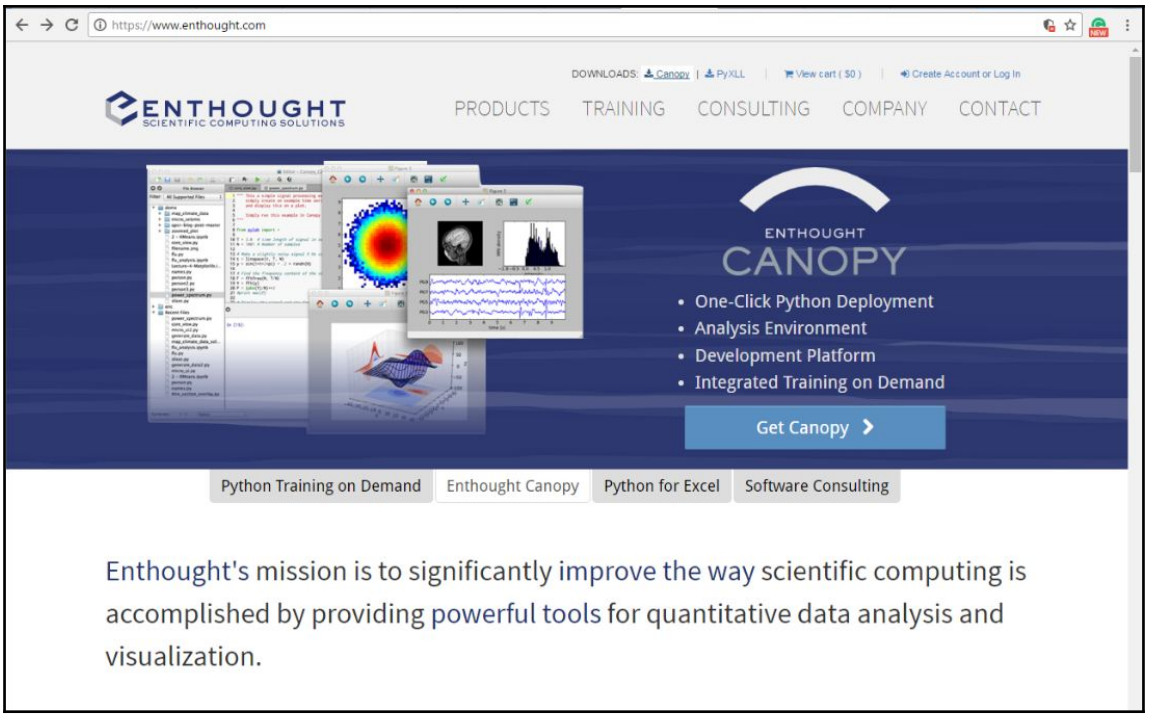
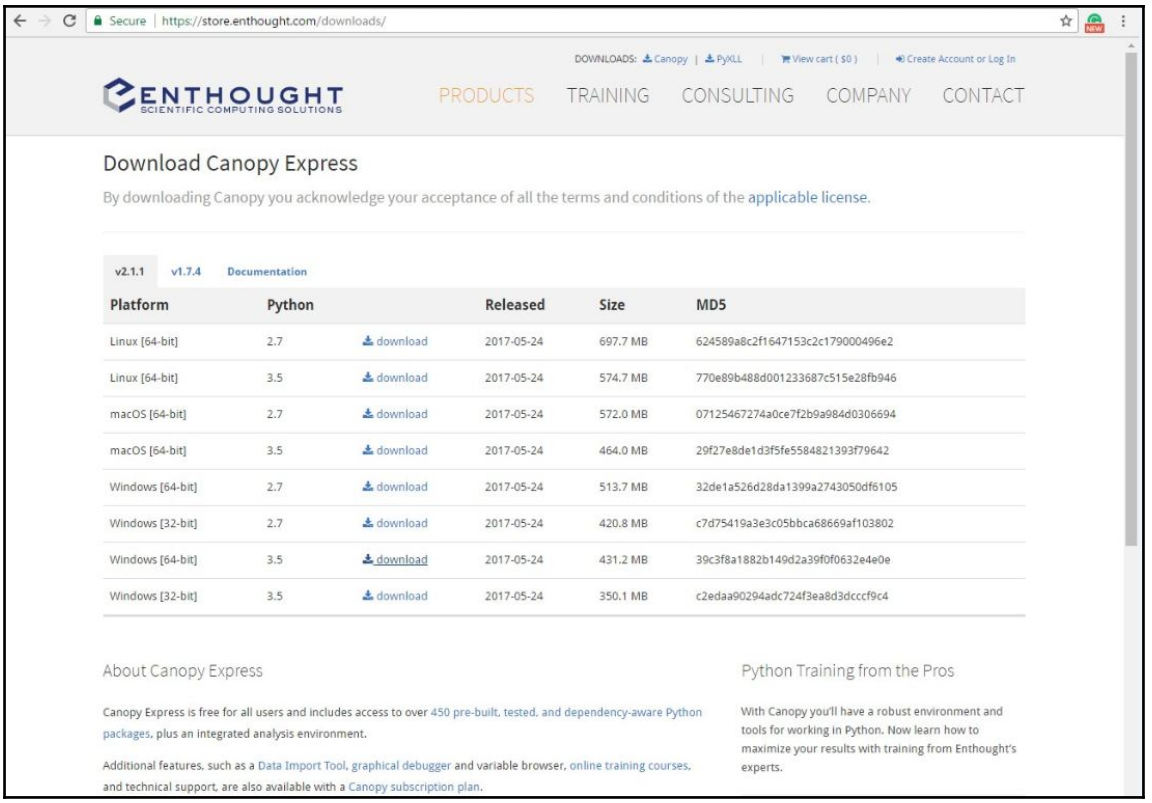
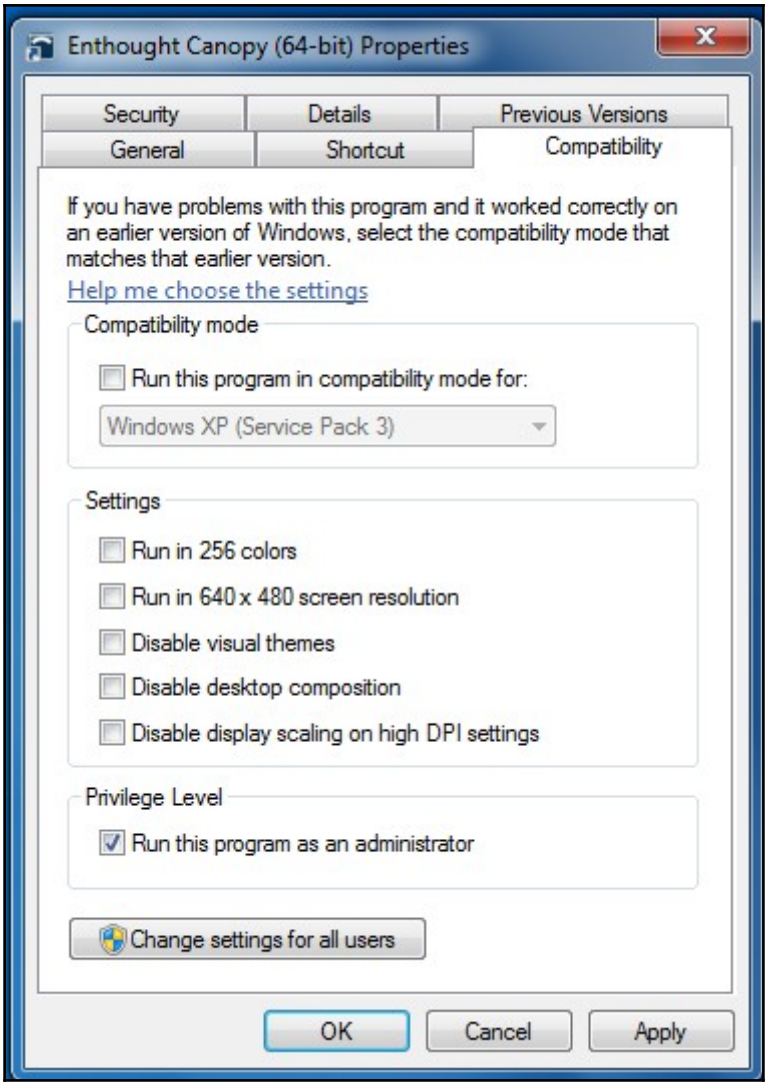
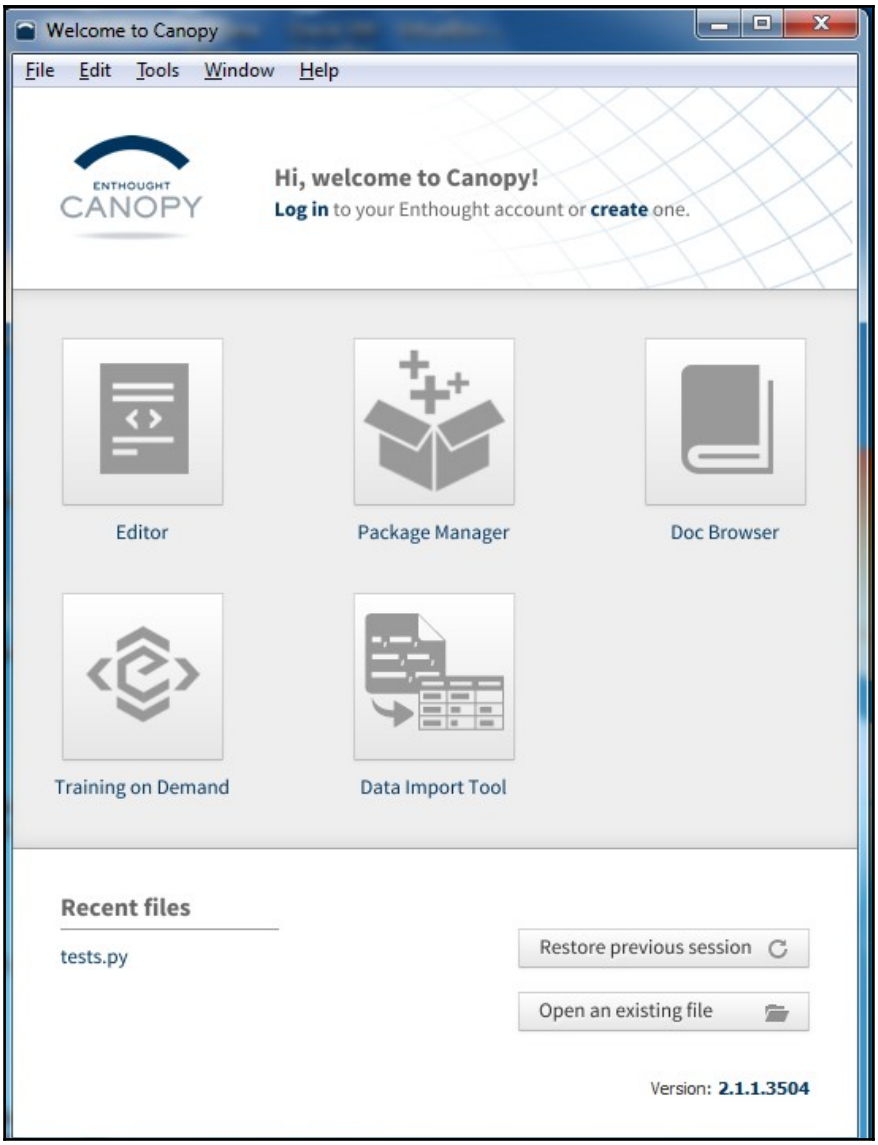
Chapter 1

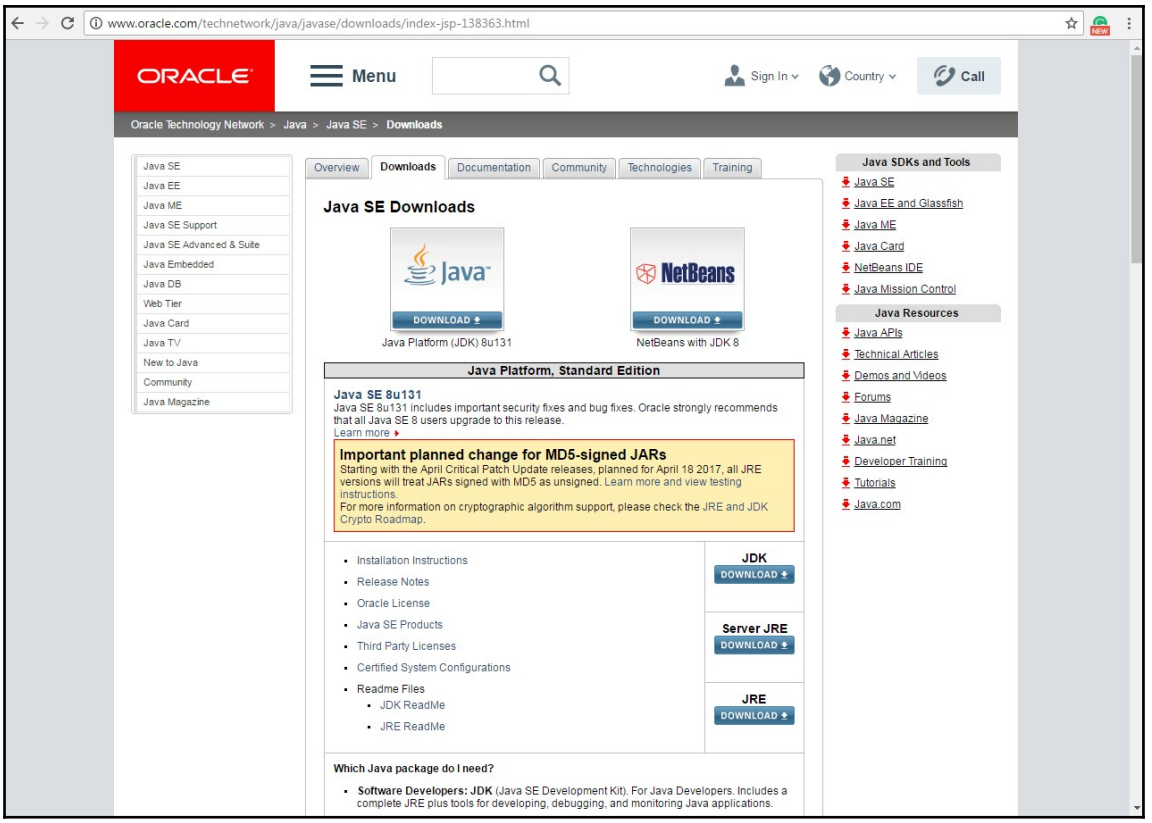

https://www.enthought.com
DOWNLOADS: Canopy | PyXLL
View cart ($0)
Create Account or Log In
ENTHOUGHT
SCIENTIFIC COMPUTING SOLUTIONS
PRODUCTS
TRAINING
CONSULTING
COMPANY
CONTACT
PyXLL
The Power of Python in Excel
Create powerful Excel add-ins
Easily deploy to others
Mitigate risk through version control
Learn More
Python Training on Demand
Enthought Canopy
Python for Excel
Software Consulting
Enthought's mission is to significantly improve the way scientific computing is accomplished by providing powerful tools for quantitative data analysis and visualization.

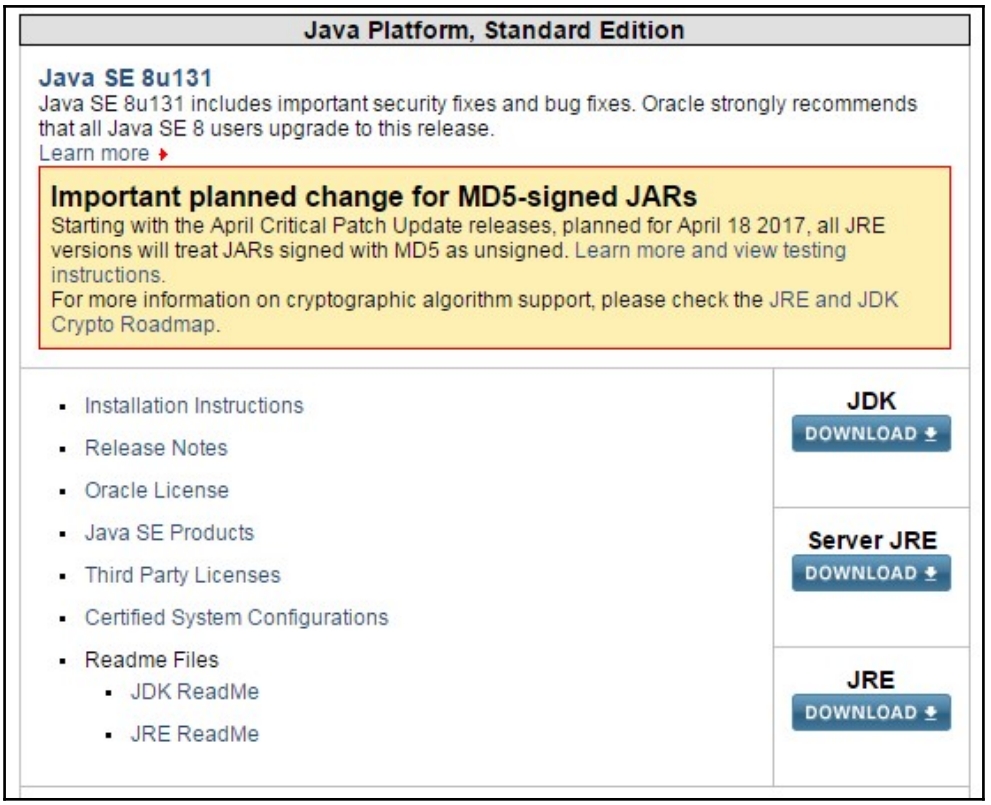

https://www.enthought.com
DOWNLOADS: Canopy | PyXLL
View cart ($0)
Create Account or Log In
ENTHOUGHT
SCIENTIFIC COMPUTING SOLUTIONS
PRODUCTS
TRAINING
CONSULTING
COMPANY
CONTACT
ENTHOUGHT
CANOPY
One-Click Python Deployment
Analysis Environment
Development Platform
Integrated Training on Demand
Get Canopy
Python Training on Demand
Enthought Canopy
Python for Excel
Software Consulting
Enthought's mission is to significantly improve the way scientific computing is accomplished by providing powerful tools for quantitative data analysis and visualization.


Secure | https://store.enthought.com/downloads/
DOWNLOADS: Canopy | PyXLL | View cart ($0) | Create Account or Log In
ENTHOUGHT
SCIENTIFIC COMPUTING SOLUTIONS
PRODUCTS TRAINING CONSULTING COMPANY CONTACT
Download Canopy Express
By downloading Canopy you acknowledge your acceptance of all the terms and conditions of the applicable license.
v2.1.1 v1.7.4 Documentation
Platform Python Released Size MD5
Linux [64-bit] 2.7 download 2017-05-24 697.7 MB 624589a8c2f1647153c2c179000496e2
Linux [64-bit] 3.5 download 2017-05-24 574.7 MB 770e89b488d001233687c515e28fb946
macOS [64-bit] 2.7 download 2017-05-24 572.0 MB 07125467274a0ce7f2b9a984d0306694
macOS [64-bit] 3.5 download 2017-05-24 464.0 MB 29f27e8de1d3f5fe5584821393f79642
Windows [64-bit] 2.7 download 2017-05-24 513.7 MB 32de1a526d28da1399a2743050df6105
Windows [32-bit] 2.7 download 2017-05-24 420.8 MB c7d75419a3e3c05bbca68669af103802
Windows [64-bit] 3.5 download 2017-05-24 431.2 MB 39c3f8a1882b149d2a39f0f0632e4e0e
Windows [32-bit] 3.5 download 2017-05-24 350.1 MB c2edaa90294adc724f3ea8d3dcccf9c4
About Canopy Express
Canopy Express is free for all users and includes access to over 450 pre-built, tested, and dependency-aware Python packages, plus an integrated analysis environment.
Additional features, such as a Data Import Tool, graphical debugger and variable browser, online training courses, and technical support, are also available with a Canopy subscription plan.
Python Training from the Pros
With Canopy you'll have a robust environment and tools for working in Python. Now learn how to maximize your results with training from Enthought's experts.


Enthought Canopy (64-bit) Properties
Security
Details
Previous Versions
General
Shortcut
Compatibility
If you have problems with this program and it worked correctly on an earlier version of Windows, select the compatibility mode that matches that earlier version.
Help me choose the settings
Compatibility mode
Run this program in compatibility mode for:
Windows XP (Service Pack 3)
Settings
Run in 256 colors
Run in 640 x 480 screen resolution
Disable visual themes
Disable desktop composition
Disable display scaling on high DPI settings
Privilege Level
Run this program as an administrator
Change settings for all users
OK
Cancel
Apply


Welcome to Canopy
File Edit Tools Window Help
ENTHOUGHT
CANOPY
Hi, welcome to Canopy!
Log in to your Enthought account or create one.
Editor
Package Manager
Doc Browser
Training on Demand
Data Import Tool
Recent files
tests.py
Restore previous session
Open an existing file
Version: 2.1.1.3504

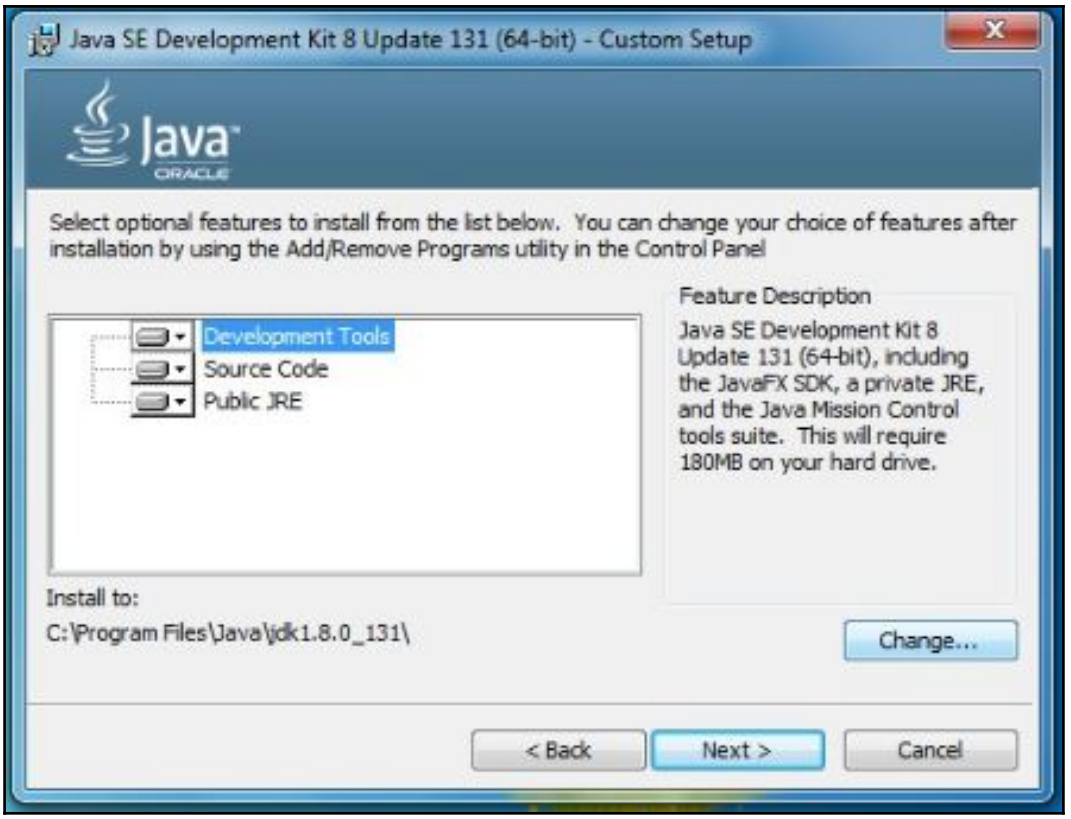

www.oracle.com/technetwork/java/javase/downloads/index-jsp-138363.html
ORACLE
Menu
Sign In
Country
Call
Oracle Technology Network > Java > Java SE > Downloads
Java SE
Java EE
Java ME
Java SE Support
Java SE Advanced & Suite
Java Embedded
Java DB
Web Tier
Java Card
Java TV
New to Java
Community
Java Magazine
Overview
Downloads
Documentation
Community
Technologies
Training
Java SE Downloads
Java
DOWNLOAD
Java Platform (JDK) 8u131
NetBeans
DOWNLOAD
NetBeans with JDK 8
Java Platform, Standard Edition
Java SE 8u131
Java SE 8u131 includes important security fixes and bug fixes. Oracle strongly recommends that all Java SE 8 users upgrade to this release.
Learn more
Important planned change for MD5-signed JARs
Starting with the April Critical Patch Update releases, planned for April 18 2017, all JRE versions will treat JARs signed with MD5 as unsigned. Learn more and view testing instructions.
For more information on cryptographic algorithm support, please check the JRE and JDK Crypto Roadmap.
Installation Instructions
Release Notes
Oracle License
Java SE Products
Third Party Licenses
Certified System Configurations
Readme Files
JDK ReadMe
JRE ReadMe
JDK
DOWNLOAD
Server JRE
DOWNLOAD
JRE
DOWNLOAD
Which Java package do I need?
Software Developers: JDK (Java SE Development Kit). For Java Developers. Includes a complete JRE plus tools for developing, debugging, and monitoring Java applications.
Java SDKs and Tools
Java SE
Java EE and Glassfish
Java ME
Java Card
NetBeans IDE
Java Mission Control
Java Resources
Java APIs
Technical Articles
Demos and Videos
Forums
Java Magazine
Java.net
Developer Training
Tutorials
Java.com

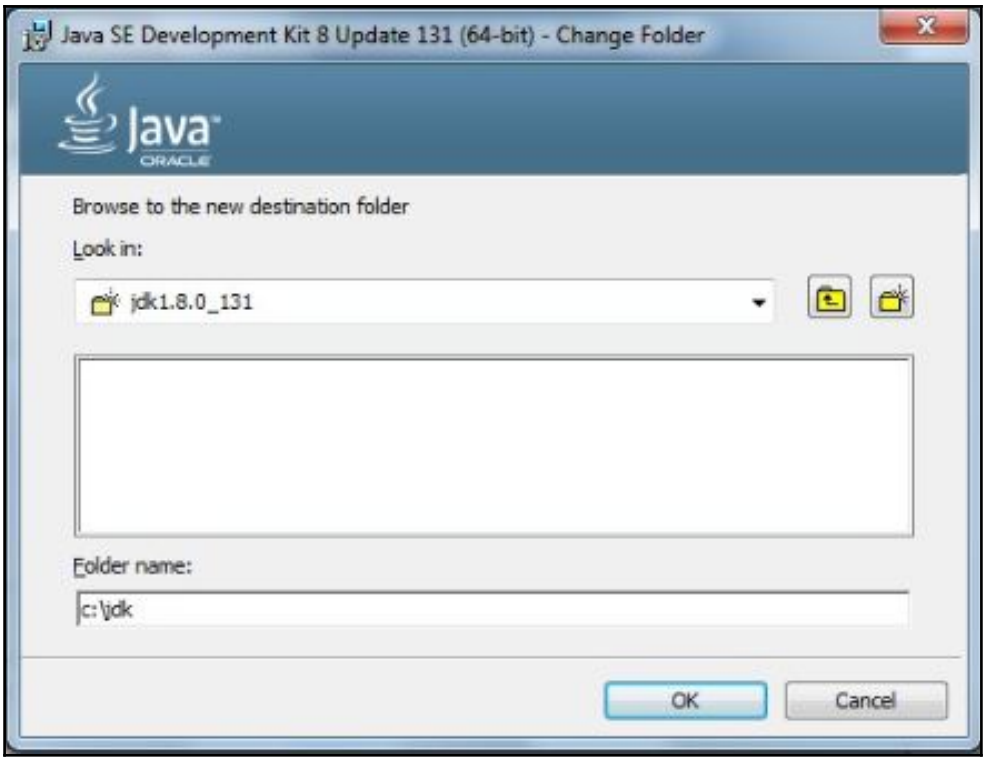

Java Platform, Standard Edition
Java SE 8u131
Java SE 8u131 includes important security fixes and bug fixes. Oracle strongly recommends that all Java SE 8 users upgrade to this release.
Learn more
Important planned change for MD5-signed JARs
Starting with the April Critical Patch Update releases, planned for April 18 2017, all JRE versions will treat JARs signed with MD5 as unsigned. Learn more and view testing instructions.
For more information on cryptographic algorithm support, please check the JRE and JDK Crypto Roadmap.
Installation Instructions
Release Notes
Oracle License
Java SE Products
Third Party Licenses
Certified System Configurations
Readme Files
JDK ReadMe
JRE ReadMe
JDK
DOWNLOAD
Server JRE
DOWNLOAD
JRE
DOWNLOAD

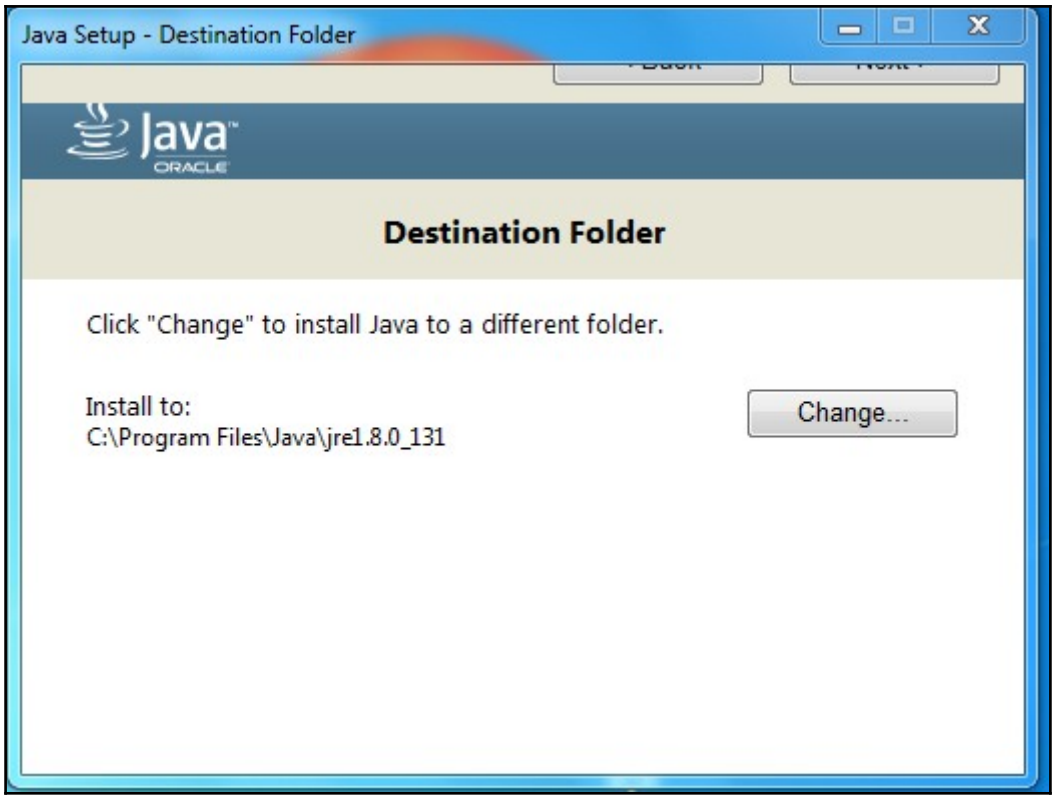
Java SE Development Kit 8u131

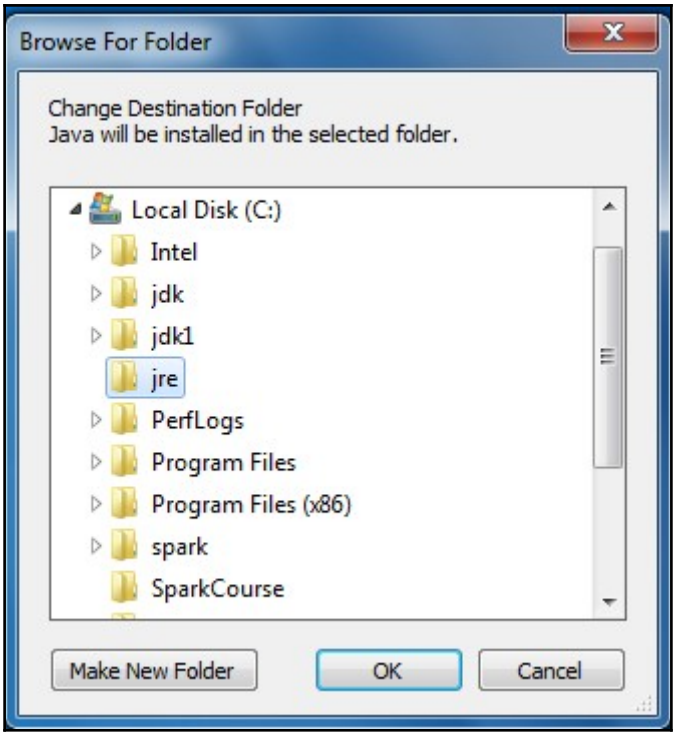
You must accept the Oracle Binary Code License Agreement for Java SE to download this software.
Thank you for accepting the Oracle Binary Code License Agreement for Java SE; you may now download this software.

Product / File Description	File Size	Download
Linux ARM 32 Hard Float ABI	77.87 MB	jdk-8u131-linux-arm32-vfp-hflt.tar.gz
Linux ARM 64 Hard Float ABI	74.81 MB	jdk-8u131-linux-arm64-vfp-hflt.tar.gz
Linux x86	164.66 MB	jdk-8u131-linux-i586.rpm
Linux x86	179.39 MB	jdk-8u131-linux-i586.tar.gz
Linux x64	162.11 MB	jdk-8u131-linux-x64.rpm
Linux x64	176.95 MB	jdk-8u131-linux-x64.tar.gz
Mac OS X	226.57 MB	jdk-8u131-macosx-x64.dmg
Solaris SPARC 64-bit	139.79 MB	jdk-8u131-solaris-sparcv9.tar.Z
Solaris SPARC 64-bit	99.13 MB	jdk-8u131-solaris-sparcv9.tar.gz
Solaris x64	140.51 MB	jdk-8u131-solaris-x64.tar.Z
Solaris x64	96.96 MB	jdk-8u131-solaris-x64.tar.gz
Windows x86	191.22 MB	jdk-8u131-windows-i586.exe
Windows x64	198.03 MB	jdk-8u131-windows-x64.exe


Java SE Development Kit 8 Update 131 (64-bit) - Custom Setup
Java
ORACLE
Select optional features to install from the list below. You can change your choice of features after installation by using the Add/Remove Programs utility in the Control Panel
Development Tools
Source Code
Public JRE
Feature Description
Java SE Development Kit 8 Update 131 (64-bit), including the JavaFX SDK, a private JRE, and the Java Mission Control tools suite. This will require 180MB on your hard drive.
Install to:
C:\Program Files\Java\jdk1.8.0_131\
Change...
< Back
Next >
Cancel


Java SE Development Kit 8 Update 131 (64-bit) - Change Folder
Java
ORACLE
Browse to the new destination folder
Look in:
jdk1.8.0_131
Folder name:
c:\jdk
OK
Cancel


Java Setup - Destination Folder
Java
ORACLE
Destination Folder
Click "Change" to install Java to a different folder.
Install to:
C:\Program Files\Java\jre1.8.0_131
Change...


Browse For Folder
Change Destination Folder
Java will be installed in the selected folder.
Local Disk (C:)
Intel
jdk
jdk1
jre
PerfLogs
Program Files
Program Files (x86)
spark
SparkCourse
Make New Folder
OK
Cancel

spark.apache.org

Apache Spark

Lightning-fast cluster computing

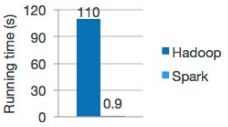
Download | Libraries | Documentation | Examples | Community | Developers | Apache Software Foundation

Apache Spark™ is a fast and general engine for large-scale data processing.

Speed

Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Apache Spark has an advanced DAG execution engine that supports acyclic data flow and in-memory computing.



Logistic regression in Hadoop and Spark

Ease of Use

Write applications quickly in Java, Scala, Python, R.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it *interactively* from the Scala, Python and R shells.

```
text_file = spark.textFile("hdfs://...")

text_file.flatMap(lambda line: line.split())
    .map(lambda word: (word, 1))
    .reduceByKey(lambda a, b: a+b)
```

Word count in Spark's Python API

Latest News

Spark 2.1.1 released (May 02, 2017)

Spark Summit (June 5-7th, 2017, San Francisco) agenda posted (Mar 31, 2017)

Spark Summit East (Feb 7-9th, 2017, Boston) agenda posted (Jan 04, 2017)

Spark 2.1.0 released (Dec 28, 2016)

Archive

Download Spark

Built-in Libraries:

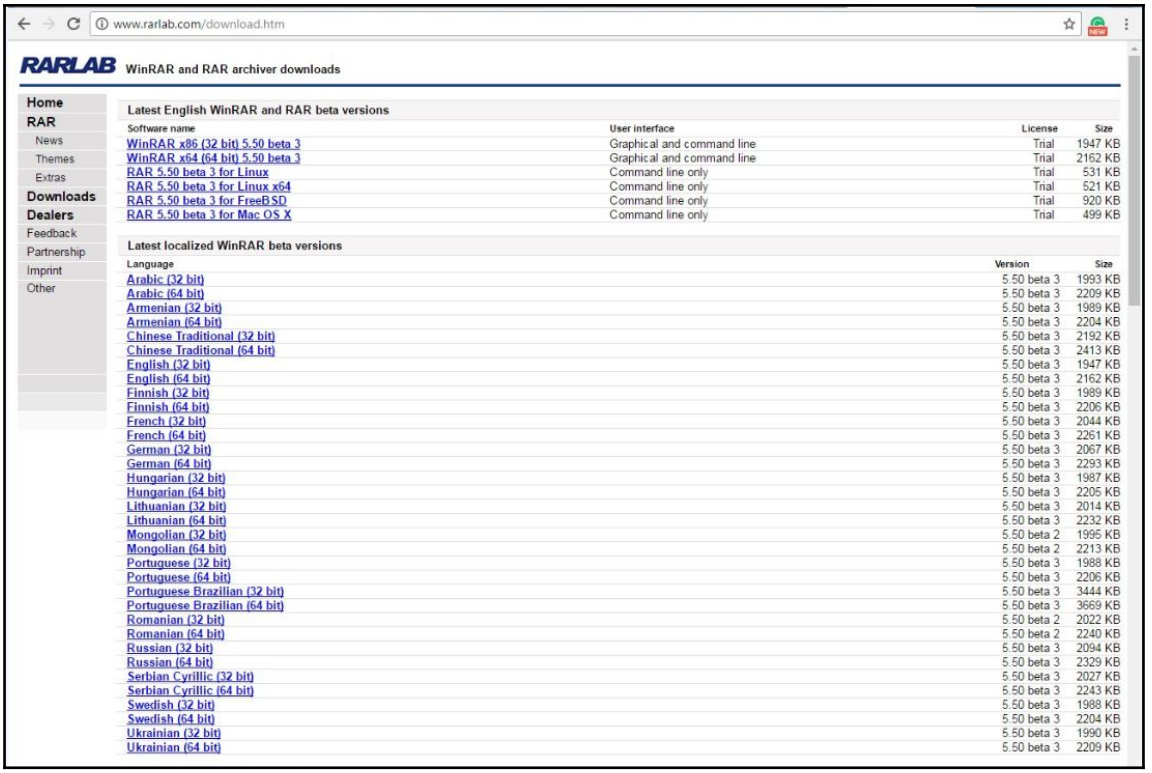
SQL and DataFrames
Spark Streaming
MLlib (machine learning)
GraphX (graph)

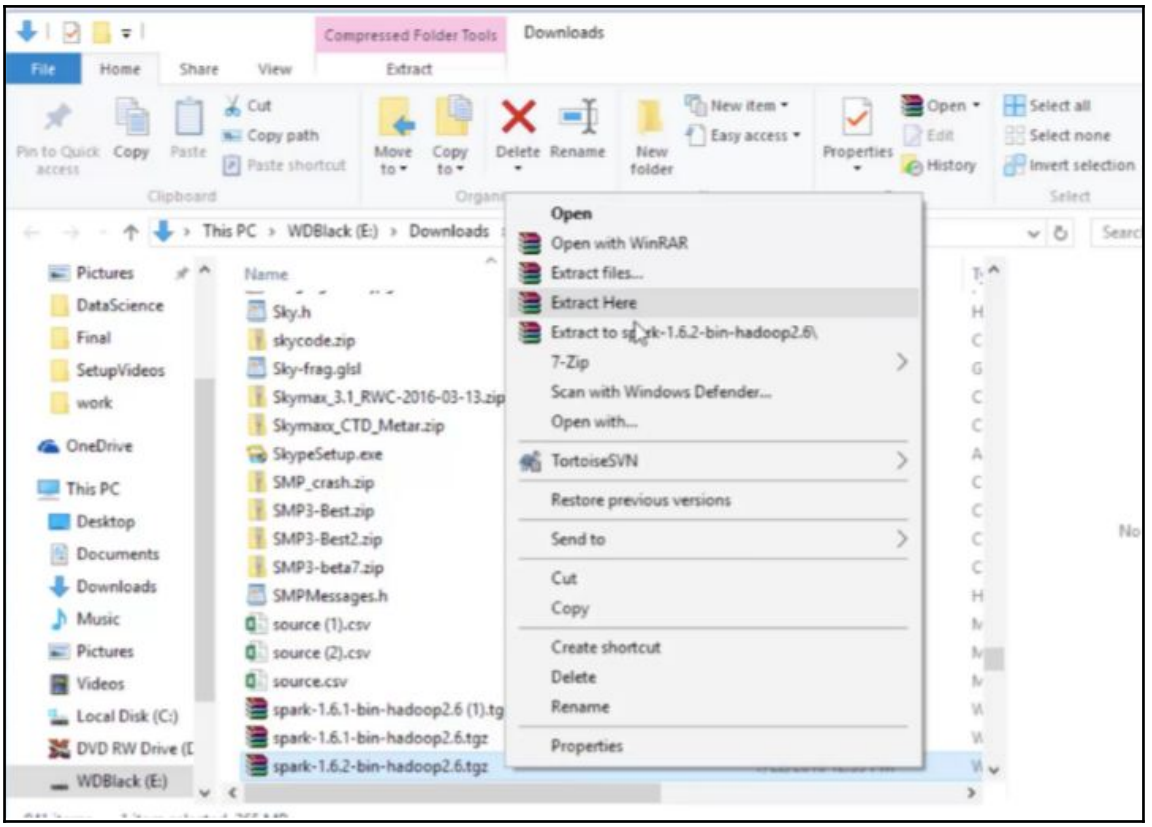
Third-Party Projects

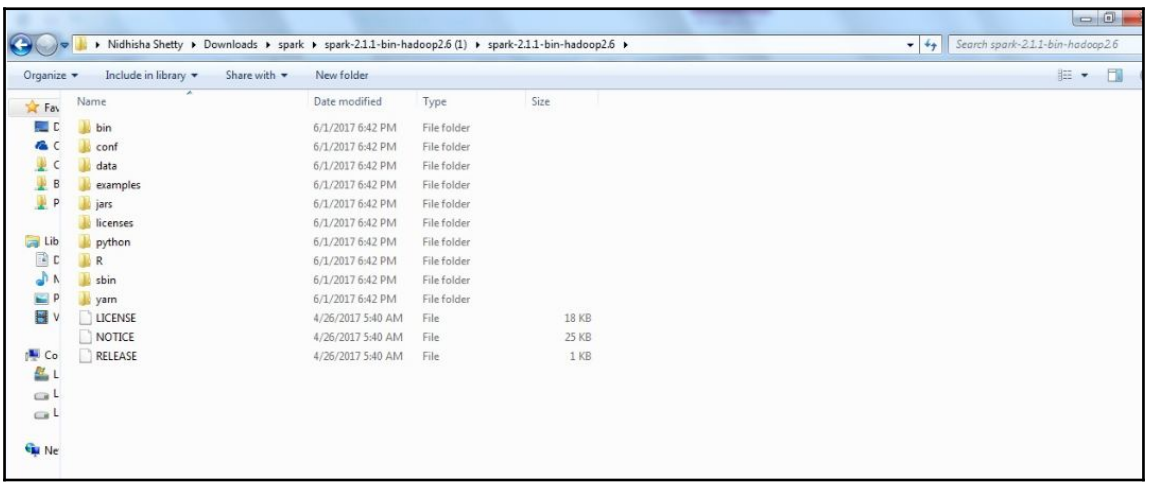
Download Apache Spark™

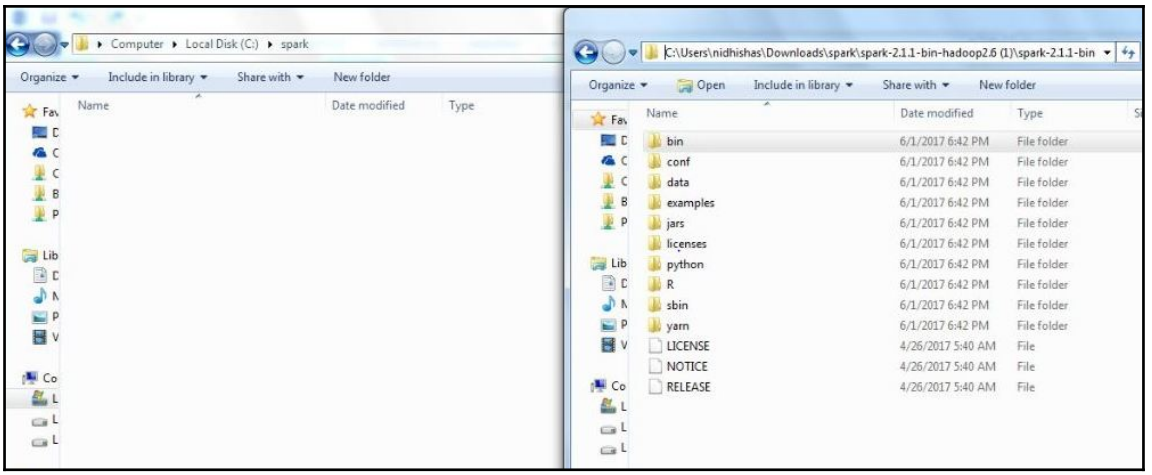
1. Choose a Spark release: 2.1.1 (May 02 2017)
2. Choose a package type: Pre-built for Apache Hadoop 2.7 and later
3. Choose a download type: Direct Download
4. Download Spark: spark-2.1.1-bin-hadoop2.7.tgz
5. Verify this release using the 2.1.1 signatures and checksums and project release KEYS.

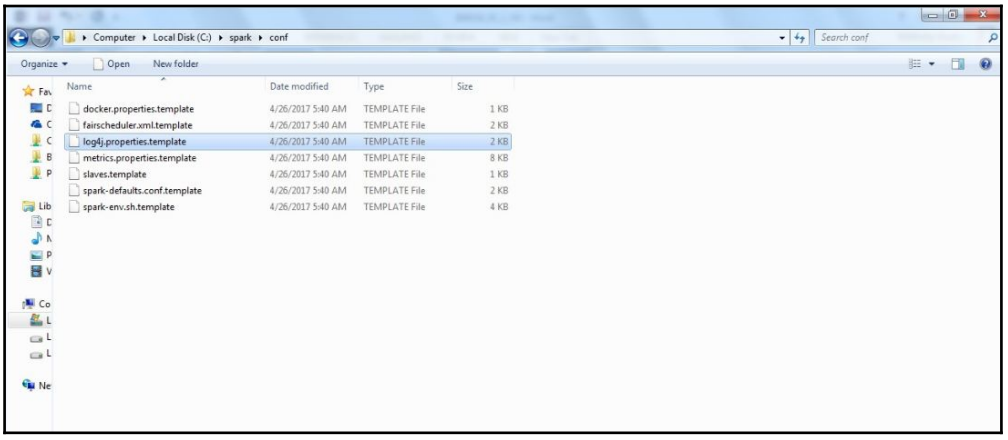
Note: Starting version 2.0, Spark is built with Scala 2.11 by default. Scala 2.10 users should download the Spark source package and build with Scala 2.10 support.

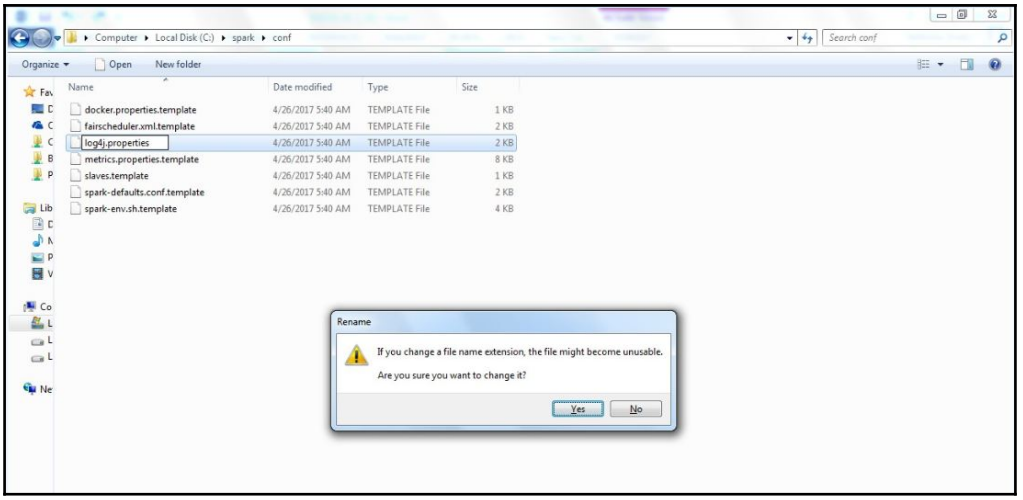

www.rarlab.com/download.htm
RARLAB
WinRAR and RAR archiver downloads
Home
RAR
News
Themes
Extras
Downloads
Dealers
Feedback
Partnership
Imprint
Other
Latest English WinRAR and RAR beta versions
Software name
User interface
License
Size
WinRAR x86 (32 bit) 5.50 beta 3
Graphical and command line
Trial
1947 KB
WinRAR x64 (64 bit) 5.50 beta 3
Graphical and command line
Trial
2162 KB
RAR 5.50 beta 3 for Linux
Command line only
Trial
531 KB
RAR 5.50 beta 3 for Linux x64
Command line only
Trial
521 KB
RAR 5.50 beta 3 for FreeBSD
Command line only
Trial
920 KB
RAR 5.50 beta 3 for Mac OS X
Command line only
Trial
499 KB
Latest localized WinRAR beta versions
Language
Version
Size
Arabic (32 bit) 5.50 beta 3 1993 KB
Arabic (64 bit) 5.50 beta 3 2209 KB
Armenian (32 bit) 5.50 beta 3 1989 KB
Armenian (64 bit) 5.50 beta 3 2204 KB
Chinese Traditional (32 bit) 5.50 beta 3 2192 KB
Chinese Traditional (64 bit) 5.50 beta 3 2413 KB
English (32 bit) 5.50 beta 3 1947 KB
English (64 bit) 5.50 beta 3 2162 KB
Finnish (32 bit) 5.50 beta 3 1989 KB
Finnish (64 bit) 5.50 beta 3 2206 KB
French (32 bit) 5.50 beta 3 2044 KB
French (64 bit) 5.50 beta 3 2261 KB
German (32 bit) 5.50 beta 3 2067 KB
German (64 bit) 5.50 beta 3 2293 KB
Hungarian (32 bit) 5.50 beta 3 1987 KB
Hungarian (64 bit) 5.50 beta 3 2205 KB
Lithuanian (32 bit) 5.50 beta 3 2014 KB
Lithuanian (64 bit) 5.50 beta 3 2232 KB
Mongolian (32 bit) 5.50 beta 2 1995 KB
Mongolian (64 bit) 5.50 beta 2 2213 KB
Portuguese (32 bit) 5.50 beta 3 1988 KB
Portuguese (64 bit) 5.50 beta 3 2206 KB
Portuguese Brazilian (32 bit) 5.50 beta 3 3444 KB
Portuguese Brazilian (64 bit) 5.50 beta 3 3669 KB
Romanian (32 bit) 5.50 beta 2 2022 KB
Romanian (64 bit) 5.50 beta 2 2240 KB
Russian (32 bit) 5.50 beta 3 2094 KB
Russian (64 bit) 5.50 beta 3 2329 KB
Serbian Cyrillic (32 bit) 5.50 beta 3 2027 KB
Serbian Cyrillic (64 bit) 5.50 beta 3 2243 KB
Swedish (32 bit) 5.50 beta 3 1988 KB
Swedish (64 bit) 5.50 beta 3 2204 KB
Ukrainian (32 bit) 5.50 beta 3 1990 KB
Ukrainian (64 bit) 5.50 beta 3 2209 KB


Compressed Folder Tools
Downloads
File
Home
Share
View
Extract
Cut
Copy path
Paste shortcut
Pin to Quick access
Copy
Paste
Clipboard
Move to
Copy to
Delete
Rename
New folder
New item
Easy access
Properties
Open
Edit
History
Select all
Select none
Invert selection
Select
This PC › WDBlack (E:) › Downloads
Pictures
DataScience
Final
SetupVideos
work
OneDrive
This PC
Desktop
Documents
Downloads
Music
Pictures
Videos
Local Disk (C:)
DVD RW Drive (D
WDBlack (E:)
Name
Sky.h
skycode.zip
Sky-frag.glsl
Skymax_3.1_RWC-2016-03-13.zip
Skymaxx_CTD_Metar.zip
SkypeSetup.exe
SMP_crash.zip
SMP3-Best.zip
SMP3-Best2.zip
SMP3-beta7.zip
SMPMessages.h
source (1).csv
source (2).csv
source.csv
spark-1.6.1-bin-hadoop2.6 (1).tg
spark-1.6.1-bin-hadoop2.6.tgz
spark-1.6.2-bin-hadoop2.6.tgz
Open
Open with WinRAR
Extract files...
Extract Here
Extract to sp k-1.6.2-bin-hadoop2.6\
7-Zip
Scan with Windows Defender...
Open with...
TortoiseSVN
Restore previous versions
Send to
Cut
Copy
Create shortcut
Delete
Rename
Properties

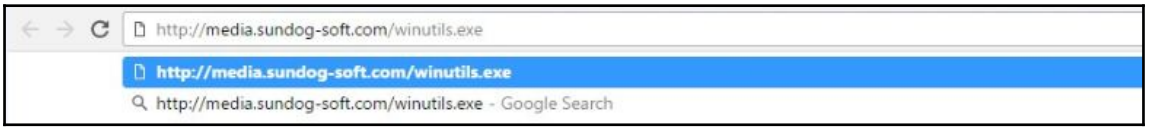

Nidhisha Shetty › Downloads › spark › spark-2.1.1-bin-hadoop2.6 (1) › spark-2.1.1-bin-hadoop2.6 ›
Search spark-2.1.1-bin-hadoop2.6
Organize
Include in library
Share with
New folder
Name
Date modified
Type
Size
bin 6/1/2017 6:42 PM File folder
conf 6/1/2017 6:42 PM File folder
data 6/1/2017 6:42 PM File folder
examples 6/1/2017 6:42 PM File folder
jars 6/1/2017 6:42 PM File folder
licenses 6/1/2017 6:42 PM File folder
python 6/1/2017 6:42 PM File folder
R 6/1/2017 6:42 PM File folder
sbin 6/1/2017 6:42 PM File folder
yarn 6/1/2017 6:42 PM File folder
LICENSE 4/26/2017 5:40 AM File 18 KB
NOTICE 4/26/2017 5:40 AM File 25 KB
RELEASE 4/26/2017 5:40 AM File 1 KB

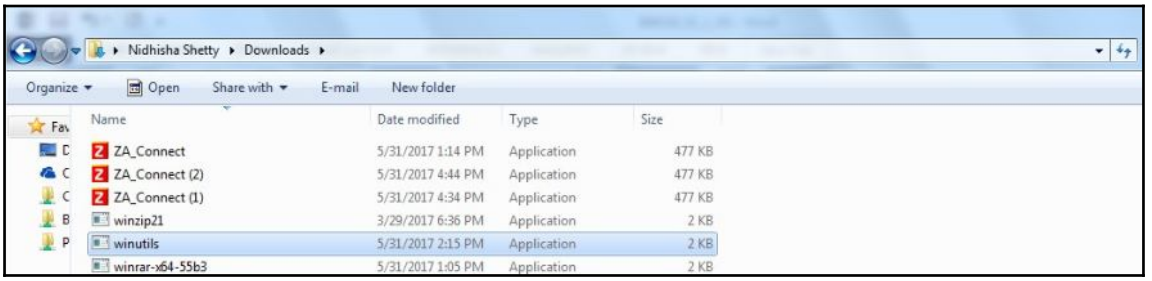

Computer ▸ Local Disk (C:) ▸ spark
Organize
Include in library
Share with
New folder
Name
Date modified
Type
C:\Users\nidhishas\Downloads\spark\spark-2.1.1-bin-hadoop2.6 (1)\spark-2.1.1-bin
Open
bin 6/1/2017 6:42 PM File folder
conf 6/1/2017 6:42 PM File folder
data 6/1/2017 6:42 PM File folder
examples 6/1/2017 6:42 PM File folder
jars 6/1/2017 6:42 PM File folder
licenses 6/1/2017 6:42 PM File folder
python 6/1/2017 6:42 PM File folder
R 6/1/2017 6:42 PM File folder
sbin 6/1/2017 6:42 PM File folder
yarn 6/1/2017 6:42 PM File folder
LICENSE 4/26/2017 5:40 AM File
NOTICE 4/26/2017 5:40 AM File
RELEASE 4/26/2017 5:40 AM File

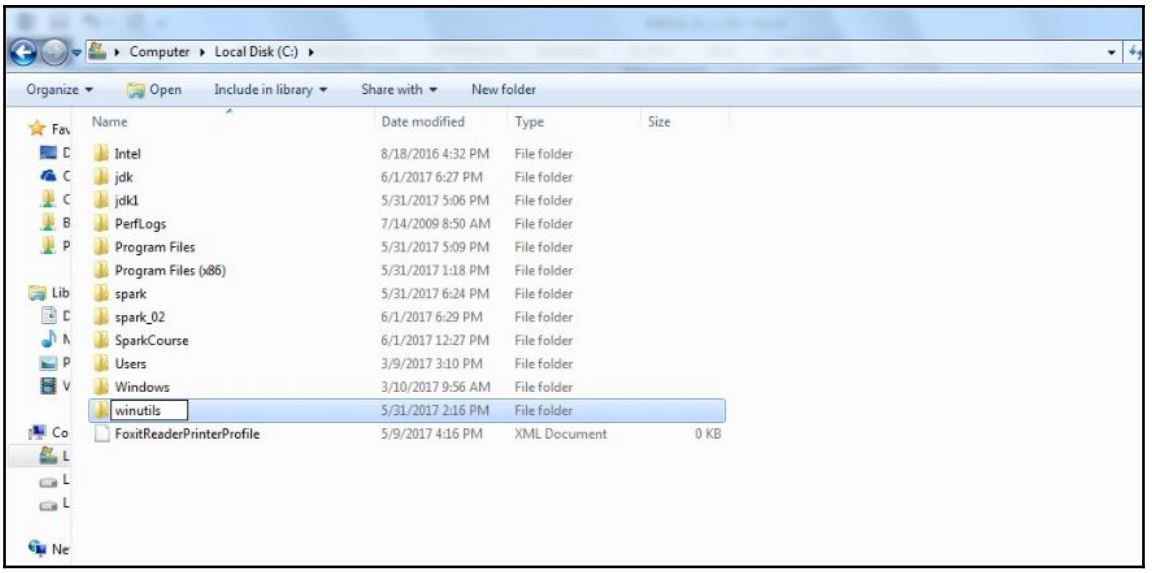

Computer ▸ Local Disk (C:) ▸ spark ▸ conf
Search conf
Organize
Open
New folder
Name
Date modified
Type
Size
docker.properties.template 4/26/2017 5:40 AM TEMPLATE File 1 KB
fairscheduler.xml.template 4/26/2017 5:40 AM TEMPLATE File 2 KB
log4j.properties.template 4/26/2017 5:40 AM TEMPLATE File 2 KB
metrics.properties.template 4/26/2017 5:40 AM TEMPLATE File 8 KB
slaves.template 4/26/2017 5:40 AM TEMPLATE File 1 KB
spark-defaults.conf.template 4/26/2017 5:40 AM TEMPLATE File 2 KB
spark-env.sh.template 4/26/2017 5:40 AM TEMPLATE File 4 KB

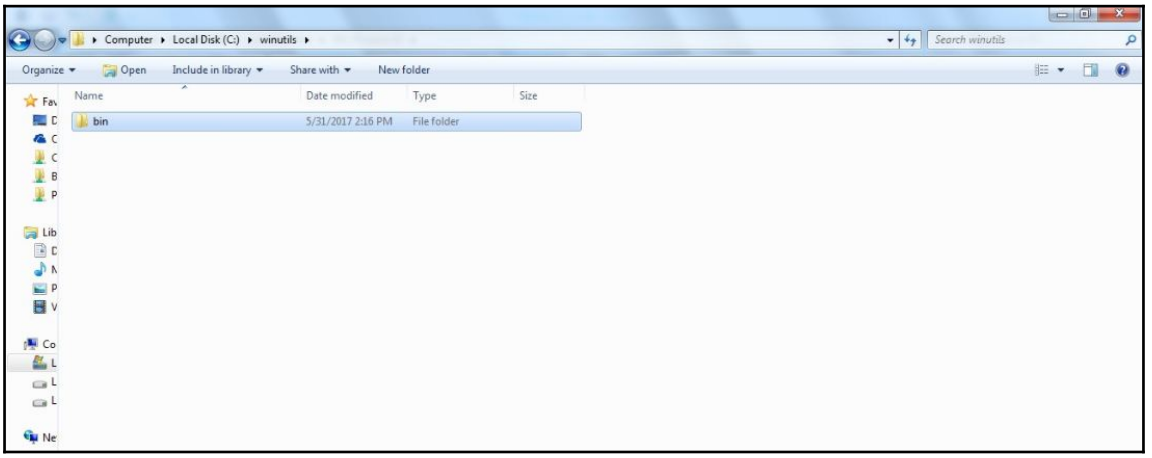


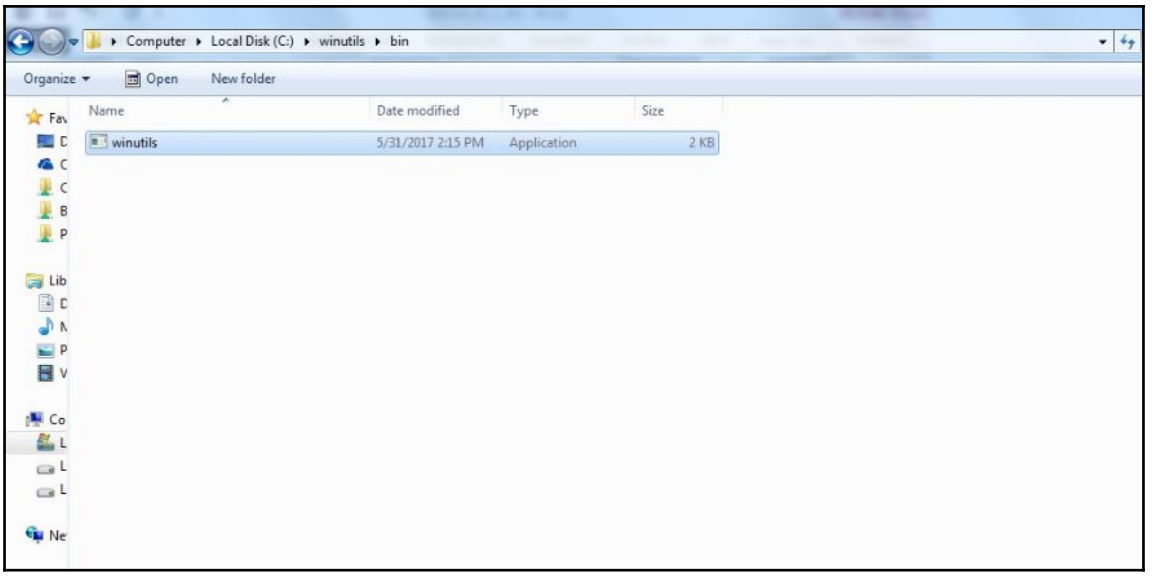

```
# Set everything to be logged to the console
log4j.rootCategory=INFO, console
log4j.appender.console=org.apache.log4j.ConsoleAppender
log4j.appender.console.target=System.err
log4j.appender.console.layout=org.apache.log4j.PatternLayout
log4j.appender.console.layout.ConversionPattern=%d{yy/MM/dd HH:mm:ss}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

```

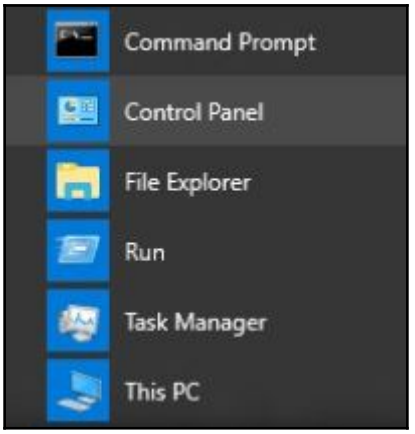




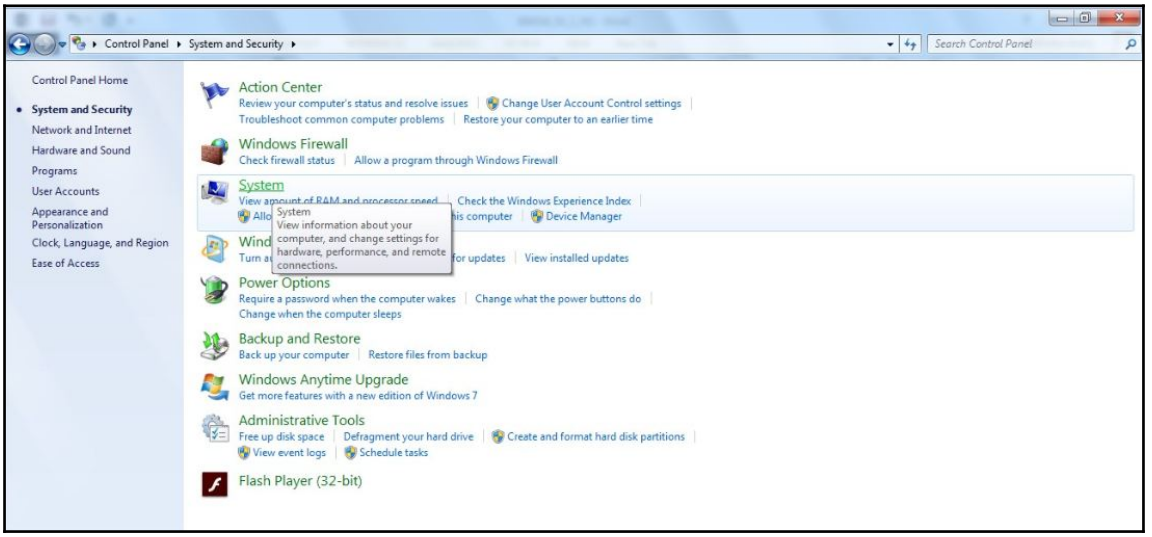

Computer ▸ Local Disk (C:) ▸
Organize
Open
Include in library
Share with
New folder
Name
Date modified
Type
Size
Intel 8/18/2016 4:32 PM File folder
jdk 6/1/2017 6:27 PM File folder
jdk1 5/31/2017 5:06 PM File folder
PerfLogs 7/14/2009 8:50 AM File folder
Program Files 5/31/2017 5:09 PM File folder
Program Files (x86) 5/31/2017 1:18 PM File folder
spark 5/31/2017 6:24 PM File folder
spark_02 6/1/2017 6:29 PM File folder
SparkCourse 6/1/2017 12:27 PM File folder
Users 3/9/2017 3:10 PM File folder
Windows 3/10/2017 9:56 AM File folder
winutils 5/31/2017 2:16 PM File folder
FoxitReaderPrinterProfile 5/9/2017 4:16 PM XML Document 0 KB

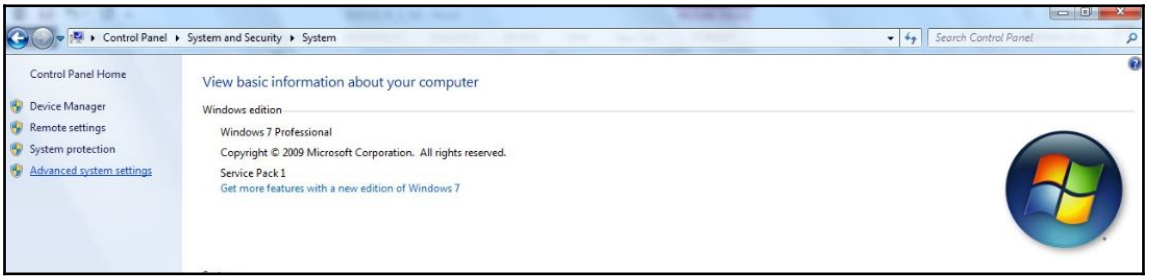

Computer ▸ Local Disk (C:) ▸ winutils ▸
Search winutils
Organize
Open
Include in library
Share with
New folder
Name
Date modified
Type
Size
bin 5/31/2017 2:16 PM File folder

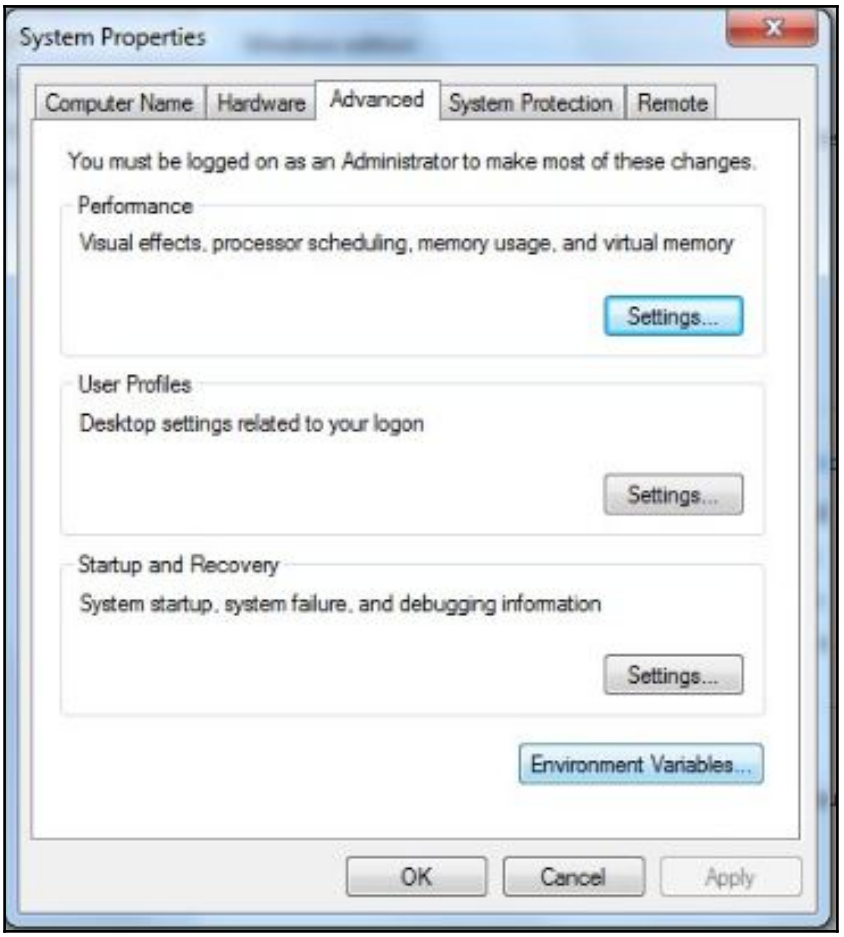

Computer ▸ Local Disk (C:) ▸ winutils ▸ bin
Organize
Open
New folder
Name
Date modified
Type
Size
winutils
5/31/2017 2:15 PM
Application
2 KB

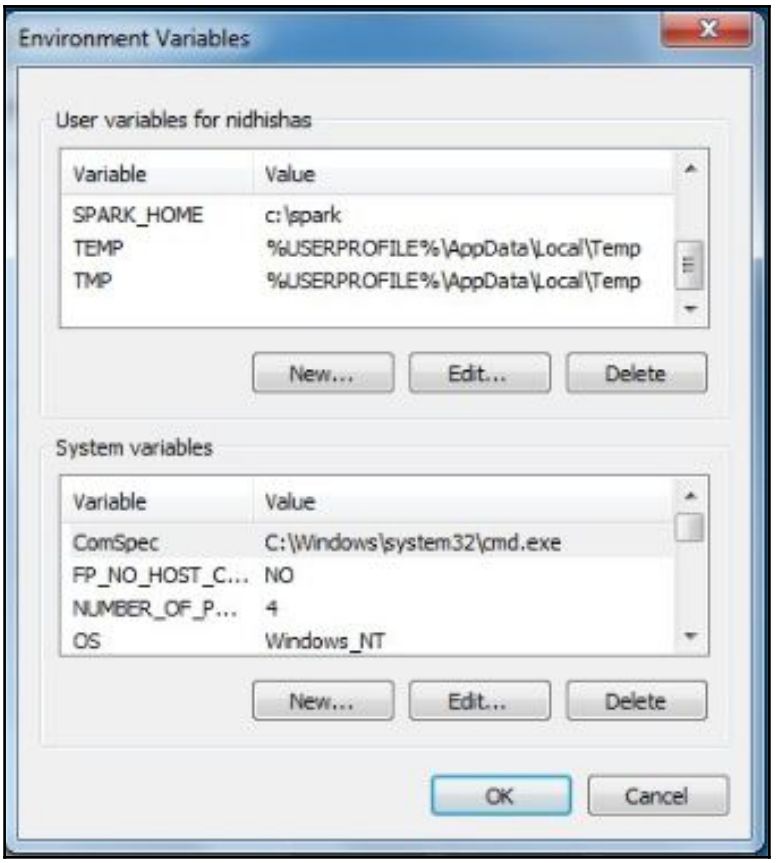

Command Prompt
Control Panel
File Explorer
Run
Task Manager
This PC

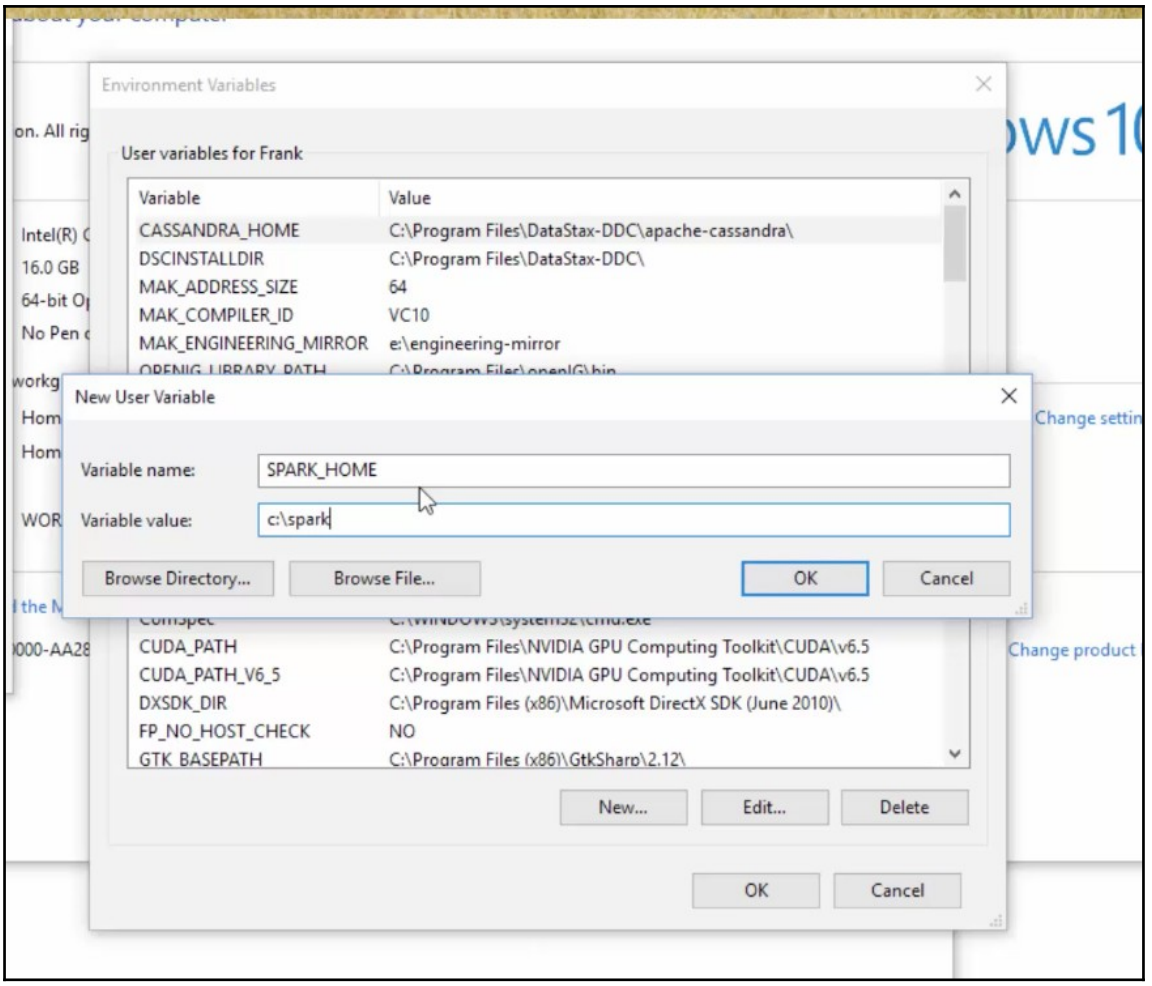

Control Panel
Search Control Panel
Adjust your computer's settings
View by: Category
System and Security
Review your computer's status
Back up your compute
Find and fix problems
System and Security
View and change system and security status, back up and restore file and system settings, update your computer, view RAM and processor speed, check firewall, and more.
Network and Inte
View network status ar
Choose homegroup ar
Hardware and Sound
View devices and printers
Add a device
Programs
Uninstall a program
Get programs
User Accounts
Change account type
Appearance and Personalization
Change the theme
Change desktop background
Adjust screen resolution
Clock, Language, and Region
Change keyboards or other input methods
Ease of Access
Let Windows suggest settings
Optimize visual display

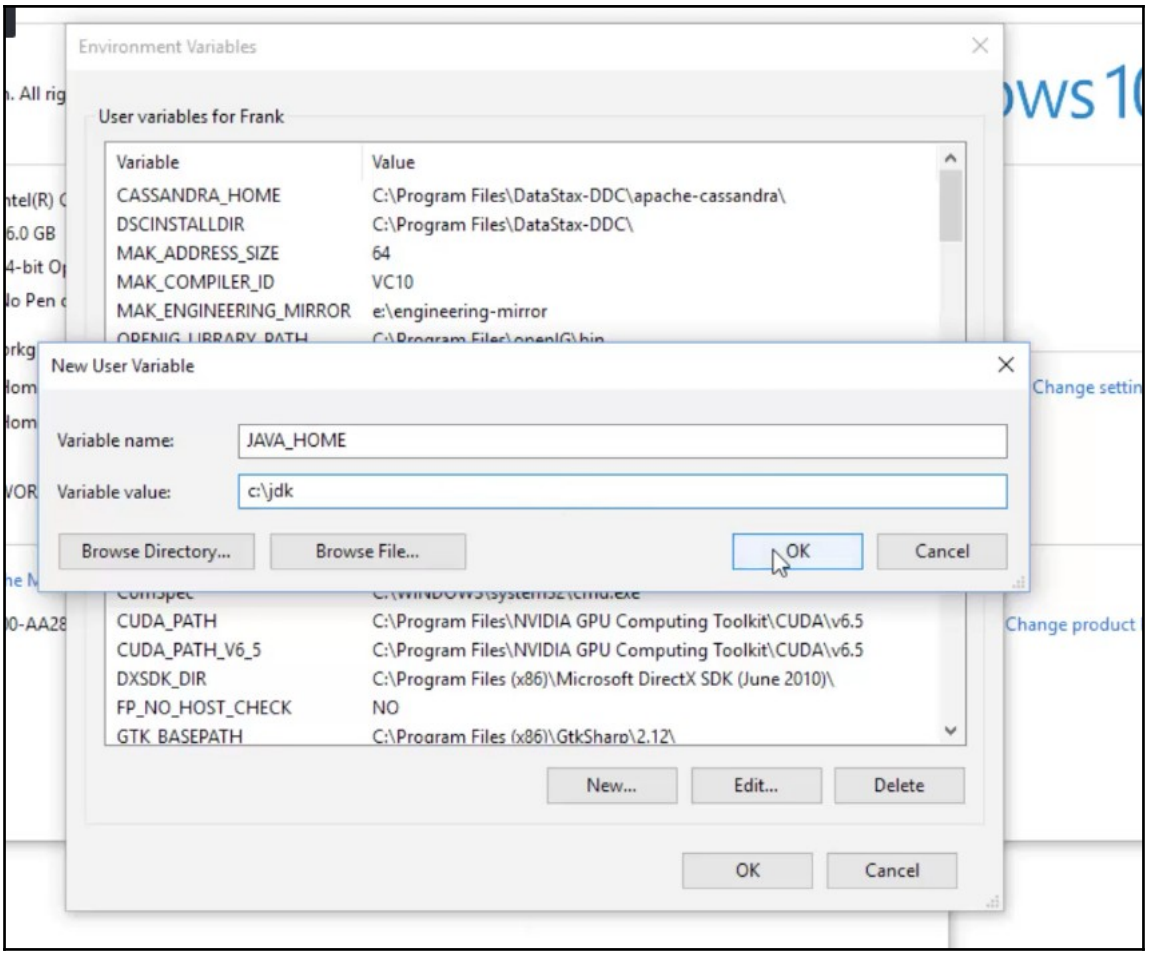

Control Panel
System and Security
Search Control Panel
Control Panel Home
System and Security
Network and Internet
Hardware and Sound
Programs
User Accounts
Appearance and Personalization
Clock, Language, and Region
Ease of Access
Action Center
Review your computer's status and resolve issues
Change User Account Control settings
Troubleshoot common computer problems
Restore your computer to an earlier time
Windows Firewall
Check firewall status
Allow a program through Windows Firewall
System
Check the Windows Experience Index
Device Manager
System
View information about your computer, and change settings for hardware, performance, and remote connections.
View installed updates
Power Options
Require a password when the computer wakes
Change what the power buttons do
Change when the computer sleeps
Backup and Restore
Back up your computer
Restore files from backup
Windows Anytime Upgrade
Get more features with a new edition of Windows 7
Administrative Tools
Free up disk space
Defragment your hard drive
Create and format hard disk partitions
View event logs
Schedule tasks
Flash Player (32-bit)

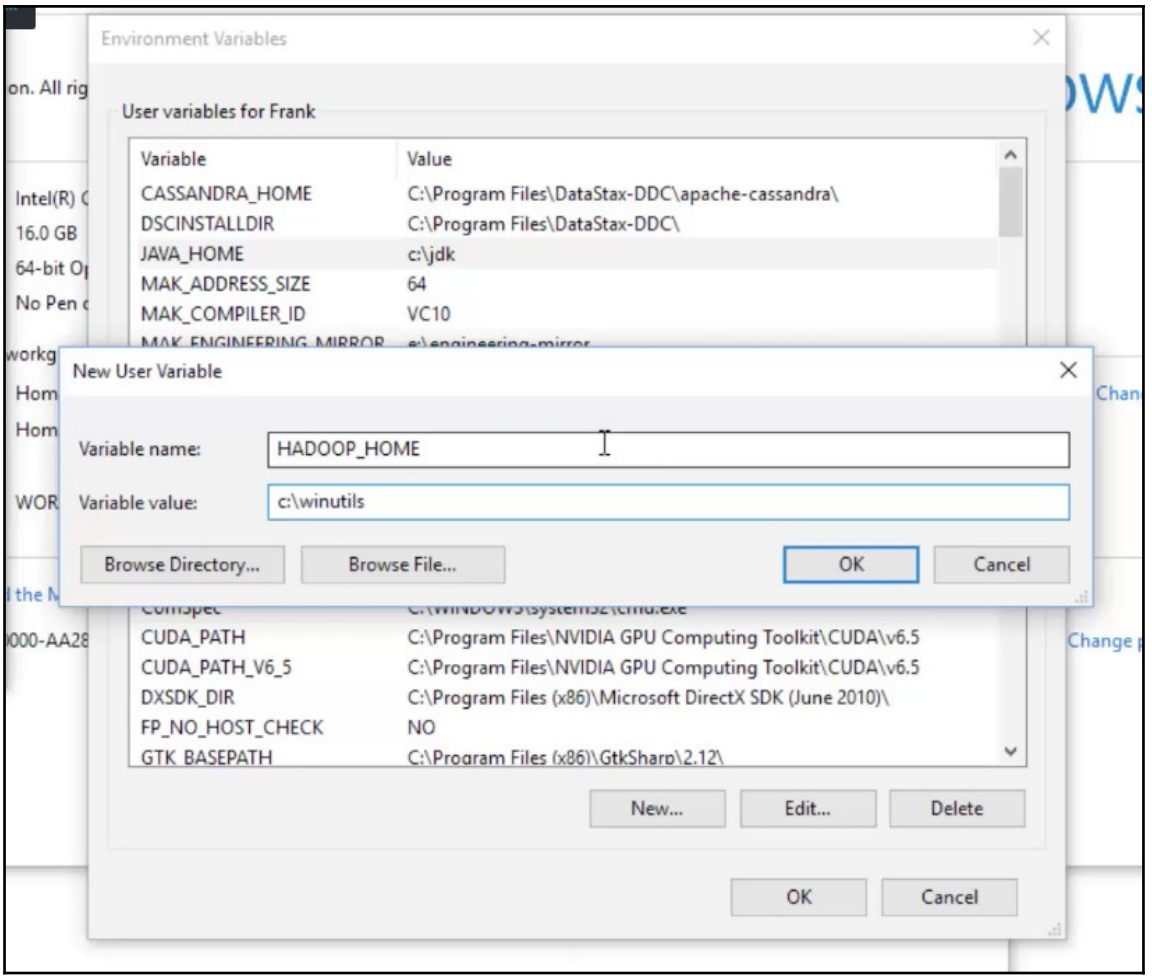

Control Panel
System and Security
System
Search Control Panel
Control Panel Home
Device Manager
Remote settings
System protection
Advanced system settings
View basic information about your computer
Windows edition
Windows 7 Professional
Copyright © 2009 Microsoft Corporation. All rights reserved.
Service Pack 1
Get more features with a new edition of Windows 7

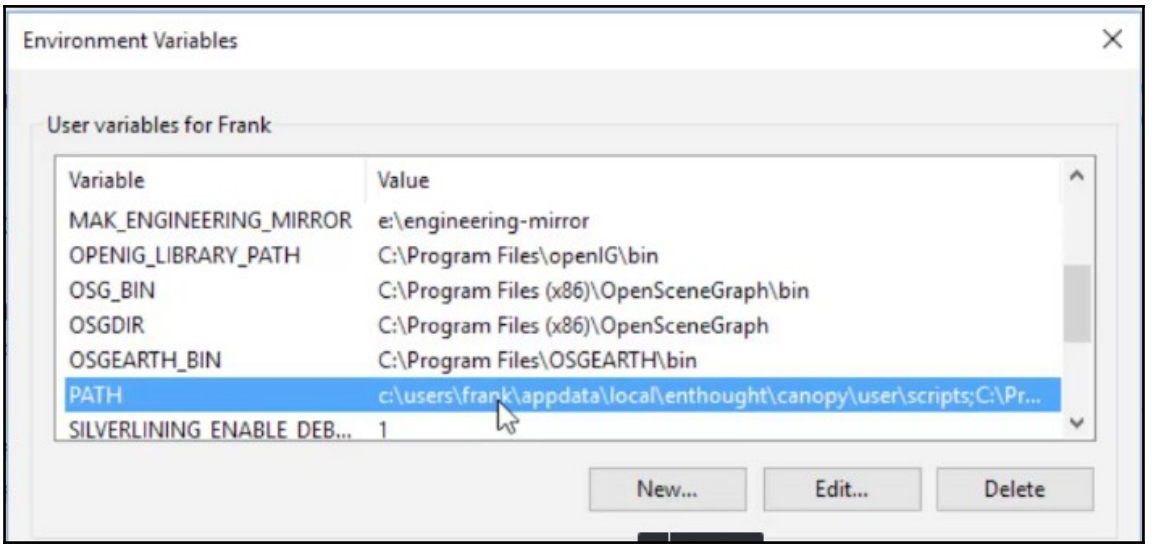

System Properties
Computer Name
Hardware
Advanced
System Protection
Remote
You must be logged on as an Administrator to make most of these changes.
Performance
Visual effects, processor scheduling, memory usage, and virtual memory
Settings...
User Profiles
Desktop settings related to your logon
Settings...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

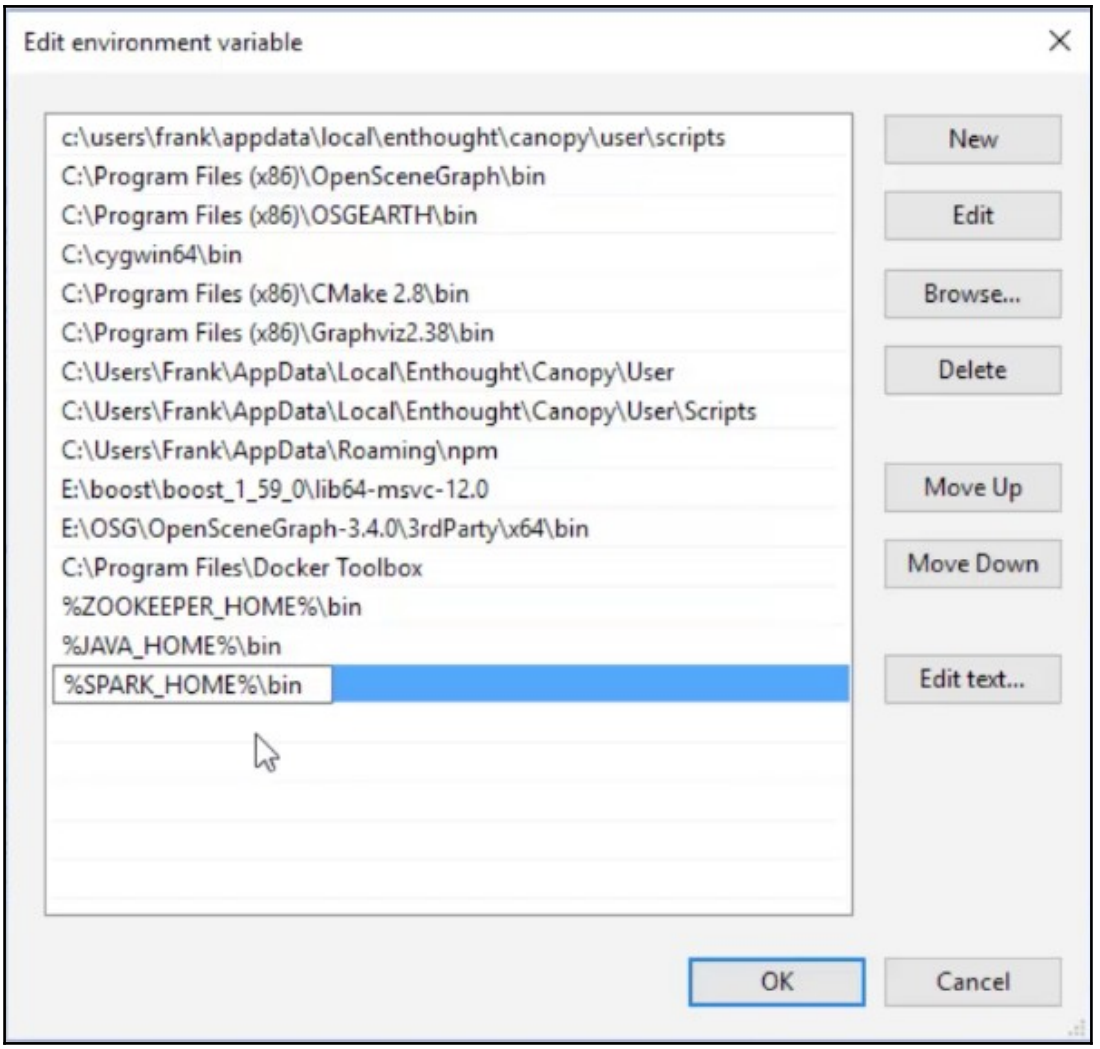

Environment Variables
User variables for nidhishas
Variable
Value
SPARK_HOME
c:\spark
TEMP
%USERPROFILE%\AppData\Local\Temp
TMP
%USERPROFILE%\AppData\Local\Temp
New...
Edit...
Delete
System variables
Variable
Value
ComSpec
C:\Windows\system32\cmd.exe
FP_NO_HOST_C...
NO
NUMBER_OF_P...
4
OS
Windows_NT
New...
Edit...
Delete
OK
Cancel


Environment Variables
User variables for Frank
Variable
Value
CASSANDRA_HOME
C:\Program Files\DataStax-DDC\apache-cassandra\
DSCINSTALLDIR
C:\Program Files\DataStax-DDC\
MAK_ADDRESS_SIZE
64
MAK_COMPILER_ID
VC10
MAK_ENGINEERING_MIRROR
e:\engineering-mirror
New User Variable
Variable name:
SPARK_HOME
Variable value:
c:\spark
Browse Directory...
Browse File...
OK
Cancel
CUDA_PATH
C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v6.5
CUDA_PATH_V6_5
C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v6.5
DXSDK_DIR
C:\Program Files (x86)\Microsoft DirectX SDK (June 2010)\
FP_NO_HOST_CHECK
NO
New...
Edit...
Delete
OK
Cancel

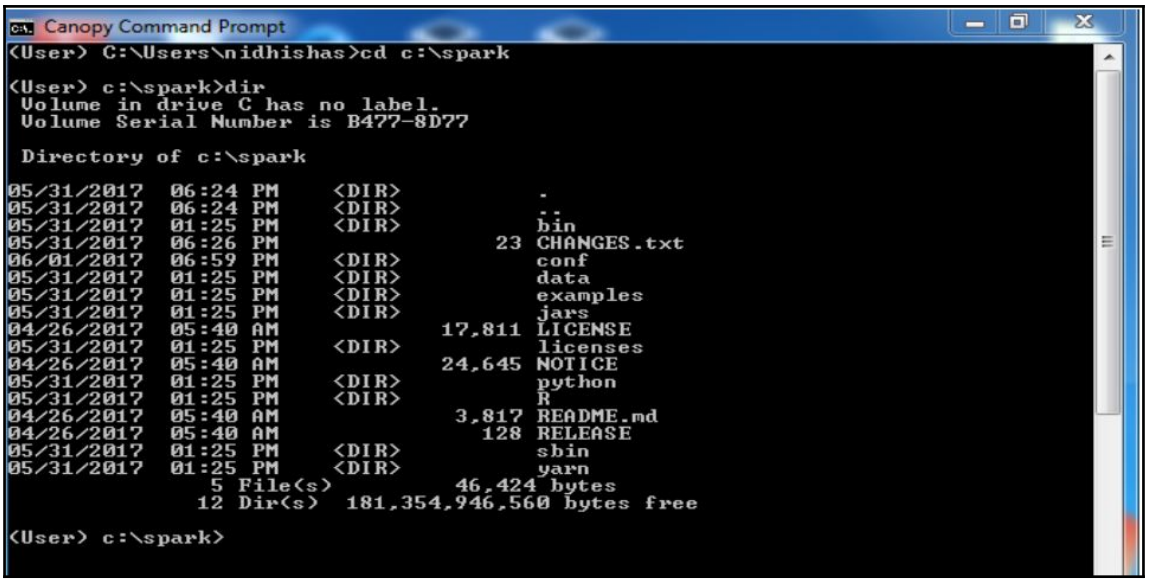

Environment Variables
User variables for Frank
Variable
Value
CASSANDRA_HOME
C:\Program Files\DataStax-DDC\apache-cassandra\
DSCINSTALLDIR
C:\Program Files\DataStax-DDC\
MAK_ADDRESS_SIZE
64
MAK_COMPILER_ID
VC10
MAK_ENGINEERING_MIRROR
e:\engineering-mirror
New User Variable
Variable name:
JAVA_HOME
Variable value:
c:\jdk
Browse Directory...
Browse File...
OK
Cancel
CUDA_PATH
C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v6.5
CUDA_PATH_V6_5
C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v6.5
DXSDK_DIR
C:\Program Files (x86)\Microsoft DirectX SDK (June 2010)\
FP_NO_HOST_CHECK
NO
New...
Edit...
Delete
OK
Cancel


Environment Variables
User variables for Frank
Variable
Value
CASSANDRA_HOME
C:\Program Files\DataStax-DDC\apache-cassandra\
DSCINSTALLDIR
C:\Program Files\DataStax-DDC\
JAVA_HOME
c:\jdk
MAK_ADDRESS_SIZE
64
MAK_COMPILER_ID
VC10
New User Variable
Variable name:
HADOOP_HOME
Variable value:
c:\winutils
Browse Directory...
Browse File...
OK
Cancel
CUDA_PATH
C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v6.5
CUDA_PATH_V6_5
C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v6.5
DXSDK_DIR
C:\Program Files (x86)\Microsoft DirectX SDK (June 2010)\
FP_NO_HOST_CHECK
NO
New...
Edit...
Delete
OK
Cancel

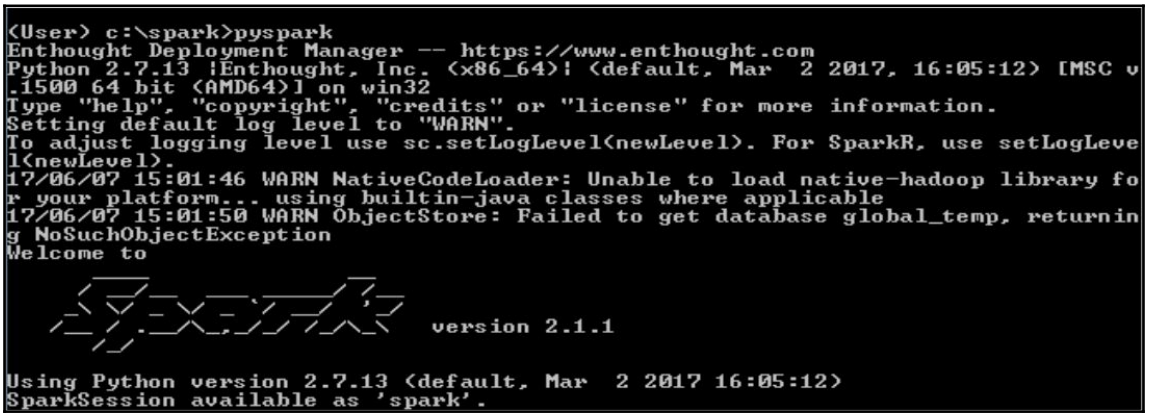

Environment Variables
User variables for Frank
Variable
Value
MAK_ENGINEERING_MIRROR
e:\engineering-mirror
OPENIG_LIBRARY_PATH
C:\Program Files\openIG\bin
OSG_BIN
C:\Program Files (x86)\OpenSceneGraph\bin
OSGDIR
C:\Program Files (x86)\OpenSceneGraph
OSGEARTH_BIN
C:\Program Files\OSGEARTH\bin
PATH
c:\users\frank\appdata\local\enthought\canopy\user\scripts;C:\Pr...
SILVERLINING_ENABLE_DEB...
1
New...
Edit...
Delete


Edit environment variable
c:\users\frank\appdata\local\enthought\canopy\user\scripts
C:\Program Files (x86)\OpenSceneGraph\bin
C:\Program Files (x86)\OSGEARTH\bin
C:\cygwin64\bin
C:\Program Files (x86)\CMake 2.8\bin
C:\Program Files (x86)\Graphviz2.38\bin
C:\Users\Frank\AppData\Local\Enthought\Canopy\User
C:\Users\Frank\AppData\Local\Enthought\Canopy\User\Scripts
C:\Users\Frank\AppData\Roaming\npm
E:\boost\boost_1_59_0\lib64-msvc-12.0
E:\OSG\OpenSceneGraph-3.4.0\3rdParty\x64\bin
C:\Program Files\Docker Toolbox
%ZOOKEEPER_HOME%\bin
%JAVA_HOME%\bin
%SPARK_HOME%\bin
New
Edit
Browse...
Delete
Move Up
Move Down
Edit text...
OK
Cancel


Canopy Command Prompt
(User) C:\Users\nidhishas>cd c:\spark

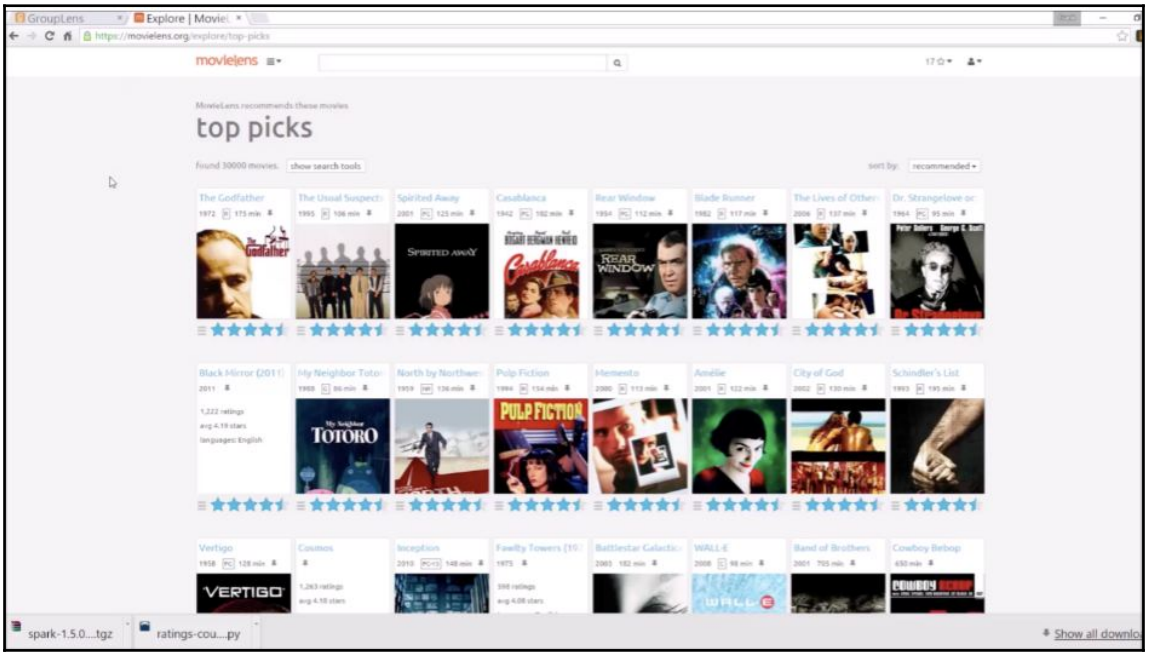

Canopy Command Prompt
(User) C:\Users\nidhishas>cd c:\spark
(User) c:\spark>dir
Volume in drive C has no label.
Volume Serial Number is B477-8D77
Directory of c:\spark
05/31/2017 06:24 PM <DIR> .
05/31/2017 06:24 PM <DIR> ..
05/31/2017 01:25 PM <DIR> bin
05/31/2017 06:26 PM 23 CHANGES.txt
06/01/2017 06:59 PM <DIR> conf
05/31/2017 01:25 PM <DIR> data
05/31/2017 01:25 PM <DIR> examples
05/31/2017 01:25 PM <DIR> jars
04/26/2017 05:40 AM 17,811 LICENSE
05/31/2017 01:25 PM <DIR> licenses
04/26/2017 05:40 AM 24,645 NOTICE
05/31/2017 01:25 PM <DIR> python
05/31/2017 01:25 PM <DIR> R
04/26/2017 05:40 AM 3,817 README.md
04/26/2017 05:40 AM 128 RELEASE
05/31/2017 01:25 PM <DIR> sbin
05/31/2017 01:25 PM <DIR> yarn
5 File(s) 46,424 bytes
12 Dir(s) 181,354,946,560 bytes free
(User) c:\spark>


(User) c:\spark>pyspark

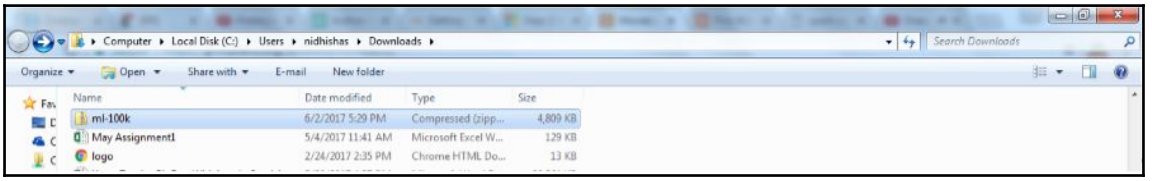

(User) c:\spark>pyspark
Enthought Deployment Manager -- https://www.enthought.com
Python 2.7.13 |Enthought, Inc. (x86_64)| (default, Mar 2 2017, 16:05:12) [MSC v
.1500 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
17/06/07 15:01:46 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
17/06/07 15:01:50 WARN ObjectStore: Failed to get database global_temp, returnin
g NoSuchObjectException
Welcome to
version 2.1.1
Using Python version 2.7.13 (default, Mar 2 2017 16:05:12)
SparkSession available as 'spark'.

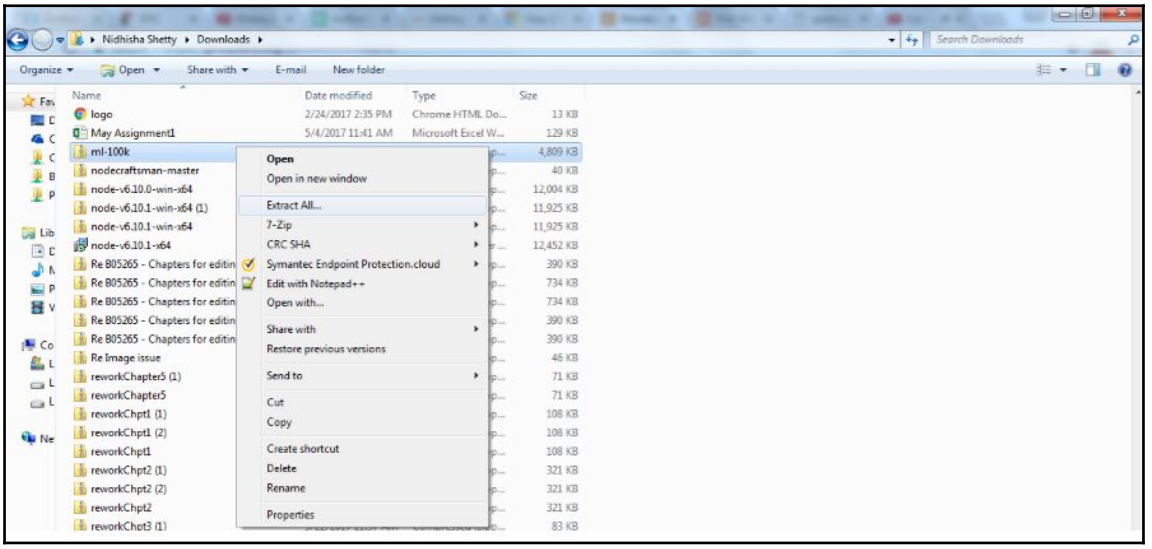

>>> rdd = sc.textFile("README.md")

```
>>> rdd = sc.textFile("README.md")
>>> rdd.count()
[Stage 0:>                                                          (0 + 2) / 2]

104
>>>
```

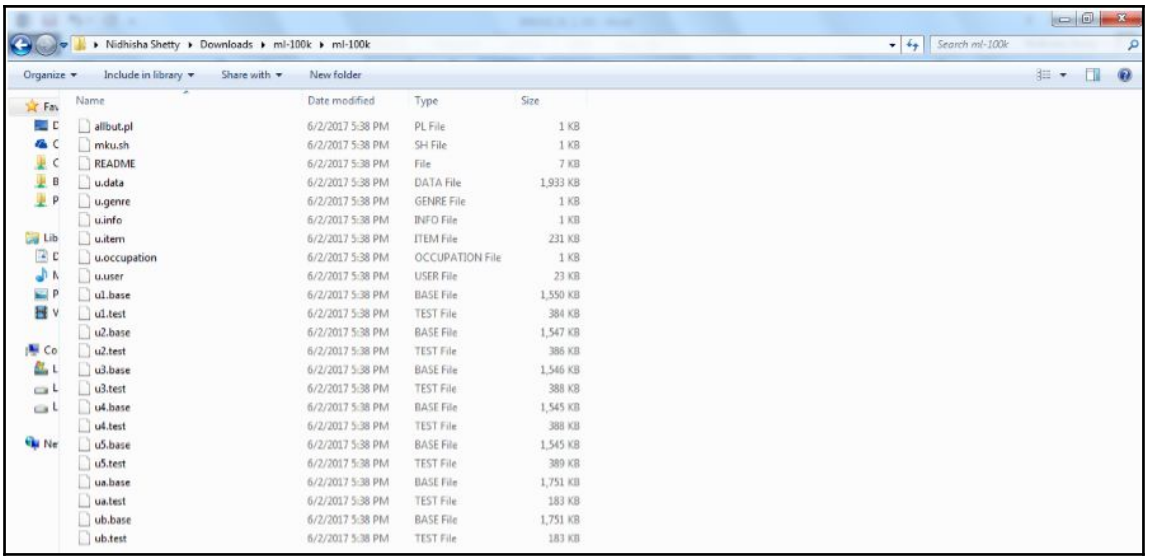


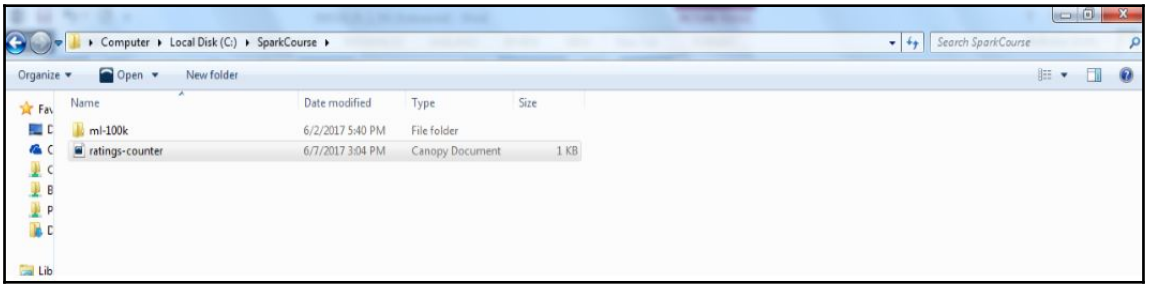

Secure | https://grouplens.org/datasets/movielens/
older datasets
MovieLens 100K Dataset
Stable benchmark dataset. 100,000 ratings from 1000 users on 1700 movies. Released 4/1998.
README.txt
ml-100k.zip (size: 5 MB, checksum)
Index of unzipped files
Permalink: http://grouplens.org/datasets/movielens/100k/

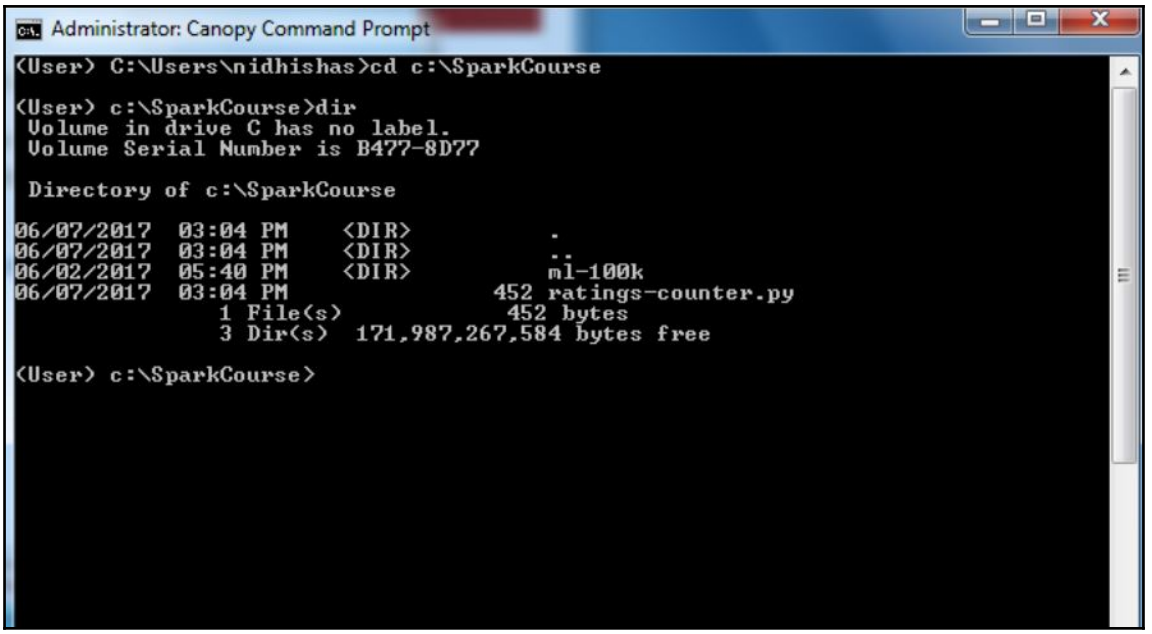

Computer ▸ Local Disk (C:) ▸ Users ▸ nidhishas ▸ Downloads ▸
Search Downloads
Organize
Open
Share with
E-mail
New folder
Name
Date modified
Type
Size
ml-100k
6/2/2017 5:29 PM
Compressed (zipp...
4,809 KB
May Assignment1
5/4/2017 11:41 AM
Microsoft Excel W...
129 KB
logo
2/24/2017 2:35 PM
Chrome HTML Do...
13 KB


Nidhisha Shetty ▸ Downloads ▸
Search Downloads
Organize
Open
Share with
E-mail
New folder
Name
Date modified
Type
Size
logo
2/24/2017 2:35 PM
Chrome HTML Do...
13 KB
May Assignment1
5/4/2017 11:41 AM
Microsoft Excel W...
129 KB
ml-100k
4,809 KB
nodecraftsman-master
40 KB
node-v6.10.0-win-x64
12,004 KB
node-v6.10.1-win-x64 (1)
11,925 KB
node-v6.10.1-win-x64
11,925 KB
node-v6.10.1-x64
12,452 KB
Re 805265 - Chapters for editin
390 KB
Re 805265 - Chapters for editin
734 KB
Re 805265 - Chapters for editin
734 KB
Re 805265 - Chapters for editin
390 KB
Re 805265 - Chapters for editin
390 KB
Re Image issue
46 KB
reworkChapter5 (1)
71 KB
reworkChapter5
71 KB
reworkChpt1 (1)
108 KB
reworkChpt1 (2)
108 KB
reworkChpt1
108 KB
reworkChpt2 (1)
321 KB
reworkChpt2 (2)
321 KB
reworkChpt2
321 KB
reworkChpt3 (1)
83 KB
Open
Open in new window
Extract All...
7-Zip
CRC SHA
Symantec Endpoint Protection.cloud
Edit with Notepad++
Open with...
Share with
Restore previous versions
Send to
Cut
Copy
Create shortcut
Delete
Rename
Properties

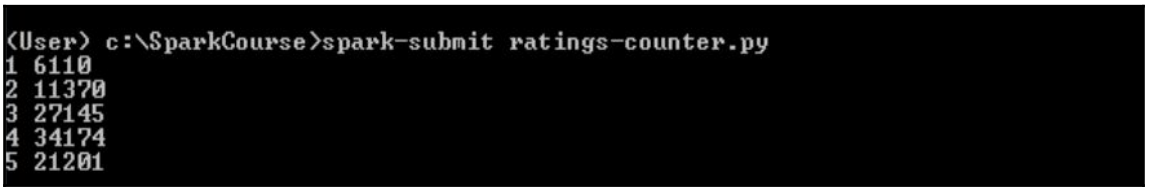

Nidhisha Shetty ▸ Downloads ▸ ml-100k ▸
Search ml-100k
Organize
Include in library
Share with
New folder
Name
Date modified
Type
Size
ml-100k
6/2/2017 5:38 PM
File folder


Nidhisha Shetty ▸ Downloads ▸ ml-100k ▸ ml-100k
Search ml-100k
Organize
Include in library
Share with
New folder
Name
Date modified
Type
Size
allbut.pl 6/2/2017 5:38 PM PL File 1 KB
mku.sh 6/2/2017 5:38 PM SH File 1 KB
README 6/2/2017 5:38 PM File 7 KB
u.data 6/2/2017 5:38 PM DATA File 1,933 KB
u.genre 6/2/2017 5:38 PM GENRE File 1 KB
u.info 6/2/2017 5:38 PM INFO File 1 KB
u.item 6/2/2017 5:38 PM ITEM File 231 KB
u.occupation 6/2/2017 5:38 PM OCCUPATION File 1 KB
u.user 6/2/2017 5:38 PM USER File 23 KB
u1.base 6/2/2017 5:38 PM BASE File 1,550 KB
u1.test 6/2/2017 5:38 PM TEST File 384 KB
u2.base 6/2/2017 5:38 PM BASE File 1,547 KB
u2.test 6/2/2017 5:38 PM TEST File 386 KB
u3.base 6/2/2017 5:38 PM BASE File 1,546 KB
u3.test 6/2/2017 5:38 PM TEST File 388 KB
u4.base 6/2/2017 5:38 PM BASE File 1,545 KB
u4.test 6/2/2017 5:38 PM TEST File 388 KB
u5.base 6/2/2017 5:38 PM BASE File 1,545 KB
u5.test 6/2/2017 5:38 PM TEST File 389 KB
ua.base 6/2/2017 5:38 PM BASE File 1,751 KB
ua.test 6/2/2017 5:38 PM TEST File 183 KB
ub.base 6/2/2017 5:38 PM BASE File 1,751 KB
ub.test 6/2/2017 5:38 PM TEST File 183 KB

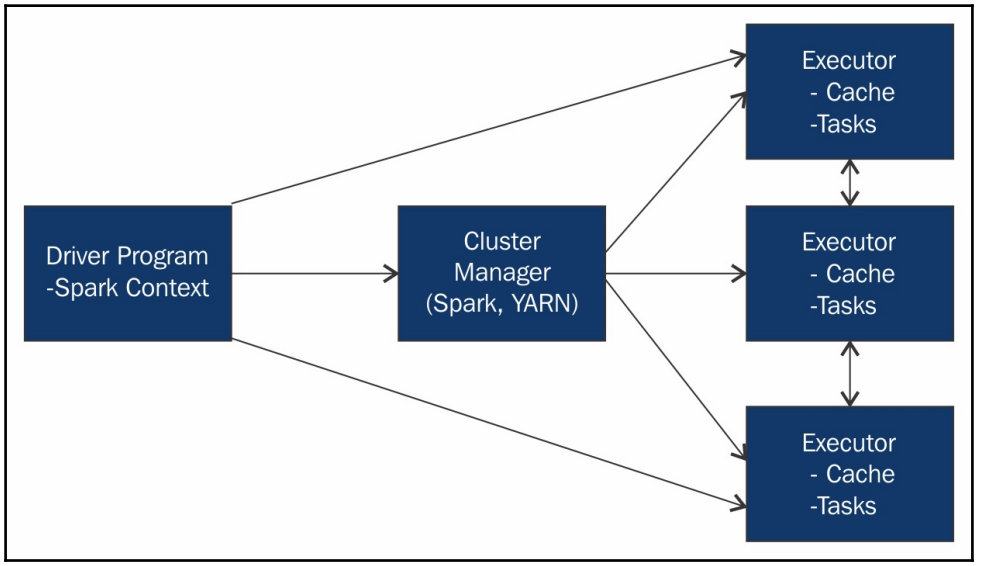

Computer ▸ Local Disk (C:) ▸ SparkCourse ▸
Search SparkCourse
Organize
Open
New folder
Name
Date modified
Type
Size
ml-100k 6/2/2017 5:40 PM File folder
ratings-counter 6/7/2017 3:04 PM Canopy Document 1 KB

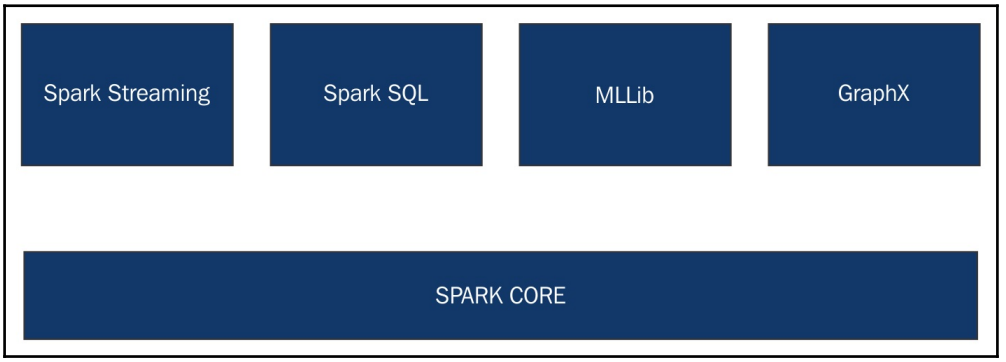

Administrator: Canopy Command Prompt
(User) C:\Users\nidhishas>cd c:\SparkCourse
(User) c:\SparkCourse>dir
Volume in drive C has no label.
Volume Serial Number is B477-8D77
Directory of c:\SparkCourse
06/07/2017 03:04 PM <DIR> .
06/07/2017 03:04 PM <DIR> ..
06/02/2017 05:40 PM <DIR> ml-100k
06/07/2017 03:04 PM 452 ratings-counter.py
1 File(s) 452 bytes
3 Dir(s) 171,987,267,584 bytes free
(User) c:\SparkCourse>


(User) c:\SparkCourse>spark-submit ratings-counter.py


(User) c:\SparkCourse>spark-submit ratings-counter.py
1 6110
2 11370
3 27145
4 34174
5 21201

Chapter 2


Executor
- Cache
-Tasks
Driver Program
-Spark Context
Cluster
Manager
(Spark, YARN)
Executor
- Cache
-Tasks
Executor
- Cache
-Tasks


Spark Streaming
Spark SQL
MLLib
GraphX
SPARK CORE

Python code to square number in a data set:

```
nums=sc.parallelize ([1, 2, 3, 4])
squared=nums.map (lambda x: x * x). collect()
```

Scala code to square numbers in a data set:

```
val nums=sc.parallelize(List(1, 2, 3, 4))
val squared=input.map(x=>x * x). collect()
```

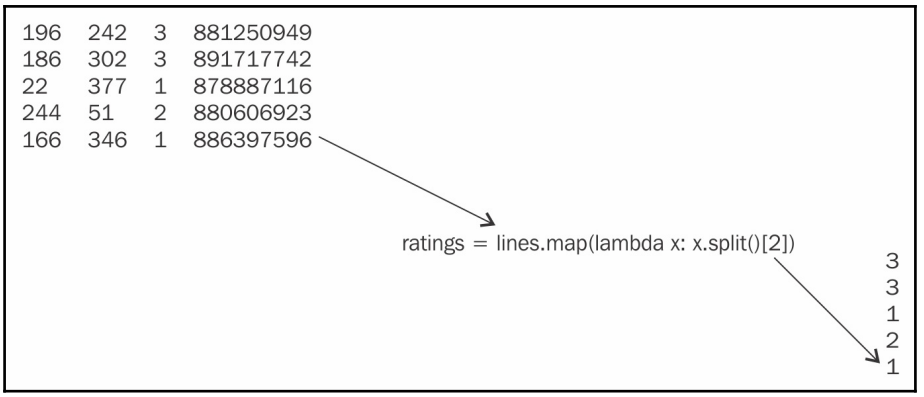
```
196  242  3  881250949
186  302  3  891717742
22   377  1  878887116
244  51   2  880606923
166  346  1  886397596
```

```
196  242  3  881250949
```

```
186  302  3  891717742
```

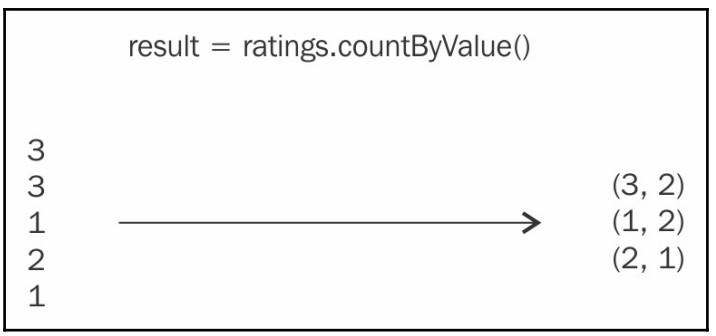
```
22  377  1  878887116
```

```
196  242  3  881250949
186  302  3  891717742
22   377  1  878887116
244  51   2  880606923
166  346  1  886397596
```



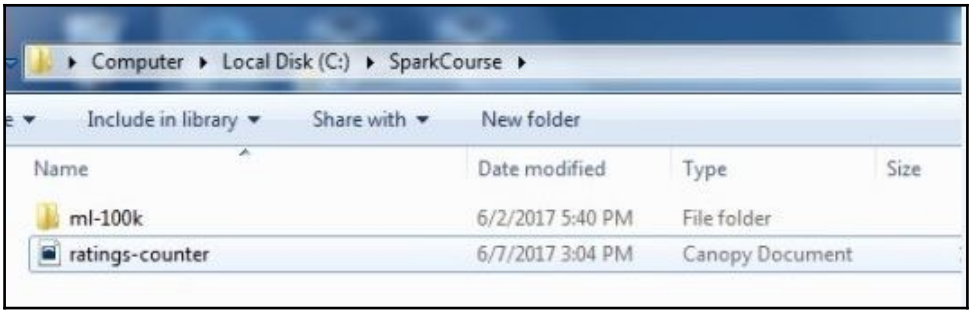
196	242	3	881250949
186	302	3	891717742
22	377	1	878887116
244	51	2	880606923
166	346	1	886397596

3
3
1
2
1



(3, 2)
(1, 2)
(2, 1)

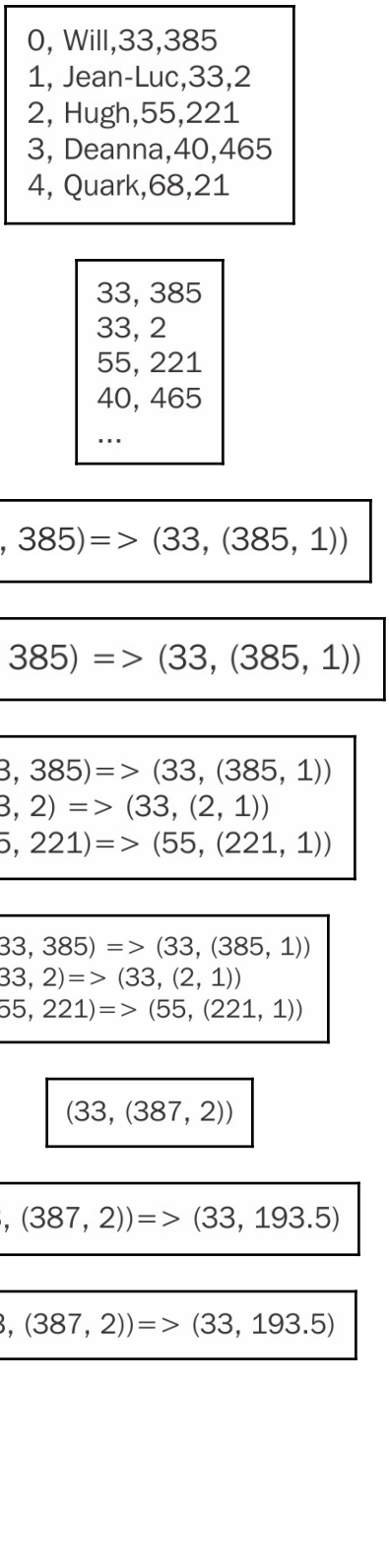
12
21
32




```
(User) c:\SparkCourse>spark-submit ratings-counter.py
1 6110
2 11370
3 27145
4 34174
5 21201
```

▷ Input Data: ID, name, age, number of friends

0, Will,33,385
1, Jean-Luc,33,2
2, Hugh,55,221
3, Deanna,40,465
4, QUARK,68,21


0, Will,33,385
1, Jean-Luc,33,2
2, Hugh,55,221
3, Deanna,40,465
4, Quark,68,21
33, 385
33, 2
55, 221
40, 465
...
(33, 385)=> (33, (385, 1))
(33, 385) => (33, (385, 1))
(33, 385)=> (33, (385, 1))
(33, 2) => (33, (2, 1))
(55, 221)=> (55, (221, 1))
(33, 385) => (33, (385, 1))
(33, 2)=> (33, (2, 1))
(55, 221)=> (55, (221, 1))
(33, (387, 2))
(33, (387, 2))=> (33, 193.5)
(33, (387, 2))=> (33, 193.5)

(33, (387, 2)) => (33, 193.5)

Name	Date modified	Type	Size
ml-100k	6/29/2017 12:37 PM	File folder	
fakefriends	6/8/2017 11:59 AM	Microsoft Excel C...	9 KB
friends-by-age	6/8/2017 11:54 AM	Canopy Document	1 KB
ratings-counter	6/7/2017 3:04 PM	Canopy Document	1 KB

friends-by-age.py

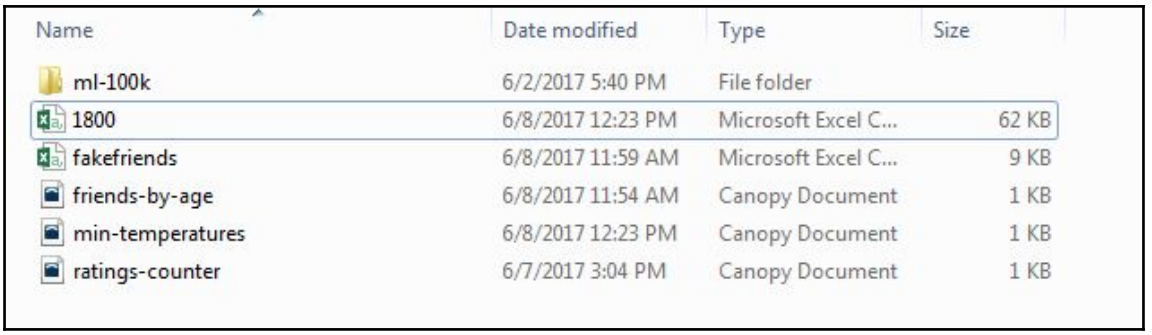
```
1 from pyspark import SparkConf, SparkContext
2
3 conf = SparkConf().setMaster("local").setAppName("FriendsByAge")
4 sc = SparkContext(conf = conf)
5
6 def parseLine(line):
7     fields = line.split(',')
8     age = int(fields[2])
9     numFriends = int(fields[3])
10    return (age, numFriends)
11
12 lines = sc.textFile("file:///SparkCourse/fakefriends.csv")
13 rdd = lines.map(parseLine)
14 totalsByAge = rdd.mapValues(lambda x: (x, 1)).reduceByKey(lambda x, y: (x[0] + y[0], x[1] + y[1]))
15 averagesByAge = totalsByAge.mapValues(lambda x: x[0] / x[1])
16 results = averagesByAge.collect()
17 for result in results:
18     print(result)
19
```

```
(47, 233)
(48, 281)
(49, 184)
(50, 254)
(51, 302)
(52, 340)
(53, 222)
(54, 278)
(55, 295)
(56, 306)
(57, 258)
(58, 116)
(59, 220)
(60, 202)
(61, 256)
(62, 220)
(63, 384)
(64, 281)
(65, 298)
(66, 276)
(67, 214)
(68, 269)
(69, 235)

(User) c:\SparkCourse>
```

```
ITE00100554,18000101,TMAX,-75,,,E,
ITE00100554,18000101,TMIN,-148,,,E,
GM000010962,18000101,PRCP,0,,,E,
EZE00100082,18000101,TMAX,-86,,,E,
EZE00100082,18000101,TMIN,-135,,,E,
```

```
ITE00100554,18000101,TMAX,-75,,,E,
ITE00100554,18000101,TMIN,-148,,,E,
GM000010962,18000101,PRCP,0,,,E,
EZE00100082,18000101,TMAX,-86,,,E,
EZE00100082,18000101,TMIN,-135,,,E,
```

```
ITE00100554,18000101,TMIN,-148,,,E,
```

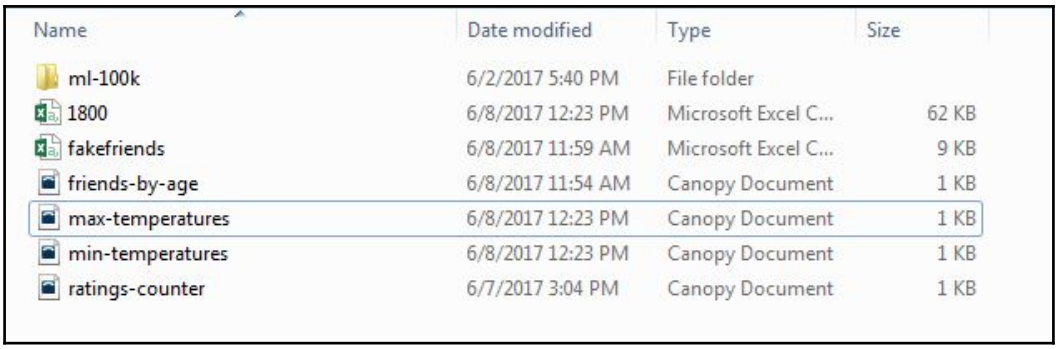


Name	Date modified	Type	Size
ml-100k	6/2/2017 5:40 PM	File folder	
1800	6/8/2017 12:23 PM	Microsoft Excel C...	62 KB
fakefriends	6/8/2017 11:59 AM	Microsoft Excel C...	9 KB
friends-by-age	6/8/2017 11:54 AM	Canopy Document	1 KB
min-temperatures	6/8/2017 12:23 PM	Canopy Document	1 KB
ratings-counter	6/7/2017 3:04 PM	Canopy Document	1 KB

```python
from pyspark import SparkConf, SparkContext

conf = SparkConf().setMaster("local").setAppName("MinTemperatures")
sc = SparkContext(conf = conf)

def parseLine(line):
    fields = line.split(',')
    stationID = fields[0]
    entryType = fields[2]
    temperature = float(fields[3]) * 0.1 * (9.0 / 5.0) + 32.0
    return (stationID, entryType, temperature)

lines = sc.textFile("file:///SparkCourse/1800.csv")
parsedLines = lines.map(parseLine)
minTemps = parsedLines.filter(lambda x: "TMIN" in x[1])
stationTemps = minTemps.map(lambda x: (x[0], x[2]))
minTemps = stationTemps.reduceByKey(lambda x, y: min(x,y))
results = minTemps.collect();

for result in results:
    print(result[0] + "\t{:.2f}F".format(result[1]))

```

```
(User) c:\SparkCourse>spark-submit min-temperatures.py
```

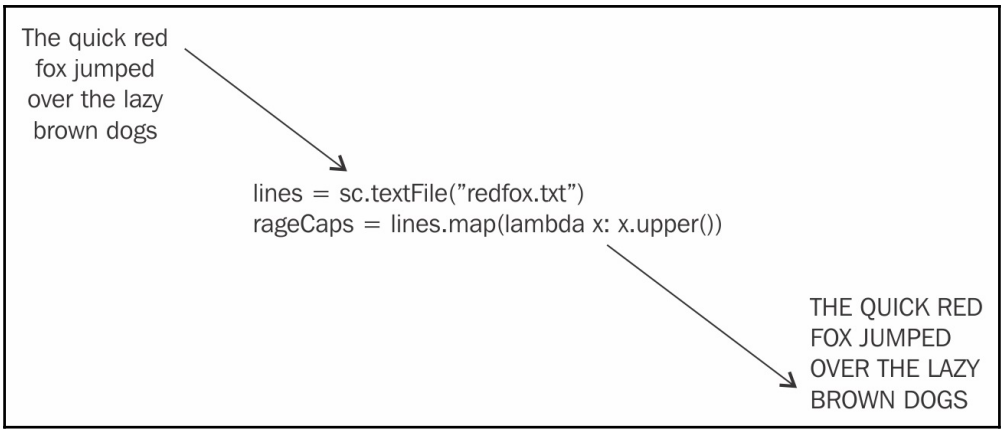
```
(User) c:\SparkCourse>spark-submit min-temperatures.py
ITE00100554     5.36F
EZE00100082     7.70F
```

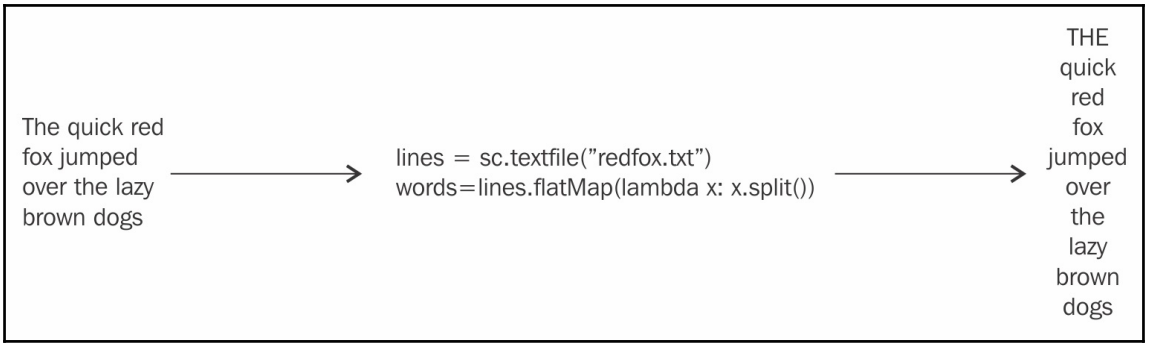


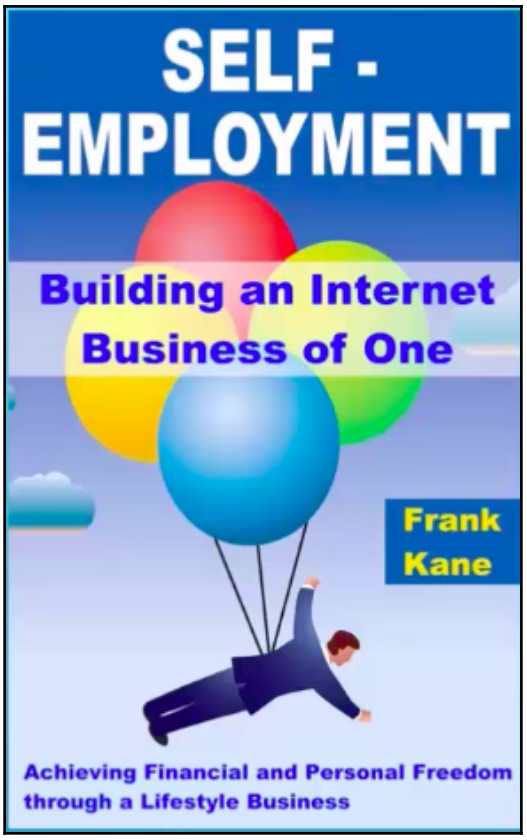
Name	Date modified	Type	Size
ml-100k	6/2/2017 5:40 PM	File folder	
1800	6/8/2017 12:23 PM	Microsoft Excel C...	62 KB
fakefriends	6/8/2017 11:59 AM	Microsoft Excel C...	9 KB
friends-by-age	6/8/2017 11:54 AM	Canopy Document	1 KB
max-temperatures	6/8/2017 12:23 PM	Canopy Document	1 KB
min-temperatures	6/8/2017 12:23 PM	Canopy Document	1 KB
ratings-counter	6/7/2017 3:04 PM	Canopy Document	1 KB

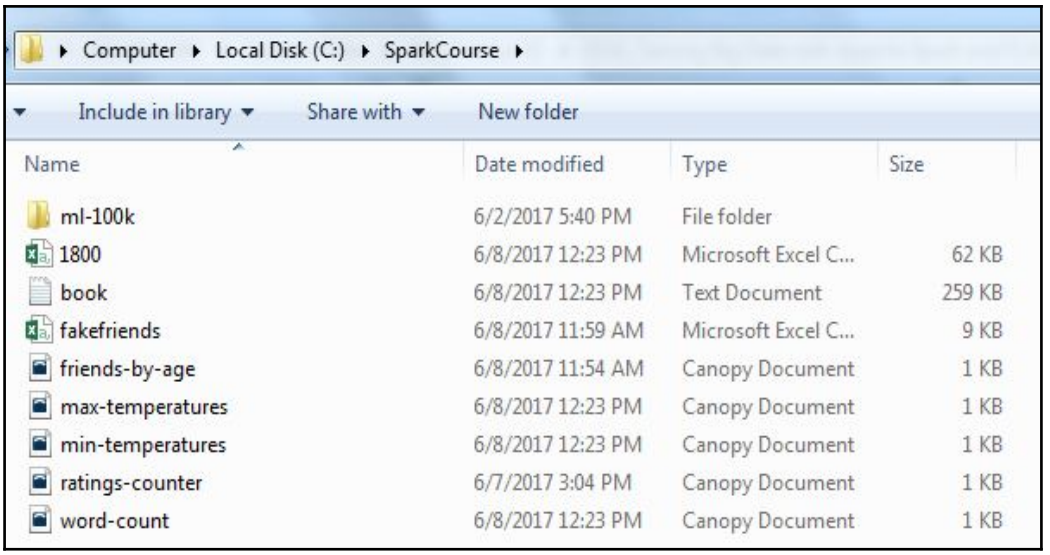
```python
from pyspark import SparkConf, SparkContext

conf = SparkConf().setMaster("local").setAppName("MaxTemperatures")
sc = SparkContext(conf = conf)

def parseLine(line):
    fields = line.split(',')
    stationID = fields[0]
    entryType = fields[2]
    temperature = float(fields[3]) * 0.1 * (9.0 / 5.0) + 32.0
    return (stationID, entryType, temperature)

lines = sc.textFile("file:///SparkCourse/1800.csv")
parsedLines = lines.map(parseLine)
maxTemps = parsedLines.filter(lambda x: "TMAX" in x[1])
stationTemps = maxTemps.map(lambda x: (x[0], x[2]))
maxTemps = stationTemps.reduceByKey(lambda x, y: max(x,y))
results = maxTemps.collect();

for result in results:
    print(result[0] + "\t{:.2f}F".format(result[1]))

```

```
(User) c:\SparkCourse>spark-submit max-temperatures.py
ITE00100554     90.14F
EZE00100082     90.14F
```


The quick red fox jumped over the lazy brown dogs
lines = sc.textFile(”redfox.txt”)
rageCaps = lines.map(lambda x: x.upper())
THE QUICK RED FOX JUMPED OVER THE LAZY BROWN DOGS

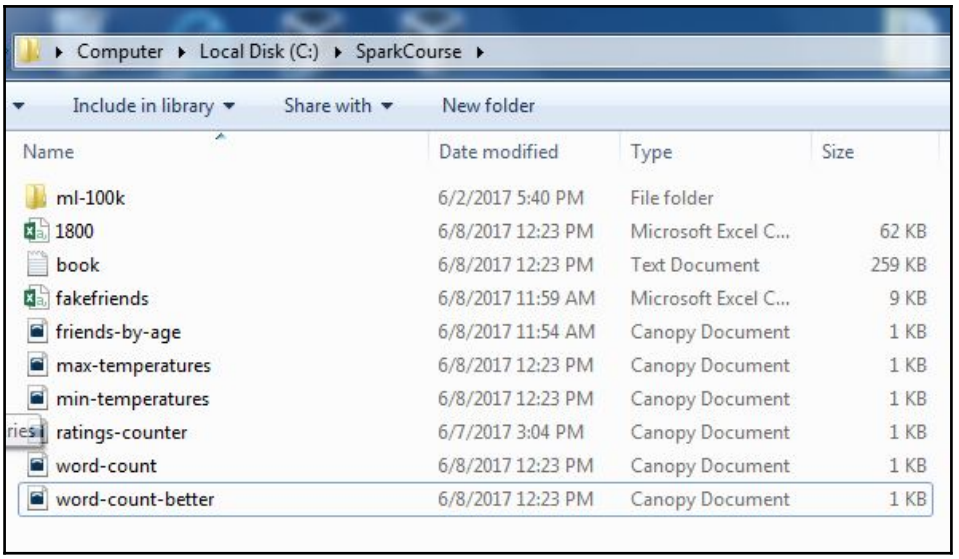

The quick red fox jumped over the lazy brown dogs
lines = sc.textfile(”redfox.txt”)
words=lines.flatMap(lambda x: x.split())
THE quick red fox jumped over the lazy brown dogs


SELF - EMPLOYMENT
Building an Internet Business of One
Frank Kane
Achieving Financial and Personal Freedom through a Lifestyle Business



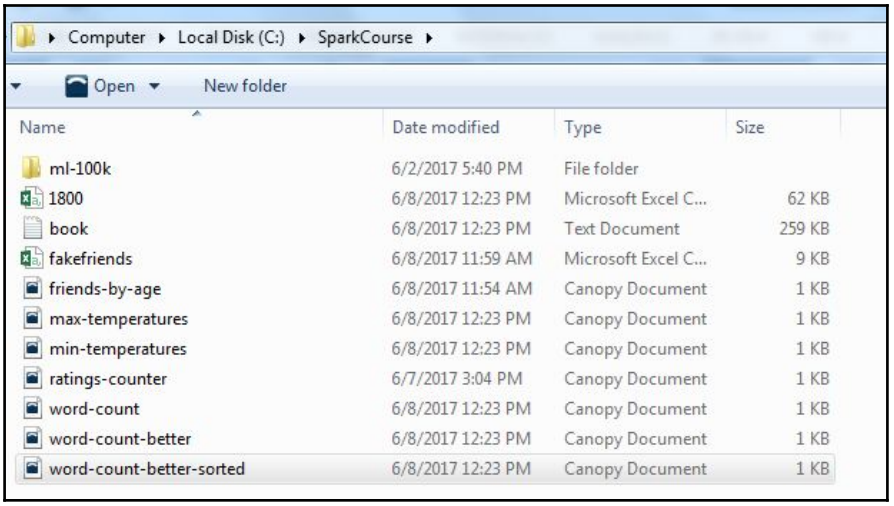

```
1  from pyspark import SparkConf, SparkContext
2
3  conf = SparkConf().setMaster("local").setAppName("WordCount")
4  sc = SparkContext(conf = conf)
5
6  input = sc.textFile("file:///sparkcourse/book.txt")
7  words = input.flatMap(lambda x: x.split())
8  wordCounts = words.countByValue()
9
10 for word, count in wordCounts.items():
11     cleanWord = word.encode('ascii', 'ignore')
12     if (cleanWord):
13         print(cleanWord.decode() + " " + str(count))
14
```

```
DESIGNING 1
clients, 2
clients. 2
made 12
whether 21
this, 10
distract 1
this. 2
below 2
USING 1
this; 1
this: 2
intimidating 1
inadequate 1
meaningless 2
highlighted 1
kind. 1
avoided 1
improving 1
Sessions 1
other 71
incredibly 2
Banner 2
clicks 2
junk 1
kinds 3
webpage, 1
PLAN 3
S-corp 1
incurred 1
extort 1
click, 1
Company) 1
LEECHES 1
click" 1
Site 1
intentionally 2
entirely. 1

(User) c:\SparkCourse>
```

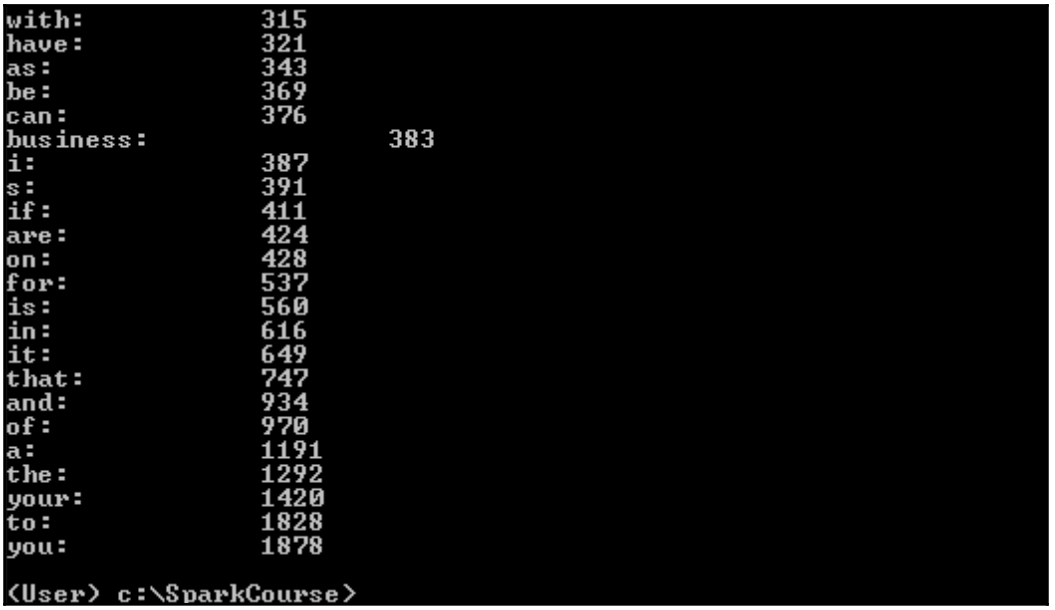



```
1  import re
2  from pyspark import SparkConf, SparkContext
3
4  def normalizeWords(text):
5      return re.compile(r'\W+', re.UNICODE).split(text.lower())
6
7  conf = SparkConf().setMaster("local").setAppName("WordCount")
8  sc = SparkContext(conf = conf)
9
10 input = sc.textFile("file:///sparkcourse/book.txt")
11 words = input.flatMap(normalizeWords)
12 wordCounts = words.countByValue()
13
14 for word, count in wordCounts.items():
15     cleanWord = word.encode('ascii', 'ignore')
16     if (cleanWord):
17         print(cleanWord.decode() + " " + str(count))
18
```

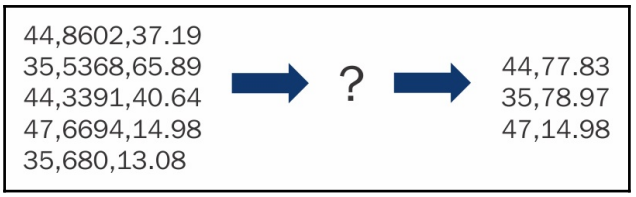
```
forgivable 1
details 5
normal 1
welcomes 1
mass 5
out 161
conversational 1
clicks 3
disposing 1
troll 1
junk 1
star 1
shown 4
variation 2
stay 7
chance 12
workaholic 1
spreadsheet 2
gap 2
friends 10
incurred 1
exposure 2
shock 1
ended 10
lasted 3
```



```python
import re
from pyspark import SparkConf, SparkContext

def normalizeWords(text):
    return re.compile(r'\W+', re.UNICODE).split(text.lower())

conf = SparkConf().setMaster("local").setAppName("WordCount")
sc = SparkContext(conf = conf)

input = sc.textFile("file:///sparkcourse/book.txt")
words = input.flatMap(normalizeWords)

wordCounts = words.map(lambda x: (x, 1)).reduceByKey(lambda x, y: x + y)
wordCountsSorted = wordCounts.map(lambda x: (x[1], x[0])).sortByKey()
results = wordCountsSorted.collect()

for result in results:
    count = str(result[0])
    word = result[1].encode('ascii', 'ignore')
    if (word):
        print(word.decode() + ":\t\t" + count)

```




```
44,8602,37.19
35,5368,65.89
44,3391,40.64
47,6694,14.98
35,680,13.08
```




```python
from pyspark import SparkConf, SparkContext

conf = SparkConf().setMaster("local").setAppName("SpendByCustomer")
sc = SparkContext(conf = conf)

def extractCustomerPricePairs(line):
    fields = line.split(',')
    return (int(fields[0]), float(fields[2]))

input = sc.textFile("file:///sparkcourse/customer-orders.csv")
mappedInput = input.map(extractCustomerPricePairs)
totalByCustomer = mappedInput.reduceByKey(lambda x, y: x + y)

results = totalByCustomer.collect();
for result in results:
    print(result)
```

```
(77, 4327.729999999999)
(78, 4524.509999999999)
(79, 3790.570000000001)
(80, 4727.860000000001)
(81, 5112.709999999999)
(82, 4812.489999999998)
(83, 4635.799999999997)
(84, 4652.939999999999)
(85, 5503.43)
(86, 4908.81)
(87, 5206.4)
(88, 4830.549999999999)
(89, 4851.479999999999)
(90, 5290.409999999998)
(91, 4642.259999999999)
(92, 5379.280000000002)
(93, 5265.750000000001)
(94, 4475.569999999999)
(95, 4876.840000000002)
(96, 3924.230000000001)
(97, 5977.189999999995)
(98, 4297.260000000001)
(99, 4172.289999999998)

(User) c:\SparkCourse>
```

```
(84, 4652.939999999999)
```

```python
from pyspark import SparkConf, SparkContext

conf = SparkConf().setMaster("local").setAppName("SpendByCustomerSorted")
sc = SparkContext(conf = conf)

def extractCustomerPricePairs(line):
    fields = line.split(',')
    return (int(fields[0]), float(fields[2]))

input = sc.textFile("file:///sparkcourse/customer-orders.csv")
mappedInput = input.map(extractCustomerPricePairs)
totalByCustomer = mappedInput.reduceByKey(lambda x, y: x + y)

#Changed for Python 3 compatibility:
#flipped = totalByCustomer.map(lambda (x,y):(y,x))
flipped = totalByCustomer.map(lambda x: (x[1], x[0]))

totalByCustomerSorted = flipped.sortByKey()

results = totalByCustomerSorted.collect();
for result in results:
    print(result)

```

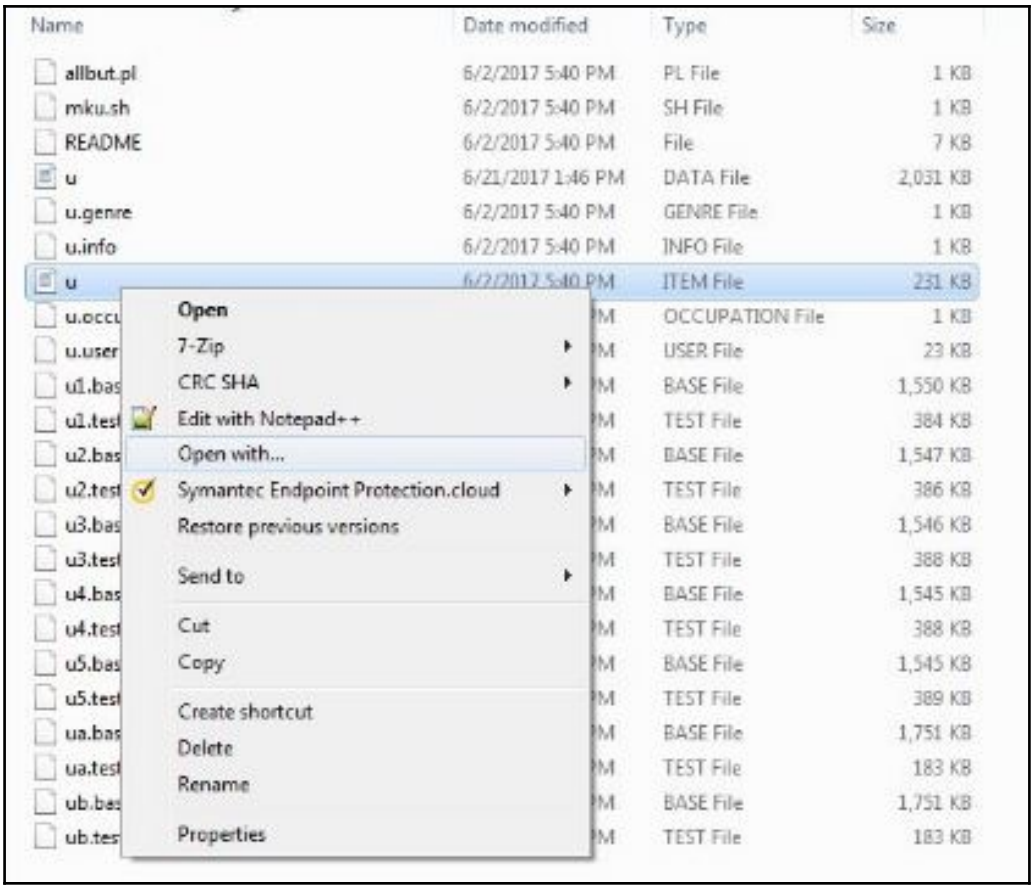
```
(5368.249999999999, 70)
(5368.83, 43)
(5379.280000000002, 92)
(5397.879999999998, 6)
(5413.510000000001, 15)
(5415.150000000001, 63)
(5437.730000000005, 58)
(5496.050000000004, 32)
(5497.479999999998, 61)
(5503.43, 85)
(5517.240000000001, 8)
(5524.949999999998, 0)
(5637.62, 41)
(5642.89, 59)
(5696.840000000003, 42)
(5963.109999999999, 46)
(5977.189999999995, 97)
(5994.59, 2)
(5995.660000000003, 71)
(6065.389999999999, 54)
(6193.109999999999, 39)
(6206.199999999999, 73)
(6375.449999999997, 68)

(User) c:\SparkCourse>
```

Chapter 3

```
196   242   3   881250949
186   302   3   891717742
22    377   1   878887116
244    51   2   880606923
166   346   1   886397596
298   474   4   884182806
```

```
(User) c:\SparkCourse>spark-submit popular-movies.py
```

```
(336, 79)
(344, 405)
(350, 204)
(350, 313)
(365, 222)
(367, 172)
(378, 117)
(384, 237)
(390, 98)
(392, 7)
(394, 56)
(413, 127)
(420, 174)
(429, 121)
(431, 300)
(452, 1)
(470, 288)
(481, 286)
(485, 294)
(507, 181)
(508, 100)
(509, 258)
(583, 50)

(User) c:\SparkCourse>_
```

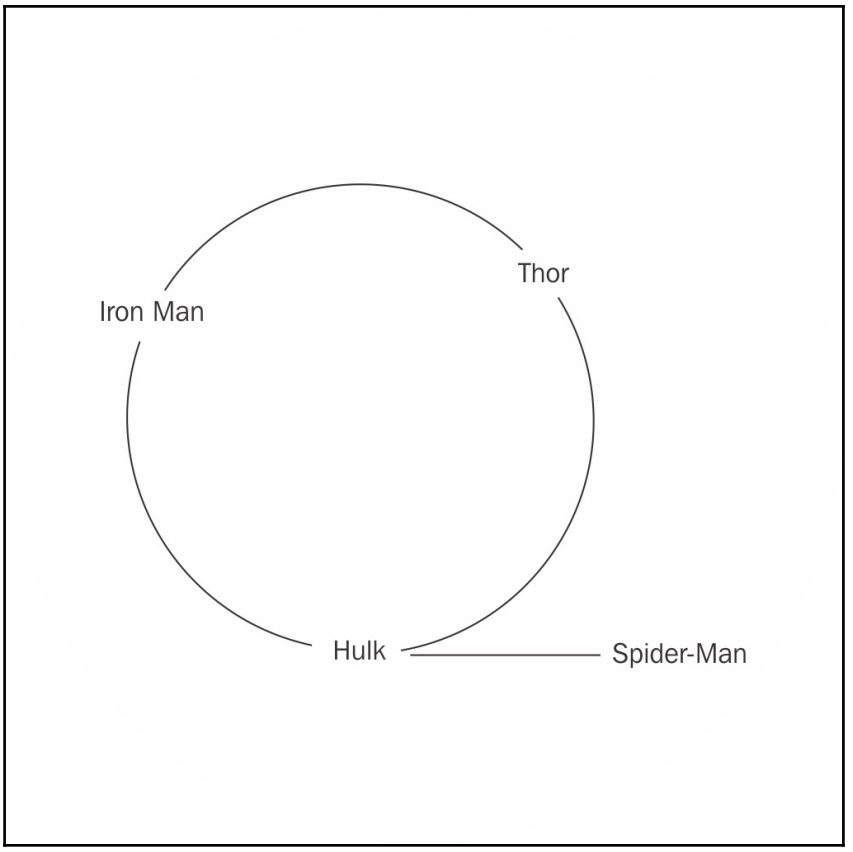

Name
Date modified
Type
Size
allbut.pl
6/2/2017 5:40 PM
PL File
1 KB
mku.sh
6/2/2017 5:40 PM
SH File
1 KB
README
6/2/2017 5:40 PM
File
7 KB
u
6/21/2017 1:46 PM
DATA File
2,031 KB
u.genre
6/2/2017 5:40 PM
GENRE File
1 KB
u.info
6/2/2017 5:40 PM
INFO File
1 KB
u
6/2/2017 5:40 PM
ITEM File
231 KB
u.occu
OCCUPATION File
1 KB
u.user
USER File
23 KB
u1.bas
BASE File
1,550 KB
u1.test
TEST File
384 KB
u2.bas
BASE File
1,547 KB
u2.test
TEST File
386 KB
u3.bas
BASE File
1,546 KB
u3.test
TEST File
388 KB
u4.bas
BASE File
1,545 KB
u4.test
TEST File
388 KB
u5.bas
BASE File
1,545 KB
u5.test
TEST File
389 KB
ua.bas
BASE File
1,751 KB
ua.test
TEST File
183 KB
ub.bas
BASE File
1,751 KB
ub.test
TEST File
183 KB
Open
7-Zip
CRC SHA
Edit with Notepad++
Open with...
Symantec Endpoint Protection.cloud
Restore previous versions
Send to
Cut
Copy
Create shortcut
Delete
Rename
Properties

```
exact?Disclosure%20(1994)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|1|0|
0
44|Dolores Claiborne (1994)|01-Jan-1994
||http://us.imdb.com/M/title-exact?Dolores%20Claiborne%20
(1994)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|1|0|0
45|Eat Drink Man Woman (1994)|01-Jan-1994
||http://us.imdb.com/M/title-exact?Yinshi%20Nan%20Nu%20(1994)|
0|0|0|0|0|0|1|0|0|1|0|0|0|0|0|0|0|0|0|0
46|Exotica (1994)|01-Jan-1994||http://us.imdb.com/M/title-
exact?Exotica%20(1994)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0
47|Ed Wood (1994)|01-Jan-1994||http://us.imdb.com/M/title-
exact?Ed%20Wood%20(1994)|0|0|0|0|0|1|0|0|1|0|0|0|0|0|0|0|0|0
48|Hoop Dreams (1994)|01-Jan-1994||http://us.imdb.com/M/title-
exact?Hoop%20Dreams%20(1994)|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0
|0|0
49|I.Q. (1994)|01-Jan-1994||http://us.imdb.com/M/title-exact?
I.Q.%20(1994)|0|0|0|0|0|1|0|0|0|0|0|0|0|1|0|0|0|0
50|Star Wars (1977)|01-Jan-1977||http://us.imdb.com/M/title-
exact?Star%20Wars%20(1977)|0|1|1|0|0|0|0|0|0|0|0|0|0|1|1|0|1
|0
51|Legends of the Fall (1994)|01-Jan-1994
||http://us.imdb.com/M/title-exact?Legends%20of%20the%20Fall%
20(1994)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|1|0|0|1|1
52|Madness of King George, The (1994)|01-Jan-1994
||http://us.imdb.com/M/title-exact?Madness%20of%20King%
20George,%20The%20(1994)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0
53|Natural Born Killers (1994)|01-Jan-1994
||http://us.imdb.com/M/title-exact?Natural%20Born%20Killers%20
(1994)|0|1|0|0|0|0|0|0|0|0|0|0|0|0|0|1|0|0
54|Outbreak (1995)|01-Jan-1995||http://us.imdb.com/M/title-
exact?Outbreak%20(1995)|0|1|0|0|0|0|0|0|1|0|0|0|0|0|0|0|1|0|0
55|Professional, The (1994)|01-Jan-1994
```

```python
from pyspark import SparkConf, SparkContext

def loadMovieNames():
    movieNames = {}
    with open("ml-100k/u.ITEM") as f:
        for line in f:
            fields = line.split('|')
            movieNames[int(fields[0])] = fields[1]
    return movieNames

conf = SparkConf().setMaster("local").setAppName("PopularMovies")
sc = SparkContext(conf = conf)

nameDict = sc.broadcast(loadMovieNames())

lines = sc.textFile("file:///SparkCourse/ml-100k/u.data")
movies = lines.map(lambda x: (int(x.split()[1]), 1))
movieCounts = movies.reduceByKey(lambda x, y: x + y)

flipped = movieCounts.map( lambda (x, y) : (y, x))
sortedMovies = flipped.sortByKey()

sortedMoviesWithNames = sortedMovies.map(lambda (count, movie) : (nameDict.value[movie], count))

results = sortedMoviesWithNames.collect()

for result in results:
    print(result)
```

```
(User) c:\SparkCourse>spark-submit popular-movies-nicer.py
```

```
('Fugitive, The (1993)', 336)
('Mission: Impossible (1996)', 344)
('Back to the Future (1985)', 350)
('Titanic (1997)', 350)
('Star Trek: First Contact (1996)', 365)
('Empire Strikes Back, The (1980)', 367)
('Rock, The (1996)', 378)
('Jerry Maguire (1996)', 384)
('Silence of the Lambs, The (1991)', 390)
('Twelve Monkeys (1995)', 392)
('Pulp Fiction (1994)', 394)
('Godfather, The (1972)', 413)
('Raiders of the Lost Ark (1981)', 420)
('Independence Day (ID4) (1996)', 429)
('Air Force One (1997)', 431)
('Toy Story (1995)', 452)
('Scream (1996)', 478)
('English Patient, The (1996)', 481)
('Liar Liar (1997)', 485)
('Return of the Jedi (1983)', 507)
('Fargo (1996)', 508)
('Contact (1997)', 509)
('Star Wars (1977)', 583)

(User) c:\SparkCourse>
```



4395 2237 1767 472 4997 5931 6235 1478 1369 806 3994 6232
3519 4704 2460 763 1602 5306 5358 6121 6160 2459 3173 4963 6166
3518 5409

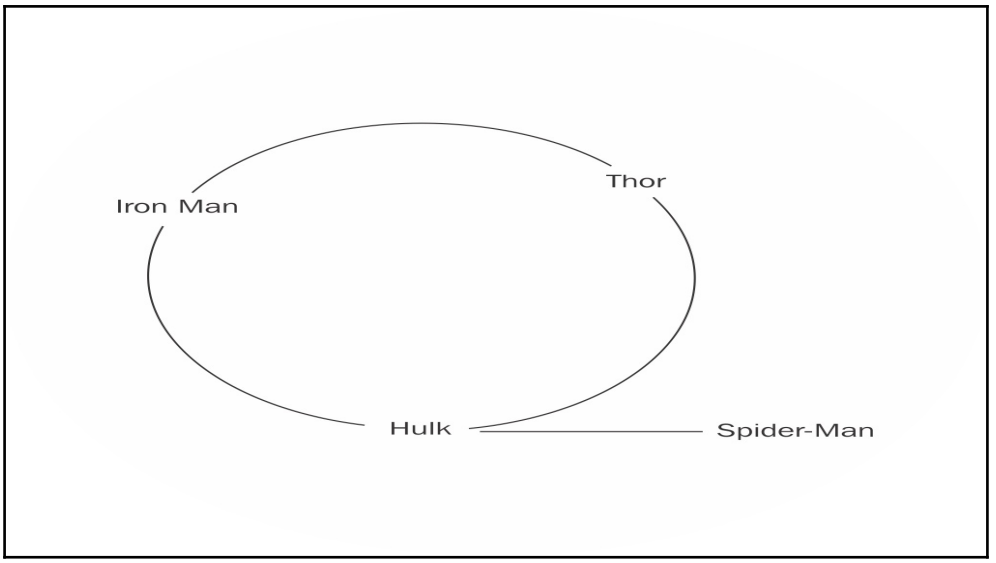
5300 “SPENCER, TRACY”
5301 “SPERZEL, ANTON”
5302 “SPETSBURO, GEN. YURI”
5303 “SPHINX”
5304 “SPHINX II”
5305 “SPHINX III”
5306 “SPIDER-MAN/PETER PAR”
5307 “SPIDER-MAN III/MARTH”
5308 “SPIDER CLONE/BEN”
5309 “SPIDER-WOMAN/JESSICA”

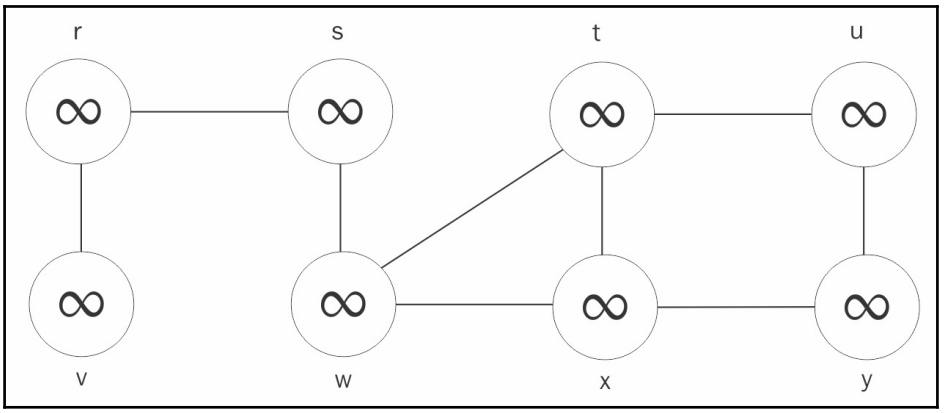
most-popular-superhero.py

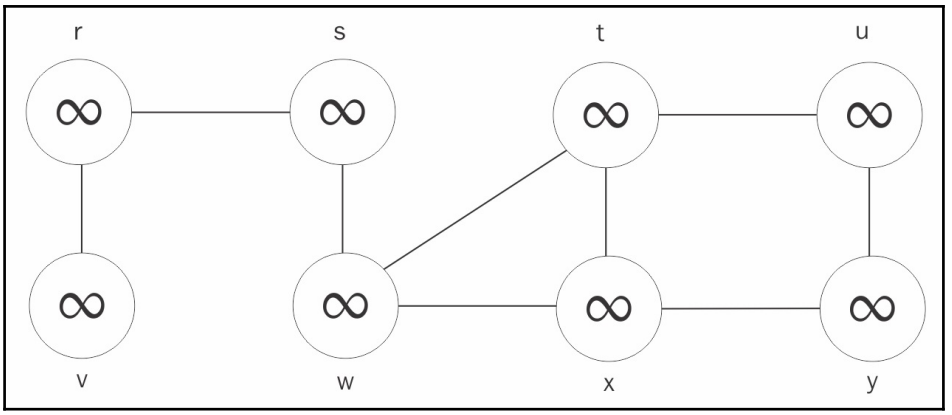
```python
from pyspark import SparkConf, SparkContext

conf = SparkConf().setMaster("local").setAppName("PopularHero")
sc = SparkContext(conf = conf)

def countCoOccurences(line):
    elements = line.split()
    return (int(elements[0]), len(elements) - 1)

def parseNames(line):
    fields = line.split('\"')
    return (int(fields[0]), fields[1].encode("utf8"))

names = sc.textFile("file:///SparkCourse/marvel-names.txt")
namesRdd = names.map(parseNames)

lines = sc.textFile("file:///SparkCourse/marvel-graph.txt")

pairings = lines.map(countCoOccurences)
totalFriendsByCharacter = pairings.reduceByKey(lambda x, y : x + y)
flipped = totalFriendsByCharacter.map(lambda (x,y) : (y,x))

mostPopular = flipped.max()

mostPopularName = namesRdd.lookup(mostPopular[1])[0]

print(mostPopularName + " is the most popular superhero, with " + \
    str(mostPopular[0]) + " co-appearances.")

```

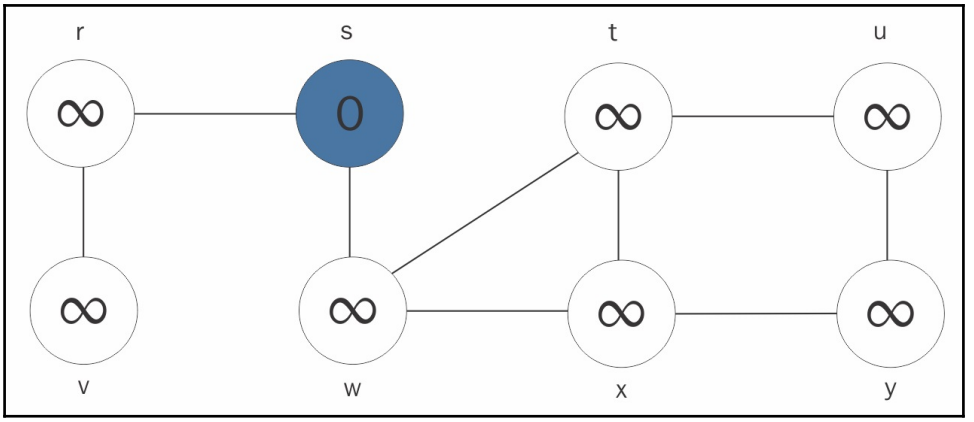
```
(User) c:\SparkCourse>spark-submit most-popular-superhero.py
```

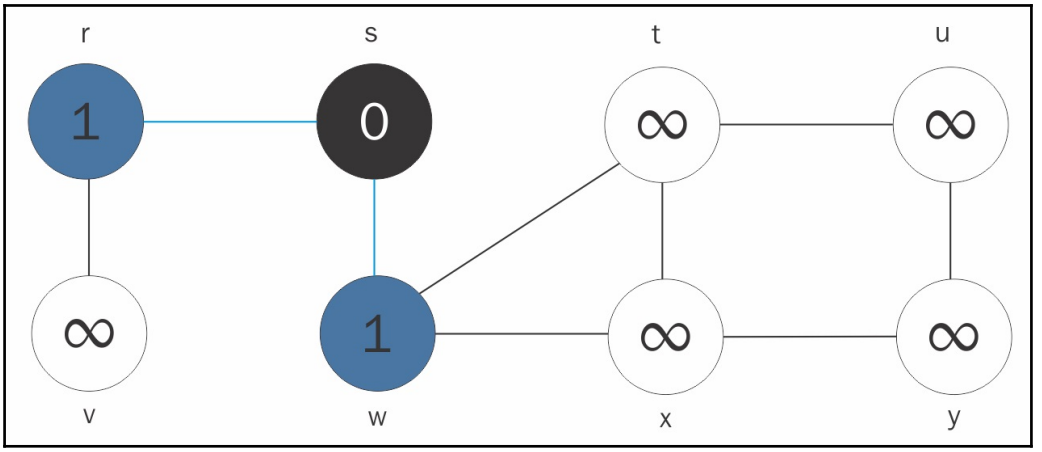
```
CAPTAIN AMERICA is the most popular superhero, with 1933 co-appearances.

(User) c:\SparkCourse>
```

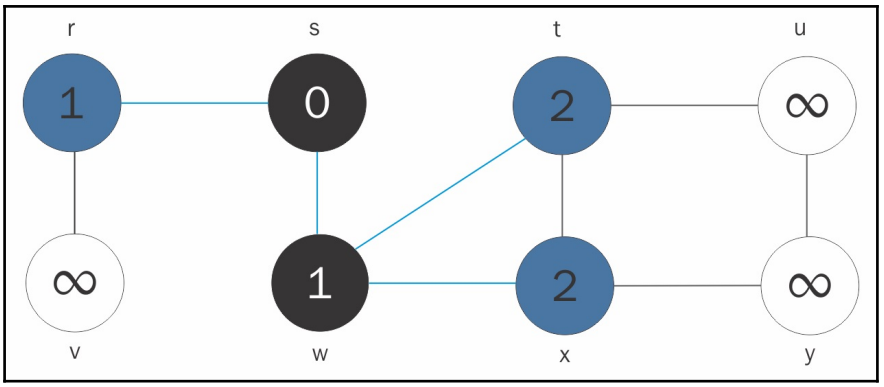

Iron Man
Thor
Hulk
Spider-Man

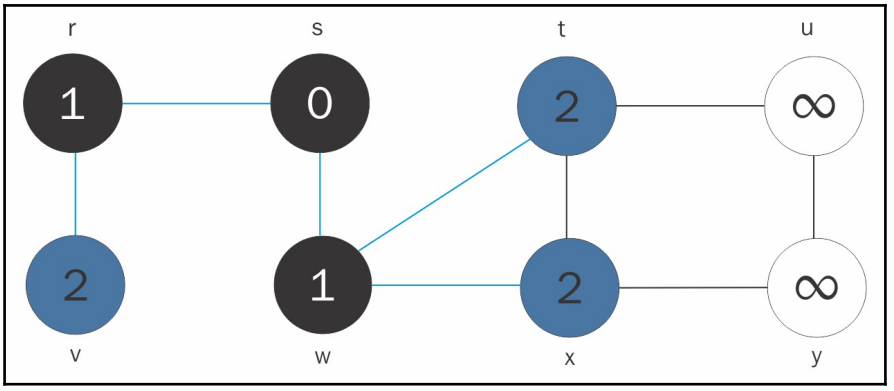

r
s
t
u
∞
∞
∞
∞
∞
∞
∞
∞
v
w
x
y

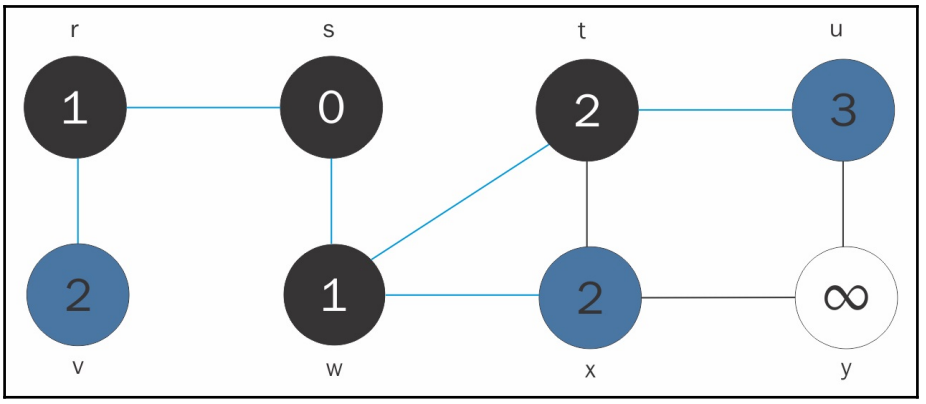

r
s
t
u
∞
∞
∞
∞
∞
∞
∞
∞
v
w
x
y

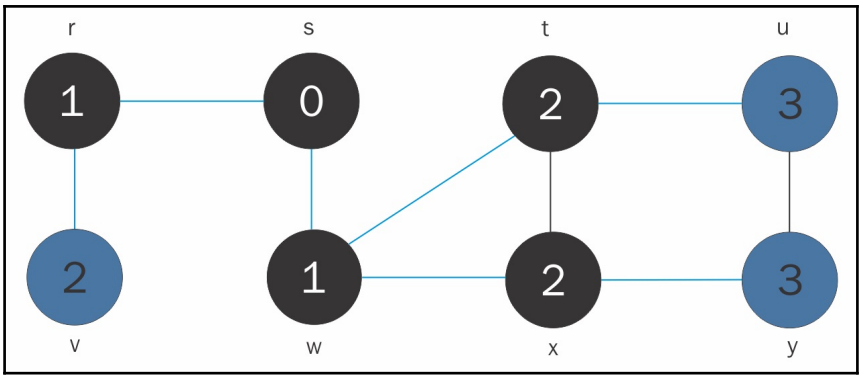

r
s
t
u
∞
0
∞
∞
∞
∞
∞
∞
v
w
x
y

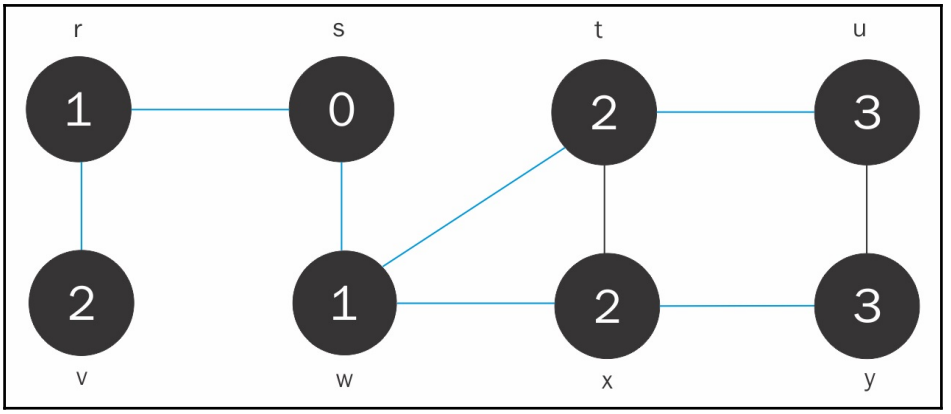

r
s
t
u
1
0
∞
∞
∞
1
∞
∞
v
w
x
y


r
s
t
u
1
0
2
∞
∞
1
2
∞
v
w
x
y


r
s
t
u
1
0
2
∞
2
1
2
∞
v
w
x
y


r
s
t
u
1
0
2
3
2
1
2
∞
v
w
x
y





5983 1165 3836 4361 1282

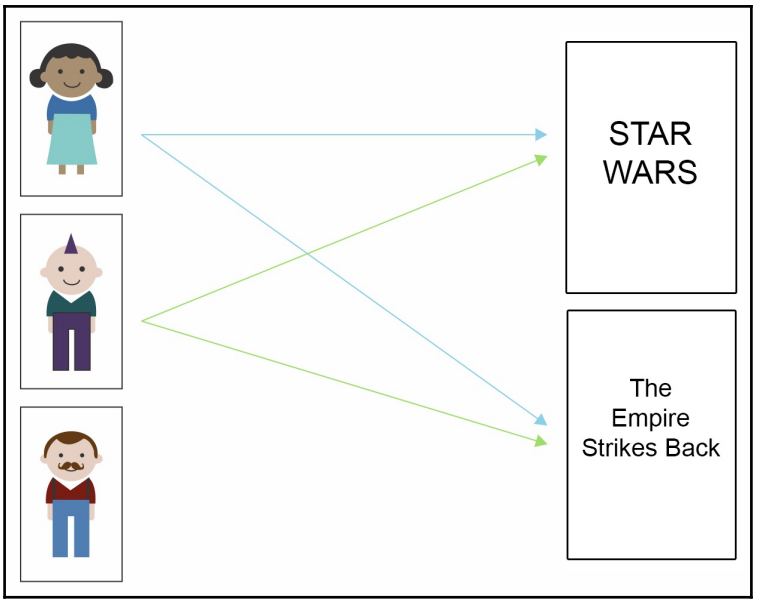
(5983, (1165, 3836, 4361, 1282), 9999, WHITE)

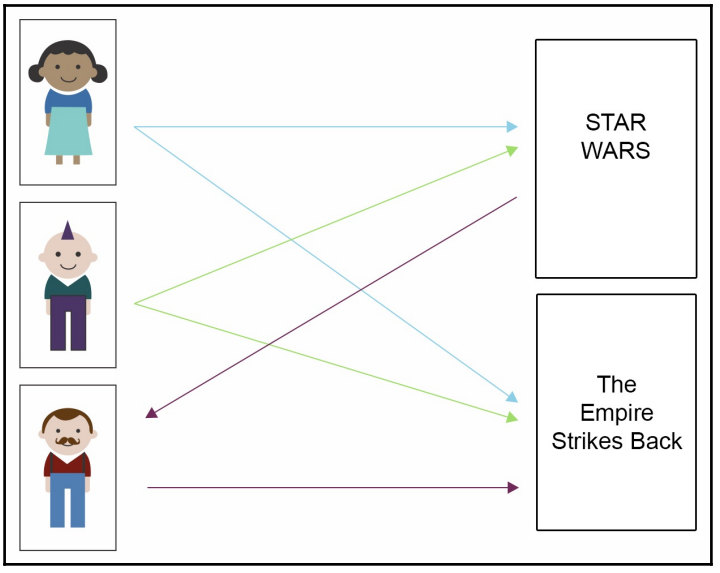
(5983, (1165, 3836, 4361, 1282), 9999, WHITE)

```python
#Boilerplate stuff:
from pyspark import SparkConf, SparkContext

conf = SparkConf().setMaster("local").setAppName("DegreesOfSeparation")
sc = SparkContext(conf = conf)

# The characters we wish to find the degree of separation between:
startCharacterID = 5306 #SpiderMan
targetCharacterID = 14  #ADAM 3,031 (who?)

# Our accumulator, used to signal when we find the target character during
# our BFS traversal.
hitCounter = sc.accumulator(0)

def convertToBFS(line):
    fields = line.split()
    heroID = int(fields[0])
    connections = []
    for connection in fields[1:]:
        connections.append(int(connection))

    color = 'WHITE'
    distance = 9999

    if (heroID == startCharacterID):
        color = 'GRAY'
        distance = 0

    return (heroID, (connections, distance, color))


def createStartingRdd():
    inputFile = sc.textFile("file:///sparkcourse/marvel-graph.txt")
    return inputFile.map(convertToBFS)

def bfsMap(node):
    characterID = node[0]
    data = node[1]
    connections = data[0]
```

```
(User) c:\SparkCourse>spark-submit degrees-of-separation.py
```

```
Running BFS iteration# 1
Processing 8330 values.
Running BFS iteration# 2
Processing 220615 values.
Hit the target character! From 1 different direction(s).

(User) c:\SparkCourse>
```





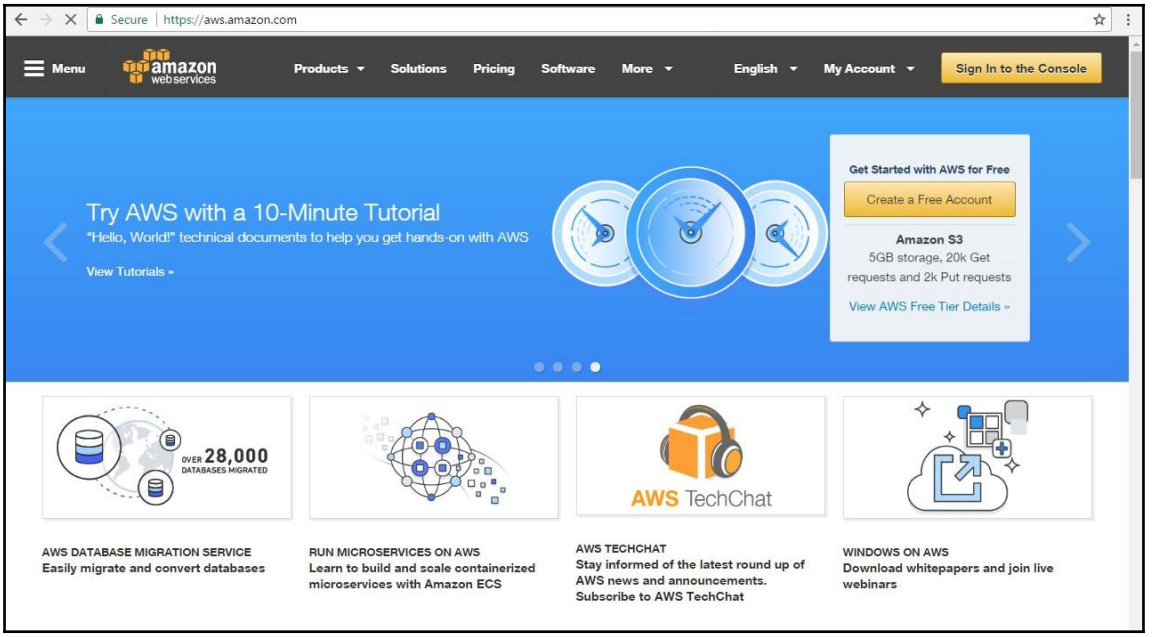

```
(User) c:\SparkCourse>spark-submit movie-similarities.py 50
```

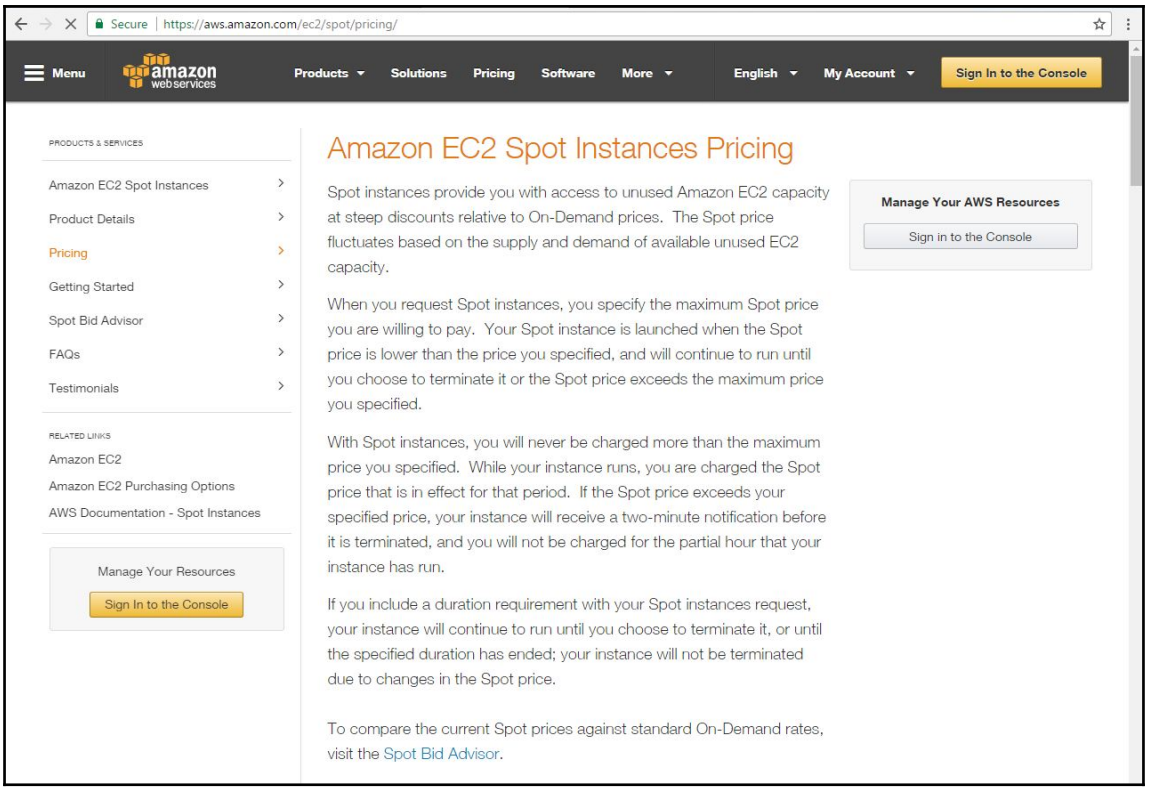
```
Loading movie names...
[Stage 0:>                                                          (0 + 0) / 4]
[Stage 0:>                                                          (0 + 4) / 4]
[Stage 0:============================>                            (2 + 2) / 4]
[Stage 0:=========================================>               (3 + 1) / 4]
[Stage 1:>                                                          (0 + 4) / 4]
[Stage 1:==============>                                          (1 + 3) / 4]
[Stage 1:============================>                            (2 + 2) / 4]
[Stage 1:=========================================>               (3 + 1) / 4]
[Stage 2:>                                                          (0 + 4) / 4]
[Stage 2:=========================================>               (3 + 1) / 4]

[Stage 5:>                                                          (0 + 4) / 4]
[Stage 5:============================>                            (2 + 2) / 4]
[Stage 5:=========================================>               (3 + 1) / 4]

[Stage 8:>                                                          (0 + 4) / 4]
[Stage 8:==============>                                          (1 + 3) / 4]
[Stage 8:=========================================>               (3 + 1) / 4]
```

```
Top 10 similar movies for Star Wars (1977)
Empire Strikes Back, The (1980) score: 0.989552207839   strength: 345
Return of the Jedi (1983)       score: 0.985723086125   strength: 480
Raiders of the Lost Ark (1981)  score: 0.981760098873   strength: 380
20,000 Leagues Under the Sea (1954)     score: 0.97893856055    strength: 68
12 Angry Men (1957)     score: 0.977657612045   strength: 109
Close Shave, A (1995)   score: 0.977594829105   strength: 92
African Queen, The (1951)       score: 0.976469222267   strength: 138
Sting, The (1973)       score: 0.975151293774   strength: 204
Wrong Trousers, The (1993)      score: 0.974868135546   strength: 103
Wallace & Gromit: The Best of Aardman Animation (1996)  score: 0.97418161283
strength: 58

(User) c:\SparkCourse>
```

Chapter 4

.

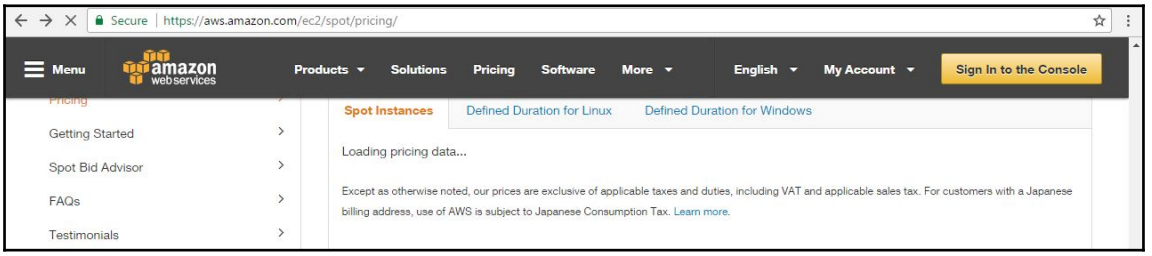

Secure | https://aws.amazon.com
Menu
amazon web services
Products
Solutions
Pricing
Software
More
English
My Account
Sign In to the Console
Try AWS with a 10-Minute Tutorial
"Hello, World!" technical documents to help you get hands-on with AWS
View Tutorials »
Get Started with AWS for Free
Create a Free Account
Amazon S3
5GB storage, 20k Get requests and 2k Put requests
View AWS Free Tier Details »
OVER 28,000 DATABASES MIGRATED
AWS TechChat
AWS DATABASE MIGRATION SERVICE
Easily migrate and convert databases
RUN MICROSERVICES ON AWS
Learn to build and scale containerized microservices with Amazon ECS
AWS TECHCHAT
Stay informed of the latest round up of AWS news and announcements. Subscribe to AWS TechChat
WINDOWS ON AWS
Download whitepapers and join live webinars

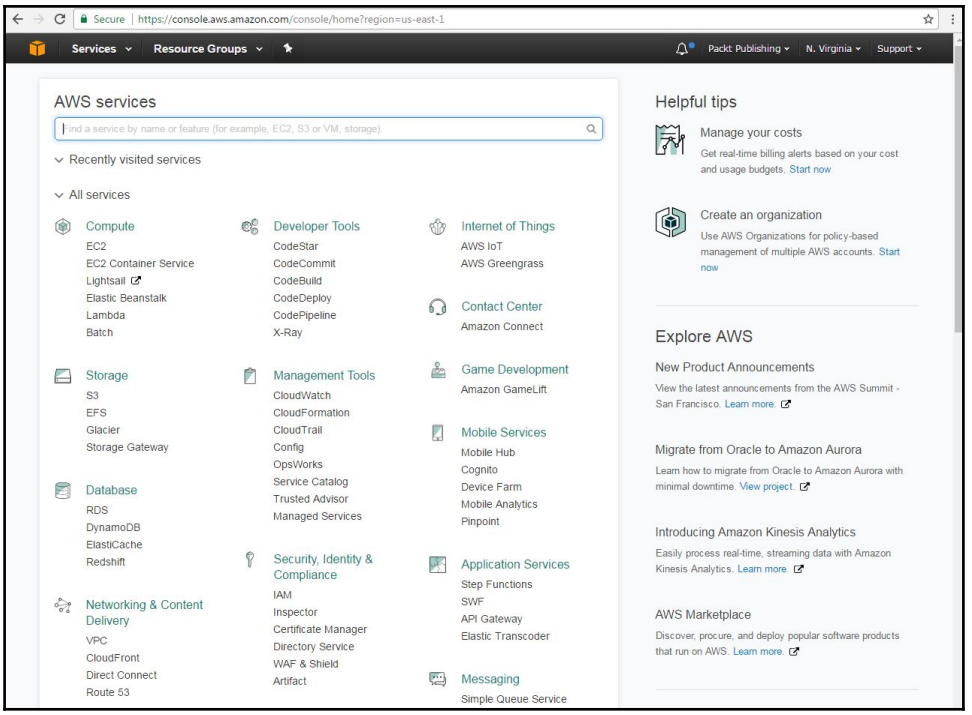

Secure | https://aws.amazon.com/ec2/spot/pricing/
Menu
amazon web services
Products
Solutions
Pricing
Software
More
English
My Account
Sign In to the Console
PRODUCTS & SERVICES
Amazon EC2 Spot Instances
Product Details
Pricing
Getting Started
Spot Bid Advisor
FAQs
Testimonials
RELATED LINKS
Amazon EC2
Amazon EC2 Purchasing Options
AWS Documentation - Spot Instances
Manage Your Resources
Sign In to the Console
Amazon EC2 Spot Instances Pricing
Spot instances provide you with access to unused Amazon EC2 capacity at steep discounts relative to On-Demand prices. The Spot price fluctuates based on the supply and demand of available unused EC2 capacity.
When you request Spot instances, you specify the maximum Spot price you are willing to pay. Your Spot instance is launched when the Spot price is lower than the price you specified, and will continue to run until you choose to terminate it or the Spot price exceeds the maximum price you specified.
With Spot instances, you will never be charged more than the maximum price you specified. While your instance runs, you are charged the Spot price that is in effect for that period. If the Spot price exceeds your specified price, your instance will receive a two-minute notification before it is terminated, and you will not be charged for the partial hour that your instance has run.
If you include a duration requirement with your Spot instances request, your instance will continue to run until you choose to terminate it, or until the specified duration has ended; your instance will not be terminated due to changes in the Spot price.
To compare the current Spot prices against standard On-Demand rates, visit the Spot Bid Advisor.
Manage Your AWS Resources
Sign in to the Console

t2.small	1	Variable	2	EBS Only	$0.026 per Hour
t2.medium	2	Variable	4	EBS Only	$0.052 per Hour
t2.large	2	Variable	8	EBS Only	$0.104 per Hour
m4.large	2	6.5	8	EBS Only	$0.126 per Hour
m4.xlarge	4	13	16	EBS Only	$0.252 per Hour
m4.2xlarge	8	26	32	EBS Only	$0.504 per Hour
m4.4xlarge	16	53.5	64	EBS Only	$1.008 per Hour
m4.10xlarge	40	124.5	160	EBS Only	$2.52 per Hour
m3.medium	1	3	3.75	1 x 4 SSD	$0.067 per Hour
m3.large	2	6.5	7.5	1 x 32 SSD	$0.133 per Hour
m3.xlarge	4	13	15	2 x 40 SSD	$0.266 per Hour
m3.2xlarge	8	26	30	2 x 80 SSD	$0.532 per Hour

Compute Optimized - Current Generation

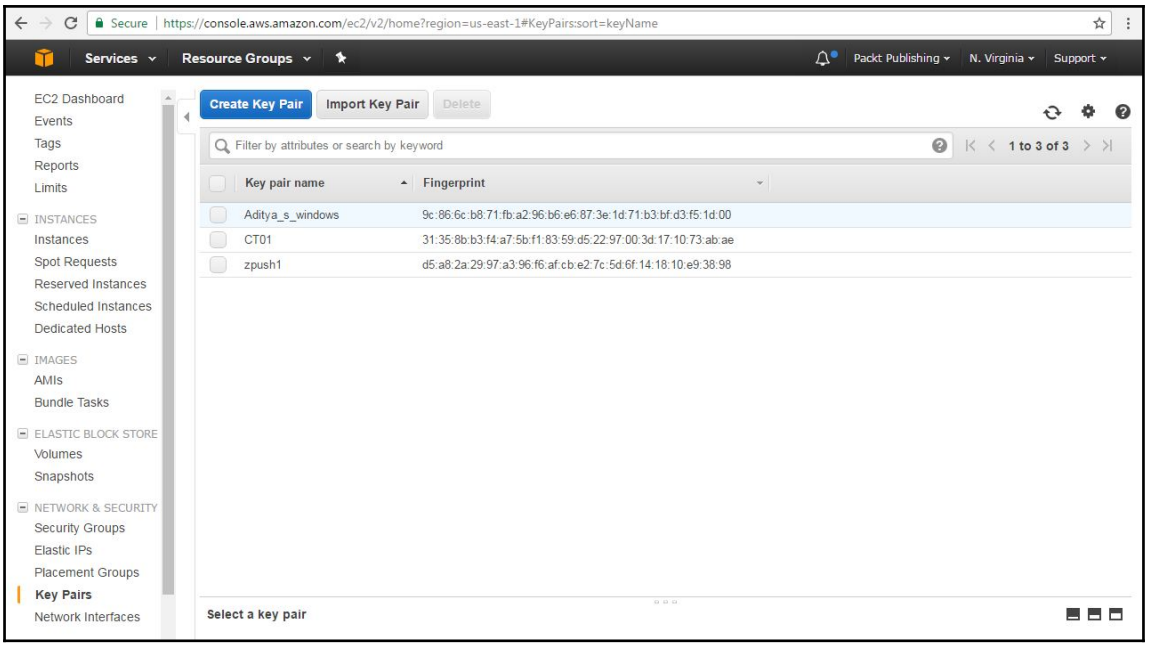
m3.xlarge	4	13	15	2 x 40 SSD	$0.266 per Hour

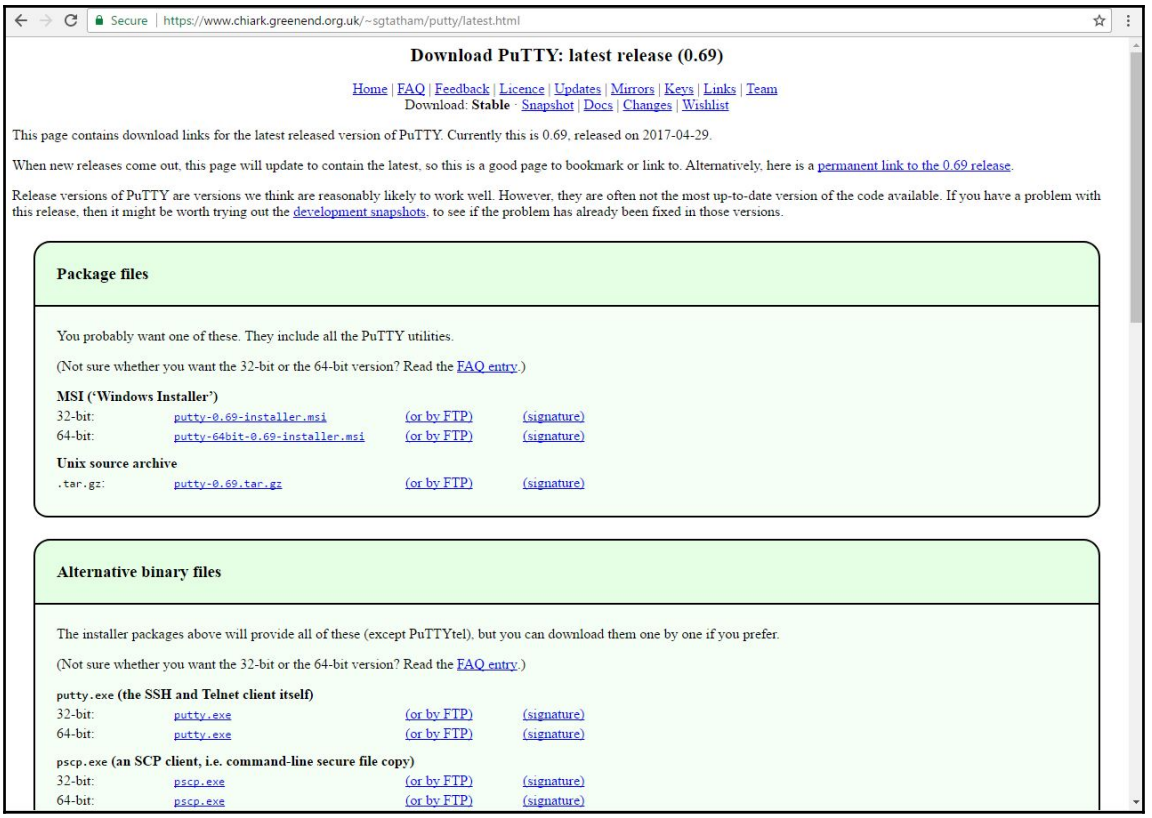


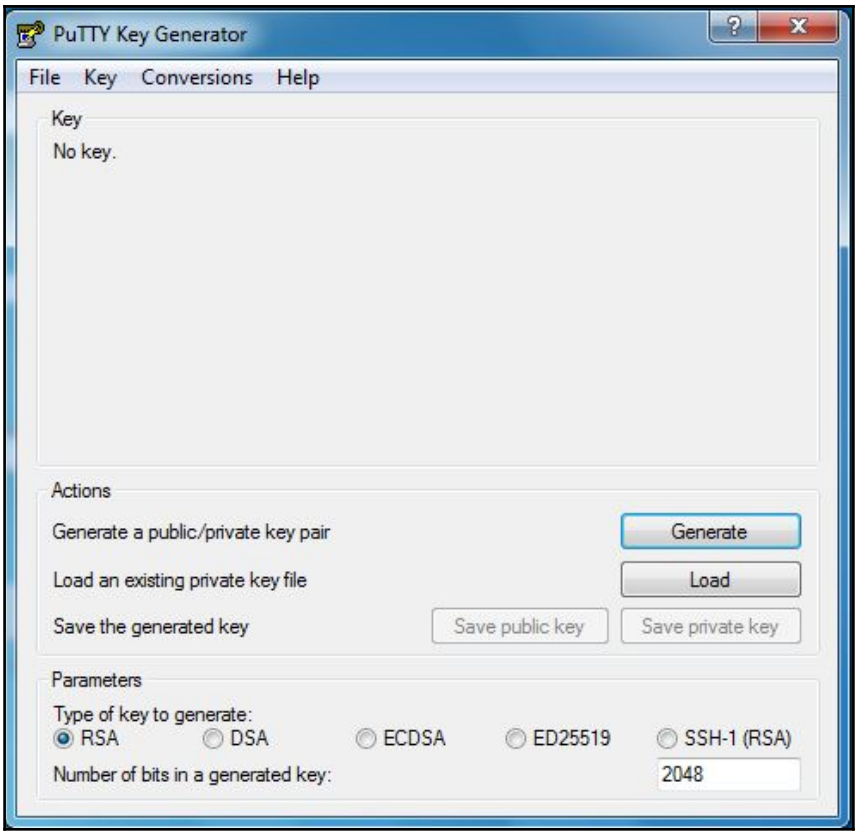

Secure | https://console.aws.amazon.com/console/home?region=us-east-1
Services
Resource Groups
Packt Publishing
N. Virginia
Support
AWS services
Find a service by name or feature (for example, EC2, S3 or VM, storage).
Recently visited services
All services
Compute
EC2
EC2 Container Service
Lightsail
Elastic Beanstalk
Lambda
Batch
Developer Tools
CodeStar
CodeCommit
CodeBuild
CodeDeploy
CodePipeline
X-Ray
Internet of Things
AWS IoT
AWS Greengrass
Contact Center
Amazon Connect
Storage
S3
EFS
Glacier
Storage Gateway
Management Tools
CloudWatch
CloudFormation
CloudTrail
Config
OpsWorks
Service Catalog
Trusted Advisor
Managed Services
Game Development
Amazon GameLift
Mobile Services
Mobile Hub
Cognito
Device Farm
Mobile Analytics
Pinpoint
Database
RDS
DynamoDB
ElastiCache
Redshift
Security, Identity & Compliance
IAM
Inspector
Certificate Manager
Directory Service
WAF & Shield
Artifact
Application Services
Step Functions
SWF
API Gateway
Elastic Transcoder
Networking & Content Delivery
VPC
CloudFront
Direct Connect
Route 53
Messaging
Simple Queue Service
Helpful tips
Manage your costs
Get real-time billing alerts based on your cost and usage budgets. Start now
Create an organization
Use AWS Organizations for policy-based management of multiple AWS accounts. Start now
Explore AWS
New Product Announcements
View the latest announcements from the AWS Summit - San Francisco. Learn more.
Migrate from Oracle to Amazon Aurora
Learn how to migrate from Oracle to Amazon Aurora with minimal downtime. View project.
Introducing Amazon Kinesis Analytics
Easily process real-time, streaming data with Amazon Kinesis Analytics. Learn more.
AWS Marketplace
Discover, procure, and deploy popular software products that run on AWS. Learn more.

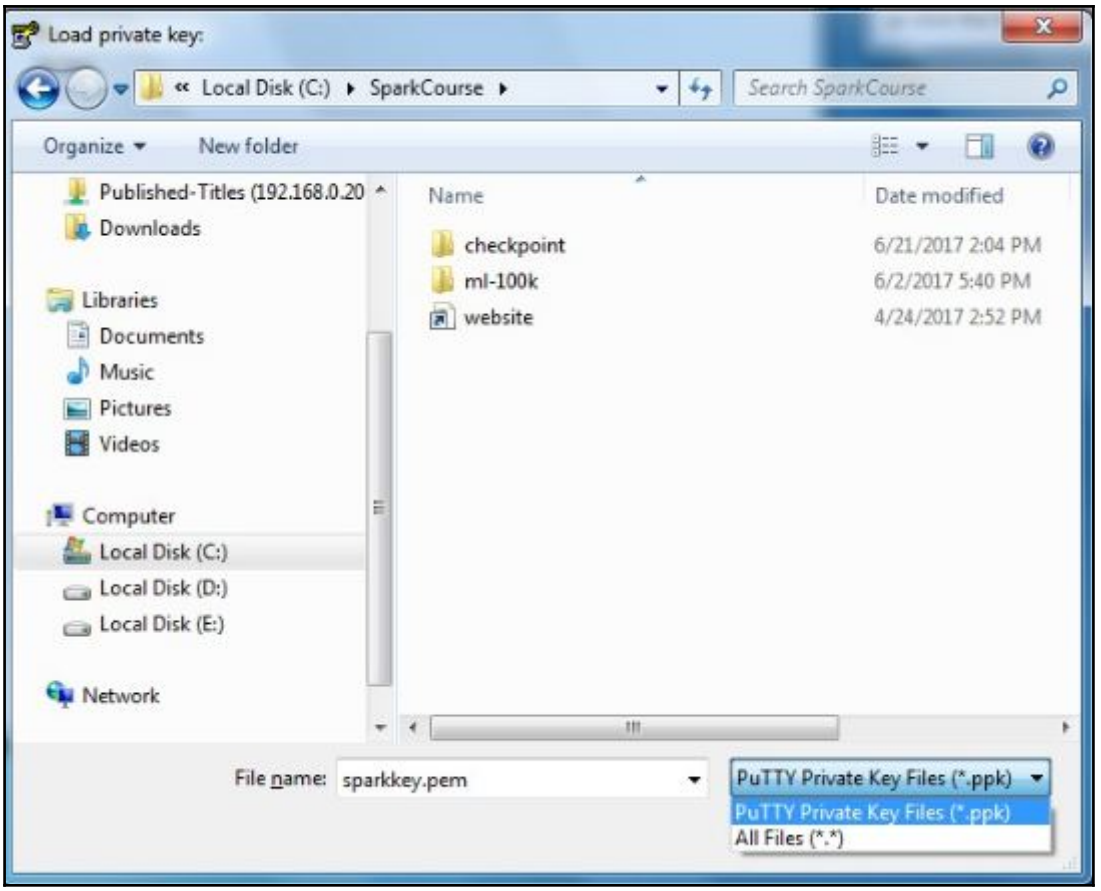

Analytics
Athena
EMR

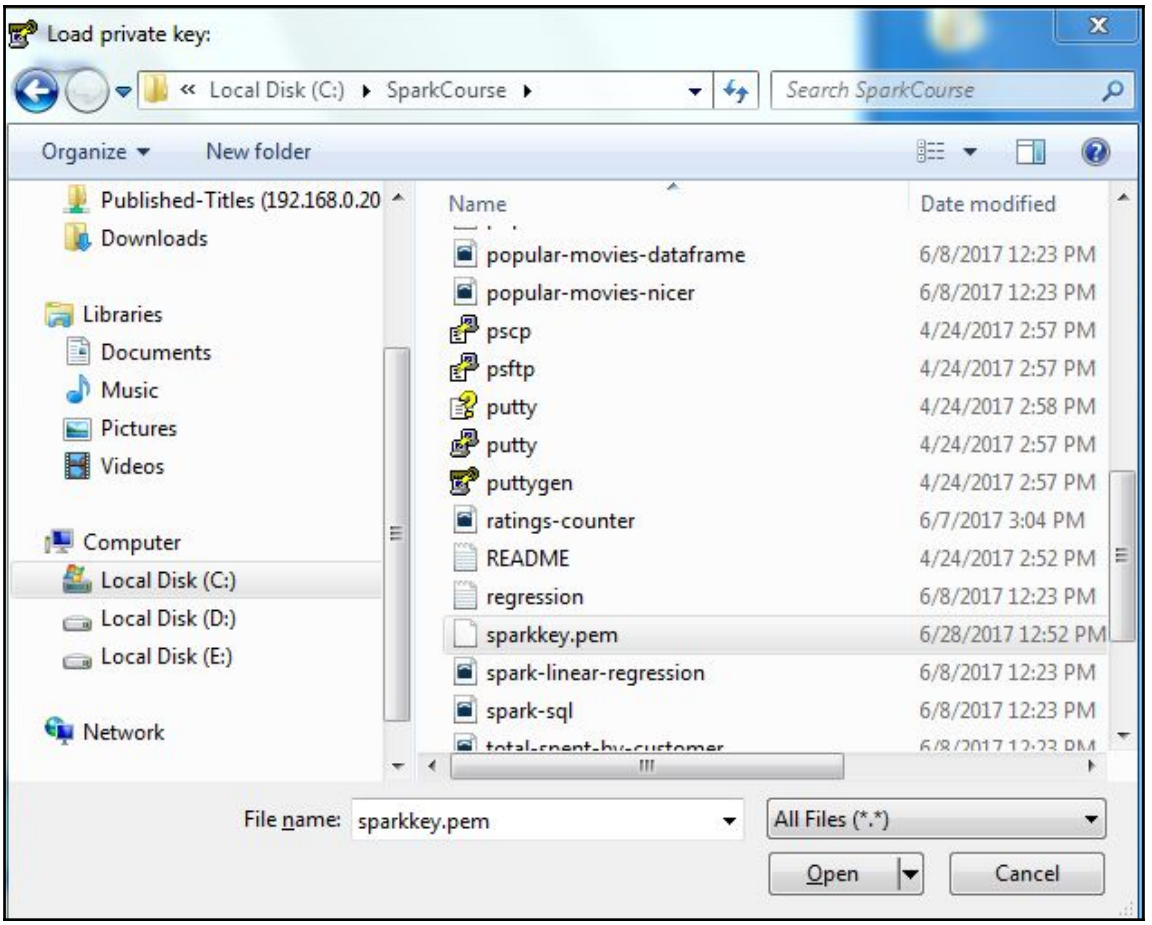

Compute
EC2

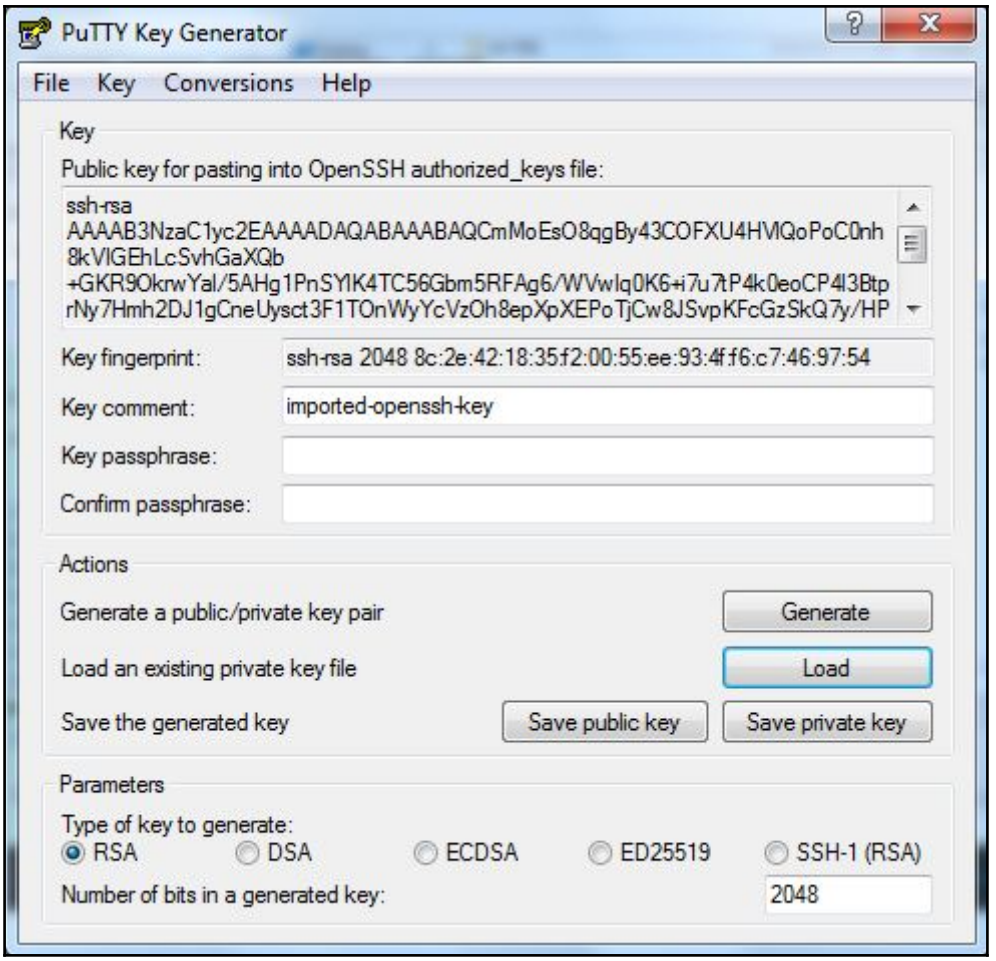

Secure | https://console.aws.amazon.com/ec2/v2/home?region=us-east-1#KeyPairs:sort=keyName
Services
Resource Groups
Packt Publishing
N. Virginia
Support
EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Spot Requests
Reserved Instances
Scheduled Instances
Dedicated Hosts
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
NETWORK & SECURITY
Security Groups
Elastic IPs
Placement Groups
Key Pairs
Network Interfaces
Create Key Pair
Import Key Pair
Delete
Filter by attributes or search by keyword
1 to 3 of 3
Key pair name
Fingerprint
Aditya_s_windows
9c:86:6c:b8:71:fb:a2:96:b6:e6:87:3e:1d:71:b3:bf:d3:f5:1d:00
CT01
31:35:8b:b3:f4:a7:5b:f1:83:59:d5:22:97:00:3d:17:10:73:ab:ae
zpush1
d5:a8:2a:29:97:a3:96:f6:af:cb:e2:7c:5d:6f:14:18:10:e9:38:98
Select a key pair

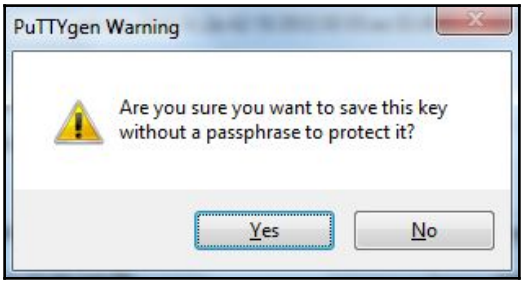

Secure | https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html
Download PuTTY: latest release (0.69)
Home | FAQ | Feedback | Licence | Updates | Mirrors | Keys | Links | Team
Download: Stable · Snapshot | Docs | Changes | Wishlist
This page contains download links for the latest released version of PuTTY. Currently this is 0.69, released on 2017-04-29.
When new releases come out, this page will update to contain the latest, so this is a good page to bookmark or link to. Alternatively, here is a permanent link to the 0.69 release.
Release versions of PuTTY are versions we think are reasonably likely to work well. However, they are often not the most up-to-date version of the code available. If you have a problem with this release, then it might be worth trying out the development snapshots, to see if the problem has already been fixed in those versions.
Package files
You probably want one of these. They include all the PuTTY utilities.
(Not sure whether you want the 32-bit or the 64-bit version? Read the FAQ entry.)
MSI ('Windows Installer')
32-bit: putty-0.69-installer.msi (or by FTP) (signature)
64-bit: putty-64bit-0.69-installer.msi (or by FTP) (signature)
Unix source archive
.tar.gz: putty-0.69.tar.gz (or by FTP) (signature)
Alternative binary files
The installer packages above will provide all of these (except PuTTYtel), but you can download them one by one if you prefer.
(Not sure whether you want the 32-bit or the 64-bit version? Read the FAQ entry.)
putty.exe (the SSH and Telnet client itself)
32-bit: putty.exe (or by FTP) (signature)
64-bit: putty.exe (or by FTP) (signature)
pscp.exe (an SCP client, i.e. command-line secure file copy)
32-bit: pscp.exe (or by FTP) (signature)
64-bit: pscp.exe (or by FTP) (signature)

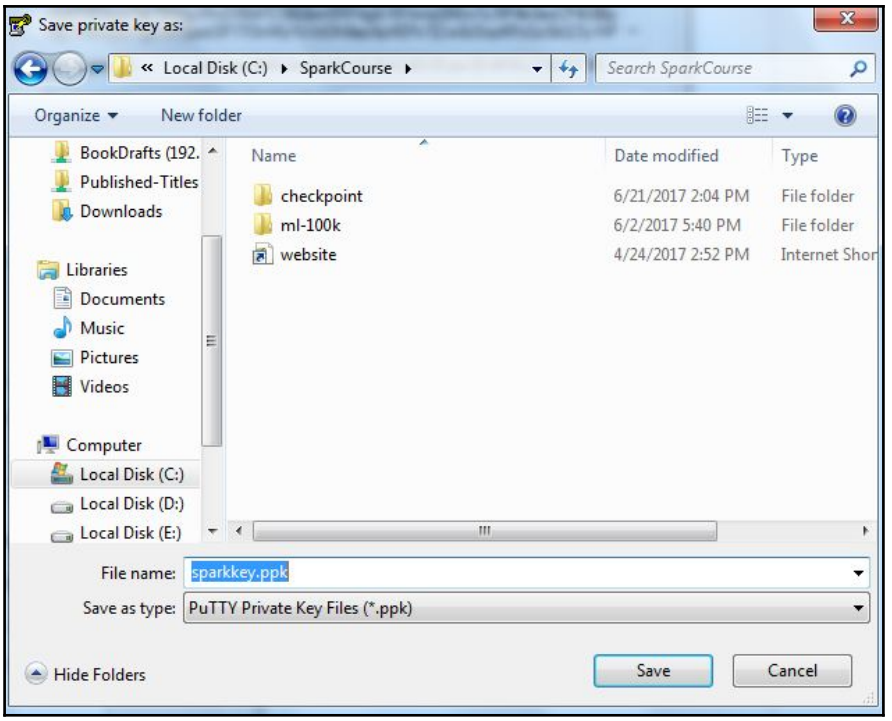

PuTTY Key Generator
File Key Conversions Help
Key
No key.
Actions
Generate a public/private key pair
Generate
Load an existing private key file
Load
Save the generated key
Save public key
Save private key
Parameters
Type of key to generate:
RSA DSA ECDSA ED25519 SSH-1 (RSA)
Number of bits in a generated key:
2048

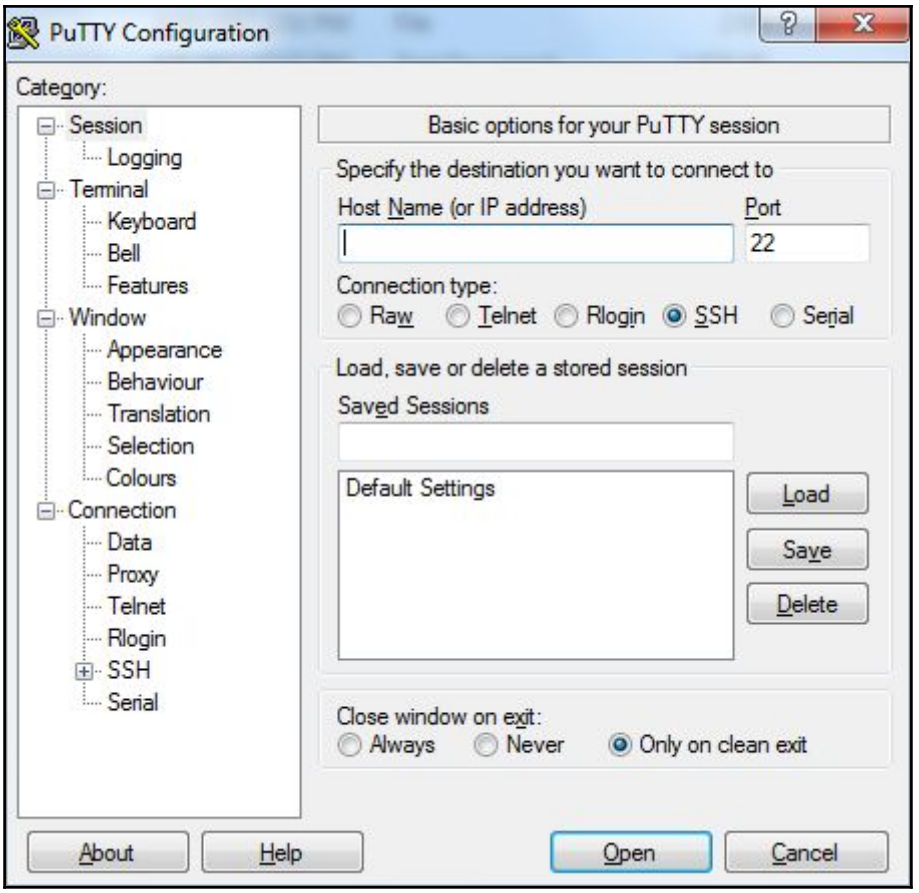

Load private key:
Local Disk (C:) ▸ SparkCourse ▸
Search SparkCourse
Organize
New folder
Published-Titles (192.168.0.20
Downloads
Libraries
Documents
Music
Pictures
Videos
Computer
Local Disk (C:)
Local Disk (D:)
Local Disk (E:)
Network
Name
Date modified
checkpoint
6/21/2017 2:04 PM
ml-100k
6/2/2017 5:40 PM
website
4/24/2017 2:52 PM
File name: sparkkey.pem
PuTTY Private Key Files (*.ppk)
PuTTY Private Key Files (*.ppk)
All Files (*.*)

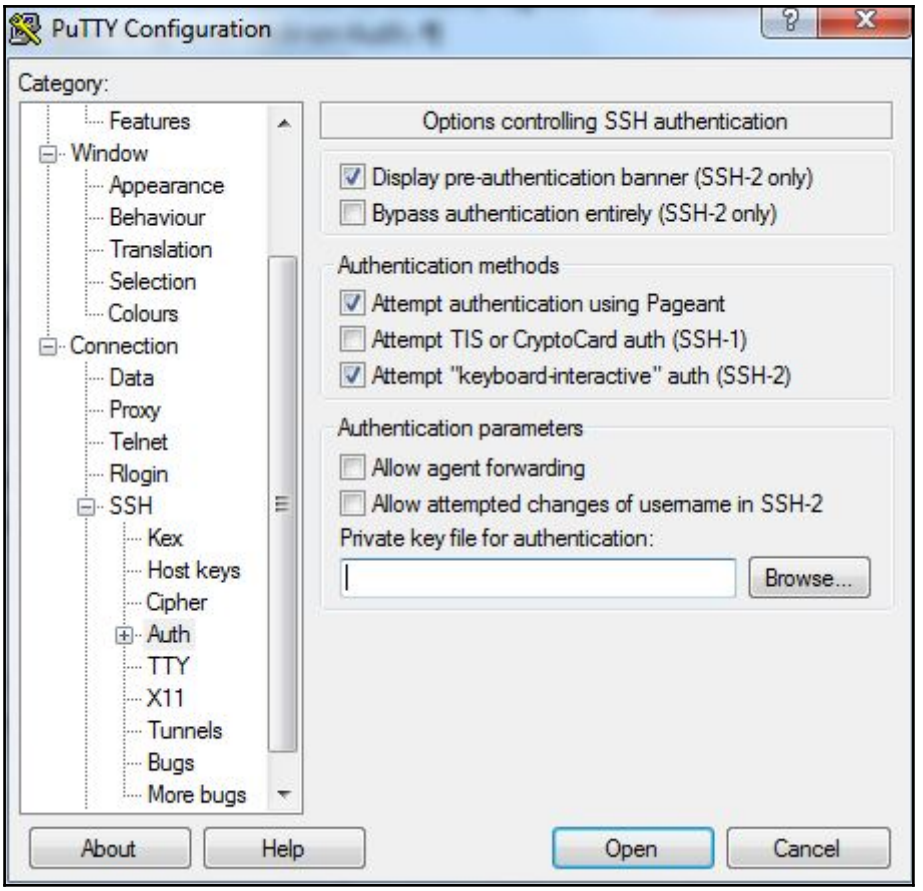

Load private key:
Local Disk (C:) ▸ SparkCourse ▸
Search SparkCourse
Organize
New folder
Published-Titles (192.168.0.20
Downloads
Libraries
Documents
Music
Pictures
Videos
Computer
Local Disk (C:)
Local Disk (D:)
Local Disk (E:)
Network
Name
Date modified
popular-movies-dataframe 6/8/2017 12:23 PM
popular-movies-nicer 6/8/2017 12:23 PM
pscp 4/24/2017 2:57 PM
psftp 4/24/2017 2:57 PM
putty 4/24/2017 2:58 PM
putty 4/24/2017 2:57 PM
puttygen 4/24/2017 2:57 PM
ratings-counter 6/7/2017 3:04 PM
README 4/24/2017 2:52 PM
regression 6/8/2017 12:23 PM
sparkkey.pem 6/28/2017 12:52 PM
spark-linear-regression 6/8/2017 12:23 PM
spark-sql 6/8/2017 12:23 PM
File name: sparkkey.pem
All Files (*.*)
Open
Cancel

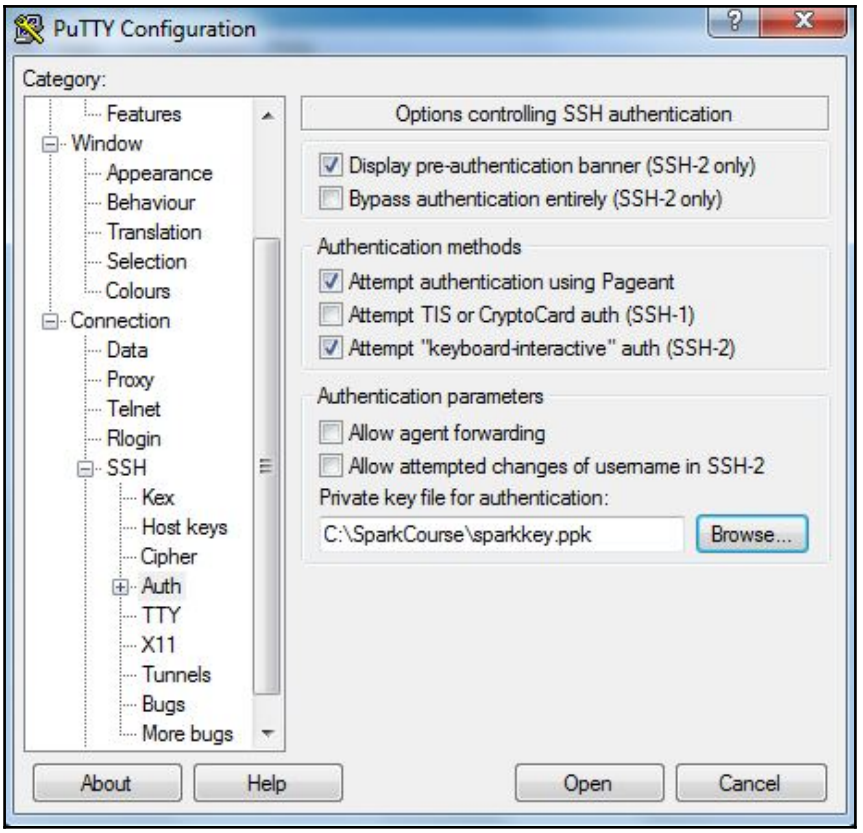

PuTTY Key Generator
File Key Conversions Help
Key
Public key for pasting into OpenSSH authorized_keys file:
ssh-rsa
AAAAB3NzaC1yc2EAAAADAQABAAABAQCmMoEsO8qgBy43COFXU4HVlQoPoC0nh
8kVlGEhLcSvhGaXQb
+GKR9OkrwYal/5AHg1PnSYlK4TC56Gbm5RFAg6/WVwlq0K6+i7u7tP4k0eoCP4l3Btp
rNy7Hmh2DJ1gCneUysct3F1TOnWyYcVzOh8epXpXEPoTjCw8JSvpKFcGzSkQ7y/HP
Key fingerprint: ssh-rsa 2048 8c:2e:42:18:35:f2:00:55:ee:93:4f:f6:c7:46:97:54
Key comment: imported-openssh-key
Key passphrase:
Confirm passphrase:
Actions
Generate a public/private key pair Generate
Load an existing private key file Load
Save the generated key Save public key Save private key
Parameters
Type of key to generate:
RSA DSA ECDSA ED25519 SSH-1 (RSA)
Number of bits in a generated key: 2048

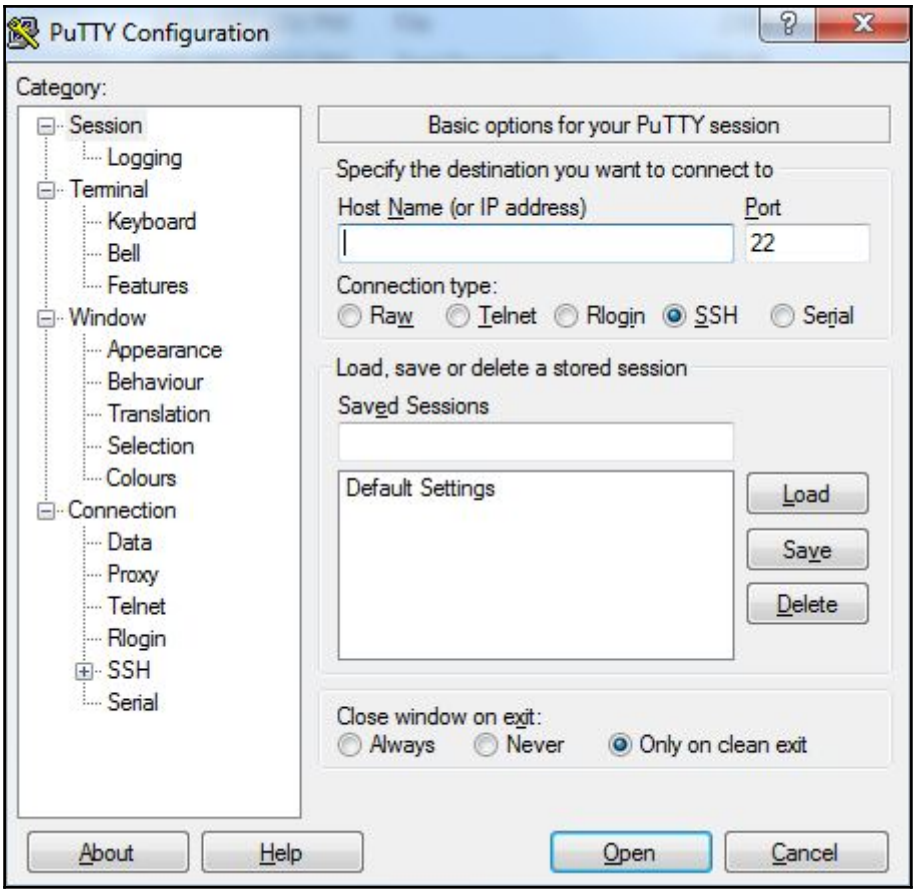

PuTTYgen Warning
Are you sure you want to save this key without a passphrase to protect it?
Yes
No


Save private key as:
Local Disk (C:) ▸ SparkCourse ▸
Search SparkCourse
Organize
New folder
BookDrafts (192.
Published-Titles
Downloads
Libraries
Documents
Music
Pictures
Videos
Computer
Local Disk (C:)
Local Disk (D:)
Local Disk (E:)
Name
Date modified
Type
checkpoint
6/21/2017 2:04 PM
File folder
ml-100k
6/2/2017 5:40 PM
File folder
website
4/24/2017 2:52 PM
Internet Shor
File name:
sparkkey.ppk
Save as type:
PuTTY Private Key Files (*.ppk)
Hide Folders
Save
Cancel

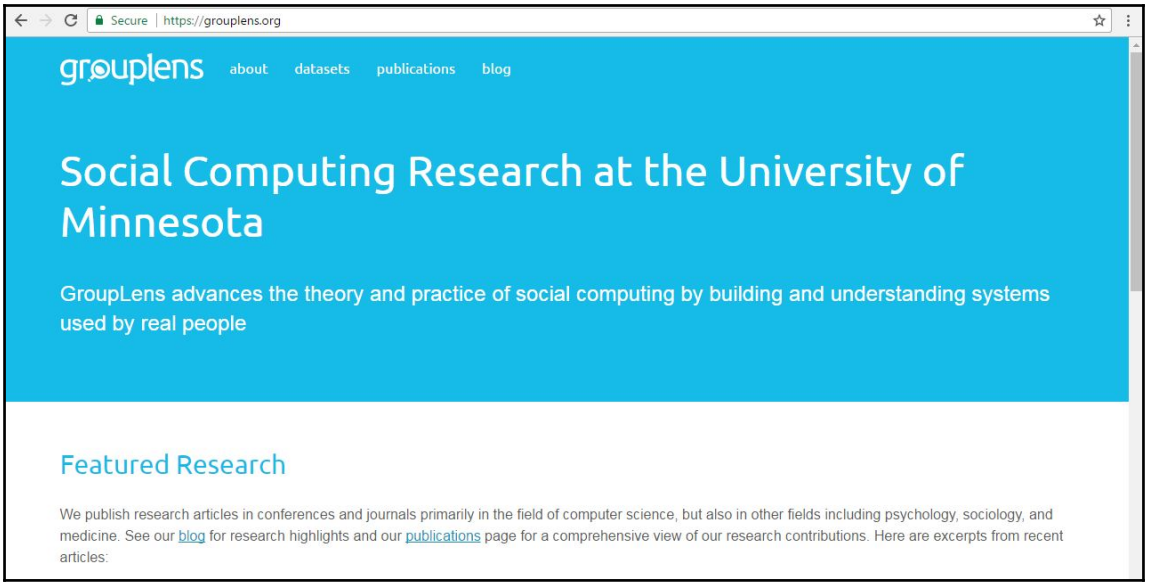

PuTTY Configuration
Category:
Session
Logging
Terminal
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Serial
Basic options for your PuTTY session
Specify the destination you want to connect to
Host Name (or IP address)
Port
22
Connection type:
Raw
Telnet
Rlogin
SSH
Serial
Load, save or delete a stored session
Saved Sessions
Default Settings
Load
Save
Delete
Close window on exit:
Always
Never
Only on clean exit
About
Help
Open
Cancel


PuTTY Configuration
Category:
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Kex
Host keys
Cipher
Auth
TTY
X11
Tunnels
Bugs
More bugs
Options controlling SSH authentication
Display pre-authentication banner (SSH-2 only)
Bypass authentication entirely (SSH-2 only)
Authentication methods
Attempt authentication using Pageant
Attempt TIS or CryptoCard auth (SSH-1)
Attempt "keyboard-interactive" auth (SSH-2)
Authentication parameters
Allow agent forwarding
Allow attempted changes of username in SSH-2
Private key file for authentication:
Browse...
About
Help
Open
Cancel


PuTTY Configuration
Category:
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Kex
Host keys
Cipher
Auth
TTY
X11
Tunnels
Bugs
More bugs
Options controlling SSH authentication
Display pre-authentication banner (SSH-2 only)
Bypass authentication entirely (SSH-2 only)
Authentication methods
Attempt authentication using Pageant
Attempt TIS or CryptoCard auth (SSH-1)
Attempt "keyboard-interactive" auth (SSH-2)
Authentication parameters
Allow agent forwarding
Allow attempted changes of username in SSH-2
Private key file for authentication:
C:\SparkCourse\sparkkey.ppk
Browse...
About
Help
Open
Cancel


PuTTY Configuration
Category:
Session
Logging
Terminal
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Serial
Basic options for your PuTTY session
Specify the destination you want to connect to
Host Name (or IP address)
Port
22
Connection type:
Raw
Telnet
Rlogin
SSH
Serial
Load, save or delete a stored session
Saved Sessions
Default Settings
Load
Save
Delete
Close window on exit:
Always
Never
Only on clean exit
About
Help
Open
Cancel

```python
import sys
from pyspark import SparkConf, SparkContext
from math import sqrt

#To run on EMR successfully + output results for Star Wars:
#aws s3 cp s3://sundog-spark/MovieSimilarities1M.py ./
#aws s3 sp c3://sundog-spark/ml-1m/movies.dat ./
#spark-submit --executor-memory 1g MovieSimilarities1M.py 260

def loadMovieNames():
    movieNames = {}
    with open("movies.dat") as f:
        for line in f:
            fields = line.split("::")
            movieNames[int(fields[0])] = fields[1].decode('ascii', 'ignore')
    return movieNames

def makePairs((user, ratings)):
    (movie1, rating1) = ratings[0]
    (movie2, rating2) = ratings[1]
    return ((movie1, movie2), (rating1, rating2))
```



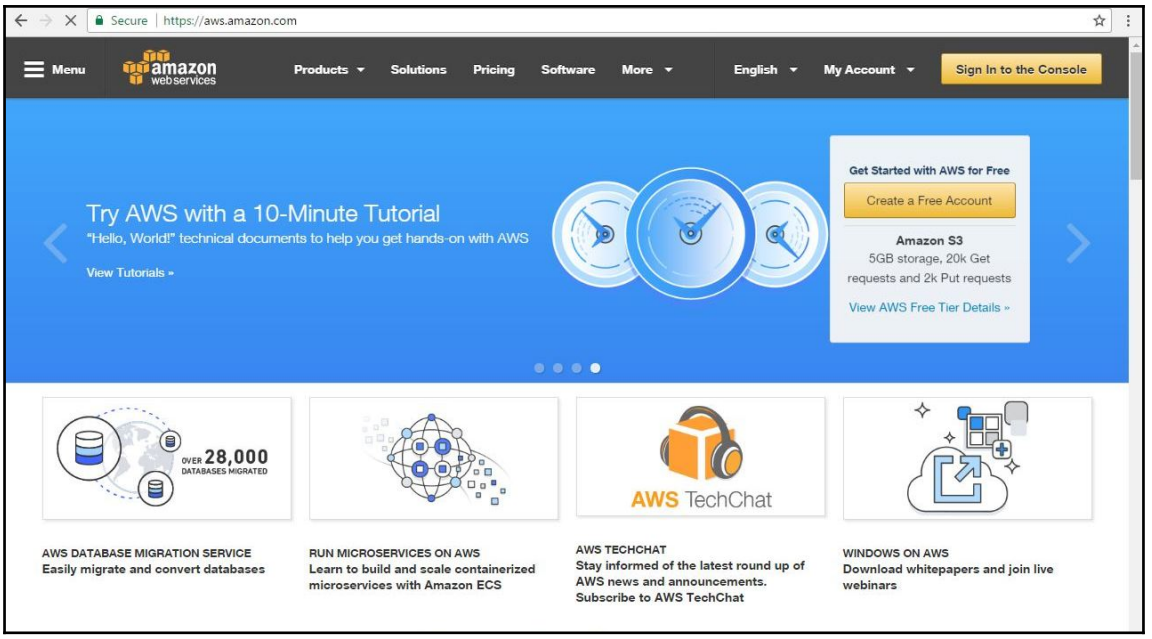
MovieLens 1M Dataset

Stable benchmark dataset. 1 million ratings from 6000 users on 4000 movies. Released 2/2003.

- README.txt
- ml-1m.zip (size: 6 MB, checksum)

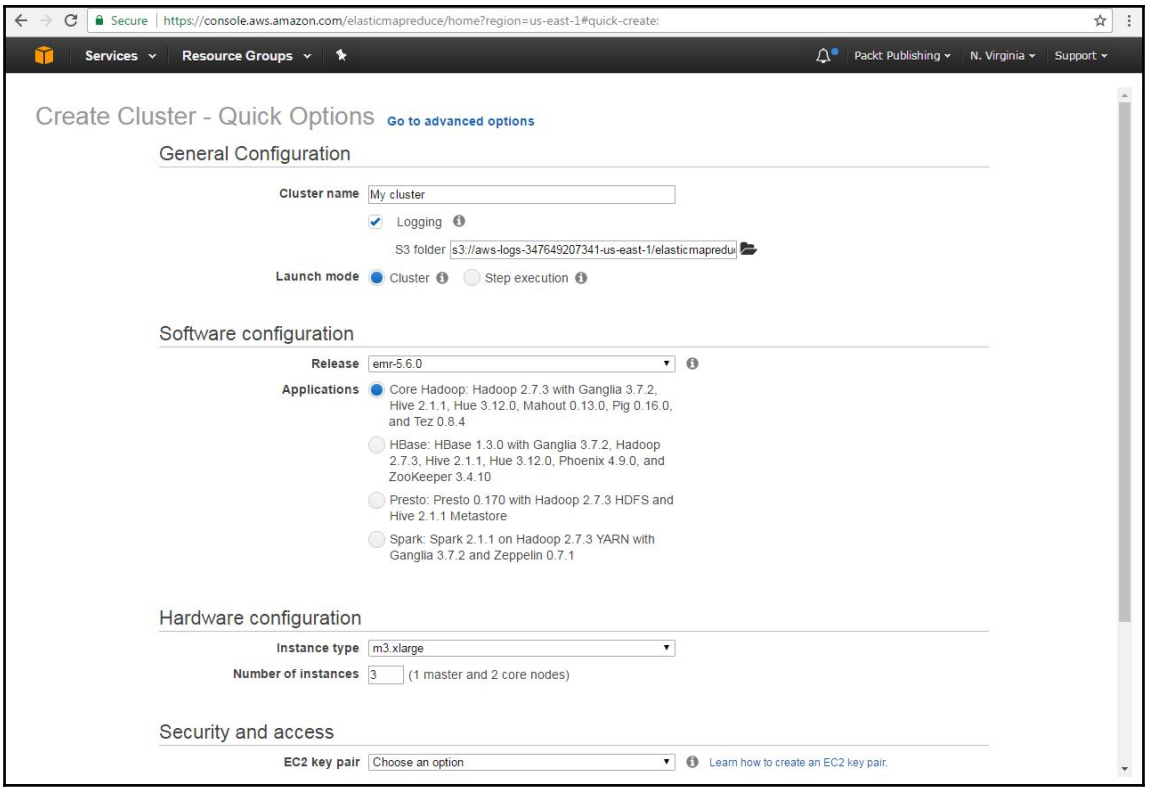
Permalink: http://grouplens.org/datasets/movielens/1m/

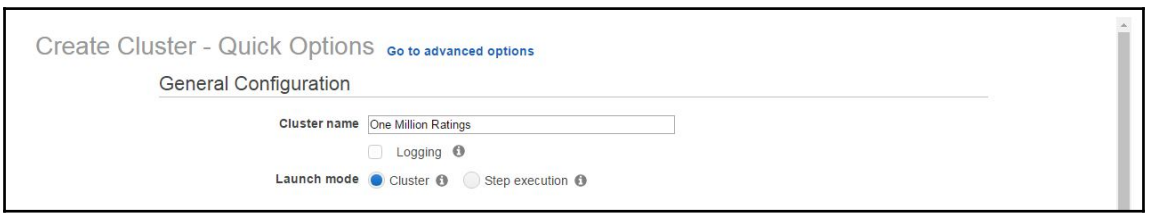
```python
import sys
from pyspark import SparkConf, SparkContext
from math import sqrt

#To run on EMR successfully + output results for Star Wars:
#aws s3 cp s3://sundog-spark/MovieSimilarities1M.py ./
#aws s3 sp c3://sundog-spark/ml-1m/movies.dat ./
#spark-submit --executor-memory 1g MovieSimilarities1M.py 260

def loadMovieNames():
    movieNames = {}
    with open("movies.dat") as f:
        for line in f:
            fields = line.split("::")
            movieNames[int(fields[0])] = fields[1].decode('ascii', 'ignore')
    return movieNames

def makePairs((user, ratings)):
    (movie1, rating1) = ratings[0]
    (movie2, rating2) = ratings[1]
    return ((movie1, movie2), (rating1, rating2))

def filterDuplicates( (userID, ratings) ):
    (movie1, rating1) = ratings[0]
    (movie2, rating2) = ratings[1]
    return movie1 < movie2

def computeCosineSimilarity(ratingPairs):
    numPairs = 0
    sum_xx = sum_yy = sum_xy = 0
    for ratingX, ratingY in ratingPairs:
        sum_xx += ratingX * ratingX
        sum_yy += ratingY * ratingY
        sum_xy += ratingX * ratingY
        numPairs += 1

    numerator = sum_xy
    denominator = sqrt(sum_xx) * sqrt(sum_yy)

    score = 0
    if (denominator):
        score = (numerator / (float(denominator)))

    return (score, numPairs)


conf = SparkConf()
sc = SparkContext(conf = conf)
```

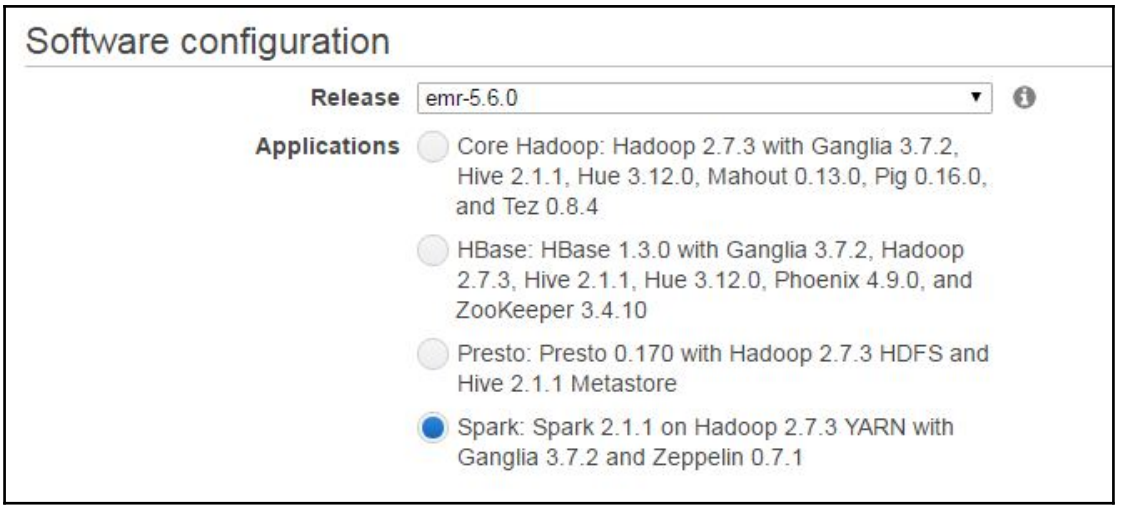

Secure | https://aws.amazon.com
Menu
amazon web services
Products
Solutions
Pricing
Software
More
English
My Account
Sign In to the Console
Try AWS with a 10-Minute Tutorial
"Hello, World!" technical documents to help you get hands-on with AWS
View Tutorials »
Get Started with AWS for Free
Create a Free Account
Amazon S3
5GB storage, 20k Get requests and 2k Put requests
View AWS Free Tier Details »
OVER 28,000 DATABASES MIGRATED
AWS TechChat
AWS DATABASE MIGRATION SERVICE
Easily migrate and convert databases
RUN MICROSERVICES ON AWS
Learn to build and scale containerized microservices with Amazon ECS
AWS TECHCHAT
Stay informed of the latest round up of AWS news and announcements. Subscribe to AWS TechChat
WINDOWS ON AWS
Download whitepapers and join live webinars

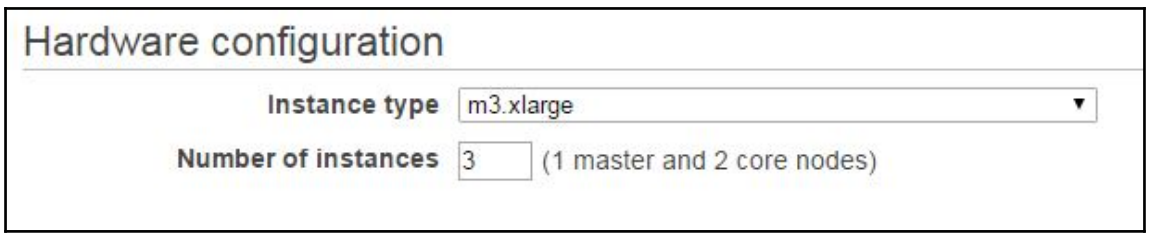

Analytics
Athena
EMR

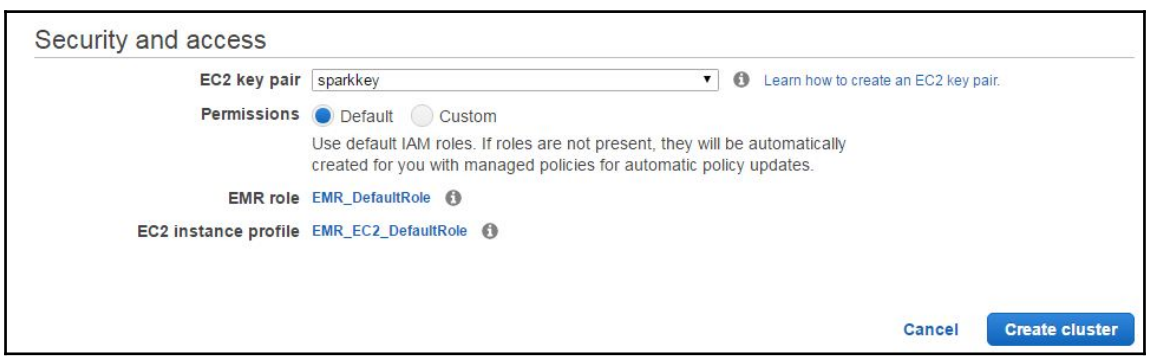

Welcome to Amazon Elastic MapReduce
Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.
You do not appear to have any clusters. Create one now:
Create cluster

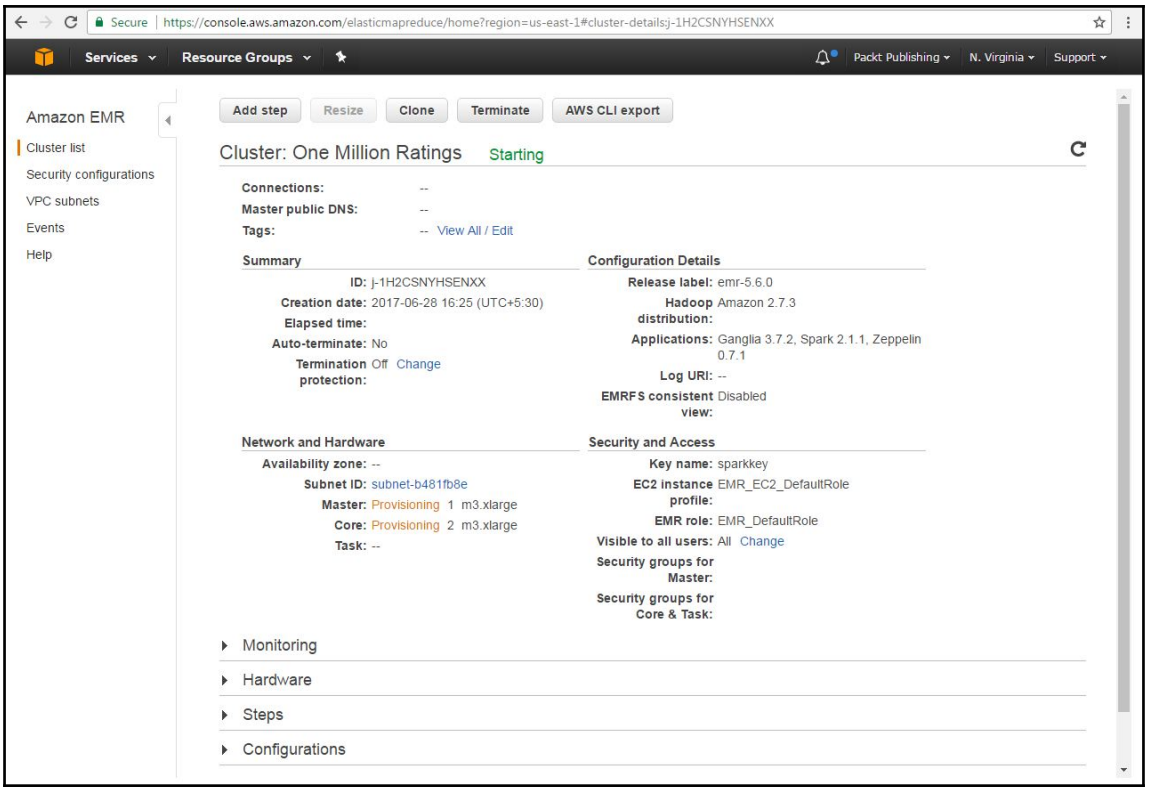

Secure | https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#quick-create:
Services
Resource Groups
Packt Publishing
N. Virginia
Support
Create Cluster - Quick Options
Go to advanced options
General Configuration
Cluster name
My cluster
Logging
S3 folder
s3://aws-logs-347649207341-us-east-1/elasticmapredu
Launch mode
Cluster
Step execution
Software configuration
Release
emr-5.6.0
Applications
Core Hadoop: Hadoop 2.7.3 with Ganglia 3.7.2, Hive 2.1.1, Hue 3.12.0, Mahout 0.13.0, Pig 0.16.0, and Tez 0.8.4
HBase: HBase 1.3.0 with Ganglia 3.7.2, Hadoop 2.7.3, Hive 2.1.1, Hue 3.12.0, Phoenix 4.9.0, and ZooKeeper 3.4.10
Presto: Presto 0.170 with Hadoop 2.7.3 HDFS and Hive 2.1.1 Metastore
Spark: Spark 2.1.1 on Hadoop 2.7.3 YARN with Ganglia 3.7.2 and Zeppelin 0.7.1
Hardware configuration
Instance type
m3.xlarge
Number of instances
3
(1 master and 2 core nodes)
Security and access
EC2 key pair
Choose an option
Learn how to create an EC2 key pair.


Create Cluster - Quick Options
Go to advanced options
General Configuration
Cluster name
One Million Ratings
Logging
Launch mode
Cluster
Step execution

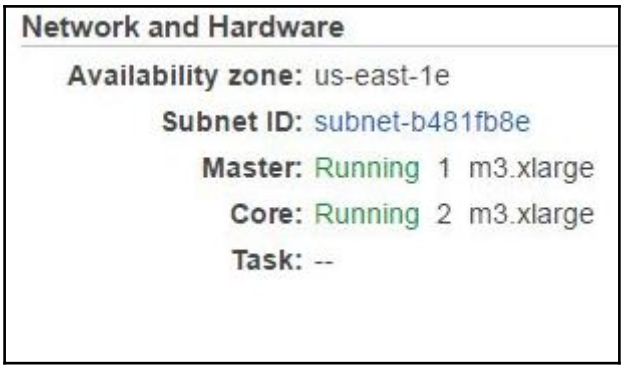

Software configuration
Release
emr-5.6.0
Applications
Core Hadoop: Hadoop 2.7.3 with Ganglia 3.7.2, Hive 2.1.1, Hue 3.12.0, Mahout 0.13.0, Pig 0.16.0, and Tez 0.8.4
HBase: HBase 1.3.0 with Ganglia 3.7.2, Hadoop 2.7.3, Hive 2.1.1, Hue 3.12.0, Phoenix 4.9.0, and ZooKeeper 3.4.10
Presto: Presto 0.170 with Hadoop 2.7.3 HDFS and Hive 2.1.1 Metastore
Spark: Spark 2.1.1 on Hadoop 2.7.3 YARN with Ganglia 3.7.2 and Zeppelin 0.7.1

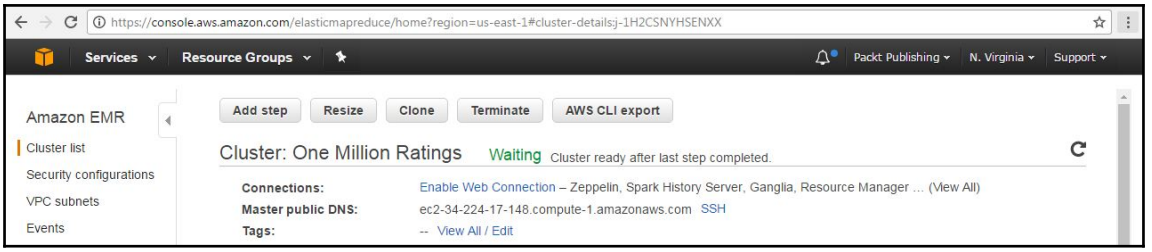

Hardware configuration
Instance type
m3.xlarge
Number of instances
3
(1 master and 2 core nodes)

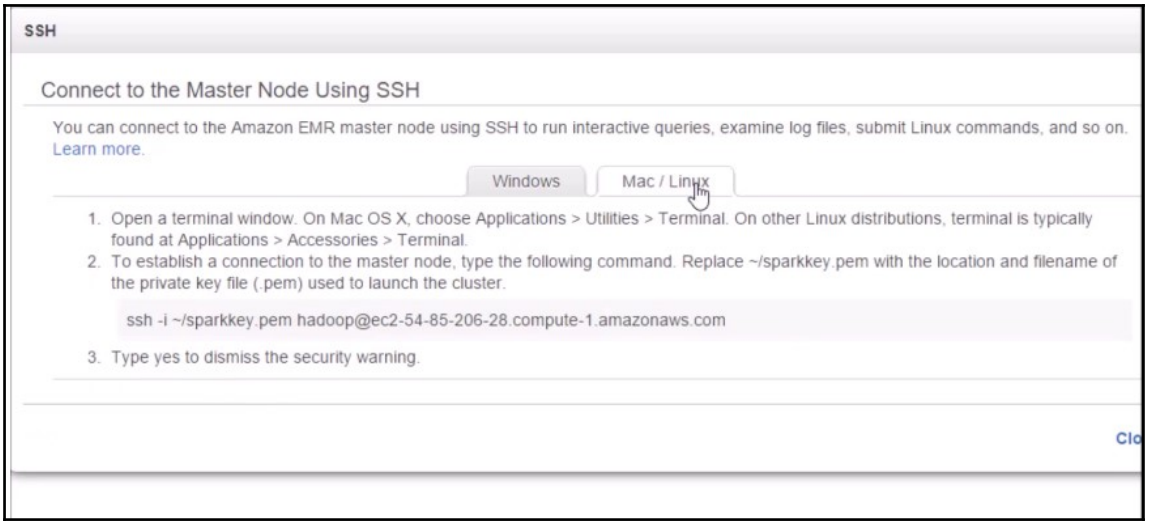

Security and access
EC2 key pair
sparkkey
Learn how to create an EC2 key pair.
Permissions
Default
Custom
Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.
EMR role
EMR_DefaultRole
EC2 instance profile
EMR_EC2_DefaultRole
Cancel
Create cluster

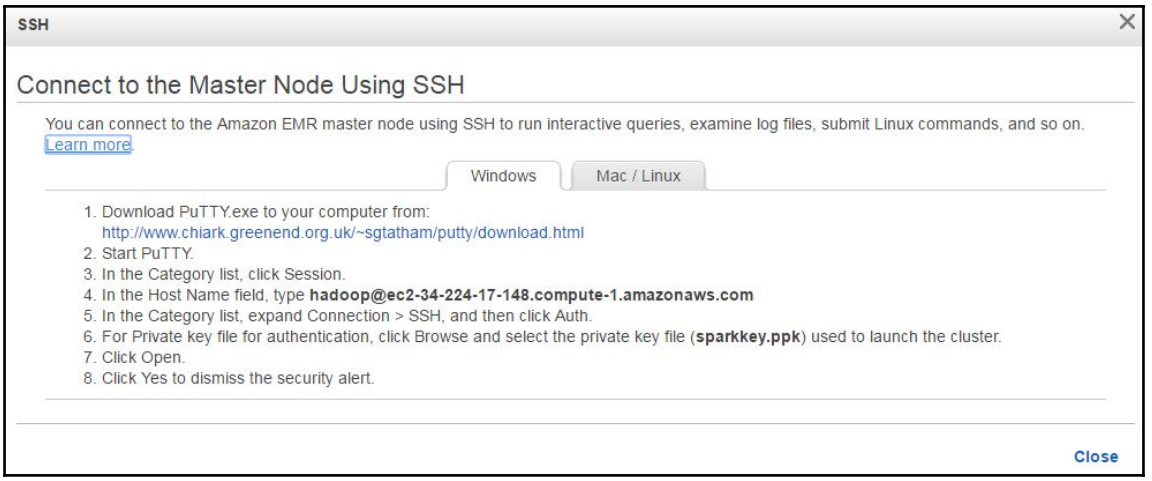

Secure | https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-1H2CSNYHSENXX
Services
Resource Groups
Packt Publishing
N. Virginia
Support
Amazon EMR
Cluster list
Security configurations
VPC subnets
Events
Help
Add step
Resize
Clone
Terminate
AWS CLI export
Cluster: One Million Ratings
Starting
Connections: --
Master public DNS: --
Tags: -- View All / Edit
Summary
ID: j-1H2CSNYHSENXX
Creation date: 2017-06-28 16:25 (UTC+5:30)
Elapsed time:
Auto-terminate: No
Termination protection: Off Change
Configuration Details
Release label: emr-5.6.0
Hadoop distribution: Amazon 2.7.3
Applications: Ganglia 3.7.2, Spark 2.1.1, Zeppelin 0.7.1
Log URI: --
EMRFS consistent view: Disabled
Network and Hardware
Availability zone: --
Subnet ID: subnet-b481fb8e
Master: Provisioning 1 m3.xlarge
Core: Provisioning 2 m3.xlarge
Task: --
Security and Access
Key name: sparkkey
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master:
Security groups for Core & Task:
Monitoring
Hardware
Steps
Configurations

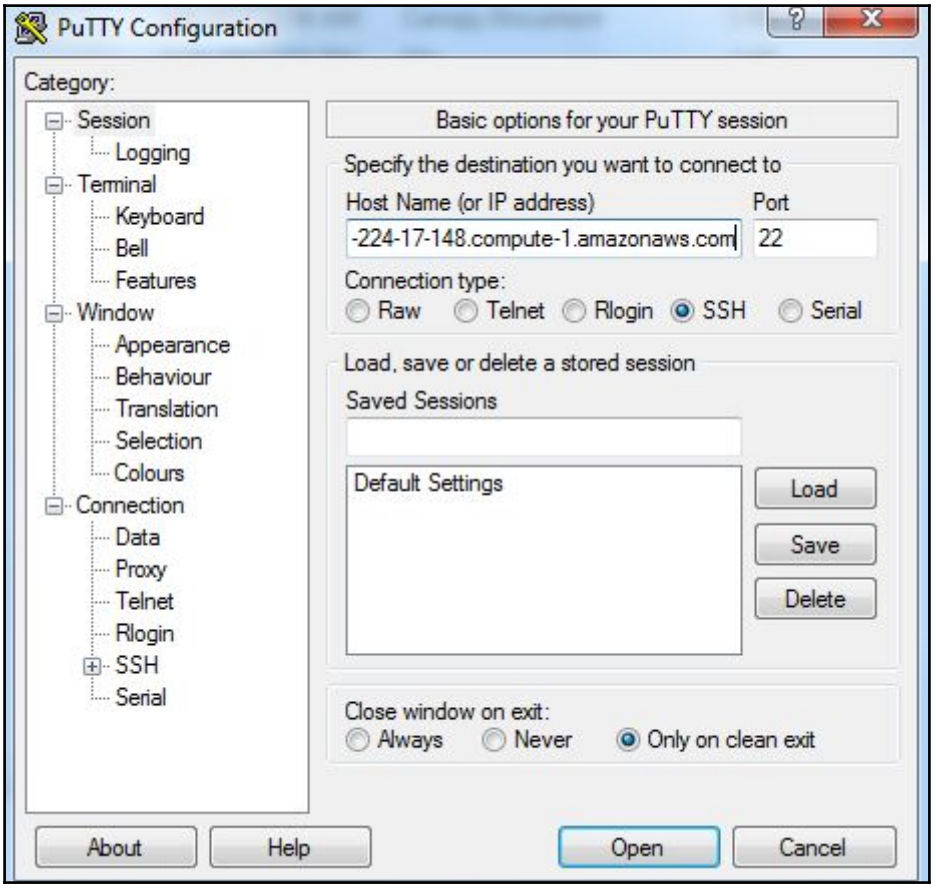

Network and Hardware
Availability zone: --
Subnet ID: subnet-b481fb8e
Master: Provisioning 1 m3.xlarge
Core: Provisioning 2 m3.xlarge
Task: --

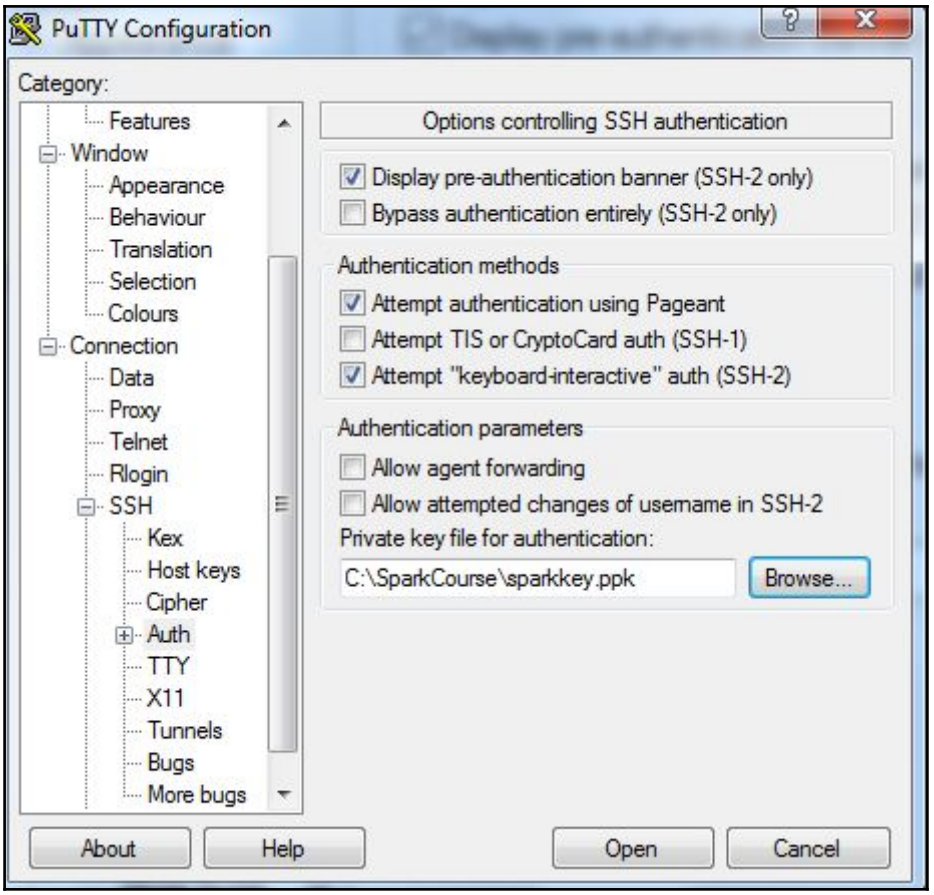

Network and Hardware
Availability zone: us-east-1e
Subnet ID: subnet-b481fb8e
Master: Running 1 m3.xlarge
Core: Running 2 m3.xlarge
Task: --

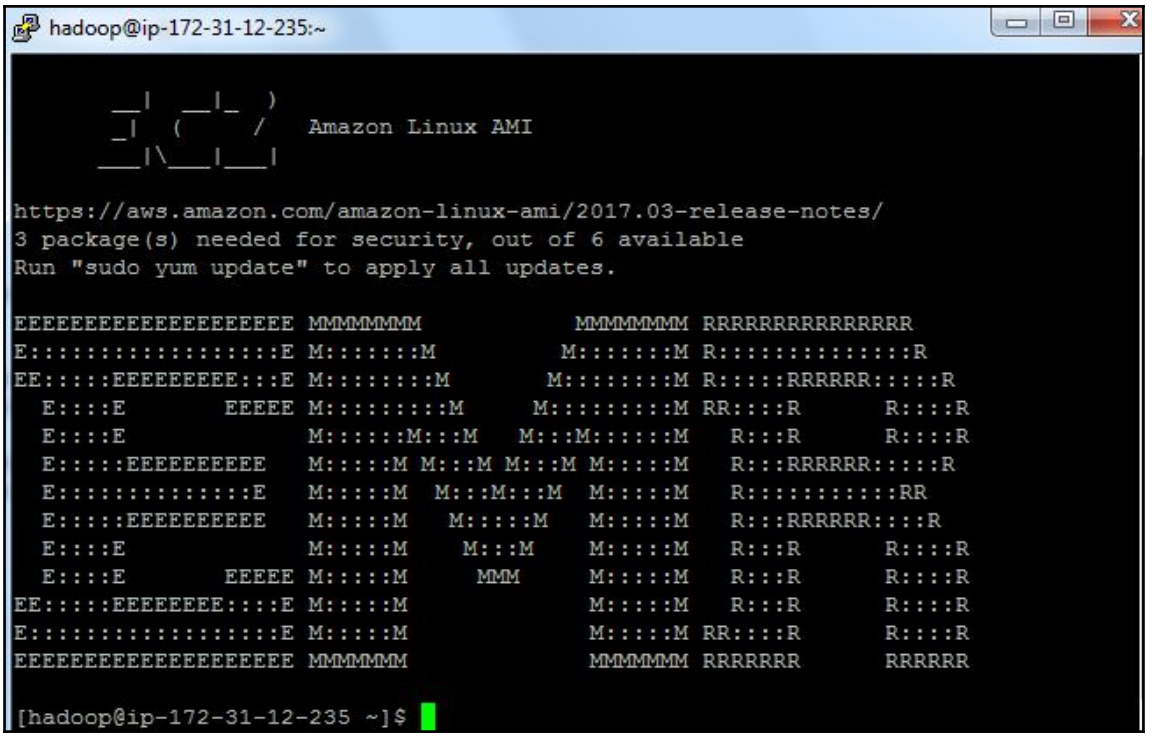

https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1#cluster-details:j-1H2CSNYHSENXX
Services
Resource Groups
Packt Publishing
N. Virginia
Support
Amazon EMR
Cluster list
Security configurations
VPC subnets
Events
Add step
Resize
Clone
Terminate
AWS CLI export
Cluster: One Million Ratings
Waiting
Cluster ready after last step completed.
Connections:
Enable Web Connection – Zeppelin, Spark History Server, Ganglia, Resource Manager ... (View All)
Master public DNS:
ec2-34-224-17-148.compute-1.amazonaws.com
SSH
Tags:
-- View All / Edit


SSH
Connect to the Master Node Using SSH
You can connect to the Amazon EMR master node using SSH to run interactive queries, examine log files, submit Linux commands, and so on.
Learn more.
Windows
Mac / Linux
1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found at Applications > Accessories > Terminal.
2. To establish a connection to the master node, type the following command. Replace ~/sparkkey.pem with the location and filename of the private key file (.pem) used to launch the cluster.
ssh -i ~/sparkkey.pem hadoop@ec2-54-85-206-28.compute-1.amazonaws.com
3. Type yes to dismiss the security warning.


SSH
Connect to the Master Node Using SSH
You can connect to the Amazon EMR master node using SSH to run interactive queries, examine log files, submit Linux commands, and so on.
Learn more
Windows
Mac / Linux
1. Download PuTTY.exe to your computer from:
http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html
2. Start PuTTY.
3. In the Category list, click Session.
4. In the Host Name field, type hadoop@ec2-34-224-17-148.compute-1.amazonaws.com
5. In the Category list, expand Connection > SSH, and then click Auth.
6. For Private key file for authentication, click Browse and select the private key file (sparkkey.ppk) used to launch the cluster.
7. Click Open.
8. Click Yes to dismiss the security alert.
Close


PuTTY Configuration
Category:
Session
Logging
Terminal
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Serial
Basic options for your PuTTY session
Specify the destination you want to connect to
Host Name (or IP address)
Port
-224-17-148.compute-1.amazonaws.com
22
Connection type:
Raw
Telnet
Rlogin
SSH
Serial
Load, save or delete a stored session
Saved Sessions
Default Settings
Load
Save
Delete
Close window on exit:
Always
Never
Only on clean exit
About
Help
Open
Cancel


PuTTY Configuration
Category:
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
Telnet
Rlogin
SSH
Kex
Host keys
Cipher
Auth
TTY
X11
Tunnels
Bugs
More bugs
Options controlling SSH authentication
Display pre-authentication banner (SSH-2 only)
Bypass authentication entirely (SSH-2 only)
Authentication methods
Attempt authentication using Pageant
Attempt TIS or CryptoCard auth (SSH-1)
Attempt "keyboard-interactive" auth (SSH-2)
Authentication parameters
Allow agent forwarding
Allow attempted changes of username in SSH-2
Private key file for authentication:
C:\SparkCourse\sparkkey.ppk
Browse...
About
Help
Open
Cancel




```
[hadoop@ip-172-31-12-235 ~]$ aws s3 cp s3://sundog-spark/MovieSimilarities1M.py
./
```

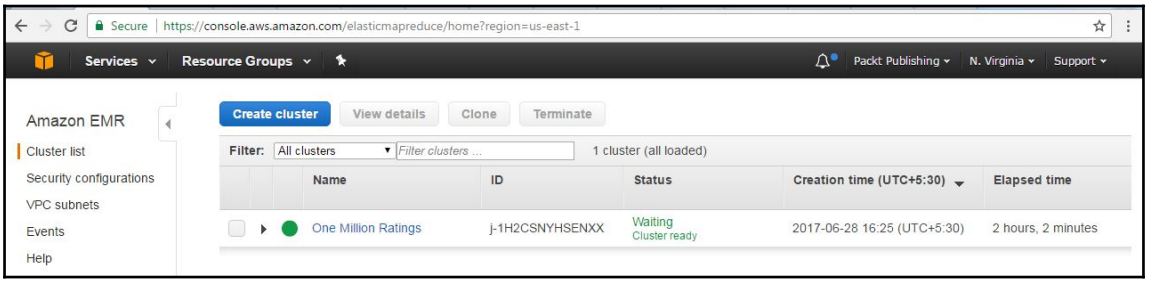
```
download: s3://sundog-spark/MovieSimilarities1M.py to ./MovieSimilarities1M.py
[hadoop@ip-172-31-12-235 ~]$
```

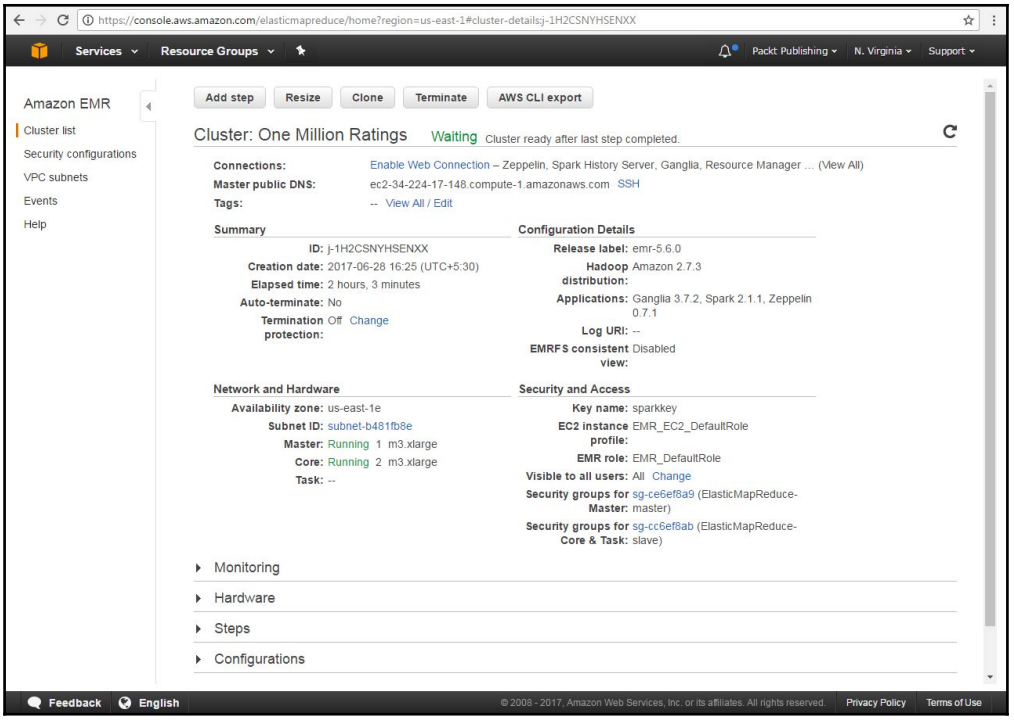
```
[hadoop@ip-172-31-12-235 ~]$ aws s3 cp s3://sundog-spark/ml-1m/movies.dat ./
download: s3://sundog-spark/ml-1m/movies.dat to ./movies.dat
```

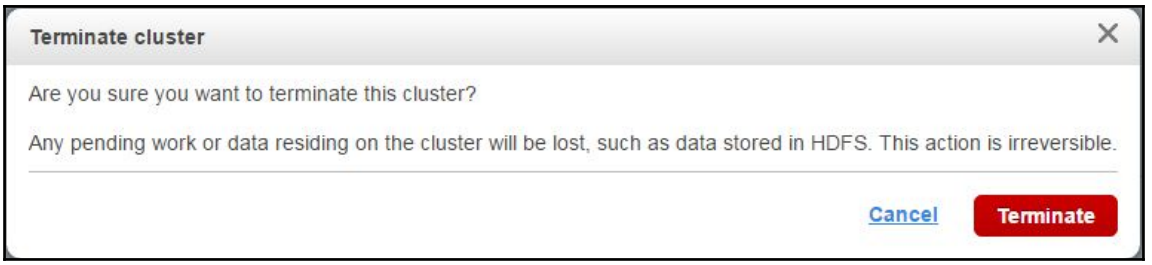
```
[hadoop@ip-172-31-12-235 ~]$
```

```
[hadoop@ip-172-31-12-235 ~]$ spark-submit --executor-memory 1g MovieSimilarities
1M.py 260
```

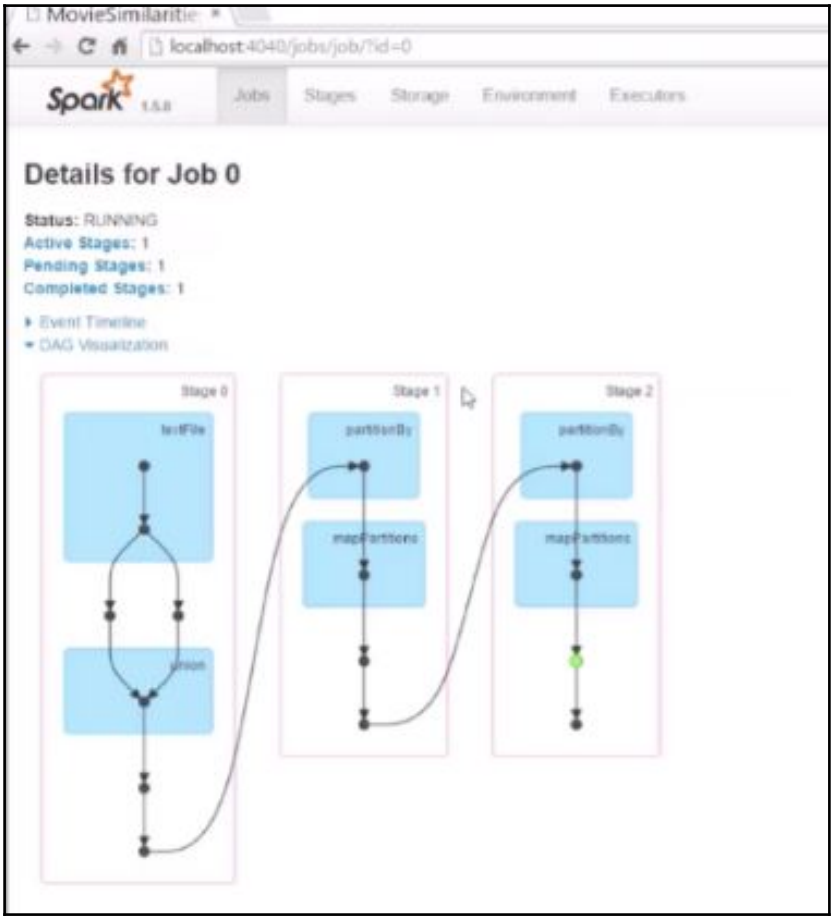
```
Top 10 similar movies for Star Wars: Episode IV - A New Hope (1977)
Star Wars: Episode V - The Empire Strikes Back (1980)    score: 0.989791710657    strength: 2355
Sanjuro (1962)  score: 0.987715715754    strength: 60
Raiders of the Lost Ark (1981)  score: 0.985554827857    strength: 1972
Star Wars: Episode VI - Return of the Jedi (1983)        score: 0.984124835993    strength: 2113
Run Silent, Run Deep (1958)      score: 0.979146338933    strength: 145
Laura (1944)     score: 0.978729003724    strength: 187
Close Shave, A (1995)    score: 0.978216762084    strength: 436
Wrong Trousers, The (1993)       score: 0.978051224484    strength: 596
Captain Horatio Hornblower (1951)        score: 0.977892172004    strength: 81
Indiana Jones and the Last Crusade (1989)        score: 0.977444002865    strength: 1397
```

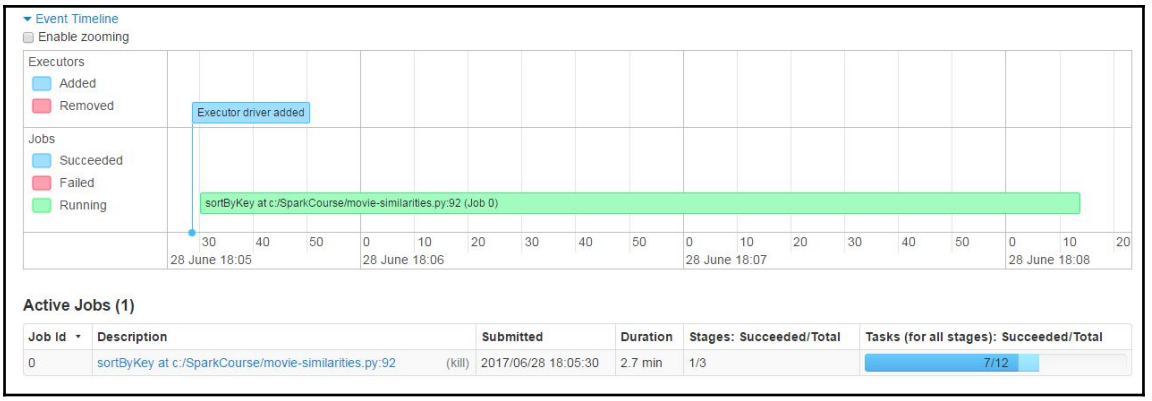






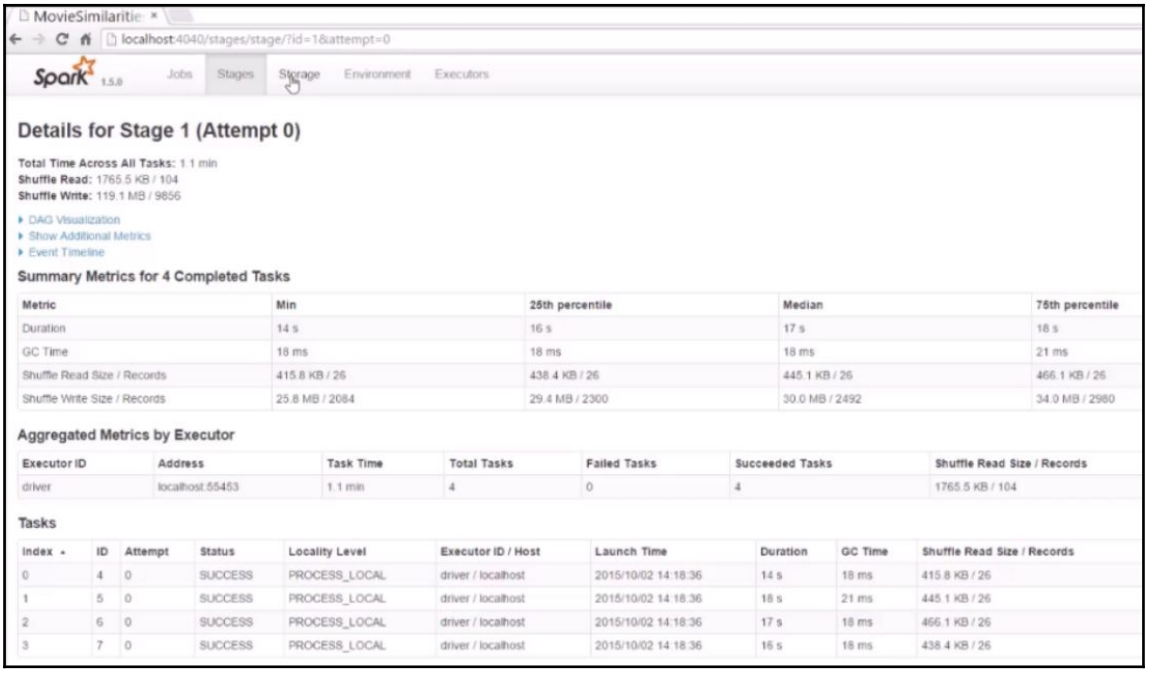

```
(User) c:\SparkCourse>spark-submit movie-similarities.py 50
```





Active Jobs (1)

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	sortByKey at c:/SparkCourse/movie-similarities.py:92 (kill)	2017/06/28 18:05:30	2.7 min	1/3	7/12



Details for Stage 1 (Attempt 0)

Total Time Across All Tasks: 1.1 min
Shuffle Read: 1765.5 KB / 104
Shuffle Write: 119.1 MB / 9856

- DAG Visualization
- Show Additional Metrics
- Event Timeline

Summary Metrics for 4 Completed Tasks

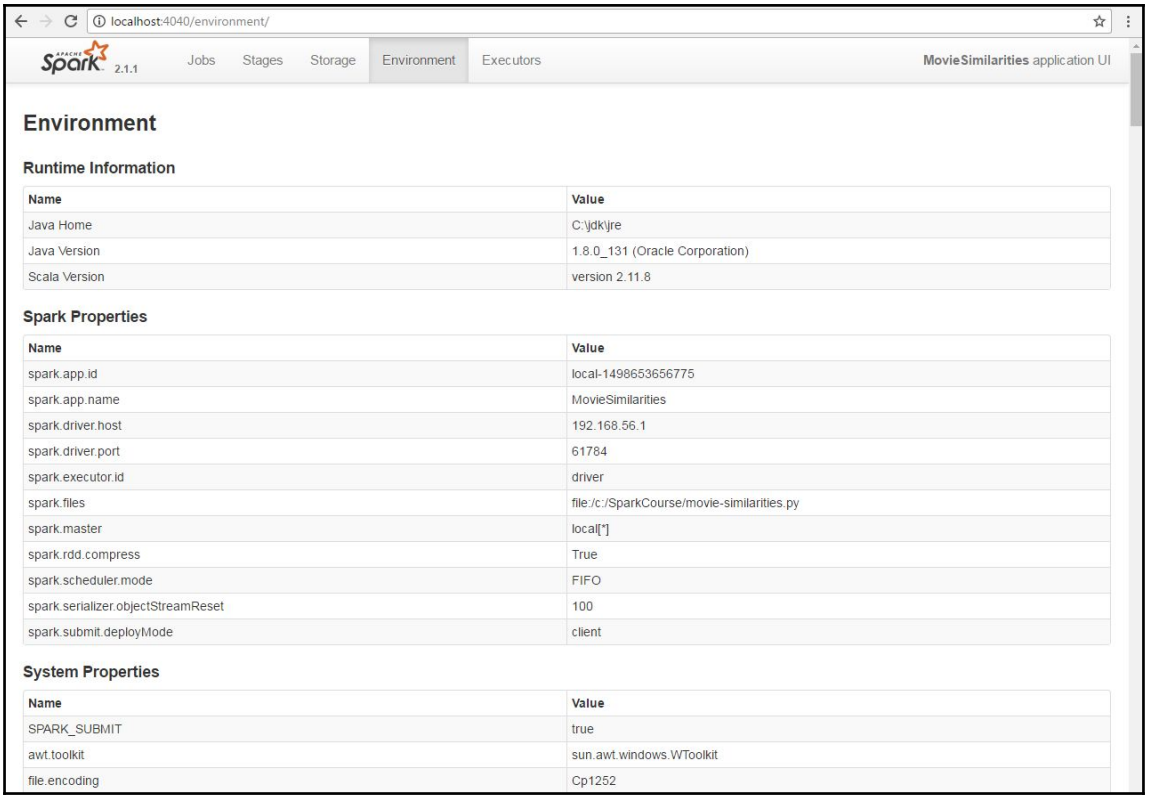
Metric	Min	25th percentile	Median	75th percentile
Duration	14 s	16 s	17 s	18 s
GC Time	18 ms	18 ms	18 ms	21 ms
Shuffle Read Size / Records	415.8 KB / 26	438.4 KB / 26	445.1 KB / 26	466.1 KB / 26
Shuffle Write Size / Records	25.8 MB / 2084	29.4 MB / 2300	30.0 MB / 2492	34.0 MB / 2980

Aggregated Metrics by Executor

Executor ID	Address	Task Time	Total Tasks	Failed Tasks	Succeeded Tasks	Shuffle Read Size / Records
driver	localhost:55453	1.1 min	4	0	4	1765.5 KB / 104

Tasks

Index	ID	Attempt	Status	Locality Level	Executor ID / Host	Launch Time	Duration	GC Time	Shuffle Read Size / Records
0	4	0	SUCCESS	PROCESS_LOCAL	driver / localhost	2015/10/02 14:18:36	14 s	18 ms	415.8 KB / 26
1	5	0	SUCCESS	PROCESS_LOCAL	driver / localhost	2015/10/02 14:18:36	18 s	21 ms	445.1 KB / 26
2	6	0	SUCCESS	PROCESS_LOCAL	driver / localhost	2015/10/02 14:18:36	17 s	18 ms	466.1 KB / 26
3	7	0	SUCCESS	PROCESS_LOCAL	driver / localhost	2015/10/02 14:18:36	16 s	18 ms	438.4 KB / 26



Environment

Runtime Information

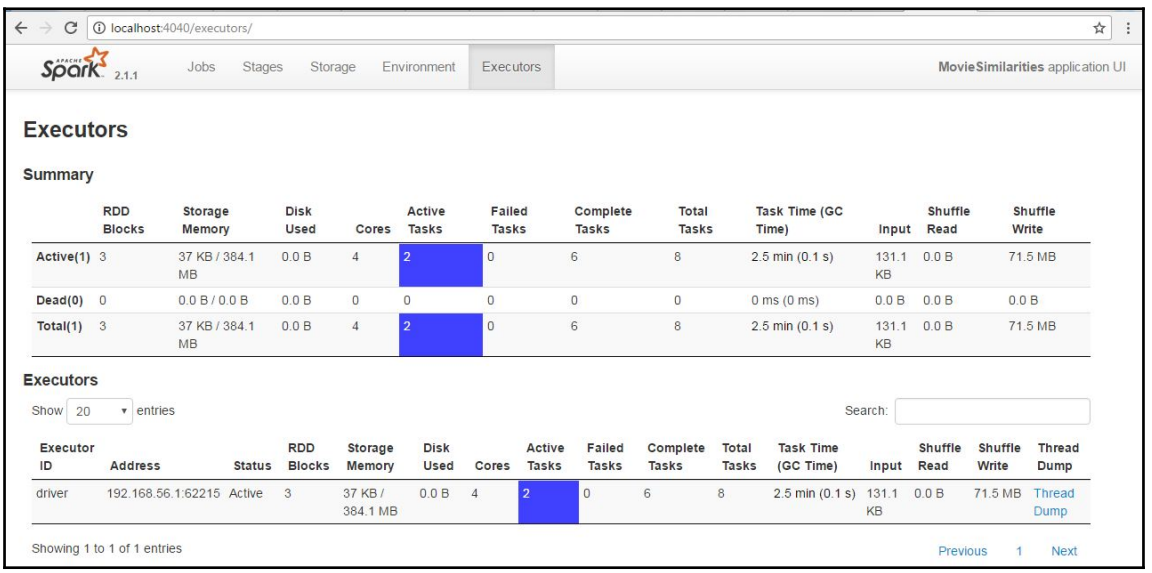
Name	Value
Java Home	C:\jdk\jre
Java Version	1.8.0_131 (Oracle Corporation)
Scala Version	version 2.11.8

Spark Properties

Name	Value
spark.app.id	local-1498653656775
spark.app.name	MovieSimilarities
spark.driver.host	192.168.56.1
spark.driver.port	61784
spark.executor.id	driver
spark.files	file:/c:/SparkCourse/movie-similarities.py
spark.master	local[*]
spark.rdd.compress	True
spark.scheduler.mode	FIFO
spark.serializer.objectStreamReset	100
spark.submit.deployMode	client

System Properties

Name	Value
SPARK_SUBMIT	true
awt.toolkit	sun.awt.windows.WToolkit
file.encoding	Cp1252



Executors

Summary

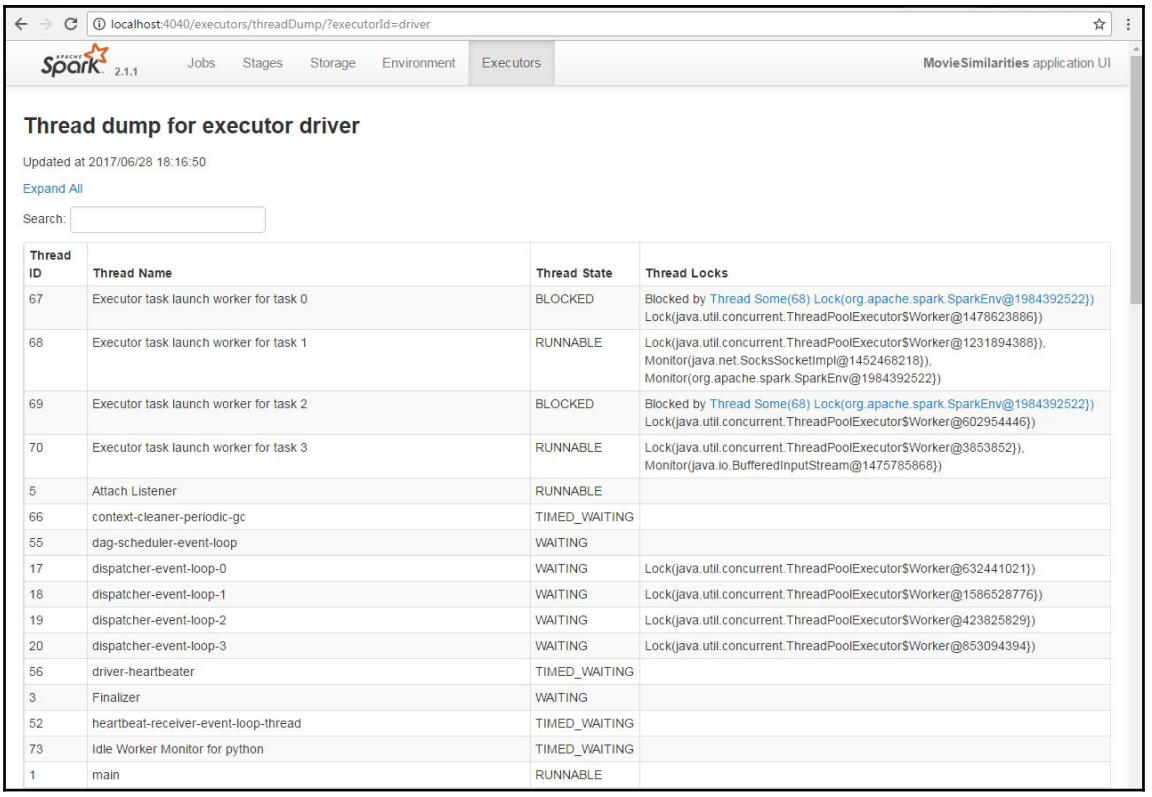
	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write
Active(1)	3	37 KB / 384.1 MB	0.0 B	4	2	0	6	8	2.5 min (0.1 s)	131.1 KB	0.0 B	71.5 MB
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0 ms (0 ms)	0.0 B	0.0 B	0.0 B
Total(1)	3	37 KB / 384.1 MB	0.0 B	4	2	0	6	8	2.5 min (0.1 s)	131.1 KB	0.0 B	71.5 MB

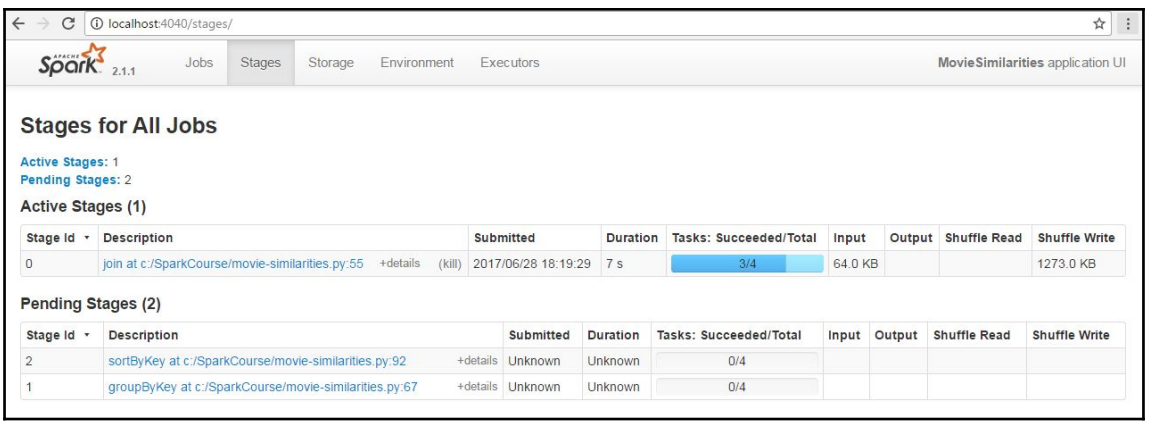
Executors

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	192.168.56.1:62215	Active	3	37 KB / 384.1 MB	0.0 B	4	2	0	6	8	2.5 min (0.1 s)	131.1 KB	0.0 B	71.5 MB	Thread Dump

Showing 1 to 1 of 1 entries Previous 1 Next

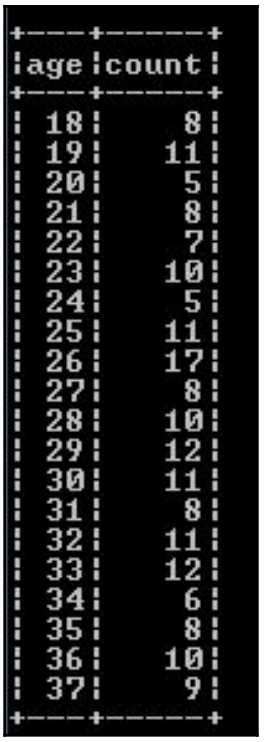

localhost:4040/executors/threadDump/?executorId=driver
Spark 2.1.1
Jobs
Stages
Storage
Environment
Executors
MovieSimilarities application UI
Thread dump for executor driver
Updated at 2017/06/28 18:16:50
Expand All
Search:
Thread ID
Thread Name
Thread State
Thread Locks
67
Executor task launch worker for task 0
BLOCKED
Blocked by Thread Some(68) Lock(org.apache.spark.SparkEnv@1984392522})
Lock(java.util.concurrent.ThreadPoolExecutor$Worker@1478623886})
68
Executor task launch worker for task 1
RUNNABLE
Lock(java.util.concurrent.ThreadPoolExecutor$Worker@1231894388}),
Monitor(java.net.SocksSocketImpl@1452468218}),
Monitor(org.apache.spark.SparkEnv@1984392522})
69
Executor task launch worker for task 2
BLOCKED
Blocked by Thread Some(68) Lock(org.apache.spark.SparkEnv@1984392522})
Lock(java.util.concurrent.ThreadPoolExecutor$Worker@602954446})
70
Executor task launch worker for task 3
RUNNABLE
Lock(java.util.concurrent.ThreadPoolExecutor$Worker@3853852}),
Monitor(java.io.BufferedInputStream@1475785868})
5
Attach Listener
RUNNABLE
66
context-cleaner-periodic-gc
TIMED_WAITING
55
dag-scheduler-event-loop
WAITING
17
dispatcher-event-loop-0
WAITING
Lock(java.util.concurrent.ThreadPoolExecutor$Worker@632441021})
18
dispatcher-event-loop-1
WAITING
Lock(java.util.concurrent.ThreadPoolExecutor$Worker@1586528776})
19
dispatcher-event-loop-2
WAITING
Lock(java.util.concurrent.ThreadPoolExecutor$Worker@423825829})
20
dispatcher-event-loop-3
WAITING
Lock(java.util.concurrent.ThreadPoolExecutor$Worker@853094394})
56
driver-heartbeater
TIMED_WAITING
3
Finalizer
WAITING
52
heartbeat-receiver-event-loop-thread
TIMED_WAITING
73
Idle Worker Monitor for python
TIMED_WAITING
1
main
RUNNABLE


localhost:4040/stages/
Spark 2.1.1
Jobs
Stages
Storage
Environment
Executors
MovieSimilarities application UI
Stages for All Jobs
Active Stages: 1
Pending Stages: 2
Active Stages (1)
Stage Id
Description
Submitted
Duration
Tasks: Succeeded/Total
Input
Output
Shuffle Read
Shuffle Write
0
join at c:/SparkCourse/movie-similarities.py:55 +details (kill)
2017/06/28 18:19:29
7 s
3/4
64.0 KB
1273.0 KB
Pending Stages (2)
Stage Id
Description
Submitted
Duration
Tasks: Succeeded/Total
Input
Output
Shuffle Read
Shuffle Write
2
sortByKey at c:/SparkCourse/movie-similarities.py:92 +details
Unknown
Unknown
0/4
1
groupByKey at c:/SparkCourse/movie-similarities.py:67 +details
Unknown
Unknown
0/4

Chapter 5

```
from pyspark.sql import SparkSession
from pyspark.sql import Row

import collections

# Create a SparkSession (Note, the config section is only for Windows!)
spark = SparkSession.builder.config("spark.sql.warehouse.dir", "file:///C:/temp").appName("SparkSQL").getOrCreate()

def mapper(line):
    fields = line.split(',')
    return Row(ID=int(fields[0]), name=str(fields[1].encode("utf-8")), age=int(fields[2]), numFriends=int(fields[3]))

lines = spark.sparkContext.textFile("fakefriends.csv")
people = lines.map(mapper)

# Infer the schema, and register the DataFrame as a table.
schemaPeople = spark.createDataFrame(people).cache()
schemaPeople.createOrReplaceTempView("people")

# SQL can be run over DataFrames that have been registered as a table.
teenagers = spark.sql("SELECT * FROM people WHERE age >= 13 AND age <= 19")

# The results of SQL queries are RDDs and support all the normal RDD operations.
for teen in teenagers.collect():
  print(teen)

# We can also use functions instead of SQL queries:
schemaPeople.groupBy("age").count().orderBy("age").show()

spark.stop()

```

```
Row(ID=21, age=19, name=u'Miles', numFriends=268)
Row(ID=52, age=19, name=u'Beverly', numFriends=269)
Row(ID=54, age=19, name=u'Brunt', numFriends=5)
Row(ID=106, age=18, name=u'Beverly', numFriends=499)
Row(ID=115, age=18, name=u'Dukat', numFriends=397)
Row(ID=133, age=19, name=u'Quark', numFriends=265)
Row(ID=136, age=19, name=u'Will', numFriends=335)
Row(ID=225, age=19, name=u'Elim', numFriends=106)
Row(ID=304, age=19, name=u'Will', numFriends=404)
Row(ID=341, age=18, name=u'Data', numFriends=326)
Row(ID=366, age=19, name=u'Keiko', numFriends=119)
Row(ID=373, age=19, name=u'Quark', numFriends=272)
Row(ID=377, age=18, name=u'Beverly', numFriends=418)
Row(ID=404, age=18, name=u'Kasidy', numFriends=24)
Row(ID=409, age=19, name=u'Nog', numFriends=267)
Row(ID=439, age=18, name=u'Data', numFriends=417)
Row(ID=444, age=18, name=u'Keiko', numFriends=472)
Row(ID=492, age=19, name=u'Dukat', numFriends=36)
Row(ID=494, age=18, name=u'Kasidy', numFriends=194)
```



age	count
18	8
19	11
20	5
21	8
22	7
23	10
24	5
25	11
26	17
27	8
28	10
29	12
30	11
31	8
32	11
33	12
34	6
35	8
36	10
37	9

```python
from pyspark.sql import SparkSession
from pyspark.sql import Row
from pyspark.sql import functions

def loadMovieNames():
    movieNames = {}
    with open("ml-100k/u.ITEM") as f:
        for line in f:
            fields = line.split('|')
            movieNames[int(fields[0])] = fields[1]
    return movieNames

# Create a SparkSession (the config bit is only for Windows!)
spark = SparkSession.builder.config("spark.sql.warehouse.dir", "file:///C:/temp").appName("PopularMovies").getOrCreate()

# Load up our movie ID -> name dictionary
nameDict = loadMovieNames()

# Get the raw data
lines = spark.sparkContext.textFile("file:///SparkCourse/ml-100k/u.data")
# Convert it to a RDD of Row objects
movies = lines.map(lambda x: Row(movieID =int(x.split()[1])))
```

```
+-------+-----+
|movieID|count|
+-------+-----+
|     50|  584|
|    258|  509|
|    100|  508|
|    181|  507|
|    294|  485|
|    286|  481|
|    288|  478|
|      1|  452|
|    300|  431|
|    121|  429|
|    174|  420|
|    127|  413|
|     56|  394|
|      7|  392|
|     98|  390|
|    237|  384|
|    117|  378|
|    172|  368|
|    222|  365|
|    313|  350|
+-------+-----+
only showing top 20 rows


Star Wars (1977): 584
Contact (1997): 509
Fargo (1996): 508
Return of the Jedi (1983): 507
Liar Liar (1997): 485
English Patient, The (1996): 481
Scream (1996): 478
Toy Story (1995): 452
Air Force One (1997): 431
Independence Day (ID4) (1996): 429
```

Chapter 6

0	50	5	881250949
0	172	5	881250949
0	133	1	881250949

```
(User) c:\SparkCourse>spark-submit movie-recommendations-als.py 0
```

```
Ratings for user ID 0:
Star Wars (1977): 5.0
Empire Strikes Back, The (1980): 5.0
Gone with the Wind (1939): 1.0

Top 10 recommendations:
Love in the Afternoon (1957) score 6.42090083536
Roommates (1995) score 6.39431215726
Burnt Offerings (1976) score 6.38702183096
Lost in Space (1998) score 6.38680899253
Endless Summer 2, The (1994) score 6.30275992511
Primary Colors (1998) score 6.03035775839
Drunks (1995) score 5.92894606542
Cronos (1992) score 5.71380632161
unknown score 5.676838214
Double Team (1997) score 5.65588319517

(User) c:\SparkCourse>
```

```
Ratings for user ID 0:
[Stage 280:>                                                        (0 + 2) / 2]
[Stage 280:============================>                          (1 + 1) / 2]

Star Wars (1977): 5.0
Empire Strikes Back, The (1980): 5.0
Gone with the Wind (1939): 1.0

Top 10 recommendations:
Roommates (1995) score 7.87966702947
I'll Do Anything (1994) score 7.57841013131
Shall We Dance? (1937) score 7.23874848332
Don't Be a Menace to South Central While Drinking Your Juice in the Hood (1996)
score 6.72436905195
Low Down Dirty Shame, A (1994) score 6.13396930989
Army of Darkness (1993) score 5.98367809308
Underneath, The (1995) score 5.9643946162
Lord of Illusions (1995) score 5.95305643224
Hard Eight (1996) score 5.93277528025
In the Line of Duty 2 (1987) score 5.88337368104

(Canopy 64bit) C:\SparkCourse>
```

```
Top 10 recommendations:
War, The (1994) score 6.65716239806
Low Down Dirty Shame, A (1994) score 6.44548993994
Lost in Space (1998) score 6.27515939994
Love in the Afternoon (1957) score 5.60112839882
Schizopolis (1996) score 5.56638126463
Meet John Doe (1941) score 5.11439598351
Star Wars (1977) score 5.04373210278
Addiction, The (1995) score 4.96306972202
Empire Strikes Back, The (1980) score 4.92202227603
Fast, Cheap & Out of Control (1997) score 4.91837364129
```

```
(User) c:\SparkCourse>spark-submit spark-linear-regression.py
```

```
(0.8061518555150741, 1.19)
(0.956678266497083, 1.25)
(0.8634952501748869, 1.27)
(0.9423424178321297, 1.34)
(1.049861282819279, 1.36)
(0.9136707205022232, 1.44)
(1.0785329801491854, 1.45)
(1.0928688288141386, 1.52)
(1.1358763748089984, 1.53)
(1.1502122234739516, 1.54)
(1.1502122234739516, 1.55)
(1.1358763748089984, 1.59)
(1.0928688288141386, 1.74)
(1.3939216507781564, 1.78)
(1.2648990127935775, 1.82)
(1.3007386344559606, 1.85)
(1.2290593911311944, 1.93)
(1.5086084400977824, 1.95)
(1.479936742767876, 1.98)
(1.329410331785867, 2.0)
(1.594623532087502, 2.08)

(User) c:\SparkCourse>
```