

Dokumen Desain Data Pipeline

overview

Proyek ETL Data Pipeline Optimization ini bertujuan untuk membangun sebuah pipeline data end-to-end yang dapat mengintegrasikan berbagai sumber data tambang, termasuk:

- - Log Produksi Harian (berasal dari file SQL dump),
- - Sensor Alat Berat (dari file CSV),
- - Data Cuaca (menggunakan API dari Open-Meteo).

Pipeline ini akan:

Extract data dari sumber-sumber di atas, Transform data menjadi berbagai metrik analitis penting, seperti:

- - Total produksi harian (total_production_daily)
- - Rata-rata kualitas batubara (average_quality_grade)
- - Utilisasi alat (equipment_utilization)
- - Efisiensi bahan bakar (fuel_efficiency)
- - Dampak cuaca terhadap produksi (weather_impact)

Load hasil akhirnya ke dalam ClickHouse, yaitu database OLAP yang cepat dan cocok untuk analisis skala besar.

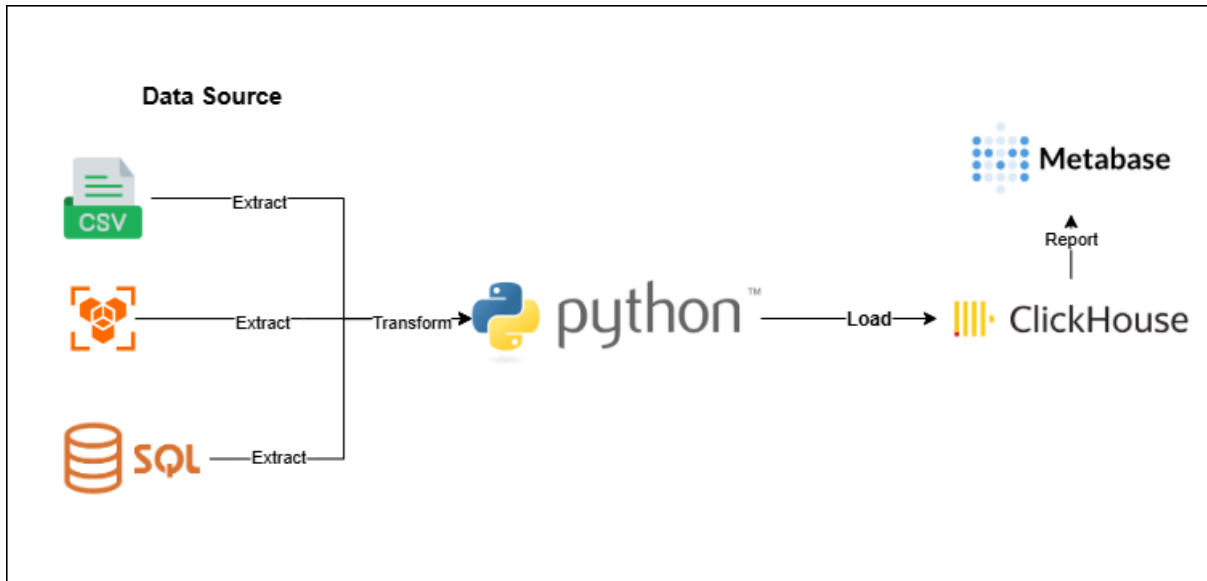
Seluruh proses berjalan otomatis dan dapat dijalankan di lingkungan lokal menggunakan Docker untuk memastikan replikasi yang konsisten.

Pipeline ini juga terhubung ke Metabase untuk menampilkan dashboard visualisasi interaktif, sehingga tim analis atau manajemen bisa mengambil keputusan berdasarkan data yang sudah bersih dan terintegrasi.

Tujuan

Tujuan dari proyek ini adalah membangun sebuah pipeline data ETL yang dapat mengintegrasikan data log produksi tambang, data sensor alat berat, dan data cuaca harian untuk menghasilkan metrik analitis yang berguna bagi pengambilan keputusan operasional dan strategis.

Arsitektur Project



Komponen Utama:

- **Data Source:**
 - File SQL (`production_logs.sql`)
 - File CSV (`equipment_sensors.csv`)
 - Open-Meteo API (JSON)
- **ETL Script:**
 - Dibangun menggunakan Python
 - Menggunakan `pandas`, `aiohttp`, dan `asyncio`
 - Transformasi data dilakukan untuk menghitung metrik
- **Data Warehouse:**
 - **ClickHouse** sebagai storage untuk hasil akhir
- **Dashboard:**
 - **Metabase** sebagai visualisasi data analitis

ETL Pipeline

[SQL + CSV + API] --> [Python ETL Script] --> [ClickHouse] --> [Metabase Dashboard]

Extract:

- Membaca file SQL untuk data produksi (**production_logs**)
- Membaca CSV sensor alat berat (**equipment_sensors.csv**)
- Mengambil data cuaca dari API berdasarkan tanggal produksi

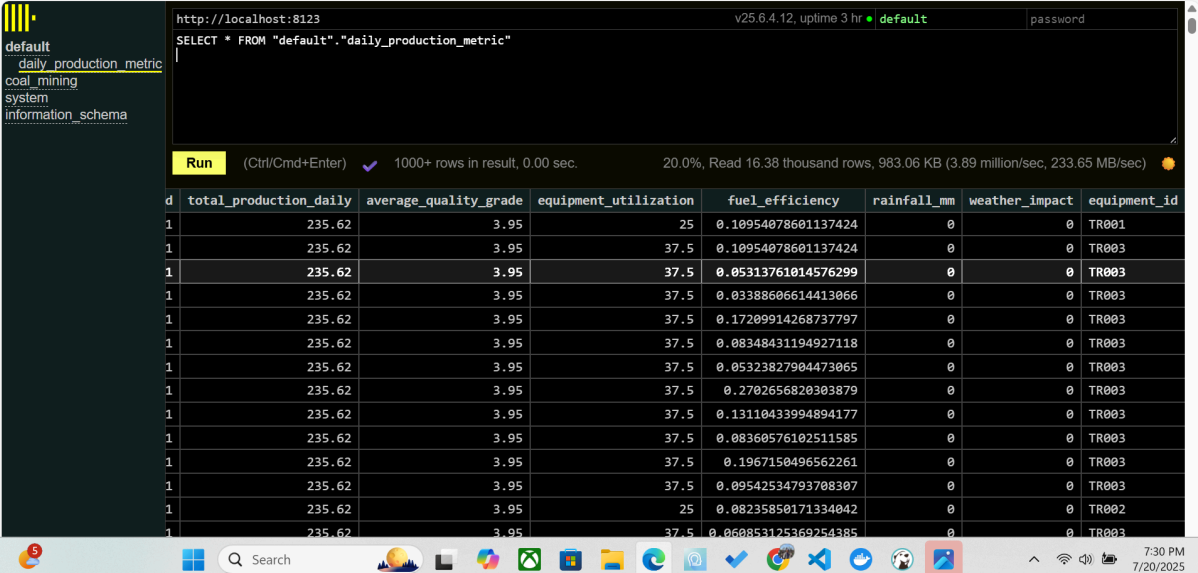
Transform:

- Pembersihan data (mis. **tons_extracted < 0** diubah ke 0)
- Menghitung:
 - **total_production_daily**
 - **average_quality_grade**
 - **equipment_utilization**
 - **fuel_efficiency**
 - **weather_impact**
- Korelasi cuaca dan produksi dihitung sebagai selisih produksi terhadap rolling average 3 hari

Load:

- Menyimpan hasil akhir ke dalam tabel **daily_production_metric** di ClickHouse

Skema Tabel daily_production_metric



The screenshot shows a ClickHouse query interface. The query executed is `SELECT * FROM "default"."daily_production_metric"`. The results show a table with 8 columns: `total_production_daily`, `average_quality_grade`, `equipment_utilization`, `fuel_efficiency`, `rainfall_mm`, `weather_impact`, and `equipment_id`. The data is grouped by `equipment_id`, with each group containing 10 rows of aggregated data.

	total_production_daily	average_quality_grade	equipment_utilization	fuel_efficiency	rainfall_mm	weather_impact	equipment_id
1	235.62	3.95	25	0.10954078601137424	0	0	TR001
1	235.62	3.95	37.5	0.10954078601137424	0	0	TR003
1	235.62	3.95	37.5	0.05313761014576299	0	0	TR003
1	235.62	3.95	37.5	0.03388606614413066	0	0	TR003
1	235.62	3.95	37.5	0.17209914268737797	0	0	TR003
1	235.62	3.95	37.5	0.08348431194927118	0	0	TR003
1	235.62	3.95	37.5	0.05323827904473065	0	0	TR003
1	235.62	3.95	37.5	0.2702656820303879	0	0	TR003
1	235.62	3.95	37.5	0.13110433994894177	0	0	TR003
1	235.62	3.95	37.5	0.08360576102511585	0	0	TR003
1	235.62	3.95	37.5	0.1967150496562261	0	0	TR003
1	235.62	3.95	37.5	0.09542534793708307	0	0	TR003
1	235.62	3.95	25	0.08235850171334042	0	0	TR002
1	235.62	3.95	37.5	0.060853125369254385	0	0	TR003

Skema Tabel daily_production_metric

Kolom	Tipe Data	Deskripsi
date	DateTime	Tanggal produksi
mine_id	Integer	ID tambang
total_production_daily	Float	Total produksi (ton)
average_quality_grade	Float	Rata-rata kualitas batubara
equipment_utilization	Float (%)	Persentase alat aktif
fuel_efficiency	Float	Rasio konsumsi BBM per ton
rainfall_mm	Float	Curah hujan (mm)
weather_impact	Float	Dampak cuaca terhadap deviasi produksi harian

Cara Menggunakan Project

Clone Repository

```
git clone https://github.com/ahmadarbain/data_mining_optimization.git
```

selanjutnya masuk ke direktori project

```
cd data_mining_optimization
```

Instalasi Docker

Jika belum, kamu bisa install Docker Desktop di:

<https://www.docker.com/products/docker-desktop>

Jalankan docker-compose.yaml

Pada tahap ini jalankan perintah berikut untuk menggunakan clickhouse dan metabase pada docker

```
docker-compose up -d
```

Ini akan menjalankan:

- ClickHouse (Database OLAP)
- Metabase (Visualisasi Dashboard)

Siapkan Environment Project (venv)

Gunakan Python 3.9, misalnya dengan pyenv atau venv.

install pypi

```
pip install py
py -3.9 -m venv venv
```

```
# aktifkan env
source venv/bin/activate # Linux/Mac
venv\Scripts\activate    # Windows
```

```
# instalasi requirements di dalam env
pip install -r requirements.txt
```

Running etl proses

Jalankan perintah berikut untuk menjalankan seluruh proses ETL

```
python main.py
```

Struktur Proyek

```
.
├── datasets/                # Folder dataset mentah
├── clickhouse/              # Konfigurasi Docker untuk ClickHouse
├── metabase/                # Konfigurasi Docker untuk Metabase
├── src/
│   ├── usecase/
│   │   ├── daily_production.py # Proses utama ETL
│   │   └── interface/
│   │       └── database.py     # Abstraksi untuk koneksi database
├── requirements.txt
├── docker-compose.yml
└── main.py
```

Report



mine_coal_dashboard



Date
Previous 30 days

SUMMARY MINES DATA PRODUCTION

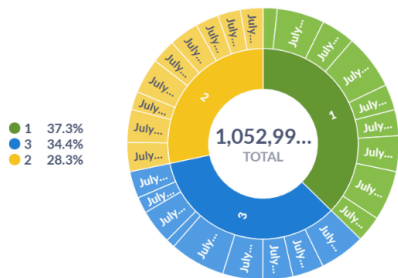
235.62

Daily Production Metric

3.95

Daily Production Metric

Daily Production Metric



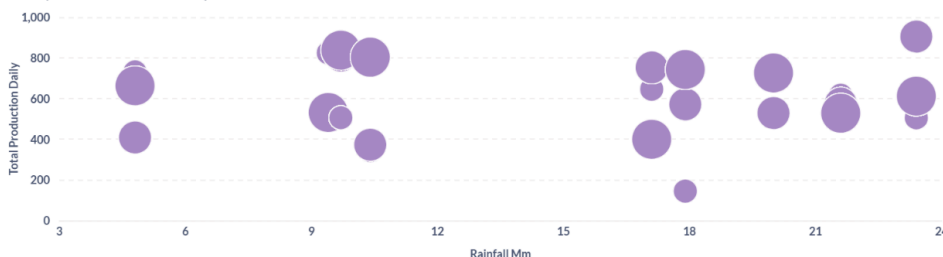
total mine daily production



Bar chart comparison average quality grade across mines



Scatter plot for corr between total production and rainfall



Daily Production Metric

Date	Mine ID	Total Production Daily	Average Quality Grade	Equipment Utilization	Fuel Efficiency	Rainfall Mm	Weather Impact
July 1, 2024	1	235.62	3.95	25	0.11	0	
July 1, 2024	1	235.62	3.95	37.5	0.11	0	
July 1, 2024	1	235.62	3.95	37.5	0.053	0	
July 1, 2024	1	235.62	3.95	37.5	0.034	0	
July 1, 2024	1	235.62	3.95	37.5	0.17	0	
July 1, 2024	1	235.62	3.95	37.5	0.083	0	
July 1, 2024	1	235.62	3.95	37.5	0.053	0	
July 1, 2024	1	235.62	3.95	37.5	0.27	0	
July 1, 2024	1	235.62	3.95	37.5	0.13	0	
July 1, 2024	1	235.62	3.95	37.5	0.084	0	
July 1, 2024	1	235.62	3.95	37.5	0.2	0	

2,000 rows