CAPSTONE PROJECT MODULE 3

GENERAL INFORMATION

Pada Modul 3 siswa telah mempelajari Database dan SQL, beberapa materi yang telah dipelajari di antaranya adalah :

- Docker Basics
- Docker Advanced
- Kubernetes Introduction
- Kubernetes Advanced
- Scala Advanced
- Scala Functional Programming
- ETL and Data Pipeline
- Google Cloud Platform (GCP) Basics
- GCP Services
- Google Cloud Storage
- GCP Databases
- Google BigQuery
- BigQuery Operations
- Advanced BigQuery
- Airflow Basics
- Advanced Airflow
- Data Warehousing and Modeling
- Data Governance
- System Design and SDLC
- Spark Basics
- Spark Structured API
- Spark Data Manipulation
- GCP DataProc Basics
- DataProc Notebook and Pub/Sub
- Cloud Dataflow
- Stream Data Pipelines 1
- Stream Data Pipelines 2

Capstone Project Module 3 ini bertujuan untuk mengukur seberapa jauh kemampuan siswa dalam mendesain alur project data pipeline orchestration untuk bidang pekerjaan Data Engineer.

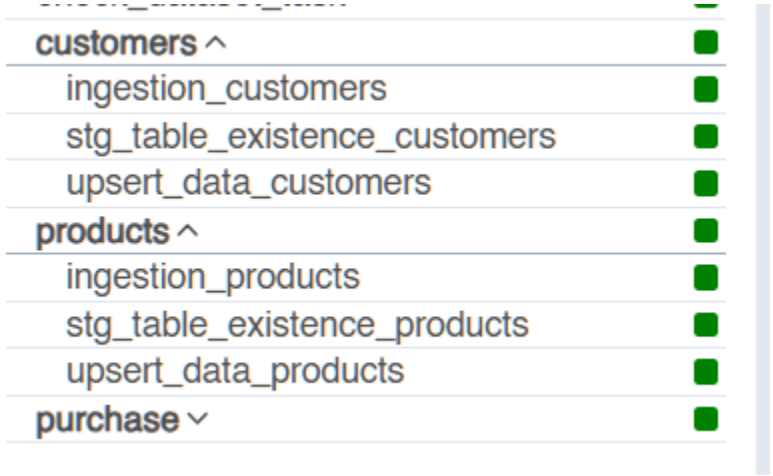
Pada project capstone 3 berikut tidak menganut dari keseluruhan materi yang telah di ajarkan, akan tetapi soal-soal yang di berikan diambil berdasarkan projek inti di bidang Data Engineering.

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

- Membuat 2 dag Airflow, dengan spesifikasi pengerjaan sebagai berikut :

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 di isi 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 di buat 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 table di filter per hari, dengan kriteria current day - 1 / H-1 ketika schedule dag di jalankan. 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 = purwadika / spreadsheet.

3. Buatlah dokumentasi dan video penjelasan dengan durasi 10 - 30 menit, dimana berisi penjelasan hasil project dan tahapan project anda.
    - Jelaskan hasil dari project yang anda kerjakan.
    - Jelaskan tahapan pengerjaan dari project yang anda kerjakan.

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

Poin Penilaian

- Document Penjelasan dan keseluruhan file project (convert menjadi file .zip dan kirim ke email / berupa link **Github Project**). : maksimal 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 : folder airflow dan folder postgres.
- Video Penjelasan : Maksimal 20 Points.
    - Maksimal durasi Video 30 menit.
- Load data tabel kedalam postgres database : maksimal 30 point
    - Jalankan docker-compose postgres.
    - Generate data dummy, minimum 3 tabel.
- Ingest data dari database postgres ke dalam BigQuery Data Warehouse : maksimal 40 points.
    - Jalankan docker-compose Airflow.
    - Read data dari postgres database.
    - Gunakan konsep functional programming dalam bangunan skrip python.
    - Load data ke dalam BigQuery.

Case Study / Reference Pembuatan Tabel Dummy.

- Data nilai siswa
- Penjualan barang toko
- Gudang (data stok)
- Rental mobil
- Perpustakaan (peminjaman buku)
- Data karyawan perusahaan
- Data pasien rumah sakit
- Yellow pages (data kontak telepon)
- Siswa dapat mengajukan tema case study sendiri

Waktu Pengerjaan
Lama waktu pengerjaan Capstone Project Module 3 adalah **10 hari kerja**. Pengerjaan akan terhitung sejak H+1 setelah pengumuman Guideline Project.
Contoh :

- Pengumuman capstone tanggal 09 Desember 2024, Hari Senin.
- Pengerjaan mulai dari 10 Desember 2024, Hari Selasa.
- Pengumpulan maksimal 23 Desember 2024 23:59:59, Hari Senin.

Metode Pengumpulan
Pengumpulan dilakukan dengan cara:

- Unggah video penjelasan project ke dalam cloud storage (Youtube, Google Drive, Dropbox) masing-masing siswa. Buka hak akses untuk publik.

Kirim ke email Mentor : Samsudiney@gmail.com dan melalui google form yang akan di sediakan oleh team purwadika.

- Video penjelasan maksimal berdurasi 30 menit dan wajib mengaktifkan kamera depan atau webcam, sehingga wajah siswa ada dalam rekaman video.

- Mengisi Google Forms yang telah disediakan oleh operasional untuk mencantumkan link video, link google document / cloud storage.

- Pastikan siswa menerima email konfirmasi bahwa siswa telah sukses melakukan pengisian dan pengumpulan Google Forms Capstone Project Module 3 yang dikirim secara otomatis oleh sistem. Cek folder spam apabila e-mail tidak ada di folder inbox.

Catatan

- Jika siswa mengumpulkan Capstone Project Module 3 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

- Segala bentuk plagiarisme tidak akan ditoleransi dan mutlak diberikan nilai 0.
- Di larang plagiat dan copy paste secara saklek dari AI.