CS 686: Special Topics in Big Data

Spark Setup

Lecture 27

Daily Cluster Status Report

- Our cluster **does** have Spark installed!
- However, it is an ancient version
  - (like all the software on bass)
- In the interest of learning relevant skills, I would **strongly discourage** using Spark on bass
  - It'd be like learning Windows 98

Today's Goal: Setting up Spark

- Project 3 will shift our focus towards Spark
- I would recommend setting up a single-node installation on your own machine
- The first half of Project 3 will use the NAM dataset
- The second half tasks you with finding a dataset you're interested in
  - You will have the opportunity to complete this part in groups (up to 5 people per group)

Spark Setup – Mac OS

- Prerequisite: Java
- Install via homebrew:

  brew install apache-spark
- You're done! Run:

  spark-shell (Scala interactive shell)

  pyspark (Python interactive shell)

Spark Setup – Linux

- Prerequisite: Java, Scala
- Download at:

  https://spark.apache.org/downloads.html
- Extract:

  tar xvf spark-2.2.0-bin-hadoop2.7.tgz
- You're done! Run:

  spark-shell (Scala interactive shell)

  pyspark (Python interactive shell)

Windows

- Haven't tried yet, but this tutorial looks promising: https://medium.com/@GalarnykMichael/install-spark-on-windows-pyspark-4498a5d8d66c
- However, there is always the second option: use your AWS credits to launch a Linux machine and install there
  - Shut it down when you're not using it to save credits

AWS

- You should all receive a link to set up your educational AWS accounts
- Let me know if you run into problems
- For Project 3, I'd encourage you to get some experience working with the cloud if you haven't already
  - Elastic MapReduce
  - EC2

Recommended: Jupyter

- If you haven't used Jupyter Notebooks yet, then now might be a great time to dive in!
- With your python package manager (I use pip) install jupyter
- Next we need to configure pyspark:

```
export PYSPARK_DRIVER_PYTHON=jupyter
export PYSPARK_DRIVER_PYTHON_OPTS='notebook'
```

- When you run pyspark it'll start the notebook server

Recommended: Plotly

- Install the plotly python package to create interactive plots right from your Jupyter notebooks
- Great tool for visualization

Other Recommendations

- As cool as Jupyter is, I hate working anywhere other than my terminal
  - That's a thing, right? Someone else agrees, right??
  - ptpython is a somewhat similar terminal interface, has good vi keybindings, etc.
- Also, matplotlib is another great option for visualization

Diving In

- A few usage examples are available here:
- https://github.com/cs686-bigdata/beginning-spark.git