MLOPS

Assignment 2

Harris Aamir

20i-0943

SE

This Python script automates a data pipeline using Airflow.

- It creates an Airflow DAG named "my-dag" to manage the workflow.
- It imports necessary libraries for interacting with Airflow, making web requests (requests), parsing HTML (BeautifulSoup), data manipulation (pandas), and version control (DVC and Git).
- The script defines three functions:
    - `extract_data`: retrieves article data from specified URLs, parses the HTML content, and saves it to a CSV file.
    - `transform_data`: cleans and transforms the extracted data using pandas and saves the cleaned data to another CSV file.
    - `load`: uses DVC and Git commands to add, commit, and push the cleaned data file to a version control system.
- It sets up a list of URLs to scrape and defines file paths for input and output data.
- Within the Airflow DAG, it creates tasks using PythonOperator and BashOperator to perform the following actions:
    - Extract data from websites using the `extract_data` function.
    - Transform the extracted data using the `transform_data` function.
    - Load the transformed data and push it to version control using the `load` function.
- Tasks are linked sequentially using dependencies, ensuring each step completes before the next starts.
- The DAG is scheduled to run daily, starting from the current time, but it will only process data for future dates (catchup is disabled).

This script essentially automates the process of collecting data from websites, cleaning and transforming it, and storing it in a version-controlled location using Airflow.