Install_Apache_Spark_PySpark_Mac

September 4, 2016

Install PySpark on Mac
Youtube tutorial available at: https://www.youtube.com/watch?v=I5JtvpyM14U
Download Spark

1) Go to the Apache Sparks.
http://spark.apache.org/downloads.html

```
<li> a) Choose a Spark release (I prefer 1.6.0)</li>
<li> b) Choose a package type: (this installation prefers "Pre-built for Hadoop 2.6 and later")</li>
<li> c) Choose a download type: (Direct Download) </li>
<li> d) Download Spark:
<a href="http://d3kbcqa49mib13.cloudfront.net/spark-1.6.0-bin-hadoop2.6.tgz" onclick="trackOutboundLink
```

2) Make sure you have java installed on your machine.

3) Go to your home directory. (You can use the command in red)

cd ~

4) Unzip the folder in your home directory using the following command.

tar -zxvf spark-1.6.0-bin-hadoop2.6.tgz

5) Use the following command to see that you have a .bash_profile

ls -a

6) Next, we will edit our .bash_profile so we can open a spark notebook in any directory.

nano .bash_profile

7) Don't remove anything in your .bash_profile. Only add

1 setting path for spark

export SPARK_PATH=~/spark-1.6.0-bin-hadoop2.6 export PYSPARK_DRIVER_PYTHON="jupyter" export PYSPARK_DRIVER_PYTHON_OPTS="notebook" alias snotebook='$SPARK_PATH/bin/pyspark –master local[2]'

Notes

The PYSPARK_DRIVER_PYTHON parameter and the PYSPARK_DRIVER_PYTHON_OPTS parameter are used to launch the PySpark shell in Jupyter Notebook. The –master parameter is used for setting the master node address. Here we launch Spark locally on 2 cores for local testing.

For Python 3 Users

You have to add the line in red before you use alias snotebook='$SPARK_PATH/bin/pyspark –master local[2]' line or you will get the error in the image above. export PYSPARK_PYTHON=python3

Other useful PySpark tutorials

https://www.dataquest.io/blog/installing-pyspark/