Assignment: Implement an Airflow DAG to Load Daily JSON Data from GCP Bucket to Hive on Dataproc

Objective: Create an Apache Airflow DAG that fetches a daily JSON file from a GCP bucket and appends the data into a Hive table on GCP Dataproc.

1. Airflow Setup:

   - Ensure that the necessary Airflow providers for Google Cloud Platform and Hive are installed.

2. Airflow Connections:

   - GCP Connection: Set up a connection in the Airflow UI for accessing the GCP bucket.

   - Hive Connection: Set up a connection in Airflow UI to interface with Hive on GCP Dataproc.

3. DAG Implementation:

   - DAG Initialization: Define a DAG with a daily schedule_interval.

   - Operators:
     - Use GCSToLocalFilesystemOperator to download the daily JSON file from the GCP bucket to the local Airflow directory.
     - Use HiveOperator to load the downloaded JSON data into the Hive table on Dataproc.

- Task Dependencies: Ensure the correct execution order for the tasks.

4. Testing:

- DAG Execution: Trigger your DAG manually and verify its successful execution.

- Data Validation: Query the Hive table on Dataproc to ensure that the data from the JSON file is correctly appended.

5. Submission:

- Provide the DAG Python script.
- Write a brief report on your implementation steps and any challenges faced.