Çankaya University
Department of Computer Engineering

CENG 595 Distributed Data Processing and Analysis
Spring 2017

Homework 1

Due by Mar 7 (Tuesday) 6pm

Subject: Hadoop environment and Word Count problem

You will run Hadoop on a personal computer (or a server if you have) in a container (Docker) environment. You will first (1) install Docker (if you do not have it), (2) download a dockerized version of Hadoop to Docker environment, (3) develop WordCount example in Eclipse IDE and export the jar file to dockerized Hadoop environment, (4) download a large text file to test WordCount on, (5) run WordCount and check the results.

1) **Download Docker**
   Containers are the latest "virtualization" technology and Docker is the most popular container technology these days. So, we chose Docker to run a virtual Hadoop cluster (single node :-) on. Download and install Docker (Community Ed.) to your computer from https://www.docker.com

2) Download a dockerized Hadoop distribution. There are a few, we chose this one from Docker Store: https://store.docker.com/community/images/sequenceiq/hadoop-docker
   Follow the instructions and install Hadoop on your Docker environment. That is:
   a. Download Hadoop for Docker:
      docker pull sequenceiq/hadoop-docker:2.7.0
   b. Start the container:
      docker run -it –v /tmp:/tmp1 sequenceiq/hadoop-docker:2.7.0 /etc/bootstrap.sh -bash
      This will start a **shell command line** that you will use to run Hadoop programs. Here /tmp:/tmp1 is the shared directory between your computer and dockerized Hadoop. /tmp should be on your computer, and /tmp1 will be on dockerized Hadoop. You can drop files in either and you will see them in the other. Now copy the following files (you will need to develop Hadoop programs in Eclipse) in the command line in docker:
      cd $HADOOP_PREFIX
      cp share/hadoop/mapreduce/hadoop-mapreduce-client-core-2.7.0.jar /tmp1
      cp share/hadoop/common/hadoop-common-2.7.0.jar /tmp1

3) Download and install **Eclipse** from: www.eclipse.org
   Download **WordCount.java** from course piazza site (Resources/Assignment).
   In Eclipse:
   a. File/New Project/Java Project (name it **WordCount**)
   b. Add **WordCount.java** to the src directory
   c. Project/Properties, Libraries, Add External Jars, and select the following two:
      hadoop-common-2.7.0.jar
      hadoop-mapreduce-client-core-2.7.0.jar

d. File/Export/Java/Jar File, name export destination: **/tmp/WordCount.jar**
(you just compiled and constructed a jar file, and copied it to the shared directory, to docker environment)

4) Download the following text file:
http://www.gutenberg.org/cache/epub/100/pg100.txt
(The Complete Works of William Shakespeare)
And save it to the **/tmp/pg100.txt** (shared directory)

5) Now ready to run your first Hadoop program? Go to shell command line of Dockerized Hadoop and run the following:
   a. Copy the input text file to HDFS:
   ./bin/hadoop fs -put /tmp1/pg100.txt input/
   b. Run WordCount on Hadoop:
   ./bin/hadoop jar /tmp1/WordCount.jar WordCount input output
   WordCount will run on all text files in input directory and write the output to output directory.
   c. Check the output files:
   ./bin/hadoop fs -cat output/*

6) Submit the output file via webonline by the deadline.