CSC 4760/6760 Big Data Programming

Assignment 1

Due Date: 11:59 pm, Wednesday, Feb 17, 2021

1. (100 points) (Setting up Hadoop and running the WordCount example)

This assignment aims at letting you setup Hadoop on your KVM on Qubit. After the installation of Hadoop, you need to run the WordCount example on your KVM.

Please follow the instructions provided in the slides "Setup Hadoop On Ubuntu.pptx". If you have any questions, please talk with the instructor or the TAs. We will help you setup Hadoop on your KVM.

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 may take screenshots. You may also explain the commands. You may include the following key steps.

1) Setup Hadoop

2) Setup HDFS and upload the datasets "test.txt" and "peterpan.txt" into HDFS.

3) Create a project in VS Code, import the "WordCount.java", configure the project, and export the "WordCount.jar" file.

4) Open a terminal, and run the "WordCount.jar" file on "test.txt" and "peterpan.txt" respectively.

Required submission materials:

a) The report should be a PDF file. Please use a text editor, such as Microsoft Word, to write a report. Please transfer the file into a PDF file and then submit it.

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

The output file of "test.txt" should be renamed as "test_output.txt".

The output file of "peterpan.txt" should be renamed as "peterpan_output.txt".