**Lab 8 – Spark Configuration Guide and Recommendation System**

**Step 1: Anaconda / packages**

Please make sure you have the **newest version of Jupyter Notebook** installed. Then, connect to your instances and follow the steps to install other packages.

**Note:** You probably need to download the latest version of Anaconda to avoid conflictions.

Use the following commands (replace the old repo in Lab 6 & 7) when installing the Anaconda package to download the newest version which released in 2020.

**wget [http://repo.continuum.io/archive/](http://repo.continuum.io/archive/Anaconda3-4.0.0-Linux-x86_64.sh" \t "_blank)**[**[Anaconda3-2020.11-Linux-x86\_64.sh](http://repo.continuum.io/archive/Anaconda3-4.0.0-Linux-x86_64.sh" \t "_blank)**](https://repo.continuum.io/archive/Anaconda3-2020.02-Linux-x86_64.sh)

**bash** [**Anaconda3-2020.11-Linux-x86\_64.sh**](https://repo.continuum.io/archive/Anaconda3-2020.02-Linux-x86_64.sh)

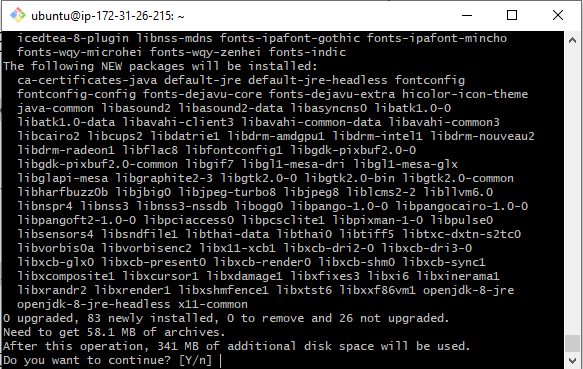
**Step 2: Install Java**

We will need java to install scala which we need for Spark. Go to the command line of your instance and type:

**sudo apt-get update**

Then install Java with:

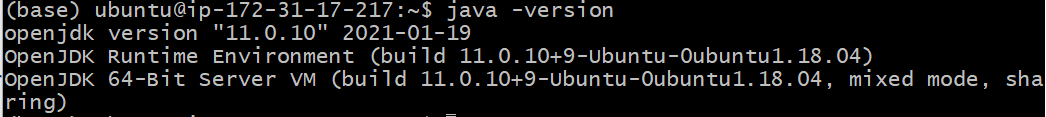
**sudo apt-get install default-jre**



Type Y to continue.

Check to see if you have successfully installed Java:

**java -version**



If you see the same as above, you are ready to go to the next step.

**Step 3: Install Scala**

Now, we can install Scala by running:

**sudo apt-get install scala**

Check if you have successfully installed scala:

**scala -version**



If you are able to see something similar to this, it means you have installed Scala successfully.

**Step 4: Install py4j**

We need to install the python library py4j, in order to this we need to make sure that pip install is connected to our Anaconda installation of Python instead of Ubuntu’s default. In the console we will export the path for pip:

**export PATH=$PATH:$HOME/anaconda3/bin**

Go to **Jupyter Notebook** installed on your instance and open a new Terminal on Jupyter Notebook not the ssh Terminal.

Now, install the py4j with pip:

**pip install py4j**

**Step 5: Install Spark and Hadoop**

Use the following commands to download and install Spark and Hadoop:

Download the package for installing Hadoop:

**wget** [**http://archive.apache.org/dist/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.7.tgz**](http://archive.apache.org/dist/spark/spark-2.4.5/spark-2.4.5-bin-hadoop2.7.tgz)

Please wait for the package to be downloaded completed and then extract the zip file:

**sudo tar -zxvf spark-2.4.5-bin-hadoop2.7.tgz**

Next, install pyspark:

**pip install pyspark**

Finally we need to set our Paths for Spark so Python can find it:

**export SPARK\_HOME='/home/ubuntu/spark-2.4.5-bin-hadoop2.7'**

**export PATH=$SPARK\_HOME:$PATH**

**export PYTHONPATH=$SPARK\_HOME/python:$PYTHONPATH**

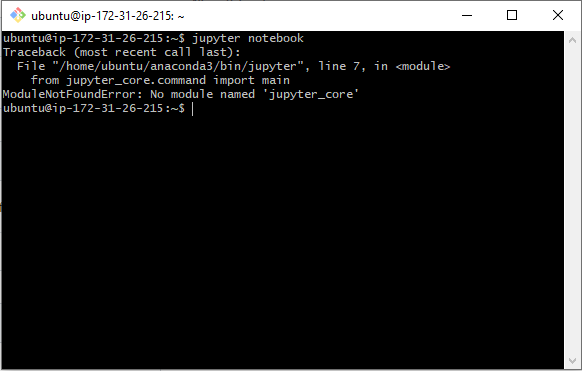
**Step 6: Create new .ipynb file**

You should have everything set up to run programs with Spark! Create a new Python3 file and follow the demos to build a recommendation system.

Skip this part if you did not encounter such error.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

If you are facing the issue below which stops you from launch the Jupyter Notebook:



Try the following command:

**pip install –upgrade jupyterthemes**

Then, try to launch again and you should be fine.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Step 7: Recommender System with Spark DEMO**

Please go to the Github repos to check out the recommender system examples with ALS model:

AWS example:

<https://github.com/ZhihanY/BYGB7990/blob/master/AWS%20Spark>

GCP Example:

<https://github.com/boomsaka/spring2020-bigdata-rec-spark-gcp-demo>