livelessons
video instruction from technology experts

Lesson 1: Introduction to the Spark Environment

1.6 Getting Set Up: Scientific Python

Installation: Scientific Python

- http://continuum.io/downloads

Installation: Scientific Python

- http://continuum.io/downloads
- Find download for your OS (make sure it is Python 2.7)
- Follow install instructions/wizard

Installation: Scientific Python

- http://continuum.io/downloads
- Find download for your OS (make sure it is Python 2.7)
- Follow install instructions/wizard

To make sure it installed correctly:

```
ipython notebook
```

And finally: `pip install py4j`

Installation: Test It All Out

```
jonathan$ ipython
```

```
IPython 3.1.0 -- An enhanced Interactive Python.
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details.
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload

In [1]: import pyspark as ps

In [2]: sc = ps.Sp
ps.SparkConf       ps.SparkContext    ps.SparkFiles      ps.SparkJobInfo    ps.SparkStageInfo

In [2]: sc = ps.SparkContext('local')
15/08/03 16:17:15 INFO spark.SparkContext: Running Spark version 1.4.1
15/08/03 16:17:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
15/08/03 16:17:16 INFO spark.SecurityManager: Changing view acls to: jonathandinu
15/08/03 16:17:16 INFO spark.SecurityManager: Changing modify acls to: jonathandinu
15/08/03 16:17:16 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(jonathandinu); users with modify permissions: Set(jonathandinu)
15/08/03 16:17:17 INFO slf4j.Slf4jLogger: Slf4jLogger started
15/08/03 16:17:17 INFO Remoting: Starting remoting
15/08/03 16:17:17 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@10.3.35.20:52356]
15/08/03 16:17:17 INFO util.Utils: Successfully started service 'sparkDriver' on port 52356.
15/08/03 16:17:17 INFO spark.SparkEnv: Registering MapOutputTracker
15/08/03 16:17:17 INFO spark.SparkEnv: Registering BlockManagerMaster
15/08/03 16:17:17 INFO storage.DiskBlockManager: Created local directory at /private/var/folders/wf/h5v75xw52g1b7wz7lwrllwrm0000gn/T/spark-8319075d-92e4-4cf1-a54f-bf6ab9415af8/blockmgr-a2ee72cc-67c6-468a-9e44-81bb5993da0b
15/08/03 16:17:17 INFO storage.MemoryStore: MemoryStore started with capacity 265.1 MB
15/08/03 16:17:17 INFO spark.HttpFileServer: HTTP File server directory is /private/var/folders/wf/h5v75xw52g1b7wz7lwrllwrm0000gn/T/spark-8319075d-92e4-4cf1-a54f-bf6ab9415af8/httpd-0f51c80a-c355-45e1-90bb-841559fadfa5
15/08/03 16:17:17 INFO spark.HttpServer: Starting HTTP Server
15/08/03 16:17:17 INFO server.Server: jetty-8.y.z-SNAPSHOT
15/08/03 16:17:17 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0:52357
15/08/03 16:17:17 INFO util.Utils: Successfully started service 'HTTP file server' on port 52357.
15/08/03 16:17:17 INFO spark.SparkEnv: Registering OutputCommitCoordinator
```

Installation: Requirements

- Spark binary
- Java JDK 6/7
- Scientific Python (and Jupyter notebook)
- `py4j`
- (Optional) IRKernel (for Jupyter)