SOLO DATA ENGINEERING PROJECT

Week 1: Project Setup and Data Ingestion

Objectives:

- **Set up the development environment.**
- **Establish connections with data sources and Snowflake.**
- **Begin data ingestion using Kafka.**

Deliverables:

1. **Environment Setup:**
    - **Install and configure Spark, Kafka, Airflow, and Snowflake connectors.**
    - **Create a Python virtual environment for the project.**
    - **Document the setup process for reproducibility.**
2. **Data Source Identification:**
    - **Select a comprehensive dataset (e.g., e-commerce transactions, social media data, etc.).**
    - **Document data sources and structures.**
3. **Kafka for Data Ingestion:**
    - **Set up Kafka topics for real-time data streaming.**
    - **Implement producers in Python to send data to Kafka.**
    - **Document Kafka setup and producer configurations.**

Week 2: Data Processing with Spark and Airflow Scheduling

Objectives:

- **Process ingested data using Spark.**
- **Set up Airflow for workflow management.**

Deliverables:

1. **Spark Data Processing:**
    - **Develop Spark jobs in Python to process streamed data.**
    - **Implement transformations and aggregations suitable for the dataset.**
    - **Document the Spark job configurations and logic.**
2. **Airflow Workflow Management:**
    - **Set up Airflow DAGs to schedule and monitor Spark jobs.**

- **Ensure error handling and retry mechanisms in workflows.**
- **Document Airflow setup and DAG configurations.**

Week 3: Data Storage and Visualization, and Final Documentation

Objectives:

- **Store processed data in Snowflake.**
- **Create visualizations and reports.**
- **Finalize project documentation.**

Deliverables:

1. **Snowflake Integration:**
    - **Configure Snowflake as the data warehouse.**
    - **Store the processed data in Snowflake tables.**
    - **Document the Snowflake schema and integration steps.**
2. **Data Visualization and Reporting:**
    - **Develop SQL queries for data analysis.**
    - **Create visualizations/reports based on the processed data.**
    - **Document the reports and their business significance.**
3. **Final Documentation and ETL Pipeline Instructions:**
    - **Compile a comprehensive project report.**
    - **Include instructions for setting up and running the ETL pipeline.**
    - **Detail troubleshooting and maintenance guidelines.**
4. **Project Review and Optimization:**
    - **Review the entire pipeline for efficiency and scalability.**
    - **Make necessary optimizations.**
    - **Document any changes and their impact.**

Additional Notes:

- **Regularly commit code to a version control system (e.g., Git).**
- **Ensure that data privacy and security practices are followed.**
- **Test the system rigorously at each stage.**
- **Consider using containerization (like Docker) for easier deployment and scalability.**

This project offers a hands-on experience with a full-stack data engineering workflow, providing valuable insights into real-world data processing and management scenarios.