

MLOPS

Assignment2

Submitted to: Sir Pir Samiullah Shah

Roll number: 20i-2654

Section: DS-N

This Airflow DAG automates the extraction of news articles from the Dawn and BBC websites, organizes the data, and sets up a Data Version Control (DVC) environment with Git integration. Below is an overview of the key components and their functionalities:

# Default Arguments

The `default\_args` dictionary sets common parameters for all tasks, such as the owner, start date, retry behavior, and email notifications.

# DAG Definition

The DAG, named `Assignment2`, is configured to run once . It is designed to manage the entire workflow of data extraction and DVC setup.

# Task: Print Current Working Directory

This task prints and returns the current working directory, which is useful for verifying the environment and setting paths for subsequent tasks.

# Task: Extract Data

This task extracts news articles from the Dawn and BBC websites. It creates necessary directories, performs the data extraction, and saves the results in CSV files.

## extract\_dawn:

Fetches and parses articles from the Dawn website, saving details like title, URL, and excerpts to a CSV file.

## extract\_bbc:

Fetches and parses articles from the BBC website, saving details including title, URL, and the first paragraph to a CSV file.

## fetch\_article\_paragraph:

Helps extract the first paragraph of an article from the BBC website.

# Task: Run DVC Commands

This task runs a series of commands to initialize a Git repository and DVC, add the extracted data files to DVC, configure remote storage, and push the data to a remote Git repository.

# Workflow

The tasks are executed in the following sequence:

1. Print Current Working Directory: Ensures the environment is correctly set up.

2. Extract Data: Collects and saves news articles from specified websites.

3. Run DVC Commands: Sets up DVC and Git, and pushes data to remote storage.

This setup facilitates the automated extraction, organization, and version control of news article data using DVC and Git.