Assignment 1

A) In this week's module, we learned how to download, install, configure, and get Hadoop up and running on our private VM. Since, we are going to need a Hadoop environment for our rest of the assignments, let us set up a Hadoop environment for our use. For this week's assignment, please execute the following:

1) Download and Set up VirtualBox (video B). Download the VirtualBox 6.1.36 and install it in your operating system.
2) Download and set up Ubuntu (Video C). Download Ubuntu 20.0.4 and install it in your operating system. The website is https://releases.ubuntu.com/20.04.4/ and the download for Desktop distribution ISO file is https://releases.ubuntu.com/20.04.4/ubuntu-20.04.4-desktop-amd64.iso .
3) Set up Hadoop. (Video D) Follow the instructions listed in the document Steps-To-Configure-A-Single-Node-Hadoop-3-YARN-Cluster.pdf to setup Hadoop. You should install Apache Hadoop 3.3.1. You can get this from the archives - https://dlcdn.apache.org/hadoop/common/
4) In step 28, you downloaded one book (Moby Dick). After completing the installation, download 3 more books. They are
   - http://www.gutenberg.org/files/74/74-0.txt (The Adventures of Tom Sawyer)
   - http://www.gutenberg.org/files/98/98-0.txt (A Tale of Two Cities)
   - http://www.gutenberg.org/files/1400/1400-0.txt (Great Expectations)
5) Run the wordcount example with a total of 4 books in /in folder (which is 4 files).

What to turn in?

Once you have the environment up, daemons are running and executed the WordCount job successfully, please turn in the following as a word document.

1) NameNode UI console first page screen shot
2) ResourceManager UI console first page screen shot
3) JobHistoryServer UI console first page screen shot
4) A screen shot of jps command output. Note that when you run the jps command, the first column is the Unix process_id.
5) What ports are the various Hadoop daemons listening on? In order to find this out, you should ensure you first install net-tools
   $ sudo apt install net-tools
   { it may ask for password – hdadmin}

   And,

   then execute a command similar to the one below:

   ```
   $ netstat –anp | grep LISTEN | grep process_id/
   ```

   where *process*_id is the pid in the output of jps. Also, ensure you place a slash(/) at the end the process_id. The command will list out into standard output a bunch of lines. The output in the 4th column of every line should have IP address such as 127.0.0.1 or 0.0.0.0 followed by a colons followed by a number. The number is the port number. Note: There may be multiple ports in which a daemon may be listening on.

6) How many times does the word “family” (to be precise the sequence of characters “family”) appear across all the books (4 books – If you recall you downloaded one book as part of installation, but you run the program with 4 books)? So, you need to run your job with 4 books.

When you run Yarn MapReduce job such as

```
$ yarn jar /usr/local/hadoop/hadoop-3.3.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.1.jar wordcount /in /out
```

wordcount is the name the main class in the `hadoop-mapreduce-examples-3.3.1.jar` , **/in** is the name of the input folder, **/out** is the name of the output folder. The input and output folder names can be arbitrary. The requirement is that the input folder must exist when the wordcount job runs, but the output folder should not exist when the job runs. If it already exists, then it will give you an error. In such a case,
You can run with a new output folder that does not exist such as:

```
$ yarn jar /usr/local/hadoop/hadoop-3.3.1/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.1.jar wordcount /in /out2
```

or

delete the existing output folder

```
$ hdfs dfs –rm –r /out
```

When you run the job the second time, Hadoop requires that the output folder not exist. So, you need to delete the existing output folder or give the output folder to be a different name. Please note that we are looking for the

word “family” literally in lowercase, without any punctuation.
Note that in Unix you can pipe commands such as the following

```
$ hdfs dfs –cat /out/part-r-00000 | grep “dance”
```

will list all the lines that have dance.

Please use Assignment 1 Solution template to turn in the assignment