CS 686: Special Topics in Big Data

Hadoop Cluster & Setup Info

Lecture 23

Today’s Schedule

- Off-topic bits
- bass cluster
- Hadoop setup

Today's Schedule

- **Off-topic bits**
- bass cluster
- Hadoop setup

Neat Thing of The Day

- Recall our discussion on Bitcoin and the moral dilemma of wasting energy to produce imaginary money
- Well, here's one way to use that energy:
- https://qz.com/1117836/bitcoin-mining-heats-homes-for-free-in-siberia/

Today's Schedule

- Off-topic bits
- **bass cluster**
- Hadoop setup

So…

- We started the semester with 24 nodes
- Power outage 1: lost 5 power supplies
- Power outage 2: lost a power distribution unit (PDU)
- There’s 9 (???) nodes running now
    - As we are all well aware, the default replication level of 3 for HDFS ain’t gonna save us here…

My Initial Reaction: Panic Button

- I spent most of last evening in the fetal position in the corner of my office
- Luckily, the sun still came up this morning

Solution

1. Modifying the project requirements a bit

2. Doing analysis on our own machines

3. Eventually moving our work to the cloud

Project Requirements

- Since we're running out of time and now our cluster is dead, I will drop Deliverable II from this project
- If you started working on it already, don't worry – it'll be a part of Project 3

Local Analysis

- I’ll provide a link for you to download the dataset
  - However, it is quite large… Probably too large for many of our machines
- I’ll also provide a new, 25% sample version of the dataset
- For those of us without powerful machines, we’ll set up single-node Hadoop on our CS accounts

Cloud Analysis

- If you're itching to get some real-world experience with doing analysis in the cloud, get an AWS student account
  - Visit: https://aws.amazon.com/education/awseducate/ '
- This will be useful to have for P3 as well, but is not required

Today's Schedule

- Off-topic bits
- bass cluster
- **Hadoop setup**

Hadoop Setup: Mac

- Install Java, homebrew (http://brew.sh)
- brew install hadoop
- You know the rest:

```
yarn -jar project-2.jar
```

- This will run the job locally on your machine
- Done!

Hadoop Setup: Linux

- Install Java. If using dept machines skip this step!
- Download Hadoop from: https://hadoop.apache.org/releases.html
- Extract it in your home directory
- You'll need to modify your ~/.bash_profile
  - `export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64/`
  - `export PATH="${PATH}:${HOME}/hadoop-2.8.2/bin"`

Hadoop Setup: Windows

- ssh to a kudlick/g12 machine
- See previous slide :-)

Setting Up

- Let’s all make sure we can run a job on our local machines
- Download the mini dataset here:
  - http://www.cs.usfca.edu/~mmalensek/courses/cs686/projects/nam_mini.tdv.gz
- **Note**: Hadoop can read gzipped archives directly!
  - (and many other formats)