Assignment 02 Report

20I0730

BILAL AHMAD

CS-B

Overview:

- **Data Extraction:** Utilized web scraping techniques to extract links from the landing pages of dawn.com and BBC.com using Python libraries such as requests and BeautifulSoup.

  Created a Python script (`data_extraction.py`) to extract links from dawn.com and BBC.com and saved them in CSV files (`dawn_links.csv` and `bbc_links.csv`).

- **Data Transformation:** Processed the extracted links to extract titles and descriptions from the articles displayed on the homepages of both websites. Implemented retry strategies and timeouts to handle potential network issues during data extraction.

  Created a Python script (`data_extraction.py`) to extract links from dawn.com and BBC.com and saved them in CSV files (`dawn_links.csv` and `bbc_links.csv`).

- **Data Storage and Version Control:** Stored the processed data in CSV format and implemented Data Version Control (DVC) to track versions of the data. Integrated Google Drive for data storage and version control, ensuring each version was accurately recorded.

- **Apache Airflow DAG Development:** Developed an Apache Airflow DAG (`dag.py`) to automate the entire data pipeline, including extraction, transformation, and storage tasks. The DAG was designed to handle task dependencies and error management effectively.

  Developed an Apache Airflow DAG (`data_extraction_transformation_dag.py`) that incorporated the data extraction and transformation tasks, ensuring seamless automation of the data pipeline.

Challenges Faced:

- **Airflow Configuration on Windows:** Encountered issues related to Airflow configuration on Windows, particularly with SQLite connection paths. Resolved by updating the Airflow configuration to use absolute paths for SQLite.

- **Data Extraction Handling:** Faced challenges in handling potential network errors during data extraction. Implemented retry strategies and timeouts to improve reliability.

Results and Outcomes:

- Successfully automated the data extraction, transformation, and version-controlled storage processes using Apache Airflow.

- Ensured efficient handling of extracted data and improved data pipeline reliability.