Executando o Spark Submit.

```
(base) [hadoop@dataserver Downloads]$ spark-submit gastos-cliente.py
2020-07-23 17:32:28,783 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2020-07-23 17:32:29,245 INFO spark.SparkContext: Running Spark version 2.4.4
2020-07-23 17:32:29,258 INFO spark.SparkContext: Submitted application: GastosPorCliente
2020-07-23 17:32:29,291 INFO spark.SecurityManager: Changing view acls to: hadoop
2020-07-23 17:32:29,291 INFO spark.SecurityManager: Changing modify acls to: hadoop
2020-07-23 17:32:29,291 INFO spark.SecurityManager: Changing view acls groups to:
2020-07-23 17:32:29,291 INFO spark.SecurityManager: Changing modify acls groups to:
2020-07-23 17:32:29,291 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(hadoop); gro
ups with view permissions: Set(); users  with modify permissions: Set(hadoop); groups with modify permissions: Set()
2020-07-23 17:32:29,465 INFO util.Utils: Successfully started service 'sparkDriver' on port 12177.
2020-07-23 17:32:29,480 INFO spark.SparkEnv: Registering MapOutputTracker
2020-07-23 17:32:29,490 INFO spark.SparkEnv: Registering BlockManagerMaster
2020-07-23 17:32:29,492 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
2020-07-23 17:32:29,492 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
2020-07-23 17:32:29,497 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-6ce588e3-cbd9-45e1-9eab-b40f7850b5d3
2020-07-23 17:32:29,508 INFO memory.MemoryStore: MemoryStore started with capacity 366.3 MB
2020-07-23 17:32:29,517 INFO spark.SparkEnv: Registering OutputCommitCoordinator
2020-07-23 17:32:29,563 INFO util.log: Logging initialized @1559ms
2020-07-23 17:32:29,606 INFO server.Server: jetty-9.3.z-SNAPSHOT, build timestamp: unknown, git hash: unknown
2020-07-23 17:32:29,616 INFO server.Server: Started @1612ms
2020-07-23 17:32:29,626 INFO server.AbstractConnector: Started ServerConnector@605963ac{HTTP/1.1,[http/1.1]}{192.168.0.57:4040}
2020-07-23 17:32:29,627 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
2020-07-23 17:32:29,645 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@7358f690{/jobs,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,646 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@720a591b{/jobs/json,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,646 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@3ea4377a{/jobs/job,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,647 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@5b0ad865{/jobs/job/json,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,648 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@7b68809e{/stages,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,648 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@6468776e{/stages/json,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,648 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@6640f08a{/stages/stage,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,650 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@73a31aba{/stages/stage/json,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,650 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@128670e4{/stages/pool,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,651 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@5f61a50e{/stages/pool/json,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,651 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@35fec7b7{/storage,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,652 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@5ee37d78{/storage/json,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,652 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@7f432710{/storage/rdd,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,653 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@2dd8a423{/storage/rdd/json,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,653 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@6ed81d50{/environment,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,654 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@2b95b54e{/environment/json,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,654 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@4d087c8b{/executors,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,655 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@251930b5{/executors/json,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,656 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@b058bdb{/executors/threadDump,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,656 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@d3c6030{/executors/threadDump/json,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,660 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@3deb938{/static,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,661 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@12354aa2{/,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,662 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@46858f7e{/api,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,662 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@19227154{/jobs/job/kill,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,663 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@5fa792b2{/stages/stage/kill,null,AVAILABLE,@Spark}
2020-07-23 17:32:29,664 INFO ui.SparkUI: Bound SparkUI to 192.168.0.57, and started at http://dataserver:4040
2020-07-23 17:32:29,760 INFO executor.Executor: Starting executor ID driver on host localhost
2020-07-23 17:32:29,801 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 19113.
2020-07-23 17:32:29,802 INFO netty.NettyBlockTransferService: Server created on dataserver:19113
2020-07-23 17:32:29,803 INFO storage.BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
2020-07-23 17:32:29,818 INFO storage.BlockManagerMaster: Registering BlockManager BlockManagerId(driver, dataserver, 19113, None)
2020-07-23 17:32:29,821 INFO storage.BlockManagerMasterEndpoint: Registering block manager dataserver:19113 with 366.3 MB RAM, BlockManagerId(driver, dataserv
er, 19113, None)
2020-07-23 17:32:29,823 INFO storage.BlockManagerMaster: Registered BlockManager BlockManagerId(driver, dataserver, 19113, None)
2020-07-23 17:32:29,823 INFO storage.BlockManager: Initialized BlockManager: BlockManagerId(driver, dataserver, 19113, None)
2020-07-23 17:32:29,900 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@3ed00bc0{/metrics/json,null,AVAILABLE,@Spark}
2020-07-23 17:32:30,263 INFO memory.MemoryStore: Block broadcast_0 stored as values in memory (estimated size 241.6 KB, free 366.1 MB)
2020-07-23 17:32:30,300 INFO memory.MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 23.2 KB, free 366.0 MB)
2020-07-23 17:32:30,302 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on dataserver:19113 (size: 23.2 KB, free: 366.3 MB)
2020-07-23 17:32:30,304 INFO spark.SparkContext: Created broadcast 0 from textFile at NativeMethodAccessorImpl.java:0
2020-07-23 17:32:30,587 INFO mapred.FileInputFormat: Total input paths to process : 1
2020-07-23 17:32:30,678 INFO spark.SparkContext: Starting job: collect at /home/hadoop/Downloads/gastos-cliente.py:20
2020-07-23 17:32:30,690 INFO scheduler.DAGScheduler: Registering RDD 3 (reduceByKey at /home/hadoop/Downloads/gastos-cliente.py:17)
2020-07-23 17:32:30,691 INFO scheduler.DAGScheduler: Got job 0 (collect at /home/hadoop/Downloads/gastos-cliente.py:20) with 1 output partitions
2020-07-23 17:32:30,692 INFO scheduler.DAGScheduler: Final stage: ResultStage 1 (collect at /home/hadoop/Downloads/gastos-cliente.py:20)
2020-07-23 17:32:30,692 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 0)
2020-07-23 17:32:30,693 INFO scheduler.DAGScheduler: Missing parents: List(ShuffleMapStage 0)
2020-07-23 17:32:30,696 INFO scheduler.DAGScheduler: Submitting ShuffleMapStage 0 (PairwiseRDD[3] at reduceByKey at /home/hadoop/Downloads/gastos-cliente.py:1
7), which has no missing parents
2020-07-23 17:32:30,719 INFO memory.MemoryStore: Block broadcast_1 stored as values in memory (estimated size 10.6 KB, free 366.0 MB)
2020-07-23 17:32:30,721 INFO memory.MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 6.9 KB, free 366.0 MB)
2020-07-23 17:32:30,724 INFO storage.BlockManagerInfo: Added broadcast_1_piece0 in memory on dataserver:19113 (size: 6.9 KB, free: 366.3 MB)
2020-07-23 17:32:30,724 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1161
2020-07-23 17:32:30,732 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 0 (PairwiseRDD[3] at reduceByKey at /home/hadoop/Download
s/gastos-cliente.py:17) (first 15 tasks are for partitions Vector(0))
2020-07-23 17:32:30,733 INFO scheduler.TaskSchedulerImpl: Adding task set 0.0 with 1 tasks
2020-07-23 17:32:30,771 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, executor driver, partition 0, ANY, 7898 bytes)
2020-07-23 17:32:30,777 INFO executor.Executor: Running task 0.0 in stage 0.0 (TID 0)
2020-07-23 17:32:30,819 INFO rdd.HadoopRDD: Input split: hdfs://localhost:9000/clientes/gastos-cliente.csv:0+146855
2020-07-23 17:32:31,365 INFO python.PythonRunner: Times: total = 390, boot = 296, init = 60, finish = 34
2020-07-23 17:32:31,380 INFO executor.Executor: Finished task 0.0 in stage 0.0 (TID 0). 1759 bytes result sent to driver
2020-07-23 17:32:31,386 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 631 ms on localhost (executor driver) (1/1)
2020-07-23 17:32:31,387 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
2020-07-23 17:32:31,389 INFO python.PythonAccumulatorV2: Connected to AccumulatorServer at host: 127.0.0.1 port: 13208
2020-07-23 17:32:31,392 INFO scheduler.DAGScheduler: ShuffleMapStage 0 (reduceByKey at /home/hadoop/Downloads/gastos-cliente.py:17) finished in 0.682 s
2020-07-23 17:32:31,392 INFO scheduler.DAGScheduler: looking for newly runnable stages
2020-07-23 17:32:31,393 INFO scheduler.DAGScheduler: running: Set()
2020-07-23 17:32:31,393 INFO scheduler.DAGScheduler: waiting: Set(ResultStage 1)
2020-07-23 17:32:31,393 INFO scheduler.DAGScheduler: failed: Set()
2020-07-23 17:32:31,395 INFO scheduler.DAGScheduler: Submitting ResultStage 1 (PythonRDD[6] at collect at /home/hadoop/Downloads/gastos-cliente.py:20), which
has no missing parents
2020-07-23 17:32:31,399 INFO memory.MemoryStore: Block broadcast_2 stored as values in memory (estimated size 7.2 KB, free 366.0 MB)
2020-07-23 17:32:31,400 INFO memory.MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 4.7 KB, free 366.0 MB)
2020-07-23 17:32:31,401 INFO storage.BlockManagerInfo: Added broadcast_2_piece0 in memory on dataserver:19113 (size: 4.7 KB, free: 366.3 MB)
2020-07-23 17:32:31,402 INFO spark.SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1161
2020-07-23 17:32:31,403 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (PythonRDD[6] at collect at /home/hadoop/Downloads/gastos-c
liente.py:20) (first 15 tasks are for partitions Vector(0))
2020-07-23 17:32:31,403 INFO scheduler.TaskSchedulerImpl: Adding task set 1.0 with 1 tasks
2020-07-23 17:32:31,405 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, executor driver, partition 0, ANY, 7662 bytes)
2020-07-23 17:32:31,406 INFO executor.Executor: Running task 0.0 in stage 1.0 (TID 1)
2020-07-23 17:32:31,419 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks including 1 local blocks and 0 remote blocks
2020-07-23 17:32:31,420 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 5 ms
2020-07-23 17:32:31,476 INFO python.PythonRunner: Times: total = 45, boot = -198, init = 242, finish = 1
2020-07-23 17:32:31,479 INFO executor.Executor: Finished task 0.0 in stage 1.0 (TID 1). 3233 bytes result sent to driver
2020-07-23 17:32:31,481 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 77 ms on localhost (executor driver) (1/1)
2020-07-23 17:32:31,481 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
2020-07-23 17:32:31,482 INFO scheduler.DAGScheduler: ResultStage 1 (collect at /home/hadoop/Downloads/gastos-cliente.py:20) finished in 0.085 s
2020-07-23 17:32:31,486 INFO scheduler.DAGScheduler: Job 0 finished: collect at /home/hadoop/Downloads/gastos-cliente.py:20, took 0.808270 s
```

Gastos por clientes (ID).

```
(23, 4042.6499999999987)
(72, 5337.44)
(19, 5059.4299999999985)
(40, 5186.429999999999)
(49, 4394.599999999999)
(83, 4635.799999999997)
(70, 5368.249999999999)
(65, 5140.3499999999985)
(46, 5963.109999999999)
(80, 4727.860000000001)
(16, 4979.06)
(26, 5250.4)
(29, 5032.529999999999)
(9, 5322.649999999999)
(86, 4908.81)
(6, 5397.879999999998)
(31, 4765.05)
(22, 5019.449999999999)
(4, 4815.050000000002)
(18, 4921.27)
(59, 5642.89)
(71, 5995.660000000003)
(44, 4756.8899999999985)
(75, 4178.500000000001)
(79, 3790.570000000001)
(5, 4561.069999999999)
(51, 4975.22)
(85, 5503.43)
(90, 5290.409999999998)
(39, 6193.109999999999)
(92, 5379.280000000002)
(93, 5265.750000000001)
(61, 5497.479999999998)
(63, 5415.150000000001)
(8, 5517.240000000001)
(3, 4659.63)
(53, 4945.299999999999)
(30, 4990.72)
(58, 5437.7300000000005)
(91, 4642.259999999999)
(37, 4735.200000000002)
(10, 4819.700000000001)
(97, 5977.189999999995)
(34, 5330.8)
(35, 5155.419999999999)
(2, 5994.59)
(47, 4316.299999999999)
(14, 4735.030000000001)
(42, 5696.840000000003)
(50, 4517.27)
(20, 4836.859999999999)
(15, 5413.510000000001)
(48, 4384.33)
(36, 4278.049999999997)
(57, 4628.4)
(12, 4664.589999999998)
(54, 6065.389999999999)
(0, 5524.949999999998)
(88, 4830.549999999999)
(13, 4367.62)
(98, 4297.260000000001)
(55, 5298.090000000002)
(69, 5123.010000000001)
(1, 4958.600000000001)
(64, 5288.689999999996)
(82, 4812.489999999998)
(99, 4172.289999999998)
(73, 6206.199999999999)
(74, 4647.129999999999)
(56, 4701.019999999999)
(68, 6375.449999999997)
(11, 5152.290000000002)
(41, 5637.62)
(87, 5206.4)
(17, 5032.679999999999)
(33, 5254.659999999998)
(62, 5253.3200000000015)
(76, 4904.209999999999)
(66, 4681.919999999999)
(43, 5368.83)
(52, 5245.059999999999)
(77, 4327.729999999999)
(81, 5112.709999999999)
(84, 4652.939999999999)
(89, 4851.479999999999)
(45, 3309.38)
(24, 5259.920000000003)
(96, 3924.230000000001)
(67, 4505.79)
(94, 4475.569999999999)
(32, 5496.050000000004)
(38, 4898.460000000002)
(28, 5000.709999999998)
2020-07-23 17:32:31,530 INFO spark.SparkContext: Invoking stop() from shutdown hook
2020-07-23 17:32:31,533 INFO server.AbstractConnector: Stopped Spark@605963ac{HTTP/1.1,[http/1.1]}{192.168.0.57:4040}
2020-07-23 17:32:31,535 INFO ui.SparkUI: Stopped Spark web UI at http://dataserver:4040
2020-07-23 17:32:31,541 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
2020-07-23 17:32:31,548 INFO memory.MemoryStore: MemoryStore cleared
2020-07-23 17:32:31,548 INFO storage.BlockManager: BlockManager stopped
2020-07-23 17:32:31,556 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
2020-07-23 17:32:31,559 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
2020-07-23 17:32:31,561 INFO spark.SparkContext: Successfully stopped SparkContext
2020-07-23 17:32:31,561 INFO util.ShutdownHookManager: Shutdown hook called
2020-07-23 17:32:31,562 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-a98092a3-b235-4d39-8e08-e70b409bc0aa
2020-07-23 17:32:31,563 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-a98092a3-b235-4d39-8e08-e70b409bc0aa/pyspark-af5e00cf-00cc-41d6-beb5-8c05
8cb2956e
2020-07-23 17:32:31,563 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-844296a1-b1c9-472b-85c2-759769a0a104
```