Data Pipeline & Data Wrangling (ETL)

Usecase

Perancangan dan implementasi pipeline data ETL (Extract, Transform, Load) *end-to-end* untuk memproses data mentah.

- Lingkup: Mengelola set data awal yang terdiri dari 1000 catatan (rows) yang memerlukan pembersihan intensif. Data dapat dibagi menjadi beberapa bagian untuk melakukan pengujian pipeline.
- Stack Teknologi: Membangun pipeline menggunakan *stack* ETL *open-source* (seperti Apache Airflow, Dagster, Prefect, atau Apache NiFi) untuk orkestrasi dan eksekusi. Tugas ini tidak boleh diselesaikan hanya menggunakan skrip Python biasa.
- Analisis: Melakukan analisis data eksploratif (EDA) awal untuk mengidentifikasi anomali, nilai yang hilang, dan inkonsistensi format.
- Transformasi (Wrangling): Menerapkan aturan data wrangling yang kompleks untuk pembersihan, validasi, dan standardisasi data secara otomatis.
- Normalisasi: Menyeragamkan format yang tidak konsisten di berbagai kolom (khususnya timestamp dan device_platform).
- Pembersihan: Menangani nilai yang hilang (missing values) dan nilai yang tidak valid (cth: durasi negatif) menggunakan logika bisnis yang telah ditentukan.
- Rekayasa Fitur: Membuat kolom data baru (device_type) yang berasal dari data yang sudah dibersihkan.
- Output: Memastikan integritas dan kualitas data pada hasil akhir (output) pipeline agar siap digunakan untuk analisis atau pelaporan.

Aturan Transformasi & Pembersihan (Rules)

Dataset mentah harus diproses dengan menerapkan aturan-aturan berikut secara berurutan:

1. Filter Event Tidak Relevan Beberapa event adalah "noise" (derau) sistem. Hapus semua baris di mana event_type adalah "system_heartbeat", "ad_load", atau None.
2. Validasi User ID Kunci analisis adalah pengguna. Hapus semua baris di mana user_id adalah None, string kosong "", atau "guest".
3. Standardisasi User ID Beberapa user_id tercatat sebagai integer (cth: 12345) dan lainnya sebagai string (cth: "U-12346"). Ubah semua user_id menjadi format string dan gunakan standar penulisan yang sama.
4. Standardisasi Platform Data device_platform tidak konsisten. Lakukan pemetaan ulang (mapping):
    - "android", "Android", "google" → menjadi "Android"
    - "ios", "iOS", "Apple" → menjadi "iOS"
    - "web", "WebApp" → menjadi "Web"
5. Pembersihan Durasi Sesi Kolom session_duration_sec (durasi dalam detik) memiliki masalah:
    - Beberapa nilainya adalah string dengan akhiran "s" (cth: "300s"). Hapus "s" dan ubah ke integer.
    - Beberapa nilainya None atau nilai negatif (cth: -99). Ubah nilai-nilai ini menjadi 0.
    - Pastikan semua nilai akhir adalah integer.
6. Normalisasi Timestamp (Paling Rumit) Kolom timestamp memiliki tiga format berbeda:
    - ISO 8601 (String): "2025-10-20T10:00:00Z" (sudah bagus)
    - Unix Timestamp (Integer/String): 1760923200 atau "1760923200"
    - Format Eropa (String): "20/10/2025 10:00:00" (Format: "DD/MM/YYYY HH:MM:SS")
    - Tugas Anda adalah mengubah *semua* format ini menjadi satu format standar: ISO 8601 UTC (contoh: "2025-10-20T10:00:00Z").
7. Ekstraksi Fitur (Feature Engineering) Buat *key* baru bernama device_type berdasarkan kolom device_platform (yang sudah dibersihkan):

- Jika device_platform adalah "Android" atau "iOS", set device_type ke "Mobile".
- Jika device_platform adalah "Web", set device_type ke "Desktop".
- Jika lainnya, set ke "Other".