Name: Muhammad Ahmed

Roll No.: 20I-0498

Department: Computer Science

Batch: 2020

Section: DS-N

Course Title: MLOPs

Assignment Number: 2

Topic: Apache Airflow

Submitted To: Sir Pir Sami Ullah

Date of Submission: 5/12/2024

Implementation:

1. This Python script sets up an Airflow DAG named "my-dag" to automate a data pipeline.

2. It imports necessary modules including `DAG` from Airflow, operators for Python and Bash tasks, and libraries like `requests` for making HTTP requests and `BeautifulSoup` for web scraping.

3. The `extract_data` function extracts article data from specified URLs, parses HTML content using BeautifulSoup, and appends the data to a CSV file.

4. The `transform_data` function reads the extracted data from the CSV file, performs cleaning and transformation using pandas, and saves the cleaned data to another CSV file.

5. The `load` function adds, commits, and pushes the cleaned dataset file using DVC and Git commands.

6. It defines a list of URLs to scrape data from and specifies file paths for input and output CSV files.

7. Within the Airflow DAG, it sets up tasks using PythonOperator and BashOperator to execute the data extraction, transformation, loading, and Git push steps sequentially.

8. Tasks are interconnected using the `>>` operator to define the task dependencies.

9. The DAG is scheduled to run daily, starting from the current datetime, with catchup disabled to only execute tasks for future dates.

10. Overall, this script orchestrates a data pipeline using Airflow, automating the process of extracting, transforming, and loading data from web sources into a Git-tracked dataset.

Result: