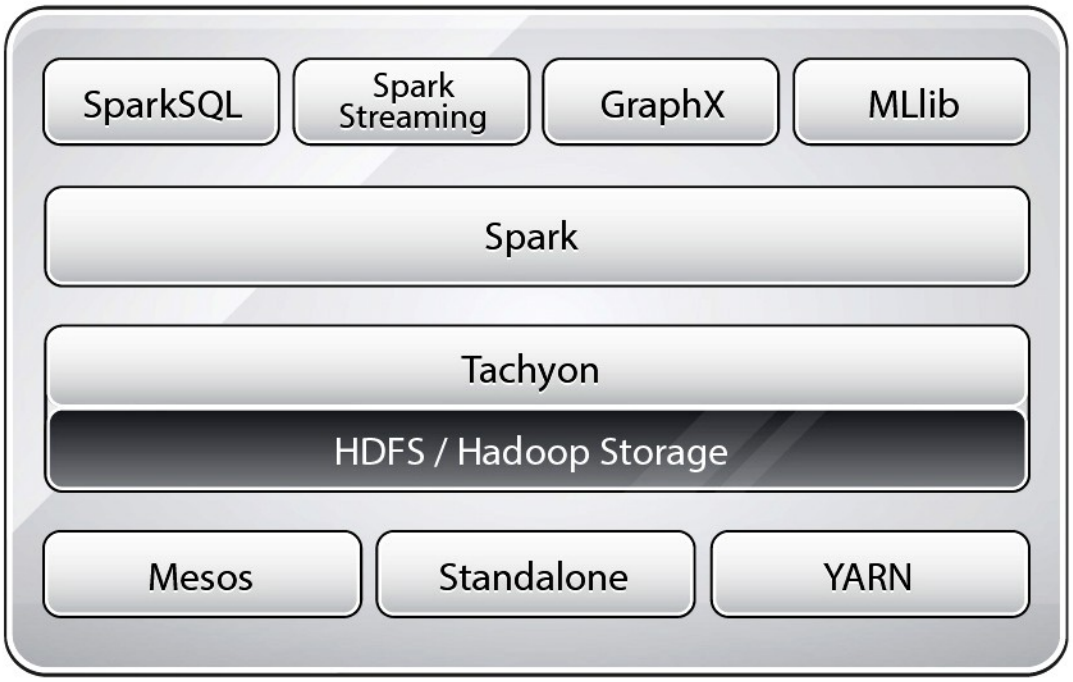
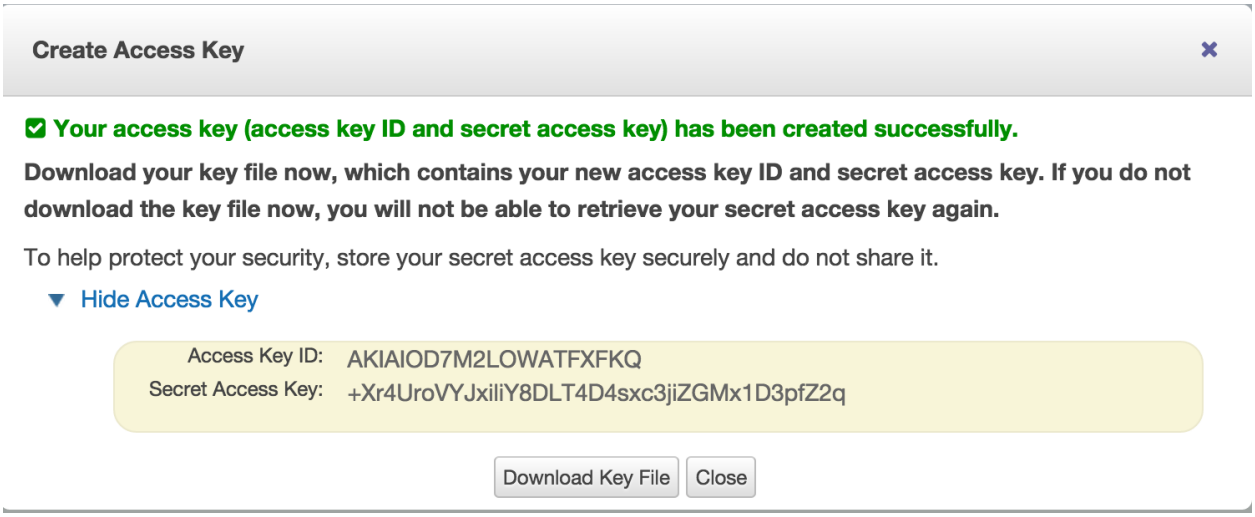
Chapter 1: Getting Started with Apache Spark





```
Connection to ec2-54-211-128-216.compute-1.amazonaws.com closed.
Spark standalone cluster started at http://ec2-54-211-128-216.compute-1.amazonaws.com:8080
Ganglia started at http://ec2-54-211-128-216.compute-1.amazonaws.com:5080/ganglia
Done!
```

Spark Master at spark://ec2-54-211-128-216.compute-1.amazonaws.com:7077

URL: spark://ec2-54-211-128-216.compute-1.amazonaws.com:7077
Workers: 3
Cores: 6 Total, 0 Used
Memory: 18.8 GB Total, 0.0 B Used
Applications: 0 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers

Id	Address	State	Cores	Memory
worker-20141130022618-ip-10-170-6-91.ec2.internal-59489	ip-10-170-6-91.ec2.internal:59489	ALIVE	2 (0 Used)	6.3 GB (0.0 B Used)
worker-20141130022618-ip-10-182-148-55.ec2.internal-51719	ip-10-182-148-55.ec2.internal:51719	ALIVE	2 (0 Used)	6.3 GB (0.0 B Used)
worker-20141130022618-ip-10-182-183-44.ec2.internal-46837	ip-10-182-183-44.ec2.internal:46837	ALIVE	2 (0 Used)	6.3 GB (0.0 B Used)

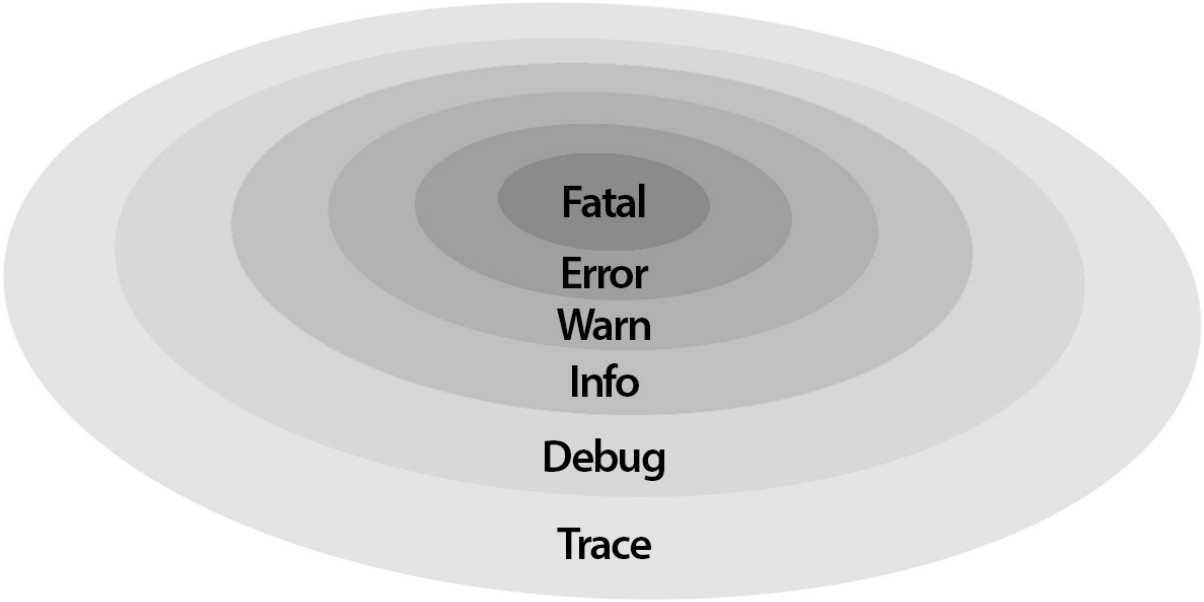
Running Applications

ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration

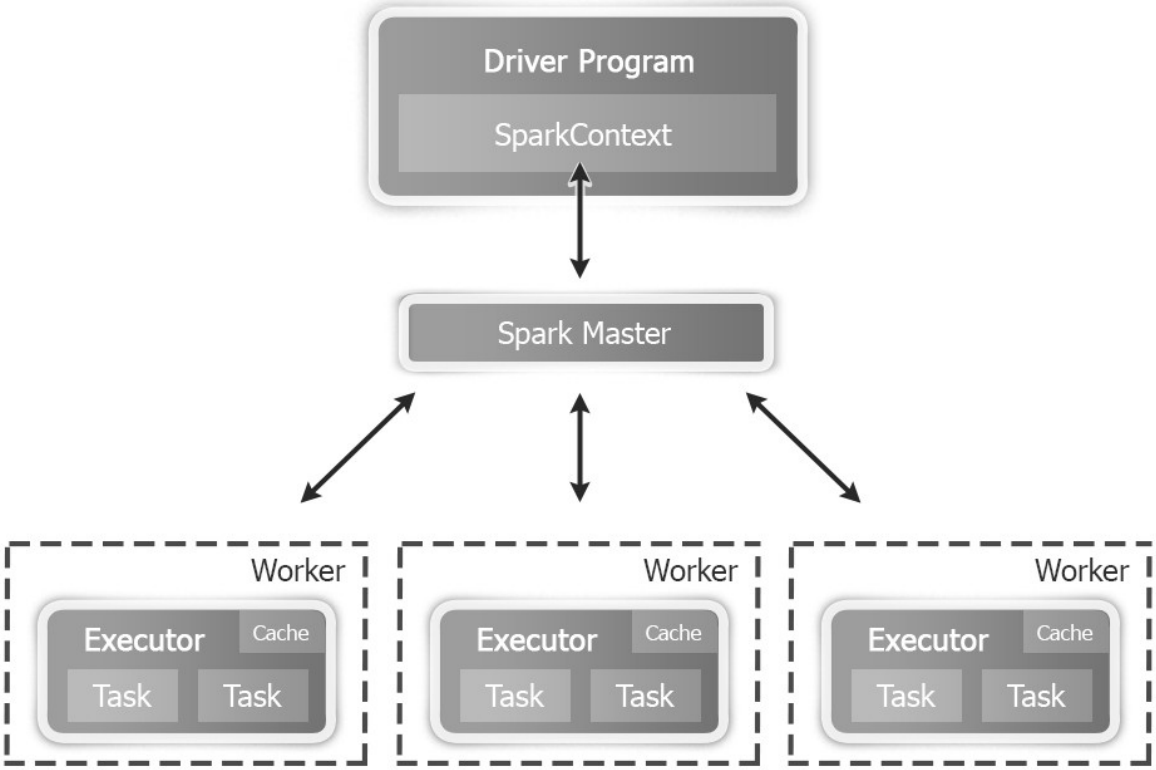
Completed Applications

ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration

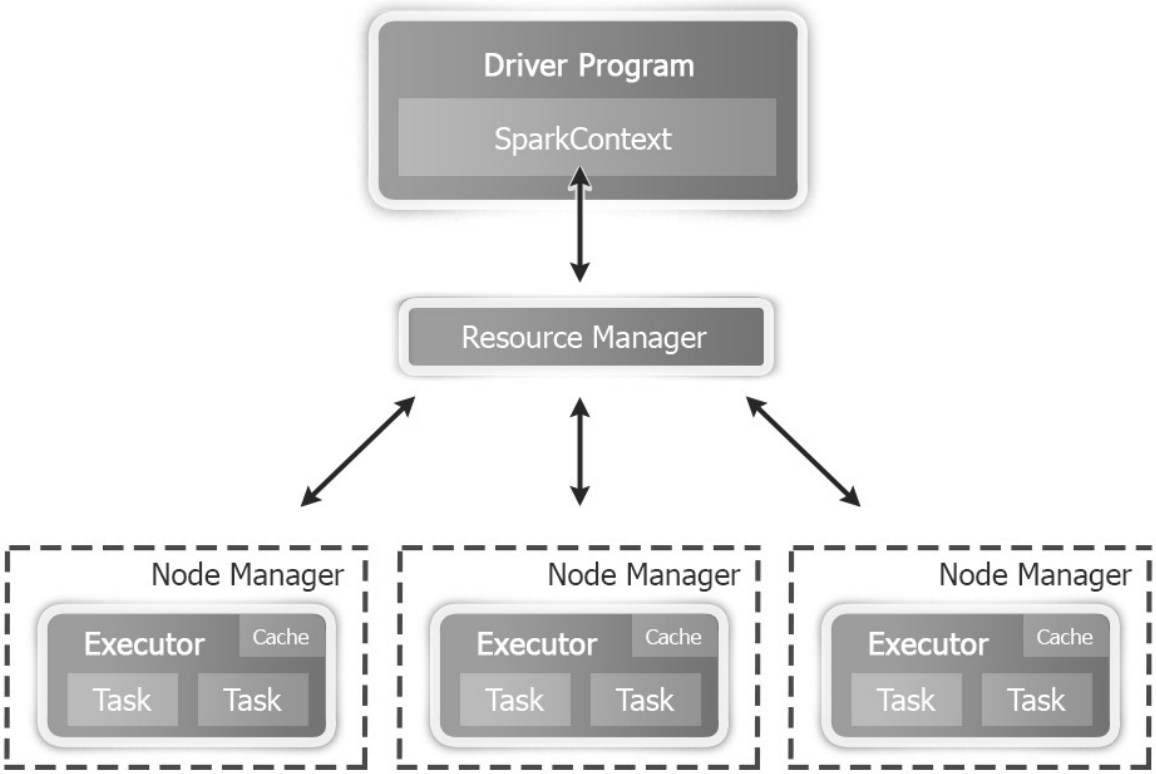
```
hduser@infoobjects:~$ spark-ec2 -k spark-kp1 -i /home/hduser/kp/spark-kp1.pem  login spark-cluster
Searching for existing cluster spark-cluster...
Found 1 master(s), 3 slaves
Logging into master ec2-54-211-128-216.compute-1.amazonaws.com...
Last login: Sun Nov 30 02:22:36 2014 from c-73-162-232-122.hsd1.ca.comcast.net

       __|  __|_  )
       _|  (     /   Amazon Linux AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-ami/2013.03-release-notes/
There are 75 security update(s) out of 282 total update(s) available
Run "sudo yum update" to apply all updates.
Amazon Linux version 2014.09 is available.
root@ip-10-182-135-159 ~]$ ls
ephemeral-hdfs  hadoop-native  mapreduce  persistent-hdfs  scala  shark  spark  spark-ec2  tachyon
```

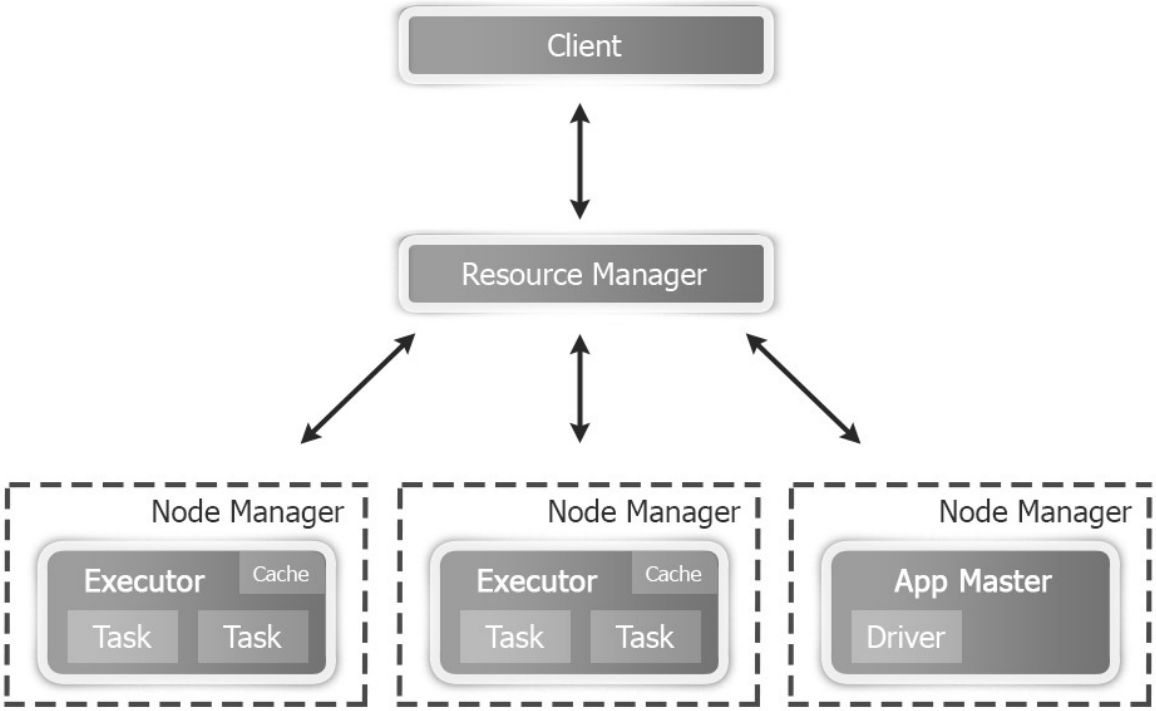


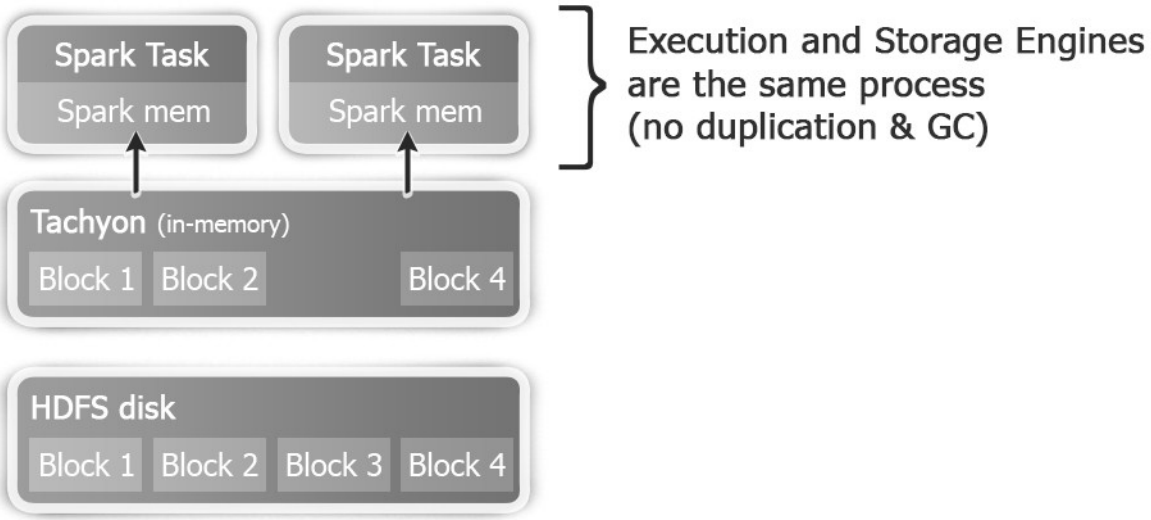
```
root@ip-10-168-32-181 ~]$ spark-ec2/copy-dir spark/conf/
RSYNC'ing /root/spark/conf to slaves...
ec2-174-129-51-11.compute-1.amazonaws.com
ec2-107-20-52-62.compute-1.amazonaws.com
ec2-54-224-17-251.compute-1.amazonaws.com
```



```
#!/usr/bin/env bash

# This file contains environment variables required to run Spark. Copy it as
# spark-env.sh and edit that to configure Spark for your site.
#
# The following variables can be set in this file:
# - SPARK_LOCAL_IP, to set the IP address Spark binds to on this node
# - MESOS_NATIVE_LIBRARY, to point to your libmesos.so if you use Mesos
# - SPARK_JAVA_OPTS, to set node-specific JVM options for Spark. Note that
#   we recommend setting app-wide options in the application's driver program.
#     Examples of node-specific options : -Dspark.local.dir, GC options
#     Examples of app-wide options : -Dspark.serializer
#
# If using the standalone deploy mode, you can also set variables for it here:
# - SPARK_MASTER_IP, to bind the master to a different IP address or hostname
# - SPARK_MASTER_PORT / SPARK_MASTER_WEBUI_PORT, to use non-default ports
# - SPARK_WORKER_CORES, to set the number of cores to use on this machine
# - SPARK_WORKER_MEMORY, to set how much memory to use (e.g. 1000m, 2g)
# - SPARK_WORKER_PORT / SPARK_WORKER_WEBUI_PORT
# - SPARK_WORKER_INSTANCES, to set the number of worker processes per node
# - SPARK_WORKER_DIR, to set the working directory of worker processes
export HADOOP_CONF_DIR=/opt/infoobjects/hadoop/etc/hadoop
export YARN_CONF_DIR=/opt/infoobjects/hadoop/etc/hadoop
export SPARK_LOG_DIR=/var/log/spark
export SPARK_WORKER_DIR=/var/spark/worker
```




Client
Resource Manager
Node Manager
Executor
Cache
Task
Task
Node Manager
Executor
Cache
Task
Task
Node Manager
App Master
Driver


Spark Task
Spark mem
Spark Task
Spark mem
Execution and Storage Engines
are the same process
(no duplication & GC)
Tachyon (in-memory)
Block 1
Block 2
Block 4
HDFS disk
Block 1
Block 2
Block 3
Block 4

TACHYON | Overview | System Configuration | Browse File System | In Memory Files

Tachyon Summary

Master Address:	localhost/127.0.0.1:19998
Started:	12-03-2014 07:34:14:914
Uptime:	0 day(s), 0 hour(s), 1 minute(s), and 32 second(s)
Version:	0.5.0
Running Workers:	1

Cluster Usage Summary

Memory Capacity:	1024.00 MB
Memory Free / Used:	1024.00 MB / 0.00 B
UnderFS Capacity:	29.33 GB
UnderFS Free / Used:	17.94 GB / 11.39 GB

Detailed Nodes Summary

Node Name	Last Heartbeat	State	Memory Usage
localhost	0	In Service	100%Free

Tachyon is an open source project developed at the UC Berkeley AMPLab.

```
hduser@localhost:~$ tachyon runTest Basic CACHE_THROUGH
/BasicFile_CACHE_THROUGH has been removed
2014-12-03 07:11:06,149 INFO   (TachyonFS.java:connect) - Trying to connect master @ localhost/127.0.0.1:19998
2014-12-03 07:11:06,204 INFO   (MasterClient.java:getUserId) - User registered at the master localhost/127.0.0.1:19998 got UserId 2
2014-12-03 07:11:06,206 INFO   (TachyonFS.java:connect) - Trying to get local worker host : localhost
2014-12-03 07:11:06,219 INFO   (TachyonFS.java:connect) - Connecting local worker @ localhost/127.0.0.1:29998
2014-12-03 07:11:06,270 INFO   (CommonUtils.java:printTimeTakenMs) - createFile with fileId 2 took 122 ms.
2014-12-03 07:11:06,333 INFO   (TachyonFS.java:createAndGetUserTempFolder) - Folder /var/tachyon/ramdisk/tachyonworker/users/2 was created!
2014-12-03 07:11:06,342 INFO   (BlockOutStream.java:<init>) - /var/tachyon/ramdisk/tachyonworker/users/2/2147483648 was created!
```

AWS | Services | Edit

EC2 Dashboard
Events
Tags
Reports
Limits

INSTANCES
Instances
Spot Requests
Reserved Instances

IMAGES
AMIs
Bundle Tasks

ELASTIC BLOCK STORE
Volumes
Snapshots

NETWORK & SECURITY
Security Groups
Elastic IPs
Placement Groups
Load Balancers
Key Pairs
Network Interfaces

Create Key Pair | Import Key Pair | Delete

Filter by attributes or search by keyword

You do not have any Key Pairs in this region.

Click the "Create Key Pair" button to create your first Key Pair.

Create Key Pair

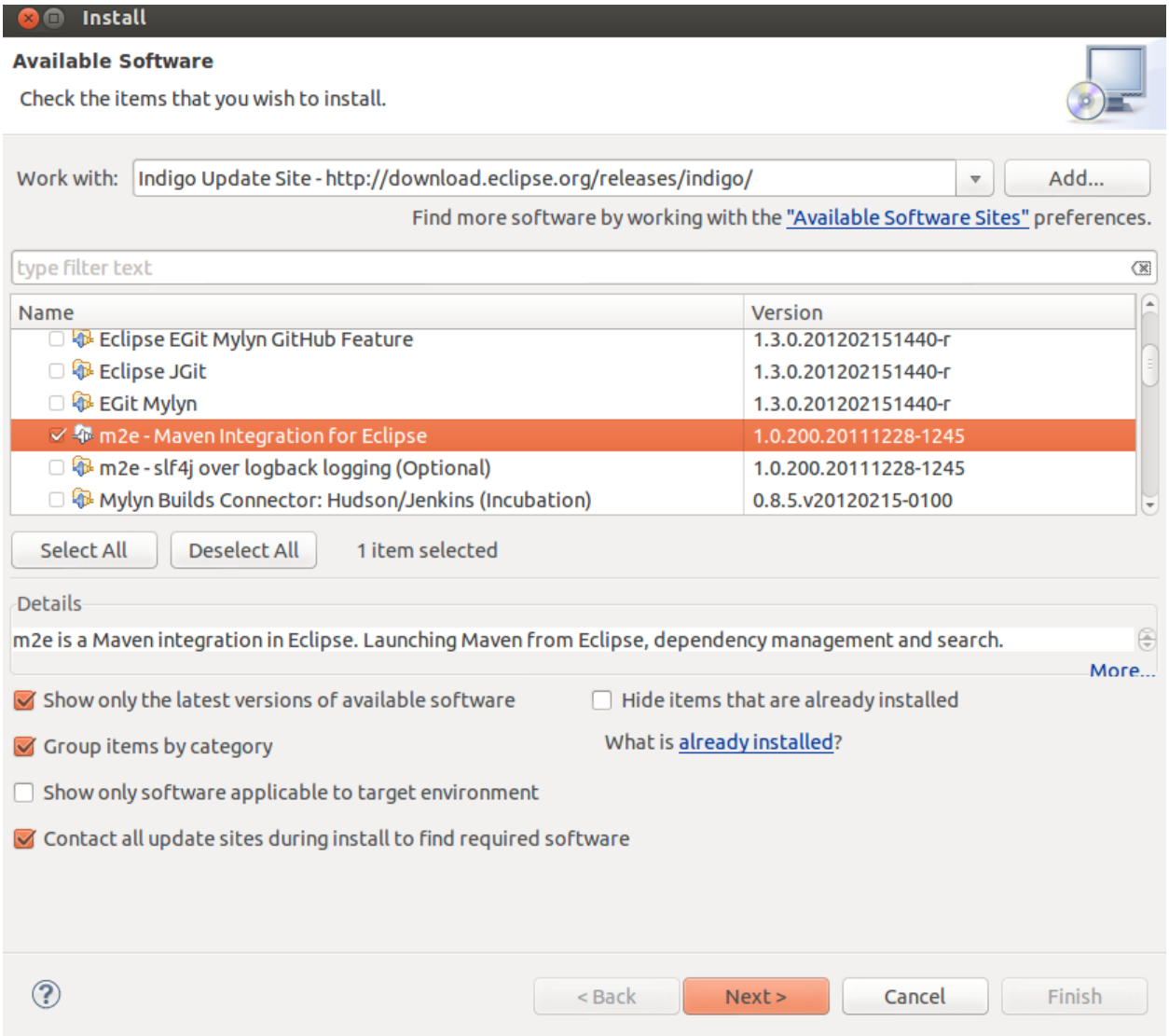
Create Key Pair

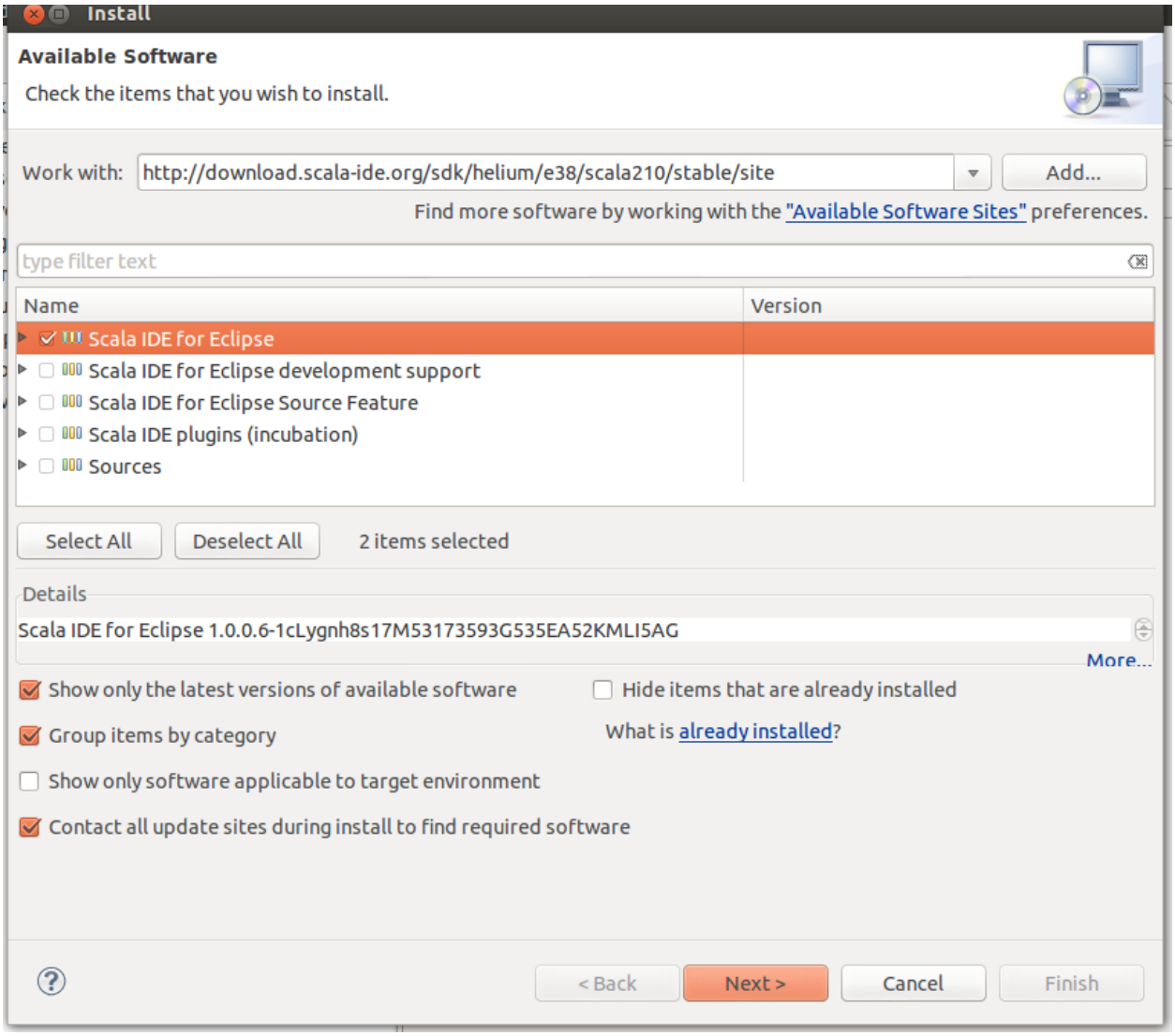
Key pair name: kp-spark

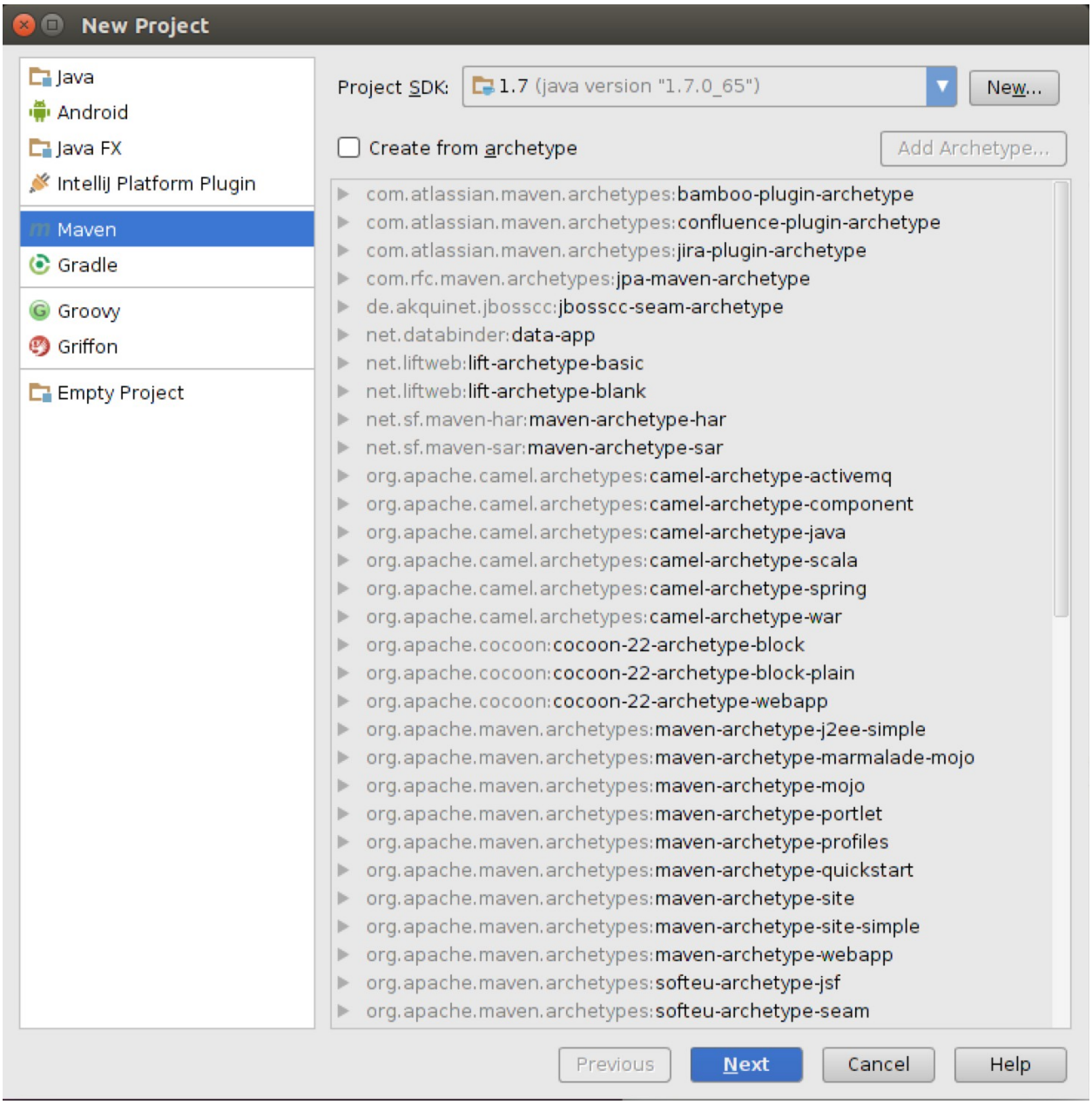
Cancel | Create

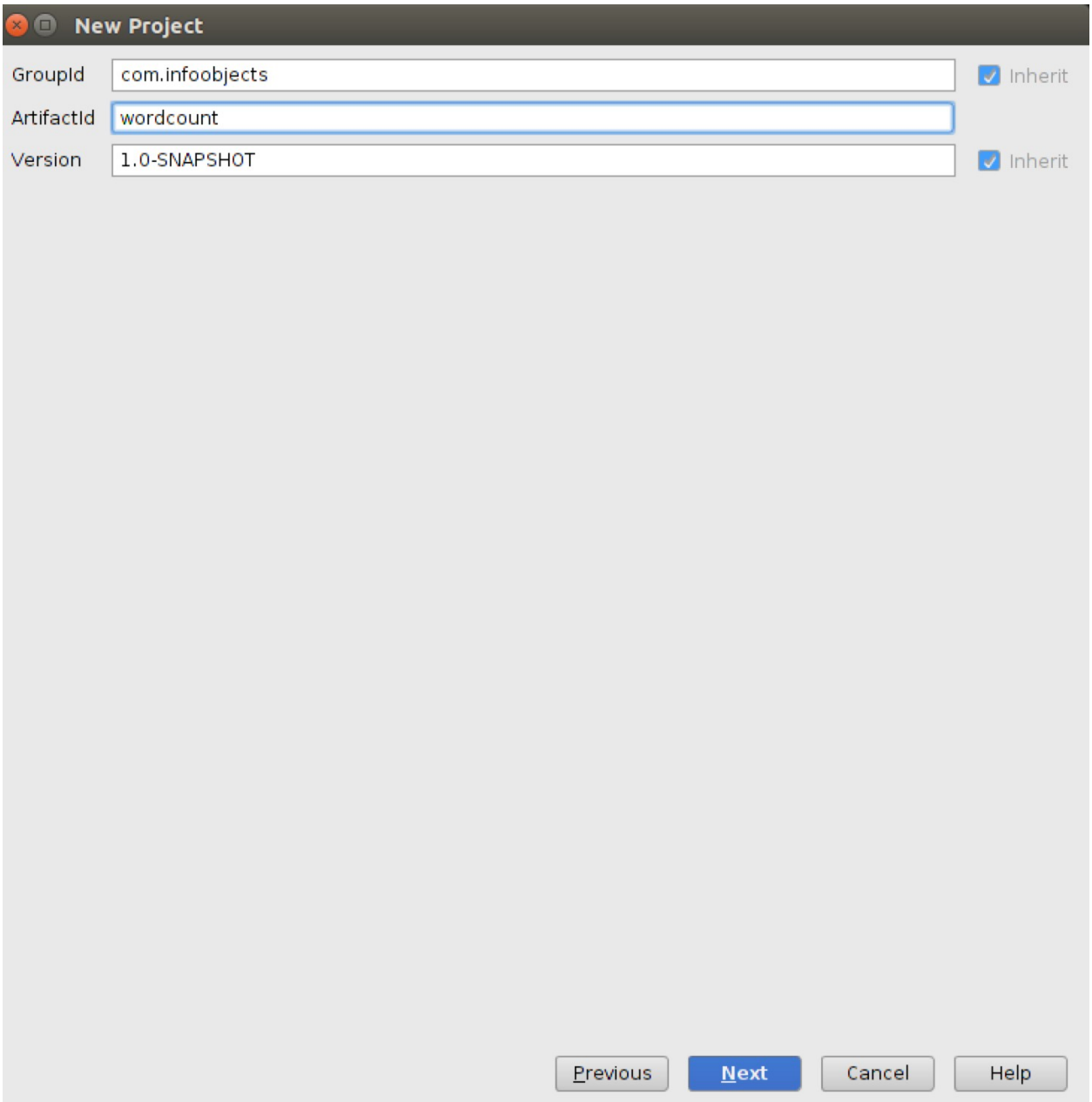
Select a key pair

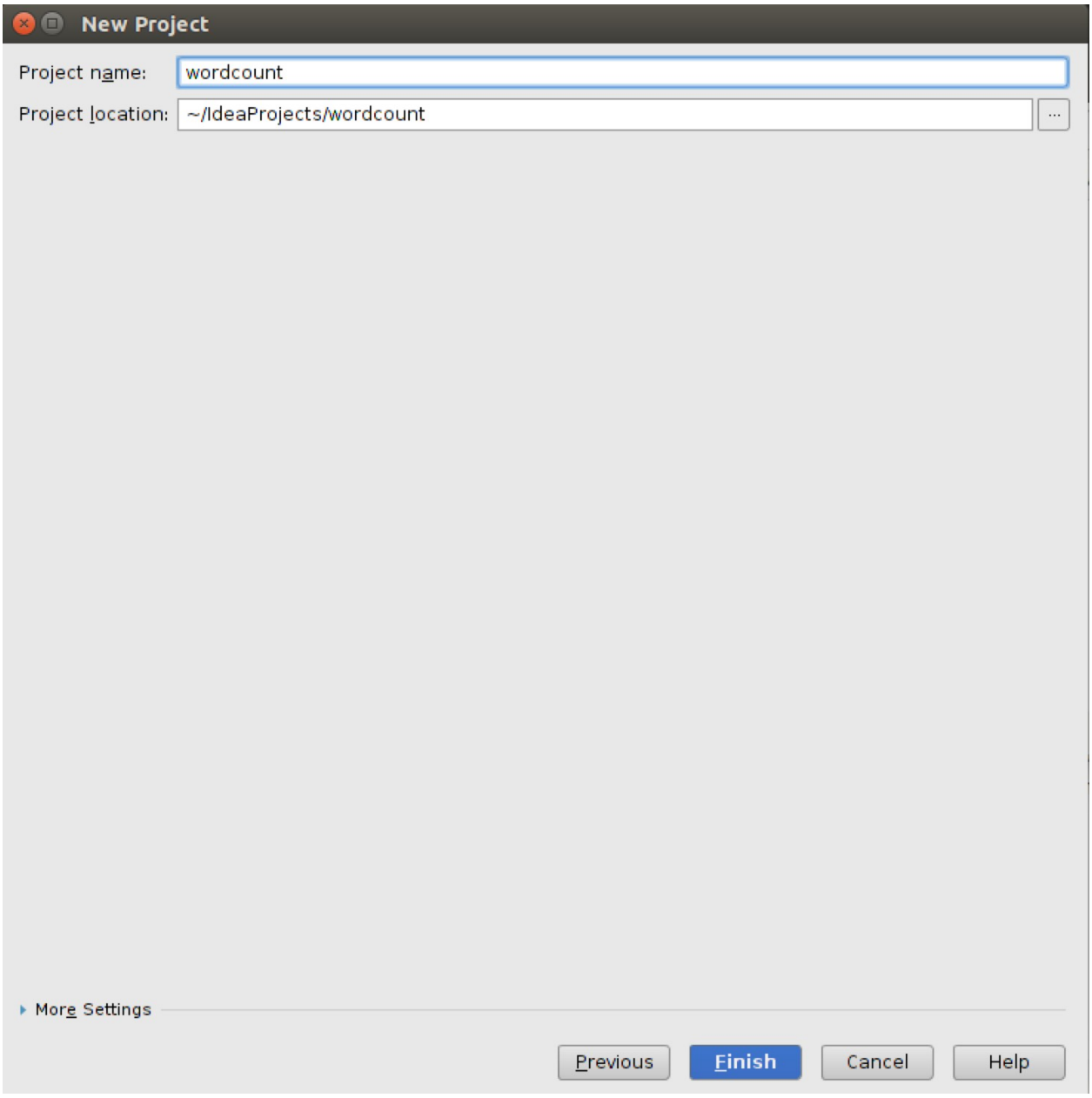
Chapter 2: Developing Applications with Spark




Install
Available Software
Check the items that you wish to install.
Work with:
http://download.scala-ide.org/sdk/helium/e38/scala210/stable/site
Add...
Find more software by working with the "Available Software Sites" preferences.
type filter text
Name
Version
Scala IDE for Eclipse
Scala IDE for Eclipse development support
Scala IDE for Eclipse Source Feature
Scala IDE plugins (incubation)
Sources
Select All
Deselect All
2 items selected
Details
Scala IDE for Eclipse 1.0.0.6-1cLygnh8s17M53173593G535EA52KMLI5AG
More...
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 Project
Java
Android
Java FX
Intellij Platform Plugin
Maven
Gradle
Groovy
Griffon
Empty Project
Project SDK:
1.7 (java version "1.7.0_65")
New...
Create from archetype
Add Archetype...
com.atlassian.maven.archetypes:bamboo-plugin-archetype
com.atlassian.maven.archetypes:confluence-plugin-archetype
com.atlassian.maven.archetypes:jira-plugin-archetype
com.rfc.maven.archetypes:jpa-maven-archetype
de.akquinet.jbosscc:jbosscc-seam-archetype
net.databinder:data-app
net.liftweb:lift-archetype-basic
net.liftweb:lift-archetype-blank
net.sf.maven-har:maven-archetype-har
net.sf.maven-sar:maven-archetype-sar
org.apache.camel.archetypes:camel-archetype-activemq
org.apache.camel.archetypes:camel-archetype-component
org.apache.camel.archetypes:camel-archetype-java
org.apache.camel.archetypes:camel-archetype-scala
org.apache.camel.archetypes:camel-archetype-spring
org.apache.camel.archetypes:camel-archetype-war
org.apache.cocoon:cocoon-22-archetype-block
org.apache.cocoon:cocoon-22-archetype-block-plain
org.apache.cocoon:cocoon-22-archetype-webapp
org.apache.maven.archetypes:maven-archetype-j2ee-simple
org.apache.maven.archetypes:maven-archetype-marmalade-mojo
org.apache.maven.archetypes:maven-archetype-mojo
org.apache.maven.archetypes:maven-archetype-portlet
org.apache.maven.archetypes:maven-archetype-profiles
org.apache.maven.archetypes:maven-archetype-quickstart
org.apache.maven.archetypes:maven-archetype-site
org.apache.maven.archetypes:maven-archetype-site-simple
org.apache.maven.archetypes:maven-archetype-webapp
org.apache.maven.archetypes:softeu-archetype-jsf
org.apache.maven.archetypes:softeu-archetype-seam
Previous
Next
Cancel
Help


New Project
GroupId
com.infoobjects
Inherit
ArtifactId
wordcount
Version
1.0-SNAPSHOT
Inherit
Previous
Next
Cancel
Help



Chapter 3: External Data Sources

```
(to,2)
(not,1)
(be,2)
(or,1)
```

ftp://ftp.ncdc.noaa.gov/pub/data/noaa/

Index of /pub/data/noaa/

Name	Size	Date Modified
[parent directory]		
1901/		11/22/04, 12:00:00 AM
1902/		11/22/04, 12:00:00 AM
1903/		11/22/04, 12:00:00 AM
1904/		11/22/04, 12:00:00 AM
1905/		11/22/04, 12:00:00 AM
1906/		11/22/04, 12:00:00 AM
1907/		11/22/04, 12:00:00 AM
1908/		11/22/04, 12:00:00 AM
1909/		11/22/04, 12:00:00 AM
1910/		11/22/04, 12:00:00 AM
1911/		11/22/04, 12:00:00 AM
1912/		11/22/04, 12:00:00 AM

ftp://ftp.ncdc.noaa.gov/pub/data/noaa/1901/

Index of /pub/data/noaa/1901/

Name	Size	Date Modified
[parent directory]		
029070-99999-1901.gz	11.2 kB	11/22/04, 12:00:00 AM
029500-99999-1901.gz	10.9 kB	11/22/04, 12:00:00 AM
029600-99999-1901.gz	11.4 kB	11/22/04, 12:00:00 AM
029720-99999-1901.gz	10.7 kB	11/22/04, 12:00:00 AM
029810-99999-1901.gz	11.7 kB	11/22/04, 12:00:00 AM
227070-99999-1901.gz	10.9 kB	11/22/04, 12:00:00 AM

```
(United States of America,US Dollar)
(Canada,Canadian Dollar)
(Mexico,Peso)
```

Create a Bucket - Select a Bucket Name and Region

Cancel

A bucket is a container for objects stored in Amazon S3. When creating a bucket, you can choose a Region to optimize for latency, minimize costs, or address regulatory requirements. For more information regarding bucket naming conventions, please visit the Amazon S3 documentation.

Bucket Name:

Region: Select a Region

- US Standard
- Oregon
- Northern California
- Ireland
- Singapore
- Tokyo
- Sydney
- Sao Paulo
- Frankfurt

Set Up Logging > | Create | Cancel

AWS | Services | Edit | Rishi Raj Yadav | Global | Support

Upload | Create Folder | Actions | None | Properties | Transfers

All Buckets / com.infoobjects.wordcount

Name	Storage Class	Size	Last Modified

The bucket 'com.infoobjects.wordcount' is empty

Upload - Select Files and Folders

Cancel

Upload to: All Buckets / com.infoobjects.wordcount / words

To upload files (up to 5 TB each) to Amazon S3, click **Add Files**. You can also drag and drop files and folders to the area below. To remove files already selected, click the **X** to the far right of the file name.

Drag and drop files and folders to upload here.

sh.txt **(19 bytes)** X

Add Files | **Remove Selected Files**

Number of files: **1** Total upload size: **19 bytes**

Set Details > | Start Upload | Cancel

AWS | Services | Edit | Rishi Raj Yadav | Global | Support

Upload | Create Folder | Actions | None | Properties | Transfers

All Buckets / com.infoobjects.wordcount / words

Name	Storage Class	Size	Last Modified
sh.txt	Standard	19 bytes	Fri Dec 26 14:06:47 GMT-800 2014

Object: sh.txt

Bucket: com.infoobjects.wordcount
Folder: words
Name: sh.txt
Link: https://s3-us-west-1.amazonaws.com/com.infoobjects.wordcount/words/sh.txt
Size: 19
Last Modified: Fri Dec 26 14:06:47 GMT-800 2014
Owner: Me
ETag: f9d804763c3031cc22323d79e165b562
Expiry Date: None
Expiration Rule: N/A

- Details
- Permissions
- Metadata

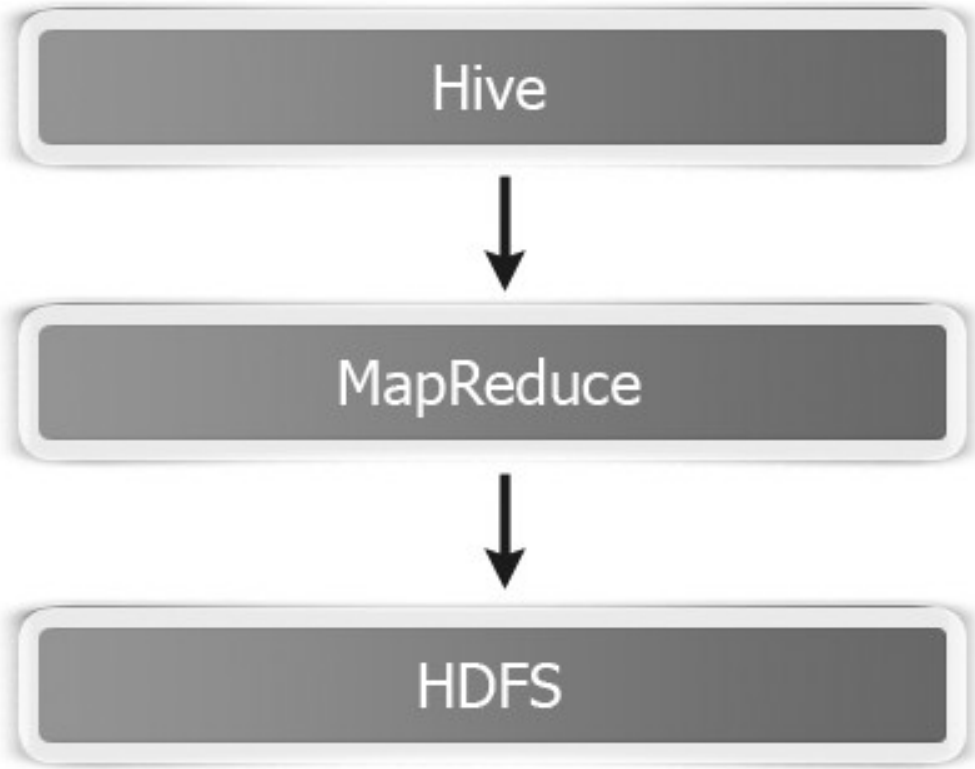
```
hduser@localhost:~/uber$ ls
build.sbt  project  target
```

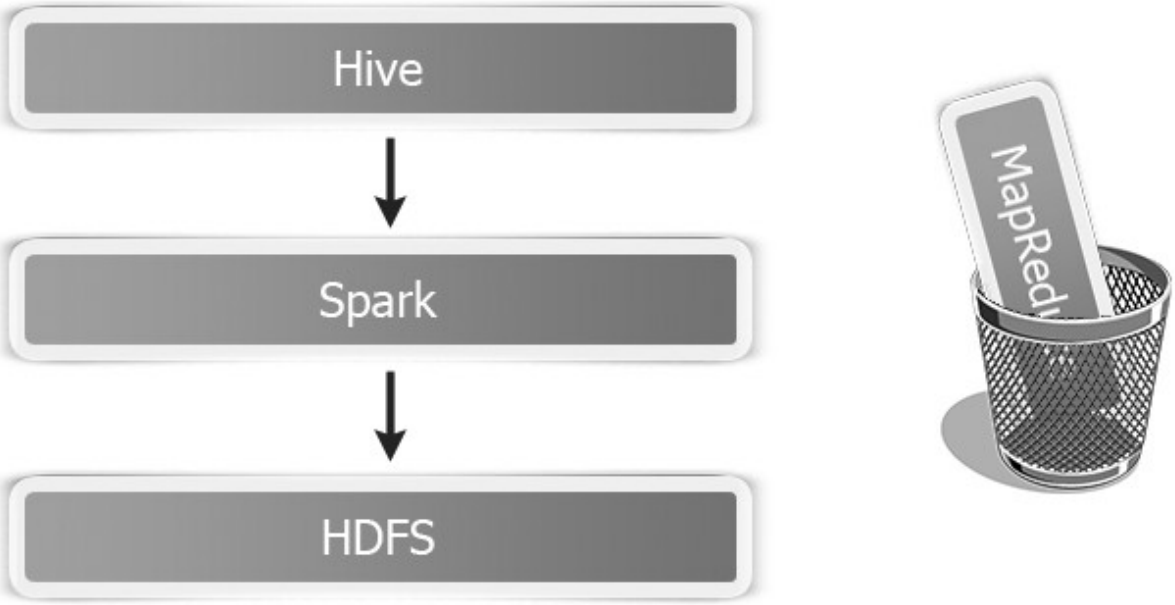
```
name := "sc-uber"

libraryDependencies += "com.datastax.spark" %% "spark-cassandra-connector" % "1.1.0"
```

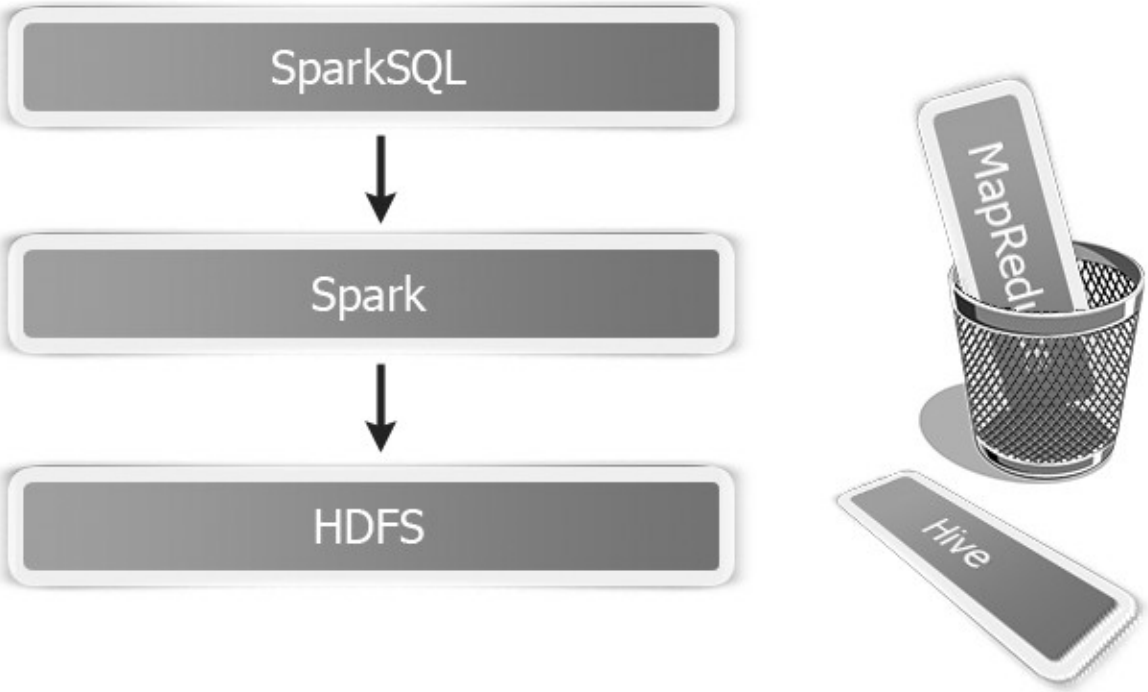

```
name := "sc-uber"

libraryDependencies += "com.datastax.spark" %% "spark-cassandra-connector" % "1.1.0"
```

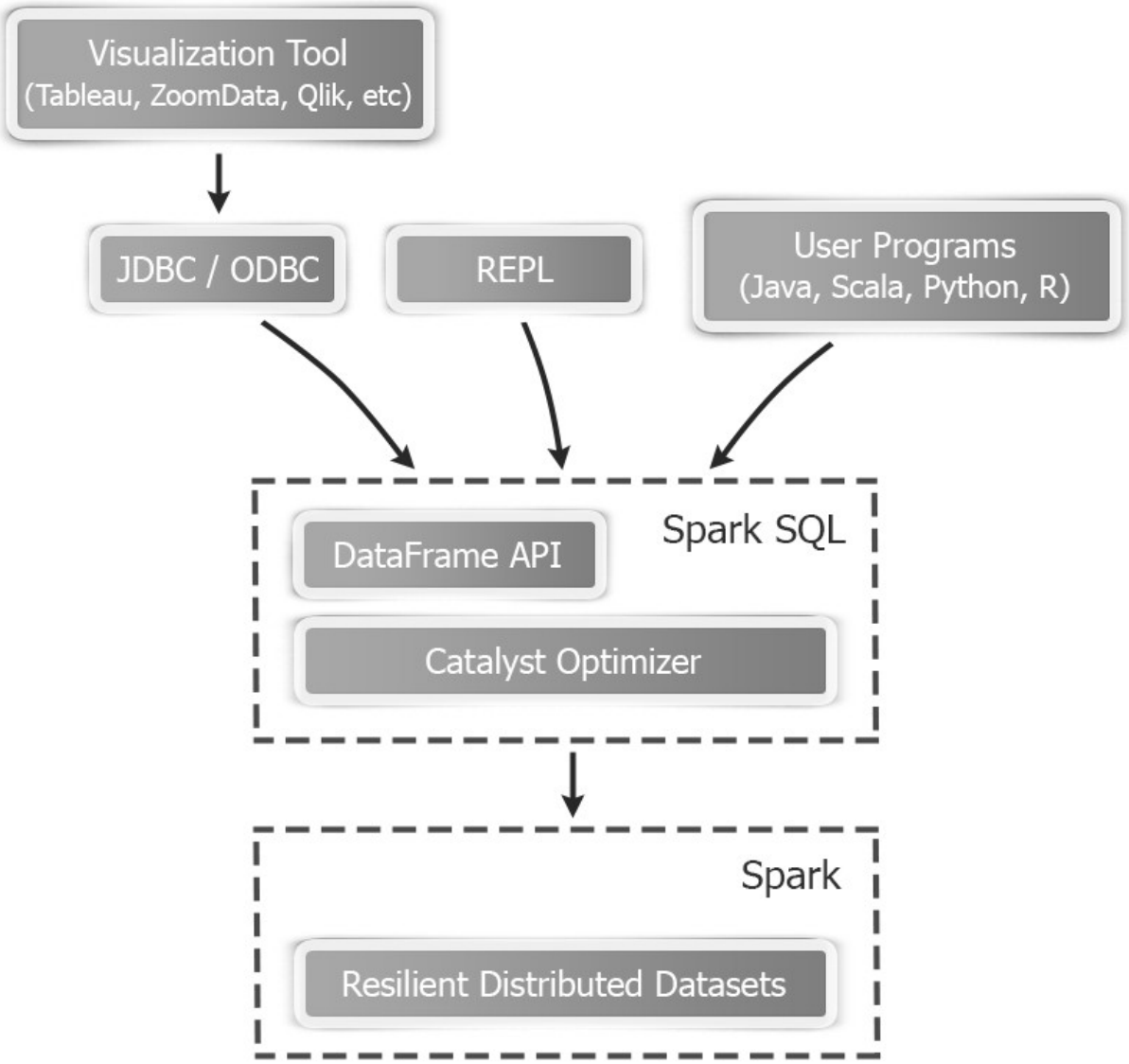

```
assemblyMergeStrategy in assembly := {
 case PathList("META-INF", xs @ _*) =>
    (xs map {_.toLowerCase}) match {
      case ("manifest.mf" :: Nil) | ("index.list" :: Nil) | ("dependencies" :: Nil) => MergeStrategy.discard
      case _ => MergeStrategy.discard
    }
 case _  => MergeStrategy.first
}
```

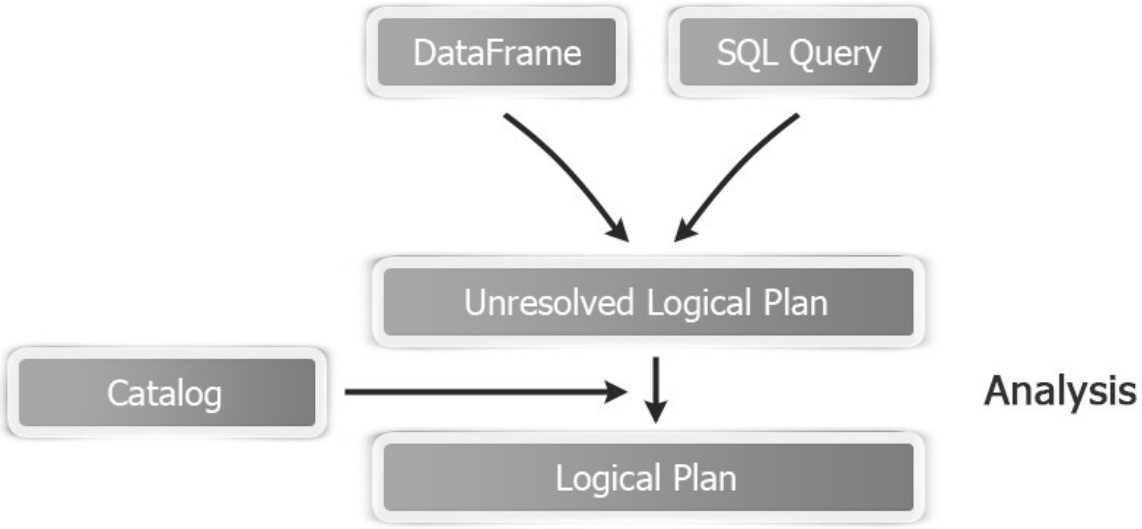
Chapter 4: Spark SQL

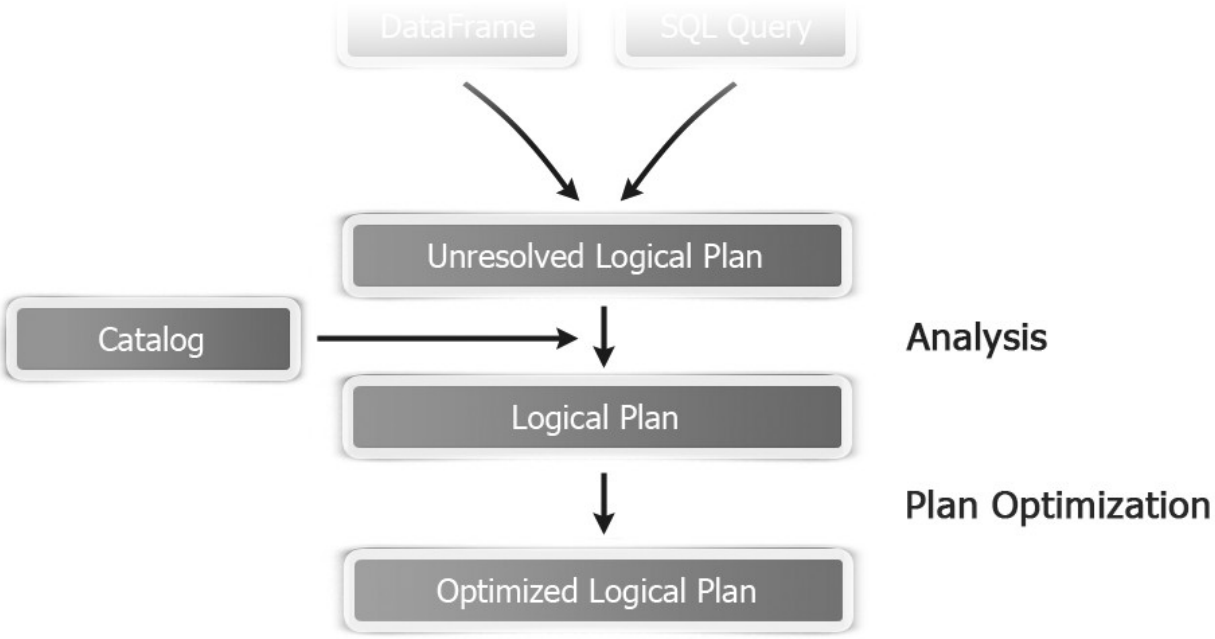


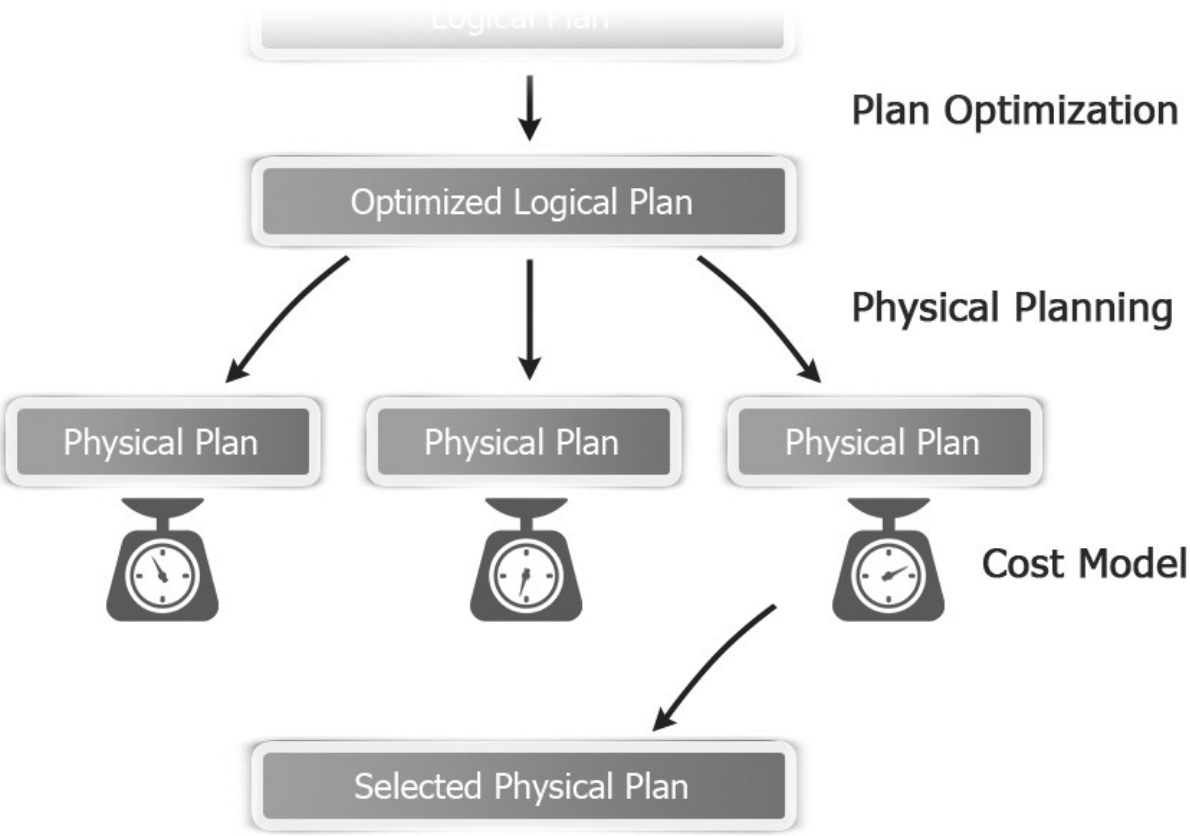

Hive
Spark
HDFS
MapRed


SparkSQL
Spark
HDFS
MapRed
Hive

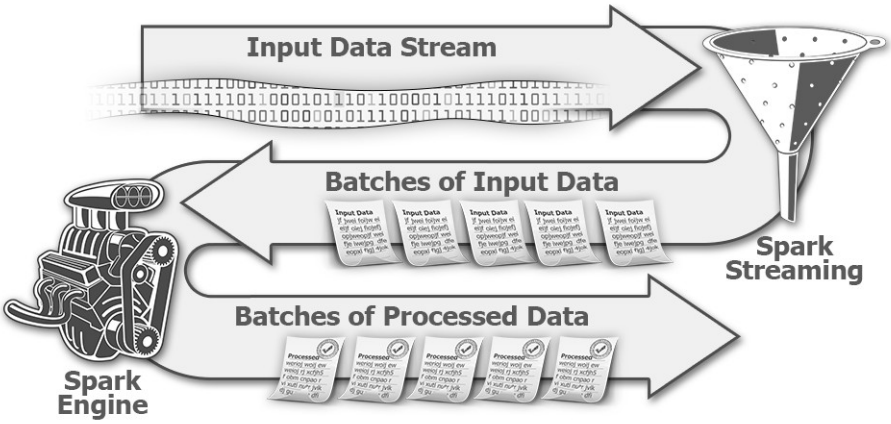

Visualization Tool
(Tableau, ZoomData, Qlik, etc)
JDBC / ODBC
REPL
User Programs
(Java, Scala, Python, R)
Spark SQL
DataFrame API
Catalyst Optimizer
Spark
Resilient Distributed Datasets


DataFrame
SQL Query
Unresolved Logical Plan
Catalog
Analysis
Logical Plan


DataFrame
SQL Query
Unresolved Logical Plan
Catalog
Analysis
Logical Plan
Plan Optimization
Optimized Logical Plan



Chapter 5: Spark Streaming



```
-------------------------------------------
Time: 1421458202000 ms
-------------------------------------------
(not,1)
(or,1)
(be,2)
(to,2)
```

Application Management

Create an application

Application Details

Name *

spark-cookbook-app

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

Streaming example

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *

http://www.infoobjects.com

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

Callback URL

Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

Developer Agreement

Last Update: October 22, 2014.

This Twitter Developer Agreement ("**Agreement**") is made between you (either an individual or an entity, referred to herein as "**you**") and Twitter, Inc., on behalf of itself and its worldwide affiliates (collectively, "**Twitter**") and governs your access to and use of the Licensed Material (as defined below).

PLEASE READ THE TERMS AND CONDITIONS OF THIS AGREEMENT CAREFULLY, INCLUDING WITHOUT LIMITATION ANY LINKED TERMS AND CONDITIONS APPEARING OR REFERENCED BELOW, WHICH ARE HEREBY MADE PART OF THIS LICENSE AGREEMENT. BY USING THE LICENSED MATERIAL, YOU ARE AGREEING THAT YOU HAVE READ, AND THAT YOU AGREE TO COMPLY WITH AND TO BE BOUND BY THE TERMS AND CONDITIONS OF THIS AGREEMENT AND ALL APPLICABLE LAWS AND REGULATIONS IN THEIR ENTIRETY WITHOUT LIMITATION OR QUALIFICATION. IF YOU DO NOT AGREE TO BE BOUND BY THIS AGREEMENT, THEN YOU MAY NOT ACCESS OR OTHERWISE USE THE LICENSED MATERIAL. THIS AGREEMENT IS EFFECTIVE AS OF THE FIRST DATE THAT YOU USE THE LICENSED MATERIAL ("**EFFECTIVE DATE**").

IF YOU ARE AN INDIVIDUAL REPRESENTING AN ENTITY, YOU ACKNOWLEDGE THAT YOU HAVE THE APPROPRIATE AUTHORITY TO ACCEPT THIS AGREEMENT ON BEHALF OF SUCH ENTITY. YOU MAY NOT USE THE LICENSED MATERIAL AND MAY NOT ACCEPT THIS AGREEMENT IF YOU ARE NOT OF LEGAL AGE TO FORM A BINDING CONTRACT WITH TWITTER, OR YOU ARE BARRED FROM USING OR RECEIVING THE LICENSED MATERIAL UNDER APPLICABLE LAW.

☐ Yes, I agree

Create your Twitter application

spark-cookbook-app

Test OAuth

Details | Settings | Keys and Access Tokens | Permissions

Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	sSRET7x8yNid8C6jMQ6r1qrkt
Consumer Secret (API Secret)	M4ruHV1nTuP5RfrG4X97vlHbdDKmogRzi76t67Mb3ht74viL1C
Access Level	Read-only (modify app permissions)
Owner	meditativesoul
Owner ID	31548859

Application Actions

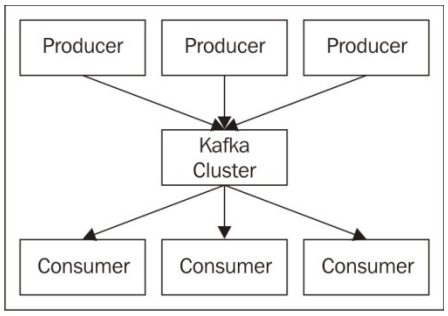
Regenerate Consumer Key and Secret | Change App Permissions

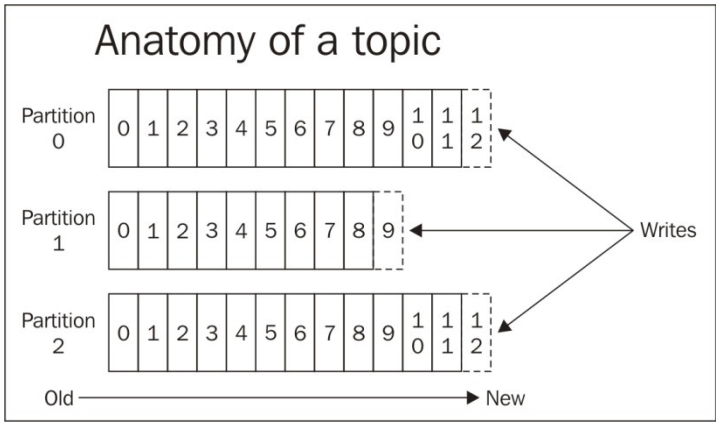
Your Access Token

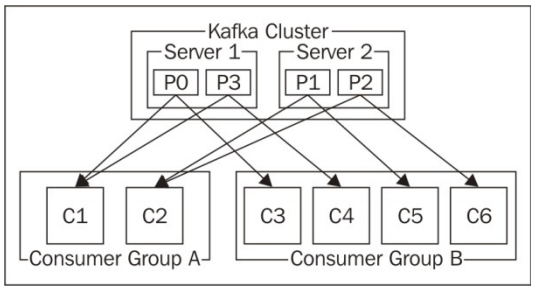
This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	31548859-sB9zJN9e6N70hZmQJbmbGDyYPbhBhyRH8cEw6ocbi
Access Token Secret	ni5hJqnLu6gsxqBUKSo5S1RVFEwxDTHaChq5R3yWWXm8H
Access Level	Read-only
Owner	meditativesoul
Owner ID	31548859

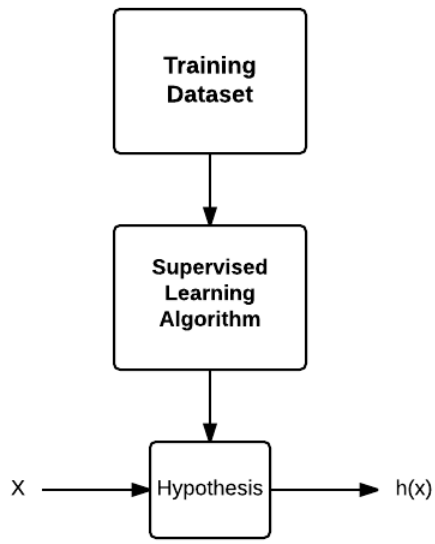
```
-------------------------------------------
Time: 1421629706000 ms
-------------------------------------------
(not,1)
(or,1)
(be,2)
(to,2)
```







Chapter 7: Supervised Learning with MLlib – Regression



$$y = \theta_0 + \theta_1 x$$

$$h(x) = \theta_0 + \theta_1 x$$

$$h(x) = \theta_0 + \theta_1 x$$

$$\left(x^i - x^i\right)^2 + \left(h\left(x^i\right) - y^i\right)^2$$
$$= \left(h\left(x^i\right) - y^i\right)^2$$

$$\frac{1}{2m}\sum_{i=1}^{m}\left(h(x)^i - y^i\right)^2$$

$$J(\theta_0,\theta_1)=\frac{1}{2m}\sum_{i=1}^{m}\left(h(x)^i-y^i\right)^2$$

$$\left(x^1,y^1\right)=(1,1)$$
$$\left(x^2,y^2\right)=(2,2)$$
$$\left(x^3,y^3\right)=(3,3)$$

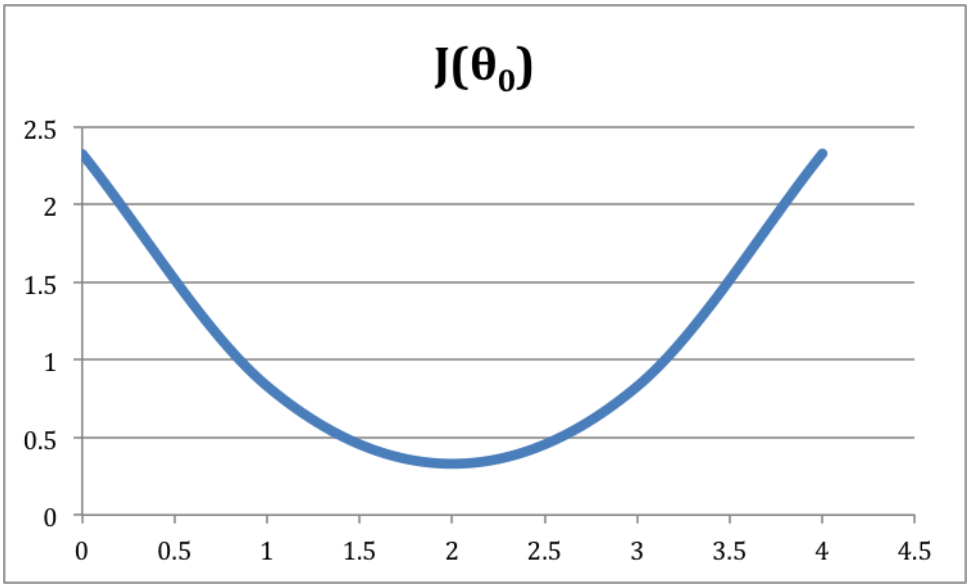
$$(\theta_0,\theta_1)=(0,0)$$
$$J(\theta_0)=\frac{1}{2\times 3}\sum_{i=1}^{3}\left(y^i\right)^2$$
$$=\frac{1}{2\times 3}(1+4+9)=\frac{14}{6}=2.33$$

$$(\theta_0,\theta_1)=(1,0)$$
$$J(\theta_0)=\frac{1}{2\times 3}\sum_{i=1}^{3}\left(1-y^i\right)^2$$
$$=\frac{1}{2\times 3}(0+1+4)=\frac{5}{6}=0.83$$

$$(\theta_0,\theta_1)=(2,0)$$
$$J(\theta_0)=\frac{1}{2\times 3}\sum_{i=1}^{3}\left(2-y^i\right)^2$$
$$=\frac{1}{2\times 3}(1+0+1)=\frac{2}{6}=0.33$$

$$(\theta_0,\theta_1)=(3,0)$$
$$J(\theta_0)=\frac{1}{2\times 3}\sum_{i=1}^{3}\left(3-y^i\right)^2$$
$$=\frac{1}{2\times 3}(4+1+0)=\frac{5}{6}=0.83$$

$$(\theta_0,\theta_1)=(4,0)$$
$$J(\theta_0)=\frac{1}{2\times 3}\sum_{i=1}^{3}\left(4-y^i\right)^2$$
$$=\frac{1}{2\times 3}(9+4+1)=\frac{14}{6}=2.33$$



$$(\theta_0, \theta_1) = (0, 0)$$

$$J(\theta_1) = \frac{1}{2\times 3}\sum_{i=1}^{3}\left(y^i\right)^2$$

$$= \frac{1}{2\times 3}(1+4+9) = \frac{14}{6} = 2.33$$

$$(\theta_0, \theta_1) = (0, 0.5)$$

$$J(\theta_1) = \frac{1}{2\times 3}\sum_{i=1}^{3}\left(0.5x^i - y^i\right)^2$$

$$= \frac{1}{2\times 3}(0.25+0+2.25) = \frac{2.5}{6} = 0.41$$

$$(\theta_0, \theta_1) = (0, 1)$$

$$J(\theta_1) = \frac{1}{2\times 3}\sum_{i=1}^{3}\left(x^i - y^i\right)^2$$

$$= \frac{1}{2\times 3}(0+0+0) = 0$$

$$(\theta_0, \theta_1) = (0, 1.5)$$

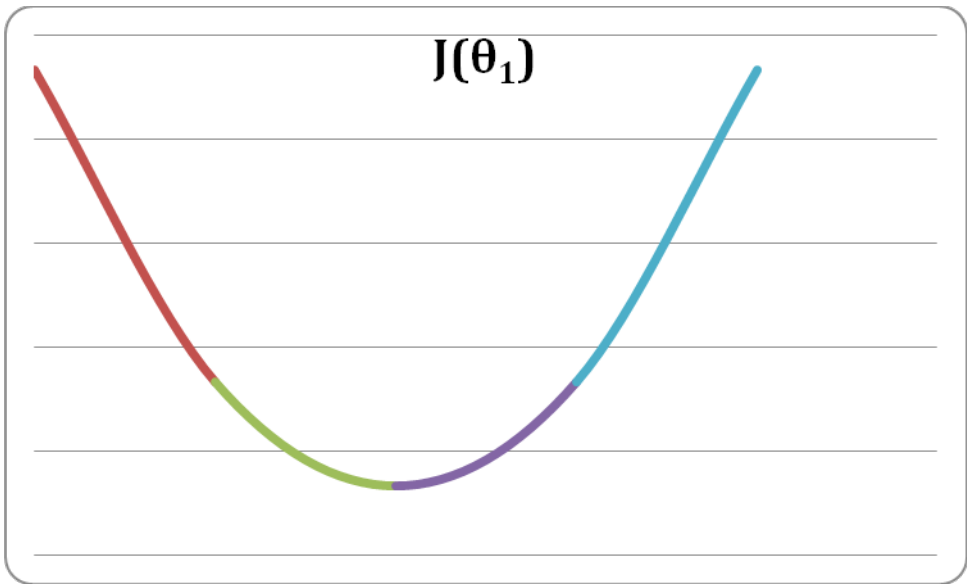
$$J(\theta_1) = \frac{1}{2\times 3}\sum_{i=1}^{3}\left(1.5x^i - y^i\right)^2$$

$$= \frac{1}{2\times 3}(0.25+1+2.25) = \frac{3.5}{6} = 0.58$$

$$(\theta_0, \theta_1) = (0, 2.0)$$

$$J(\theta_1) = \frac{1}{2\times 3}\sum_{i=1}^{3}\left(2x^i - y^i\right)^2$$

$$= \frac{1}{2\times 3}(1+4+9) = \frac{14}{6} = 2.33$$



$$h(x) = \theta_0 + \theta_1 x$$

$$h(x) = \theta_0 + \theta_1 x_1$$

$$x_0$$

$$h(x) = \theta_0 x_0 + \theta_1 x_1$$

$$h(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

$$X = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$$

$$\theta^T x = \begin{bmatrix} \theta_0 & \theta_1 \end{bmatrix} \times \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \theta_0 x_0 + \theta_1 x_1$$

$$h(x) = \theta^T x$$

Chapter 8: Supervised Learning with MLlib – Classification

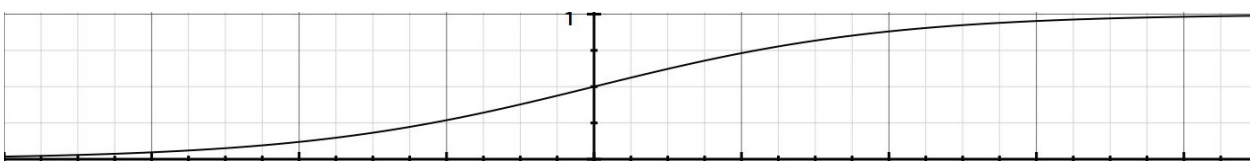
$$y \in \{0,1\}$$

$$1 \geq h(x) \geq 0$$

$$h(x) = \theta^T x$$

$$h(x) = g\left(\theta^T x\right)$$

$$g(t) = \frac{1}{1+e^{-t}}$$



$$h(x) = \frac{1}{1+e^{-\theta^T x}}$$

$$h(x) = P(y=1 \mid x;\theta)$$

$$\theta$$

$$h(x) \geq 0.5$$

$$t \geq 0 \geq 0.5$$

$$h(x) = g\left(\theta^T x\right)$$

$$\theta^T x \geq 0$$

$$\theta^T x \geq 0$$
$$\theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 \geq 0$$

$$\theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 = 0$$

$$h(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2$$

$$h(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h(x^i) - y^i \right)^2$$

$$Cost\left(h(x^i) - y^i \right) = \frac{\left(h(x^i) - y^i \right)^2}{2}$$

$$J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} Cost\left(h(x^i) - y^i \right)$$

$$Cost\left(h(x) - y \right) = \frac{\left(h(x) - y \right)^2}{2}$$

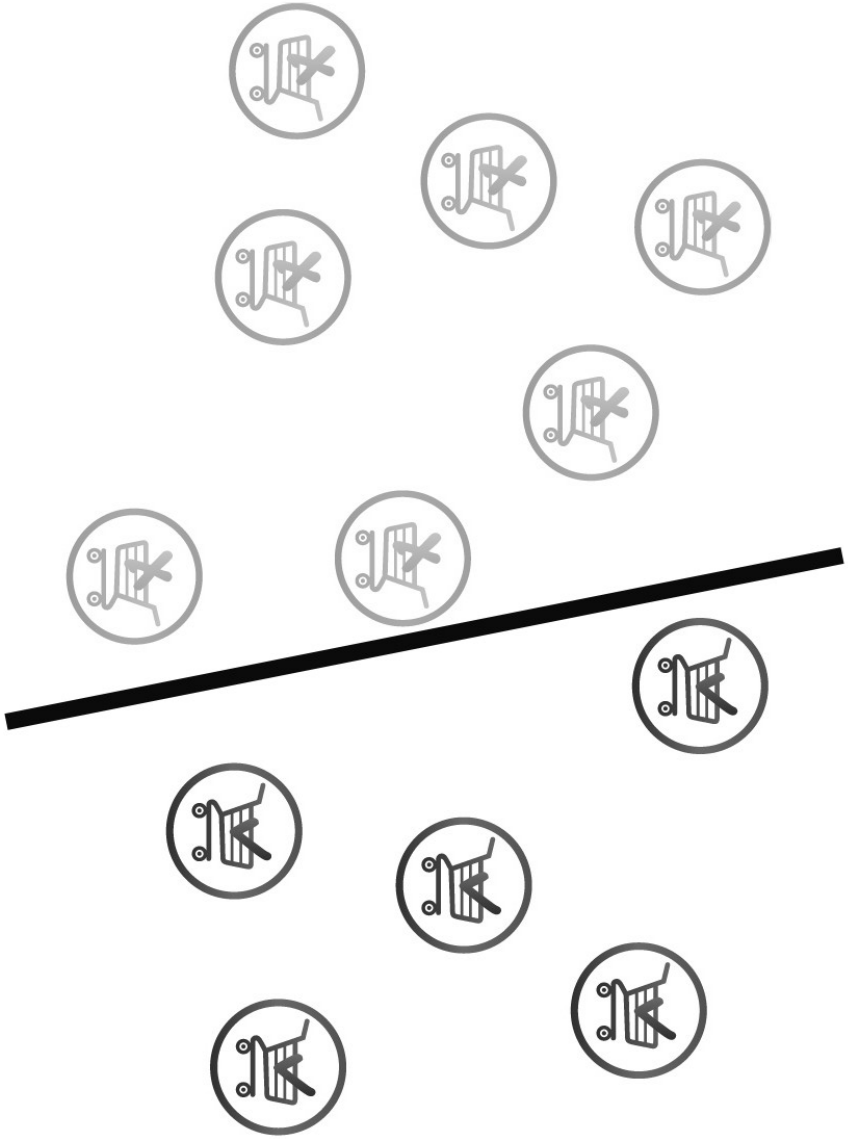
$$Cost\left(h(x), y \right) = -\log\left(h(x) \right) \text{ // for positive class}$$

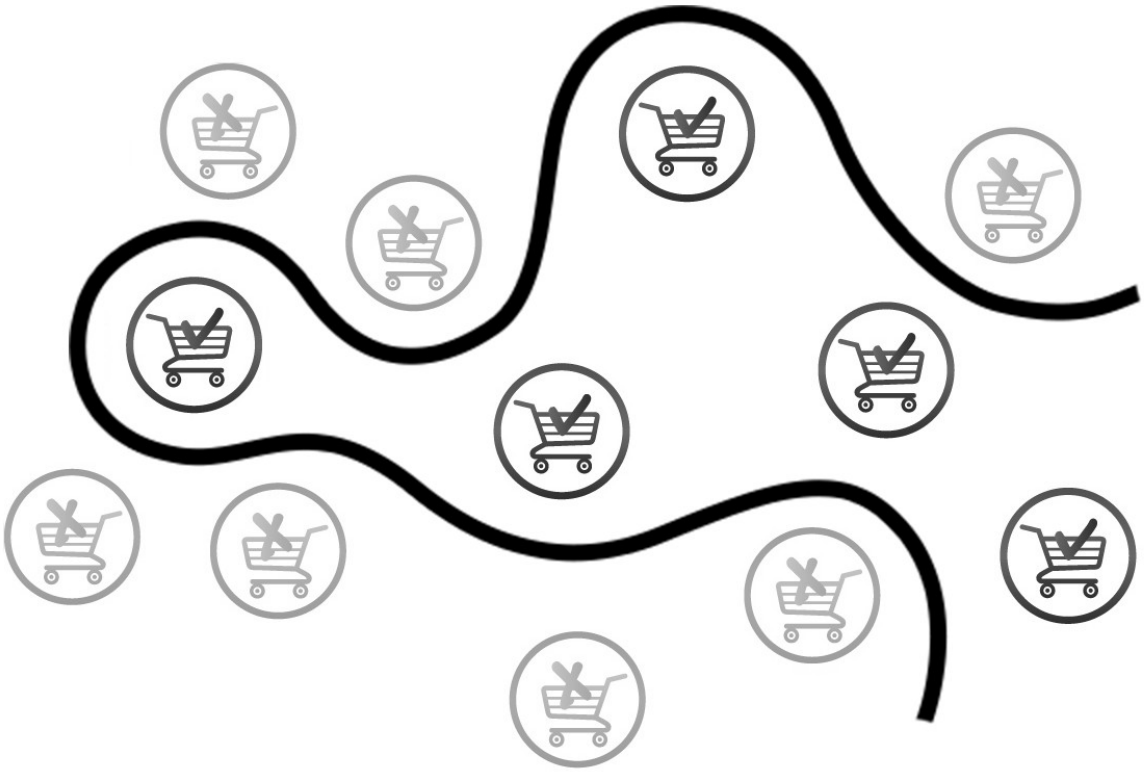
$$Cost\left(h(x), y \right) = -\log\left(1 - h(x) \right) \text{ // for negative class}$$

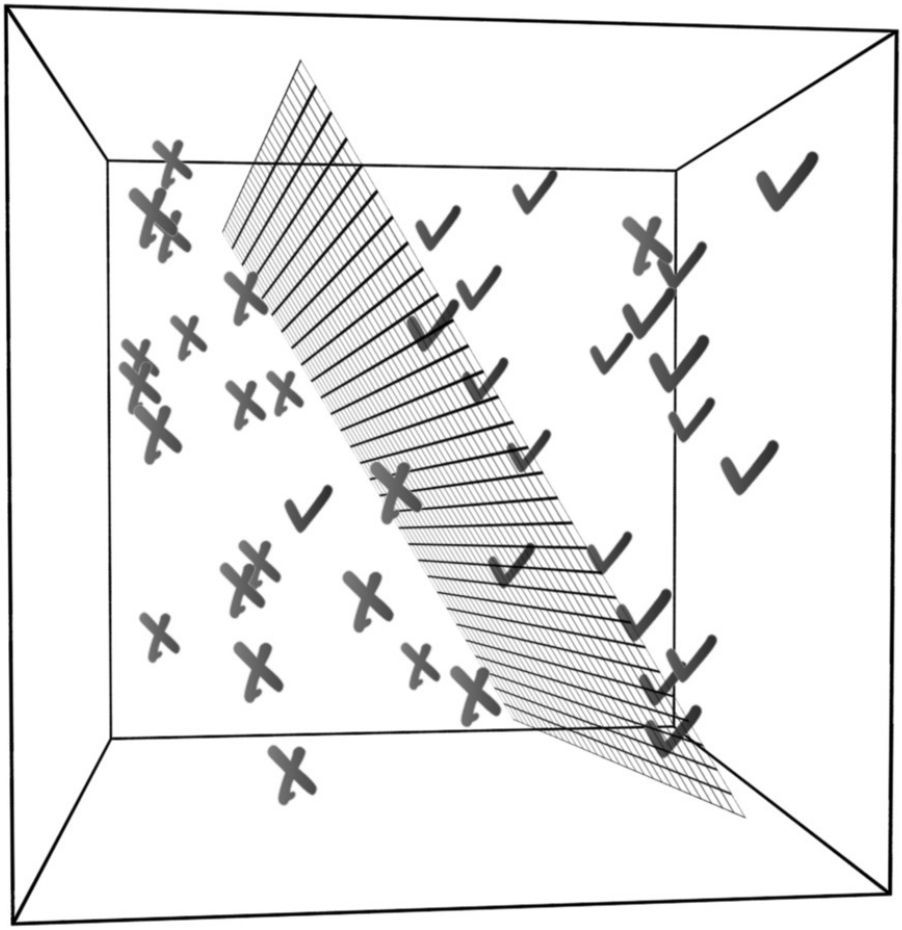
$$Cost\left(h(x), y \right) = -y \log\left(h(x) \right) - (1 - y) \log\left(1 - h(x) \right)$$

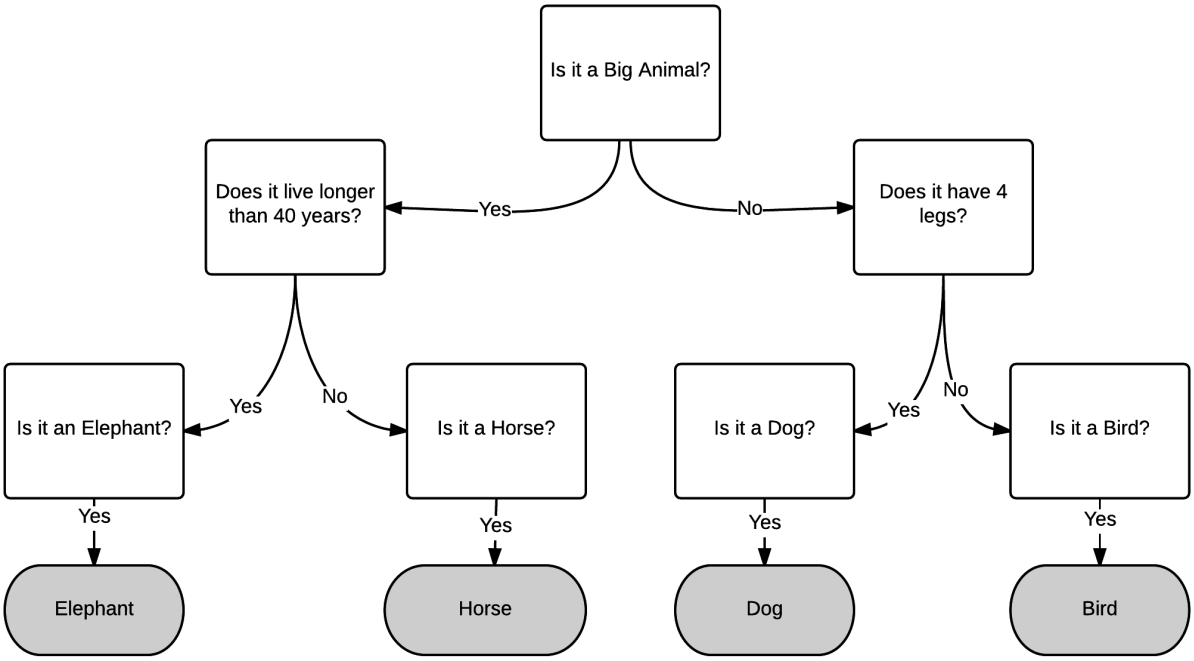
$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left(y^i \log h(x^i) + (1 - y^i) \log\left(1 - h(x^i) \right) \right)$$

$$J(\theta)$$

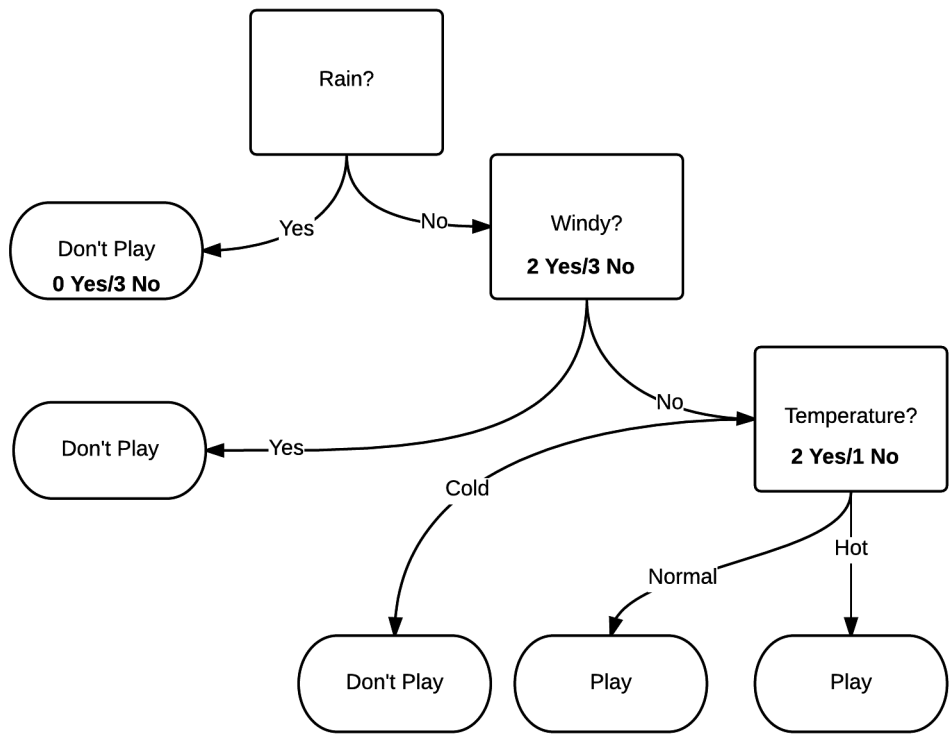







Is it a Big Animal?
Yes
No
Does it live longer than 40 years?
Does it have 4 legs?
Yes
No
Yes
No
Is it an Elephant?
Is it a Horse?
Is it a Dog?
Is it a Bird?
Yes
Yes
Yes
Yes
Elephant
Horse
Dog
Bird

$$Rain\{Yes, No\} => \{2.0, 1.0\}$$
$$Windy\{Yes, No\} => \{2.0, 1.0\}$$
$$Temperature\{Hot, Normal, Cold\} => \{3.0, 2.0, 1.0\}$$



$$Entropy(S) = -p_+ log_2 p_+ - p_- log_2 p_-$$

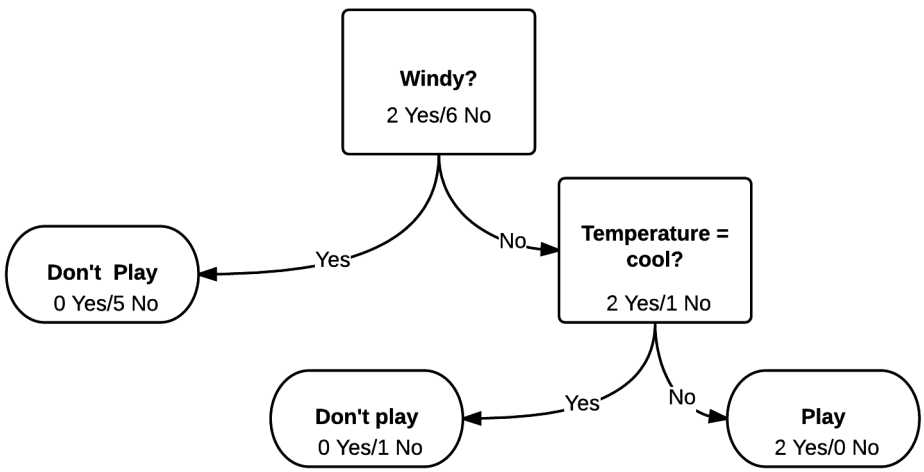
$$Entropy(S) = -0 - 1 log 1 = 0$$

$$\begin{aligned} Entropy(S) &= -0.4 log_2 04 - 0.6 log_2 0.6 \\ &= -0.4 \times (-1.32) - 0.6 \times (-0.736) \\ &= 0.528 + 0.4416 \\ &= 0.967 \end{aligned}$$

$$IG(rain, s1) = Impurity(rain) - \left(\frac{N_{no\ rain}}{N_{rain}}\right) Impurity(no\ rain)$$
$$- \left(\frac{N_{wind}}{N_{rain}}\right) Impurity(wind)$$

$$
\begin{aligned}
Entropy(rain) &= -\left(\frac{2}{8}\right)log_2\left(\frac{2}{8}\right)-\left(\frac{6}{8}\right)log_2\left(\frac{6}{8}\right) \\
&= -\left(\frac{1}{4}\right)\times(-2)-\left(\frac{3}{4}\right)\times(-0.41) \\
&= 0.8
\end{aligned}
$$

$$
\begin{aligned}
IG(rain, s1) &= Impurity(rain)-\left(\frac{N_{no\ rain}}{N_{rain}}\right)Impurity(no\ rain) \\
&\quad -\left(\frac{N_{wind}}{N_{rain}}\right)Impurity(wind) \\
&= 0.8-\left(\frac{5}{8}\right)\times 0.967 \\
&= 0.2
\end{aligned}
$$

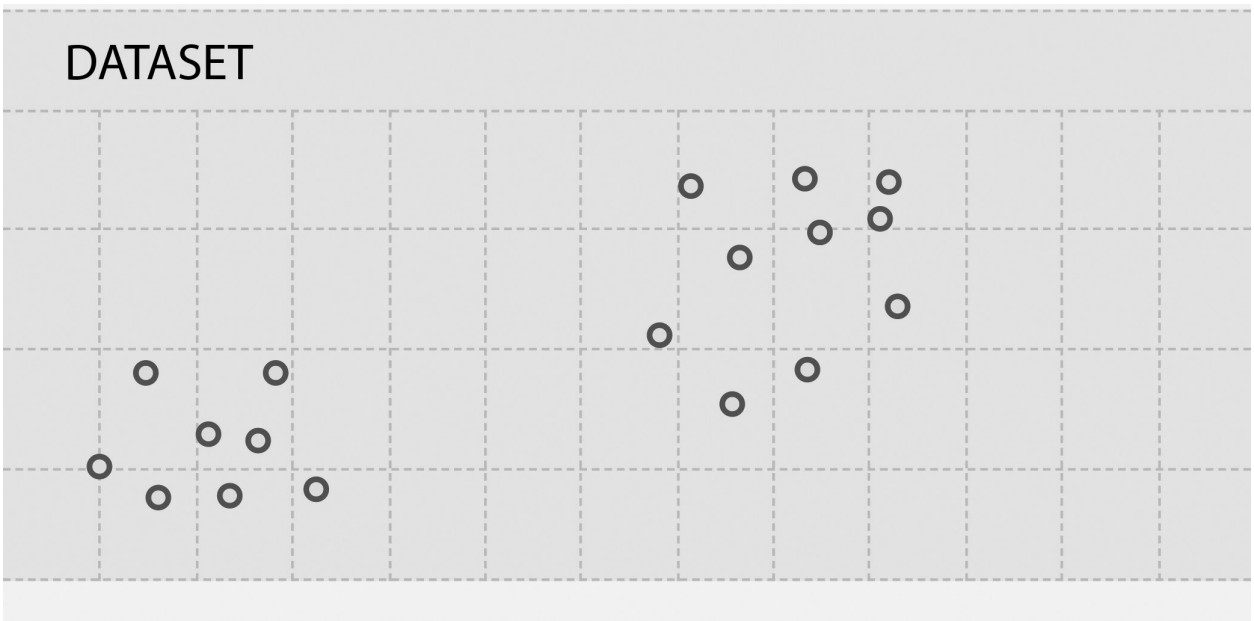


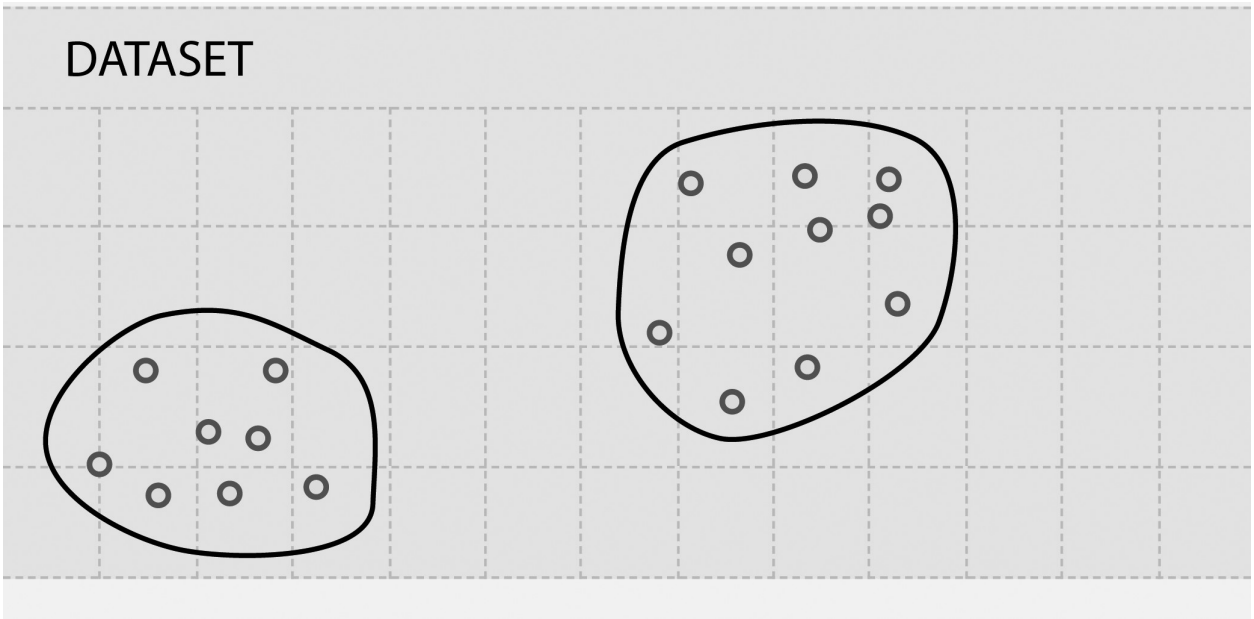
$$
\begin{aligned}
IG_{windy?}, s1 &= Impurity_{windy?} - N_{no\ wind}N_{windy?}Impurity(no\ wind) \\
&\quad - N_{windy}N_{windy?}Impurity(windy) \\
&= 0.44
\end{aligned}
$$

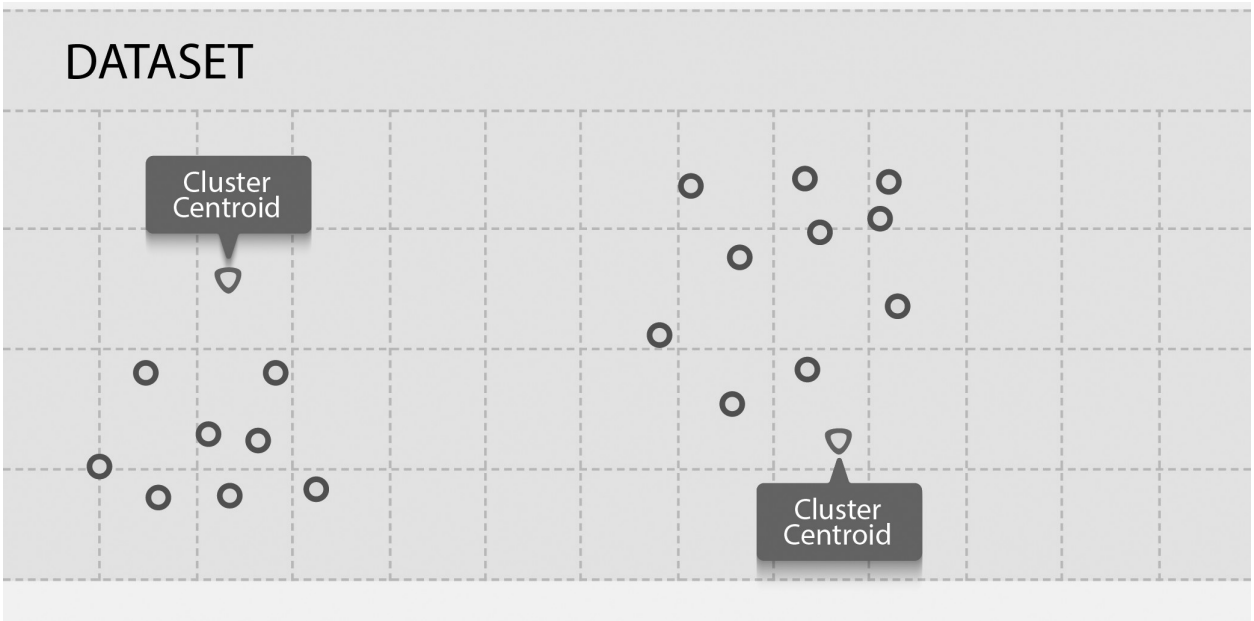
$$y \in \{0,1\}$$

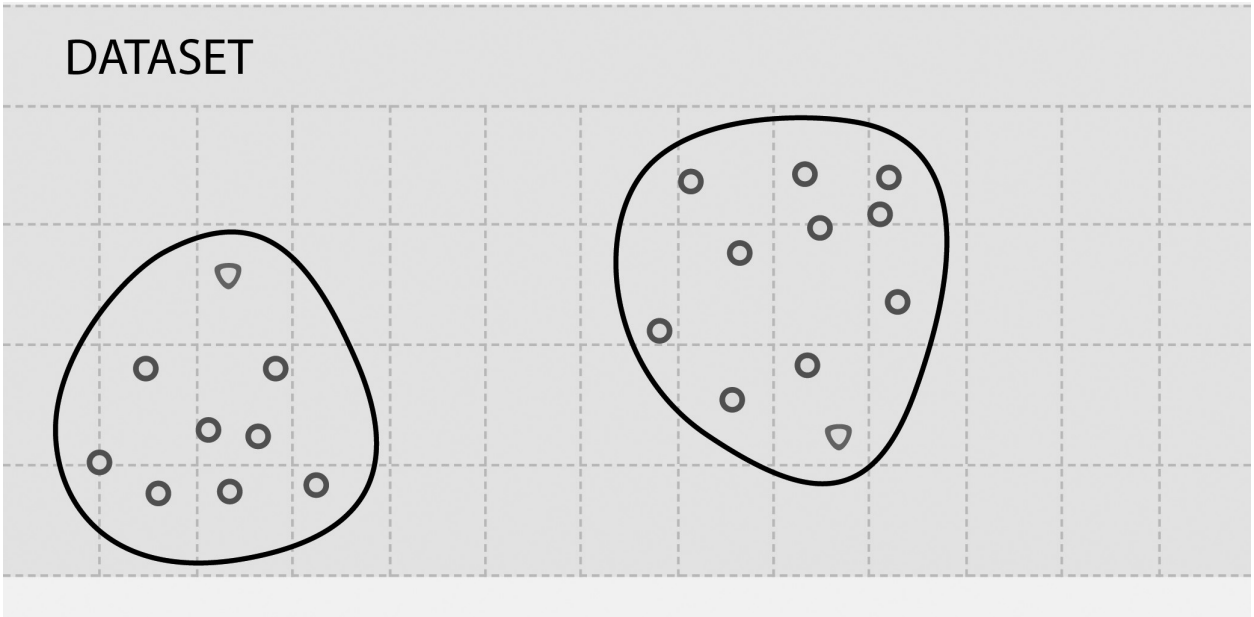
$$x = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 1 \\ \dots \\ 1 \\ \dots \\ 1 \\ \dots \end{bmatrix} \begin{matrix} a \\ aard-vark \\ \dots \\ online \\ \dots \\ pharmacy \\ \dots \\ sale \\ \dots \end{matrix}$$

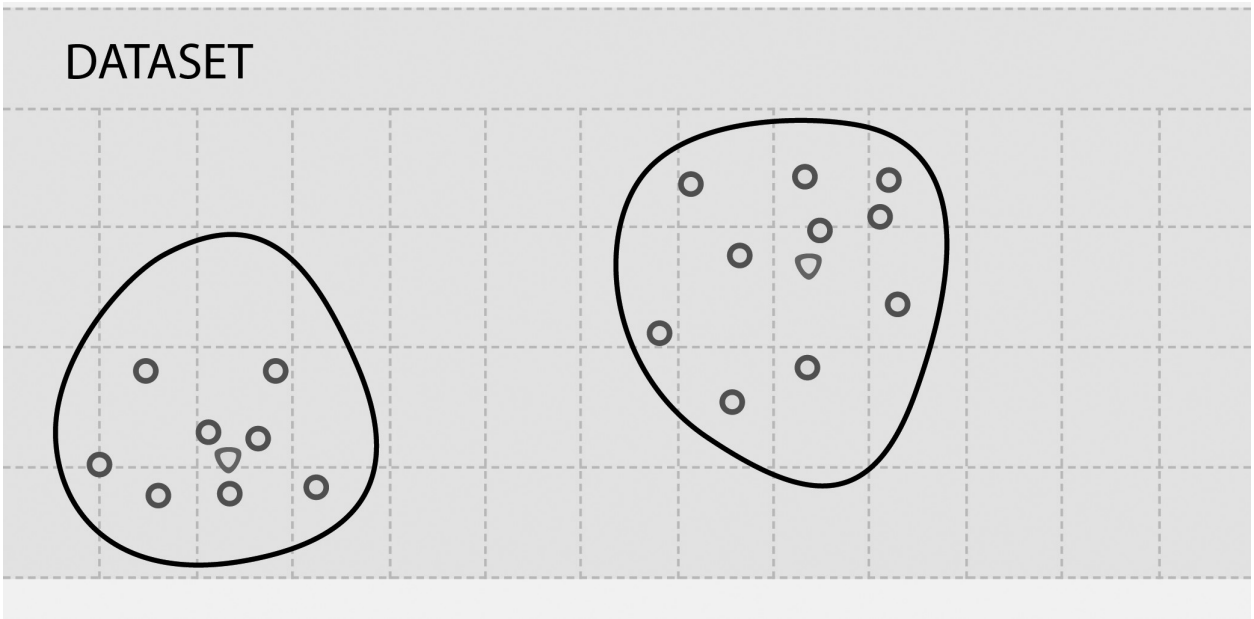
Chapter 9: Unsupervised Learning with MLlib

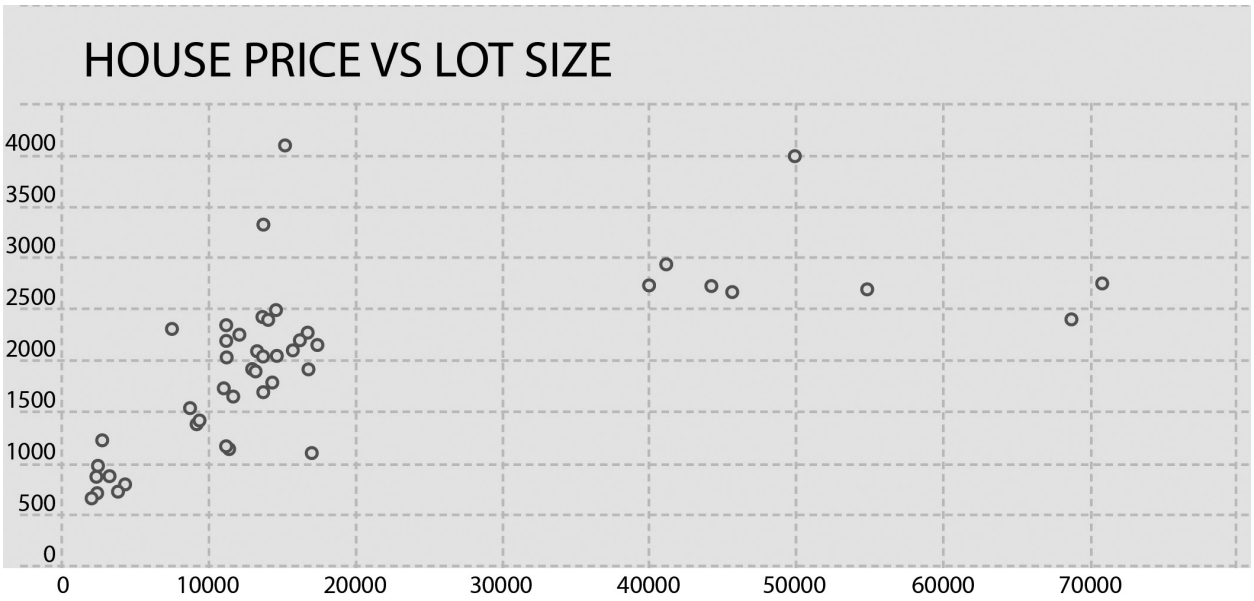


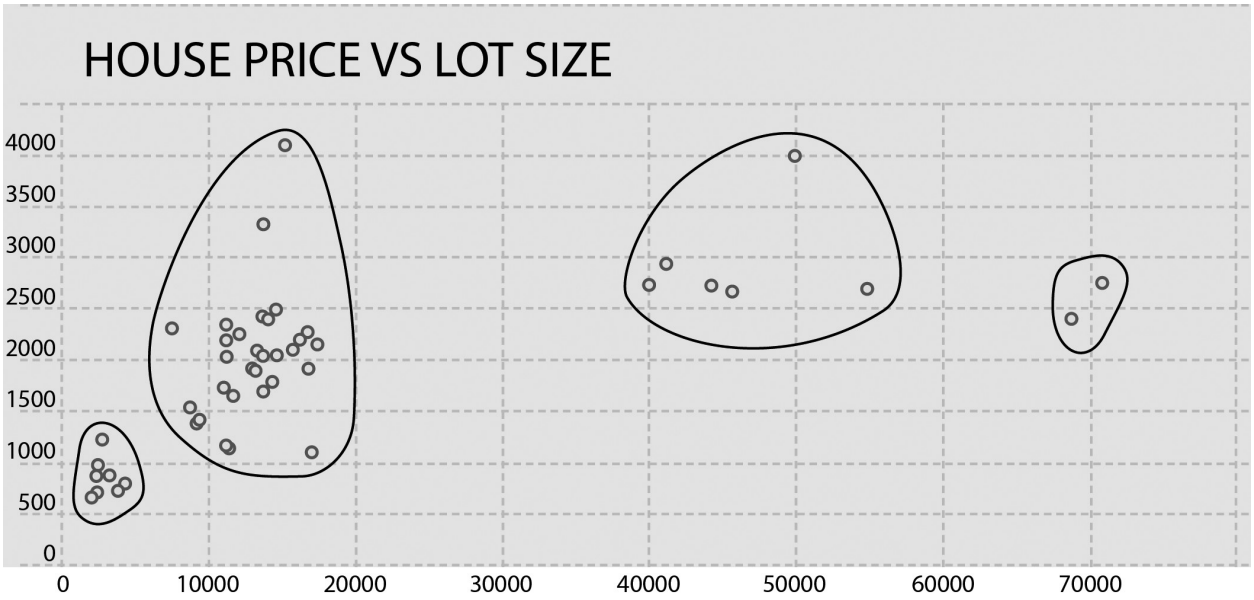

DATASET

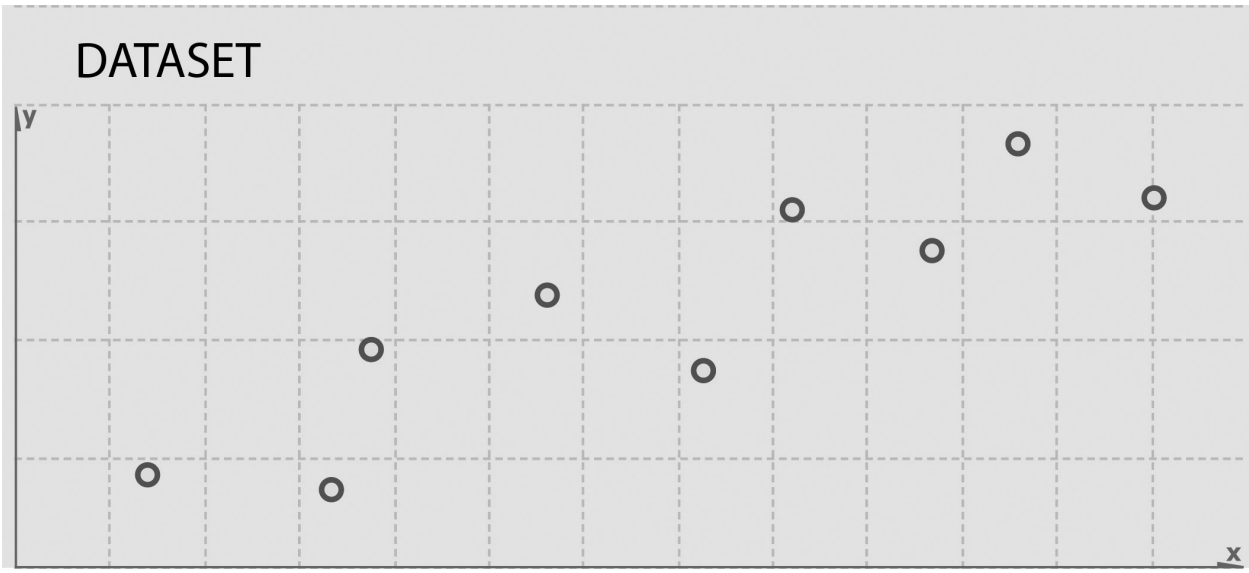

DATASET
Cluster Centroid
Cluster Centroid

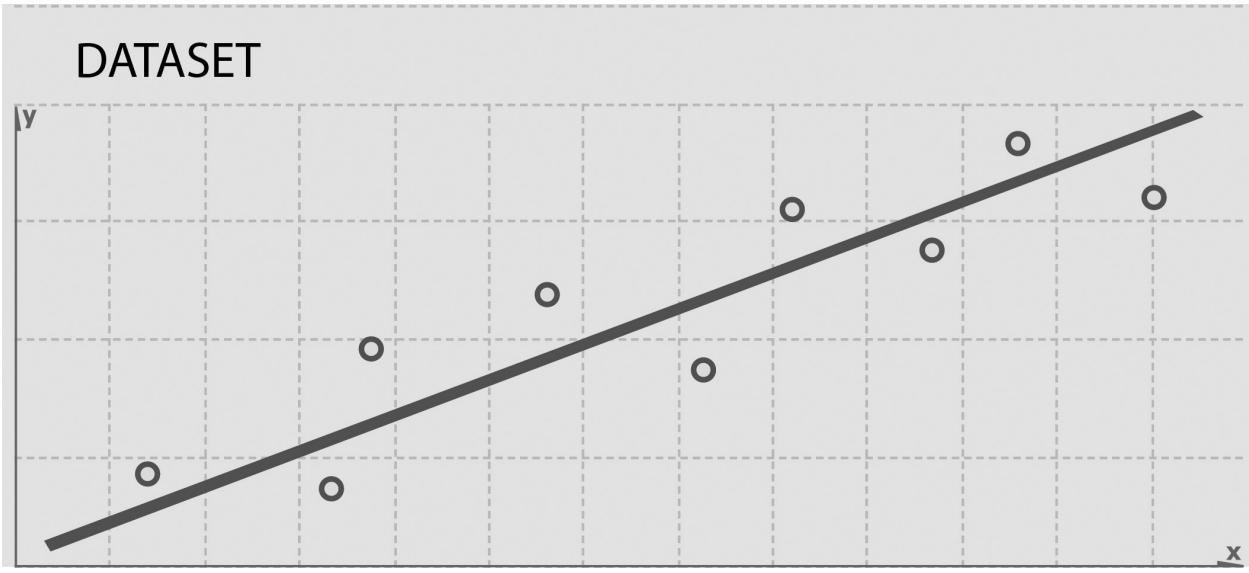

DATASET

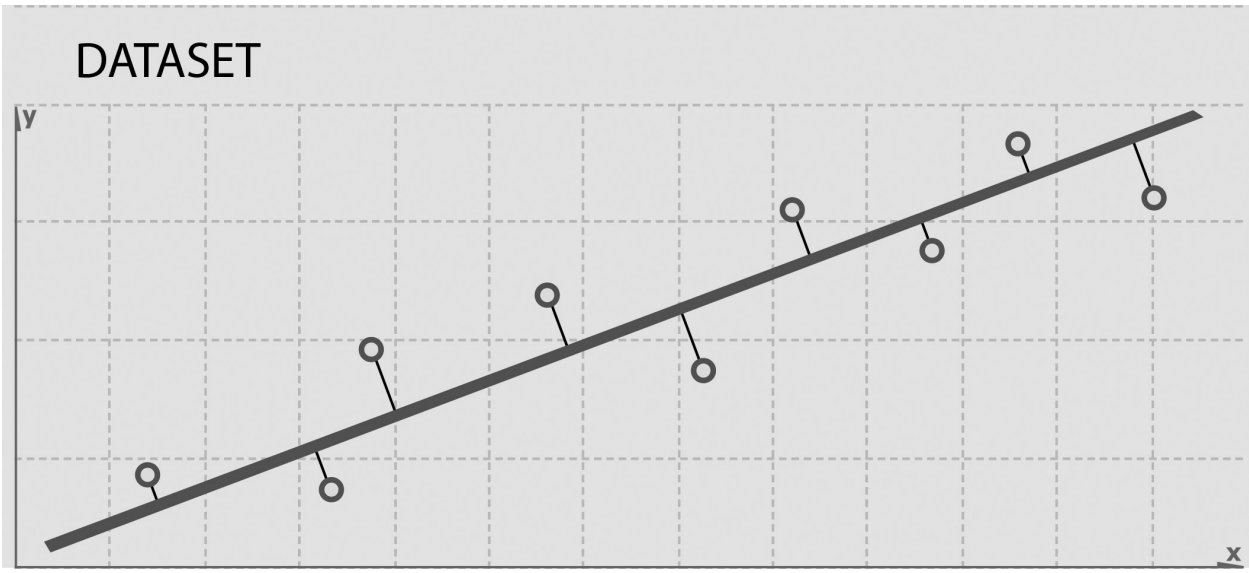

DATASET

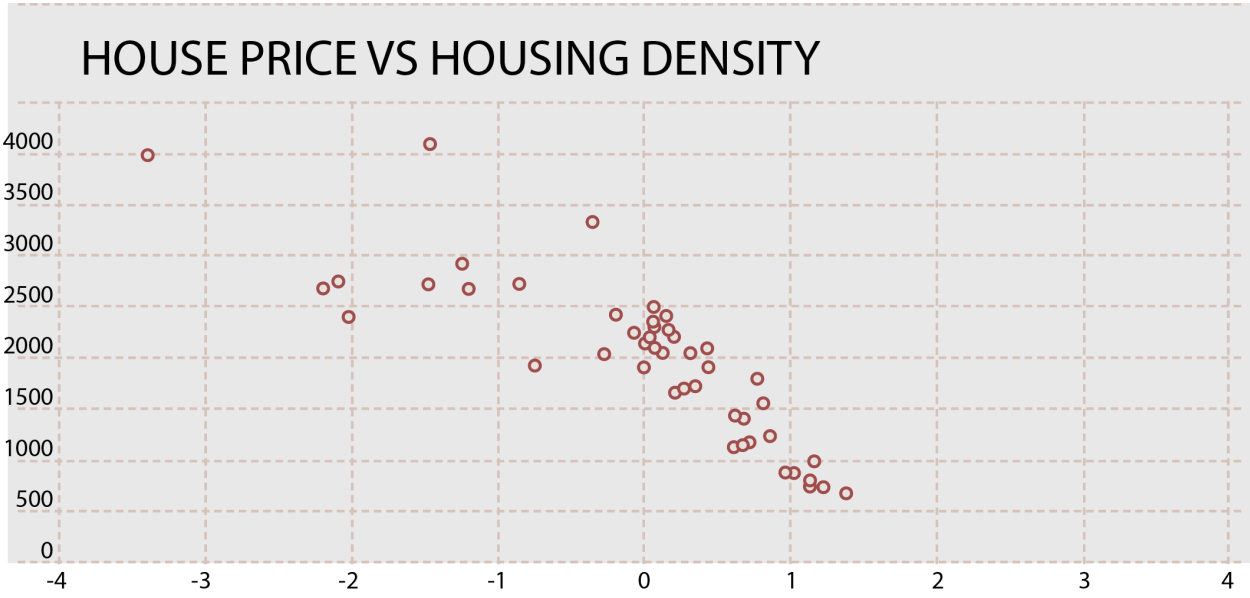

HOUSE PRICE VS LOT SIZE
4000
3500
3000
2500
2000
1500
1000
500
0
0
10000
20000
30000
40000
50000
60000
70000

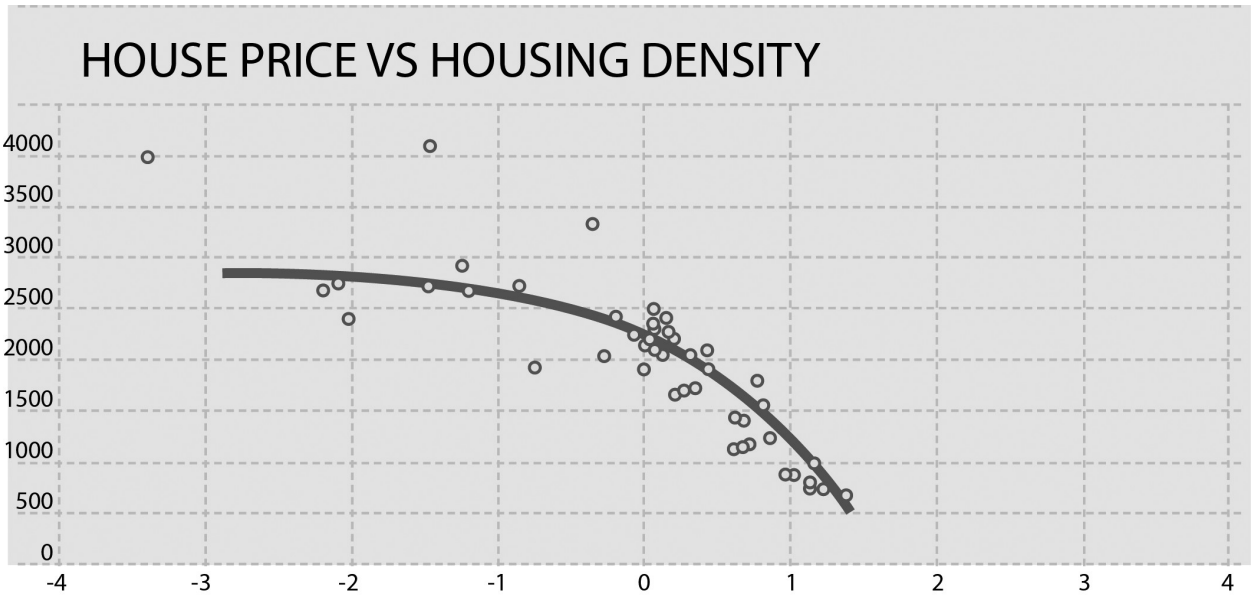

HOUSE PRICE VS LOT SIZE
4000
3500
3000
2500
2000
1500
1000
500
0
0
10000
20000
30000
40000
50000
60000
70000

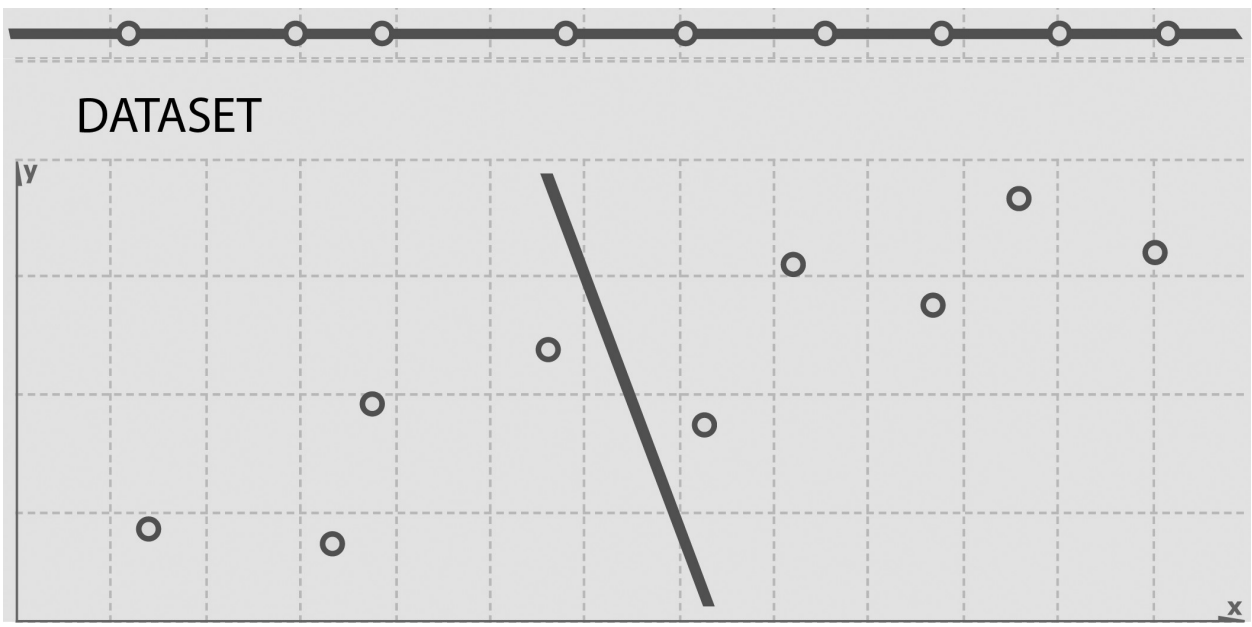

DATASET
y
x


DATASET
y
x


DATASET
y
x


HOUSE PRICE VS HOUSING DENSITY
4000
3500
3000
2500
2000
1500
1000
500
0
-4
-3
-2
-1
0
1
2
3
4





$$A = USV^T$$

$$U^T U = 1$$

$$V^T V = 1$$

$$AA^T$$

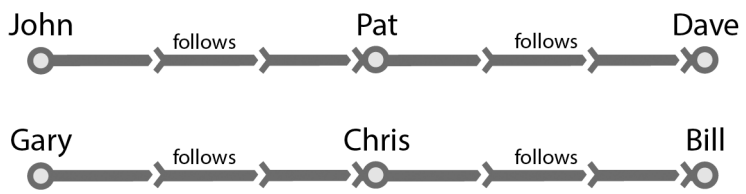
$$A^T A$$

$$npr\ \ fox$$

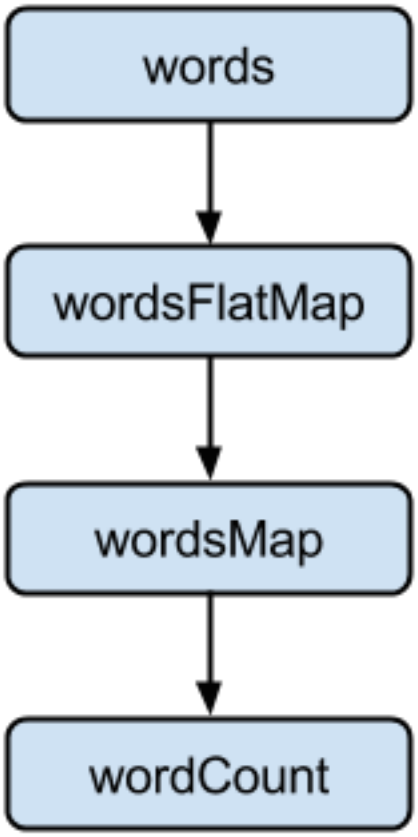
ChrisChristie	1	2
JebBush	2	3
MikeHuckabee	1	4
GeorgePataki	1	0
RickSantorum	1	0
LindseyGraham	1	3
TedCruz	1	2
ScottWalker	1	0
RickScott	1	2
HillaryClinton	0	3
MarkRubio	0	1
RickPerry	0	2

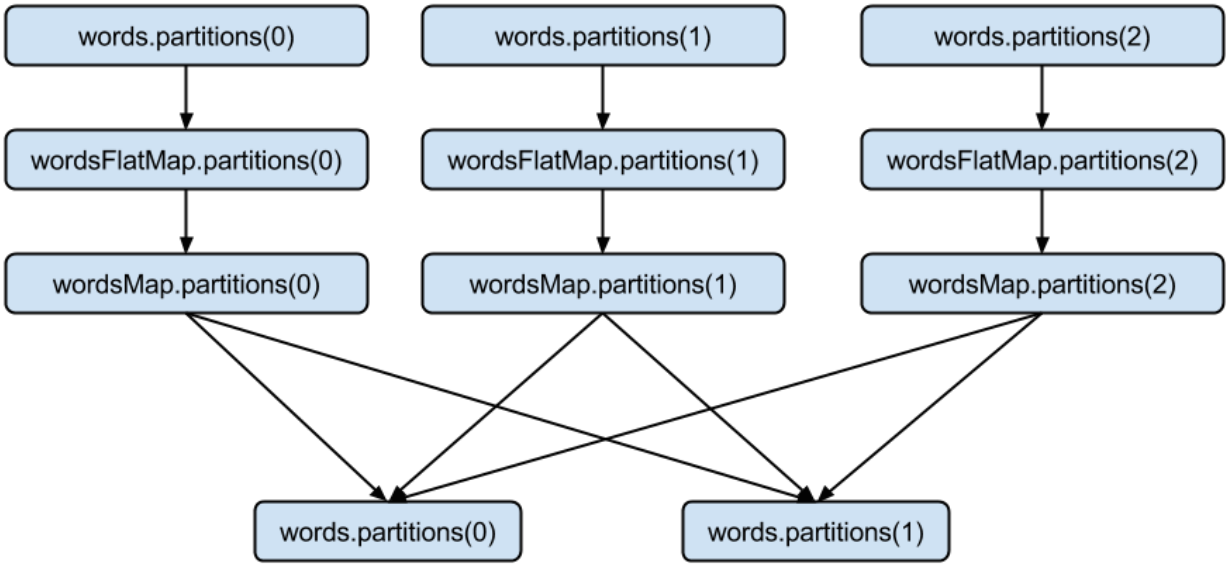
$$-1 \leq x \leq 1$$

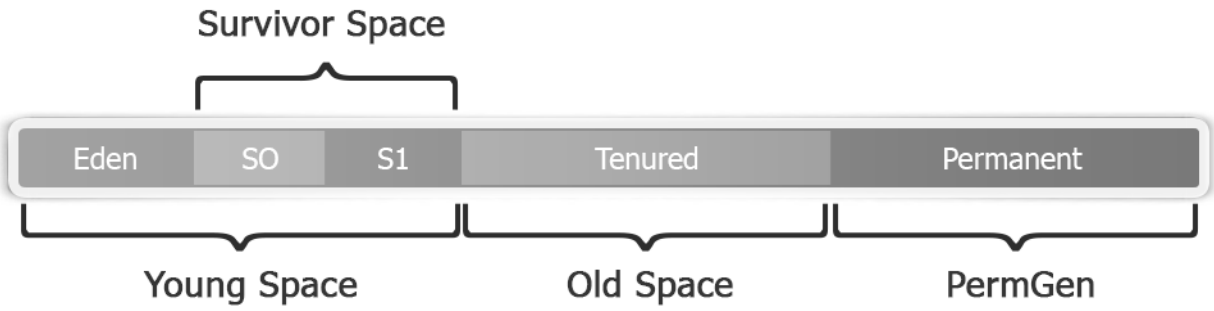
Chapter 11: Graph Processing Using GraphX

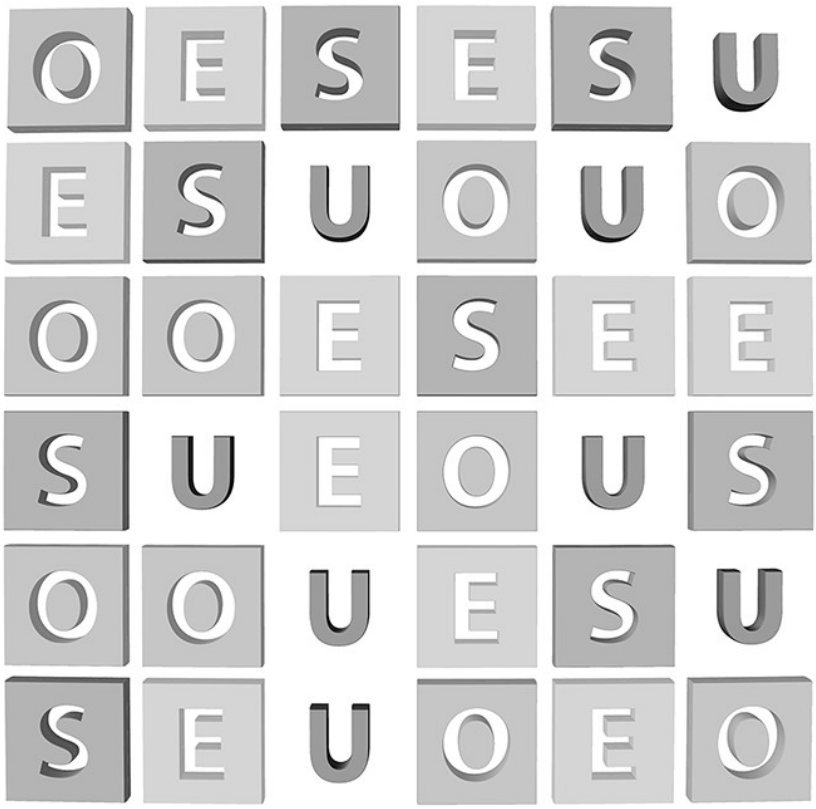


Chapter 12: Optimizations and Performance Tuning

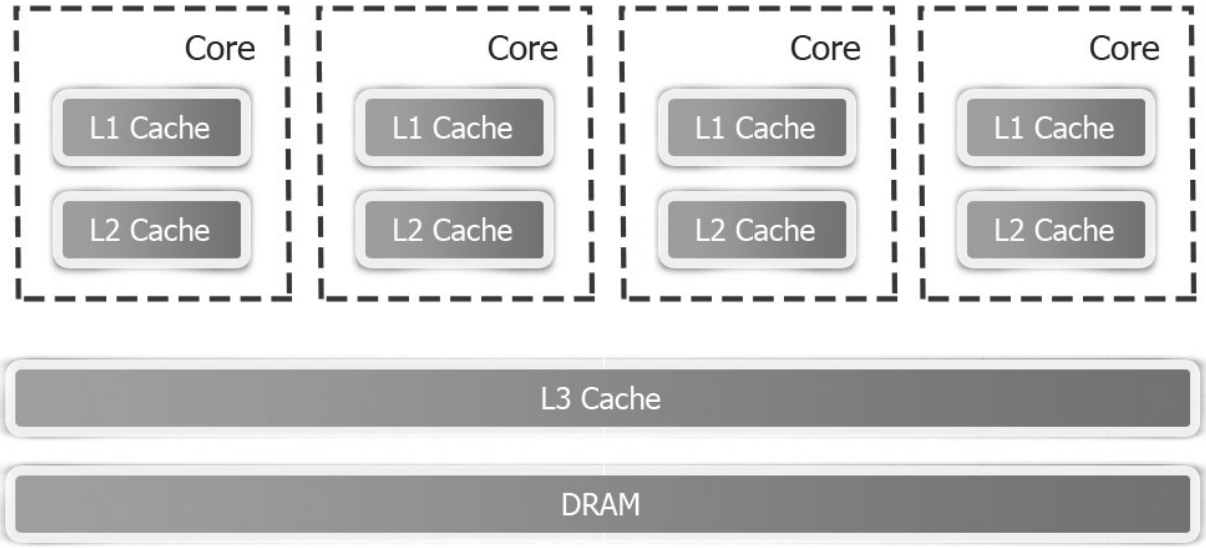

words
wordsFlatMap
wordsMap
wordCount


words.partitions(0)
words.partitions(1)
words.partitions(2)
wordsFlatMap.partitions(0)
wordsFlatMap.partitions(1)
wordsFlatMap.partitions(2)
wordsMap.partitions(0)
wordsMap.partitions(1)
wordsMap.partitions(2)
words.partitions(0)
words.partitions(1)





O = Old Generation
E = Eden Space
S = Survivor Space
U = Unused


Core
L1 Cache
L2 Cache
Core
L1 Cache
L2 Cache
Core
L1 Cache
L2 Cache
Core
L1 Cache
L2 Cache
L3 Cache
DRAM