Map and Reduce Lab

Due Tuesday, January 28, 2020

Setup a single node Hadoop cluster and run the sample program

1. For MAC or Linux computers, please follow the instructions below:

https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html

2. For Windows computers, please follow the instructions below:

https://muhammadbilalyar.github.io/blogs/How-to-install-Hadoop-on-Window-10/

Note1: the step 4 in Configuration section, the **hdfs-site.xml** should be similar to the following:

```
<configuration>
	<property>
		<name>dfs.replication</name>
		<value>1</value>
	</property>
	<property>
		<name>dfs.namenode.name.dir</name>
		<value>/c:/hadoop/data/namenode</value>
	</property>
	<property>
		<name>dfs.datanode.data.dir</name>
		<value>/c:/hadoop/data/datanode</value>

	</property>

</configuration>
```

Note3, in the step 4 in Testing section, the URL should be http://localhost:9870 instead (depending on which version you installed).

Following the instruction below, run the sample program.
https://muhammadbilalyar.github.io/blogs/How-to-Run-Hadoop-wordcount-MapReduce-Example-on-Windows-10/

3. Download, compile, and run the sample program **`WordCount.java`**, using **`warandpeace.txt`** as the input. Save the final output in a text file.

To compile and run the program, the **`classpath`** environment variable should be similar to the following:

```
classpath=C:\hadoop\share\hadoop\common\hadoop-common-3.0.0.jar;C:\hadoop\share\hadoop\mapreduce\hadoop-mapreduce-client-core-3.0.0.jar
```

If you want to compile the program using eclipse, you should add the above two jar files in the referenced libraries of your eclipse project.

Run the sample program on an AWS Elastic MapReduce (EMR) cluster

1. Please sign up for an AWS Educate for Students account (wait for more instructions later) (**https://aws.amazon.com/education/awseducate/**).

You should be able to receive $50 AWS credit, which should be enough for you to complete all your assignments, if you use it carefully. **However, you are responsible for all the charges over the credits. We cannot reimburse you for any charges**.

2. Following the instructions in the video below, start a AWS EMR cluster and run your jar file. Download the log files: **syslog.txt** and **controller.txt**.

https://www.youtube.com/watch?v=JDk-LYJMzEU

Report

1. Write a brief report (one page) to describe the major steps you did and a brief reflection on what you learned.
2. Save the console output of your run at your local Hadoop single node cluster, include it in your report.
3. Download the AWS log files: syslog.txt and controller.txt, include them in your report.

Please submit your report to the "P1" folder on the Canvas.

Notes

1. Map Reduce Tutorials:

https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

2. AWS EMR Documentation:

https://aws.amazon.com/documentation/emr/