**Assignment no # 2**

**Course:** Machine Learning Operations (MLOPS)

**Assignment:** MLOps Implementation with Apache Airflow

**Department:** Software Engineering

**Section**: R

**Submitted by**: Sehrish khan , i200817

**Submitted on**: 5/12/2024

**Submitted to:** Sir Pir Sami Ullah Shah

**Workflow OverView:**

The workflow for this assignment consisted of three major steps , Data Extraction , Data Transformation and Data Storage and Version Controlling. Data extraction was carried out to extract relevant data from the bbc and dawn websites. After the data was collected , information was extracted through it by transforming the data to some useful format. After all these steps , DVC (Data Versioning Control) to store the data on a local storage and keep its version updated. Below each step is describe d in detail. Later on , an Airflow DAG script was written in the dags folder to automate the process of extraction , transformation and storage.

**1. Data Extraction:**

* Utilized dawn.com and BBC.com as data sources. This was followed by extracting links from the landing pages. Also , extracted titles and descriptions from articles displayed on their homepages.

**2. Data Transformation:**

* Preprocessed the extracted text data by removing HTML tags, special characters, and converting text to lowercase for further analysis.

**3. Data Storage and Version Control:**

* Stored the processed data on Google Drive for easy access and sharing and implemented Data Version Control (DVC) to track versions of the data . The metadata was also versioned against each DVC push to the GitHub repository for traceability and reproducibility.

**4. Apache Airflow DAG Development:**

* Developed an Airflow DAG to automate the processes of extraction, transformation, and storage. It was ensured that the DAG handles task dependencies and error management effectively for a robust workflow.

**Challenges Encountered:**

Several Challenges were posed while executing the workflow in terms handling dynamic web content, managing data versioning and while installing Apache Airflow.

**1. Handling Dynamic Web Content:**

* Extracting links and article information from dynamic web content posed a challenge due to variations in HTML structure. I tried to address this challenge by utilizing robust parsing libraries like BeautifulSoup and adjusting extraction logic to handle different scenarios.

**2. Managing Data Versioning:**

* Implementing DVC for version control required careful consideration of data file structures and metadata tracking. I tried to ensure proper configuration to accurately record changes and maintain a clear history of data versions. Also it was mandatory to ensure all necessary files existed in the folder such as requirement.txt , Pipfile.lock and dvc.yaml. Furthermore it created confusion sometimes regarding which dvc command to run first followed by another command. This issue was resolved by taking help from the internet and notes.

**3. DAG Development and Error Handling:**

* Before the DAG script , installing Apache Airflow required careful implementation of the commands on the ubuntu terminal. If a single command was incorrect , several steps were re-executed uptill the latest command which was very time consuming. However careful execution , reduced the chances of error in typing. Developing an Airflow DAG that efficiently handles task dependencies and their automatic execution was carefully handled. Error management too required careful planning.

The implementation of Apache Airflow for MLOps automation, coupled with effective data extraction, transformation, storage, and version control, enhanced the efficiency and reproducibility of the data pipeline. Despite encountering challenges, thorough planning and utilization of appropriate tools enabled the execution of the workflow.