BDCC Assignment: Mini Project

This deliverable has 40% weightage in the Consolidated Total Score.

Deliverables:

- 2 solution notebooks are required. The purpose is explained in the Assignment Instructions below. The solution notebook can be exported in .***dbc*** or ***.ipynb*** format and submitted.
- The dataset should be submitted along with the code or the link to the dataset should be mentioned in the data description of the notebook.
- ***There may be penalty if code clarity and documentation is not proper.***

Assessment Weightage:

A 70% weightage will be given for accomplishing the tasks in the assignment. Another 30% will be given for the following:

1. Code clarity
2. Documentation (use markdown)
3. Dataset description
4. Problem or insight description
5. 3 keys Lessons learnt in the assignment.

Note: The Assignment submission form should be submitted as well, as a separate copy. Your submission will not be accepted or penalized without the Assignment Submission form being submitted.

Instructions for the assignment:

The objective of the assignment is to design and build a data lake and relevant pipelines to enable downstream applications to find key insights.

1) Dataset:

1. The teams have freedom to choose problems and datasets as per their familiarity with specific domain.
2. Choose a dataset with reasonable size with at least two data files/datasets.
3. One of the datasets should have at least 6 columns and 1 MB in size.
4. Define the schema and load the data as per schema.
5. Ensure handling of bad record, if any, should be designed in the pipeline.

2) Business Insights:

1. Define at least 5 key business insights that will be derived from the data.

3) Data Lake Creation:

1. Copy/Ingest the raw files into the landing zone.
2. Apply required schema and store the dataset in the staging zone. Decide on the data formats and partitions required based on the analysis required to create business insights.
3. Apply required data transformation before ingesting data into the staging zone.
4. Create curated (aggregated) tables required to create specified "Insights".
5. Precompute the aggregated statistics or measures as required. Mention the design decisions.

4) Pipeline Creation:

1. Design and develop appropriate pipelines created to transform data for both the zones as required.
2. Precompute the aggregates using either cube, rollups or grouping sets.
3. Choose one or two business insights (out of 5 insights) that will involve the below complex operations (minimum two of the below operations):
    - Partition over ranking function
    - Moving average
    - Lead or lag
4. The pipelines can be created either using Spark DataFrame APIs or Spark SQLs depending on your comfort and skill level.

5) Dashboards/Charts

1. Create one dashboard for each insight mentioned in the "Business Insights" section.
2. Create charts whenever necessary using Databricks features or Seaborn or Matplotlib APIs. Mention the specific insights being created and write those queries (in SQL) for each specific insight.

6) Development Requirements

The solution should be developed:

1. On Databricks community edition and use spark DataFrame and/or Spark SQL as required.
2. Code should follow PEP8 coding guidelines for code formatting.

3. The code should be segregated into separate sections with proper headings. Proper sections should be created for clarity (like heading 2/3).
4. The project name, team members (with IDs) and description at the top.
5. Clear description of the problem and dataset. Each section should be explained clearly (objective and approach). Each design decision should be documented in the appropriate sections.
6. There should be a conclusion section with a summary of accomplishments and a few bulleted points of lessons learnt in developing the project.
7. *Create two notebooks:*
    a. One for ingesting data from landing zone into staging zone. This notebook should have the dataset description.
    b. The second one is to create aggregates from staging zone into curated zone. This notebook should have a description of the business insights.

General Instructions:

1. This is a group assignment. Max. 5 members per project.
2. Do NOT submit .zip files, otherwise, the submission will not be considered.
3. Please include your team members' names and PGIDs in the submission. Marks will not be awarded to the team member/s if the name(s) are missing.
4. One person should submit on behalf of all the team members.
5. Any late submission will attract a penalty as mentioned in the course outline.
6. The honour code for this submission is **2N-c.**
7. **Please look through the honor code restrictions carefully before attempting the assignment as there will be strong consequences for breaking them.**
8. **Please adhere to the given instructions, otherwise, your submission will not be accepted, or a severe penalty will be applied.**

Deadline: 11th March 2023, 11:55 PM.