Data Project 1 Brief – ETL Pipeline with Apache Airflow

Overview:

The goal of this project is to create an ETL Pipeline using Apache Airflow.

You will choose a data source, and utilise everything you learned on the course to produce a pipeline to ingest that data into an appropriate data store.

Project Components:

- **Data pipeline** showcasing the automation of an ETL process
- **GitHub repo containing all code and scripts for above, and a README File with**
    - Detailed setup instructions
    - User guide
    - Diagrams showing your architecture and pipelines
    - Contribution guidelines for future developers.
- **Project Management:** Maintain a project board (e.g., Trello, Jira or GitHub Project) to manage tasks

Deliverables:

A 5-minute video which explains what you did and what you learned in this project. The target audience is a non-technical stakeholder.

- Project board – how did you conduct the project?
- Discussion of the components
- Demonstration of the working system showing any outputs

Documentation Package:

- PDF copy of the README
- Link/invitation to the project repo

Tools Required:

- **Workflow**: Apache Airflow
- **Containerisation**: Docker
- **Data Source:** your choice
- **Data Store:** Database System of your choice
- **Documentation**: Markdown for README and other guides.
- **Project Management**: Jira, Trello or GitHub Project for task management and tracking.