Jawad Ahmed

20I0945

DS-N

Assignment -2

Construct a data pipeline using airflow considering all those steps that we discussed in class.

**Data Pipeline Report**

**Overview:**

This Airflow DAG (**extract\_and\_store\_data**) is designed to scrape articles from specified sources, extract their links, titles, and descriptions, store this data in a CSV file, add it to DVC (Data Version Control), and push the updated CSV file to a GitHub repository. Let's delve into its components:

**1. DAG Structure:**

* **Name:** **extract\_and\_store\_data**
* **Description:** A DAG for extracting links, titles, and descriptions from articles and storing the data.
* **Tags:** Web Scraping

**2. Tasks:**

1. **extract\_links (PythonOperator):**
   * Callable: **extract\_links\_and\_metadata()**
   * Description: Scrapes articles from specified sources, extracts link, title, description, and source.
2. **store\_data\_in\_csv (PythonOperator):**
   * Callable: **store\_data\_in\_csv(\*\*kwargs)**
   * Description: Stores extracted data in a CSV file.
3. **push\_to\_dvc (PythonOperator):**
   * Callable: **push\_to\_dvc()**
   * Description: Adds the CSV file to DVC and pushes it.
4. **push\_to\_github (PythonOperator):**
   * Callable: **git\_push()**
   * Description: Pushes the updated CSV file to a GitHub repository.

**3. Workflow:**

1. The **extract\_links** task extracts article data from specified sources.
2. The **store\_data\_in\_csv** task stores the extracted data in a CSV file.
3. The **push\_to\_dvc** task adds the CSV file to DVC and pushes it.
4. The **push\_to\_github** task pushes the updated CSV file to a GitHub repository.

The provided Airflow DAG, "extract\_and\_store\_data," orchestrates a data pipeline for scraping articles from designated sources, extracting pertinent information, storing it in a CSV file, and then versioning and pushing it to both DVC (Data Version Control) and a GitHub repository. This pipeline streamlines the process of collecting and managing article data, facilitating further analysis or dissemination.