MLOps Assignment #2 Report

Authored by Muhammad Huzaifa, DS-N

1. Introduction:

This technical report presents an overview of an Airflow Data Scraping Pipeline designed to extract data from specified URLs, preprocess the extracted data, save it to a CSV file, and then push it to a version control system using DVC (Data Version Control).

2. Pipeline Overview:

The pipeline consists of the following main components:

Extract Data: Fetches HTML content from specified URLs, parses them using BeautifulSoup to extract article data and links.

Preprocess Data: Cleans the extracted data by removing HTML tags, non-alphabetic characters, and converting text to lowercase.

Save to CSV: Saves the preprocessed data to a CSV file.

DVC Push: Uses DVC to track changes in the CSV file and pushes them to a remote Git repository.

3. Code Walkthrough:

Imports and Setup: The code starts with necessary imports such as `DAG` and `PythonOperator` from Airflow, along with imports for web scraping (`requests` and `BeautifulSoup`), data manipulation (`csv`), regular expressions (`re`), and system operations (`os`).

Data Extraction (`extract_data`): The `extract_data` function takes a list of URLs as input, iterates over each URL, fetches HTML content, extracts article data and links using BeautifulSoup, and returns a list containing dictionaries of article information.

Data Preprocessing (`preprocess` and `clean_data`): The `preprocess` function removes HTML tags and non-alphabetic characters from text and converts it to lowercase. The `clean_data` function applies preprocessing to each article's title and description.

Save to CSV (`save_to_csv`): Saves the preprocessed article data to a CSV file specified by the `file_name`.

DVC Push (`dvc_push`): Automates the process of adding the CSV file to DVC, pushing changes to the remote repository, and committing the changes to Git.

DAG Definition: Defines the Airflow DAG (`mlops-dag`) with default arguments and description.

Airflow Operators: Defines PythonOperator tasks within the DAG, specifying the callable functions (`extract_data`, `clean_data`, `save_to_csv`, `dvc_push`) and their dependencies.

Task Execution Order: Specifies the execution order of tasks using the bitshift (`>>`) operator.