Assignment #3 - Hadoop & Spark

Due: 30th June 2024

Content Covered

Hadoop , Page -rank , Spark

Homework Overview:

This assignment will provide you with hands-on experience in writing and executing Spark code, as well as conducting subsequent analysis. You will begin by installing and configuring Spark, then proceed to write fundamental programs including a Dijkstra's shortest path algorithm, and a Page-Rank algorithm. Following the implementation phase, you will be analyzing the data flow within Spark applications, gaining insights into how Spark processes and manipulates data across distributed computing environments. Through this comprehensive approach, you'll not only develop practical programming skills but also deepen your understanding of distributed data processing concepts using Spark.

General Homework Requirements:

- **Work Environment**: This homework can be written in PySpark or Scala.
- **Programming**: You can use Jupyter Notebook, Jupyter Lab or Google collab
- **Academic Integrity**: You will get an automatic F for the course if you violate the academic integrity policy.
- **Teams**: This assignment is a team work. You are permitted to work with your team mates .

Submission Format:

1. Source Directory: All input data files, and code implementations should be organized within a specific directory named "src/" This directory will contain both the input datasets and the code files required for the assignment.
2. Report: Prepare a comprehensive report containing answers to all questions posed in the assignment. Each answer should include suitable proofs or evidence to validate the authenticity of your submission and demonstrate that the outputs are legitimate. This report should be well-structured and provide clear explanations for each question, along with any necessary supporting materials, documentation, and required screenshots.

By adhering to these refined requirements, you will ensure that your submission is well-organized, thoroughly documented, and adequately substantiated, thereby demonstrating your proficiency in completing the assignment successfully. **Failure to adhere to the specified submission format will result in a deduction of 3 marks. All submissions must follow the prescribed structure to ensure consistency and clarity. Submissions must be made on Brightspace by 30th June 11:59 PM. We will consider 1 day late submission with 15% penalty. After 1 day We will not accept**

Any submission. Even a delay of 1 minute will result in a score of 0 for the assignment. It's recommended to upload your assignment at least a few hours before the due date and time to avoid any last-minute issues.

Setup:

- To prepare your development environment for this homework you must first install and set up PySpark. To install PySpark, follow the instructions here: https://spark.apache.org/docs/latest/api/python/getting_started/install.html
- Use the question1.txt file for your Part-1 of the assignment for calculating the shortest path.
- Run the code file "code_2.py" attached and you will get a text file named question2.txt indicating page ranks of different nodes which is used for Part – 2 of the assignment.

Questions:

Part-1 Review paper on Hadoop/HDFS .

[30 Marks]

1. Write a comprehensive 500-word review of the paper on Hadoop **" Hadoop Distributed File System"** Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler
This paper is available on Brightspace under the resources tab.

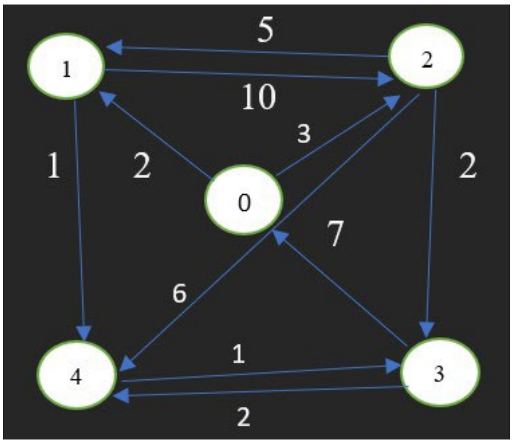
• Usage of any other online sources other than provided in this course will result in an automatic 0 in this assignment.

• Usage of any Artificial Intelligence tools to generate text will result in a -5 in the course

Part-2 Implement and analyze Dijkstra's Shortest Path algorithm.

[4X10=40 Marks]

1. You must write a basic Dijkstra's shortest path algorithm for the text file provided as part of the assignment. Where the first column of each row is the initial node, the second column of each row is the destination, and the third column is the weight associated with the connection. The graph representation of the nodes from the question1.txt file is as follows:



2. The algorithm should read the file, compute the shortest path between the first node **indicated by 0** to the node **indicated by 4** and print out the value. Additionally, save the distances to all the nodes in a text file named output_1.txt.
3. Additionally, save the shortest distances to all the nodes in a text file named

output_1.txt and provide a screenshot of the same.

4. How many stages is execution broken up into? Explain why. Include a screenshot of the DAG visualization from Spark’s Web UI.

Part-2 Implement and analyze Page-Rank algorithm.

[3X10=30 Marks]

1. You must write a basic page-rank algorithm considering the text file that is generated (question2.txt). It is a simulated network of 100 pages and its hyperlink.
   The algorithm should take the network provided and evaluate the page rank for all the webpages or nodes.
2. Find the node with the highest and the lowest page rank and provide a screenshot of the same. Explain with the practical approach of why your highest and lowest page ranks are correct.
3. How many stages is execution broken up into? Explain why. Include a screenshot of the DAG visualization from Spark's Web UI.