# I200521\_Muhammad Arbaz Ishfaq

# Assignment02

# MLOps Implementation with Apache Airflow

# Introduction

The objective of this project was to implement Apache Airflow to automate the processes of data extraction, transformation, and version-controlled storage. The project involved extracting data from dawn.com and BBC.com, preprocessing the extracted text data, storing the processed data on Google Drive, implementing Data Version Control (DVC) to track versions of the data, and writing a fully functional Airflow DAG script to automate the workflow.

# Workflow

## 1. Data Extraction

- Source: Utilized dawn.com and BBC.com as data sources.

- Extracted Links: Extracted links on the landing pages of both websites.

- Extracted Article Data: Extracted titles and descriptions from articles displayed on the homepages.

## 2. Data Transformation

- Preprocessing: Preprocessed the extracted text data by cleaning and formatting it appropriately for further analysis.

- Cleaning: Removed HTML tags, special characters, and performed tokenization and lemmatization.

## 3. Data Storage and Version Control

- Google Drive: Stored the processed data on Google Drive for easy access and sharing.

- DVC Setup: Implemented DVC to track versions of the data. Each version was accurately recorded as changes were made.

- Metadata Versioning: Versioned metadata alongside each DVC push to the GitHub repository.

## 4. Apache Airflow DAG Development

- DAG Script: Developed a fully functional Airflow DAG script to automate the processes of extraction, transformation, and storage.

- Task Dependencies: Defined task dependencies to ensure the workflow ran smoothly and efficiently.

- Error Management: Implemented error handling to gracefully handle any failures during the execution of tasks.

## Challenges Encountered

- Data Extraction: Parsing HTML content and extracting relevant information posed challenges due to the dynamic nature of web pages.

- Preprocessing: Designing an effective preprocessing pipeline required experimentation and fine-tuning to achieve optimal results.

- DVC Integration: Integrating DVC with Google Drive and GitHub required careful configuration and troubleshooting to ensure seamless version control.

## Conclusion

Overall, the implementation of Apache Airflow for MLOps proved to be a valuable tool for automating data workflows. By leveraging Airflow's scheduling and task orchestration capabilities, along with DVC for version control, the project achieved efficient and reproducible data processing and storage. While some challenges were encountered during the implementation, they were successfully overcome through experimentation and collaboration.