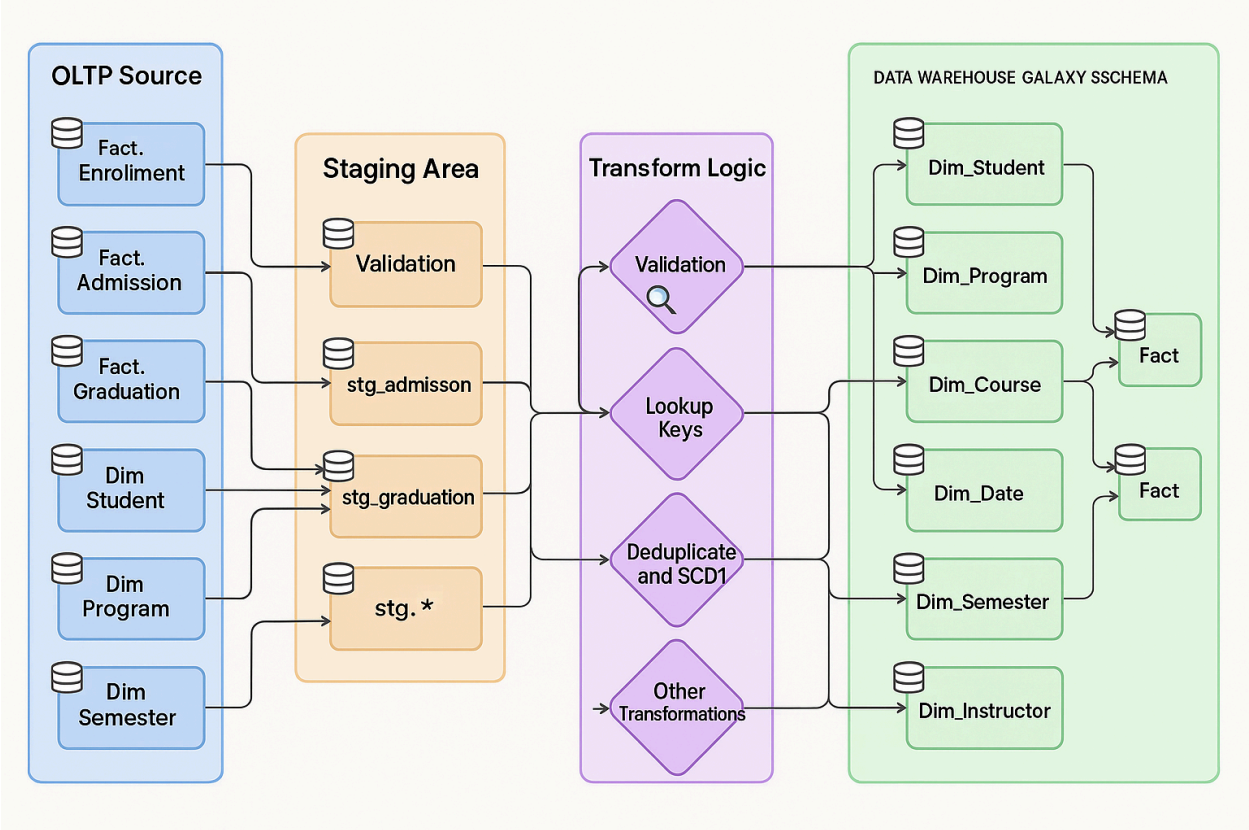
ETL ARCHITECTURE

1. Pendahuluan

Dokumentasi Teknis ETL



1.1 Tujuan dokumen

Dokumen ini dibuat untuk mendeskripsikan arsitektur ETL (Extract, Transform, Load) pada Data Warehouse Akademik Kampus. Dokumen ini memberikan pedoman teknis bagi tim Data Engineer, DBA, dan Developer dalam membangun, mengelola, dan memelihara pipeline ETL.

1.2 Ruang lingkup ETL

ETL mencakup seluruh proses integrasi data akademik dari sistem operasional menjadi Data Warehouse, meliputi data:

- Mahasiswa
- Dosen
- Program Studi

- Mata Kuliah
- Jadwal Kuliah
- KRS (Kartu Rencana Studi)
- KHS (Kartu Hasil Studi)
- Nilai
- Kehadiran
- Registrasi dan Status Akademik

1.3 Sistem sumber dan target

- Source System: SIAKAD (Database Operasional)
- Staging Area: Database *stg*
- Integration Layer: Database *int*
- Data Warehouse: Database *dw* (berisi dim & fact schema)

2. Gambaran Umur Arsitektur ETL

2.1 Diagram Alur ETL

Alur ETL berjalan dalam tahapan berikut:
Source → Staging → Integration → Data Warehouse

2.2 Penjelasan Layer

a. Source Layer

Layer ini merupakan sumber data utama yaitu database operasional SIAKAD yang menyimpan data akademik harian.

b. Staging Layer

Di dalam schema stg, data yang diambil dari SIAKAD disimpan apa adanya (raw data) tanpa transformasi. Layer ini menjadi buffer dan meminimalkan beban pada database operasional.

c. Integration Layer

Pada layer ini dilakukan:

- Data cleansing
- Normalisasi format
- Mapping nilai
- Validasi integritas

- Penggabungan data antar tabel

d. Data Warehouse Layer

Data yang telah bersih akan dimuat ke dalam tabel dimensi dan fakta, antara lain:

- Dimensi: Program, Student, Course, Instructor, Date, Semester
- Fakta: Admission, Graduation, Enrollment

3. ETL Architecture Detail

3.1 Extract Process

- Sumber data (nama database, tabel, koneksi)
- Jenis extract: full extract / incremental extract
- Jadwal extract (harian, mingguan, bulanan)
- Mekanisme logging extract

3.2 Transform Process

Rules Pembersihan Data (Data Cleansing)

- Menghapus duplikasi NIM atau Kode MK
- Validasi format tanggal pada data enrollment, admission, graduation
- Normalisasi gender menjadi ‘M’/’F’
- Validasi foreign key seperti ProgramCode, InstructorCode, CourseCode

Standardization Rules

- Nama → UPPER(TRIM(Name))
- Program, Course Code → kapital seluruhnya
- Format AcademicYear → YYYY/YYYY

Join/Merge Rules

- Student + Program → Dim_Student
- Course + Program → Dim_Course
- Enrollment + Student + Course + Instructor → Fact_Enrollment
- Admission/Graduation + Student → Fact_Admission & Fact_Graduation

Business Logic

- Mapping gender: “Laki-laki” → “M”; “Perempuan” → “F”
- Menghitung SKS total dan biaya jika atribut tersedia
- Menentukan status mahasiswa berdasarkan semester aktif/pelajaran selesai

Error Handling & Data Validation

- Record invalid disimpan di stg.ErrorLog
- Load dihentikan jika error melebihi threshold
- Validasi constraint sebelum load DW

3.3 Load Process

- Load ke Staging
- Load dari Staging ke Integration
- Load ke Dimensi dan Faktual:
    - Mode load: insert/update (SCD Type 1/2 untuk dimension)
    - Dependency antar tabel
    - Mekanisme delete/invalid data

4. Logical Architecture Diagram

- Sumber → Staging → Integration → DW
- Proses ETL Tools (jika ada)
- Storage/DB yang digunakan

5. Physical Architecture

Database Platform

- SQL Server 2019
- Server ETL terpisah dari server operasional

Schema

- dbo (sorce SIAKAD)
- stg (staging area)
- int (integration)
  dw (data warehouse)

Tabel DW

- Dimensi: Dim_Student, Dim_Program, Dim_Course, Dim_Instructor
- Fakta: Fact_Enrollment, Fact_Grades, Fact_TuitionFee

Storage File

Jika bergantung pada dump CSV:

- /extract/mahasiswa_*.csv
- /extract/enrollment_*.csv

6. Job Scheduling & Workflow

Urutan Job ETL

1. Extract mahasiswa
2. Extract program studi
3. Extract dosen
4. Extract mata kuliah
5. Extract KRS/KHS
6. Transform data per entitas
7. Load dimensi
8. Load fakta
9. Logging dan notifikasi

Dependensi

- Student dan Program harus selesai sebelum Enrollment
- Mata Kuliah harus selesai sebelum Fact_Enrollment
- Dosen harus selesai sebelum Fact_Enrollmen**t**

Penjadwalan (cron, SQL Server Agent)

- Menggunakan SQL Server Agent / SSIS Scheduling
- Dijalankan harian/mingguan sesuai kebutuhan operasional

Retry Strategy

- Jika gagal, retry maksimal 3 kali
- Jika masih gagal, kirim notifikasi email admin

7. Error Handling & Logging

Error ditangani melalui mekanisme pencatatan (logging), baik pada level staging maupun integration. Error meliputi:

- Null pada kolom wajib
- Format data tidak sesuai
- Referensi foreign key tidak ditemukan
- Duplikasi yang tidak diizinkan

Setiap error dicatat pada log file atau tabel log, lalu dapat diproses ulang setelah diperbaiki.

8. Data Quality Rules

Data Quality digunakan untuk memastikan data valid sebelum masuk ke DW. Aturan ini mencakup:

- Validasi mandatory: NIM, CourseCode, InstructorCode
- Format field: tanggal valid, gender valid
- Constraint: ProgramCode harus ada di Dim_Program
- Duplicate handling: record duplikat disaring di staging atau integration layer

9. Metadata Management

Metadata digunakan untuk pelacakan dan audit.

- Audit columns: LoadDate, LastUpdate
- Lineage: data berasal dari tabel apa
- Tracking perubahan menggunakan SCD sesuai jenis dimensi

10. Security & Access Control

Keamanan data dikelola dengan membatasi akses pada schema tertentu.

- Schema stg hanya dapat diakses oleh user ETL.
- User BI hanya memiliki hak baca (read-only) pada schema dw.
- Jika ada data sensitif (misalnya tanggal lahir), enkripsi dapat diterapkan.
  Role pengguna didefinisikan sesuai kebutuhan operasional.

11. Deployment & Maintenance

Deployment dilakukan secara bertahap dari development ke production. Setiap perubahan struktur tabel harus diadaptasi mulai dari staging hingga data warehouse.

Prosedur backup dilakukan secara rutin, sedangkan recovery menggunakan mekanisme full backup dan differential backup. Monitoring performa ETL juga dilakukan secara berkala untuk menjaga kualitas dan konsistensi data.