Setting Up Spark, PySpark and Notebook

• • •

Setting up your workstation

Session Outline

We'll

- Set up your system
- Run "Hello World"

Setting up

Your System

- Ubuntu 16.04LTS
- 64-bit
- Python3 (Anaconda)

What we'll set-up

- Spark2.0
- findspark

Hello World

We'll

- Start a local Spark server
- Use pyspark to run a program
- Understand the Spark MasterWebUI

Setting Up

Install Spark

We'll use Spark 2.0.0, prebuilt for Hadoop 2.7 or later

Download link

 http://d3kbcqa49mib13.cloudfron t.net/spark-2.0.0-bin-hadoop2.7.tg
 z

Spark Download Page

http://spark.apache.org/download
 s.html

PySpark

How to talk to PySpark from Jupyter Notebooks

- PySpark isn't on sys.path by default
 - This means the Python kernel in Jupyter Notebook doesn't know where to look for PySpark
- You can address this by either
 - symlinking pyspark into your site-packages, or
 - adding pyspark to sys.path at runtime
 - by passing the path diretly
 - by looking at a running instance
- findspark adds pyspark to sys.path at runtime

PySpark

How to talk to PySpark from Jupyter Notebooks

findspark homepage

 https://github.com/minrk/findspa rk

Install

pip install findspark

Hello World

Install Spark

Just extract the files and folders from the compressed file and you are done.

If you've used the link in the last slide to download Spark, then

- go to the folder it has been downloaded in
- > tar xvzf spark-2.0.0-bin-hadoop2.7.tgz
- > mv spark-2.0.0-bin-hadoop2.7 spark2
 - Start a local (master) server
- > cd spark2/sbin
- > ./start-master.sh

The pySpark Shell

```
→ bin ./pyspark

Python 3.5.2 | Anaconda 4.2.0 (x86_64)| (default, Jul 2 2016, 17:52:12)

[GCC 4.2.1 Compatible Apple LLVM 4.2 (clang-425.0.28)] on darwin

Type "help", "copyright", "credits" or "license" for more information.

Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties

Setting default log level to "WARN".

To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).

17/07/10 22:45:27 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

17/07/10 22:45:36 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.

17/07/10 22:45:52 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is not enabled so recording the schema version 1.2.0

17/07/10 22:45:55 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException

17/07/10 22:45:55 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException

Welcome to
```

Using Python version 3.5.2 (default, Jul 2 2016 17:52:12) SparkSession available as 'spark'.

- → sbin ./start-master.sh starting org.apache.spark.deploy.master.Master, logging to /Users/soumendra/spark2/logs/spark-soumendra-org.apache.spark.deploy.master.Master-1-Soumendras-mucbook-Air.local.out
- → sbin cd ../logs
- → logs ls
- spark-soumendra-org.apache.spark.deploy.master.Master-1-Soumendras-MacBook-Air.local.out
- → logs tail spark-soumendra-org.apache.spark.deploy.master.Master-1-Soumendras-MacBook-Air.local.out
- 16/10/18 15:35:11 INFO SecurityManager: Changing modify acls groups to:
- 16/10/18 15:35:11 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(soumendra); groups with view permissions: Set(); users with modify permissions: Set(soumendra); groups with modify permissions: Set()
- 16/10/18 15:35:12 INFO Utils: Successfully started service 'specification on port roll.
- 16/10/18 15:35:12 INFO Master: Starting Spark master at spark://Soumendras-MacBook-Air.local:7077
- 16/10/18 15:35:12 INFO Master: Running Spark version 2.0.0
- 16/10/18 15:35:13 INFO Utils: Successfully started service 'MasterUI' on port 8080.
- 16/10/18 15:35:13 INFO MasterWebUI: Bound MasterWebUI to 0.0.0.0, and started at http://10.6.21.185:8080
- 16/10/18 15:35:13 INFO Utils: Successfully starced service on part 6066.
- 16/10/18 15:35:13 INFO StandaloneRestServer: Started REST server or submitting applications on port 6066
- 16/10/18 15:35:14 INFO Master: I have been elected leader: New state: ALIVE

localhost:8080



Spark Master at spark://Soumendras-MacBook-Air.local:7077

URL: spark://Soumendras-MacBook-Air.local:7077

REST URL: spark://Soumendras-MacBook-Air.local:6066 (cluster mode)

Alive Workers: 0

Cores in use: 0 Total, 0 Used

Memory in use: 0.0 B Total, 0.0 B Used Applications: 0 Running, 0 Completed Drivers: 0 Running, 0 Completed

Status: ALIVE

Workers

unning Application	s						
Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration

Completed Applications

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

Hello World in Spark (counting words)

import findspark

```
# provide path to your spark directory directly
findspark.init("/home/soumendra/downloads/spark2")
import pyspark
sc = pyspark.SparkContext(appName="helloworld")
# let's test our setup by counting the number of lines in a text file
lines = sc.textFile('/home/soumendra/helloworld')
lines_nonempty = lines.filter( lambda x: len(x) > 0 )
lines_nonempty.count()
```

Hello World in Spark (counting words)

Spark_Activities_01_Basics.ipynb: Activity 1