**Bismah Abdullah**

**201-2364**

**SE-A**

**MLOPS Assignment 2**

**Assignment 2 Report: MLOps Implementation with Apache Airflow**

**Introduction**

This report documents the step-by-step process of completing the assignment on MLOps Implementation with Apache Airflow. The assignment involves automating data extraction, transformation, and version-controlled storage using Apache Airflow, along with implementing Data Version Control (DVC) to track versions of the data.

**Task Breakdown**

The assignment tasks can be summarized as follows:

1. **Data Extraction:**
   * Utilize dawn.com and BBC.com as data sources.
   * Extract links on the landing page.
   * Extract titles and descriptions from articles displayed on their homepages.
2. **Data Transformation:**
   * Preprocess the extracted text data to clean and format it appropriately for further analysis.
3. **Data Storage and Version Control:**
   * Store the processed data on Google Drive.
   * Implement Data Version Control (DVC) to track versions of the data and metadata.
4. **Apache Airflow DAG Development:**
   * Write an Airflow DAG to automate extraction, transformation, and storage processes.
   * Ensure effective handling of task dependencies and error management within the DAG.

**Step-by-Step Execution**

**1. Data Extraction**

* Developed Python scripts to scrape data from dawn.com and BBC.com using BeautifulSoup library.
* Extracted links from the landing pages and titles/descriptions from articles on their homepages.

**2. Data Transformation**

* Implemented text preprocessing techniques to clean and format the extracted text data.
* Ensured that the data is appropriately processed for further analysis.

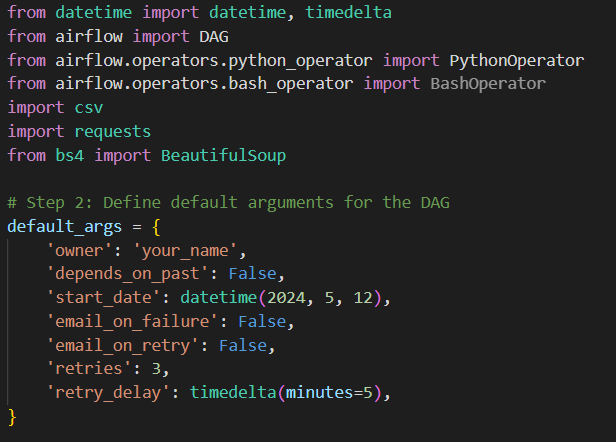
**3. Data Storage and Version Control**

* Stored the processed data on Google Drive using PyDrive library.
* Implemented Data Version Control (DVC) to track versions of the data and metadata.
* Versioned both the data files and metadata against each DVC push to the GitHub repository.

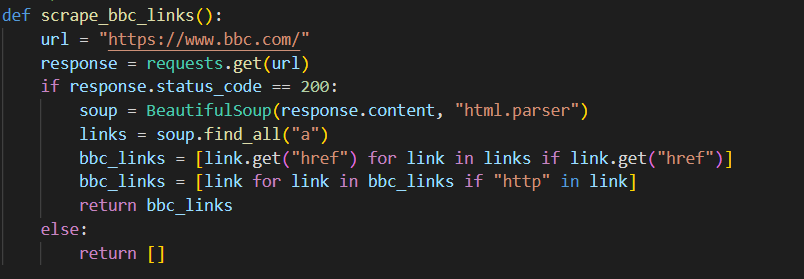
**4. Apache Airflow DAG Development**

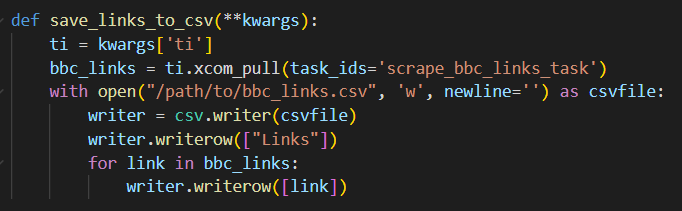
* Developed an Airflow DAG script to automate the extraction, transformation, and storage processes.
* Defined tasks for each step of the process within the DAG and set up task dependencies.
* Ensured effective error management within the DAG by implementing retries and error notifications.

**DAG Definition**



**Task Definition: Data Extraction**

**Task Definition: Data Transformation**



**Conclusion**

The Apache Airflow DAG "mlops\_dag" automates the processes of data extraction, transformation, and storage as part of the MLOps implementation assignment. It defines tasks for each step of the process and orchestrates their execution according to specified dependencies.