Big Data Lab

DS-GA 1004, Spring 2016
Monday, 4:55-7:35pm, 19 Univ. Pl. Room 102

January 25, 2016

Labs Overview

- Each week, some portion of the class will be dedicated to the lab portion
  - Bring your laptops!
- Some labs may span multiple weeks
- Rough overview:
  - Intro and setup
  - Relational algebra
  - SQL
  - Hadoop on AWS
  - Hadoop on HPC
  - NoSQL
  - Reproducibility
  - Visualization and Spatio-Temporal data
  - More TBD

Computing Systems for the Semester

- HPC Cluster at NYU
  - High-performance computing resources maintained by NYU
  - Dumbo: HPC's 44-node Hadoop cluster, running Cloudera CDH 5.4.5

- AWS
  - 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 Lab1 PDF and download Lab1 Resources folder from NYU Classes

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 0: Setup

- Make sure your have your HPC account username and password and your AWS username and password
- Make sure you have SSH/SCP tools (Windows, MacOS)
- Using the Linux command line - see links in lab document for tutorial and command cheatsheet

Part 1: Accessing HPC and Setup

- Logging in to HPC cluster
- Logging in to dumbo
- Create command aliases

Part 2: Running Hadoop 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

Part 3: Running Spark on HPC

- **Apache Spark** is an open source cluster computing framework
- Hadoop: two-stage disk-based MapReduce paradigm; Spark: multi-stage in-memory primitives (provides performance up to 100 times faster for certain applications)
- Well-suited to machine learning algorithms

- In this part of the lab, we will
  - Run a Spark job on dumbo

Part 4: AWS: Set up a Key Pair

- 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)

Part 5: Run a Hadoop Job on AWS

- We will now run our wordcount program on an Amazon EMR (elastic MapReduce) cluster
- EMR: Tool that automates provisioning of the Hadoop cluster, running and terminating jobs, and handling data transfer between EC2(VM) and S3(Object Storage)

- In this part, we will
  - Create an S3 bucket (Amazon's cloud storage) so we can access the files we need
  - Start up an Amazon EMR cluster
  - Run the wordcount program as a Hadoop streaming job

Part 6: Terminating your Cluster

- Incredibly important to do!
- Failing to terminate will use up all your class credits and then start charging your credit card!