

Chapter 01: Installing Pyspark and Setting up Your Development Environment

```
C:\Users\Admin>gow --list
Available executables:
```

```
awk, basename, bash, bc, bison, bunzip2, bzip2, bzip2recover, cat,
chgrp, chmod, chown, chroot, cksum, clear, cp, csplit, curl, cut, dc,
dd, df, diff, diff3, dirname, dos2unix, du, egrep, env, expand, expr,
factor, fgrep, flex, fmt, fold, gawk, gfind, gow, grep, gsar, gsort,
gzip, head, hostid, hostname, id, indent, install, join, jwhois, less,
lesskey, ln, ls, m4, make, md5sum, mkdir, mkfifo, mknod, mv, nano,
ncftp, nl, od, pageant, paste, patch, pathchk, plink, pr, printenv,
printf, pscp, psftp, putty, puttygen, pwd, rm, rmdir, scp, sdiff, sed,
seq, sftp, sha1sum, shar, sleep, split, ssh, su, sum, sync, tac, tail,
tar, tee, test, touch, tr, uname, unexpand, uniq, unix2dos, unlink,
unrar, unshar, uuencode, uuencode, vim, wc, wget, whereis, which,
whoami, xargs, yes, zip
```

```
C:\Users\Admin>
```

Download Apache Spark™

1. Choose a Spark release:
2. Choose a package type:
3. Download Spark: [spark-2.3.3-bin-hadoop2.7.tgz](#)
4. Verify this release using the 2.3.3 [signatures](#), [checksums](#) and [project release KEYS](#).

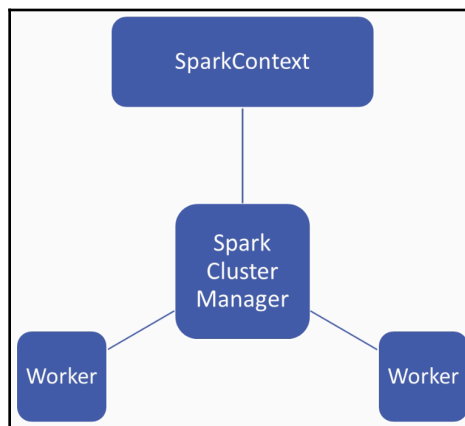
```
C:\>java -version
java version "1.8.0_201"
Java(TM) SE Runtime Environment (build 1.8.0_201-b09)
Java HotSpot(TM) Client VM (build 25.201-b09, mixed mode, sharing)

C:\>_
```

```
C:\>python --version
Python 3.7.0
```

```
c:\opt\spark\spark-2.1.0-bin-hadoop2.7>.\bin\pyspark
[TerminalIPythonApp] WARNING | Subcommand `ipython notebook` is deprecated and will be removed in future versions.
[TerminalIPythonApp] WARNING | You likely want to use `jupyter notebook` in the future
[I 18:58:16.992 NotebookApp] [nb_conda_kernels] enabled, 0 kernels found
[I 18:58:17.060 NotebookApp] JupyterLab extension loaded from C:\Users\Admin\Anaconda3\lib\site-packages\jupyterlab
[I 18:58:17.060 NotebookApp] JupyterLab application directory is C:\Users\Admin\Anaconda3\share\jupyter\lab
[I 18:58:17.409 NotebookApp] [nb_conda] enabled
[I 18:58:17.410 NotebookApp] Serving notebooks from local directory: c:\opt\spark\spark-2.1.0-bin-hadoop2.7
[I 18:58:17.415 NotebookApp] The Jupyter Notebook is running at:
[I 18:58:17.415 NotebookApp] http://localhost:8888/?token=9d6838b0bacad0c6083861cf7378a1379719f1ccc22879a7
[I 18:58:17.416 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 18:58:17.476 NotebookApp]

Copy/paste this URL into your browser when you connect for the first time,
to login with a token:
    http://localhost:8888/?token=9d6838b0bacad0c6083861cf7378a1379719f1ccc22879a7
[I 18:58:17.767 NotebookApp] Accepting one-time-token-authenticated connection from ::1
```



```
10-26 13:27:44 spark-2.3.2-bin-hadoop2.7
> .\bin\pyspark
Python 3.6.2 (v3.6.2:5fd33b5, Jul 8 2017, 04:57:36) [MSC v.1900 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (f
ile:/C:/spark/spark-2.3.2-bin-hadoop2.7/spark-2.3.2-bin-hadoop2.7/jars/hadoop-auth-2.7.3.jar) to met
hod sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authenticat
ion.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operation
s
WARNING: All illegal access operations will be denied in a future release
2018-10-26 13:28:15 WARN NativeCodeLoader:62 - Unable to load native-hadoop library for your platfo
rm... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Welcome to

  ____
 /  __ \
/   /  \
/_____/

 version 2.3.2

Using Python version 3.6.2 (v3.6.2:5fd33b5, Jul 8 2017 04:57:36)
SparkSession available as 'spark'.
>>>
```

Chapter 02: Getting Your Big Data into the Spark Environment Using RDDs

Index of /ml/machine-learning-databases/kddcup99-ml

Name	Last modified	Size	Description
Parent Directory		-	
corrected.gz	16-May-2000 08:50	1.3M	
kddcup_data.gz	26-Jun-2007 12:59	17M	
kddcup_data_10_percent.gz	26-Jun-2007 12:59	2.0M	
kddcup_names	23-Oct-2012 11:03	1.3K	
kddcup.newtestdata_10_percent_unlabeled.gz	28-Oct-1999 11:23	1.3M	
kddcup.testdata.unlabeled.gz	28-Oct-1999 11:23	11M	
kddcup.testdata.unlabeled_10_percent.gz	28-Oct-1999 11:23	1.3M	
kddcup99.html	23-Oct-2012 11:03	2.4K	
task.html	23-Oct-2012 11:03	10K	
training_attack_types	23-Oct-2012 11:03	272	
typo-correction.txt	23-Oct-2012 11:03	825	

Apache/2.2.15 (CentOS) Server at archive.ics.uci.edu Port 443

```
>>> contains_normal.count()
972781
>>>
```

```
>>> split_file.take(5)
[['0', 'tcp', 'http', 'SF', '215', '45076', '0', '0', '0', '0', '1', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '1', '1', '0.00', '0.00', '0.00', '0.00', '1.00', '0.00', '0.00', '0', '0', '0.00', '0.00', '0.00', '0.00', '0.00', '0.00', '0.00', '0.00', 'normal.'], ['0', 'tcp', 'http', 'SF', '162', '4528', '0', '0', '0', '0', '0', '0', '1', '0', '0', '0', '0', '0', '0', '0', '0', '2', '2', '0.00', '0.00', '0.00', '0.00', '1.0', '0', '0.00', '0.00', '1', '1', '1.00', '0.00', '1.00', '0.00', '0.00', '0.00', '0.00', '0.00', 'normal.'], ['0', 'tcp', 'http', 'SF', '236', '1228', '0', '0', '0', '0', '0', '0', '1', '0', '0', '0', '0', '0', '0', '0', '0', '1', '1', '0.00', '0.00', '0.00', '0.00', '1.00', '0.00', '0.00', '2', '2', '1.00', '0.00', '0.50', '0.00', '0.00', '0.00', '0.00', '0.00', 'normal.'], ['0', 'tcp', 'http', 'SF', '233', '2032', '0', '0', '0', '0', '1', '0', '0', '0', '0', '0', '0', '0', '0', '2', '2', '0.00', '0.00', '0.00', '0.00', '1.00', '0.00', '0.00', '0.00', '3', '3', '1.00', '0.00', '0.33', '0.00', '0.00', '0.00', '0.00', '0.00', 'normal.'], ['0', 'tcp', 'http', 'SF', '239', '486', '0', '0', '0', '0', '1', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '0', '3', '3', '0.00', '0.00', '0.00', '0.00', '1.00', '0.00', '0.00', '0.00', '4', '4', '1.00', '0.00', '0.25', '0.00', '0.00', '0.00', '0.00', '0.00', '0.00', 'normal.']]
>>>
```

Chapter 03: Big Data Cleaning and Wrangling with Spark Notebooks

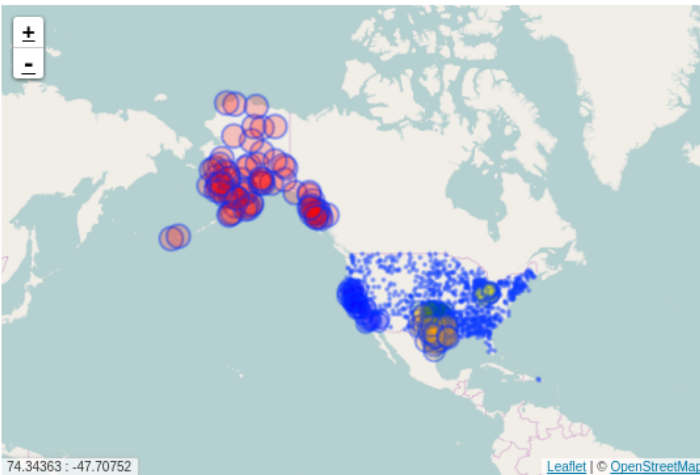
Spark Notebook

gitter join chat ❌ FAILED 1.0

The Spark Notebook is the open source notebook aimed at enterprise environments, providing Data Scientists and Data Engineers with an interactive web-based editor that can combine Scala code, SQL queries, Markup and JavaScript in a collaborative manner to explore, analyse and learn from massive data sets.

```
GeoPointsChart(airportsDFWithStyles,  
  latLonFields=Some(("lat", "long")),  
  rField = Some("r"),  
  colorField = Some("c"))
```

```
airportsDFWithStyles: org.apache.spark.sql.DataFrame = [airport: string, city: string, country: string,  
  iata: string, lat: double, long: double, state: string, r: int, c: string]  
res46: notebook.front.widgets.charts.GeoPointsChart[org.apache.spark.sql.DataFrame] = <GeoPointsChart w  
idget>
```



Took: 1 second 531 milliseconds, at 2016-7-12 19:34

🔗 Reactive

All components in the Spark Notebook are dynamic and reactive.

The Spark Notebook comes with dynamic charts and most (if not all) components can be listened for and can react to events. This is very helpful in many cases, for example:

- data entering the system live at runtime
- visually plots of events
- multiple interconnected visual components Dynamic and reactive components mean that you don't have write the html, js, server code just for basic use cases.

Quick Start

Go to [Quick Start](#) for our 5-minutes guide to get up and running with the Spark Notebook.

C'mon on to [Gitter](#) to discuss things, to get some help, or to start contributing!

Learn more

Documentation

Quick Start Guide

Start using the Spark Notebook in less than 5 minutes? Take these steps:

Requirements:

- Make sure you are running at least Java 7

Steps

- Go to [spark-notebook.io](#)
 - Choose your release, spark version and additional packages according to your specific needs
 - Read more on: [Using Releases](#)
 - If trying Spark for the first time, pick the latest release
- Download one of the packaged builds (TGZ or ZIP)
- Extract the file somewhere convenient
- Open a terminal/command window
- Change to the root directory of the expanded distribution
- Execute the command `bin/spark-notebook` (*NIX) or `bin\spark-notebook` (Windows)
- Open your browser to [localhost:9001](#)

tgz		deb		zip	
Notebook 0.8.3		Notebook 0.7.0-pre2		Notebook 0.8.3	
Scala 2.11	Scala 2.10	Scala 2.11	Scala 2.10	Scala 2.11	Scala 2.10
Spark 2.2.2 and Hadoop	Spark 2.2.2 and Hadoop	Spark 1.6.3 and Hadoop	Spark 1.6.0 and Hadoop	Spark 2.2.2 and Hadoop	Spark 2.2.2 and Hadoop
<ul style="list-style-type: none"> ◦ 2.7.2 parquet 	<ul style="list-style-type: none"> ◦ 2.7.2 parquet 	<ul style="list-style-type: none"> ◦ 2.7.2 	<ul style="list-style-type: none"> ◦ 2.6.0 hive parquet 	<ul style="list-style-type: none"> ◦ 2.6.0 hive parquet ◦ 2.7.2 hive parquet ◦ 2.7.2 parquet ◦ 2.7.3 hive parquet 	<ul style="list-style-type: none"> ◦ 1.0.3 parquet ◦ 2.7.2 parquet

The screenshot shows a Jupyter Notebook window titled "Chapter 3" with a last checkpoint 6 minutes ago. The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help), a toolbar with icons for file operations and execution, and a code editor area. The code editor contains the following text:

```

export PYSPARK_DRIVER_PYTHON=/path/to/anaconda3/bin/jupyter
export PYSPARK_DRIVER_PYTHON_OPTS="notebook --NotebookApp.open_browser=False --NotebookApp.ip='*' --NotebookApp.port=8888"

```

Below the code editor, the input prompt "In []:" is followed by the number "1".

```

In [29]: 1 f2
Out[29]: ['http',
          'finger',
          'auth',
          'domain_u',
          'smtp',
          'ftp',
          'telnet',
          'eco_i',
          'ntp_u',
          'ecr_i',
          'other',
          'private',
          'pop_3',

```

Chapter 04: Aggregating and Summarizing Data into Useful Reports

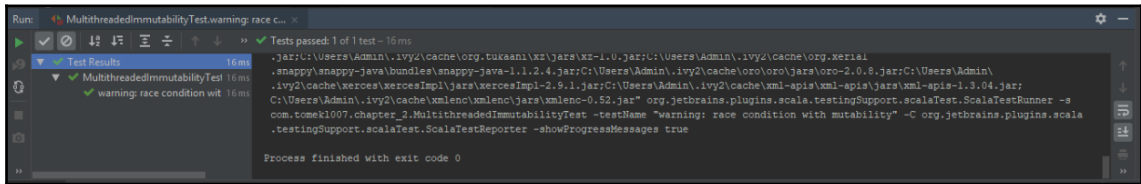
```
[('normal.', 211895753.0),
 ('buffer_overflow.', 2751.0),
 ('loadmodule.', 326.0),
 ('perl.', 124.0),
 ('neptune.', 2.0),
 ('smurf.', 0.0),
 ('guess_passwd.', 144.0),
 ('pod.', 0.0),
 ('teardrop.', 0.0),
 ('portsweep.', 24257982.0),
 ('ipsweep.', 13049.0),
 ('land.', 0.0),
 ('ftp_write.', 259.0),
 ('back.', 284.0),
 ('imap.', 72.0),
 ('satan.', 500.0),
 ('phf.', 18.0),
 ('nmap.', 0.0),
 ('multihop.', 1288.0),
 ('warezmaster.', 301.0),
 ('warezclient.', 627563.0),
 ('spy.', 636.0),
 ('rootkit.', 1008.0)]
```

```
defaultdict(int,
             {'back.': 2203,
              'buffer_overflow.': 30,
              'ftp_write.': 8,
              'guess_passwd.': 53,
              'imap.': 12,
              'ipsweep.': 12481,
              'land.': 21,
              'loadmodule.': 9,
              'multihop.': 7,
              'neptune.': 1072017,
              'nmap.': 2316,
              'normal.': 972781,
              'perl.': 3,
              'phf.': 4,
              'pod.': 264,
              'portsweep.': 10413,
              'rootkit.': 10,
              'satan.': 15892,
```

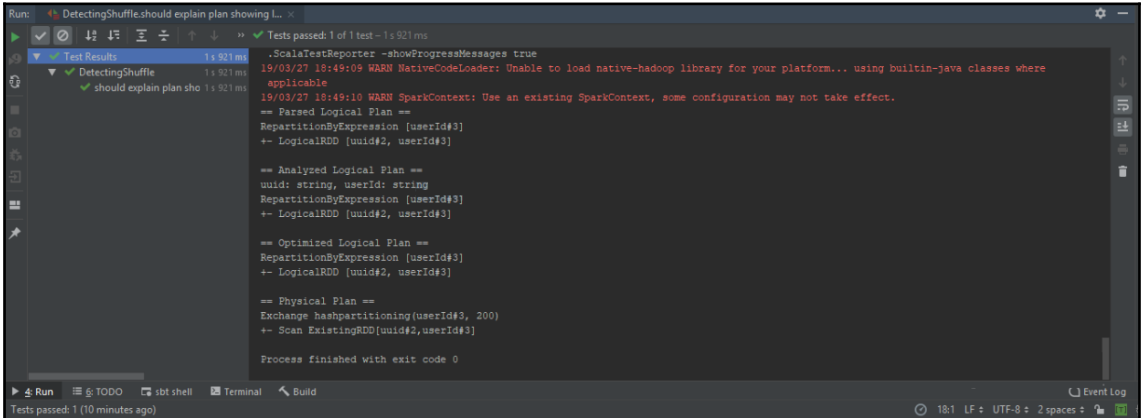
Chapter 05: Powerful Exploratory Data Analysis with MLlib

```
Chi squared test summary:  
method: pearson  
degrees of freedom = 4  
statistic = 0.5852136752136753  
pValue = 0.9646925263439344  
No presumption against null hypothesis: observed follows the same distribution as expected..
```

Chapter 08: Immutable Design



Chapter 09: Avoiding Shuffle and Reducing Operational Expenses



```
Run: DetectingShuffle.should explain plan showing L...
Tests passed: 1 of 1 test - 1 s 921 ms

Test Results
  DetectingShuffle
    should explain plan sho...

ScalaTestReporter -showProgressMessages true
19/03/27 18:49:09 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/03/27 18:49:10 WARN SparkContext: Use an existing SparkContext, some configuration may not take effect.

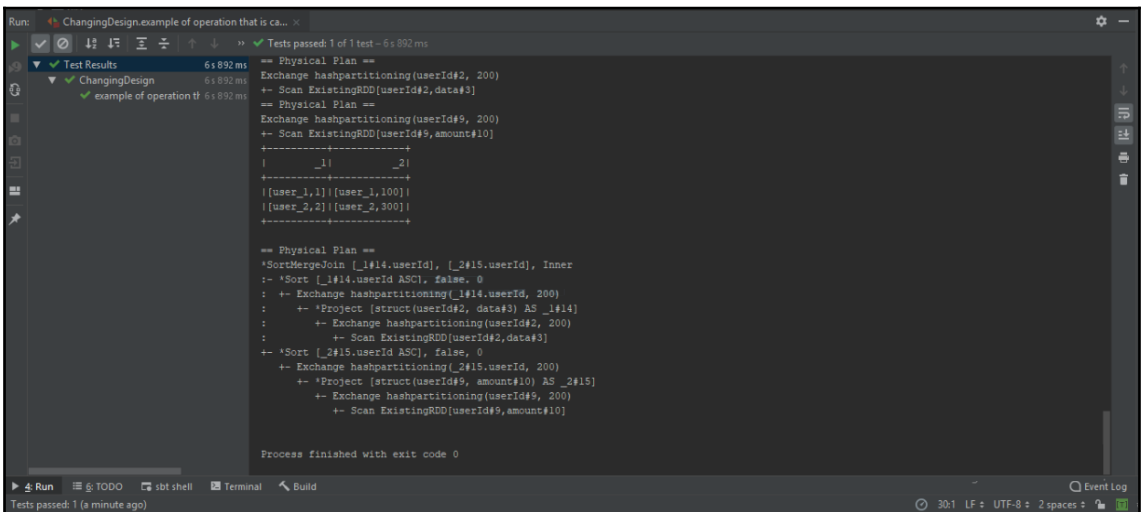
== Parsed Logical Plan ==
RepartitionByExpression [userId#3]
+- LogicalRDD [uid#2, userId#3]

== Analyzed Logical Plan ==
uid#2: string, userId: string
RepartitionByExpression [userId#3]
+- LogicalRDD [uid#2, userId#3]

== Optimized Logical Plan ==
RepartitionByExpression [userId#3]
+- LogicalRDD [uid#2, userId#3]

== Physical Plan ==
Exchange hashpartitioning(userId#3, 200)
+- Scan ExistingRDD[uid#2,userId#3]

Process finished with exit code 0
```



```
Run: ChangingDesign.example of operation that is ca...
Tests passed: 1 of 1 test - 6 s 892 ms

Test Results
  ChangingDesign
    example of operation th...

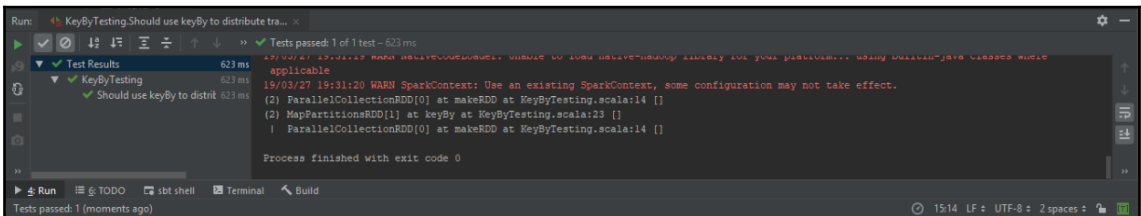
== Physical Plan ==
Exchange hashpartitioning(userId#2, 200)
+- Scan ExistingRDD[userId#2,data#3]

== Physical Plan ==
Exchange hashpartitioning(userId#9, 200)
+- Scan ExistingRDD[userId#9,amount#10]

-----+-----
|      _1      |      _2      |
| [user_1,1] | [user_1,100] |
| [user_2,2] | [user_2,300] |
-----+-----

== Physical Plan ==
*SortMergeJoin [_1#14.userId], [_2#15.userId], Inner
:- *Sort [_1#14.userId ASC], false, 0
: +- Exchange hashpartitioning(_1#14.userId, 200)
:    +- *Project [struct(userId#2, data#3) AS _1#14]
:      +- Exchange hashpartitioning(userId#2, 200)
:        +- Scan ExistingRDD[userId#2,data#3]
+- *Sort [_2#15.userId ASC], false, 0
: +- Exchange hashpartitioning(_2#15.userId, 200)
:    +- *Project [struct(userId#9, amount#10) AS _2#15]
:      +- Exchange hashpartitioning(userId#9, 200)
:        +- Scan ExistingRDD[userId#9,amount#10]

Process finished with exit code 0
```



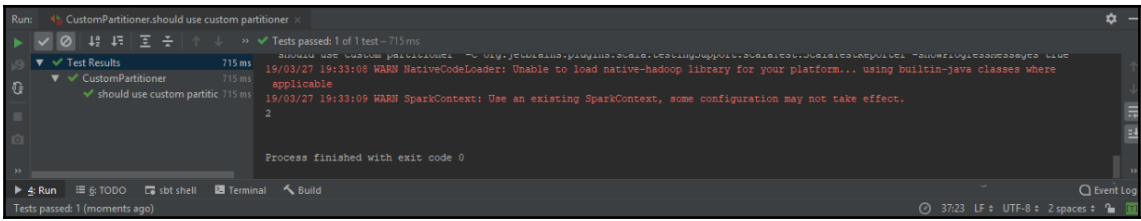
```
Run: KeyByTesting.Should use keyBy to distribute tra...
Tests passed: 1 of 1 test - 623 ms

Test Results
  KeyByTesting
    Should use keyBy to distr...

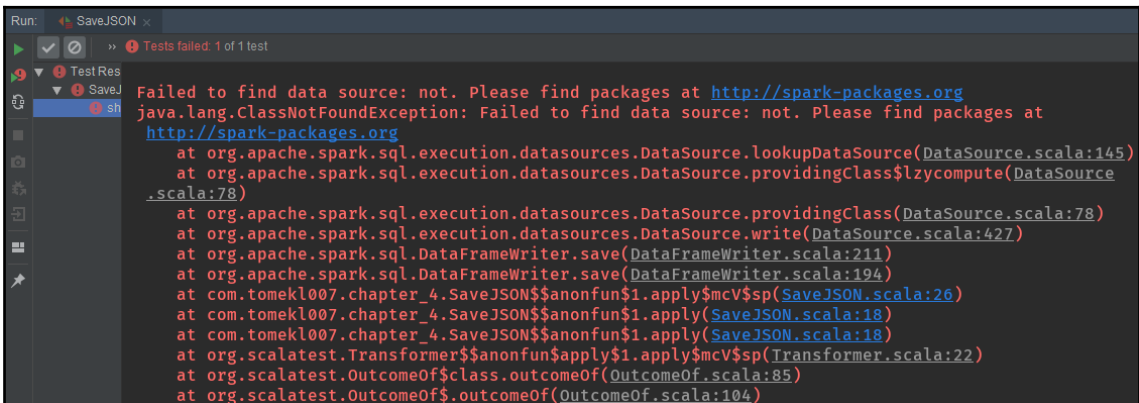
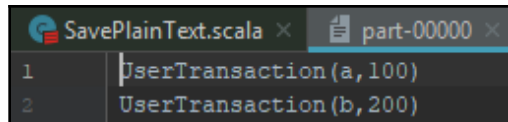
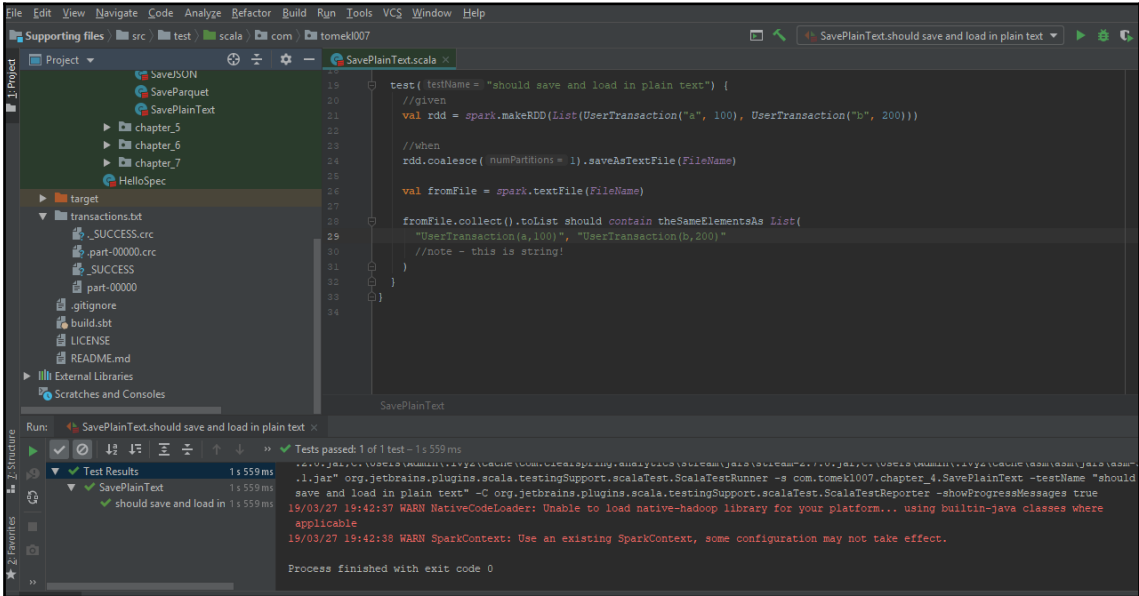
19/03/27 19:31:20 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/03/27 19:31:20 WARN SparkContext: Use an existing SparkContext, some configuration may not take effect.

(2) ParallelCollectionRDD[0] at makeRDD at KeyByTesting.scala:14 []
(2) MapPartitionsRDD[1] at keyBy at KeyByTesting.scala:23 []
  1 ParallelCollectionRDD[0] at makeRDD at KeyByTesting.scala:14 []

Process finished with exit code 0
```



Chapter 10: Saving Data in the Correct Format



```
Run: SaveCSV.should save and load CSV with hea... x
Test Results
  SaveCSV
    should
      +---+---+
      | a|100|
      +---+---+
      | b|200|
      +---+---+

1 did not equal 2
ScalaTestFailureLocation: com.tomekl007.chapter_4.SaveCSV$$anonfun$1 at (SaveCSV.scala:34)
Expected :2
Actual   :1
<Click to see difference>

org.scalatest.exceptions.TestFailedException: 1 did not equal 2
    at org.scalatest.Assertions$class.newAssertionFailedException(Assertions.scala:500)
    at org.scalatest.FunSuite.newAssertionFailedException(FunSuite.scala:1555)
```

```
part-r-00000-552e2fca-10d3-455c-a0aa-1198bebe7a4b.avro x part-r-00001-552e2fca-10d3-455c-a0aa-1198bebe7a4b.avro x
File was loaded in the wrong encoding: 'UTF-8' Reload in another encoding
1 Obj[avro.schema]{"type":"record","name":"topLevelRecord","fields":[{"name":"u$
```

Chapter 11: Working with the Spark Key/Value API

Run: TransformationsOnPairs.should use... x

Tests passed: 1 of 1 test – 862 ms

- Test Results (862 ms)
 - TransformationsOnPairs (862 ms)
 - should use transformatic (862 ms)

```
Testing started at 11:56 AM ...
"C:\Program Files\Java\jdk1.8.0_65\bin\java.exe" "-javaagent:C:\Program F
19/03/28 11:56:30 WARN NativeCodeLoader: Unable to load native-hadoop lib
19/03/28 11:56:31 WARN SparkContext: Use an existing SparkContext, some c

Process finished with exit code 0
```

Run | TODO | sbt shell | Terminal | Build | Event Log

Run: AggregateByKey x

Tests passed: 1 of 1 test – 828 ms

- Test Results (828 ms)
 - AggregateByKey (828 ms)
 - should use aggregateBy (828 ms)

```
Testing started at 11:59 AM ...
"C:\Program Files\Java\jdk1.8.0_65\bin\java.exe" "-javaagent:C:\Program F
19/03/28 11:59:32 WARN NativeCodeLoader: Unable to load native-hadoop lib
19/03/28 11:59:33 WARN SparkContext: Use an existing SparkContext, some c

Process finished with exit code 0
```

Run | TODO | sbt shell | Terminal | Build | Event Log

Run: ActionsOnPairs.should use action t... x

Tests passed: 1 of 1 test – 624 ms

- Test Results (624 ms)
 - ActionsOnPairs (624 ms)
 - should use action to see kv data (624 ms)

```
Testing started at 11:47 AM ...
"C:\Program Files\Java\jdk1.8.0_65\bin\java.exe" "-javaagent:C:\Program Files
19/03/28 11:47:08 WARN NativeCodeLoader: Unable to load native-hadoop library
19/03/28 11:47:09 WARN SparkContext: Use an existing SparkContext, some conf:

Process finished with exit code 0
```

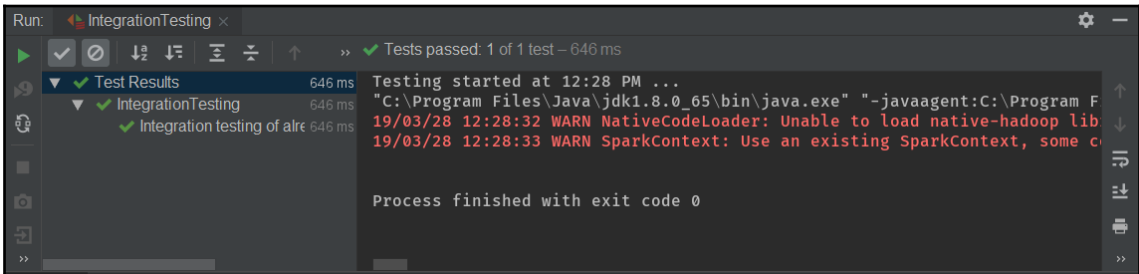
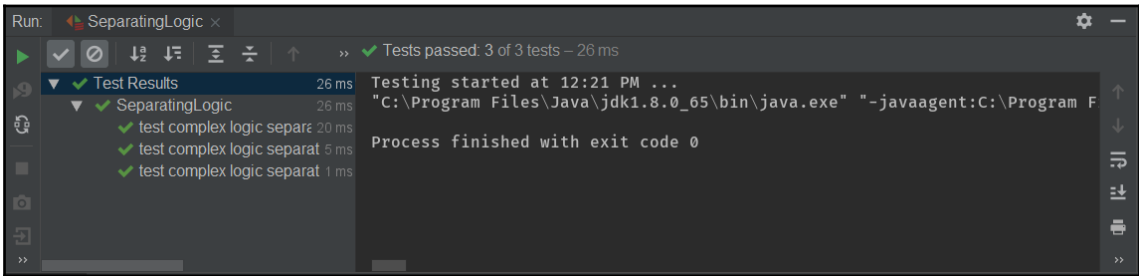
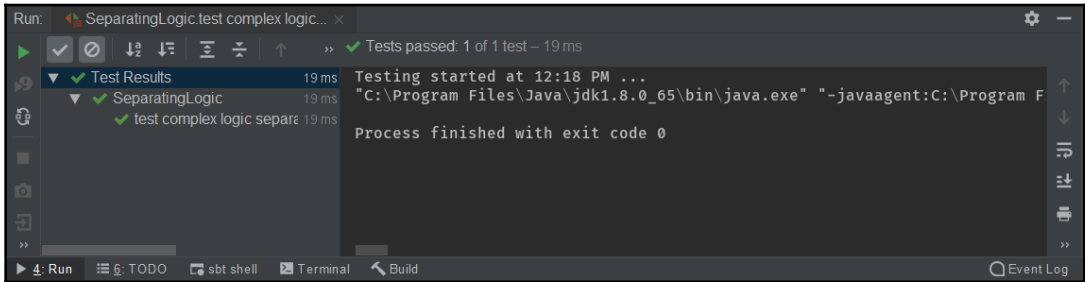
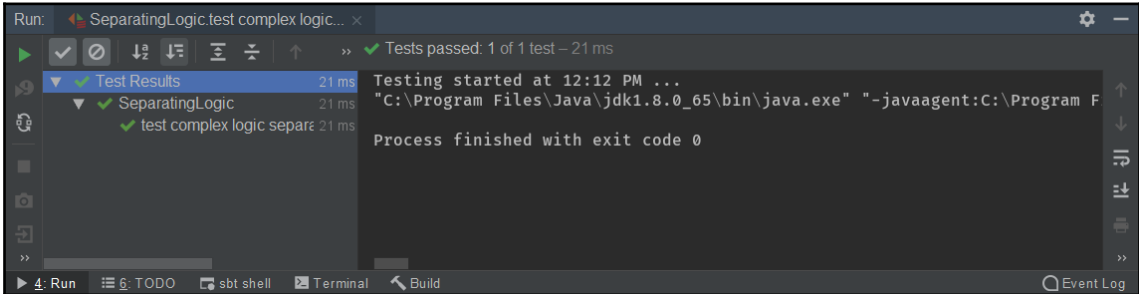
Run | TODO | sbt shell | Terminal | Build | Event Log

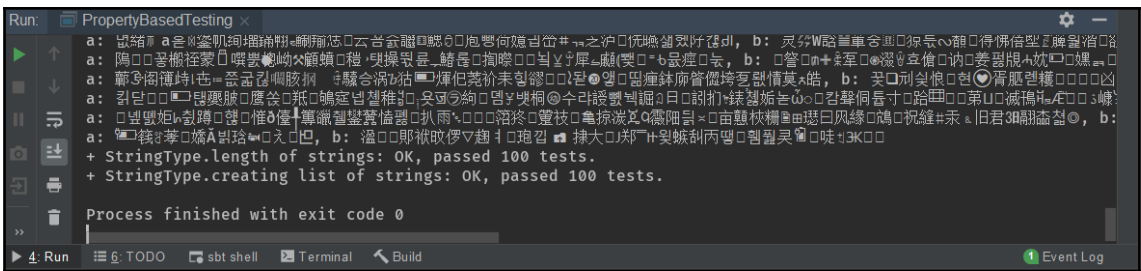
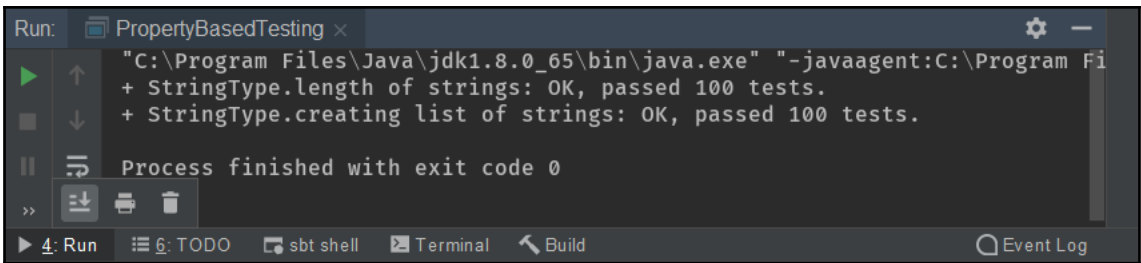
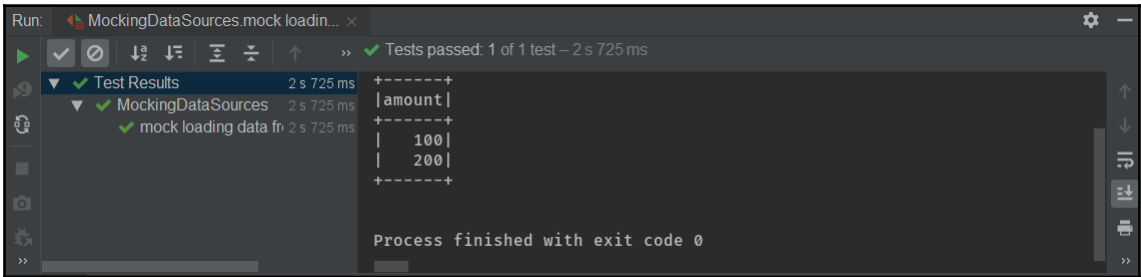
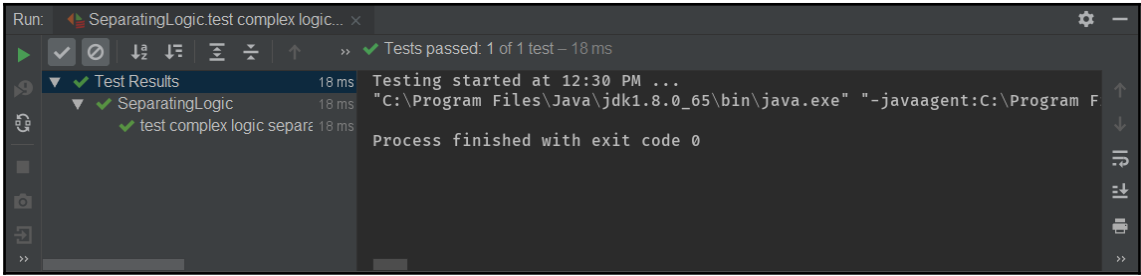
Tests passed: 1 (a minute ago) | 16:35 LF : UTF-8 : 2 spaces

```
Run: UsePartitioner x
Tests passed: 1 of 1 test - 598 ms
Test Results 598 ms
  UsePartitioner 598 ms
    should use different part 598 ms
19/03/28 11:51:19 WARN SparkContext: Use an existing SparkContext, some c
ShuffledRDD[2] at partitionBy at UsePartitioner.scala:28
ShuffledRDD[5] at partitionBy at UsePartitioner.scala:32
Process finished with exit code 0
Run | 6: TODO | sbt shell | Terminal | Build | Event Log
Tests passed: 1 (a minute ago) | 10:1 LF | UTF-8 | 2 spaces |
```

```
Run: CustomRangePartitionerTest.shoul... x
Tests passed: 1 of 1 test - 682 ms
Test Results 682 ms
  CustomRangePartitionerTe 682 ms
    should use custom rang 682 ms
19/03/28 11:54:13 WARN NativeCodeLoader: Unable to load native-hadoop lib
19/03/28 11:54:14 WARN SparkContext: Use an existing SparkContext, some c
for key: 100001 return 2
for key: 100 return 1
for key: 10 return 0
for key: 4 return 0
for key: 10 return 0
Process finished with exit code 0
Run | 6: TODO | sbt shell | Terminal | Build | Event Log
Tests passed: 1 (a minute ago) | 13:16 LF | UTF-8 | 2 spaces |
```

Chapter 12: Testing Apache Spark Jobs





Chapter 13: Leveraging the Spark GraphX API

```
Run: CreatingGraph.should load graph fr... x
>> Tests passed: 1 of 1 test - 1 s 195 ms

1 s 195 ms Testing started at 1:20 PM ...
1 s 195 ms "C:\Program Files\Java\jdk1.8.0_65\bin\java.exe" "-javaagent:C:\Program Files\JetBrains\IntelliJ I
19/03/28 13:20:03 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform...
19/03/28 13:20:04 WARN SparkContext: Use an existing SparkContext, some configuration may not take
((2,1),(3,1),1)
((3,1),(5,1),1)
((1,1),(2,1),1)
((1,1),(3,1),1)

Process finished with exit code 0
```

```
resolution-cache 29
30 val graph = Graph(users, relationships)
Choose Declaration
Graph (org.apache.spark.graphx) sbt: org.apache.spark:spark-graphx_2.11:2.0.0:jar (spark-graphx_2.11-2.0.0.jar)
apply (org.apache.spark.graphx.Graph) sbt: org.apache.spark:spark-graphx_2.11:2.0.0:jar (spark-graphx_2.11-2.0.0.jar)
```

```
VertexAPI.scala x Graph.scala x
48 */
49 val vertices: VertexRDD[VD]
50
51 /**
52 * An RDD containing the edges and their associated attributes. The entries in the RDD contain
53 * just the source id and target id along with the edge data.
54 *
55 * @return an RDD containing the edges in this graph
56 *
57 * @see [[Edge]] for the edge type.
58 * @see [[Graph#triplets]] to get an RDD which contains all the edges
59 * along with their vertex data.
60 *
61 */
62 val edges: EdgeRDD[ED]
63
64 /**
65 * An RDD containing the edge triplets, which are edges along with the vertex data associated with
66 * the adjacent vertices. The caller should use [[edges]] if the vertex data are not needed, i.e.
67 * if only the edge data and adjacent vertex ids are needed.
68 *
69 * @return an RDD containing edge triplets
70 *
71 * @example This operation might be used to evaluate a graph
72 * coloring where we would like to check that both vertices are a
73 * different color.
74 * {{{
75 * type Color = Int
76 * val graph: Graph[Color, Int] = GraphLoader.edgeListFile("hdfs://file.tsv")
77 * val numInvalid = graph.triplets.map(e => if (e.src.data == e.dst.data) 1 else 0).sum
78 * }}}
79 */
```

```

Run: VertexAPI.Should use Vertex API x
Tests passed: 1 of 1 test - 1 s 2 ms
Testing started at 1:33 PM ...
"C:\Program Files\Java\jdk1.8.0_65\bin\java.exe" "-javaagent:C:\Program Files\JetBrains\IntelliJ I...
19/03/28 13:33:52 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform...
19/03/28 13:33:53 WARN SparkContext: Use an existing SparkContext, some configuration may not take effect
[Stage 0:>
[rdd_14_0]
19/03/28 13:33:54 WARN Executor: 1 block locks were not released by TID = 5:
[rdd_14_1]
List((4,D), (2,B), (1,A), (3,C))

```

```

val resFromFilter = graph.edges.filter((e1) => e1.attr == "friend")
println(resFromFilter)

val res = graph.mapEdges(e => e.attr.toUpperCase)
println(res.edges.collect().toList)

```

Edge API > λ(e1: Any)

Vertex API x

ed: 1 of 1 test - 1 s 2 ms

attr	String
dstId	graphx.VertexId
srcId	graphx.VertexId
otherVertexId(vid: graphx.VertexId, relativeDirection: graphx.EdgeDirection)	graphx.VertexId
relativeDirection(vid: graphx.VertexId, otherVertexId: graphx.VertexId)	EdgeDirection
copy[ED](srcId: graphx.VertexId = Edge.dstId, dstId: graphx.VertexId = Edge.srcId)	Edge[ED]
clone()	AnyRef
hashCode()	Int
toString()	String

```

Run: EdgeAPI.Should use Edge API x
Tests passed: 1 of 1 test - 885 ms
Testing started at 2:51 PM ...
"C:\Program Files\Java\jdk1.8.0_65\bin\java.exe" "-javaagent:C:\Program Files\JetBrains\IntelliJ I...
19/03/28 14:51:19 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform...
19/03/28 14:51:20 WARN SparkContext: Use an existing SparkContext, some configuration may not take effect
[Stage 0:>
[rdd_12_0]
19/03/28 14:51:21 WARN Executor: 1 block locks were not released by TID = 1:
[rdd_12_1]
List(Edge(1,2,friend), Edge(1,3,friend))
19/03/28 14:51:21 WARN Executor: 1 block locks were not released by TID = 2:
[rdd_12_0]
19/03/28 14:51:21 WARN Executor: 1 block locks were not released by TID = 3:
[rdd_12_1]
List(Edge(1,2,FRIEND), Edge(1,3,FRIEND), Edge(2,4,WIFE))

Process finished with exit code 0

```

```

Run: CalculateDegreeTest.should calculate degree x
Tests passed: 1 of 1 test - 983 ms
Testing started at 3:06 PM ...
"C:\Program Files\Java\jdk1.8.0_65\bin\java.exe" "-javaagent:C:\Program Files\JetBrains\IntelliJ IDEA Co...
19/03/28 15:06:21 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using
19/03/28 15:06:22 WARN SparkContext: Use an existing SparkContext, some configuration may not take effect
[Stage 0:>
[rdd_9_1]
19/03/28 15:06:23 WARN Executor: 1 block locks were not released by TID = 6:
[rdd_9_0]

Process finished with exit code 0

```

```
Run: CalculateDegreeTest x
>> Tests passed: 1 of 1 test - 1 s 47 ms
1 s 47 ms Testing started at 3:09 PM ...
1 s 47 ms "C:\Program Files\Java\jdk1.8.0_65\bin\java.exe" "-javaagent:C:\Program Files\JetBrains\IntelliJ IDEA Co
19/03/28 15:09:33 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using
19/03/28 15:09:34 WARN SparkContext: Use an existing SparkContext, some configuration may not take effect
[Stage 0:> (0 + 0) / 2]19/03/28 15:09:35 WARN E
[rdd_9_0]
19/03/28 15:09:35 WARN Executor: 1 block locks were not released by TID = 7:
[rdd_9_1]

Process finished with exit code 0
```

```
Run: CalculateDegreeTest x
>> Tests passed: 1 of 1 test - 964 ms
964 ms "C:\Program Files\Java\jdk1.8.0_65\bin\java.exe" "-javaagent:C:\Program Files\JetBrains\IntelliJ IDEA Co
964 ms 19/03/28 15:11:29 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using
964 ms 19/03/28 15:11:29 WARN SparkContext: Use an existing SparkContext, some configuration may not take effect
19/03/28 15:11:30 WARN Executor: 1 block locks were not released by TID = 6:
[rdd_9_0]
19/03/28 15:11:30 WARN Executor: 1 block locks were not released by TID = 7:
[rdd_9_1]

Process finished with exit code 0
```

1	2	1
2	4	1
3	1	2
4	6	3
5	7	3
6	7	6
7	6	7
8	3	7

1	1,BarackObama,Barack Obama
2	2,ladygaga,Goddess of Love
3	3,jeresig,John Resig
4	4,justinbieber,Justin Bieber
5	6,matei_zaharia,Matei Zaharia
6	7,odersky,Martin Odersky
7	8,anonsys

```
Run: PageRankTest should calculate pa... x
>> Tests passed: 1 of 1 test - 7 s 423 ms
7 s 423 ms [rdd_896_0]
7 s 423 ms List((BarackObama,1.4588814096664682), (ladygaga,1.390049198216498), (odersky,1.2973176314422592)
7 s 423 ms , (jeresig,0.9993442038507723), (matei_zaharia,0.7013599933629602), (justinbieber,0.15))
Process finished with exit code 0
```