MSDS 694 Distributed Computing Group Project

Diane Woodbridge, Ph.D.
Fall 2021

Task 1. Create/Join a Team and Select Data Sets. (5 pt)

1. Create a team of 3-5 students and make sure that everyone joins the Canvas’ Project Group.
2. Submit a 1-page data description.
    - a. Each group member proposes one or two topics.
        - i. List Student Name, Data titles, List data sources (URLs), Size, Reasons you chose, Possible analytic goals.
        - ii. If you plan to collect your own data (web crawling, smartphone application, IoT application, etc.), please describe your specific plans and timeline.
        - iii. Data should be at least 2GB and more than 1M records/rows.
    - b. Each group chooses one data set.
        - i. Describe the reasons why you chose.

Task 2. Loading and Saving Data on AWS S3 (5 pt)

Watch the video on “Group Project -> Task 2 – Loading and Saving Data on AWS S3” and submit a 1-page (.pdf) including the following.

1. A screenshot of your S3 bucket.
2. A screenshot of the bucket size (and the number of items - optional).

```
aws s3api list-objects --bucket BUCKETNAME --output json --query
"[sum(Contents[].Size), length(Contents[])]”
```

3. A screenshot of your code shows you can connect and retrieve data from your S3 bucket locally.

Task 3. Data Processing and Visualization on EMR (10 pt)

Deliverables

1. **7 minutes presentation (.mp4) and slides (.pdf)**
    - Data Description
    - Data Processing Goal
    - Preprocessing/analytics outcome
    - Each member’s cluster setting and execution time comparison
        - Ex. 1) 1,2,3,4,5 node cluster 2) 3 node clusters with different specs (CPU, Memory, Disk, etc.)
    - Lesson Learned
2. **Jupyter notebook code (.ipynb)**
    - For 1) EDA and 2) Visualization
3. **Plot.ly HTML file (.html)**

Work Policy

Everyone is expected to work the same amount. After Task 3, everyone will submit a peer-review that reflects each member’s contribution. **The final grades will be given based on your contribution.**