Lab Assignment 1: MapReduce / Hadoop

Notes

You can work on this assignment in teams of two.

Objectives

- Understand the MapReduce programming model.
- Setting up Hadoop on a single node and on a cluster of nodes.

Overview

It is required to install Hadoop on both single node cluster and multiple nodes cluster. Next, you will practice running few HDFS commands and executing Hadoop jobs. You can use the following command to download Hadoop on your machine:
`wget http://www-eu.apache.org/dist/hadoop/common/hadoop-2.7.3/hadoop-2.7.3.tar.gz`
You can extract the downloaded file using:
`tar -xzvf hadoop-2.7.3.tar.gz`

Setting up Hadoop

- You will need to download the latest stable version of Hadoop (2.7.3) from this link: `http://hadoop.apache.org/releases.html`.
- Setup the downloaded Hadoop version on your machine. These are the steps that you will need to follow: `https://goo.gl/8KVyGJ`. You can do this step before the lab time.
- During the lab, you will need to setup Hadoop on a cluster of machines using the following steps: `https://goo.gl/KLyzFU`. You have the choice to use one of the following options to setup the Hadoop clusters: (1) AWS EC2 instances; (2) Lab machines; or (3) your laptops.

Useful resources: This tutorial can help you setup hadoop: Part I and Part II

HDFS

- Create a directory called `input` in your home directory.
- Download the following text files from the Gutenberg project, in Plain Text UTF-8 format (hint: you can use wget):
  - The Outline of Science, Vol. 1 (of 4) by J. Arthur Thomson
  - The Notebooks of Leonardo Da Vinci

  - Ulysses by James Joyce
  - The Art of War by 6th cent. B.C. Sunzi
  - The Adventures of Sherlock Holmes by Sir Arthur Conan Doyle
  - Encyclopaedia Britannica, 11th Edition, "Brquigny, Louis Georges Oudard Feudrix

- Download the above data and store them to the `input` directory on your machine.
- Create a new file `mydata.txt` in the `input` directory. Open the file and write to it this line: `CS432 FirstStudentID SecondStudentID`. Repeat this line four times in the file.
- Copy the `input` directory from your local disk to HDFS. You can use the command: hadoop fs -copyToLocal /home/userid/input /home/userid/input. The first path is the source, which is on your local disk. The second path is the destination, which is on HDFS.
- Now check that the files were already copied using this command: hadoop fs -ls /home/userid/input

Running Hadoop Jobs

At this step, you run Hadoop jobs on the data loaded on HDFS.

- You need to build the WordCount example described in this tutorial. Name the created jar file `wc.jar`.
- You are now ready to run the jar file using:
  `hadoop jar wc.jar /home/userid/input /home/userid/output`
- Check the output files created in the `/home/userid/output`.
- Copy the output directory to your local disk using:
  `hadoop fs –get /home/userid/output /home/userid/output`. You can also use `copyToLocal` or `getmerge`
- In the output file, check that `CS432` and your SIDs are counted four times. Check the word count for various words that appeared in the input files.
- You can now update the words count to filter the words based on a second input file. Therefore, only the words that appear in that dictionary file will be counted.

Resources

- HDFS shell commands
- MapReduce Tutorial