

LAB 12 BIG DATA

Submitted by:

ABHISHEK ARYA

NET ID: axa220149

```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System Tue Aug 1, 12:48 PM
training@localhost:~
File Edit View Search Terminal Help
23/08/01 12:41:04 INFO storage.MemoryStore: Block broadcast 5 piece0 stored as bytes in memory (estimated size 20.7 KB, free 266.6 MB)
23/08/01 12:41:04 INFO storage.BlockManagerInfo: Added broadcast_5_piece0 in memory on localhost:42552 (size: 20.7 KB, free: 267.2 MB)
23/08/01 12:41:04 INFO storage.BlockManagerMaster: Updated info of block broadcast_5_piece0
23/08/01 12:41:04 INFO spark.SparkContext: Created broadcast 5 from textFile at NativeMethodAccessorImpl.java:-2

In [8]: accounthits = accounts.join(userreqs)
23/08/01 12:41:33 INFO mapred.FileInputFormat: Total input paths to process : 19

In [9]: for (userid,(values,count)) in accounthits.take(5) :
...:     print userid, count, values[3],values[4]
...:
23/08/01 12:42:35 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:356
23/08/01 12:42:35 INFO spark.MapOutputTrackerMaster: Size of output statuses for shuffle 0 is 743 bytes
23/08/01 12:42:35 INFO scheduler.DAGScheduler: Registering RDD 21 (join at <ipython-input-8-b8d0361ddb3f>:1)
23/08/01 12:42:35 INFO scheduler.DAGScheduler: Got job 2 (runJob at PythonRDD.scala:356) with 1 output partitions (allowLocal=true)
23/08/01 12:42:35 INFO scheduler.DAGScheduler: Final stage: Stage 6(runJob at PythonRDD.scala:356)
23/08/01 12:42:35 INFO scheduler.DAGScheduler: Parents of final stage: List(Stage 5)
23/08/01 12:42:35 INFO scheduler.DAGScheduler: Missing parents: List(Stage 5)
23/08/01 12:42:36 INFO scheduler.DAGScheduler: Submitting Stage 5 (PairwiseRDD[21] at join at <ipython-input-8-b8d0361ddb3f>:1), which has no missing parents
23/08/01 12:42:36 INFO storage.MemoryStore: ensureFreeSpace(12264) called with curMem=649619, maxMem=280248975
23/08/01 12:42:36 INFO storage.MemoryStore: Block broadcast 6 stored as values in memory (estimated size 12.0 KB, free 266.6 MB)
23/08/01 12:42:36 INFO storage.MemoryStore: ensureFreeSpace(6606) called with curMem=661883, maxMem=280248975
23/08/01 12:42:36 INFO storage.MemoryStore: Block broadcast 6 piece0 stored as bytes in memory (estimated size 6.5 KB, free 266.6 MB)
23/08/01 12:42:36 INFO storage.BlockManagerInfo: Added broadcast_6_piece0 in memory on localhost:42552 (size: 6.5 KB, free: 267.2 MB)
23/08/01 12:42:36 INFO storage.BlockManagerMaster: Updated info of block broadcast_6_piece0
23/08/01 12:42:36 INFO spark.SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:839
23/08/01 12:42:36 INFO scheduler.DAGScheduler: Submitting 51 missing tasks from Stage 5 (PairwiseRDD[21] at join at <ipython-input-8-b8d0361ddb3f>:1)
23/08/01 12:42:36 INFO scheduler.TaskSchedulerImpl: Adding task set 5.0 with 51 tasks
23/08/01 12:42:36 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 5.0 (TID 97, localhost, PROCESS_LOCAL, 1414 bytes)
23/08/01 12:42:36 INFO executor.Executor: Running task 0.0 in stage 5.0 (TID 97)
23/08/01 12:42:36 INFO rdd.HadoopRDD: Input split: hdfs://localhost:8020/loudacre/accounts/part-m-000000:0+4706617
23/08/01 12:42:37 INFO python.PythonRDD: Times: total = 1371, boot = 8, init = 41, finish = 1322
23/08/01 12:42:38 INFO python.PythonRDD: Times: total = 2544, boot = 9, init = 145, finish = 2390
training@localhost:~
```

```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 17 Player (Non-commercial use only)
Player
Applications Places System Tue Aug 1, 12:47 PM
training@localhost:~
File Edit View Search Terminal Help
23/08/01 12:42:56 INFO executor.Executor: Finished task 0.0 in stage 6.0 (TID 148). 2153 bytes result sent to driver
23/08/01 12:42:56 INFO scheduler.DAGScheduler: Stage 6 (runJob at PythonRDD.scala:356) finished in 0.190 s
23/08/01 12:42:56 INFO scheduler.DAGScheduler: Job 2 finished: runJob at PythonRDD.scala:356, took 21.021583 s
26378 4 Jamie Castillo
5986 10 David Moore
32839 2 Nellie Harris
61170 2 James Spies
61100 10 Melissa Young

In [10]: 23/08/01 12:42:56 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 6.0 (TID 148) in 189 ms on localhost (1/1)
23/08/01 12:42:56 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 6.0, whose tasks have all completed, from pool

In [10]: accountsByPCode = sc.textFile("/loudacre/accounts") \
.....: .map(lambda s: s.split(','))\
.....: .keyBy(lambda account: account[8])
23/08/01 12:45:14 INFO storage.MemoryStore: ensureFreeSpace(280243) called with curMem=657167, maxMem=280248975
23/08/01 12:45:14 INFO storage.MemoryStore: Block broadcast 8 stored as values in memory (estimated size 273.7 KB, free 266.4 MB)
23/08/01 12:45:14 INFO storage.MemoryStore: ensureFreeSpace(21204) called with curMem=937410, maxMem=280248975
23/08/01 12:45:14 INFO storage.MemoryStore: Block broadcast 8 piece0 stored as bytes in memory (estimated size 20.7 KB, free 266.4 MB)
23/08/01 12:45:14 INFO storage.BlockManagerInfo: Added broadcast 8 piece0 in memory on localhost:42552 (size: 20.7 KB, free: 267.2 MB)
23/08/01 12:45:14 INFO storage.BlockManagerMaster: Updated info of block broadcast 8 piece0
23/08/01 12:45:14 INFO spark.SparkContext: Created broadcast 8 from textFile at NativeMethodAccessorImpl.java:-2
```