Welcome to ANLY 502

ANLY 502 combines three elements --- architectures for working with massive data (plumbing), massive-data algorithms and programming (practice), and how big data impacts our society (policy). This two-week assignment is designed to help you think about these three Ps.

1. The standard Hadoop “Hello World” program is a program called WordCount that prints a histogram of the most frequent words. Run the standard WordCount demo in Java on your Cloudera VM. This should be straightforward, as there are many WordCount demos out there. Load the Shakespear plays from the ZIP file into HDFS on your VM and generate a report of the 10 most frequent words in the corpus. Please submit:
   1. Your Java source file(s).
   2. The command line you used to run the job.
   3. The console messages created by your Hadoop job.
   4. The 10 most common words and their frequency count.
2. Re-implement the WordCount in Python using the Streaming API. Submit:
   1. Your python source file(s).
   2. The command line you used to run the job.
   3. The console messages created by your Hadoop Streaming API job.
   4. The 10 most common words and their frequency count.
3. Re-implement the WordCount using mrjob. The mrjob program gives you a higher level abstraction for working with Hadoop in Python, and it allows you to run in three ways: 1) locally, without using Hadoop at all. 2) Locally on Hadoop. 3) Remotely on Amazon EC2. Submit:
   1. Your python source file(s).
   2. The command line you used to run the job locally, without Hadoop.
   3. The non-Hadoop console output.
   4. The 10 most frequent words.
   5. The command line you used to run the job locally with Hadoop.
   6. The Hadoop console output.
   7. The 10 most frequent words.

Extra ideas:

* Count the number of words in the Wikipedia, and graph the top 100.
* Perform a Join in Hadoop
* Calculate PI with monte carlo simulation

References:

* Apache MapReduce Tutorial (WordCount in Java): <https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html>