Aidetic Data Engineer Assignment

Assignment Description

In this assignment, you will demonstrate your ability to work with **Pyspark**.

Assignment Instructions:

- Set up your development environment with the necessary libraries.
- Access the provided dataset from **here** and save it into your local directory.
- Tasks:

  1.Load the dataset into a PySpark DataFrame.

  2.Convert the Date and Time columns into a timestamp column named Timestamp.

  3.Filter the dataset to include only earthquakes with a magnitude greater than 5.0.

  4.Calculate the average depth and magnitude of earthquakes for each earthquake type.

  5.Implement a UDF to categorize the earthquakes into levels (e.g., Low, Moderate, High) based on their magnitudes.

  6.Calculate the distance of each earthquake from a reference location (e.g., (0, 0)).

  7.Visualize the geographical distribution of earthquakes on a world map using appropriate libraries (e.g., Basemap or Folium).

  8.Please include the final csv in the repository.

Dataset Definition:

Use the earthquake dataset with the following columns:

- Date (string): Date of the earthquake.
- Time (string): Time of the earthquake.
- Latitude (float): Latitude of the earthquake location.
- Longitude (float): Longitude of the earthquake location.
- Type (string): Type of earthquake.
- Depth (float): Depth of the earthquake.
- Magnitude (float): Magnitude of the earthquake.

Aidetic Data Engineer Assignment

Submission:

- Organize your code into a well-structured PySpark application.
- Include comments to explain your code and any assumptions made.
- Provide a README file that explains how to run your PySpark application and any additional notes.

Additional Information:

- Use PySpark version 3.0 or later.
- Assume the CSV file is well-formatted with a header row.
- You can use any additional PySpark functions that you find suitable for the tasks.
- Make sure to handle any potential errors or edge cases gracefully.

Evaluation Criteria:

- Correctness of the code.
- Efficient use of PySpark functions.
- Clarity and organization of the code.
- Proper handling of transformations and calculations.
- This assignment is designed to test the candidate's proficiency in working with PySpark and their ability to perform data engineering tasks. It covers reading data, transforming it, and writing the results in a different format, which are common tasks for a Data Engineer working with distributed data processing frameworks.

Deliverables:

- Push your script to Github and make it publicly accessible and share the corresponding link **here**.
- Please record a Loom video (sharing your screen) and explain the overall workflow.
- Simply upload the file(s) to your personal github repo. Make it public and share it with us.