CSC 4760/6760, DSCI 4760 Big Data Programming

Assignment 1

Due Date: 10 PM ET, 10/1/2022

1. (**100** points) (Setting up Hadoop on Ubuntu Linux and running the WordCount example)

This assignment aims at letting you setup Hadoop on Ubuntu Linux – via the command line. After the installation of Hadoop, you need to run the WordCount example.

Source Code and Datasets:

The java source code is given in the file "WordCount.java". You need to run it on two datasets:

1) test.txt

2) peterpan.txt

Report:

Please write a report to explain the key steps. You must take screenshots. You must also explain the commands (one line each on which command does what). You must include the following key steps in your PDF report.

1) Setup Hadoop on Ubuntu Linux [Command Line installation is important from the (open-book) Exam perspective].

2) For CSC 6760: How to Setup HDFS and upload the datasets "test.txt" and "peterpan.txt" into HDFS.

2) **For CSC/DSCI 4760: Not required to setup/use HDFS. Use the datasets "test.txt" and "peterpan.txt" from local File System (FS).**

3) Create a project in Eclipse or VS Code or any IDE that you prefer, import the "WordCount.java", configure the project, and export the "WordCount.jar" file.

4) Run the "WordCount.jar" file on "test.txt" and "peterpan.txt" respectively.

Required submission materials:

a) The report should be a PDF file.

b) Please also submit the output files, which should be renamed as follows.

The output file of "test.txt" must be "test_output.txt".

The output file of "peterpan.txt" must be "peterpan_output.txt".