

Lab #8 – Spark

1. Answer the Review Questions for Module 1,2 and 3. Take a screenshot of both answered reviews.

The image displays two screenshots of the 'Spark Fundamentals I' course interface on the Cognitive Class platform. The top screenshot shows 'REVIEW QUESTION 1' with the question 'What gives Spark its speed advantage for complex applications?'. The correct answer, 'Spark makes extensive use of in-memory computations', is selected and marked with a green checkmark. The bottom screenshot shows 'REVIEW QUESTION 2' with the question 'For what purpose would an Engineer use Spark? Select all that apply.'. The selected answers are 'Programming with Spark's API', 'Developing a data processing system', and 'Tuning an application for a business use case', all marked with green checkmarks. The interface includes a sidebar with course modules, a top navigation bar, and a bottom taskbar.

Spark Fundamentals I
Cognitive Class BD0211EN

Hands-on lab exercise: Getting started

Graded Review Questions

Module 2: Resilient Distributed Dataset and DataFrames

Module 3: Spark application programming

Module 4: Introduction to the Spark libraries

Module 5: Spark configuration, monitoring and tuning

Final Exam

Course Survey and Feedback

Completion Certificate and Badge

REVIEW QUESTION 1 (1/1 point)

What gives Spark its speed advantage for complex applications?

☐ Spark can cover a wide range of workloads under one system

☒ Spark makes extensive use of in-memory computations ✓

☐ Spark extends the MapReduce model

☐ Various libraries provide Spark with additional functionality

☐ All of the above

REVIEW QUESTION 2 (1/1 point)

For what purpose would an Engineer use Spark? Select all that apply.

☐ Analyzing data to obtain insights

☒ Programming with Spark's API

☐ Transforming data into a useable form for analysis

☒ Developing a data processing system

☒ Tuning an application for a business use case ✓

RESET You have used 1 of 2 submissions

You have used 2 of 2 submissions

The image displays two screenshots of a web-based course interface for 'Spark Fundamentals I' by Cognitive Class. The top screenshot shows 'REVIEW QUESTION 3' with a point value of 1/1. The question asks which statements are true of the Resilient Distributed Dataset (RDD). The options are: 'There are three types of RDD operations.', 'RDDs allow Spark to reconstruct transformations' (checked), 'RDDs only add a small amount of code due to tight integration' (checked), 'RDD action operations do not return a value', and 'RDD is a distributed collection of elements parallelized across the cluster.' (checked). A 'RESET' button is visible below the question. The bottom screenshot shows 'REVIEW QUESTION 1' (1/1 point) asking which methods can be used to create an RDD. The options are: 'Creating a directed acyclic graph (DAG)', 'Parallelizing an existing Spark collection' (checked), 'Referencing a Hadoop-supported dataset' (checked), 'Using data that resides in Spark', and 'Transforming an existing RDD to form a new one' (checked). Below this is 'REVIEW QUESTION 2' (1/1 point) asking 'What happens when an action is executed?'. A sidebar on the left of the bottom screenshot lists course modules: 'Module 3: Spark application programming', 'Module 4: Introduction to the Spark libraries', 'Module 5: Spark configuration, monitoring and tuning', 'Final Exam', 'Course Survey and Feedback', and 'Completion Certificate and Badge'. The user's name 'PiyushBajaj' is visible in the top right of both screenshots.

COGNITIVE CLASS Spark Fundamentals I Cognitive Class BD0211EN Explore new learning opportunities PiyushBajaj

You have used 2 of 2 submissions

REVIEW QUESTION 3 (1/1 point)

Which of the following statements are true of the Resilient Distributed Dataset (RDD)? Select all that apply.

- ☐ There are three types of RDD operations.
- ☒ RDDs allow Spark to reconstruct transformations
- ☒ RDDs only add a small amount of code due to tight integration
- ☐ RDD action operations do not return a value
- ☒ RDD is a distributed collection of elements parallelized across the cluster.

RESET You have used 1 of 2 submissions

COGNITIVE CLASS Spark Fundamentals I Cognitive Class BD0211EN Explore new learning opportunities PiyushBajaj

Review Questions

- Module 3: Spark application programming
- Module 4: Introduction to the Spark libraries
- Module 5: Spark configuration, monitoring and tuning
- Final Exam
- Course Survey and Feedback
- Completion Certificate and Badge

REVIEW QUESTION 1 (1/1 point)

Which of the following methods can be used to create a Resilient Distributed Dataset (RDD)? Select all that apply.

- ☐ Creating a directed acyclic graph (DAG)
- ☒ Parallelizing an existing Spark collection
- ☒ Referencing a Hadoop-supported dataset
- ☐ Using data that resides in Spark
- ☒ Transforming an existing RDD to form a new one

RESET You have used 1 of 2 submissions

REVIEW QUESTION 2 (1/1 point)

What happens when an action is executed?

The screenshot displays a web browser window with multiple tabs open, including 'Application', 'Files', 'Instructions | Graded Re...', 'LinkedIn', and 'salammagari - Facebook'. The active tab shows a quiz page for 'Spark Fundamentals I' on the 'COGNITIVE CLASS' platform. The URL is <https://courses.cognitiveclass.ai/courses/course-v1:BigDataUniversity+BD0211EN+2016/courseware/8fdff50a06d44cd2a01cf0277e131f02/07d708d630d74dd96abbaf927d3183/>. The user is identified as 'PiyushBajaj'.

The quiz interface shows two questions:

REVIEW QUESTION 2 (1/1 point)
What happens when an action is executed?

- ☐ Executors prepare the data for operation in parallel
- ☐ A cache is created for storing partial results in memory
- ☐ The driver sends code to be executed on each block
- ☐ Data is partitioned into different blocks across the cluster
- ☒ All of the above ✓

REVIEW QUESTION 3 (1/1 point)
Which of the following statements is true of RDD persistence? Select all that apply.

- ☒ Persistence through caching provides fault tolerance
- ☒ Future actions can be performed significantly faster
- ☐ Each partition is replicated on two cluster nodes
- ☐ RDD persistence always improves space efficiency
- ☐ By default, objects that are too big for memory are stored on the disk

A green checkmark indicates the correct answer. Below the question, it says 'You have used 2 of 2 submissions'. Navigation arrows are visible at the bottom of the question area.

The screenshot displays the 'Spark Fundamentals I' course page on the Cognitive Class platform. The interface includes a sidebar with course modules, a main content area with instructions and review questions, and a bottom taskbar with various application icons.

Course Sidebar:

- Application Programming - Part 1 (5:12)
- Application Programming - Part 2 (5:42)
- Application Programming - Part 3 (5:37)
- Optional hands-on lab exercise
- Graded Review Questions (Review Questions)
- Module 4: Introduction to the Spark libraries
- Module 5: Spark configuration, monitoring and tuning
- Final Exam
- Course Survey and Feedback
- Completion Certificate and Badge

Instructions:

- One attempt - For True/False questions
- Two attempts - For any question other than True/False
- Clicking the "Final Check" button when it appears, means your submission is **FINAL**. You will **NOT** be able to resubmit your answer for that question ever again
- Check your grades in the course at any time by clicking on the "Progress" tab

REVIEW QUESTION 1 (1/1 point)

What is SparkContext?

- ☐ A programming language for applications
- ☐ A tool for linking to nodes
- ☐ The built-in shell for the Spark engine
- ☐ A tool that provides fault tolerance
- ☒ An object that represents the connection to a Spark cluster ✓

RESET You have used 1 of 2 submissions

REVIEW QUESTION 2 (1/1 point)

Which of the following methods can be used to pass functions to Spark? Select all that apply.

- ☐ Transformations and actions
- ☒ Passing by reference
- ☒ Static methods in a global singleton
- ☐ Import statements
- ☒ Anonymous function syntax

✓

RESET You have used 1 of 2 submissions

REVIEW QUESTION 3 (1 point possible)

Which of the following is a main component of a Spark application's source code?

- ☐ Business Logic

The screenshot shows a web browser window with the URL <https://courses.cognitiveclass.ai/courses/course-v1:BigDataUniversity+BD0211EN+2016/courseware/dd62d684b714d279e5945bb7fe44459/3de96dc4fc3449fbec777cb1177fb0c/>. The page is titled "Spark Fundamentals I" and "Cognitive Class BD0211EN". A "RESET" button is visible, along with the text "You have used 1 of 2 submissions". The main content is "REVIEW QUESTION 3 (1/1 point)" which asks: "Which of the following is a main component of a Spark application's source code?". The options are:

- ☐ SparkContext object
- ☐ Import statements
- ☐ Business Logic
- ☐ Transformations and actions
- ☒ All of the above

 The "All of the above" option is selected and marked with a green checkmark. Below the options, it says "You have used 2 of 2 submissions". The browser's taskbar at the bottom shows various application icons and the system clock indicating 02:40 on 01/11/2017.

2. b. Run the Sample Scala application to calculate the value of Pi. Take a screenshot of the output of the run of spark-submit.

The screenshot shows a terminal window with the output of a `spark-submit` command. The output is a log of messages from the Spark driver and executors. Key messages include:

- INFO Remoting: Starting remoting
- INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@10.0.2.15:44849]
- INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 44849.
- INFO SparkEnv: Registering MapOutputTracker
- INFO SparkEnv: Registering BlockManagerMaster
- INFO DiskBlockManager: Created local directory at /tmp/blockmgr-351f6015-3a2d-4b0c-841c-abfa6dc7af36
- INFO MemoryStore: MemoryStore started with capacity 517.4 MB
- INFO SparkEnv: Registering OutputCommitCoordinator
- INFO Utils: Successfully started service 'SparkUI' on port 4040.
- INFO HttpFileServer: HTTP File server directory is /tmp/spark-672615d3-01b6-41c2-a910-a13e3c9b7d01/httpd-b3581be5-bd28-4e89-8230-f7822e80adea
- INFO HttpServer: Starting HTTP Server
- INFO Utils: Successfully started service 'HTTP file server' on port 40806.
- INFO SparkContext: Added JAR file:/home/osboxes/spark/target/scala-2.10/sparkpl-project_2.10-1.0.jar at http://10.0.2.15:40065/jars/sparkpl-project_2.10-1.0.jar with timestamp 150978240215
- INFO Executor: Starting executor ID driver on host localhost
- INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 34821.
- INFO BlockManagerMaster: Trying to register BlockManager
- INFO BlockManagerMasterEndpoint: Registering block manager localhost:34821 with 517.4 MB RAM, BlockManagerId(driver, localhost, 34821)
- INFO BlockManagerMaster: Registered BlockManager
- INFO SparkContext: Starting job: reduce at Spark.scala:18
- INFO DAGScheduler: Got job 0 (reduce at Spark.scala:18) with 2 output partitions
- INFO DAGScheduler: Final stage: ResultStage 0 (reduce at Spark.scala:18)
- INFO DAGScheduler: Parents of final stage: List()
- INFO DAGScheduler: Missing parents: List()
- INFO DAGScheduler: Submitting ResultStage 0 (MapPartitionsRDD[1] at map at Spark.scala:14), which has no missing parents
- INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 1202.0 B, free 3.0 KB)
- INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:34821 (size: 1202.0 B, free: 517.4 MB)
- INFO SparkContext: Created broadcast 0 from broadcast at DAGScheduler.scala:1006
- INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 0 (MapPartitionsRDD[1] at map at Spark.scala:14)
- INFO TaskSchedulerImpl: Adding task set 0.0 with 2 tasks
- INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, partition 0, PROCESS_LOCAL, 2144 bytes)
- INFO TaskSetManager: Starting task 1.0 in stage 0.0 (TID 1, localhost, partition 1, PROCESS_LOCAL, 2144 bytes)
- INFO Executor: Running task 0.0 in stage 0.0 (TID 0)
- INFO Executor: Running task 1.0 in stage 0.0 (TID 1)
- INFO Executor: Fetching http://10.0.2.15:40065/jars/sparkpl-project_2.10-1.0.jar with timestamp 150978240215
- INFO Utils: Fetching http://10.0.2.15:40065/jars/sparkpl-project_2.10-1.0.jar to /tmp/spark-672615d3-01b6-41c2-a910-a13e3c9b7d01/userFiles-a21fcc38-08a7-4500-8bf4-71dcac97e618/sparkpl-project_2.10-1.0.jar to class lo
- INFO Executor: Adding file:/tmp/spark-672615d3-01b6-41c2-a910-a13e3c9b7d01/userFiles-a21fcc38-08a7-4500-8bf4-71dcac97e618/sparkpl-project_2.10-1.0.jar to class lo
- INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1031 bytes result sent to driver
- INFO Executor: Finished task 1.0 in stage 0.0 (TID 1). 1031 bytes result sent to driver
- INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 498 ms on localhost (1/2)
- INFO TaskSetManager: Finished task 1.0 in stage 0.0 (TID 1) in 444 ms on localhost (2/2)
- INFO DAGScheduler: ResultStage 0 (reduce at Spark.scala:18) finished in 0.552 s
- INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
- INFO DAGScheduler: Job 0 finished: reduce at Spark.scala:18, took 1.185784 s
- INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
- INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
- INFO MemoryStore: MemoryStore cleared
- INFO BlockManager: BlockManager stopped
- INFO BlockManagerMaster: BlockManagerMaster stopped
- INFO OutputCommitCoordinator\$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
- INFO RemoterActorRefProvider\$RemotingTerminator: Shutting down remote daemon.
- INFO RemoterActorRefProvider\$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.
- INFO SparkContext: Successfully stopped SparkContext
- INFO ShutdownHookManager: Shutdown hook called
- INFO ShutdownHookManager: Deleting directory /tmp/spark-672615d3-01b6-41c2-a910-a13e3c9b7d01/httpd-b3581be5-bd28-4e89-8230-f7822e80adea
- INFO ShutdownHookManager: Deleting directory /tmp/spark-672615d3-01b6-41c2-a910-a13e3c9b7d01

3. Spark can be run with Java or Python as well. Continuing the same hands-on lab exercise, setup/write the WordCount program in Java based on the sample given. Run the wordcount program and take a screenshot of the output. Try the program with your own input textfile afterwards.

```

ln: 5
-DskipTests: 1
downloaded: 1
versions: 1
online: 1
Guide[http://spark.apache.org/docs/latest/configuration.html]: 1
comes: 1
[building: 1
Python: 2
Many: 1
building: 2
Running: 1
from: 1
ways: 1
Online: 1
site: 1
other: 1
Example: 1
analysis: 1
sc.parallelize(range(1000)).count(): 1
yous: 4
runs: 1
Building: 1
higher-level: 1
protocols: 1
guidance: 2
a: 8
guide: 1
name: 1
fast: 1
SQL: 2
will: 1
Terminal
core: 1
: 67
web: 1
"local[N]": 1
programs: 2
package: 1
that: 2
MLlib: 1
["Building: 1
shell: 2
Scala: 1
and: 10
command: 2
./dev/run-tests: 1
sample: 1
17/11/04 01:57:39 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
17/11/04 01:57:39 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
17/11/04 01:57:39 INFO MemoryStore: MemoryStore cleared
17/11/04 01:57:39 INFO BlockManager: BlockManager stopped
17/11/04 01:57:39 INFO BlockManagerMaster: BlockManagerMaster stopped
17/11/04 01:57:39 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
17/11/04 01:57:39 INFO RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
17/11/04 01:57:39 INFO RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.
17/11/04 01:57:39 INFO SparkContext: Successfully stopped SparkContext
17/11/04 01:57:39 INFO ShutdownHookManager: Shutdown hook called
17/11/04 01:57:39 INFO ShutdownHookManager: Deleting directory /tmp/spark-2be41112-7195-4a03-980c-780a252ae407/httpd-25e40253-Sab1-4e09-b6b5-26b810c44b2c
17/11/04 01:57:39 INFO ShutdownHookManager: Deleting directory /tmp/spark-2be41112-7195-4a03-980c-780a252ae407
osboxes@osboxes:~$

```

Full output:

```

package: 1
For: 2
Programs: 1
processing.: 1
Because: 1
The: 1
cluster.: 1
its: 1
[run: 1
APIs: 1
have: 1
Try: 1
computation: 1
through: 1
several: 1
This: 2
graph: 1
Hive: 2
storage: 1
["Specifying: 1
To: 2

```

page](http://spark.apache.org/documentation.html): 1
Once: 1
"yarn": 1
prefer: 1
SparkPi: 2
engine: 1
version: 1
file: 1
documentation,: 1
processing,: 1
the: 21
are: 1
systems.: 1
params: 1
not: 1
different: 1
refer: 2
Interactive: 2
R,: 1
given.: 1
17/11/04 01:57:39 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 192 ms on localhost (1/1)
if: 4
build: 3
17/11/04 01:57:39 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
when: 1
be: 2
Tests: 1
Apache: 1
./bin/run-example: 2
programs,: 1
including: 3
Spark.: 1
package.: 1
1000).count(): 1
Versions: 1
HDFS: 1
Data.: 1
>>>: 1
programming: 1
Testing: 1
module,: 1
Streaming: 1
environment: 1
run:: 1
clean: 1
1000:: 2
rich: 1
GraphX: 1
Please: 3
is: 6
run: 7
URL,: 1
threads.: 1
same: 1
MASTER=spark://host:7077: 1
on: 5
built: 1
against: 1

[Apache: 1
tests: 2
examples: 2
at: 2
optimized: 1
usage: 1
using: 2
graphs: 1
talk: 1
Shell: 2
class: 2
abbreviated: 1
directory.: 1
README: 1
computing: 1
overview: 1
`examples`: 2
example:: 1
##: 8
N: 1
set: 2
use: 3
Hadoop-supported: 1
tests](https://cwiki.apache.org/confluence/display/SPARK/Useful+Developer+Tools).: 1
running: 1
find: 1
contains: 1
project: 1
Pi: 1
need: 1
or: 3
Big: 1
Java,: 1
high-level: 1
uses: 1
<class>: 1
Hadoop,: 2
available: 1
requires: 1
(You: 1
see: 1
Documentation: 1
of: 5
tools: 1
using:: 1
cluster: 2
must: 1
supports: 2
built,: 1
system: 1
build/mvn: 1
Hadoop: 3
this: 1
Version"](<http://spark.apache.org/docs/latest/building-spark.html#specifying-the-hadoop-version>): 1
particular: 2
Python: 2
Spark: 13
general: 2

YARN,: 1
pre-built: 1
[Configuration: 1
locally: 2
library: 1
A: 1
locally.: 1
sc.parallelize(1: 1
only: 1
Configuration: 1
following: 2
basic: 1
#: 1
changed: 1
More: 1
which: 2
learning,: 1
first: 1
./bin/pyspark: 1
also: 4
should: 2
for: 11
[params]`: 1
documentation: 3
[project: 2
mesos://: 1
Maven](http://maven.apache.org/): 1
setup: 1
<http://spark.apache.org/>: 1
latest: 1
your: 1
MASTER: 1
example: 3
scala>: 1
DataFrames,: 1
provides: 1
configure: 1
distributions.: 1
can: 6
About: 1
instructions.: 1
do: 2
easiest: 1
no: 1
how: 2
`./bin/run-example: 1
Note: 1
individual: 1
spark://: 1
It: 2
Scala: 2
Alternatively,: 1
an: 3
variable: 1
submit: 1
machine: 1
thread,: 1
them,: 1

detailed: 2
stream: 1
And: 1
distribution: 1
return: 2
Thriftserver: 1
./bin/spark-shell: 1
"local": 1
start: 1
You: 3
Spark](#building-spark).: 1
one: 2
help: 1
with: 3
print: 1
Spark"](<http://spark.apache.org/docs/latest/building-spark.html>): 1
data: 1
wiki](<https://cwiki.apache.org/confluence/display/SPARK>): 1
in: 5
-DskipTests: 1
downloaded: 1
versions: 1
online: 1
Guide](<http://spark.apache.org/docs/latest/configuration.html>): 1
comes: 1
[building: 1
Python,: 2
Many: 1
building: 2
Running: 1
from: 1
way: 1
Online: 1
site,: 1
other: 1
Example: 1
analysis.: 1
sc.parallelize(range(1000)).count(): 1
you: 4
runs.: 1
Building: 1
higher-level: 1
protocols: 1
guidance: 2
a: 8
guide,: 1
name: 1
fast: 1
SQL: 2
will: 1
instance:: 1
to: 14
core: 1
: 67
web: 1
"local[N]": 1
programs: 2
package.): 1

that: 2
 MLib: 1
 ["Building: 1
 shell:: 2
 Scala,: 1
 and: 10
 command,: 2
 ./dev/run-tests: 1
 sample: 1

Extra Credit:

5b. b. Create your own input text file. Run the wordcount program. Paste the contents of your input textfile along with the contents of parts-00000 and part-00001.

```

Machine View Input Devices Help
Terminal
osboxes@osboxes: ~/Spark/output
drwxrwxr-x 2 osboxes osboxes 4096 Nov 4 02:57 output
drwxrwxr-x 3 osboxes osboxes 4096 Nov 4 02:08 project
-rw-rw-r-- 1 osboxes osboxes 136 Nov 4 02:06 spark.sbt
drwxrwxr-x 3 osboxes osboxes 4096 Nov 4 02:01 src
drwxrwxr-x 4 osboxes osboxes 4096 Nov 4 02:09 target
osboxes@osboxes:~/Spark$ cd output
osboxes@osboxes:~/Spark/output$ ls -l
total 4
-rw-r--r-- 1 osboxes osboxes 102 Nov 4 02:57 part-00000
-rw-r--r-- 1 osboxes osboxes 0 Nov 4 02:57 _SUCCESS
osboxes@osboxes:~/Spark/output$ cat part-00000
(long,1)
(you,1)
(its,2)
(a,2)
(here,1)
(with,1)
(I,1)
(day,2)
(wish,1)
(were,1)
(beautiful,1)
(me,1)
osboxes@osboxes:~/Spark/output$

Using Scala version 2.10.5 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_151)
Type in expressions to have them evaluated.
Type :help for more information.
17/11/04 02:52:12 WARN Utils: Your hostname, osboxes resolves to a loopback address: 127.0.1.1; using 1
17/11/04 02:52:12 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Spark context available as sc.
17/11/04 02:52:17 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependence
17/11/04 02:52:18 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependence
17/11/04 02:52:24 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.v
17/11/04 02:52:24 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException
17/11/04 02:52:27 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependence
17/11/04 02:52:27 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependence
17/11/04 02:52:32 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.v
17/11/04 02:52:32 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException
SQL context available as sqlContext.

scala> val inputfile = sc.textFile("input.txt")
inputfile: org.apache.spark.rdd.RDD[String] = input.txt MapPartitionsRDD[1] at textFile at <console>:27

scala> val counts = inputfile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey( + );

```