Informe: Diseño de flujo ETL batch para consolidación de clientes multi-origen

Objetivos

- Unificar diariamente los registros de clientes (tienda física, web, app móvil) en un repositorio único y normalizado.
- Eliminar duplicados, resolver inconsistencias y normalizar nombres, correos y teléfonos.
- Dejar datos listos para analítica y gobierno (trazabilidad y calidad).

Arquitectura (alto nivel)

- Landing/Raw: recibe archivos CSV/JSON de cada sistema.
- ETL Batch orquestado (Apache NiFi o Talend / Airflow como scheduler).
- Área Curated/DW: base de datos centralizada para el cliente.
- Metastore/Catálogo + Logs & Métricas.

Proceso ETL

1) Extracción

- Fuentes: CSV desde POS, API web, BD de app móvil.
- Frecuencia: diaria (01:00 AM).
- Controles: esquema, tamaño, checksum, nombre patrón.

2) Transformación

- Limpieza: normalización de tildes, fechas, correos.
- Validaciones: regex de emails, teléfonos obligatorios, claves no nulas.
- Deduplicación: reglas determinísticas + fuzzy matching.
- Enriquecimiento: geocodificación, canal de alta.
- Conformado: generación de cliente_hash y surrogate key.

3) Carga

- Destino: tabla dim_cliente y tablas auxiliares.
- Estrategia: upsert por cliente_hash.
- Particionado y clustering para mejorar consultas.
- Índices en email, teléfono, cliente_hash.

Herramienta ETL recomendada

Se recomienda Apache NiFi por su facilidad para ingestas de archivos/API/SFTP, trazabilidad y control de flujos. Talend es una alternativa si se requiere mapeo complejo. Orquestación con Airflow/cron fuera del horario operativo.

Buenas prácticas

- Trazabilidad completa con etl_run_id.
- Idempotencia de procesos.
- Validaciones de calidad con Great Expectations/Deequ.
- Monitoreo con métricas y alertas.
- Seguridad: cifrado, vaults y enmascaramiento.
- Optimización: compresión, particiones y paralelismo controlado.

Paso a paso (ejecución)

1. Ingesta a landing con GetSFTP, InvokeHTTP y JDBC Query.
2. Validación de entrada con ValidateRecord.
3. Normalización de datos con UpdateRecord o mapeo.
4. Deduplicación y Golden Record.
5. Upsert a DW.
6. Registro de rechazos en tabla de errores.
7. Publicación en catálogo.
8. Reportes de ejecución y alertas.