**Muhammad Kazim**

**20I-2310**

**SE**

**MLOPS – Assignment#2**

***Report***

**Workflow:**

1. **Data Extraction:** Articles are extracted from BBC and Dawn websites using web scraping techniques. This involves sending HTTP requests to the websites, parsing the HTML content, and extracting relevant information such as article titles, links, and descriptions.
2. **Data Cleaning:** To ensure data quality, any articles with empty descriptions are removed from the dataset. This step helps in eliminating incomplete or irrelevant data, ensuring that only meaningful information is processed further.
3. **Data Storage:** The cleaned data is stored in a CSV file named **articles.csv**. This file serves as a structured format for storing the extracted article data, facilitating easy access, and analysis.
4. **Version Control:** Both DVC and Git are utilized for versioning the data. DVC manages the data files, allowing for efficient versioning and collaboration, while Git handles the versioning of the entire project, including code, configuration files, and data files.
5. **Automation:** The entire workflow is automated using Airflow DAG. With scheduled execution and proper task dependencies, the data extraction, cleaning, and storage processes are executed seamlessly, ensuring timely updates to the dataset.

**Challenges Encountered:**

* **Error Handling**: Implementing robust error handling mechanisms was crucial, especially during web scraping and HTTP requests. Handling potential network errors, timeouts, and malformed HTML structures required careful consideration to maintain the reliability of the workflow.
* **Dependency Management**: Managing dependencies and ensuring proper task dependencies in the Airflow DAG was essential for orchestrating the workflow effectively. Establishing the correct order of execution and handling inter-task communication ensured the smooth execution of the pipeline.
* **Integration with DVC and Git**: Configuring DVC and Git integration within the Airflow environment posed challenges. Ensuring seamless integration, proper authentication, and handling credentials securely were important considerations to maintain data integrity and security.