in stage 2836.0 (TID 1820)
(172.30.115.138, executor driver, partition 0, ANY, 7412 bytes)
23/07/08 18:12:44 INFO TaskSetManager: Starting task 1.0 in stage 2836.0 (TID 1821)
(172.30.115.138, executor driver, partition 1, ANY, 7412 bytes)
23/07/08 18:12:44 INFO Executor: Running task 0.0 in stage 2836.0 (TID 1820)
23/07/08 18:12:44 INFO Executor: Running task 1.0 in stage 2836.0 (TID 1821)
23/07/08 18:12:44 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/CanalDeVe
nta.csv:0+29
23/07/08 18:12:44 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/CanalDeVe
nta.csv:29+29
23/07/08 18:12:44 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:12:44 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:12:44 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:12:45 INFO PythonRunner: Times: total = 134, boot = -1123, init = 1257,
finish = 0
23/07/08 18:12:45 INFO Executor: Finished task 0.0 in stage 2836.0 (TID 1820). 2481
bytes result sent to driver
23/07/08 18:12:45 INFO TaskSetManager: Finished task 0.0 in stage 2836.0 (TID 1820)
in 171 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:12:45 INFO PythonRunner: Times: total = 136, boot = -988, init = 1124, f
inish = 0
23/07/08 18:12:45 INFO Executor: Finished task 1.0 in stage 2836.0 (TID 1821). 2481
bytes result sent to driver
23/07/08 18:12:45 INFO TaskSetManager: Finished task 1.0 in stage 2836.0 (TID 1821)
in 202 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:12:45 INFO TaskSchedulerImpl: Removed TaskSet 2836.0, whose tasks have a
ll completed, from pool
```

```
23/07/08 18:12:45 INFO DAGScheduler: ShuffleMapStage 2836 (showString at <unknown>:
0) finished in 0.208 s
23/07/08 18:12:45 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:45 INFO DAGScheduler: running: Set(ShuffleMapStage 2835, ShuffleMapSt
age 2834)
23/07/08 18:12:45 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:45 INFO DAGScheduler: failed: Set()
23/07/08 18:12:45 INFO BlockManagerInfo: Removed broadcast_1621_piece0 on 172.30.11
5.138:43839 in memory (size: 147.7 KiB, free: 362.7 MiB)
23/07/08 18:12:45 INFO PythonRunner: Times: total = 220, boot = -1112, init = 1275,
finish = 57
23/07/08 18:12:45 INFO PythonRunner: Times: total = 232, boot = -1127, init = 1296,
finish = 63
23/07/08 18:12:45 INFO Executor: Finished task 1.0 in stage 2834.0 (TID 1817). 2524
bytes result sent to driver
23/07/08 18:12:45 INFO TaskSetManager: Finished task 1.0 in stage 2834.0 (TID 1817)
in 314 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:12:45 INFO Executor: Finished task 0.0 in stage 2834.0 (TID 1816). 2524
bytes result sent to driver
23/07/08 18:12:45 INFO TaskSetManager: Finished task 0.0 in stage 2834.0 (TID 1816)
in 318 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:12:45 INFO TaskSchedulerImpl: Removed TaskSet 2834.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:45 INFO DAGScheduler: ShuffleMapStage 2834 (showString at <unknown>:
0) finished in 0.325 s
23/07/08 18:12:45 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:45 INFO DAGScheduler: running: Set(ShuffleMapStage 2835)
23/07/08 18:12:45 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:45 INFO DAGScheduler: failed: Set()
23/07/08 18:12:45 INFO ShufflePartitionsUtil: For shuffle(682), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:45 INFO SparkContext: Starting job: $anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266
23/07/08 18:12:45 INFO DAGScheduler: Got job 1330 ($anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266) with 1 output partitions
23/07/08 18:12:45 INFO DAGScheduler: Final stage: ResultStage 2838 ($anonfun$withThr
eadLocalCaptured$1 at FutureTask.java:266)
23/07/08 18:12:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 28
37)
23/07/08 18:12:45 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:45 INFO DAGScheduler: Submitting ResultStage 2838 (MapPartitionsRDD[4
100] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:266), which has no mis
sing parents
23/07/08 18:12:45 INFO MemoryStore: Block broadcast_1625 stored as values in memory
(estimated size 8.2 KiB, free 321.0 MiB)
23/07/08 18:12:45 INFO MemoryStore: Block broadcast_1625_piece0 stored as bytes in m
emory (estimated size 4.2 KiB, free 321.0 MiB)
23/07/08 18:12:45 INFO BlockManagerInfo: Added broadcast_1625_piece0 in memory on 17
2.30.115.138:43839 (size: 4.2 KiB, free: 362.7 MiB)
23/07/08 18:12:45 INFO SparkContext: Created broadcast 1625 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 283
8 (MapPartitionsRDD[4100] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:2
66) (first 15 tasks are for partitions Vector(0))
23/07/08 18:12:45 INFO TaskSchedulerImpl: Adding task set 2838.0 with 1 tasks resour
ce profile 0
```

```
23/07/08 18:12:45 INFO TaskSetManager: Starting task 0.0 in stage 2838.0 (TID 1822)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7379 bytes)
23/07/08 18:12:45 INFO Executor: Running task 0.0 in stage 2838.0 (TID 1822)
23/07/08 18:12:45 INFO ShuffleBlockFetcherIterator: Getting 2 (234.6 KiB) non-empty
blocks including 2 (234.6 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:12:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:45 INFO PythonRunner: Times: total = 282, boot = -1071, init = 1220,
finish = 133
23/07/08 18:12:45 INFO Executor: Finished task 0.0 in stage 2838.0 (TID 1822). 17174
1 bytes result sent to driver
23/07/08 18:12:45 INFO TaskSetManager: Finished task 0.0 in stage 2838.0 (TID 1822)
in 13 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:45 INFO TaskSchedulerImpl: Removed TaskSet 2838.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:45 INFO DAGScheduler: ResultStage 2838 ($anonfun$withThreadLocalCaptu
red$1 at FutureTask.java:266) finished in 0.017 s
23/07/08 18:12:45 INFO DAGScheduler: Job 1330 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:12:45 INFO TaskSchedulerImpl: Killing all running tasks in stage 2838: S
tage finished
23/07/08 18:12:45 INFO DAGScheduler: Job 1330 finished: $anonfun$withThreadLocalCapt
ured$1 at FutureTask.java:266, took 0.019118 s
23/07/08 18:12:45 INFO MemoryStore: Block broadcast_1626 stored as values in memory
(estimated size 2.5 MiB, free 318.5 MiB)
23/07/08 18:12:45 INFO BlockManagerInfo: Removed broadcast_1625_piece0 on 172.30.11
5.138:43839 in memory (size: 4.2 KiB, free: 362.7 MiB)
23/07/08 18:12:45 INFO MemoryStore: Block broadcast_1626_piece0 stored as bytes in m
emory (estimated size 299.6 KiB, free 318.2 MiB)
23/07/08 18:12:45 INFO BlockManagerInfo: Added broadcast_1626_piece0 in memory on 17
2.30.115.138:43839 (size: 299.6 KiB, free: 362.4 MiB)
23/07/08 18:12:45 INFO SparkContext: Created broadcast 1626 from $anonfun$withThread
LocalCaptured$1 at FutureTask.java:266
23/07/08 18:12:45 INFO Executor: Finished task 0.0 in stage 2835.0 (TID 1818). 2524
bytes result sent to driver
23/07/08 18:12:45 INFO TaskSetManager: Finished task 0.0 in stage 2835.0 (TID 1818)
in 382 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:12:45 INFO PythonRunner: Times: total = 315, boot = -1040, init = 1209,
finish = 146
23/07/08 18:12:45 INFO Executor: Finished task 1.0 in stage 2835.0 (TID 1819). 2524
bytes result sent to driver
23/07/08 18:12:45 INFO TaskSetManager: Finished task 1.0 in stage 2835.0 (TID 1819)
in 418 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:12:45 INFO TaskSchedulerImpl: Removed TaskSet 2835.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:45 INFO DAGScheduler: ShuffleMapStage 2835 (showString at <unknown>:
0) finished in 0.426 s
23/07/08 18:12:45 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:45 INFO DAGScheduler: running: Set()
23/07/08 18:12:45 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:45 INFO DAGScheduler: failed: Set()
23/07/08 18:12:45 INFO ShufflePartitionsUtil: For shuffle(683), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:45 INFO DAGScheduler: Registering RDD 4103 (showString at <unknown>:
0) as input to shuffle 685
23/07/08 18:12:45 INFO DAGScheduler: Got map stage job 1331 (showString at <unknown
```

```
>:0) with 1 output partitions
23/07/08 18:12:45 INFO DAGScheduler: Final stage: ShuffleMapStage 2840 (showString a
t <unknown>:0)
23/07/08 18:12:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 28
39)
23/07/08 18:12:45 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:45 INFO DAGScheduler: Submitting ShuffleMapStage 2840 (MapPartitionsR
DD[4103] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:45 INFO MemoryStore: Block broadcast_1627 stored as values in memory
(estimated size 15.3 KiB, free 318.2 MiB)
23/07/08 18:12:45 INFO MemoryStore: Block broadcast_1627_piece0 stored as bytes in m
emory (estimated size 7.4 KiB, free 318.2 MiB)
23/07/08 18:12:45 INFO BlockManagerInfo: Added broadcast_1627_piece0 in memory on 17
2.30.115.138:43839 (size: 7.4 KiB, free: 362.4 MiB)
23/07/08 18:12:45 INFO SparkContext: Created broadcast 1627 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:45 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
2840 (MapPartitionsRDD[4103] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0))
23/07/08 18:12:45 INFO TaskSchedulerImpl: Adding task set 2840.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:45 INFO TaskSetManager: Starting task 0.0 in stage 2840.0 (TID 1823)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7368 bytes)
23/07/08 18:12:45 INFO Executor: Running task 0.0 in stage 2840.0 (TID 1823)
23/07/08 18:12:45 INFO ShuffleBlockFetcherIterator: Getting 2 (484.3 KiB) non-empty
blocks including 2 (484.3 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:12:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:45 INFO Executor: Finished task 0.0 in stage 2840.0 (TID 1823). 4301
bytes result sent to driver
23/07/08 18:12:45 INFO TaskSetManager: Finished task 0.0 in stage 2840.0 (TID 1823)
in 40 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:45 INFO TaskSchedulerImpl: Removed TaskSet 2840.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:45 INFO DAGScheduler: ShuffleMapStage 2840 (showString at <unknown>:
0) finished in 0.045 s
23/07/08 18:12:45 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:45 INFO DAGScheduler: running: Set()
23/07/08 18:12:45 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:45 INFO DAGScheduler: failed: Set()
23/07/08 18:12:45 INFO ShufflePartitionsUtil: For shuffle(685, 684), advisory target
size: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:45 INFO DAGScheduler: Registering RDD 4110 (showString at <unknown>:
0) as input to shuffle 686
23/07/08 18:12:45 INFO DAGScheduler: Got map stage job 1332 (showString at <unknown
>:0) with 1 output partitions
23/07/08 18:12:45 INFO DAGScheduler: Final stage: ShuffleMapStage 2844 (showString a
t <unknown>:0)
23/07/08 18:12:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 28
42, ShuffleMapStage 2843)
23/07/08 18:12:45 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:45 INFO DAGScheduler: Submitting ShuffleMapStage 2844 (MapPartitionsR
DD[4110] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:45 INFO MemoryStore: Block broadcast_1628 stored as values in memory
(estimated size 92.8 KiB, free 318.1 MiB)
23/07/08 18:12:45 INFO MemoryStore: Block broadcast_1628_piece0 stored as bytes in m
```

```
emory (estimated size 40.2 KiB, free 318.1 MiB)
23/07/08 18:12:45 INFO BlockManagerInfo: Added broadcast_1628_piece0 in memory on 17
2.30.115.138:43839 (size: 40.2 KiB, free: 362.3 MiB)
23/07/08 18:12:45 INFO SparkContext: Created broadcast 1628 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:45 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
2844 (MapPartitionsRDD[4110] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0))
23/07/08 18:12:45 INFO TaskSchedulerImpl: Adding task set 2844.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:45 INFO TaskSetManager: Starting task 0.0 in stage 2844.0 (TID 1824)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7634 bytes)
23/07/08 18:12:45 INFO Executor: Running task 0.0 in stage 2844.0 (TID 1824)
23/07/08 18:12:45 INFO ShuffleBlockFetcherIterator: Getting 1 (406.5 KiB) non-empty
blocks including 1 (406.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-mer
ged-local and 0 (0.0 B) remote blocks
23/07/08 18:12:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:45 INFO ShuffleBlockFetcherIterator: Getting 2 (274.0 B) non-empty bl
ocks including 2 (274.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-
local and 0 (0.0 B) remote blocks
23/07/08 18:12:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:45 INFO Executor: Finished task 0.0 in stage 2844.0 (TID 1824). 10354
bytes result sent to driver
23/07/08 18:12:45 INFO TaskSetManager: Finished task 0.0 in stage 2844.0 (TID 1824)
in 54 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:45 INFO TaskSchedulerImpl: Removed TaskSet 2844.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:45 INFO DAGScheduler: ShuffleMapStage 2844 (showString at <unknown>:
0) finished in 0.060 s
23/07/08 18:12:45 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:45 INFO DAGScheduler: running: Set()
23/07/08 18:12:45 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:45 INFO DAGScheduler: failed: Set()
23/07/08 18:12:45 INFO ShufflePartitionsUtil: For shuffle(686), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:45 INFO HashAggregateExec: spark.sql.codegen.aggregate.map.twolevel.e
nabled is set to true, but current version of codegened fast hashmap does not suppor
t this aggregate.
23/07/08 18:12:45 INFO DAGScheduler: Registering RDD 4113 (showString at <unknown>:
0) as input to shuffle 687
23/07/08 18:12:45 INFO DAGScheduler: Got map stage job 1333 (showString at <unknown
>:0) with 1 output partitions
23/07/08 18:12:45 INFO DAGScheduler: Final stage: ShuffleMapStage 2849 (showString a
t <unknown>:0)
23/07/08 18:12:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 28
48)
23/07/08 18:12:45 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:45 INFO DAGScheduler: Submitting ShuffleMapStage 2849 (MapPartitionsR
DD[4113] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:45 INFO MemoryStore: Block broadcast_1629 stored as values in memory
(estimated size 84.8 KiB, free 318.0 MiB)
23/07/08 18:12:45 INFO MemoryStore: Block broadcast_1629_piece0 stored as bytes in m
emory (estimated size 32.5 KiB, free 318.0 MiB)
23/07/08 18:12:45 INFO BlockManagerInfo: Added broadcast_1629_piece0 in memory on 17
2.30.115.138:43839 (size: 32.5 KiB, free: 362.3 MiB)
23/07/08 18:12:45 INFO SparkContext: Created broadcast 1629 from broadcast at DAGSch
```

```
eduler.scala:1535
23/07/08 18:12:45 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
2849 (MapPartitionsRDD[4113] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0))
23/07/08 18:12:45 INFO TaskSchedulerImpl: Adding task set 2849.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:45 INFO TaskSetManager: Starting task 0.0 in stage 2849.0 (TID 1825)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7352 bytes)
23/07/08 18:12:45 INFO Executor: Running task 0.0 in stage 2849.0 (TID 1825)
23/07/08 18:12:45 INFO ShuffleBlockFetcherIterator: Getting 1 (1739.0 B) non-empty b
locks including 1 (1739.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:12:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
```

DESCRIPCION	AÑO	VentasAnuales	rank	total_ventas_format
Presencial	2015	8.30127421800004E7	1	83,012,742.18
Telefónica	2016	2.779175894000002E7	1	27,791,758.94
Telefónica	2017	2.9340872620000027E7	1	29,340,872.62
OnLine	2018	9.791991473000018E7	1	97,919,914.73
OnLine	2019	3.3156581069999978E7	1	33,156,581.07
OnLine	2020	6.407009900000002E7	1	64,070,099.00

```
23/07/08 18:12:45 INFO Executor: Finished task 0.0 in stage 2849.0 (TID 1825). 11978
bytes result sent to driver
23/07/08 18:12:45 INFO TaskSetManager: Finished task 0.0 in stage 2849.0 (TID 1825)
in 17 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:45 INFO TaskSchedulerImpl: Removed TaskSet 2849.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:45 INFO DAGScheduler: ShuffleMapStage 2849 (showString at <unknown>:
0) finished in 0.023 s
23/07/08 18:12:45 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:12:45 INFO DAGScheduler: running: Set()
23/07/08 18:12:45 INFO DAGScheduler: waiting: Set()
23/07/08 18:12:45 INFO DAGScheduler: failed: Set()
23/07/08 18:12:45 INFO ShufflePartitionsUtil: For shuffle(687), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:12:45 INFO CodeGenerator: Code generated in 14.26777 ms
23/07/08 18:12:45 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:12:45 INFO DAGScheduler: Got job 1334 (showString at <unknown>:0) with 1
output partitions
23/07/08 18:12:45 INFO DAGScheduler: Final stage: ResultStage 2855 (showString at <u
nknown>:0)
23/07/08 18:12:45 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 28
54)
23/07/08 18:12:45 INFO DAGScheduler: Missing parents: List()
23/07/08 18:12:45 INFO DAGScheduler: Submitting ResultStage 2855 (MapPartitionsRDD[4
118] at showString at <unknown>:0), which has no missing parents
23/07/08 18:12:45 INFO MemoryStore: Block broadcast_1630 stored as values in memory
(estimated size 89.9 KiB, free 317.9 MiB)
23/07/08 18:12:45 INFO MemoryStore: Block broadcast_1630_piece0 stored as bytes in m
emory (estimated size 36.8 KiB, free 317.8 MiB)
23/07/08 18:12:45 INFO BlockManagerInfo: Added broadcast_1630_piece0 in memory on 17
2.30.115.138:43839 (size: 36.8 KiB, free: 362.3 MiB)
23/07/08 18:12:45 INFO SparkContext: Created broadcast 1630 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:12:45 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 285
5 (MapPartitionsRDD[4118] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0))
23/07/08 18:12:45 INFO TaskSchedulerImpl: Adding task set 2855.0 with 1 tasks resour
ce profile 0
23/07/08 18:12:45 INFO TaskSetManager: Starting task 0.0 in stage 2855.0 (TID 1826)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7363 bytes)
23/07/08 18:12:45 INFO Executor: Running task 0.0 in stage 2855.0 (TID 1826)
23/07/08 18:12:45 INFO ShuffleBlockFetcherIterator: Getting 1 (1026.0 B) non-empty b
locks including 1 (1026.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:12:45 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:12:45 INFO Executor: Finished task 0.0 in stage 2855.0 (TID 1826). 13660
bytes result sent to driver
23/07/08 18:12:45 INFO TaskSetManager: Finished task 0.0 in stage 2855.0 (TID 1826)
in 17 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:12:45 INFO TaskSchedulerImpl: Removed TaskSet 2855.0, whose tasks have a
ll completed, from pool
23/07/08 18:12:45 INFO DAGScheduler: ResultStage 2855 (showString at <unknown>:0) fi
nished in 0.023 s
23/07/08 18:12:45 INFO DAGScheduler: Job 1334 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:12:45 INFO TaskSchedulerImpl: Killing all running tasks in stage 2855: S
```

```
tage finished
23/07/08 18:12:45 INFO DAGScheduler: Job 1334 finished: showString at <unknown>:0, t
ook 0.025323 s
```

e) Listar las familias de productos, sucursales, canales de venta y empleados ordenados por mes,año y total de venta alcanzada.

In [212…

```
from pyspark.sql.functions import desc, format_number, col

# Realizar las relaciones adecuadas
r1 = df_producto.join(df_venta, df_producto["ID_PRODUCTO"] == df_venta["IdProducto"
r2 = r1.join(df_sucursal, r1["IdSucursal"] == df_sucursal["ID"])
r3 = r2.join(df_canal, r2["IdCanal"] == df_canal["CODIGO"])
r4 = r3.join(df_empleado, r3["IdEmpleado"] == df_empleado["ID_empleado"])
r5 = r4.join(df_DIMDATE, r4["Fecha"] == df_DIMDATE["FECHA"])

# Agrupamos por mes, año, producto, sucursal, canal y empleado
r = r5.groupBy(
    df_producto['Tipo'].alias("TipoFamilia"),
    df_sucursal['Sucursal'],
    df_canal['DESCRIPCION'].alias('Descripción_Canal'),
    df_DIMDATE['AÑO'],
    df_DIMDATE['MES'],
    df_empleado['Nombre'].alias("Nombre_Empl"),
    df_empleado['Apellido'].alias("Apellido_Empl"),
).agg(
    (sum(df_venta['total_venta'])).alias('TotalVentas')
).orderBy(desc('AÑO'), desc('MES'), desc('TotalVentas'))

# Guardando el Archivo
r.write.csv("/datos/salida/salida5", header=True, mode="overwrite")

# Mostramos el resultado
r.show()
```

```
23/07/08 18:54:35 INFO DAGScheduler: Registering RDD 4950 (csv at <unknown>:0) as in
put to shuffle 878
23/07/08 18:54:35 INFO DAGScheduler: Got map stage job 1615 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:54:35 INFO DAGScheduler: Final stage: ShuffleMapStage 3626 (csv at <unkn
own>:0)
23/07/08 18:54:35 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:54:35 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:35 INFO DAGScheduler: Submitting ShuffleMapStage 3626 (MapPartitionsR
DD[4950] at csv at <unknown>:0), which has no missing parents
23/07/08 18:54:35 INFO MemoryStore: Block broadcast_1971 stored as values in memory
(estimated size 20.6 KiB, free 333.6 MiB)
23/07/08 18:54:35 INFO MemoryStore: Block broadcast_1971_piece0 stored as bytes in m
emory (estimated size 10.6 KiB, free 333.6 MiB)
23/07/08 18:54:35 INFO BlockManagerInfo: Added broadcast_1971_piece0 in memory on 17
2.30.115.138:43839 (size: 10.6 KiB, free: 363.6 MiB)
23/07/08 18:54:35 INFO SparkContext: Created broadcast 1971 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:35 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
3626 (MapPartitionsRDD[4950] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:54:35 INFO TaskSchedulerImpl: Adding task set 3626.0 with 2 tasks resour
ce profile 0
23/07/08 18:54:35 INFO TaskSetManager: Starting task 0.0 in stage 3626.0 (TID 2267)
(172.30.115.138, executor driver, partition 0, ANY, 7408 bytes)
23/07/08 18:54:35 INFO TaskSetManager: Starting task 1.0 in stage 3626.0 (TID 2268)
(172.30.115.138, executor driver, partition 1, ANY, 7408 bytes)
23/07/08 18:54:35 INFO Executor: Running task 1.0 in stage 3626.0 (TID 2268)
23/07/08 18:54:35 INFO Executor: Running task 0.0 in stage 3626.0 (TID 2267)
23/07/08 18:54:35 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Producto.
csv:8440+8441
23/07/08 18:54:35 INFO DAGScheduler: Registering RDD 4952 (csv at <unknown>:0) as in
put to shuffle 879
23/07/08 18:54:35 INFO DAGScheduler: Got map stage job 1616 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:54:35 INFO DAGScheduler: Final stage: ShuffleMapStage 3627 (csv at <unkn
own>:0)
23/07/08 18:54:35 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:54:35 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:35 INFO DAGScheduler: Submitting ShuffleMapStage 3627 (MapPartitionsR
DD[4952] at csv at <unknown>:0), which has no missing parents
23/07/08 18:54:35 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Producto.
csv:0+8440
23/07/08 18:54:35 INFO MemoryStore: Block broadcast_1972 stored as values in memory
(estimated size 24.5 KiB, free 333.5 MiB)
23/07/08 18:54:35 INFO MemoryStore: Block broadcast_1972_piece0 stored as bytes in m
emory (estimated size 11.6 KiB, free 333.5 MiB)
23/07/08 18:54:35 INFO BlockManagerInfo: Added broadcast_1972_piece0 in memory on 17
2.30.115.138:43839 (size: 11.6 KiB, free: 363.6 MiB)
23/07/08 18:54:35 INFO SparkContext: Created broadcast 1972 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:35 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
3627 (MapPartitionsRDD[4952] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:54:35 INFO TaskSchedulerImpl: Adding task set 3627.0 with 2 tasks resour
ce profile 0
```

```
23/07/08 18:54:35 INFO DAGScheduler: Registering RDD 4954 (csv at <unknown>:0) as in
put to shuffle 880
23/07/08 18:54:35 INFO DAGScheduler: Got map stage job 1617 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:54:35 INFO DAGScheduler: Final stage: ShuffleMapStage 3628 (csv at <unkn
own>:0)
23/07/08 18:54:35 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:54:35 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:35 INFO TaskSetManager: Starting task 0.0 in stage 3627.0 (TID 2269)
(172.30.115.138, executor driver, partition 0, ANY, 7406 bytes)
23/07/08 18:54:35 INFO TaskSetManager: Starting task 1.0 in stage 3627.0 (TID 2270)
(172.30.115.138, executor driver, partition 1, ANY, 7406 bytes)
23/07/08 18:54:35 INFO Executor: Running task 1.0 in stage 3627.0 (TID 2270)
23/07/08 18:54:35 INFO Executor: Running task 0.0 in stage 3627.0 (TID 2269)
23/07/08 18:54:35 INFO DAGScheduler: Submitting ShuffleMapStage 3628 (MapPartitionsR
DD[4954] at csv at <unknown>:0), which has no missing parents
23/07/08 18:54:35 INFO MemoryStore: Block broadcast_1973 stored as values in memory
(estimated size 21.1 KiB, free 333.5 MiB)
23/07/08 18:54:35 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:0+1310749
23/07/08 18:54:35 INFO MemoryStore: Block broadcast_1973_piece0 stored as bytes in m
emory (estimated size 10.6 KiB, free 333.5 MiB)
23/07/08 18:54:35 INFO BlockManagerInfo: Added broadcast_1973_piece0 in memory on 17
2.30.115.138:43839 (size: 10.6 KiB, free: 363.6 MiB)
23/07/08 18:54:35 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:1310749+1310750
23/07/08 18:54:35 INFO SparkContext: Created broadcast 1973 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:35 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
3628 (MapPartitionsRDD[4954] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:54:35 INFO TaskSchedulerImpl: Adding task set 3628.0 with 2 tasks resour
ce profile 0
23/07/08 18:54:35 INFO DAGScheduler: Registering RDD 4956 (csv at <unknown>:0) as in
put to shuffle 881
23/07/08 18:54:35 INFO DAGScheduler: Got map stage job 1618 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:54:35 INFO DAGScheduler: Final stage: ShuffleMapStage 3629 (csv at <unkn
own>:0)
23/07/08 18:54:35 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:54:35 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:35 INFO TaskSetManager: Starting task 0.0 in stage 3628.0 (TID 2271)
(172.30.115.138, executor driver, partition 0, ANY, 7408 bytes)
23/07/08 18:54:35 INFO TaskSetManager: Starting task 1.0 in stage 3628.0 (TID 2272)
(172.30.115.138, executor driver, partition 1, ANY, 7408 bytes)
23/07/08 18:54:35 INFO DAGScheduler: Submitting ShuffleMapStage 3629 (MapPartitionsR
DD[4956] at csv at <unknown>:0), which has no missing parents
23/07/08 18:54:35 INFO Executor: Running task 0.0 in stage 3628.0 (TID 2271)
23/07/08 18:54:35 INFO Executor: Running task 1.0 in stage 3628.0 (TID 2272)
23/07/08 18:54:35 INFO MemoryStore: Block broadcast_1974 stored as values in memory
(estimated size 20.1 KiB, free 333.5 MiB)
23/07/08 18:54:35 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Sucursal.
csv:1266+1266
23/07/08 18:54:35 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Sucursal.
csv:0+1266
23/07/08 18:54:35 INFO MemoryStore: Block broadcast_1974_piece0 stored as bytes in m
```

```
emory (estimated size 10.4 KiB, free 333.5 MiB)
23/07/08 18:54:35 INFO BlockManagerInfo: Added broadcast_1974_piece0 in memory on 17
2.30.115.138:43839 (size: 10.4 KiB, free: 363.6 MiB)
23/07/08 18:54:35 INFO SparkContext: Created broadcast 1974 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:35 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
3629 (MapPartitionsRDD[4956] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:54:35 INFO TaskSchedulerImpl: Adding task set 3629.0 with 2 tasks resour
ce profile 0
23/07/08 18:54:35 INFO DAGScheduler: Registering RDD 4958 (csv at <unknown>:0) as in
put to shuffle 882
23/07/08 18:54:35 INFO DAGScheduler: Got map stage job 1619 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:54:35 INFO DAGScheduler: Final stage: ShuffleMapStage 3630 (csv at <unkn
own>:0)
23/07/08 18:54:35 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:54:35 INFO TaskSetManager: Starting task 0.0 in stage 3629.0 (TID 2273)
(172.30.115.138, executor driver, partition 0, ANY, 7412 bytes)
23/07/08 18:54:35 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:35 INFO TaskSetManager: Starting task 1.0 in stage 3629.0 (TID 2274)
(172.30.115.138, executor driver, partition 1, ANY, 7412 bytes)
23/07/08 18:54:35 INFO DAGScheduler: Submitting ShuffleMapStage 3630 (MapPartitionsR
DD[4958] at csv at <unknown>:0), which has no missing parents
23/07/08 18:54:35 INFO Executor: Running task 0.0 in stage 3629.0 (TID 2273)
23/07/08 18:54:35 INFO Executor: Running task 1.0 in stage 3629.0 (TID 2274)
23/07/08 18:54:35 INFO MemoryStore: Block broadcast_1975 stored as values in memory
(estimated size 21.5 KiB, free 333.4 MiB)
23/07/08 18:54:35 INFO MemoryStore: Block broadcast_1975_piece0 stored as bytes in m
emory (estimated size 10.7 KiB, free 333.4 MiB)
23/07/08 18:54:35 INFO BlockManagerInfo: Added broadcast_1975_piece0 in memory on 17
2.30.115.138:43839 (size: 10.7 KiB, free: 363.6 MiB)
23/07/08 18:54:35 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/CanalDeVe
nta.csv:29+29
23/07/08 18:54:35 INFO SparkContext: Created broadcast 1975 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:35 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
3630 (MapPartitionsRDD[4958] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:54:35 INFO TaskSchedulerImpl: Adding task set 3630.0 with 2 tasks resour
ce profile 0
23/07/08 18:54:35 INFO DAGScheduler: Registering RDD 4960 (csv at <unknown>:0) as in
put to shuffle 883
23/07/08 18:54:35 INFO DAGScheduler: Got map stage job 1620 (csv at <unknown>:0) wit
h 2 output partitions
23/07/08 18:54:35 INFO DAGScheduler: Final stage: ShuffleMapStage 3631 (csv at <unkn
own>:0)
23/07/08 18:54:35 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:54:35 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:35 INFO TaskSetManager: Starting task 0.0 in stage 3630.0 (TID 2275)
(172.30.115.138, executor driver, partition 0, ANY, 7408 bytes)
23/07/08 18:54:35 INFO DAGScheduler: Submitting ShuffleMapStage 3631 (MapPartitionsR
DD[4960] at csv at <unknown>:0), which has no missing parents
23/07/08 18:54:35 INFO TaskSetManager: Starting task 1.0 in stage 3630.0 (TID 2276)
(172.30.115.138, executor driver, partition 1, ANY, 7408 bytes)
23/07/08 18:54:35 INFO Executor: Running task 1.0 in stage 3630.0 (TID 2276)
```

```
23/07/08 18:54:35 INFO Executor: Running task 0.0 in stage 3630.0 (TID 2275)
23/07/08 18:54:35 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Empleado.
csv:8119+8119
23/07/08 18:54:35 INFO MemoryStore: Block broadcast_1976 stored as values in memory
(estimated size 22.8 KiB, free 333.4 MiB)
23/07/08 18:54:35 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:54:35 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Empleado.
csv:0+8119
23/07/08 18:54:35 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/CanalDeVe
nta.csv:0+29
23/07/08 18:54:35 INFO MemoryStore: Block broadcast_1976_piece0 stored as bytes in m
emory (estimated size 10.9 KiB, free 333.4 MiB)
23/07/08 18:54:35 INFO BlockManagerInfo: Added broadcast_1976_piece0 in memory on 17
2.30.115.138:43839 (size: 10.9 KiB, free: 363.5 MiB)
23/07/08 18:54:35 INFO SparkContext: Created broadcast 1976 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:35 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
3631 (MapPartitionsRDD[4960] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:54:35 INFO TaskSchedulerImpl: Adding task set 3631.0 with 2 tasks resour
ce profile 0
23/07/08 18:54:35 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:54:35 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:54:36 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:54:36 INFO PythonRunner: Times: total = 161, boot = -18835, init = 1899
6, finish = 0
23/07/08 18:54:36 INFO Executor: Finished task 1.0 in stage 3628.0 (TID 2272). 2524
bytes result sent to driver
23/07/08 18:54:36 INFO TaskSetManager: Starting task 0.0 in stage 3631.0 (TID 2277)
(172.30.115.138, executor driver, partition 0, ANY, 7411 bytes)
23/07/08 18:54:36 INFO Executor: Running task 0.0 in stage 3631.0 (TID 2277)
23/07/08 18:54:36 INFO TaskSetManager: Finished task 1.0 in stage 3628.0 (TID 2272)
in 277 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:36 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:0+450275
23/07/08 18:54:36 INFO PythonRunner: Times: total = 163, boot = -18814, init = 1897
6, finish = 1
23/07/08 18:54:36 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:54:36 INFO PythonRunner: Times: total = 152, boot = -18898, init = 1904
9, finish = 1
23/07/08 18:54:36 INFO Executor: Finished task 0.0 in stage 3630.0 (TID 2275). 2524
bytes result sent to driver
23/07/08 18:54:36 INFO PythonRunner: Times: total = 170, boot = -18833, init = 1900
3, finish = 0
23/07/08 18:54:36 INFO TaskSetManager: Starting task 1.0 in stage 3631.0 (TID 2278)
(172.30.115.138, executor driver, partition 1, ANY, 7411 bytes)
23/07/08 18:54:36 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:54:36 INFO TaskSetManager: Finished task 0.0 in stage 3630.0 (TID 2275)
in 295 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:36 INFO Executor: Running task 1.0 in stage 3631.0 (TID 2278)
23/07/08 18:54:36 INFO Executor: Finished task 0.0 in stage 3628.0 (TID 2271). 2524
bytes result sent to driver
23/07/08 18:54:36 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:450275+450275
23/07/08 18:54:36 INFO TaskSetManager: Finished task 0.0 in stage 3628.0 (TID 2271)
in 318 ms on 172.30.115.138 (executor driver) (2/2)
```

```
23/07/08 18:54:36 INFO TaskSchedulerImpl: Removed TaskSet 3628.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:36 INFO DAGScheduler: ShuffleMapStage 3628 (csv at <unknown>:0) finis
hed in 0.322 s
23/07/08 18:54:36 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:36 INFO DAGScheduler: running: Set(ShuffleMapStage 3631, ShuffleMapSt
age 3629, ShuffleMapStage 3626, ShuffleMapStage 3630, ShuffleMapStage 3627)
23/07/08 18:54:36 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:36 INFO DAGScheduler: failed: Set()
23/07/08 18:54:36 INFO PythonRunner: Times: total = 185, boot = -18854, init = 1903
8, finish = 1
23/07/08 18:54:36 INFO PythonRunner: Times: total = 280, boot = -18818, init = 1909
7, finish = 1
23/07/08 18:54:36 INFO PythonRunner: Times: total = 159, boot = -18765, init = 1892
4, finish = 0
23/07/08 18:54:36 INFO BlockManagerInfo: Removed broadcast_1970_piece0 on 172.30.11
5.138:43839 in memory (size: 45.4 KiB, free: 363.6 MiB)
23/07/08 18:54:36 INFO Executor: Finished task 1.0 in stage 3629.0 (TID 2274). 2524
bytes result sent to driver
23/07/08 18:54:36 INFO Executor: Finished task 1.0 in stage 3626.0 (TID 2268). 2524
bytes result sent to driver
23/07/08 18:54:36 INFO TaskSetManager: Finished task 1.0 in stage 3629.0 (TID 2274)
in 335 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:36 INFO Executor: Finished task 0.0 in stage 3626.0 (TID 2267). 2524
bytes result sent to driver
23/07/08 18:54:36 INFO TaskSetManager: Finished task 1.0 in stage 3626.0 (TID 2268)
in 376 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:36 INFO TaskSetManager: Finished task 0.0 in stage 3626.0 (TID 2267)
in 376 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:36 INFO TaskSchedulerImpl: Removed TaskSet 3626.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:36 INFO DAGScheduler: ShuffleMapStage 3626 (csv at <unknown>:0) finis
hed in 0.383 s
23/07/08 18:54:36 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:36 INFO DAGScheduler: running: Set(ShuffleMapStage 3631, ShuffleMapSt
age 3629, ShuffleMapStage 3630, ShuffleMapStage 3627)
23/07/08 18:54:36 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:36 INFO DAGScheduler: failed: Set()
23/07/08 18:54:36 INFO Executor: Finished task 1.0 in stage 3630.0 (TID 2276). 2524
bytes result sent to driver
23/07/08 18:54:36 INFO TaskSetManager: Finished task 1.0 in stage 3630.0 (TID 2276)
in 333 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:36 INFO TaskSchedulerImpl: Removed TaskSet 3630.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:36 INFO DAGScheduler: ShuffleMapStage 3630 (csv at <unknown>:0) finis
hed in 0.338 s
23/07/08 18:54:36 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:36 INFO DAGScheduler: running: Set(ShuffleMapStage 3631, ShuffleMapSt
age 3629, ShuffleMapStage 3627)
23/07/08 18:54:36 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:36 INFO DAGScheduler: failed: Set()
23/07/08 18:54:36 INFO ShufflePartitionsUtil: For shuffle(878), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:36 INFO ShufflePartitionsUtil: For shuffle(882), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:36 INFO SparkContext: Starting job: $anonfun$withThreadLocalCaptured
```

```
$1 at FutureTask.java:266
23/07/08 18:54:36 INFO PythonRunner: Times: total = 143, boot = -18892, init = 1903
5, finish = 0
23/07/08 18:54:36 INFO DAGScheduler: Got job 1621 ($anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266) with 1 output partitions
23/07/08 18:54:36 INFO DAGScheduler: Final stage: ResultStage 3633 ($anonfun$withThr
eadLocalCaptured$1 at FutureTask.java:266)
23/07/08 18:54:36 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
32)
23/07/08 18:54:36 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:36 INFO DAGScheduler: Submitting ResultStage 3633 (MapPartitionsRDD[4
962] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:266), which has no mis
sing parents
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1977 stored as values in memory
(estimated size 8.2 KiB, free 333.6 MiB)
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1977_piece0 stored as bytes in m
emory (estimated size 4.2 KiB, free 333.6 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Added broadcast_1977_piece0 in memory on 17
2.30.115.138:43839 (size: 4.2 KiB, free: 363.6 MiB)
23/07/08 18:54:36 INFO SparkContext: Created broadcast 1977 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:36 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 363
3 (MapPartitionsRDD[4962] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:2
66) (first 15 tasks are for partitions Vector(0))
23/07/08 18:54:36 INFO TaskSchedulerImpl: Adding task set 3633.0 with 1 tasks resour
ce profile 0
23/07/08 18:54:36 INFO TaskSetManager: Starting task 0.0 in stage 3633.0 (TID 2279)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7379 bytes)
23/07/08 18:54:36 INFO Executor: Running task 0.0 in stage 3633.0 (TID 2279)
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Getting 2 (20.5 KiB) non-empty b
locks including 2 (20.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:36 INFO Executor: Finished task 0.0 in stage 3629.0 (TID 2273). 2567
bytes result sent to driver
23/07/08 18:54:36 INFO SparkContext: Starting job: $anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266
23/07/08 18:54:36 INFO TaskSetManager: Finished task 0.0 in stage 3629.0 (TID 2273)
in 444 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:36 INFO TaskSchedulerImpl: Removed TaskSet 3629.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:36 INFO Executor: Finished task 0.0 in stage 3633.0 (TID 2279). 6937
bytes result sent to driver
23/07/08 18:54:36 INFO DAGScheduler: ShuffleMapStage 3629 (csv at <unknown>:0) finis
hed in 0.458 s
23/07/08 18:54:36 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:36 INFO DAGScheduler: running: Set(ShuffleMapStage 3631, ResultStage
3633, ShuffleMapStage 3627)
23/07/08 18:54:36 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:36 INFO DAGScheduler: failed: Set()
23/07/08 18:54:36 INFO DAGScheduler: Got job 1622 ($anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266) with 1 output partitions
23/07/08 18:54:36 INFO TaskSetManager: Finished task 0.0 in stage 3633.0 (TID 2279)
in 7 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:54:36 INFO DAGScheduler: Final stage: ResultStage 3635 ($anonfun$withThr
eadLocalCaptured$1 at FutureTask.java:266)
```

```
23/07/08 18:54:36 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
34)
23/07/08 18:54:36 INFO TaskSchedulerImpl: Removed TaskSet 3633.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:36 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:36 INFO DAGScheduler: Submitting ResultStage 3635 (MapPartitionsRDD[4
964] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:266), which has no mis
sing parents
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1978 stored as values in memory
(estimated size 8.2 KiB, free 333.6 MiB)
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1978_piece0 stored as bytes in m
emory (estimated size 4.2 KiB, free 333.6 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Added broadcast_1978_piece0 in memory on 17
2.30.115.138:43839 (size: 4.2 KiB, free: 363.6 MiB)
23/07/08 18:54:36 INFO SparkContext: Created broadcast 1978 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:36 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 363
5 (MapPartitionsRDD[4964] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:2
66) (first 15 tasks are for partitions Vector(0))
23/07/08 18:54:36 INFO TaskSchedulerImpl: Adding task set 3635.0 with 1 tasks resour
ce profile 0
23/07/08 18:54:36 INFO PythonRunner: Times: total = 357, boot = -18816, init = 1901
9, finish = 154
23/07/08 18:54:36 INFO DAGScheduler: ResultStage 3633 ($anonfun$withThreadLocalCaptu
red$1 at FutureTask.java:266) finished in 0.016 s
23/07/08 18:54:36 INFO DAGScheduler: Job 1621 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:54:36 INFO TaskSetManager: Starting task 0.0 in stage 3635.0 (TID 2280)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7379 bytes)
23/07/08 18:54:36 INFO TaskSchedulerImpl: Killing all running tasks in stage 3633: S
tage finished
23/07/08 18:54:36 INFO Executor: Running task 0.0 in stage 3635.0 (TID 2280)
23/07/08 18:54:36 INFO DAGScheduler: Job 1621 finished: $anonfun$withThreadLocalCapt
ured$1 at FutureTask.java:266, took 0.023501 s
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Getting 2 (22.5 KiB) non-empty b
locks including 2 (22.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:36 INFO Executor: Finished task 0.0 in stage 3635.0 (TID 2280). 10564
bytes result sent to driver
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1979 stored as values in memory
 (estimated size 2.0 MiB, free 331.5 MiB)
23/07/08 18:54:36 INFO TaskSetManager: Finished task 0.0 in stage 3635.0 (TID 2280)
 in 10 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:54:36 INFO TaskSchedulerImpl: Removed TaskSet 3635.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1979_piece0 stored as bytes in m
emory (estimated size 5.9 KiB, free 331.5 MiB)
23/07/08 18:54:36 INFO DAGScheduler: ResultStage 3635 ($anonfun$withThreadLocalCaptu
red$1 at FutureTask.java:266) finished in 0.015 s
23/07/08 18:54:36 INFO BlockManagerInfo: Added broadcast_1979_piece0 in memory on 17
2.30.115.138:43839 (size: 5.9 KiB, free: 363.6 MiB)
23/07/08 18:54:36 INFO DAGScheduler: Job 1622 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:54:36 INFO SparkContext: Created broadcast 1979 from $anonfun$withThread
LocalCaptured$1 at FutureTask.java:266
```

```
23/07/08 18:54:36 INFO TaskSchedulerImpl: Killing all running tasks in stage 3635: S
tage finished
23/07/08 18:54:36 INFO DAGScheduler: Job 1622 finished: $anonfun$withThreadLocalCapt
ured$1 at FutureTask.java:266, took 0.024061 s
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1980 stored as values in memory
 (estimated size 2.0 MiB, free 329.5 MiB)
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1980_piece0 stored as bytes in m
emory (estimated size 8.5 KiB, free 329.5 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Added broadcast_1980_piece0 in memory on 17
2.30.115.138:43839 (size: 8.5 KiB, free: 363.6 MiB)
23/07/08 18:54:36 INFO SparkContext: Created broadcast 1980 from $anonfun$withThread
LocalCaptured$1 at FutureTask.java:266
23/07/08 18:54:36 INFO Executor: Finished task 1.0 in stage 3627.0 (TID 2270). 2524
 bytes result sent to driver
23/07/08 18:54:36 INFO TaskSetManager: Finished task 1.0 in stage 3627.0 (TID 2270)
 in 504 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:36 INFO PythonRunner: Times: total = 198, boot = 3, init = 144, finis
h = 51
23/07/08 18:54:36 INFO Executor: Finished task 0.0 in stage 3631.0 (TID 2277). 2524
 bytes result sent to driver
23/07/08 18:54:36 INFO PythonRunner: Times: total = 403, boot = -18872, init = 1913
3, finish = 142
23/07/08 18:54:36 INFO TaskSetManager: Finished task 0.0 in stage 3631.0 (TID 2277)
 in 247 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:36 INFO PythonRunner: Times: total = 198, boot = -75, init = 223, fin
ish = 50
23/07/08 18:54:36 INFO Executor: Finished task 0.0 in stage 3627.0 (TID 2269). 2524
 bytes result sent to driver
23/07/08 18:54:36 INFO TaskSetManager: Finished task 0.0 in stage 3627.0 (TID 2269)
 in 555 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:36 INFO TaskSchedulerImpl: Removed TaskSet 3627.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:36 INFO DAGScheduler: ShuffleMapStage 3627 (csv at <unknown>:0) finis
hed in 0.568 s
23/07/08 18:54:36 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:36 INFO DAGScheduler: running: Set(ShuffleMapStage 3631)
23/07/08 18:54:36 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:36 INFO DAGScheduler: failed: Set()
23/07/08 18:54:36 INFO Executor: Finished task 1.0 in stage 3631.0 (TID 2278). 2524
 bytes result sent to driver
23/07/08 18:54:36 INFO TaskSetManager: Finished task 1.0 in stage 3631.0 (TID 2278)
 in 242 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:36 INFO TaskSchedulerImpl: Removed TaskSet 3631.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:36 INFO DAGScheduler: ShuffleMapStage 3631 (csv at <unknown>:0) finis
hed in 0.535 s
23/07/08 18:54:36 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:36 INFO DAGScheduler: running: Set()
23/07/08 18:54:36 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:36 INFO DAGScheduler: failed: Set()
23/07/08 18:54:36 INFO ShufflePartitionsUtil: For shuffle(879), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:36 INFO DAGScheduler: Registering RDD 4967 (csv at <unknown>:0) as in
put to shuffle 884
23/07/08 18:54:36 INFO DAGScheduler: Got map stage job 1623 (csv at <unknown>:0) wit
h 1 output partitions
```

```
23/07/08 18:54:36 INFO DAGScheduler: Final stage: ShuffleMapStage 3637 (csv at <unkn
own>:0)
23/07/08 18:54:36 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
36)
23/07/08 18:54:36 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:36 INFO DAGScheduler: Submitting ShuffleMapStage 3637 (MapPartitionsR
DD[4967] at csv at <unknown>:0), which has no missing parents
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1981 stored as values in memory
 (estimated size 16.5 KiB, free 329.5 MiB)
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1981_piece0 stored as bytes in m
emory (estimated size 7.7 KiB, free 329.5 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Added broadcast_1981_piece0 in memory on 17
2.30.115.138:43839 (size: 7.7 KiB, free: 363.6 MiB)
23/07/08 18:54:36 INFO SparkContext: Created broadcast 1981 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:36 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
3637 (MapPartitionsRDD[4967] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0))
23/07/08 18:54:36 INFO TaskSchedulerImpl: Adding task set 3637.0 with 1 tasks resour
ce profile 0
23/07/08 18:54:36 INFO BlockManagerInfo: Removed broadcast_1977_piece0 on 172.30.11
5.138:43839 in memory (size: 4.2 KiB, free: 363.6 MiB)
23/07/08 18:54:36 INFO TaskSetManager: Starting task 0.0 in stage 3637.0 (TID 2281)
 (172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7368 bytes)
23/07/08 18:54:36 INFO Executor: Running task 0.0 in stage 3637.0 (TID 2281)
23/07/08 18:54:36 INFO BlockManagerInfo: Removed broadcast_1978_piece0 on 172.30.11
5.138:43839 in memory (size: 4.2 KiB, free: 363.6 MiB)
23/07/08 18:54:36 INFO ShufflePartitionsUtil: For shuffle(883), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Getting 2 (894.6 KiB) non-empty
 blocks including 2 (894.6 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:36 INFO SparkContext: Starting job: $anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266
23/07/08 18:54:36 INFO DAGScheduler: Got job 1624 ($anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266) with 1 output partitions
23/07/08 18:54:36 INFO DAGScheduler: Final stage: ResultStage 3639 ($anonfun$withThr
eadLocalCaptured$1 at FutureTask.java:266)
23/07/08 18:54:36 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
38)
23/07/08 18:54:36 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:36 INFO DAGScheduler: Submitting ResultStage 3639 (MapPartitionsRDD[4
969] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:266), which has no mis
sing parents
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1982 stored as values in memory
 (estimated size 8.2 KiB, free 329.5 MiB)
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1982_piece0 stored as bytes in m
emory (estimated size 4.2 KiB, free 329.5 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Added broadcast_1982_piece0 in memory on 17
2.30.115.138:43839 (size: 4.2 KiB, free: 363.6 MiB)
23/07/08 18:54:36 INFO SparkContext: Created broadcast 1982 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:36 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 363
9 (MapPartitionsRDD[4969] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:2
66) (first 15 tasks are for partitions Vector(0))
```

```
23/07/08 18:54:36 INFO TaskSchedulerImpl: Adding task set 3639.0 with 1 tasks resour
ce profile 0
23/07/08 18:54:36 INFO TaskSetManager: Starting task 0.0 in stage 3639.0 (TID 2282)
 (172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7379 bytes)
23/07/08 18:54:36 INFO Executor: Running task 0.0 in stage 3639.0 (TID 2282)
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Getting 2 (324.5 KiB) non-empty
 blocks including 2 (324.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:36 INFO Executor: Finished task 0.0 in stage 3639.0 (TID 2282). 27427
7 bytes result sent to driver
23/07/08 18:54:36 INFO TaskSetManager: Finished task 0.0 in stage 3639.0 (TID 2282)
 in 13 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:54:36 INFO TaskSchedulerImpl: Removed TaskSet 3639.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:36 INFO DAGScheduler: ResultStage 3639 ($anonfun$withThreadLocalCaptu
red$1 at FutureTask.java:266) finished in 0.016 s
23/07/08 18:54:36 INFO DAGScheduler: Job 1624 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:54:36 INFO TaskSchedulerImpl: Killing all running tasks in stage 3639: S
tage finished
23/07/08 18:54:36 INFO DAGScheduler: Job 1624 finished: $anonfun$withThreadLocalCapt
ured$1 at FutureTask.java:266, took 0.017649 s
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1983 stored as values in memory
 (estimated size 2.5 MiB, free 327.0 MiB)
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1983_piece0 stored as bytes in m
emory (estimated size 405.3 KiB, free 326.6 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Added broadcast_1983_piece0 in memory on 17
2.30.115.138:43839 (size: 405.3 KiB, free: 363.2 MiB)
23/07/08 18:54:36 INFO SparkContext: Created broadcast 1983 from $anonfun$withThread
LocalCaptured$1 at FutureTask.java:266
23/07/08 18:54:36 INFO Executor: Finished task 0.0 in stage 3637.0 (TID 2281). 4301
 bytes result sent to driver
23/07/08 18:54:36 INFO TaskSetManager: Finished task 0.0 in stage 3637.0 (TID 2281)
 in 63 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:54:36 INFO TaskSchedulerImpl: Removed TaskSet 3637.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:36 INFO DAGScheduler: ShuffleMapStage 3637 (csv at <unknown>:0) finis
hed in 0.080 s
23/07/08 18:54:36 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:36 INFO DAGScheduler: running: Set()
23/07/08 18:54:36 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:36 INFO DAGScheduler: failed: Set()
23/07/08 18:54:36 INFO ShufflePartitionsUtil: For shuffle(884, 880), advisory target
size: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:36 INFO DAGScheduler: Registering RDD 4976 (csv at <unknown>:0) as in
put to shuffle 885
23/07/08 18:54:36 INFO DAGScheduler: Got map stage job 1625 (csv at <unknown>:0) wit
h 1 output partitions
23/07/08 18:54:36 INFO DAGScheduler: Final stage: ShuffleMapStage 3643 (csv at <unkn
own>:0)
23/07/08 18:54:36 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
42, ShuffleMapStage 3641)
23/07/08 18:54:36 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:36 INFO DAGScheduler: Submitting ShuffleMapStage 3643 (MapPartitionsR
DD[4976] at csv at <unknown>:0), which has no missing parents
```

```
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1984 stored as values in memory
 (estimated size 77.6 KiB, free 326.5 MiB)
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1984_piece0 stored as bytes in m
emory (estimated size 34.3 KiB, free 326.5 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Added broadcast_1984_piece0 in memory on 17
2.30.115.138:43839 (size: 34.3 KiB, free: 363.1 MiB)
23/07/08 18:54:36 INFO SparkContext: Created broadcast 1984 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:36 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
3643 (MapPartitionsRDD[4976] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0))
23/07/08 18:54:36 INFO TaskSchedulerImpl: Adding task set 3643.0 with 1 tasks resour
ce profile 0
23/07/08 18:54:36 INFO TaskSetManager: Starting task 0.0 in stage 3643.0 (TID 2283)
 (172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7634 bytes)
23/07/08 18:54:36 INFO Executor: Running task 0.0 in stage 3643.0 (TID 2283)
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Getting 1 (825.8 KiB) non-empty
 blocks including 1 (825.8 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Getting 2 (2.7 KiB) non-empty bl
ocks including 2 (2.7 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-
local and 0 (0.0 B) remote blocks
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:36 INFO Executor: Finished task 0.0 in stage 3643.0 (TID 2283). 9924
 bytes result sent to driver
23/07/08 18:54:36 INFO TaskSetManager: Finished task 0.0 in stage 3643.0 (TID 2283)
 in 73 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:54:36 INFO TaskSchedulerImpl: Removed TaskSet 3643.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:36 INFO DAGScheduler: ShuffleMapStage 3643 (csv at <unknown>:0) finis
hed in 0.079 s
23/07/08 18:54:36 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:36 INFO DAGScheduler: running: Set()
23/07/08 18:54:36 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:36 INFO DAGScheduler: failed: Set()
23/07/08 18:54:36 INFO ShufflePartitionsUtil: For shuffle(885, 881), advisory target
size: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:36 INFO DAGScheduler: Registering RDD 4983 (csv at <unknown>:0) as in
put to shuffle 886
23/07/08 18:54:36 INFO DAGScheduler: Got map stage job 1626 (csv at <unknown>:0) wit
h 1 output partitions
23/07/08 18:54:36 INFO DAGScheduler: Final stage: ShuffleMapStage 3649 (csv at <unkn
own>:0)
23/07/08 18:54:36 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
47, ShuffleMapStage 3648)
23/07/08 18:54:36 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:36 INFO DAGScheduler: Submitting ShuffleMapStage 3649 (MapPartitionsR
DD[4983] at csv at <unknown>:0), which has no missing parents
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1985 stored as values in memory
 (estimated size 141.9 KiB, free 326.4 MiB)
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1985_piece0 stored as bytes in m
emory (estimated size 59.0 KiB, free 326.3 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Added broadcast_1985_piece0 in memory on 17
2.30.115.138:43839 (size: 59.0 KiB, free: 363.1 MiB)
23/07/08 18:54:36 INFO SparkContext: Created broadcast 1985 from broadcast at DAGSch
```

```
eduler.scala:1535
23/07/08 18:54:36 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
3649 (MapPartitionsRDD[4983] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0))
23/07/08 18:54:36 INFO TaskSchedulerImpl: Adding task set 3649.0 with 1 tasks resour
ce profile 0
23/07/08 18:54:36 INFO TaskSetManager: Starting task 0.0 in stage 3649.0 (TID 2284)
 (172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7634 bytes)
23/07/08 18:54:36 INFO Executor: Running task 0.0 in stage 3649.0 (TID 2284)
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Getting 1 (958.5 KiB) non-empty
 blocks including 1 (958.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Getting 2 (274.0 B) non-empty bl
ocks including 2 (274.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-
local and 0 (0.0 B) remote blocks
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:36 INFO Executor: Finished task 0.0 in stage 3649.0 (TID 2284). 17074
bytes result sent to driver
23/07/08 18:54:36 INFO TaskSetManager: Finished task 0.0 in stage 3649.0 (TID 2284)
 in 177 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:54:36 INFO TaskSchedulerImpl: Removed TaskSet 3649.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:36 INFO DAGScheduler: ShuffleMapStage 3649 (csv at <unknown>:0) finis
hed in 0.184 s
23/07/08 18:54:36 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:36 INFO DAGScheduler: running: Set()
23/07/08 18:54:36 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:36 INFO DAGScheduler: failed: Set()
23/07/08 18:54:36 INFO ShufflePartitionsUtil: For shuffle(886), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:36 INFO HashAggregateExec: spark.sql.codegen.aggregate.map.twolevel.e
nabled is set to true, but current version of codegened fast hashmap does not suppor
t this aggregate.
23/07/08 18:54:36 INFO SparkContext: Starting job: csv at <unknown>:0
23/07/08 18:54:36 INFO DAGScheduler: Got job 1627 (csv at <unknown>:0) with 2 output
partitions
23/07/08 18:54:36 INFO DAGScheduler: Final stage: ResultStage 3656 (csv at <unknown
>:0)
23/07/08 18:54:36 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
55)
23/07/08 18:54:36 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:36 INFO DAGScheduler: Submitting ResultStage 3656 (MapPartitionsRDD[4
988] at csv at <unknown>:0), which has no missing parents
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1986 stored as values in memory
 (estimated size 132.2 KiB, free 326.2 MiB)
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1986_piece0 stored as bytes in m
emory (estimated size 44.9 KiB, free 326.1 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Added broadcast_1986_piece0 in memory on 17
2.30.115.138:43839 (size: 44.9 KiB, free: 363.0 MiB)
23/07/08 18:54:36 INFO SparkContext: Created broadcast 1986 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:36 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 365
6 (MapPartitionsRDD[4988] at csv at <unknown>:0) (first 15 tasks are for partitions
 Vector(0, 1))
23/07/08 18:54:36 INFO TaskSchedulerImpl: Adding task set 3656.0 with 2 tasks resour
```

```
ce profile 0
23/07/08 18:54:36 INFO TaskSetManager: Starting task 0.0 in stage 3656.0 (TID 2285)
 (172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7363 bytes)
23/07/08 18:54:36 INFO TaskSetManager: Starting task 1.0 in stage 3656.0 (TID 2286)
 (172.30.115.138, executor driver, partition 1, NODE_LOCAL, 7363 bytes)
23/07/08 18:54:36 INFO Executor: Running task 0.0 in stage 3656.0 (TID 2285)
23/07/08 18:54:36 INFO Executor: Running task 1.0 in stage 3656.0 (TID 2286)
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Getting 1 (1378.9 KiB) non-empty
blocks including 1 (1378.9 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Getting 1 (1033.8 KiB) non-empty
blocks including 1 (1033.8 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:36 INFO MemoryStore: Block taskresult_2285 stored as bytes in memory
 (estimated size 1275.0 KiB, free 324.9 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Removed broadcast_1984_piece0 on 172.30.11
5.138:43839 in memory (size: 34.3 KiB, free: 363.1 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Added taskresult_2285 in memory on 172.30.1
15.138:43839 (size: 1275.0 KiB, free: 361.8 MiB)
23/07/08 18:54:36 INFO Executor: Finished task 0.0 in stage 3656.0 (TID 2285). 13056
27 bytes result sent via BlockManager)
23/07/08 18:54:36 INFO BlockManagerInfo: Removed broadcast_1982_piece0 on 172.30.11
5.138:43839 in memory (size: 4.2 KiB, free: 361.8 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Removed broadcast_1981_piece0 on 172.30.11
5.138:43839 in memory (size: 7.7 KiB, free: 361.8 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Removed broadcast_1985_piece0 on 172.30.11
5.138:43839 in memory (size: 59.0 KiB, free: 361.9 MiB)
23/07/08 18:54:36 INFO MemoryStore: Block taskresult_2286 stored as bytes in memory
 (estimated size 1683.6 KiB, free 323.6 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Added taskresult_2286 in memory on 172.30.1
15.138:43839 (size: 1683.6 KiB, free: 360.2 MiB)
23/07/08 18:54:36 INFO Executor: Finished task 1.0 in stage 3656.0 (TID 2286). 17240
50 bytes result sent via BlockManager)
23/07/08 18:54:36 INFO BlockManagerInfo: Removed taskresult_2285 on 172.30.115.138:4
3839 in memory (size: 1275.0 KiB, free: 361.5 MiB)
23/07/08 18:54:36 INFO TaskSetManager: Finished task 0.0 in stage 3656.0 (TID 2285)
 in 80 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:36 INFO TaskSetManager: Finished task 1.0 in stage 3656.0 (TID 2286)
 in 94 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:36 INFO TaskSchedulerImpl: Removed TaskSet 3656.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:36 INFO BlockManagerInfo: Removed taskresult_2286 on 172.30.115.138:4
3839 in memory (size: 1683.6 KiB, free: 363.1 MiB)
23/07/08 18:54:36 INFO DAGScheduler: ResultStage 3656 (csv at <unknown>:0) finished
 in 0.101 s
23/07/08 18:54:36 INFO DAGScheduler: Job 1627 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:54:36 INFO TaskSchedulerImpl: Killing all running tasks in stage 3656: S
tage finished
23/07/08 18:54:36 INFO DAGScheduler: Job 1627 finished: csv at <unknown>:0, took 0.1
02469 s
23/07/08 18:54:36 INFO DAGScheduler: Registering RDD 4989 (csv at <unknown>:0) as in
put to shuffle 887
23/07/08 18:54:36 INFO DAGScheduler: Got map stage job 1628 (csv at <unknown>:0) wit
```

```
h 2 output partitions
23/07/08 18:54:36 INFO DAGScheduler: Final stage: ShuffleMapStage 3663 (csv at <unkn
own>:0)
23/07/08 18:54:36 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
62)
23/07/08 18:54:36 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:36 INFO DAGScheduler: Submitting ShuffleMapStage 3663 (MapPartitionsR
DD[4989] at csv at <unknown>:0), which has no missing parents
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1987 stored as values in memory
 (estimated size 145.1 KiB, free 326.3 MiB)
23/07/08 18:54:36 INFO MemoryStore: Block broadcast_1987_piece0 stored as bytes in m
emory (estimated size 47.4 KiB, free 326.3 MiB)
23/07/08 18:54:36 INFO BlockManagerInfo: Added broadcast_1987_piece0 in memory on 17
2.30.115.138:43839 (size: 47.4 KiB, free: 363.1 MiB)
23/07/08 18:54:36 INFO SparkContext: Created broadcast 1987 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:36 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
3663 (MapPartitionsRDD[4989] at csv at <unknown>:0) (first 15 tasks are for partitio
ns Vector(0, 1))
23/07/08 18:54:36 INFO TaskSchedulerImpl: Adding task set 3663.0 with 2 tasks resour
ce profile 0
23/07/08 18:54:36 INFO TaskSetManager: Starting task 0.0 in stage 3663.0 (TID 2287)
 (172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7352 bytes)
23/07/08 18:54:36 INFO TaskSetManager: Starting task 1.0 in stage 3663.0 (TID 2288)
 (172.30.115.138, executor driver, partition 1, NODE_LOCAL, 7352 bytes)
23/07/08 18:54:36 INFO Executor: Running task 0.0 in stage 3663.0 (TID 2287)
23/07/08 18:54:36 INFO Executor: Running task 1.0 in stage 3663.0 (TID 2288)
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Getting 1 (1378.9 KiB) non-empty
blocks including 1 (1378.9 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Getting 1 (1033.8 KiB) non-empty
blocks including 1 (1033.8 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:36 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:37 INFO Executor: Finished task 0.0 in stage 3663.0 (TID 2287). 18698
bytes result sent to driver
23/07/08 18:54:37 INFO TaskSetManager: Finished task 0.0 in stage 3663.0 (TID 2287)
 in 96 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:37 INFO Executor: Finished task 1.0 in stage 3663.0 (TID 2288). 18698
bytes result sent to driver
23/07/08 18:54:37 INFO TaskSetManager: Finished task 1.0 in stage 3663.0 (TID 2288)
 in 108 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:37 INFO TaskSchedulerImpl: Removed TaskSet 3663.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:37 INFO DAGScheduler: ShuffleMapStage 3663 (csv at <unknown>:0) finis
hed in 0.117 s
23/07/08 18:54:37 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:37 INFO DAGScheduler: running: Set()
23/07/08 18:54:37 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:37 INFO DAGScheduler: failed: Set()
23/07/08 18:54:37 INFO ShufflePartitionsUtil: For shuffle(887), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:37 INFO FileOutputCommitter: File Output Committer Algorithm version
 is 1
23/07/08 18:54:37 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _tempor
```

```
ary folders under output directory:false, ignore cleanup failures: false
23/07/08 18:54:37 INFO SQLHadoopMapReduceCommitProtocol: Using output committer clas
s org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
23/07/08 18:54:37 INFO SparkContext: Starting job: csv at <unknown>:0
23/07/08 18:54:37 INFO DAGScheduler: Got job 1629 (csv at <unknown>:0) with 2 output
partitions
23/07/08 18:54:37 INFO DAGScheduler: Final stage: ResultStage 3671 (csv at <unknown
>:0)
23/07/08 18:54:37 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
70)
23/07/08 18:54:37 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:37 INFO DAGScheduler: Submitting ResultStage 3671 (MapPartitionsRDD[4
992] at csv at <unknown>:0), which has no missing parents
23/07/08 18:54:37 INFO MemoryStore: Block broadcast_1988 stored as values in memory
 (estimated size 428.0 KiB, free 325.9 MiB)
23/07/08 18:54:37 INFO MemoryStore: Block broadcast_1988_piece0 stored as bytes in m
emory (estimated size 156.1 KiB, free 325.7 MiB)
23/07/08 18:54:37 INFO BlockManagerInfo: Added broadcast_1988_piece0 in memory on 17
2.30.115.138:43839 (size: 156.1 KiB, free: 362.9 MiB)
23/07/08 18:54:37 INFO SparkContext: Created broadcast 1988 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:37 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 367
1 (MapPartitionsRDD[4992] at csv at <unknown>:0) (first 15 tasks are for partitions
 Vector(0, 1))
23/07/08 18:54:37 INFO TaskSchedulerImpl: Adding task set 3671.0 with 2 tasks resour
ce profile 0
23/07/08 18:54:37 INFO TaskSetManager: Starting task 0.0 in stage 3671.0 (TID 2289)
 (172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7363 bytes)
23/07/08 18:54:37 INFO TaskSetManager: Starting task 1.0 in stage 3671.0 (TID 2290)
 (172.30.115.138, executor driver, partition 1, NODE_LOCAL, 7363 bytes)
23/07/08 18:54:37 INFO Executor: Running task 1.0 in stage 3671.0 (TID 2290)
23/07/08 18:54:37 INFO Executor: Running task 0.0 in stage 3671.0 (TID 2289)
23/07/08 18:54:37 INFO ShuffleBlockFetcherIterator: Getting 2 (1027.1 KiB) non-empty
blocks including 2 (1027.1 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:37 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:37 INFO ShuffleBlockFetcherIterator: Getting 2 (1618.7 KiB) non-empty
blocks including 2 (1618.7 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:37 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:37 INFO FileOutputCommitter: File Output Committer Algorithm version
 is 1
23/07/08 18:54:37 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _tempor
ary folders under output directory:false, ignore cleanup failures: false
23/07/08 18:54:37 INFO SQLHadoopMapReduceCommitProtocol: Using output committer clas
s org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
23/07/08 18:54:37 INFO FileOutputCommitter: File Output Committer Algorithm version
 is 1
23/07/08 18:54:37 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _tempor
ary folders under output directory:false, ignore cleanup failures: false
23/07/08 18:54:37 INFO SQLHadoopMapReduceCommitProtocol: Using output committer clas
s org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
23/07/08 18:54:37 INFO FileOutputCommitter: Saved output of task 'attempt_2023070818
54376964530498379514657_3671_m_000000_2289' to hdfs://127.0.0.1:9000/datos/salida/sa
lida5/_temporary/0/task_202307081854376964530498379514657_3671_m_000000
23/07/08 18:54:37 INFO SparkHadoopMapRedUtil: attempt_202307081854376964530498379514
```

```
657_3671_m_000000_2289: Committed. Elapsed time: 8 ms.
23/07/08 18:54:37 INFO Executor: Finished task 0.0 in stage 3671.0 (TID 2289). 20785
bytes result sent to driver
23/07/08 18:54:37 INFO TaskSetManager: Finished task 0.0 in stage 3671.0 (TID 2289)
 in 526 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:37 INFO FileOutputCommitter: Saved output of task 'attempt_2023070818
5437696453049837951 4657_3671_m_000001_2290' to hdfs://127.0.0.1:9000/datos/salida/sa
lida5/_temporary/0/task_20230708185437696453049837951 4657_3671_m_000001
23/07/08 18:54:37 INFO SparkHadoopMapRedUtil: attempt_20230708185437696453049837951 4
657_3671_m_000001_2290: Committed. Elapsed time: 7 ms.
23/07/08 18:54:37 INFO Executor: Finished task 1.0 in stage 3671.0 (TID 2290). 20785
bytes result sent to driver
23/07/08 18:54:37 INFO TaskSetManager: Finished task 1.0 in stage 3671.0 (TID 2290)
 in 552 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:37 INFO TaskSchedulerImpl: Removed TaskSet 3671.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:37 INFO DAGScheduler: ResultStage 3671 (csv at <unknown>:0) finished
 in 0.579 s
23/07/08 18:54:37 INFO DAGScheduler: Job 1629 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:54:37 INFO TaskSchedulerImpl: Killing all running tasks in stage 3671: S
tage finished
23/07/08 18:54:37 INFO DAGScheduler: Job 1629 finished: csv at <unknown>:0, took 0.5
81141 s
23/07/08 18:54:37 INFO FileFormatWriter: Start to commit write Job 5c78f6d5-0361-4c5
a-ad4a-eda8aa2cf587.
23/07/08 18:54:37 INFO FileFormatWriter: Write Job 5c78f6d5-0361-4c5a-ad4a-eda8aa2cf
587 committed. Elapsed time: 39 ms.
23/07/08 18:54:37 INFO FileFormatWriter: Finished processing stats for write job 5c7
8f6d5-0361-4c5a-ad4a-eda8aa2cf587.
23/07/08 18:54:37 INFO BlockManagerInfo: Removed broadcast_1988_piece0 on 172.30.11
5.138:43839 in memory (size: 156.1 KiB, free: 363.1 MiB)
23/07/08 18:54:37 INFO BlockManagerInfo: Removed broadcast_1987_piece0 on 172.30.11
5.138:43839 in memory (size: 47.4 KiB, free: 363.1 MiB)
23/07/08 18:54:37 INFO DAGScheduler: Registering RDD 4994 (showString at <unknown>:
0) as input to shuffle 888
23/07/08 18:54:37 INFO DAGScheduler: Got map stage job 1630 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:54:37 INFO DAGScheduler: Final stage: ShuffleMapStage 3672 (showString a
t <unknown>:0)
23/07/08 18:54:37 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:54:37 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:37 INFO DAGScheduler: Submitting ShuffleMapStage 3672 (MapPartitionsR
DD[4994] at showString at <unknown>:0), which has no missing parents
23/07/08 18:54:37 INFO MemoryStore: Block broadcast_1989 stored as values in memory
 (estimated size 20.6 KiB, free 326.5 MiB)
23/07/08 18:54:37 INFO MemoryStore: Block broadcast_1989_piece0 stored as bytes in m
emory (estimated size 10.6 KiB, free 326.4 MiB)
23/07/08 18:54:37 INFO BlockManagerInfo: Added broadcast_1989_piece0 in memory on 17
2.30.115.138:43839 (size: 10.6 KiB, free: 363.1 MiB)
23/07/08 18:54:37 INFO SparkContext: Created broadcast 1989 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:37 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
3672 (MapPartitionsRDD[4994] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:54:37 INFO TaskSchedulerImpl: Adding task set 3672.0 with 2 tasks resour
```

```
ce profile 0
23/07/08 18:54:37 INFO TaskSetManager: Starting task 0.0 in stage 3672.0 (TID 2291)
 (172.30.115.138, executor driver, partition 0, ANY, 7408 bytes)
23/07/08 18:54:37 INFO TaskSetManager: Starting task 1.0 in stage 3672.0 (TID 2292)
 (172.30.115.138, executor driver, partition 1, ANY, 7408 bytes)
23/07/08 18:54:37 INFO Executor: Running task 0.0 in stage 3672.0 (TID 2291)
23/07/08 18:54:37 INFO Executor: Running task 1.0 in stage 3672.0 (TID 2292)
23/07/08 18:54:37 INFO DAGScheduler: Registering RDD 4996 (showString at <unknown>:
0) as input to shuffle 889
23/07/08 18:54:37 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Producto.
csv:0+8440
23/07/08 18:54:37 INFO DAGScheduler: Got map stage job 1631 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:54:37 INFO DAGScheduler: Final stage: ShuffleMapStage 3673 (showString a
t <unknown>:0)
23/07/08 18:54:37 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:54:37 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:37 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Producto.
csv:8440+8441
23/07/08 18:54:37 INFO DAGScheduler: Submitting ShuffleMapStage 3673 (MapPartitionsR
DD[4996] at showString at <unknown>:0), which has no missing parents
23/07/08 18:54:37 INFO MemoryStore: Block broadcast_1990 stored as values in memory
 (estimated size 24.5 KiB, free 326.4 MiB)
23/07/08 18:54:37 INFO MemoryStore: Block broadcast_1990_piece0 stored as bytes in m
emory (estimated size 11.6 KiB, free 326.4 MiB)
23/07/08 18:54:37 INFO BlockManagerInfo: Added broadcast_1990_piece0 in memory on 17
2.30.115.138:43839 (size: 11.6 KiB, free: 363.1 MiB)
23/07/08 18:54:37 INFO SparkContext: Created broadcast 1990 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:37 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
3673 (MapPartitionsRDD[4996] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:54:37 INFO TaskSchedulerImpl: Adding task set 3673.0 with 2 tasks resour
ce profile 0
23/07/08 18:54:37 INFO DAGScheduler: Registering RDD 4998 (showString at <unknown>:
0) as input to shuffle 890
23/07/08 18:54:37 INFO DAGScheduler: Got map stage job 1632 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:54:37 INFO DAGScheduler: Final stage: ShuffleMapStage 3674 (showString a
t <unknown>:0)
23/07/08 18:54:37 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:54:37 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:37 INFO DAGScheduler: Submitting ShuffleMapStage 3674 (MapPartitionsR
DD[4998] at showString at <unknown>:0), which has no missing parents
23/07/08 18:54:37 INFO TaskSetManager: Starting task 0.0 in stage 3673.0 (TID 2293)
 (172.30.115.138, executor driver, partition 0, ANY, 7406 bytes)
23/07/08 18:54:37 INFO TaskSetManager: Starting task 1.0 in stage 3673.0 (TID 2294)
 (172.30.115.138, executor driver, partition 1, ANY, 7406 bytes)
23/07/08 18:54:37 INFO Executor: Running task 0.0 in stage 3673.0 (TID 2293)
23/07/08 18:54:37 INFO Executor: Running task 1.0 in stage 3673.0 (TID 2294)
23/07/08 18:54:37 INFO MemoryStore: Block broadcast_1991 stored as values in memory
 (estimated size 21.1 KiB, free 326.4 MiB)
23/07/08 18:54:37 INFO MemoryStore: Block broadcast_1991_piece0 stored as bytes in m
emory (estimated size 10.6 KiB, free 326.4 MiB)
23/07/08 18:54:37 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:0+1310749
```

```
23/07/08 18:54:37 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Ventas.cs
v:1310749+1310750
23/07/08 18:54:37 INFO BlockManagerInfo: Added broadcast_1991_piece0 in memory on 17
2.30.115.138:43839 (size: 10.6 KiB, free: 363.1 MiB)
23/07/08 18:54:37 INFO SparkContext: Created broadcast 1991 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:37 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
3674 (MapPartitionsRDD[4998] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:54:37 INFO TaskSchedulerImpl: Adding task set 3674.0 with 2 tasks resour
ce profile 0
23/07/08 18:54:37 INFO DAGScheduler: Registering RDD 5000 (showString at <unknown>:
0) as input to shuffle 891
23/07/08 18:54:37 INFO DAGScheduler: Got map stage job 1633 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:54:37 INFO DAGScheduler: Final stage: ShuffleMapStage 3675 (showString a
t <unknown>:0)
23/07/08 18:54:37 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:54:37 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:37 INFO DAGScheduler: Submitting ShuffleMapStage 3675 (MapPartitionsR
DD[5000] at showString at <unknown>:0), which has no missing parents
23/07/08 18:54:37 INFO MemoryStore: Block broadcast_1992 stored as values in memory
 (estimated size 20.1 KiB, free 326.4 MiB)
23/07/08 18:54:37 INFO MemoryStore: Block broadcast_1992_piece0 stored as bytes in m
emory (estimated size 10.4 KiB, free 326.3 MiB)
23/07/08 18:54:37 INFO BlockManagerInfo: Added broadcast_1992_piece0 in memory on 17
2.30.115.138:43839 (size: 10.4 KiB, free: 363.1 MiB)
23/07/08 18:54:37 INFO SparkContext: Created broadcast 1992 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:37 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
3675 (MapPartitionsRDD[5000] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:54:37 INFO TaskSchedulerImpl: Adding task set 3675.0 with 2 tasks resour
ce profile 0
23/07/08 18:54:37 INFO DAGScheduler: Registering RDD 5002 (showString at <unknown>:
0) as input to shuffle 892
23/07/08 18:54:37 INFO DAGScheduler: Got map stage job 1634 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:54:37 INFO DAGScheduler: Final stage: ShuffleMapStage 3676 (showString a
t <unknown>:0)
23/07/08 18:54:37 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:54:37 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:37 INFO DAGScheduler: Submitting ShuffleMapStage 3676 (MapPartitionsR
DD[5002] at showString at <unknown>:0), which has no missing parents
23/07/08 18:54:37 INFO MemoryStore: Block broadcast_1993 stored as values in memory
 (estimated size 21.5 KiB, free 326.3 MiB)
23/07/08 18:54:37 INFO MemoryStore: Block broadcast_1993_piece0 stored as bytes in m
emory (estimated size 10.7 KiB, free 326.3 MiB)
23/07/08 18:54:37 INFO BlockManagerInfo: Added broadcast_1993_piece0 in memory on 17
2.30.115.138:43839 (size: 10.7 KiB, free: 363.1 MiB)
23/07/08 18:54:37 INFO TaskSetManager: Starting task 0.0 in stage 3674.0 (TID 2295)
 (172.30.115.138, executor driver, partition 0, ANY, 7408 bytes)
23/07/08 18:54:37 INFO TaskSetManager: Starting task 1.0 in stage 3674.0 (TID 2296)
 (172.30.115.138, executor driver, partition 1, ANY, 7408 bytes)
23/07/08 18:54:37 INFO TaskSetManager: Starting task 0.0 in stage 3675.0 (TID 2297)
 (172.30.115.138, executor driver, partition 0, ANY, 7412 bytes)
```

```
23/07/08 18:54:37 INFO TaskSetManager: Starting task 1.0 in stage 3675.0 (TID 2298)
 (172.30.115.138, executor driver, partition 1, ANY, 7412 bytes)
23/07/08 18:54:37 INFO Executor: Running task 1.0 in stage 3675.0 (TID 2298)
23/07/08 18:54:37 INFO Executor: Running task 0.0 in stage 3675.0 (TID 2297)
23/07/08 18:54:37 INFO Executor: Running task 0.0 in stage 3674.0 (TID 2295)
23/07/08 18:54:37 INFO SparkContext: Created broadcast 1993 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:37 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/CanalDeVe
nta.csv:0+29
23/07/08 18:54:37 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Sucursal.
csv:0+1266
23/07/08 18:54:37 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/CanalDeVe
nta.csv:29+29
23/07/08 18:54:37 INFO Executor: Running task 1.0 in stage 3674.0 (TID 2296)
23/07/08 18:54:37 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Sucursal.
csv:1266+1266
23/07/08 18:54:37 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:54:37 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
3676 (MapPartitionsRDD[5002] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:54:37 INFO TaskSchedulerImpl: Adding task set 3676.0 with 2 tasks resour
ce profile 0
23/07/08 18:54:37 INFO DAGScheduler: Registering RDD 5004 (showString at <unknown>:
0) as input to shuffle 893
23/07/08 18:54:37 INFO DAGScheduler: Got map stage job 1635 (showString at <unknown
>:0) with 2 output partitions
23/07/08 18:54:37 INFO DAGScheduler: Final stage: ShuffleMapStage 3677 (showString a
t <unknown>:0)
23/07/08 18:54:37 INFO DAGScheduler: Parents of final stage: List()
23/07/08 18:54:37 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:37 INFO DAGScheduler: Submitting ShuffleMapStage 3677 (MapPartitionsR
DD[5004] at showString at <unknown>:0), which has no missing parents
23/07/08 18:54:37 INFO MemoryStore: Block broadcast_1994 stored as values in memory
 (estimated size 22.8 KiB, free 326.3 MiB)
23/07/08 18:54:37 INFO MemoryStore: Block broadcast_1994_piece0 stored as bytes in m
emory (estimated size 10.9 KiB, free 326.3 MiB)
23/07/08 18:54:37 INFO TaskSetManager: Starting task 0.0 in stage 3676.0 (TID 2299)
 (172.30.115.138, executor driver, partition 0, ANY, 7408 bytes)
23/07/08 18:54:37 INFO TaskSetManager: Starting task 1.0 in stage 3676.0 (TID 2300)
 (172.30.115.138, executor driver, partition 1, ANY, 7408 bytes)
23/07/08 18:54:37 INFO Executor: Running task 0.0 in stage 3676.0 (TID 2299)
23/07/08 18:54:37 INFO Executor: Running task 1.0 in stage 3676.0 (TID 2300)
23/07/08 18:54:37 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:54:37 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Empleado.
csv:8119+8119
23/07/08 18:54:37 INFO BlockManagerInfo: Added broadcast_1994_piece0 in memory on 17
2.30.115.138:43839 (size: 10.9 KiB, free: 363.1 MiB)
23/07/08 18:54:37 INFO SparkContext: Created broadcast 1994 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:37 INFO DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage
3677 (MapPartitionsRDD[5004] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0, 1))
23/07/08 18:54:37 INFO TaskSchedulerImpl: Adding task set 3677.0 with 2 tasks resour
ce profile 0
23/07/08 18:54:37 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/Empleado.
csv:0+8119
```

```
23/07/08 18:54:37 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:54:37 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:54:37 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:54:38 INFO PythonRunner: Times: total = 200, boot = -1727, init = 1926,
 finish = 1
23/07/08 18:54:38 INFO Executor: Finished task 0.0 in stage 3675.0 (TID 2297). 2481
 bytes result sent to driver
23/07/08 18:54:38 INFO TaskSetManager: Starting task 0.0 in stage 3677.0 (TID 2301)
 (172.30.115.138, executor driver, partition 0, ANY, 7411 bytes)
23/07/08 18:54:38 INFO TaskSetManager: Finished task 0.0 in stage 3675.0 (TID 2297)
 in 223 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:38 INFO Executor: Running task 0.0 in stage 3677.0 (TID 2301)
23/07/08 18:54:38 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:0+450275
23/07/08 18:54:38 INFO PythonRunner: Times: total = 193, boot = -1539, init = 1731,
 finish = 1
23/07/08 18:54:38 INFO PythonRunner: Times: total = 266, boot = -1805, init = 2067,
 finish = 4
23/07/08 18:54:38 INFO PythonRunner: Times: total = 241, boot = -1662, init = 1903,
 finish = 0
23/07/08 18:54:38 INFO LineRecordReader: Found UTF-8 BOM and skipped it
23/07/08 18:54:38 INFO PythonRunner: Times: total = 228, boot = -1643, init = 1871,
 finish = 0
23/07/08 18:54:38 INFO Executor: Finished task 1.0 in stage 3676.0 (TID 2300). 2481
 bytes result sent to driver
23/07/08 18:54:38 INFO TaskSetManager: Starting task 1.0 in stage 3677.0 (TID 2302)
 (172.30.115.138, executor driver, partition 1, ANY, 7411 bytes)
23/07/08 18:54:38 INFO TaskSetManager: Finished task 1.0 in stage 3676.0 (TID 2300)
 in 265 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:38 INFO Executor: Running task 1.0 in stage 3677.0 (TID 2302)
23/07/08 18:54:38 INFO Executor: Finished task 1.0 in stage 3675.0 (TID 2298). 2481
 bytes result sent to driver
23/07/08 18:54:38 INFO HadoopRDD: Input split: hdfs://127.0.0.1:9000/datos/dim/DIMDA
TE.csv:450275+450275
23/07/08 18:54:38 INFO TaskSetManager: Finished task 1.0 in stage 3675.0 (TID 2298)
 in 325 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:38 INFO TaskSchedulerImpl: Removed TaskSet 3675.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:38 INFO PythonRunner: Times: total = 243, boot = -1557, init = 1800,
 finish = 0
23/07/08 18:54:38 INFO DAGScheduler: ShuffleMapStage 3675 (showString at <unknown>:
0) finished in 0.432 s
23/07/08 18:54:38 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:38 INFO DAGScheduler: running: Set(ShuffleMapStage 3672, ShuffleMapSt
age 3676, ShuffleMapStage 3673, ShuffleMapStage 3677, ShuffleMapStage 3674)
23/07/08 18:54:38 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:38 INFO DAGScheduler: failed: Set()
23/07/08 18:54:38 INFO Executor: Finished task 1.0 in stage 3674.0 (TID 2296). 2524
 bytes result sent to driver
23/07/08 18:54:38 INFO TaskSetManager: Finished task 1.0 in stage 3674.0 (TID 2296)
 in 427 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:38 INFO Executor: Finished task 1.0 in stage 3672.0 (TID 2292). 2524
 bytes result sent to driver
23/07/08 18:54:38 INFO TaskSetManager: Finished task 1.0 in stage 3672.0 (TID 2292)
 in 460 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:38 INFO PythonRunner: Times: total = 388, boot = -1610, init = 1998,
```

```
 finish = 0
23/07/08 18:54:38 INFO PythonRunner: Times: total = 353, boot = -1647, init = 1999,
 finish = 1
23/07/08 18:54:38 INFO Executor: Finished task 0.0 in stage 3674.0 (TID 2295). 2524
 bytes result sent to driver
23/07/08 18:54:38 INFO Executor: Finished task 0.0 in stage 3672.0 (TID 2291). 2524
 bytes result sent to driver
23/07/08 18:54:38 INFO Executor: Finished task 0.0 in stage 3676.0 (TID 2299). 2524
 bytes result sent to driver
23/07/08 18:54:38 INFO TaskSetManager: Finished task 0.0 in stage 3674.0 (TID 2295)
 in 463 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:38 INFO TaskSchedulerImpl: Removed TaskSet 3674.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:38 INFO DAGScheduler: ShuffleMapStage 3674 (showString at <unknown>:
0) finished in 0.481 s
23/07/08 18:54:38 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:38 INFO DAGScheduler: running: Set(ShuffleMapStage 3672, ShuffleMapSt
age 3676, ShuffleMapStage 3673, ShuffleMapStage 3677)
23/07/08 18:54:38 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:38 INFO DAGScheduler: failed: Set()
23/07/08 18:54:38 INFO TaskSetManager: Finished task 0.0 in stage 3672.0 (TID 2291)
 in 490 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:38 INFO TaskSchedulerImpl: Removed TaskSet 3672.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:38 INFO TaskSetManager: Finished task 0.0 in stage 3676.0 (TID 2299)
 in 456 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:38 INFO TaskSchedulerImpl: Removed TaskSet 3676.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:38 INFO DAGScheduler: ShuffleMapStage 3672 (showString at <unknown>:
0) finished in 0.497 s
23/07/08 18:54:38 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:38 INFO DAGScheduler: running: Set(ShuffleMapStage 3676, ShuffleMapSt
age 3673, ShuffleMapStage 3677)
23/07/08 18:54:38 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:38 INFO DAGScheduler: failed: Set()
23/07/08 18:54:38 INFO DAGScheduler: ShuffleMapStage 3676 (showString at <unknown>:
0) finished in 0.471 s
23/07/08 18:54:38 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:38 INFO DAGScheduler: running: Set(ShuffleMapStage 3673, ShuffleMapSt
age 3677)
23/07/08 18:54:38 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:38 INFO DAGScheduler: failed: Set()
23/07/08 18:54:38 INFO ShufflePartitionsUtil: For shuffle(888), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:38 INFO ShufflePartitionsUtil: For shuffle(892), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:38 INFO PythonRunner: Times: total = 232, boot = 42, init = 137, fini
sh = 53
23/07/08 18:54:38 INFO PythonRunner: Times: total = 406, boot = -1797, init = 1972,
 finish = 231
23/07/08 18:54:38 INFO SparkContext: Starting job: $anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266
23/07/08 18:54:38 INFO DAGScheduler: Got job 1636 ($anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266) with 1 output partitions
23/07/08 18:54:38 INFO DAGScheduler: Final stage: ResultStage 3679 ($anonfun$withThr
eadLocalCaptured$1 at FutureTask.java:266)
```

```
23/07/08 18:54:38 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
78)
23/07/08 18:54:38 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:38 INFO DAGScheduler: Submitting ResultStage 3679 (MapPartitionsRDD[5
006] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:266), which has no mis
sing parents
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_1995 stored as values in memory
 (estimated size 8.2 KiB, free 326.3 MiB)
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_1995_piece0 stored as bytes in m
emory (estimated size 4.2 KiB, free 326.3 MiB)
23/07/08 18:54:38 INFO BlockManagerInfo: Added broadcast_1995_piece0 in memory on 17
2.30.115.138:43839 (size: 4.2 KiB, free: 363.1 MiB)
23/07/08 18:54:38 INFO SparkContext: Created broadcast 1995 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:38 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 367
9 (MapPartitionsRDD[5006] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:2
66) (first 15 tasks are for partitions Vector(0))
23/07/08 18:54:38 INFO TaskSchedulerImpl: Adding task set 3679.0 with 1 tasks resour
ce profile 0
23/07/08 18:54:38 INFO TaskSetManager: Starting task 0.0 in stage 3679.0 (TID 2303)
 (172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7379 bytes)
23/07/08 18:54:38 INFO Executor: Running task 0.0 in stage 3679.0 (TID 2303)
23/07/08 18:54:38 INFO SparkContext: Starting job: $anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Getting 2 (20.5 KiB) non-empty b
locks including 2 (20.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:38 INFO DAGScheduler: Got job 1637 ($anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266) with 1 output partitions
23/07/08 18:54:38 INFO DAGScheduler: Final stage: ResultStage 3681 ($anonfun$withThr
eadLocalCaptured$1 at FutureTask.java:266)
23/07/08 18:54:38 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
80)
23/07/08 18:54:38 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:38 INFO DAGScheduler: Submitting ResultStage 3681 (MapPartitionsRDD[5
008] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:266), which has no mis
sing parents
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_1996 stored as values in memory
 (estimated size 8.2 KiB, free 326.3 MiB)
23/07/08 18:54:38 INFO Executor: Finished task 0.0 in stage 3679.0 (TID 2303). 6937
 bytes result sent to driver
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_1996_piece0 stored as bytes in m
emory (estimated size 4.2 KiB, free 326.3 MiB)
23/07/08 18:54:38 INFO BlockManagerInfo: Added broadcast_1996_piece0 in memory on 17
2.30.115.138:43839 (size: 4.2 KiB, free: 363.1 MiB)
23/07/08 18:54:38 INFO SparkContext: Created broadcast 1996 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:38 INFO TaskSetManager: Finished task 0.0 in stage 3679.0 (TID 2303)
 in 30 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:54:38 INFO TaskSchedulerImpl: Removed TaskSet 3679.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:38 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 368
1 (MapPartitionsRDD[5008] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:2
66) (first 15 tasks are for partitions Vector(0))
23/07/08 18:54:38 INFO TaskSchedulerImpl: Adding task set 3681.0 with 1 tasks resour
```

```
ce profile 0
23/07/08 18:54:38 INFO TaskSetManager: Starting task 0.0 in stage 3681.0 (TID 2304)
 (172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7379 bytes)
23/07/08 18:54:38 INFO DAGScheduler: ResultStage 3679 ($anonfun$withThreadLocalCaptu
red$1 at FutureTask.java:266) finished in 0.035 s
23/07/08 18:54:38 INFO DAGScheduler: Job 1636 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:54:38 INFO TaskSchedulerImpl: Killing all running tasks in stage 3679: S
tage finished
23/07/08 18:54:38 INFO Executor: Running task 0.0 in stage 3681.0 (TID 2304)
23/07/08 18:54:38 INFO DAGScheduler: Job 1636 finished: $anonfun$withThreadLocalCapt
ured$1 at FutureTask.java:266, took 0.036690 s
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Getting 2 (22.5 KiB) non-empty b
locks including 2 (22.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merge
d-local and 0 (0.0 B) remote blocks
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_1997 stored as values in memory
 (estimated size 2.0 MiB, free 324.2 MiB)
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_1997_piece0 stored as bytes in m
emory (estimated size 5.9 KiB, free 324.2 MiB)
23/07/08 18:54:38 INFO Executor: Finished task 0.0 in stage 3681.0 (TID 2304). 10564
bytes result sent to driver
23/07/08 18:54:38 INFO Executor: Finished task 0.0 in stage 3677.0 (TID 2301). 2524
 bytes result sent to driver
23/07/08 18:54:38 INFO BlockManagerInfo: Added broadcast_1997_piece0 in memory on 17
2.30.115.138:43839 (size: 5.9 KiB, free: 363.1 MiB)
23/07/08 18:54:38 INFO SparkContext: Created broadcast 1997 from $anonfun$withThread
LocalCaptured$1 at FutureTask.java:266
23/07/08 18:54:38 INFO TaskSetManager: Finished task 0.0 in stage 3677.0 (TID 2301)
 in 326 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:38 INFO TaskSetManager: Finished task 0.0 in stage 3681.0 (TID 2304)
 in 9 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:54:38 INFO TaskSchedulerImpl: Removed TaskSet 3681.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:38 INFO DAGScheduler: ResultStage 3681 ($anonfun$withThreadLocalCaptu
red$1 at FutureTask.java:266) finished in 0.037 s
23/07/08 18:54:38 INFO DAGScheduler: Job 1637 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:54:38 INFO TaskSchedulerImpl: Killing all running tasks in stage 3681: S
tage finished
23/07/08 18:54:38 INFO DAGScheduler: Job 1637 finished: $anonfun$withThreadLocalCapt
ured$1 at FutureTask.java:266, took 0.038838 s
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_1998 stored as values in memory
 (estimated size 2.0 MiB, free 322.2 MiB)
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_1998_piece0 stored as bytes in m
emory (estimated size 8.5 KiB, free 322.2 MiB)
23/07/08 18:54:38 INFO BlockManagerInfo: Added broadcast_1998_piece0 in memory on 17
2.30.115.138:43839 (size: 8.5 KiB, free: 363.1 MiB)
23/07/08 18:54:38 INFO SparkContext: Created broadcast 1998 from $anonfun$withThread
LocalCaptured$1 at FutureTask.java:266
23/07/08 18:54:38 INFO Executor: Finished task 0.0 in stage 3673.0 (TID 2293). 2524
 bytes result sent to driver
23/07/08 18:54:38 INFO TaskSetManager: Finished task 0.0 in stage 3673.0 (TID 2293)
 in 591 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:38 INFO PythonRunner: Times: total = 215, boot = 12, init = 139, fini
sh = 64
```

```
23/07/08 18:54:38 INFO PythonRunner: Times: total = 401, boot = -1810, init = 2026,
 finish = 185
23/07/08 18:54:38 INFO Executor: Finished task 1.0 in stage 3677.0 (TID 2302). 2524
 bytes result sent to driver
23/07/08 18:54:38 INFO TaskSetManager: Finished task 1.0 in stage 3677.0 (TID 2302)
 in 321 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:38 INFO TaskSchedulerImpl: Removed TaskSet 3677.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:38 INFO DAGScheduler: ShuffleMapStage 3677 (showString at <unknown>:
0) finished in 0.589 s
23/07/08 18:54:38 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:38 INFO DAGScheduler: running: Set(ShuffleMapStage 3673)
23/07/08 18:54:38 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:38 INFO DAGScheduler: failed: Set()
23/07/08 18:54:38 INFO Executor: Finished task 1.0 in stage 3673.0 (TID 2294). 2524
 bytes result sent to driver
23/07/08 18:54:38 INFO TaskSetManager: Finished task 1.0 in stage 3673.0 (TID 2294)
 in 619 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:38 INFO TaskSchedulerImpl: Removed TaskSet 3673.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:38 INFO DAGScheduler: ShuffleMapStage 3673 (showString at <unknown>:
0) finished in 0.624 s
23/07/08 18:54:38 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:38 INFO DAGScheduler: running: Set()
23/07/08 18:54:38 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:38 INFO DAGScheduler: failed: Set()
23/07/08 18:54:38 INFO ShufflePartitionsUtil: For shuffle(893), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:38 INFO ShufflePartitionsUtil: For shuffle(889), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:38 INFO SparkContext: Starting job: $anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266
23/07/08 18:54:38 INFO DAGScheduler: Got job 1638 ($anonfun$withThreadLocalCaptured
$1 at FutureTask.java:266) with 1 output partitions
23/07/08 18:54:38 INFO DAGScheduler: Final stage: ResultStage 3683 ($anonfun$withThr
eadLocalCaptured$1 at FutureTask.java:266)
23/07/08 18:54:38 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
82)
23/07/08 18:54:38 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:38 INFO DAGScheduler: Submitting ResultStage 3683 (MapPartitionsRDD[5
010] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:266), which has no mis
sing parents
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_1999 stored as values in memory
 (estimated size 8.2 KiB, free 322.2 MiB)
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_1999_piece0 stored as bytes in m
emory (estimated size 4.2 KiB, free 322.2 MiB)
23/07/08 18:54:38 INFO BlockManagerInfo: Added broadcast_1999_piece0 in memory on 17
2.30.115.138:43839 (size: 4.2 KiB, free: 363.0 MiB)
23/07/08 18:54:38 INFO SparkContext: Created broadcast 1999 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:38 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 368
3 (MapPartitionsRDD[5010] at $anonfun$withThreadLocalCaptured$1 at FutureTask.java:2
66) (first 15 tasks are for partitions Vector(0))
23/07/08 18:54:38 INFO TaskSchedulerImpl: Adding task set 3683.0 with 1 tasks resour
ce profile 0
23/07/08 18:54:38 INFO TaskSetManager: Starting task 0.0 in stage 3683.0 (TID 2305)
```

```
 (172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7379 bytes)
23/07/08 18:54:38 INFO Executor: Running task 0.0 in stage 3683.0 (TID 2305)
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Getting 2 (324.5 KiB) non-empty
 blocks including 2 (324.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:38 INFO DAGScheduler: Registering RDD 5013 (showString at <unknown>:
0) as input to shuffle 894
23/07/08 18:54:38 INFO DAGScheduler: Got map stage job 1639 (showString at <unknown
>:0) with 1 output partitions
23/07/08 18:54:38 INFO DAGScheduler: Final stage: ShuffleMapStage 3685 (showString a
t <unknown>:0)
23/07/08 18:54:38 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
84)
23/07/08 18:54:38 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:38 INFO DAGScheduler: Submitting ShuffleMapStage 3685 (MapPartitionsR
DD[5013] at showString at <unknown>:0), which has no missing parents
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_2000 stored as values in memory
 (estimated size 16.5 KiB, free 322.2 MiB)
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_2000_piece0 stored as bytes in m
emory (estimated size 7.7 KiB, free 322.2 MiB)
23/07/08 18:54:38 INFO BlockManagerInfo: Added broadcast_2000_piece0 in memory on 17
2.30.115.138:43839 (size: 7.7 KiB, free: 363.0 MiB)
23/07/08 18:54:38 INFO SparkContext: Created broadcast 2000 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:38 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
3685 (MapPartitionsRDD[5013] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0))
23/07/08 18:54:38 INFO TaskSchedulerImpl: Adding task set 3685.0 with 1 tasks resour
ce profile 0
23/07/08 18:54:38 INFO Executor: Finished task 0.0 in stage 3683.0 (TID 2305). 27423
4 bytes result sent to driver
23/07/08 18:54:38 INFO TaskSetManager: Starting task 0.0 in stage 3685.0 (TID 2306)
 (172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7368 bytes)
23/07/08 18:54:38 INFO Executor: Running task 0.0 in stage 3685.0 (TID 2306)
23/07/08 18:54:38 INFO TaskSetManager: Finished task 0.0 in stage 3683.0 (TID 2305)
 in 12 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:54:38 INFO TaskSchedulerImpl: Removed TaskSet 3683.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:38 INFO DAGScheduler: ResultStage 3683 ($anonfun$withThreadLocalCaptu
red$1 at FutureTask.java:266) finished in 0.016 s
23/07/08 18:54:38 INFO DAGScheduler: Job 1638 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:54:38 INFO TaskSchedulerImpl: Killing all running tasks in stage 3683: S
tage finished
23/07/08 18:54:38 INFO DAGScheduler: Job 1638 finished: $anonfun$withThreadLocalCapt
ured$1 at FutureTask.java:266, took 0.017303 s
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Getting 2 (894.6 KiB) non-empty
 blocks including 2 (894.6 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_2001 stored as values in memory
 (estimated size 2.5 MiB, free 319.7 MiB)
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_2001_piece0 stored as bytes in m
emory (estimated size 405.3 KiB, free 319.3 MiB)
23/07/08 18:54:38 INFO BlockManagerInfo: Added broadcast_2001_piece0 in memory on 17
```

```
2.30.115.138:43839 (size: 405.3 KiB, free: 362.6 MiB)
23/07/08 18:54:38 INFO SparkContext: Created broadcast 2001 from $anonfun$withThread
LocalCaptured$1 at FutureTask.java:266
23/07/08 18:54:38 INFO Executor: Finished task 0.0 in stage 3685.0 (TID 2306). 4301
 bytes result sent to driver
23/07/08 18:54:38 INFO TaskSetManager: Finished task 0.0 in stage 3685.0 (TID 2306)
 in 61 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:54:38 INFO TaskSchedulerImpl: Removed TaskSet 3685.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:38 INFO DAGScheduler: ShuffleMapStage 3685 (showString at <unknown>:
0) finished in 0.066 s
23/07/08 18:54:38 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:38 INFO DAGScheduler: running: Set()
23/07/08 18:54:38 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:38 INFO DAGScheduler: failed: Set()
23/07/08 18:54:38 INFO ShufflePartitionsUtil: For shuffle(894, 890), advisory target
size: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:38 INFO DAGScheduler: Registering RDD 5020 (showString at <unknown>:
0) as input to shuffle 895
23/07/08 18:54:38 INFO DAGScheduler: Got map stage job 1640 (showString at <unknown
>:0) with 1 output partitions
23/07/08 18:54:38 INFO DAGScheduler: Final stage: ShuffleMapStage 3689 (showString a
t <unknown>:0)
23/07/08 18:54:38 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
87, ShuffleMapStage 3688)
23/07/08 18:54:38 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:38 INFO DAGScheduler: Submitting ShuffleMapStage 3689 (MapPartitionsR
DD[5020] at showString at <unknown>:0), which has no missing parents
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_2002 stored as values in memory
 (estimated size 77.6 KiB, free 319.2 MiB)
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_2002_piece0 stored as bytes in m
emory (estimated size 34.2 KiB, free 319.2 MiB)
23/07/08 18:54:38 INFO BlockManagerInfo: Added broadcast_2002_piece0 in memory on 17
2.30.115.138:43839 (size: 34.2 KiB, free: 362.6 MiB)
23/07/08 18:54:38 INFO SparkContext: Created broadcast 2002 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:38 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
3689 (MapPartitionsRDD[5020] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0))
23/07/08 18:54:38 INFO TaskSchedulerImpl: Adding task set 3689.0 with 1 tasks resour
ce profile 0
23/07/08 18:54:38 INFO TaskSetManager: Starting task 0.0 in stage 3689.0 (TID 2307)
 (172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7634 bytes)
23/07/08 18:54:38 INFO Executor: Running task 0.0 in stage 3689.0 (TID 2307)
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Getting 1 (825.8 KiB) non-empty
 blocks including 1 (825.8 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Getting 2 (2.7 KiB) non-empty bl
ocks including 2 (2.7 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-
local and 0 (0.0 B) remote blocks
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:38 INFO Executor: Finished task 0.0 in stage 3689.0 (TID 2307). 9924
 bytes result sent to driver
23/07/08 18:54:38 INFO TaskSetManager: Finished task 0.0 in stage 3689.0 (TID 2307)
 in 81 ms on 172.30.115.138 (executor driver) (1/1)
```

```
23/07/08 18:54:38 INFO TaskSchedulerImpl: Removed TaskSet 3689.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:38 INFO DAGScheduler: ShuffleMapStage 3689 (showString at <unknown>:
0) finished in 0.086 s
23/07/08 18:54:38 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:38 INFO DAGScheduler: running: Set()
23/07/08 18:54:38 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:38 INFO DAGScheduler: failed: Set()
23/07/08 18:54:38 INFO ShufflePartitionsUtil: For shuffle(895, 891), advisory target
size: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:38 INFO DAGScheduler: Registering RDD 5027 (showString at <unknown>:
0) as input to shuffle 896
23/07/08 18:54:38 INFO DAGScheduler: Got map stage job 1641 (showString at <unknown
>:0) with 1 output partitions
23/07/08 18:54:38 INFO DAGScheduler: Final stage: ShuffleMapStage 3695 (showString a
t <unknown>:0)
23/07/08 18:54:38 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 36
93, ShuffleMapStage 3694)
23/07/08 18:54:38 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:38 INFO DAGScheduler: Submitting ShuffleMapStage 3695 (MapPartitionsR
DD[5027] at showString at <unknown>:0), which has no missing parents
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_2003 stored as values in memory
 (estimated size 142.0 KiB, free 319.0 MiB)
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_2003_piece0 stored as bytes in m
emory (estimated size 59.0 KiB, free 319.0 MiB)
23/07/08 18:54:38 INFO BlockManagerInfo: Added broadcast_2003_piece0 in memory on 17
2.30.115.138:43839 (size: 59.0 KiB, free: 362.6 MiB)
23/07/08 18:54:38 INFO SparkContext: Created broadcast 2003 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:38 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage
3695 (MapPartitionsRDD[5027] at showString at <unknown>:0) (first 15 tasks are for p
artitions Vector(0))
23/07/08 18:54:38 INFO TaskSchedulerImpl: Adding task set 3695.0 with 1 tasks resour
ce profile 0
23/07/08 18:54:38 INFO TaskSetManager: Starting task 0.0 in stage 3695.0 (TID 2308)
 (172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7634 bytes)
23/07/08 18:54:38 INFO Executor: Running task 0.0 in stage 3695.0 (TID 2308)
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Getting 1 (958.5 KiB) non-empty
 blocks including 1 (958.5 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Getting 2 (274.0 B) non-empty bl
ocks including 2 (274.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-
local and 0 (0.0 B) remote blocks
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:38 INFO BlockManagerInfo: Removed broadcast_1999_piece0 on 172.30.11
5.138:43839 in memory (size: 4.2 KiB, free: 362.6 MiB)
23/07/08 18:54:38 INFO BlockManagerInfo: Removed broadcast_2002_piece0 on 172.30.11
5.138:43839 in memory (size: 34.2 KiB, free: 362.6 MiB)
23/07/08 18:54:38 INFO BlockManagerInfo: Removed broadcast_2000_piece0 on 172.30.11
5.138:43839 in memory (size: 7.7 KiB, free: 362.6 MiB)
23/07/08 18:54:38 INFO BlockManagerInfo: Removed broadcast_1996_piece0 on 172.30.11
5.138:43839 in memory (size: 4.2 KiB, free: 362.6 MiB)
```

```
+-----------+------------------+-----------------+----+---+-----------+-------------
+------------------+
|TipoFamilia|          Sucursal|Descripción_Canal| AÑO|MES|Nombre_Empl|Apellido_Empl
|       TotalVentas|
+-----------+------------------+-----------------+----+---+-----------+-------------
+------------------+
|INFORMATICA|         Palermo 2|           OnLine|2020|  9|       Juan|       Arango
|         2103486.0|
|      AUDIO|          Rosario1|           OnLine|2020|  9|       José|      Vallejo
|          853974.0|
| ESTUCHERIA|          La Plata|           OnLine|2020|  9|   Leonardo|        Uribe
|          448182.0|
|           |    Córdoba Centro|           OnLine|2020|  9|      Elena|     Arroyave
|          334500.0|
|INFORMATICA|           Caseros|           OnLine|2020|  9|    Marcela|    De santis
|          327448.0|
| ESTUCHERIA|           Alberdi|       Telefónica|2020|  9|     Evelyn|        Diaz
|          236400.0|
| ESTUCHERIA|          Mendoza1|           OnLine|2020|  9|     Manuel|       Rojas
|          231800.0|
|           |          Rosario1|           OnLine|2020|  9|     Felipe|       Guerra
|          218100.0|
|  IMPRESIÓN|           Quilmes|           OnLine|2020|  9|   Isabella|      Simanca
|          154500.0|
|  IMPRESIÓN|        San Isidro|           OnLine|2020|  9|    Alberto|        Lopez
|          130500.0|
|  IMPRESIÓN|           Almagro|           OnLine|2020|  9|     Melisa|        Uribe
|          123569.0|
|  IMPRESIÓN|           Almagro|           OnLine|2020|  9|     Carlos|        Gomez
|          121220.0|
|INFORMATICA|             Moron|           OnLine|2020|  9|     Carmen|        Uribe
|           91564.0|
|      BASES|    Córdoba Quiroz|           OnLine|2020|  9|     Carlos|      Jimenez
|           79600.0|
|  IMPRESIÓN|    Córdoba Quiroz|           OnLine|2020|  9|    Gonzalo|       Florez
|           73919.0|
|      AUDIO|             Lanus|           OnLine|2020|  9|     Julian|       Burgos
|           70664.0|
|INFORMATICA|         Palermo 2|           OnLine|2020|  9|   Virginia|  Saldarriaga
|          57441.78|
|INFORMATICA|         Palermo 2|           OnLine|2020|  9|   Federico|     Arroyave
|          57441.78|
|  IMPRESIÓN|        San Isidro|           OnLine|2020|  9|  Sebastian|        Perez
|           52600.0|
|INFORMATICA|Cerro de las Rosas|           OnLine|2020|  9|      Carla|     Betancur
|39469.979999999996|
+-----------+------------------+-----------------+----+---+-----------+-------------
+------------------+
only showing top 20 rows
```

```
23/07/08 18:54:38 INFO Executor: Finished task 0.0 in stage 3695.0 (TID 2308). 17117
bytes result sent to driver
23/07/08 18:54:38 INFO TaskSetManager: Finished task 0.0 in stage 3695.0 (TID 2308)
in 194 ms on 172.30.115.138 (executor driver) (1/1)
23/07/08 18:54:38 INFO TaskSchedulerImpl: Removed TaskSet 3695.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:38 INFO DAGScheduler: ShuffleMapStage 3695 (showString at <unknown>:
0) finished in 0.203 s
23/07/08 18:54:38 INFO DAGScheduler: looking for newly runnable stages
23/07/08 18:54:38 INFO DAGScheduler: running: Set()
23/07/08 18:54:38 INFO DAGScheduler: waiting: Set()
23/07/08 18:54:38 INFO DAGScheduler: failed: Set()
23/07/08 18:54:38 INFO ShufflePartitionsUtil: For shuffle(896), advisory target siz
e: 67108864, actual target size 1048576, minimum partition size: 1048576
23/07/08 18:54:38 INFO HashAggregateExec: spark.sql.codegen.aggregate.map.twolevel.e
nabled is set to true, but current version of codegened fast hashmap does not suppor
t this aggregate.
23/07/08 18:54:38 INFO SparkContext: Starting job: showString at <unknown>:0
23/07/08 18:54:38 INFO DAGScheduler: Got job 1642 (showString at <unknown>:0) with 2
output partitions
23/07/08 18:54:38 INFO DAGScheduler: Final stage: ResultStage 3702 (showString at <u
nknown>:0)
23/07/08 18:54:38 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 37
01)
23/07/08 18:54:38 INFO DAGScheduler: Missing parents: List()
23/07/08 18:54:38 INFO DAGScheduler: Submitting ResultStage 3702 (MapPartitionsRDD[5
031] at showString at <unknown>:0), which has no missing parents
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_2004 stored as values in memory
(estimated size 131.4 KiB, free 319.0 MiB)
23/07/08 18:54:38 INFO MemoryStore: Block broadcast_2004_piece0 stored as bytes in m
emory (estimated size 44.7 KiB, free 319.0 MiB)
23/07/08 18:54:38 INFO BlockManagerInfo: Added broadcast_2004_piece0 in memory on 17
2.30.115.138:43839 (size: 44.7 KiB, free: 362.6 MiB)
23/07/08 18:54:38 INFO SparkContext: Created broadcast 2004 from broadcast at DAGSch
eduler.scala:1535
23/07/08 18:54:38 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 370
2 (MapPartitionsRDD[5031] at showString at <unknown>:0) (first 15 tasks are for part
itions Vector(0, 1))
23/07/08 18:54:38 INFO TaskSchedulerImpl: Adding task set 3702.0 with 2 tasks resour
ce profile 0
23/07/08 18:54:38 INFO TaskSetManager: Starting task 0.0 in stage 3702.0 (TID 2309)
(172.30.115.138, executor driver, partition 0, NODE_LOCAL, 7363 bytes)
23/07/08 18:54:38 INFO TaskSetManager: Starting task 1.0 in stage 3702.0 (TID 2310)
(172.30.115.138, executor driver, partition 1, NODE_LOCAL, 7363 bytes)
23/07/08 18:54:38 INFO Executor: Running task 1.0 in stage 3702.0 (TID 2310)
23/07/08 18:54:38 INFO Executor: Running task 0.0 in stage 3702.0 (TID 2309)
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Getting 1 (1378.9 KiB) non-empty
blocks including 1 (1378.9 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Getting 1 (1033.8 KiB) non-empty
blocks including 1 (1033.8 KiB) local and 0 (0.0 B) host-local and 0 (0.0 B) push-me
rged-local and 0 (0.0 B) remote blocks
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:38 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
23/07/08 18:54:38 INFO Executor: Finished task 0.0 in stage 3702.0 (TID 2309). 21628
bytes result sent to driver
```

```
23/07/08 18:54:38 INFO TaskSetManager: Finished task 0.0 in stage 3702.0 (TID 2309)
in 37 ms on 172.30.115.138 (executor driver) (1/2)
23/07/08 18:54:38 INFO Executor: Finished task 1.0 in stage 3702.0 (TID 2310). 21636
bytes result sent to driver
23/07/08 18:54:38 INFO TaskSetManager: Finished task 1.0 in stage 3702.0 (TID 2310)
in 44 ms on 172.30.115.138 (executor driver) (2/2)
23/07/08 18:54:38 INFO TaskSchedulerImpl: Removed TaskSet 3702.0, whose tasks have a
ll completed, from pool
23/07/08 18:54:38 INFO DAGScheduler: ResultStage 3702 (showString at <unknown>:0) fi
nished in 0.051 s
23/07/08 18:54:38 INFO DAGScheduler: Job 1642 is finished. Cancelling potential spec
ulative or zombie tasks for this job
23/07/08 18:54:38 INFO TaskSchedulerImpl: Killing all running tasks in stage 3702: S
tage finished
23/07/08 18:54:38 INFO DAGScheduler: Job 1642 finished: showString at <unknown>:0, t
ook 0.053609 s
```

Imprimir la estructura de cada dataset.

In [200…
```
df_canal.printSchema()
```

```
root
 |-- CODIGO: string (nullable = true)
 |-- DESCRIPCION: string (nullable = true)
```

In [201…
```
df_cliente.printSchema()
```

```
root
 |-- ID: string (nullable = true)
 |-- Provincia: string (nullable = true)
 |-- Nombre_y_Apellido: string (nullable = true)
 |-- Domicilio: string (nullable = true)
 |-- Telefono: string (nullable = true)
 |-- Edad: string (nullable = true)
 |-- Localidad: string (nullable = true)
 |-- X: string (nullable = true)
 |-- Y: string (nullable = true)
 |-- col10: string (nullable = true)
```

In [202…
```
df_empleado.printSchema()
```

```
root
 |-- ID_empleado: string (nullable = true)
 |-- Apellido: string (nullable = true)
 |-- Nombre: string (nullable = true)
 |-- Sucursal: string (nullable = true)
 |-- Sector: string (nullable = true)
 |-- Cargo: string (nullable = true)
 |-- Salario: string (nullable = true)
```

In [134…
```
df_producto.printSchema()
```

```
root
 |-- ID_PRODUCTO: string (nullable = true)
 |-- Concepto: string (nullable = true)
 |-- Tipo: string (nullable = true)
 |--  Precio : string (nullable = true)
```

In [135…
```
df_sucursal.printSchema()
```

```
root
 |-- ID: string (nullable = true)
 |-- Sucursal: string (nullable = true)
 |-- Direccion: string (nullable = true)
 |-- Localidad: string (nullable = true)
 |-- Provincia: string (nullable = true)
 |-- Latitud: string (nullable = true)
 |-- Longitud: string (nullable = true)
```

In [136…
```
df_venta.printSchema()
```

```
root
 |-- IdVenta: string (nullable = true)
 |-- Fecha: string (nullable = true)
 |-- Fecha_Entrega: string (nullable = true)
 |-- IdCanal: string (nullable = true)
 |-- IdCliente: string (nullable = true)
 |-- IdSucursal: string (nullable = true)
 |-- IdEmpleado: string (nullable = true)
 |-- IdProducto: string (nullable = true)
 |-- Precio: string (nullable = true)
 |-- Cantidad: string (nullable = true)
 |-- total_venta: double (nullable = true)
```

In [137…
```
df_DIMDATE_DATAONLY.printSchema()
```

```
root
 |-- ID: string (nullable = true)
 |-- FECHA: string (nullable = true)
 |-- DIA: string (nullable = true)
 |-- MES: string (nullable = true)
 |-- AÑO: string (nullable = true)
 |-- SEMANA: string (nullable = true)
 |-- BIMESTRE: string (nullable = true)
 |-- TRIMESTRE: string (nullable = true)
 |-- CUATRIMESTRE: string (nullable = true)
 |-- SEMESTRE: string (nullable = true)
 |-- FINDESEMANA: string (nullable = true)
 |-- DIADELAÑO: string (nullable = true)
 |-- AÑO: string (nullable = true)
 |-- DIASEMANA: string (nullable = true)
```

In [138…
```
df_DIMDATE.printSchema()
```

```
root
 |-- ID: string (nullable = true)
 |-- FECHA: string (nullable = true)
 |-- DIA: string (nullable = true)
 |-- MES: string (nullable = true)
 |-- AÑO: string (nullable = true)
 |-- SEMANA: string (nullable = true)
 |-- BIMESTRE: string (nullable = true)
 |-- TRIMESTRE: string (nullable = true)
 |-- CUATRIMESTRE: string (nullable = true)
 |-- SEMESTRE: string (nullable = true)
 |-- FINDESEMANA: string (nullable = true)
 |-- DIADELAÑO: string (nullable = true)
 |-- AÑO2: string (nullable = true)
 |-- DIASEMANA: string (nullable = true)
 |-- DIASEMANATEXTO: string (nullable = true)
```

In [139…
```
df_DIMTIME_DATAONLY.printSchema()
```

```
root
 |-- ID: string (nullable = true)
 |-- TIEMPO: string (nullable = true)
 |-- HORA: string (nullable = true)
 |-- MINUTO: string (nullable = true)
 |-- BLOQUEMEDIAHORA: string (nullable = true)
 |-- HORAOFICINA: string (nullable = true)
 |-- AM-PM: string (nullable = true)
```

In [140…
```
df_DIMTIME.printSchema()
```

```
root
 |-- ID: string (nullable = true)
 |-- TIEMPO: string (nullable = true)
 |-- HORA: string (nullable = true)
 |-- MINUTO: string (nullable = true)
 |-- BLOQUEMEDIAHORA: string (nullable = true)
 |-- HORAOFICINA: string (nullable = true)
 |-- AM-PM: string (nullable = true)
```