

Dokumen Strategi Penjadwalan DAG

Deskripsi DAG

DAG ini bertanggung jawab untuk menjalankan pipeline ETL yang mengolah dan memuat data dari pasar saham Brazil ke dalam data warehouse berbasis DuckDB. Pipeline terdiri dari tiga task utama:

1. Inisialisasi penuh (full load),
2. Update data harian untuk koin kripto (kolom `keyCoin` pada tabel `dimCoin`),
3. Update data harian untuk data perusahaan (kolom `keyCompany` pada tabel `dimCompany`).

Karakteristik Data

Dataset dari Kaggle berisi data historis saham yang tersedia dalam file CSV. Meskipun bersifat statis (tidak diupdate *real-time*), skenario yang dibangun mensimulasikan kondisi operasional nyata di mana:

- Data saham umumnya dirilis setelah jam pasar tutup
- Update harian dilakukan dari sumber eksternal
- Data historis dapat diperbarui jika ditemukan koreksi.

Strategi Penjadwalan DAG

1. Initial Load (run_full_pipeline)

- Fungsi: Memuat seluruh data awal ke dalam warehouse
- Frekuensi: Sekali saja saat awal deployment
- Penjadwalan: Tidak perlu dijadwalkan otomatis, cukup dijalankan manual. Initial Load dipisahkan dari update harian menghindari duplikasi atau reprocessing yang tidak perlu.
-

2. Daily Update Tasks (update_coin_data & update_stock_data)

- Fungsi: Menambahkan data baru harian ke dalam warehouse
- Frekuensi: Sekali sehari (simulasi update real-time).
- Waktu ideal: Tengah malam (00:00) waktu server.
- Penjadwalan (schedule_interval): DAG dijadwalkan harian pada pukul 00:00 memungkinkan proses berjalan saat server dalam beban rendah, dan data untuk hari sebelumnya telah tersedia sepenuhnya.