FINAL PROJECT

Subject : Data Engineer

Batch : 5

Nama : Muhammad Wildan Hardiyanto

Tema : Data pipeline Russia loses equipment in Ukraine war

* *Student mampu memilih dataset yang berkualitas sesuai dengan sumber referensi yang sudah disediakan.*
* *Student mampu merancang data pipeline dan data modeling dengan tepat.*
* *Student mampu melakukan proses batch atau streaming dengan stack yang sesuai.*
* *Student mampu membuat repository GIT.*
* *Student mampu melakukan dokumentasi project dengan rapi dan bagus.*

=================================================================

**Project Data Engineering: Russia Losses Equipment During Ukranian War**

**1. Pendahuluan**

Proyek ini bertujuan untuk merancang dan mengimplementasikan solusi rekayasa data menggunakan dataset "Russia Losses Equipment" yang diambil sumber nya dari Kaggle.com. Proyek ini mencakup pembuatan pipeline data, pemodelan data, serta proses ETL (Extract, Transform, Load) untuk mengelola data baik dalam mode batch maupun streaming.

### 2. Environment dan tools yang digunakan

* **Bahasa Pemrograman:** Python
* **Ekstraksi dan Transformasi Data:** Pandas, PySpark
* **Penyimpanan Data:** PostgreSQL, Amazon S3 (opsional)
* **Streaming Data:** Apache Kafka, Apache Spark Streaming
* **Orkestrasi:** Apache Airflow
* **Pemodelan Data:** dbt (data build tool)
* **Visualisasi:** Tableau, Power BI, atau alat BI lainnya

### 3. Langkah-langkah Project

#### 1. Set Environment dan Library

Instalasi Library:

1. Pandas

pip install pandas

1. PySpark

pip install pyspark

1. PostgreSQL dan Kafka

Instalasi manual di web resmi PostgreSQL serta Kafka

1. Apache Airflow

pip install apache-airflow

1. Dbt

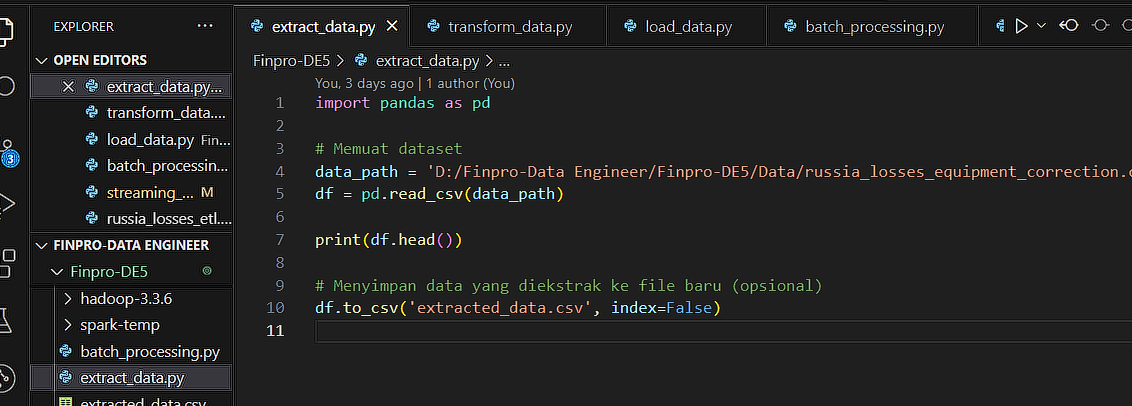
pip install dbt

1. BI tools

Instal Tableau atau Power BI sesuai kebutuhan melalui web resmi Tableau

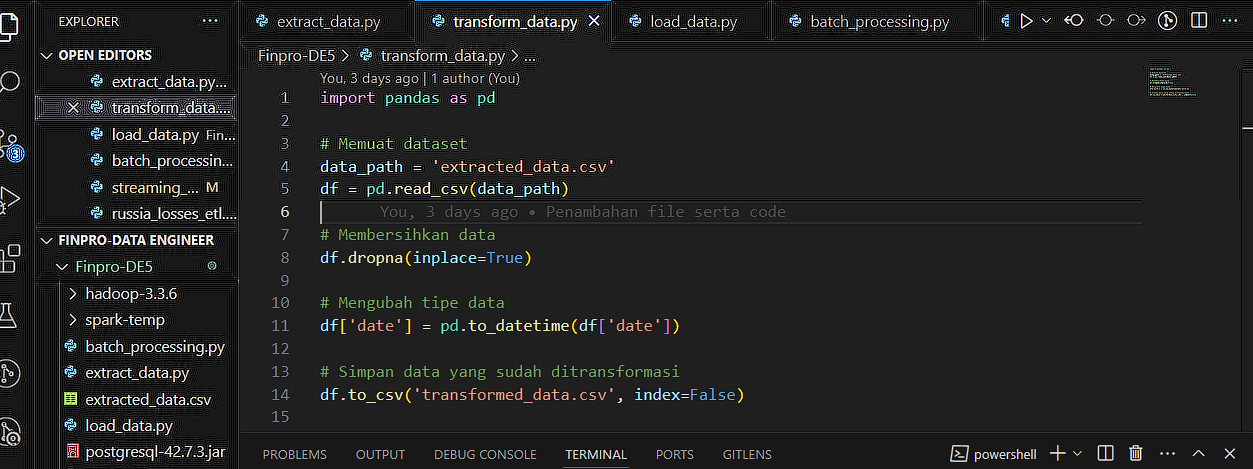
#### 2. Ekstraksi Data

Unduh dataset dari Kaggle dan simpan ke dalam direktori. Berikut adalah contoh skrip untuk memuat data ke dalam DataFrame:



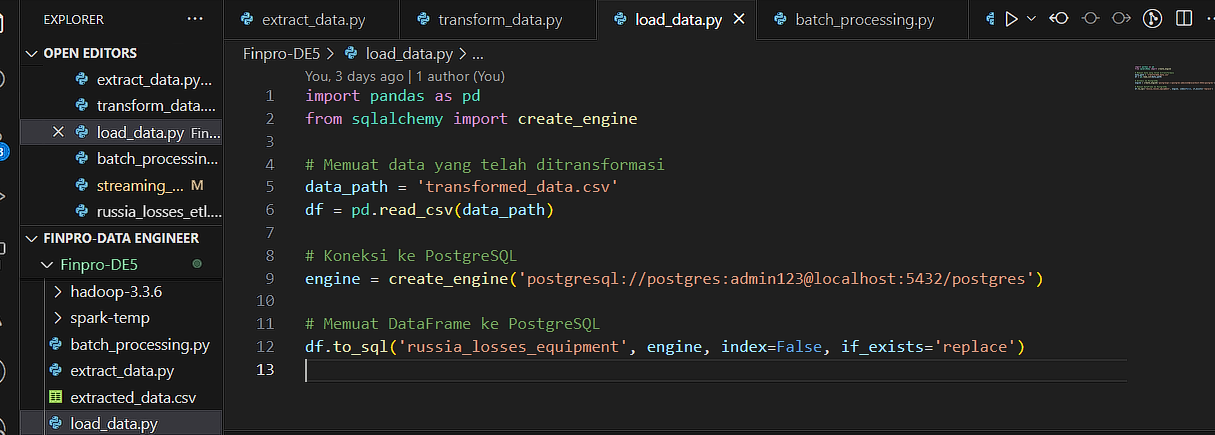
#### 3. Transformasi Data

Bersihkan dan transformasi data menggunakan library Pandas. Misalnya, handle nilai yang hilang dan mengubah tipe data:



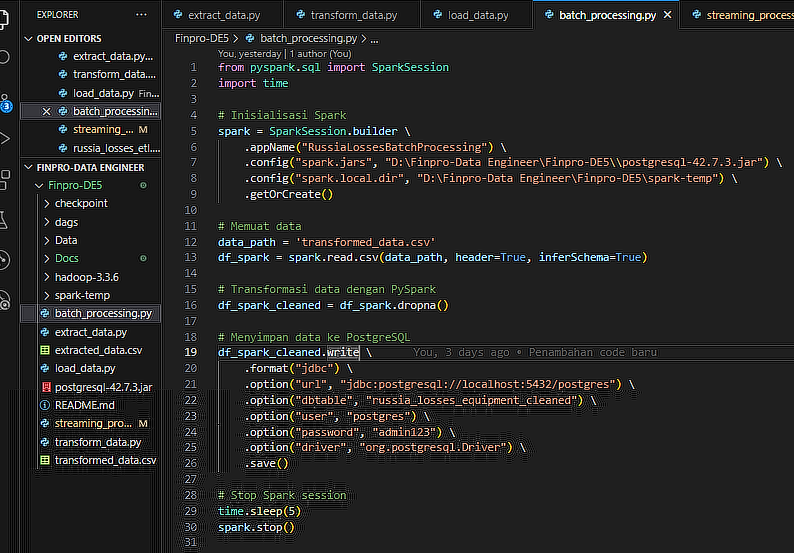
#### 4. Load Data

Desain skema database (skema bintang atau skema salju) dan buat tabel di PostgreSQL. Misalnya, membuat tabel dengan SQLAlchemy:



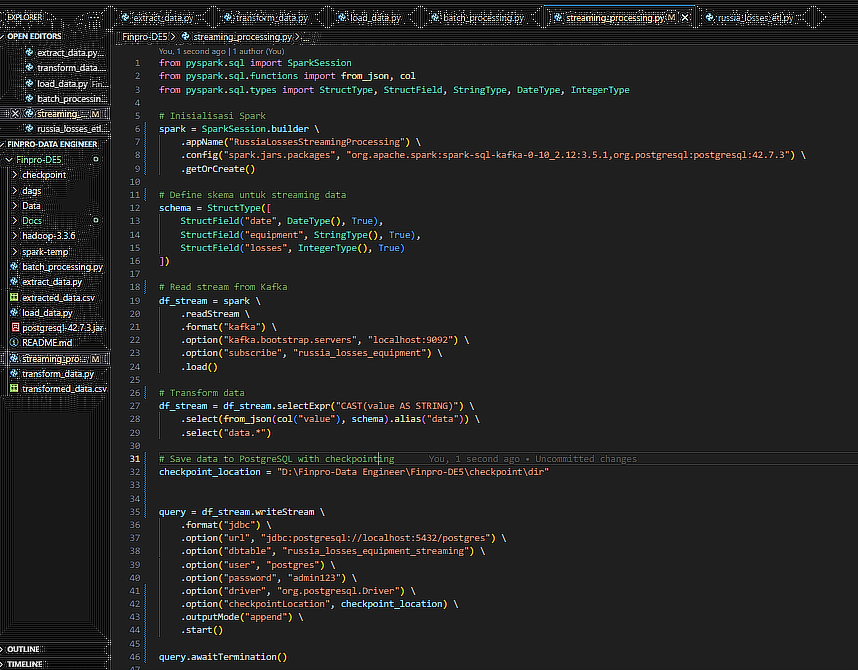
#### 5. Batch processing

Skrip PySpark untuk memproses data secara batch:



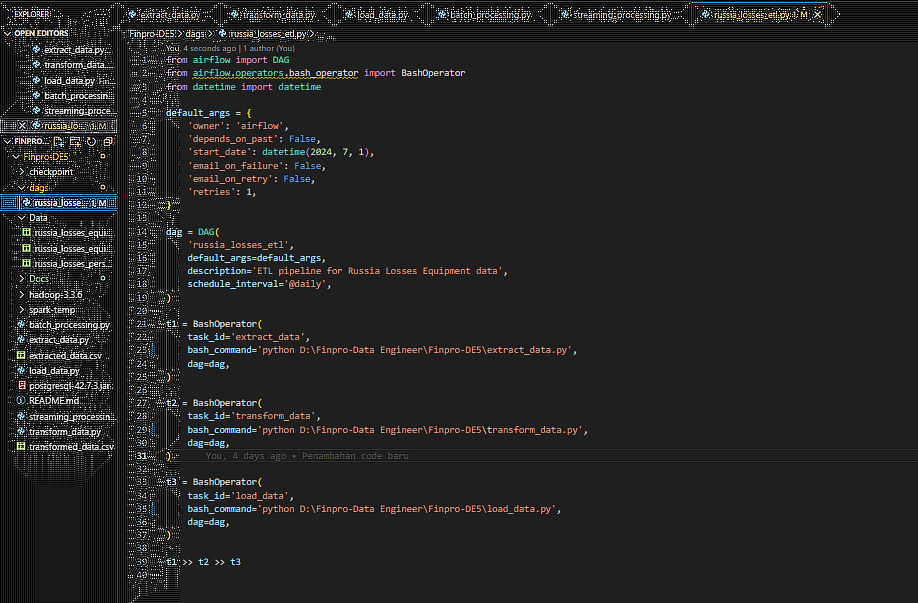
#### Streaming processing

Set Apache Kafka dan skrip untuk Spark Streaming:



#### Orkestrasi ETL

Membuat DAG Airflow untuk mengelola proses ETL:



### 4. Kesimpulan

Proyek ini mencakup seluruh tahapan penting dalam rekayasa data, mulai dari ekstraksi data hingga visualisasi akhir. Setiap tahap diuji dan divalidasi untuk memastikan bahwa data diproses dan dianalisis dengan benar. Dengan panduan ini, Anda dapat menjalankan proyek ini secara end-to-end.

### 5. Visualisasi

### Load Data ke PostgreSQL

### 

### Hasil batch processing

### 