FINAL PROJECT
DATA ENGINEERING

GENERAL INFORMATION

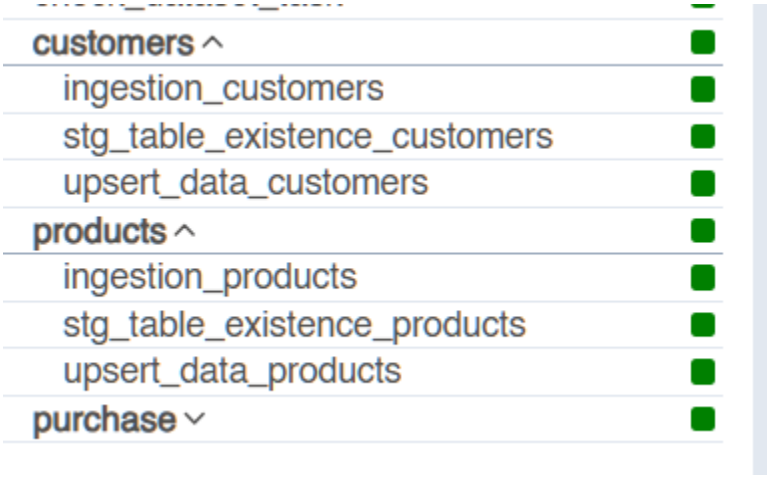
Final Project ini bertujuan untuk mengukur seberapa jauh kemampuan siswa dalam mendesain alur project data pipeline orchestration untuk bidang pekerjaan Data Engineer.

Setiap siswa akan mengerjakan semua case study, dengan project soal seperti berikut :

Mengembangkan capstone project 3 yang sudah dikerjakan kemarin dengan menambahkan:

1. Buatlah Dag Airflow untuk insert data otomatis ke dalam postgres database. dan running per jam. Dengan kriteria sebagai berikut :
    - Buatlah minimal 3 tabel, dan tabel saling keterkaitan. Dimana memiliki primary key dan foreign key, serta memiliki tipe kolom data yang jelas. Misal id (type = INTEGER), nama (type = STRING). Dan seterusnya.
    - Setiap table wajib memiliki kolom created_at.
    - Tema tabel bebas, bisa diambil dari referensi Case Study di bawah.
    - Data bisa diisi dengan random value, asalkan tetap sesuai dengan type dari kolom data tabel.
    - Data di bangun menggunakan Python, dan di schedule setiap 1 jam sekali, kemudian load ke dalam postgre database.
    - Database dibuat menggunakan docker compose, dengan image postgres, dan di koneksikan ke dalam aplikasi DBEaver.
    - Screenshot hasil task airflow dan hasil output dari postgres di DBEaver, minimal 2 hari, dimana perharinya ada minimal 1 task yang berhasil load data ke dalam postgres.

2. Buatlah Dag Airflow otomatis untuk ingest data dari database ke dalam Data Warehouse, secara daily. Dan kriteria pengerjaan sebagai berikut :
    - Airflow di install menggunakan docker-compose.
    - Data sources di ambil dari hasil jawaban no 1, dimana database dijadikan sebagai sumber data yang akan di ingest ke dalam data warehouse.
    - Data pada setiap tabel di filter per hari, dengan kriteria current day - 1 / H-1 ketika schedule dag dijalankan. Data dapat di filter menggunakan kolom created_at.
    - Buatlah 1 operator fungsi yang dapat meng-extract semua tabel dari 1 sumber database, dan dijadikan 1 task di dalam dag. Kemudian dilanjutkan dengan 1 task terpisah untuk proses load data ke dalam tabel BigQuery.
      Contoh referensi : under task customers terdapat sub-task proses extract data, dan load data / insert data kedalam BigQuery.



- Buatlah schema incremental ketika load data ke dalam BigQuery.
- Buatlah schema partition pada setiap tabel ketika load data kedalam BigQuery.
- Buatlah 1 dataset untuk 1 database source di dalam BigQuery. Contoh : ihsan_perpustakaan_capstone3
- Load data table ke dalam BigQuery project = purwadhika.
- Buatlah alert notification yang akan dikirimkan ke channel discord / telegram untuk mengetahui jika dag airflow gagal running / proses retry.

3. Setelah data berhasil ingested ke dalam data warehouse, buatlah data mart dengan mengolah data-data yang sudah tersedia di Data Warehouse. Kriteria pembuatan data warehouse sebaiknya seperti berikut :
   - Buatlah beberapa layer berikut :
     1. Preparation Layer
     2. Dim and Fact table Layer
     3. Datamart Layer
   - Install dan setup Data Build Tools (DBT), dan jalankan menggunakan tools DBT tersebut.

4. Buatlah Scraping salah satu website berikut menggunakan Airflow, dan di load ke dalam data warehouse setiap 1 hari sekali. Bisa pilih dari salah satu website berikut :
   - Adakami : https://www.adakami.id/about > data statistik
   - Adapundi : https://www.adapundi.com/ >
   - Asetku : https://www.asetku.co.id/#! >
   - FIndaya : https://www.findaya.co.id/ >
   - Rupiah Cepat : https://www.rupiahcepat.co.id/about/index.html >

5. Buatlah dokumentasi dimana berisi penjelasan hasil project dan tahapan project anda.
   - Jelaskan hasil dari project yang anda kerjakan.
   - Jelaskan tahapan project yang anda kerjakan, bisa di highlight beberapa poin yang penting.

Setelah mengikut tahapan pengerjaan, setiap tahapan akan memiliki bobot penilaian, diantaranya sebagai berikut :

Poin Penilaian

1. Document Penjelasan dan keseluruhan file project (convert menjadi file .zip dan kirim ke email / berupa link **Github Project**). : 10 points.
    - Google Document, berisi tentang background project (bisa berisi masalah dan tujuan dari project), tahapan project, dan hasil dari pembahasan project. Bisa ditambahkan kendala yang dihadapi, atau masukan dalam proses optimasi project.
    - File project : keseluruhan file projects.
2. Presentasi final project : Maksimal 20 Points.
    - Maksimal waktu presentasi 20 menit.
    - Tanya jawab 40 menit
3. Load data tabel ke dalam postgres database : maksimal 10 point (3 point dari file project, 7 point dari tes tanya jawab / presentasi).
    - Jalankan docker-compose postgres.
    - Generate data dummy, minimum 3 tabel.
4. Ingest data dari database postgres ke dalam BigQuery Data Warehouse : maksimal 20 points. (5 point dari file project, 15 point dari tes tanya jawab / presentasi).
    - Jalankan docker-compose Airflow.
    - Read data dari postgres database.
    - Gunakan konsep functional programming dalam bangunan skrip python.
    - Load data ke dalam BigQuery.
5. Pembuatan datamart dan schema data warehouse : maksimal 20 points (5 point dari dari file project, 15 point dari tes tanya jawab / presentasi)
6. Scraping website menggunakan Airflow : maksimal 10 point (3 point dari dari file project, 7 point dari tes tanya jawab / presentasi).

Case Study/Reference Pembuatan Tabel Dummy

1. Data nilai siswa
2. Penjualan barang toko
3. Gudang (data stok)
4. Rental mobil
5. Perpustakaan (peminjaman buku)
6. Data karyawan perusahaan
7. Data pasien rumah sakit
8. Yellow pages (data kontak telepon)
9. Siswa dapat mengajukan tema case study sendiri

Waktu Pengerjaan

Lama waktu pengerjaan Final Project adalah **10 hari kerja**. Pengerjaan akan terhitung sejak H+1 setelah pengumuman Guideline Project.
Contoh :

1. Pengumuman final project tanggal 29 Januari 2025, Hari Rabu.
2. Pengerjaan mulai dari 30 Januari 2025, Hari Kamis.
3. Pengumpulan maksimal 12 Februari 2025 23:59:59, Hari Rabu.
4. Persiapan presentasi mulai dari 13 Februari 2025 - 19 Februari 2025

Metode Pengumpulan

Pengumpulan dilakukan dengan cara:

1. Unggah video penjelasan project ke dalam cloud storage (Youtube, Google Drive, Dropbox) masing-masing siswa. Buka hak akses untuk publik.
2. Kirim ke email Mentor: samsudiney@gmail.com dan melalui google form yang akan disediakan oleh team purwadhika.
3. Mengisi Google Forms yang telah disediakan oleh operasional untuk mencantumkan link video, link google document / cloud storage.
4. Pastikan siswa menerima email konfirmasi bahwa siswa telah sukses melakukan pengisian dan pengumpulan Google Forms Final Project yang dikirim secara otomatis oleh sistem. Cek folder spam apabila email tidak ada di folder inbox.

Metode Presentasi

1. Setiap siswa akan melakukan presentasi yang akan dijadwalkan oleh team purwadhika
2. Setiap siswa akan melakukan presentasi hasil kerja project dengan waktu maksimal 20 menit. Lakukan presentasi dengan open camera dan jelaskan poin-poin penting dari tahapan project maupun hasil dari project.
3. Setelah presentasi akan ada sesi tanya jawab yang akan dilakukan oleh tim penilai selama 40 menit.

Catatan

1. Jika siswa mengumpulkan Final project melewati tenggat waktu yang sudah ditentukan, maka akan ada pengurangan poin untuk nilai akhir sebagai berikut:
    - Telat 1 detik sampai 24 jam: nilai akhir dikurangi 10 poin
    - Telat 24 jam sampai 72 jam: nilai akhir dikurangi 20 poin
    - Telat lebih dari 72 jam: nilai akhir menjadi 0
2. Segala bentuk plagiarisme tidak akan ditoleransi dan mutlak diberikan nilai 0.
3. Dilarang plagiat dan copy paste secara saklek dari AI.