Lab 1

Big Data, Spring 2017

January 23, 2017

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
- (Usually) a deliverable every week, to be submitted to NYU Classes before the next class period

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 Cluster at NYU
  - High-performance computing resources maintained by NYU
  - Dumbo: HPC's 48-node Hadoop cluster

https://wikis.nyu.edu/display/NYUHPC/Clusters+-+Dumbo

Hadoop Cluster

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 Hadoop 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 Lab1 PDF from NYU Classes

Main objectives:

- HPC and Linux/Unix introduction
- Accessing dumbo cluster
- Running a Hadoop job
- Setting up AWS account

Part 1: Introduction to HPC and Unix/Linux

Tutorial:

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

Part 2: Running a Hadoop job on dumbo

- Logging in to dumbo
- Create command aliases
- Run an example job

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
- 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

Running the Hadoop job

- **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
  - Copy files to/from HDFS
  - Run a Hadoop job using Hadoop Streaming
  - Copy output files to local machine

Deliverable

Deliverable for today: submit your examplemerged.out file to NYU Classes under “Assignments”->”Lab 1”.

Part 3: AWS Setup

- In this part, we will
  - Set up an AWS account
  - Sign up for AWS Educate
  - 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)