

MP5: MapReduce for Tweet Analysis

150 points

Due: May 2nd, 2022, at 11:59pm

1 Overview

For this assignment you will use a Docker-based Hadoop container. Please follow the steps below:

1. Download and install “Docker Desktop” on your machine, from here.
2. Download and install the “sequenceiq/hadoop-docker” container image from here. You do not need to build the image. Simply “pull” the image, as described at the URL above.
3. From a command line window in your system, start the container:

```
$ docker run -it --volume <local path where you store  
data and files>:/mnt/<docker folder name where you will  
access local files> sequenceiq/hadoop-docker:2.7.0  
/etc/bootstrap.sh -bash
```

Pay attention to the local (i.e., local to your computer) and “docker folder name” so that you are able to access local files from inside the container. By default, the container is not allowed to access local files.

4. Once you start the container, you will see something like:

```
Starting sshd: [ OK ]  
Starting namenodes on [1ff7118197e8]  
1ff7118197e8: starting namenode, logging to /usr/local/hadoop/logs/hadoop-root-namenode-1ff7118197e8.out  
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-root-datanode-1ff7118197e8.out  
Starting secondary namenodes [0.0.0.0]  
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-root-secondarynamenode-1ff7118197e8.out  
starting yarn daemons  
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn--resourcemanager-1ff7118197e8.out  
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-root-nodemanager-1ff7118197e8.out
```

And you will obtain a shell prompt. The hadoop environment is located in /usr/local/hadoop inside the container. You may find useful to execute the following commands, inside the container:

```
$ export HADOOP_CLASSPATH=/usr/java/default/lib/tools.jar
$ export PATH=$PATH:/usr/local/hadoop/bin
```

5. Download a “Tweets” data file from (it is about 2.7GB) [here](#).

Each input file contains a series of Twitter tweets in the following format:

```
T 2009-06-01 00:00:00
U http://twitter.com/testuser
W Post content
Empty line
```

where the first line is the time of the tweet, the second line is the user who posted the tweet, the third line is the actual content of the post, and the fourth line is an empty line.

6. Assuming the downloaded file is “/mnt/your_docker_folder_name/where_you_will_access_local_files/tweets.txt”, put the file in HDFS by executing the following commands:

```
$ hdfs dfs -mkdir /user/root/data
$ hdfs dfs -put \mnt/<docker folder name where you will
access local files>/tweets.txt /user/root/data
```

Now you can check if the file is there:

```
$ hdfs dfs -ls /user/root/data
```

7. Now you are ready to compile (before compilation, check Step 8 below, regarding the applications you need to write) your MapReduce application. Let’s assume it is called WordCount.java. Execute the following commands:

```
$ hadoop com.sun.tools.javac.Main WordCount.java
$ jar -cvf WordCount.jar WordCount*.class
```

8. Now you are ready to execute your MapReduce application (Step 6) using the data from Step 5:

```
$hadoop jar WordCount.jar WordCount /user/root/data /user/root/output
```

2 Time of Day Most Often Tweets

Write a Mapreduce application and use the given datasets to analyze what time in a day do users post tweets most often?

- Divide a day into 24 hours, and answer the question: how many tweets are posted during each hour, e.g. 0:00 - 0:59, 1:00 - 1:59, ..., 23:00 - 23:59?
- Plot a graph that shows the histogram from the above result, i.e. x-axis is the time (e.g. 0:00 - 0:59) and y-axis is the total number of tweets posted during this hour in the dataset.

3 Time of Day When Usually People Go To Sleep

Write a Mapreduce application and use the given datasets to analyze when do people usually go to sleep?

- Here we make an assumption that people may post tweets that contain the keyword sleep before they go to sleep.
- Using a similar approach as above, answer the question: how many tweets that contain the keyword “sleep” are posted during each hour, e.g. 0:00 - 0:59, 1:00 - 1:59, ..., 23:00 - 23:59?
- Plot a graph that shows the histogram from the above result, i.e. x-axis is the time (e.g. 0:00 - 0:59) and y-axis is the total number of tweets posted during this hour in the dataset.
- Note: this may require a custom RecordReader class, as the default one in Mapreduce reads the file line-by-line while here multiple lines constitutes a single tweet record.