

Data Mart Portal Satu Data ITERA

Misi 1: Desain Konseptual & Logikal

Kelompok 10

Institut Teknologi Sumatera

Daftar Isi

- 1 Pendahuluan
- 2 Analisis Kebutuhan Bisnis
- 3 Sumber Data
- 4 Desain Konseptual (ERD)
- 5 Desain Logikal (Dimensional Model)
- 6 Contoh Query Analitik
- 7 Kesimpulan

Portal Satu Data ITERA

Katalog dataset terpusat yang menyediakan akses ke berbagai dataset institusional (akademik, kepegawaian, keuangan, penelitian)

Tantangan:

- Belum ada sistem analytics
- Monitoring operasional terbatas
- Sulit tracking kualitas data

Kebutuhan:

- Analisis akses dataset
- Monitoring kualitas data
- Perilaku pengguna portal

Fokus Proyek

Merancang Data Mart untuk **analisis operasional Portal Satu Data ITERA**

Tujuan Proyek

- 1 Menganalisis kebutuhan bisnis Portal Satu Data ITERA
- 2 Mengidentifikasi sumber data
- 3 Merancang ERD (Entity Relationship Diagram) untuk entitas portal
- 4 Merancang dimensional model (Star Schema) untuk analytics
- 5 Membuat data dictionary lengkap

Expected Outcomes

Dashboard monitoring operasional portal untuk:

- Tim Pengelola Portal Satu Data
- Data Stewards
- Management ITERA

Stakeholder Primer:

- Tim Pengelola Portal Satu Data
- Data Stewards (per unit organisasi)
- Management ITERA
- Pengguna Eksternal (publik, peneliti)

Stakeholder Sekunder:

- Mahasiswa (pengguna data)
- Dosen & Peneliti
- Bagian Perencanaan & QA
- Media & Publik

① Pengelolaan Dataset

- Upload, validasi, publikasi dataset
- *KPI*: Jumlah dataset baru/bulan, waktu publikasi, dataset aktif

② Akses & Download Dataset

- User view, download, atau akses via API
- *KPI*: Total akses, unique users, download rate, API calls, response time

③ Data Quality Management

- Monitoring kualitas dataset (completeness, accuracy, timeliness)
- *KPI*: Avg quality score, datasets dengan warning, update compliance

④ Search & Discovery

- Pencarian dataset oleh user
- *KPI*: Total searches, top keywords, zero-result rate, CTR

Dataset Management Analytics:

- Dataset mana yang paling sering diakses?
- Kategori apa yang paling populer?
- Waktu respons rata-rata download?

User Behavior Analytics:

- Tipe user mana yang paling aktif?
- Pola akses harian/mingguan?
- User retention rate?

Institution Metrics:

- Unit organisasi mana yang paling produktif?
- Performa publikasi per fakultas?

Data Quality Analytics:

- Dataset mana yang perlu improvement?
- Trend kualitas data per kategori?
- Compliance dengan SLA update?

Identifikasi Sumber Data

Sumber Data	Tipe	Deskripsi
Dataset Catalog DB	PostgreSQL	Metadata dataset (nama, kategori, format)
Access Log DB	PostgreSQL	Log akses user (view, download, API)
Search Query Log	Log Files	Keyword pencarian & hasil
Quality Metrics DB	PostgreSQL	Skor kualitas dataset
User Activity DB	PostgreSQL	Profil & aktivitas user
Dataset Lineage	Metadata	Sumber & riwayat dataset
System Performance	Monitoring	Response time, uptime

Integration Architecture

ETL process: Extract dari 7 sumber → Transform (cleaning, validation) → Load ke Data Mart (Star Schema)

Data Profiling & Karakteristik

Volume Data

- Total dataset aktif
- Total log akses/bulan
- Jumlah pencarian/hari
- Total user terdaftar

Isu Kualitas Data

- Persentase data hilang
- Duplicate records rate
- Dataset tidak diperbarui
- Format data tidak konsisten

Update Frequency

- Access logs: *Real-time*
- Dataset metadata: *On-change*
- Quality assessment: *Monthly*

ETL Strategy

- Incremental load (daily)
- Full refresh (monthly)
- SCD Type 2 untuk Dim_Dataset

10 Entitas Utama Portal

No	Entitas	No	Entitas
1	DATASET <i>Katalog dataset</i>	6	METADATA_KUALITAS <i>Metrik kualitas</i>
2	KATEGORI_DATASET <i>Klasifikasi dataset</i>	7	SUMBER_DATA <i>Asal dataset</i>
3	USER_PORTAL <i>Pengguna portal</i>	8	UNIT_ORGANISASI <i>Struktur ITERA</i>
4	AKSES_DATASET <i>Log akses</i>	9	INSTITUSI_METRICS <i>Agregat metrik</i>
5	PENCARIAN <i>Log pencarian</i>	10	PERIODE_WAKTU <i>Dimensi waktu</i>

Dataset Publication Rules

- Dataset harus melewati validasi kualitas
- Metadata lengkap wajib
- Update sesuai frekuensi

Access Control Rules

- User eksternal: Public only
- User internal: Public + Internal
- Log akses wajib dicatat

Data Quality Rules

- Completeness \geq 80%: Warning
- Missing values \leq 10%: Review
- Assessment: Monthly
- Accuracy \geq 70%: Block publish

Search Rules

- Query disimpan untuk analisis
- Zero-result \rightarrow kandidat dataset baru
- CTR untuk ranking

4 Fact Tables + 6 Dimension Tables

Fact Tables (1-2)

1. Fact_Dataset_Access

- *Grain*: Per access event
- *Measures*: Jumlah akses, download, file size, response time

2. Fact_Dataset_Quality

- *Grain*: Per dataset per assessment date
- *Measures*: Quality scores, missing values, duplicates

Fact Tables (3-4)

3. Fact_Search_Query

- *Grain*: Per search query
- *Measures*: Jumlah pencarian, hasil, CTR

4. Fact_Institution_Metrics

- *Grain*: Per organization per period
- *Measures*: Total datasets, downloads, quality avg

Dimension Tables

Dimension	SCD Type	Key Attributes
Dim_Dataset	Type 2	Nama, kategori, format, tingkat akses, status
Dim_User	Type 1	Username, tipe user, unit organisasi
Dim_Category	-	Nama kategori, parent (hierarki), level
Dim_Organization	-	Nama unit, tipe, parent (hierarki)
Dim_Data_Source	-	Nama sumber, tipe, unit pengelola
Dim_Time	-	Tanggal, tahun, bulan, kuartal, hari kerja

SCD Type 2 Implementation

Dim_Dataset menggunakan SCD Type 2 untuk tracking perubahan metadata (status, tingkat akses, deskripsi) dengan fields: Effective_Date, End_Date, Is_Current

Contoh Query Analitik

Query 1: Top 10 Dataset Paling Populer

- Mengetahui dataset yang paling sering diakses
- Prioritas maintenance dan resource allocation

Query 2: Analisis Kualitas per Kategori

- Monitoring kualitas dataset per kategori
- Identifikasi area yang perlu improvement

Query 3: Zero-Result Search Analysis

- Identifikasi keyword pencarian tanpa hasil
- Kandidat dataset baru yang perlu dibuat

Insight

Keyword dengan banyak zero-result queries menunjukkan demand untuk dataset yang belum tersedia

Technology Stack Recommendations

Component	Technology
Database Engine	PostgreSQL (open-source, analytics-optimized)
ETL Tool	Python + Apache Airflow (orchestration)
BI Dashboard	Apache Superset / Metabase
Data Quality	Great Expectations / dbt tests
Monitoring	Prometheus + Grafana
Version Control	Git + GitHub
Documentation	Markdown + MkDocs

Why Open-Source?

Cost-effective, community support, flexibility, no vendor lock-in

Dokumen Analisis:

- Business Requirements
- Data Sources Analysis
- ERD Portal (10 entities)
- Dimensional Model (Star Schema)
- Data Dictionary (120+ columns)

Dokumentasi Teknis:

- README.md (comprehensive)
- Laporan LaTeX
- Mermaid diagrams (ERD, Star Schema)
- Sample queries
- Implementation roadmap

GitHub Repository

https://github.com/sains-data/Kelompok10_SatuData/tree/main

Key Takeaways

Fokus Proyek

- Data Mart fokus pada **analisis operasional Portal Satu Data ITERA**
- Bukan sistem akademik transaksional

4 Fact Tables

- Dataset access patterns
- Data quality metrics
- Search behavior
- Institution performance

Desain & Tech Stack

- Star Schema untuk query optimal
- SCD Type 2 tracking historis
- Open-source stack (cost-effective)

Terima Kasih!