**Drexel University**

**College of Computing and Informatics**

**INFO 323 – Cloud Computing and Big Data**

**Assignment 2**

**Due Date: Sunday, January 30, 2022**

**A. Requirements**

1. Provide full and detailed answer to each question.
2. Work individually. The Drexel University Academic Honesty Rules and Procedures (as stated in the student handbook) will be adhered to strictly.
3. There are 5 questions. The total marks are 100.

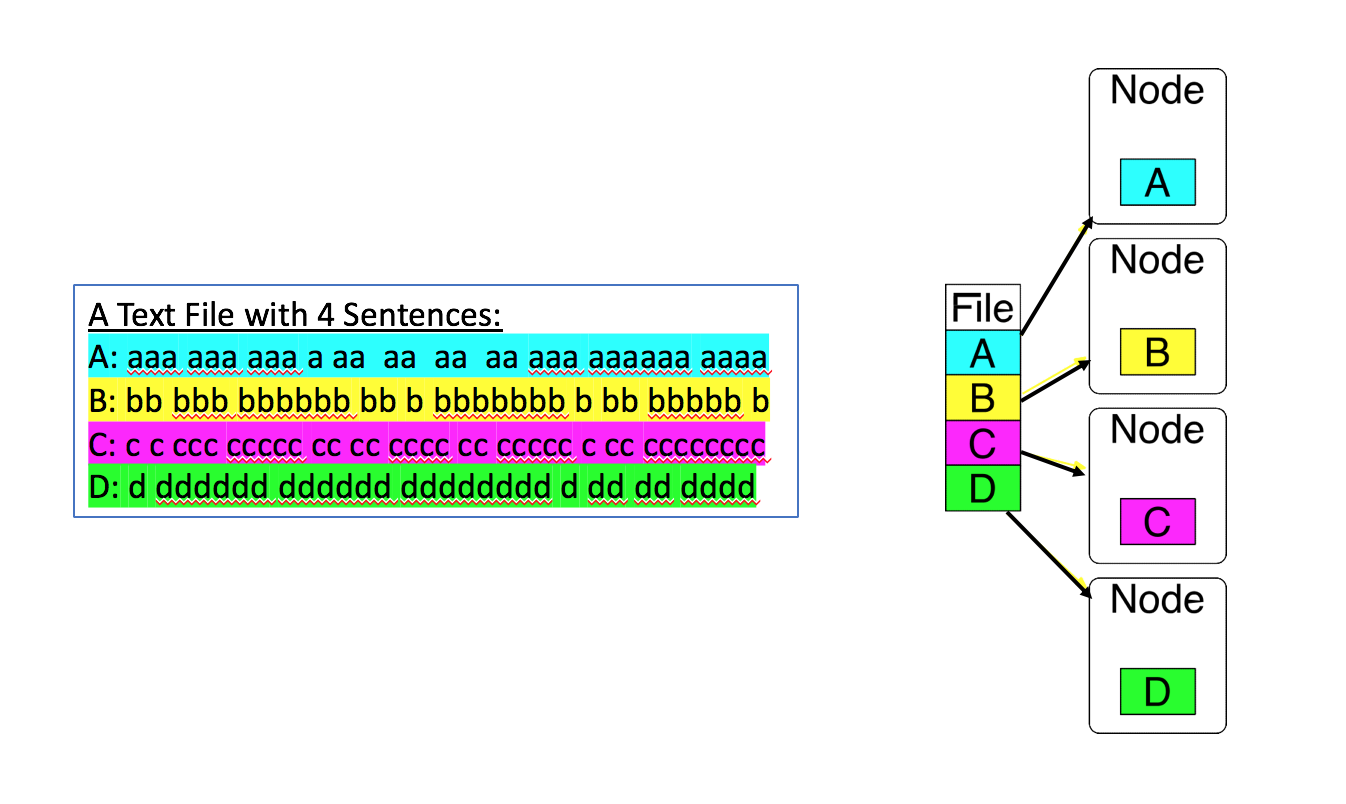
**B. Questions**

**Question 1 [15 marks].** Explain the following terms and discuss the roles they play in cloud computing:

1. Distributed file systems
2. Commodity clusters
3. MapReduce

**Question 2 [15 marks]**. Describe the key components and workflow of MapReduce’s parallel computing process. Describe how MapReduce coordinates communication and handles machine failures.

**Question 3 [15 marks]**. Create a text file containing 4 real sentences. Assume you have a cluster with 4 nodes. Initially, the 4 sentences are distributed onto the 4 nodes. Each sentence is on a node (see Figure 1.) Describe the MapReduce process for computing the frequency of each word in the file. List the steps and show the intermediate and final results.



**Figure 1: Initial Setting of a 4-node Cluster for Computing Word Counts by MapReduce**

**Question 4 [15 marks]**. **Hands-On Exercise.**

Create a GCP Dataproc cluster and SSH connect to the master node. Complete the following hands-on exercises in a terminal of the master node. At each step, you are required to capture screenshots to show the commands and results.

**Step 1**. In the terminal of the master node, create a folder named “info323-assign2” in the home directory on the master node.

**Step 2**. Download a novel from https://www.gutenberg.org/browse/scores/top. Upload it to the folder “info323-assign2” as “a-novel.txt”.

**Step 3**. Create a directory /user/wc-input in HDFS and copy the file a-novel.txt to the HDFS folder:

hdfs dfs -put ~/info323-assign2/a-novel.txt /user/wc-input/

**Step 4**. List the file in the HDFS folder:

hdfs dfs -ls /user/wc-input

**Step 5**. Check the Hadoop MapReduce example programs: Hadoop comes with several example MapReduce applications. You can check the examples by running:

hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar.

*We are interested in running WordCount*.

**Step 6.** Check the WordCount command line arguments: You can learn how to run WordCount by examining its command-line arguments:

hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount

**Step 7.** Run WordCount on “a-novel.txt”. If the output folder /user/wc-output exists in HDFS, you need delete the folder first.

hadoop jar /usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount /user/wc-input/a-novel.txt /user/wc-output

**Step 8.** Look inside the output directory /user/wc-output**:** The directory created by WordCount contains several files. List the content of the directory:

hdfs dfs -ls /user/wc-output

**Step 9.** Copy the output folder in the HDFS to a folder named output in the master node:

hdfs dfs -get /user/wc-output output

**Step 10**. Combine all the partitions in the output folder as a file named counts.txt:

cat part-r-0000\* > counts.txt

**Step 11.** Check how many lines in the file counts.txt**.**

**Step 12.** View some of the WordCount results**:** Run a command to print out 20 lines **in the middle of the file** counts.txt.

**Question 5 [40 marks]**. **Programming Exercises in Spark**.

In this exercise, you are asked to write Spark program to count the flights’ destination countries in the given file “flights-dest-origin.csv”. Each line in the file consists of 3 fields, Destination, Origin, and NumberOfFlights. The Spark program should parse the lines in the file and output each unique destination country name along with its total number of occurrences in the Destination field.

You are asked to complete this question in the Jupyter notebook “INFO323-Assignment2-Question5.ipynb”:

* Create a Dataproc cluster which links to a Google storage bucket.
* Download the Jupyter notebook and the data file “flights-dest-origin.csv” to your local file system.
* Upload the Jupypter notebook to the folder ‘notebooks/jupyter’ in the Google storage bucket that has been linked to the Dataproc cluster.
* Open the Jupyter notebook in the Dataproc cluster.
* Follow the steps in the notebook to complete the exercises.

**Submit your completed Jupyter notebook and the two output files (no screenshot needed). Answer the following question**:

* Discuss the differences between the two exercises of using Google File System and HDFS in terms of data and resource management (such as data partitioning, data replication, fault tolerance management, resources management, etc).

**C. What to Hand In**

1. A Microsoft Word document including the following items:
2. Your name
3. Course number and title
4. Assignment number
5. Assignment questions and detailed answers to the questions
6. The completed Jupyter notebook and required outputs for the programming exercise.

**D. How to Hand In**

1. Please name your assignment WORD file as **INFO323-assign2-yourFirstName-yourLastName.docx**.
2. Submit your assignment files through the course website in the **Blackboard Learn** system.

**E. When to Hand In**

1. Submit your assignment no later than **11:59pm** in the due date.
2. There will be a 10% (absolute value) deduction for each day of lateness, to a maximum of 3 days; assignments will not be accepted beyond that point. Missing work will earn a zero grade.

**F. Written Presentation Requirements**

Images must be clear and legible. Assignments will be judged on the basis of visual appearance, grammatical correctness, and quality of writing, as well as their contents. Please make sure that the text of your assignments is well-structured, using paragraphs, full sentences, and other features of well-written presentation. Text font size should be at least 11 points.