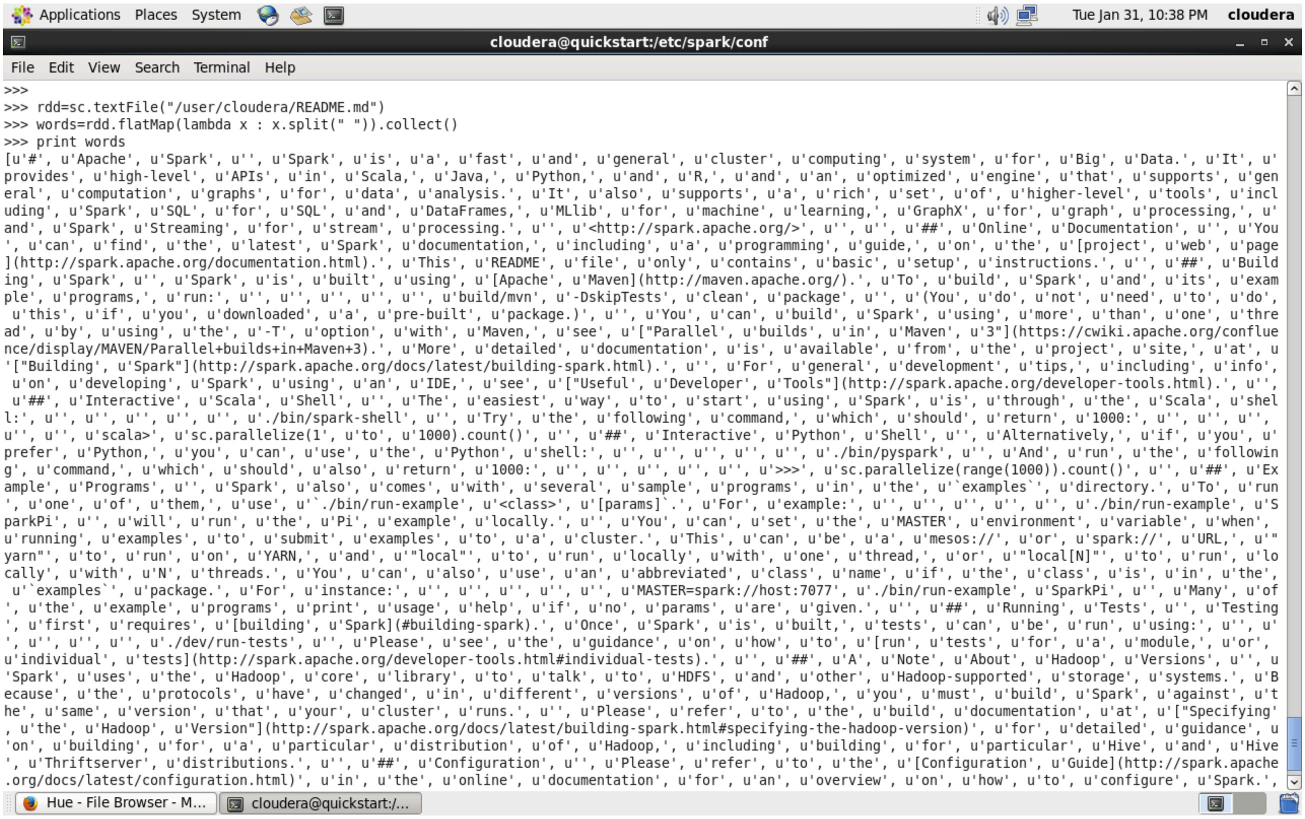
2017-02-01 Spark 學習筆記

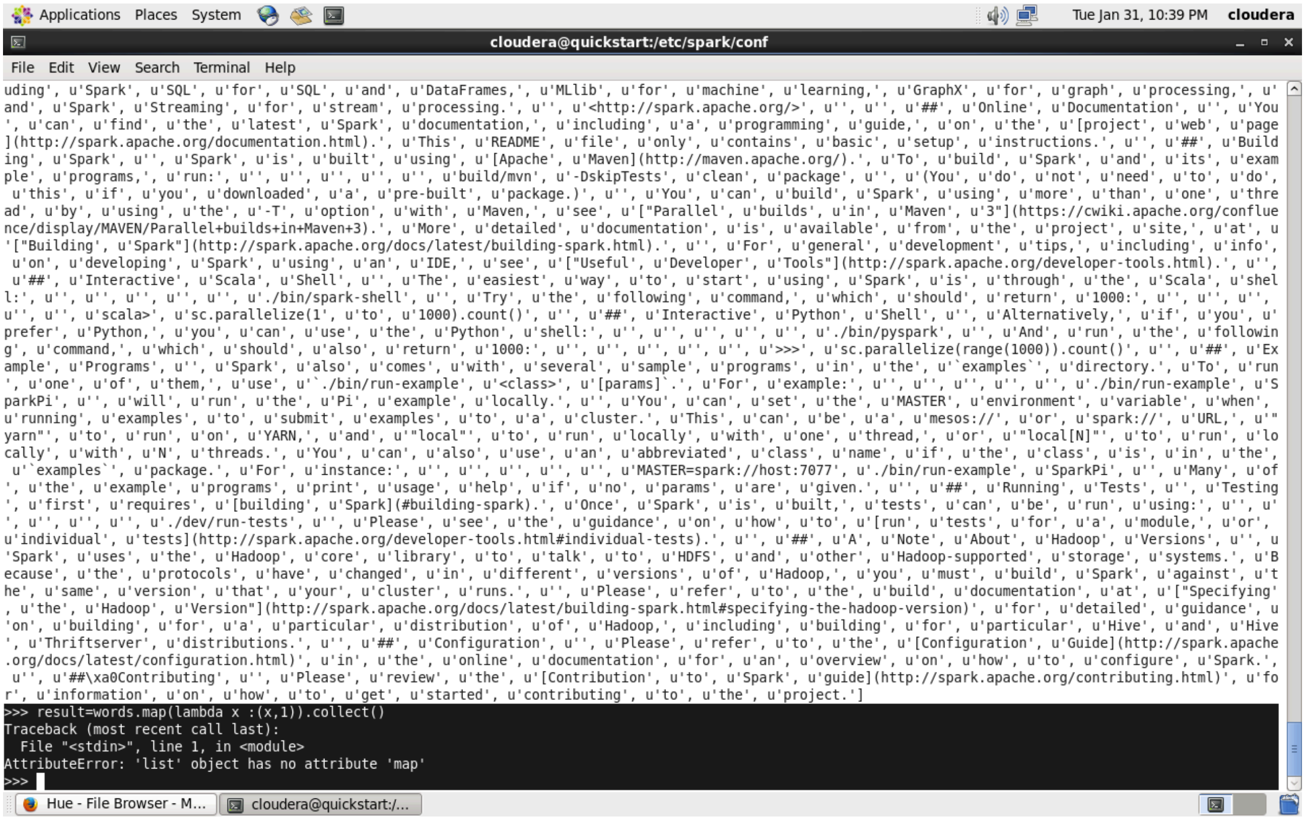
=================
用flatMap+ReduceByKey做WordCount計算
參考：Spark學習手冊 p.53
=================

Step1：先將README.md上傳到HDFS上面
Step2：sc.textFile()讀取檔案，並且建立一個RDD字串
Step3：使用flatMap()產生一個單字以及數字1的pairRDD，並使用collect()接收整個RDD


Applications Places System
Tue Jan 31, 10:38 PM cloudera
cloudera@quickstart:/etc/spark/conf
File Edit View Search Terminal Help
Hue - File Browser - M...
cloudera@quickstart:/...

```
>>>
>>> rdd=sc.textFile("/user/cloudera/README.md")
>>> words=rdd.flatMap(lambda x : x.split(" ")).collect()
>>> print words
[u'#', u'Apache', u'Spark', u'', u'Spark', u'is', u'a', u'fast', u'and', u'general', u'cluster', u'computing', u'system', u'for', u'Big', u'Data.', u'It', u'
provides', u'high-level', u'APIs', u'in', u'Scala,', u'Java,', u'Python,', u'and', u'R,', u'and', u'an', u'optimized', u'engine', u'that', u'supports', u'gen
eral', u'computation', u'graphs', u'for', u'data', u'analysis.', u'It', u'also', u'supports', u'a', u'rich', u'set', u'of', u'higher-level', u'tools', u'incl
uding', u'Spark', u'SQL', u'for', u'SQL', u'and', u'DataFrames,', u'MLlib', u'for', u'machine', u'learning,', u'GraphX', u'for', u'graph', u'processing,', u'
and', u'Spark', u'Streaming', u'for', u'stream', u'processing.', u'', u'<http://spark.apache.org/>', u'', u'', u'##', u'Online', u'Documentation', u'', u'You
', u'can', u'find', u'the', u'latest', u'Spark', u'documentation,', u'including', u'a', u'programming', u'guide,', u'on', u'the', u'[project', u'web', u'page
](http://spark.apache.org/documentation.html).', u'This', u'README', u'file', u'only', u'contains', u'basic', u'setup', u'instructions.', u'', u'##', u'Build
ing', u'Spark', u'', u'Spark', u'is', u'built', u'using', u'[Apache', u'Maven](http://maven.apache.org/).', u'To', u'build', u'Spark', u'and', u'its', u'exam
ple', u'programs,', u'run:', u'', u'', u'', u'', u'', u'build/mvn', u'-DskipTests', u'clean', u'package', u'', u'(You', u'do', u'not', u'need', u'to', u'do',
 u'this', u'if', u'you', u'downloaded', u'a', u'pre-built', u'package.)', u'', u'You', u'can', u'build', u'Spark', u'using', u'more', u'than', u'one', u'thre
ad', u'by', u'using', u'the', u'-T', u'option', u'with', u'Maven,', u'see', u'["Parallel', u'builds', u'in', u'Maven', u'3"](https://cwiki.apache.org/conflue
nce/display/MAVEN/Parallel+builds+in+Maven+3).', u'More', u'detailed', u'documentation', u'is', u'available', u'from', u'the', u'project', u'site,', u'at', u
'["Building', u'Spark"](http://spark.apache.org/docs/latest/building-spark.html).', u'', u'For', u'general', u'development', u'tips,', u'including', u'info',
 u'on', u'developing', u'Spark', u'using', u'an', u'IDE,', u'see', u'["Useful', u'Developer', u'Tools"](http://spark.apache.org/developer-tools.html).', u'',
 u'##', u'Interactive', u'Scala', u'Shell', u'', u'The', u'easiest', u'way', u'to', u'start', u'using', u'Spark', u'is', u'through', u'the', u'Scala', u'shel
l:', u'', u'', u'', u'', u'', u'./bin/spark-shell', u'', u'Try', u'the', u'following', u'command,', u'which', u'should', u'return', u'1000:', u'', u'', u'',
u'', u'', u'scala>', u'sc.parallelize(1', u'to', u'1000).count()', u'', u'##', u'Interactive', u'Python', u'Shell', u'', u'Alternatively,', u'if', u'you', u'
prefer', u'Python,', u'you', u'can', u'use', u'the', u'Python', u'shell:', u'', u'', u'', u'', u'', u'./bin/pyspark', u'', u'And', u'run', u'the', u'followin
g', u'command,', u'which', u'should', u'also', u'return', u'1000:', u'', u'', u'', u'', u'', u'>>>', u'sc.parallelize(range(1000)).count()', u'', u'##', u'Ex
ample', u'Programs', u'', u'Spark', u'also', u'comes', u'with', u'several', u'sample', u'programs', u'in', u'the', u'`examples`', u'directory.', u'To', u'run
', u'one', u'of', u'them,', u'use', u'`./bin/run-example', u'<class>', u'[params]`.', u'For', u'example:', u'', u'', u'', u'', u'', u'./bin/run-example', u'S
parkPi', u'', u'will', u'run', u'the', u'Pi', u'example', u'locally.', u'', u'You', u'can', u'set', u'the', u'MASTER', u'environment', u'variable', u'when',
u'running', u'examples', u'to', u'submit', u'examples', u'to', u'a', u'cluster.', u'This', u'can', u'be', u'a', u'mesos://', u'or', u'spark://', u'URL,', u'"
yarn"', u'to', u'run', u'on', u'YARN,', u'and', u'"local"', u'to', u'run', u'locally', u'with', u'one', u'thread,', u'or', u'"local[N]"', u'to', u'run', u'lo
cally', u'with', u'N', u'threads.', u'You', u'can', u'also', u'use', u'an', u'abbreviated', u'class', u'name', u'if', u'the', u'class', u'is', u'in', u'the',
 u'`examples`', u'package.', u'For', u'instance:', u'', u'', u'', u'', u'', u'MASTER=spark://host:7077', u'./bin/run-example', u'SparkPi', u'', u'Many', u'of
', u'the', u'example', u'programs', u'print', u'usage', u'help', u'if', u'no', u'params', u'are', u'given.', u'', u'##', u'Running', u'Tests', u'', u'Testing
', u'first', u'requires', u'[building', u'Spark](#building-spark).', u'Once', u'Spark', u'is', u'built,', u'tests', u'can', u'be', u'run', u'using:', u'', u'
', u'', u'', u'', u'./dev/run-tests', u'', u'Please', u'see', u'the', u'guidance', u'on', u'how', u'to', u'[run', u'tests', u'for', u'a', u'module,', u'or',
u'individual', u'tests](http://spark.apache.org/developer-tools.html#individual-tests).', u'', u'##', u'A', u'Note', u'About', u'Hadoop', u'Versions', u'', u
'Spark', u'uses', u'the', u'Hadoop', u'core', u'library', u'to', u'talk', u'to', u'HDFS', u'and', u'other', u'Hadoop-supported', u'storage', u'systems.', u'B
ecause', u'the', u'protocols', u'have', u'changed', u'in', u'different', u'versions', u'of', u'Hadoop,', u'you', u'must', u'build', u'Spark', u'against', u't
he', u'same', u'version', u'that', u'your', u'cluster', u'runs.', u'', u'Please', u'refer', u'to', u'the', u'build', u'documentation', u'at', u'["Specifying'
, u'the', u'Hadoop', u'Version"](http://spark.apache.org/docs/latest/building-spark.html#specifying-the-hadoop-version)', u'for', u'detailed', u'guidance', u
'on', u'building', u'for', u'a', u'particular', u'distribution', u'of', u'Hadoop,', u'including', u'building', u'for', u'particular', u'Hive', u'and', u'Hive
', u'Thriftserver', u'distributions.', u'', u'##', u'Configuration', u'', u'Please', u'refer', u'to', u'the', u'[Configuration', u'Guide](http://spark.apache
.org/docs/latest/configuration.html)', u'in', u'the', u'online', u'documentation', u'for', u'an', u'overview', u'on', u'how', u'to', u'configure', u'Spark.',
```

Step4：使用flatMap()產生一個單字以及數字1的pairRDD，並使用collect()接收整個RDD
結果就產生Error了！！！！～～～～


Applications Places System
Tue Jan 31, 10:39 PM cloudera
cloudera@quickstart:/etc/spark/conf
File Edit View Search Terminal Help
Hue - File Browser - M...
cloudera@quickstart:/...

```
uding', u'Spark', u'SQL', u'for', u'SQL', u'and', u'DataFrames,', u'MLlib', u'for', u'machine', u'learning,', u'GraphX', u'for', u'graph', u'processing,', u'
and', u'Spark', u'Streaming', u'for', u'stream', u'processing.', u'', u'<http://spark.apache.org/>', u'', u'', u'##', u'Online', u'Documentation', u'', u'You
', u'can', u'find', u'the', u'latest', u'Spark', u'documentation,', u'including', u'a', u'programming', u'guide,', u'on', u'the', u'[project', u'web', u'page
](http://spark.apache.org/documentation.html).', u'This', u'README', u'file', u'only', u'contains', u'basic', u'setup', u'instructions.', u'', u'##', u'Build
ing', u'Spark', u'', u'Spark', u'is', u'built', u'using', u'[Apache', u'Maven](http://maven.apache.org/).', u'To', u'build', u'Spark', u'and', u'its', u'exam
ple', u'programs,', u'run:', u'', u'', u'', u'', u'', u'build/mvn', u'-DskipTests', u'clean', u'package', u'', u'(You', u'do', u'not', u'need', u'to', u'do',
 u'this', u'if', u'you', u'downloaded', u'a', u'pre-built', u'package.)', u'', u'You', u'can', u'build', u'Spark', u'using', u'more', u'than', u'one', u'thre
ad', u'by', u'using', u'the', u'-T', u'option', u'with', u'Maven,', u'see', u'["Parallel', u'builds', u'in', u'Maven', u'3"](https://cwiki.apache.org/conflue
nce/display/MAVEN/Parallel+builds+in+Maven+3).', u'More', u'detailed', u'documentation', u'is', u'available', u'from', u'the', u'project', u'site,', u'at', u
'["Building', u'Spark"](http://spark.apache.org/docs/latest/building-spark.html).', u'', u'For', u'general', u'development', u'tips,', u'including', u'info',
 u'on', u'developing', u'Spark', u'using', u'an', u'IDE,', u'see', u'["Useful', u'Developer', u'Tools"](http://spark.apache.org/developer-tools.html).', u'',
 u'##', u'Interactive', u'Scala', u'Shell', u'', u'The', u'easiest', u'way', u'to', u'start', u'using', u'Spark', u'is', u'through', u'the', u'Scala', u'shel
l:', u'', u'', u'', u'', u'', u'./bin/spark-shell', u'', u'Try', u'the', u'following', u'command,', u'which', u'should', u'return', u'1000:', u'', u'', u'',
u'', u'', u'scala>', u'sc.parallelize(1', u'to', u'1000).count()', u'', u'##', u'Interactive', u'Python', u'Shell', u'', u'Alternatively,', u'if', u'you', u'
prefer', u'Python,', u'you', u'can', u'use', u'the', u'Python', u'shell:', u'', u'', u'', u'', u'', u'./bin/pyspark', u'', u'And', u'run', u'the', u'followin
g', u'command,', u'which', u'should', u'also', u'return', u'1000:', u'', u'', u'', u'', u'', u'>>>', u'sc.parallelize(range(1000)).count()', u'', u'##', u'Ex
ample', u'Programs', u'', u'Spark', u'also', u'comes', u'with', u'several', u'sample', u'programs', u'in', u'the', u'`examples`', u'directory.', u'To', u'run
', u'one', u'of', u'them,', u'use', u'`./bin/run-example', u'<class>', u'[params]`.', u'For', u'example:', u'', u'', u'', u'', u'', u'./bin/run-example', u'S
parkPi', u'', u'will', u'run', u'the', u'Pi', u'example', u'locally.', u'', u'You', u'can', u'set', u'the', u'MASTER', u'environment', u'variable', u'when',
u'running', u'examples', u'to', u'submit', u'examples', u'to', u'a', u'cluster.', u'This', u'can', u'be', u'a', u'mesos://', u'or', u'spark://', u'URL,', u'"
yarn"', u'to', u'run', u'on', u'YARN,', u'and', u'"local"', u'to', u'run', u'locally', u'with', u'one', u'thread,', u'or', u'"local[N]"', u'to', u'run', u'lo
cally', u'with', u'N', u'threads.', u'You', u'can', u'also', u'use', u'an', u'abbreviated', u'class', u'name', u'if', u'the', u'class', u'is', u'in', u'the',
 u'`examples`', u'package.', u'For', u'instance:', u'', u'', u'', u'', u'', u'MASTER=spark://host:7077', u'./bin/run-example', u'SparkPi', u'', u'Many', u'of
', u'the', u'example', u'programs', u'print', u'usage', u'help', u'if', u'no', u'params', u'are', u'given.', u'', u'##', u'Running', u'Tests', u'', u'Testing
', u'first', u'requires', u'[building', u'Spark](#building-spark).', u'Once', u'Spark', u'is', u'built,', u'tests', u'can', u'be', u'run', u'using:', u'', u'
', u'', u'', u'', u'./dev/run-tests', u'', u'Please', u'see', u'the', u'guidance', u'on', u'how', u'to', u'[run', u'tests', u'for', u'a', u'module,', u'or',
u'individual', u'tests](http://spark.apache.org/developer-tools.html#individual-tests).', u'', u'##', u'A', u'Note', u'About', u'Hadoop', u'Versions', u'', u
'Spark', u'uses', u'the', u'Hadoop', u'core', u'library', u'to', u'talk', u'to', u'HDFS', u'and', u'other', u'Hadoop-supported', u'storage', u'systems.', u'B
ecause', u'the', u'protocols', u'have', u'changed', u'in', u'different', u'versions', u'of', u'Hadoop,', u'you', u'must', u'build', u'Spark', u'against', u't
he', u'same', u'version', u'that', u'your', u'cluster', u'runs.', u'', u'Please', u'refer', u'to', u'the', u'build', u'documentation', u'at', u'["Specifying'
, u'the', u'Hadoop', u'Version"](http://spark.apache.org/docs/latest/building-spark.html#specifying-the-hadoop-version)', u'for', u'detailed', u'guidance', u
'on', u'building', u'for', u'a', u'particular', u'distribution', u'of', u'Hadoop,', u'including', u'building', u'for', u'particular', u'Hive', u'and', u'Hive
', u'Thriftserver', u'distributions.', u'', u'##', u'Configuration', u'', u'Please', u'refer', u'to', u'the', u'[Configuration', u'Guide](http://spark.apache
.org/docs/latest/configuration.html)', u'in', u'the', u'online', u'documentation', u'for', u'an', u'overview', u'on', u'how', u'to', u'configure', u'Spark.',
 u'', u'##\xa0Contributing', u'', u'Please', u'review', u'the', u'[Contribution', u'to', u'Spark', u'guide](http://spark.apache.org/contributing.html)', u'fo
r', u'information', u'on', u'how', u'to', u'get', u'started', u'contributing', u'to', u'the', u'project.']
>>> result=words.map(lambda x :(x,1)).collect()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'list' object has no attribute 'map'
>>> 
```

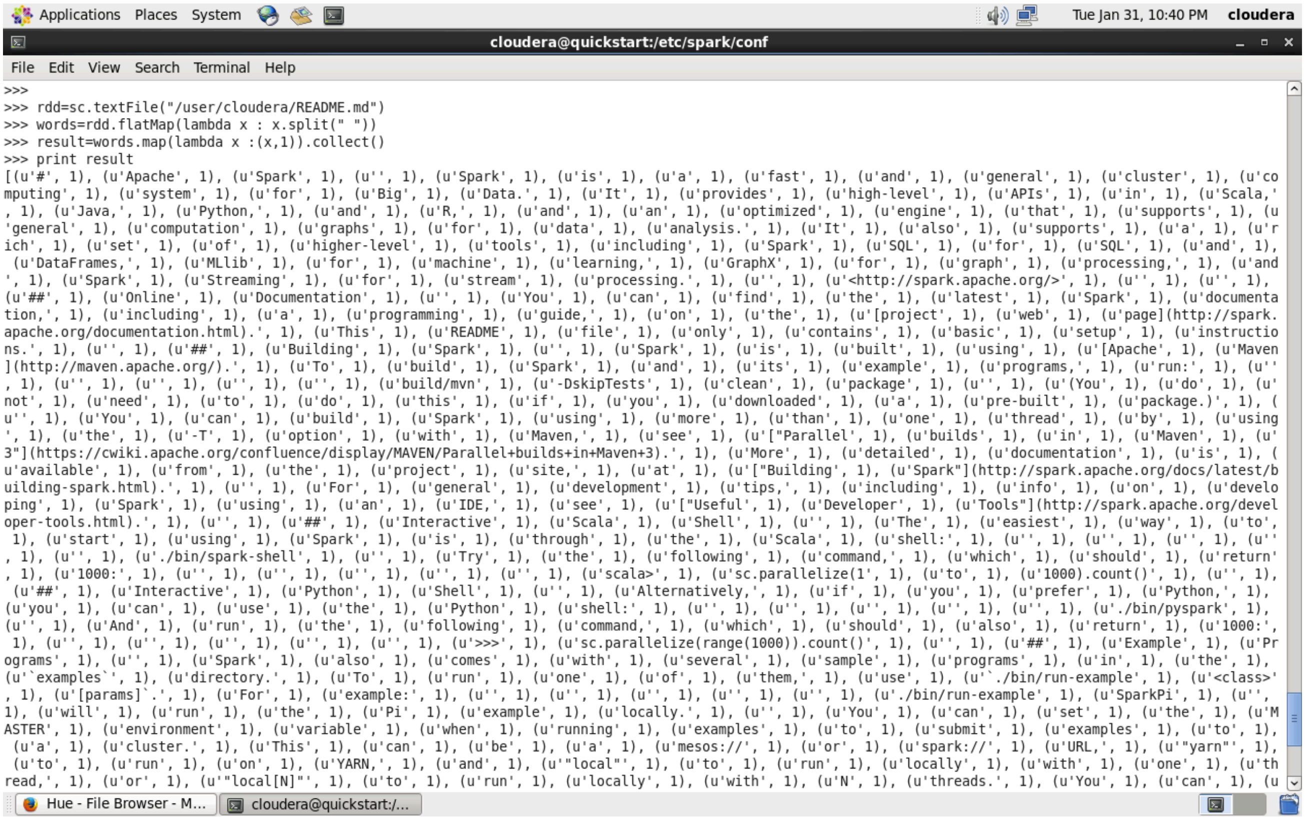
===================

Error錯誤排解

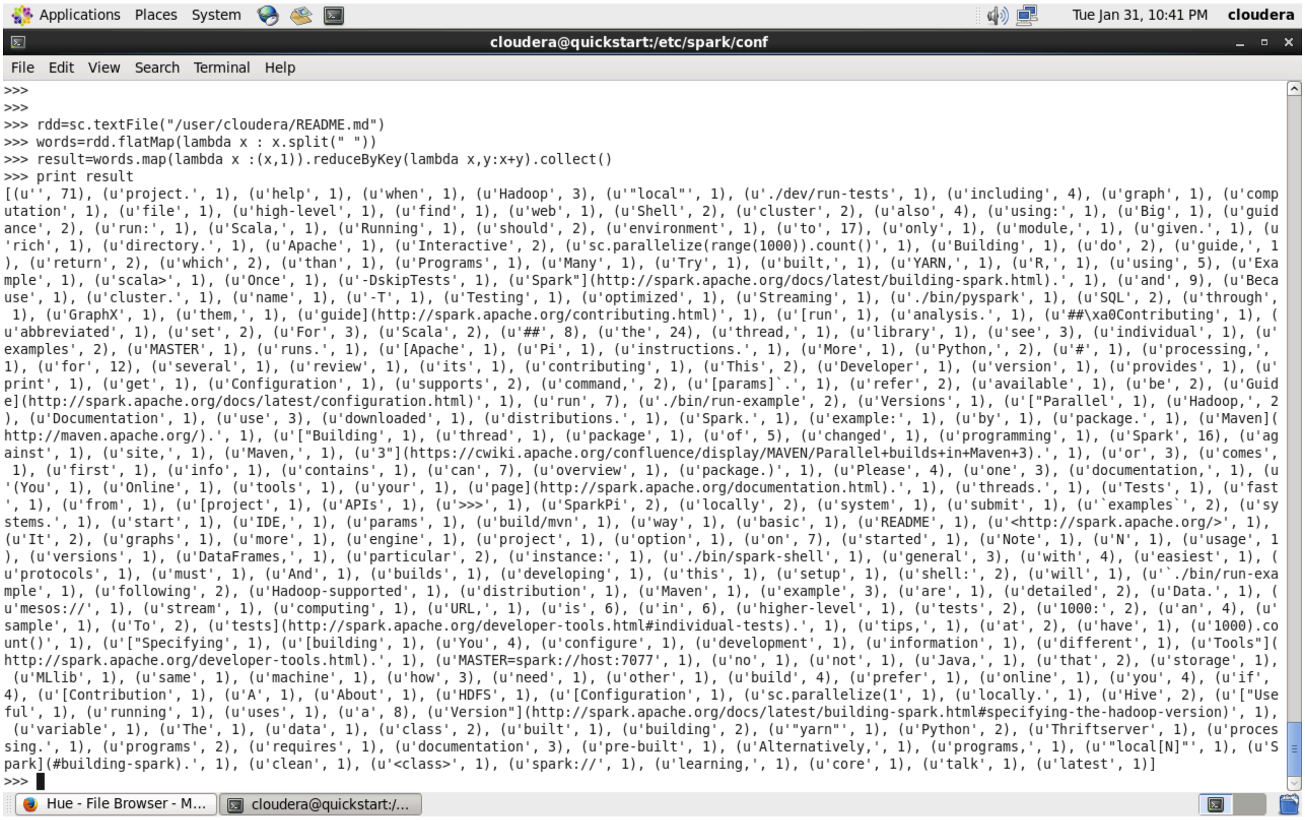
#原因：因為如果在flapMap的時候使用collect，command回傳的型態是list而不是一個RDD，所以只有把它移掉即可

#參考網站：http://stackoverflow.com/questions/40281834/list-object-has-no-attribute-map

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



```
>>>
>>> rdd=sc.textFile("/user/cloudera/README.md")
>>> words=rdd.flatMap(lambda x : x.split(" "))
>>> result=words.map(lambda x :(x,1)).collect()
>>> print result
[(u'#', 1), (u'Apache', 1), (u'Spark', 1), (u'', 1), (u'Spark', 1), (u'is', 1), (u'a', 1), (u'fast', 1), (u'and', 1), (u'general', 1), (u'cluster', 1), (u'co
mputing', 1), (u'system', 1), (u'for', 1), (u'Big', 1), (u'Data.', 1), (u'It', 1), (u'provides', 1), (u'high-level', 1), (u'APIs', 1), (u'in', 1), (u'Scala,'
, 1), (u'Java,', 1), (u'Python,', 1), (u'and', 1), (u'R,', 1), (u'and', 1), (u'an', 1), (u'optimized', 1), (u'engine', 1), (u'that', 1), (u'supports', 1), (u
'general', 1), (u'computation', 1), (u'graphs', 1), (u'for', 1), (u'data', 1), (u'analysis.', 1), (u'It', 1), (u'also', 1), (u'supports', 1), (u'a', 1), (u'r
ich', 1), (u'set', 1), (u'of', 1), (u'higher-level', 1), (u'tools', 1), (u'including', 1), (u'Spark', 1), (u'SQL', 1), (u'for', 1), (u'SQL', 1), (u'and', 1),
 (u'DataFrames,', 1), (u'MLlib', 1), (u'for', 1), (u'machine', 1), (u'learning,', 1), (u'GraphX', 1), (u'for', 1), (u'graph', 1), (u'processing,', 1), (u'and
', 1), (u'Spark', 1), (u'Streaming', 1), (u'for', 1), (u'stream', 1), (u'processing.', 1), (u'', 1), (u'<http://spark.apache.org/>', 1), (u'', 1), (u'', 1),
(u'##', 1), (u'Online', 1), (u'Documentation', 1), (u'', 1), (u'You', 1), (u'can', 1), (u'find', 1), (u'the', 1), (u'latest', 1), (u'Spark', 1), (u'documenta
tion,', 1), (u'including', 1), (u'a', 1), (u'programming', 1), (u'guide,', 1), (u'on', 1), (u'the', 1), (u'[project', 1), (u'web', 1), (u'page](http://spark.
apache.org/documentation.html).', 1), (u'This', 1), (u'README', 1), (u'file', 1), (u'only', 1), (u'contains', 1), (u'basic', 1), (u'setup', 1), (u'instructio
ns.', 1), (u'', 1), (u'##', 1), (u'Building', 1), (u'Spark', 1), (u'', 1), (u'Spark', 1), (u'is', 1), (u'built', 1), (u'using', 1), (u'[Apache', 1), (u'Maven
](http://maven.apache.org/).', 1), (u'To', 1), (u'build', 1), (u'Spark', 1), (u'and', 1), (u'its', 1), (u'example', 1), (u'programs,', 1), (u'run:', 1), (u''
, 1), (u'', 1), (u'', 1), (u'', 1), (u'', 1), (u'build/mvn', 1), (u'-DskipTests', 1), (u'clean', 1), (u'package', 1), (u'', 1), (u'(You', 1), (u'do', 1), (u'
not', 1), (u'need', 1), (u'to', 1), (u'do', 1), (u'this', 1), (u'if', 1), (u'you', 1), (u'downloaded', 1), (u'a', 1), (u'pre-built', 1), (u'package.)', 1), (
u'', 1), (u'You', 1), (u'can', 1), (u'build', 1), (u'Spark', 1), (u'using', 1), (u'more', 1), (u'than', 1), (u'one', 1), (u'thread', 1), (u'by', 1), (u'using
', 1), (u'the', 1), (u'-T', 1), (u'option', 1), (u'with', 1), (u'Maven,', 1), (u'see', 1), (u'["Parallel', 1), (u'builds', 1), (u'in', 1), (u'Maven', 1), (u'
3"](https://cwiki.apache.org/confluence/display/MAVEN/Parallel+builds+in+Maven+3).', 1), (u'More', 1), (u'detailed', 1), (u'documentation', 1), (u'is', 1), (
u'available', 1), (u'from', 1), (u'the', 1), (u'project', 1), (u'site,', 1), (u'at', 1), (u'["Building', 1), (u'Spark"](http://spark.apache.org/docs/latest/b
uilding-spark.html).', 1), (u'', 1), (u'For', 1), (u'general', 1), (u'development', 1), (u'tips,', 1), (u'including', 1), (u'info', 1), (u'on', 1), (u'develo
ping', 1), (u'Spark', 1), (u'using', 1), (u'an', 1), (u'IDE,', 1), (u'see', 1), (u'["Useful', 1), (u'Developer', 1), (u'Tools"](http://spark.apache.org/devel
oper-tools.html).', 1), (u'', 1), (u'##', 1), (u'Interactive', 1), (u'Scala', 1), (u'Shell', 1), (u'', 1), (u'The', 1), (u'easiest', 1), (u'way', 1), (u'to',
 1), (u'start', 1), (u'using', 1), (u'Spark', 1), (u'is', 1), (u'through', 1), (u'the', 1), (u'Scala', 1), (u'shell:', 1), (u'', 1), (u'', 1), (u'', 1), (u''
, 1), (u'', 1), (u'./bin/spark-shell', 1), (u'', 1), (u'Try', 1), (u'the', 1), (u'following', 1), (u'command,', 1), (u'which', 1), (u'should', 1), (u'return'
, 1), (u'1000:', 1), (u'', 1), (u'', 1), (u'', 1), (u'', 1), (u'', 1), (u'scala>', 1), (u'sc.parallelize(1', 1), (u'to', 1), (u'1000).count()', 1), (u'', 1),
 (u'##', 1), (u'Interactive', 1), (u'Python', 1), (u'Shell', 1), (u'', 1), (u'Alternatively,', 1), (u'if', 1), (u'you', 1), (u'prefer', 1), (u'Python,', 1),
(u'you', 1), (u'can', 1), (u'use', 1), (u'the', 1), (u'Python', 1), (u'shell:', 1), (u'', 1), (u'', 1), (u'', 1), (u'', 1), (u'', 1), (u'./bin/pyspark', 1),
(u'', 1), (u'And', 1), (u'run', 1), (u'the', 1), (u'following', 1), (u'command,', 1), (u'which', 1), (u'should', 1), (u'also', 1), (u'return', 1), (u'1000:',
 1), (u'', 1), (u'', 1), (u'', 1), (u'', 1), (u'', 1), (u'>>>', 1), (u'sc.parallelize(range(1000)).count()', 1), (u'', 1), (u'##', 1), (u'Example', 1), (u'Pr
ograms', 1), (u'', 1), (u'Spark', 1), (u'also', 1), (u'comes', 1), (u'with', 1), (u'several', 1), (u'sample', 1), (u'programs', 1), (u'in', 1), (u'the', 1),
(u'`examples`', 1), (u'directory.', 1), (u'To', 1), (u'run', 1), (u'one', 1), (u'of', 1), (u'them,', 1), (u'use', 1), (u'`./bin/run-example', 1), (u'<class>'
, 1), (u'[params]`.', 1), (u'For', 1), (u'example:', 1), (u'', 1), (u'', 1), (u'', 1), (u'', 1), (u'', 1), (u'./bin/run-example', 1), (u'SparkPi', 1), (u'',
1), (u'will', 1), (u'run', 1), (u'the', 1), (u'Pi', 1), (u'example', 1), (u'locally.', 1), (u'', 1), (u'You', 1), (u'can', 1), (u'set', 1), (u'the', 1), (u'M
ASTER', 1), (u'environment', 1), (u'variable', 1), (u'when', 1), (u'running', 1), (u'examples', 1), (u'to', 1), (u'submit', 1), (u'examples', 1), (u'to', 1),
 (u'a', 1), (u'cluster.', 1), (u'This', 1), (u'can', 1), (u'be', 1), (u'a', 1), (u'mesos://', 1), (u'or', 1), (u'spark://', 1), (u'URL,', 1), (u'"yarn"', 1),
 (u'to', 1), (u'run', 1), (u'on', 1), (u'YARN,', 1), (u'and', 1), (u'"local"', 1), (u'to', 1), (u'run', 1), (u'locally', 1), (u'with', 1), (u'one', 1), (u'th
read,', 1), (u'or', 1), (u'"local[N]"', 1), (u'to', 1), (u'run', 1), (u'locally', 1), (u'with', 1), (u'N', 1), (u'threads.', 1), (u'You', 1), (u'can', 1), (u
```



```
>>>
>>>
>>> rdd=sc.textFile("/user/cloudera/README.md")
>>> words=rdd.flatMap(lambda x : x.split(" "))
>>> result=words.map(lambda x :(x,1)).reduceByKey(lambda x,y:x+y).collect()
>>> print result
[(u'', 71), (u'project.', 1), (u'help', 1), (u'when', 1), (u'Hadoop', 3), (u'"local"', 1), (u'./dev/run-tests', 1), (u'including', 4), (u'graph', 1), (u'comp
utation', 1), (u'file', 1), (u'high-level', 1), (u'find', 1), (u'web', 1), (u'Shell', 2), (u'cluster', 2), (u'also', 4), (u'using:', 1), (u'Big', 1), (u'guid
ance', 2), (u'run:', 1), (u'Scala,', 1), (u'Running', 1), (u'should', 2), (u'environment', 1), (u'to', 17), (u'only', 1), (u'module,', 1), (u'given.', 1), (u
'rich', 1), (u'directory.', 1), (u'Apache', 1), (u'Interactive', 2), (u'sc.parallelize(range(1000)).count()', 1), (u'Building', 1), (u'do', 2), (u'guide,', 1
), (u'return', 2), (u'which', 2), (u'than', 1), (u'Programs', 1), (u'Many', 1), (u'Try', 1), (u'built,', 1), (u'YARN,', 1), (u'R,', 1), (u'using', 5), (u'Exa
mple', 1), (u'scala>', 1), (u'Once', 1), (u'-DskipTests', 1), (u'Spark"](http://spark.apache.org/docs/latest/building-spark.html).', 1), (u'and', 9), (u'Beca
use', 1), (u'cluster.', 1), (u'name', 1), (u'-T', 1), (u'Testing', 1), (u'optimized', 1), (u'Streaming', 1), (u'./bin/pyspark', 1), (u'SQL', 2), (u'through',
 1), (u'GraphX', 1), (u'them,', 1), (u'guide](http://spark.apache.org/contributing.html)', 1), (u'[run', 1), (u'analysis.', 1), (u'##\xa0Contributing', 1), (
u'abbreviated', 1), (u'set', 2), (u'For', 3), (u'Scala', 2), (u'##', 8), (u'the', 24), (u'thread,', 1), (u'library', 1), (u'see', 3), (u'individual', 1), (u'
examples', 2), (u'MASTER', 1), (u'runs.', 1), (u'[Apache', 1), (u'Pi', 1), (u'instructions.', 1), (u'More', 1), (u'Python,', 2), (u'#', 1), (u'processing,',
1), (u'for', 12), (u'several', 1), (u'review', 1), (u'its', 1), (u'contributing', 1), (u'This', 2), (u'Developer', 1), (u'version', 1), (u'provides', 1), (u'
print', 1), (u'get', 1), (u'Configuration', 1), (u'supports', 2), (u'command,', 2), (u'[params]`.', 1), (u'refer', 2), (u'available', 1), (u'be', 2), (u'Guid
e](http://spark.apache.org/docs/latest/configuration.html)', 1), (u'run', 7), (u'./bin/run-example', 2), (u'Versions', 1), (u'["Parallel', 1), (u'Hadoop,', 2
), (u'Documentation', 1), (u'use', 3), (u'downloaded', 1), (u'distributions.', 1), (u'Spark.', 1), (u'example:', 1), (u'by', 1), (u'package.', 1), (u'Maven](
http://maven.apache.org/).', 1), (u'["Building', 1), (u'thread', 1), (u'package', 1), (u'of', 5), (u'changed', 1), (u'programming', 1), (u'Spark', 16), (u'ag
ainst', 1), (u'site,', 1), (u'Maven,', 1), (u'3"](https://cwiki.apache.org/confluence/display/MAVEN/Parallel+builds+in+Maven+3).', 1), (u'or', 3), (u'comes',
 1), (u'first', 1), (u'info', 1), (u'contains', 1), (u'can', 7), (u'overview', 1), (u'package.)', 1), (u'Please', 4), (u'one', 3), (u'documentation,', 1), (u
'(You', 1), (u'Online', 1), (u'tools', 1), (u'your', 1), (u'page](http://spark.apache.org/documentation.html).', 1), (u'threads.', 1), (u'Tests', 1), (u'fast
', 1), (u'from', 1), (u'[project', 1), (u'APIs', 1), (u'>>>', 1), (u'SparkPi', 2), (u'locally', 2), (u'system', 1), (u'submit', 1), (u'`examples`', 2), (u'sy
stems.', 1), (u'start', 1), (u'IDE,', 1), (u'params', 1), (u'build/mvn', 1), (u'way', 1), (u'basic', 1), (u'README', 1), (u'<http://spark.apache.org/>', 1),
(u'It', 2), (u'graphs', 1), (u'more', 1), (u'engine', 1), (u'project', 1), (u'option', 1), (u'on', 7), (u'started', 1), (u'Note', 1), (u'N', 1), (u'usage', 1
), (u'versions', 1), (u'DataFrames,', 1), (u'particular', 2), (u'instance:', 1), (u'./bin/spark-shell', 1), (u'general', 3), (u'with', 4), (u'easiest', 1), (
u'protocols', 1), (u'must', 1), (u'And', 1), (u'builds', 1), (u'developing', 1), (u'this', 1), (u'setup', 1), (u'shell:', 2), (u'will', 1), (u'`./bin/run-exa
mple', 1), (u'following', 2), (u'Hadoop-supported', 1), (u'distribution', 1), (u'Maven', 1), (u'example', 3), (u'are', 1), (u'detailed', 2), (u'Data.', 1), (
u'mesos://', 1), (u'stream', 1), (u'computing', 1), (u'URL,', 1), (u'is', 6), (u'in', 6), (u'higher-level', 1), (u'tests', 2), (u'1000:', 2), (u'an', 4), (u'
sample', 1), (u'To', 2), (u'tests](http://spark.apache.org/developer-tools.html#individual-tests).', 1), (u'tips,', 1), (u'at', 2), (u'have', 1), (u'1000).co
unt()', 1), (u'["Specifying', 1), (u'[building', 1), (u'You', 4), (u'configure', 1), (u'development', 1), (u'information', 1), (u'different', 1), (u'Tools"](
http://spark.apache.org/developer-tools.html).', 1), (u'MASTER=spark://host:7077', 1), (u'no', 1), (u'not', 1), (u'Java,', 1), (u'that', 2), (u'storage', 1),
 (u'MLlib', 1), (u'same', 1), (u'machine', 1), (u'how', 3), (u'need', 1), (u'other', 1), (u'build', 4), (u'prefer', 1), (u'online', 1), (u'you', 4), (u'if',
4), (u'[Contribution', 1), (u'A', 1), (u'About', 1), (u'HDFS', 1), (u'[Configuration', 1), (u'sc.parallelize(1', 1), (u'locally.', 1), (u'Hive', 2), (u'["Use
ful', 1), (u'running', 1), (u'uses', 1), (u'a', 8), (u'Version"](http://spark.apache.org/docs/latest/building-spark.html#specifying-the-hadoop-version)', 1),
 (u'variable', 1), (u'The', 1), (u'data', 1), (u'class', 2), (u'built', 1), (u'building', 2), (u'"yarn"', 1), (u'Python', 2), (u'Thriftserver', 1), (u'proces
sing.', 1), (u'programs', 2), (u'requires', 1), (u'documentation', 3), (u'pre-built', 1), (u'Alternatively,', 1), (u'programs,', 1), (u'"local[N]"', 1), (u'S
park](#building-spark).', 1), (u'clean', 1), (u'<class>', 1), (u'spark://', 1), (u'learning,', 1), (u'core', 1), (u'talk', 1), (u'latest', 1)]
>>> 
```

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

CombineByKey()

參考網站（概

念）：http://r97846001.blog.ntu.edu.tw/2015/07/02/spark-learning-spark-%E8%A7%80%E5%BF%B5%E7%AD%86%E8%A8%98-ch4/

參考網站（概念）：http://jude90.github.io/2015/10/20/key-value-rdd.html

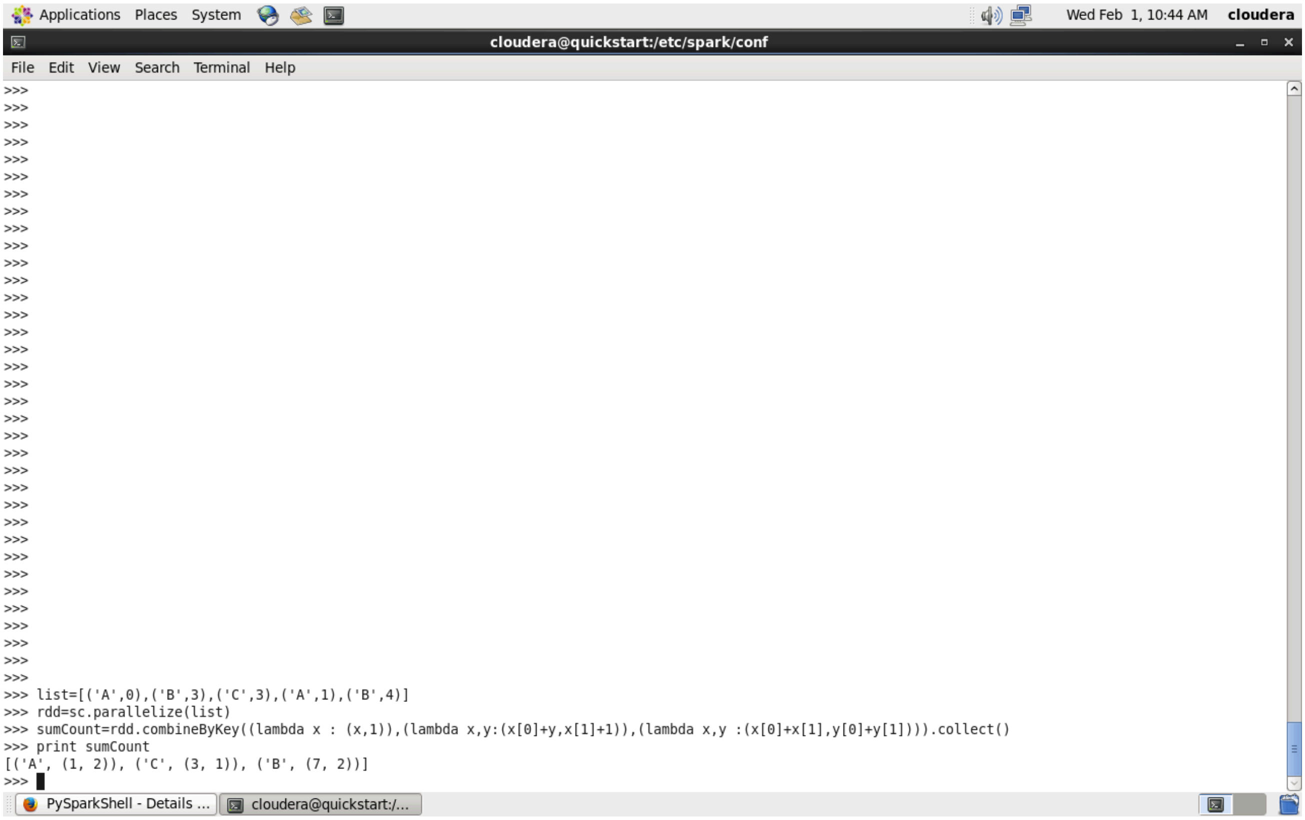
參考網站（概念）：http://bit1129.iteye.com/blog/2216002

參考網站（概念）：http://lxw1234.com/archives/2015/07/358.htm

參考網站（概念）：http://www.voidcn.com/blog/wisgood/article/p-5998930.html

參考網站（概念）：http://www.jianshu.com/p/03f63af51afe

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



```
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>> list=[('A',0),('B',3),('C',3),('A',1),('B',4)]
>>> rdd=sc.parallelize(list)
>>> sumCount=rdd.combineByKey((lambda x : (x,1)),(lambda x,y:(x[0]+y,x[1]+1)),(lambda x,y :(x[0]+x[1],y[0]+y[1]))).collect()
>>> print sumCount
[('A', (1, 2)), ('C', (3, 1)), ('B', (7, 2))]
>>> 
```

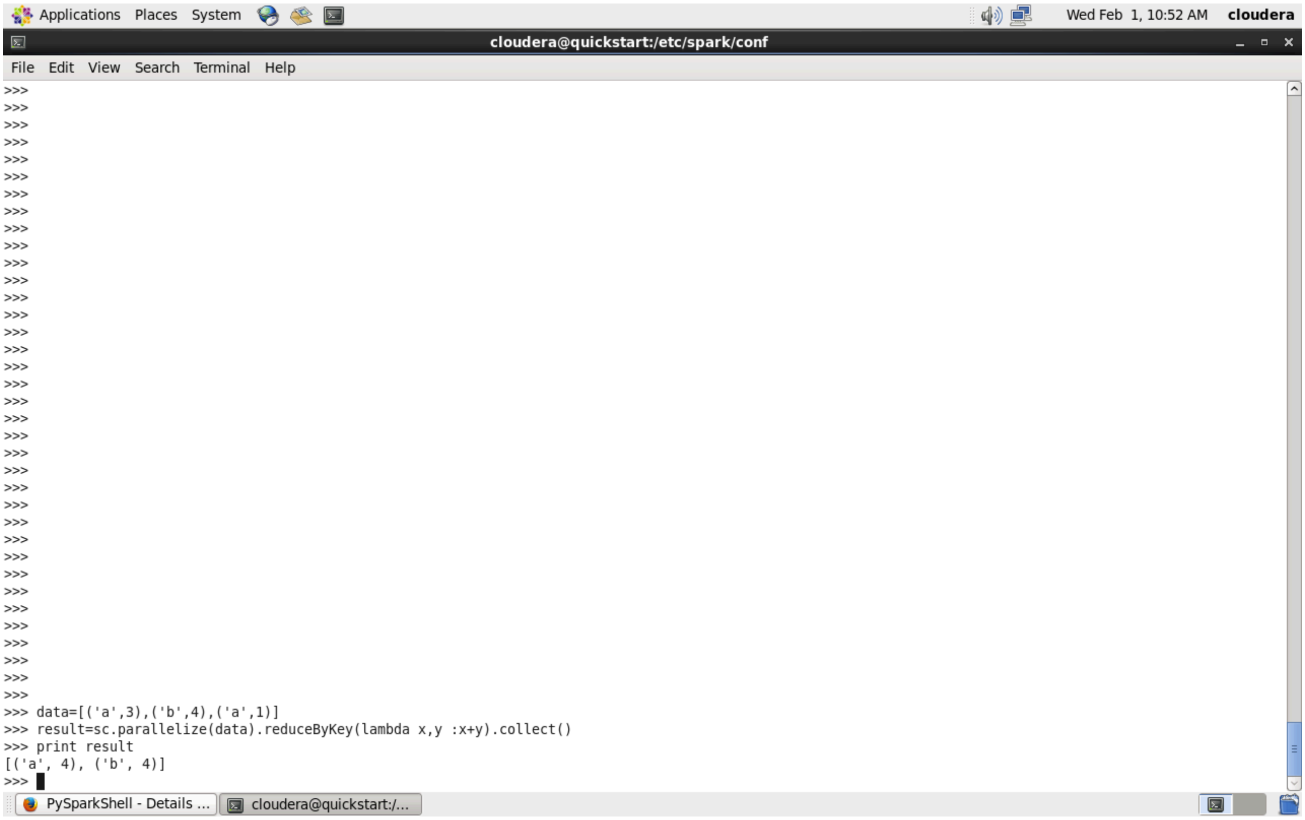
Step1 : [('A',(0,1)) ('B',(3,1)) ('C',(3,1))

step2 : ('A',(0,1))+('A',1)--- > ('A',(0+1,1+1)) —>('A',(1,2))   ('B',(3,1))+('B',4)--- > ('A',(3+4,1+1)) —>('A',(7,2))

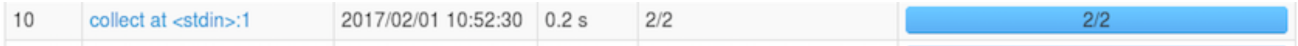
=================

調教平行度

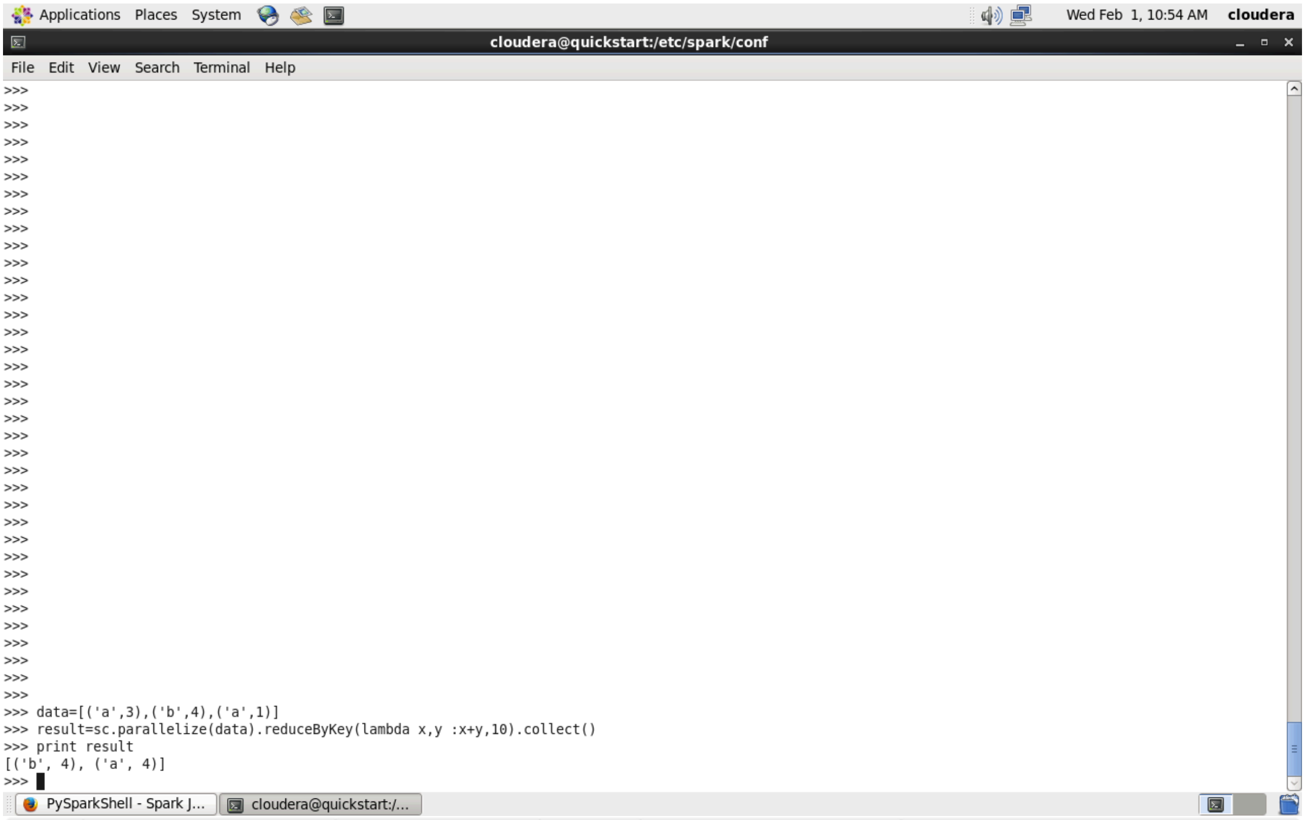
=================


Applications Places System
Wed Feb 1, 10:52 AM
cloudera
cloudera@quickstart:/etc/spark/conf
File Edit View Search Terminal Help
PySparkShell - Details ...
cloudera@quickstart:/...

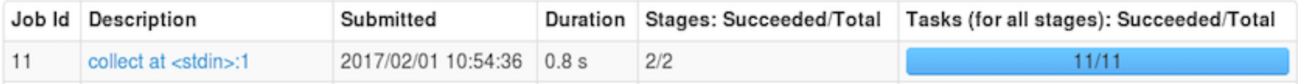
```
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>> data=[('a',3),('b',4),('a',1)]
>>> result=sc.parallelize(data).reduceByKey(lambda x,y :x+y).collect()
>>> print result
[('a', 4), ('b', 4)]
>>> 
```



10	collect at <stdin>:1	2017/02/01 10:52:30	0.2 s	2/2	2/2


Applications Places System
Wed Feb 1, 10:54 AM
cloudera
cloudera@quickstart:/etc/spark/conf
File Edit View Search Terminal Help
PySparkShell - Spark J...
cloudera@quickstart:/...

```
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>> data=[('a',3),('b',4),('a',1)]
>>> result=sc.parallelize(data).reduceByKey(lambda x,y :x+y,10).collect()
>>> print result
[('b', 4), ('a', 4)]
>>> 
```



Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
11	collect at <stdin>:1	2017/02/01 10:54:36	0.8 s	2/2	11/11

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

改變分割的需求：coalesce()、repartition

參考網站（概念）：http://lxw1234.com/archives/2015/07/341.htm

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

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

GroupBy()

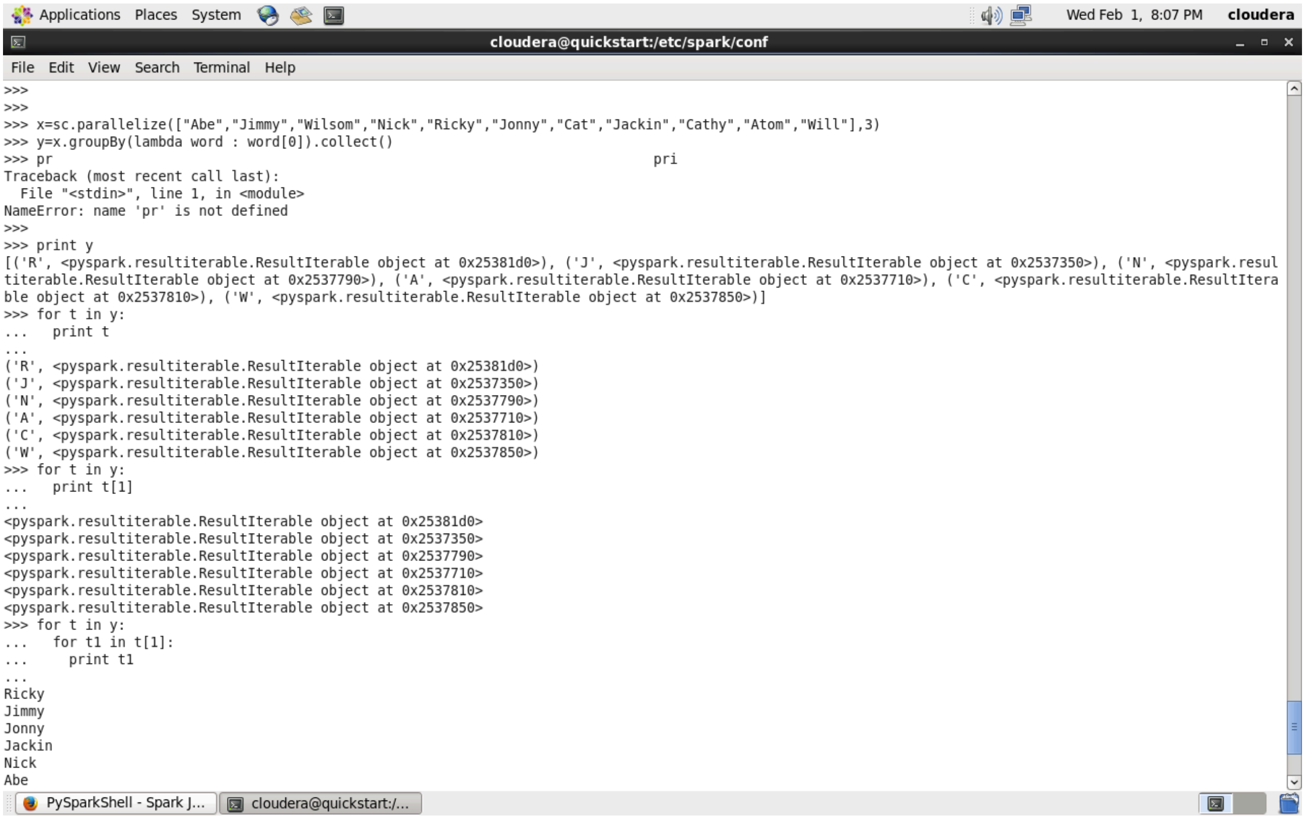
原理：將元素透過function生成相對應的Key，數據就轉換成Key-Value格式，之後將key相同的元素分為一組

GroupBy和GroupByKey差異：groupByKey對key-Value型態的RDD操作，其所分組使用的key不是由指定的function產生，而是產用元素本身的Key

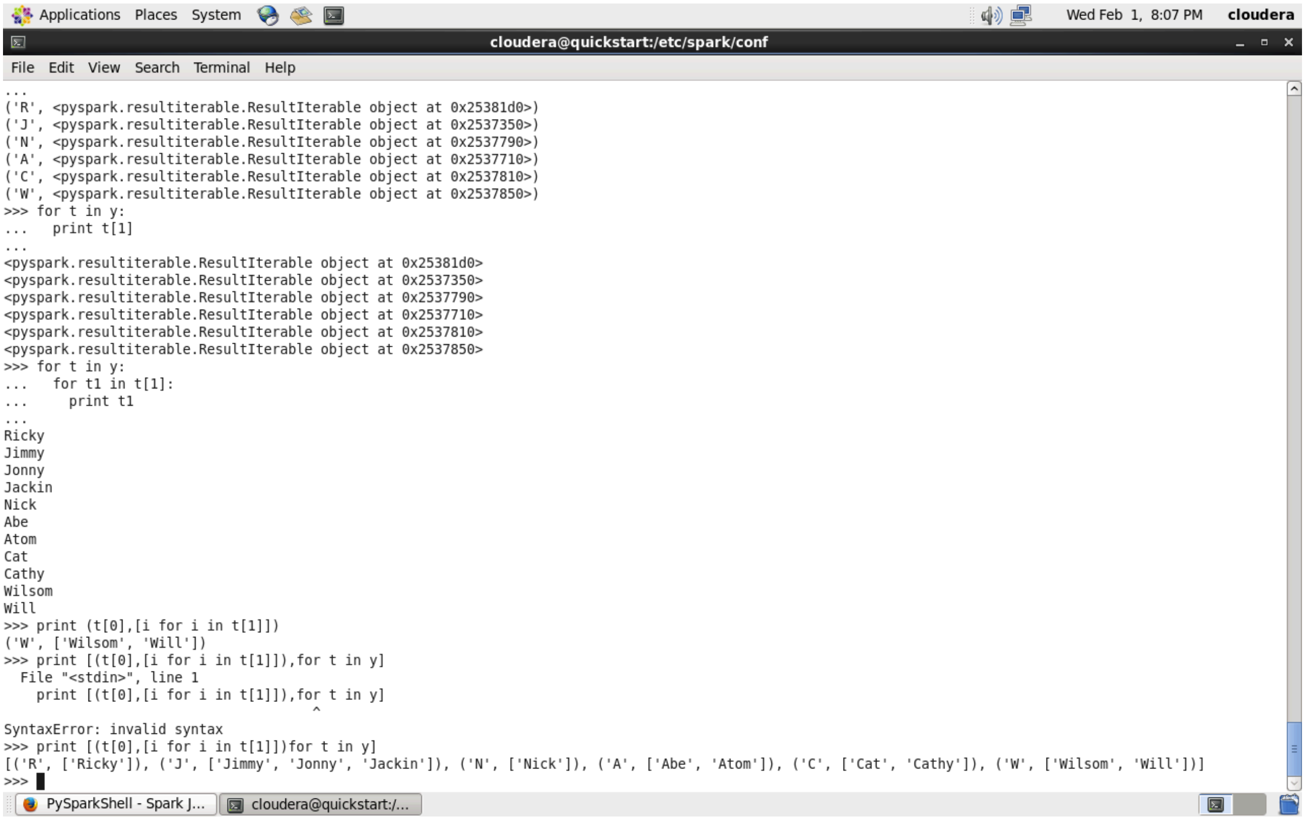
參考網站（概念）：http://backtobazics.com/big-data/spark/apache-spark-groupby-example/

參考網站（概念）：http://blog.cheyo.net/178.html

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



```
>>>
>>>
>>> x=sc.parallelize(["Abe","Jimmy","Wilsom","Nick","Ricky","Jonny","Cat","Jackin","Cathy","Atom","Will"],3)
>>> y=x.groupBy(lambda word : word[0]).collect()
>>> pr                                                                           pri
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'pr' is not defined
>>>
>>> print y
[('R', <pyspark.resultiterable.ResultIterable object at 0x25381d0>), ('J', <pyspark.resultiterable.ResultIterable object at 0x2537350>), ('N', <pyspark.resul
titerable.ResultIterable object at 0x2537790>), ('A', <pyspark.resultiterable.ResultIterable object at 0x2537710>), ('C', <pyspark.resultiterable.ResultItera
ble object at 0x2537810>), ('W', <pyspark.resultiterable.ResultIterable object at 0x2537850>)]
>>> for t in y:
...   print t
...
('R', <pyspark.resultiterable.ResultIterable object at 0x25381d0>)
('J', <pyspark.resultiterable.ResultIterable object at 0x2537350>)
('N', <pyspark.resultiterable.ResultIterable object at 0x2537790>)
('A', <pyspark.resultiterable.ResultIterable object at 0x2537710>)
('C', <pyspark.resultiterable.ResultIterable object at 0x2537810>)
('W', <pyspark.resultiterable.ResultIterable object at 0x2537850>)
>>> for t in y:
...   print t[1]
...
<pyspark.resultiterable.ResultIterable object at 0x25381d0>
<pyspark.resultiterable.ResultIterable object at 0x2537350>
<pyspark.resultiterable.ResultIterable object at 0x2537790>
<pyspark.resultiterable.ResultIterable object at 0x2537710>
<pyspark.resultiterable.ResultIterable object at 0x2537810>
<pyspark.resultiterable.ResultIterable object at 0x2537850>
>>> for t in y:
...   for t1 in t[1]:
...     print t1
...
Ricky
Jimmy
Jonny
Jackin
Nick
Abe
```


Applications Places System
Wed Feb 1, 8:07 PM
cloudera
cloudera@quickstart:/etc/spark/conf
File Edit View Search Terminal Help
PySparkShell - Spark J...
cloudera@quickstart:/...

```
...
('R', <pyspark.resultiterable.ResultIterable object at 0x25381d0>)
('J', <pyspark.resultiterable.ResultIterable object at 0x2537350>)
('N', <pyspark.resultiterable.ResultIterable object at 0x2537790>)
('A', <pyspark.resultiterable.ResultIterable object at 0x2537710>)
('C', <pyspark.resultiterable.ResultIterable object at 0x2537810>)
('W', <pyspark.resultiterable.ResultIterable object at 0x2537850>)
>>> for t in y:
...   print t[1]
...
<pyspark.resultiterable.ResultIterable object at 0x25381d0>
<pyspark.resultiterable.ResultIterable object at 0x2537350>
<pyspark.resultiterable.ResultIterable object at 0x2537790>
<pyspark.resultiterable.ResultIterable object at 0x2537710>
<pyspark.resultiterable.ResultIterable object at 0x2537810>
<pyspark.resultiterable.ResultIterable object at 0x2537850>
>>> for t in y:
...   for t1 in t[1]:
...     print t1
...
Ricky
Jimmy
Jonny
Jackin
Nick
Abe
Atom
Cat
Cathy
Wilsom
Will
>>> print (t[0],[i for i in t[1]])
('W', ['Wilsom', 'Will'])
>>> print [(t[0],[i for i in t[1]]),for t in y]
  File "<stdin>", line 1
    print [(t[0],[i for i in t[1]]),for t in y]
                                      ^

SyntaxError: invalid syntax
>>> print [(t[0],[i for i in t[1]])for t in y]
[('R', ['Ricky']), ('J', ['Jimmy', 'Jonny', 'Jackin']), ('N', ['Nick']), ('A', ['Abe', 'Atom']), ('C', ['Cat', 'Cathy']), ('W', ['Wilsom', 'Will'])]
>>> 
```

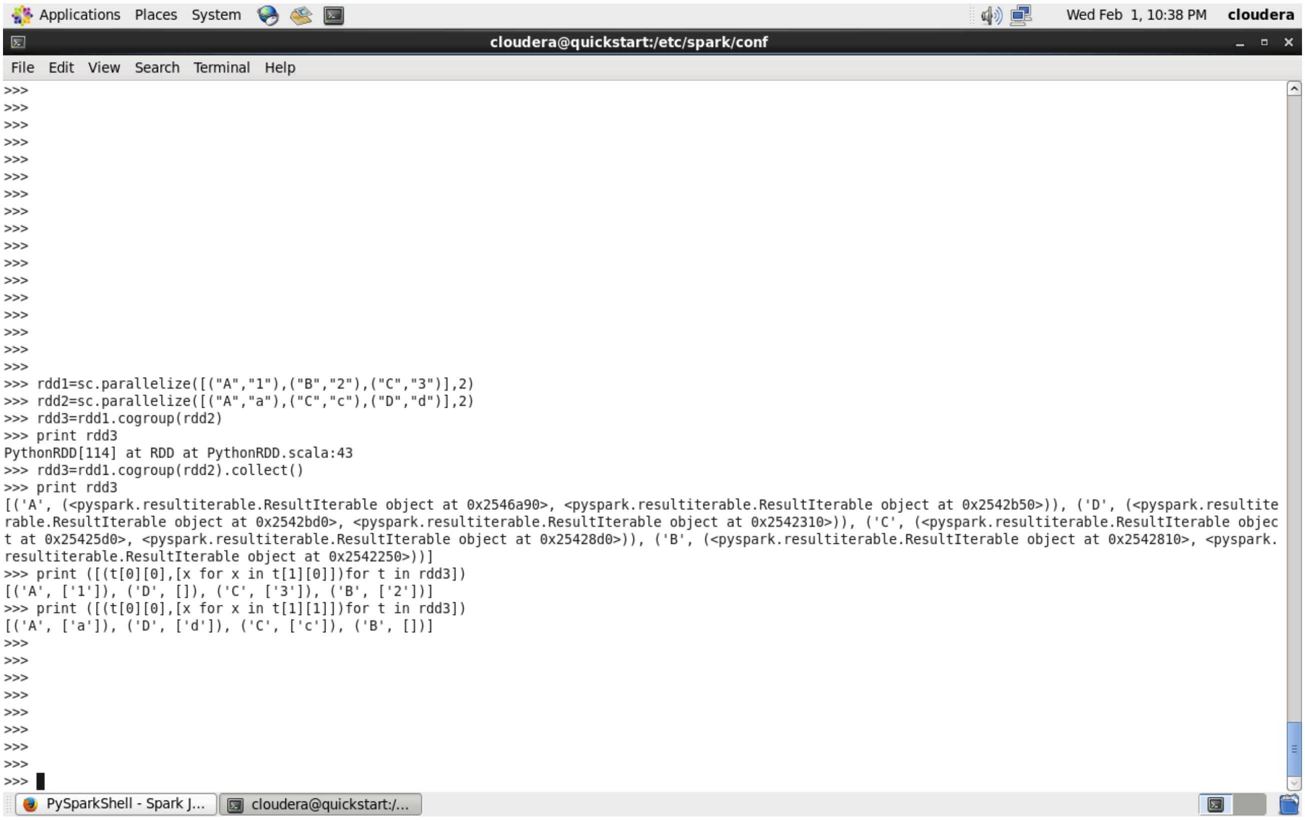
=================

cogroup()

參考網站（概念）：http://lxw1234.com/archives/2015/07/384.htm

問題：要如何把value的兩層list讀取出來

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

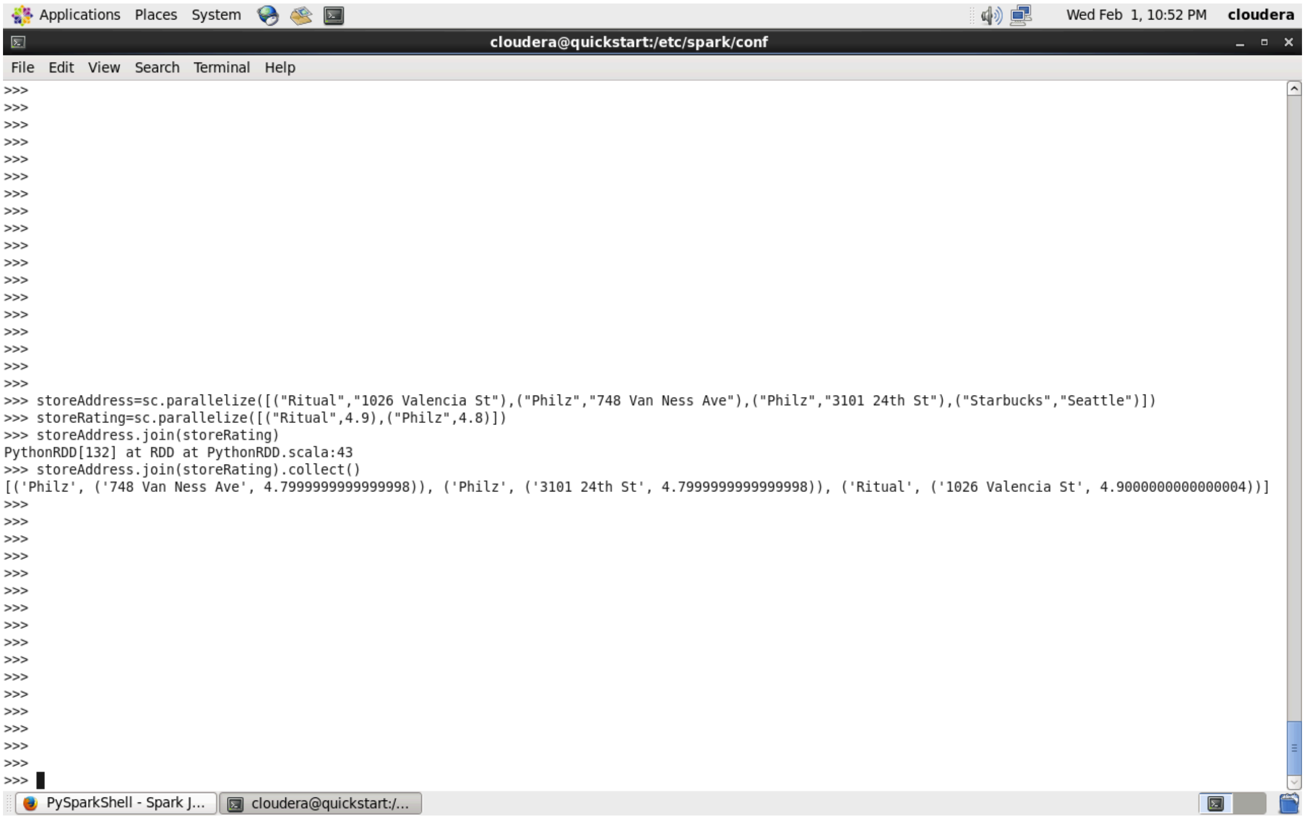

Applications Places System
Wed Feb 1, 10:38 PM cloudera
cloudera@quickstart:/etc/spark/conf
File Edit View Search Terminal Help
PySparkShell - Spark J...
cloudera@quickstart:/...

```
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>> rdd1=sc.parallelize([("A","1"),("B","2"),("C","3")],2)
>>> rdd2=sc.parallelize([("A","a"),("C","c"),("D","d")],2)
>>> rdd3=rdd1.cogroup(rdd2)
>>> print rdd3
PythonRDD[114] at RDD at PythonRDD.scala:43
>>> rdd3=rdd1.cogroup(rdd2).collect()
>>> print rdd3
[('A', (<pyspark.resultiterable.ResultIterable object at 0x2546a90>, <pyspark.resultiterable.ResultIterable object at 0x2542b50>)), ('D', (<pyspark.resultite
rable.ResultIterable object at 0x2542bd0>, <pyspark.resultiterable.ResultIterable object at 0x2542310>)), ('C', (<pyspark.resultiterable.ResultIterable objec
t at 0x25425d0>, <pyspark.resultiterable.ResultIterable object at 0x25428d0>)), ('B', (<pyspark.resultiterable.ResultIterable object at 0x2542810>, <pyspark.
resultiterable.ResultIterable object at 0x2542250>))]
>>> print ([(t[0][0],[x for x in t[1][0]])for t in rdd3])
[('A', ['1']), ('D', []), ('C', ['3']), ('B', ['2'])]
>>> print ([(t[0][0],[x for x in t[1][1]])for t in rdd3])
[('A', ['a']), ('D', ['d']), ('C', ['c']), ('B', [])]
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
```

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

Python : inner join

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

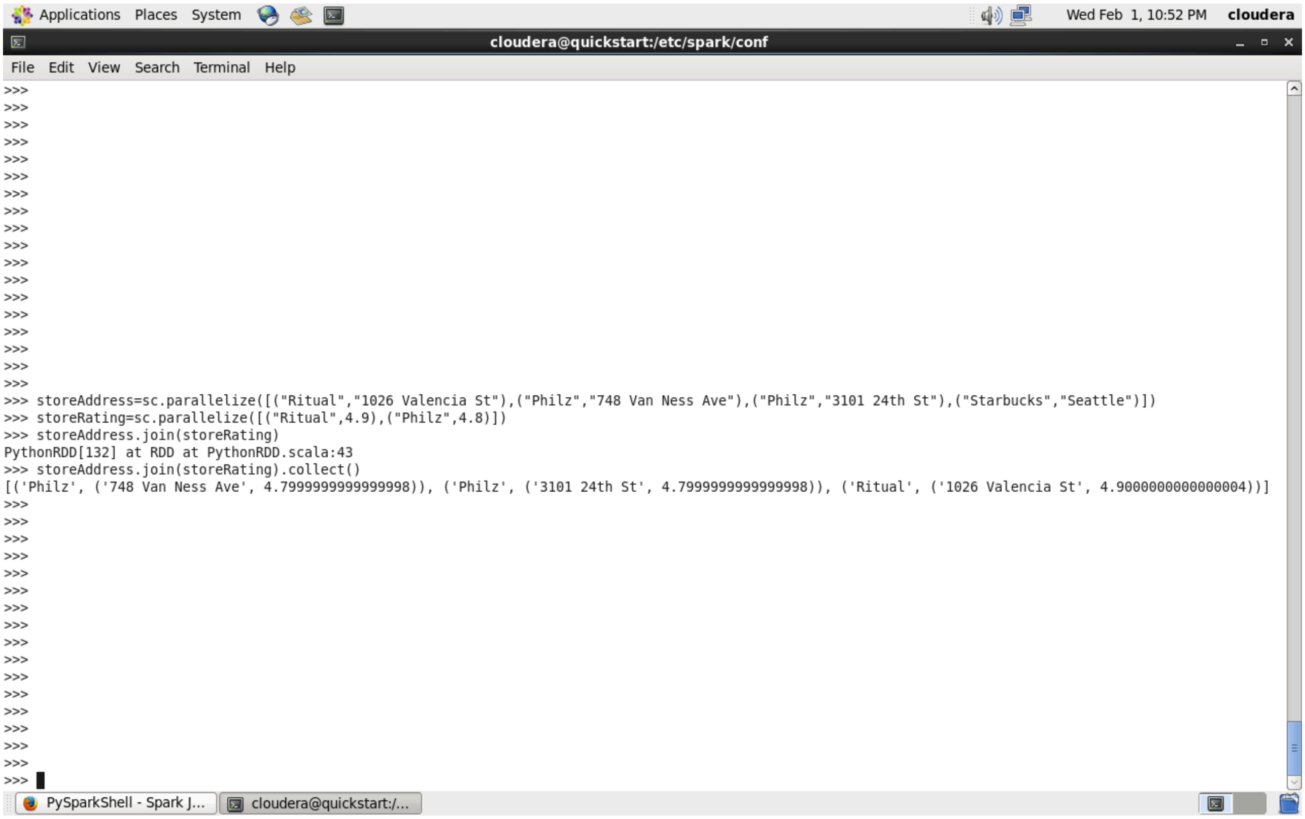

Applications Places System
Wed Feb 1, 10:52 PM cloudera
cloudera@quickstart:/etc/spark/conf
File Edit View Search Terminal Help
PySparkShell - Spark J...
cloudera@quickstart:/...

```
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>> storeAddress=sc.parallelize([("Ritual","1026 Valencia St"),("Philz","748 Van Ness Ave"),("Philz","3101 24th St"),("Starbucks","Seattle")])
>>> storeRating=sc.parallelize([("Ritual",4.9),("Philz",4.8)])
>>> storeAddress.join(storeRating)
PythonRDD[132] at RDD at PythonRDD.scala:43
>>> storeAddress.join(storeRating).collect()
[('Philz', ('748 Van Ness Ave', 4.7999999999999998)), ('Philz', ('3101 24th St', 4.7999999999999998)), ('Ritual', ('1026 Valencia St', 4.9000000000000004))]
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
```

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

Python : leftOuterjoin()

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

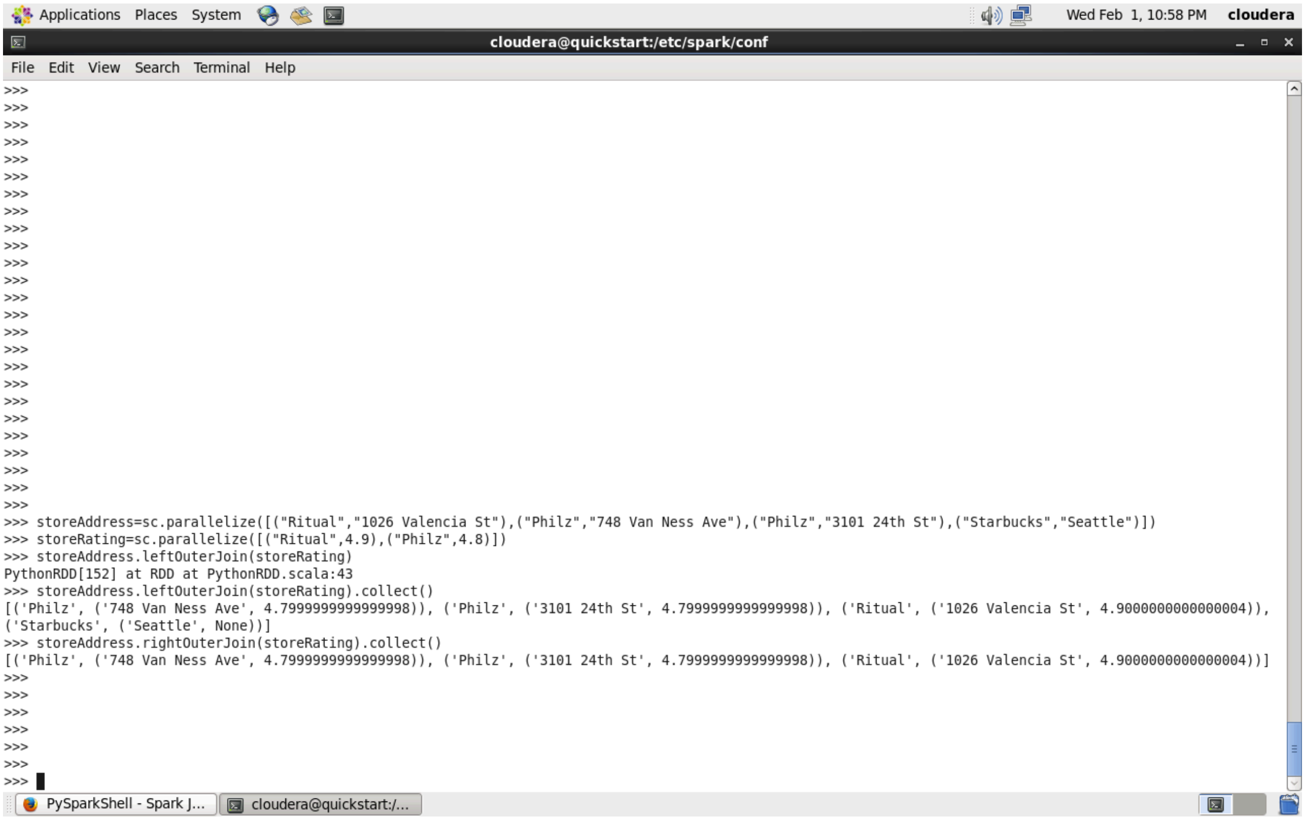


```
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>> storeAddress=sc.parallelize([("Ritual","1026 Valencia St"),("Philz","748 Van Ness Ave"),("Philz","3101 24th St"),("Starbucks","Seattle")])
>>> storeRating=sc.parallelize([("Ritual",4.9),("Philz",4.8)])
>>> storeAddress.join(storeRating)
PythonRDD[132] at RDD at PythonRDD.scala:43
>>> storeAddress.join(storeRating).collect()
[('Philz', ('748 Van Ness Ave', 4.7999999999999998)), ('Philz', ('3101 24th St', 4.7999999999999998)), ('Ritual', ('1026 Valencia St', 4.9000000000000004))]
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>> 
```

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

Python : rightOutjoin()

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


Applications Places System
Wed Feb 1, 10:58 PM cloudera
cloudera@quickstart:/etc/spark/conf
File Edit View Search Terminal Help
PySparkShell - Spark J...
cloudera@quickstart:/...

```
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>> storeAddress=sc.parallelize([("Ritual","1026 Valencia St"),("Philz","748 Van Ness Ave"),("Philz","3101 24th St"),("Starbucks","Seattle")])
>>> storeRating=sc.parallelize([("Ritual",4.9),("Philz",4.8)])
>>> storeAddress.leftOuterJoin(storeRating)
PythonRDD[152] at RDD at PythonRDD.scala:43
>>> storeAddress.leftOuterJoin(storeRating).collect()
[('Philz', ('748 Van Ness Ave', 4.7999999999999998)), ('Philz', ('3101 24th St', 4.7999999999999998)), ('Ritual', ('1026 Valencia St', 4.9000000000000004)),
('Starbucks', ('Seattle', None))]
>>> storeAddress.rightOuterJoin(storeRating).collect()
[('Philz', ('748 Van Ness Ave', 4.7999999999999998)), ('Philz', ('3101 24th St', 4.7999999999999998)), ('Ritual', ('1026 Valencia St', 4.9000000000000004))]
>>>
>>>
>>>
>>>
>>>
>>>
>>> 
```

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

Python：資料排序

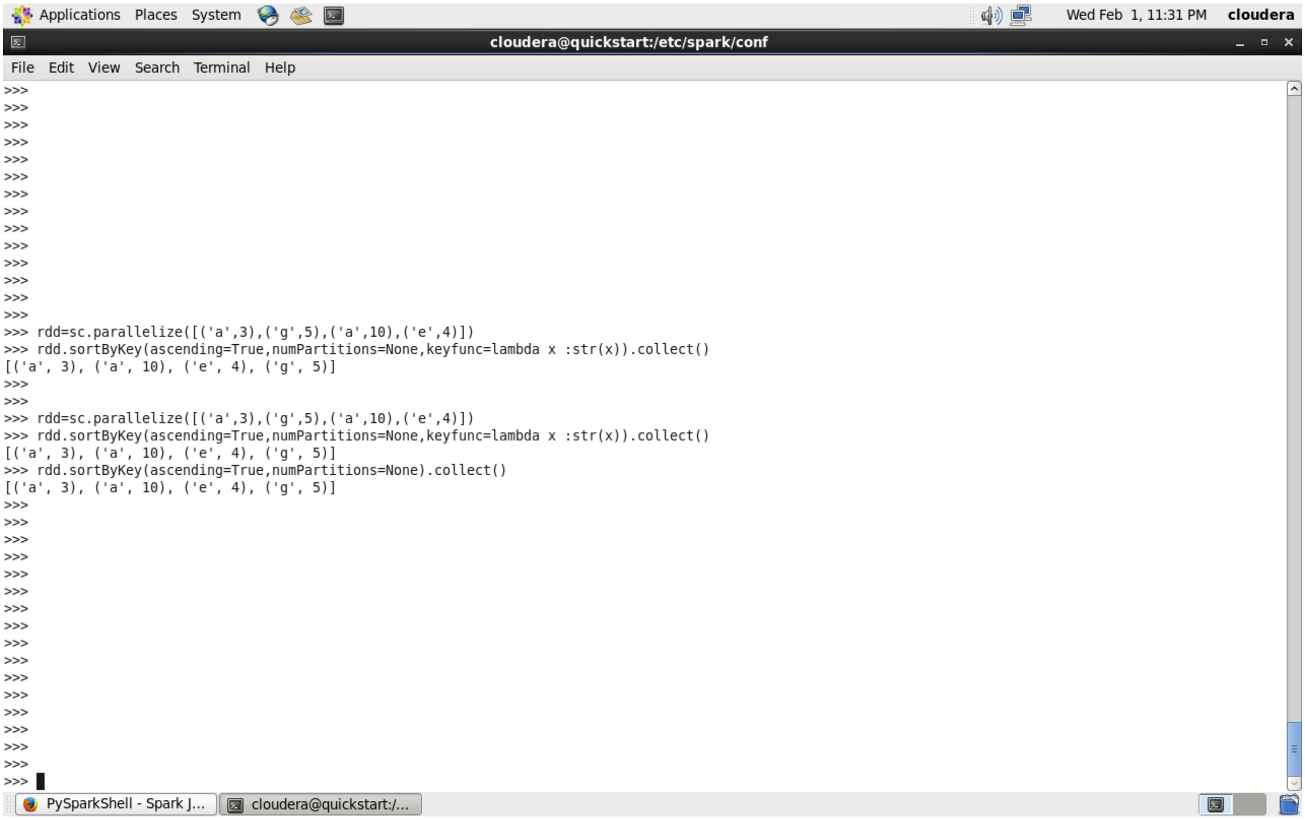
參數：

ascending（是否希望資料是倒序排列,預設是Ture)

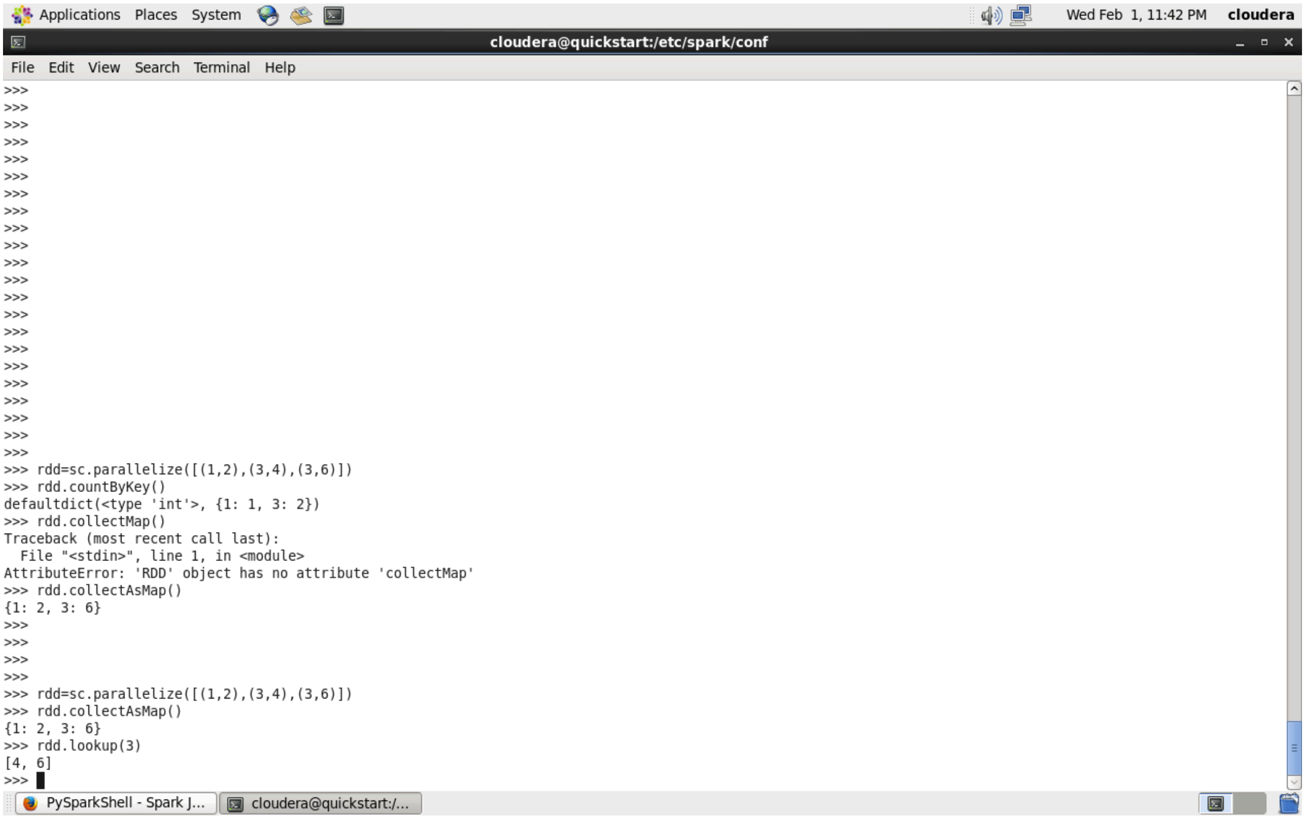
numpartitions（切片)

keyfunc（自定義比較函式)

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


Applications Places System
Wed Feb 1, 11:31 PM
cloudera
cloudera@quickstart:/etc/spark/conf
File Edit View Search Terminal Help
PySparkShell - Spark J...
cloudera@quickstart:/...

```
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>> rdd=sc.parallelize([('a',3),('g',5),('a',10),('e',4)])
>>> rdd.sortByKey(ascending=True,numPartitions=None,keyfunc=lambda x :str(x)).collect()
[('a', 3), ('a', 10), ('e', 4), ('g', 5)]
>>>
>>>
>>> rdd=sc.parallelize([('a',3),('g',5),('a',10),('e',4)])
>>> rdd.sortByKey(ascending=True,numPartitions=None,keyfunc=lambda x :str(x)).collect()
[('a', 3), ('a', 10), ('e', 4), ('g', 5)]
>>> rdd.sortByKey(ascending=True,numPartitions=None).collect()
[('a', 3), ('a', 10), ('e', 4), ('g', 5)]
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
```


Applications Places System
Wed Feb 1, 11:42 PM
cloudera
cloudera@quickstart:/etc/spark/conf
File Edit View Search Terminal Help
PySparkShell - Spark J...
cloudera@quickstart:/...

```
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>>
>>> rdd=sc.parallelize([(1,2),(3,4),(3,6)])
>>> rdd.countByKey()
defaultdict(<type 'int'>, {1: 1, 3: 2})
>>> rdd.collectMap()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'RDD' object has no attribute 'collectMap'
>>> rdd.collectAsMap()
{1: 2, 3: 6}
>>>
>>>
>>>
>>>
>>> rdd=sc.parallelize([(1,2),(3,4),(3,6)])
>>> rdd.collectAsMap()
{1: 2, 3: 6}
>>> rdd.lookup(3)
[4, 6]
>>>
```