MLOps Assignment 2 Report

Name: Muhammad Usman

Roll no: 20i-0416

Section: DS-N

Setup:

Install Apache Airflow:

Steps can be found in My GitHub Repository (20i-0416_Mlops_A2) README.md.

Setup DVC:

```
pip install beautifulsoup4 requests dvc dvc[gdrive] pandas
```

download this file:

https://dvc.org/download/linux-deb/dvc-3.50.1

Run this command:

```
sudo apt install ./dvc_3.50.1_amd64.deb
```

Introduction:

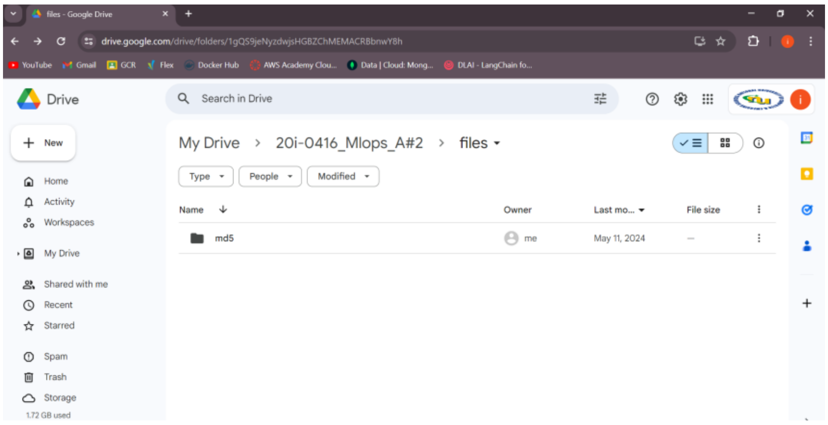
This report outlines my execution of MLOps Assignment 2 that involved data extraction from Dawn and BBC websites, followed by transformation, storage, and the development of an Apache Airflow DAG. Through this assignment, I aimed to gather relevant information, preprocess it, store it securely, and automate the entire process for enhanced efficiency.

Data Extraction: I initiated the assignment by extracting data from Dawn and BBC websites. Leveraging Python libraries like **requests** and **BeautifulSoup**, I fetched HTML content from their landing pages. Subsequently, I extracted links from anchor tags and titles/descriptions from article tags. This extracted data was then saved in a CSV file to facilitate further processing.

Data Transformation: Following data extraction, I proceeded with data transformation to ensure its suitability for analysis. Using the pandas library, I removed duplicate entries from the dataset and effectively handled null values in titles and descriptions. This step aimed to enhance data quality and integrity, laying a solid foundation for subsequent analysis.

Data Storage and Version Control: To ensure secure storage and version control, I opted to store the processed data on Google Drive and implement Data Version Control (DVC). Processed data was securely stored on Google Drive, while DVC was employed to meticulously track versions of the data. Metadata was versioned against each DVC push to the GitHub repository, ensuring comprehensive version control and reproducibility.

Google Drive:



files - Google Drive

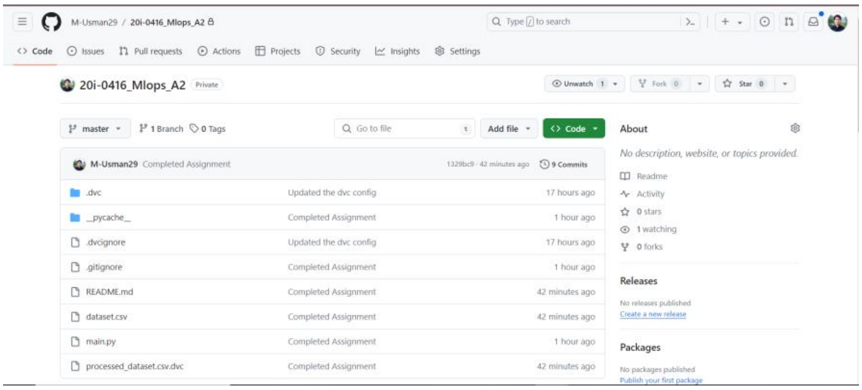
drive.google.com/drive/folders/1gQS9jeNyzdwjsHGBZChMEMACRBbnwY8h

My Drive > 20i-0416_Mlops_A#2 > files

Name	Owner	Last mo...	File size
md5	me	May 11, 2024	—

1.72 GB used

GitHub:



M-Usman29 / 20i-0416_Mlops_A2

20i-0416_Mlops_A2 Private

master 1 Branch 0 Tags

M-Usman29 Completed Assignment 1329bc9 · 42 minutes ago 9 Commits

.dvc	Updated the dvc config	17 hours ago
pycache	Completed Assignment	1 hour ago
.dvcignore	Updated the dvc config	17 hours ago
.gitignore	Completed Assignment	1 hour ago
README.md	Completed Assignment	42 minutes ago
dataset.csv	Completed Assignment	42 minutes ago
main.py	Completed Assignment	1 hour ago
processed_dataset.csv.dvc	Completed Assignment	42 minutes ago

About

No description, website, or topics provided.

Readme
Activity
0 stars
1 watching
0 forks

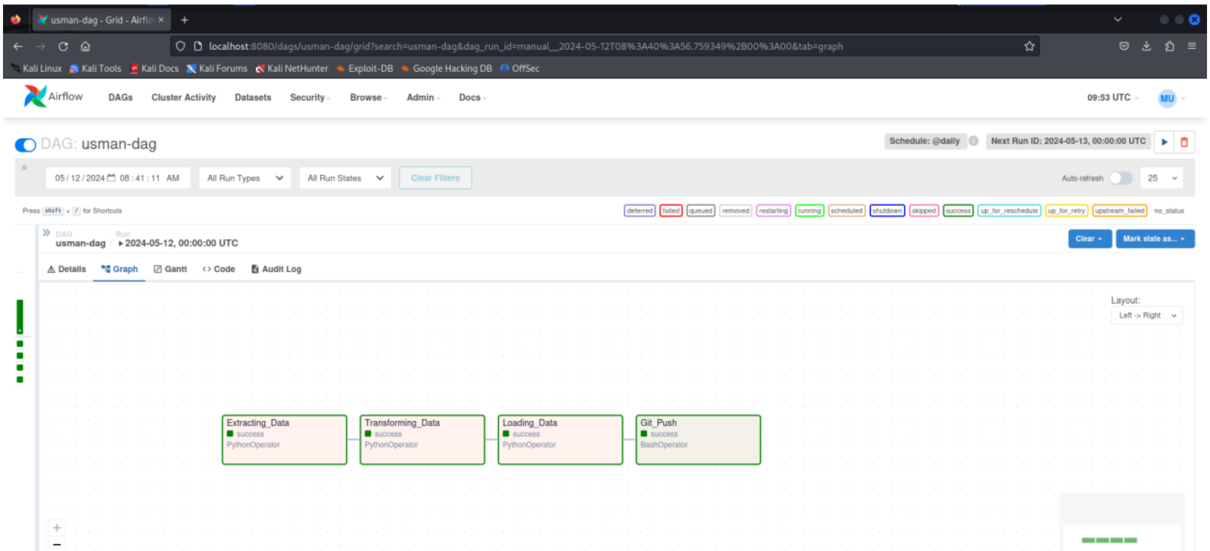
Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Apache Airflow DAG Development: The final phase of the assignment involved the development of an Apache Airflow DAG to automate the entire workflow. I created a DAG named "usman-dag" encompassing tasks for data extraction, transformation, loading, and Git push. Defining task dependencies ensured sequential execution, while implementing error management mechanisms enabled graceful handling of failures. The Airflow DAG streamlined the entire process, enhancing efficiency and reliability.



localhost:8080/dags/usman-dag/grid?search=usman-dag&dag_run_id=manual__2024-05-12T08%3A40%3A56.759349%2B00%3A00&tab=graph

DAG: usman-dag

Schedule: @daily

Next Run ID: 2024-05-13, 00:00:00 UTC

05 / 12 / 2024 08 : 41 : 11 AM

25

DAG usman-dag / Run 2024-05-12, 00:00:00 UTC

Layout: Left -> Right

Conclusion: In conclusion, the successful execution of this assignment demonstrated my ability to efficiently handle data extraction, transformation, storage, and automation using Apache Airflow. Through systematic execution and meticulous attention to detail, I achieved the assignment objectives of gathering, preprocessing, storing, and automating the workflow effectively.