Assignment: BDCC (Group Assignment)

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

Deliverables:

The deliverables should be the notebook with all outputs from the complete run of the jupyter notebook (.ipynb file) and the datasets (in a zipped file).

General Instructions:

1. This is a group assignment.
2. Any late submission will attract a penalty as mentioned in the course outline.
3. The Honour code for this submission is 2N-C

Dataset Link :

https://www.kaggle.com/datasets/datasnaek/youtube-new

1.Dataset Description

YouTube (the world-famous video sharing website) maintains a list of the top trending videos on the platform. According to Variety magazine, "To determine the year's top-trending videos, YouTube uses a combination of factors including measuring users interactions (number of views, shares, comments and likes). Note that they're not the most-viewed videos overall for the calendar year". Top performers on the YouTube trending list are music videos (such as the famously virile "Gangnam Style"), celebrity and/or reality TV performances, and the random dude-with-a-camera viral videos that YouTube is well-known for.

This dataset is a daily record of the top trending YouTube videos.

This dataset includes several months (and counting) of data on daily trending YouTube videos. Data is included for the US, GB, DE, CA, FR, RU, MX, KR, JP and IN regions (USA, Great Britain, Germany, Canada, and France, Russia, Mexico, South Korea, Japan and India respectively), with up to 200 listed trending videos per day.

Each region's data is in a separate file. Data includes the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count.

The data also includes a category_id field, which varies between regions. To retrieve the categories for a specific video, find it in the associated JSON. One such file is included for each of the five regions in the dataset.

2.Problem Description

The analysis for this assignment will be limited to the IN (India) region only.

2.1.Business Insights

Design and prepare the data pipelines such that the business analysts should be able to create the following insights.

1. Top 3 channels (titles) by views for each category for a specific month and year.
2. Top 10 titles by views of each month-Year.
3. Top 3 titles by views for each category for a specific month and year.
4. Most liked tags for a specific month in a year.
5. Average views, likes and dislikes for each category for a specific month and year.
6. Is there a correlation between view, likes, dislikes for videos for a specific category?

2.2.Data Lake Creation

The zones to be created in the

Landing Zone

1. Create a directory in Databricks file system
2. Copy/Ingest the category (JSON) and video statistics (CSV) file into the landing zone.

Staging Zone

1. Join the datasets to create one dataset.
2. Select columns that will be required for the analysis.
3. Apply schema and store the dataset. Decide on the data formats and partitions required based on the analysis required to create business insights.

4. Apply required data transformation like timestamp (or any other)

Curated Zone

1. The data need to be curated (aggregated) in such a way that running the queries specified in the "Insights" section are optimized.

2. Precompute the aggregations required. Mention the design decisions.

2.3.Pipeline Creation

Design and develop appropriate pipelines created to transform data as required.

2.4.Dashboards/Charts

Create one dashboard for each insight mentioned in the “Business Insights” section. Create charts whenever necessary. Mention the specific insight being created. And write that queries (in SQL) for each specific insight.

For example: Top 3 channels (titles) for each category for Nov '2017. (The table should show the top 3 channels by all categories for that month). It should show the actual category name (not the ID).

3.Development Requirements

The solution should be developed:

1. On Databricks community edition and use spark
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 document in the appropriate sections.
6. There should be conclusion section with summary of accomplishment and few bulleted points of lessons learnt in developing the project.

4.Assessment Weightage:

80% weightage will be given for accomplishing the above tasks. Another 20% will be given for the following.

1. Code clarity
2. Documentation (use markdown)

There may be penalty if code clarity and documentation is not proper.

Coding scheme for group project is 2N-c.

5.Deliverables:

The solution notebook can be exported in dbc or .ipynb format and submitted.

Deadline : 7th August 2022