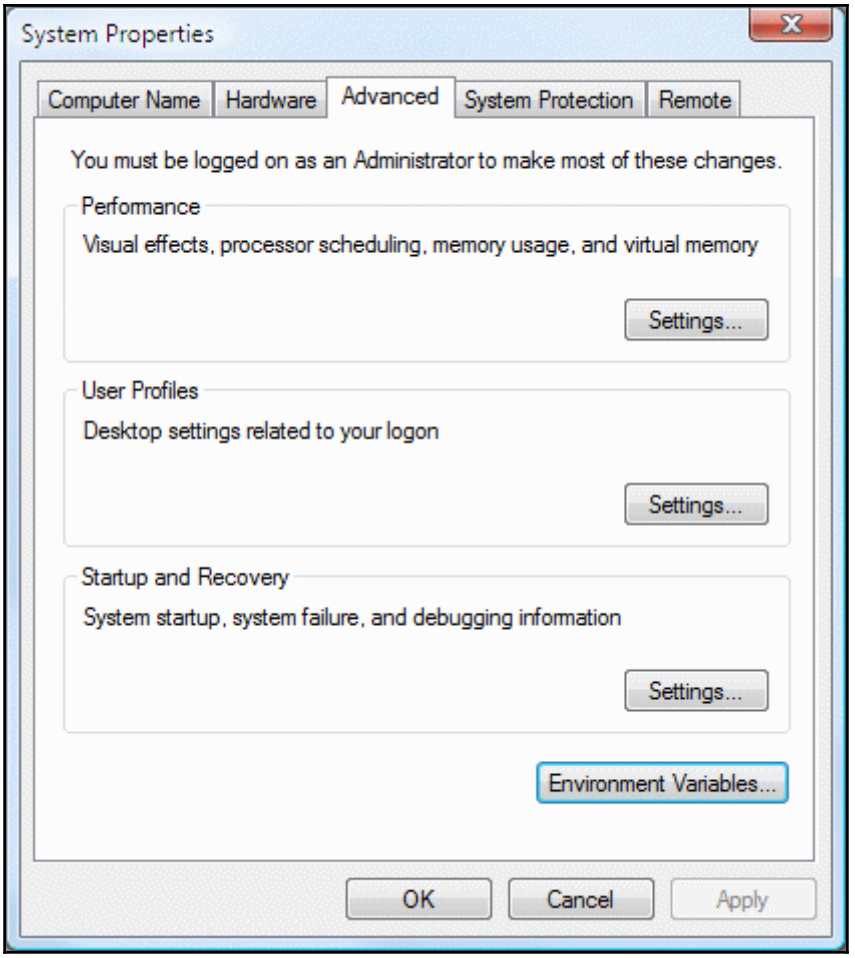
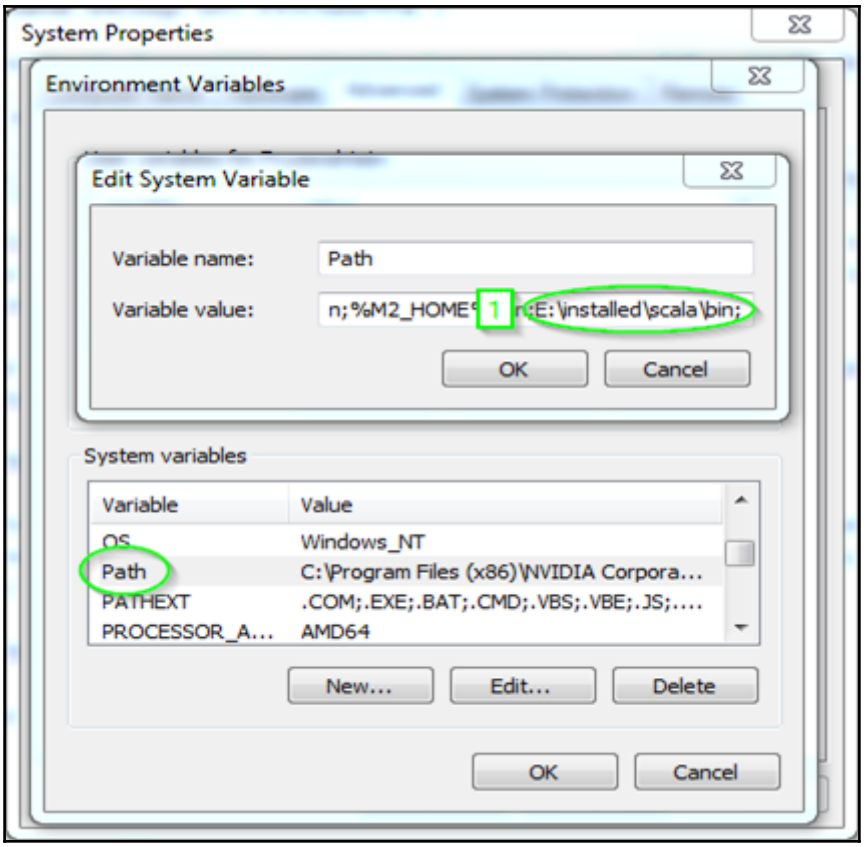
# 1
# Graphic Bundle

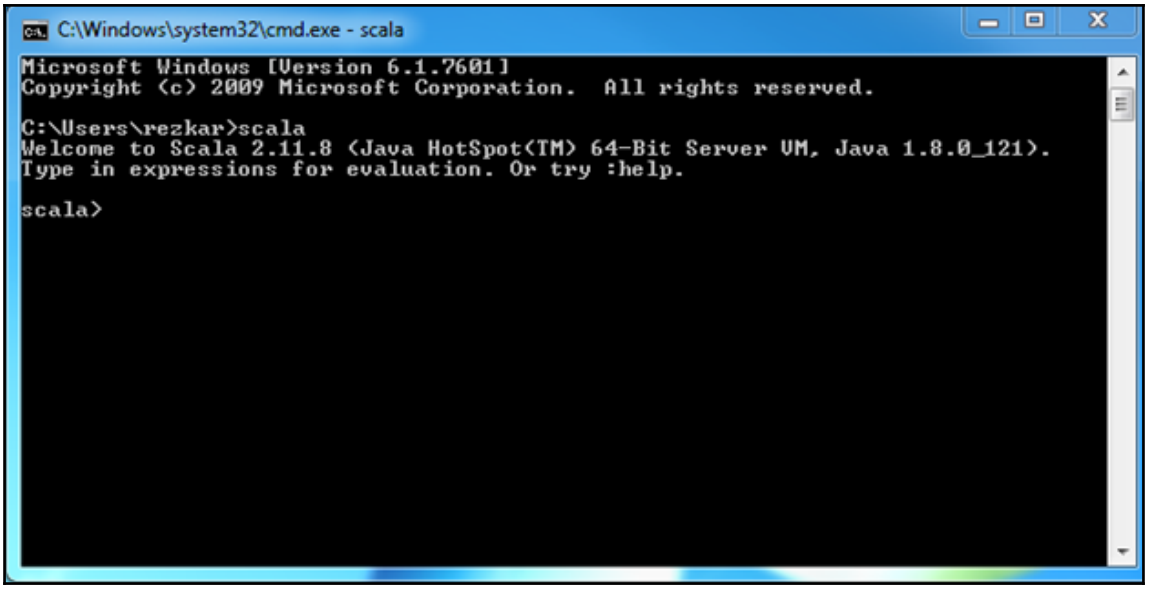
## Chapter 1: Introduction to Scala

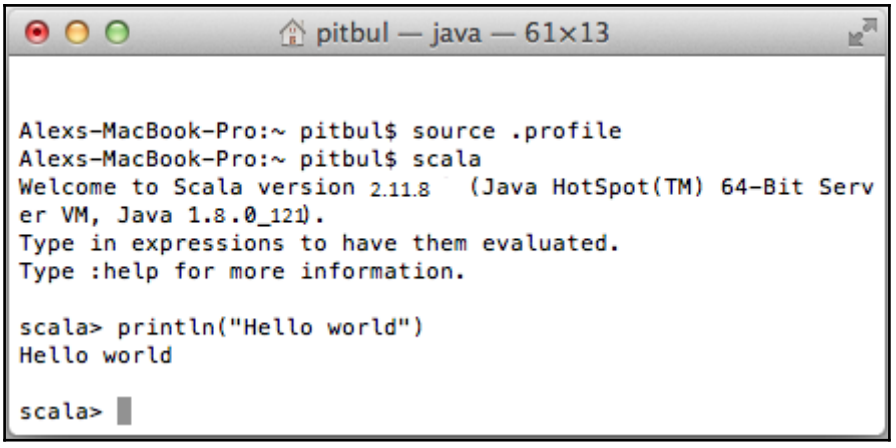
| InfoWorld Scorecard | Documentation and help system (15.0%) | Ease of use (30.0%) | Plug-in ecosystem (25.0%) | Java features (30.0%) | Overall Score (100%) |
|---|---|---|---|---|---|
| Eclipse 3.6 | 8.0 | 6.0 | 10.0 | 8.0 | 7.9 ★★★☆☆ |
| JetBrains IntelliJ IDEA 9.0.3 | 7.0 | 9.0 | 8.0 | 9.0 | 8.5 ★★★★☆ |
| NetBeans 6.9 | 8.0 | 8.0 | 8.0 | 8.0 | 8.0 ★★★★☆ |
| Oracle JDeveloper Studio 11g (11.1.1.3.0) | 9.0 | 8.0 | 5.0 | 8.0 | 7.4 ★★★☆☆ |

| Archive | System | Size |
|---|---|---|
| scala-2.11.8.tgz | Mac OS X, Unix, Cygwin | 27.35M |
| scala-2.11.8.msi | Windows (msi installer) | 109.35M |
| scala-2.11.8.zip | Windows | 27.40M |
| scala-2.11.8.deb | Debian | 76.02M |
| scala-2.11.8.rpm | RPM package | 108.16M |
| scala-docs-2.11.8.txz | API docs | 46.00M |
| scala-docs-2.11.8.zip | API docs | 84.21M |
| scala-sources-2.11.8.tar.gz | Sources | |


System Properties
Computer Name
Hardware
Advanced
System Protection
Remote
You must be logged on as an Administrator to make most of these changes.
Performance
Visual effects, processor scheduling, memory usage, and virtual memory
Settings...
User Profiles
Desktop settings related to your logon
Settings...
Startup and Recovery
System startup, system failure, and debugging information
Settings...
Environment Variables...
OK
Cancel
Apply


System Properties
Environment Variables
Edit System Variable
Variable name:
Path
Variable value:
n;%M2_HOME
1
E:\installed\scala\bin;
OK
Cancel
System variables
Variable
Value
OS
Windows_NT
Path
C:\Program Files (x86)\NVIDIA Corpora...
PATHEXT
.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;....
PROCESSOR_A...
AMD64
New...
Edit...
Delete
OK
Cancel


C:\Windows\system32\cmd.exe - scala
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.
C:\Users\rezkar>scala
Welcome to Scala 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_121).
Type in expressions for evaluation. Or try :help.
scala>


pitbul — java — 61×13
Alexs-MacBook-Pro:~ pitbul$ source .profile
Alexs-MacBook-Pro:~ pitbul$ scala
Welcome to Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Serv
er VM, Java 1.8.0_121).
Type in expressions to have them evaluated.
Type :help for more information.
scala> println("Hello world")
Hello world
scala>

```
asif@ubuntu:~$ cd /usr/local/share/
asif@ubuntu:/usr/local/share$ ls
ca-certificates  fonts  man  scala-2.11.8  sgml  texmf  xml
asif@ubuntu:/usr/local/share$ cd ~
asif@ubuntu:~$ echo "export SCALA_HOME=/usr/local/share/scala-2.11.8" >> ~/.bash
rc
asif@ubuntu:~$ echo "export PATH=$PATH:$SCALA_HOME/bin" >> ~/.bashrc
asif@ubuntu:~$ source  ~/.bashrc
asif@ubuntu:~$ scala
Welcome to Scala version 2.9.2 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_12
1).
Type in expressions to have them evaluated.
Type :help for more information.
```

```
C:\Windows\system32\cmd.exe - scala

C:\Users\rezkar>scala
Welcome to Scala 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_121).
Type in expressions for evaluation. Or try :help.

scala> var a = 10
a: Int = 10

scala> var b = 20.5
b: Double = 20.5

scala> var c = "String"
c: String = String

scala>

scala> val d = 5000
d: Int = 5000

scala> val f = 'cat'
<console>:1: error: unclosed character literal
val f = 'cat'
        ^

scala> val f = "cat"
f: String = cat

scala>
```

```
scala> val i:Int = "hello"
<console>:11: error: type mismatch;
 found   : String("hello")
 required: Int
       val i:Int = "hello"
                   ^

scala> val x = "hello"
x: String = hello

scala> x.re
reduce              reduceRight         replaceAll              reverse
reduceLeft          reduceRightOption   replaceAllLiterally     reverseIterator
reduceLeftOption    regionMatches       replaceFirst            reverseMap
reduceOption        replace             repr

scala> val x = new AnyRef{def helloWord = "Hello, world!"}
x: AnyRef{def helloWord: String} = $anon$1@58065f0c

scala> x.helloWord
   def helloWord: String

scala> x.helloWord
warning: there was one feature warning; re-run with -feature for details
res0: String = Hello, world!

scala> _
```

```
scala> trait Logging  { override def toString = "Logging " }
defined trait Logging

scala> class A extends Logging  { override def toString = "A->" + super.toString
 }
defined class A

scala> trait B extends Logging  { override def toString = "B->" + super.toString
 }
defined trait B

scala> trait C extends Logging  { override def toString = "C->" + super.toString
 }
defined trait C

scala> class D extends A with B with C { override def toString = "D->" + super.t
oString }
defined class D

scala> new D()
res4: D = D->C->B->A->Logging

scala> _
```

HelloWorld.scala

```
object HelloWorld{
        def main(args:Array[String]){
        println("Hello, world!")
    }
}
```

# Chapter 2: Object-Oriented Scala

HamOrSpamDemo.scala | DeepLearningwithH2O/po... | AirlinesWithWeatherDem...

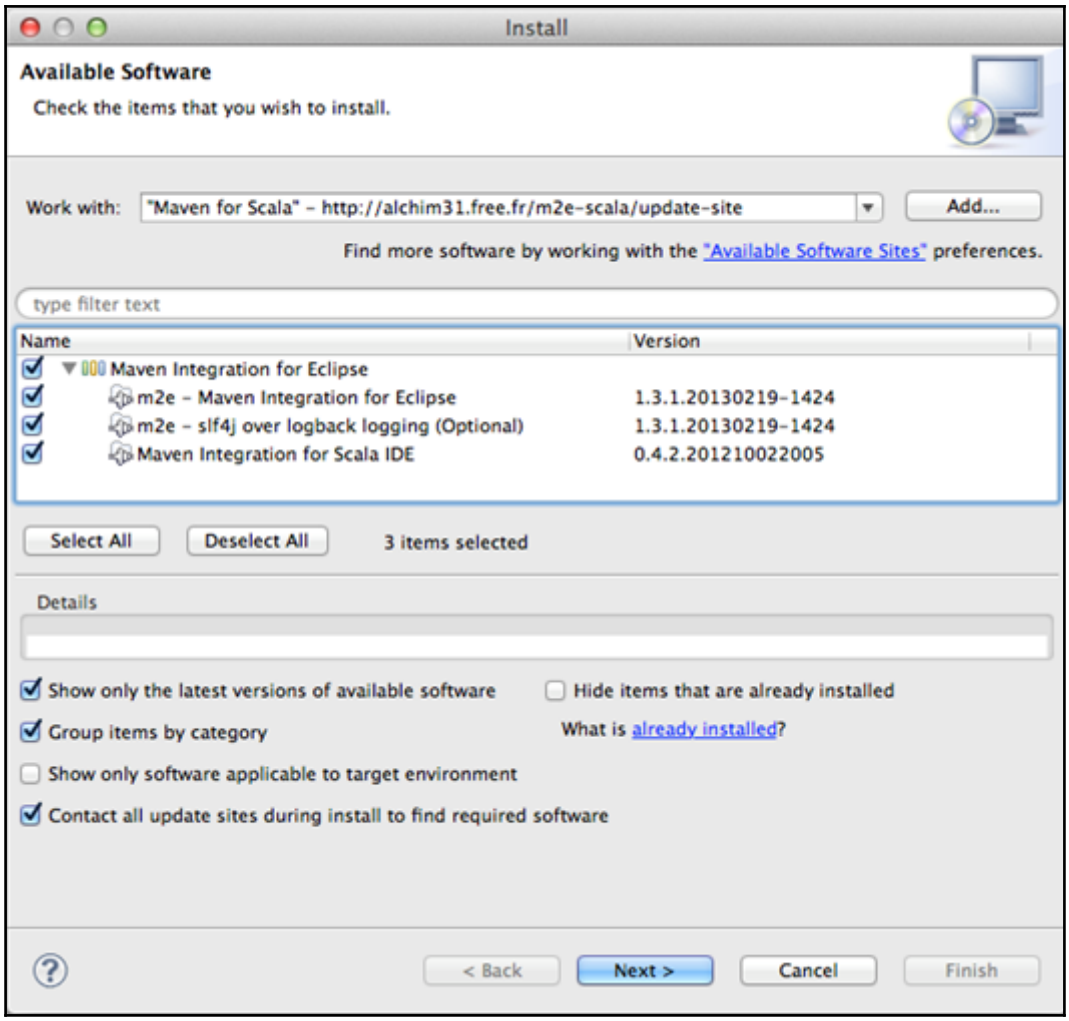
```
1  package com.chapter3.OOP
2
3  object VariablesDemo {
4     def main(args: Array[String]) {
5        var myVar : Int = 50;
6        val myVal : String = "Hello World! I've started learning Scala.";
7
8        myVar = 90;
9
10
11
12
13
14    }
15 }
```

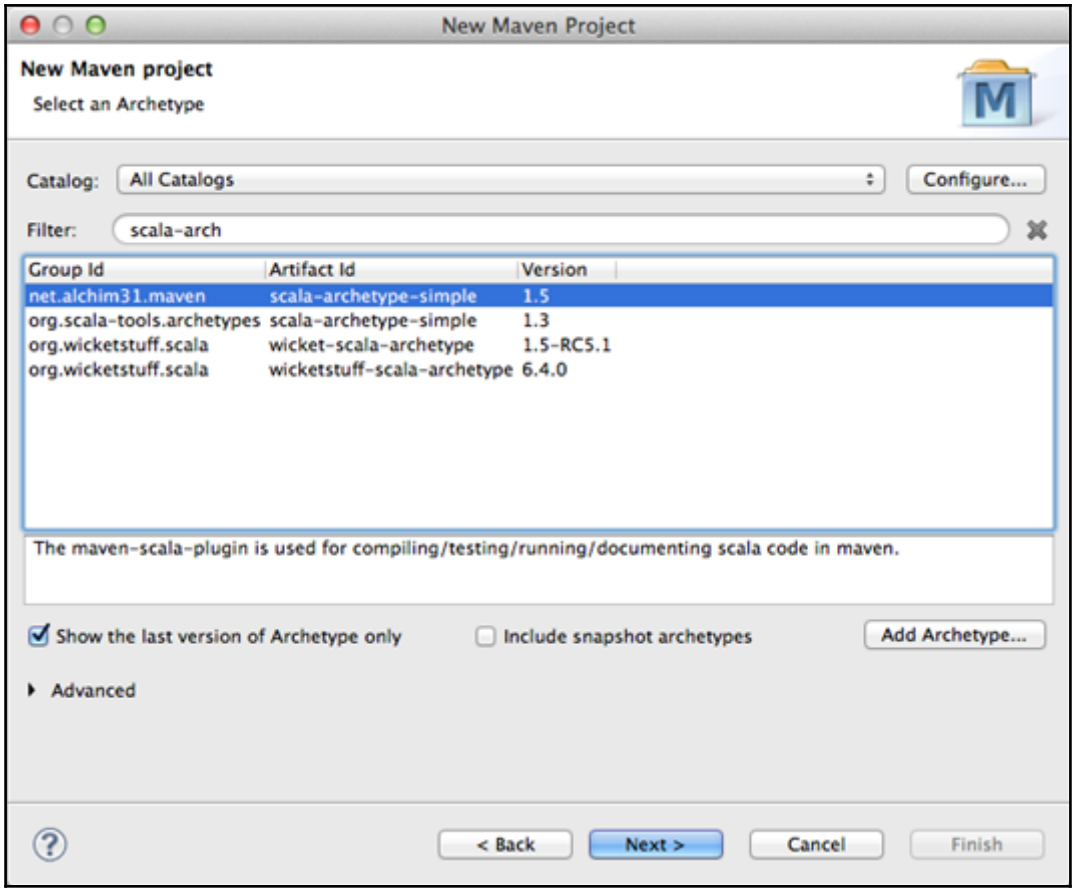
Multiple markers at this line:

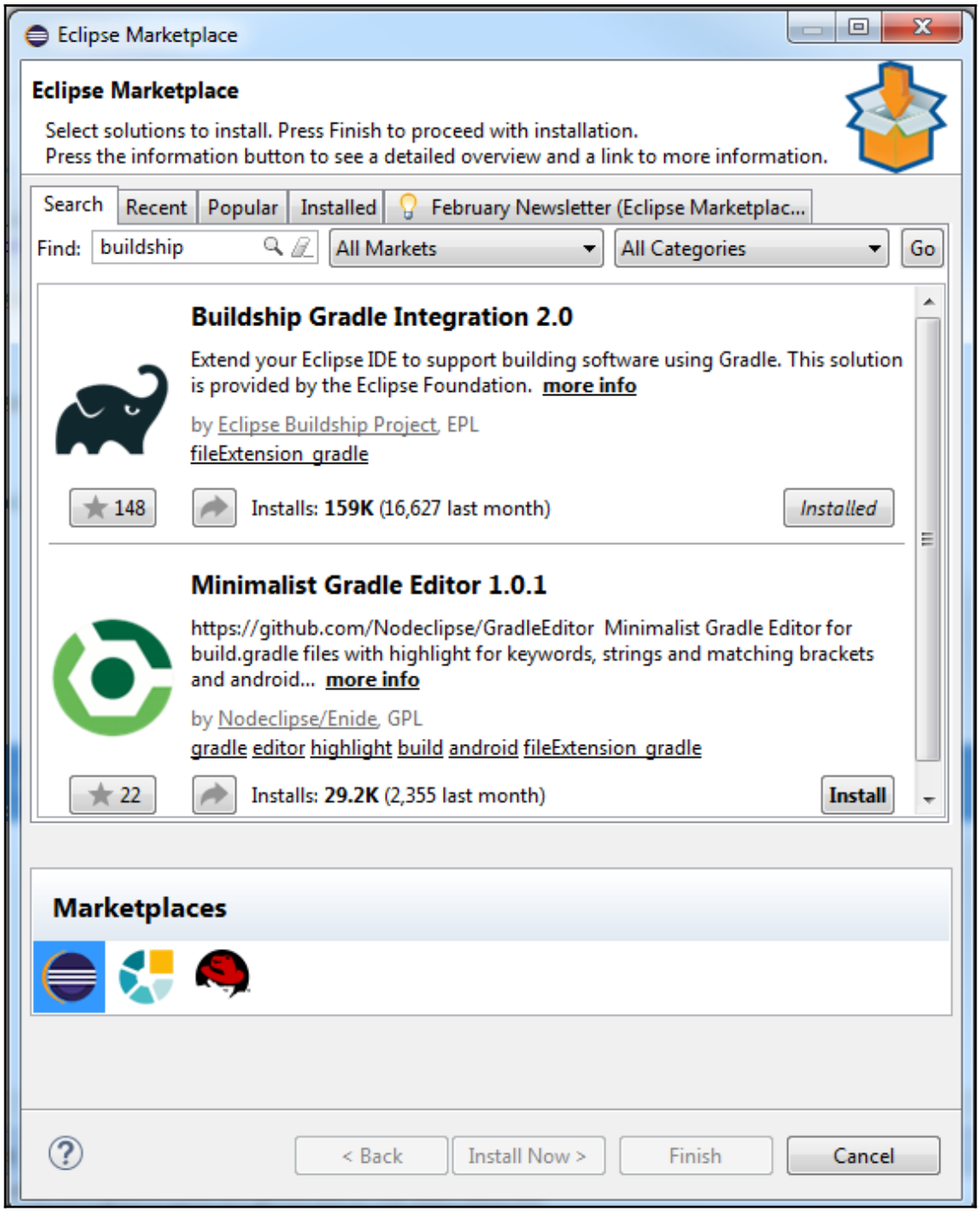
- reassignment to val
- reassignment to val

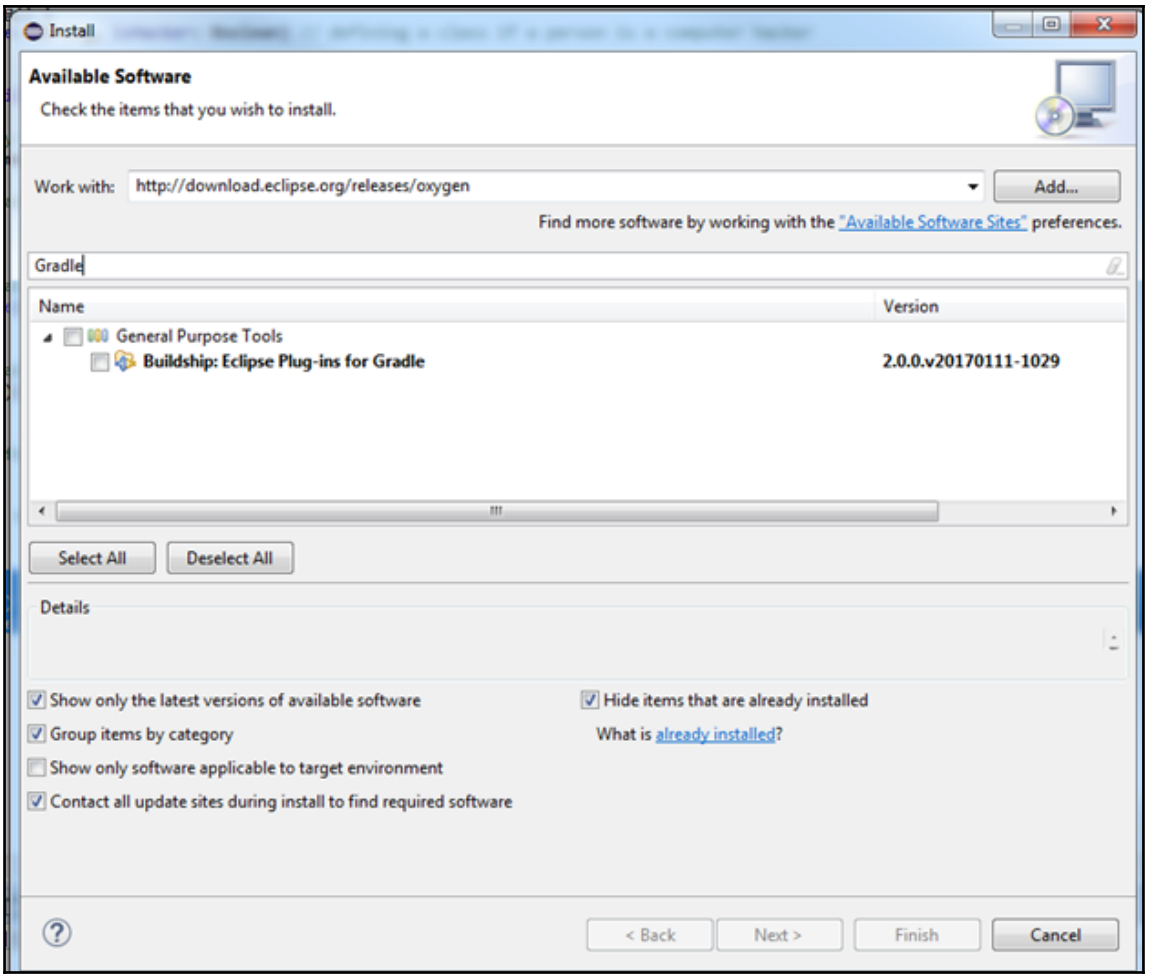
```
C:\Windows\system32\cmd.exe - scala

C:\Users\rezkar>scala
Welcome to Scala 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_121).
Type in expressions for evaluation. Or try :help.

scala> case class Character(name: String, isHacker: Boolean)
defined class Character

scala> val nail = Character("Nail", true)
nail: Character = Character(Nail,true)

scala>  val joyce = nail.copy(name = "Joyce")
joyce: Character = Character(Joyce,true)

scala>  println(nail == joyce)
false

scala>  println(nail.equals(joyce))
false

scala> println(nail.equals(nail))
true

scala> println(nail.hashCode())
-112671915

scala> println(nail.toString())
Character(Nail,true)

scala>     joyce match {
     |       case Character(x, true) => s"$x is a hacker"
     |       case Character(x, false) => s"$x is not a hacker"
     |     }
res5: String = Joyce is a hacker

scala> _
```

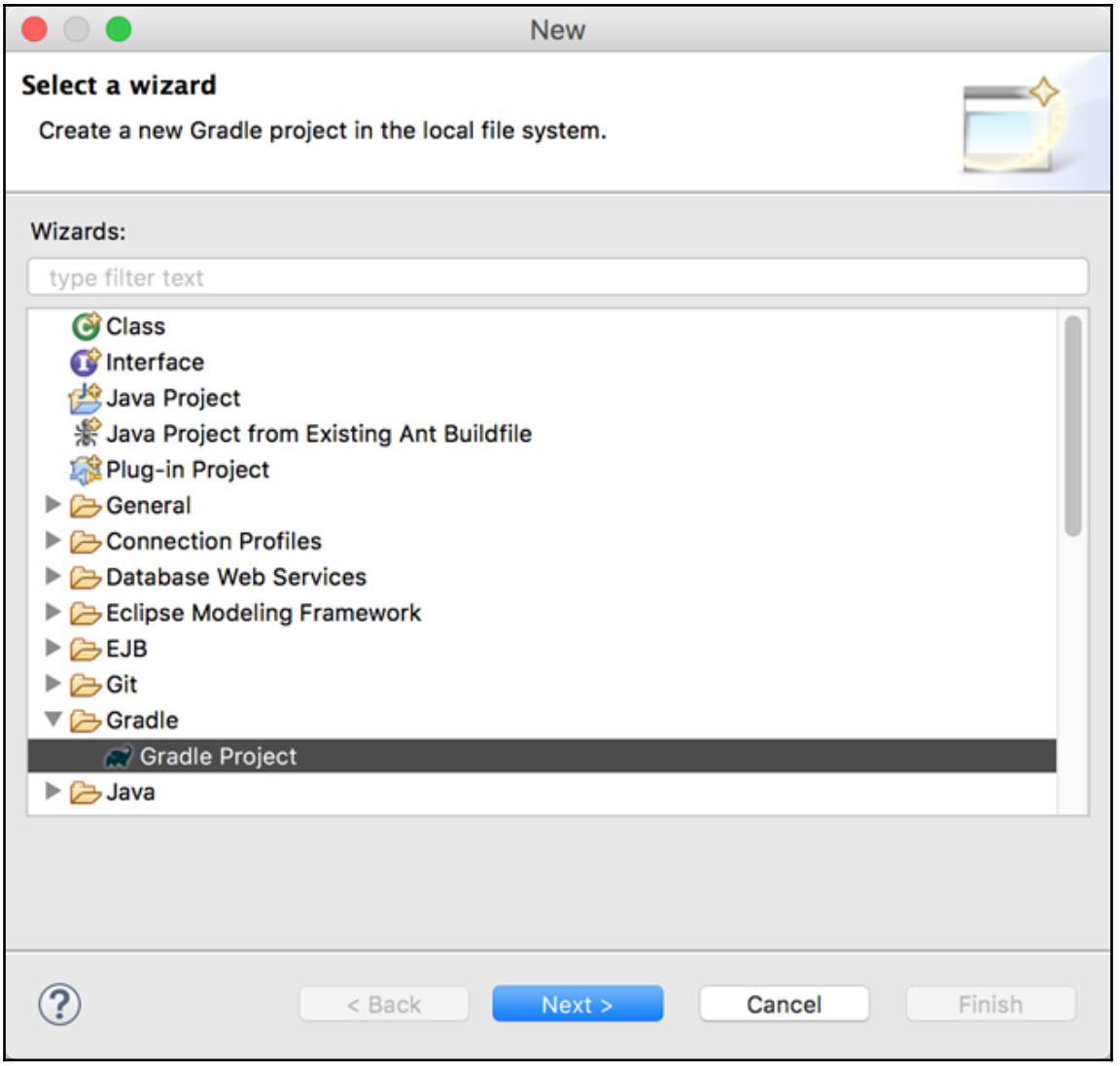
```
scala> object test { def printSomething() = {println("Inside an object")} }
defined object test

scala> test.printSomething
Inside an object

scala> val x = new test()
<console>:11: error: not found: type test
       val x = new test()
                   ^
```

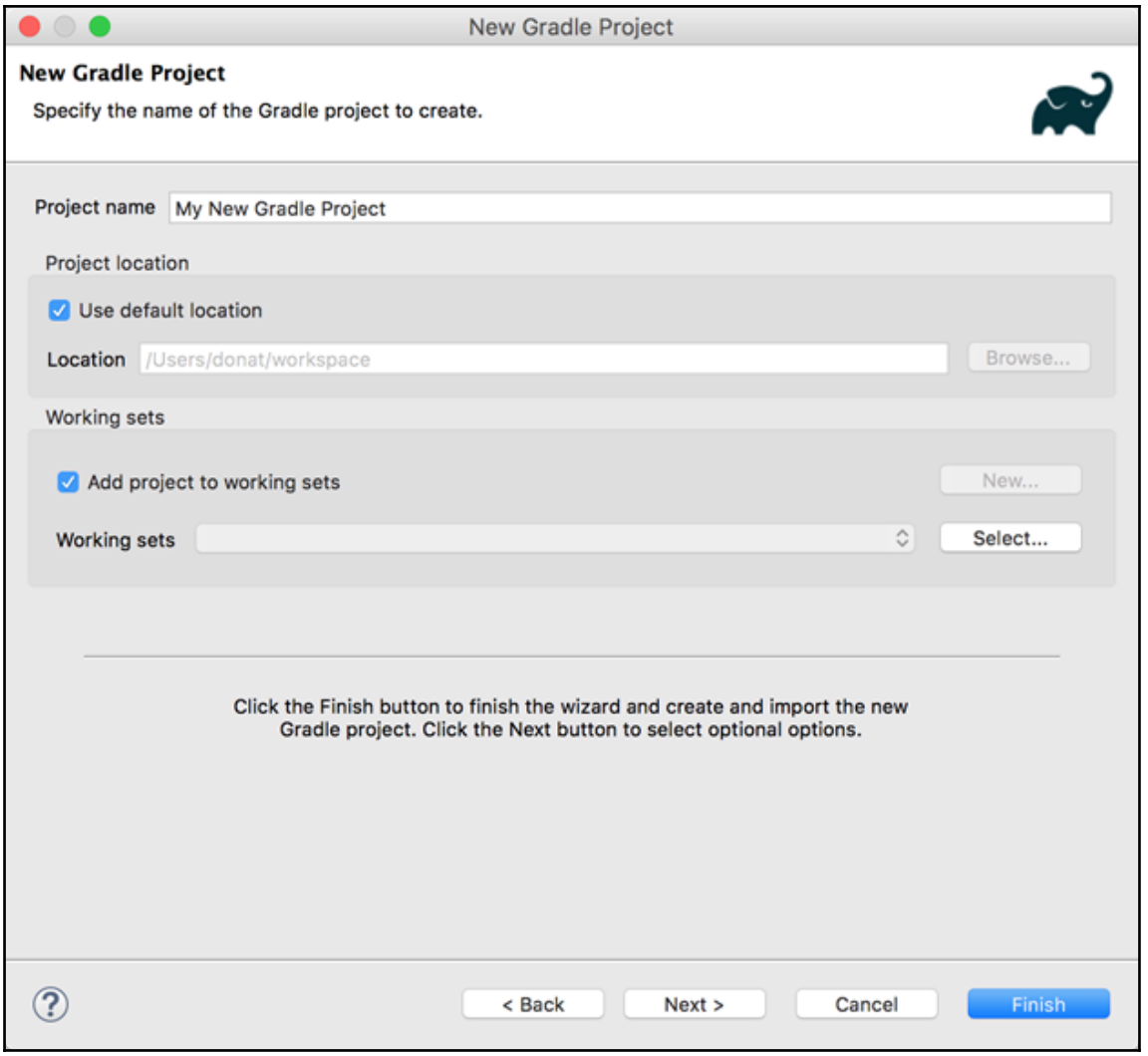

Install
Available Software
Check the items that you wish to install.
Work with:
"Maven for Scala" - http://alchim31.free.fr/m2e-scala/update-site
Add...
Find more software by working with the "Available Software Sites" preferences.
type filter text
Name
Version
Maven Integration for Eclipse
m2e - Maven Integration for Eclipse
1.3.1.20130219-1424
m2e - slf4j over logback logging (Optional)
1.3.1.20130219-1424
Maven Integration for Scala IDE
0.4.2.201210022005
Select All
Deselect All
3 items selected
Details
Show only the latest versions of available software
Hide items that are already installed
Group items by category
What is already installed?
Show only software applicable to target environment
Contact all update sites during install to find required software
< Back
Next >
Cancel
Finish

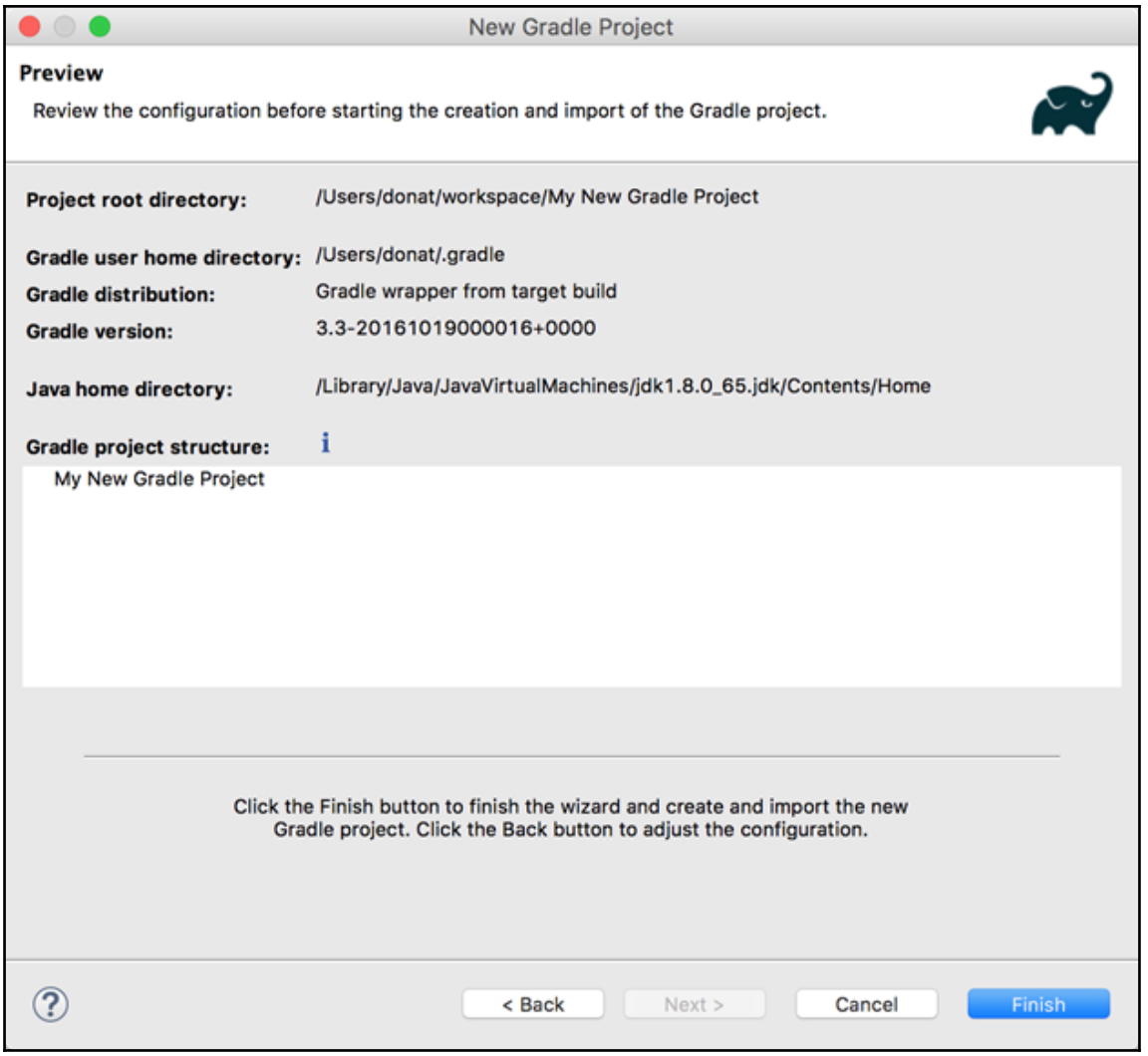

New Maven Project
New Maven project
Select an Archetype
Catalog:
All Catalogs
Configure...
Filter:
scala-arch
Group Id
Artifact Id
Version
net.alchim31.maven
scala-archetype-simple
1.5
org.scala-tools.archetypes
scala-archetype-simple
1.3
org.wicketstuff.scala
wicket-scala-archetype
1.5-RC5.1
org.wicketstuff.scala
wicketstuff-scala-archetype
6.4.0
The maven-scala-plugin is used for compiling/testing/running/documenting scala code in maven.
Show the last version of Archetype only
Include snapshot archetypes
Add Archetype...
Advanced
< Back
Next >
Cancel
Finish

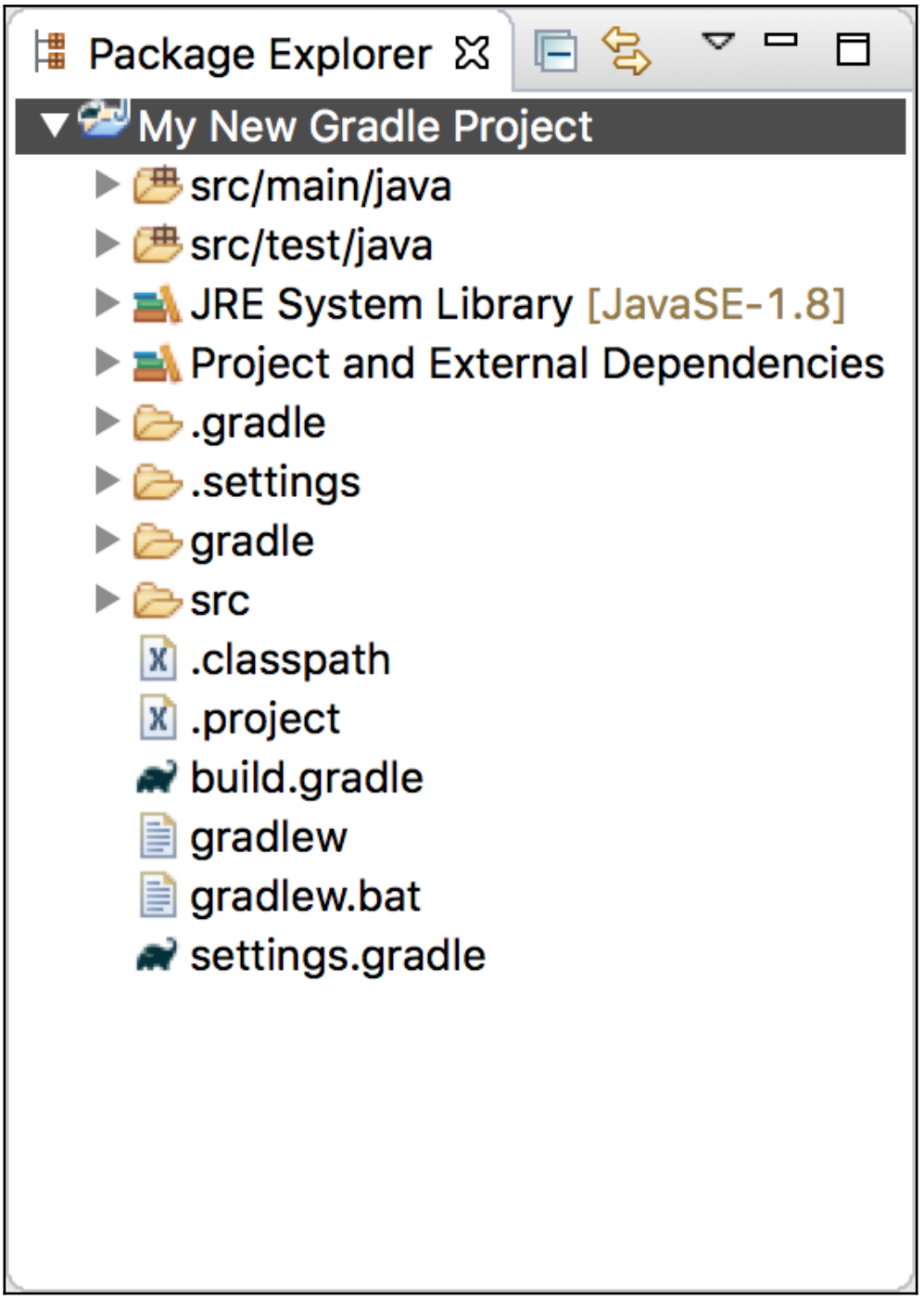

Eclipse Marketplace
Eclipse Marketplace
Select solutions to install. Press Finish to proceed with installation.
Press the information button to see a detailed overview and a link to more information.
Search
Recent
Popular
Installed
February Newsletter (Eclipse Marketplac...
Find:
buildship
All Markets
All Categories
Go
Buildship Gradle Integration 2.0
Extend your Eclipse IDE to support building software using Gradle. This solution is provided by the Eclipse Foundation. more info
by Eclipse Buildship Project, EPL
fileExtension_gradle
148
Installs: 159K (16,627 last month)
Installed
Minimalist Gradle Editor 1.0.1
https://github.com/Nodeclipse/GradleEditor Minimalist Gradle Editor for build.gradle files with highlight for keywords, strings and matching brackets and android... more info
by Nodeclipse/Enide, GPL
gradle editor highlight build android fileExtension_gradle
22
Installs: 29.2K (2,355 last month)
Install
Marketplaces
< Back
Install Now >
Finish
Cancel


Install
Available Software
Check the items that you wish to install.
Work with: http://download.eclipse.org/releases/oxygen
Add...
Find more software by working with the "Available Software Sites" preferences.
Gradle
Name
Version
General Purpose Tools
Buildship: Eclipse Plug-ins for Gradle
2.0.0.v20170111-1029
Select All
Deselect All
Details
Show only the latest versions of available software
Group items by category
Show only software applicable to target environment
Contact all update sites during install to find required software
Hide items that are already installed
What is already installed?
< Back
Next >
Finish
Cancel

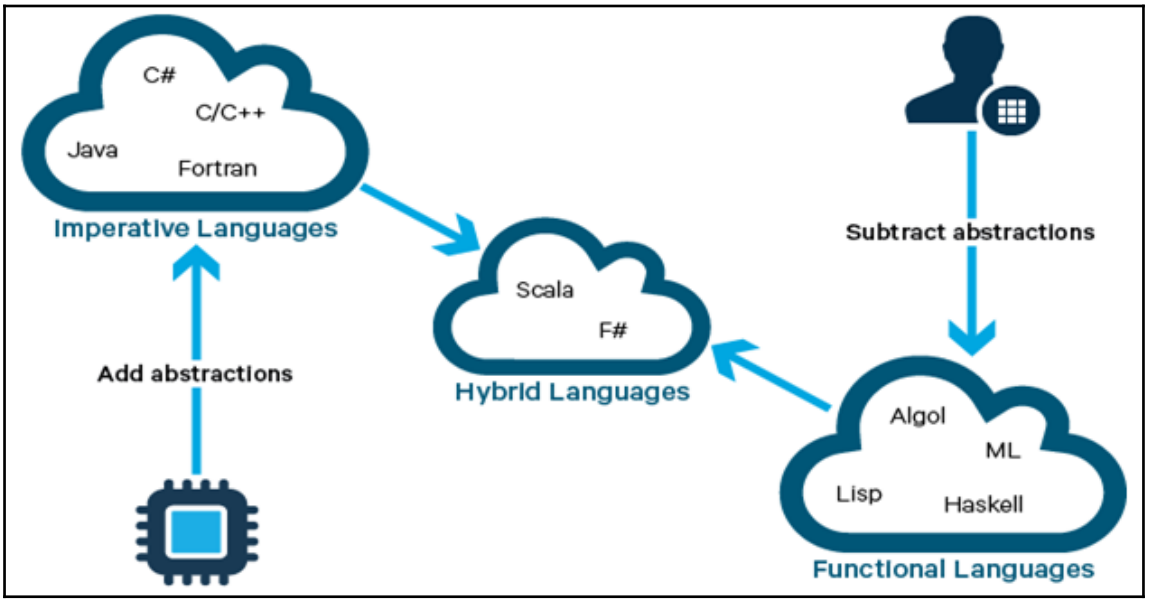

New
Select a wizard
Create a new Gradle project in the local file system.
Wizards:
type filter text
Class
Interface
Java Project
Java Project from Existing Ant Buildfile
Plug-in Project
General
Connection Profiles
Database Web Services
Eclipse Modeling Framework
EJB
Git
Gradle
Gradle Project
Java
< Back
Next >
Cancel
Finish


New Gradle Project
New Gradle Project
Specify the name of the Gradle project to create.
Project name
My New Gradle Project
Project location
Use default location
Location
/Users/donat/workspace
Browse...
Working sets
Add project to working sets
New...
Working sets
Select...
Click the Finish button to finish the wizard and create and import the new Gradle project. Click the Next button to select optional options.
< Back
Next >
Cancel
Finish


New Gradle Project
Preview
Review the configuration before starting the creation and import of the Gradle project.
Project root directory:
/Users/donat/workspace/My New Gradle Project
Gradle user home directory:
/Users/donat/.gradle
Gradle distribution:
Gradle wrapper from target build
Gradle version:
3.3-20161019000016+0000
Java home directory:
/Library/Java/JavaVirtualMachines/jdk1.8.0_65.jdk/Contents/Home
Gradle project structure:
My New Gradle Project
Click the Finish button to finish the wizard and create and import the new Gradle project. Click the Back button to adjust the configuration.
< Back
Next >
Cancel
Finish


Package Explorer
My New Gradle Project
src/main/java
src/test/java
JRE System Library [JavaSE-1.8]
Project and External Dependencies
.gradle
.settings
gradle
src
.classpath
.project
build.gradle
gradlew
gradlew.bat
settings.gradle

# Chapter 3: Functional programming concepts




```
scala> def quarterMaker(value: Int): Double = value.toDouble/4
quarterMaker: (value: Int)Double

scala> def addTwo(value: Int): Int = value + 2
addTwo: (value: Int)Int

scala> def applyFuncOnRange(begin: Int, end: Int, func: Int => AnyVal): Unit = {
     |     for (i <- begin to end)
     |         println(func(i))
     | }
applyFuncOnRange: (begin: Int, end: Int, func: Int => AnyVal)Unit

scala>

scala>

scala>

scala>

scala>

scala>
```

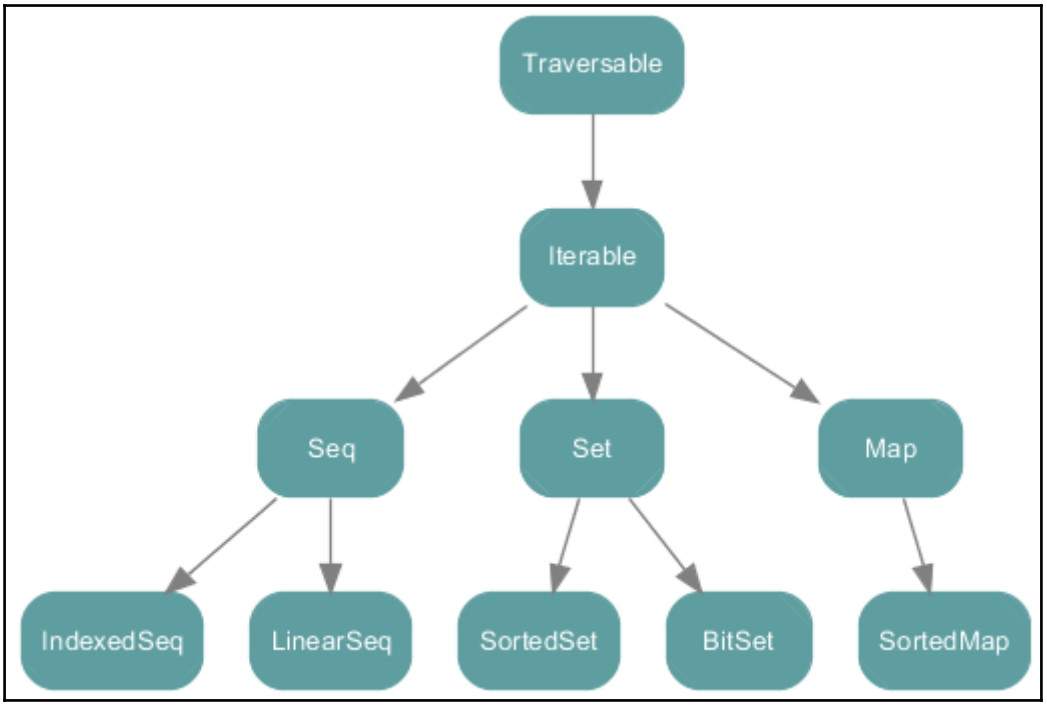
```
scala> applyFuncOnRange(1,10,quarterMaker)
0.25
0.5
0.75
1.0
1.25
1.5
1.75
2.0
2.25
2.5

scala>

scala>

scala>

scala>

scala>

scala>
```

```
scala> applyFuncOnRange(1,10,addTwo)
3
4
5
6
7
8
9
10
11
12

scala>

scala>

scala>

scala>

scala>

scala>
```

```
scala> def bankFee(amount: Double) = amount * 0.05
bankFee: (amount: Double)Double

scala> def TransferMoney(money: Double, bankFee: Double => Double): Double = {
     |     money + bankFee(money)
     | }
TransferMoney: (money: Double, bankFee: Double => Double)Double

scala> TransferMoney(100, bankFee)
res2: Double = 105.0

scala>

scala>

scala>

scala>

scala>

scala>

scala>
```

```
scala> def TransferMoney(money: Double, bankFee: Double => Double): Double = {
     |     money + bankFee(money)
     | }
TransferMoney: (money: Double, bankFee: Double => Double)Double

scala> TransferMoney(100, (amount: Double) => amount * 0.05)
res12: Double = 105.0

scala> TransferMoney(100, amount => amount * 0.05)
res13: Double = 105.0

scala>

scala>

scala>

scala>

scala>

scala>

scala>
```

```
scala> def TransferMoney(money: Double) = {
     |   if (money > 1000)
     |     (money: Double) => "Dear customer we are going to add the following amount as Fee
: "+money * 0.05
     |   else
     |     (money: Double) => "Dear customer we are going to add the following amount as Fee
: "+money * 0.1
     | }
TransferMoney: (money: Double)Double => String

scala> val returnedFunction = TransferMoney(1500)
returnedFunction: Double => String = <function1>

scala>
     | returnedFunction(1500)
res17: String = Dear customer we are going to add the following amount as Fee: 75.0

scala>

scala>

scala>

scala>
```



```
scala> val evenList = List(2,4,6,8,10)
evenList: List[Int] = List(2, 4, 6, 8, 10)

scala> evenList.map(x => x * 2 )
res18: List[Int] = List(4, 8, 12, 16, 20)

scala>

scala>

scala>

scala>

scala>

scala>

scala>

scala>

scala>
```

```
scala> def func(x : Int) = if(x > 4) Some(x) else None
func: (x: Int)Option[Int]

scala> evenList.map(x => func(x))
res19: List[Option[Int]] = List(None, None, Some(6), Some(8), Some(10))

scala>

scala>

scala>

scala>

scala>

scala>

scala>

scala>

scala>
```

```
scala> def around(x : Int) = List(x-1, x, x+1)
around: (x: Int)List[Int]

scala> evenList.map(x => around(x))
res23: List[List[Int]] = List(List(1, 2, 3), List(3, 4, 5), List(5, 6, 7), List(7, 8, 9), Li
st(9, 10, 11))

scala> evenList.flatMap( x => around(x))
res24: List[Int] = List(1, 2, 3, 3, 4, 5, 5, 6, 7, 7, 8, 9, 9, 10, 11)

scala>

scala>

scala>

scala>

scala>

scala>

scala>
```

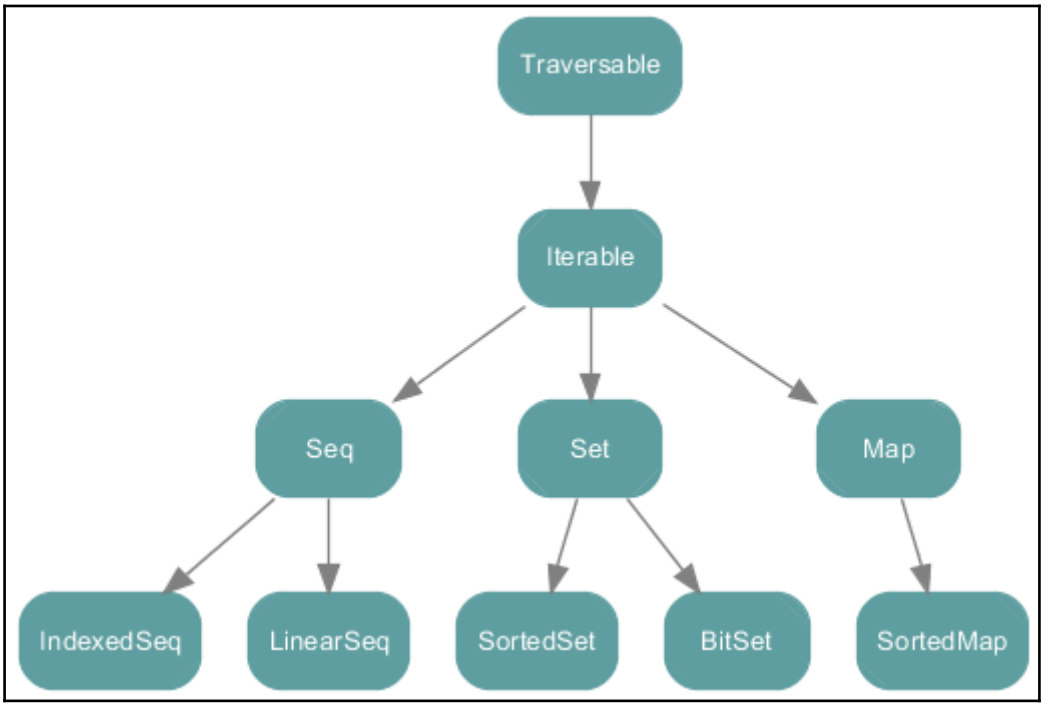
```
scala> val range = List.range(1,10)
range: List[Int] = List(1, 2, 3, 4, 5, 6, 7, 8, 9)

scala> val odds = range.filter(_ % 2 != 0)
odds: List[Int] = List(1, 3, 5, 7, 9)

scala>

scala>

scala>

scala>

scala>

scala>

scala>

scala>

scala>
```

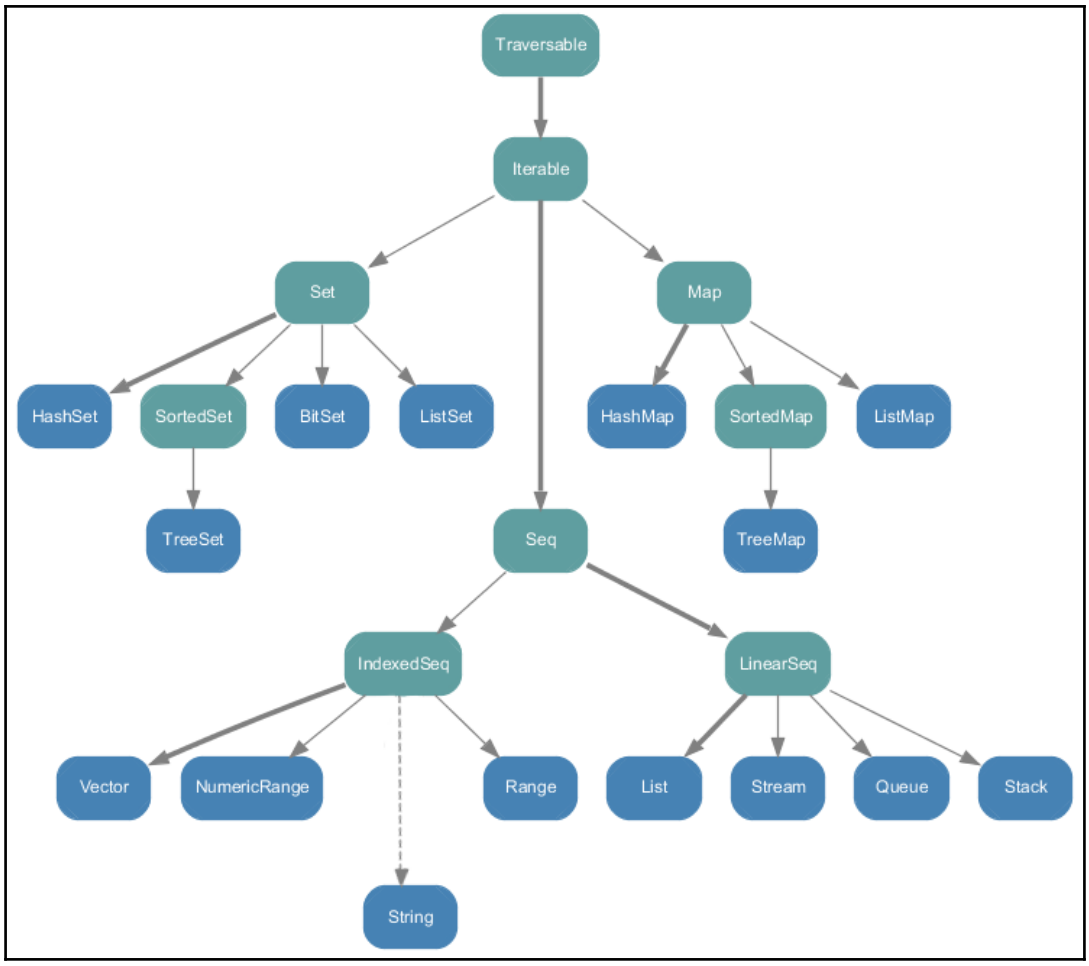
```
scala> for (x <- 10 until (0, -2))
     |     yield x
res25: scala.collection.immutable.IndexedSeq[Int] = Vector(10, 8, 6, 4, 2)

scala>

scala>

scala>

scala>

scala>

scala>

scala>

scala>

scala>

scala>
```

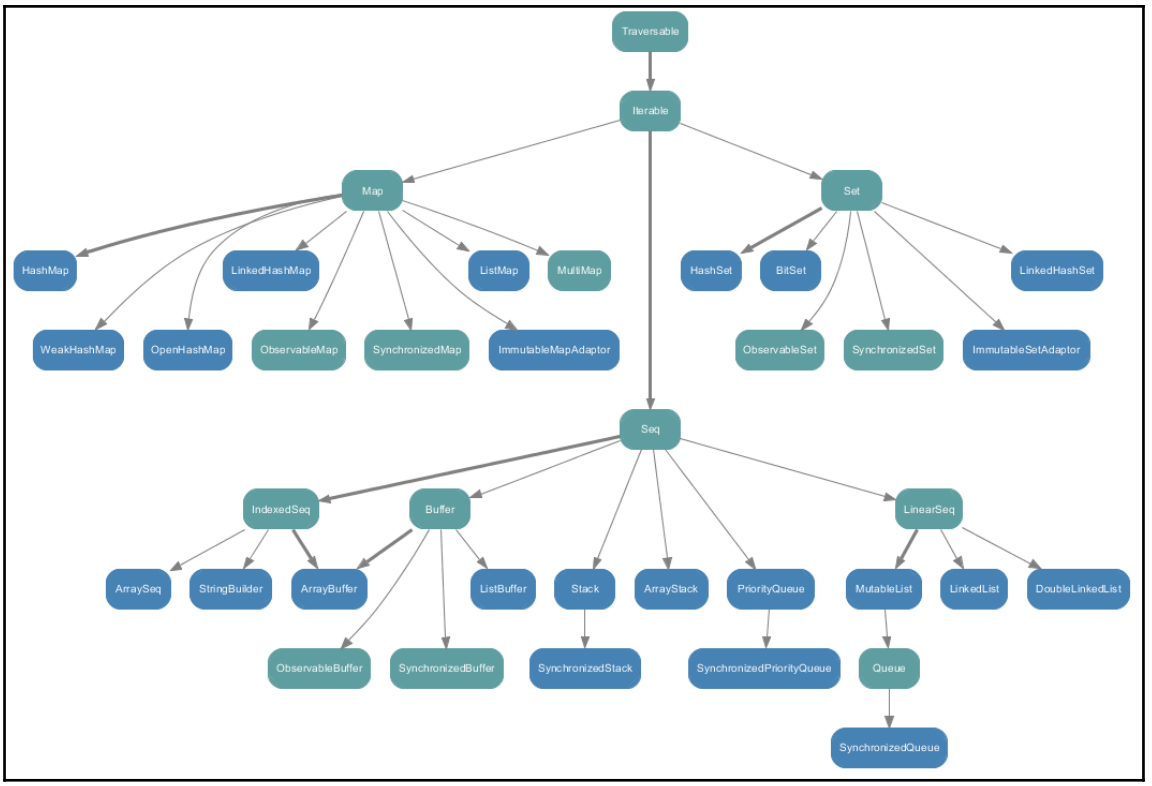
```
scala> for (x <- 1 to 10 if x % 2 == 0)
     |     yield x
res26: scala.collection.immutable.IndexedSeq[Int] = Vector(2, 4, 6, 8, 10)

scala>

scala>

scala>

scala>

scala>

scala>

scala>

scala>

scala>

scala>
```

```
scala> for (x <- 1 to 10; y <- 1 until x)
     |     yield (x, y)
res29: scala.collection.immutable.IndexedSeq[(Int, Int)] = Vector((2,1), (3,1), (3,2), (4,1)
, (4,2), (4,3), (5,1), (5,2), (5,3), (5,4), (6,1), (6,2), (6,3), (6,4), (6,5), (7,1), (7,2),
 (7,3), (7,4), (7,5), (7,6), (8,1), (8,2), (8,3), (8,4), (8,5), (8,6), (8,7), (9,1), (9,2),
(9,3), (9,4), (9,5), (9,6), (9,7), (9,8), (10,1), (10,2), (10,3), (10,4), (10,5), (10,6), (1
0,7), (10,8), (10,9))

scala>

scala> (1 to 10).flatMap(
     |     i => (1 until i).map(
     |         j => (i, j)
     |     )
     | )
res30: scala.collection.immutable.IndexedSeq[(Int, Int)] = Vector((2,1), (3,1), (3,2), (4,1)
, (4,2), (4,3), (5,1), (5,2), (5,3), (5,4), (6,1), (6,2), (6,3), (6,4), (6,5), (7,1), (7,2),
 (7,3), (7,4), (7,5), (7,6), (8,1), (8,2), (8,3), (8,4), (8,5), (8,6), (8,7), (9,1), (9,2),
(9,3), (9,4), (9,5), (9,6), (9,7), (9,8), (10,1), (10,2), (10,3), (10,4), (10,5), (10,6), (1
0,7), (10,8), (10,9))

scala>

scala>
```

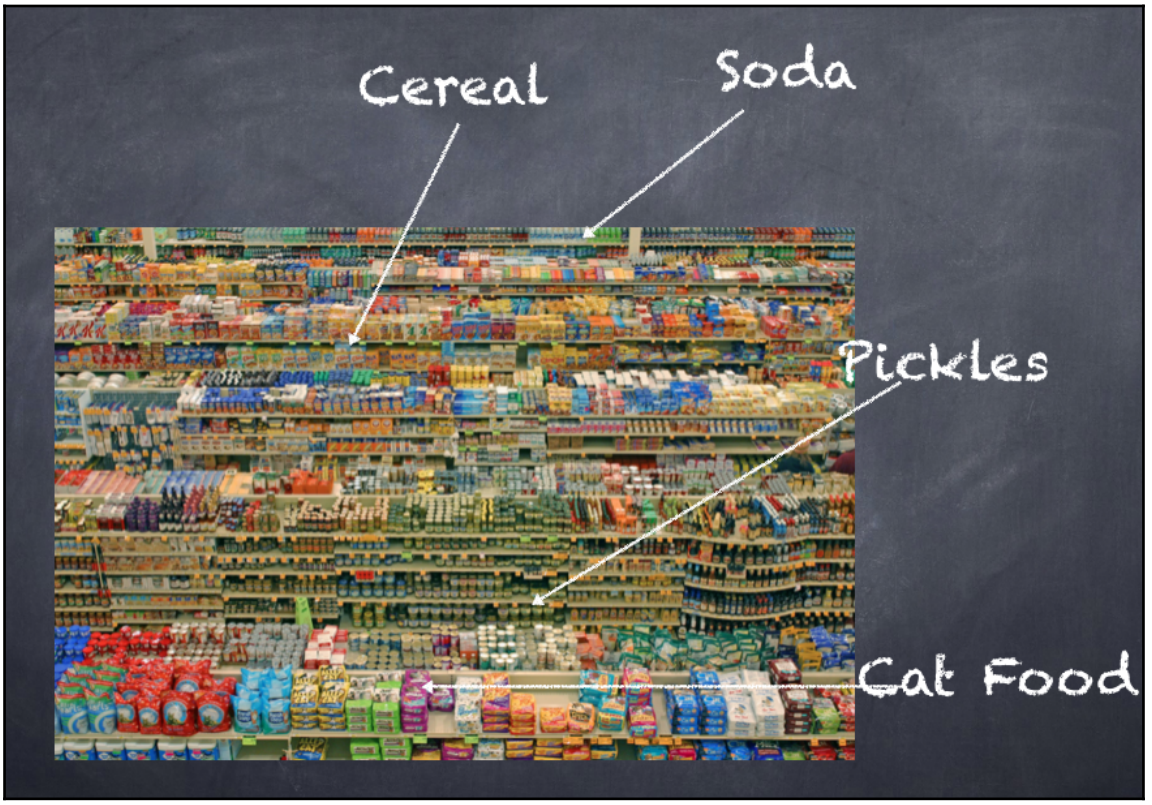
# Chapter 4: Collections APIs



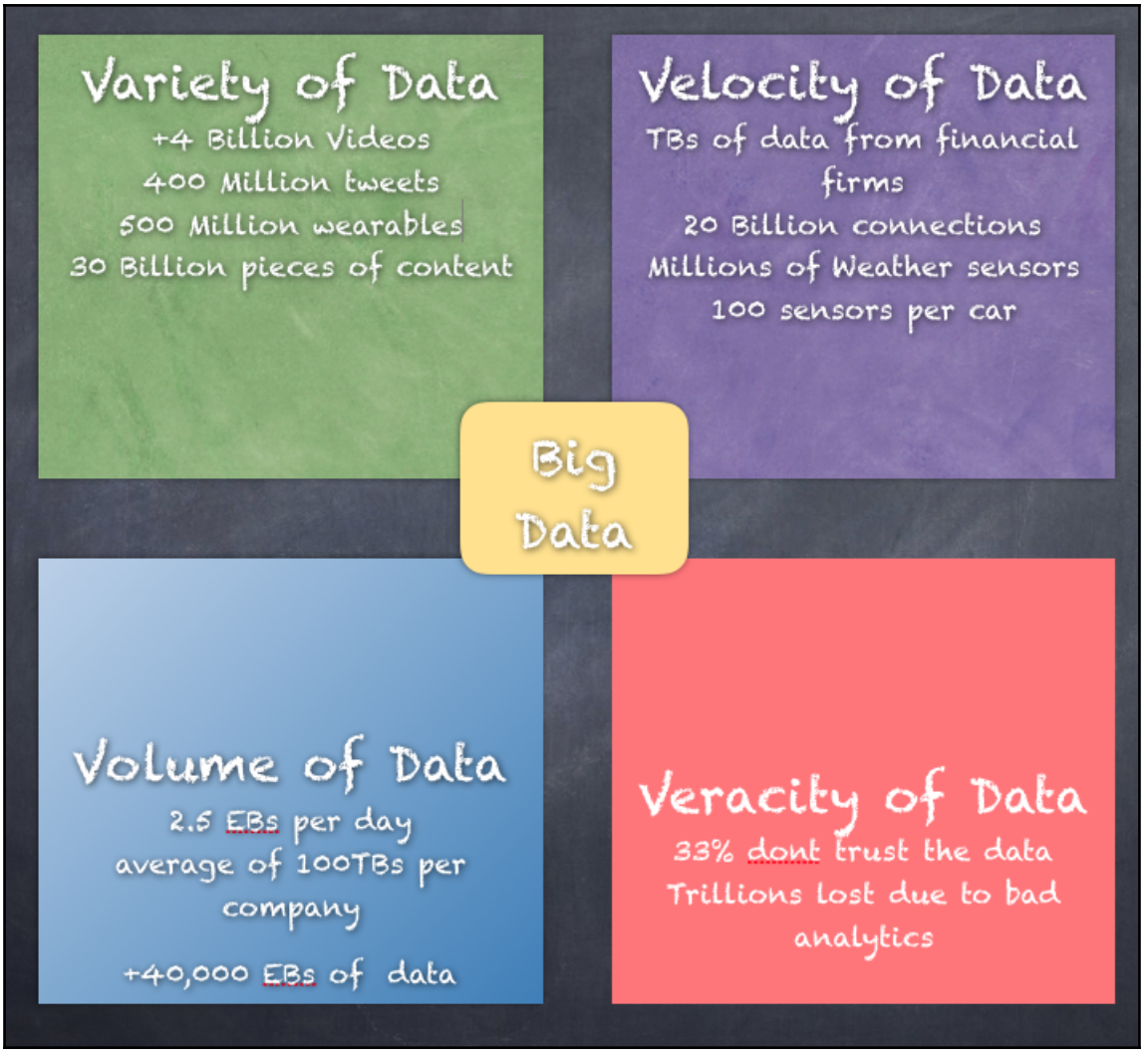

Traversable
Iterable
Set
Map
HashSet
SortedSet
BitSet
ListSet
HashMap
SortedMap
ListMap
TreeSet
Seq
TreeMap
IndexedSeq
LinearSeq
Vector
NumericRange
Range
List
Stream
Queue
Stack
String

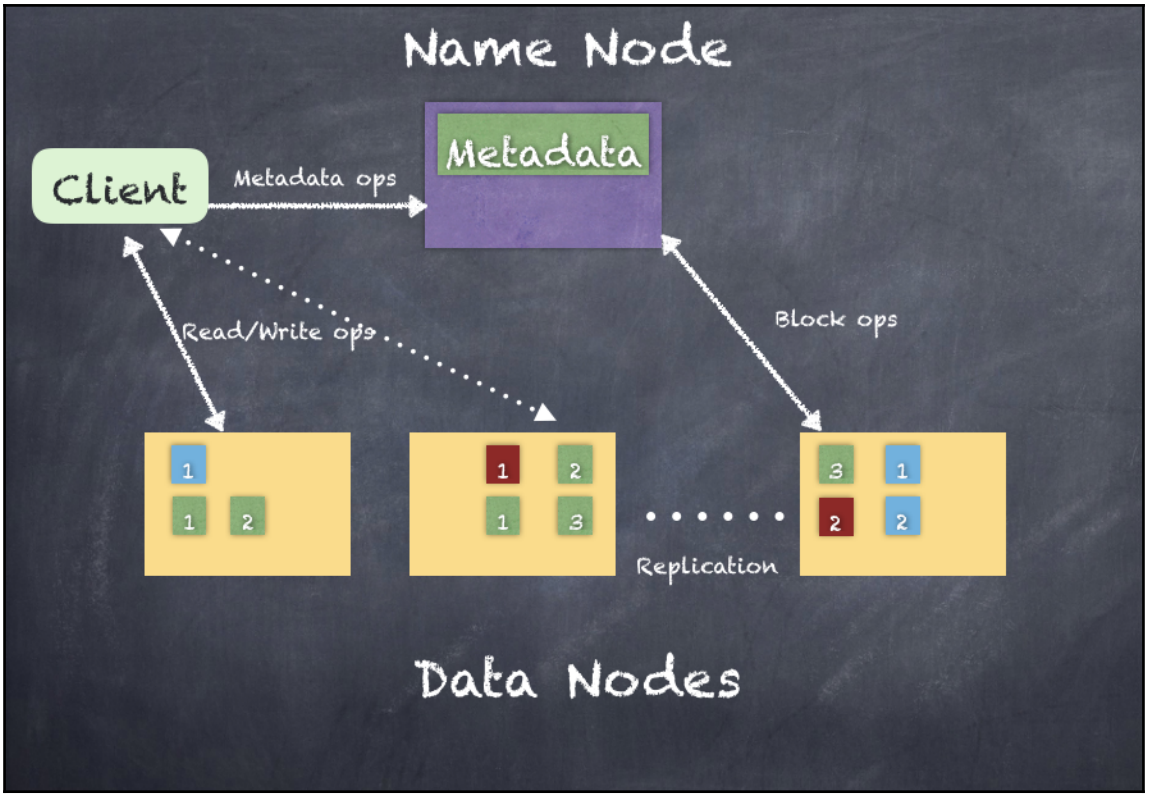

Traversable
Iterable
Map
Set
HashMap
LinkedHashMap
ListMap
MultiMap
HashSet
BitSet
LinkedHashSet
WeakHashMap
OpenHashMap
ObservableMap
SynchronizedMap
ImmutableMapAdaptor
ObservableSet
SynchronizedSet
ImmutableSetAdaptor
Seq
IndexedSeq
Buffer
LinearSeq
ArraySeq
StringBuilder
ArrayBuffer
ListBuffer
Stack
ArrayStack
PriorityQueue
MutableList
LinkedList
DoubleLinkedList
ObservableBuffer
SynchronizedBuffer
SynchronizedStack
SynchronizedPriorityQueue
Queue
SynchronizedQueue

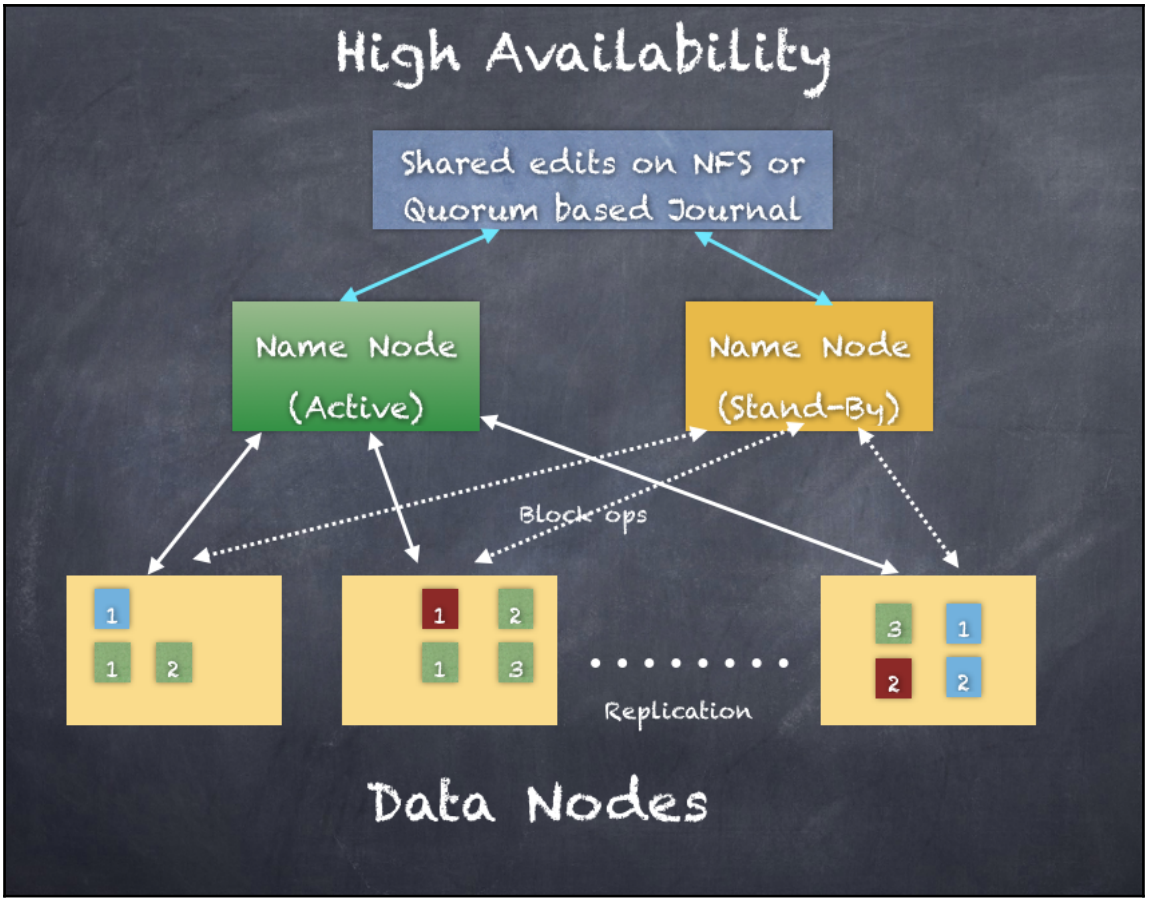
# Chapter 5: Tackle Big Data – Spark Comes to the Party

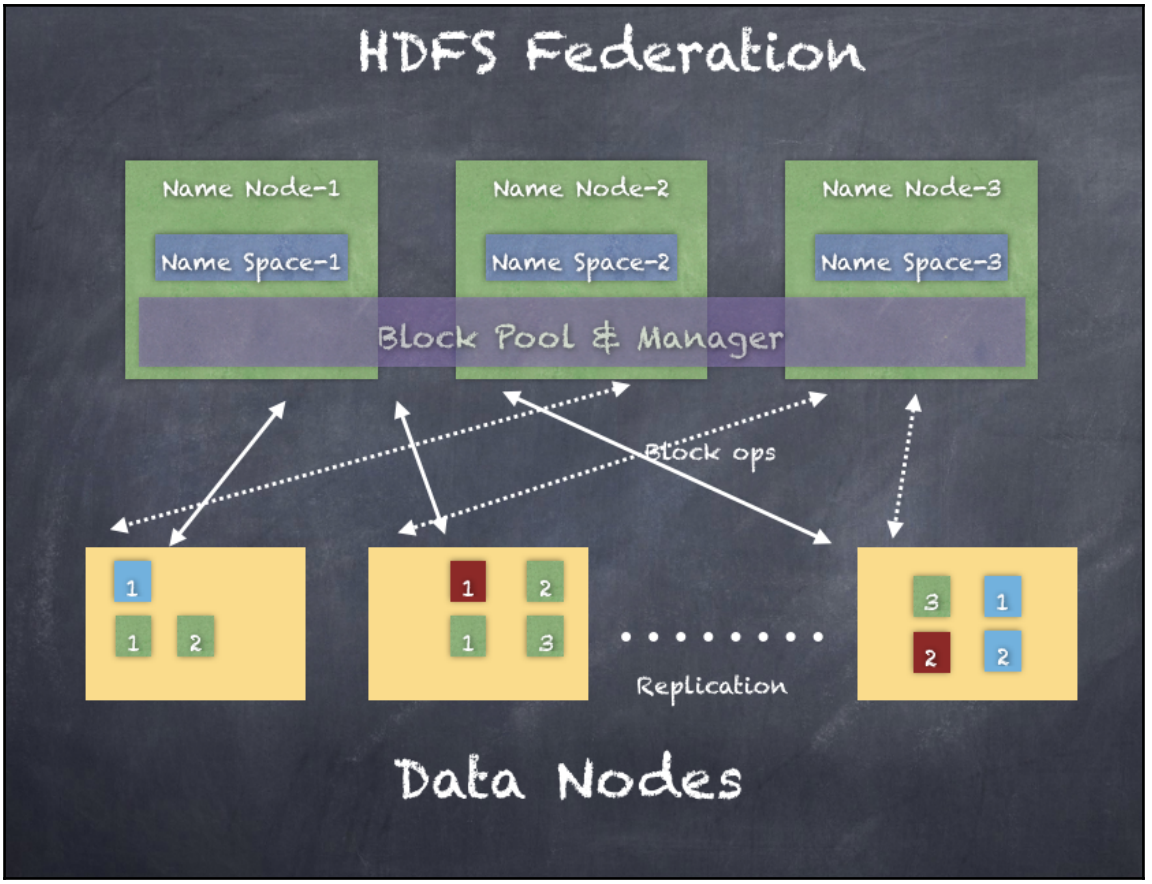


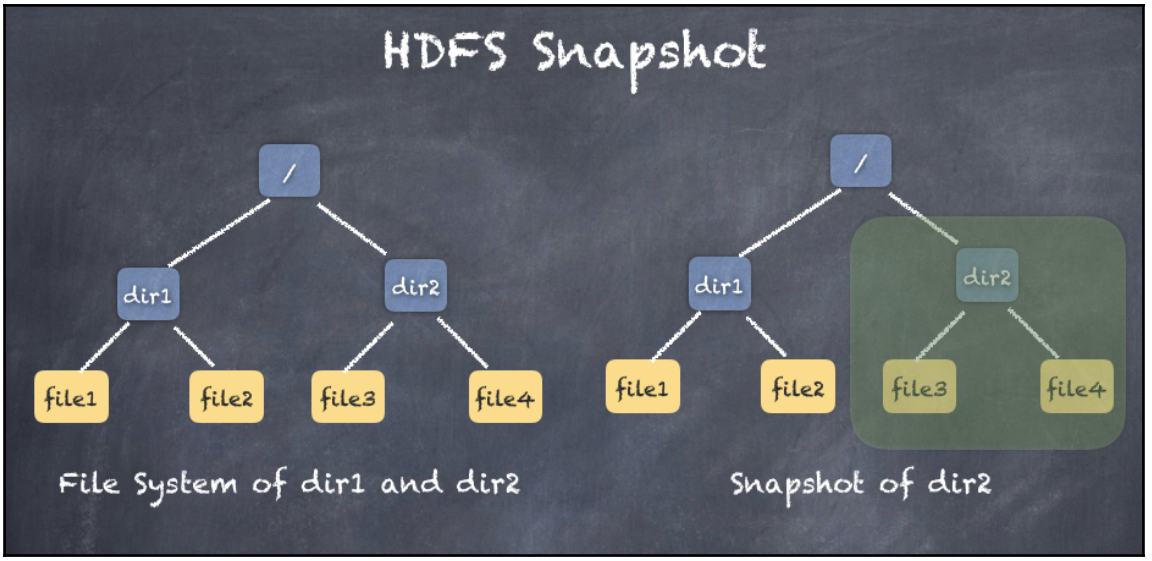

| Receipt No | Date | ItemId | Quantity | Cost$ | Sale$ |
|---|---|---|---|---|---|
| 123 | 03/01 | 24 | 3 | 9 | 12 |
| 123 | 03/01 | 25 | 2 | 6 | 7 |
| 124 | 03/02 | 12 | 4 | 7 | 9 |
| 125 | 03/02 | 25 | 1 | 3 | 4 |

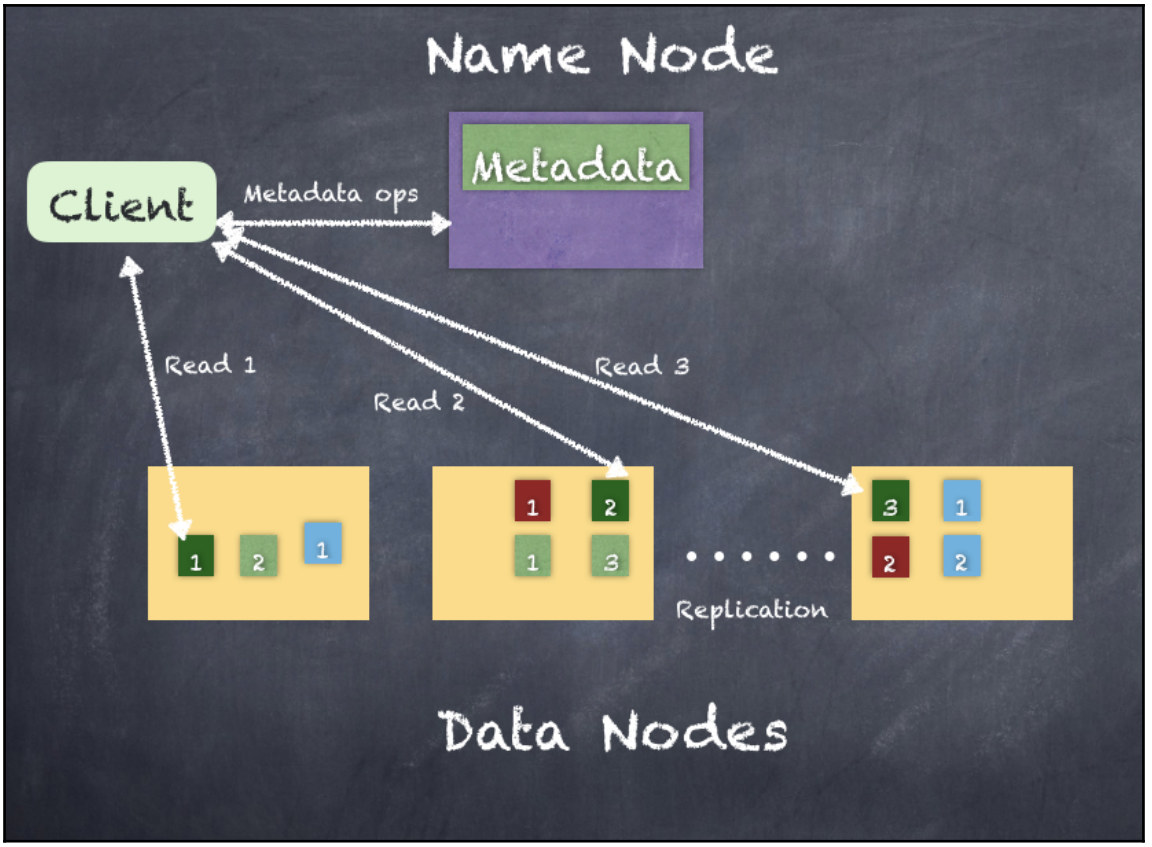

Variety of Data
+4 Billion Videos
400 Million tweets
500 Million wearables
30 Billion pieces of content
Velocity of Data
TBs of data from financial firms
20 Billion connections
Millions of Weather sensors
100 sensors per car
Big Data
Volume of Data
2.5 EBs per day
average of 100TBs per company
+40,000 EBs of data
Veracity of Data
33% dont trust the data
Trillions lost due to bad analytics

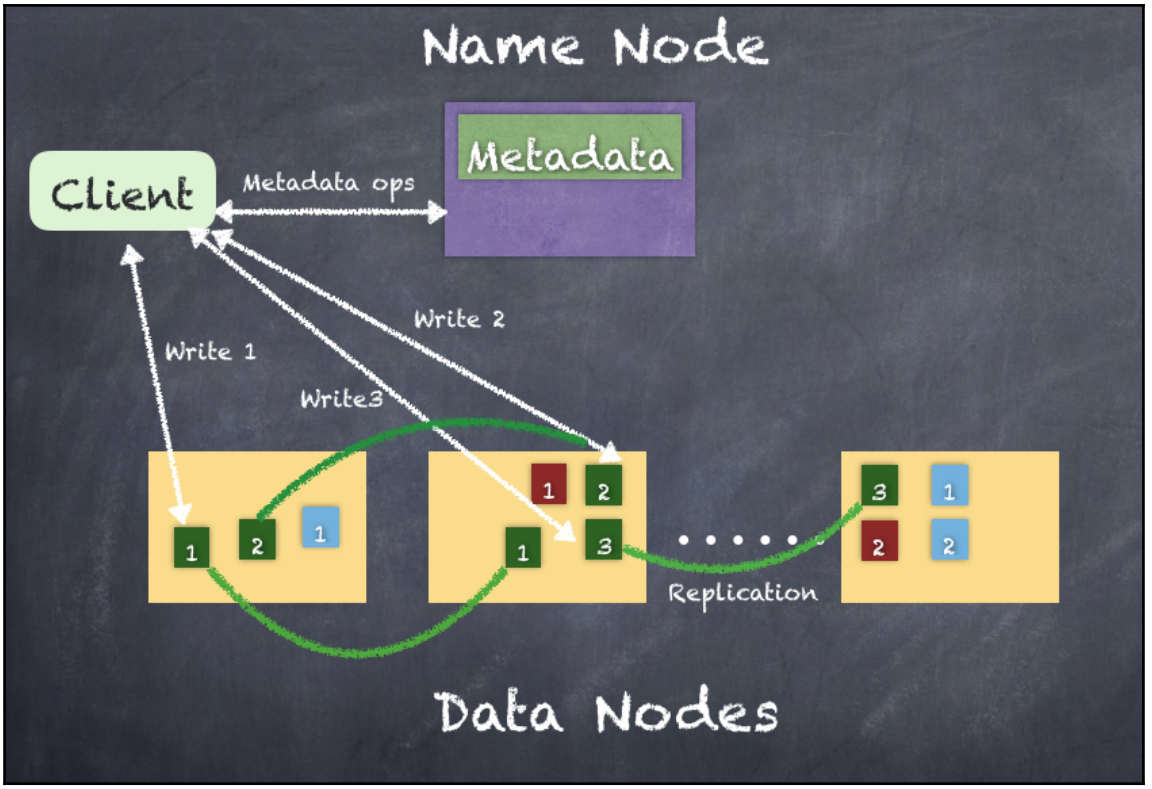

Name Node
Metadata
Client
Metadata ops
Read/Write ops
Block ops
1
1
2
1
2
1
3
3
1
2
2
Replication
Data Nodes

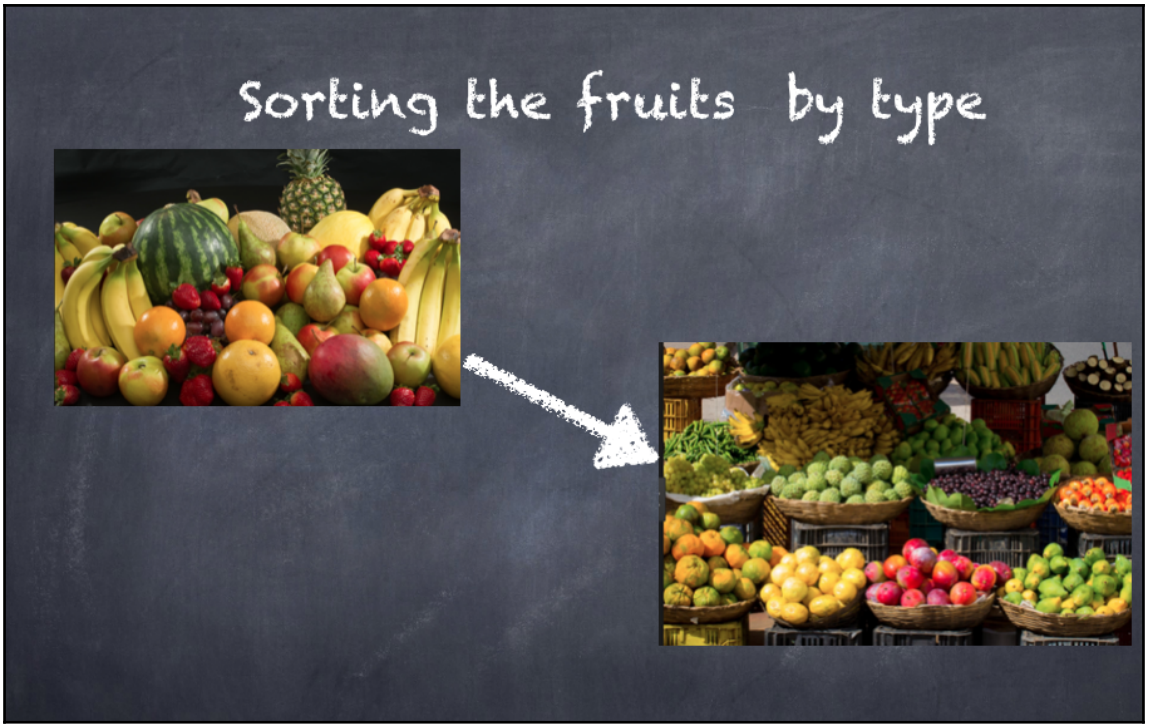

High Availability
Shared edits on NFS or
Quorum based Journal
Name Node
(Active)
Name Node
(Stand-By)
Block ops
1
1
2
1
2
1
3
3
1
2
2
Replication
Data Nodes

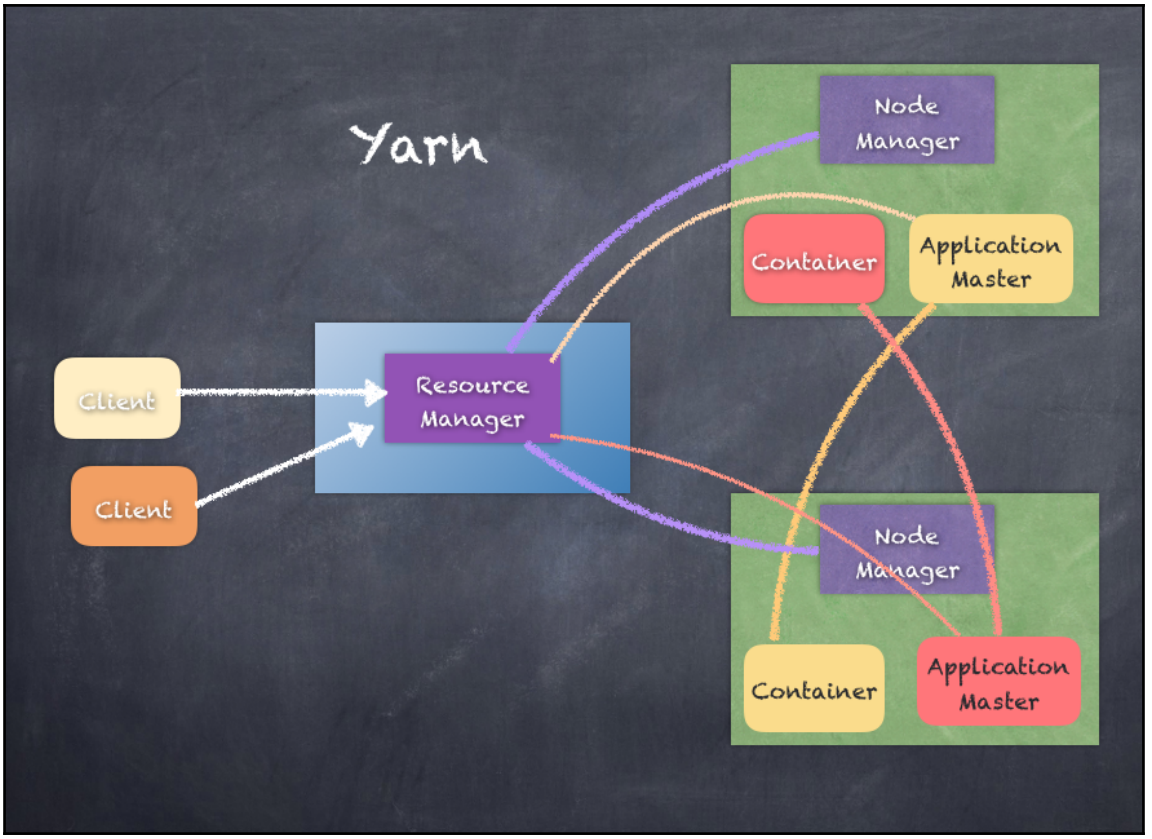

HDFS Federation
Name Node-1
Name Space-1
Name Node-2
Name Space-2
Name Node-3
Name Space-3
Block Pool & Manager
Block ops
1
1
2
1
2
1
3
3
1
2
2
Replication
Data Nodes

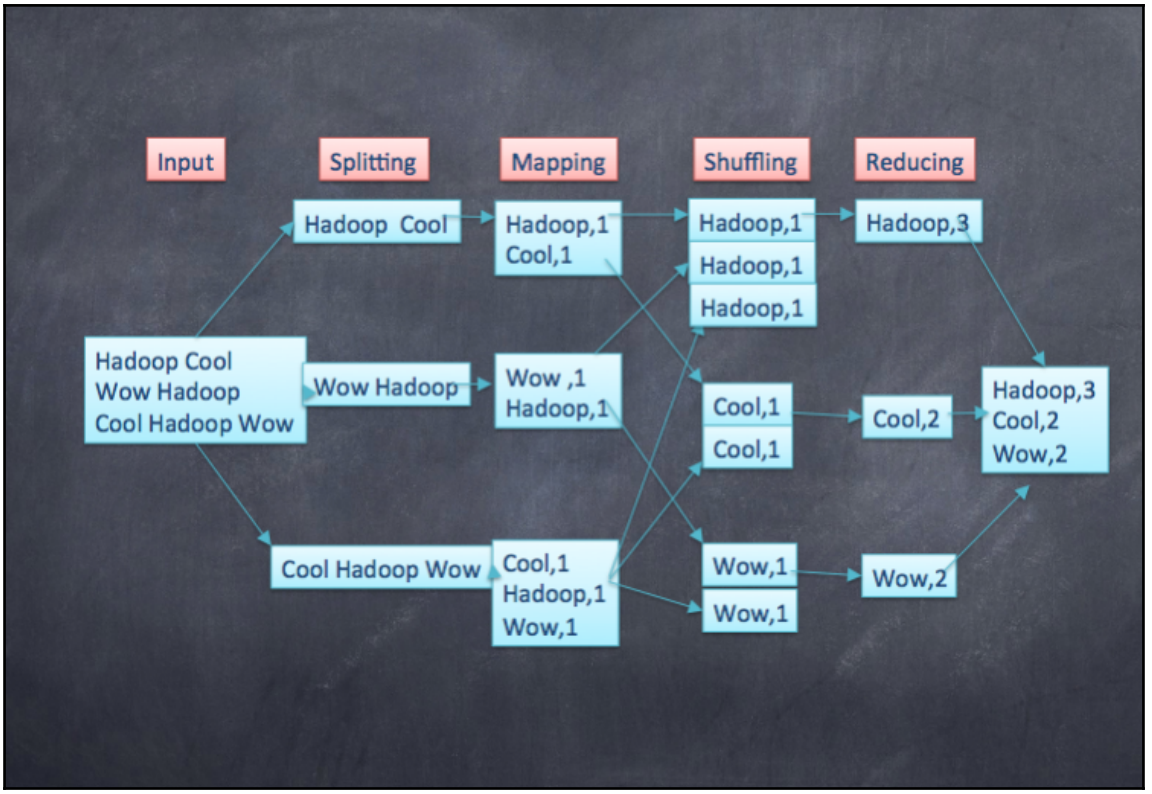

HDFS Snapshot
/
dir1
dir2
file1
file2
file3
file4
File System of dir1 and dir2
/
dir1
dir2
file1
file2
file3
file4
Snapshot of dir2

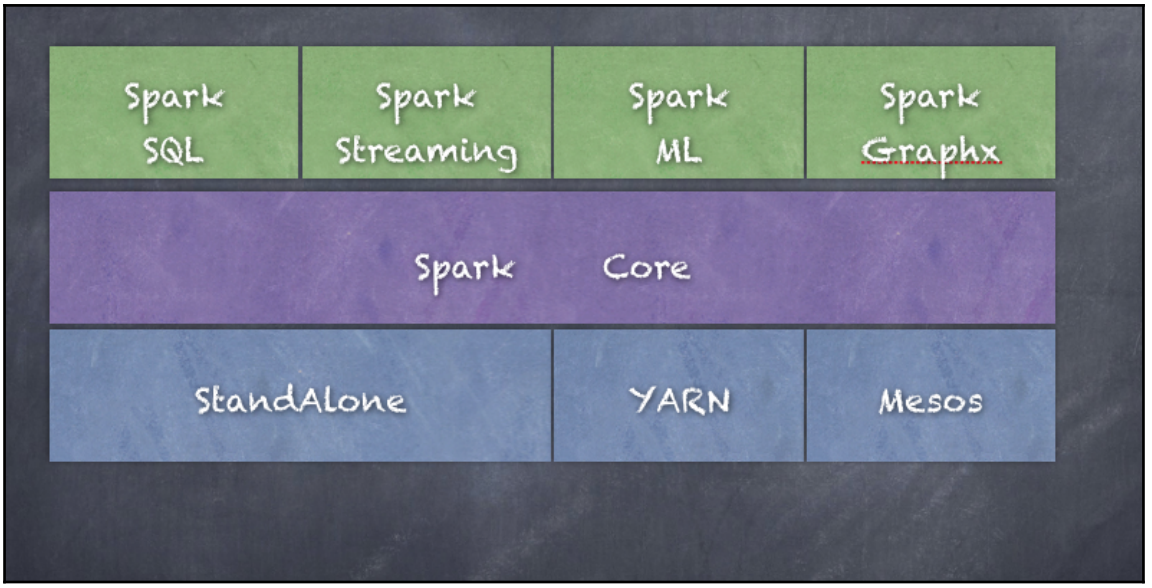

Name Node
Metadata
Client
Metadata ops
Read 1
Read 2
Read 3
1
2
1
1
2
1
3
3
1
2
2
Replication
Data Nodes

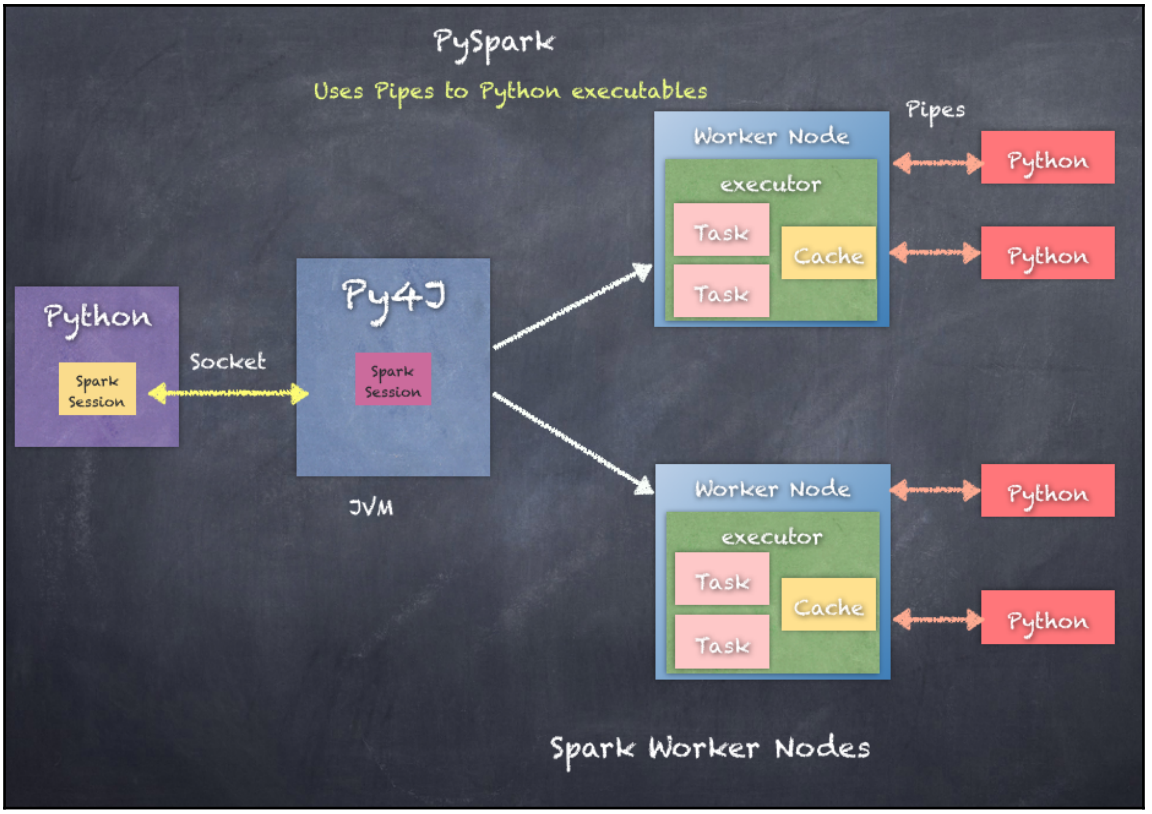

Name Node
Metadata
Client
Metadata ops
Write 1
Write 2
Write3
1
2
1
1
2
3
1
3
2
1
2
Replication
Data Nodes

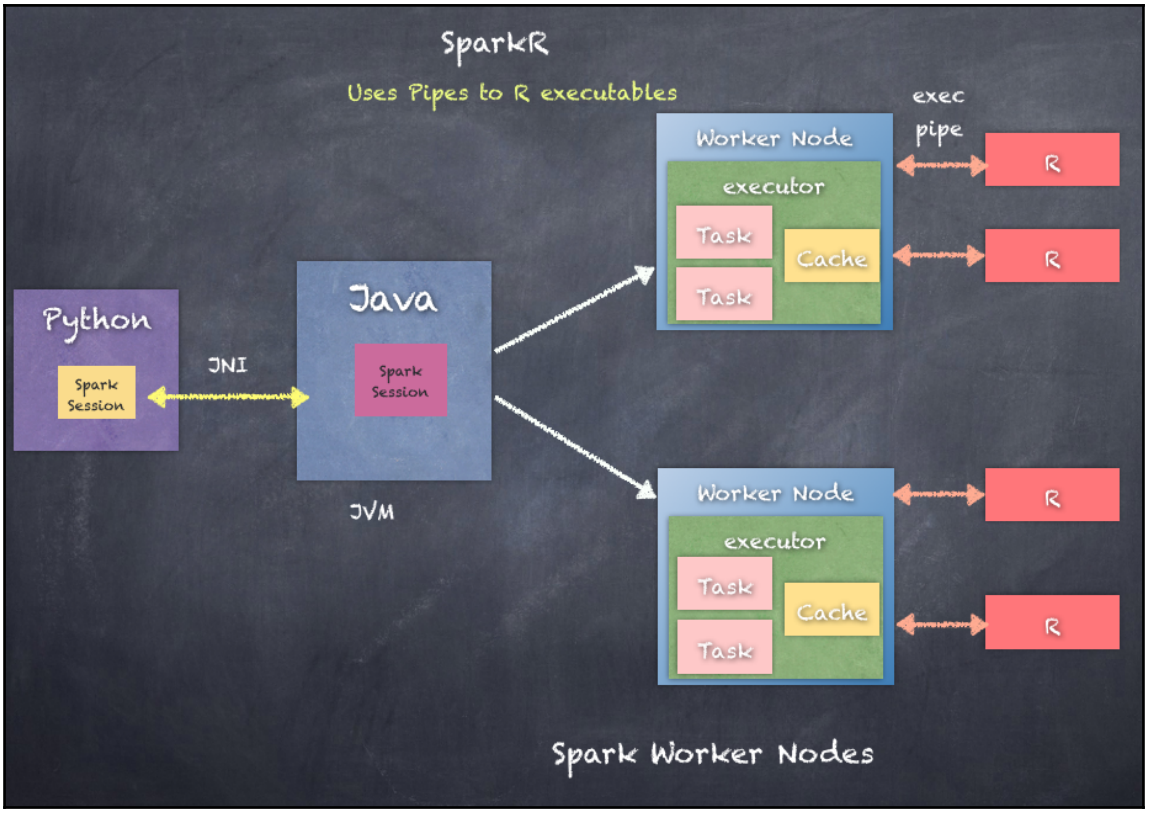

Sorting the fruits by type


Yarn
Node Manager
Container
Application Master
Client
Resource Manager
Client
Node Manager
Container
Application Master

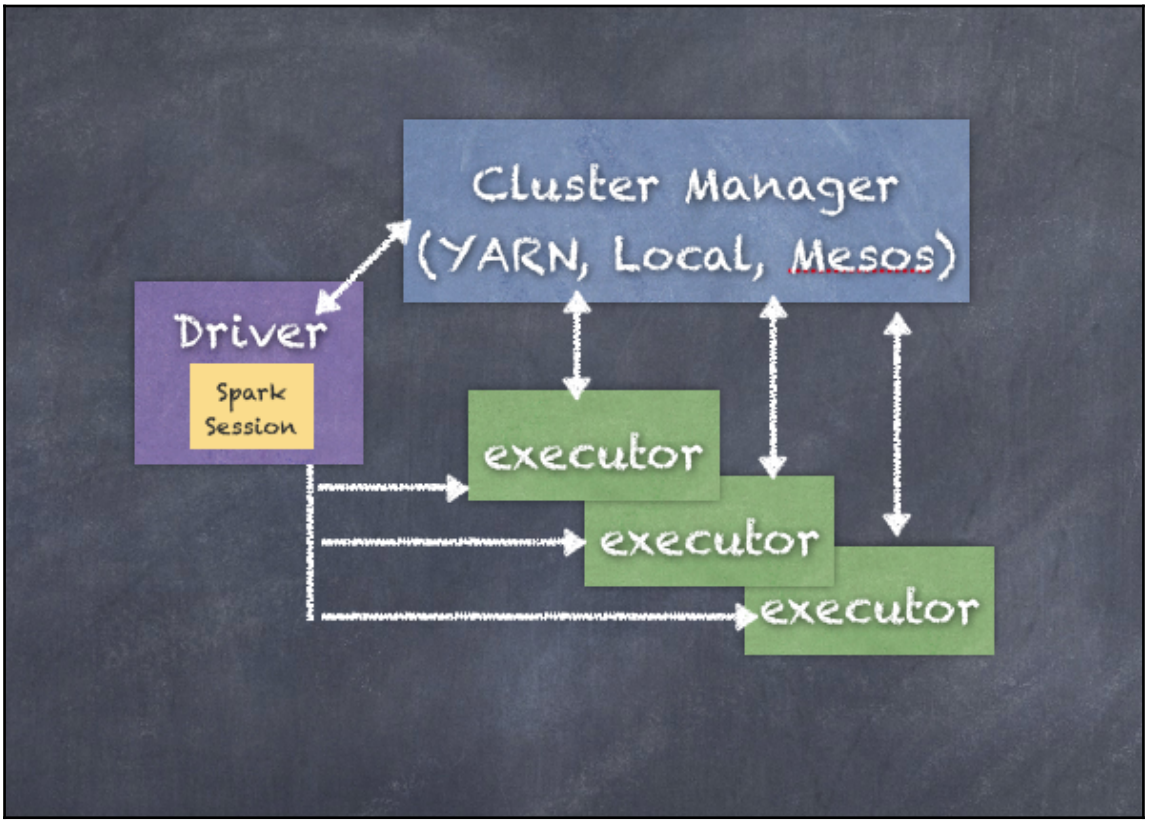

Input
Splitting
Mapping
Shuffling
Reducing
Hadoop Cool
Wow Hadoop
Cool Hadoop Wow
Hadoop Cool
Hadoop,1
Cool,1
Wow ,1
Hadoop,1
Cool Hadoop Wow
Cool,1
Hadoop,1
Wow,1
Hadoop,1
Hadoop,1
Hadoop,1
Cool,1
Cool,1
Wow,1
Wow,1
Hadoop,3
Cool,2
Wow,2
Hadoop,3
Cool,2
Wow,2

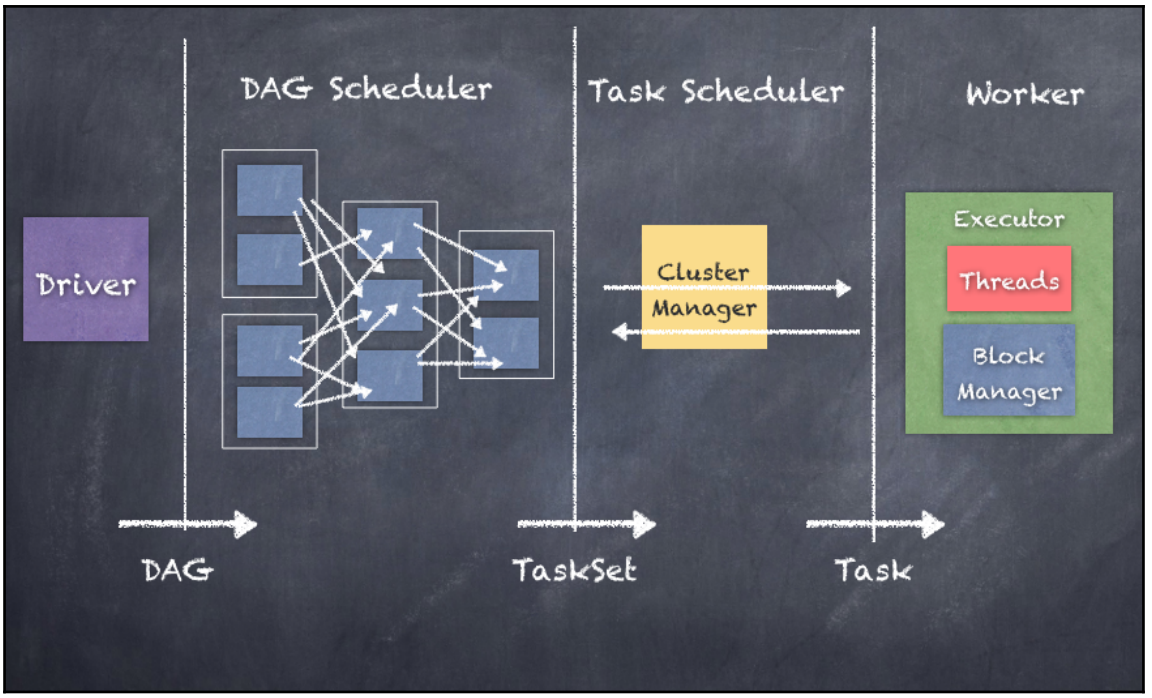

Spark SQL
Spark Streaming
Spark ML
Spark Graphx
Spark Core
StandAlone
YARN
Mesos

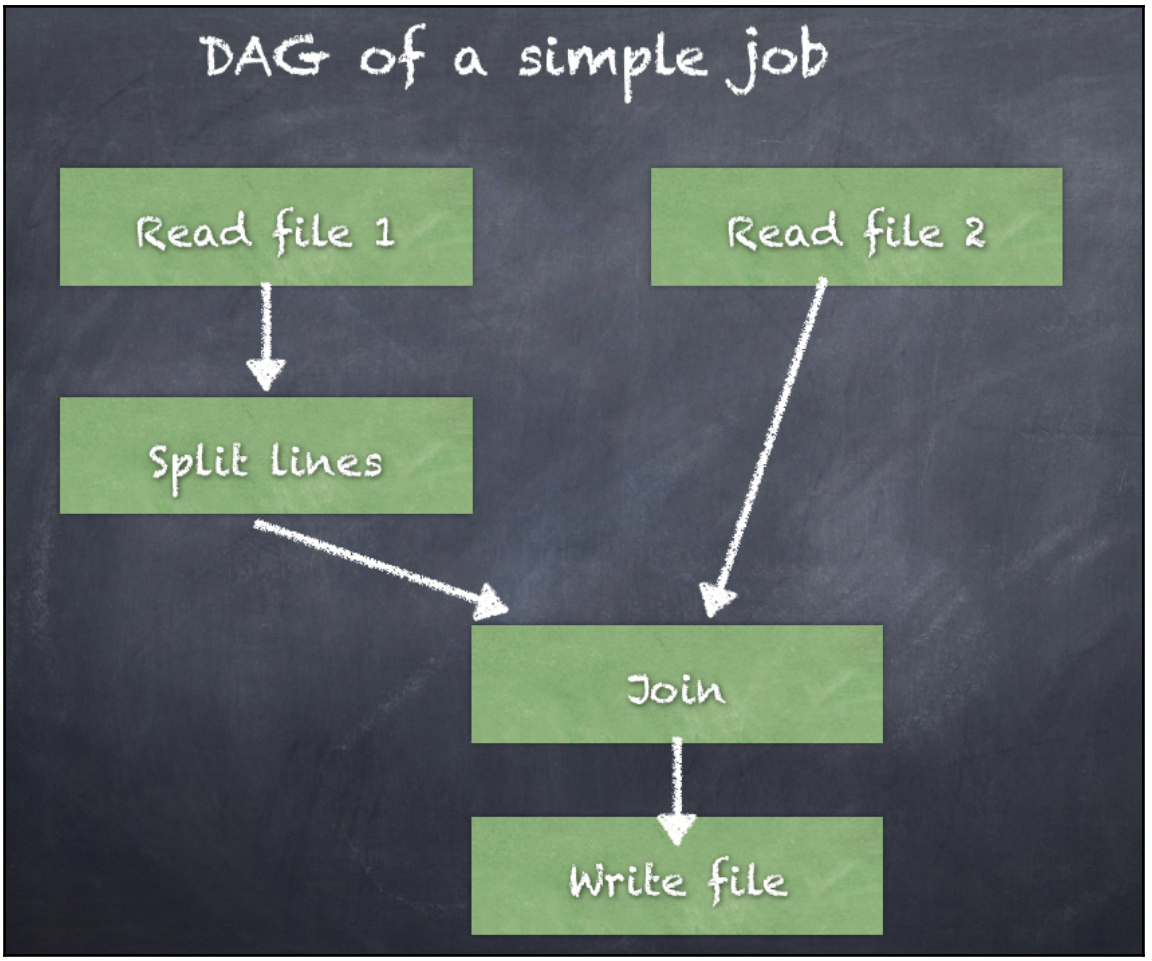

PySpark
Uses Pipes to Python executables
Pipes
Worker Node
executor
Task
Cache
Task
Python
Python
Python
Spark Session
Socket
Py4J
Spark Session
JVM
Worker Node
executor
Task
Cache
Task
Python
Python
Spark Worker Nodes

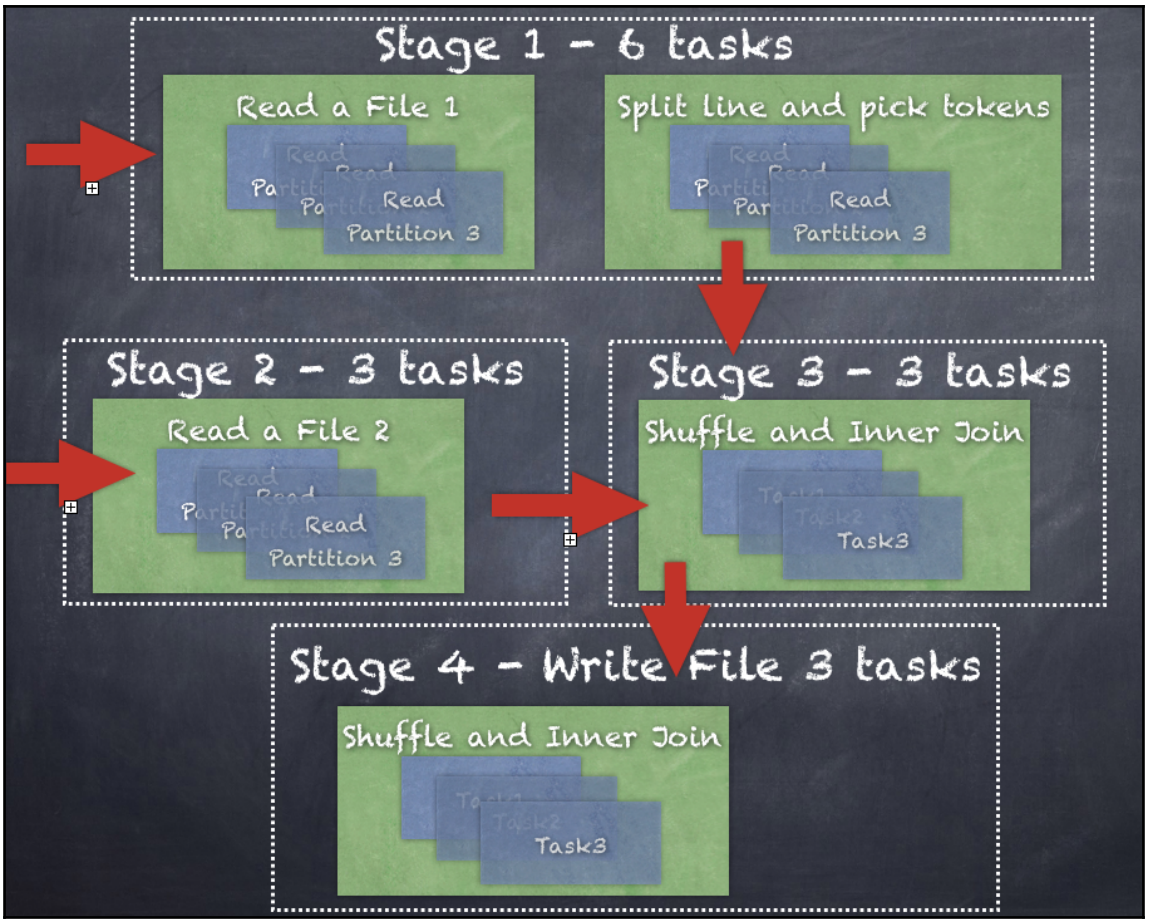

SparkR
Uses Pipes to R executables
exec
pipe
Worker Node
executor
Task
Cache
Task
R
R
Python
Spark
Session
JNI
Java
Spark
Session
JVM
Worker Node
executor
Task
Cache
Task
R
R
Spark Worker Nodes

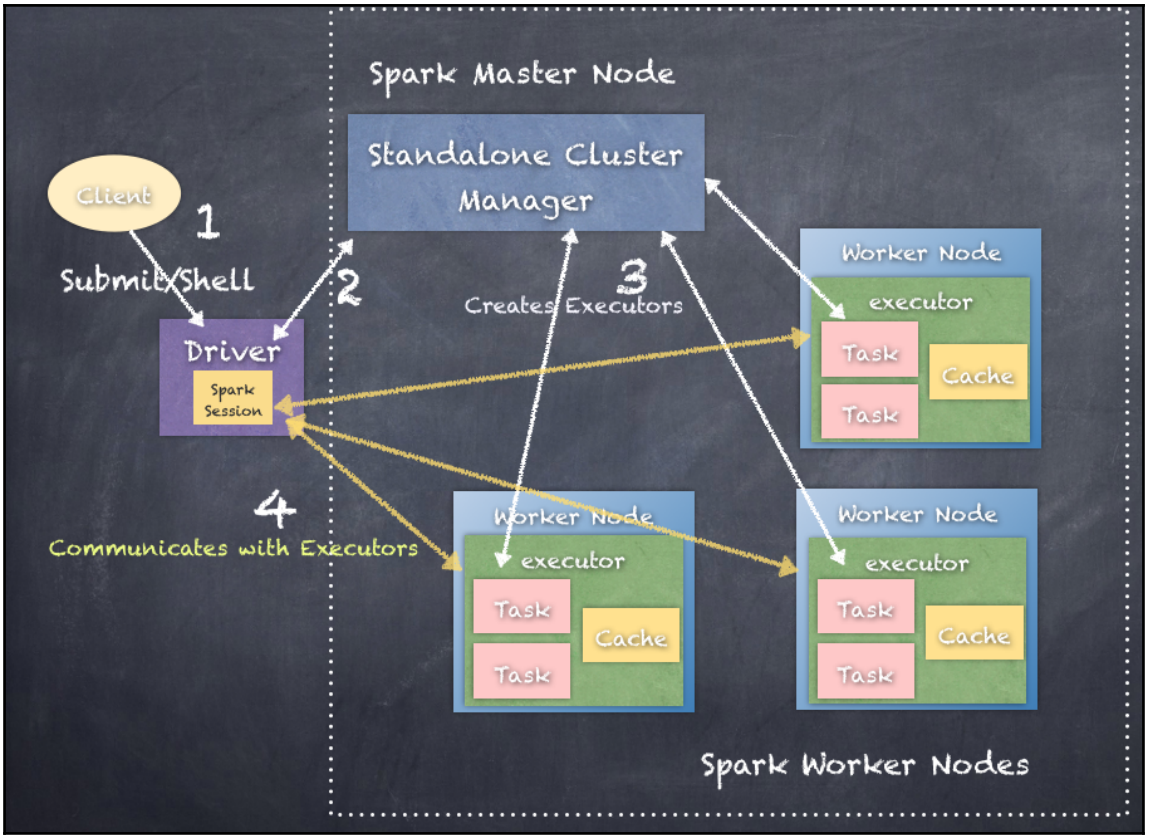
# Chapter 6: Start Working with Spark REPL and RDDs




DAG Scheduler
Task Scheduler
Worker
Driver
Cluster Manager
Executor
Threads
Block Manager
DAG
TaskSet
Task


DAG of a simple job
Read file 1
Read file 2
Split lines
Join
Write file


Stage 1 - 6 tasks
Read a File 1
Read
Partition 3
Split line and pick tokens
Read
Partition 3
Stage 2 - 3 tasks
Read a File 2
Read
Partition 3
Stage 3 - 3 tasks
Shuffle and Inner Join
Task3
Stage 4 - Write File 3 tasks
Shuffle and Inner Join
Task3


Spark Master Node
Standalone Cluster Manager
Client
1
Submit/Shell
2
Driver
Spark Session
3
Creates Executors
Worker Node
executor
Task
Task
Cache
4
Communicates with Executors
Worker Node
executor
Task
Task
Cache
Worker Node
executor
Task
Task
Cache
Spark Worker Nodes

spark.apache.org/downloads.html

APACHE Spark™ *Lightning-fast cluster computing*

Download | Libraries | Documentation | Examples | Community | Developers

## Download Apache Spark™

1. Choose a Spark release: 2.1.1 (May 02 2017)
2. Choose a package type: Pre-built for Apache Hadoop 2.7 and later
3. Choose a download type: Direct Download
4. Download Spark: spark-2.1.1-bin-hadoop2.7.tgz
5. Verify this release using the 2.1.1 signatures and checksums and project release KEYS.

*Note: Starting version 2.0, Spark is built with Scala 2.11 by default. Scala 2.10 users should download the Spark source package and build with Scala 2.10 support.*

Spark 2.1.1 | Jobs | Stages | Storage | Environment | Executors | SQL | Spark shell application UI

## Spark Jobs (?)

**User:** salla
**Total Uptime:** 33.0 h
**Scheduling Mode:** FIFO
**Completed Jobs:** 61

Event Timeline

### Completed Jobs (61)

| Job Id | Description | Submitted | Duration | Stages: Succeeded/Total | Tasks (for all stages): Succeeded/Total |
|---|---|---|---|---|---|
| 60 | take at <console>:34 | 2017/05/22 21:56:46 | 6 ms | 1/1 | 3/3 |
| 59 | take at <console>:34 | 2017/05/22 21:56:46 | 22 ms | 1/1 | 4/4 |
| 58 | take at <console>:34 | 2017/05/22 21:56:46 | 6 ms | 1/1 | 1/1 |
| 57 | take at <console>:34 | 2017/05/22 21:56:18 | 12 ms | 1/1 | 3/3 |
| 56 | take at <console>:34 | 2017/05/22 21:56:18 | 22 ms | 1/1 | 4/4 |
| 55 | take at <console>:34 | 2017/05/22 21:56:18 | 5 ms | 1/1 | 1/1 |
| 54 | take at <console>:34 | 2017/05/22 21:56:03 | 5 ms | 1/1 | 3/3 |
| 53 | take at <console>:34 | 2017/05/22 21:56:03 | 7 ms | 1/1 | 4/4 |
| 52 | take at <console>:34 | 2017/05/22 21:56:03 | 19 ms | 1/1 | 1/1 |
| 51 | take at <console>:32 | 2017/05/22 21:52:45 | 20 ms | 1/1 | 3/3 |
| 50 | take at <console>:32 | 2017/05/22 21:52:45 | 21 ms | 1/1 | 4/4 |
| 49 | take at <console>:32 | 2017/05/22 21:52:45 | 12 ms | 1/1 | 1/1 |

Spark 2.1.1 | Jobs | Stages | Storage | Environment | Executors | SQL | **Spark shell** application UI

## Executors

### Summary

| | RDD Blocks | Storage Memory | Disk Used | Cores | Active Tasks | Failed Tasks | Complete Tasks | Total Tasks | Task Time (GC Time) | Input | Shuffle Read | Shuffle Write |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Active(1)** | 4 | 93.6 KB / 384.1 MB | 0.0 B | 8 | 0 | 0 | 141 | 141 | 5 s (0 ms) | 46.3 KB | 0.0 B | 0.0 B |
| **Dead(0)** | 0 | 0.0 B / 0.0 B | 0.0 B | 0 | 0 | 0 | 0 | 0 | 0 ms (0 ms) | 0.0 B | 0.0 B | 0.0 B |
| **Total(1)** | 4 | 93.6 KB / 384.1 MB | 0.0 B | 8 | 0 | 0 | 141 | 141 | 5 s (0 ms) | 46.3 KB | 0.0 B | 0.0 B |

### Executors

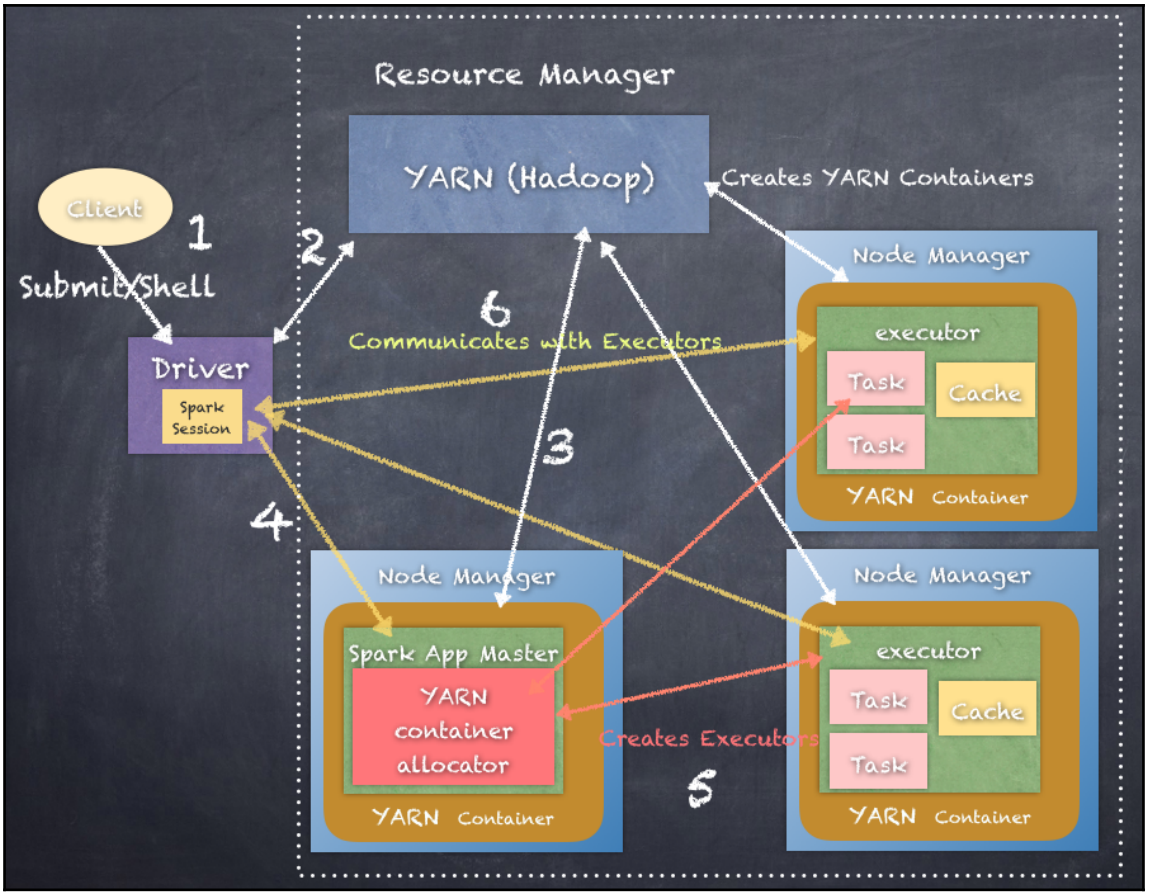
Show 20 entries

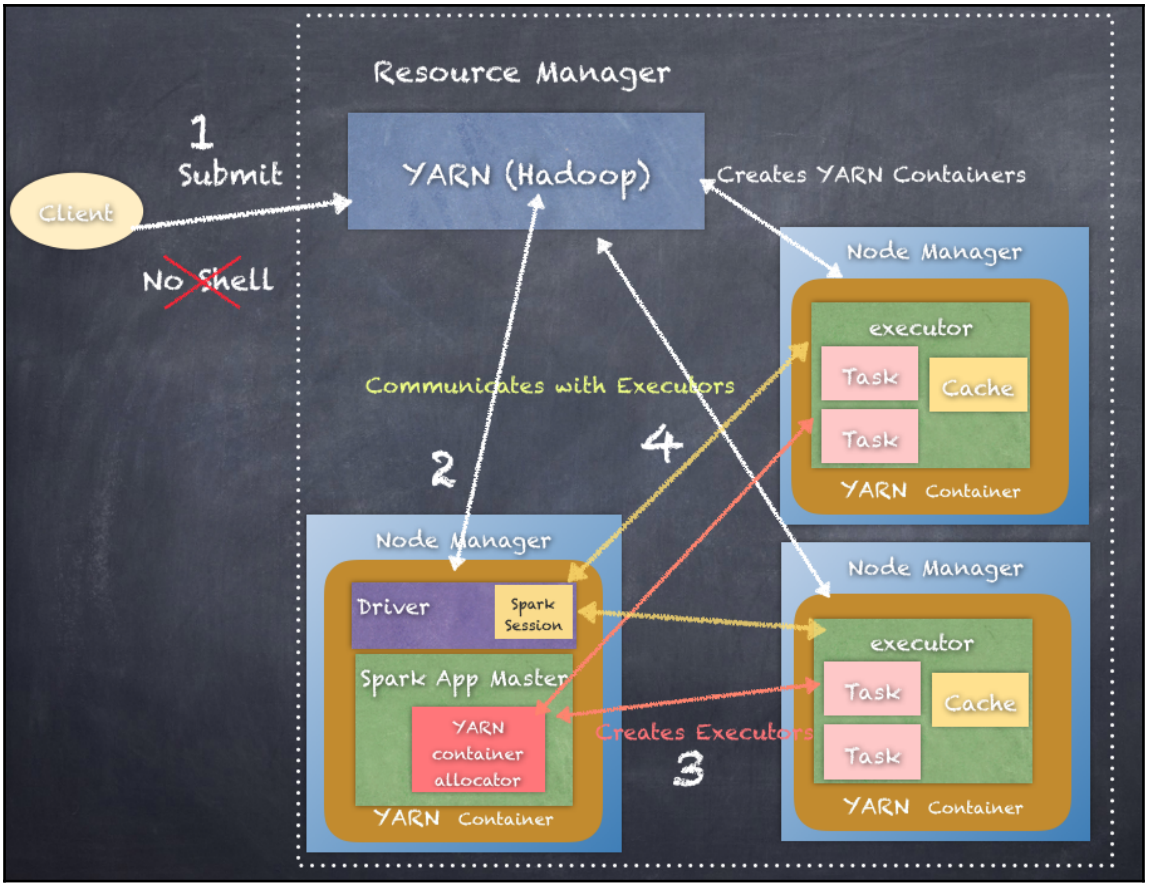
Search:

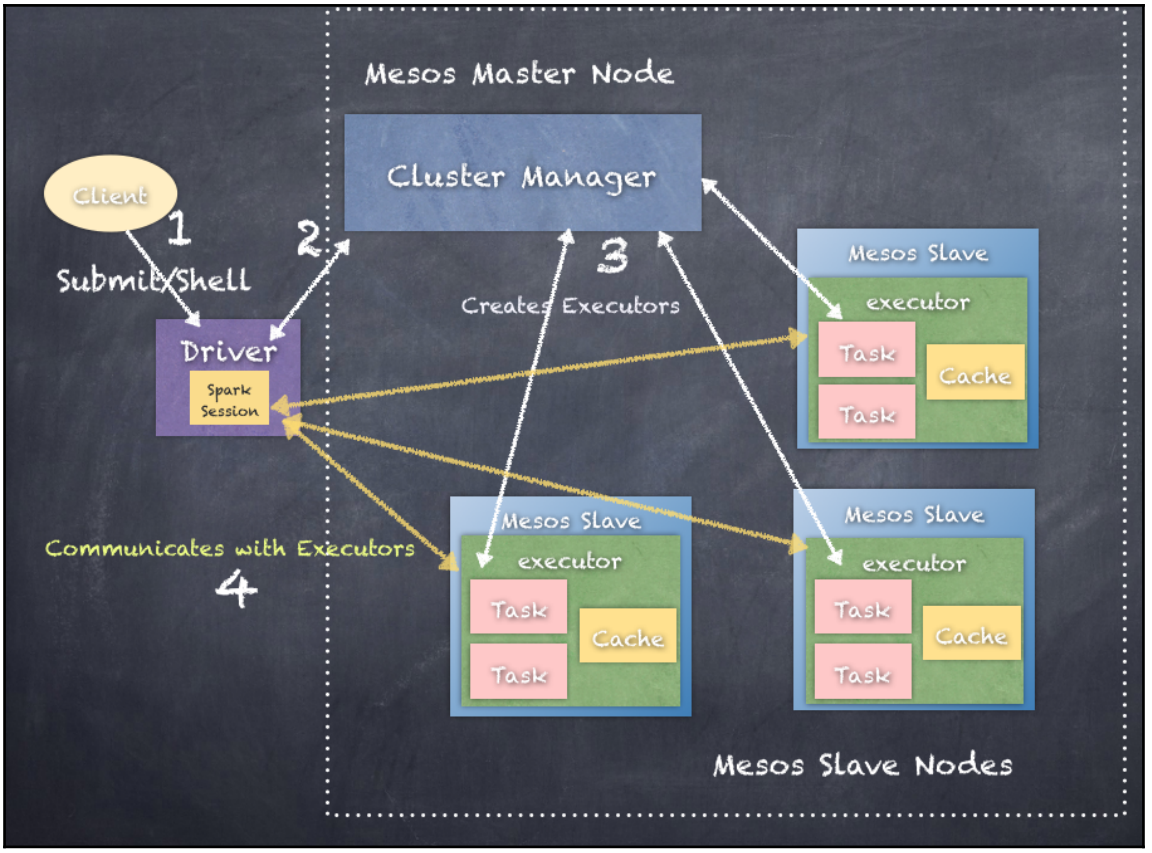
| Executor ID | Address | Status | RDD Blocks | Storage Memory | Disk Used | Cores | Active Tasks | Failed Tasks | Complete Tasks | Total Tasks | Task Time (GC Time) | Input | Shuffle Read | Shuffle Write | Thread Dump |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| driver | 192.168.1.253:60407 | Active | 4 | 93.6 KB / 384.1 MB | 0.0 B | 8 | 0 | 0 | 141 | 141 | 5 s (0 ms) | 46.3 KB | 0.0 B | 0.0 B | Thread Dump |

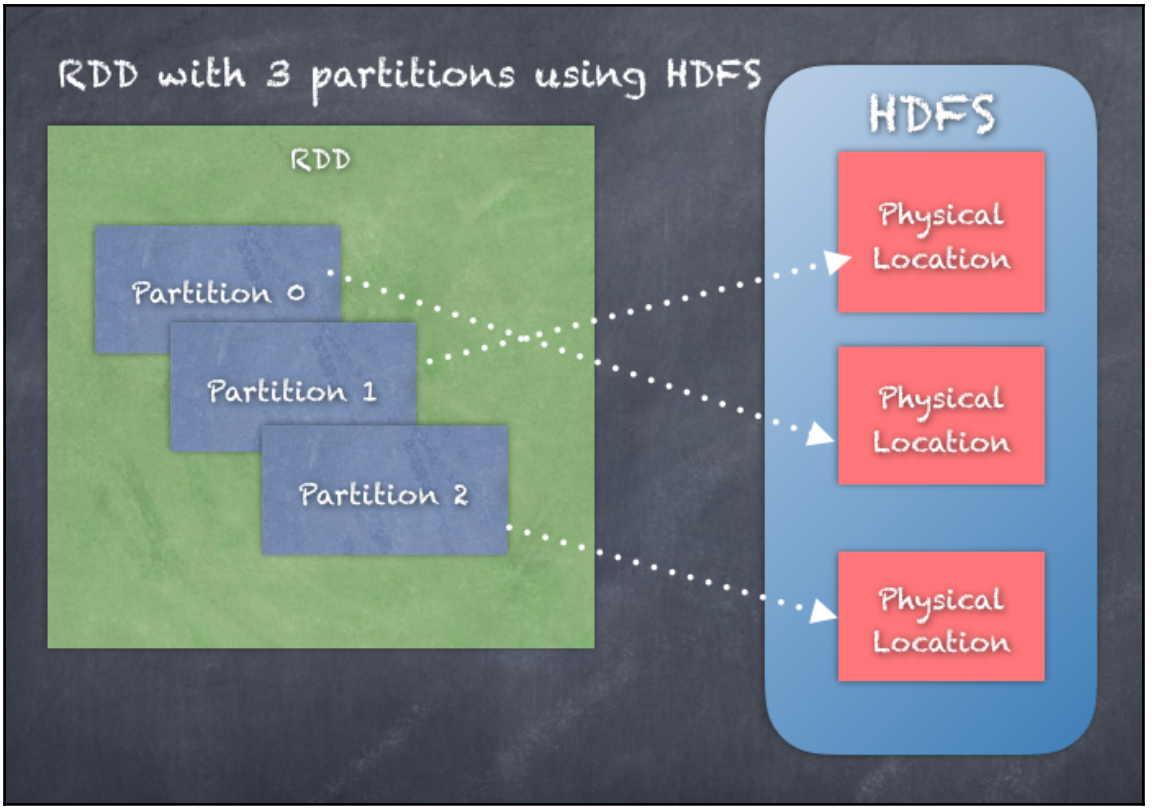
Showing 1 to 1 of 1 entries

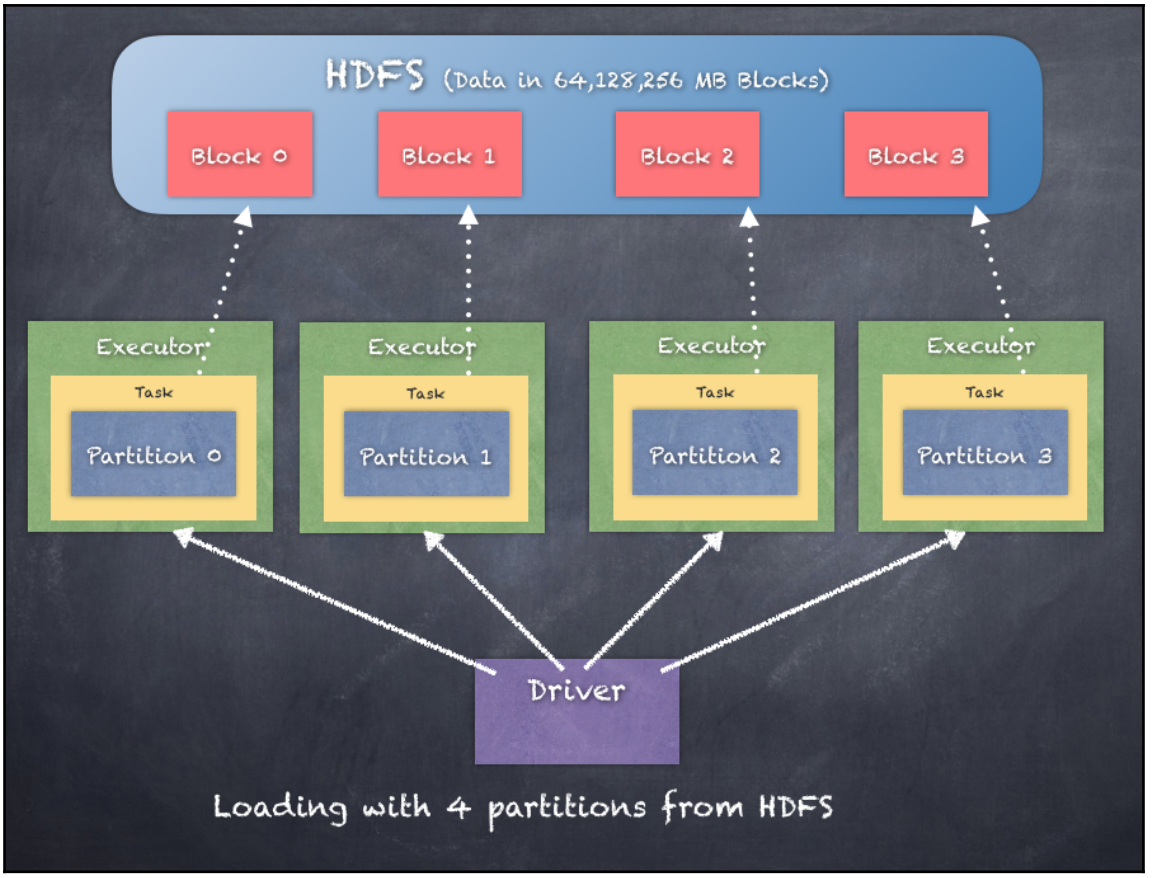
Previous 1 Next

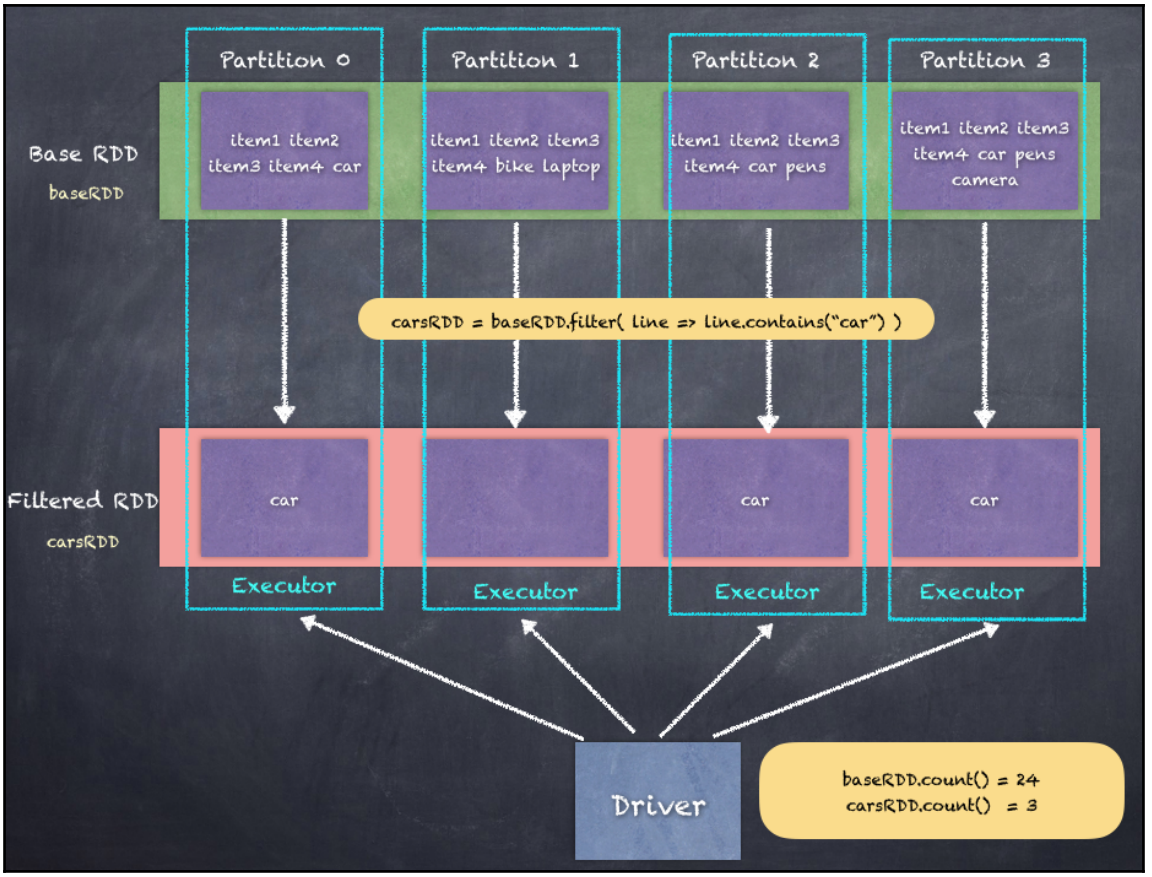

Resource Manager
YARN (Hadoop)
Creates YARN Containers
Client
1
Submit/Shell
2
Driver
Spark Session
6
Communicates with Executors
3
4
Node Manager
executor
Task
Cache
Task
YARN Container
Node Manager
Spark App Master
YARN container allocator
YARN Container
Node Manager
executor
Task
Cache
Task
YARN Container
Creates Executors
5

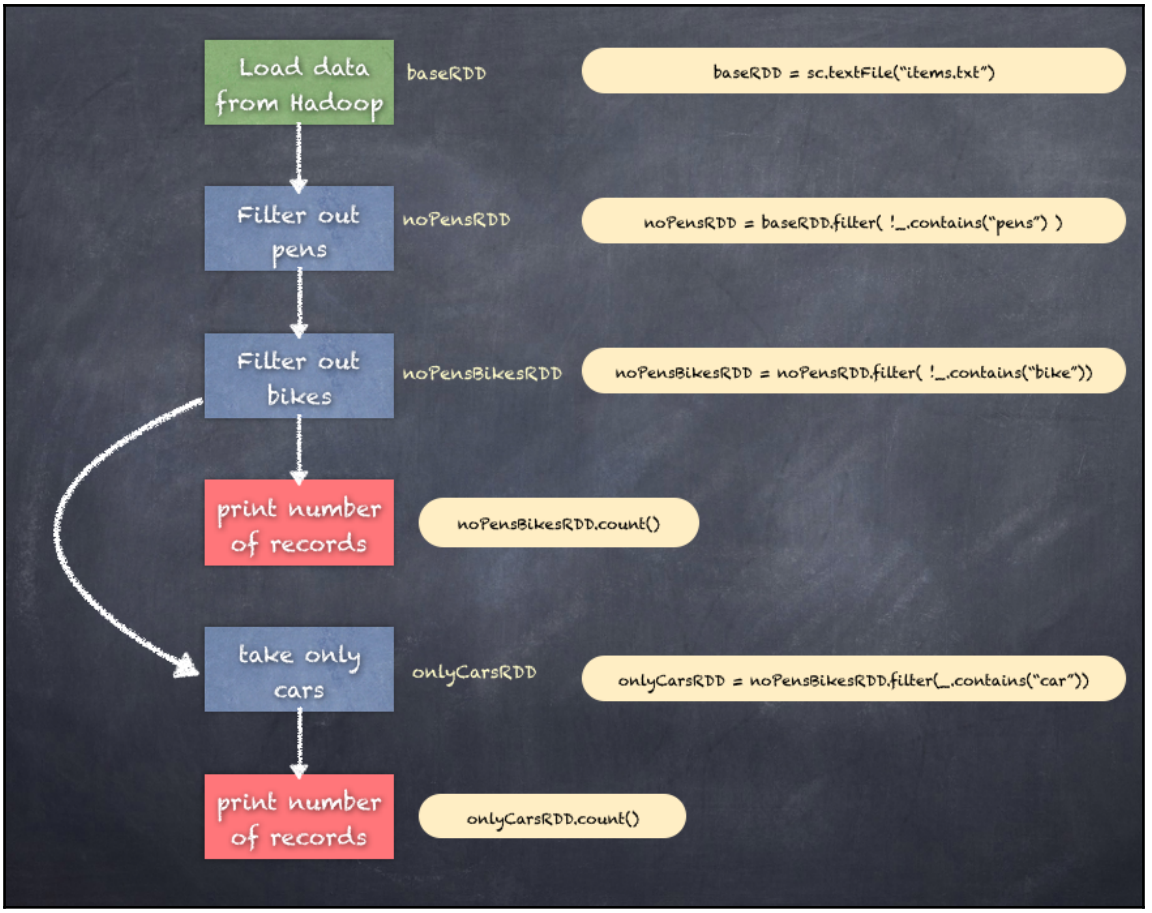

Resource Manager
1
Submit
Client
No Shell
YARN (Hadoop)
Creates YARN Containers
Node Manager
executor
Task
Cache
Task
YARN Container
Communicates with Executors
2
4
Node Manager
Driver
Spark Session
Spark App Master
YARN container allocator
YARN Container
Node Manager
executor
Task
Cache
Task
YARN Container
Creates Executors
3

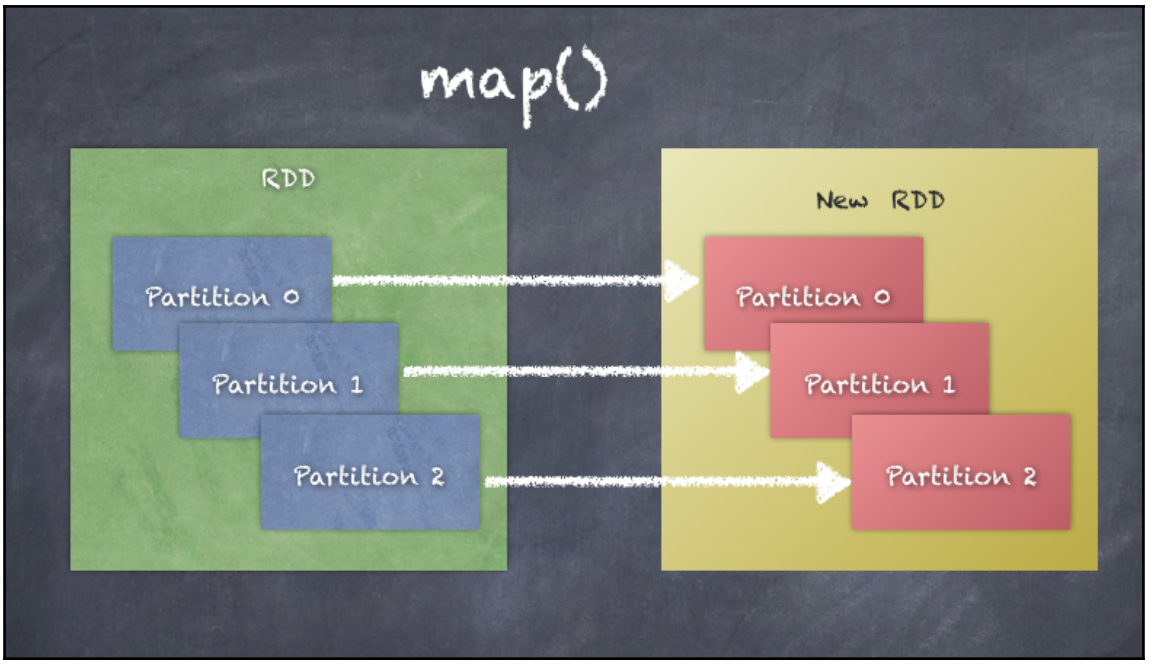

Mesos Master Node
Cluster Manager
Client
1
Submit/Shell
2
3
Creates Executors
Driver
Spark Session
Mesos Slave
executor
Task
Cache
Task
Mesos Slave
executor
Task
Cache
Task
Mesos Slave
executor
Task
Cache
Task
Communicates with Executors
4
Mesos Slave Nodes

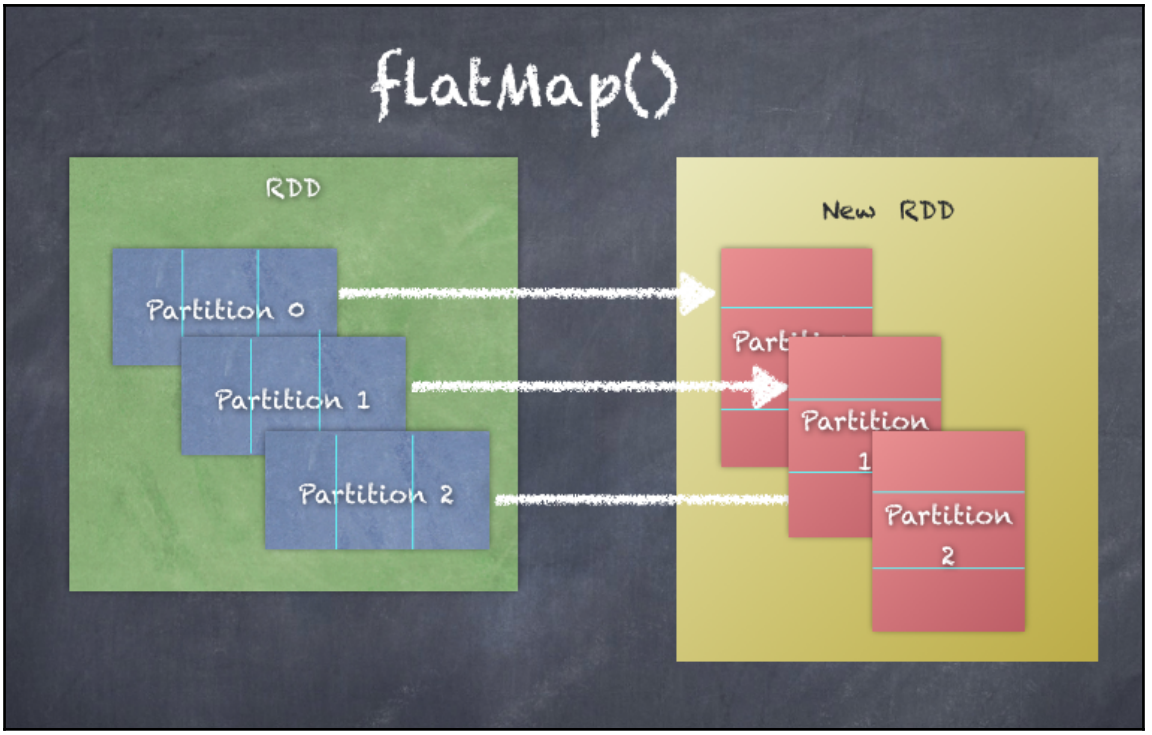

RDD with 3 partitions using HDFS
RDD
Partition 0
Partition 1
Partition 2
HDFS
Physical Location
Physical Location
Physical Location

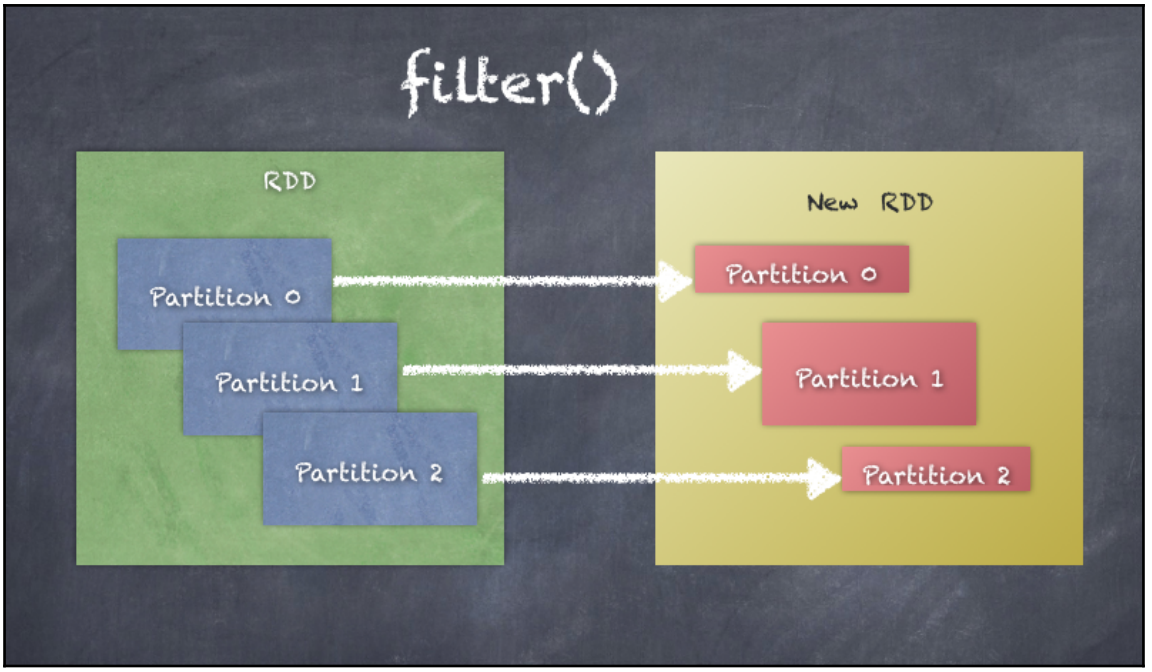

HDFS (Data in 64,128,256 MB Blocks)
Block 0
Block 1
Block 2
Block 3
Executor
Task
Partition 0
Executor
Task
Partition 1
Executor
Task
Partition 2
Executor
Task
Partition 3
Driver
Loading with 4 partitions from HDFS

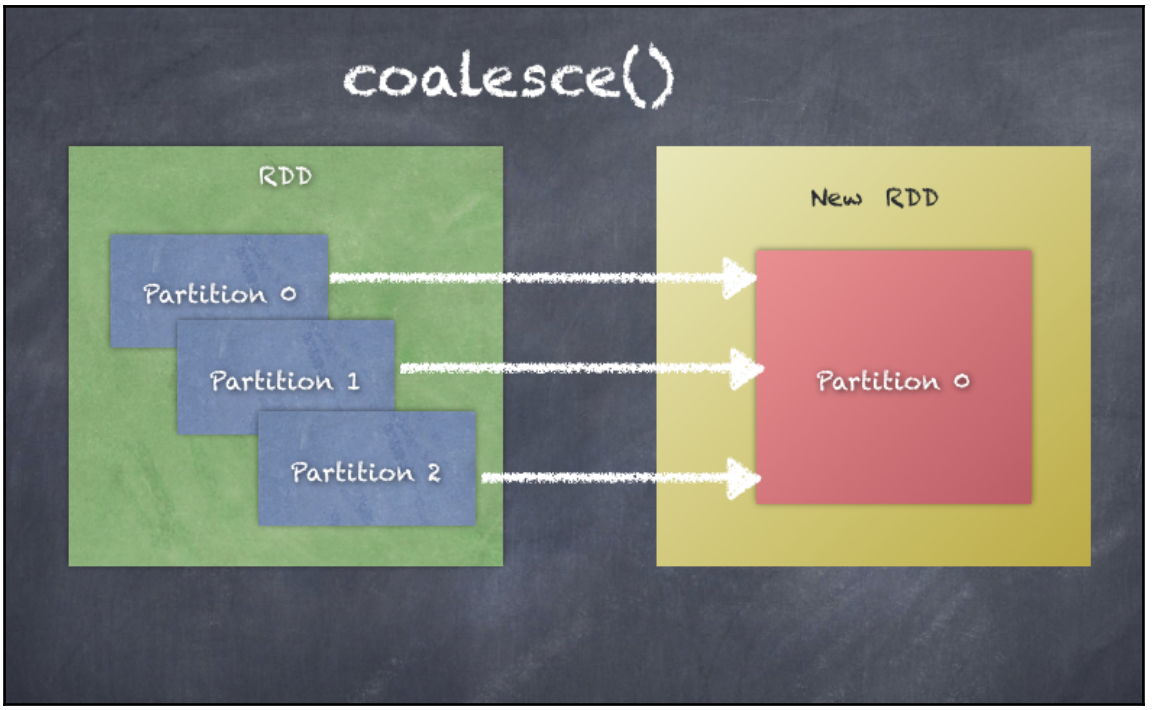

Partition 0
Partition 1
Partition 2
Partition 3
Base RDD
baseRDD
item1 item2 item3 item4 car
item1 item2 item3 item4 bike laptop
item1 item2 item3 item4 car pens
item1 item2 item3 item4 car pens camera
carsRDD = baseRDD.filter( line => line.contains("car") )
Filtered RDD
carsRDD
car
car
car
Executor
Executor
Executor
Executor
Driver
baseRDD.count() = 24
carsRDD.count() = 3

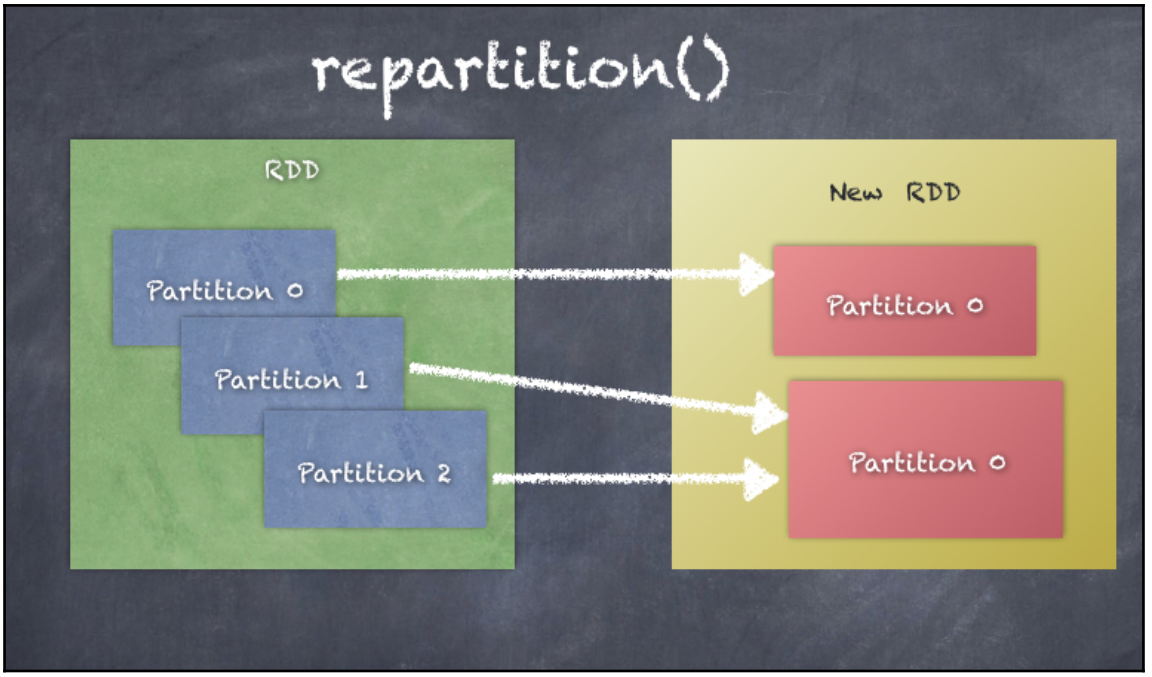

Load data from Hadoop
baseRDD
baseRDD = sc.textFile("items.txt")
Filter out pens
noPensRDD
noPensRDD = baseRDD.filter( !_.contains("pens") )
Filter out bikes
noPensBikesRDD
noPensBikesRDD = noPensRDD.filter( !_.contains("bike"))
print number of records
noPensBikesRDD.count()
take only cars
onlyCarsRDD
onlyCarsRDD = noPensBikesRDD.filter(_.contains("car"))
print number of records
onlyCarsRDD.count()

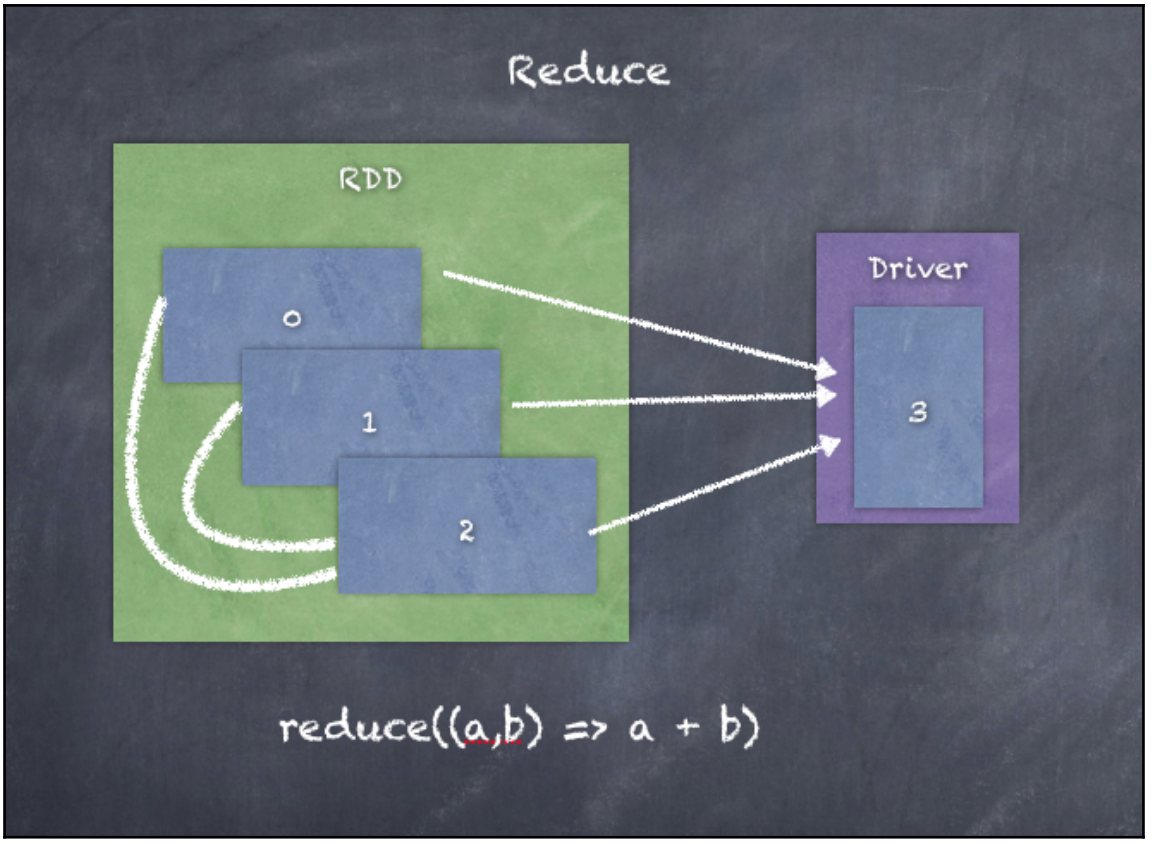

map()
RDD
Partition 0
Partition 1
Partition 2
New RDD
Partition 0
Partition 1
Partition 2

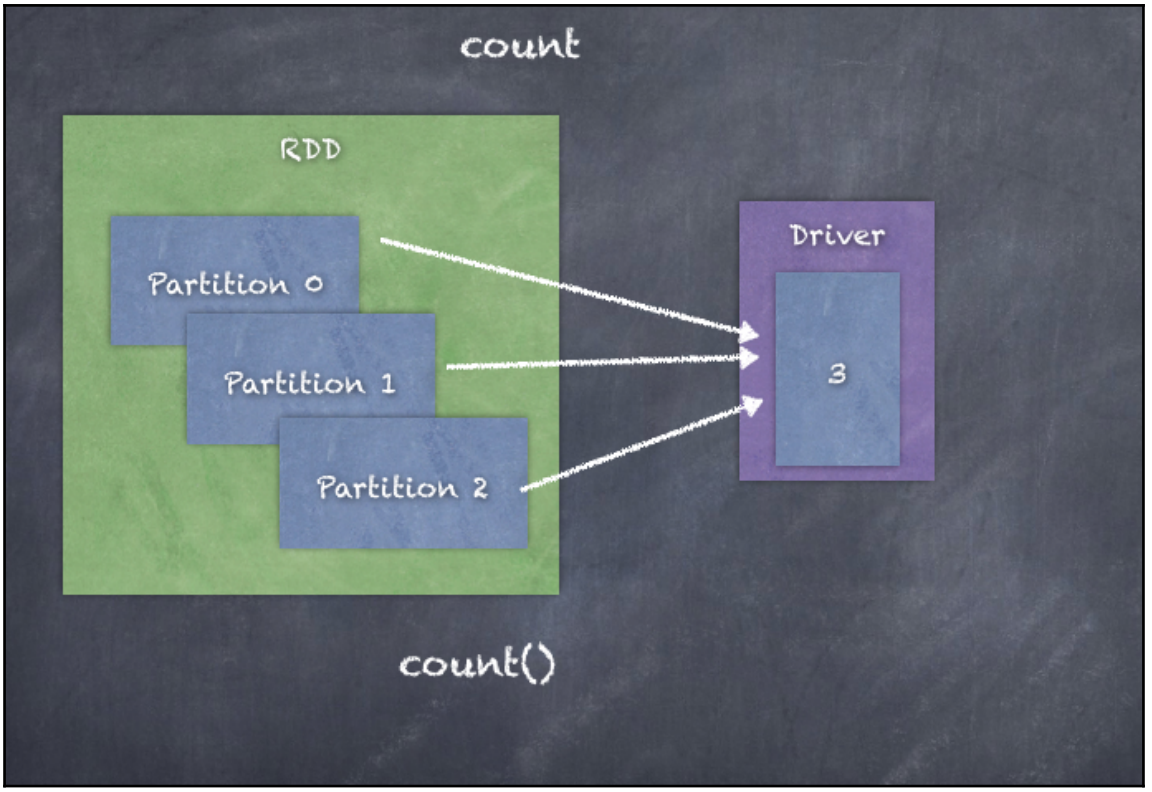

flatMap()
RDD
Partition 0
Partition 1
Partition 2
New RDD
Partition
1
Partition
2

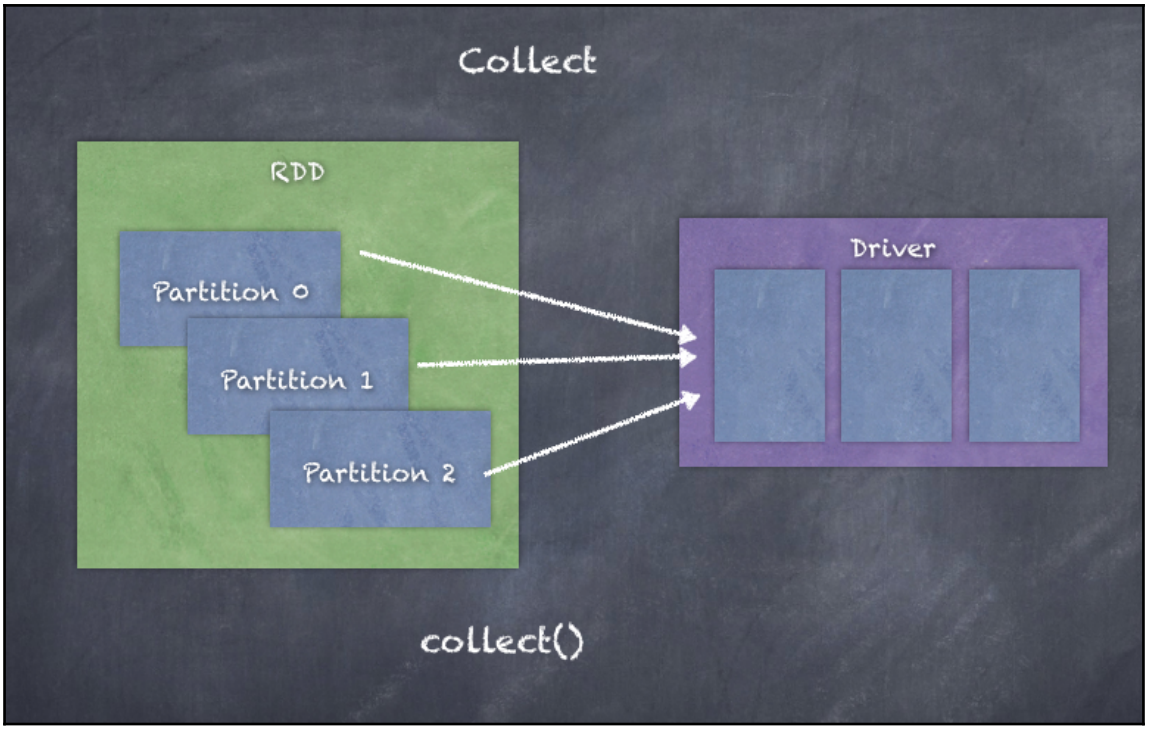

filter()
RDD
Partition 0
Partition 1
Partition 2
New RDD
Partition 0
Partition 1
Partition 2

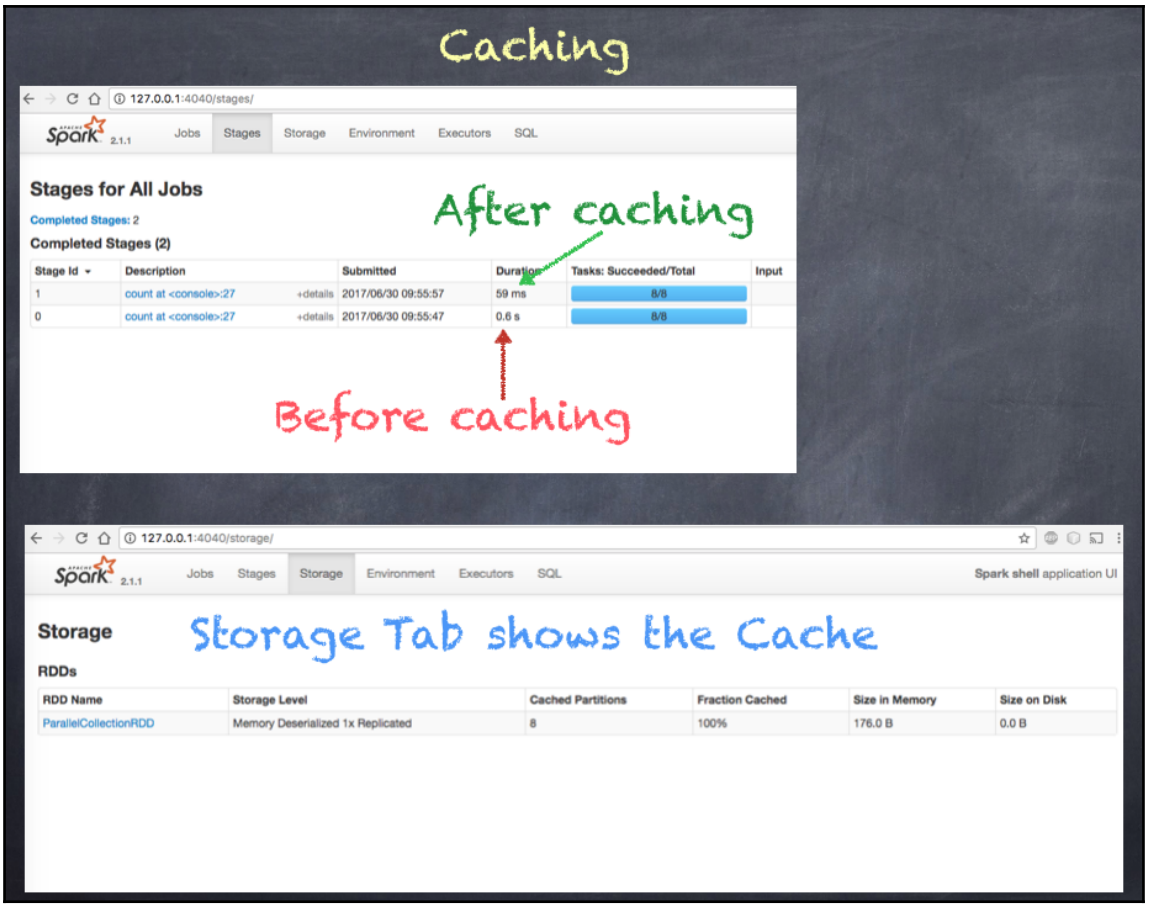

coalesce()
RDD
Partition 0
Partition 1
Partition 2
New RDD
Partition 0


repartition()
RDD
Partition 0
Partition 1
Partition 2
New RDD
Partition 0
Partition 0

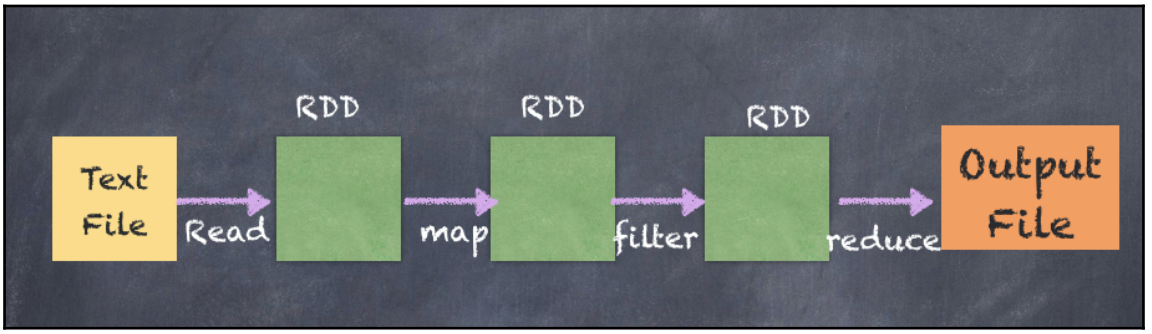

Reduce
RDD
0
1
2
Driver
3
reduce((a,b) => a + b)

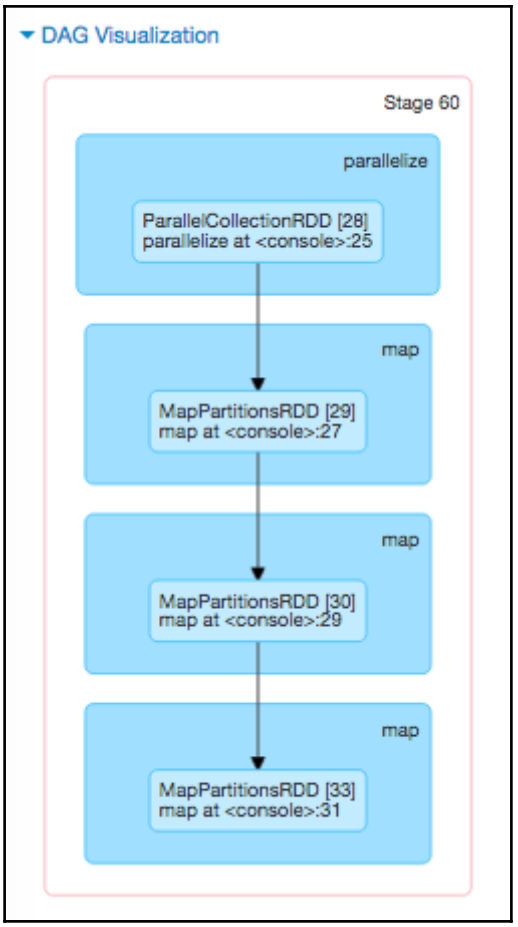

count
RDD
Partition 0
Partition 1
Partition 2
Driver
3
count()

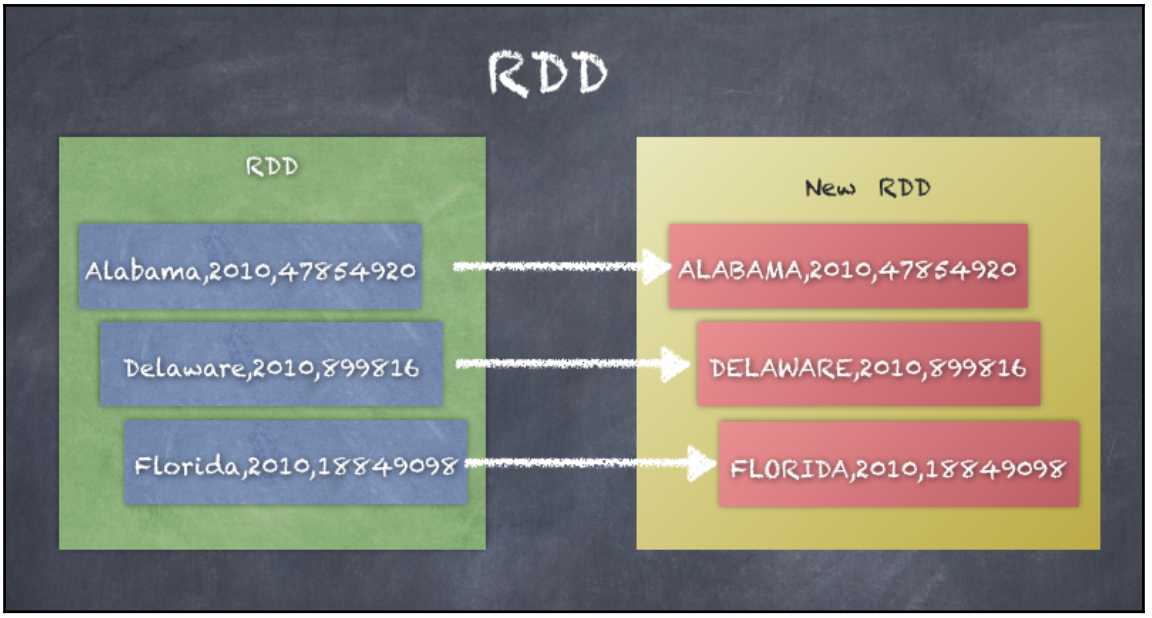

Collect
RDD
Partition 0
Partition 1
Partition 2
Driver
collect()

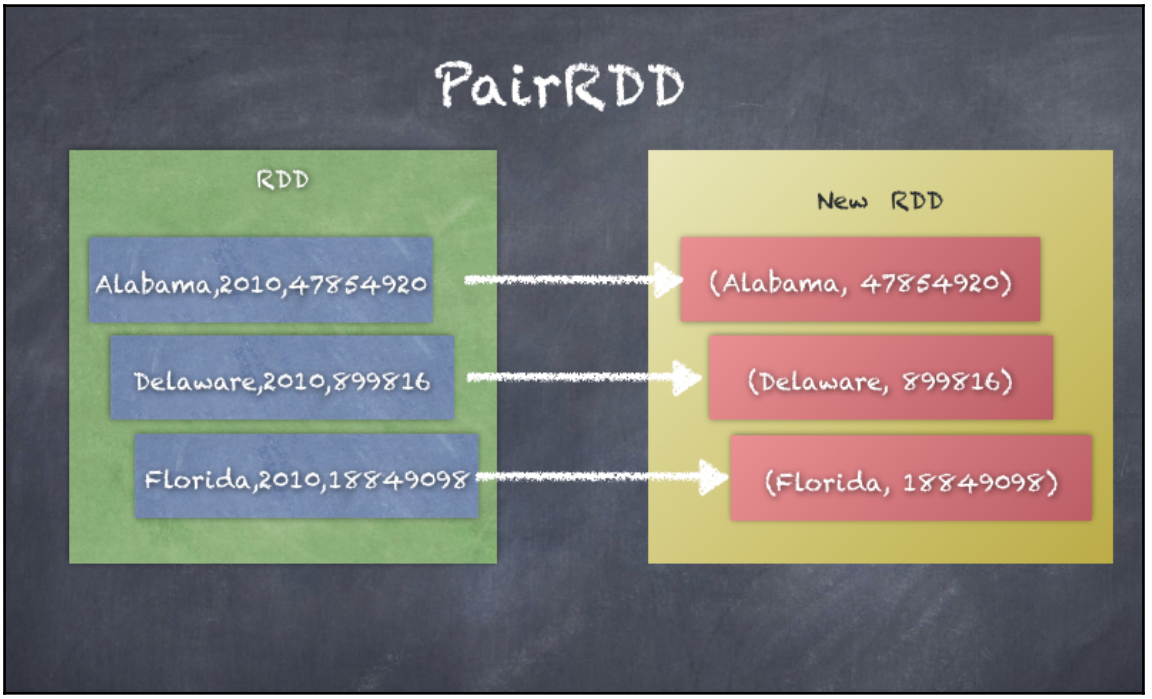

Caching
127.0.0.1:4040/stages/
Spark 2.1.1
Jobs
Stages
Storage
Environment
Executors
SQL
Stages for All Jobs
Completed Stages: 2
Completed Stages (2)
After caching
Stage Id
Description
Submitted
Duration
Tasks: Succeeded/Total
Input
1
count at <console>:27
+details
2017/06/30 09:55:57
59 ms
8/8
0
count at <console>:27
+details
2017/06/30 09:55:47
0.6 s
8/8
Before caching
127.0.0.1:4040/storage/
Spark 2.1.1
Jobs
Stages
Storage
Environment
Executors
SQL
Spark shell application UI
Storage
Storage Tab shows the Cache
RDDs
RDD Name
Storage Level
Cached Partitions
Fraction Cached
Size in Memory
Size on Disk
ParallelCollectionRDD
Memory Deserialized 1x Replicated
8
100%
176.0 B
0.0 B

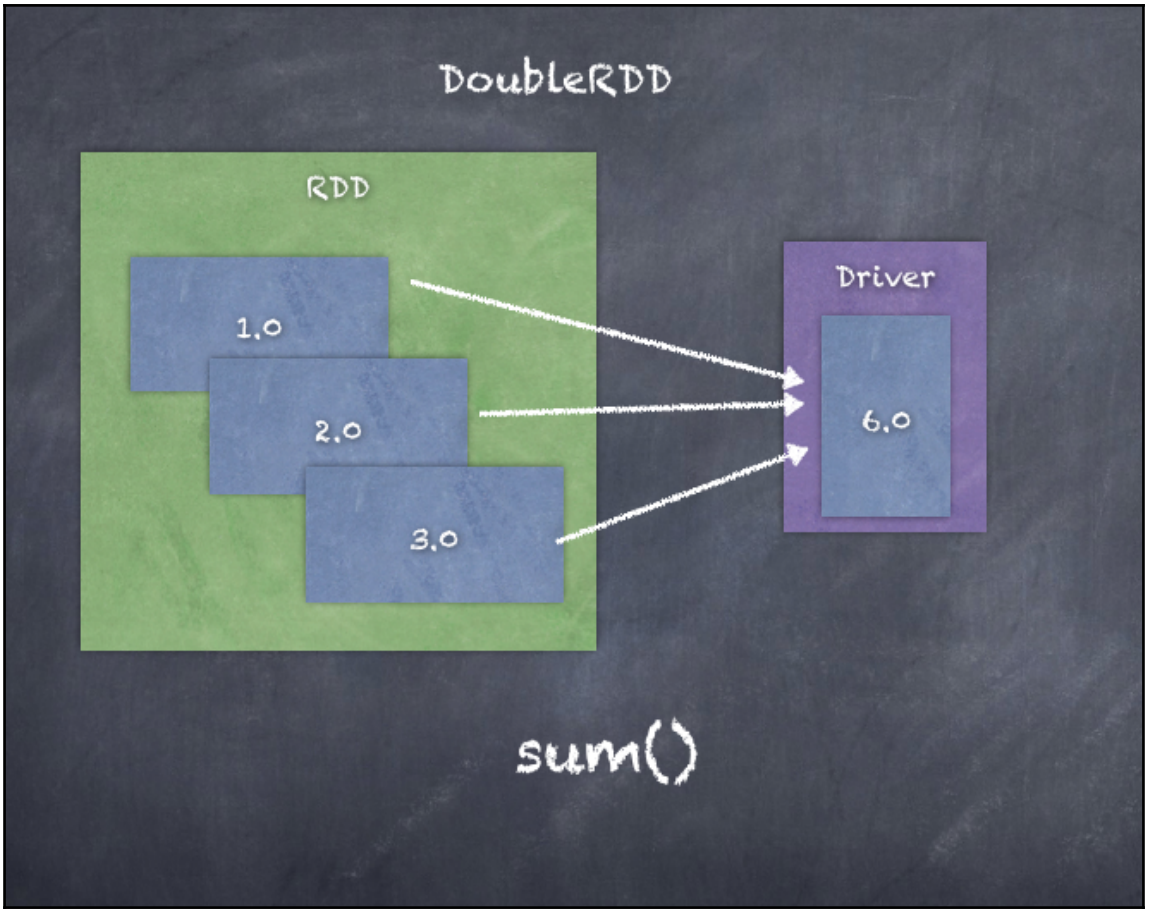
# Chapter 7: Special RDD Operations

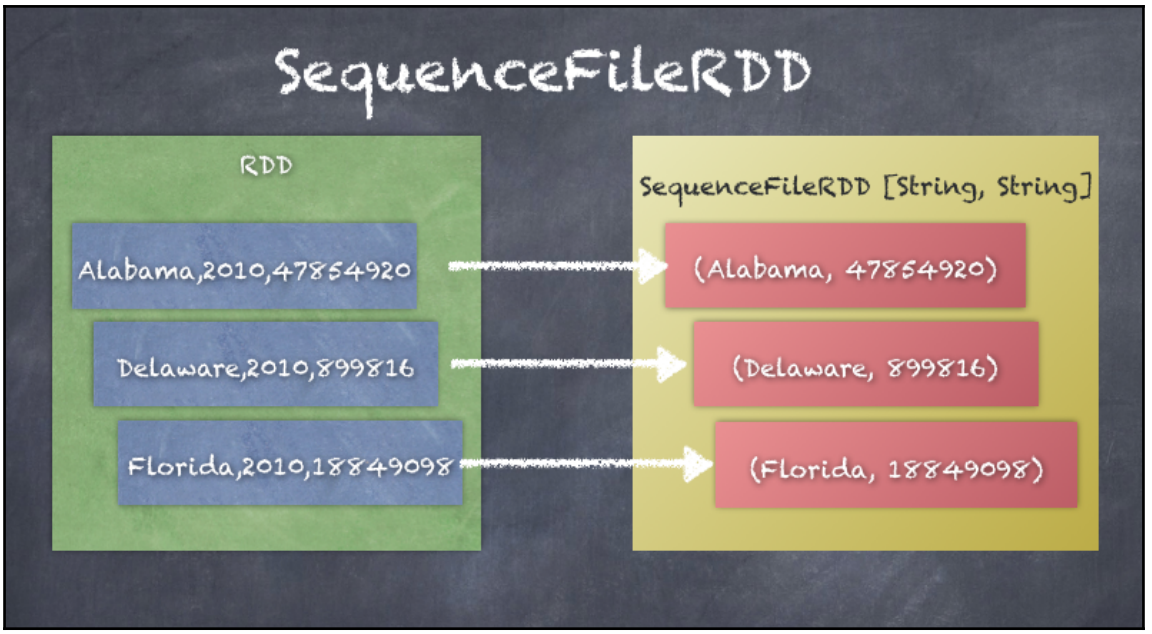


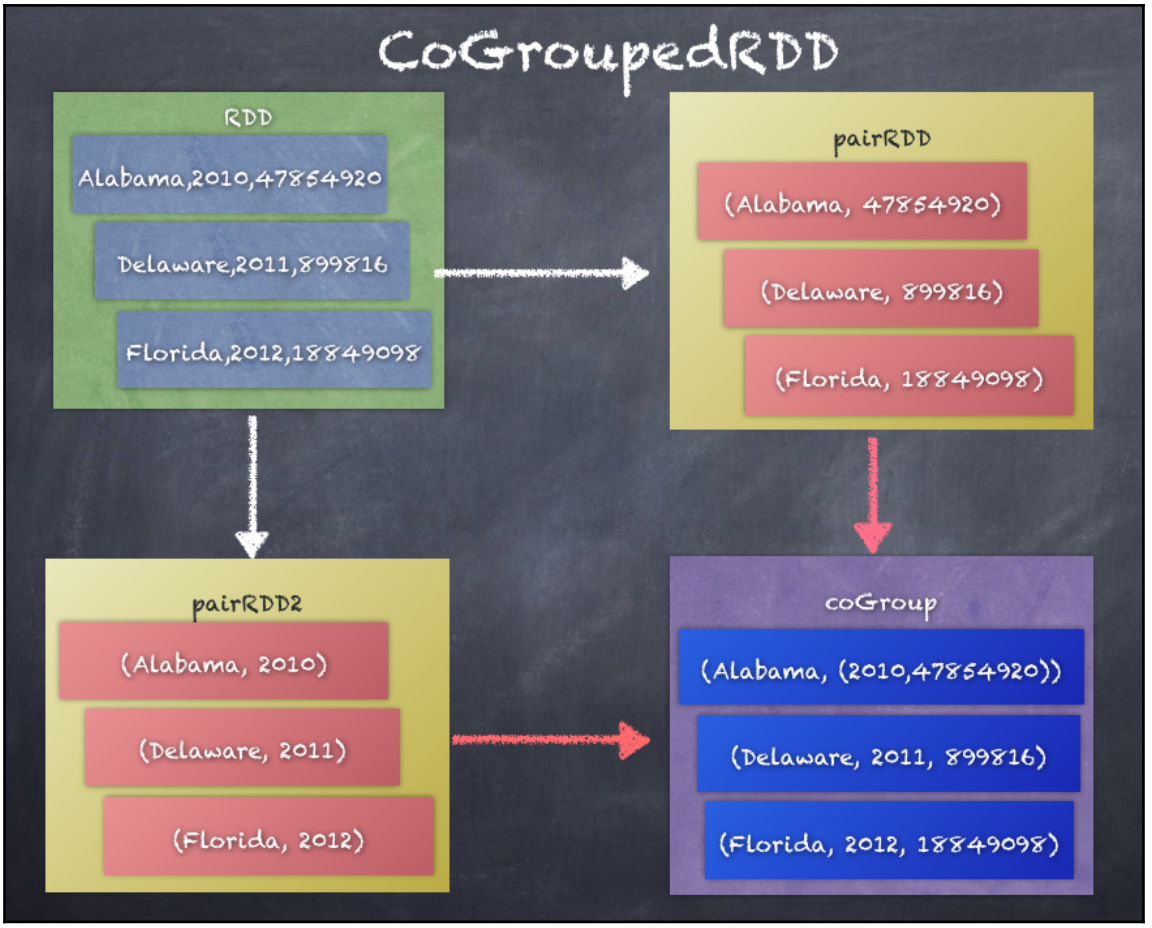

DAG Visualization
Stage 60
parallelize
ParallelCollectionRDD [28]
parallelize at <console>:25
map
MapPartitionsRDD [29]
map at <console>:27
map
MapPartitionsRDD [30]
map at <console>:29
map
MapPartitionsRDD [33]
map at <console>:31

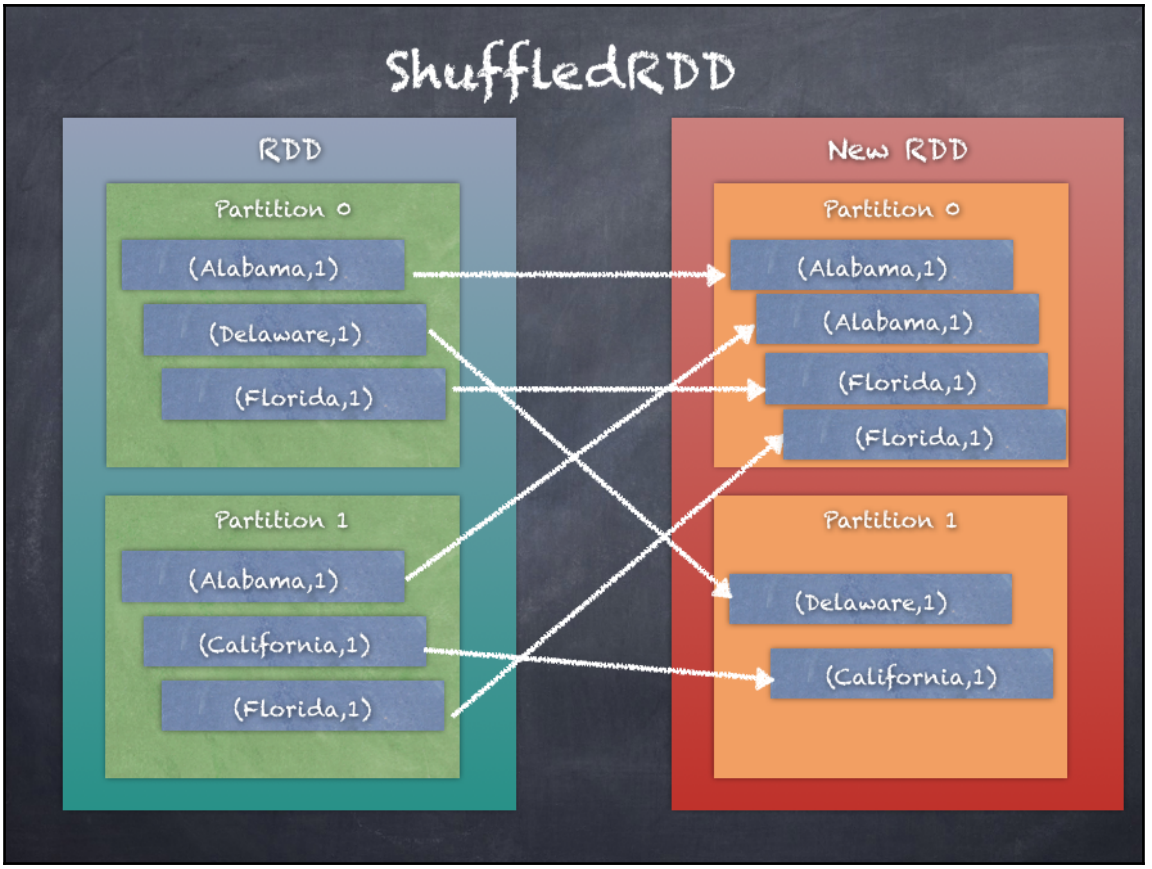

RDD
RDD
Alabama,2010,47854920
Delaware,2010,899816
Florida,2010,18849098
New RDD
ALABAMA,2010,47854920
DELAWARE,2010,899816
FLORIDA,2010,18849098

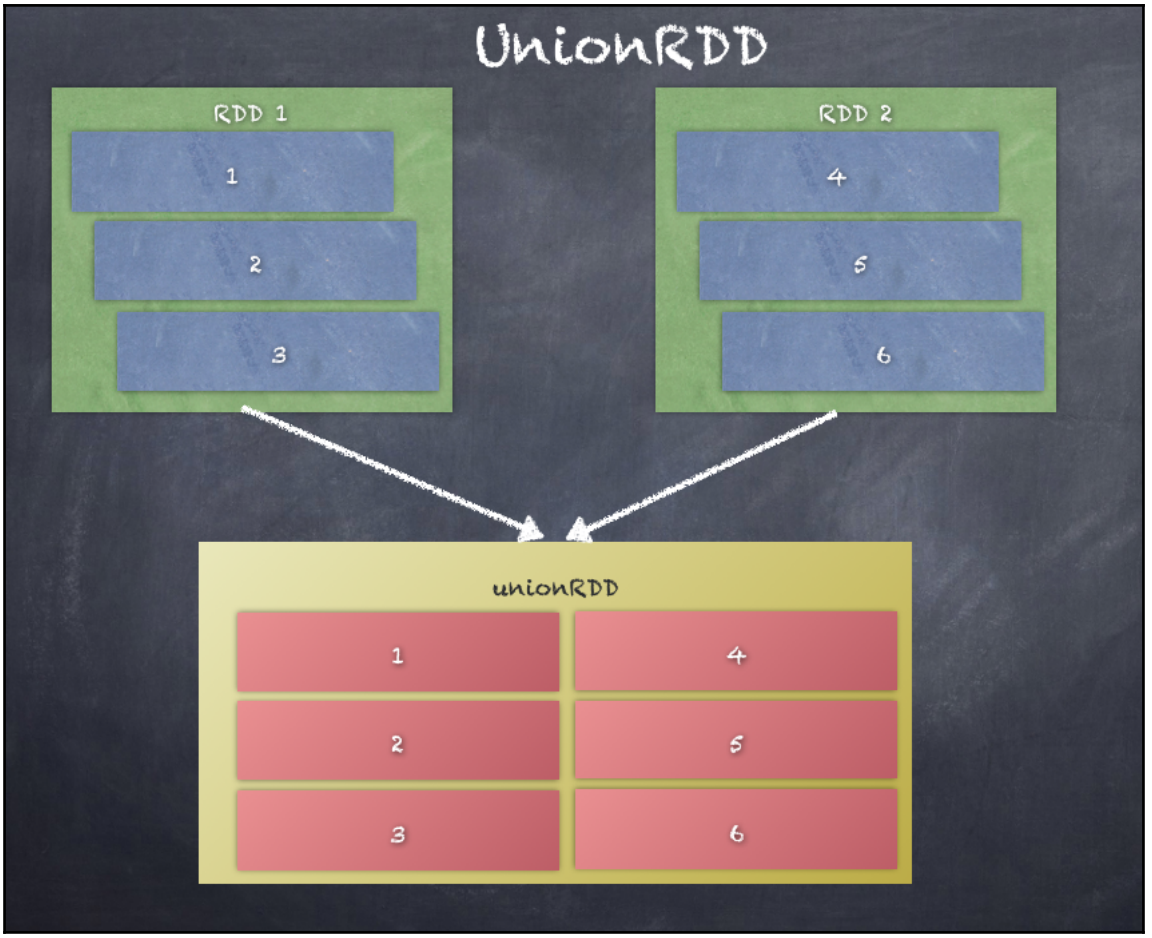

PairRDD
RDD
Alabama,2010,47854920
Delaware,2010,899816
Florida,2010,18849098
New RDD
(Alabama, 47854920)
(Delaware, 899816)
(Florida, 18849098)

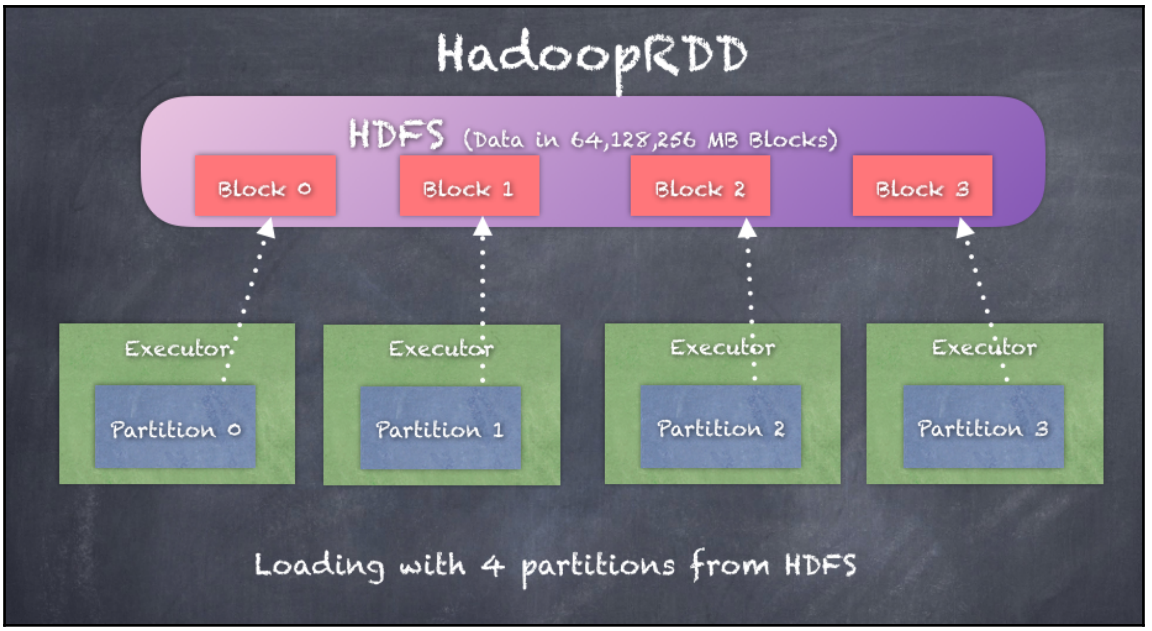

DoubleRDD
RDD
1.0
2.0
3.0
Driver
6.0
sum()

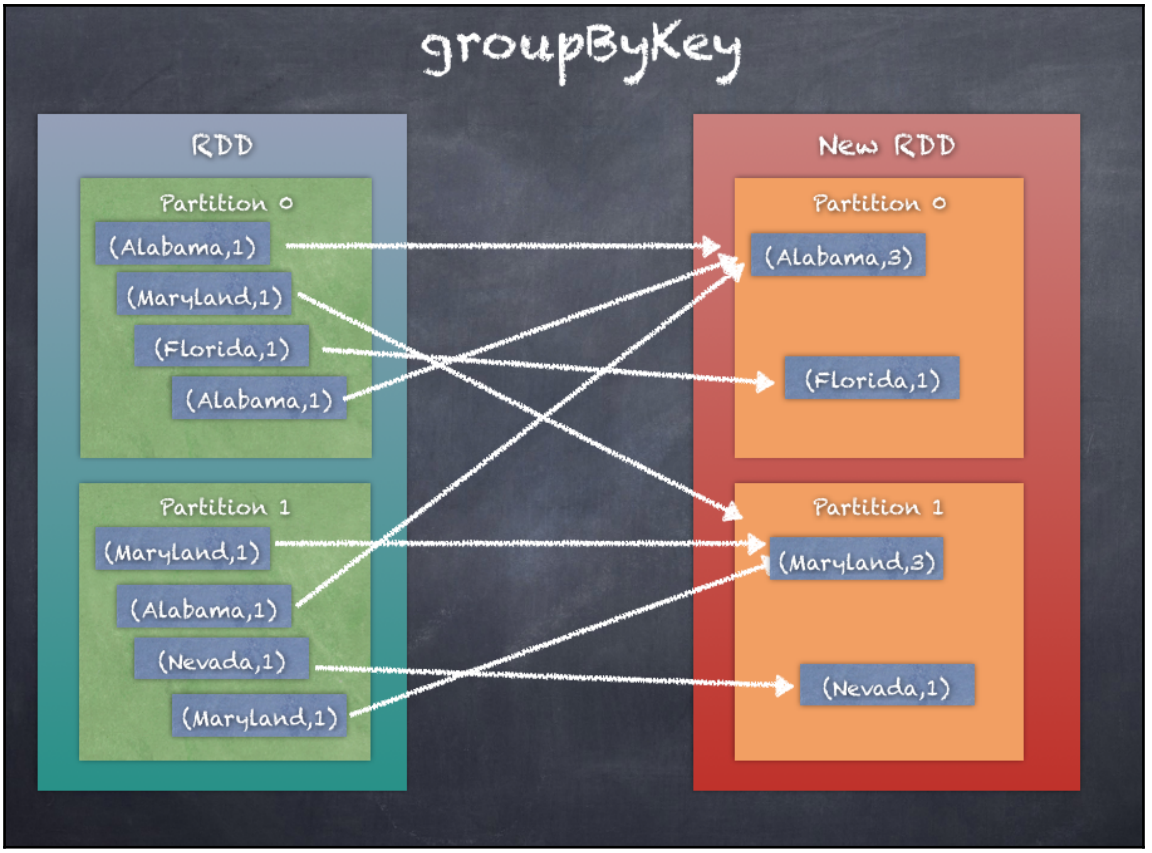

SequenceFileRDD
RDD
Alabama,2010,47854920
Delaware,2010,899816
Florida,2010,18849098
SequenceFileRDD [String, String]
(Alabama, 47854920)
(Delaware, 899816)
(Florida, 18849098)

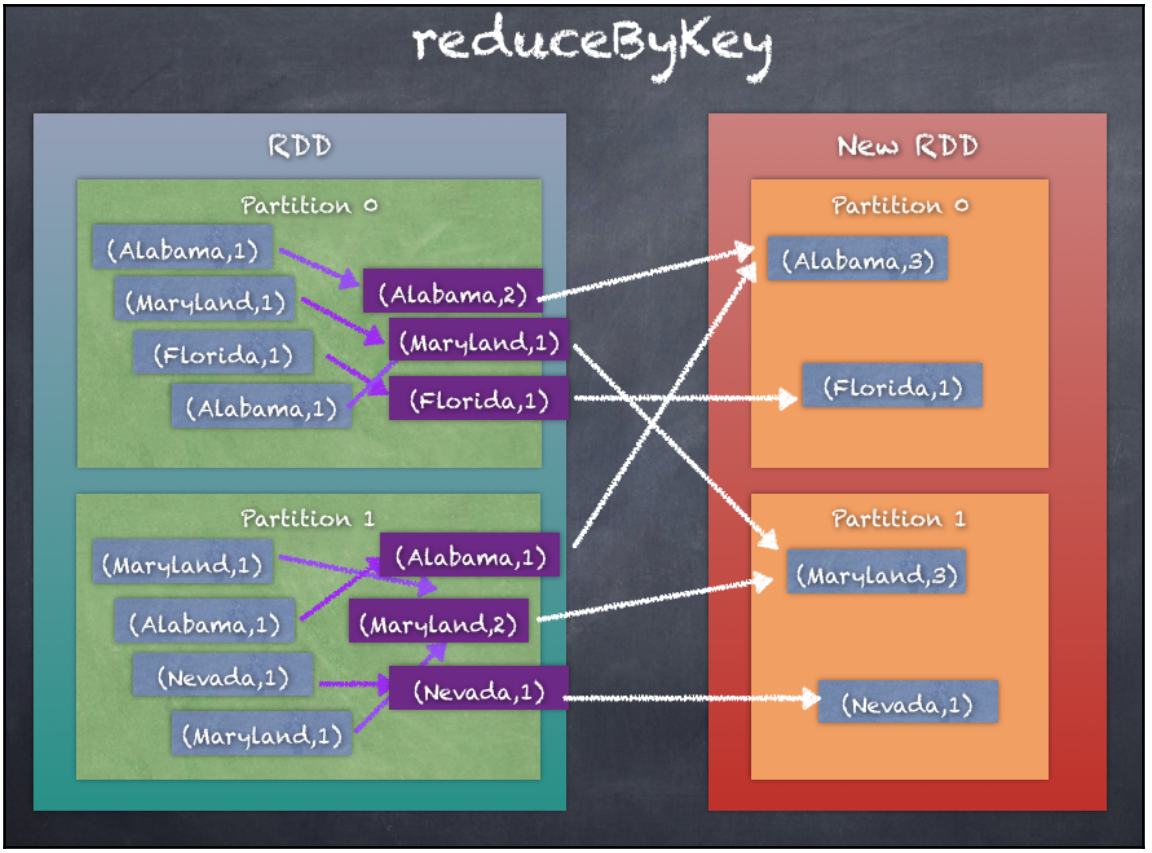

CoGroupedRDD
RDD
Alabama,2010,47854920
Delaware,2011,899816
Florida,2012,18849098
pairRDD
(Alabama, 47854920)
(Delaware, 899816)
(Florida, 18849098)
pairRDD2
(Alabama, 2010)
(Delaware, 2011)
(Florida, 2012)
coGroup
(Alabama, (2010,47854920))
(Delaware, 2011, 899816)
(Florida, 2012, 18849098)

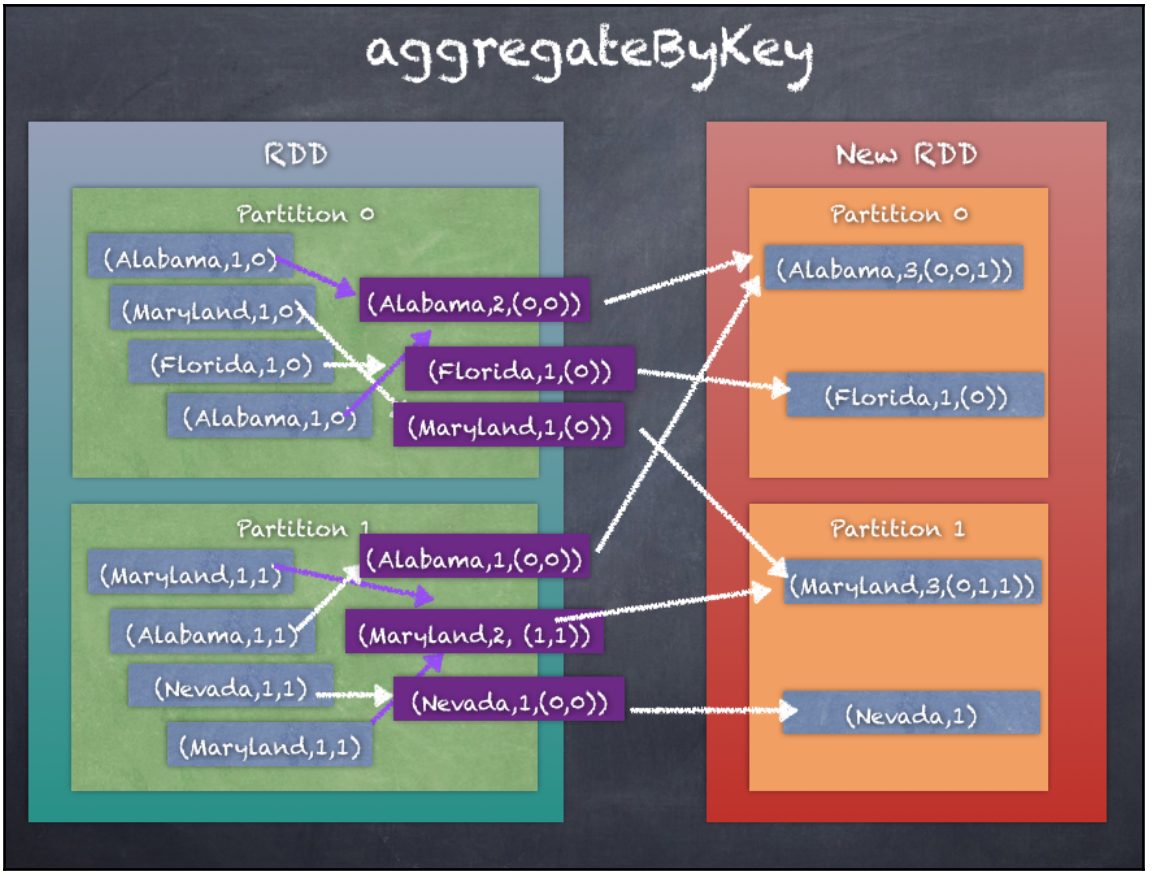

ShuffledRDD
RDD
Partition 0
(Alabama,1)
(Delaware,1)
(Florida,1)
Partition 1
(Alabama,1)
(California,1)
(Florida,1)
New RDD
Partition 0
(Alabama,1)
(Alabama,1)
(Florida,1)
(Florida,1)
Partition 1
(Delaware,1)
(California,1)

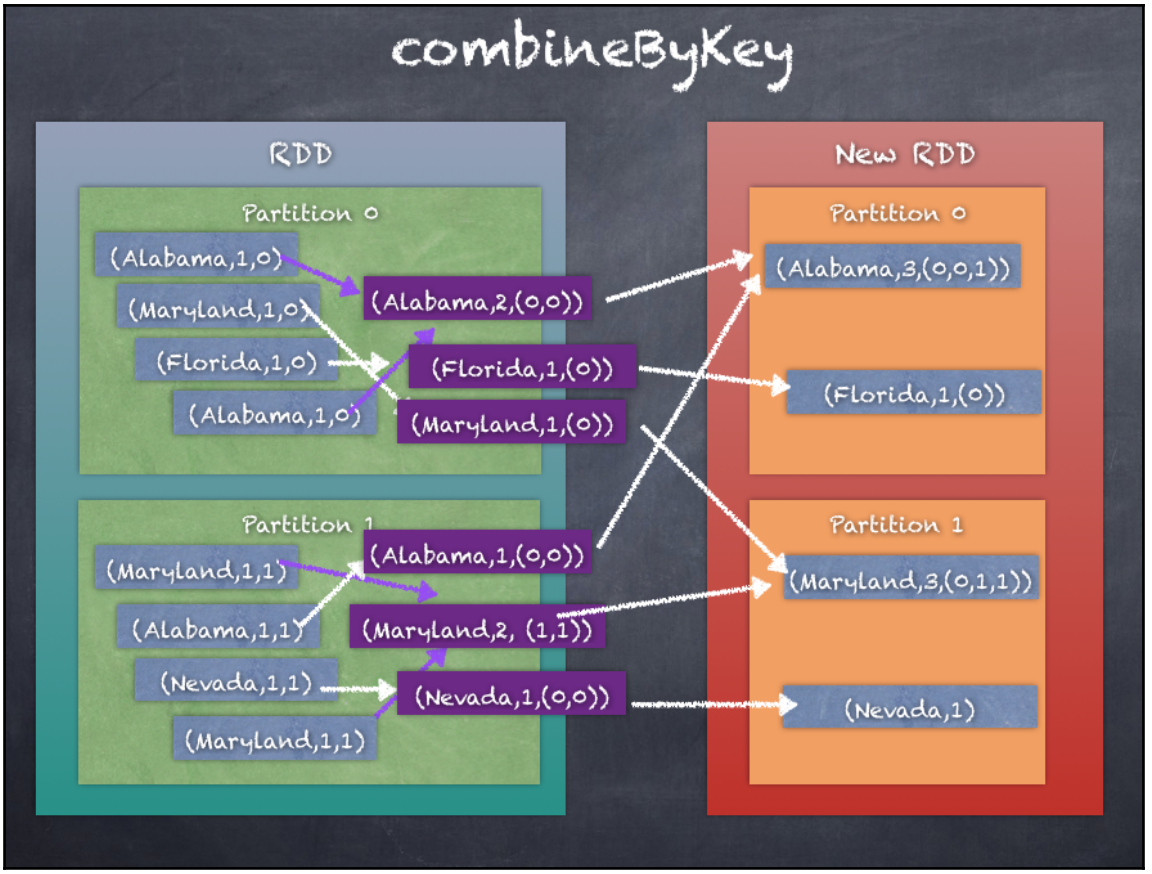

UnionRDD
RDD 1
1
2
3
RDD 2
4
5
6
unionRDD
1
4
2
5
3
6

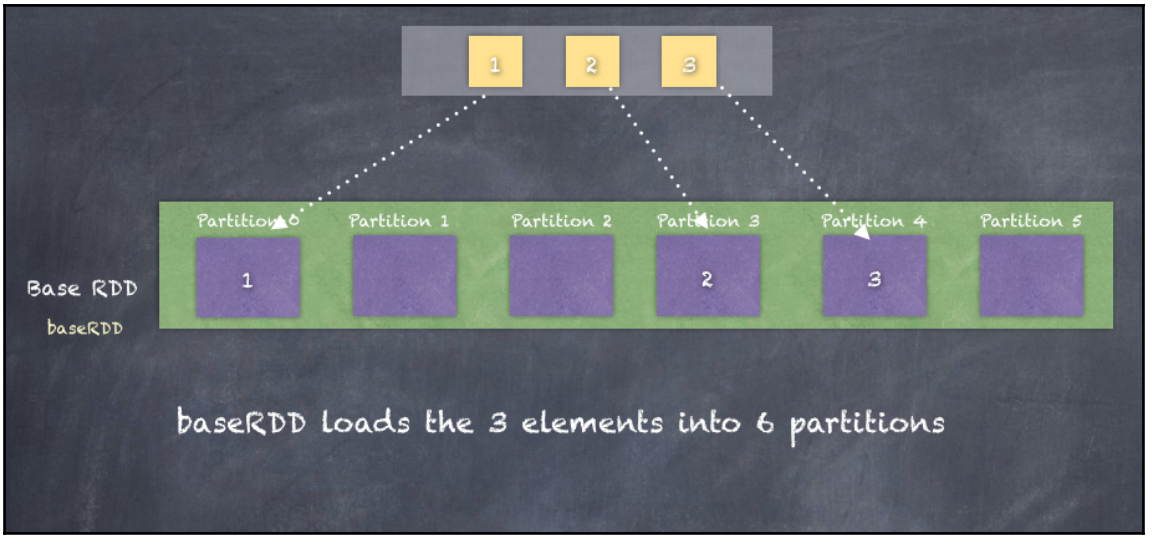

HadoopRDD
HDFS (Data in 64,128,256 MB Blocks)
Block 0
Block 1
Block 2
Block 3
Executor
Executor
Executor
Executor
Partition 0
Partition 1
Partition 2
Partition 3
Loading with 4 partitions from HDFS

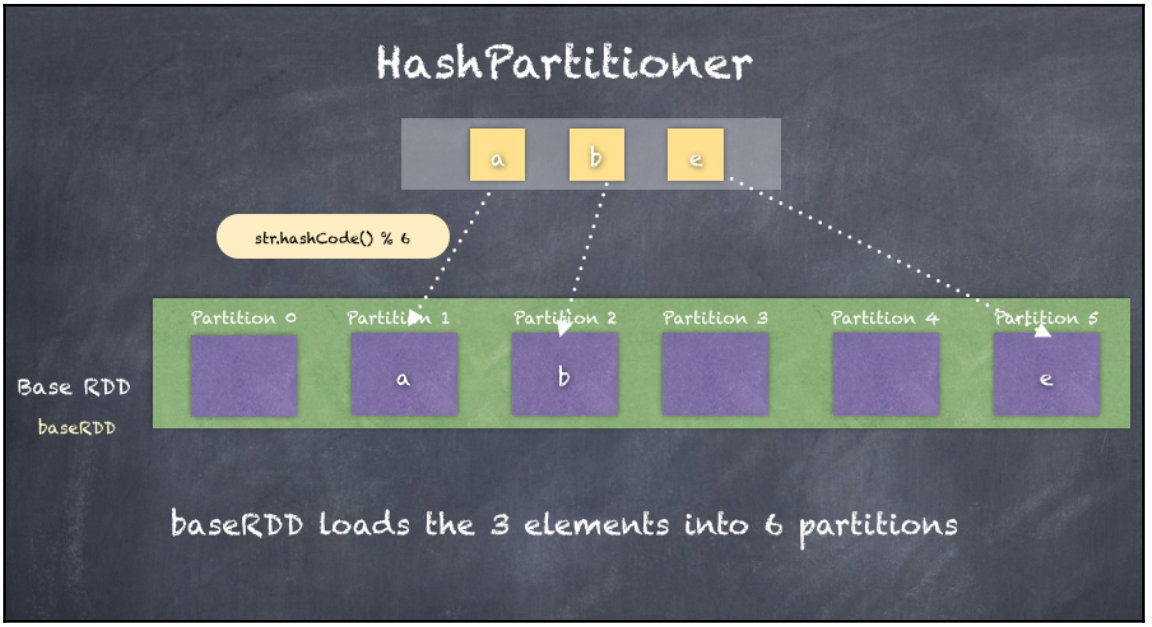

groupByKey
RDD
Partition 0
(Alabama,1)
(Maryland,1)
(Florida,1)
(Alabama,1)
Partition 1
(Maryland,1)
(Alabama,1)
(Nevada,1)
(Maryland,1)
New RDD
Partition 0
(Alabama,3)
(Florida,1)
Partition 1
(Maryland,3)
(Nevada,1)

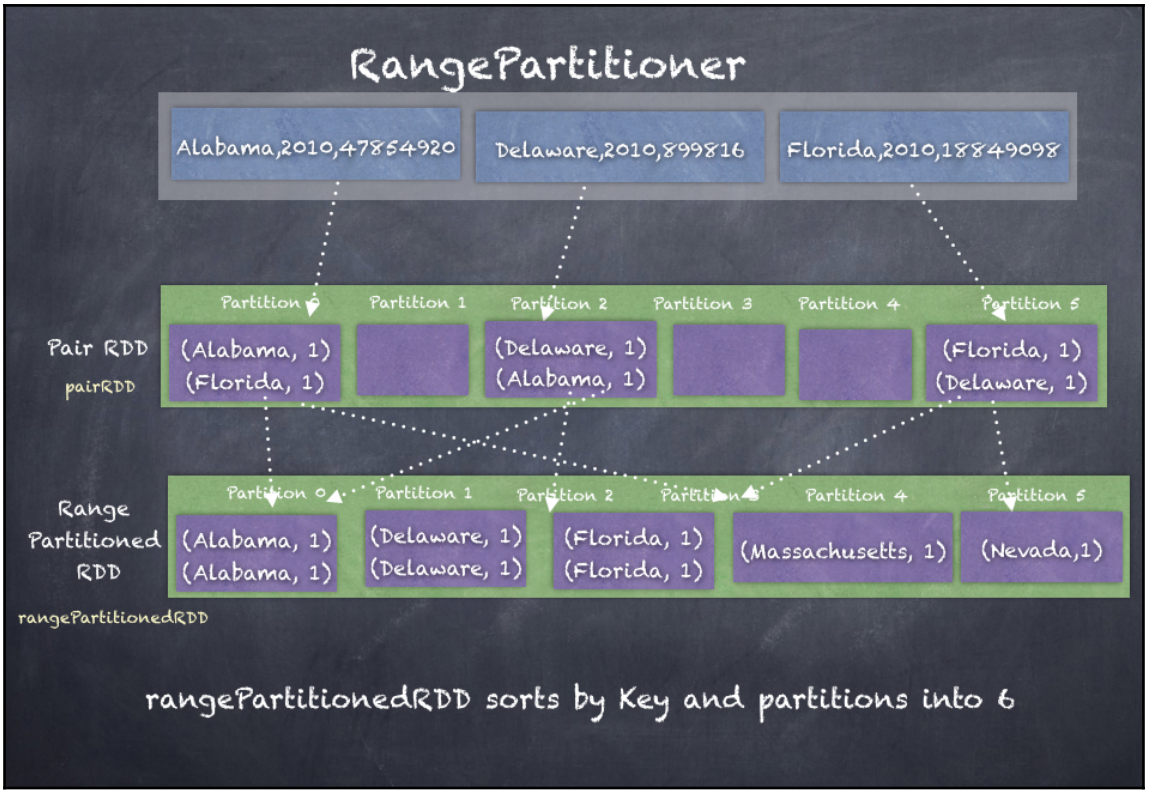

reduceByKey
RDD
Partition 0
(Alabama,1)
(Maryland,1)
(Florida,1)
(Alabama,1)
(Alabama,2)
(Maryland,1)
(Florida,1)
Partition 1
(Maryland,1)
(Alabama,1)
(Nevada,1)
(Maryland,1)
(Alabama,1)
(Maryland,2)
(Nevada,1)
New RDD
Partition 0
(Alabama,3)
(Florida,1)
Partition 1
(Maryland,3)
(Nevada,1)

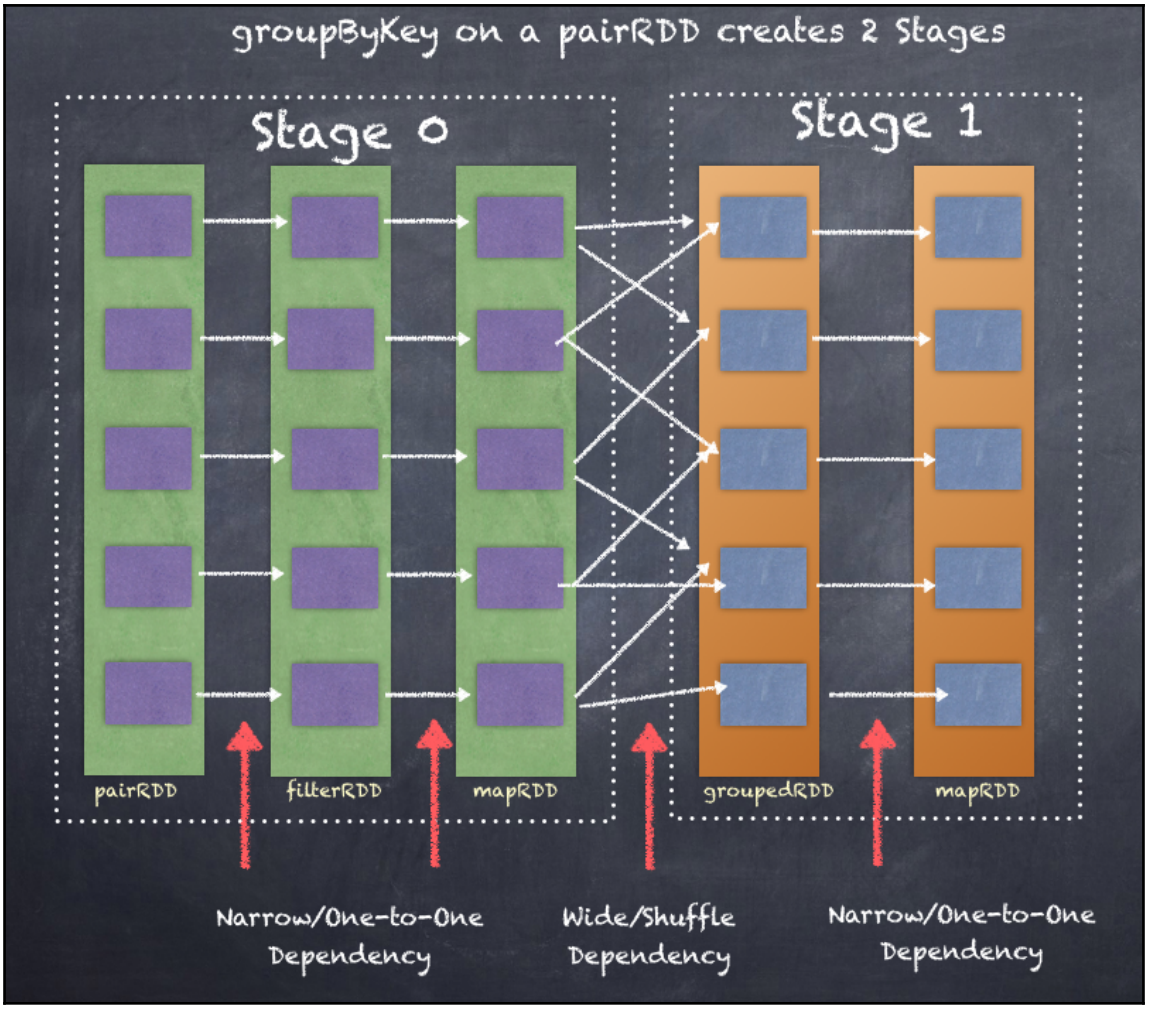

aggregateByKey
RDD
Partition 0
(Alabama,1,0)
(Maryland,1,0)
(Florida,1,0)
(Alabama,1,0)
(Alabama,2,(0,0))
(Florida,1,(0))
(Maryland,1,(0))
Partition 1
(Maryland,1,1)
(Alabama,1,1)
(Nevada,1,1)
(Maryland,1,1)
(Alabama,1,(0,0))
(Maryland,2, (1,1))
(Nevada,1,(0,0))
New RDD
Partition 0
(Alabama,3,(0,0,1))
(Florida,1,(0))
Partition 1
(Maryland,3,(0,1,1))
(Nevada,1)

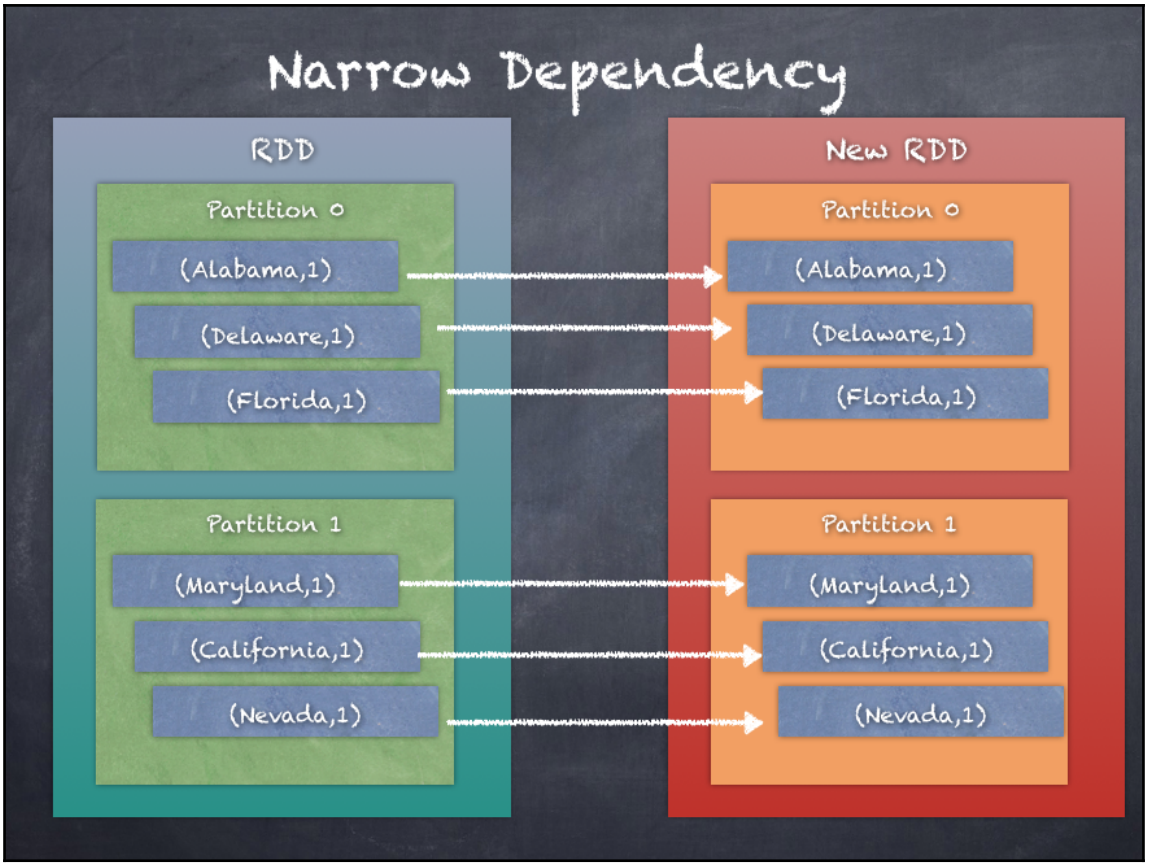

combineByKey
RDD
Partition 0
(Alabama,1,0)
(Maryland,1,0)
(Florida,1,0)
(Alabama,1,0)
(Alabama,2,(0,0))
(Florida,1,(0))
(Maryland,1,(0))
Partition 1
(Maryland,1,1)
(Alabama,1,1)
(Nevada,1,1)
(Maryland,1,1)
(Alabama,1,(0,0))
(Maryland,2, (1,1))
(Nevada,1,(0,0))
New RDD
Partition 0
(Alabama,3,(0,0,1))
(Florida,1,(0))
Partition 1
(Maryland,3,(0,1,1))
(Nevada,1)

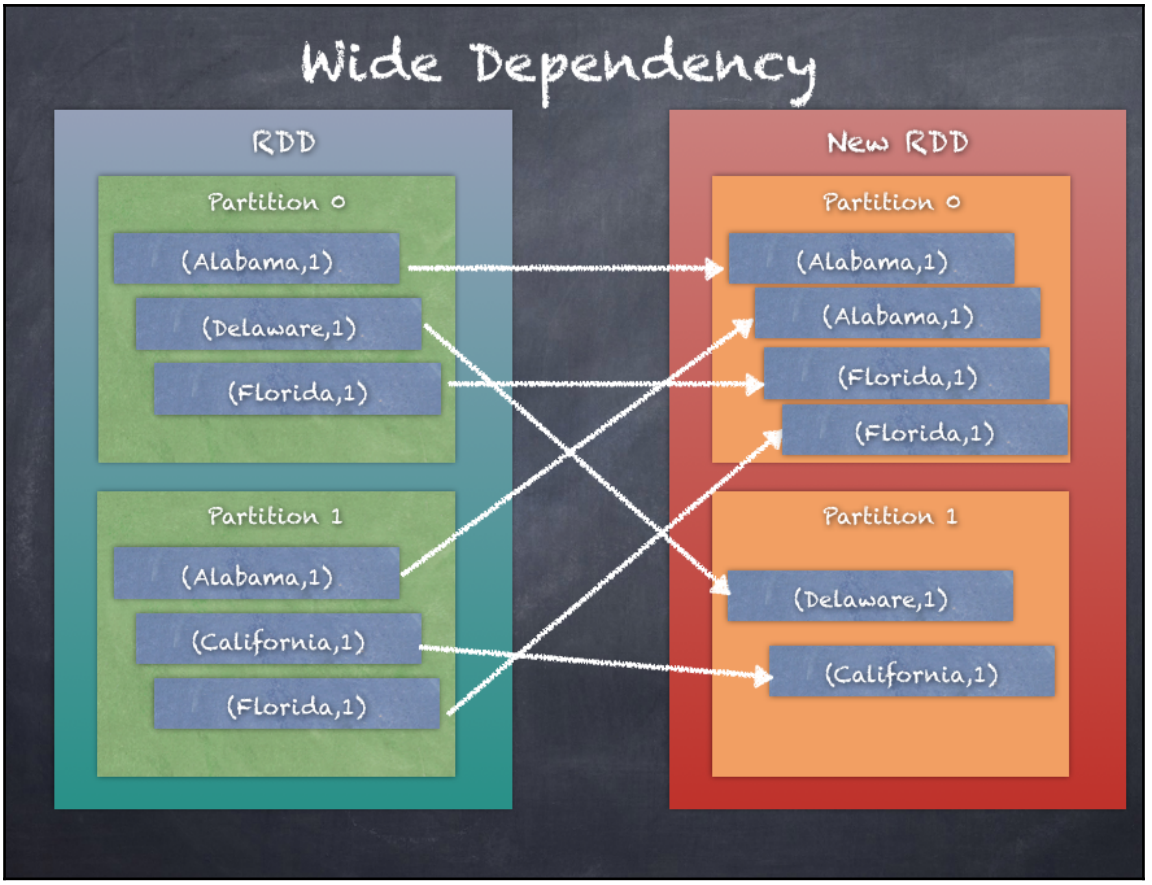

1
2
3
Partition 0
Partition 1
Partition 2
Partition 3
Partition 4
Partition 5
Base RDD
baseRDD
1
2
3
baseRDD loads the 3 elements into 6 partitions

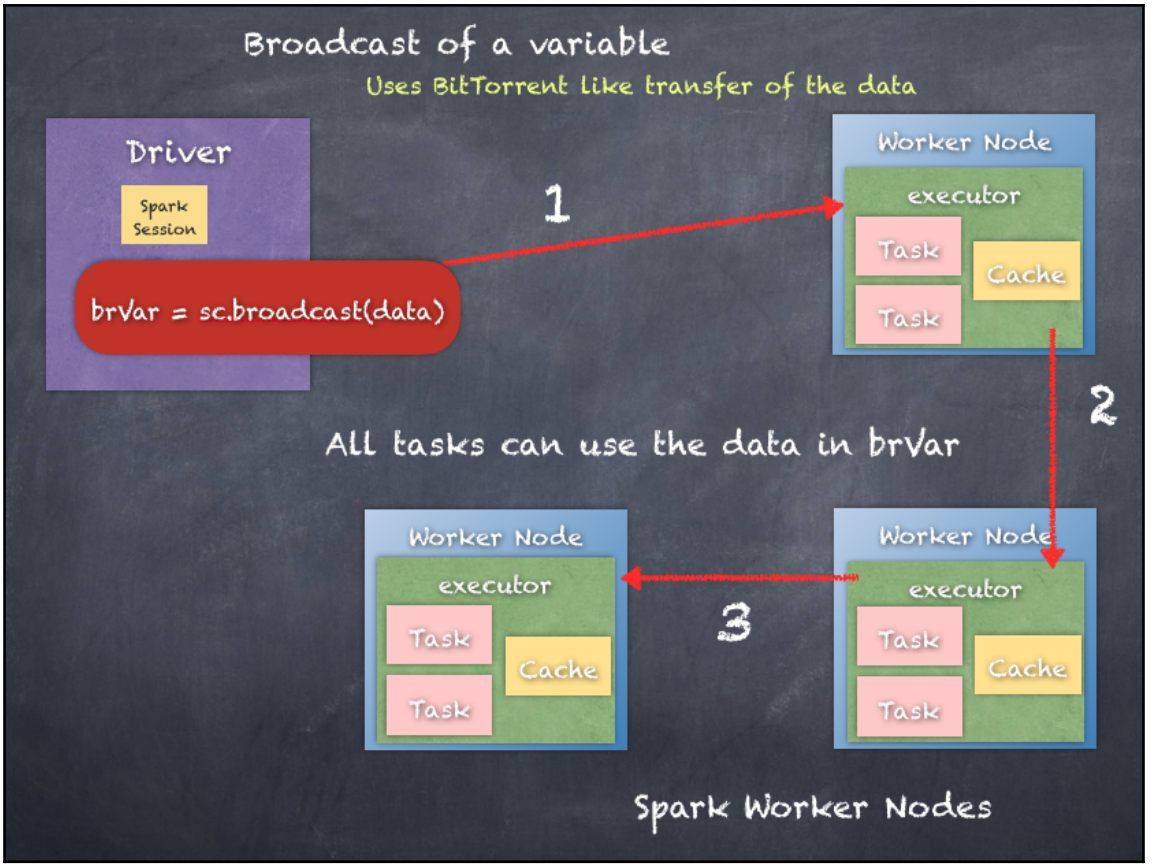

HashPartitioner
a
b
e
str.hashCode() % 6
Partition 0
Partition 1
Partition 2
Partition 3
Partition 4
Partition 5
Base RDD
baseRDD
a
b
e
baseRDD loads the 3 elements into 6 partitions


RangePartitioner
Alabama,2010,47854920
Delaware,2010,899816
Florida,2010,18849098
Pair RDD
pairRDD
Partition 0
Partition 1
Partition 2
Partition 3
Partition 4
Partition 5
(Alabama, 1)
(Florida, 1)
(Delaware, 1)
(Alabama, 1)
(Florida, 1)
(Delaware, 1)
Range Partitioned RDD
rangePartitionedRDD
Partition 0
Partition 1
Partition 2
Partition 3
Partition 4
Partition 5
(Alabama, 1)
(Alabama, 1)
(Delaware, 1)
(Delaware, 1)
(Florida, 1)
(Florida, 1)
(Massachusetts, 1)
(Nevada,1)
rangePartitionedRDD sorts by Key and partitions into 6

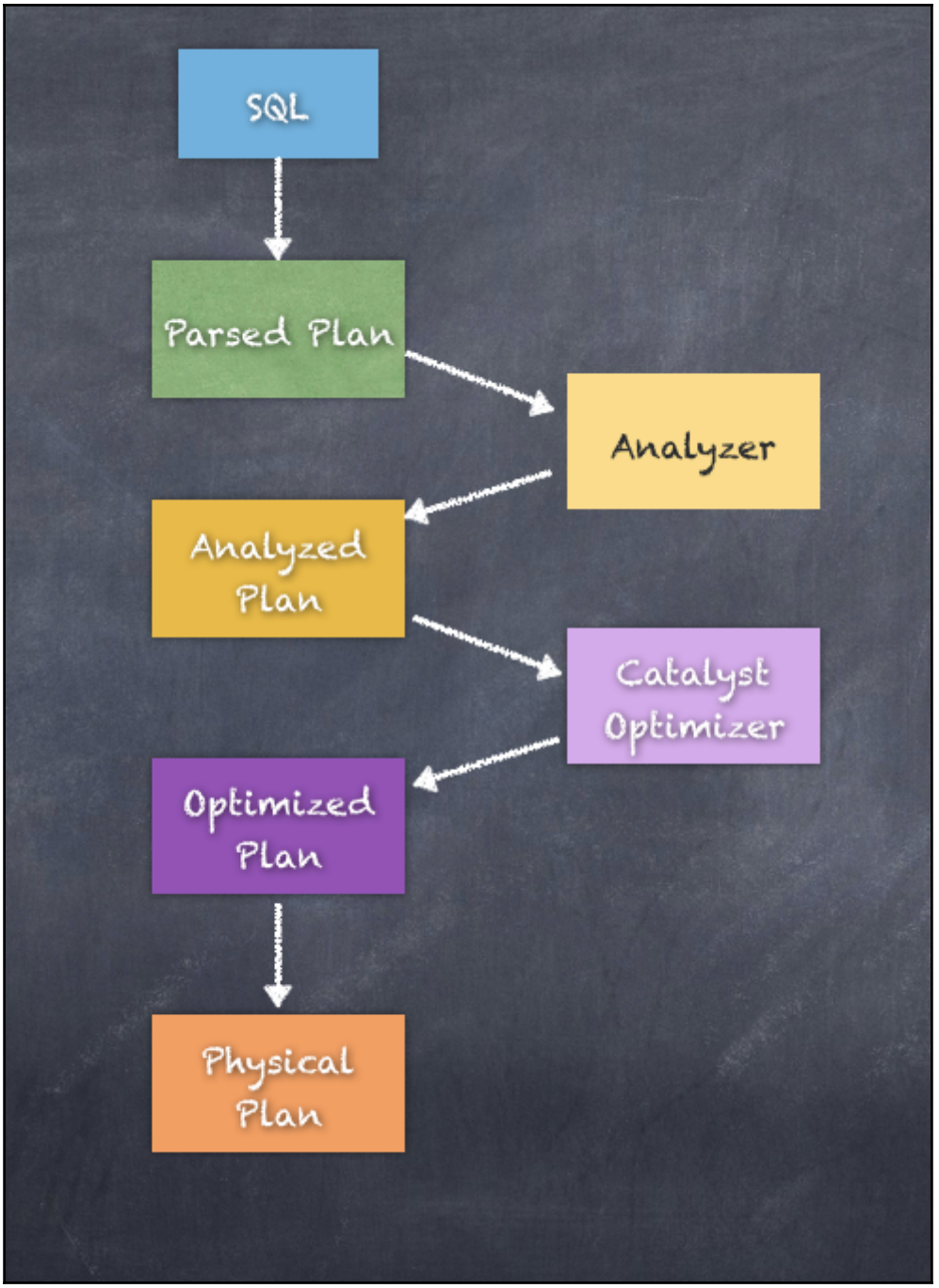

groupByKey on a pairRDD creates 2 Stages
Stage 0
Stage 1
pairRDD
filterRDD
mapRDD
groupedRDD
mapRDD
Narrow/One-to-One Dependency
Wide/Shuffle Dependency
Narrow/One-to-One Dependency


Narrow Dependency
RDD
Partition 0
(Alabama,1)
(Delaware,1)
(Florida,1)
Partition 1
(Maryland,1)
(California,1)
(Nevada,1)
New RDD
Partition 0
(Alabama,1)
(Delaware,1)
(Florida,1)
Partition 1
(Maryland,1)
(California,1)
(Nevada,1)

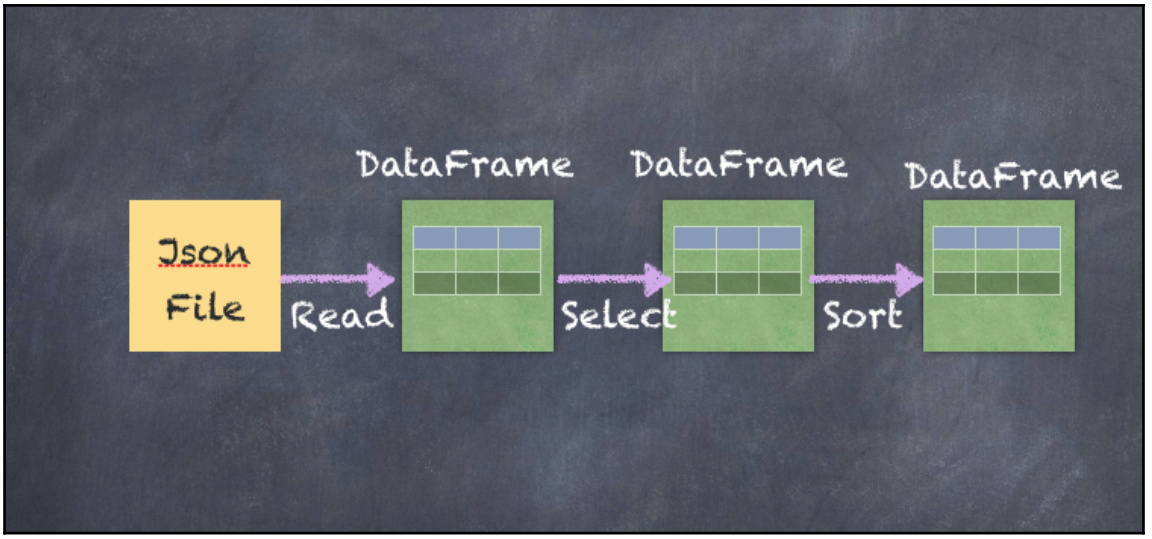

Wide Dependency
RDD
Partition 0
(Alabama,1)
(Delaware,1)
(Florida,1)
Partition 1
(Alabama,1)
(California,1)
(Florida,1)
New RDD
Partition 0
(Alabama,1)
(Alabama,1)
(Florida,1)
(Florida,1)
Partition 1
(Delaware,1)
(California,1)

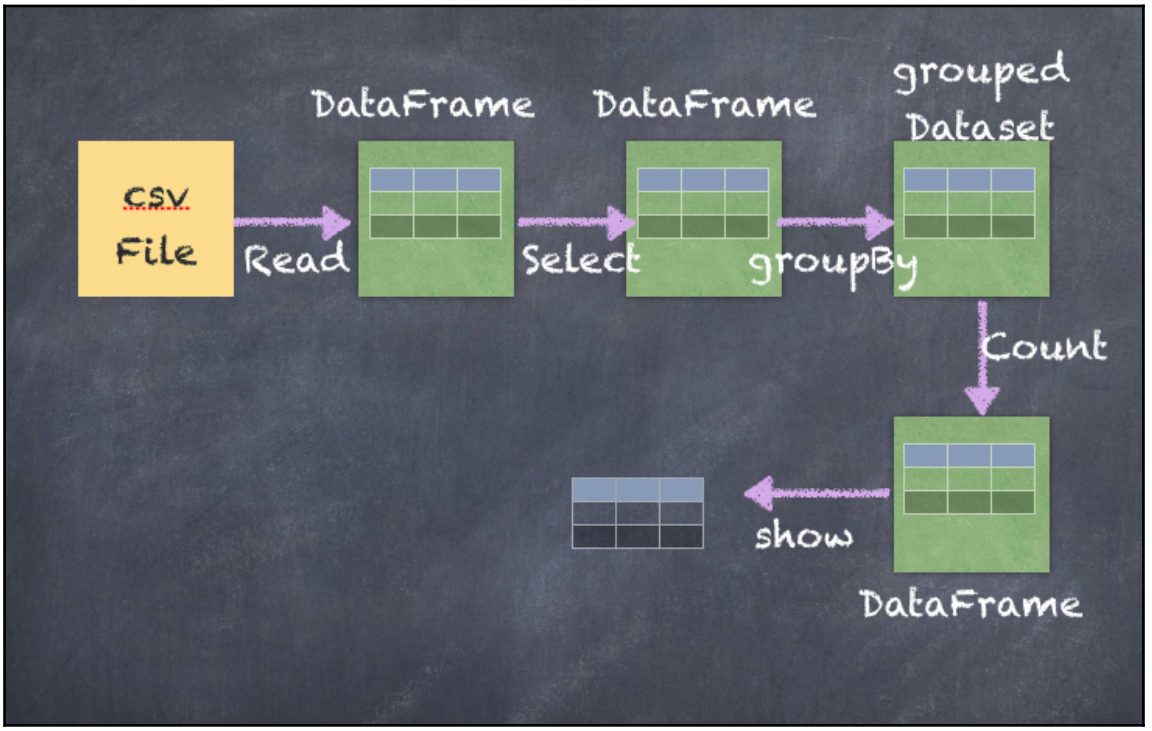

Broadcast of a variable
Uses BitTorrent like transfer of the data
Driver
Spark Session
brVar = sc.broadcast(data)
1
Worker Node
executor
Task
Cache
Task
2
All tasks can use the data in brVar
Worker Node
executor
Task
Cache
Task
3
Worker Node
executor
Task
Cache
Task
Spark Worker Nodes

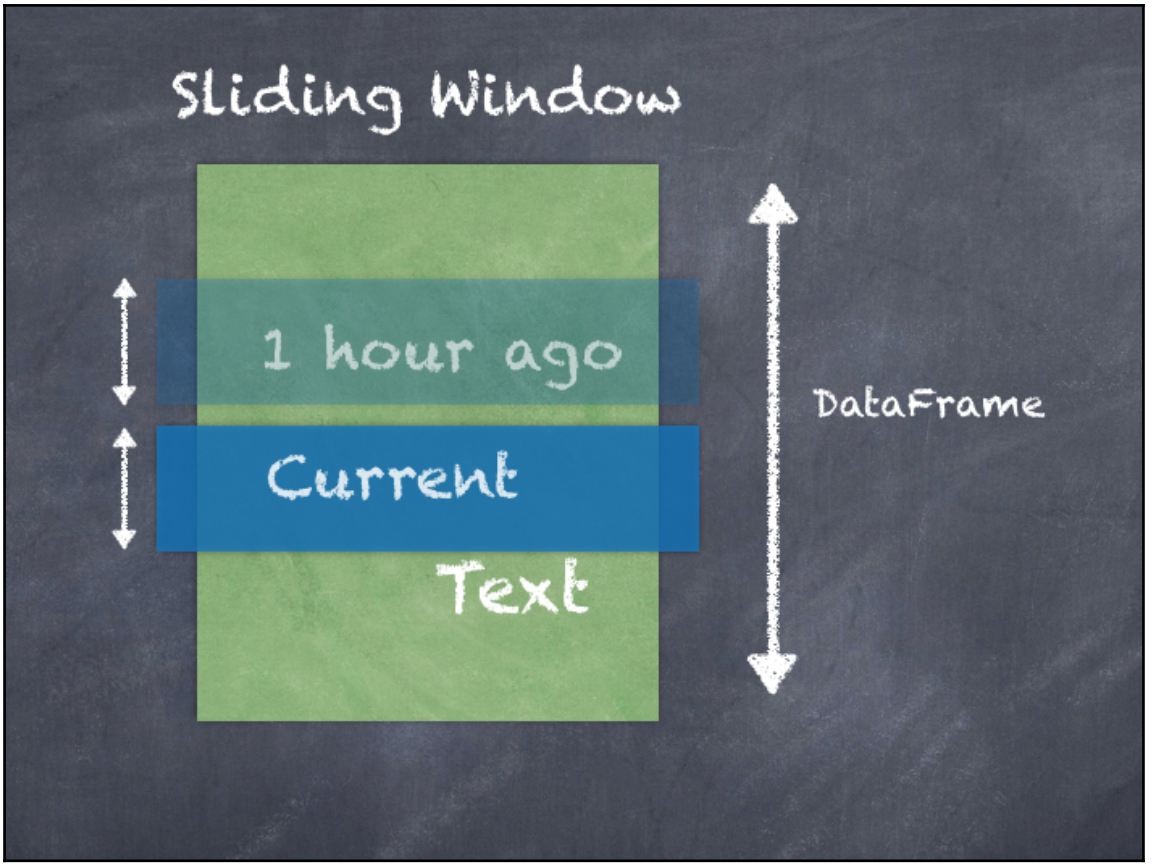
# Chapter 8: Introduce a Little Structure SparkSQL

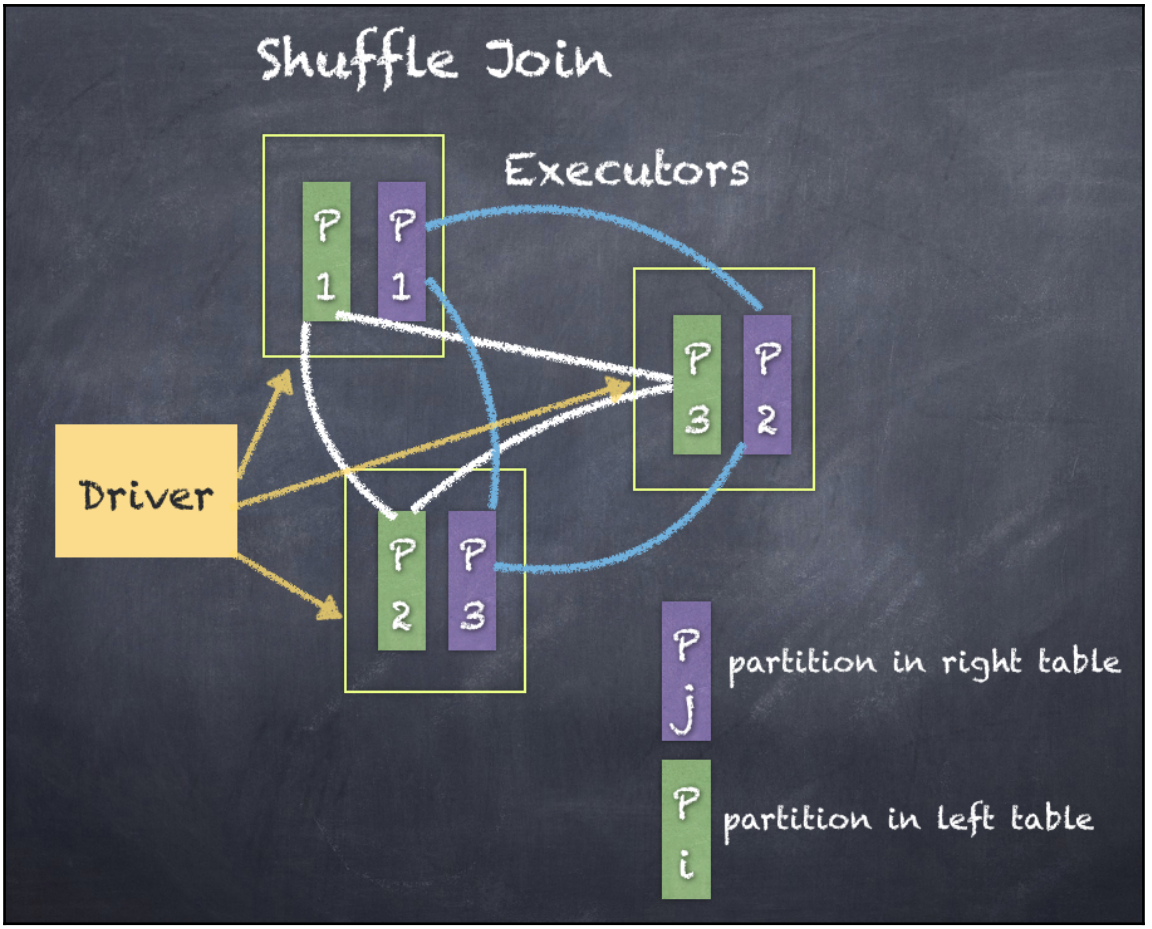

SQL
Parsed Plan
Analyzer
Analyzed Plan
Catalyst Optimizer
Optimized Plan
Physical Plan

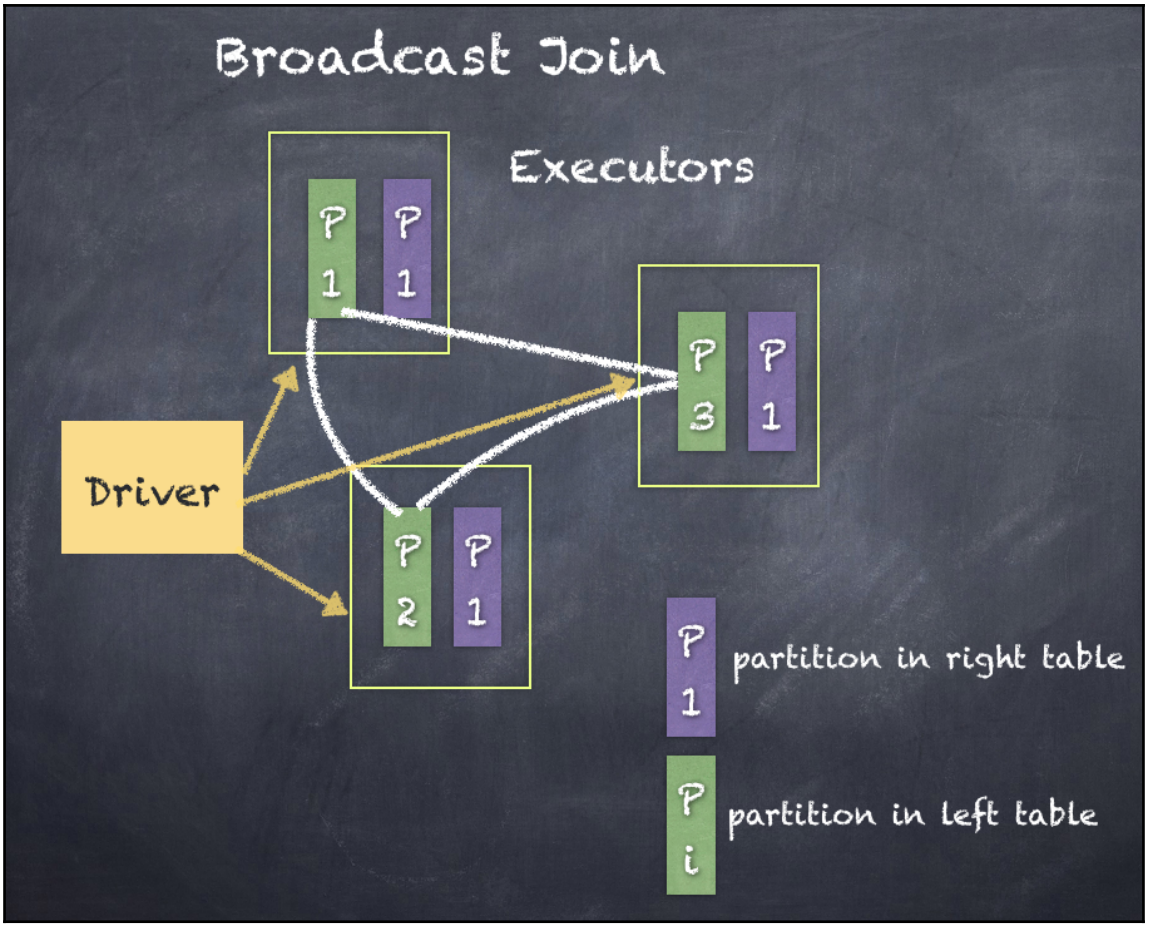

2 column table (id:Int, word:String)
(45, "hello")
0x0
45
16
5
hello
(64, "world")
0x0
64
16
5
world
filter by id > 50
(64, "world")
0x0
64
16
5
world
select first 3 chars of word
("wor")
0x0
12
3
for

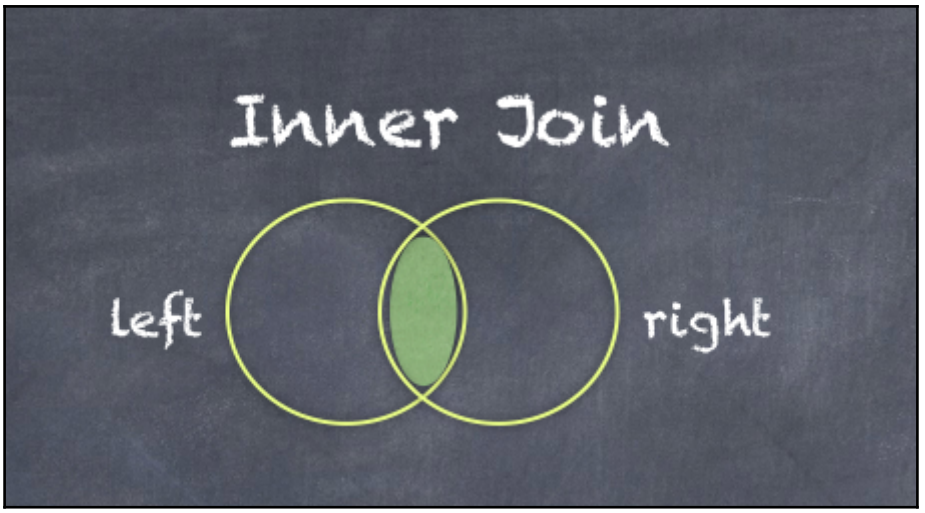

DataFrame
DataFrame
DataFrame
Json File
Read
Select
Sort

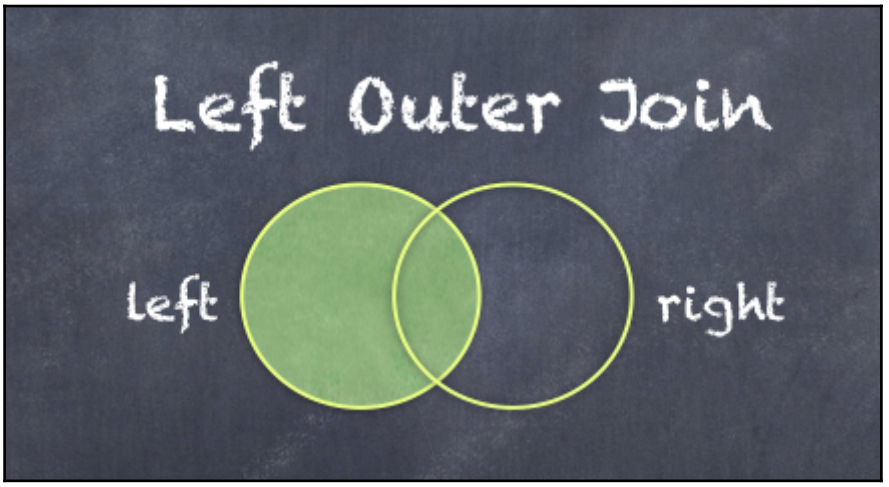

DataFrame
DataFrame
grouped Dataset
CSV File
Read
Select
groupBy
Count
show
DataFrame

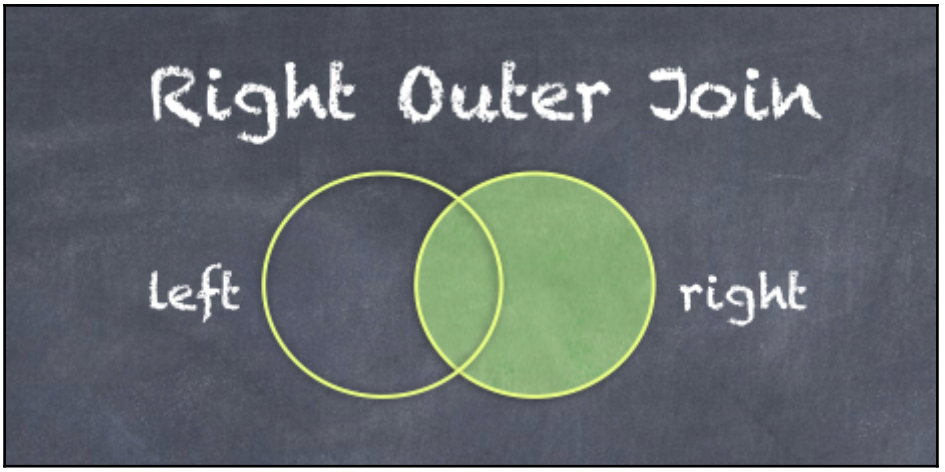

Sliding Window
1 hour ago
Current
Text
DataFrame

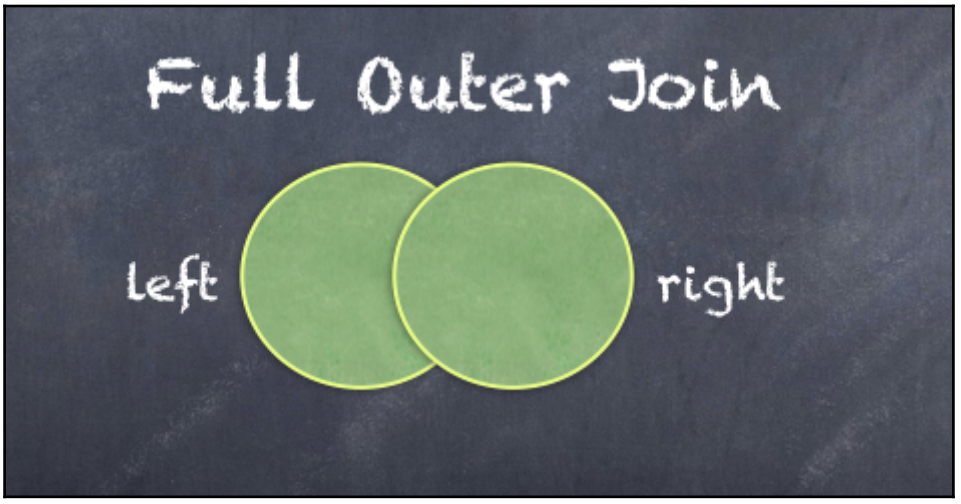

Shuffle Join
Executors
P 1
P 1
P 3
P 2
P 2
P 3
Driver
P j
partition in right table
P i
partition in left table

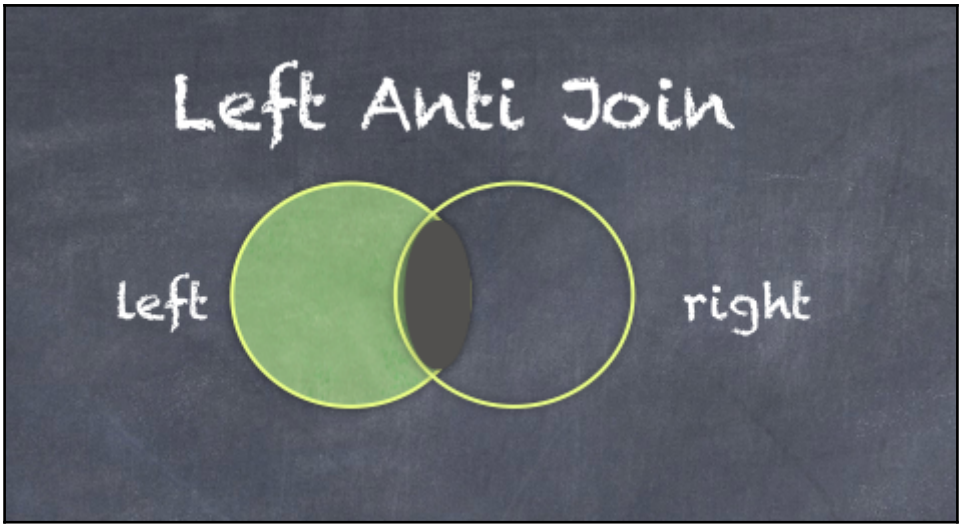

Broadcast Join
Executors
P 1
P 1
P 3
P 1
Driver
P 2
P 1
P 1
partition in right table
P i
partition in left table

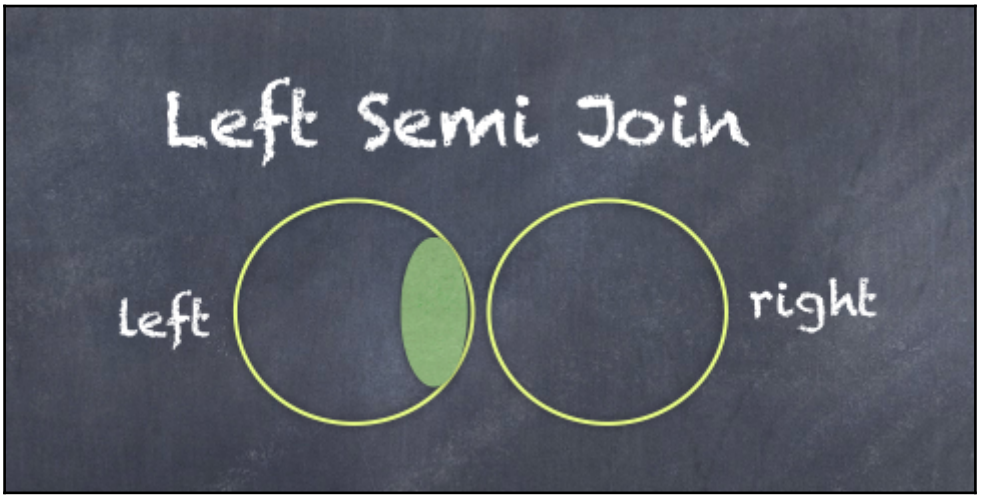

Inner Join
left
right

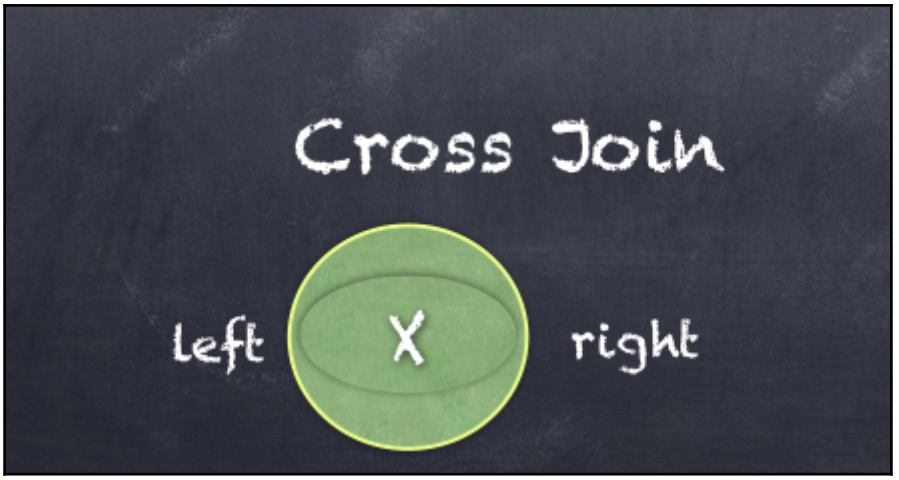

Left Outer Join
left
right


Right Outer Join
left
right

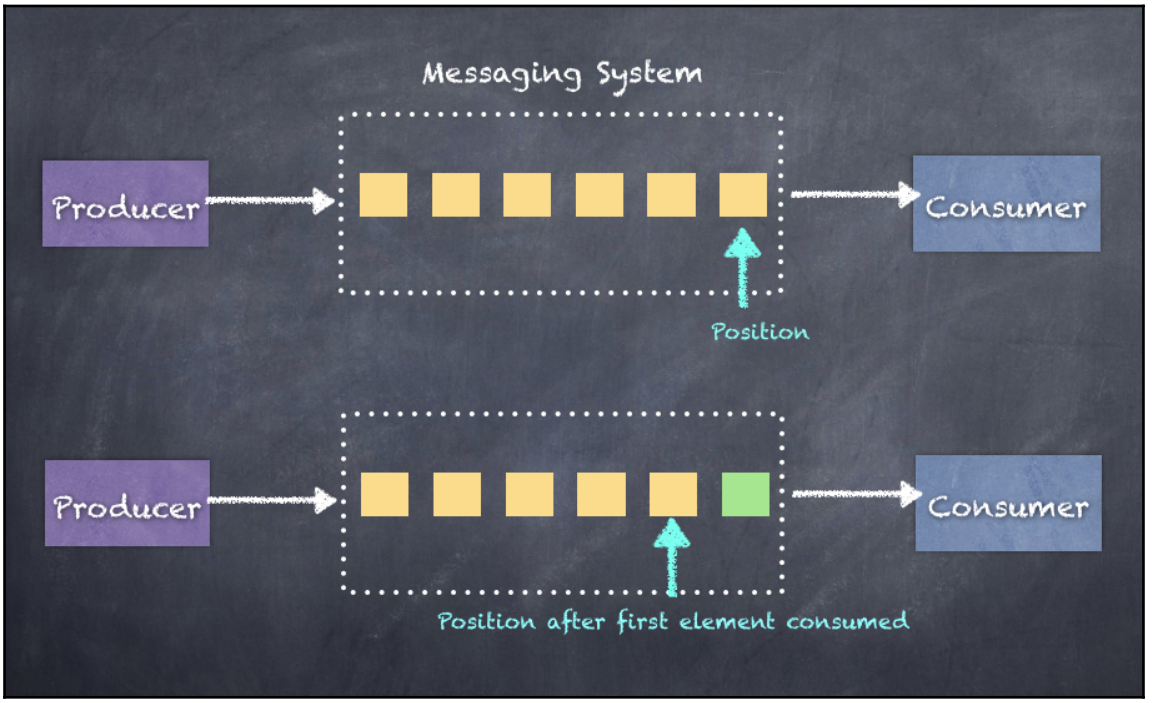

Full Outer Join
left
right

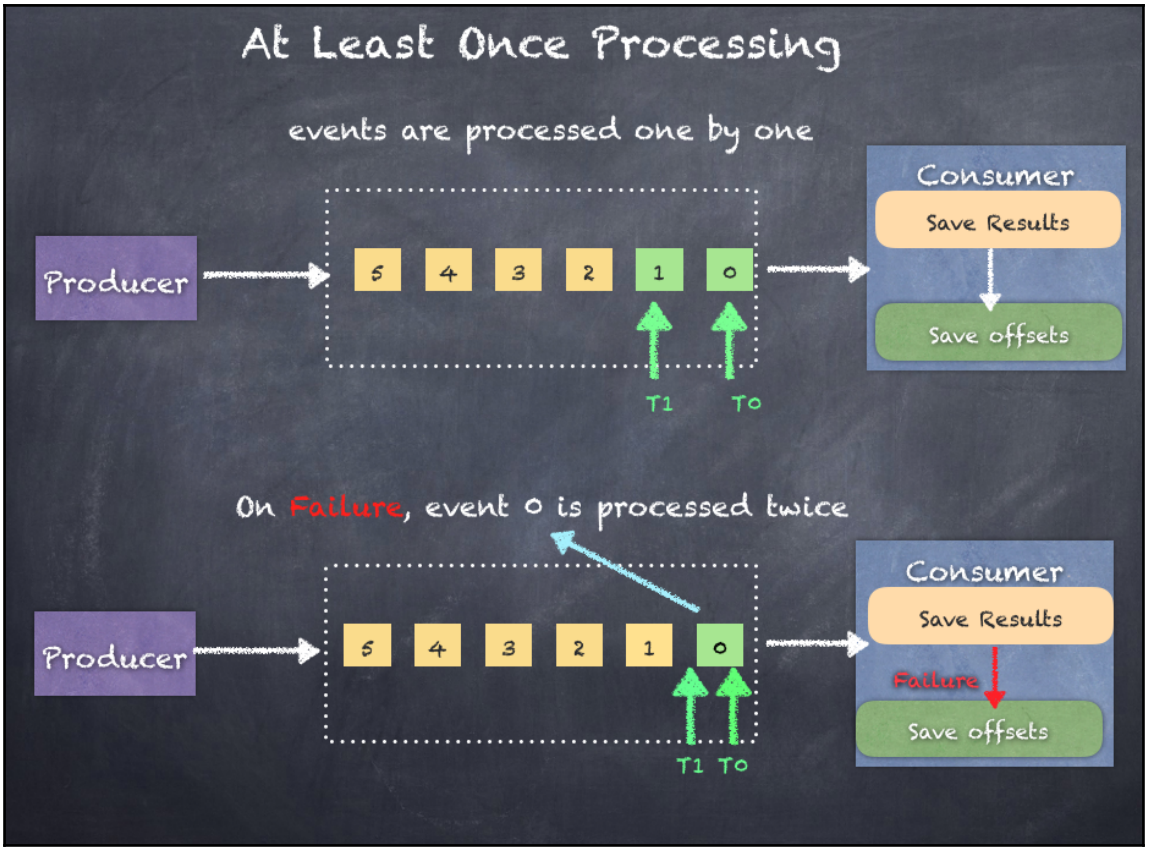

Left Anti Join
left
right

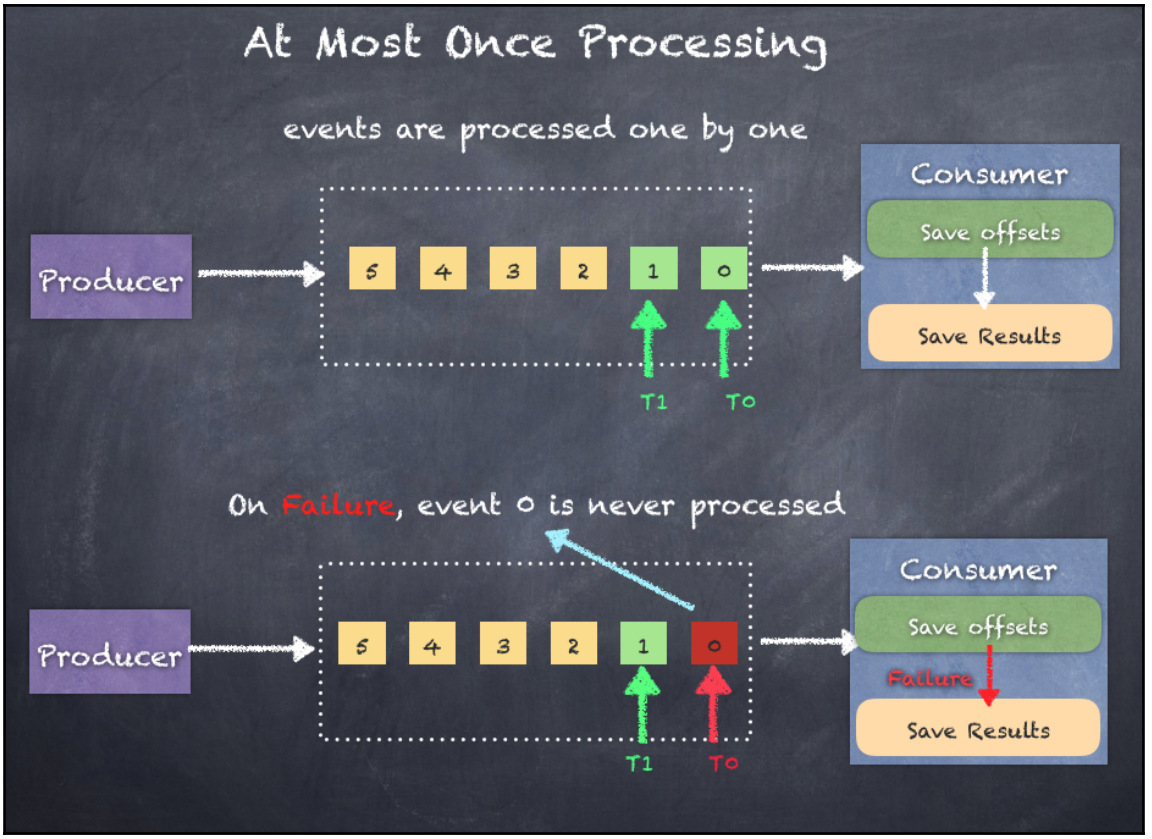

Left Semi Join
left
right

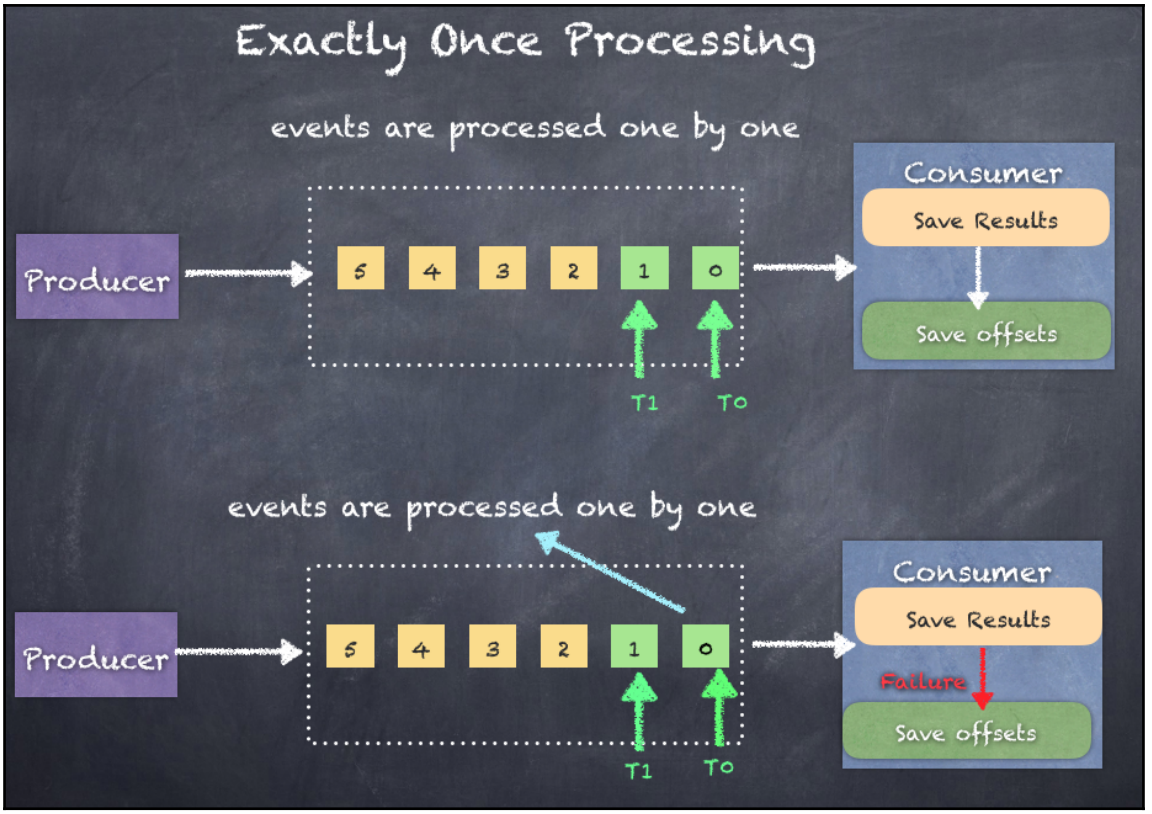

Cross Join
left
X
right

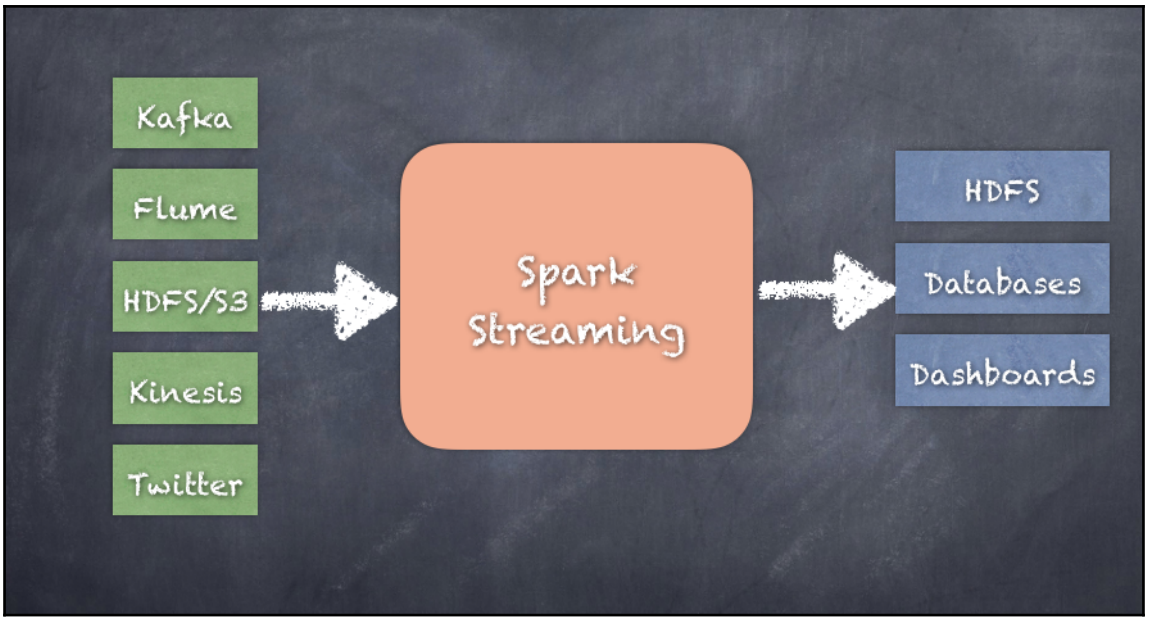
# Chapter 9: Stream Me Up Scotty - Spark Streaming

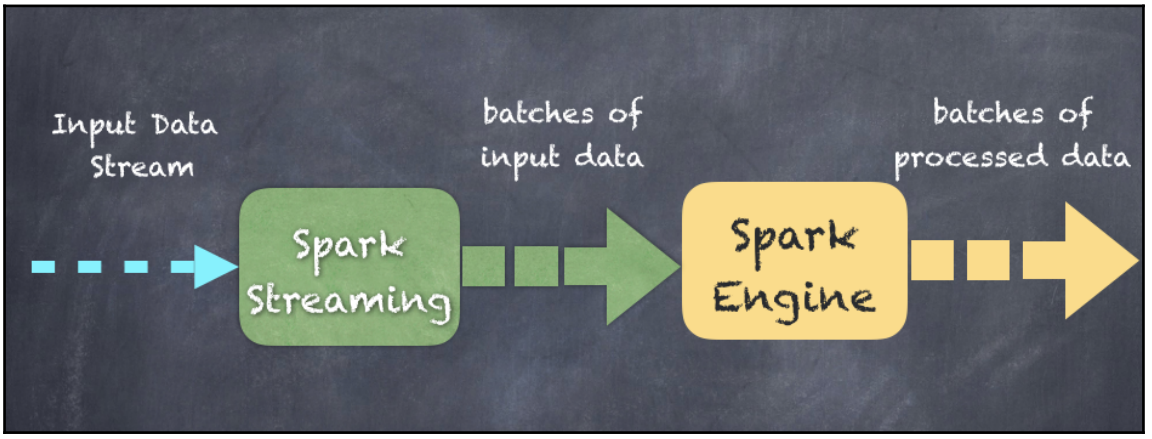


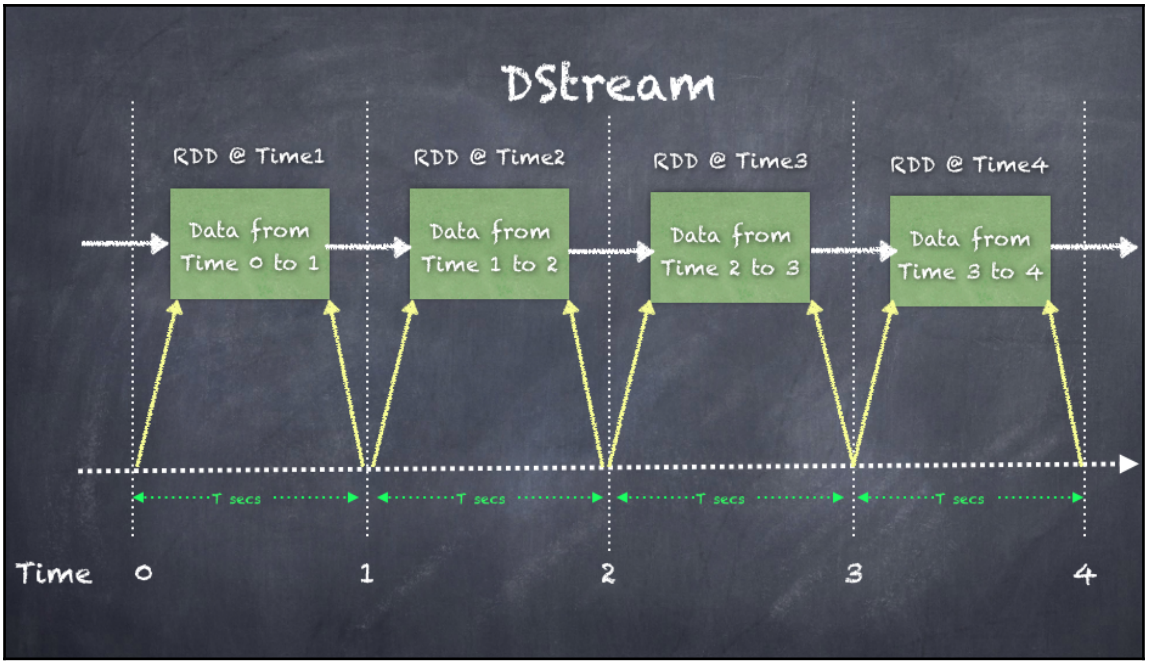

At Least Once Processing
events are processed one by one
Producer
5
4
3
2
1
0
T1
T0
Consumer
Save Results
Save offsets
On Failure, event 0 is processed twice
Producer
5
4
3
2
1
0
T1
T0
Consumer
Save Results
Failure
Save offsets

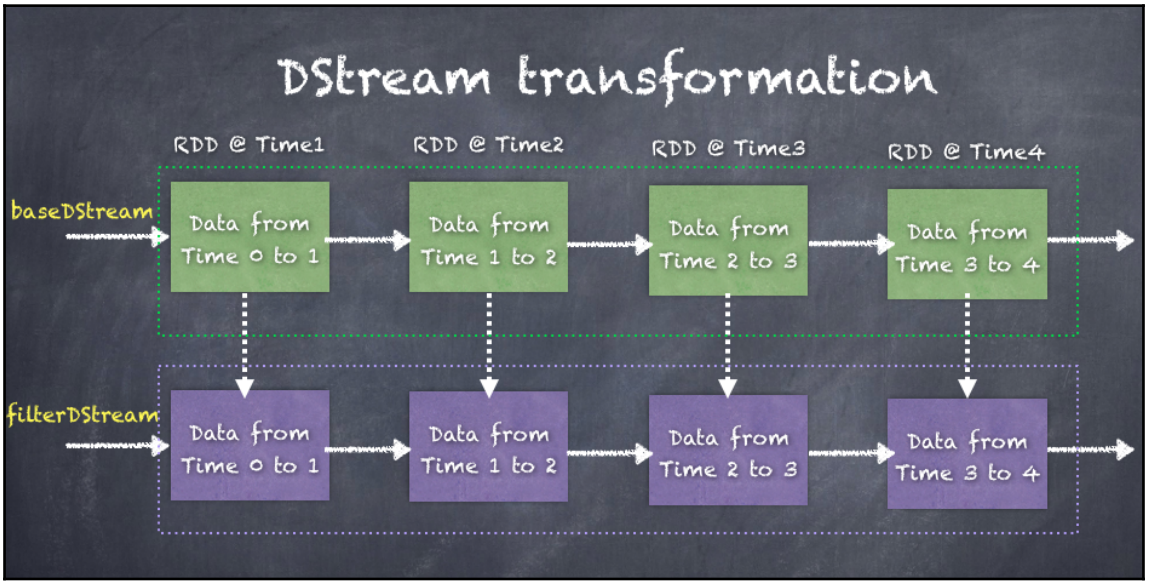

At Most Once Processing
events are processed one by one
Producer
5
4
3
2
1
0
T1
T0
Consumer
Save offsets
Save Results
On Failure, event 0 is never processed
Producer
5
4
3
2
1
0
T1
T0
Consumer
Save offsets
Failure
Save Results

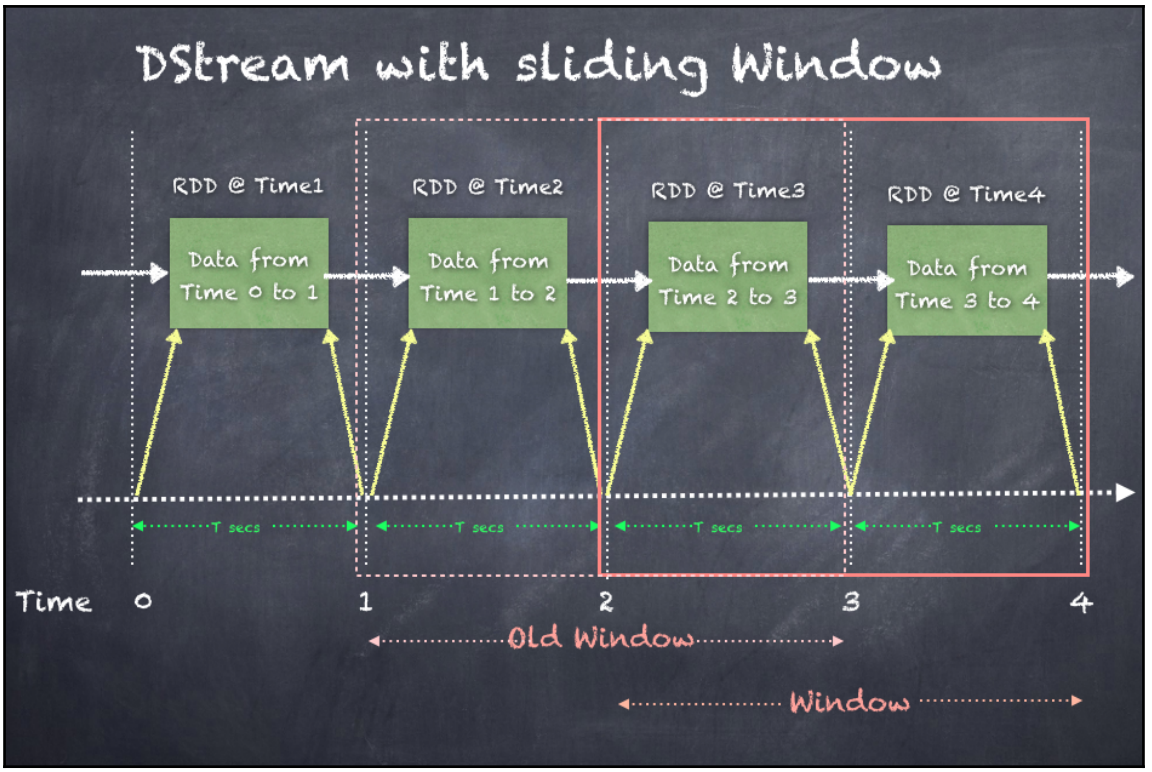

Exactly Once Processing
events are processed one by one
Producer
5
4
3
2
1
0
T1
T0
Consumer
Save Results
Save offsets
events are processed one by one
Producer
5
4
3
2
1
0
T1
T0
Consumer
Save Results
Failure
Save offsets

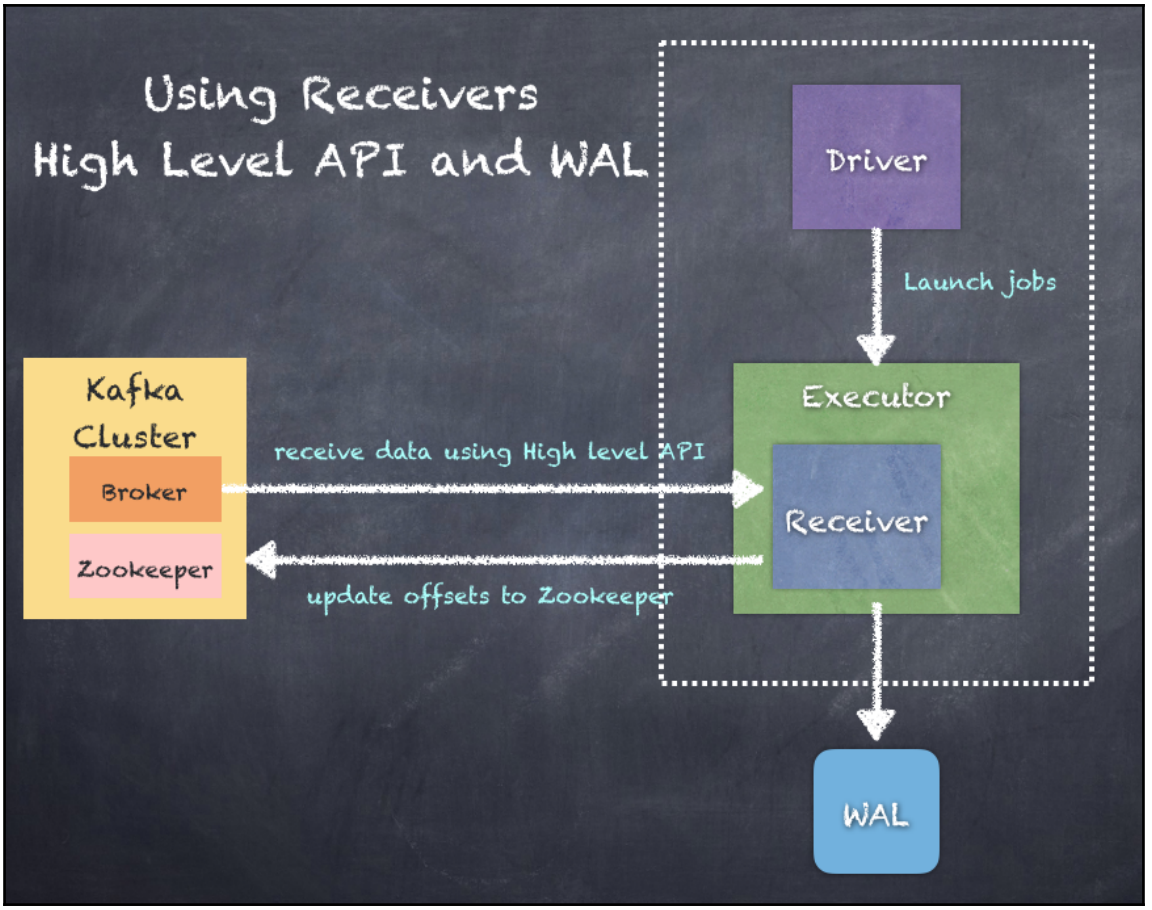

Kafka
Flume
HDFS/S3
Kinesis
Twitter
Spark Streaming
HDFS
Databases
Dashboards

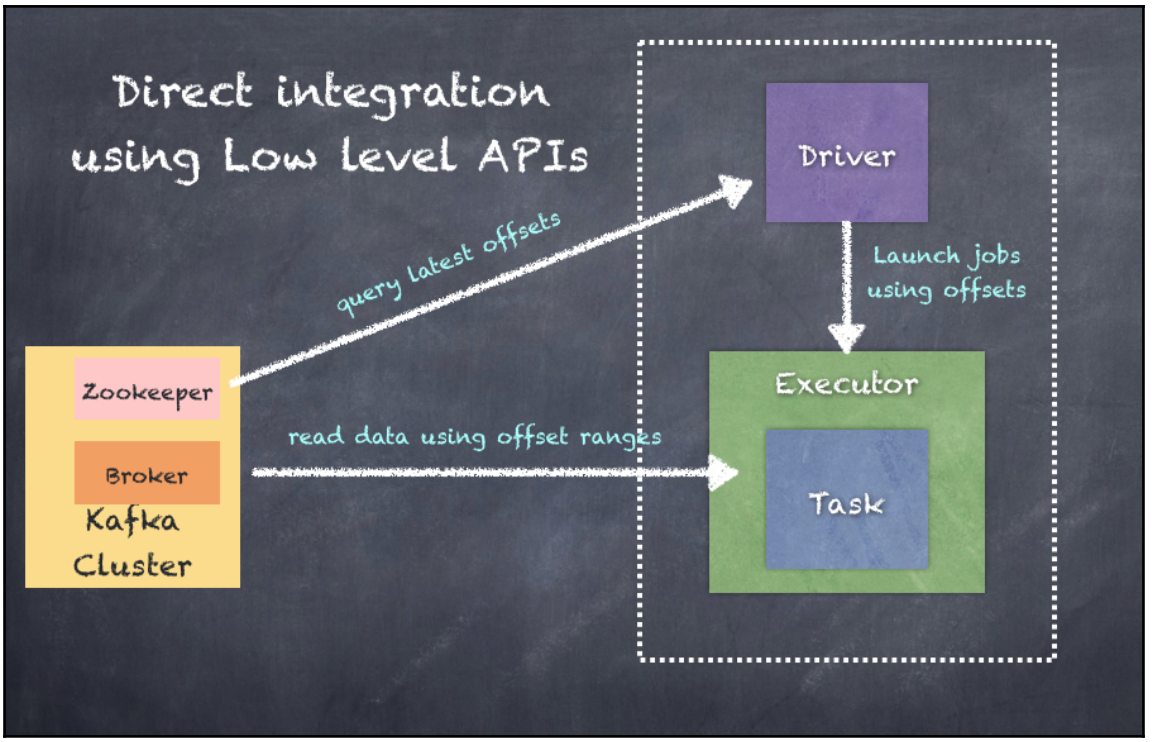

Input Data Stream
Spark Streaming
batches of input data
Spark Engine
batches of processed data


DStream
RDD @ Time1
RDD @ Time2
RDD @ Time3
RDD @ Time4
Data from Time 0 to 1
Data from Time 1 to 2
Data from Time 2 to 3
Data from Time 3 to 4
T secs
T secs
T secs
T secs
Time
0
1
2
3
4

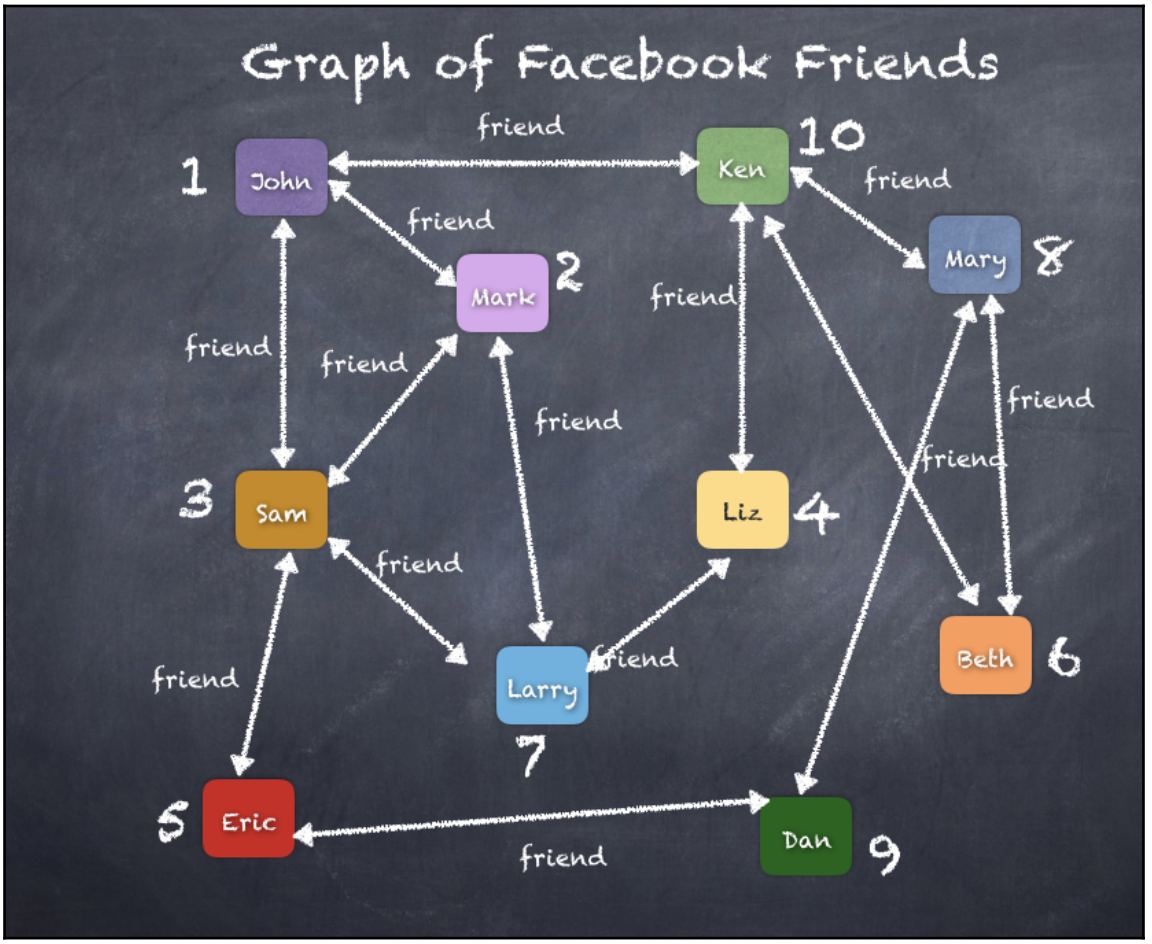

DStream transformation
RDD @ Time1
RDD @ Time2
RDD @ Time3
RDD @ Time4
baseDStream
Data from Time 0 to 1
Data from Time 1 to 2
Data from Time 2 to 3
Data from Time 3 to 4
filterDStream
Data from Time 0 to 1
Data from Time 1 to 2
Data from Time 2 to 3
Data from Time 3 to 4

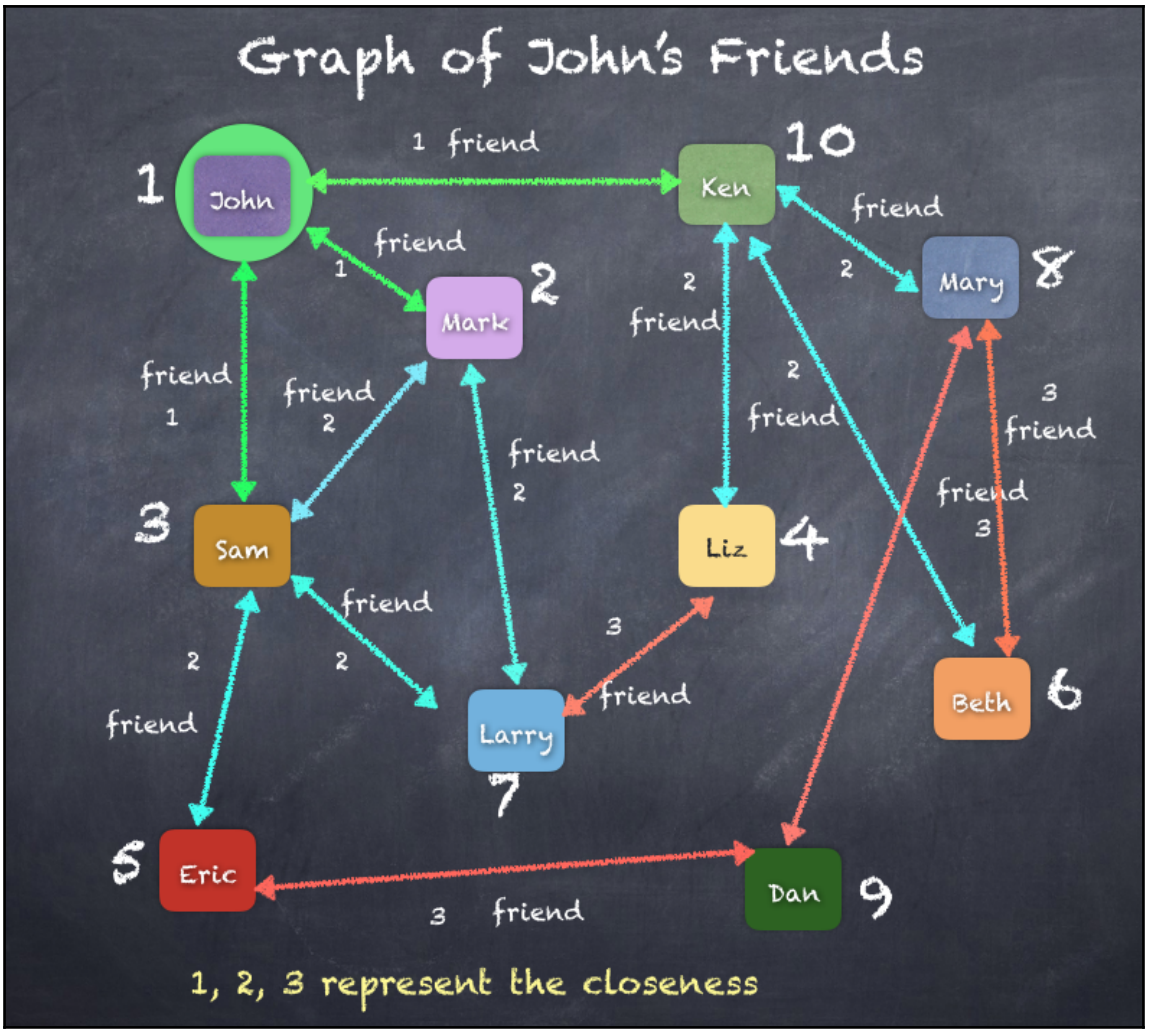

DStream with sliding Window
RDD @ Time1
RDD @ Time2
RDD @ Time3
RDD @ Time4
Data from Time 0 to 1
Data from Time 1 to 2
Data from Time 2 to 3
Data from Time 3 to 4
T secs
T secs
T secs
T secs
Time
0
1
2
3
4
Old Window
Window

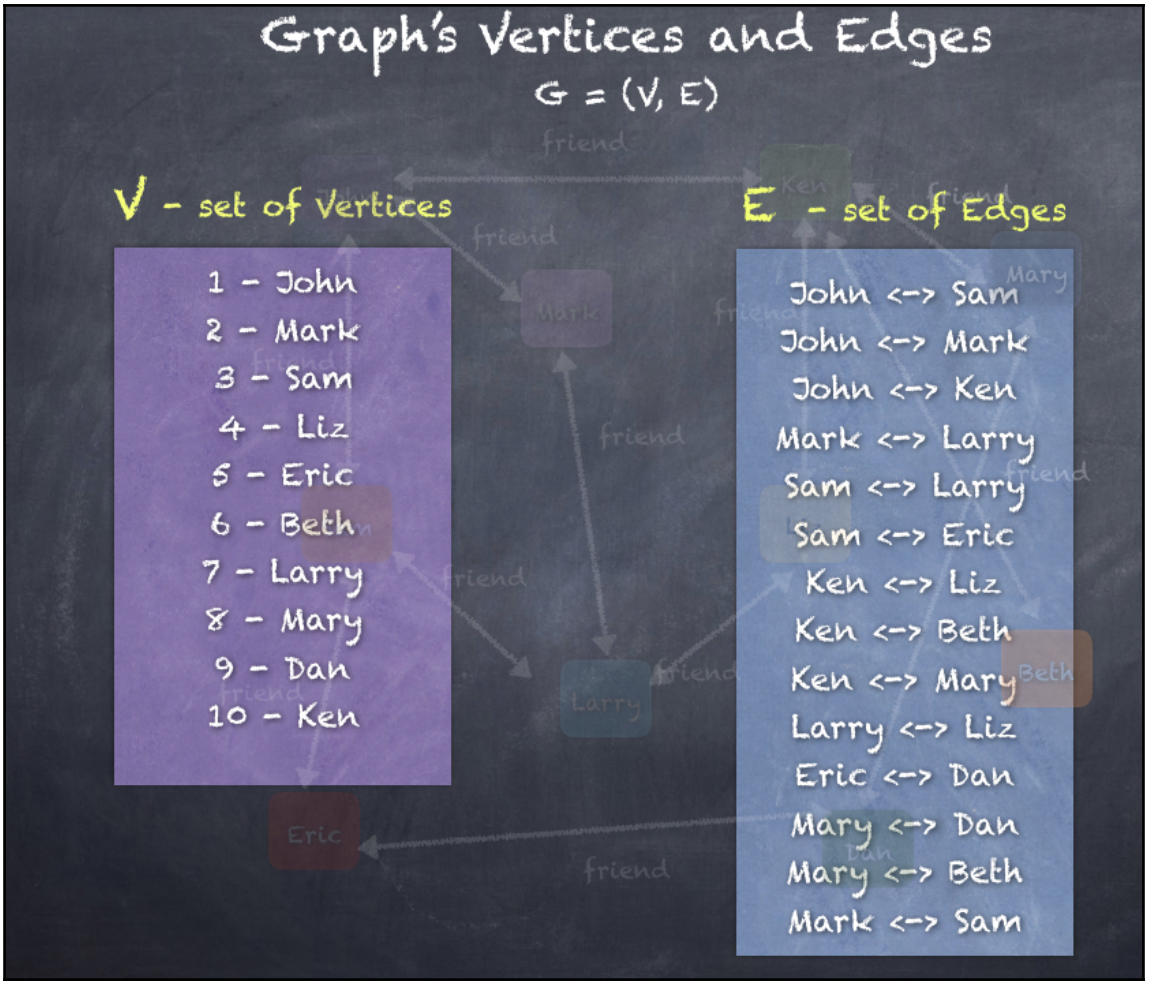

Using Receivers
High Level API and WAL
Driver
Launch jobs
Kafka
Cluster
Broker
Zookeeper
receive data using High level API
Executor
Receiver
update offsets to Zookeeper
WAL

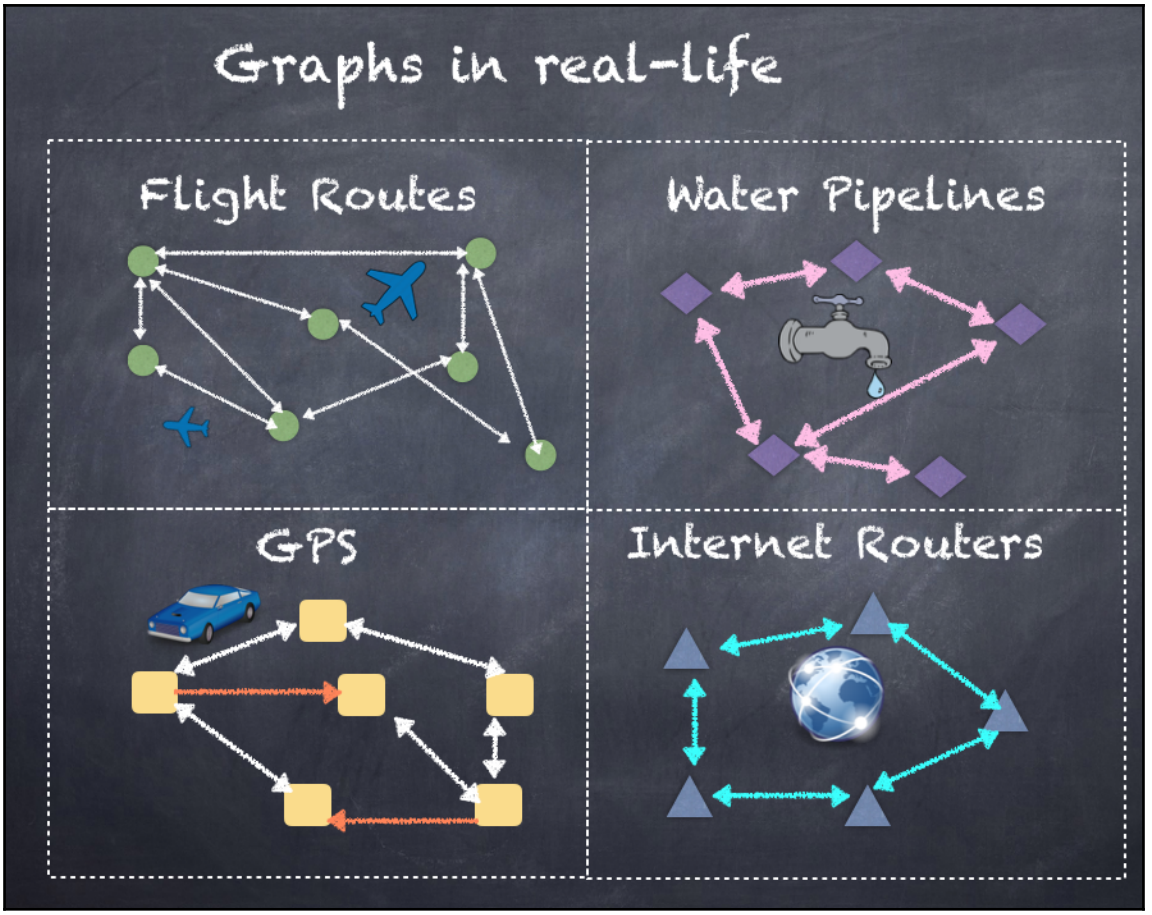

Direct integration
using Low level APIs
Driver
query latest offsets
Launch jobs
using offsets
Zookeeper
Executor
read data using offset ranges
Broker
Task
Kafka
Cluster

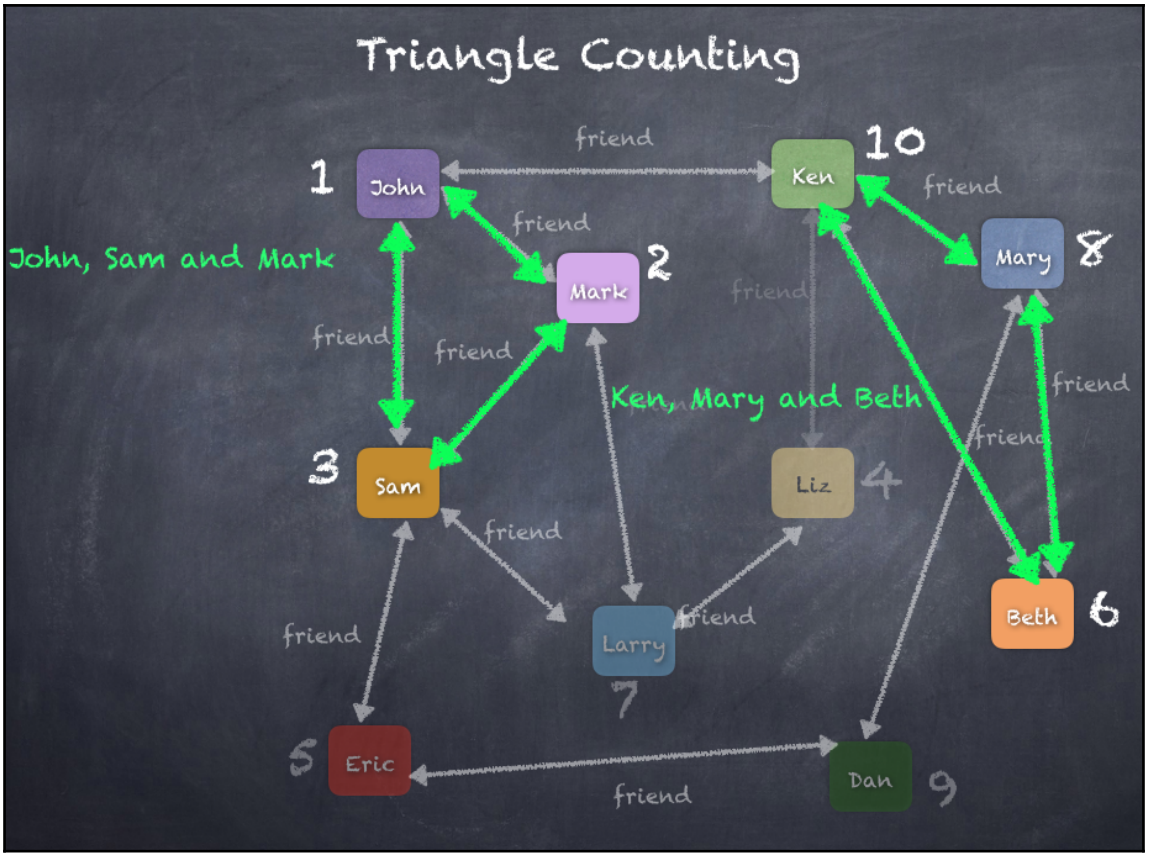
# Chapter 10: Everything is Connected - GraphX

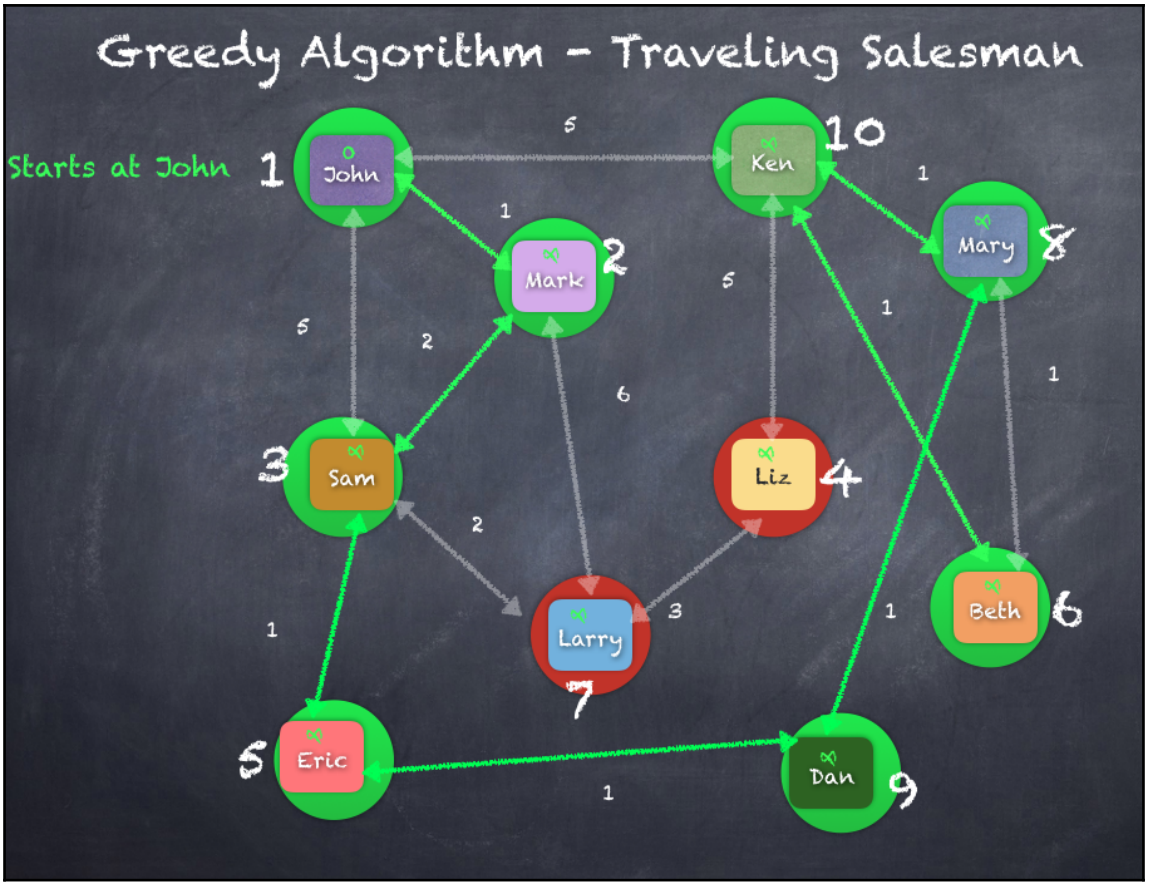


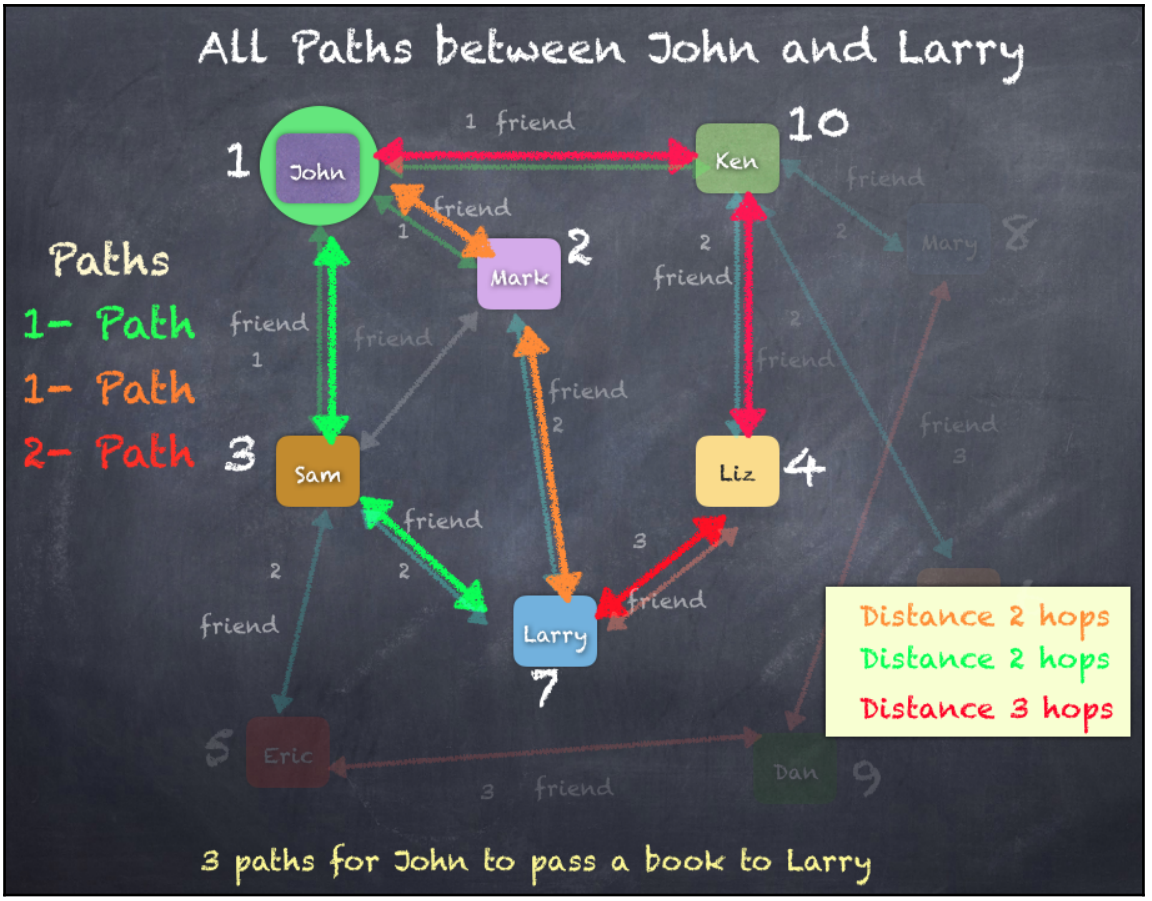

Graph of John's Friends
1 friend
1
John
10
Ken
friend
1
Mark
2
2
friend
2
Mary
8
friend
1
friend
2
friend
2
2
friend
3
friend
friend
3
3
Sam
Liz
4
friend
2
2
3
friend
Larry
7
Beth
6
friend
5
Eric
Dan
9
3 friend
1, 2, 3 represent the closeness

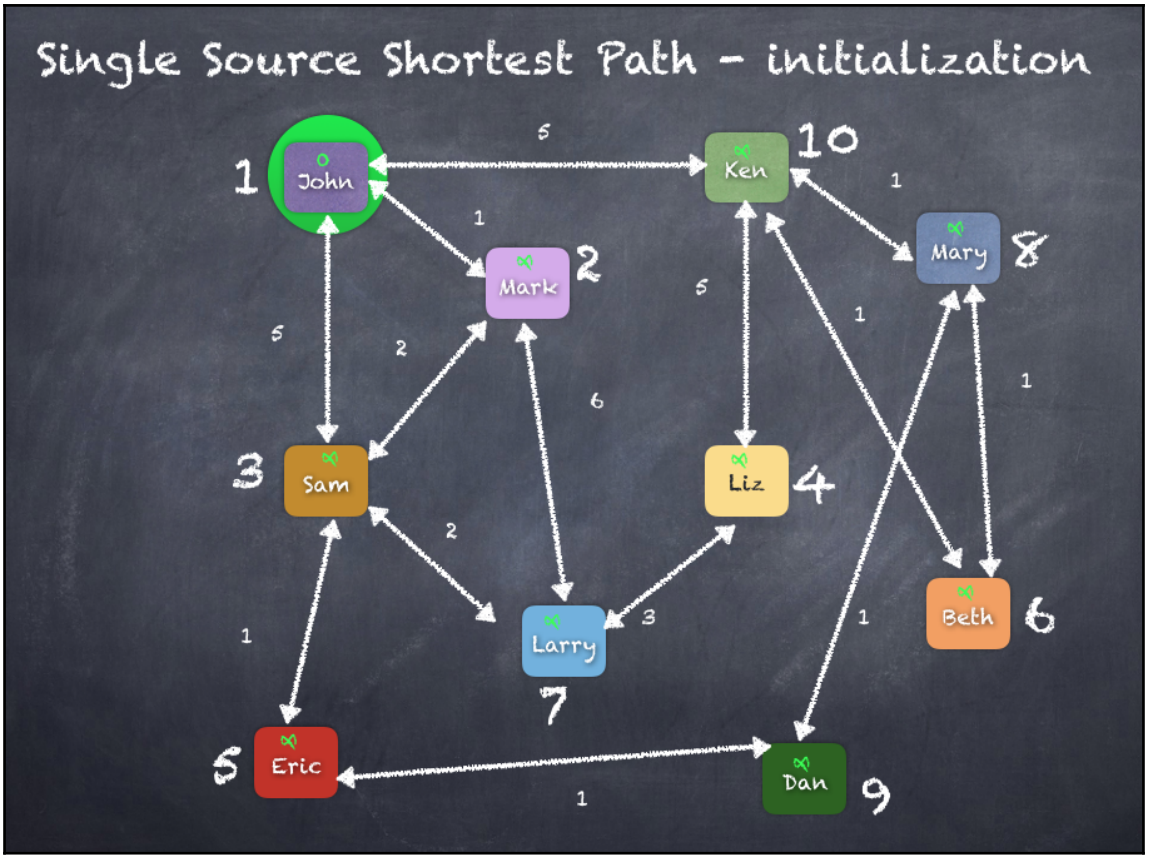

Graph's Vertices and Edges
G = (V, E)
V - set of Vertices
1 - John
2 - Mark
3 - Sam
4 - Liz
5 - Eric
6 - Beth
7 - Larry
8 - Mary
9 - Dan
10 - Ken
E - set of Edges
John <-> Sam
John <-> Mark
John <-> Ken
Mark <-> Larry
Sam <-> Larry
Sam <-> Eric
Ken <-> Liz
Ken <-> Beth
Ken <-> Mary
Larry <-> Liz
Eric <-> Dan
Mary <-> Dan
Mary <-> Beth
Mark <-> Sam

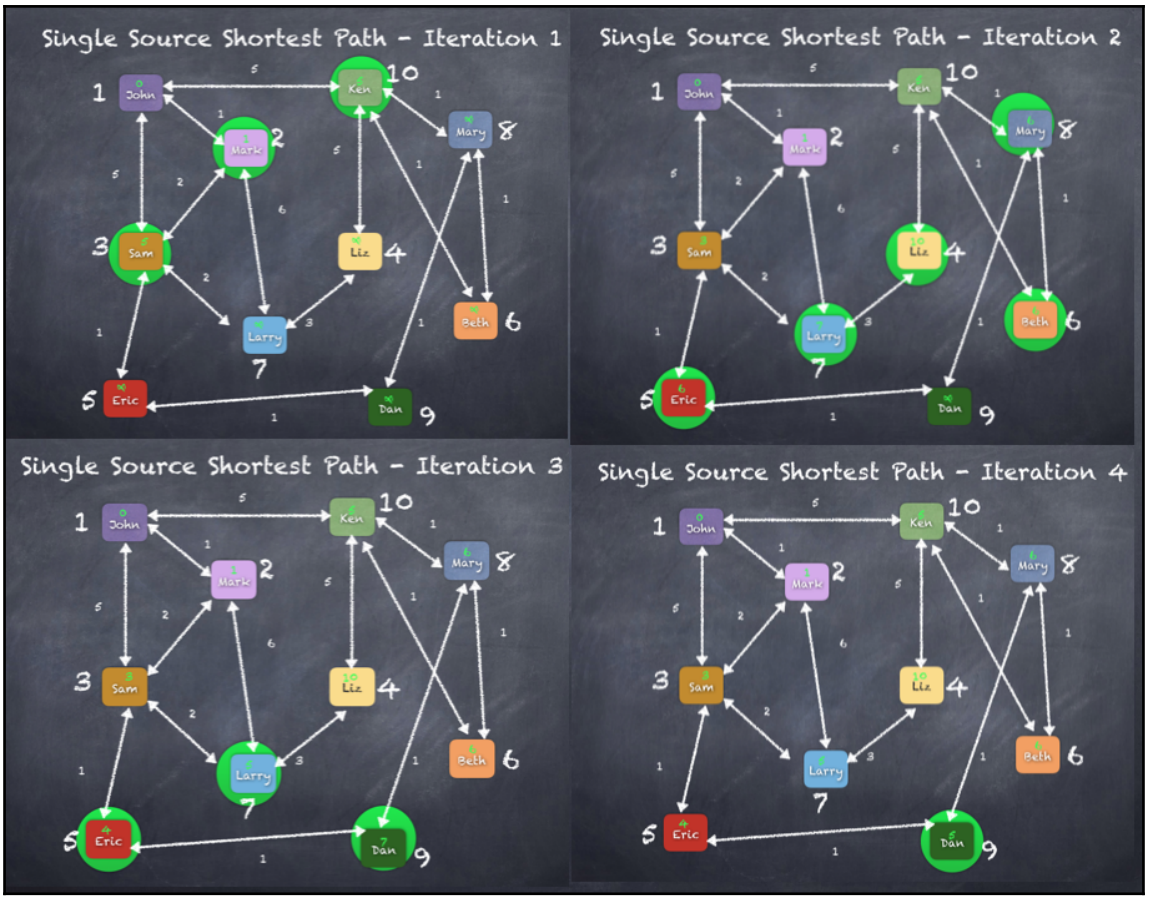

Graphs in real-life
Flight Routes
Water Pipelines
GPS
Internet Routers

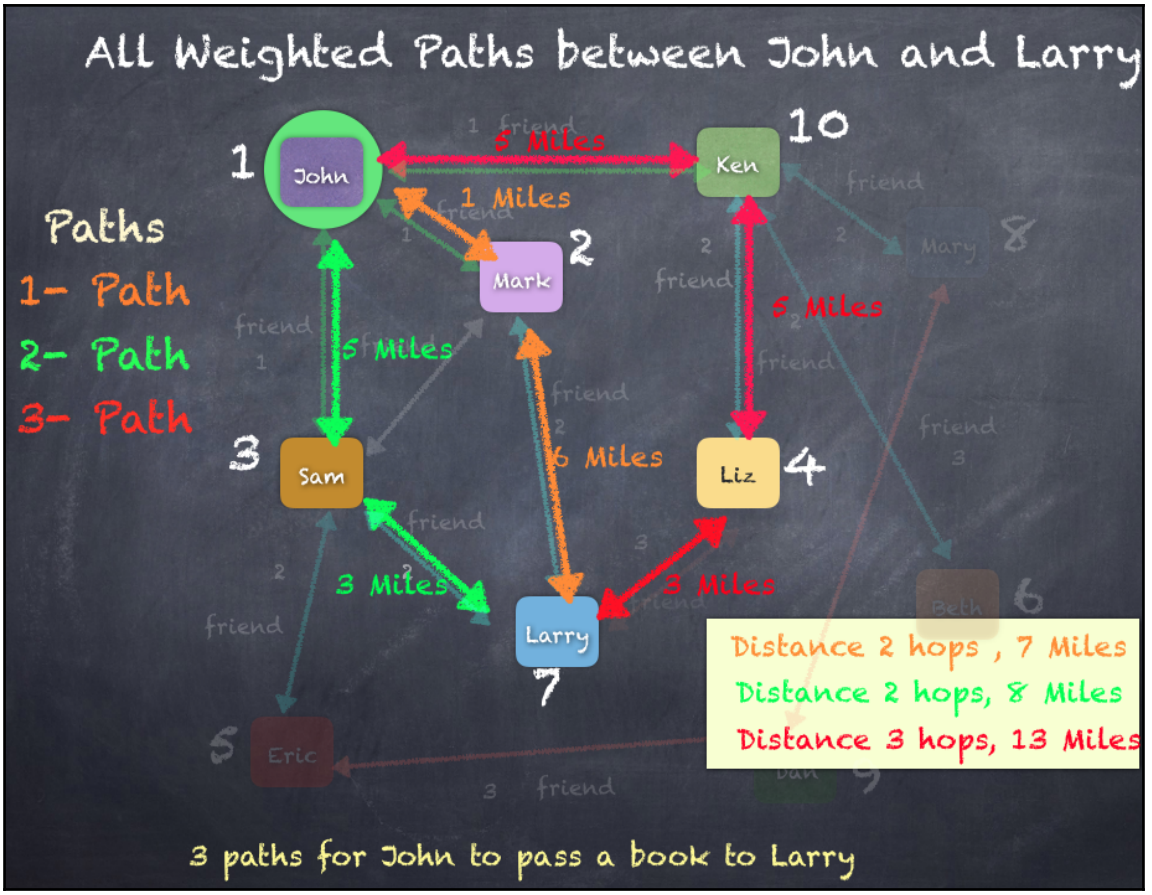

Triangle Counting
John, Sam and Mark
Ken, Mary and Beth
1
John
2
Mark
3
Sam
4
Liz
5
Eric
6
Beth
7
Larry
8
Mary
9
Dan
10
Ken
friend

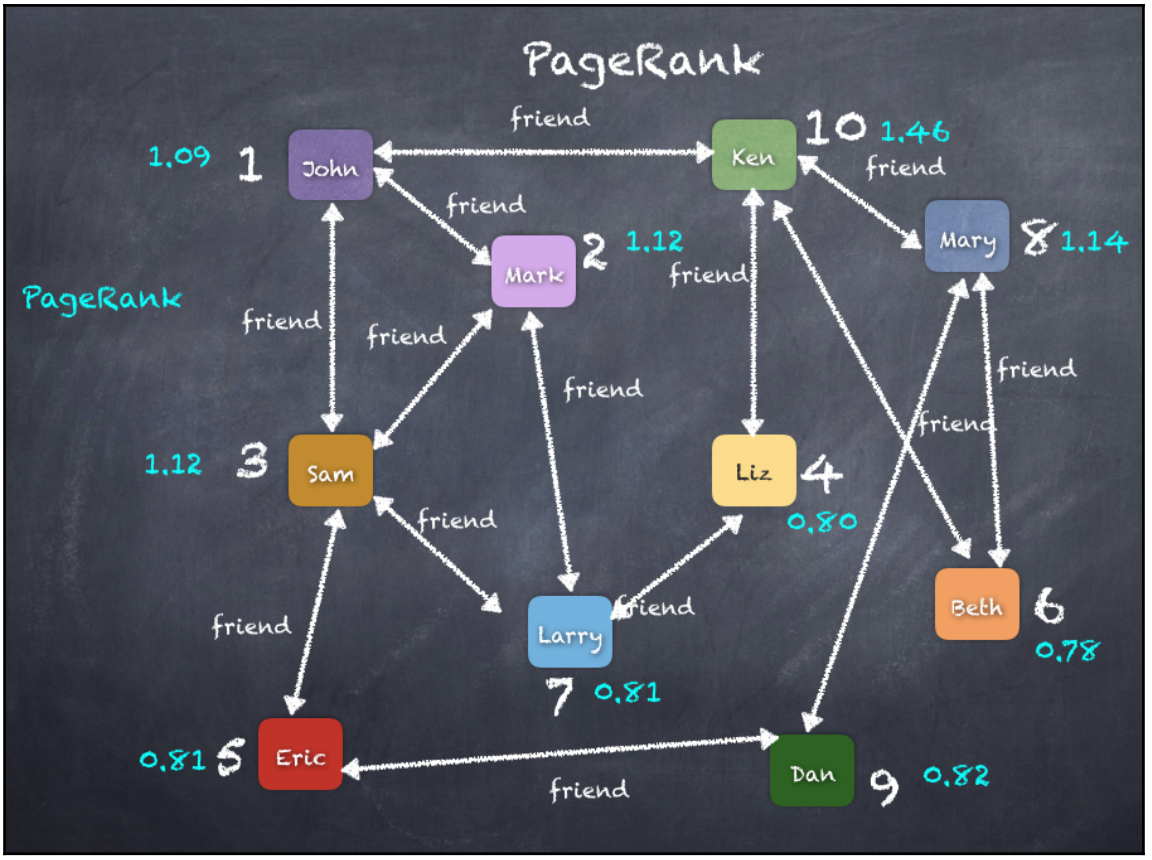

Greedy Algorithm - Traveling Salesman
Starts at John
1
0
John
5
Ken
10
1
1
Mark
2
Mary
8
5
5
2
1
6
1
3
Sam
Liz
4
2
Larry
3
1
Beth
6
7
5
Eric
Dan
9
1


All Paths between John and Larry
Paths
1- Path
1- Path
2- Path
1 John
2 Mark
3 Sam
4 Liz
5 Eric
7 Larry
8 Mary
9 Dan
10 Ken
1 friend
friend
Distance 2 hops
Distance 2 hops
Distance 3 hops
3 paths for John to pass a book to Larry

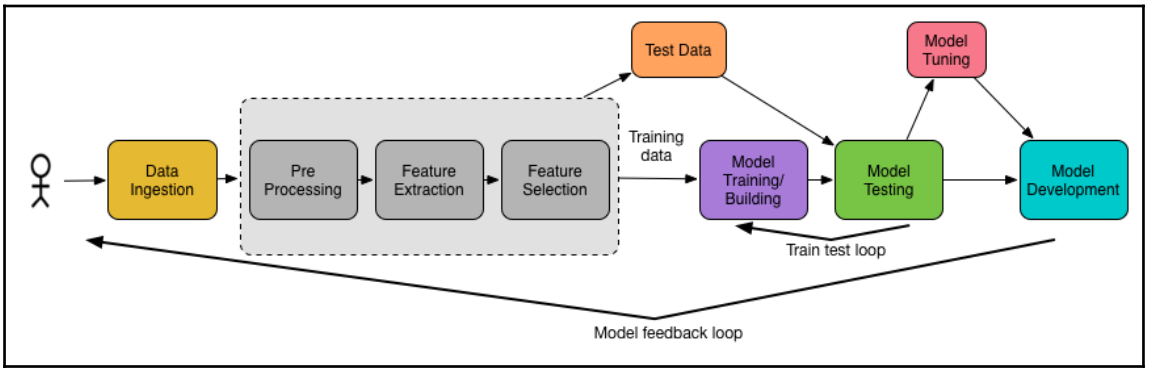

Single Source Shortest Path - initialization
1
0
John
5
Ken
10
1
1
Mark
2
Mary
8
5
5
2
1
6
1
3
Sam
Liz
4
2
1
Larry
3
1
Beth
6
7
5
Eric
Dan
9
1

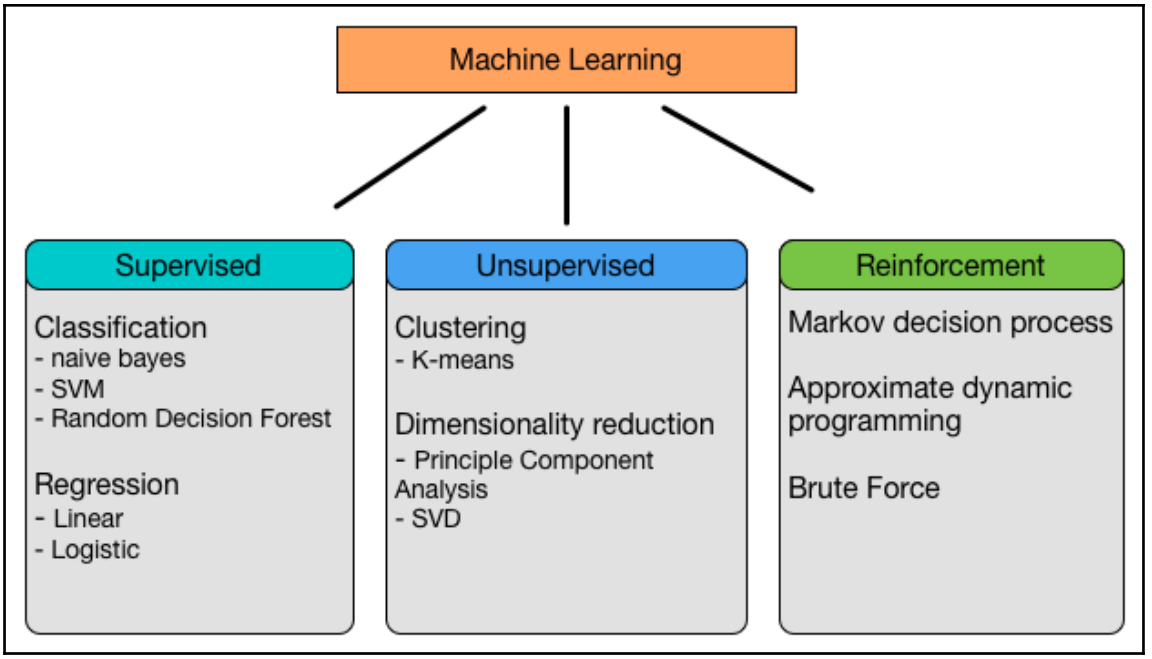

Single Source Shortest Path - Iteration 1
Single Source Shortest Path - Iteration 2
Single Source Shortest Path - Iteration 3
Single Source Shortest Path - Iteration 4
John
Ken
Mary
Mark
Sam
Liz
Larry
Beth
Eric
Dan

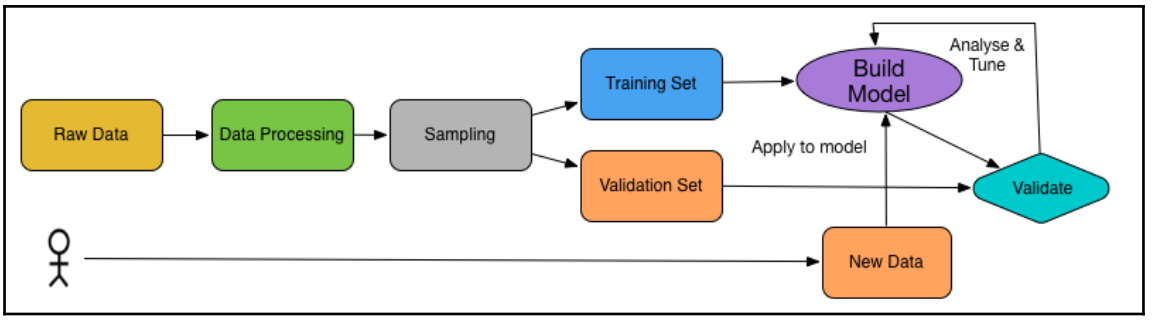

All Weighted Paths between John and Larry
Paths
1- Path
2- Path
3- Path
John
Ken
Mark
Sam
Liz
Larry
Eric
Mary
Beth
5 Miles
1 Miles
5 Miles
6 Miles
3 Miles
3 Miles
5 Miles
friend
Distance 2 hops , 7 Miles
Distance 2 hops, 8 Miles
Distance 3 hops, 13 Miles
3 paths for John to pass a book to Larry

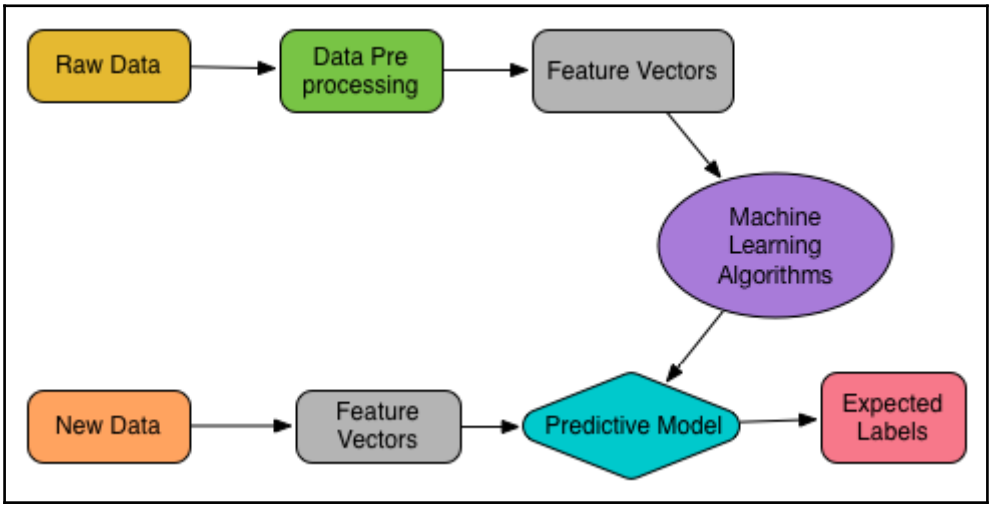

PageRank
PageRank
John
1
1.09
Ken
10
1.46
Mark
2
1.12
Mary
8
1.14
Sam
3
1.12
Liz
4
0.80
Larry
7
0.81
Beth
6
0.78
Eric
5
0.81
Dan
9
0.82
friend

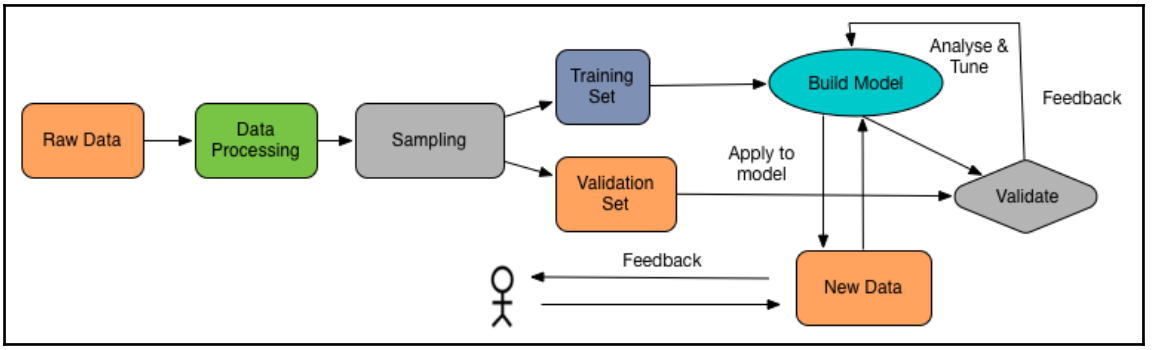
# Chapter 11: Learning Machine Learning - Spark MLlib and Spark ML

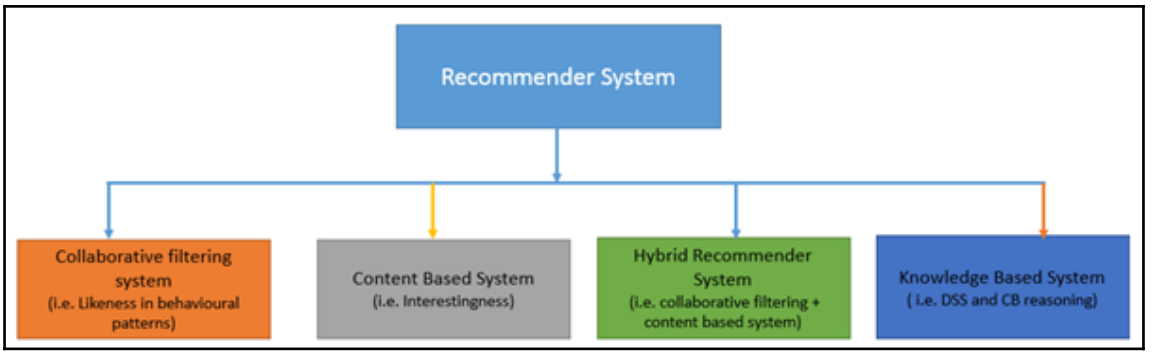



Raw Data
Data Processing
Sampling
Training Set
Validation Set
Build Model
Analyse & Tune
Apply to model
Validate
New Data


Raw Data
Data Pre processing
Feature Vectors
Machine Learning Algorithms
New Data
Feature Vectors
Predictive Model
Expected Labels


Raw Data
Data Processing
Sampling
Training Set
Validation Set
Build Model
Analyse & Tune
Feedback
Apply to model
Validate
Feedback
New Data




```
+---+---------------+
|id |name           |
+---+---------------+
|0  |[Jason, David]  |
|1  |[David, Martin] |
|2  |[Martin, Jason] |
|3  |[Jason, Daiel]  |
|4  |[Daiel, Martin] |
|5  |[Moahmed, Jason]|
|6  |[David, David]  |
|7  |[Jason, Martin] |
+---+---------------+
```

```
+---+---------------+-------------------+
|id |name           |features           |
+---+---------------+-------------------+
|0  |[Jason, David]  |(3,[0,1],[1.0,1.0])|
|1  |[David, Martin] |(3,[1,2],[1.0,1.0])|
|2  |[Martin, Jason] |(3,[0,2],[1.0,1.0])|
|3  |[Jason, Daiel]  |(3,[0],[1.0])      |
|4  |[Daiel, Martin] |(3,[2],[1.0])      |
|5  |[Moahmed, Jason]|(3,[0],[1.0])      |
|6  |[David, David]  |(3,[1],[2.0])      |
|7  |[Jason, Martin] |(3,[0,2],[1.0,1.0])|
+---+---------------+-------------------+
```

| sentence | words | tokens |
|---|---|---|
| Tokenization,is the process of enchanting words,from the raw text | [tokenization,is, the, process, of, enchanting, words,from, the, raw, text] | 9 |
| If you want,to have more advance tokenization,RegexTokenizer,is a good option | [, if, you, want,to, have, more, advance, tokenization,regextokenizer,is, a, good, option] | 11 |
| Here,will provide a sample example on how to tockenize sentences | [, here,will, provide, a, sample, example, on, how, to, tockenize, sentences] | 11 |
| This way,you can find all matching occurrences | [this, way,you, can, find, all, matching, occurrences] | 7 |

| sentence | words | tokens |
|---|---|---|
| Tokenization,is the process of enchanting words,from the raw text | [tokenization, is, the, process, of, enchanting, words, from, the, raw, text] | 11 |
| If you want,to have more advance tokenization,RegexTokenizer,is a good option | [if, you, want, to, have, more, advance, tokenization, regextokenizer, is, a, good, option] | 13 |
| Here,will provide a sample example on how to tockenize sentences | [here, will, provide, a, sample, example, on, how, to, tockenize, sentences] | 11 |
| This way,you can find all matching occurrences | [this, way, you, can, find, all, matching, occurrences] | 8 |

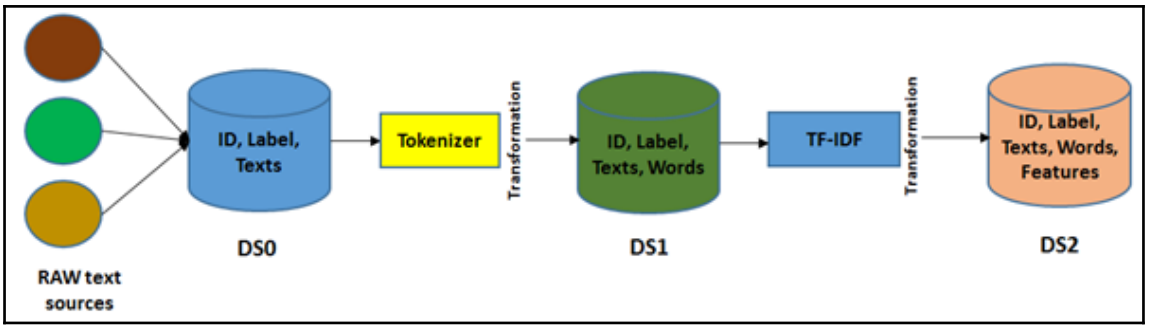
| id | filtered |
|---|---|
| 0 | [tokenization, process, enchanting, words, raw, text] |
| 1 | [want, advance, tokenization, regextokenizer, good, option] |
| 2 | [provide, sample, example, tockenize, sentences] |
| 3 | [way, find, matching, occurrences] |

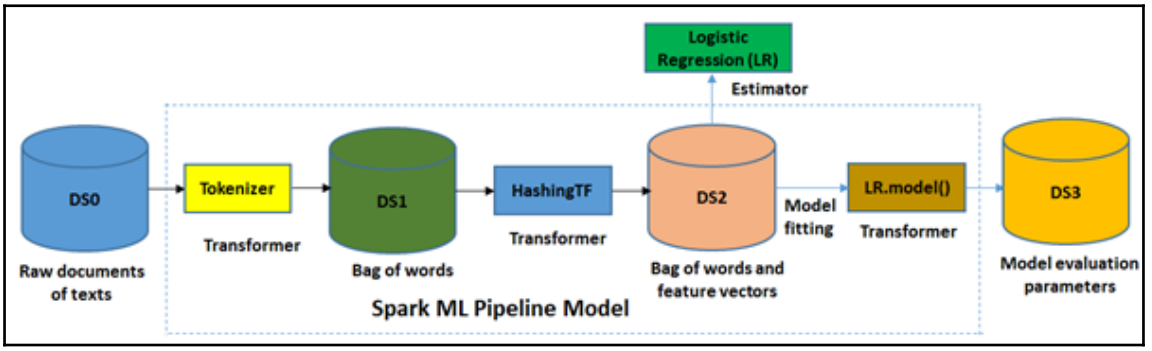
| id | name | address |
|---|---|---|
| 0 | Jason | Germany |
| 1 | David | France |
| 2 | Martin | Spain |
| 3 | Jason | USA |
| 4 | Daiel | UK |
| 5 | Moahmed | Bangladesh |
| 6 | David | Ireland |
| 7 | Jason | Netherlands |

| id | name | address | label |
|---|---|---|---|
| 0 | Jason | Germany | 0.0 |
| 1 | David | France | 1.0 |
| 2 | Martin | Spain | 3.0 |
| 3 | Jason | USA | 0.0 |
| 4 | Daiel | UK | 4.0 |
| 5 | Moahmed | Bangladesh | 2.0 |
| 6 | David | Ireland | 1.0 |
| 7 | Jason | Netherlands | 0.0 |

| id | name | address |
|---|---|---|
| 0 | Jason | Germany |
| 1 | David | France |
| 2 | Martin | Spain |
| 3 | Jason | USA |
| 4 | Daiel | UK |
| 5 | Moahmed | Bangladesh |
| 6 | David | Ireland |
| 7 | Jason | Netherlands |

| id | name | address | categoryIndex | categoryVec |
|---|---|---|---|---|
| 0 | Jason | Germany | 0.0 | (4,[0],[1.0]) |
| 1 | David | France | 1.0 | (4,[1],[1.0]) |
| 2 | Martin | Spain | 3.0 | (4,[3],[1.0]) |
| 3 | Jason | USA | 0.0 | (4,[0],[1.0]) |
| 4 | Daiel | UK | 4.0 | (4,[],[]) |
| 5 | Moahmed | Bangladesh | 2.0 | (4,[2],[1.0]) |
| 6 | David | Ireland | 1.0 | (4,[1],[1.0]) |
| 7 | Jason | Netherlands | 0.0 | (4,[0],[1.0]) |

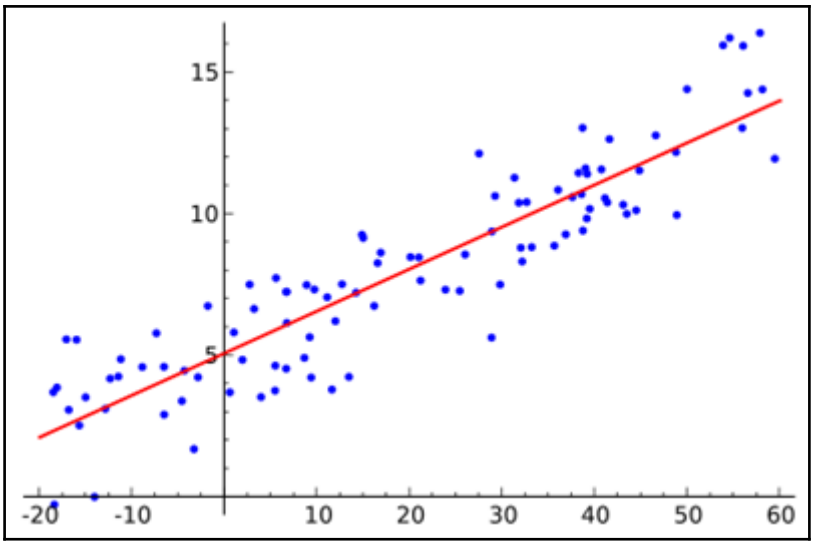


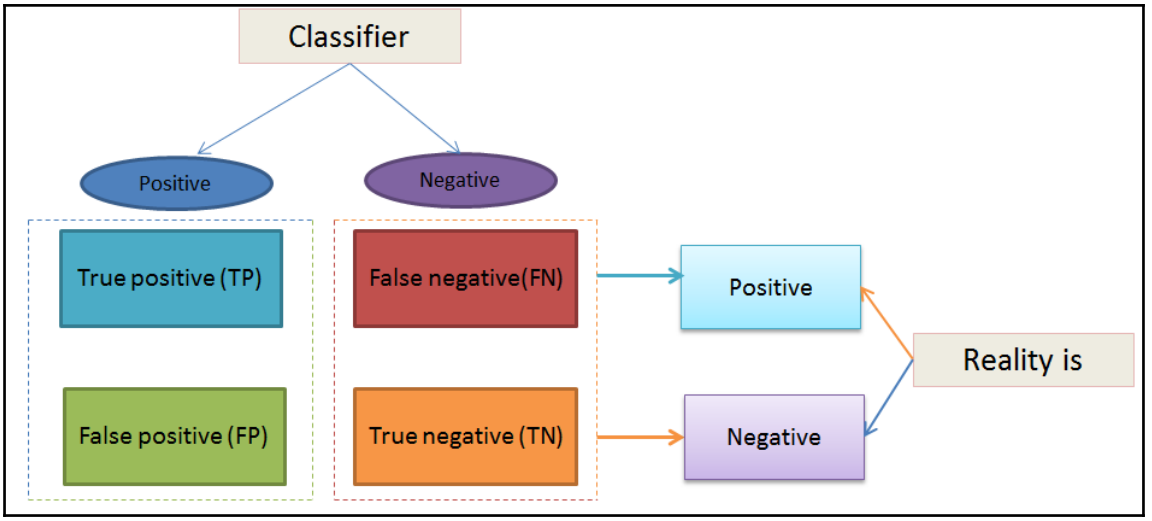

```
+-------------------------+
|features                 |
+-------------------------+
|[3.5,2.0,5.0,6.3,5.6,2.4] |
|[4.4,0.1,3.0,9.0,7.0,8.75]|
|[3.2,2.4,0.0,6.0,7.4,3.34]|
+-------------------------+
```

```
+-----------------------------------------------------------------------------+
|pcaFeatures                                                                  |
+-----------------------------------------------------------------------------+
|[-5.149253129088702,3.2157431427730385,-6.828533673828153,5.774261462142295] |
|[-12.372614091904445,0.804196667817684,-6.828533673828154,5.774261462142296] |
|[-5.649682494292658,-2.189177804885822,-6.828533673828155,5.7742614621422925]|
+-----------------------------------------------------------------------------+
```

```
|label|              features|
+-----+--------------------+
|  5.0|(780,[152,153,154...|
|  0.0|(780,[127,128,129...|
|  4.0|(780,[160,161,162...|
|  1.0|(780,[158,159,160...|
|  9.0|(780,[208,209,210...|
|  2.0|(780,[155,156,157...|
|  1.0|(780,[124,125,126...|
|  3.0|(780,[151,152,153...|
|  1.0|(780,[152,153,154...|
|  4.0|(780,[134,135,161...|
|  3.0|(780,[123,124,125...|
|  5.0|(780,[216,217,218...|
|  3.0|(780,[143,144,145...|
|  6.0|(780,[72,73,74,99...|
|  1.0|(780,[151,152,153...|
|  7.0|(780,[211,212,213...|
|  2.0|(780,[151,152,153...|
|  8.0|(780,[159,160,161...|
|  6.0|(780,[100,101,102...|
|  9.0|(780,[209,210,211...|
|  4.0|(780,[129,130,131...|
|  0.0|(780,[129,130,131...|
|  9.0|(780,[183,184,185...|
|  1.0|(780,[158,159,160...|
|  1.0|(780,[99,100,101,...|
|  2.0|(780,[124,125,126...|
|  4.0|(780,[185,186,187...|
|  3.0|(780,[150,151,152...|
|  2.0|(780,[145,146,147...|
|  7.0|(780,[240,241,242...|
+-----+--------------------+
only showing top 30 rows
```




$$RMSD = \sqrt{\frac{\sum_{t=1}^{n} \left( \hat{y}_t - y_t \right)^2}{n}}$$




| Metric | Definition |
|---|---|
| Precision (Positive Predictive Value) | $PPV = \frac{TP}{TP+FP}$ |
| Recall (True Positive Rate) | $TPR = \frac{TP}{P} = \frac{TP}{TP+FN}$ |
| F-measure | $F(\beta) = (1+\beta^2) \cdot \left(\frac{PPV \cdot TPR}{\beta^2 \cdot PPV + TPR}\right)$ |
| Receiver Operating Characteristic (ROC) | $FPR(T) = \int_T^\infty P_0(T)\,dT$<br>$TPR(T) = \int_T^\infty P_1(T)\,dT$ |
| Area Under ROC Curve | $AUROC = \int_0^1 \frac{TP}{P} d\left(\frac{FP}{N}\right)$ |
| Area Under Precision-Recall Curve | $AUPRC = \int_0^1 \frac{TP}{TP+FP} d\left(\frac{TP}{P}\right)$ |

| Metric | Definition |
|---|---|
| Confusion Matrix | $C_{ij} = \sum_{k=0}^{N-1} \hat{\delta}(\mathbf{y}_k - \ell_i) \cdot \hat{\delta}(\hat{\mathbf{y}}_k - \ell_j)$ <br> $\begin{pmatrix} \sum_{k=0}^{N-1} \hat{\delta}(\mathbf{y}_k - \ell_1) \cdot \hat{\delta}(\hat{\mathbf{y}}_k - \ell_1) & \dots & \sum_{k=0}^{N-1} \hat{\delta}(\mathbf{y}_k - \ell_1) \cdot \hat{\delta}(\hat{\mathbf{y}}_k - \ell_N) \\ \vdots & \ddots & \vdots \\ \sum_{k=0}^{N-1} \hat{\delta}(\mathbf{y}_k - \ell_N) \cdot \hat{\delta}(\hat{\mathbf{y}}_k - \ell_1) & \dots & \sum_{k=0}^{N-1} \hat{\delta}(\mathbf{y}_k - \ell_N) \cdot \hat{\delta}(\hat{\mathbf{y}}_k - \ell_N) \end{pmatrix}$ |
| Accuracy | $ACC = \frac{TP}{TP+FP} = \frac{1}{N} \sum_{i=0}^{N-1} \hat{\delta}\left(\hat{\mathbf{y}}_i - \mathbf{y}_i\right)$ |
| Precision by label | $PPV(\ell) = \frac{TP}{TP+FP} = \frac{\sum_{i=0}^{N-1} \hat{\delta}(\hat{\mathbf{y}}_i - \ell) \cdot \hat{\delta}(\mathbf{y}_i - \ell)}{\sum_{i=0}^{N-1} \hat{\delta}(\hat{\mathbf{y}}_i - \ell)}$ |
| Recall by label | $TPR(\ell) = \frac{TP}{P} = \frac{\sum_{i=0}^{N-1} \hat{\delta}(\hat{\mathbf{y}}_i - \ell) \cdot \hat{\delta}(\mathbf{y}_i - \ell)}{\sum_{i=0}^{N-1} \hat{\delta}(\mathbf{y}_i - \ell)}$ |
| F-measure by label | $F(\beta, \ell) = \left(1 + \beta^2\right) \cdot \left(\frac{PPV(\ell) \cdot TPR(\ell)}{\beta^2 \cdot PPV(\ell) + TPR(\ell)}\right)$ |
| Weighted precision | $PPV_w = \frac{1}{N} \sum_{\ell \in L} PPV(\ell) \cdot \sum_{i=0}^{N-1} \hat{\delta}(\mathbf{y}_i - \ell)$ |
| Weighted recall | $TPR_w = \frac{1}{N} \sum_{\ell \in L} TPR(\ell) \cdot \sum_{i=0}^{N-1} \hat{\delta}(\mathbf{y}_i - \ell)$ |
| Weighted F-measure | $F_w(\beta) = \frac{1}{N} \sum_{\ell \in L} F(\beta, \ell) \cdot \sum_{i=0}^{N-1} \hat{\delta}(\mathbf{y}_i - \ell)$ |

$$\hat{\delta}(x) = \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{otherwise} \end{cases}$$

$$L\left(\mathbf{w}; \mathbf{x}, y\right) := \log\left(1 + \exp\left(-y\mathbf{w}^T\mathbf{x}\right)\right)$$

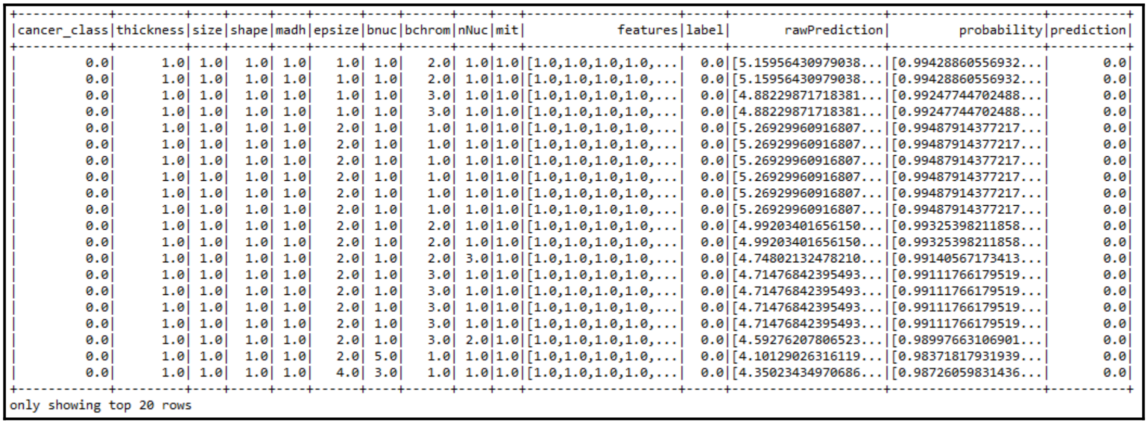
$$f(z) = 1/(1 + e - z)$$

| cancer_class | thickness | size | shape | madh | epsize | bnuc | bchrom | nNuc | mit |
|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 5.0 | 1.0 | 1.0 | 1.0 | 2.0 | 1.0 | 3.0 | 1.0 | 1.0 |
| 0.0 | 5.0 | 4.0 | 4.0 | 5.0 | 7.0 | 10.0 | 3.0 | 2.0 | 1.0 |
| 0.0 | 3.0 | 1.0 | 1.0 | 1.0 | 2.0 | 2.0 | 3.0 | 1.0 | 1.0 |
| 0.0 | 6.0 | 8.0 | 8.0 | 1.0 | 3.0 | 4.0 | 3.0 | 7.0 | 1.0 |
| 0.0 | 4.0 | 1.0 | 1.0 | 3.0 | 2.0 | 1.0 | 3.0 | 1.0 | 1.0 |
| 1.0 | 8.0 | 10.0 | 10.0 | 8.0 | 7.0 | 10.0 | 9.0 | 7.0 | 1.0 |
| 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | 2.0 | 10.0 | 3.0 | 1.0 | 1.0 |
| 0.0 | 2.0 | 1.0 | 2.0 | 1.0 | 2.0 | 1.0 | 3.0 | 1.0 | 1.0 |
| 0.0 | 2.0 | 1.0 | 1.0 | 1.0 | 2.0 | 1.0 | 1.0 | 1.0 | 5.0 |
| 0.0 | 4.0 | 2.0 | 1.0 | 1.0 | 2.0 | 1.0 | 2.0 | 1.0 | 1.0 |
| 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 3.0 | 1.0 | 1.0 |
| 0.0 | 2.0 | 1.0 | 1.0 | 1.0 | 2.0 | 1.0 | 2.0 | 1.0 | 1.0 |
| 1.0 | 5.0 | 3.0 | 3.0 | 3.0 | 2.0 | 3.0 | 4.0 | 4.0 | 1.0 |
| 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | 2.0 | 3.0 | 3.0 | 1.0 | 1.0 |
| 1.0 | 8.0 | 7.0 | 5.0 | 10.0 | 7.0 | 9.0 | 5.0 | 5.0 | 4.0 |
| 1.0 | 7.0 | 4.0 | 6.0 | 4.0 | 6.0 | 1.0 | 4.0 | 3.0 | 1.0 |
| 0.0 | 4.0 | 1.0 | 1.0 | 1.0 | 2.0 | 1.0 | 2.0 | 1.0 | 1.0 |
| 0.0 | 4.0 | 1.0 | 1.0 | 1.0 | 2.0 | 1.0 | 3.0 | 1.0 | 1.0 |
| 1.0 | 10.0 | 7.0 | 7.0 | 6.0 | 4.0 | 10.0 | 4.0 | 1.0 | 2.0 |
| 0.0 | 6.0 | 1.0 | 1.0 | 1.0 | 2.0 | 1.0 | 3.0 | 1.0 | 1.0 |

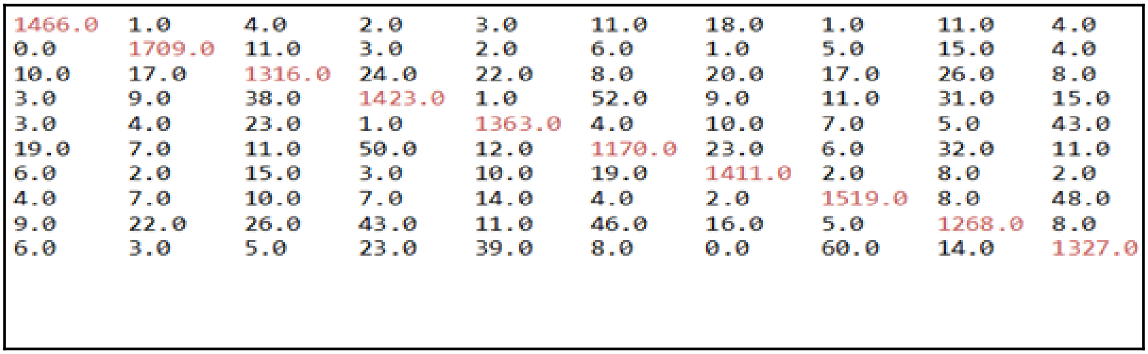
only showing top 20 rows

```
+------------+---------+----+-----+----+------+----+------+----+---+--------------------+
|cancer_class|thickness|size|shape|madh|epsize|bnuc|bchrom|nNuc|mit|            features|
+------------+---------+----+-----+----+------+----+------+----+---+--------------------+
|         0.0|      5.0| 1.0|  1.0| 1.0|   2.0| 1.0|   3.0| 1.0|1.0|[5.0,1.0,1.0,1.0,...|
|         0.0|      5.0| 4.0|  4.0| 5.0|   7.0|10.0|   3.0| 2.0|1.0|[5.0,4.0,4.0,5.0,...|
|         0.0|      3.0| 1.0|  1.0| 1.0|   2.0| 2.0|   3.0| 1.0|1.0|[3.0,1.0,1.0,1.0,...|
|         0.0|      6.0| 8.0|  8.0| 1.0|   3.0| 4.0|   3.0| 7.0|1.0|[6.0,8.0,8.0,1.0,...|
|         0.0|      4.0| 1.0|  1.0| 3.0|   2.0| 1.0|   3.0| 1.0|1.0|[4.0,1.0,1.0,3.0,...|
|         1.0|      8.0|10.0| 10.0| 8.0|   7.0|10.0|   9.0| 7.0|1.0|[8.0,10.0,10.0,8....|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0|10.0|   3.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|
|         0.0|      2.0| 1.0|  2.0| 1.0|   2.0| 1.0|   3.0| 1.0|1.0|[2.0,1.0,2.0,1.0,...|
|         0.0|      2.0| 1.0|  1.0| 1.0|   2.0| 1.0|   1.0| 1.0|5.0|[2.0,1.0,1.0,1.0,...|
|         0.0|      4.0| 2.0|  1.0| 1.0|   2.0| 1.0|   2.0| 1.0|1.0|[4.0,2.0,1.0,1.0,...|
|         0.0|      1.0| 1.0|  1.0| 1.0|   1.0| 1.0|   3.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|
|         0.0|      2.0| 1.0|  1.0| 1.0|   2.0| 1.0|   2.0| 1.0|1.0|[2.0,1.0,1.0,1.0,...|
|         1.0|      5.0| 3.0|  3.0| 3.0|   2.0| 3.0|   4.0| 4.0|1.0|[5.0,3.0,3.0,3.0,...|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 3.0|   3.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|
|         1.0|      8.0| 7.0|  5.0|10.0|   7.0| 9.0|   5.0| 5.0|4.0|[8.0,7.0,5.0,10.0...|
|         1.0|      7.0| 4.0|  6.0| 4.0|   6.0| 1.0|   4.0| 3.0|1.0|[7.0,4.0,6.0,4.0,...|
|         0.0|      4.0| 1.0|  1.0| 1.0|   2.0| 1.0|   2.0| 1.0|1.0|[4.0,1.0,1.0,1.0,...|
|         0.0|      4.0| 1.0|  1.0| 1.0|   2.0| 1.0|   3.0| 1.0|1.0|[4.0,1.0,1.0,1.0,...|
|         1.0|     10.0| 7.0|  7.0| 6.0|   4.0|10.0|   4.0| 1.0|2.0|[10.0,7.0,7.0,6.0...|
|         0.0|      6.0| 1.0|  1.0| 1.0|   2.0| 1.0|   3.0| 1.0|1.0|[6.0,1.0,1.0,1.0,...|
+------------+---------+----+-----+----+------+----+------+----+---+--------------------+
only showing top 20 rows
```

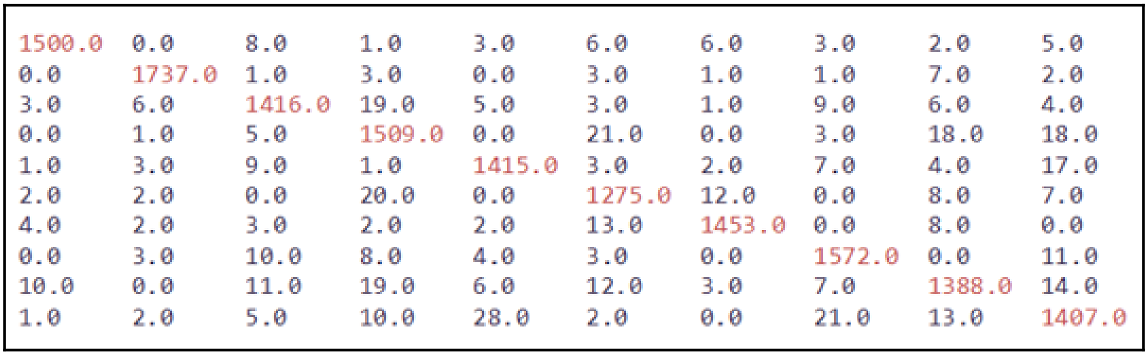
```
+------------+---------+----+-----+----+------+----+------+----+---+--------------------+-----+
|cancer_class|thickness|size|shape|madh|epsize|bnuc|bchrom|nNuc|mit|            features|label|
+------------+---------+----+-----+----+------+----+------+----+---+--------------------+-----+
|         0.0|      5.0| 1.0|  1.0| 1.0|   2.0| 1.0|   3.0| 1.0|1.0|[5.0,1.0,1.0,1.0,...|  0.0|
|         0.0|      5.0| 4.0|  4.0| 5.0|   7.0|10.0|   3.0| 2.0|1.0|[5.0,4.0,4.0,5.0,...|  0.0|
|         0.0|      3.0| 1.0|  1.0| 1.0|   2.0| 2.0|   3.0| 1.0|1.0|[3.0,1.0,1.0,1.0,...|  0.0|
|         0.0|      6.0| 8.0|  8.0| 1.0|   3.0| 4.0|   3.0| 7.0|1.0|[6.0,8.0,8.0,1.0,...|  0.0|
|         0.0|      4.0| 1.0|  1.0| 3.0|   2.0| 1.0|   3.0| 1.0|1.0|[4.0,1.0,1.0,3.0,...|  0.0|
|         1.0|      8.0|10.0| 10.0| 8.0|   7.0|10.0|   9.0| 7.0|1.0|[8.0,10.0,10.0,8....|  1.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0|10.0|   3.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|
|         0.0|      2.0| 1.0|  2.0| 1.0|   2.0| 1.0|   3.0| 1.0|1.0|[2.0,1.0,2.0,1.0,...|  0.0|
|         0.0|      2.0| 1.0|  1.0| 1.0|   2.0| 1.0|   1.0| 1.0|5.0|[2.0,1.0,1.0,1.0,...|  0.0|
|         0.0|      4.0| 2.0|  1.0| 1.0|   2.0| 1.0|   2.0| 1.0|1.0|[4.0,2.0,1.0,1.0,...|  0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   1.0| 1.0|   3.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|
|         0.0|      2.0| 1.0|  1.0| 1.0|   2.0| 1.0|   2.0| 1.0|1.0|[2.0,1.0,1.0,1.0,...|  0.0|
|         1.0|      5.0| 3.0|  3.0| 3.0|   2.0| 3.0|   4.0| 4.0|1.0|[5.0,3.0,3.0,3.0,...|  1.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 3.0|   3.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|
|         1.0|      8.0| 7.0|  5.0|10.0|   7.0| 9.0|   5.0| 5.0|4.0|[8.0,7.0,5.0,10.0...|  1.0|
|         1.0|      7.0| 4.0|  6.0| 4.0|   6.0| 1.0|   4.0| 3.0|1.0|[7.0,4.0,6.0,4.0,...|  1.0|
|         0.0|      4.0| 1.0|  1.0| 1.0|   2.0| 1.0|   2.0| 1.0|1.0|[4.0,1.0,1.0,1.0,...|  0.0|
|         0.0|      4.0| 1.0|  1.0| 1.0|   2.0| 1.0|   3.0| 1.0|1.0|[4.0,1.0,1.0,1.0,...|  0.0|
|         1.0|     10.0| 7.0|  7.0| 6.0|   4.0|10.0|   4.0| 1.0|2.0|[10.0,7.0,7.0,6.0...|  1.0|
|         0.0|      6.0| 1.0|  1.0| 1.0|   2.0| 1.0|   3.0| 1.0|1.0|[6.0,1.0,1.0,1.0,...|  0.0|
+------------+---------+----+-----+----+------+----+------+----+---+--------------------+-----+
only showing top 20 rows
```



```
+------------+---------+----+-----+----+------+----+------+----+---+--------------------+-----+--------------------+--------------------+----------+
|cancer_class|thickness|size|shape|madh|epsize|bnuc|bchrom|nNuc|mit|            features|label|       rawPrediction|         probability|prediction|
+------------+---------+----+-----+----+------+----+------+----+---+--------------------+-----+--------------------+--------------------+----------+
|         0.0|      1.0| 1.0|  1.0| 1.0|   1.0| 1.0|   2.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[5.15956430979038...|[0.99428860556932...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   1.0| 1.0|   2.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[5.15956430979038...|[0.99428860556932...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   1.0| 1.0|   3.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[4.88229871718381...|[0.99247744702488...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   1.0| 1.0|   3.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[4.88229871718381...|[0.99247744702488...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 1.0|   1.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[5.26929960916807...|[0.99487914377217...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 1.0|   1.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[5.26929960916807...|[0.99487914377217...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 1.0|   1.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[5.26929960916807...|[0.99487914377217...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 1.0|   1.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[5.26929960916807...|[0.99487914377217...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 1.0|   1.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[5.26929960916807...|[0.99487914377217...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 1.0|   1.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[5.26929960916807...|[0.99487914377217...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 1.0|   2.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[4.99203401656150...|[0.99325398211858...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 1.0|   2.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[4.99203401656150...|[0.99325398211858...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 1.0|   2.0| 3.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[4.74802132478210...|[0.99140567173413...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 1.0|   3.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[4.71476842395493...|[0.99111766179519...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 1.0|   3.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[4.71476842395493...|[0.99111766179519...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 1.0|   3.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[4.71476842395493...|[0.99111766179519...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 1.0|   3.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[4.71476842395493...|[0.99111766179519...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 1.0|   3.0| 2.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[4.59276207806523...|[0.98997663106901...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   2.0| 5.0|   1.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[4.10129026316119...|[0.98371817931939...|       0.0|
|         0.0|      1.0| 1.0|  1.0| 1.0|   4.0| 3.0|   1.0| 1.0|1.0|[1.0,1.0,1.0,1.0,...|  0.0|[4.35023434970686...|[0.98726059831436...|       0.0|
+------------+---------+----+-----+----+------+----+------+----+---+--------------------+-----+--------------------+--------------------+----------+
only showing top 20 rows
```

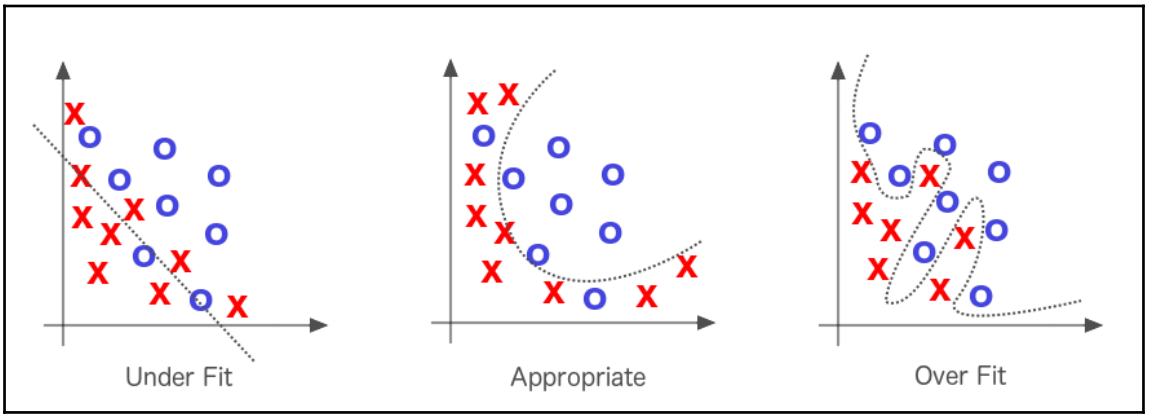


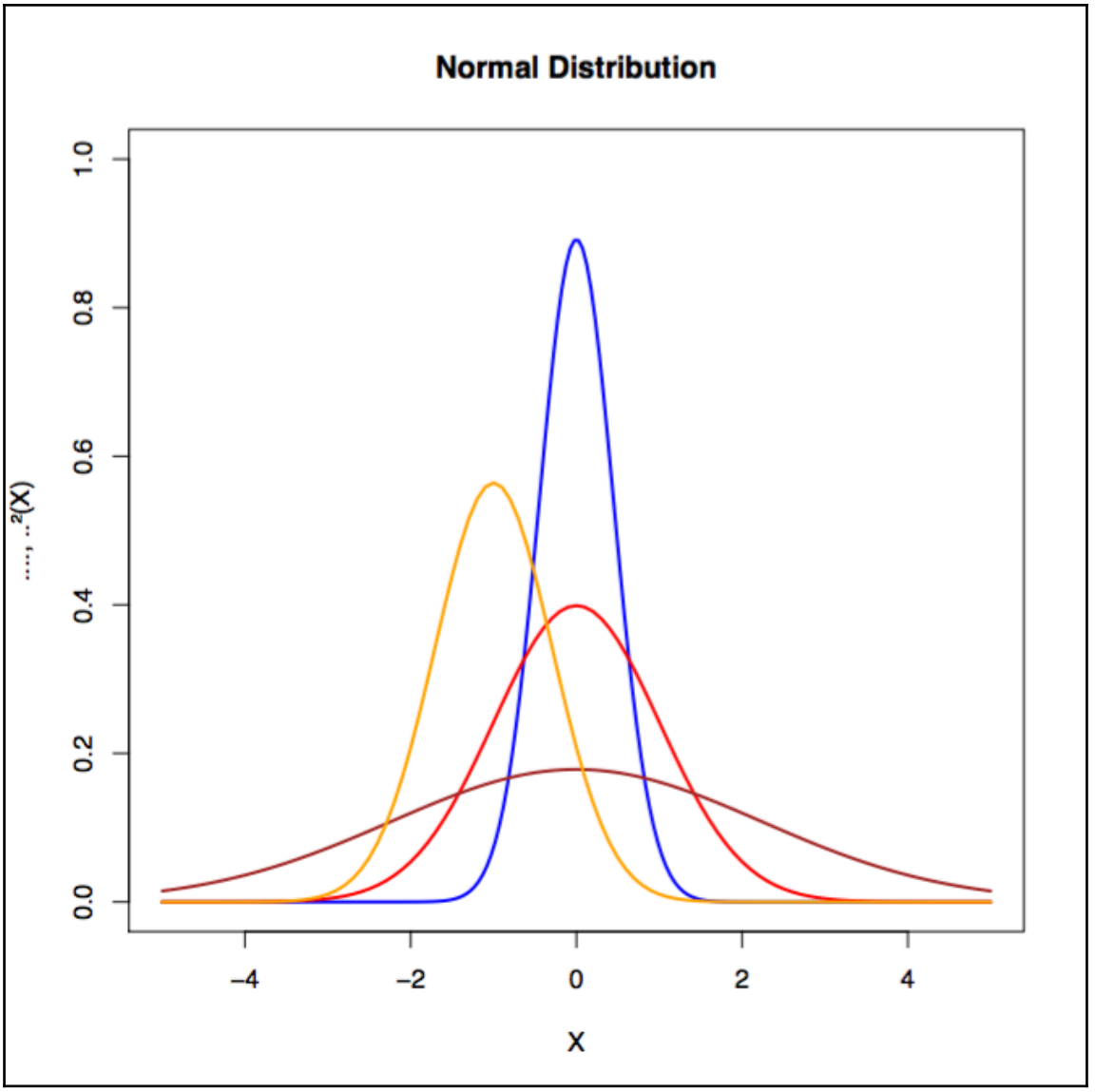
```
1466.0  1.0     4.0     2.0     3.0     11.0    18.0    1.0     11.0    4.0
0.0     1709.0  11.0    3.0     2.0     6.0     1.0     5.0     15.0    4.0
10.0    17.0    1316.0  24.0    22.0    8.0     20.0    17.0    26.0    8.0
3.0     9.0     38.0    1423.0  1.0     52.0    9.0     11.0    31.0    15.0
3.0     4.0     23.0    1.0     1363.0  4.0     10.0    7.0     5.0     43.0
19.0    7.0     11.0    50.0    12.0    1170.0  23.0    6.0     32.0    11.0
6.0     2.0     15.0    3.0     10.0    19.0    1411.0  2.0     8.0     2.0
4.0     7.0     10.0    7.0     14.0    4.0     2.0     1519.0  8.0     48.0
9.0     22.0    26.0    43.0    11.0    46.0    16.0    5.0     1268.0  8.0
6.0     3.0     5.0     23.0    39.0    8.0     0.0     60.0    14.0    1327.0
```

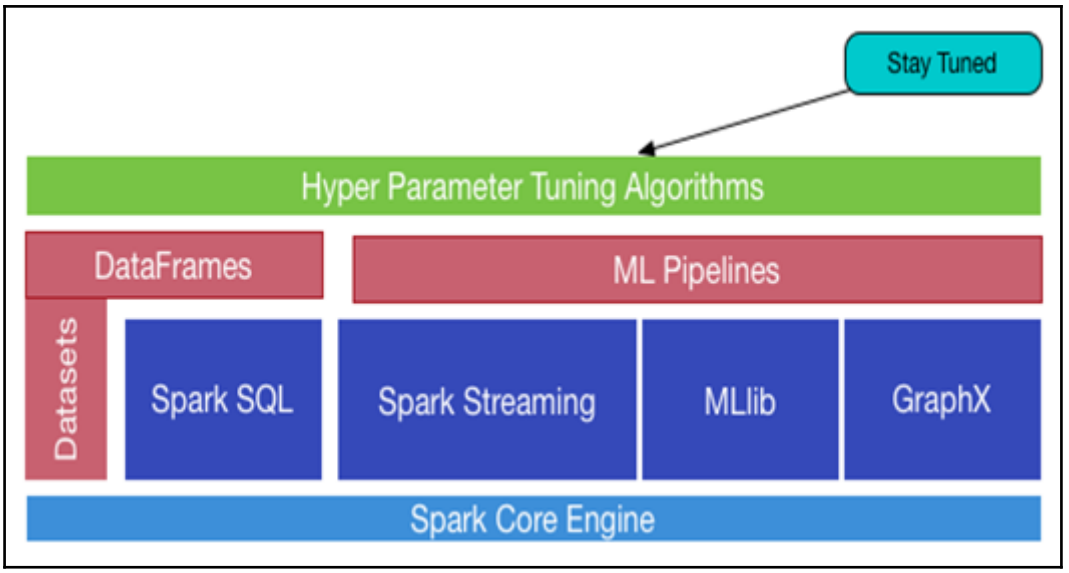


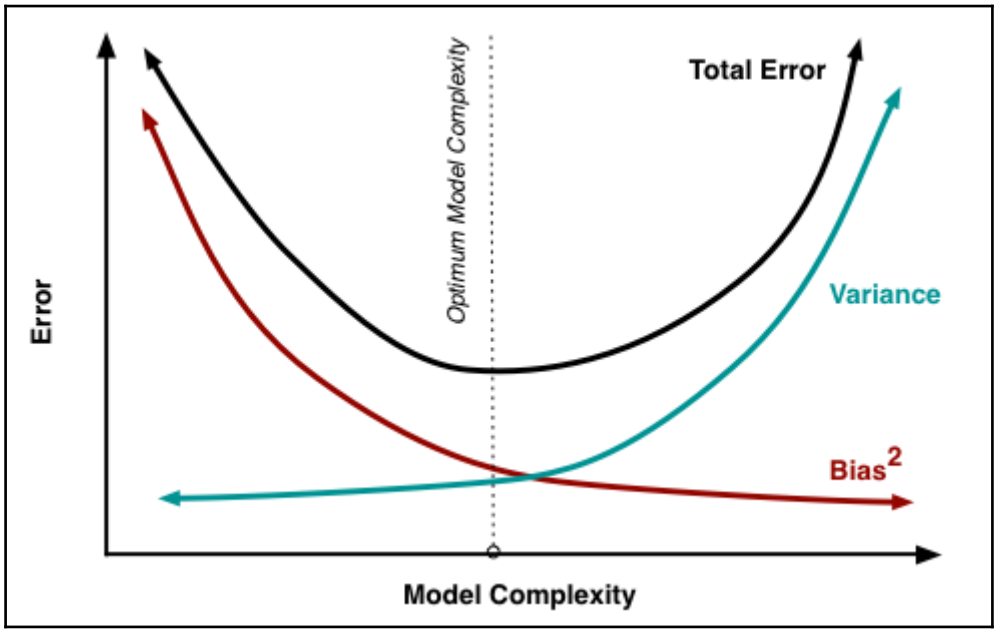
```
1500.0  0.0     8.0     1.0     3.0     6.0     6.0     3.0     2.0     5.0
0.0     1737.0  1.0     3.0     0.0     3.0     1.0     1.0     7.0     2.0
3.0     6.0     1416.0  19.0    5.0     3.0     1.0     9.0     6.0     4.0
0.0     1.0     5.0     1509.0  0.0     21.0    0.0     3.0     18.0    18.0
1.0     3.0     9.0     1.0     1415.0  3.0     2.0     7.0     4.0     17.0
2.0     2.0     0.0     20.0    0.0     1275.0  12.0    0.0     8.0     7.0
4.0     2.0     3.0     2.0     2.0     13.0    1453.0  0.0     8.0     0.0
0.0     3.0     10.0    8.0     4.0     3.0     0.0     1572.0  0.0     11.0
10.0    0.0     11.0    19.0    6.0     12.0    3.0     7.0     1388.0  14.0
1.0     2.0     5.0     10.0    28.0    2.0     0.0     21.0    13.0    1407.0
```

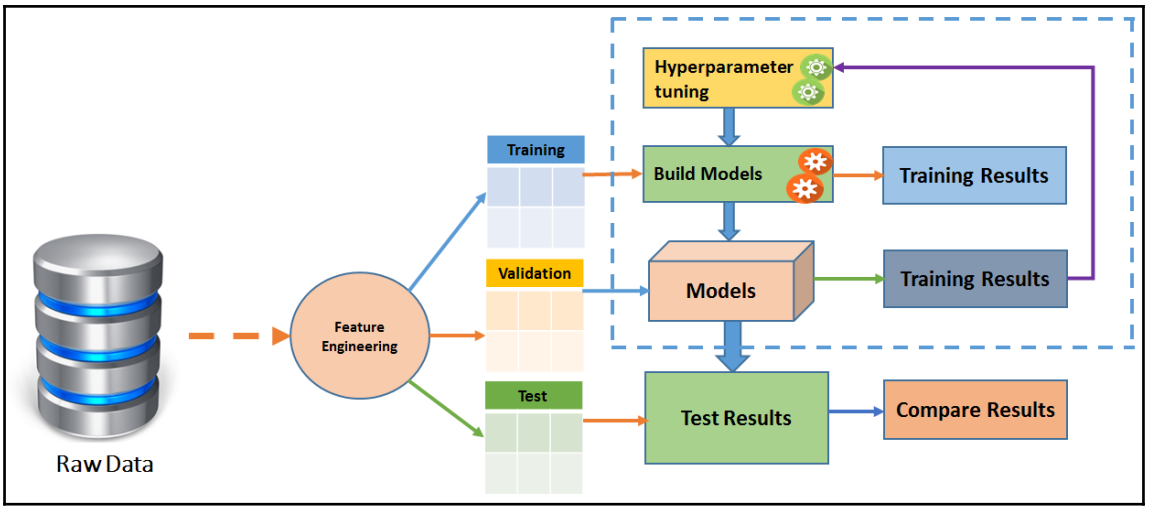
# Chapter 12: Advanced Machine Learning Best Practices

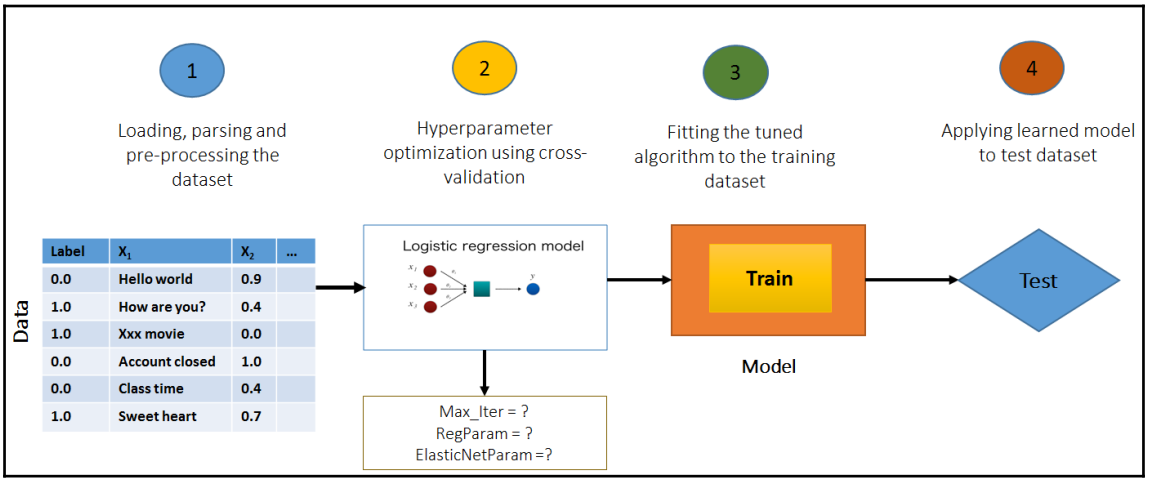



Normal Distribution
1.0
0.8
0.6
0.4
0.2
0.0
–4
–2
0
2
4
X


Stay Tuned
Hyper Parameter Tuning Algorithms
DataFrames
ML Pipelines
Datasets
Spark SQL
Spark Streaming
MLlib
GraphX
Spark Core Engine


Total Error
Optimum Model Complexity
Variance
Bias$^2$
Error
Model Complexity


Hyperparameter tuning
Training
Build Models
Training Results
Validation
Models
Training Results
Feature Engineering
Test
Test Results
Compare Results
Raw Data


1
Loading, parsing and pre-processing the dataset
2
Hyperparameter optimization using cross-validation
3
Fitting the tuned algorithm to the training dataset
4
Applying learned model to test dataset
Data
Label
X1
X2
...
0.0 Hello world 0.9
1.0 How are you? 0.4
1.0 Xxx movie 0.0
0.0 Account closed 1.0
0.0 Class time 0.4
1.0 Sweet heart 0.7
Logistic regression model
Max_Iter = ?
RegParam = ?
ElasticNetParam =?
Train
Model
Test

| creditability | balance | duration | history | purpose | amount | savings | employment | instPercent | sexMarried | guarantors | residenceDuration | assets | age | concCredit | apartment | credits | occupation | dependents | hasPhone | foreign |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.0 | 0.0 | 18.0 | 4.0 | 2.0 | 1049.0 | 0.0 | 1.0 | 4.0 | 1.0 | 0.0 | 3.0 | 1.0 | 21.0 | 2.0 | 0.0 | 0.0 | 2.0 | 0.0 | 0.0 | 0.0 |
| 1.0 | 0.0 | 9.0 | 4.0 | 0.0 | 2799.0 | 0.0 | 2.0 | 2.0 | 2.0 | 0.0 | 1.0 | 0.0 | 36.0 | 2.0 | 0.0 | 1.0 | 2.0 | 1.0 | 0.0 | 0.0 |
| 1.0 | 1.0 | 12.0 | 2.0 | 9.0 | 841.0 | 1.0 | 3.0 | 2.0 | 1.0 | 0.0 | 3.0 | 0.0 | 23.0 | 2.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 1.0 | 0.0 | 12.0 | 4.0 | 0.0 | 2122.0 | 0.0 | 2.0 | 3.0 | 2.0 | 0.0 | 1.0 | 0.0 | 39.0 | 2.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | 1.0 |
| 1.0 | 0.0 | 12.0 | 4.0 | 0.0 | 2171.0 | 0.0 | 2.0 | 4.0 | 2.0 | 0.0 | 3.0 | 1.0 | 38.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 |
| 1.0 | 0.0 | 10.0 | 4.0 | 0.0 | 2241.0 | 0.0 | 1.0 | 1.0 | 2.0 | 0.0 | 2.0 | 0.0 | 48.0 | 2.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | 1.0 |
| 1.0 | 0.0 | 8.0 | 4.0 | 0.0 | 3398.0 | 0.0 | 3.0 | 1.0 | 2.0 | 0.0 | 3.0 | 0.0 | 39.0 | 2.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 |
| 1.0 | 0.0 | 6.0 | 4.0 | 0.0 | 1361.0 | 0.0 | 1.0 | 2.0 | 2.0 | 0.0 | 3.0 | 0.0 | 40.0 | 2.0 | 1.0 | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 |
| 1.0 | 3.0 | 18.0 | 4.0 | 3.0 | 1098.0 | 0.0 | 0.0 | 4.0 | 1.0 | 0.0 | 3.0 | 2.0 | 65.0 | 2.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1.0 | 1.0 | 24.0 | 2.0 | 3.0 | 3758.0 | 2.0 | 0.0 | 1.0 | 1.0 | 0.0 | 3.0 | 3.0 | 23.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1.0 | 0.0 | 11.0 | 4.0 | 0.0 | 3905.0 | 0.0 | 2.0 | 2.0 | 2.0 | 0.0 | 1.0 | 0.0 | 36.0 | 2.0 | 0.0 | 1.0 | 2.0 | 1.0 | 0.0 | 0.0 |
| 1.0 | 0.0 | 30.0 | 4.0 | 1.0 | 6187.0 | 1.0 | 3.0 | 1.0 | 3.0 | 0.0 | 3.0 | 2.0 | 24.0 | 2.0 | 0.0 | 1.0 | 2.0 | 0.0 | 0.0 | 0.0 |
| 1.0 | 0.0 | 6.0 | 4.0 | 3.0 | 1957.0 | 0.0 | 3.0 | 1.0 | 1.0 | 0.0 | 3.0 | 2.0 | 31.0 | 2.0 | 1.0 | 0.0 | 2.0 | 0.0 | 0.0 | 0.0 |
| 1.0 | 1.0 | 48.0 | 3.0 | 10.0 | 7582.0 | 1.0 | 0.0 | 2.0 | 2.0 | 0.0 | 3.0 | 3.0 | 31.0 | 2.0 | 1.0 | 0.0 | 3.0 | 0.0 | 1.0 | 0.0 |
| 1.0 | 0.0 | 18.0 | 2.0 | 3.0 | 1936.0 | 4.0 | 3.0 | 2.0 | 3.0 | 0.0 | 3.0 | 2.0 | 23.0 | 2.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 1.0 | 0.0 | 6.0 | 2.0 | 3.0 | 2647.0 | 2.0 | 2.0 | 2.0 | 2.0 | 0.0 | 2.0 | 0.0 | 44.0 | 2.0 | 0.0 | 0.0 | 2.0 | 1.0 | 0.0 | 0.0 |
| 1.0 | 0.0 | 11.0 | 4.0 | 0.0 | 3939.0 | 0.0 | 2.0 | 1.0 | 2.0 | 0.0 | 1.0 | 0.0 | 40.0 | 2.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 |
| 1.0 | 1.0 | 18.0 | 2.0 | 3.0 | 3213.0 | 2.0 | 1.0 | 1.0 | 3.0 | 0.0 | 2.0 | 0.0 | 25.0 | 2.0 | 0.0 | 0.0 | 2.0 | 0.0 | 0.0 | 0.0 |
| 1.0 | 1.0 | 36.0 | 4.0 | 3.0 | 2337.0 | 0.0 | 4.0 | 4.0 | 2.0 | 0.0 | 3.0 | 0.0 | 36.0 | 2.0 | 1.0 | 0.0 | 2.0 | 0.0 | 0.0 | 0.0 |
| 1.0 | 3.0 | 11.0 | 4.0 | 0.0 | 7228.0 | 0.0 | 2.0 | 1.0 | 2.0 | 0.0 | 3.0 | 1.0 | 39.0 | 2.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 |

only showing top 20 rows

| creditability | avgbalance | avgamt | avgdur |
|---|---|---|---|
| 0.0 | 0.9033333333333333 | 3938.1266666666666 | 24.86 |
| 1.0 | 1.8657142857142857 | 2985.442857142857 | 19.207142857142856 |

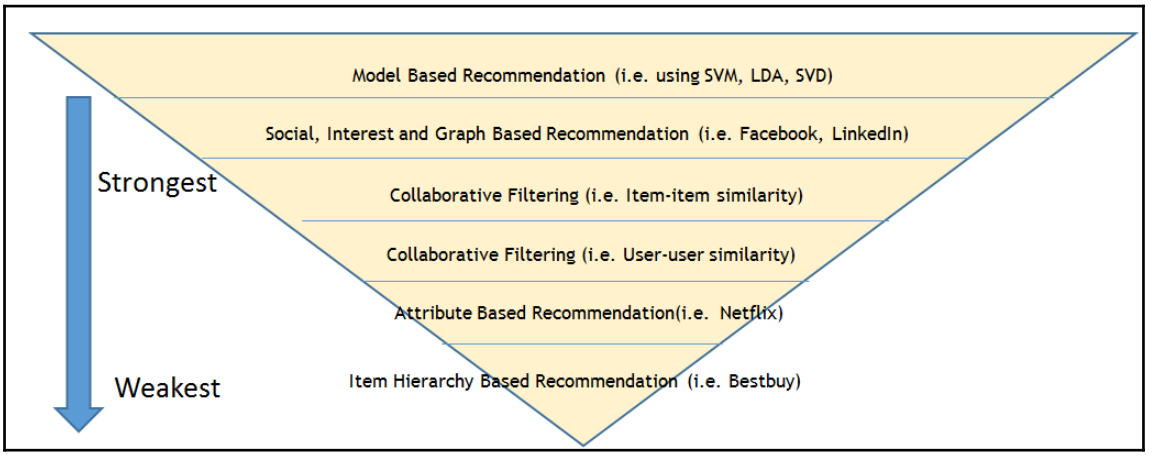
| summary | balance |
|---|---|
| count | 1000 |
| mean | 1.577 |
| stddev | 1.257637727110893 |
| min | 0.0 |
| max | 3.0 |

| creditability | avg(balance) |
|---|---|
| 0.0 | 0.9033333333333333 |
| 1.0 | 1.8657142857142857 |

```
+--------------------+
|            features|
+--------------------+
|(20,[1,2,3,4,6,7,...|
|(20,[1,2,4,6,7,8,...|
|[1.0,12.0,2.0,9.0...|
|[0.0,12.0,4.0,0.0...|
|[0.0,12.0,4.0,0.0...|
|[0.0,10.0,4.0,0.0...|
|[0.0,8.0,4.0,0.0,...|
|[0.0,6.0,4.0,0.0,...|
|[3.0,18.0,4.0,3.0...|
|(20,[0,1,2,3,4,5,...|
|(20,[1,2,4,6,7,8,...|
|[0.0,30.0,4.0,1.0...|
|[0.0,6.0,4.0,3.0,...|
|[1.0,48.0,3.0,10....|
|[0.0,18.0,2.0,3.0...|
|[0.0,6.0,2.0,3.0,...|
|[0.0,11.0,4.0,0.0...|
|[1.0,18.0,2.0,3.0...|
|[1.0,36.0,4.0,3.0...|
|[3.0,11.0,4.0,0.0...|
+--------------------+
only showing top 20 rows
```

```
+-----+--------------------+
|label|            features|
+-----+--------------------+
|  0.0|(20,[1,2,3,4,6,7,...|
|  0.0|(20,[1,2,4,6,7,8,...|
|  0.0|[1.0,12.0,2.0,9.0...|
|  0.0|[0.0,12.0,4.0,0.0...|
|  0.0|[0.0,12.0,4.0,0.0...|
|  0.0|[0.0,10.0,4.0,0.0...|
|  0.0|[0.0,8.0,4.0,0.0,...|
|  0.0|[0.0,6.0,4.0,0.0,...|
|  0.0|[3.0,18.0,4.0,3.0...|
|  0.0|(20,[0,1,2,3,4,5,...|
|  0.0|(20,[1,2,4,6,7,8,...|
|  0.0|[0.0,30.0,4.0,1.0...|
|  0.0|[0.0,6.0,4.0,3.0,...|
|  0.0|[1.0,48.0,3.0,10....|
|  0.0|[0.0,18.0,2.0,3.0...|
|  0.0|[0.0,6.0,2.0,3.0,...|
|  0.0|[0.0,11.0,4.0,0.0...|
|  0.0|[1.0,18.0,2.0,3.0...|
|  0.0|[1.0,36.0,4.0,3.0...|
|  0.0|[3.0,11.0,4.0,0.0...|
+-----+--------------------+
only showing top 20 rows
```

| label | rawPrediction | probability | prediction |
|---|---|---|---|
| 1.0 | [21.0,9.0] | [0.7,0.3] | 0.0 |
| 0.0 | [28.9868421052631... | [0.96622807017543... | 0.0 |
| 0.0 | [18.0,12.0] | [0.6,0.4] | 0.0 |
| 0.0 | [23.9873417721519... | [0.79957805907173... | 0.0 |
| 0.0 | [24.6540084388185... | [0.82180028129395... | 0.0 |
| 0.0 | [22.9868421052631... | [0.76622807017543... | 0.0 |
| 0.0 | [14.5952380952380... | [0.48650793650793... | 1.0 |
| 0.0 | [17.9547224224945... | [0.59849074741648... | 0.0 |
| 0.0 | [23.9684210526315... | [0.79894736842105... | 0.0 |
| 0.0 | [25.0,5.0] | [0.83333333333333... | 0.0 |
| 0.0 | [15.5,14.5] | [0.51666666666666... | 0.0 |
| 0.0 | [22.5,7.5] | [0.75,0.25] | 0.0 |
| 0.0 | [22.9486422749787... | [0.76495474249929... | 0.0 |
| 0.0 | [18.0,12.0] | [0.6,0.4] | 0.0 |
| 0.0 | [27.9631948664260... | [0.93210649554753... | 0.0 |
| 0.0 | [21.0,9.0] | [0.7,0.3] | 0.0 |
| 0.0 | [24.0,6.0] | [0.8,0.2] | 0.0 |
| 0.0 | [16.0,14.0] | [0.53333333333333... | 0.0 |
| 0.0 | [23.9921259842519... | [0.79973753280839... | 0.0 |
| 0.0 | [14.9890109890109... | [0.49963369963369... | 1.0 |



```
+------+-------+------+----------+
|userId|movieId|rating|timestamp |
+------+-------+------+----------+
|1     |16     |4.0   |1217897793|
|1     |24     |1.5   |1217895807|
|1     |32     |4.0   |1217896246|
|1     |47     |4.0   |1217896556|
|1     |50     |4.0   |1217896523|
|1     |110    |4.0   |1217896150|
|1     |150    |3.0   |1217895940|
|1     |161    |4.0   |1217897864|
|1     |165    |3.0   |1217897135|
|1     |204    |0.5   |1217895786|
|1     |223    |4.0   |1217897795|
|1     |256    |0.5   |1217895764|
|1     |260    |4.5   |1217895864|
|1     |261    |1.5   |1217895750|
|1     |277    |0.5   |1217895772|
|1     |296    |4.0   |1217896125|
|1     |318    |4.0   |1217895860|
|1     |349    |4.5   |1217897058|
|1     |356    |3.0   |1217896231|
|1     |377    |2.5   |1217896373|
+------+-------+------+----------+
only showing top 20 rows
```

```
+-------+-----------------------------------+-----------------------------------------+
|movieId|title                              |genres                                   |
+-------+-----------------------------------+-----------------------------------------+
|1      |Toy Story (1995)                   |Adventure|Animation|Children|Comedy|Fantasy|
|2      |Jumanji (1995)                     |Adventure|Children|Fantasy                |
|3      |Grumpier Old Men (1995)            |Comedy|Romance                           |
|4      |Waiting to Exhale (1995)           |Comedy|Drama|Romance                     |
|5      |Father of the Bride Part II (1995) |Comedy                                   |
|6      |Heat (1995)                        |Action|Crime|Thriller                    |
|7      |Sabrina (1995)                     |Comedy|Romance                           |
|8      |Tom and Huck (1995)                |Adventure|Children                        |
|9      |Sudden Death (1995)                |Action                                   |
|10     |GoldenEye (1995)                   |Action|Adventure|Thriller                |
|11     |American President, The (1995)     |Comedy|Drama|Romance                     |
|12     |Dracula: Dead and Loving It (1995) |Comedy|Horror                            |
|13     |Balto (1995)                       |Adventure|Animation|Children              |
|14     |Nixon (1995)                       |Drama                                    |
|15     |Cutthroat Island (1995)            |Action|Adventure|Romance                 |
|16     |Casino (1995)                      |Crime|Drama                              |
|17     |Sense and Sensibility (1995)       |Drama|Romance                            |
|18     |Four Rooms (1995)                  |Comedy                                   |
|19     |Ace Ventura: When Nature Calls (1995)|Comedy                                   |
|20     |Money Train (1995)                 |Action|Comedy|Crime|Drama|Thriller        |
+-------+-----------------------------------+-----------------------------------------+
only showing top 20 rows
```

```
Got 105339 ratings from 668 users on 10325 movies.
+-------------------------------------------------------------------------+----+----+----+
|title                                                                    |maxr|minr|cntu|
+-------------------------------------------------------------------------+----+----+----+
|Pulp Fiction (1994)                                                      |5.0 |0.5 |325 |
|Forrest Gump (1994)                                                      |5.0 |0.5 |311 |
|Shawshank Redemption, The (1994)                                         |5.0 |0.5 |308 |
|Jurassic Park (1993)                                                     |5.0 |1.0 |294 |
|Silence of the Lambs, The (1991)                                         |5.0 |0.5 |290 |
|Star Wars: Episode IV - A New Hope (1977)                                |5.0 |0.5 |273 |
|Matrix, The (1999)                                                       |5.0 |0.5 |261 |
|Terminator 2: Judgment Day (1991)                                        |5.0 |0.5 |253 |
|Braveheart (1995)                                                        |5.0 |0.5 |248 |
|Schindler's List (1993)                                                  |5.0 |0.5 |248 |
|Fugitive, The (1993)                                                     |5.0 |1.0 |244 |
|Toy Story (1995)                                                         |5.0 |1.0 |232 |
|Star Wars: Episode V - The Empire Strikes Back (1980)                    |5.0 |0.5 |228 |
|Usual Suspects, The (1995)                                               |5.0 |1.0 |228 |
|Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)|5.0 |1.0 |224 |
|Star Wars: Episode VI - Return of the Jedi (1983)                        |5.0 |0.5 |222 |
|Batman (1989)                                                            |5.0 |0.5 |217 |
|American Beauty (1999)                                                   |5.0 |1.0 |216 |
|Back to the Future (1985)                                                |5.0 |1.5 |213 |
|Godfather, The (1972)                                                    |5.0 |1.0 |210 |
+-------------------------------------------------------------------------+----+----+----+
only showing top 20 rows
```

```
+------+----+
|userId|ct  |
+------+----+
|668   |5678|
|575   |2837|
|458   |2086|
|232   |1421|
|310   |1287|
|475   |1249|
|128   |1231|
|224   |1182|
|607   |1176|
|63    |1107|
+------+----+
```

```
+------+-------+------+-------------------------------------------+
|userId|movieId|rating|title                                      |
+------+-------+------+-------------------------------------------+
|668   |6      |5.0   |Heat (1995)                                |
|668   |326    |4.5   |To Live (Huozhe) (1994)                    |
|668   |446    |4.5   |Farewell My Concubine (Ba wang bie ji) (1993)|
|668   |515    |4.5   |Remains of the Day, The (1993)             |
|668   |593    |4.5   |Silence of the Lambs, The (1991)           |
|668   |594    |4.5   |Snow White and the Seven Dwarfs (1937)     |
|668   |608    |4.5   |Fargo (1996)                               |
|668   |858    |5.0   |Godfather, The (1972)                      |
|668   |898    |5.0   |Philadelphia Story, The (1940)             |
|668   |907    |4.5   |Gay Divorcee, The (1934)                   |
|668   |908    |5.0   |North by Northwest (1959)                  |
|668   |910    |5.0   |Some Like It Hot (1959)                    |
|668   |912    |5.0   |Casablanca (1942)                          |
|668   |913    |5.0   |Maltese Falcon, The (1941)                 |
|668   |914    |5.0   |My Fair Lady (1964)                        |
|668   |919    |5.0   |Wizard of Oz, The (1939)                   |
|668   |927    |4.5   |Women, The (1939)                          |
|668   |930    |4.5   |Notorious (1946)                           |
|668   |945    |4.5   |Top Hat (1935)                             |
|668   |947    |4.5   |My Man Godfrey (1936)                      |
+------+-------+------+-------------------------------------------+
only showing top 20 rows
```

```
Rating:(UserID, MovieID, Rating)
----------------------------------
Rating(668,101862,4.8525842164777435)
Rating(668,5304,4.8525842164777435)
Rating(668,25961,4.8525842164777435)
Rating(668,80969,4.779325934293423)
Rating(668,93040,4.7528736838886)
Rating(668,25795,4.676957397667861)
----------------------------------
(Prediction, Rating)
(3.848087516442212,3.5)
(4.647813269020743,5.0)
(3.578002886107389,4.0)
(3.681217214985231,3.0)
(2.844685318141285,3.0)
```

```
      Topic: 0
Terms | Index | Weight
-----------------------
space    10665   0.046582
just     10667   0.034397
posted   10637   0.016093
love     10661   0.015652
photo    10639   0.013296
cosmic   10635   0.013212
angry    10656   0.012860
like     10666   0.012629
life     10640   0.012107
time     10662   0.011634
-------------------
Sum:= 0.188459750219041
```

```
      Topic: 1
Terms | Index | Weight
-----------------------
smile    10668   0.129227
just     10667   0.024922
good     10663   0.022404
hope     10645   0.017981
going    10655   0.015764
thanks   10648   0.014945
time     10662   0.014941
like     10666   0.014827
think    10659   0.014438
work     10649   0.012702
--------------------
Sum:= 0.28215004471848354
```

```
      Topic: 2
Terms | Index | Weight
-----------------------
grin     10664   0.078958
yang     10628   0.029173
kita     10574   0.017318
disgust          10618   0.016325
udah     10544   0.014584
science          10590   0.012792
space    10665   0.011765
nggak    10501   0.011290
kalo     10476   0.010203
angry    10656   0.009313
--------------------
Sum:= 0.21172148557919923
```

```
      Topic: 3
Terms | Index | Weight
-----------------------
like     10666   0.030890
just     10667   0.020093
know     10660   0.016473
good     10663   0.013343
that     10651   0.012687
people   10658   0.012137
right    10654   0.012097
think    10659   0.011395
love     10661   0.010943
does     10646   0.009002
--------------------
Sum:= 0.14905966677477597
```

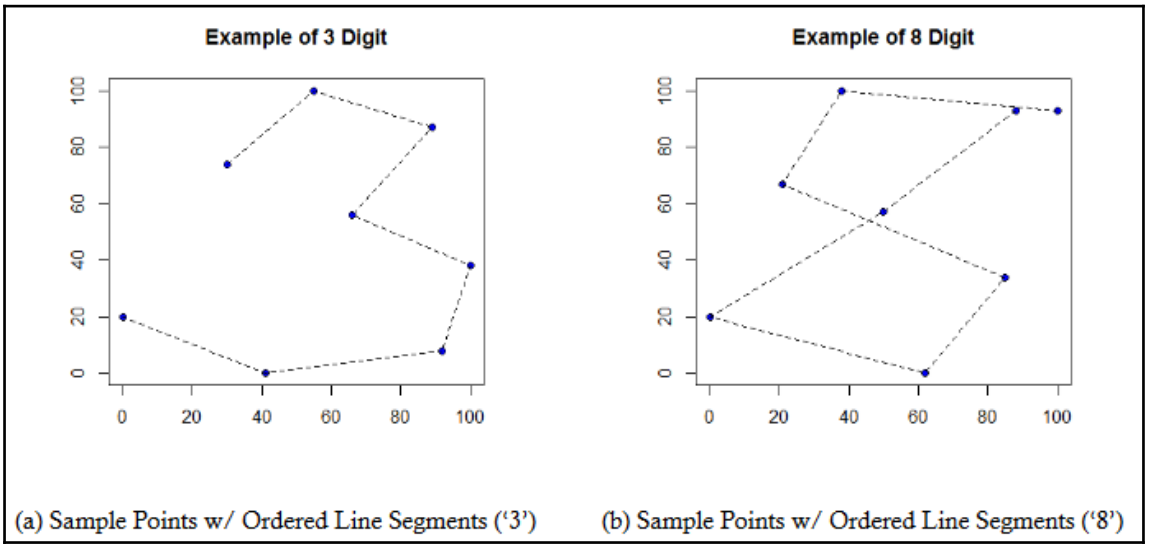
```+--------------------+
|                docs|
+--------------------+
|20 000 2000010 th...|
|fifty two  still ...|
|please this   loo...|
|gable this       ...|
|Sheridan   world ...|
|  wiretap with Ca...|
|                 ...|
|find  plea Lamb h...|
|  find  10th  aes...|
|Disney  alad10 tx...|
| Alcott  Gutenber...|
|   this  hand dom...|
|please this   loo...|
|  HORATIO breaker...|
| HORATIO  ragged ...|
| HORATIO upward s...|
| RESEARCH Electro...|
|tradition Richard...|
+--------------------+```
```

# Chapter 13: My Name is Bayes, Naive Bayes

| letter | xbox | ybox | width | height | onpix | xbar | ybar | x2bar | y2bar | xybar | x2ybar | xy2bar | xedge | xedgey | yedge | yedgex |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T | 2 | 8 | 3 | 5 | 1 | 8 | 13 | 0 | 6 | 6 | 10 | 8 | 0 | 8 | 0 | 8 |
| I | 5 | 12 | 3 | 7 | 2 | 10 | 5 | 5 | 4 | 13 | 3 | 9 | 2 | 8 | 4 | 10 |
| D | 4 | 11 | 6 | 8 | 6 | 10 | 6 | 2 | 6 | 10 | 3 | 7 | 3 | 7 | 3 | 9 |
| N | 7 | 11 | 6 | 6 | 3 | 5 | 9 | 4 | 6 | 4 | 4 | 10 | 6 | 10 | 2 | 8 |
| G | 2 | 1 | 3 | 1 | 1 | 8 | 6 | 6 | 6 | 6 | 5 | 9 | 1 | 7 | 5 | 10 |
| S | 4 | 11 | 5 | 8 | 3 | 8 | 8 | 6 | 9 | 5 | 6 | 6 | 0 | 8 | 9 | 7 |
| B | 4 | 2 | 5 | 4 | 4 | 8 | 7 | 6 | 6 | 7 | 6 | 6 | 2 | 8 | 7 | 10 |
| A | 1 | 1 | 3 | 2 | 1 | 8 | 2 | 2 | 2 | 8 | 2 | 8 | 1 | 6 | 2 | 7 |
| J | 2 | 2 | 4 | 4 | 2 | 10 | 6 | 2 | 6 | 12 | 4 | 8 | 1 | 6 | 1 | 7 |
| M | 11 | 15 | 13 | 9 | 7 | 13 | 2 | 6 | 2 | 12 | 1 | 9 | 8 | 1 | 1 | 8 |
| X | 3 | 9 | 5 | 7 | 4 | 8 | 7 | 3 | 8 | 5 | 6 | 8 | 2 | 8 | 6 | 7 |
| O | 6 | 13 | 4 | 7 | 4 | 6 | 7 | 6 | 3 | 10 | 7 | 9 | 5 | 9 | 5 | 8 |
| G | 4 | 9 | 6 | 7 | 6 | 7 | 8 | 6 | 2 | 6 | 5 | 11 | 4 | 8 | 7 | 8 |
| M | 6 | 9 | 8 | 6 | 9 | 7 | 8 | 6 | 5 | 7 | 5 | 8 | 8 | 9 | 8 | 6 |
| R | 5 | 9 | 5 | 7 | 6 | 6 | 11 | 7 | 3 | 7 | 3 | 9 | 2 | 7 | 5 | 11 |
| F | 6 | 9 | 5 | 4 | 3 | 10 | 6 | 3 | 5 | 10 | 5 | 7 | 3 | 9 | 6 | 9 |
| O | 3 | 4 | 4 | 3 | 2 | 8 | 7 | 7 | 5 | 7 | 6 | 8 | 2 | 8 | 3 | 8 |
| C | 7 | 10 | 5 | 5 | 2 | 6 | 8 | 6 | 8 | 11 | 7 | 11 | 2 | 8 | 5 | 9 |
| T | 6 | 11 | 6 | 8 | 5 | 6 | 11 | 5 | 6 | 11 | 9 | 4 | 3 | 12 | 2 | 4 |
| J | 2 | 2 | 3 | 3 | 1 | 10 | 6 | 3 | 6 | 12 | 4 | 9 | 0 | 7 | 1 | 7 |

only showing top 20 rows

```
+-----+--------------------+
|label|            features|
+-----+--------------------+
|  8.0|(17,[0,1,2,3,4,5,...|
| 10.0|(17,[0,1,2,3,4,5,...|
|  9.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
| 10.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
|  5.0|(17,[0,1,2,3,4,5,...|
|  6.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
|  7.0|(17,[0,1,2,3,4,5,...|
|  6.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
|  9.0|(17,[0,1,2,3,4,5,...|
|  4.0|(17,[0,1,2,3,4,5,...|
|  7.0|(17,[0,1,2,3,4,5,...|
|  7.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
+-----+--------------------+
only showing top 20 rows
```

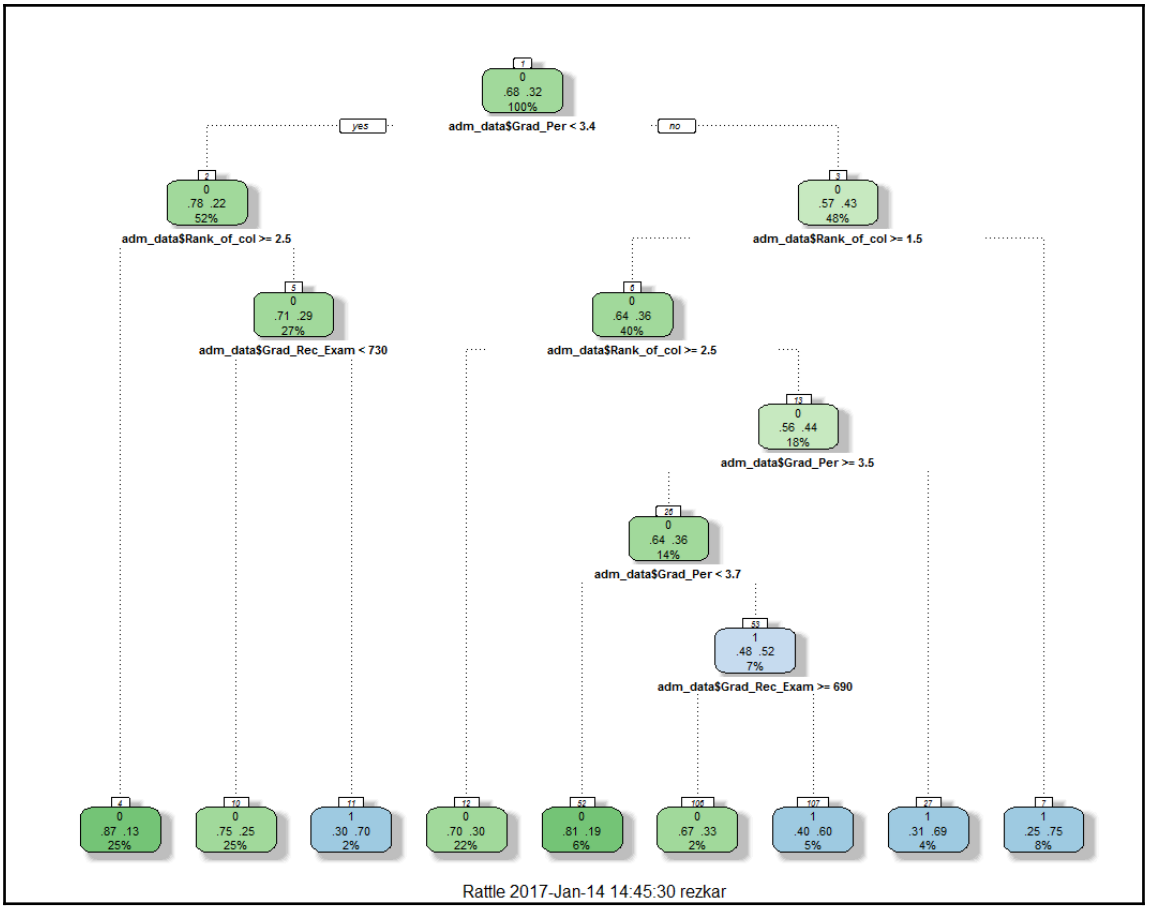


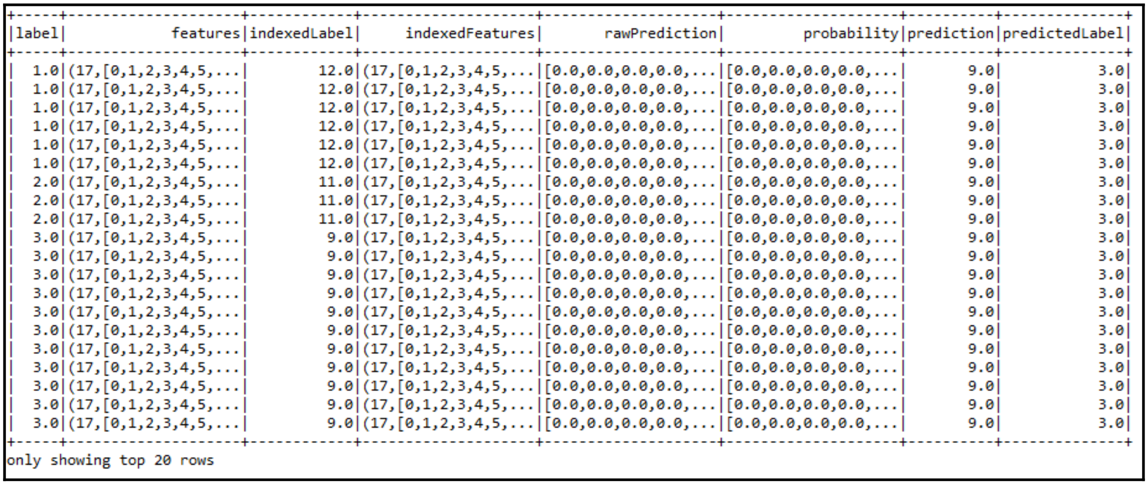

(a) Sample Points w/ Ordered Line Segments ('3') (b) Sample Points w/ Ordered Line Segments ('8')

```
+-----+--------------------+
|label|            features|
+-----+--------------------+
|  8.0|(16,[0,1,2,3,4,5,...|
|  2.0|(16,[1,2,3,4,5,6,...|
|  1.0|(16,[1,2,3,4,5,6,...|
|  4.0|(16,[1,2,3,4,5,6,...|
|  1.0|(16,[1,2,3,4,5,6,...|
|  6.0|(16,[0,1,2,3,4,5,...|
|  4.0|(16,[1,2,3,4,5,6,...|
|  0.0|(16,[1,2,3,4,5,6,...|
|  5.0|(16,[0,1,2,3,4,5,...|
|  0.0|(16,[0,1,2,3,5,6,...|
|  9.0|(16,[0,1,2,3,5,6,...|
|  8.0|(16,[0,1,2,3,4,5,...|
|  5.0|(16,[0,1,2,3,4,5,...|
|  9.0|(16,[0,1,2,3,5,6,...|
|  7.0|(16,[1,2,3,4,5,6,...|
|  3.0|(16,[0,1,2,3,4,5,...|
|  3.0|(16,[0,1,2,3,4,5,...|
|  9.0|(16,[0,1,2,3,4,5,...|
|  2.0|(16,[0,1,2,3,4,5,...|
|  2.0|(16,[1,2,3,4,5,6,...|
+-----+--------------------+
only showing top 20 rows
```

```
+-----+--------------------+--------------------+--------------------+----------+
|label|            features|       rawPrediction|         probability|prediction|
+-----+--------------------+--------------------+--------------------+----------+
|  0.0|(16,[0,1,2,3,4,5,...|[-2439.0893277449...|[1.32132340702018...|       4.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-1941.7868705353...|[1.0,1.5395790656...|       0.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-2024.4356335162...|[1.0,1.6764090944...|       0.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-1989.5775697073...|[1.0,2.2647494021...|       0.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-1706.6857288506...|[1.0,5.1940219699...|       0.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-1838.2628605334...|[1.0,7.2364926581...|       0.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-2168.4931444350...|[1.0,6.8428584454...|       0.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-2068.2067411172...|[1.0,1.1943331620...|       0.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-2132.6929489447...|[1.0,1.9943684266...|       0.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-1983.0451148771...|[1.0,4.9959906892...|       0.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-2049.2850893323...|[1.0,1.3644883115...|       0.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-1971.1755138520...|[1.0,1.6415723270...|       0.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-2216.9188759036...|[1.0,1.3805417667...|       0.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-2216.0583349043...|[1.0,7.7430733808...|       0.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-2290.1517462265...|[1.0,1.3312677171...|       0.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-2268.9492946577...|[0.01491770995335...|       6.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-2377.8867352336...|[1.27336913041488...|       8.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-2206.2037445466...|[1.20068275169939...|       6.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-2290.1662968738...|[2.82560057752915...|       8.0|
|  0.0|(16,[0,1,2,3,4,5,...|[-2662.3029788480...|[2.38039426503477...|       8.0|
+-----+--------------------+--------------------+--------------------+----------+
only showing top 20 rows
```

```
+-----+--------------------+
|label|            features|
+-----+--------------------+
|  0.0|(8287348,[592160,...|
|  0.0|(8287348,[592137,...|
|  0.0|(8287348,[592137,...|
|  0.0|(8287348,[598864,...|
|  0.0|(8287348,[670767,...|
|  0.0|(8287348,[592137,...|
|  0.0|(8287348,[592137,...|
|  1.0|(8287348,[657980,...|
|  0.0|(8287348,[592208,...|
|  0.0|(8287348,[663584,...|
|  1.0|(8287348,[592137,...|
|  1.0|(8287348,[592137,...|
|  0.0|(8287348,[592137,...|
|  1.0|(8287348,[592137,...|
|  1.0|(8287348,[657930,...|
|  0.0|(8287348,[670767,...|
|  1.0|(8287348,[598111,...|
|  0.0|(8287348,[592137,...|
|  0.0|(8287348,[592188,...|
|  0.0|(8287348,[592137,...|
+-----+--------------------+
only showing top 20 rows
```

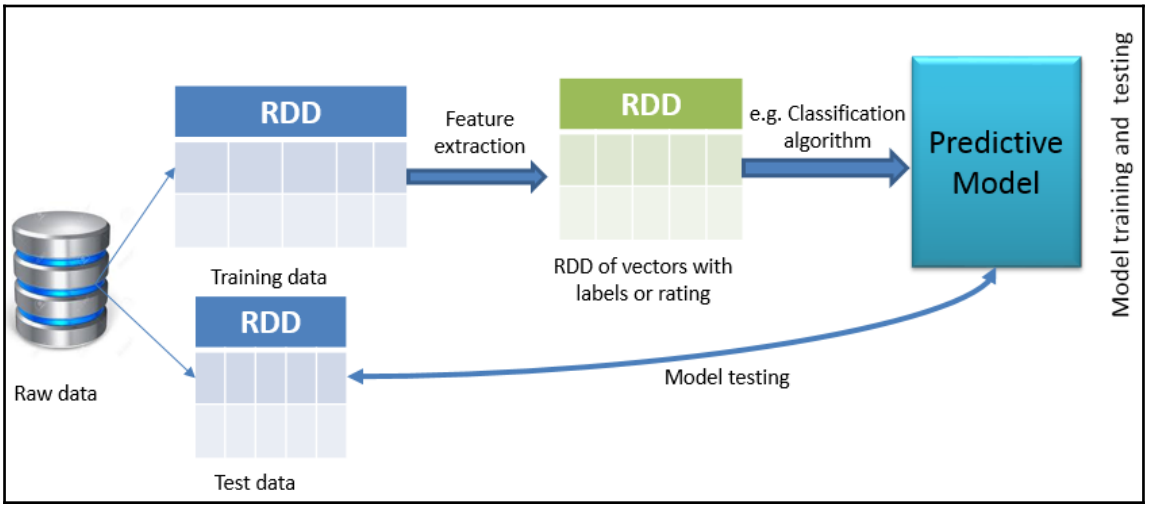

1
0
.68 .32
100%
adm_data$Grad_Per < 3.4
yes
no
2
0
.78 .22
52%
adm_data$Rank_of_col >= 2.5
3
0
.57 .43
48%
adm_data$Rank_of_col >= 1.5
5
0
.71 .29
27%
adm_data$Grad_Rec_Exam < 730
6
0
.64 .36
40%
adm_data$Rank_of_col >= 2.5
13
0
.56 .44
18%
adm_data$Grad_Per >= 3.5
26
0
.64 .36
14%
adm_data$Grad_Per < 3.7
53
1
.48 .52
7%
adm_data$Grad_Rec_Exam >= 690
4
0
.87 .13
25%
10
0
.75 .25
25%
11
1
.30 .70
2%
12
0
.70 .30
22%
52
0
.81 .19
6%
106
0
.67 .33
2%
107
1
.40 .60
5%
27
1
.31 .69
4%
7
1
.25 .75
8%
Rattle 2017-Jan-14 14:45:30 rezkar

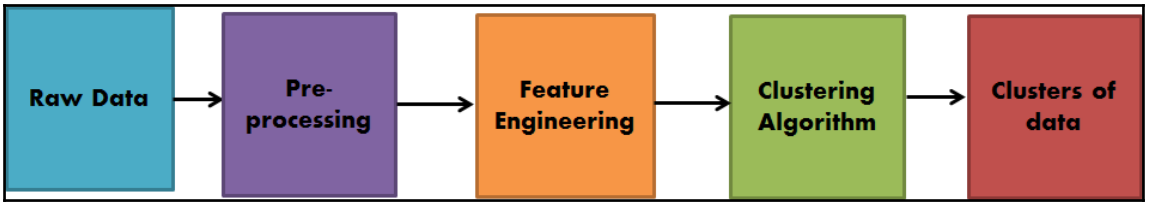


```
+-----+--------------------+------------+--------------------+--------------------+--------------------+----------+--------------+
|label|            features|indexedLabel|     indexedFeatures|       rawPrediction|         probability|prediction|predictedLabel|
+-----+--------------------+------------+--------------------+--------------------+--------------------+----------+--------------+
|  1.0|(17,[0,1,2,3,4,5,...|        12.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  1.0|(17,[0,1,2,3,4,5,...|        12.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  1.0|(17,[0,1,2,3,4,5,...|        12.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  1.0|(17,[0,1,2,3,4,5,...|        12.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  1.0|(17,[0,1,2,3,4,5,...|        12.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  1.0|(17,[0,1,2,3,4,5,...|        12.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  2.0|(17,[0,1,2,3,4,5,...|        11.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  2.0|(17,[0,1,2,3,4,5,...|        11.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  2.0|(17,[0,1,2,3,4,5,...|        11.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  3.0|(17,[0,1,2,3,4,5,...|         9.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  3.0|(17,[0,1,2,3,4,5,...|         9.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  3.0|(17,[0,1,2,3,4,5,...|         9.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  3.0|(17,[0,1,2,3,4,5,...|         9.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  3.0|(17,[0,1,2,3,4,5,...|         9.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  3.0|(17,[0,1,2,3,4,5,...|         9.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  3.0|(17,[0,1,2,3,4,5,...|         9.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  3.0|(17,[0,1,2,3,4,5,...|         9.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  3.0|(17,[0,1,2,3,4,5,...|         9.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  3.0|(17,[0,1,2,3,4,5,...|         9.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
|  3.0|(17,[0,1,2,3,4,5,...|         9.0|(17,[0,1,2,3,4,5,...|[0.0,0.0,0.0,0.0,...|[0.0,0.0,0.0,0.0,...|       9.0|           3.0|
+-----+--------------------+------------+--------------------+--------------------+--------------------+----------+--------------+
only showing top 20 rows
```

```
Learned classification tree model:
DecisionTreeClassificationModel (uid=dtc_fbc6a27aa70b) of depth 5 with 19 nodes
  If (feature 16 <= 7.0)
   If (feature 16 <= 6.0)
    If (feature 16 <= 5.0)
     If (feature 16 <= 4.0)
      If (feature 16 <= 3.0)
       Predict: 9.0
      Else (feature 16 > 3.0)
       Predict: 7.0
     Else (feature 16 > 4.0)
      Predict: 5.0
    Else (feature 16 > 5.0)
     Predict: 3.0
   Else (feature 16 > 6.0)
    Predict: 1.0
  Else (feature 16 > 7.0)
   If (feature 16 <= 8.0)
    Predict: 0.0
   Else (feature 16 > 8.0)
    If (feature 16 <= 9.0)
     Predict: 2.0
    Else (feature 16 > 9.0)
     If (feature 16 <= 10.0)
      Predict: 4.0
     Else (feature 16 > 10.0)
      If (feature 16 <= 11.0)
       Predict: 6.0
      Else (feature 16 > 11.0)
       Predict: 8.0
```

# Chapter 14: Time to Put Some Order - Cluster Your Data with Spark MLlib




| Price | LotSize | Waterfront | Age | LandValue | NewConstruct | CentralAir | FuelType | HeatType | SewerType | LivingArea | PctCollege | Bedrooms | Fireplaces | Bathrooms | rooms |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 132500.0 | 0.09 | 0.0 | 42.0 | 50000.0 | 0.0 | 0.0 | 3.0 | 4.0 | 2.0 | 906.0 | 35.0 | 2.0 | 1.0 | 1.0 | 5.0 |
| 181115.0 | 0.92 | 0.0 | 0.0 | 22300.0 | 0.0 | 0.0 | 2.0 | 3.0 | 2.0 | 1953.0 | 51.0 | 3.0 | 0.0 | 2.5 | 6.0 |
| 109000.0 | 0.19 | 0.0 | 133.0 | 7300.0 | 0.0 | 0.0 | 2.0 | 3.0 | 3.0 | 1944.0 | 51.0 | 4.0 | 1.0 | 1.0 | 8.0 |
| 155000.0 | 0.41 | 0.0 | 13.0 | 18700.0 | 0.0 | 0.0 | 2.0 | 2.0 | 2.0 | 1944.0 | 51.0 | 3.0 | 1.0 | 1.5 | 5.0 |
| 86060.0 | 0.11 | 0.0 | 0.0 | 15000.0 | 1.0 | 1.0 | 2.0 | 2.0 | 3.0 | 840.0 | 51.0 | 2.0 | 0.0 | 1.0 | 3.0 |
| 120000.0 | 0.68 | 0.0 | 31.0 | 14000.0 | 0.0 | 0.0 | 2.0 | 2.0 | 2.0 | 1152.0 | 22.0 | 4.0 | 1.0 | 1.0 | 8.0 |
| 153000.0 | 0.4 | 0.0 | 33.0 | 23300.0 | 0.0 | 0.0 | 4.0 | 3.0 | 2.0 | 2752.0 | 51.0 | 4.0 | 1.0 | 1.5 | 8.0 |
| 170000.0 | 1.21 | 0.0 | 23.0 | 14600.0 | 0.0 | 0.0 | 4.0 | 2.0 | 2.0 | 1662.0 | 35.0 | 4.0 | 1.0 | 1.5 | 9.0 |
| 90000.0 | 0.83 | 0.0 | 36.0 | 22200.0 | 0.0 | 0.0 | 3.0 | 4.0 | 2.0 | 1632.0 | 51.0 | 3.0 | 0.0 | 1.5 | 8.0 |
| 122900.0 | 1.94 | 0.0 | 4.0 | 21200.0 | 0.0 | 0.0 | 2.0 | 2.0 | 1.0 | 1416.0 | 44.0 | 3.0 | 0.0 | 1.5 | 6.0 |
| 325000.0 | 2.29 | 0.0 | 123.0 | 12600.0 | 0.0 | 0.0 | 4.0 | 2.0 | 2.0 | 2894.0 | 51.0 | 7.0 | 0.0 | 1.0 | 12.0 |
| 120000.0 | 0.92 | 0.0 | 1.0 | 22300.0 | 0.0 | 0.0 | 2.0 | 2.0 | 2.0 | 1624.0 | 51.0 | 3.0 | 0.0 | 2.0 | 6.0 |
| 85860.0 | 8.97 | 0.0 | 13.0 | 4800.0 | 0.0 | 0.0 | 3.0 | 4.0 | 2.0 | 704.0 | 41.0 | 2.0 | 0.0 | 1.0 | 4.0 |
| 97000.0 | 0.11 | 0.0 | 153.0 | 3100.0 | 0.0 | 0.0 | 2.0 | 3.0 | 3.0 | 1383.0 | 57.0 | 3.0 | 0.0 | 2.0 | 5.0 |
| 127000.0 | 0.14 | 0.0 | 9.0 | 300.0 | 0.0 | 0.0 | 4.0 | 2.0 | 2.0 | 1300.0 | 41.0 | 3.0 | 0.0 | 1.5 | 8.0 |
| 89900.0 | 0.0 | 0.0 | 88.0 | 2500.0 | 0.0 | 0.0 | 2.0 | 3.0 | 3.0 | 936.0 | 57.0 | 3.0 | 0.0 | 1.0 | 4.0 |
| 155000.0 | 0.13 | 0.0 | 9.0 | 300.0 | 0.0 | 0.0 | 4.0 | 2.0 | 2.0 | 1300.0 | 41.0 | 3.0 | 0.0 | 1.5 | 7.0 |
| 253750.0 | 2.0 | 0.0 | 0.0 | 49800.0 | 0.0 | 1.0 | 2.0 | 2.0 | 1.0 | 2816.0 | 71.0 | 4.0 | 1.0 | 2.5 | 12.0 |
| 60000.0 | 0.21 | 0.0 | 82.0 | 8500.0 | 0.0 | 0.0 | 4.0 | 3.0 | 2.0 | 924.0 | 35.0 | 2.0 | 0.0 | 1.0 | 6.0 |
| 87500.0 | 0.88 | 0.0 | 17.0 | 19400.0 | 0.0 | 0.0 | 4.0 | 2.0 | 2.0 | 1092.0 | 35.0 | 3.0 | 0.0 | 1.0 | 6.0 |

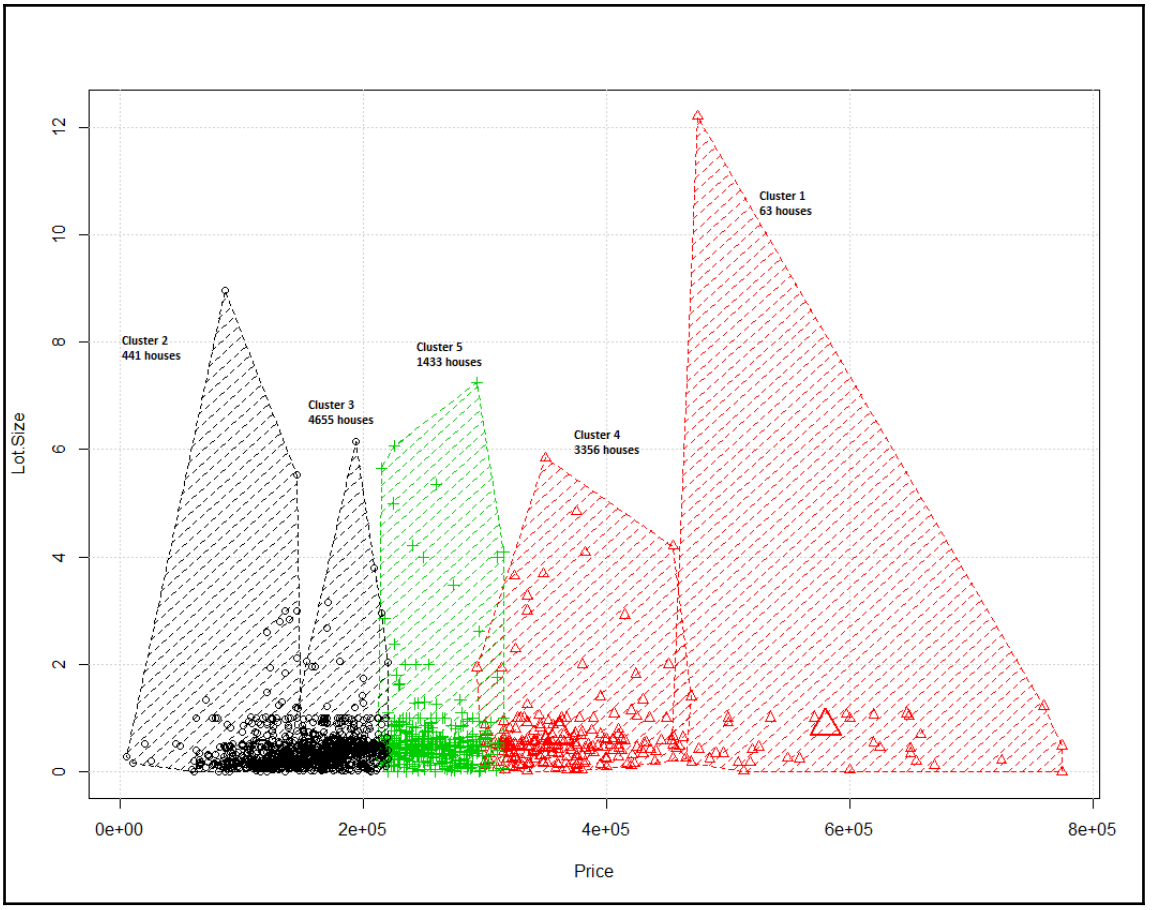
only showing top 20 rows

| Price | CLUSTER |
|---|---|
| 132500.0 | 4 |
| 181115.0 | 3 |
| 109000.0 | 0 |
| 155000.0 | 0 |
| 86060.0 | 0 |
| 120000.0 | 0 |
| 153000.0 | 3 |
| 170000.0 | 0 |
| 90000.0 | 3 |
| 122900.0 | 0 |
| 325000.0 | 0 |
| 120000.0 | 3 |
| 85860.0 | 0 |
| 97000.0 | 0 |
| 127000.0 | 0 |
| 89900.0 | 0 |
| 155000.0 | 0 |
| 253750.0 | 4 |
| 60000.0 | 0 |
| 87500.0 | 0 |

only showing top 20 rows

| Price | LotSize | Waterfront | Age | LandValue | NewConstruct | CentralAir | FuelType | HeatType | SewerType | LivingArea | PctCollege | Bedrooms | Fireplaces | Bathrooms | rooms | CLUSTER |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 132500.0 | 0.21 | 0.0 | 77.0 | 3500.0 | 0.0 | 0.0 | 2.0 | 2.0 | 3.0 | 1379.0 | 36.0 | 3.0 | 0.0 | 1.0 | 7.0 | 4 |
| 132500.0 | 0.37 | 0.0 | 19.0 | 13000.0 | 0.0 | 0.0 | 3.0 | 4.0 | 3.0 | 1988.0 | 63.0 | 2.0 | 0.0 | 1.0 | 5.0 | 4 |
| 132500.0 | 0.37 | 0.0 | 19.0 | 13000.0 | 0.0 | 0.0 | 3.0 | 4.0 | 3.0 | 1988.0 | 63.0 | 2.0 | 0.0 | 1.0 | 4.0 | 4 |
| 132500.0 | 0.09 | 0.0 | 42.0 | 50000.0 | 0.0 | 0.0 | 3.0 | 4.0 | 2.0 | 906.0 | 35.0 | 2.0 | 1.0 | 1.0 | 5.0 | 4 |
| 253750.0 | 2.0 | 0.0 | 0.0 | 49800.0 | 0.0 | 1.0 | 2.0 | 2.0 | 1.0 | 2816.0 | 71.0 | 4.0 | 1.0 | 2.5 | 12.0 | 4 |
| 290000.0 | 0.66 | 0.0 | 15.0 | 31200.0 | 0.0 | 1.0 | 2.0 | 2.0 | 2.0 | 2305.0 | 51.0 | 4.0 | 1.0 | 2.5 | 11.0 | 4 |
| 290000.0 | 0.46 | 0.0 | 22.0 | 48000.0 | 0.0 | 1.0 | 2.0 | 2.0 | 3.0 | 2030.0 | 64.0 | 4.0 | 1.0 | 2.5 | 10.0 | 4 |
| 290000.0 | 0.61 | 0.0 | 34.0 | 32300.0 | 0.0 | 0.0 | 2.0 | 3.0 | 3.0 | 2728.0 | 64.0 | 4.0 | 1.0 | 2.5 | 10.0 | 4 |
| 290000.0 | 0.12 | 0.0 | 3.0 | 108300.0 | 0.0 | 1.0 | 2.0 | 2.0 | 3.0 | 1620.0 | 57.0 | 3.0 | 1.0 | 2.5 | 7.0 | 4 |
| 290000.0 | 1.0 | 1.0 | 33.0 | 21700.0 | 0.0 | 0.0 | 4.0 | 2.0 | 2.0 | 944.0 | 27.0 | 1.0 | 1.0 | 1.0 | 4.0 | 4 |
| 290000.0 | 0.15 | 0.0 | 13.0 | 400.0 | 0.0 | 1.0 | 2.0 | 2.0 | 3.0 | 1758.0 | 47.0 | 2.0 | 1.0 | 2.5 | 6.0 | 4 |
| 290000.0 | 0.51 | 0.0 | 7.0 | 39100.0 | 0.0 | 0.0 | 2.0 | 2.0 | 3.0 | 2362.0 | 64.0 | 4.0 | 1.0 | 2.5 | 8.0 | 4 |
| 290000.0 | 0.71 | 1.0 | 73.0 | 61000.0 | 0.0 | 0.0 | 4.0 | 2.0 | 2.0 | 1838.0 | 71.0 | 4.0 | 0.0 | 2.0 | 8.0 | 4 |
| 205980.0 | 0.14 | 0.0 | 1.0 | 45200.0 | 1.0 | 1.0 | 2.0 | 2.0 | 3.0 | 1983.0 | 64.0 | 3.0 | 1.0 | 2.5 | 5.0 | 4 |
| 275000.0 | 0.54 | 0.0 | 19.0 | 30200.0 | 0.0 | 0.0 | 2.0 | 3.0 | 3.0 | 2175.0 | 64.0 | 4.0 | 1.0 | 2.5 | 10.0 | 4 |
| 275000.0 | 0.47 | 0.0 | 35.0 | 27800.0 | 0.0 | 0.0 | 2.0 | 3.0 | 3.0 | 2588.0 | 64.0 | 4.0 | 1.0 | 2.5 | 10.0 | 4 |
| 275000.0 | 0.37 | 0.0 | 14.0 | 31200.0 | 0.0 | 1.0 | 2.0 | 2.0 | 2.0 | 2011.0 | 40.0 | 4.0 | 1.0 | 2.5 | 8.0 | 4 |
| 275000.0 | 0.61 | 0.0 | 21.0 | 16100.0 | 0.0 | 1.0 | 2.0 | 2.0 | 2.0 | 2486.0 | 62.0 | 4.0 | 1.0 | 2.5 | 11.0 | 4 |
| 275000.0 | 0.46 | 0.0 | 7.0 | 18400.0 | 0.0 | 0.0 | 2.0 | 2.0 | 3.0 | 1865.0 | 57.0 | 3.0 | 0.0 | 2.5 | 8.0 | 4 |
| 275000.0 | 0.03 | 0.0 | 16.0 | 27000.0 | 0.0 | 1.0 | 2.0 | 2.0 | 3.0 | 1812.0 | 57.0 | 2.0 | 1.0 | 2.5 | 7.0 | 4 |

only showing top 20 rows




| summary | Price | LotSize | Waterfront | CLUSTER |
|---|---|---|---|---|
| count | 4655 | 4655 | 4655 | 4655 |
| mean | 162537.34135338347 | 0.4691321160042959 | 0.003007518796992 | 0.0 |
| stddev | 51449.17174680274 | 0.6264212879059081 | 0.05476420278337016 | 0.0 |
| min | 10300.0 | 0.0 | 0.0 | 0 |
| max | 600000.0 | 8.97 | 1.0 | 0 |

```
+-------+------------------+------------------+--------------------+--------+
|summary|             Price|           LotSize|          Waterfront| CLUSTER|
+-------+------------------+------------------+--------------------++-------+
|  count|              3356|              3356|                3356|    3356|
|   mean| 208313.6853396901|0.5529678188319437|0.006555423122765197|     0.0|
| stddev| 55025.18531388466|0.6481204374941402| 0.08071177527503304|     0.0|
|    min|            5000.0|              0.01|                 0.0|       0|
|    max|           600000.0|              7.24|                 1.0|       0|
-------+------------------+------------------+--------------------++---------
```

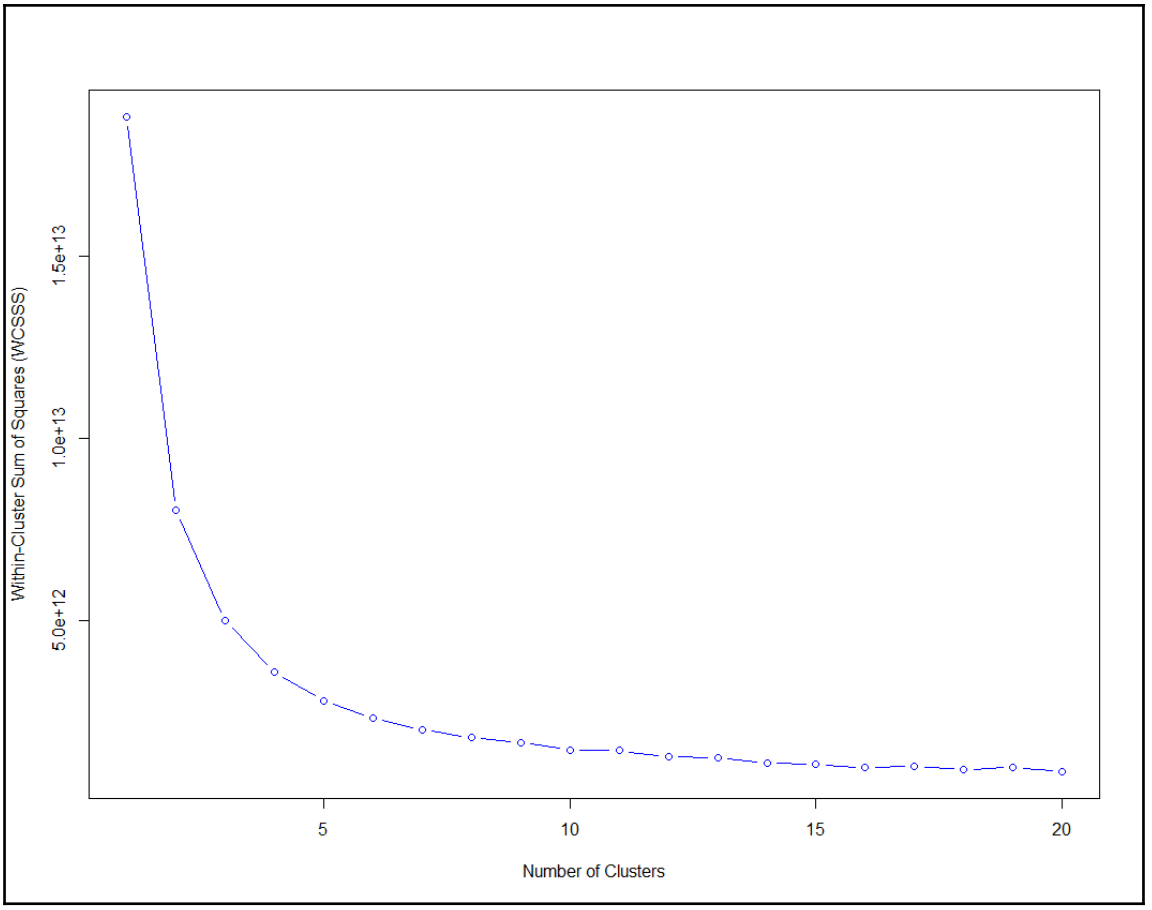
```
weight=0.062914
mu=[0.7808989073647417,0.027594804693120447,43.592594389644596,
sigma=
1.3102878656532657     -0.008781556531403927  ... (15 total)
-0.008781556531403927  0.026833331447068984   ...
-7.778315613332572     -0.22791269559972674   ...
-1994.3759086646505    953.5473866305449      ...
-0.047375946352474545  -0.004141141861757399  ...
0.037134027030446944   -8.540323116658249E-4  ...
0.18951000788031888    0.006470217824087922   ...
0.09271897737610613    0.0036900139379503023  ...
-0.18856936310448533   0.002380060277798572   ...
68.41825719674146      -4.1514710237897745    ...
0.03436783576161607    -0.1939395215895821    ...
-0.013693506665595243  -0.020101496691984813  ...
0.07766511419505753    -7.837804459298127E-4  ...
0.05849863330847962    -4.91629596665423E-4   ...
0.21074653933320067    -0.03620197853953523   ...
```

```
weight=0.062916
mu=[0.7808963058537262,0.027594300109707637,43.59271302953474,
sigma=
1.310268484362083       -0.008781324169799203  ... (15 total)
-0.008781324169799203   0.026832854711163028   ...
-7.778307903368862      -0.22791180189966537   ...
-1994.312293144824      953.5408564069132      ...
-0.047374773188269      -0.004140996766905095  ...
0.03713394906123063     -8.538912936897718E-4  ...
0.18950712664774483     0.006469719690237634   ...
0.09271774632030978     0.003690014536237195   ...
-0.18856667725951917    0.002380147914518987   ...
68.41855307532745       -4.151157768413957     ...
0.03435556638123301     -0.19393666789368957   ...
-0.013692860075417968   -0.020100969638081837  ...
0.07766539257767785     -7.836426721759263E-4  ...
0.058499301884249254    -4.913759714689525E-4  ...
0.21074627737377502     -0.03620069422593031   ...
```

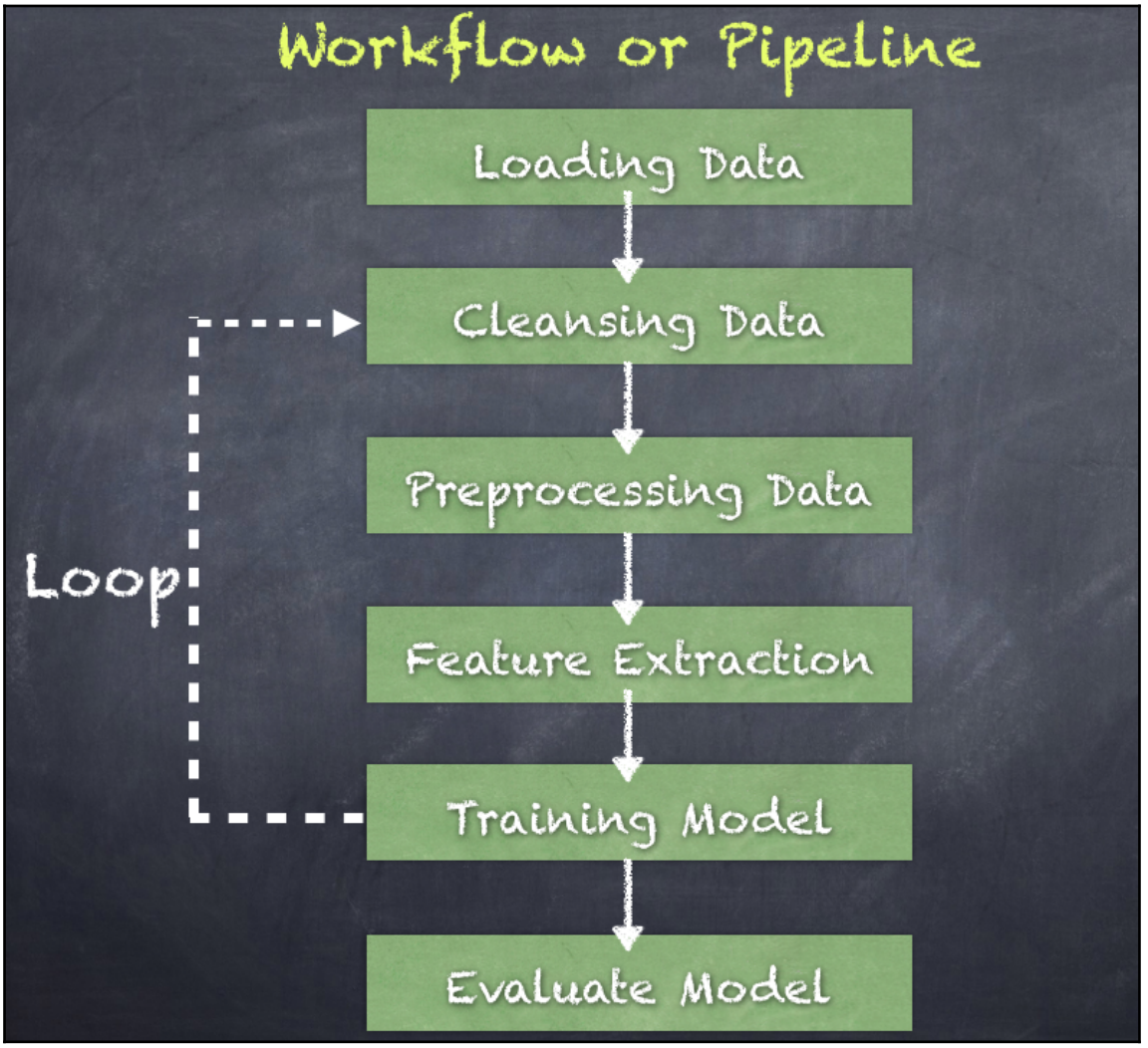
```
weight=0.062914
mu=[0.7808992393728132,0.027594857592962586,43.592578423292814,
sigma=
1.3102898209522704      -0.008781582527563107  ... (15 total)
-0.008781582527563107   0.026833381427386702   ...
-7.778313620309155      -0.2279126919235134    ...
-1994.3796587358433     953.5483992884145      ...
-0.04737604387775983    -0.004141154489440008  ...
0.03713404011490542 6   -8.540440057795036E-4  ...
0.18951016842223162     0.006470255770598361   ...
0.0927190997773798      0.003690015897543276   ...
-0.18856955926488908    0.0023800577384774702  ...
68.41823645079883       -4.151497911732236     ...
0.03436931403004916     -0.1939398262567097    ...
-0.013693555188677108   -0.020101546506840207  ...
0.07766507105646259     -7.837925362348146E-4  ...
0.05849857421583572     -4.916501535027908E-4  ...
0.21074661417458015     -0.036202094871478796  ...
```

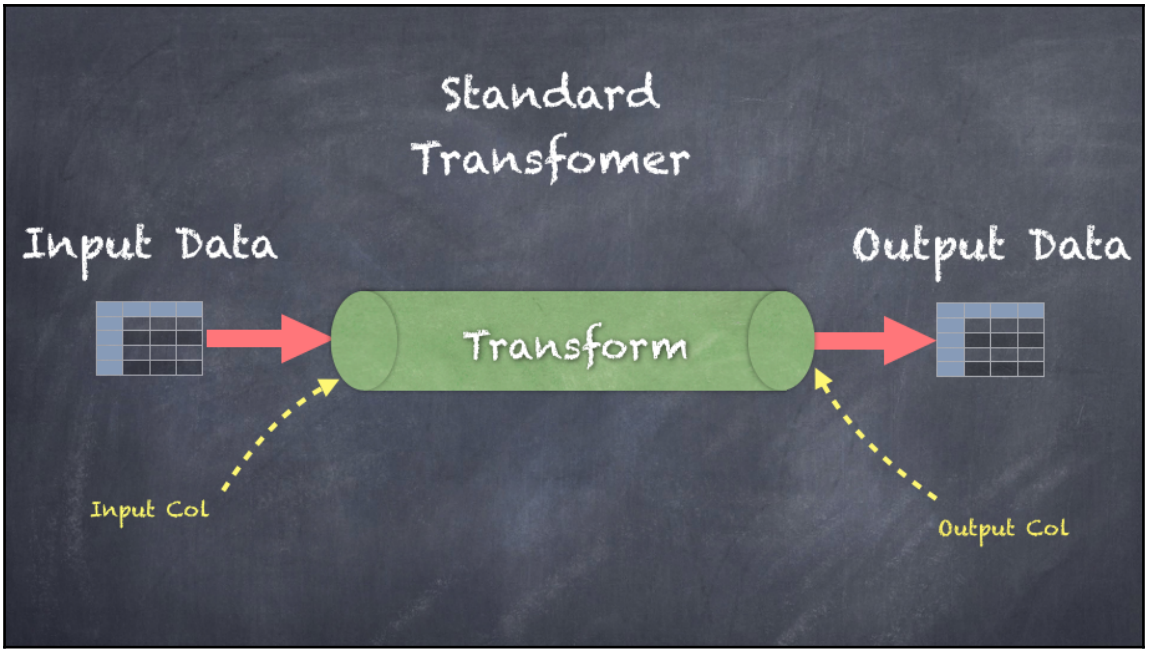
```
weight=0.062914
mu=[0.7808992393728132,0.027594857592962586,43.592578423292814,
sigma=
1.3102898209522704     -0.008781582527563107  ... (15 total)
-0.008781582527563107  0.026833381427386702   ...
-7.778313620309155     -0.2279126919235134    ...
-1994.3796587358433    953.5483992884145      ...
-0.04737604387775983   -0.004141154489440008  ...
0.037134040114905426   -8.540440057795036E-4  ...
0.18951016842223162    0.006470255770598361   ...
0.0927190997773798     0.003690015897543276   ...
-0.18856955926488908   0.0023800577384774702  ...
68.41823645079883      -4.151497911732236     ...
0.03436931403004916    -0.1939398262567097    ...
-0.013693555188677108  -0.020101546506840207  ...
0.07766507105646259    -7.837925362348146E-4  ...
0.05849857421583572    -4.916501535027908E-4  ...
0.21074661417458015    -0.036202094871478796  ...
```

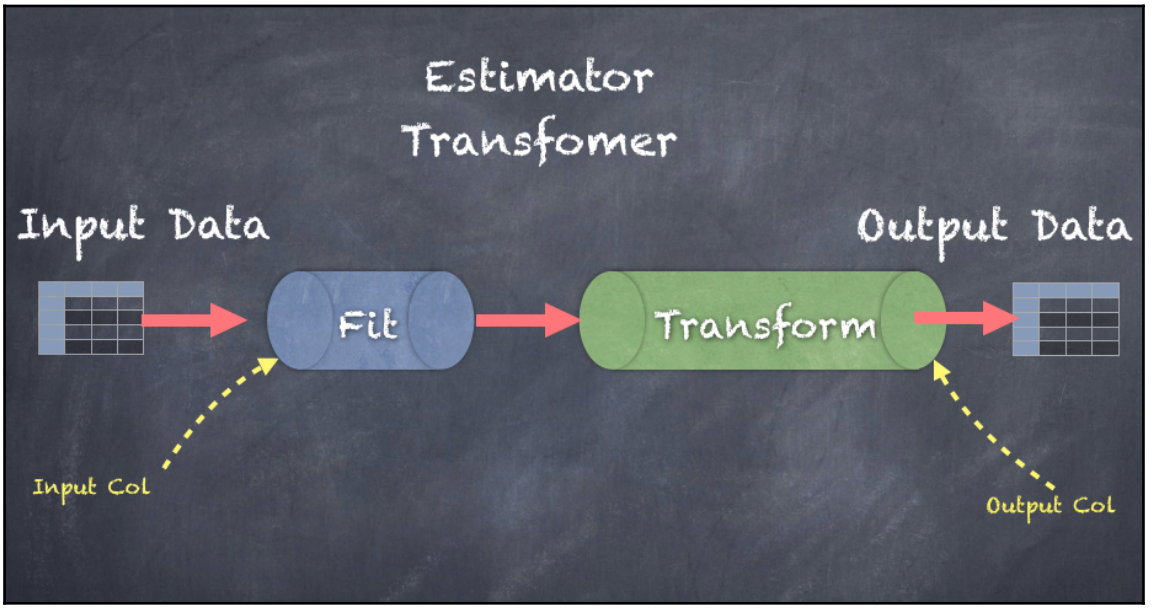
```
weight=0.748341
mu=[0.4058231162813968,0.0023199464802859684,22.644237040865264,
sigma=
0.17589935013824443    1.318739954100798E-4    ... (15 total)
1.318739954100798E-4   0.0023145643286145772   ...
0.1927311852469183     0.029438024228507845    ...
1108.2263125353527     116.97771555806634      ...
7.447161249472683E-4   -2.8238392691917215E-5  ...
-0.007840809295101704  -2.8837901025969534E-4  ...
0.023475650829693853   0.0019903868910481287   ...
-0.007865077010407383  -8.710340648340969E-5   ...
-0.06405665959152722   -5.685489003490323E-4   ...
54.81575059498132      0.06205695119789516     ...
0.08164968346291952    -0.02645130673658966    ...
0.07800513455809353    -0.001357416457442531   ...
0.023165924984143067   -1.9025716915325663E-4  ...
0.03441444210703747    -8.997788887508309E-5   ...
0.17565249306476233    -0.0017924278494132312  ...
```

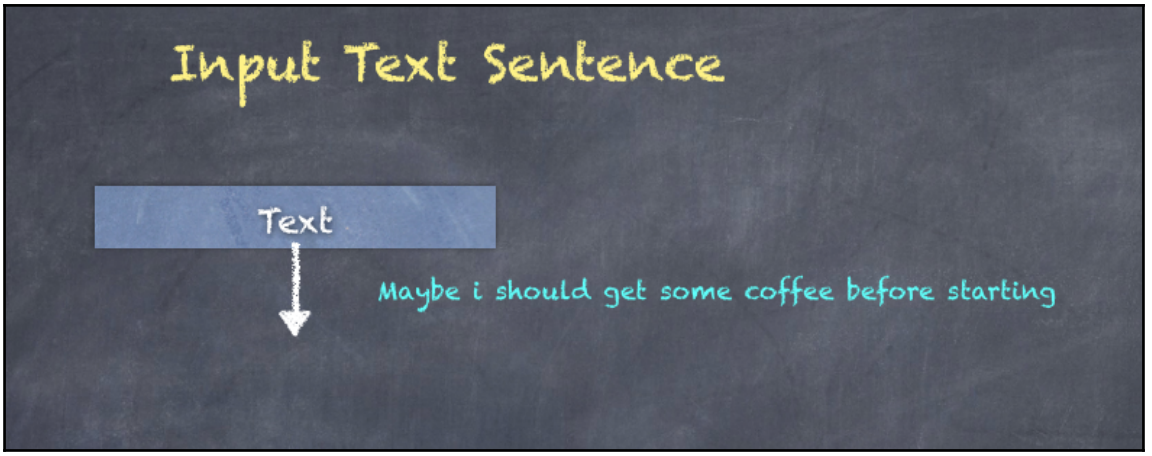

Within-Cluster Sum of Squares (WCSSS)
5.0e+12
1.0e+13
1.5e+13
5
10
15
20
Number of Clusters

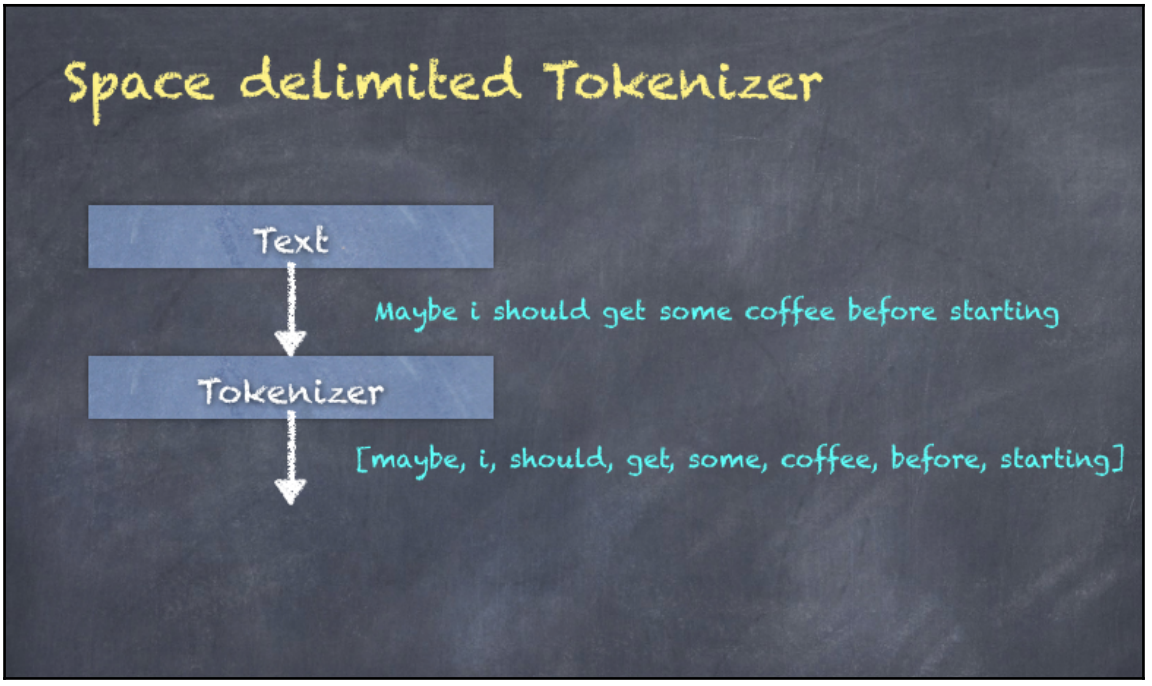
# Chapter 15: Text Analytics Using Spark ML

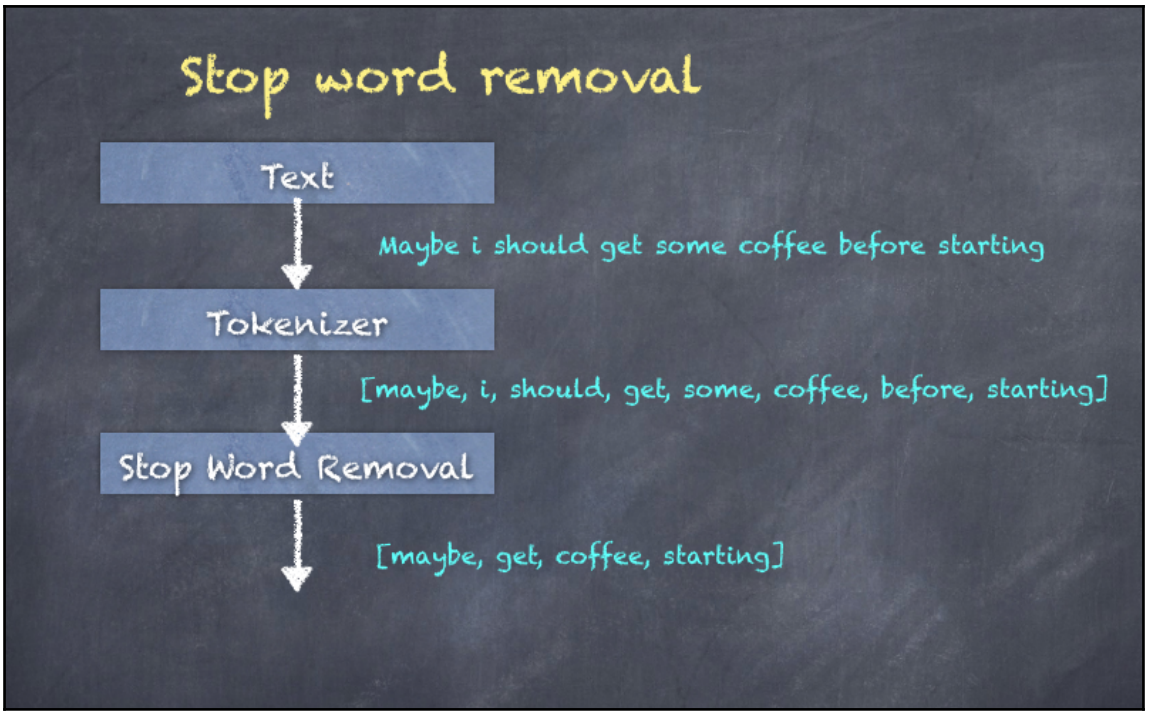


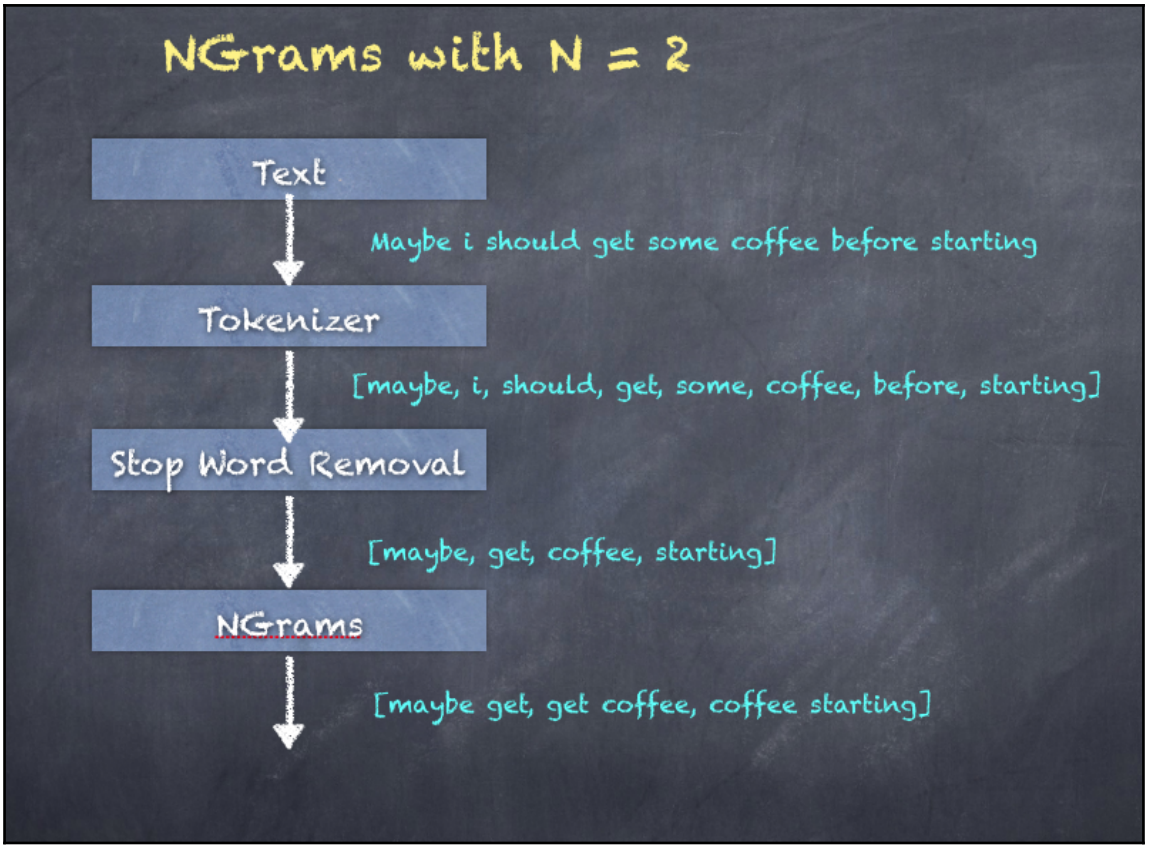

Standard
Transfomer
Input Data
Output Data
Transform
Input Col
Output Col

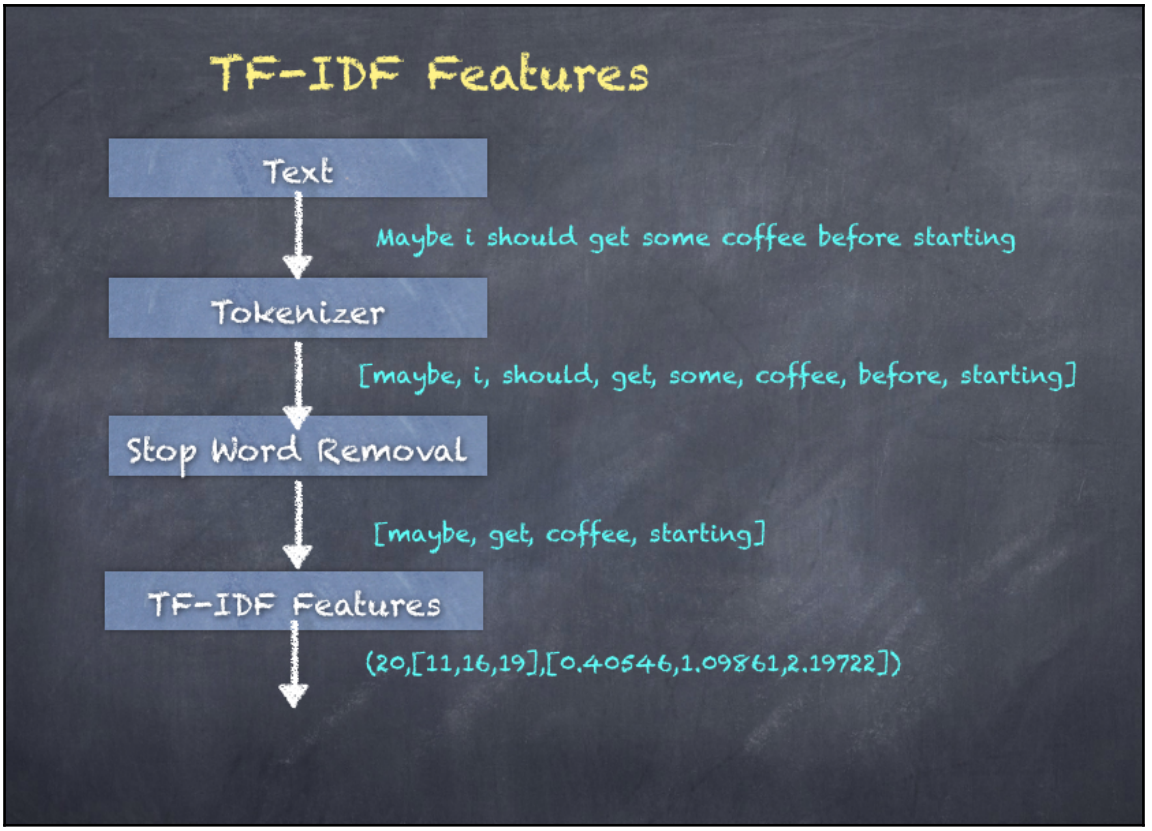

Estimator
Transfomer
Input Data
Output Data
Fit
Transform
Input Col
Output Col

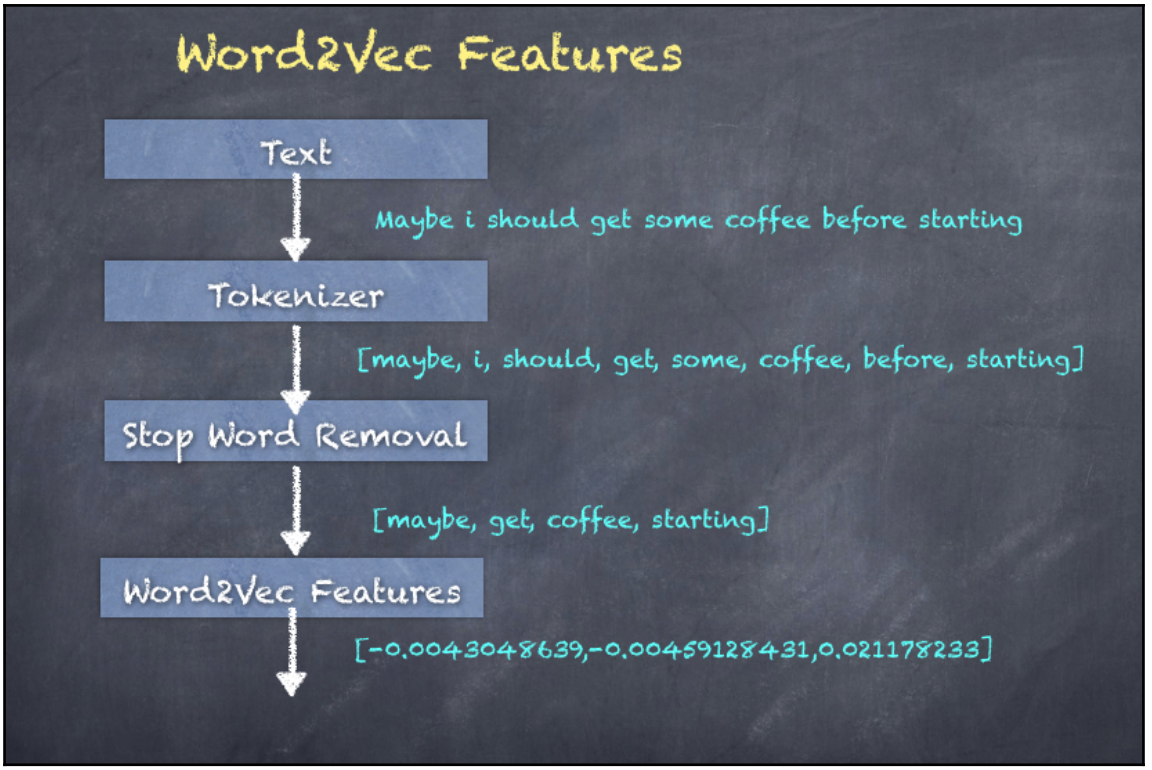

Input Text Sentence
Text
Maybe i should get some coffee before starting

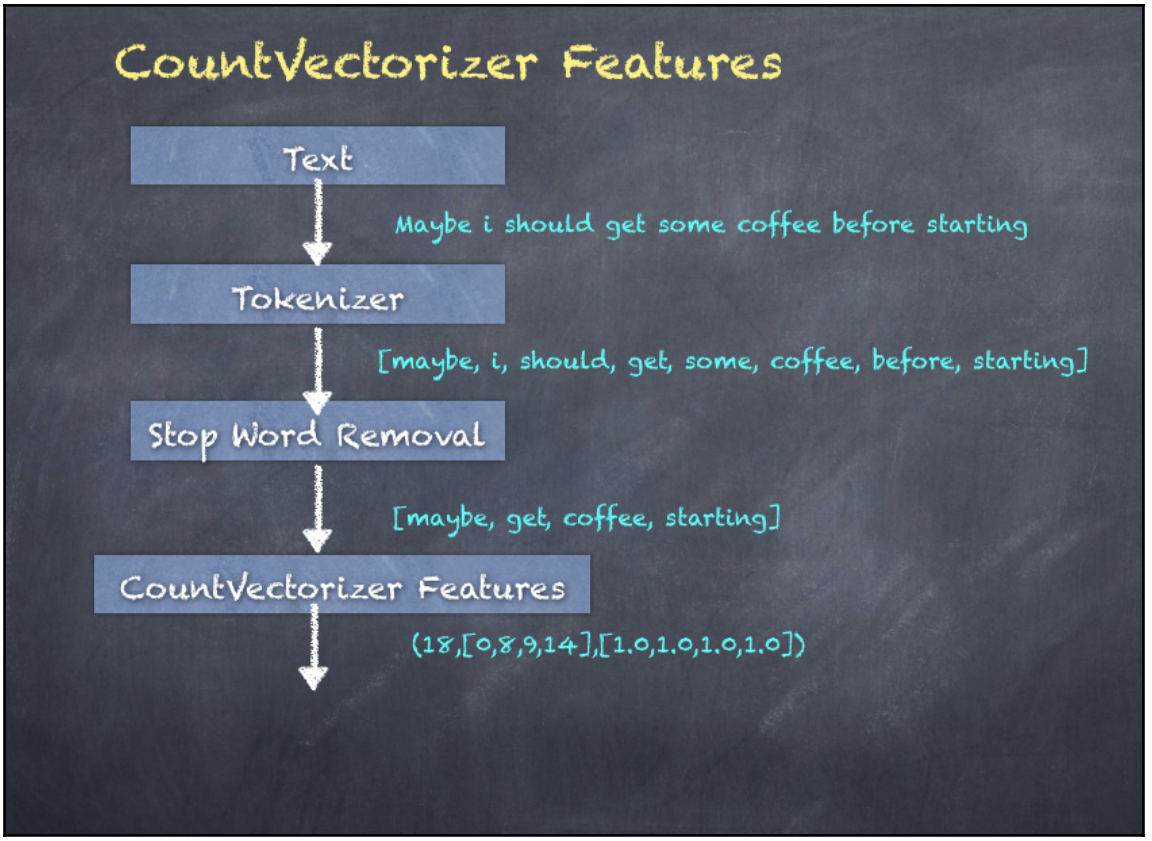

Space delimited Tokenizer
Text
Maybe i should get some coffee before starting
Tokenizer
[maybe, i, should, get, some, coffee, before, starting]

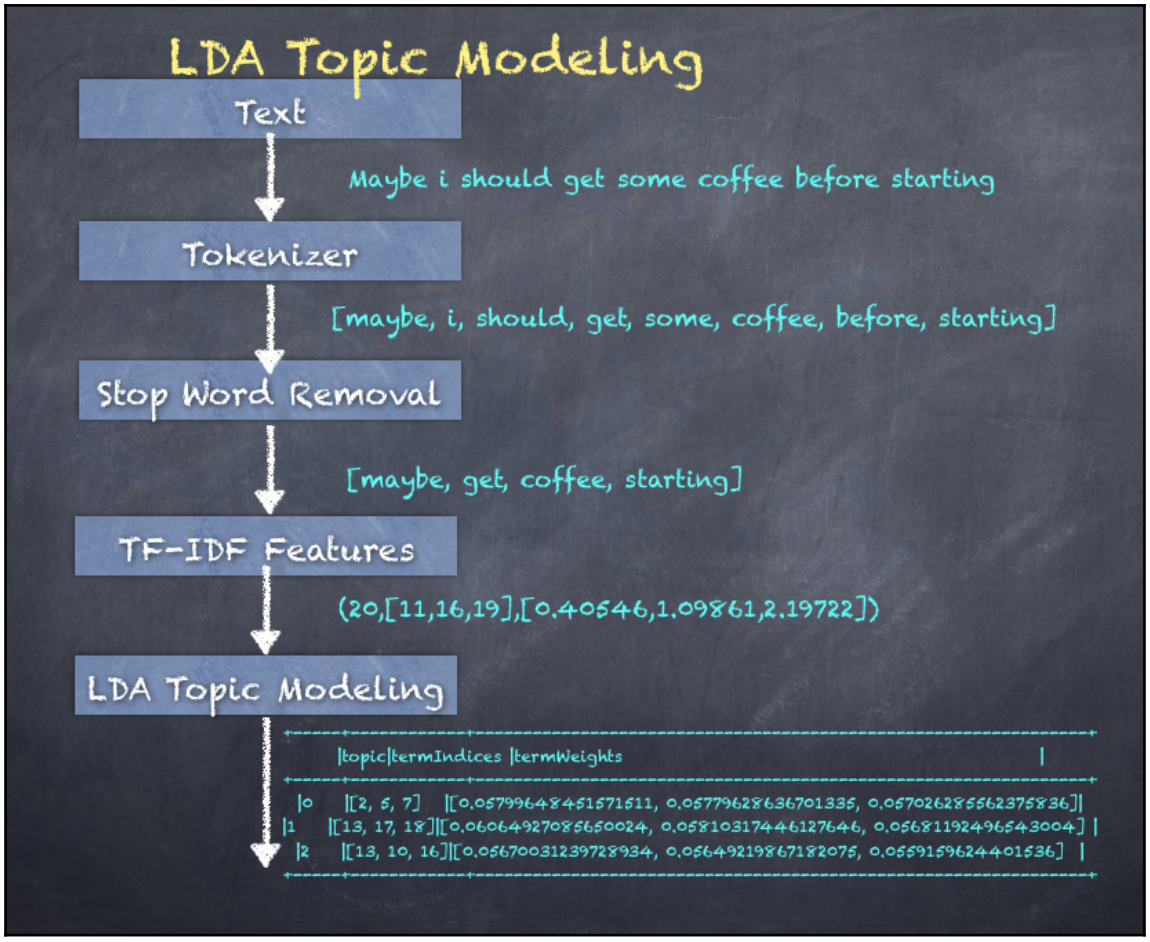

Stop word removal
Text
Maybe i should get some coffee before starting
Tokenizer
[maybe, i, should, get, some, coffee, before, starting]
Stop Word Removal
[maybe, get, coffee, starting]

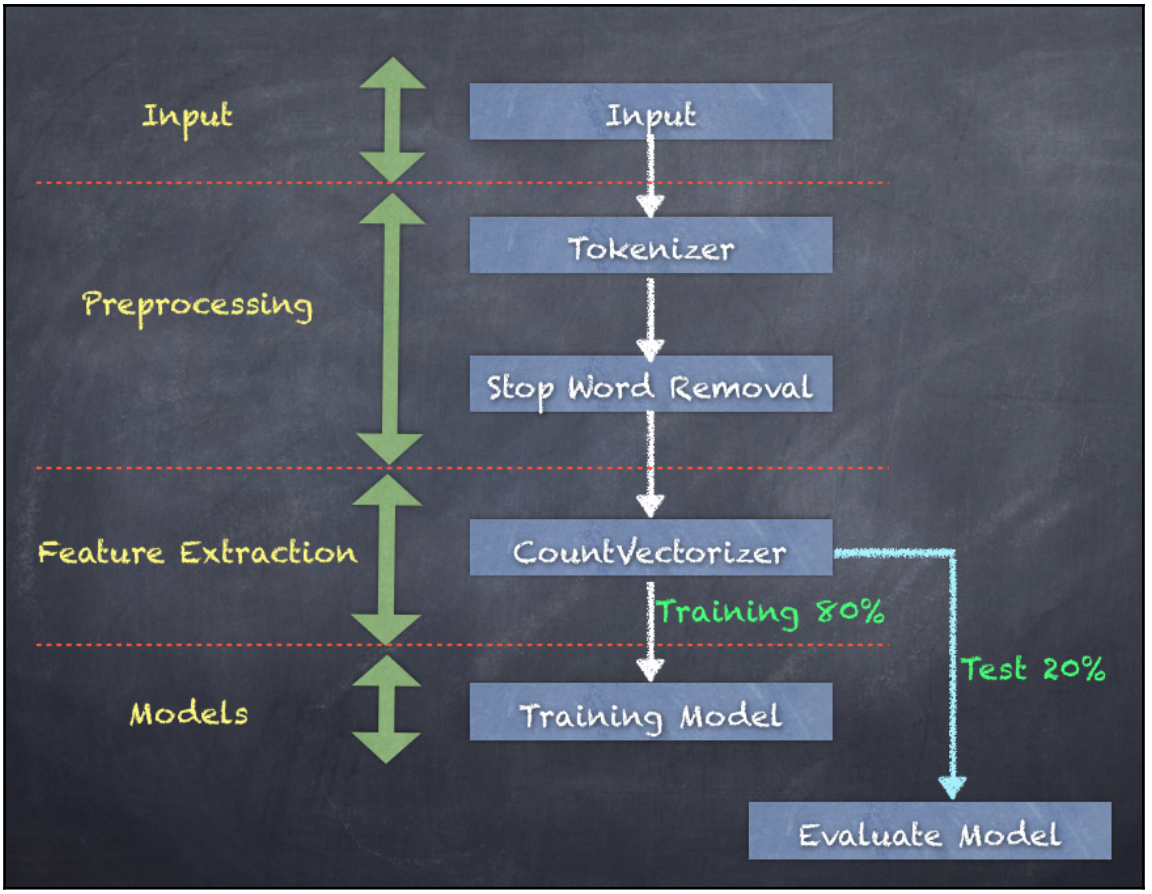

NGrams with N = 2
Text
Maybe i should get some coffee before starting
Tokenizer
[maybe, i, should, get, some, coffee, before, starting]
Stop Word Removal
[maybe, get, coffee, starting]
NGrams
[maybe get, get coffee, coffee starting]


TF-IDF Features
Text
Maybe i should get some coffee before starting
Tokenizer
[maybe, i, should, get, some, coffee, before, starting]
Stop Word Removal
[maybe, get, coffee, starting]
TF-IDF Features
(20,[11,16,19],[0.40546,1.09861,2.19722])

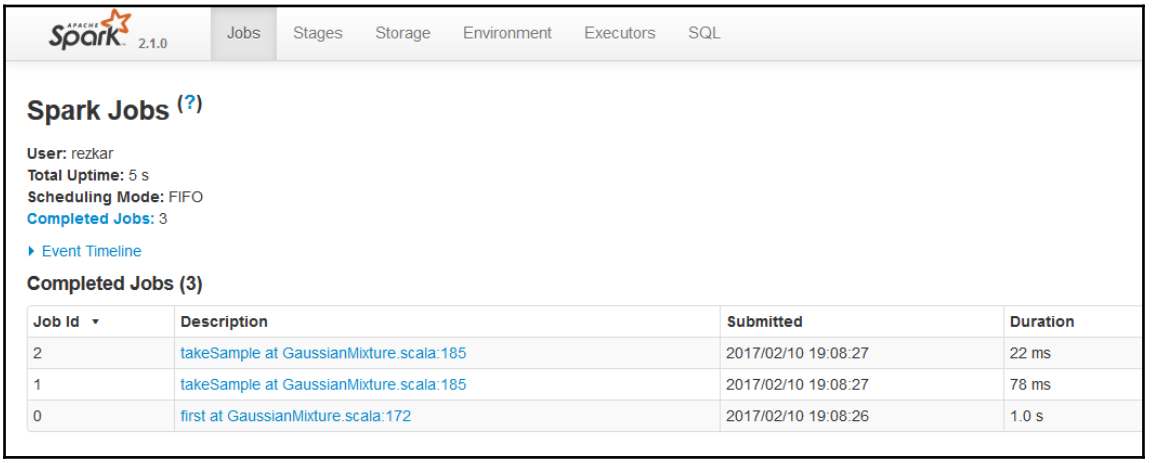

Word2Vec Features
Text
Maybe i should get some coffee before starting
Tokenizer
[maybe, i, should, get, some, coffee, before, starting]
Stop Word Removal
[maybe, get, coffee, starting]
Word2Vec Features
[-0.0043048639,-0.00459128431,0.021178233]

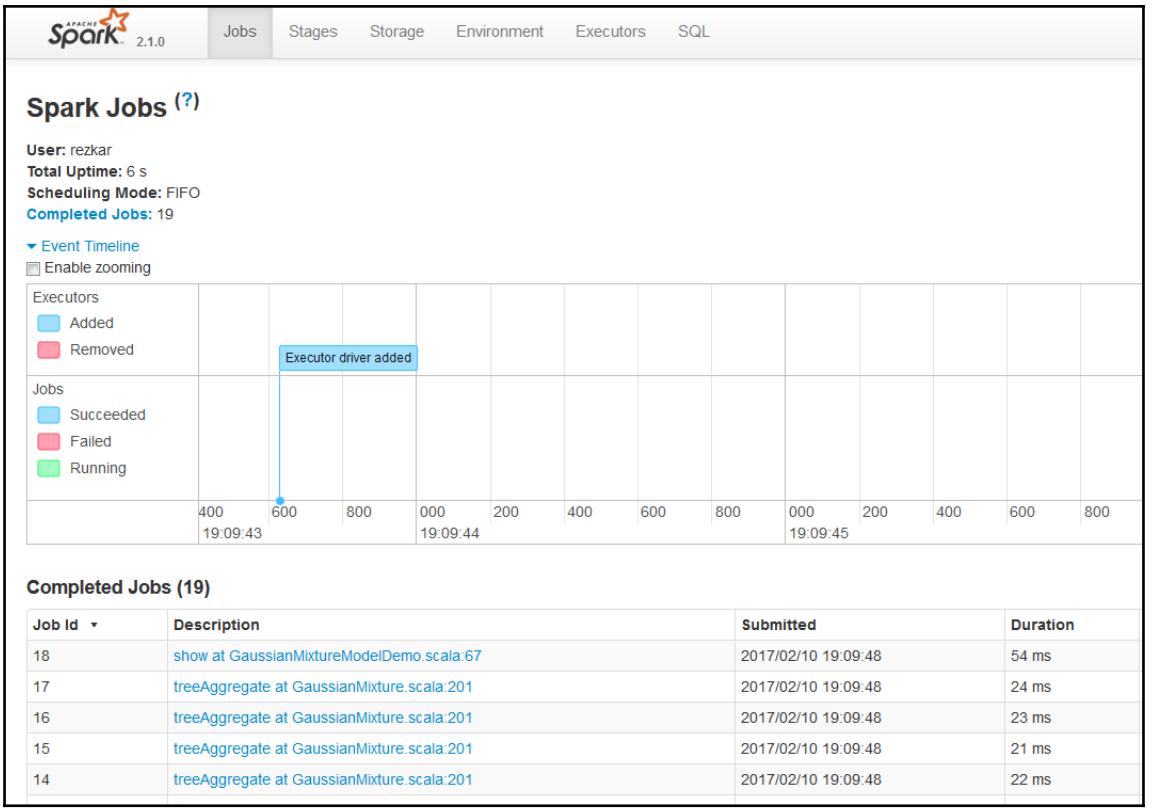

CountVectorizer Features
Text
Maybe i should get some coffee before starting
Tokenizer
[maybe, i, should, get, some, coffee, before, starting]
Stop Word Removal
[maybe, get, coffee, starting]
CountVectorizer Features
(18,[0,8,9,14],[1.0,1.0,1.0,1.0])

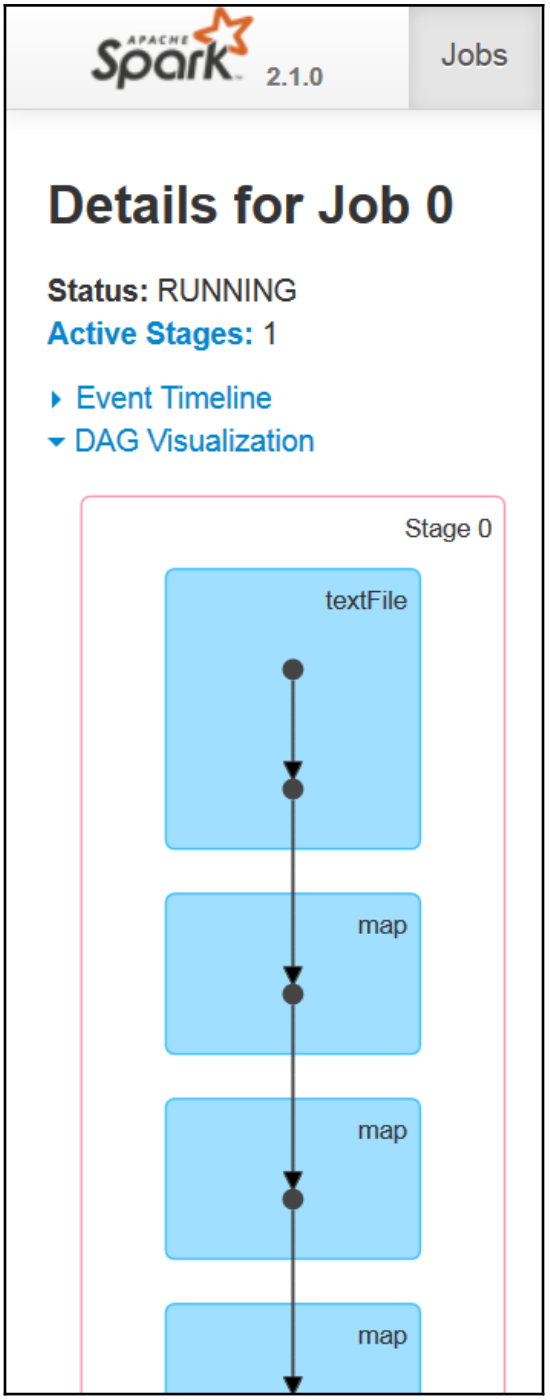

LDA Topic Modeling
Text
Maybe i should get some coffee before starting
Tokenizer
[maybe, i, should, get, some, coffee, before, starting]
Stop Word Removal
[maybe, get, coffee, starting]
TF-IDF Features
(20,[11,16,19],[0.40546,1.09861,2.19722])
LDA Topic Modeling
|topic|termIndices |termWeights |
|0 |[2, 5, 7] |[0.05799648451571511, 0.05779628636701335, 0.05702628556237583 6]|
|1 |[13, 17, 18]|[0.06064927085650024, 0.05810317446127646, 0.05681192496543004] |
|2 |[13, 10, 16]|[0.05670031239728934, 0.05649219867182075, 0.05591596244015 36] |

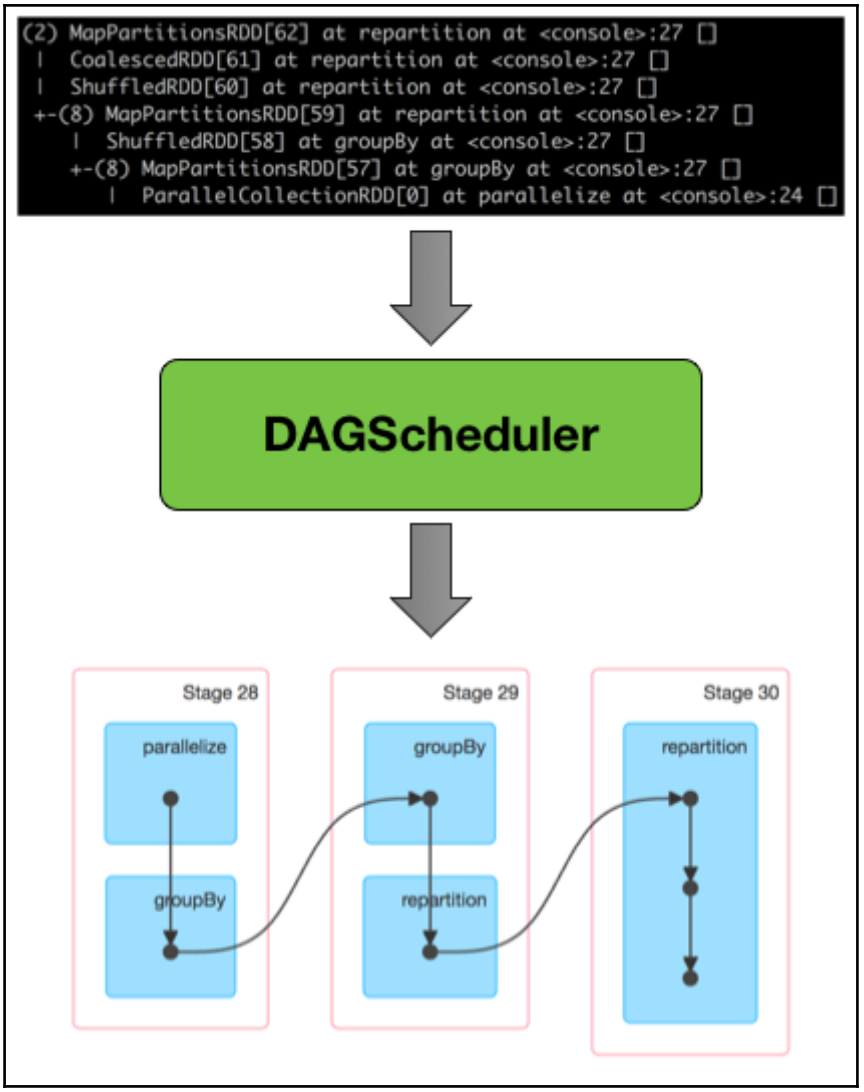

Input
Preprocessing
Feature Extraction
Models
Input
Tokenizer
Stop Word Removal
CountVectorizer
Training 80%
Test 20%
Training Model
Evaluate Model

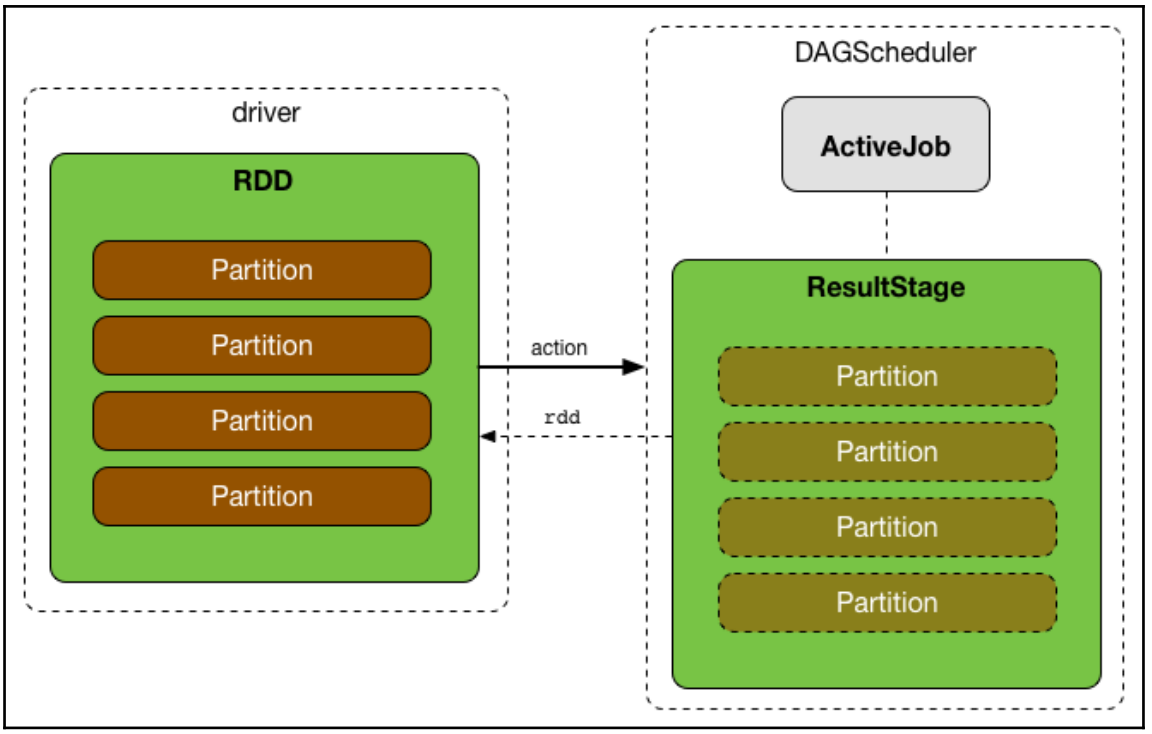
# Chapter 16: Spark Tuning

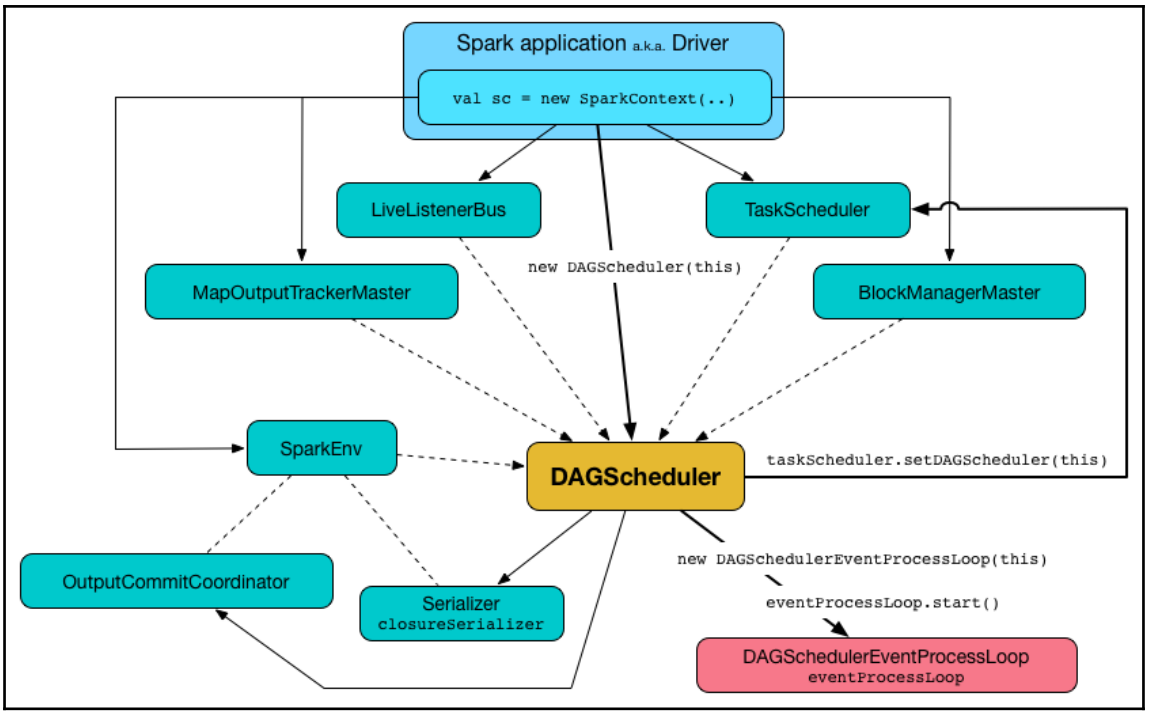


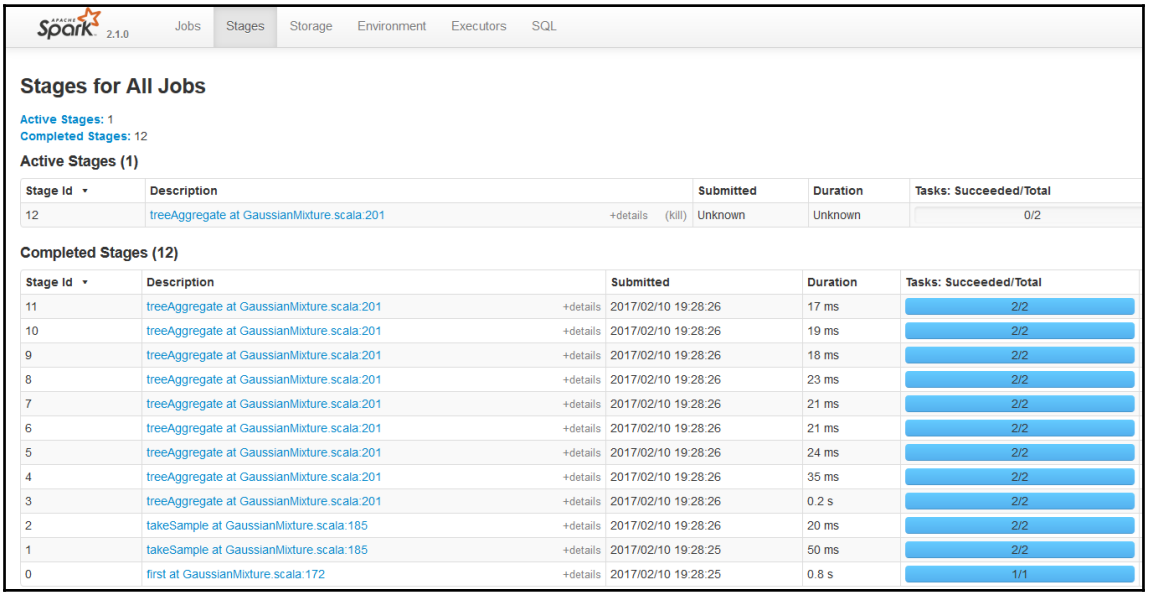

Spark 2.1.0
Jobs
Stages
Storage
Environment
Executors
SQL
Spark Jobs (?)
User: rezkar
Total Uptime: 6 s
Scheduling Mode: FIFO
Completed Jobs: 19
Event Timeline
Enable zooming
Executors
Added
Removed
Executor driver added
Jobs
Succeeded
Failed
Running
400
600
800
000
200
400
600
800
000
200
400
600
800
19:09:43
19:09:44
19:09:45
Completed Jobs (19)
Job Id
Description
Submitted
Duration
18
show at GaussianMixtureModelDemo.scala:67
2017/02/10 19:09:48
54 ms
17
treeAggregate at GaussianMixture.scala:201
2017/02/10 19:09:48
24 ms
16
treeAggregate at GaussianMixture.scala:201
2017/02/10 19:09:48
23 ms
15
treeAggregate at GaussianMixture.scala:201
2017/02/10 19:09:48
21 ms
14
treeAggregate at GaussianMixture.scala:201
2017/02/10 19:09:48
22 ms


Spark 2.1.0
Jobs
Details for Job 0
Status: RUNNING
Active Stages: 1
Event Timeline
DAG Visualization
Stage 0
textFile
map
map
map


(2) MapPartitionsRDD[62] at repartition at <console>:27 []
| CoalescedRDD[61] at repartition at <console>:27 []
| ShuffledRDD[60] at repartition at <console>:27 []
+-(8) MapPartitionsRDD[59] at repartition at <console>:27 []
| ShuffledRDD[58] at groupBy at <console>:27 []
+-(8) MapPartitionsRDD[57] at groupBy at <console>:27 []
| ParallelCollectionRDD[0] at parallelize at <console>:24 []
DAGScheduler
Stage 28
parallelize
groupBy
Stage 29
groupBy
repartition
Stage 30
repartition


DAGScheduler
driver
ActiveJob
RDD
Partition
Partition
Partition
Partition
action
rdd
ResultStage
Partition
Partition
Partition
Partition


Spark application a.k.a. Driver
val sc = new SparkContext(..)
LiveListenerBus
TaskScheduler
new DAGScheduler(this)
MapOutputTrackerMaster
BlockManagerMaster
SparkEnv
DAGScheduler
taskScheduler.setDAGScheduler(this)
new DAGSchedulerEventProcessLoop(this)
OutputCommitCoordinator
Serializer
closureSerializer
eventProcessLoop.start()
DAGSchedulerEventProcessLoop
eventProcessLoop


Spark 2.1.0
Jobs
Stages
Storage
Environment
Executors
SQL
Stages for All Jobs
Active Stages: 1
Completed Stages: 12
Active Stages (1)
Stage Id
Description
Submitted
Duration
Tasks: Succeeded/Total
12
treeAggregate at GaussianMixture.scala:201
+details
(kill)
Unknown
Unknown
0/2
Completed Stages (12)
Stage Id
Description
Submitted
Duration
Tasks: Succeeded/Total
11 treeAggregate at GaussianMixture.scala:201 +details 2017/02/10 19:28:26 17 ms 2/2
10 treeAggregate at GaussianMixture.scala:201 +details 2017/02/10 19:28:26 19 ms 2/2
9 treeAggregate at GaussianMixture.scala:201 +details 2017/02/10 19:28:26 18 ms 2/2
8 treeAggregate at GaussianMixture.scala:201 +details 2017/02/10 19:28:26 23 ms 2/2
7 treeAggregate at GaussianMixture.scala:201 +details 2017/02/10 19:28:26 21 ms 2/2
6 treeAggregate at GaussianMixture.scala:201 +details 2017/02/10 19:28:26 21 ms 2/2
5 treeAggregate at GaussianMixture.scala:201 +details 2017/02/10 19:28:26 24 ms 2/2
4 treeAggregate at GaussianMixture.scala:201 +details 2017/02/10 19:28:26 35 ms 2/2
3 treeAggregate at GaussianMixture.scala:201 +details 2017/02/10 19:28:26 0.2 s 2/2
2 takeSample at GaussianMixture.scala:185 +details 2017/02/10 19:28:26 20 ms 2/2
1 takeSample at GaussianMixture.scala:185 +details 2017/02/10 19:28:25 50 ms 2/2
0 first at GaussianMixture.scala:172 +details 2017/02/10 19:28:25 0.8 s 1/1

**Summary Metrics for 2 Completed Tasks**

| Metric | Min | 25th percentile | Median |
|---|---|---|---|
| Duration | 0.2 s | 0.2 s | 0.2 s |
| GC Time | 0 ms | 0 ms | 0 ms |
| Input Size / Records | 27.6 KB / 1 | 27.6 KB / 1 | 28.6 KB / 1 |

**Aggregated Metrics by Executor**

| Executor ID | Address | Task Time | Total Tasks | Failed Tasks | Killed Tasks |
|---|---|---|---|---|---|
| driver | 10.2.17.13:53512 | 0.5 s | 2 | 0 | 0 |

**Tasks (2)**

| Index | ID | Attempt | Status | Locality Level | Executor ID / Host | Launch Time |
|---|---|---|---|---|---|---|
| 0 | 4 | 0 | SUCCESS | PROCESS_LOCAL | driver / localhost | 2017/02/04 12:57:01 |
| 1 | 5 | 0 | SUCCESS | PROCESS_LOCAL | driver / localhost | 2017/02/04 12:57:01 |

Spark 2.1.0 Jobs Stages Storage Environment Executors SQL SparkDFeb

**Storage**

**RDDs**

| RDD Name | Storage Level | Cached Partitions | Fraction Cached | Size in Memory |
|---|---|---|---|---|
| MapPartitionsRDD | Memory Deserialized 1x Replicated | 2 | 100% | 281.8 KB |
| *SerializeFromObject [assertnotnull(input[0, com.chapter13.Clustering.GaussianMixtureModelDemo$Land, true], top level Product input object).Price AS Price#17, assertnotnull(input[0, com.chapter13.Clustering.GaussianMixtureModelDemo$Land, true], top level Product input object).LotSize AS LotSize#18, assertnotnull(input[0, com.chapter13.Clustering.GaussianMixtureModelDemo$Land, true], top level Product input object).Waterfront AS Waterfront#19, assertnotnull(input[0, com.chapter13.Clustering.GaussianMixtureModelDemo$Land, true], top level Product input object).Age AS Age#20, assertnotnull(input[0, com.chapter13.Clustering.GaussianMixtureModelDemo$Land, true], top level Product input object).LandValue AS LandValue#21, assertnotnull(input[0, com.chapter13.Clustering.GaussianMixtureModelDemo$Land, true], top level Product input object).NewConstruct AS NewConstruct#22, assertnotnull(input[0, com.chapter13.Clustering.GaussianMixtureModelDemo$Land, true], top level Product input object).CentralAir AS CentralAir#23... | Memory Deserialized 1x Replicated | 2 | 100% | 220.5 KB |
| MapPartitionsRDD | Memory Deserialized 1x Replicated | 2 | 100% | 244.3 KB |

Spark 2.1.0 Jobs Stages Storage Environment Executors SQL

## RDD Storage Info for MapPartitionsRDD

**Storage Level:** Memory Deserialized 1x Replicated
**Cached Partitions:** 1
**Total Partitions:** 2
**Memory Size:** 5.3 MB
**Disk Size:** 0.0 B

### Data Distribution on 1 Executors

| Host | Memory Usage |
|---|---|
| 192.168.192.1:52341 | 5.3 MB (4.0 GB Remaining) |

### 1 Partitions

| Block Name ▲ | Storage Level | Size in Memory |
|---|---|---|
| rdd_4_0 | Memory Deserialized 1x Replicated | 5.3 MB |

Spark 2.1.0 Jobs Stages Storage Environment Executors SQL

## Environment

### Runtime Information

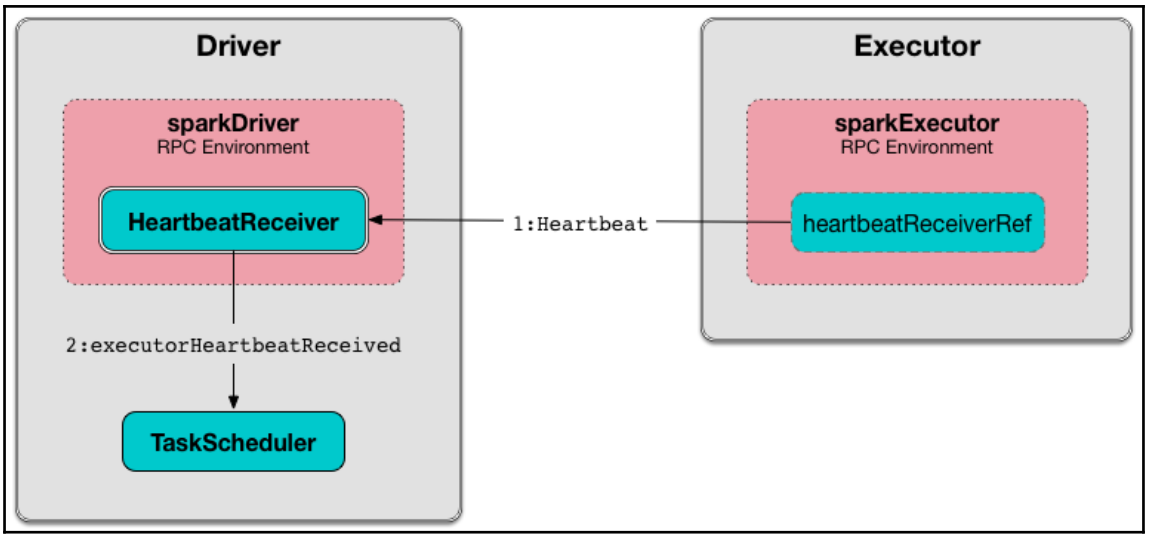
| Name | Value |
|---|---|
| Java Home | C:\Program Files\Java\jdk1.8.0_102\jre |
| Java Version | 1.8.0_102 (Oracle Corporation) |
| Scala Version | version 2.11.8 |

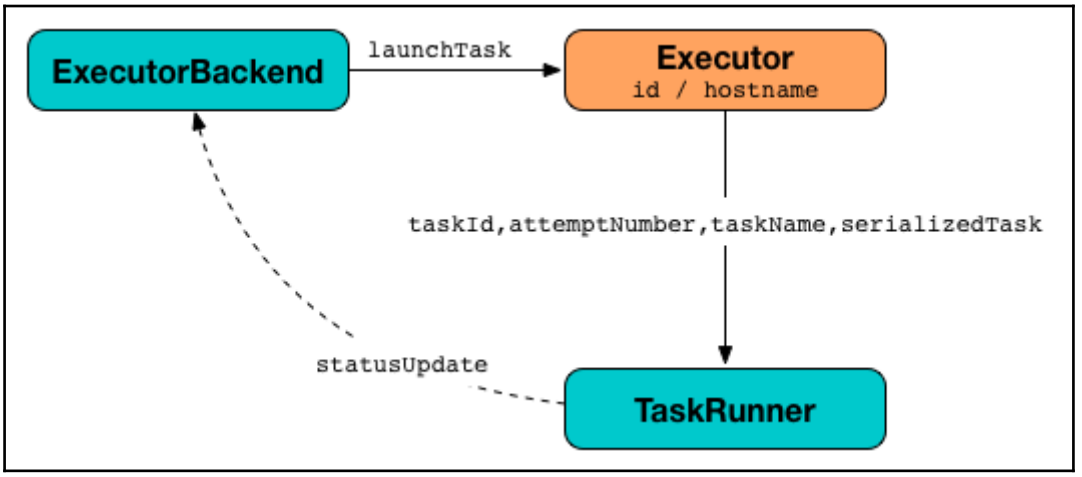
### Spark Properties

| Name | Value |
|---|---|
| spark.app.id | local-1485627820379 |
| spark.app.name | SparkDFebay |
| spark.driver.host | 10.2.16.255 |
| spark.driver.port | 64649 |
| spark.executor.id | driver |
| spark.master | local[*] |
| spark.scheduler.mode | FIFO |

### System Properties

| Name | Value |
|---|---|
| awt.toolkit | sun.awt.windows.WToolkit |
| file.encoding | UTF-8 |
| file.encoding.pkg | sun.io |


Driver
sparkDriver
RPC Environment
HeartbeatReceiver
2:executorHeartbeatReceived
TaskScheduler
1:Heartbeat
Executor
sparkExecutor
RPC Environment
heartbeatReceiverRef


ExecutorBackend
launchTask
Executor
id / hostname
taskId,attemptNumber,taskName,serializedTask
statusUpdate
TaskRunner

Spark 2.1.0 — Jobs | Stages | Storage | Environment | Executors

## Executors

### Summary

| | RDD Blocks | Storage Memory | Disk Used | Cores | Active Tasks | Failed Tasks | Complete Tasks | Total Tasks | Task Time (GC Time) |
|---|---|---|---|---|---|---|---|---|---|
| **Active(1)** | 4 | 263.4 KB / 4.3 GB | 0.0 B | 8 | 1 | 0 | 0 | 1 | 0 ms (0 ms) |
| **Dead(0)** | 0 | 0.0 B / 0.0 B | 0.0 B | 0 | 0 | 0 | 0 | 0 | 0 ms (0 ms) |
| **Total(1)** | 4 | 263.4 KB / 4.3 GB | 0.0 B | 8 | 1 | 0 | 0 | 1 | 0 ms (0 ms) |

### Executors

Show 20 entries

| Executor ID | Address | Status | RDD Blocks | Storage Memory | Disk Used | Cores | Active Tasks | Failed Tasks | Complete Tasks | Total Tasks | Task Time (GC Time) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| driver | 10.2.16.255:64829 | Active | 4 | 263.4 KB / 4.3 GB | 0.0 B | 8 | 1 | 0 | 0 | 1 | 0 ms (0 ms) |

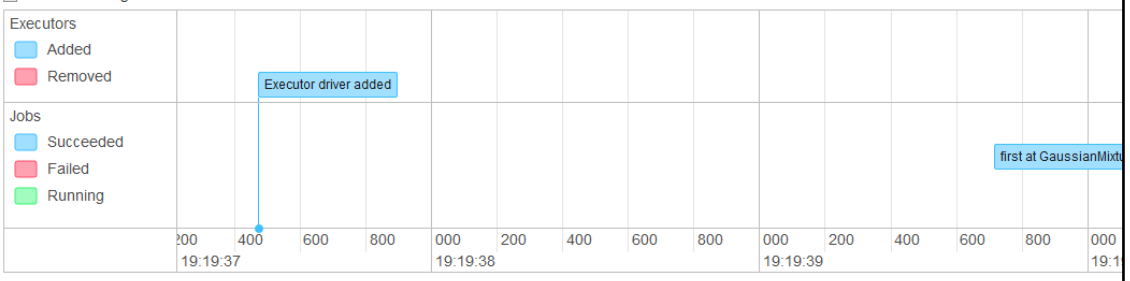
Showing 1 to 1 of 1 entries

Spark 2.1.0 — Jobs | Stages | Storage | Environment | Executors | SQL

## Spark Jobs (?)

**User:** rezkar
**Total Uptime:** 7 s
**Scheduling Mode:** FIFO
**Completed Jobs:** 25

Event Timeline
Enable zooming




### Completed Jobs (25)

| Job Id | Description | Submitted | Duration |
|---|---|---|---|
| 24 | show at GaussianMixtureModelDemo.scala:75 | 2017/02/10 19:19:42 | 14 ms |
| 23 | run at ThreadPoolExecutor.java:1142 | 2017/02/10 19:19:42 | 12 ms |
| 22 | show at GaussianMixtureModelDemo.scala:74 | 2017/02/10 19:19:42 | 15 ms |
| 21 | run at ThreadPoolExecutor.java:1142 | 2017/02/10 19:19:42 | 13 ms |
| 20 | show at GaussianMixtureModelDemo.scala:71 | 2017/02/10 19:19:42 | 13 ms |
| 19 | run at ThreadPoolExecutor.java:1142 | 2017/02/10 19:19:42 | 23 ms |

Spark 2.1.0 | Jobs | Stages | Storage | Environment | Executors | SQL

## Spark Jobs (?)

**User:** rezkar
**Total Uptime:** 6 s
**Scheduling Mode:** FIFO
**Active Jobs:** 1
**Completed Jobs:** 12

Event Timeline

### Active Jobs (1)

| Job Id | Description | Submitted | Duration | Stages: Succeeded/Total |
|---|---|---|---|---|
| 12 | treeAggregate at GaussianMixture.scala:201 (kill) | 2017/02/10 19:32:16 | 30 ms | 0/1 |

### Completed Jobs (12)

| Job Id | Description | Submitted | Duration | Stages: Succeeded/Total |
|---|---|---|---|---|
| 11 | treeAggregate at GaussianMixture.scala:201 | 2017/02/10 19:32:16 | 57 ms | 1/1 |
| 10 | treeAggregate at GaussianMixture.scala:201 | 2017/02/10 19:32:16 | 35 ms | 1/1 |
| 9 | treeAggregate at GaussianMixture.scala:201 | 2017/02/10 19:32:16 | 68 ms | 1/1 |
| 8 | treeAggregate at GaussianMixture.scala:201 | 2017/02/10 19:32:16 | 28 ms | 1/1 |
| 7 | treeAggregate at GaussianMixture.scala:201 | 2017/02/10 19:32:16 | 33 ms | 1/1 |
| 6 | treeAggregate at GaussianMixture.scala:201 | 2017/02/10 19:32:16 | 34 ms | 1/1 |
| 5 | treeAggregate at GaussianMixture.scala:201 | 2017/02/10 19:32:16 | 46 ms | 1/1 |
| 4 | treeAggregate at GaussianMixture.scala:201 | 2017/02/10 19:32:16 | 44 ms | 1/1 |
| 3 | treeAggregate at GaussianMixture.scala:201 | 2017/02/10 19:32:15 | 0.2 s | 1/1 |
| 2 | takeSample at GaussianMixture.scala:185 | 2017/02/10 19:32:15 | 29 ms | 1/1 |
| 1 | takeSample at GaussianMixture.scala:185 | 2017/02/10 19:32:15 | 69 ms | 1/1 |

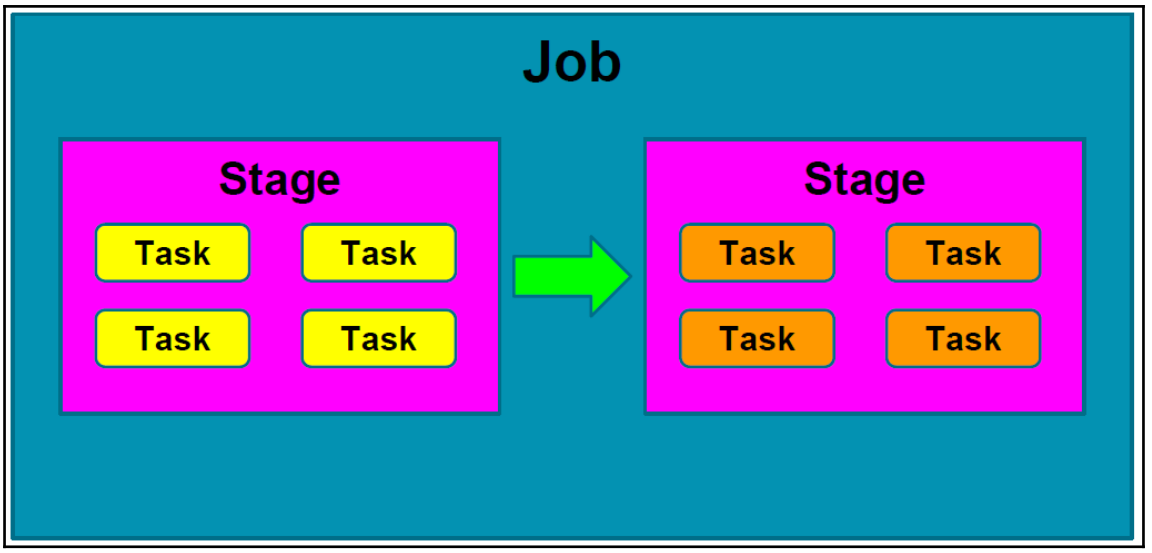
```
# Set everything to be logged to the console
log4j.rootCategory=INFO, console
log4j.appender.console=org.apache.log4j.ConsoleAppender
log4j.appender.console.target=System.err
log4j.appender.console.layout=org.apache.log4j.PatternLayout
log4j.appender.console.layout.ConversionPattern=%d{yy/MM/dd HH:mm:ss} %p %c{1}: %m%n

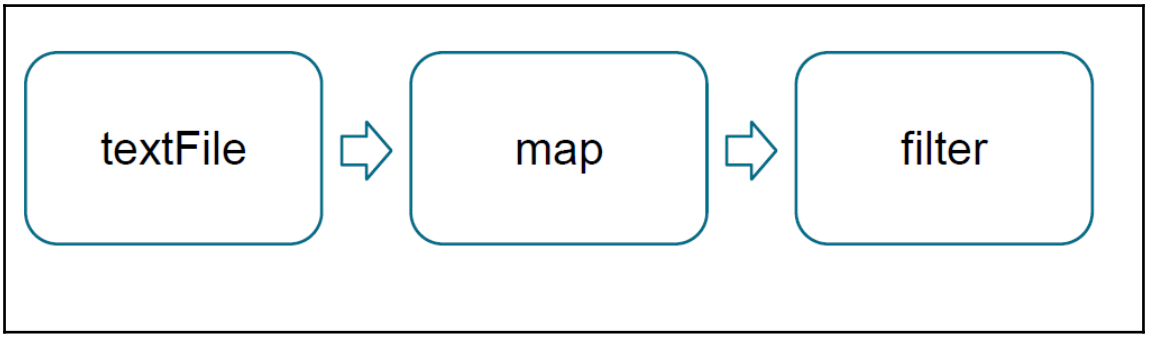
# Set the default spark-shell log level to WARN. When running the spark-shell, the
# log level for this class is used to overwrite the root logger's log level, so that
# the user can have different defaults for the shell and regular Spark apps.
log4j.logger.org.apache.spark.repl.Main=WARN

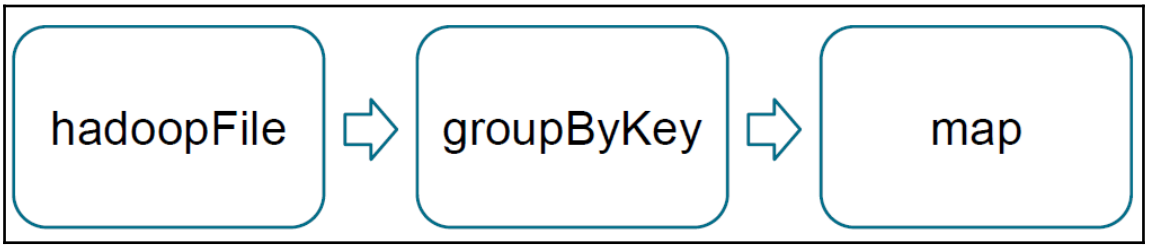
# Settings to quiet third party logs that are too verbose
log4j.logger.org.spark_project.jetty=WARN
log4j.logger.org.spark_project.jetty.util.component.AbstractLifeCycle=ERROR
log4j.logger.org.apache.spark.repl.SparkIMain$exprTyper=INFO
log4j.logger.org.apache.spark.repl.SparkILoop$SparkILoopInterpreter=INFO
log4j.logger.org.apache.parquet=ERROR
log4j.logger.parquet=ERROR

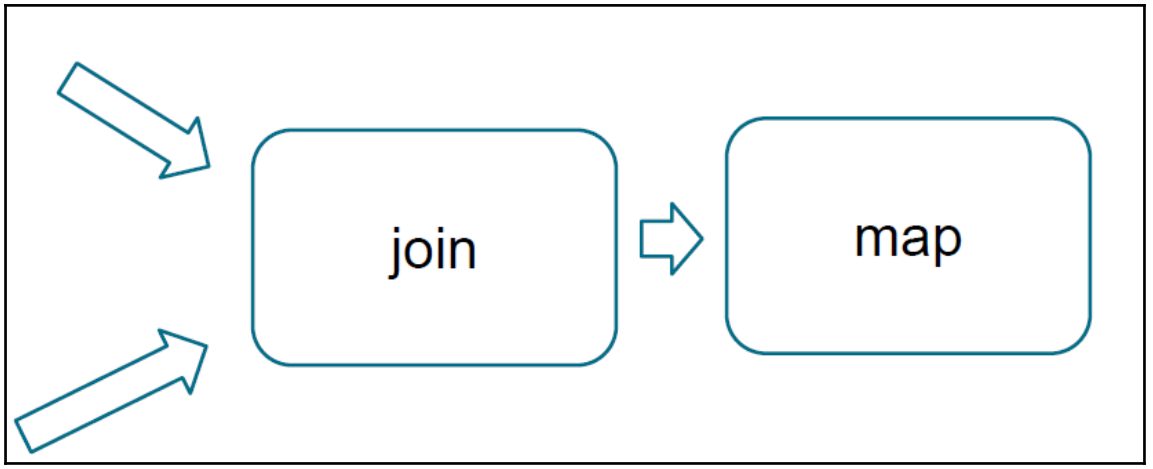
# SPARK-9183: Settings to avoid annoying messages when looking up nonexistent UDFs in SparkSQL with Hive support
log4j.logger.org.apache.hadoop.hive.metastore.RetryingHMSHandler=FATAL
log4j.logger.org.apache.hadoop.hive.ql.exec.FunctionRegistry=ERROR
```

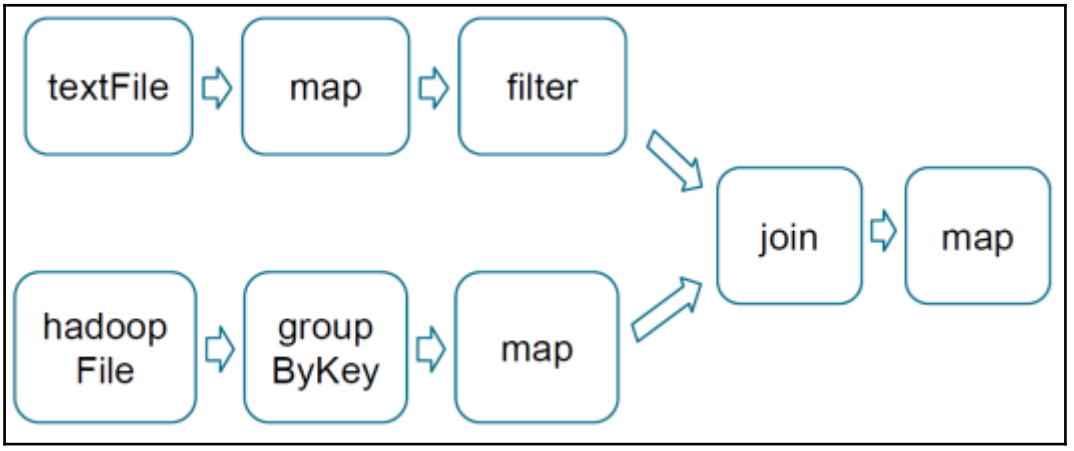
| Environment Variable | Meaning |
|---|---|
| SPARK_MASTER_HOST | Bind the master to a specific hostname or IP address, for example a public one. |
| SPARK_MASTER_PORT | Start the master on a different port (default: 7077). |
| SPARK_MASTER_WEBUI_PORT | Port for the master web UI (default: 8080). |
| SPARK_MASTER_OPTS | Configuration properties that apply only to the master in the form "-Dx=y" (default: none). See below for a list of possible options. |
| SPARK_LOCAL_DIRS | Directory to use for "scratch" space in Spark, including map output files and RDDs that get stored on disk. This should be on a fast, local disk in your system. It can also be a comma-separated list of multiple directories on different disks. |
| SPARK_WORKER_CORES | Total number of cores to allow Spark applications to use on the machine (default: all available cores). |
| SPARK_WORKER_MEMORY | Total amount of memory to allow Spark applications to use on the machine, e.g. `1000m`, `2g` (default: total memory minus 1 GB); note that each application's *individual* memory is configured using its `spark.executor.memory` property. |
| SPARK_WORKER_PORT | Start the Spark worker on a specific port (default: random). |
| SPARK_WORKER_WEBUI_PORT | Port for the worker web UI (default: 8081). |
| SPARK_WORKER_DIR | Directory to run applications in, which will include both logs and scratch space (default: SPARK_HOME/work). |
| SPARK_WORKER_OPTS | Configuration properties that apply only to the worker in the form "-Dx=y" (default: none). See below for a list of possible options. |
| SPARK_DAEMON_MEMORY | Memory to allocate to the Spark master and worker daemons themselves (default: 1g). |
| SPARK_DAEMON_JAVA_OPTS | JVM options for the Spark master and worker daemons themselves in the form "-Dx=y" (default: none). |
| SPARK_PUBLIC_DNS | The public DNS name of the Spark master and workers (default: none). |


Job
Stage
Task
Task
Task
Task
Stage
Task
Task
Task
Task


textFile
map
filter


hadoopFile
groupByKey
map

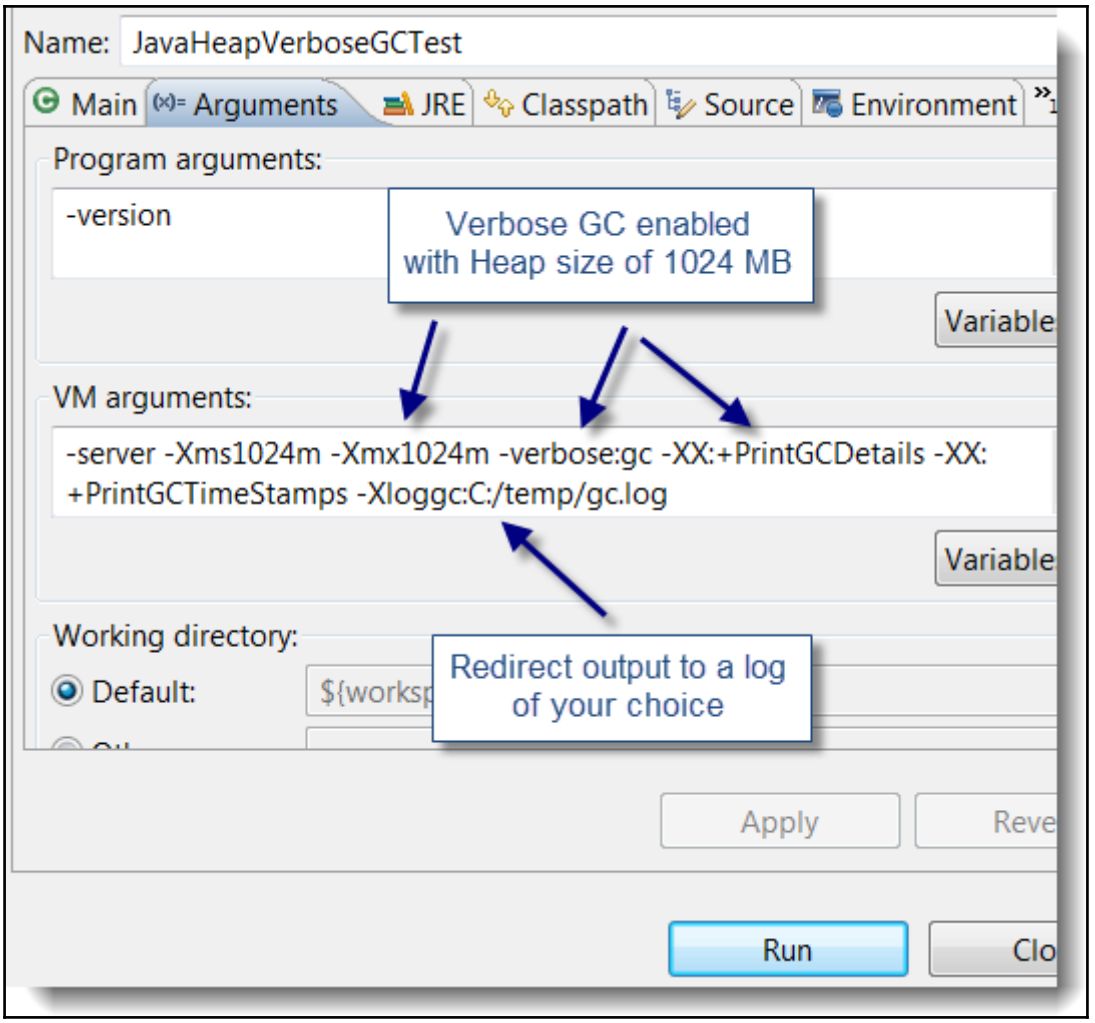

join
map

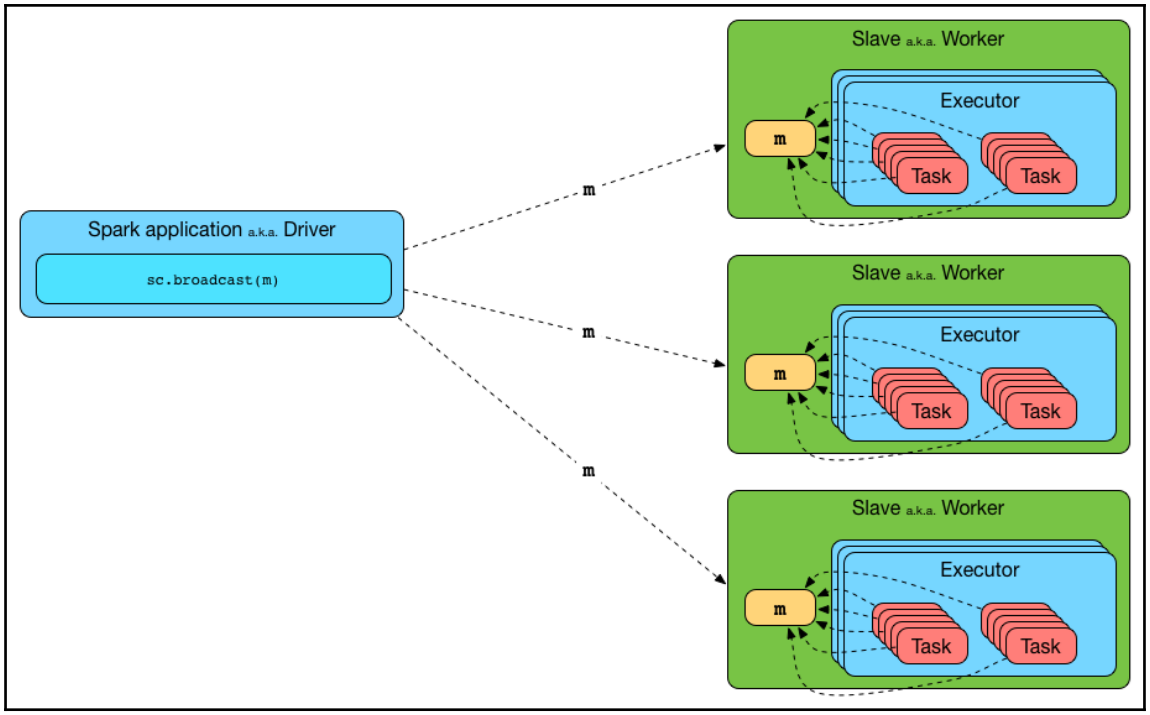

textFile
map
filter
hadoop File
group ByKey
map
join
map

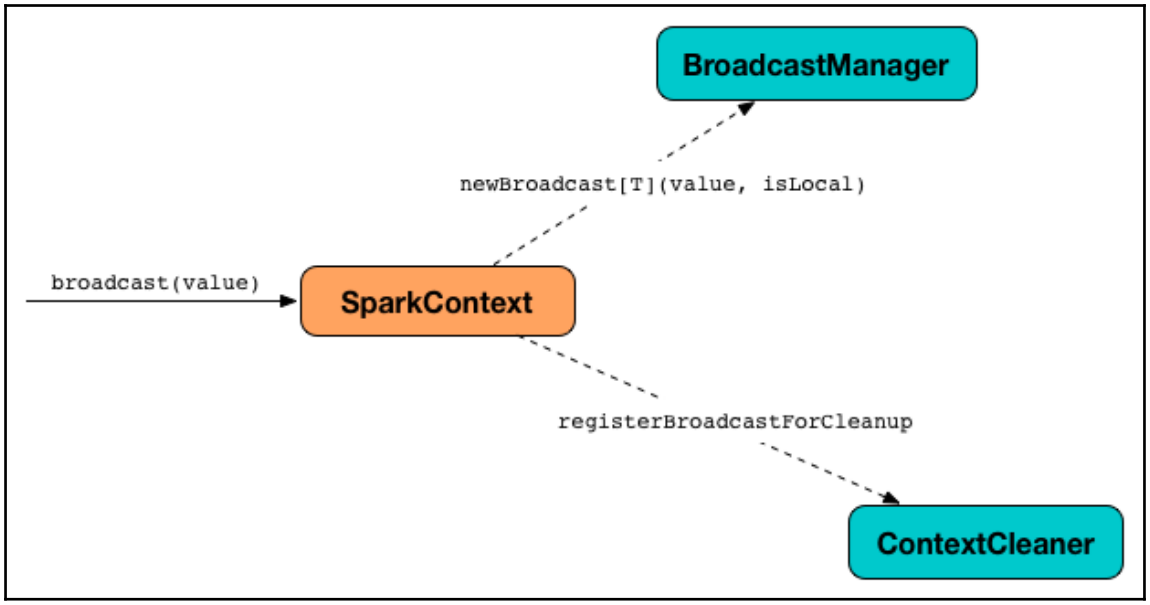
```
  /11/20 17:20:58 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
  /11/20 17:20:58 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
  /11/20 17:20:58 INFO DAGScheduler: Failed to run collect at ReceiverTracker.scala:270
  /11/20 17:20:58 INFO TaskSchedulerImpl: Cancelling stage 1
Exception in thread "Thread-53" org.apache.spark.SparkException: Job aborted due to stage failure: All masters are unresponsive!
Giving up.
        at org.apache.spark.scheduler.DAGScheduler.org$apache$spark$scheduler$DAGScheduler$$failJobAndIndependentStages
(DAGScheduler.scala:1033)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(DAGScheduler.scala:1017)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(DAGScheduler.scala:1015)
        at scala.collection.mutable.ResizableArray$class.foreach(ResizableArray.scala:59)
        at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:47)
        at org.apache.spark.scheduler.DAGScheduler.abortStage(DAGScheduler.scala:1015)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskSetFailed$1.apply(DAGScheduler.scala:633)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskSetFailed$1.apply(DAGScheduler.scala:633)
        at scala.Option.foreach(Option.scala:236)
        at org.apache.spark.scheduler.DAGScheduler.handleTaskSetFailed(DAGScheduler.scala:633)
        at org.apache.spark.scheduler.DAGSchedulerEventProcessActor$$anonfun$receive$2.applyOrElse(DAGScheduler.scala:1207)
        at akka.actor.ActorCell.receiveMessage(ActorCell.scala:498)
        at akka.actor.ActorCell.invoke(ActorCell.scala:456)
        at akka.dispatch.Mailbox.processMailbox(Mailbox.scala:237)
        at akka.dispatch.Mailbox.run(Mailbox.scala:219)
        at akka.dispatch.ForkJoinExecutorConfigurator$AkkaForkJoinTask.exec(AbstractDispatcher.scala:386)
        at scala.concurrent.forkjoin.ForkJoinTask.doExec(ForkJoinTask.java:260)
        at scala.concurrent.forkjoin.ForkJoinPool$WorkQueue.runTask(ForkJoinPool.java:1339)
        at scala.concurrent.forkjoin.ForkJoinPool.runWorker(ForkJoinPool.java:1979)
        at scala.concurrent.forkjoin.ForkJoinWorkerThread.run(ForkJoinWorkerThread.java:107)
  /11/20 17:20:58 INFO DAGScheduler: Failed to run take at DStream.scala:593
  /11/20 17:20:58 INFO TaskSchedulerImpl: Cancelling stage 2
  /11/20 17:20:58 INFO JobScheduler: Starting job streaming job 1416484202000 ms.0 from job set of time 1416484202000 ms
  /11/20 17:20:58 INFO SparkContext: Starting job: take at DStream.scala:593
  /11/20 17:20:58 ERROR JobScheduler: Error running job streaming job 1416484200000 ms.0
org.apache.spark.SparkException: Job aborted due to stage failure: All masters are unresponsive! Giving up.
        at org.apache.spark.scheduler.DAGScheduler.org$apache$spark$scheduler$DAGScheduler$$failJobAndIndependentStages
(DAGScheduler.scala:1033)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(DAGScheduler.scala:1017)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(DAGScheduler.scala:1015)
        at scala.collection.mutable.ResizableArray$class.foreach(ResizableArray.scala:59)
        at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:47)
        at org.apache.spark.scheduler.DAGScheduler.abortStage(DAGScheduler.scala:1015)
```

| Job Scheduling Information | Diagnostic Info |
|---|---|
| NA | Job initialization failed: java.io.IOException: Filesystem closed at org.apache.hadoop.hdfs.DFSClient.checkOpen(DFSClient.java:241) at org.apache.hadoop.hdfs.DFSClient.access$800(DFSClient.java:74) at org.apache.hadoop.hdfs.DFSClient$DFSOutputStream.closeInternal(DFSClient.java:3667) at org.apache.hadoop.hdfs.DFSClient$DFSOutputStream.close(DFSClient.java:3626) at org.apache.hadoop.fs.FSDataOutputStream$PositionCache.close(FSDataOutputStream.java:61) at org.apache.hadoop.fs.FSDataOutputStream.close(FSDataOutputStream.java:86) at org.apache.hadoop.security.Credentials.writeTokenStorageFile(Credentials.java:171) at org.apache.hadoop.mapred.JobInProgress.generateAndStoreTokens(JobInProgress.java:3528) at org.apache.hadoop.mapred.JobInProgress.initTasks(JobInProgress.java:696) at org.apache.hadoop.mapred.JobTracker.initJob(JobTracker.java:4207) at org.apache.hadoop.mapred.FairScheduler$JobInitializer$InitJob.run(FairScheduler.java:291) at java.util.concurrent.ThreadPoolExecutor$Worker.runTask(ThreadPoolExecutor.java:886) at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:908) at java.lang.Thread.run(Thread.java:662) |

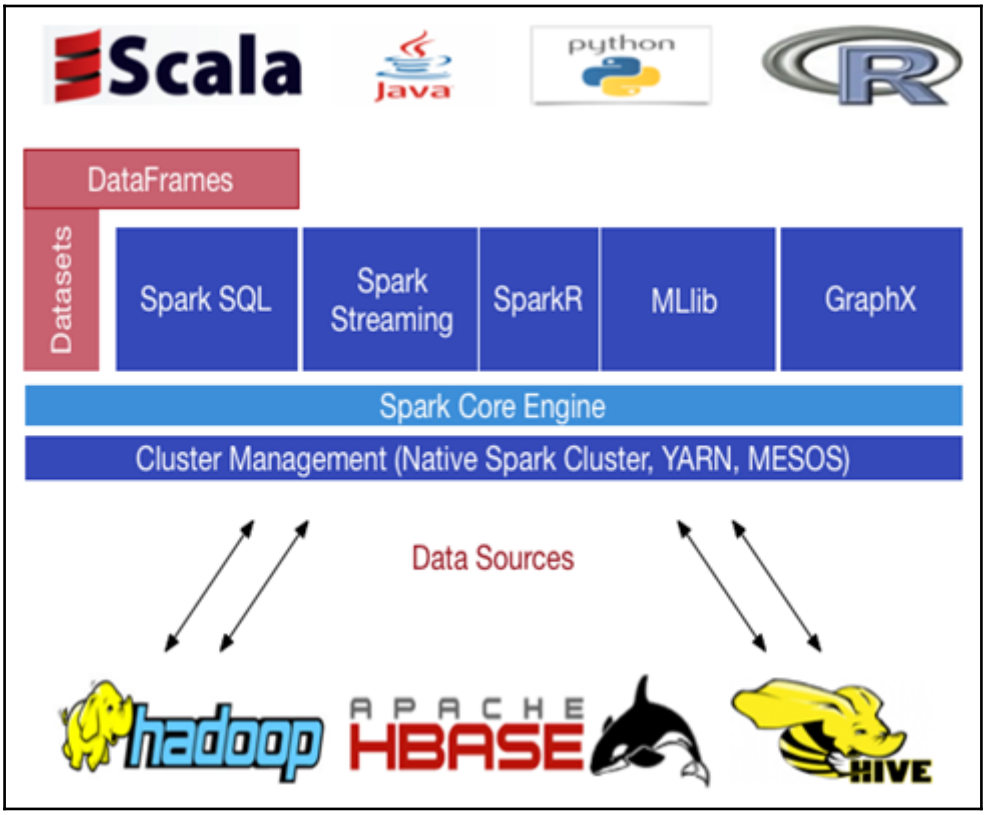
| Task Index | Task ID | Status | Locality Level | Executor | Launch Time | Duration | GC Time |
|---|---|---|---|---|---|---|---|
| 1 | 0 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 12.82 h | 9.59 h |
| 2 | 1 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 12.00 h | 8.97 h |
| 3 | 2 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 12.39 h | 9.16 h |
| 0 | 3 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 12.09 h | 8.88 h |
| 6 | 4 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 11.65 h | 8.54 h |
| 4 | 5 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 11.68 h | 8.62 h |
| 7 | 6 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 12.19 h | 9.12 h |
| 12 | 7 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 11.62 h | 8.50 h |
| 8 | 8 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 12.57 h | 9.40 h |
| 9 | 9 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 12.02 h | 8.98 h |
| 5 | 10 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 12.24 h | 9.04 h |
| 11 | 11 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 11.11 h | 8.15 h |
| 10 | 12 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 11.84 h | 8.68 h |
| 13 | 13 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 11.85 h | 8.74 h |
| 18 | 14 | SUCCESS | NODE_LOCAL | [redacted] | 2014/06/13 13:14:16 | 12.26 h | 9.17 h |

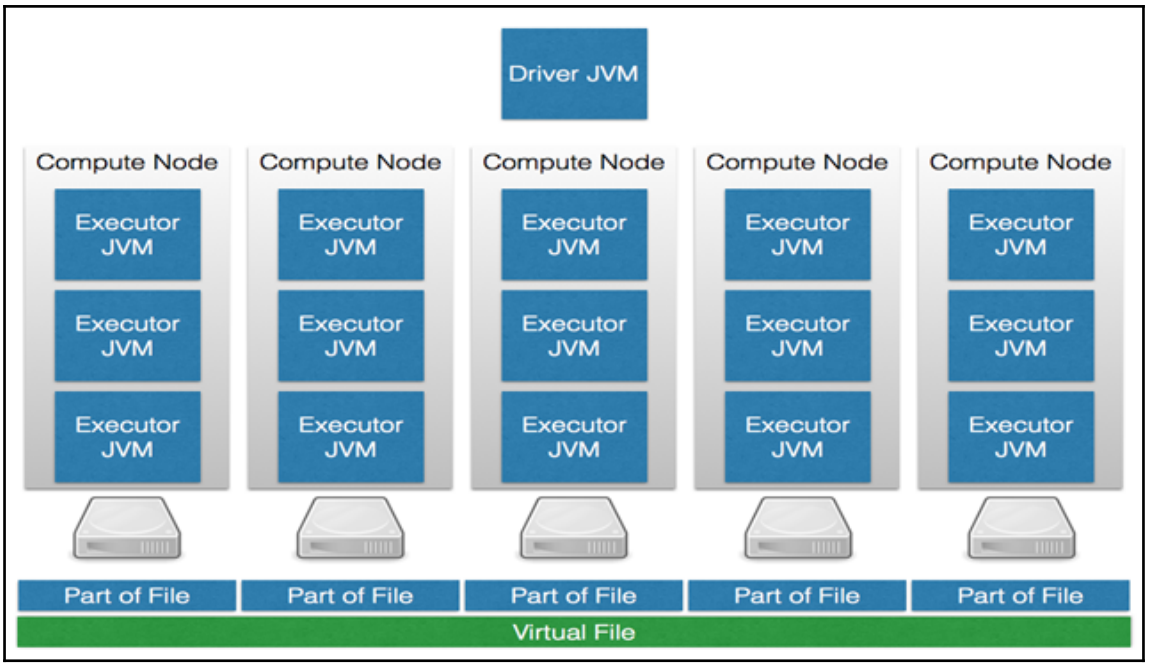

Name: JavaHeapVerboseGCTest
Main
Arguments
JRE
Classpath
Source
Environment
Program arguments:
-version
Verbose GC enabled
with Heap size of 1024 MB
Variable
VM arguments:
-server -Xms1024m -Xmx1024m -verbose:gc -XX:+PrintGCDetails -XX:
+PrintGCTimeStamps -Xloggc:C:/temp/gc.log
Variable
Working directory:
Default:
Redirect output to a log
of your choice
Apply
Run

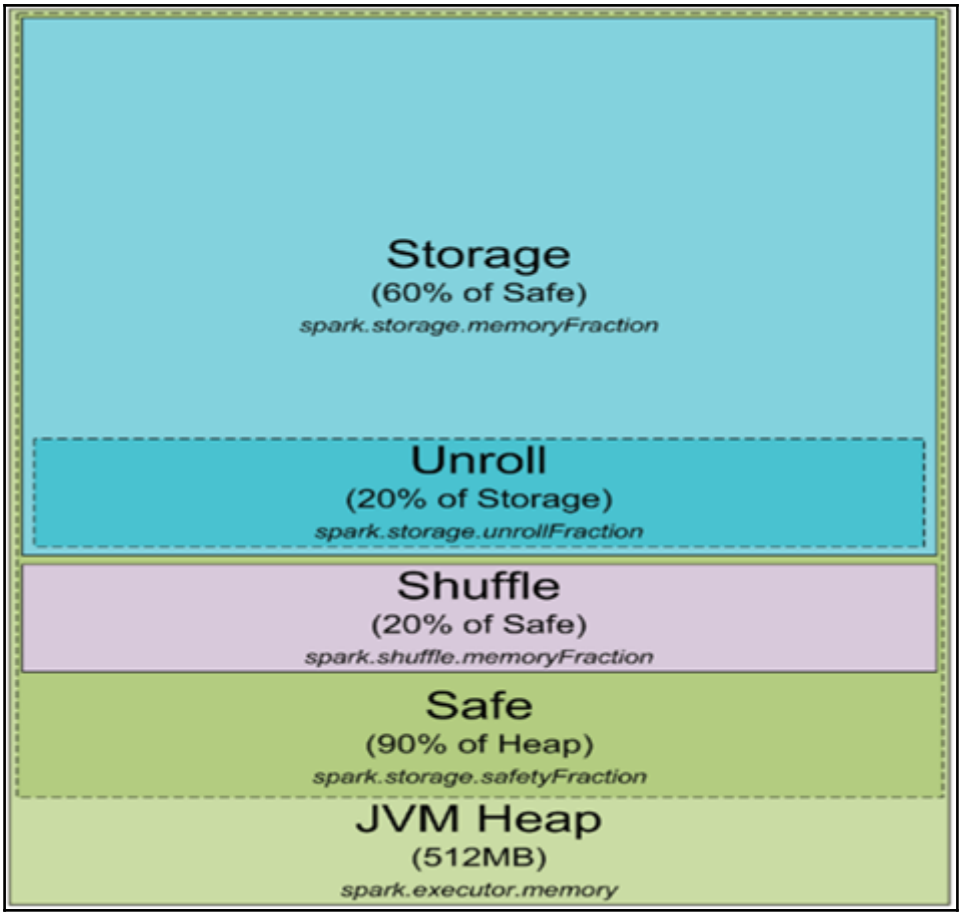

Spark application a.k.a. Driver
sc.broadcast(m)
m
m
m
Slave a.k.a. Worker
Executor
m
Task
Task
Slave a.k.a. Worker
Executor
m
Task
Task
Slave a.k.a. Worker
Executor
m
Task
Task

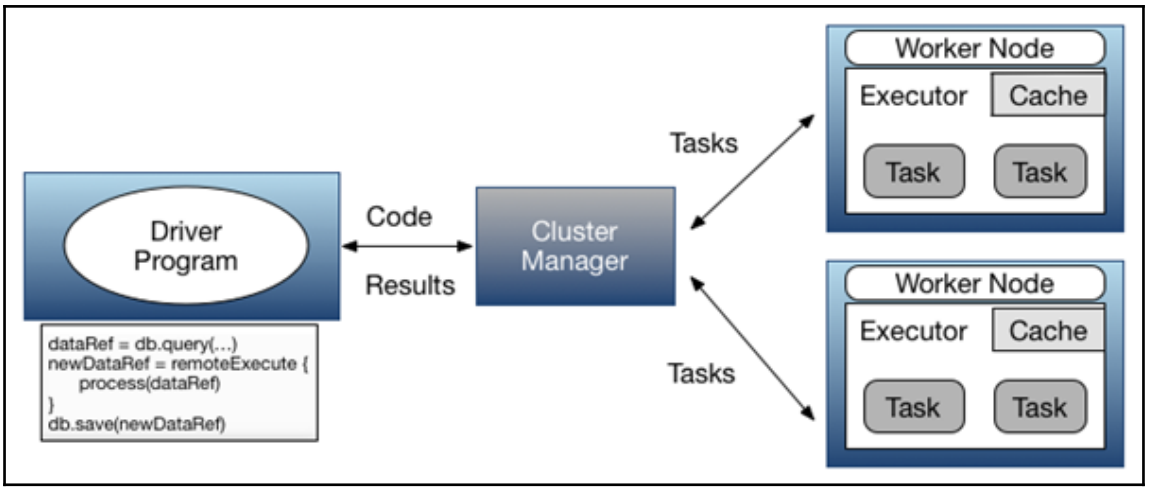

BroadcastManager
newBroadcast[T](value, isLocal)
broadcast(value)
SparkContext
registerBroadcastForCleanup
ContextCleaner

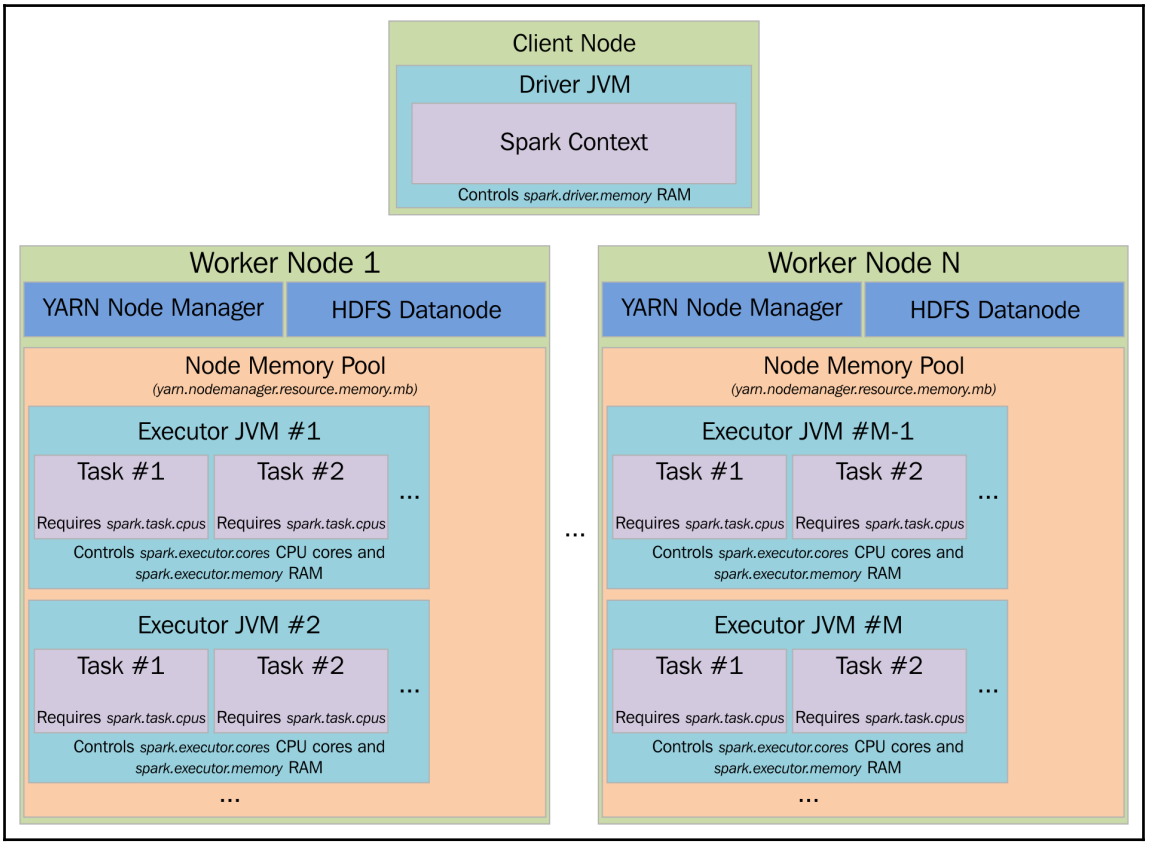
# Chapter 17: Time to Go to ClusterLand - Deploying Spark on a Cluster

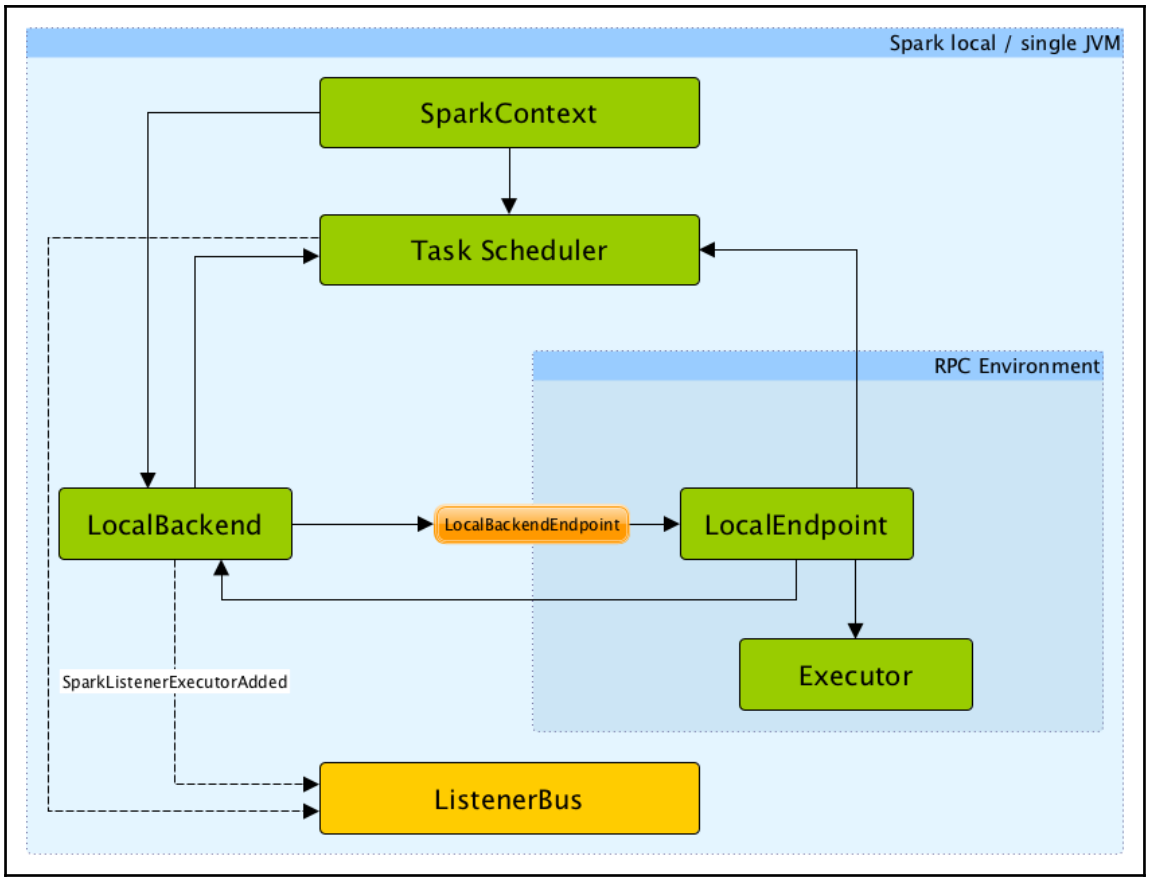


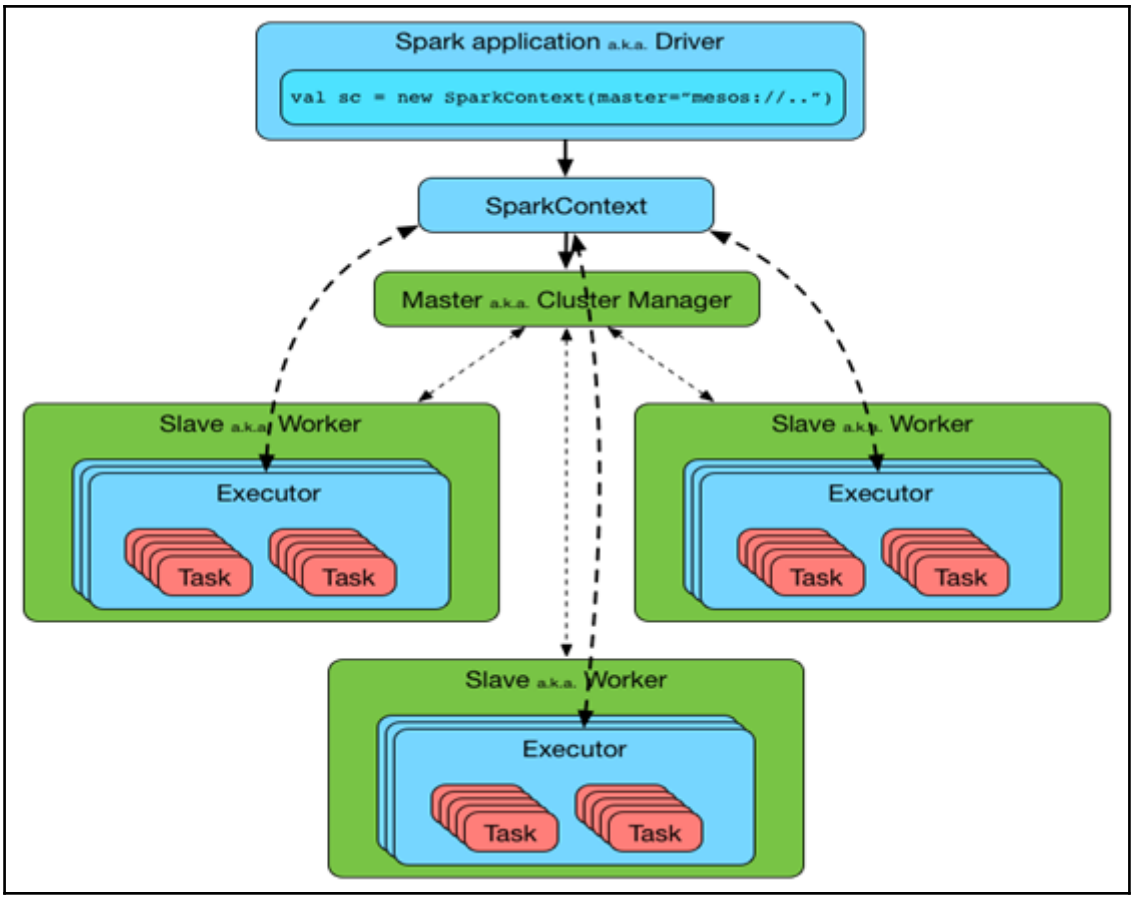

Driver JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Compute Node
Executor JVM
Executor JVM
Executor JVM
Part of File
Part of File
Part of File
Part of File
Part of File
Virtual File


Storage
(60% of Safe)
spark.storage.memoryFraction
Unroll
(20% of Storage)
spark.storage.unrollFraction
Shuffle
(20% of Safe)
spark.shuffle.memoryFraction
Safe
(90% of Heap)
spark.storage.safetyFraction
JVM Heap
(512MB)
spark.executor.memory


Driver Program
dataRef = db.query(...)
newDataRef = remoteExecute {
process(dataRef)
}
db.save(newDataRef)
Code
Results
Cluster Manager
Tasks
Tasks
Worker Node
Executor
Cache
Task
Task
Worker Node
Executor
Cache
Task
Task


Client Node
Driver JVM
Spark Context
Controls spark.driver.memory RAM
Worker Node 1
YARN Node Manager
HDFS Datanode
Node Memory Pool
(yarn.nodemanager.resource.memory.mb)
Executor JVM #1
Task #1
Requires spark.task.cpus
Task #2
Requires spark.task.cpus
...
Controls spark.executor.cores CPU cores and spark.executor.memory RAM
Executor JVM #2
Task #1
Requires spark.task.cpus
Task #2
Requires spark.task.cpus
...
Controls spark.executor.cores CPU cores and spark.executor.memory RAM
...
...
Worker Node N
YARN Node Manager
HDFS Datanode
Node Memory Pool
(yarn.nodemanager.resource.memory.mb)
Executor JVM #M-1
Task #1
Requires spark.task.cpus
Task #2
Requires spark.task.cpus
...
Controls spark.executor.cores CPU cores and spark.executor.memory RAM
Executor JVM #M
Task #1
Requires spark.task.cpus
Task #2
Requires spark.task.cpus
...
Controls spark.executor.cores CPU cores and spark.executor.memory RAM
...


Spark local / single JVM
SparkContext
Task Scheduler
RPC Environment
LocalBackend
LocalBackendEndpoint
LocalEndpoint
Executor
SparkListenerExecutorAdded
ListenerBus


Spark application a.k.a. Driver
val sc = new SparkContext(master="mesos://..")
SparkContext
Master a.k.a. Cluster Manager
Slave a.k.a. Worker
Executor
Task
Task
Slave a.k.a. Worker
Executor
Task
Task
Slave a.k.a. Worker
Executor
Task
Task

| Term | Meaning |
|---|---|
| Application | User program built on Spark. Consists of a *driver program* and *executors* on the cluster. |
| Application jar | A jar containing the user's Spark application. In some cases users will want to create an "uber jar" containing their application along with its dependencies. The user's jar should never include Hadoop or Spark libraries, however, these will be added at runtime. |
| Driver program | The process running the main() function of the application and creating the SparkContext |
| Cluster manager | An external service for acquiring resources on the cluster (e.g. standalone manager, Mesos, YARN) |
| Deploy mode | Distinguishes where the driver process runs. In "cluster" mode, the framework launches the driver inside of the cluster. In "client" mode, the submitter launches the driver outside of the cluster. |
| Worker node | Any node that can run application code in the cluster |
| Executor | A process launched for an application on a worker node, that runs tasks and keeps data in memory or disk storage across them. Each application has its own executors. |
| Task | A unit of work that will be sent to one executor |
| Job | A parallel computation consisting of multiple tasks that gets spawned in response to a Spark action (e.g. `save`, `collect`); you'll see this term used in the driver's logs. |
| Stage | Each job gets divided into smaller sets of tasks called *stages* that depend on each other (similar to the map and reduce stages in MapReduce); you'll see this term used in the driver's logs. |

| Master URL | Meaning |
|---|---|
| `local` | Run Spark locally with one worker thread (i.e. no parallelism at all). |
| `local[K]` | Run Spark locally with K worker threads (ideally, set this to the number of cores on your machine). |
| `local[*]` | Run Spark locally with as many worker threads as logical cores on your machine. |
| `spark://HOST:PORT` | Connect to the given Spark standalone cluster master. The port must be whichever one your master is configured to use, which is 7077 by default. |
| `mesos://HOST:PORT` | Connect to the given Mesos cluster. The port must be whichever one your is configured to use, which is 5050 by default. Or, for a Mesos cluster using ZooKeeper, use `mesos://zk://...`. To submit with `--deploy-mode cluster`, the HOST:PORT should be configured to connect to the MesosClusterDispatcher. |
| `yarn` | Connect to a YARN cluster in `client` or `cluster` mode depending on the value of `--deploy-mode`. The cluster location will be found based on the `HADOOP_CONF_DIR` or `YARN_CONF_DIR` variable. |

```
17/02/14 12:31:02 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 3343 bytes result sent to driver
17/02/14 12:31:02 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 215 ms on localhost (executor driver) (1/1)
17/02/14 12:31:02 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
17/02/14 12:31:02 INFO DAGScheduler: ResultStage 0 (show at KMeansDemo.scala:56) finished in 0.225 s
17/02/14 12:31:02 INFO DAGScheduler: Job 0 finished: show at KMeansDemo.scala:56, took 0.322031 s
17/02/14 12:31:02 INFO CodeGenerator: Code generated in 19.812394 ms
+--------+-------+----------+-----+---------+------------+----------+--------+--------+----------+----------+----------+--------+----------+---------+-----+
|   Price|LotSize|Waterfront|  Age|LandValue|NewConstruct|CentralAir|FuelType|HeatType|SewerType|LivingArea|PctCollege|Bedrooms|Fireplaces|Bathrooms|rooms|
+--------+-------+----------+-----+---------+------------+----------+--------+--------+----------+----------+----------+--------+----------+---------+-----+
|132500.0|   0.09|       0.0| 42.0|  50000.0|         0.0|       0.0|     3.0|     4.0|       2.0|     906.0|      35.0|     2.0|       1.0|      1.0|  5.0|
|181115.0|   0.92|       0.0|  0.0|  22300.0|         0.0|       0.0|     2.0|     3.0|       2.0|    1953.0|      51.0|     3.0|       0.0|      2.5|  6.0|
|109000.0|   0.19|       0.0|133.0|   7300.0|         0.0|       0.0|     2.0|     3.0|       3.0|    1944.0|      51.0|     4.0|       1.0|      1.0|  8.0|
|155000.0|   0.41|       0.0| 13.0|  18700.0|         0.0|       0.0|     2.0|     2.0|       2.0|    1944.0|      51.0|     3.0|       1.0|      1.5|  5.0|
| 86060.0|   0.11|       0.0|  0.0|  15000.0|         1.0|       1.0|     2.0|     2.0|       3.0|     840.0|      51.0|     2.0|       0.0|      1.0|  3.0|
|120000.0|   0.68|       0.0| 31.0|  14000.0|         0.0|       0.0|     2.0|     2.0|       2.0|    1152.0|      22.0|     4.0|       1.0|      1.0|  8.0|
|153000.0|    0.4|       0.0| 33.0|  23300.0|         0.0|       0.0|     4.0|     3.0|       2.0|    2752.0|      51.0|     4.0|       1.0|      1.5|  8.0|
|170000.0|   1.21|       0.0| 23.0|  14600.0|         0.0|       0.0|     4.0|     2.0|       2.0|    1662.0|      35.0|     4.0|       1.0|      1.5|  9.0|
| 90000.0|   0.83|       0.0| 36.0|  22200.0|         0.0|       0.0|     3.0|     4.0|       2.0|    1632.0|      51.0|     3.0|       0.0|      1.5|  8.0|
|122900.0|   1.94|       0.0|  4.0|  21200.0|         0.0|       0.0|     2.0|     2.0|       1.0|    1416.0|      44.0|     3.0|       0.0|      1.5|  6.0|
|325000.0|   2.29|       0.0|123.0|  12600.0|         0.0|       0.0|     4.0|     2.0|       2.0|    2894.0|      51.0|     7.0|       0.0|      1.0| 12.0|
|120000.0|   0.92|       0.0|  1.0|  22300.0|         0.0|       0.0|     2.0|     2.0|       2.0|    1624.0|      51.0|     3.0|       0.0|      2.0|  6.0|
| 85860.0|   8.97|       0.0| 13.0|   4800.0|         0.0|       0.0|     3.0|     4.0|       2.0|     704.0|      41.0|     2.0|       0.0|      1.0|  4.0|
| 97000.0|   0.11|       0.0|153.0|   3100.0|         0.0|       0.0|     2.0|     3.0|       3.0|    1383.0|      57.0|     3.0|       0.0|      2.0|  5.0|
|127000.0|   0.14|       0.0|  9.0|    300.0|         0.0|       0.0|     4.0|     2.0|       2.0|    1300.0|      41.0|     3.0|       0.0|      1.5|  8.0|
| 89900.0|    0.0|       0.0| 88.0|   2500.0|         0.0|       0.0|     2.0|     3.0|       3.0|     936.0|      57.0|     3.0|       0.0|      1.0|  4.0|
|155000.0|   0.13|       0.0|  9.0|    300.0|         0.0|       0.0|     4.0|     2.0|       2.0|    1300.0|      41.0|     3.0|       0.0|      1.5|  7.0|
|253750.0|    2.0|       0.0|  0.0|  49800.0|         0.0|       1.0|     2.0|     2.0|       1.0|    2816.0|      71.0|     4.0|       1.0|      2.5| 12.0|
| 60000.0|   0.21|       0.0| 82.0|   8500.0|         0.0|       0.0|     4.0|     3.0|       2.0|     924.0|      35.0|     2.0|       0.0|      1.0|  6.0|
| 87500.0|   0.88|       0.0| 17.0|  19400.0|         0.0|       0.0|     4.0|     2.0|       2.0|    1092.0|      35.0|     3.0|       0.0|      1.0|  6.0|
+--------+-------+----------+-----+---------+------------+----------+--------+--------+----------+----------+----------+--------+----------+---------+-----+
only showing top 20 rows

17/02/14 12:31:02 INFO ContextCleaner: Cleaned accumulator 3
17/02/14 12:31:03 INFO BlockManagerInfo: Removed broadcast_1_piece0 on 10.2.16.255:53581 in memory (size: 9.4 KB, free: 4.0 GB)
17/02/14 12:31:03 INFO SparkContext: Starting job: takeSample at KMeans.scala:353
17/02/14 12:31:03 INFO DAGScheduler: Got job 1 (takeSample at KMeans.scala:353) with 2 output partitions
17/02/14 12:31:03 INFO DAGScheduler: Final stage: ResultStage 1 (takeSample at KMeans.scala:353)
17/02/14 12:31:03 INFO DAGScheduler: Parents of final stage: List()
```

Spark 2.1.0 **Spark Master at spark://ubuntu:7077**

**URL:** spark://ubuntu:7077
**REST URL:** spark://ubuntu:6066 *(cluster mode)*
**Alive Workers:** 0
**Cores in use:** 0 Total, 0 Used
**Memory in use:** 0.0 B Total, 0.0 B Used
**Applications:** 0 Running, 0 Completed
**Drivers:** 0 Running, 0 Completed
**Status:** ALIVE

**Workers**

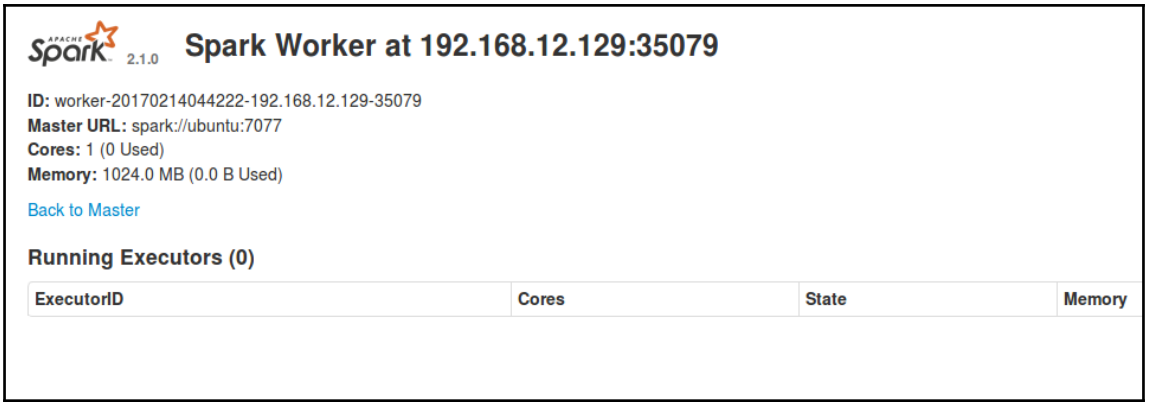
| Worker Id | Address | State |
|---|---|---|

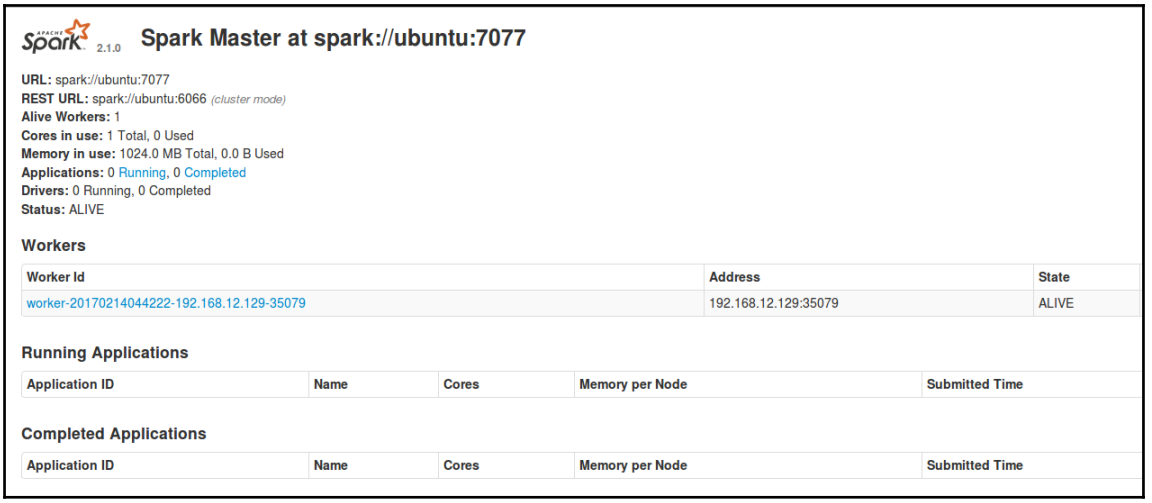
**Running Applications**

| Application ID | Name | Cores | Memory per Node | Submitted Time |
|---|---|---|---|---|

**Completed Applications**

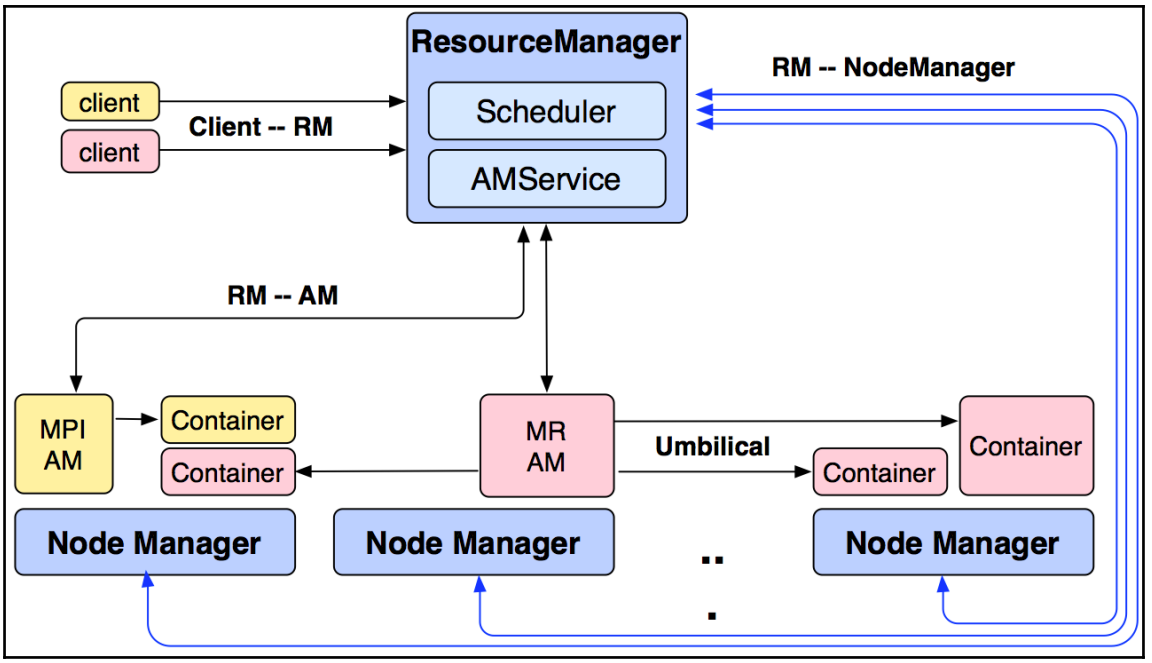
| Application ID | Name | Cores | Memory per Node | Submitted Time |
|---|---|---|---|---|

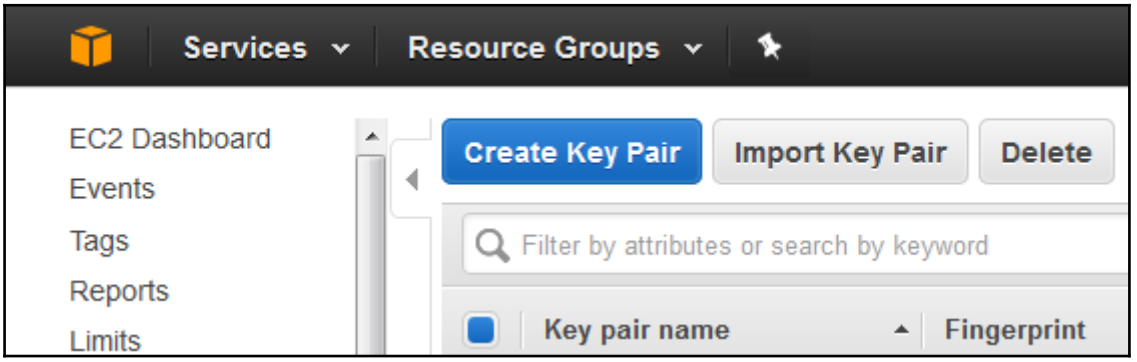

Spark 2.1.0
Spark Worker at 192.168.12.129:35079
ID: worker-20170214044222-192.168.12.129-35079
Master URL: spark://ubuntu:7077
Cores: 1 (0 Used)
Memory: 1024.0 MB (0.0 B Used)
Back to Master
Running Executors (0)
ExecutorID
Cores
State
Memory


Spark 2.1.0
Spark Master at spark://ubuntu:7077
URL: spark://ubuntu:7077
REST URL: spark://ubuntu:6066 (cluster mode)
Alive Workers: 1
Cores in use: 1 Total, 0 Used
Memory in use: 1024.0 MB Total, 0.0 B Used
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE
Workers
Worker Id
Address
State
worker-20170214044222-192.168.12.129-35079
192.168.12.129:35079
ALIVE
Running Applications
Application ID
Name
Cores
Memory per Node
Submitted Time
Completed Applications
Application ID
Name
Cores
Memory per Node
Submitted Time

| Argument | Meaning |
|---|---|
| `-h HOST`, `--host HOST` | Hostname to listen on |
| `-i HOST`, `--ip HOST` | Hostname to listen on (deprecated, use -h or --host) |
| `-p PORT`, `--port PORT` | Port for service to listen on (default: 7077 for master, random for worker) |
| `--webui-port PORT` | Port for web UI (default: 8080 for master, 8081 for worker) |
| `-c CORES`, `--cores CORES` | Total CPU cores to allow Spark applications to use on the machine (default: all available); only on worker |
| `-m MEM`, `--memory MEM` | Total amount of memory to allow Spark applications to use on the machine, in a format like 1000M or 2G (default: your machine's total RAM minus 1 GB); only on worker |
| `-d DIR`, `--work-dir DIR` | Directory to use for scratch space and job output logs (default: SPARK_HOME/work); only on worker |
| `--properties-file FILE` | Path to a custom Spark properties file to load (default: conf/spark-defaults.conf) |

- `sbin/start-master.sh` - Starts a master instance on the machine the script is executed on.
- `sbin/start-slaves.sh` - Starts a slave instance on each machine specified in the `conf/slaves` file.
- `sbin/start-slave.sh` - Starts a slave instance on the machine the script is executed on.
- `sbin/start-all.sh` - Starts both a master and a number of slaves as described above.
- `sbin/stop-master.sh` - Stops the master that was started via the `bin/start-master.sh` script.
- `sbin/stop-slaves.sh` - Stops all slave instances on the machines specified in the `conf/slaves` file.
- `sbin/stop-all.sh` - Stops both the master and the slaves as described above.


ResourceManager
Scheduler
AMService
client
client
Client -- RM
RM -- NodeManager
RM -- AM
MPI AM
Container
Container
MR AM
Umbilical
Container
Container
Node Manager
Node Manager
Node Manager


Services
Resource Groups
EC2 Dashboard
Events
Tags
Reports
Limits
Create Key Pair
Import Key Pair
Delete
Filter by attributes or search by keyword
Key pair name
Fingerprint

ec2-52-19-229-38.eu-west-1.compute.amazonaws.com:8080

Spark 2.1.0

## Spark Master at spark://ec2-52-19-229-38.eu-west-1.compute.amazonaws.com:7077

**URL:** spark://ec2-52-19-229-38.eu-west-1.compute.amazonaws.com:7077
**REST URL:** spark://ec2-52-19-229-38.eu-west-1.compute.amazonaws.com:6066 *(cluster mode)*
**Alive Workers:** 2
**Cores in use:** 16 Total, 0 Used
**Memory in use:** 56.4 GB Total, 0.0 B Used
**Applications:** 0 Running, 0 Completed
**Drivers:** 0 Running, 0 Completed
**Status:** ALIVE

### Workers

| Worker Id | Address | State | Cores | Memory |
|---|---|---|---|---|
| worker-20160325211558-172.31.8.198-45995 | 172.31.8.198:45995 | ALIVE | 8 (0 Used) | 28.2 GB (0.0 B Used) |
| worker-20160325211601-172.31.8.197-44556 | 172.31.8.197:44556 | ALIVE | 8 (0 Used) | 28.2 GB (0.0 B Used) |

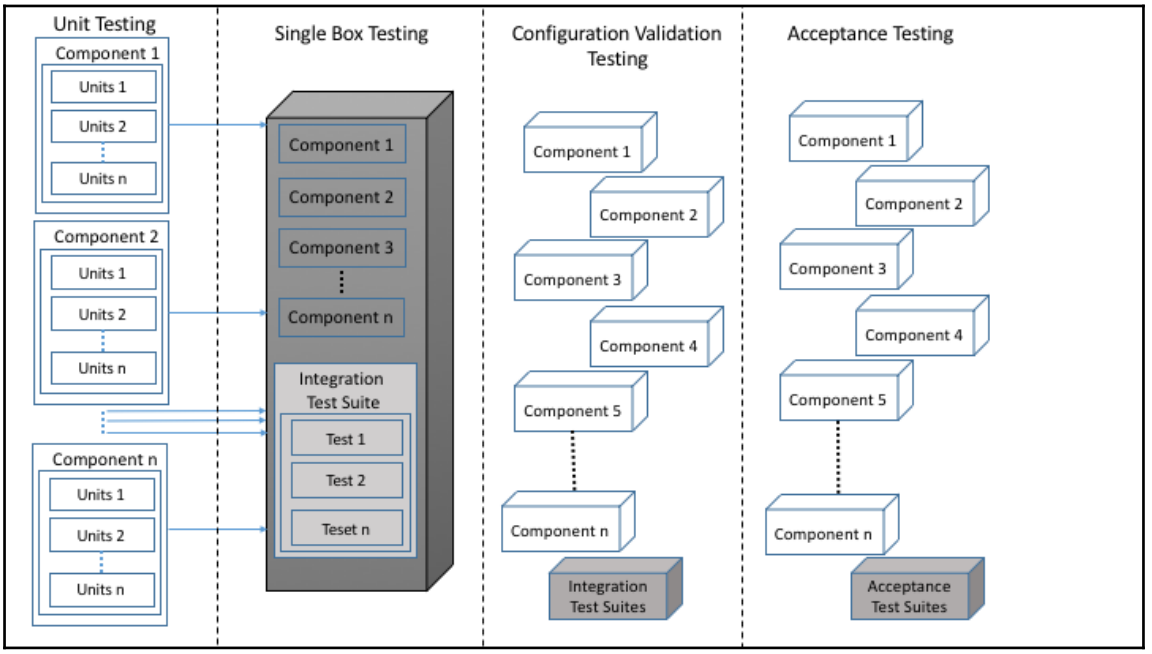
### Running Applications

| Application ID | Name | Cores | Memory per Node | Submitted Time | User | State | Duration |
|---|---|---|---|---|---|---|---|

### Completed Applications

| Application ID | Name | Cores | Memory per Node | Submitted Time | User | State | Duration |
|---|---|---|---|---|---|---|---|

# Chapter 18: Testing and Debugging Spark



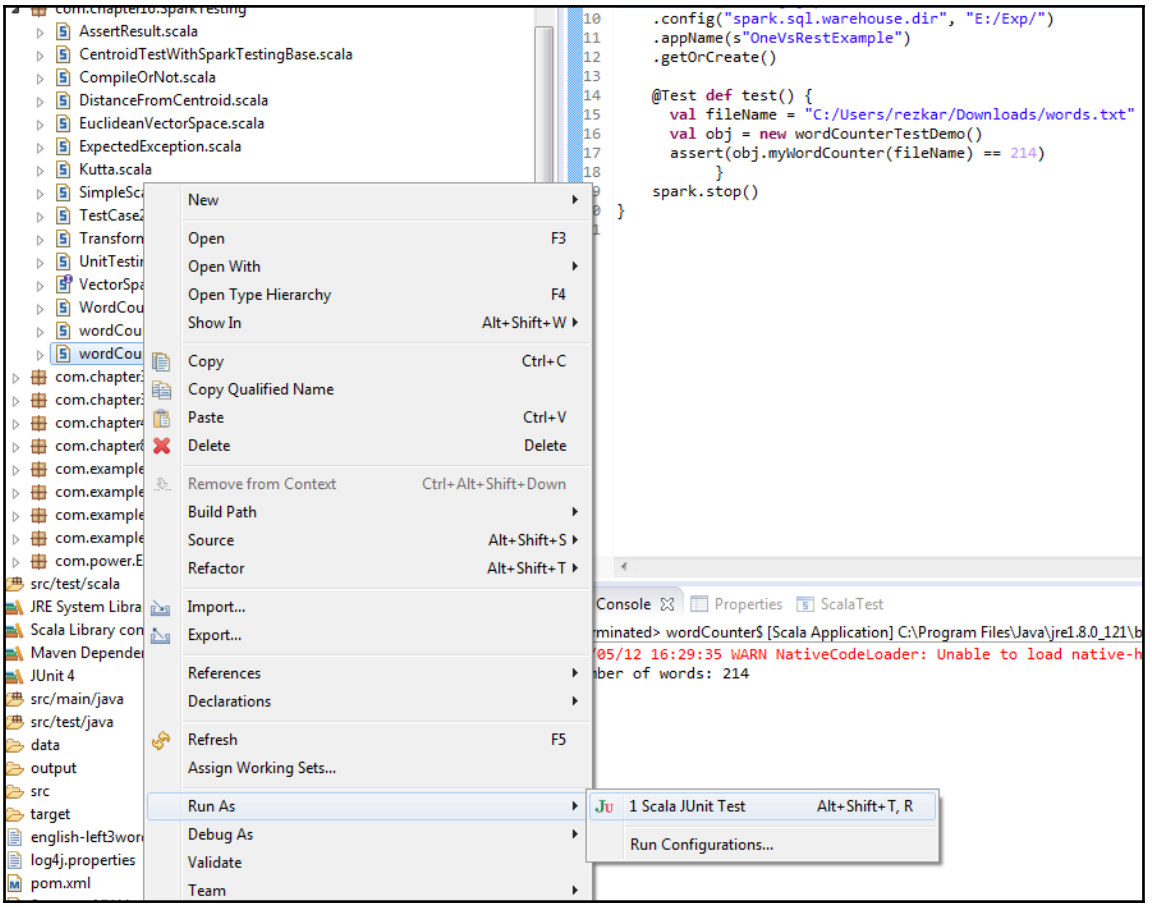

```
Exception in thread "main" java.lang.AssertionError: assertion failed
    at scala.Predef$.assert(Predef.scala:156)
    at com.chapter16.SparkTesting.SimpleScalaTest$.main(SimpleScalaTest.scala:7)
    at com.chapter16.SparkTesting.SimpleScalaTest.main(SimpleScalaTest.scala)
```

```
Exception in thread "main" org.scalatest.exceptions.TestFailedException: 2 did not equal 1
    at org.scalatest.Assertions$class.newAssertionFailedException(Assertions.scala:500)
    at org.scalatest.Assertions$.newAssertionFailedException(Assertions.scala:1538)
    at org.scalatest.Assertions$AssertionsHelper.macroAssert(Assertions.scala:466)
    at com.chapter16.SparkTesting.SimpleScalaTest$.main(SimpleScalaTest.scala:8)
    at com.chapter16.SparkTesting.SimpleScalaTest.main(SimpleScalaTest.scala)
```

```
Exception in thread "main" org.scalatest.exceptions.TestFailedException: Expected 3, but
got 4
    at org.scalatest.Assertions$class.newAssertionFailedException(Assertions.scala:495)
    at org.scalatest.Assertions$.newAssertionFailedException(Assertions.scala:1538)
    at org.scalatest.Assertions$class.assertResult(Assertions.scala:1226)
    at org.scalatest.Assertions$.assertResult(Assertions.scala:1538)
    at com.chapter16.SparkTesting.AssertResult$.main(AssertResult.scala:8)
    at com.chapter16.SparkTesting.AssertResult.main(AssertResult.scala)
```

```
Exception in thread "main" org.scalatest.exceptions.TestFailedException
     at org.scalatest.Assertions$class.newAssertionFailedException(Assertions.scala:493)
     at org.scalatest.Assertions$.newAssertionFailedException(Assertions.scala:1538)
     at org.scalatest.Assertions$class.fail(Assertions.scala:1313)
     at org.scalatest.Assertions$.fail(Assertions.scala:1538)
     at com.chapter16.SparkTesting.ExpectedException$.main(ExpectedException.scala:9)
     at com.chapter16.SparkTesting.ExpectedException.main(ExpectedException.scala)
```

```
AssertDoesNotCompile True
AssertTypeError True
AssertCompiles True
Exception in thread "main" org.scalatest.exceptions.TestFailedException: Expected a
compiler error, but got none for code: val a: Int = 1
    at com.chapter16.SparkTesting.CompileOrNot$.main(CompileOrNot.scala:15)
    at com.chapter16.SparkTesting.CompileOrNot.main(CompileOrNot.scala)
```


AssertResult.scala
CentroidTestWithSparkTestingBase.scala
CompileOrNot.scala
DistanceFromCentroid.scala
EuclideanVectorSpace.scala
ExpectedException.scala
Kutta.scala
src/test/scala
JRE System Libra
Scala Library con
Maven Depende
JUnit 4
src/main/java
src/test/java
data
output
src
target
log4j.properties
pom.xml
New
Open F3
Open With
Open Type Hierarchy F4
Show In Alt+Shift+W
Copy Ctrl+C
Copy Qualified Name
Paste Ctrl+V
Delete Delete
Remove from Context Ctrl+Alt+Shift+Down
Build Path
Source Alt+Shift+S
Refactor Alt+Shift+T
Import...
Export...
References
Declarations
Refresh F5
Assign Working Sets...
Run As
Debug As
Validate
Team
1 Scala JUnit Test Alt+Shift+T, R
Run Configurations...
.config("spark.sql.warehouse.dir", "E:/Exp/")
.appName(s"OneVsRestExample")
.getOrCreate()
@Test def test() {
val fileName = "C:/Users/rezkar/Downloads/words.txt"
val obj = new wordCounterTestDemo()
assert(obj.myWordCounter(fileName) == 214)
}
spark.stop()
}
Console
Properties
ScalaTest
ber of words: 214

File Edit Refactor Navigate Search Project Run Window Help

Package Explorer | JUnit

Finished after 2.597 seconds

Runs: 1/1 Errors: 0 Failures: 0

com.chapter16.SparkTesting.wordCountTest [Runner: Scala JUnit 4] (0.859 s)

wordCounterTestDemo.scala | wordCountTest.scala | WordCounter1.scala

```scala
1  package com.chapter16.SparkTesting
2  import org.scalatest.Assertions._
3  import org.junit.Test
4  import org.apache.spark.sql.SparkSession
5
6  class wordCountTest {
7    val spark = SparkSession
8      .builder
9      .master("local[*]")
10     .config("spark.sql.warehouse.dir", "E:/Exp/")
11     .appName(s"OneVsRestExample")
12     .getOrCreate()
13
14     @Test def test() {
15       val fileName = "C:/Users/rezkar/Downloads/words.txt"
16       val obj = new wordCounterTestDemo()
17       assert(obj.myWordCounter(fileName) == 214)
18           }
19     spark.stop()
20 }
21
```

File Edit Refactor Navigate Search Project Run Window Help

Package Explorer | JUnit

Finished after 2.602 seconds

Runs: 1/1 Errors: 1 Failures: 0

com.chapter16.SparkTesting.wordCountTest [Runner: Scala JUnit 4] (0.897 s)

test (0.897 s)

Failure Trace

org.scalatest.exceptions.TestFailedException: 214 did not equal 210

at org.scalatest.Assertions$class.newAssertionFailedException(Assertions.scala:500)

at org.scalatest.Assertions$.newAssertionFailedException(Assertions.scala:1538)

at org.scalatest.Assertions$AssertionsHelper.macroAssert(Assertions.scala:466)

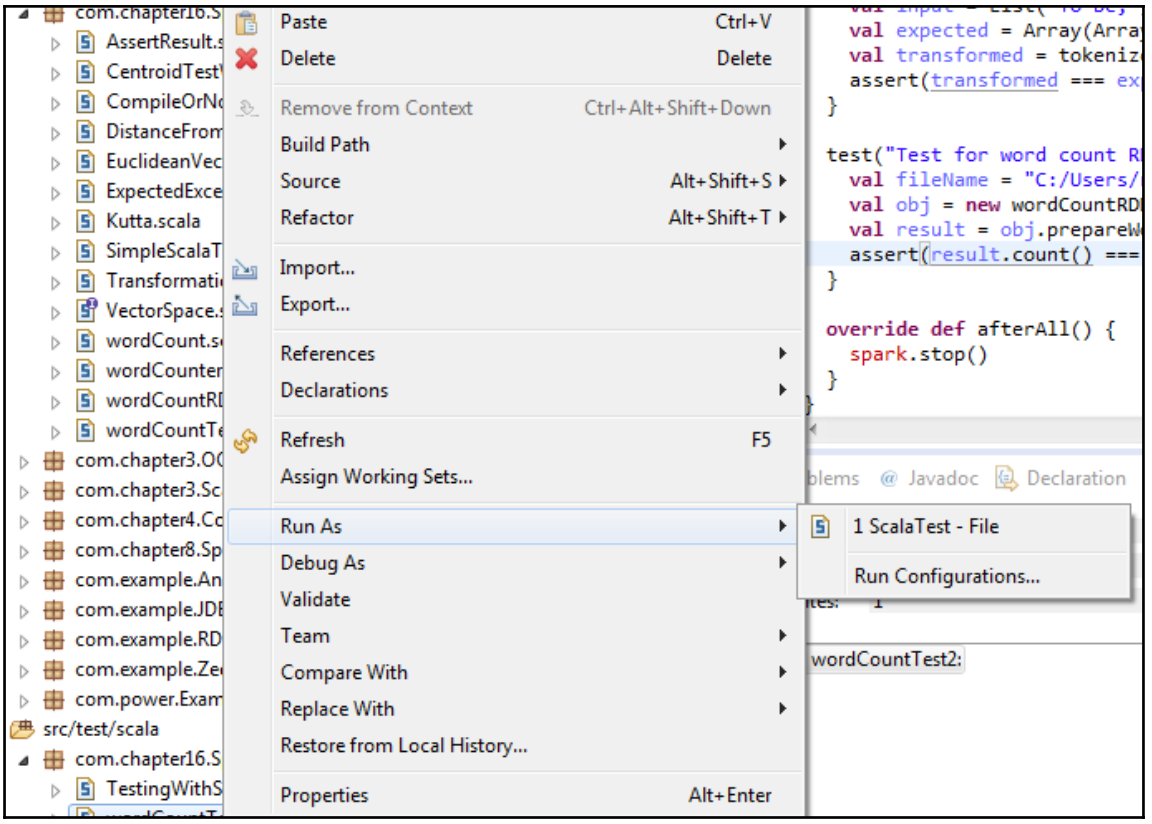
at com.chapter16.SparkTesting.wordCountTest.test(wordCountTest.scala:17)

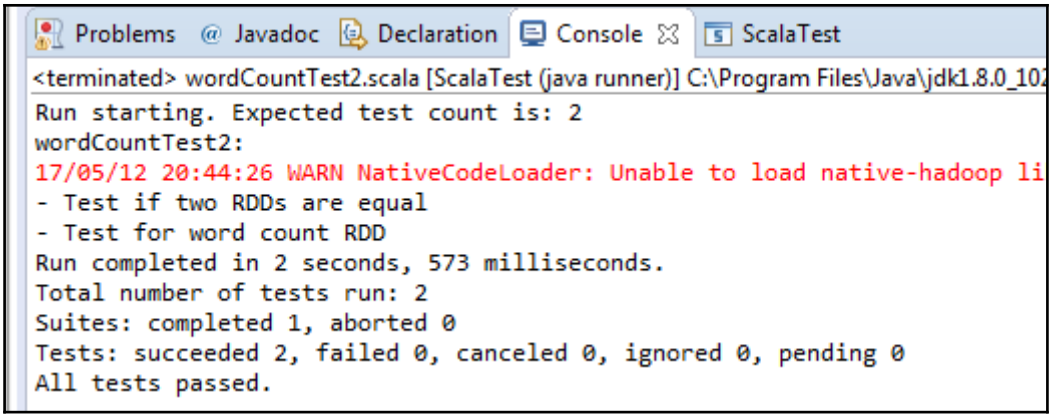
wordCounterTestDemo.scala | wordCountTest.scala | wordCount.

```scala
1  package com.chapter16.SparkTesting
2  import org.scalatest.Assertions._
3  import org.junit.Test
4  import org.apache.spark.sql.SparkSession
5
6  class wordCountTest {
7    val spark = SparkSession
8      .builder
9      .master("local[*]")
10     .config("spark.sql.warehouse.dir", "E:/Exp/")
11     .appName(s"OneVsRestExample")
12     .getOrCreate()
13
14     @Test def test() {
15       val fileName = "C:/Users/rezkar/Downloads/words.txt"
16       val obj = new wordCounterTestDemo()
17       assert(obj.myWordCounter(fileName) == 210)
18           }
19     spark.stop()
20 }
21
```

Console | Properties | ScalaTest

<terminated> wordCountTest [JUnit] C:\Program Files\Java\jre1.8.0_121\bin\javaw.ex


Paste Ctrl+V
Delete Delete
Remove from Context Ctrl+Alt+Shift+Down
Build Path
Source Alt+Shift+S
Refactor Alt+Shift+T
Import...
Export...
References
Declarations
Refresh F5
Assign Working Sets...
Run As
Debug As
Validate
Team
Compare With
Replace With
Restore from Local History...
Properties Alt+Enter
1 ScalaTest - File
Run Configurations...
wordCountTest2:

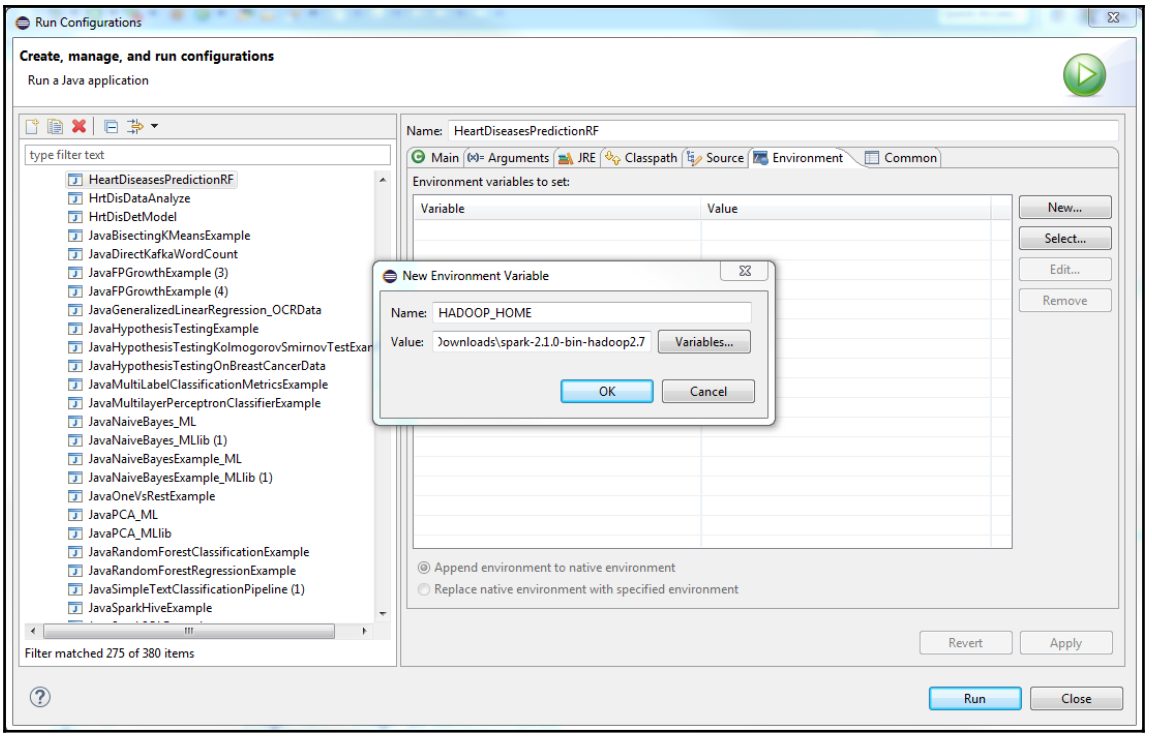

Problems
Javadoc
Declaration
Console
ScalaTest
<terminated> wordCountTest2.scala [ScalaTest (java runner)] C:\Program Files\Java\jdk1.8.0_102
Run starting. Expected test count is: 2
wordCountTest2:
17/05/12 20:44:26 WARN NativeCodeLoader: Unable to load native-hadoop li
- Test if two RDDs are equal
- Test for word count RDD
Run completed in 2 seconds, 573 milliseconds.
Total number of tests run: 2
Suites: completed 1, aborted 0
Tests: succeeded 2, failed 0, canceled 0, ignored 0, pending 0
All tests passed.

```
Problems  @ Javadoc  Declaration  Console  ScalaTest
<terminated> wordCountTest2.scala [ScalaTest (java runner)] C:\Program Files\Java\jdk1.8.0_102\bin\javaw.exe (12 May 2017, 20:47:04)
Run starting. Expected test count is: 2
wordCountTest2:
17/05/12 20:47:05 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

[Stage 0:>                                                          (0 + 0) / 8]
                                                                    - Test if two RDDs are equal *** FAILED ***
  Array(Array("To", "be,"), Array("or", "not", "to", "be:"), Array("that", "is", "the", "question-"), Array("William", "Shakespeare")) did not equal Array(Array("To", "be")
- Test for word count RDD *** FAILED ***
  214 did not equal 210 (wordCountTest2.scala:33)
Run completed in 2 seconds, 749 milliseconds.
Total number of tests run: 2
Suites: completed 1, aborted 0
Tests: succeeded 0, failed 2, canceled 0, ignored 0, pending 0
*** 2 TESTS FAILED ***
```

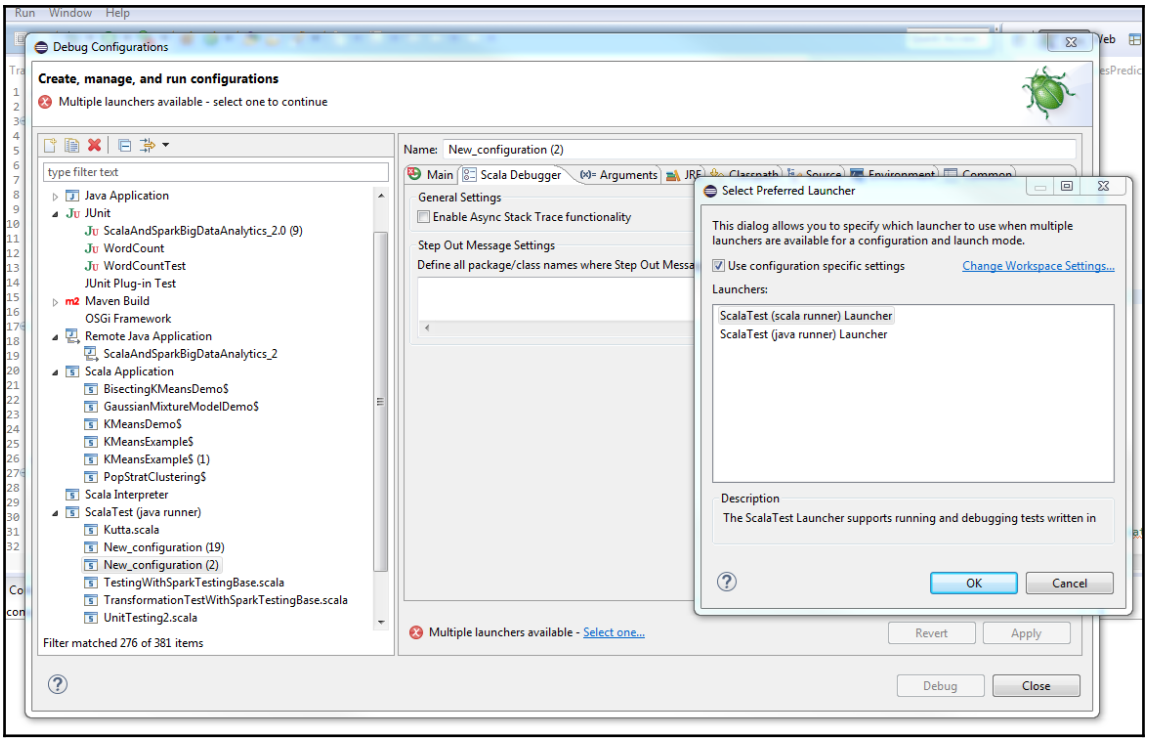
```
Console  Problems
<terminated> TransformationTestWithSparkTestingBase.scala [ScalaTest (java runner)] C:\Program Files\Java\jdk1.8.0_102\bin\javaw.exe (19 Feb 2017, 00:49:48)
Run starting. Expected test count is: 3
TransformationTestWithSparkTestingBase:
17/02/19 00:49:49 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
- works, obviously!

[Stage 0:>                                                          (0 + 0) / 8]
                                                                    - broken
- Testing RDD transformations using a shared Spark Context
Run completed in 2 seconds, 534 milliseconds.
Total number of tests run: 3
Suites: completed 1, aborted 0
Tests: succeeded 3, failed 0, canceled 0, ignored 0, pending 0
All tests passed.
```

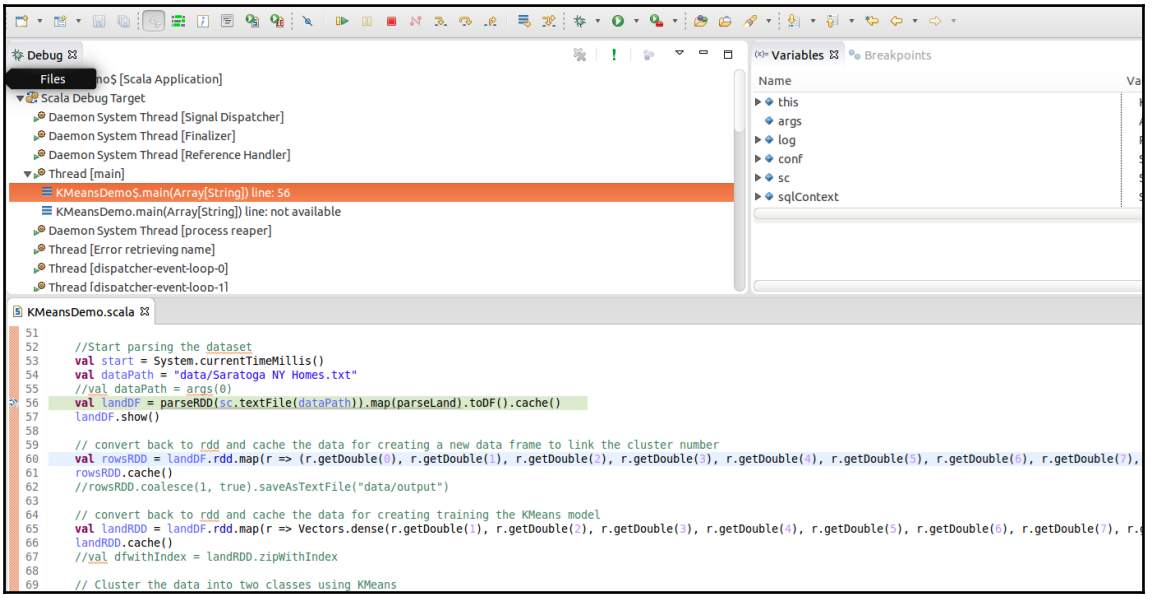
```
Console  Problems  ScalaTest
Tests:      3/3     Succeeded:  3     Failed:     0
Ignored:    0       Pending:    0     Canceled:   0
Suites:     1       Aborted:    0
TransformationTestWithSparkTestingBase (2.179 s)        Stack Trace
```

```
17/02/26 13:22:00 ERROR Shell: Failed to locate the winutils binary in the hadoop binary path
java.io.IOException: Could not locate executable null\bin\winutils.exe in the Hadoop binaries.
	at org.apache.hadoop.util.Shell.getQualifiedBinPath(Shell.java:278)
	at org.apache.hadoop.util.Shell.getWinUtilsPath(Shell.java:300)
	at org.apache.hadoop.util.Shell.<clinit>(Shell.java:293)
	at org.apache.hadoop.util.StringUtils.<clinit>(StringUtils.java:76)
	at org.apache.hadoop.mapred.FileInputFormat.setInputPaths(FileInputFormat.java:362)
	at org.apache.spark.SparkContext$$anonfun$hadoopFile$1$$anonfun$30.apply(SparkContext.scala:1014)
	at org.apache.spark.SparkContext$$anonfun$hadoopFile$1$$anonfun$30.apply(SparkContext.scala:1014)
	at org.apache.spark.rdd.HadoopRDD$$anonfun$getJobConf$6.apply(HadoopRDD.scala:179)
	at org.apache.spark.rdd.HadoopRDD$$anonfun$getJobConf$6.apply(HadoopRDD.scala:179)
	at scala.Option.foreach(Option.scala:257)
	at org.apache.spark.rdd.HadoopRDD.getJobConf(HadoopRDD.scala:179)
	at org.apache.spark.rdd.HadoopRDD.getPartitions(HadoopRDD.scala:198)
	at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:252)
	at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:250)
	at scala.Option.getOrElse(Option.scala:121)
	at org.apache.spark.rdd.RDD.partitions(RDD.scala:250)
	at org.apache.spark.rdd.MapPartitionsRDD.getPartitions(MapPartitionsRDD.scala:35)
	at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:252)
	at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:250)
	at scala.Option.getOrElse(Option.scala:121)
	at org.apache.spark.rdd.RDD.partitions(RDD.scala:250)
	at org.apache.spark.rdd.MapPartitionsRDD.getPartitions(MapPartitionsRDD.scala:35)
	at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:252)
	at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:250)
	at scala.Option.getOrElse(Option.scala:121)
	at org.apache.spark.rdd.RDD.partitions(RDD.scala:250)
	at org.apache.spark.rdd.MapPartitionsRDD.getPartitions(MapPartitionsRDD.scala:35)
	at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:252)
	at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:250)
	at scala.Option.getOrElse(Option.scala:121)
	at org.apache.spark.rdd.RDD.partitions(RDD.scala:250)
	at org.apache.spark.rdd.MapPartitionsRDD.getPartitions(MapPartitionsRDD.scala:35)
	at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:252)
	at org.apache.spark.rdd.RDD$$anonfun$partitions$2.apply(RDD.scala:250)
	at scala.Option.getOrElse(Option.scala:121)
```




```
# Set everything to be logged to the console
log4j.rootCategory=INFO, console
log4j.appender.console=org.apache.log4j.ConsoleAppender
log4j.appender.console.target=System.err
log4j.appender.console.layout=org.apache.log4j.PatternLayout
log4j.appender.console.layout.ConversionPattern=%d{yy/MM/dd HH:mm:ss} %p %c{1}: %m%n

# Set the default spark-shell log level to WARN. When running the spark-shell, the
# log level for this class is used to overwrite the root logger's log level, so that
# the user can have different defaults for the shell and regular Spark apps.
log4j.logger.org.apache.spark.repl.Main=WARN

# Settings to quiet third party logs that are too verbose
log4j.logger.org.spark_project.jetty=WARN
log4j.logger.org.spark_project.jetty.util.component.AbstractLifeCycle=ERROR
log4j.logger.org.apache.spark.repl.SparkIMain$exprTyper=INFO
log4j.logger.org.apache.spark.repl.SparkILoop$SparkILoopInterpreter=INFO
log4j.logger.org.apache.parquet=ERROR
log4j.logger.parquet=ERROR

# SPARK-9183: Settings to avoid annoying messages when looking up nonexistent UDFs in SparkSQL with Hive support
log4j.logger.org.apache.hadoop.hive.metastore.RetryingHMSHandler=FATAL
log4j.logger.org.apache.hadoop.hive.ql.exec.FunctionRegistry=ERROR
```


Run Window Help
Debug Configurations
Create, manage, and run configurations
Multiple launchers available - select one to continue
type filter text
Java Application
JUnit
ScalaAndSparkBigDataAnalytics_2.0 (9)
WordCount
WordCountTest
JUnit Plug-in Test
Maven Build
OSGi Framework
Remote Java Application
ScalaAndSparkBigDataAnalytics_2
Scala Application
BisectingKMeansDemo$
GaussianMixtureModelDemo$
KMeansDemo$
KMeansExample$
KMeansExample$ (1)
PopStratClustering$
Scala Interpreter
ScalaTest (java runner)
Kutta.scala
New_configuration (19)
New_configuration (2)
TestingWithSparkTestingBase.scala
TransformationTestWithSparkTestingBase.scala
UnitTesting2.scala
Filter matched 276 of 381 items
Name: New_configuration (2)
Main
Scala Debugger
Arguments
General Settings
Enable Async Stack Trace functionality
Step Out Message Settings
Define all package/class names where Step Out Messa
Multiple launchers available - Select one...
Revert
Apply
Debug
Close
Select Preferred Launcher
This dialog allows you to specify which launcher to use when multiple launchers are available for a configuration and launch mode.
Use configuration specific settings
Change Workspace Settings...
Launchers:
ScalaTest (scala runner) Launcher
ScalaTest (java runner) Launcher
Description
The ScalaTest Launcher supports running and debugging tests written in
OK
Cancel


Debug
Files
Scala Debug Target
Daemon System Thread [Signal Dispatcher]
Daemon System Thread [Finalizer]
Daemon System Thread [Reference Handler]
Thread [main]
KMeansDemo$.main(Array[String]) line: 56
KMeansDemo.main(Array[String]) line: not available
Daemon System Thread [process reaper]
Thread [Error retrieving name]
Thread [dispatcher-event-loop-0]
Variables
Breakpoints
Name
this
args
log
conf
sc
sqlContext
KMeansDemo.scala
//Start parsing the dataset
val start = System.currentTimeMillis()
val dataPath = "data/Saratoga NY Homes.txt"
//val dataPath = args(0)
val landDF = parseRDD(sc.textFile(dataPath)).map(parseLand).toDF().cache()
landDF.show()
// convert back to rdd and cache the data for creating a new data frame to link the cluster number
val rowsRDD = landDF.rdd.map(r => (r.getDouble(0), r.getDouble(1), r.getDouble(2), r.getDouble(3), r.getDouble(4), r.getDouble(5), r.getDouble(6), r.getDouble(7),
rowsRDD.cache()
//rowsRDD.coalesce(1, true).saveAsTextFile("data/output")
// convert back to rdd and cache the data for creating training the KMeans model
val landRDD = landDF.rdd.map(r => Vectors.dense(r.getDouble(1), r.getDouble(2), r.getDouble(3), r.getDouble(4), r.getDouble(5), r.getDouble(6), r.getDouble(7), r.
landRDD.cache()
//val dfwithIndex = landRDD.zipWithIndex
// Cluster the data into two classes using KMeans

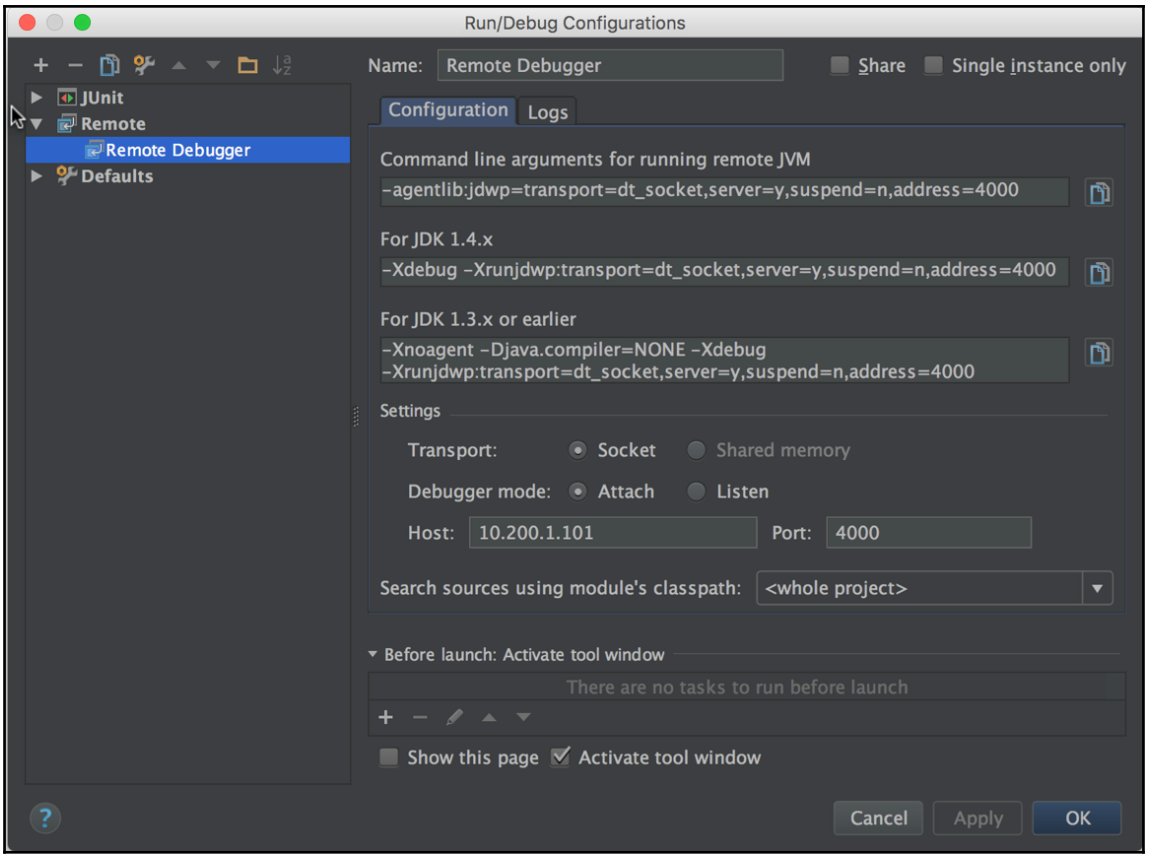
```
<terminated>KMeansDemo$ [Scala Application]
<terminated, exit value: 0>/usr/lib/jvm/java-8-oracle/bin/java (Feb 19, 2017, 12:34:48 PM)
```

KMeansDemo.scala

```
    val end = System.currentTimeMillis()
    println("Model building and prediction time: "+ {end - start} + "ms")

    // Compute and print the prediction accuracy for each house
    model.predict(landRDD).foreach(println)
    landDF.show()

    // Get the prediction from the model with the ID so we can link them back to other information
    val predictions = rowsRDD.map{r => (r._1, model.predict(Vectors.dense(r._2, r._3, r._4, r._5, r._6, r._7, r._
    val conMat = predictions.collect().toMap.values
    println(conMat)


    // convert the rdd to a dataframe
    val predCluster = predictions.toDF("Price", "CLUSTER")
    predCluster.show()


    // Join the prediction DataFrame with the old dataframe to know the individual cluster number for each house
    val newDF = landDF.join(predCluster, "Price")
```

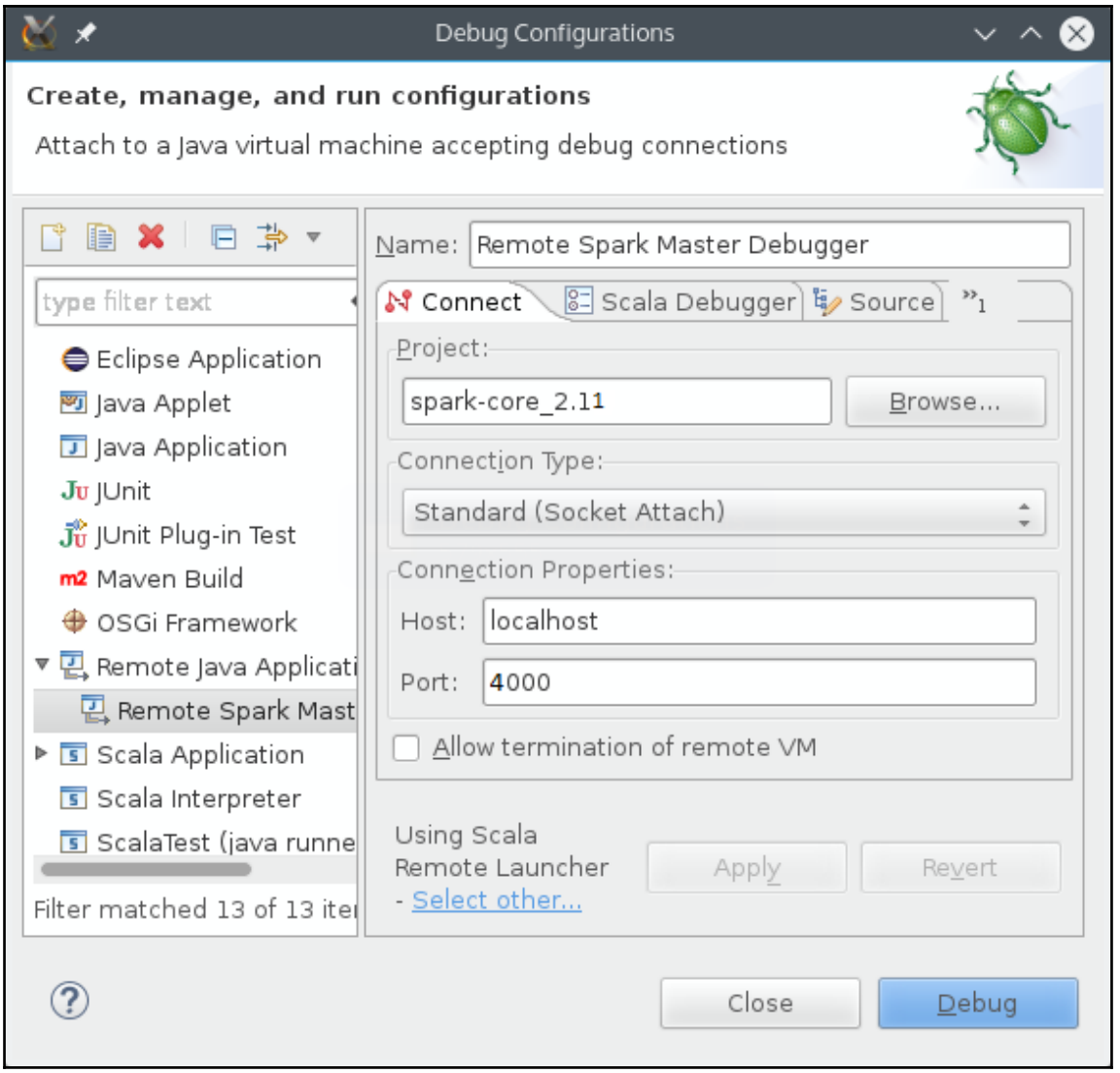
Console  Tasks  Scala Expression Evaluator

```
<terminated> KMeansDemo$ [Scala Application] /usr/lib/jvm/java-8-oracle/bin/java (Feb 19, 2017, 12:34:48 PM)
|253750.0|      3|
| 60000.0|      2|
| 87500.0|      2|
+--------+-------+
only showing top 20 rows

MapLike(2, 2, 2, 1, 0, 2, 0, 0, 0, 0, 2, 1, 3, 3, 2, 3, 0, 2, 2, 0, 3, 2, 2, 2, 1, 0, 0, 0, 3, 2, 3, 3, 3, 2, 0, 1, 3,
17/02/19 12:35:09 WARN root: Finshed
```


Run/Debug Configurations
JUnit
Remote
Remote Debugger
Defaults
Name: Remote Debugger
Share
Single instance only
Configuration
Logs
Command line arguments for running remote JVM
-agentlib:jdwp=transport=dt_socket,server=y,suspend=n,address=4000
For JDK 1.4.x
-Xdebug -Xrunjdwp:transport=dt_socket,server=y,suspend=n,address=4000
For JDK 1.3.x or earlier
-Xnoagent -Djava.compiler=NONE -Xdebug
-Xrunjdwp:transport=dt_socket,server=y,suspend=n,address=4000
Settings
Transport:
Socket
Shared memory
Debugger mode:
Attach
Listen
Host:
10.200.1.101
Port:
4000
Search sources using module's classpath:
<whole project>
Before launch: Activate tool window
There are no tasks to run before launch
Show this page
Activate tool window
Cancel
Apply
OK

SparkStreamingToHive.java - finserv-application-blueprint - [~/development/finserv-application-blueprint]

```
System.out.println("Waiting for messages...");

// Get the raw ticks, cast them to Tick objects, and insert them into Hive
// Note, this foreachRDD is an output operator with at-least once semantics,
// so this may persist duplicate records into Hive if the worker fails/restarts.
// Reference: http://spark.apache.org/docs/latest/streaming-programming-guide
.html#semantics-of-output-operations
directKafkaStream.foreachRDD(
        (VoidFunction) (rdd) -> {  rdd: JavaPairRDD@5340
                if (rdd.count() > 0) {
                        System.out.println("RDD batch size = " + rdd.count());  rdd: JavaPairRDD@5340
                        // Parse each row using the helper methods in the Tick data structure
                        JavaRDD<String> tick = rdd.map(
                                (x) -> { return new String(x._2); });

                        JavaRDD<Row> rowRDD = tick.map(
                                (record) -> {
                                        if (record.trim().length() > 0) {
                                                Tick t = new Tick(record.trim());
```


Debug Configurations
Create, manage, and run configurations
Attach to a Java virtual machine accepting debug connections
type filter text
Eclipse Application
Java Applet
Java Application
JUnit
JUnit Plug-in Test
Maven Build
OSGi Framework
Remote Java Applicati
Remote Spark Mast
Scala Application
Scala Interpreter
ScalaTest (java runne
Filter matched 13 of 13 iter
Name: Remote Spark Master Debugger
Connect
Scala Debugger
Source
Project:
spark-core_2.11
Browse...
Connection Type:
Standard (Socket Attach)
Connection Properties:
Host: localhost
Port: 4000
Allow termination of remote VM
Using Scala Remote Launcher - Select other...
Apply
Revert
Close
Debug

# Chapter 19: PySpark and SparkR

```
asif@ubuntu:~$ cd $SPARK_HOME
asif@ubuntu:~/Spark$ ./bin/pyspark
Python 2.7.6 (default, Oct 26 2016, 20:30:19)
[GCC 4.8.4] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
17/04/24 09:49:02 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... usin
17/04/24 09:49:02 WARN Utils: Your hostname, ubuntu resolves to a loopback address: 127.0.1.1; using 19
17/04/24 09:49:02 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
17/04/24 09:49:06 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.1.0
      /_/

Using Python version 2.7.6 (default, Oct 26 2016 20:30:19)
SparkSession available as 'spark'.
>>> 
```


```
asif@ubuntu:~$ cd $SPARK_HOME
asif@ubuntu:~/Spark$ ./bin/pyspark
Python 2.7.6 (default, Oct 26 2016, 20:30:19)
[GCC 4.8.4] on linux2
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
17/04/24 09:49:02 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... usin
17/04/24 09:49:02 WARN Utils: Your hostname, ubuntu resolves to a loopback address: 127.0.1.1; using 19
17/04/24 09:49:02 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
17/04/24 09:49:06 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.1.0
      /_/

Using Python version 2.7.6 (default, Oct 26 2016 20:30:19)
SparkSession available as 'spark'.
>>> 
```

MyPySpark.py | pca_example.py | pySparkDemo.py | Query.py | svm_with_sgd_example.py | PySparkDemo2.py

```
if __name__ == ...
from __future__ import print_function
import os
import sys

# Path for spark source folder
os.environ['SPARK_HOME'] = "C:/Users/rezkar/Downloads/spark-2.1.0-bin-hadoop2.7/"
os.environ['HADOOP_HOME'] = "C:/Users/rezkar/Downloads/spark-2.1.0-bin-hadoop2.7"

# Append pyspark  to Python Path
sys.path.append("C:/Users/rezkar/Downloads/spark-2.1.0-bin-hadoop2.7/python/")
sys.path.append("C:/Users/rezkar/Downloads/spark-2.1.0-bin-hadoop2.7/python/lib/py4j-0.10.4-src.zip")

try:
    from pyspark.ml.feature import PCA
    from pyspark.ml.linalg import Vectors
    from pyspark.sql import SparkSession
    print ("Successfully imported Spark Modules")

except ImportError as e:
    print ("Can not import Spark Modules", e)
    sys.exit(1)
```

Run pca_example

```
C:\Python27\python.exe C:/Users/rezkar/PycharmProjects/SparkWithPython/pca_example.py
Successfully imported Spark Modules
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
17/04/24 18:11:15 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using built
[Stage 6:>                                                          (0 + 8) / 8]17/04/24 18:11:25 WARN BLAS: Fa
17/04/24 18:11:25 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
17/04/24 18:11:25 WARN LAPACK: Failed to load implementation from: com.github.fommil.netlib.NativeSystemLAPACK
17/04/24 18:11:25 WARN LAPACK: Failed to load implementation from: com.github.fommil.netlib.NativeRefLAPACK
+-------------------------------------------------------------+
```

```
Run  pca_example
C:\Python27\python.exe C:/Users/rezkar/PycharmProjects/SparkWithPython/pca_example.py
Successfully imported Spark Modules
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
17/04/26 14:50:59 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform..
[Stage 6:>                                                          (0 + 8) / 8]17/04/26 14:51:10
17/04/26 14:51:10 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeR
17/04/26 14:51:10 WARN LAPACK: Failed to load implementation from: com.github.fommil.netlib.Nativ
17/04/26 14:51:10 WARN LAPACK: Failed to load implementation from: com.github.fommil.netlib.Nativ
+-----------------------------------------------------------+
|pcaFeatures                                                |
+-----------------------------------------------------------+
|[1.6485728230883807,-4.013282700516296,-5.524543751369388] |
|[-4.645104331781534,-1.1167972663619026,-5.524543751369387]|
|[-6.428880535676489,-5.337951427775355,-5.524543751369389] |
+-----------------------------------------------------------+

SUCCESS: The process with PID 7416 (child process of PID 4100) has been terminated.
SUCCESS: The process with PID 4100 (child process of PID 6488) has been terminated.
SUCCESS: The process with PID 6488 (child process of PID 1200) has been terminated.

Process finished with exit code 0
```

```
+-----+--------------------+
|label|            features|
+-----+--------------------+
|  8.0|(17,[0,1,2,3,4,5,...|
| 10.0|(17,[0,1,2,3,4,5,...|
|  9.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
| 10.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
|  5.0|(17,[0,1,2,3,4,5,...|
|  6.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
|  7.0|(17,[0,1,2,3,4,5,...|
|  6.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
|  9.0|(17,[0,1,2,3,4,5,...|
|  4.0|(17,[0,1,2,3,4,5,...|
|  7.0|(17,[0,1,2,3,4,5,...|
|  7.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
|  8.0|(17,[0,1,2,3,4,5,...|
+-----+--------------------+
only showing top 20 rows
```

```
root
 |-- year: string (nullable = true)
 |-- month: string (nullable = true)
 |-- day: string (nullable = true)
 |-- dep_time: string (nullable = true)
 |-- dep_delay: string (nullable = true)
 |-- arr_time: string (nullable = true)
 |-- arr_delay: string (nullable = true)
 |-- carrier: string (nullable = true)
 |-- tailnum: string (nullable = true)
 |-- flight: string (nullable = true)
 |-- origin: string (nullable = true)
 |-- dest: string (nullable = true)
 |-- air_time: string (nullable = true)
 |-- distance: string (nullable = true)
 |-- hour: string (nullable = true)
 |-- minute: string (nullable = true)
```

| year | month | day | dep_time | dep_delay | arr_time | arr_delay | carrier | tailnum | flight | origin | dest | air_time | distance | hour | minute |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2013 | 1 | 1 | 517 | 2 | 830 | 11 | UA | N14228 | 1545 | EWR | IAH | 227 | 1400 | 5 | 17 |
| 2013 | 1 | 1 | 533 | 4 | 850 | 20 | UA | N24211 | 1714 | LGA | IAH | 227 | 1416 | 5 | 33 |
| 2013 | 1 | 1 | 542 | 2 | 923 | 33 | AA | N619AA | 1141 | JFK | MIA | 160 | 1089 | 5 | 42 |
| 2013 | 1 | 1 | 544 | -1 | 1004 | -18 | B6 | N804JB | 725 | JFK | BQN | 183 | 1576 | 5 | 44 |
| 2013 | 1 | 1 | 554 | -6 | 812 | -25 | DL | N668DN | 461 | LGA | ATL | 116 | 762 | 5 | 54 |
| 2013 | 1 | 1 | 554 | -4 | 740 | 12 | UA | N39463 | 1696 | EWR | ORD | 150 | 719 | 5 | 54 |
| 2013 | 1 | 1 | 555 | -5 | 913 | 19 | B6 | N516JB | 507 | EWR | FLL | 158 | 1065 | 5 | 55 |
| 2013 | 1 | 1 | 557 | -3 | 709 | -14 | EV | N829AS | 5708 | LGA | IAD | 53 | 229 | 5 | 57 |
| 2013 | 1 | 1 | 557 | -3 | 838 | -8 | B6 | N593JB | 79 | JFK | MCO | 140 | 944 | 5 | 57 |
| 2013 | 1 | 1 | 558 | -2 | 753 | 8 | AA | N3ALAA | 301 | LGA | ORD | 138 | 733 | 5 | 58 |
| 2013 | 1 | 1 | 558 | -2 | 849 | -2 | B6 | N793JB | 49 | JFK | PBI | 149 | 1028 | 5 | 58 |
| 2013 | 1 | 1 | 558 | -2 | 853 | -3 | B6 | N657JB | 71 | JFK | TPA | 158 | 1005 | 5 | 58 |
| 2013 | 1 | 1 | 558 | -2 | 924 | 7 | UA | N29129 | 194 | JFK | LAX | 345 | 2475 | 5 | 58 |
| 2013 | 1 | 1 | 558 | -2 | 923 | -14 | UA | N53441 | 1124 | EWR | SFO | 361 | 2565 | 5 | 58 |
| 2013 | 1 | 1 | 559 | -1 | 941 | 31 | AA | N3DUAA | 707 | LGA | DFW | 257 | 1389 | 5 | 59 |
| 2013 | 1 | 1 | 559 | 0 | 702 | -4 | B6 | N708JB | 1806 | JFK | BOS | 44 | 187 | 5 | 59 |
| 2013 | 1 | 1 | 559 | -1 | 854 | -8 | UA | N76515 | 1187 | EWR | LAS | 337 | 2227 | 5 | 59 |
| 2013 | 1 | 1 | 600 | 0 | 851 | -7 | B6 | N595JB | 371 | LGA | FLL | 152 | 1076 | 6 | 0 |
| 2013 | 1 | 1 | 600 | 0 | 837 | 12 | MQ | N542MQ | 4650 | LGA | ATL | 134 | 762 | 6 | 0 |
| 2013 | 1 | 1 | 601 | 1 | 844 | -6 | B6 | N644JB | 343 | EWR | PBI | 147 | 1023 | 6 | 1 |

only showing top 20 rows

```
+------+------------+--------------+-----------+-----+----------+
|number|product_name| ransaction_id|    website|price|      date|
+------+------------+--------------+-----------+-----+----------+
|     0|       jeans|30160906182001|   ebay.com|  100|12-02-2016|
|     1|      camera|70151231120504| amazon.com|  450|09-08-2017|
|     2|      laptop|90151231120504|    ebay.ie| 1500|07--5-2016|
|     3|        book|80151231120506|  packt.com|   45|03-12-2016|
|     4|       drone| 8876531120508|alibaba.com|  120|01-05-2017|
+------+------------+--------------+-----------+-----+----------+
```

```
+------+------------+--------------+-----------+-----+----------+
|number|product_name| ransaction_id|    website|price|      date|
+------+------------+--------------+-----------+-----+----------+
|     0|       jeans|30160906182001|   ebay.com|  100|12-02-2016|
|     1|      camera|70151231120504| amazon.com|  450|09-08-2017|
|     2|      laptop|90151231120504|    ebay.ie| 1500|07--5-2016|
|     3|        book|80151231120506|  packt.com|   45|03-12-2016|
|     4|       drone| 8876531120508|alibaba.com|  120|01-05-2017|
+------+------------+--------------+-----------+-----+----------+

+------------+-----+
|product_name|price|
+------------+-----+
|      laptop| 1500|
+------------+-----+

+------------+-----+
|product_name|price|
+------------+-----+
|      camera|  450|
|        book|   45|
|      laptop| 1500|
|       drone|  120|
|       jeans|  100|
+------------+-----+
```

```
+-------+----------+-----+
|Student|    Course|Score|
+-------+----------+-----+
|  Jason|      Math|   87|
|  Jason|   Science|   32|
|  Jason| Geography|  126|
|  Jason|   History|   12|
|  Jason|        IT|   17|
|  Jason|Statistics|   37|
|   John|      Math|  143|
|   John|   Science|   54|
|   John| Geography|  146|
|   John|   History|   54|
|   John|        IT|   26|
|   John|Statistics|  171|
| Geroge|      Math|  102|
| Geroge|   Science|  146|
| Geroge| Geography|    5|
| Geroge|   History|  112|
| Geroge|        IT|  163|
| Geroge|Statistics|  175|
|  David|      Math|   27|
|  David|   Science|    4|
+-------+----------+-----+
only showing top 20 rows
```

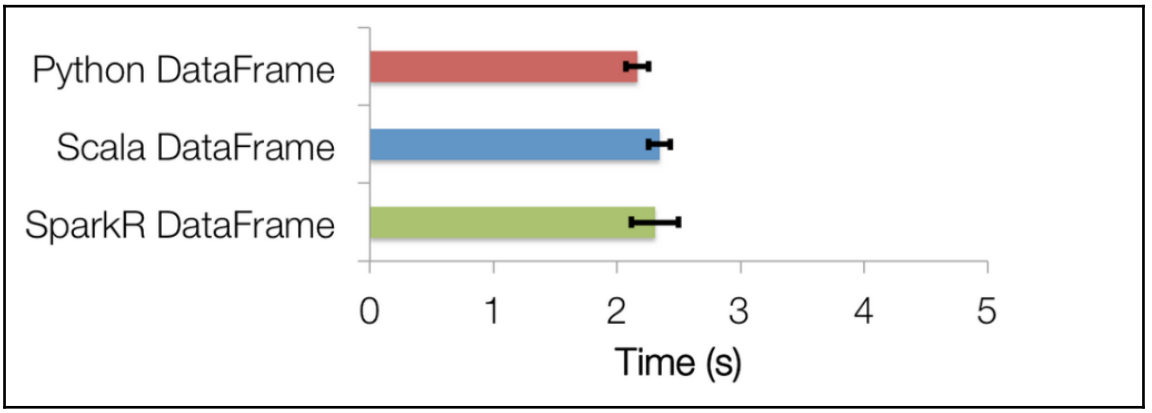
| Student | Course | Score | Grade |
|---|---|---|---|
| Jason | Math | 87 | B |
| Jason | Science | 32 | D |
| Jason | Geography | 126 | A |
| Jason | History | 12 | D |
| Jason | IT | 17 | D |
| Jason | Statistics | 37 | D |
| John | Math | 143 | A |
| John | Science | 54 | D |
| John | Geography | 146 | A |
| John | History | 54 | D |
| John | IT | 26 | D |
| John | Statistics | 171 | A |
| Geroge | Math | 102 | A |
| Geroge | Science | 146 | A |
| Geroge | Geography | 5 | D |
| Geroge | History | 112 | A |
| Geroge | IT | 163 | A |
| Geroge | Statistics | 175 | A |
| David | Math | 27 | D |
| David | Science | 4 | D |
| David | Geography | 1 | D |
| David | History | 13 | D |
| David | IT | 60 | C |
| David | Statistics | 19 | D |

```
+-------+-----+-----+
|Student|Score|Grade|
+-------+-----+-----+
|  Jason|   42|    D|
|  Jason|  153|    A|
|  Jason|  120|    A|
|  Jason|   99|    A|
|  Jason|  110|    A|
|  Jason|  150|    A|
|   John|   21|    D|
|   John|   45|    D|
|   John|    1|    D|
|   John|  138|    A|
|   John|  168|    A|
|   John|   90|    A|
| Geroge|   84|    B|
| Geroge|   84|    B|
| Geroge|  192|    A|
| Geroge|  192|    A|
| Geroge|   10|    D|
| Geroge|  132|    A|
|  David|   93|    A|
|  David|  127|    A|
+-------+-----+-----+
only showing top 20 rows
```

```
+---+---+----+---+-------+
| _1| _2|  _3| _4|    _42|
+---+---+----+---+-------+
|  0|tcp|http| SF|normal.|
|  0|tcp|http| SF|normal.|
|  0|tcp|http| SF|normal.|
|  0|tcp|http| SF|normal.|
|  0|tcp|http| SF|normal.|
+---+---+----+---+-------+
only showing top 5 rows
```

```
smurf. 2807886
neptune. 1072017
normal. 972781
satan. 15892
ipsweep. 12481
portsweep. 10413
nmap. 2316
back. 2203
warezclient. 1020
teardrop. 979
pod. 264
guess_passwd. 53
buffer_overflow. 30
land. 21
warezmaster. 20
imap. 12
rootkit. 10
loadmodule. 9
ftp_write. 8
multihop. 7
phf. 4
perl. 3
spy. 2
```

```
Final centers: [array([  4.10612163e+00,   6.36522840e-02,   4.85948958e-02,
        -2.21319176e-03,  -1.51849176e-02,   1.59666681e-02,
        -1.37464150e-02,   4.63552710e-03,  -2.80722691e-01,
         1.01178785e-01,   7.90818282e-02,   1.62820689e-01,
         1.08778945e-01,   3.21998554e-01,  -8.41384069e-03,
         6.05393588e-02,   0.00000000e+00,   3.30078588e-02,
        -2.46237569e-02,  -1.14832651e+00,  -1.19575475e+00,
        -3.71645499e-01,  -3.67973482e-01,   8.19357206e-01,
         8.14955084e-01,  -3.26320418e-01,   4.33755203e+00,
        -1.82859395e-01,   1.79392516e-01,  -1.71925941e+00,
        -1.75521881e+00,   6.82285609e+00,   2.23215018e-01,
        -1.16133090e-01,  -3.68177485e-01,  -3.66477378e-01,
         8.07658804e-01,   8.18438116e-01]), array([ -6.69802290e-02,  -1.36283222e-03,  -1.65369293e-03,
        -2.21319176e-03,  -1.51849176e-02,  -1.64391576e-03,
        -2.65266109e-02,  -4.38631465e-03,  -4.09296131e-01,
        -2.00370428e-03,  -8.21527723e-03,  -4.60861589e-03,
        -3.04988915e-03,  -9.62851412e-03,  -8.41384069e-03,
        -2.85810713e-02,   0.00000000e+00,  -5.21653093e-04,
        -2.88684412e-02,   6.87674624e-01,   7.54010775e-01,
        -4.65800760e-01,  -4.65512939e-01,  -2.48364764e-01,
        -2.48177638e-01,   5.39551929e-01,  -2.55781037e-01,
        -2.01125081e-01,   3.42806366e-01,   6.19909484e-01,
         5.98368428e-01,  -2.82739959e-01,   8.20664819e-01,
        -1.56479158e-01,  -4.66075407e-01,  -4.65194517e-01,
        -2.50690649e-01,  -2.49676723e-01]), array([ -6.69767578e-02,  -1.86749297e-03,  -1.65012194e-03,
        -2.21319176e-03,  -1.51849176e-02,  -1.64391576e-03,
        -2.64973873e-02,  -4.38631465e-03,  -4.09177709e-01,
        -1.99486560e-03,  -8.21527723e-03,  -4.60861589e-03,
```



| Property Name | Property group | `spark-submit` equivalent |
|---|---|---|
| `spark.master` | Application Properties | `--master` |
| `spark.yarn.keytab` | Application Properties | `--keytab` |
| `spark.yarn.principal` | Application Properties | `--principal` |
| `spark.driver.memory` | Application Properties | `--driver-memory` |
| `spark.driver.extraClassPath` | Runtime Environment | `--driver-class-path` |
| `spark.driver.extraJavaOptions` | Runtime Environment | `--driver-java-options` |
| `spark.driver.extraLibraryPath` | Runtime Environment | `--driver-library-path` |

Filter

| | year | month | day | dep_time | dep_delay | arr_time | arr_delay | carrier | tailnum | flight | origin | dest | air_time | distance | hour | minute |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2013 | 1 | 1 | 517 | 2 | 830 | 11 | UA | N14228 | 1545 | EWR | IAH | 227 | 1400 | 5 | 17 |
| 2 | 2013 | 1 | 1 | 533 | 4 | 850 | 20 | UA | N24211 | 1714 | LGA | IAH | 227 | 1416 | 5 | 33 |
| 3 | 2013 | 1 | 1 | 542 | 2 | 923 | 33 | AA | N619AA | 1141 | JFK | MIA | 160 | 1089 | 5 | 42 |
| 4 | 2013 | 1 | 1 | 544 | -1 | 1004 | -18 | B6 | N804JB | 725 | JFK | BQN | 183 | 1576 | 5 | 44 |
| 5 | 2013 | 1 | 1 | 554 | -6 | 812 | -25 | DL | N668DN | 461 | LGA | ATL | 116 | 762 | 5 | 54 |
| 6 | 2013 | 1 | 1 | 554 | -4 | 740 | 12 | UA | N39463 | 1696 | EWR | ORD | 150 | 719 | 5 | 54 |
| 7 | 2013 | 1 | 1 | 555 | -5 | 913 | 19 | B6 | N516JB | 507 | EWR | FLL | 158 | 1065 | 5 | 55 |
| 8 | 2013 | 1 | 1 | 557 | -3 | 709 | -14 | EV | N829AS | 5708 | LGA | IAD | 53 | 229 | 5 | 57 |
| 9 | 2013 | 1 | 1 | 557 | -3 | 838 | -8 | B6 | N593JB | 79 | JFK | MCO | 140 | 944 | 5 | 57 |
| 10 | 2013 | 1 | 1 | 558 | -2 | 753 | 8 | AA | N3ALAA | 301 | LGA | ORD | 138 | 733 | 5 | 58 |
| 11 | 2013 | 1 | 1 | 558 | -2 | 849 | -2 | B6 | N793JB | 49 | JFK | PBI | 149 | 1028 | 5 | 58 |
| 12 | 2013 | 1 | 1 | 558 | -2 | 853 | -3 | B6 | N657JB | 71 | JFK | TPA | 158 | 1005 | 5 | 58 |
| 13 | 2013 | 1 | 1 | 558 | -2 | 924 | 7 | UA | N29129 | 194 | JFK | LAX | 345 | 2475 | 5 | 58 |
| 14 | 2013 | 1 | 1 | 558 | -2 | 923 | -14 | UA | N53441 | 1124 | EWR | SFO | 361 | 2565 | 5 | 58 |
| 15 | 2013 | 1 | 1 | 559 | -1 | 941 | 31 | AA | N3DUAA | 707 | LGA | DFW | 257 | 1389 | 5 | 59 |
| 16 | 2013 | 1 | 1 | 559 | 0 | 702 | -4 | B6 | N708JB | 1806 | JFK | BOS | 44 | 187 | 5 | 59 |
| 17 | 2013 | 1 | 1 | 559 | -1 | 854 | -8 | UA | N76515 | 1187 | EWR | LAS | 337 | 2227 | 5 | 59 |
| 18 | 2013 | 1 | 1 | 600 | 0 | 851 | -7 | B6 | N595JB | 371 | LGA | FLL | 152 | 1076 | 6 | 0 |
| 19 | 2013 | 1 | 1 | 600 | 0 | 837 | 12 | MQ | N542MQ | 4650 | LGA | ATL | 134 | 762 | 6 | 0 |
| 20 | 2013 | 1 | 1 | 601 | 1 | 844 | -6 | B6 | N644JB | 343 | EWR | PBI | 147 | 1023 | 6 | 1 |

```
root
 |-- year: integer (nullable = true)
 |-- month: integer (nullable = true)
 |-- day: integer (nullable = true)
 |-- dep_time: string (nullable = true)
 |-- dep_delay: string (nullable = true)
 |-- arr_time: string (nullable = true)
 |-- arr_delay: string (nullable = true)
 |-- carrier: string (nullable = true)
 |-- tailnum: string (nullable = true)
 |-- flight: integer (nullable = true)
 |-- origin: string (nullable = true)
 |-- dest: string (nullable = true)
 |-- air_time: string (nullable = true)
 |-- distance: integer (nullable = true)
 |-- hour: string (nullable = true)
 |-- minute: string (nullable = true)
```

| year | month | day | dep_time | dep_delay | arr_time | arr_delay | carrier | tailnum | flight | origin | dest | air_time | distance | hour | minute |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2013 | 1 | 1 | 517 | 2 | 830 | 11 | UA | N14228 | 1545 | EWR | IAH | 227 | 1400 | 5 | 17 |
| 2013 | 1 | 1 | 533 | 4 | 850 | 20 | UA | N24211 | 1714 | LGA | IAH | 227 | 1416 | 5 | 33 |
| 2013 | 1 | 1 | 542 | 2 | 923 | 33 | AA | N619AA | 1141 | JFK | MIA | 160 | 1089 | 5 | 42 |
| 2013 | 1 | 1 | 544 | -1 | 1004 | -18 | B6 | N804JB | 725 | JFK | BQN | 183 | 1576 | 5 | 44 |
| 2013 | 1 | 1 | 554 | -6 | 812 | -25 | DL | N668DN | 461 | LGA | ATL | 116 | 762 | 5 | 54 |
| 2013 | 1 | 1 | 554 | -4 | 740 | 12 | UA | N39463 | 1696 | EWR | ORD | 150 | 719 | 5 | 54 |
| 2013 | 1 | 1 | 555 | -5 | 913 | 19 | B6 | N516JB | 507 | EWR | FLL | 158 | 1065 | 5 | 55 |
| 2013 | 1 | 1 | 557 | -3 | 709 | -14 | EV | N829AS | 5708 | LGA | IAD | 53 | 229 | 5 | 57 |
| 2013 | 1 | 1 | 557 | -3 | 838 | -8 | B6 | N593JB | 79 | JFK | MCO | 140 | 944 | 5 | 57 |
| 2013 | 1 | 1 | 558 | -2 | 753 | 8 | AA | N3ALAA | 301 | LGA | ORD | 138 | 733 | 5 | 58 |

only showing top 10 rows

```
root
 |-- year: integer (nullable = true)
 |-- month: integer (nullable = true)
 |-- day: integer (nullable = true)
 |-- dep_time: string (nullable = true)
 |-- dep_delay: string (nullable = true)
 |-- arr_time: string (nullable = true)
 |-- arr_delay: string (nullable = true)
 |-- carrier: string (nullable = true)
 |-- tailnum: string (nullable = true)
 |-- flight: integer (nullable = true)
 |-- origin: string (nullable = true)
 |-- dest: string (nullable = true)
 |-- air_time: string (nullable = true)
 |-- distance: integer (nullable = true)
 |-- hour: string (nullable = true)
 |-- minute: string (nullable = true)
```

| year | month | day | dep_time | dep_delay | arr_time | arr_delay | carrier | tailnum | flight | origin | dest | air_time | distance | hour | minute |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2013 | 1 | 1 | 517 | 2 | 830 | 11 | UA | N14228 | 1545 | EWR | IAH | 227 | 1400 | 5 | 17 |
| 2013 | 1 | 1 | 533 | 4 | 850 | 20 | UA | N24211 | 1714 | LGA | IAH | 227 | 1416 | 5 | 33 |
| 2013 | 1 | 1 | 542 | 2 | 923 | 33 | AA | N619AA | 1141 | JFK | MIA | 160 | 1089 | 5 | 42 |
| 2013 | 1 | 1 | 544 | -1 | 1004 | -18 | B6 | N804JB | 725 | JFK | BQN | 183 | 1576 | 5 | 44 |
| 2013 | 1 | 1 | 554 | -6 | 812 | -25 | DL | N668DN | 461 | LGA | ATL | 116 | 762 | 5 | 54 |
| 2013 | 1 | 1 | 554 | -4 | 740 | 12 | UA | N39463 | 1696 | EWR | ORD | 150 | 719 | 5 | 54 |
| 2013 | 1 | 1 | 555 | -5 | 913 | 19 | B6 | N516JB | 507 | EWR | FLL | 158 | 1065 | 5 | 55 |
| 2013 | 1 | 1 | 557 | -3 | 709 | -14 | EV | N829AS | 5708 | LGA | IAD | 53 | 229 | 5 | 57 |
| 2013 | 1 | 1 | 557 | -3 | 838 | -8 | B6 | N593JB | 79 | JFK | MCO | 140 | 944 | 5 | 57 |
| 2013 | 1 | 1 | 558 | -2 | 753 | 8 | AA | N3ALAA | 301 | LGA | ORD | 138 | 733 | 5 | 58 |

only showing top 10 rows

| year | month | day | dep_time | dep_delay | arr_time | arr_delay | carrier | tailnum | flight | origin | dest | air_time | distance | hour | minute |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2013 | 1 | 1 | 542 | 2 | 923 | 33 | AA | N619AA | 1141 | JFK | MIA | 160 | 1089 | 5 | 42 |
| 2013 | 1 | 1 | 606 | -4 | 858 | -12 | AA | N633AA | 1895 | EWR | MIA | 152 | 1085 | 6 | 6 |
| 2013 | 1 | 1 | 607 | 0 | 858 | -17 | UA | N53442 | 1077 | EWR | MIA | 157 | 1085 | 6 | 7 |
| 2013 | 1 | 1 | 623 | 13 | 920 | 5 | AA | N3EMAA | 1837 | LGA | MIA | 153 | 1096 | 6 | 23 |
| 2013 | 1 | 1 | 655 | -5 | 1002 | -18 | DL | N997DL | 2003 | LGA | MIA | 161 | 1096 | 6 | 55 |
| 2013 | 1 | 1 | 659 | -1 | 1008 | -7 | AA | N3EKAA | 2279 | LGA | MIA | 159 | 1096 | 6 | 59 |
| 2013 | 1 | 1 | 753 | -2 | 1056 | -14 | AA | N3HMAA | 2267 | LGA | MIA | 157 | 1096 | 7 | 53 |
| 2013 | 1 | 1 | 759 | -1 | 1057 | -30 | DL | N955DL | 1843 | JFK | MIA | 158 | 1089 | 7 | 59 |
| 2013 | 1 | 1 | 826 | 71 | 1136 | 51 | AA | N3GVAA | 443 | JFK | MIA | 160 | 1089 | 8 | 26 |
| 2013 | 1 | 1 | 856 | -4 | 1222 | -10 | DL | N970DL | 2143 | LGA | MIA | 158 | 1096 | 8 | 56 |

only showing top 10 rows

```
+------+---------+------+----+
|flight|dep_delay|origin|dest|
+------+---------+------+----+
|  1545|        2|   EWR| IAH|
|  1714|        4|   LGA| IAH|
|   496|       -4|   LGA| IAH|
|   473|       -4|   LGA| IAH|
|  1479|        0|   EWR| IAH|
|  1220|        0|   EWR| IAH|
|  1004|        2|   LGA| IAH|
|   455|       -1|   EWR| IAH|
|  1086|      134|   LGA| IAH|
|  1461|        5|   EWR| IAH|
+------+---------+------+----+
only showing top 10 rows
```

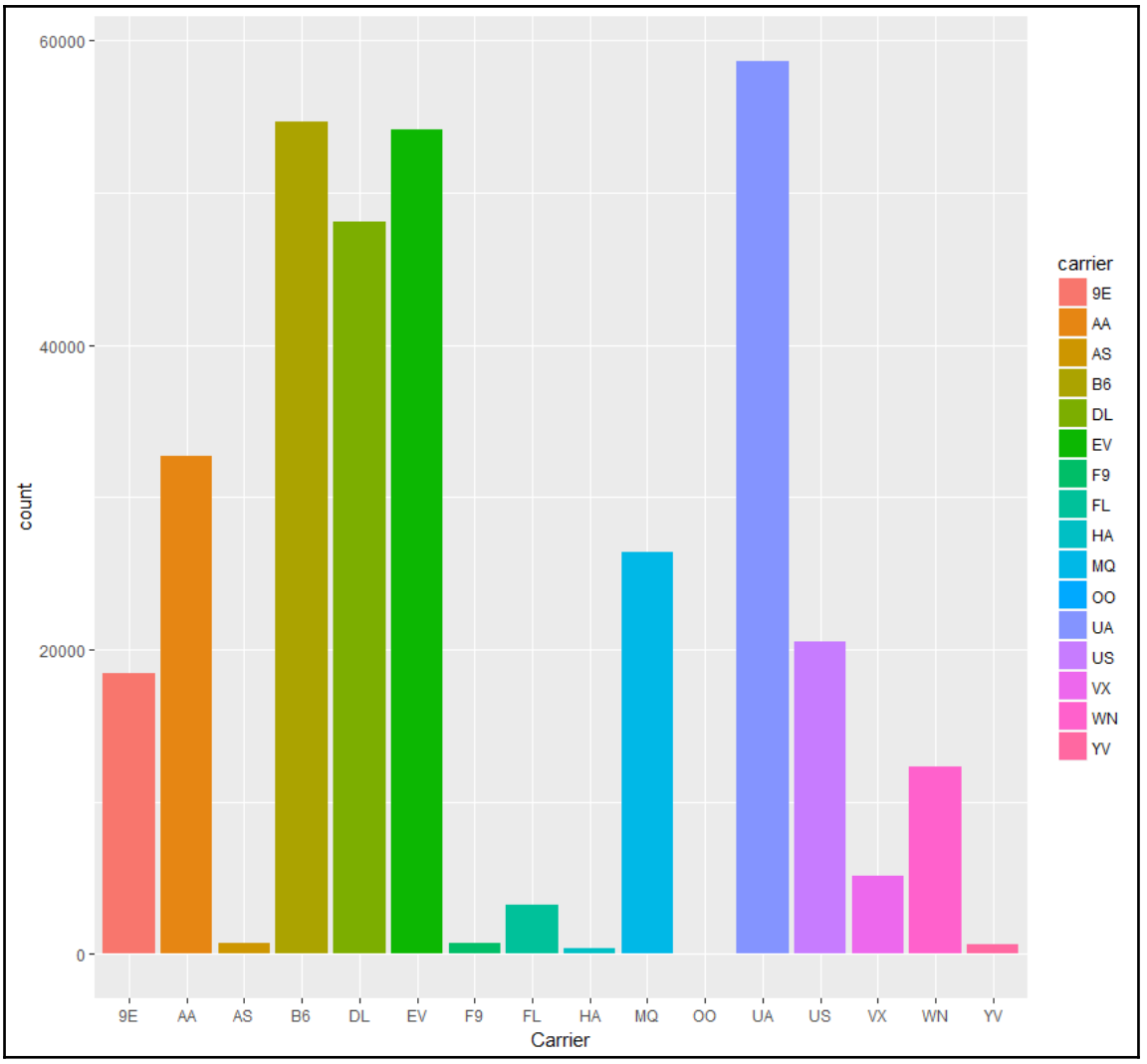
```
  day avg(dep_delay) avg(arr_delay)
1  31       9.506521       3.359225
2  28      15.743213       8.183567
3  26       9.748002       3.656098
4  27      12.083969       3.331213
5  12      15.177765      11.138973
6  22      18.712073      17.404916
```

```
  dest NUM_FLIGHTS AVG_DELAY MAX_DELAY MIN_DELAY
1  PSE         365  7.871508        NA        -1
2  MSY        3799  6.490175        NA        -1
3  BUR         371  8.175676        NA        -1
4  SNA         825 -7.868227        NA        -1
5  GRR         765 18.189560        NA        -1
6  GSO        1606 14.112601        NA        -1
```

```
+----+------+-------+
|dest|origin|carrier|
+----+------+-------+
| IAH|   EWR|     UA|
| IAH|   LGA|     UA|
| MIA|   JFK|     AA|
| BQN|   JFK|     B6|
| ATL|   LGA|     DL|
| ORD|   EWR|     UA|
| FLL|   EWR|     B6|
| IAD|   LGA|     EV|
| MCO|   JFK|     B6|
| ORD|   LGA|     AA|
+----+------+-------+
only showing top 10 rows
```

```
+----+------+---------+
|dest|origin|arr_delay|
+----+------+---------+
| CLT|   LGA|      137|
| BWI|   JFK|      851|
| BOS|   EWR|      123|
| IAH|   LGA|      145|
| RIC|   EWR|      127|
| MCO|   EWR|      125|
| MCI|   EWR|      136|
| IAD|   JFK|      123|
| DAY|   EWR|      123|
| BNA|   LGA|      138|
+----+------+---------+
only showing top 10 rows
```

```
+------+----+---------+
|origin|dest|arr_delay|
+------+----+---------+
|   JFK| IAH|      783|
|   LGA| IAH|      435|
|   LGA| IAH|      390|
|   EWR| IAH|      374|
|   EWR| IAH|      373|
|   LGA| IAH|      370|
|   LGA| IAH|      363|
|   EWR| IAH|      338|
|   LGA| IAH|      324|
|   LGA| IAH|      321|
|   LGA| IAH|      312|
|   LGA| IAH|      309|
|   EWR| IAH|      302|
|   LGA| IAH|      301|
|   EWR| IAH|      297|
|   LGA| IAH|      294|
|   EWR| IAH|      292|
|   EWR| IAH|      288|
|   EWR| IAH|      283|
|   LGA| IAH|      278|
+------+----+---------+
```


60000
40000
20000
0
count
9E
AA
AS
B6
DL
EV
F9
FL
HA
MQ
OO
UA
US
VX
WN
YV
Carrier
carrier
9E
AA
AS
B6
DL
EV
F9
FL
HA
MQ
OO
UA
US
VX
WN
YV

```
+-------+-----+
|carrier|  cnt|
+-------+-----+
|     UA|58665|
|     B6|54635|
|     EV|54173|
|     DL|48110|
|     AA|32729|
|     MQ|26397|
|     US|20536|
|     9E|18460|
|     WN|12275|
|     VX| 5162|
|     FL| 3260|
|     AS|  714|
|     F9|  685|
|     YV|  601|
|     HA|  342|
|     OO|   32|
+-------+-----+
```