Lab 1

Big Data, Spring 2018

January 29, 2018

Labs Overview

- Some portion of most class meetings will be dedicated to the lab
  - Bring your laptops!

- Some labs may span multiple weeks

- Most labs have a **deliverable, due before the next class**, submitted via NYU Classes
  - You can talk with your classmates but need to submit your own deliverable.

Labs Overview

- Rough overview of topics/tools:
  - Relational algebra
  - SQL
  - Hadoop
  - Spark
  - Hive
  - AWS
  - NoSQL
  - Reproducibility
  - Visualization and Spatio-Temporal data
  - More TBD…

Computing Systems for the Semester

HPC@NYU: High-performance computing resources maintained by NYU
https://wikis.nyu.edu/display/NYUHPC/High+Performance+Computing+at+NYU

Dumbo: a 48-node cluster: https://wikis.nyu.edu/display/NYUHPC/Clusters+-+Dumbo

- Dumbo is old, used mainly for instructional purposes
- Lab 1: we will use Dumbo as a Hadoop cluster

Hadoop Clusters

Apache Hadoop: a framework for distributed storage and processing of very large data sets; runs on commodity computer clusters

Apache Hadoop consists of a storage part (HDFS: Hadoop Distributed File System) and a processing part which uses a MapReduce programming model

You specify the functions to execute, the Apache framework takes care of partitioning, distributing, and moving around data for you

Computing Systems for the Semester

- AWS (Amazon Web Services)
  - A collection of cloud computing services, also called web services, that make up a cloud-computing platform offered by Amazon.com
  - Service to provide large computing capacity quickly and cheaply

Lab 1: Setup and Intro

Open **NYU Classes -> Resources -> Labs -> Lab1.pdf**

Make sure your have your:

- HPC account username and password
- AWS username and password

See links in lab document for command-line tutorials

Part 1: HPC@NYU

https://wikis.nyu.edu/pages/viewpage.action?pageId=53859101

Part 2: Running Hadoop-Streaming on HPC

- **Apache Hadoop** is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware
- Consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce.
- Hadoop splits files into large blocks and distributes them across nodes in a cluster.
- To process data, Hadoop transfers packaged code for nodes to process in parallel based on the data that needs to be processed.

- In this part of the lab, we will
  - SCP files you need to dumbo
  - Copy files to HDFS
  - Run a Hadoop job

Running Example for Today

- We will be using a wordcount program
- Input: text file
- Output: files that count the number of occurrences of each word
- Typical example of a task well-suited to the MapReduce programming model
- We provide a couple different text files to use, but you try these exercises with your own if you'd like!
- Core idea behind MapReduce: dataset is *mapped* into a collection of (key, value) pairs, and then *reduced* over all pairs with the same key
  - Wordcount example: Each word mapped to pair (<word>, 1); reduction operation sums values for every pair with the same key, which gives the total number of occurrences of each word

Part 2: Steps

- Step 1: Log in
- Step 2: Create command aliases for Hadoop-Streaming
- Step 3: Run Hadoop-Streaming job

Part 3: Set Up AWS

- In this part, we will set up a public key/private key pair for use with AWS
- For Windows users using PuTTY, you should use PuTTYgen to convert the key AWS generates to a different file format (you won’t need this today, but may for future labs)