Exercise Z

Using Apache Zeppelin to manage your Spark Experience

Prior Knowledge

Previous Apache Spark lessons

Learning Objectives

Using a notebook

Software Requirements

(see separate document for installation of these)

- Apache Spark 2.0.0
- Python 2.7.x
- Nano text editor or other text editor

Steps

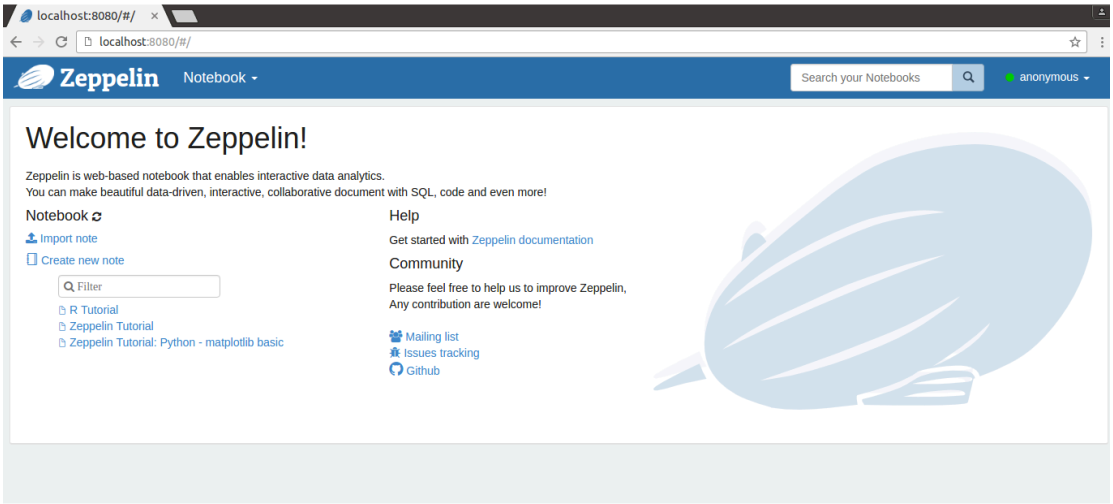
1. Install Zeppelin:
   ```
   curl –L http://freo.me/zeppinst | sh
   ```
   (Wait a bit - it’s a fairly large download)

2. Tell Zeppelin to use CSV package: (All on one line)
   ```
   export SPARK_SUBMIT_OPTIONS="--packages com.databricks:spark-csv_2.10:1.2.0"
   ```

3. Start Zeppelin
   ```
   cd ~/zepp
   bin/zeppelin-daemon.sh start
   ```

   You should see:
   ```
   Log dir doesn't exist, create /home/oxclo/zepp/logs
   Pid dir doesn't exist, create /home/oxclo/zepp/run
   Zeppelin start                                            [  OK  ]
   ```

4. Browse to http://localhost:8080
   You should see:



Welcome to Zeppelin!

Zeppelin is web-based notebook that enables interactive data analytics.
You can make beautiful data-driven, interactive, collaborative document with SQL, code and even more!

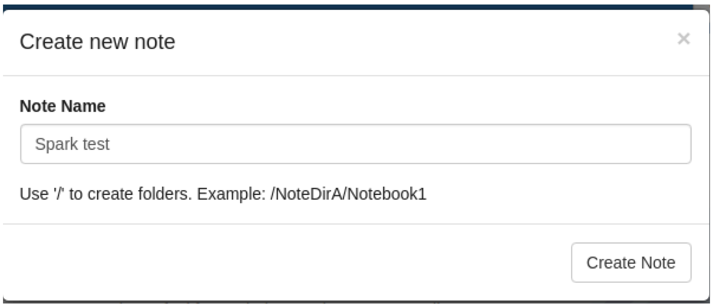
Get started with Zeppelin documentation

Community

Please feel free to help us to improve Zeppelin,
Any contribution are welcome!

5. Click on **Create New Note**

6. Give it a name:



Note Name: Spark test

Use '/' to create folders. Example: /NoteDirA/Notebook1

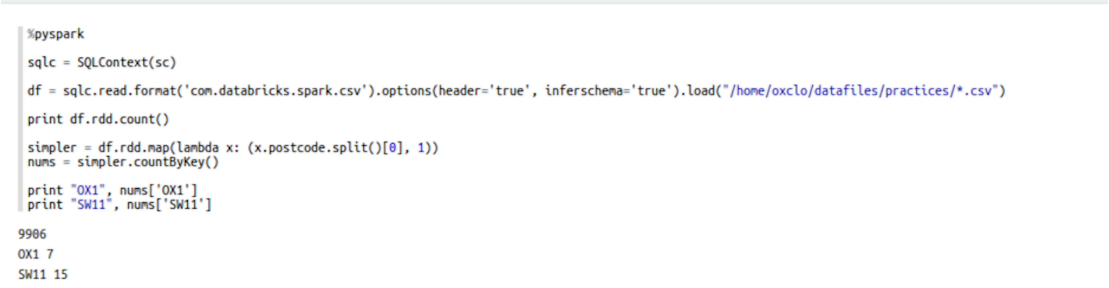
7. From this page http://freo.me/zepp-pc paste the following:

```
%pyspark

sqlc = SQLContext(sc)
df =
sqlc.read.format('com.databricks.spark.csv').options(header=
'true',
inferschema='true').load("/home/oxclo/datafiles/practices/*.
csv")

simpler = df.rdd.map(lambda x: (x.postcode.split()[0], 1))
nums = simpler.countByKey()

print "OX1", nums['OX1']
print "SW11", nums['SW11']
```

8. The first line is a hint to Zeppelin that it needs to use the Pyspark interpreter (Zeppelin supports multiple different backends)

9. Hit the Run button (the little “Play Arrow”)

10. You should see:



```
%pyspark

sqlc = SQLContext(sc)

df = sqlc.read.format('com.databricks.spark.csv').options(header='true', inferschema='true').load("/home/oxclo/datafiles/practices/*.csv")

print df.rdd.count()

simpler = df.rdd.map(lambda x: (x.postcode.split()[0], 1))
nums = simpler.countByKey()

print "OX1", nums['OX1']
print "SW11", nums['SW11']
```

```
9906
OX1 7
SW11 15
```

11. That’s all!