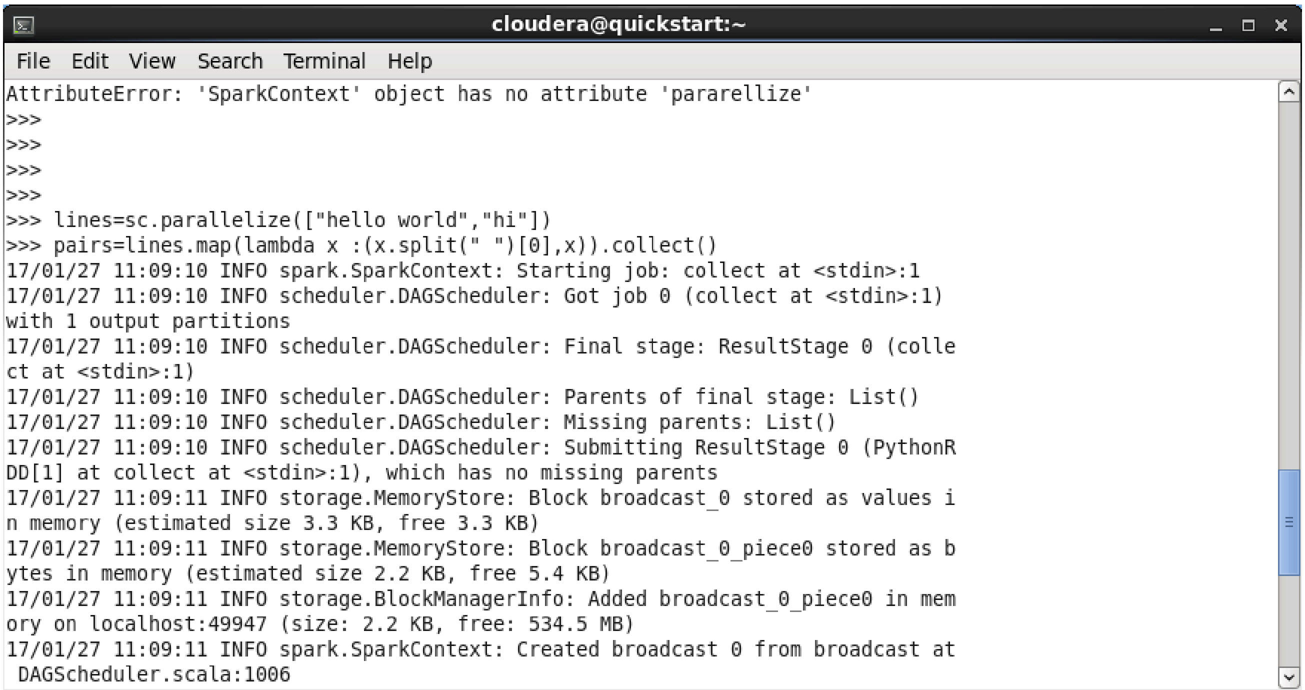
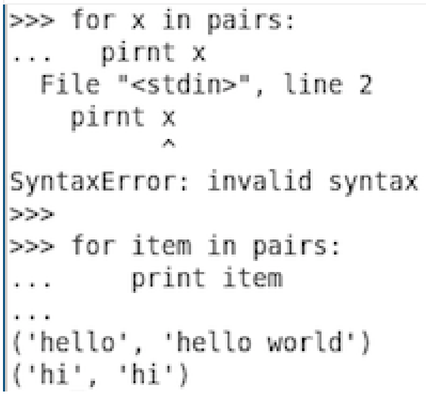
2017-01-27 Spark 學習筆記

=================
建立PairRDD
=================

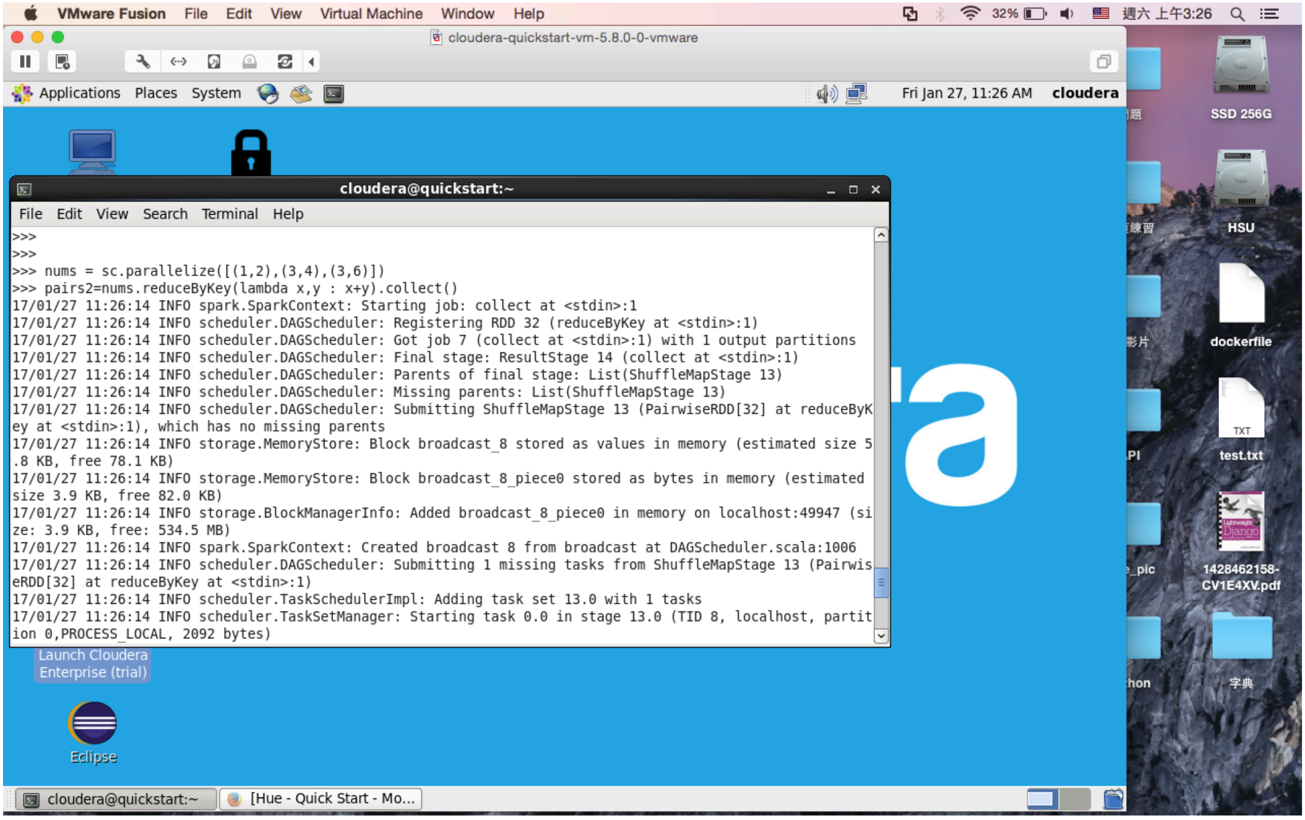


```
AttributeError: 'SparkContext' object has no attribute 'pararellize'
>>>
>>>
>>>
>>>
>>> lines=sc.parallelize(["hello world","hi"])
>>> pairs=lines.map(lambda x :(x.split(" ")[0],x)).collect()
17/01/27 11:09:10 INFO spark.SparkContext: Starting job: collect at <stdin>:1
17/01/27 11:09:10 INFO scheduler.DAGScheduler: Got job 0 (collect at <stdin>:1)
with 1 output partitions
17/01/27 11:09:10 INFO scheduler.DAGScheduler: Final stage: ResultStage 0 (colle
ct at <stdin>:1)
17/01/27 11:09:10 INFO scheduler.DAGScheduler: Parents of final stage: List()
17/01/27 11:09:10 INFO scheduler.DAGScheduler: Missing parents: List()
17/01/27 11:09:10 INFO scheduler.DAGScheduler: Submitting ResultStage 0 (PythonR
DD[1] at collect at <stdin>:1), which has no missing parents
17/01/27 11:09:11 INFO storage.MemoryStore: Block broadcast_0 stored as values i
n memory (estimated size 3.3 KB, free 3.3 KB)
17/01/27 11:09:11 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as b
ytes in memory (estimated size 2.2 KB, free 5.4 KB)
17/01/27 11:09:11 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in mem
ory on localhost:49947 (size: 2.2 KB, free: 534.5 MB)
17/01/27 11:09:11 INFO spark.SparkContext: Created broadcast 0 from broadcast at
 DAGScheduler.scala:1006
```

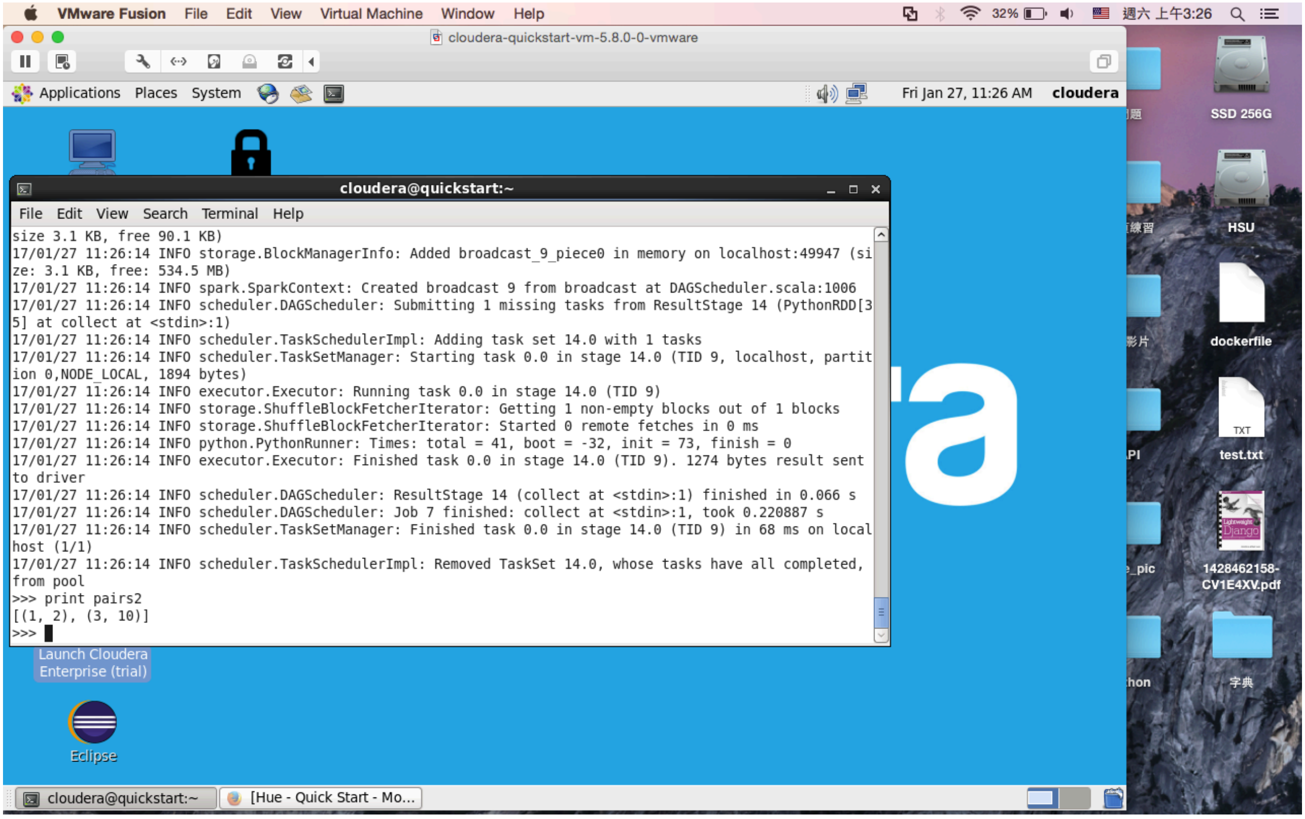


```
>>> for x in pairs:
...    pirnt x
  File "<stdin>", line 2
    pirnt x
          ^
SyntaxError: invalid syntax
>>>
>>> for item in pairs:
...      print item
...
('hello', 'hello world')
('hi', 'hi')
```

=================
reduceByKey
=================


VMware Fusion File Edit View Virtual Machine Window Help
32%
週六 上午3:26
cloudera-quickstart-vm-5.8.0-0-vmware
Applications Places System
Fri Jan 27, 11:26 AM
cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
Launch Cloudera Enterprise (trial)
Eclipse
SSD 256G
HSU
dockerfile
test.txt
1428462158-CV1E4XV.pdf
字典
cloudera@quickstart:~
[Hue - Quick Start - Mo...

```
>>>
>>>
>>> nums = sc.parallelize([(1,2),(3,4),(3,6)])
>>> pairs2=nums.reduceByKey(lambda x,y : x+y).collect()
17/01/27 11:26:14 INFO spark.SparkContext: Starting job: collect at <stdin>:1
17/01/27 11:26:14 INFO scheduler.DAGScheduler: Registering RDD 32 (reduceByKey at <stdin>:1)
17/01/27 11:26:14 INFO scheduler.DAGScheduler: Got job 7 (collect at <stdin>:1) with 1 output partitions
17/01/27 11:26:14 INFO scheduler.DAGScheduler: Final stage: ResultStage 14 (collect at <stdin>:1)
17/01/27 11:26:14 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 13)
17/01/27 11:26:14 INFO scheduler.DAGScheduler: Missing parents: List(ShuffleMapStage 13)
17/01/27 11:26:14 INFO scheduler.DAGScheduler: Submitting ShuffleMapStage 13 (PairwiseRDD[32] at reduceByK
ey at <stdin>:1), which has no missing parents
17/01/27 11:26:14 INFO storage.MemoryStore: Block broadcast_8 stored as values in memory (estimated size 5
.8 KB, free 78.1 KB)
17/01/27 11:26:14 INFO storage.MemoryStore: Block broadcast_8_piece0 stored as bytes in memory (estimated
size 3.9 KB, free 82.0 KB)
17/01/27 11:26:14 INFO storage.BlockManagerInfo: Added broadcast_8_piece0 in memory on localhost:49947 (si
ze: 3.9 KB, free: 534.5 MB)
17/01/27 11:26:14 INFO spark.SparkContext: Created broadcast 8 from broadcast at DAGScheduler.scala:1006
17/01/27 11:26:14 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 13 (Pairwis
eRDD[32] at reduceByKey at <stdin>:1)
17/01/27 11:26:14 INFO scheduler.TaskSchedulerImpl: Adding task set 13.0 with 1 tasks
17/01/27 11:26:14 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 13.0 (TID 8, localhost, partit
ion 0,PROCESS_LOCAL, 2092 bytes)
```

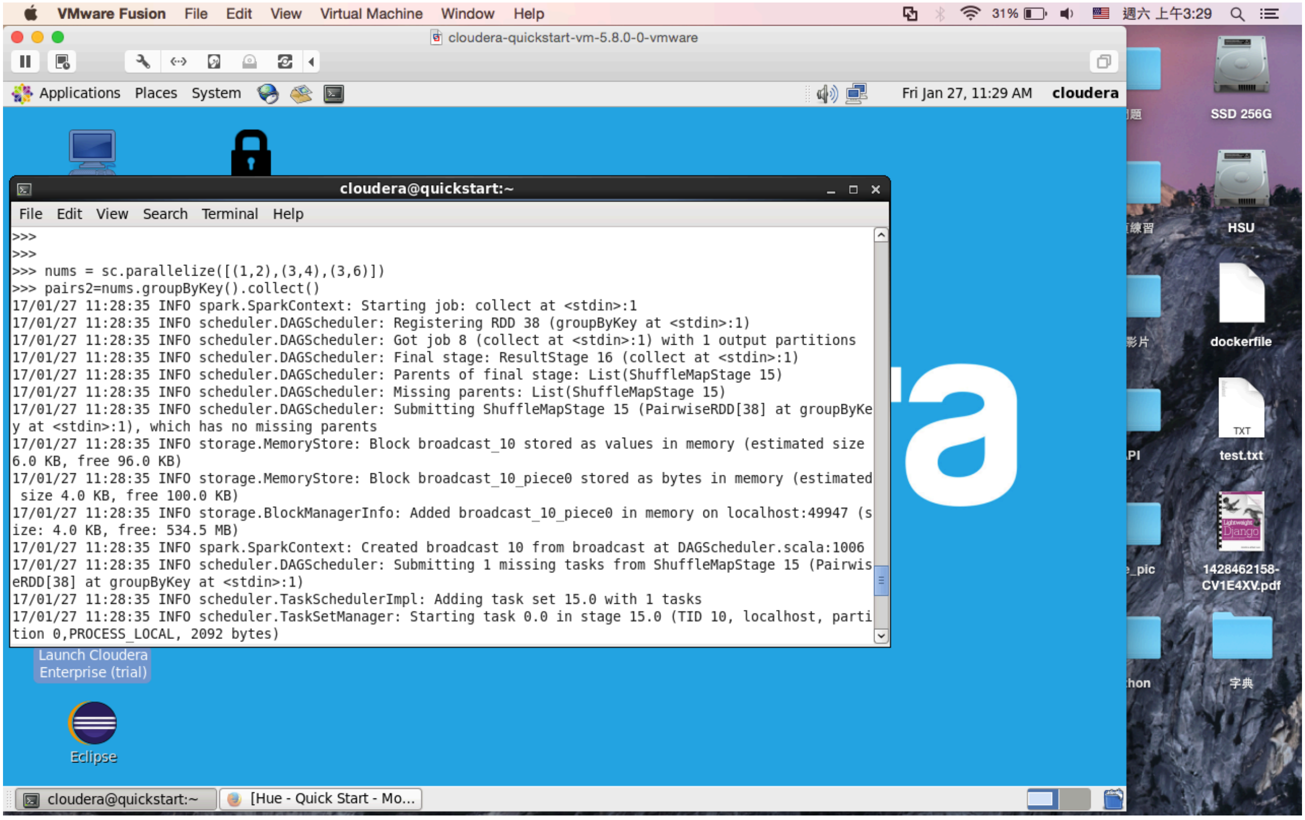

VMware Fusion File Edit View Virtual Machine Window Help
32%
週六 上午3:26
cloudera-quickstart-vm-5.8.0-0-vmware
Applications Places System
Fri Jan 27, 11:26 AM
cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
Launch Cloudera Enterprise (trial)
Eclipse
SSD 256G
HSU
dockerfile
test.txt
1428462158-CV1E4XV.pdf
字典
cloudera@quickstart:~
[Hue - Quick Start - Mo...

```
size 3.1 KB, free 90.1 KB)
17/01/27 11:26:14 INFO storage.BlockManagerInfo: Added broadcast_9_piece0 in memory on localhost:49947 (si
ze: 3.1 KB, free: 534.5 MB)
17/01/27 11:26:14 INFO spark.SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:1006
17/01/27 11:26:14 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 14 (PythonRDD[3
5] at collect at <stdin>:1)
17/01/27 11:26:14 INFO scheduler.TaskSchedulerImpl: Adding task set 14.0 with 1 tasks
17/01/27 11:26:14 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 14.0 (TID 9, localhost, partit
ion 0,NODE_LOCAL, 1894 bytes)
17/01/27 11:26:14 INFO executor.Executor: Running task 0.0 in stage 14.0 (TID 9)
17/01/27 11:26:14 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
17/01/27 11:26:14 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
17/01/27 11:26:14 INFO python.PythonRunner: Times: total = 41, boot = -32, init = 73, finish = 0
17/01/27 11:26:14 INFO executor.Executor: Finished task 0.0 in stage 14.0 (TID 9). 1274 bytes result sent
to driver
17/01/27 11:26:14 INFO scheduler.DAGScheduler: ResultStage 14 (collect at <stdin>:1) finished in 0.066 s
17/01/27 11:26:14 INFO scheduler.DAGScheduler: Job 7 finished: collect at <stdin>:1, took 0.220887 s
17/01/27 11:26:14 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 14.0 (TID 9) in 68 ms on local
host (1/1)
17/01/27 11:26:14 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 14.0, whose tasks have all completed,
from pool
>>> print pairs2
[(1, 2), (3, 10)]
>>> 
```

================

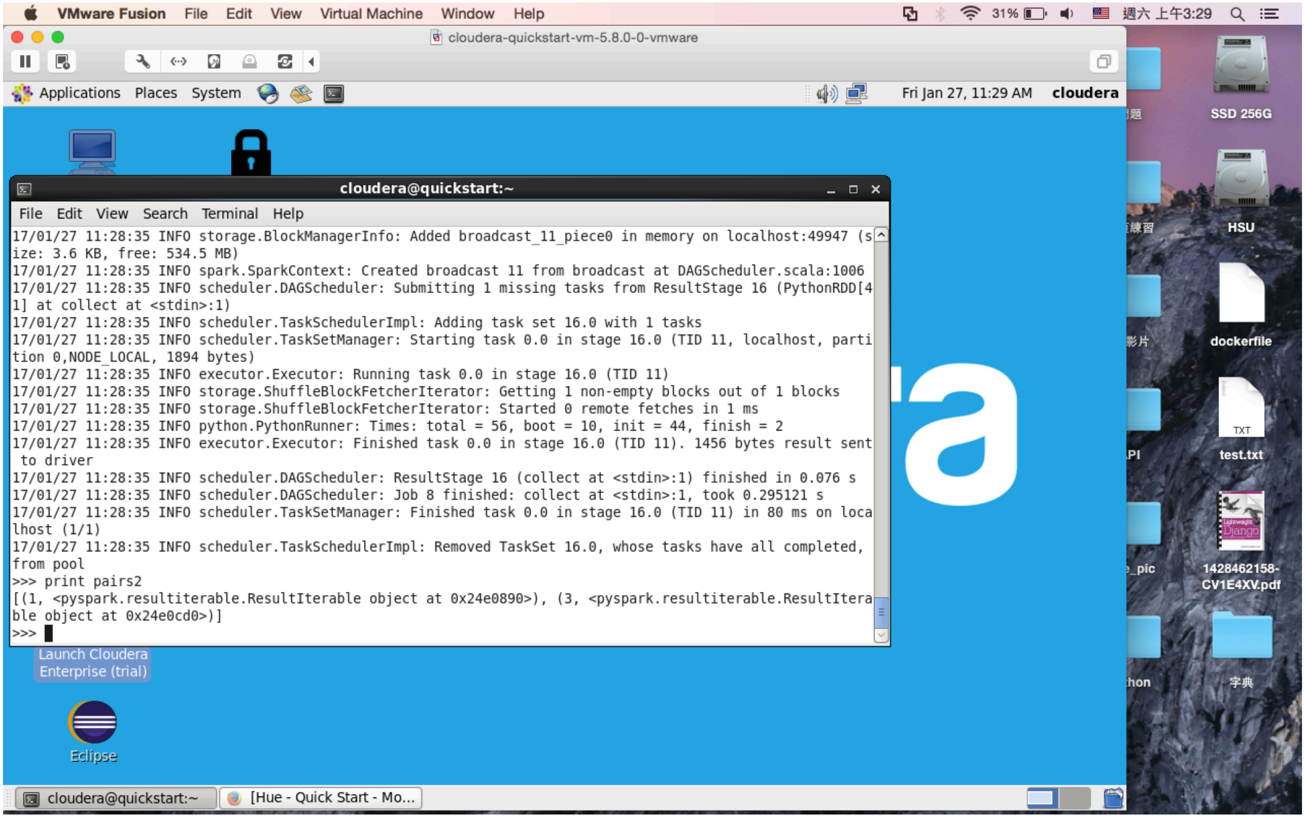
groupByKey

================



VMware Fusion File Edit View Virtual Machine Window Help
cloudera-quickstart-vm-5.8.0-0-vmware
Applications Places System
Fri Jan 27, 11:29 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
Launch Cloudera Enterprise (trial)
Eclipse
cloudera@quickstart:~
[Hue - Quick Start - Mo...
SSD 256G
HSU
dockerfile
test.txt
1428462158-CV1E4XV.pdf
字典

```
>>>
>>>
>>> nums = sc.parallelize([(1,2),(3,4),(3,6)])
>>> pairs2=nums.groupByKey().collect()
17/01/27 11:28:35 INFO spark.SparkContext: Starting job: collect at <stdin>:1
17/01/27 11:28:35 INFO scheduler.DAGScheduler: Registering RDD 38 (groupByKey at <stdin>:1)
17/01/27 11:28:35 INFO scheduler.DAGScheduler: Got job 8 (collect at <stdin>:1) with 1 output partitions
17/01/27 11:28:35 INFO scheduler.DAGScheduler: Final stage: ResultStage 16 (collect at <stdin>:1)
17/01/27 11:28:35 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 15)
17/01/27 11:28:35 INFO scheduler.DAGScheduler: Missing parents: List(ShuffleMapStage 15)
17/01/27 11:28:35 INFO scheduler.DAGScheduler: Submitting ShuffleMapStage 15 (PairwiseRDD[38] at groupByKe
y at <stdin>:1), which has no missing parents
17/01/27 11:28:35 INFO storage.MemoryStore: Block broadcast_10 stored as values in memory (estimated size
6.0 KB, free 96.0 KB)
17/01/27 11:28:35 INFO storage.MemoryStore: Block broadcast_10_piece0 stored as bytes in memory (estimated
 size 4.0 KB, free 100.0 KB)
17/01/27 11:28:35 INFO storage.BlockManagerInfo: Added broadcast_10_piece0 in memory on localhost:49947 (s
ize: 4.0 KB, free: 534.5 MB)
17/01/27 11:28:35 INFO spark.SparkContext: Created broadcast 10 from broadcast at DAGScheduler.scala:1006
17/01/27 11:28:35 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 15 (Pairwis
eRDD[38] at groupByKey at <stdin>:1)
17/01/27 11:28:35 INFO scheduler.TaskSchedulerImpl: Adding task set 15.0 with 1 tasks
17/01/27 11:28:35 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 15.0 (TID 10, localhost, parti
tion 0,PROCESS_LOCAL, 2092 bytes)
```



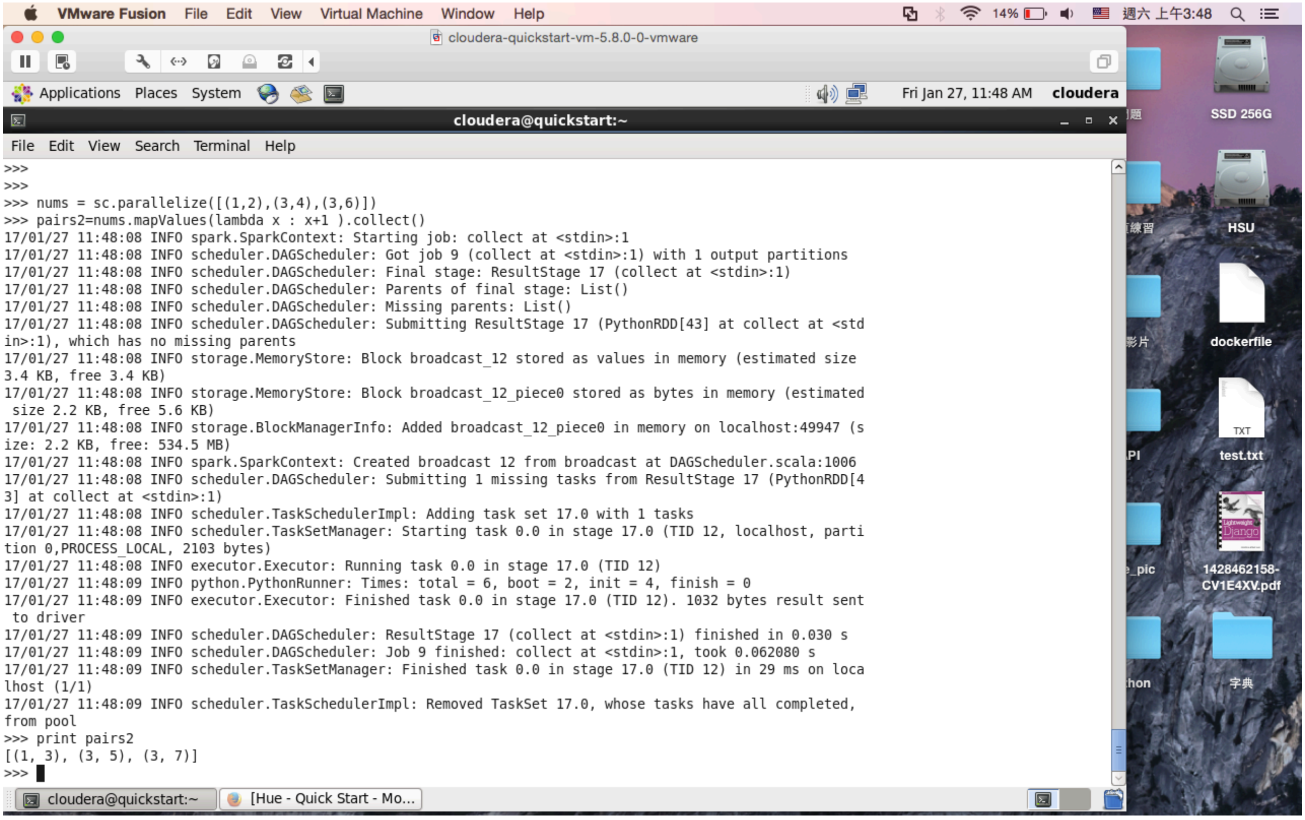
VMware Fusion File Edit View Virtual Machine Window Help
cloudera-quickstart-vm-5.8.0-0-vmware
Applications Places System
Fri Jan 27, 11:29 AM cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
Launch Cloudera Enterprise (trial)
Eclipse
cloudera@quickstart:~
[Hue - Quick Start - Mo...
SSD 256G
HSU
dockerfile
test.txt
1428462158-CV1E4XV.pdf
字典

```
17/01/27 11:28:35 INFO storage.BlockManagerInfo: Added broadcast_11_piece0 in memory on localhost:49947 (s
ize: 3.6 KB, free: 534.5 MB)
17/01/27 11:28:35 INFO spark.SparkContext: Created broadcast 11 from broadcast at DAGScheduler.scala:1006
17/01/27 11:28:35 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 16 (PythonRDD[4
1] at collect at <stdin>:1)
17/01/27 11:28:35 INFO scheduler.TaskSchedulerImpl: Adding task set 16.0 with 1 tasks
17/01/27 11:28:35 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 16.0 (TID 11, localhost, parti
tion 0,NODE_LOCAL, 1894 bytes)
17/01/27 11:28:35 INFO executor.Executor: Running task 0.0 in stage 16.0 (TID 11)
17/01/27 11:28:35 INFO storage.ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
17/01/27 11:28:35 INFO storage.ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
17/01/27 11:28:35 INFO python.PythonRunner: Times: total = 56, boot = 10, init = 44, finish = 2
17/01/27 11:28:35 INFO executor.Executor: Finished task 0.0 in stage 16.0 (TID 11). 1456 bytes result sent
 to driver
17/01/27 11:28:35 INFO scheduler.DAGScheduler: ResultStage 16 (collect at <stdin>:1) finished in 0.076 s
17/01/27 11:28:35 INFO scheduler.DAGScheduler: Job 8 finished: collect at <stdin>:1, took 0.295121 s
17/01/27 11:28:35 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 16.0 (TID 11) in 80 ms on loca
lhost (1/1)
17/01/27 11:28:35 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 16.0, whose tasks have all completed,
from pool
>>> print pairs2
[(1, <pyspark.resultiterable.ResultIterable object at 0x24e0890>), (3, <pyspark.resultiterable.ResultItera
ble object at 0x24e0cd0>)]
>>> 
```

=================

mapValues

=================



```
>>>
>>>
>>> nums = sc.parallelize([(1,2),(3,4),(3,6)])
>>> pairs2=nums.mapValues(lambda x : x+1 ).collect()
17/01/27 11:48:08 INFO spark.SparkContext: Starting job: collect at <stdin>:1
17/01/27 11:48:08 INFO scheduler.DAGScheduler: Got job 9 (collect at <stdin>:1) with 1 output partitions
17/01/27 11:48:08 INFO scheduler.DAGScheduler: Final stage: ResultStage 17 (collect at <stdin>:1)
17/01/27 11:48:08 INFO scheduler.DAGScheduler: Parents of final stage: List()
17/01/27 11:48:08 INFO scheduler.DAGScheduler: Missing parents: List()
17/01/27 11:48:08 INFO scheduler.DAGScheduler: Submitting ResultStage 17 (PythonRDD[43] at collect at <std
in>:1), which has no missing parents
17/01/27 11:48:08 INFO storage.MemoryStore: Block broadcast_12 stored as values in memory (estimated size
3.4 KB, free 3.4 KB)
17/01/27 11:48:08 INFO storage.MemoryStore: Block broadcast_12_piece0 stored as bytes in memory (estimated
 size 2.2 KB, free 5.6 KB)
17/01/27 11:48:08 INFO storage.BlockManagerInfo: Added broadcast_12_piece0 in memory on localhost:49947 (s
ize: 2.2 KB, free: 534.5 MB)
17/01/27 11:48:08 INFO spark.SparkContext: Created broadcast 12 from broadcast at DAGScheduler.scala:1006
17/01/27 11:48:08 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 17 (PythonRDD[4
3] at collect at <stdin>:1)
17/01/27 11:48:08 INFO scheduler.TaskSchedulerImpl: Adding task set 17.0 with 1 tasks
17/01/27 11:48:08 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 17.0 (TID 12, localhost, parti
tion 0,PROCESS_LOCAL, 2103 bytes)
17/01/27 11:48:08 INFO executor.Executor: Running task 0.0 in stage 17.0 (TID 12)
17/01/27 11:48:09 INFO python.PythonRunner: Times: total = 6, boot = 2, init = 4, finish = 0
17/01/27 11:48:09 INFO executor.Executor: Finished task 0.0 in stage 17.0 (TID 12). 1032 bytes result sent
 to driver
17/01/27 11:48:09 INFO scheduler.DAGScheduler: ResultStage 17 (collect at <stdin>:1) finished in 0.030 s
17/01/27 11:48:09 INFO scheduler.DAGScheduler: Job 9 finished: collect at <stdin>:1, took 0.062080 s
17/01/27 11:48:09 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 17.0 (TID 12) in 29 ms on loca
lhost (1/1)
17/01/27 11:48:09 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 17.0, whose tasks have all completed,
from pool
>>> print pairs2
[(1, 3), (3, 5), (3, 7)]
>>>
```