Diseño de Flujo de Ingesta Batch para la Consolidación de Ventas Regionales

Nombres: Jonathan Vasquez Perdomo - Juan Rojas Monserrat

Introducción

La organización cuenta con múltiples sucursales en todo el país, cada una generando archivos de ventas diarios en formato CSV, con estructuras variables según el sistema de origen. Esta diversidad genera dificultades para la consolidación y el análisis. El área de Business Intelligence solicita implementar un flujo de ingesta batch que unifique y prepare los datos de manera confiable y automática, dejándolos listos para su explotación analítica cada mañana.

Objetivo del flujo de ingesta

Consolidar datos de ventas y generar reportes unificados, asegurando que la información procesada cumpla con criterios de calidad, consistencia y trazabilidad para la correcta toma de decisiones.

Estructura del flujo ETL

Fase 1: Extracción

1. Acceder al archivo `.csv` generado por cada sucursal.
2. Verificar la calidad de datos inicial, asegurando que el archivo no esté corrupto.
3. Validar la consistencia de datos, comprobando la estructura y delimitadores.
4. Preparación de datos para transformación, almacenando los archivos en una zona Raw para auditoría.

Fase 2: Transformación

1. **Limpieza** de registros vacíos, nulos o inconsistentes.

2. **Conversión de formatos**, estandarizando fechas, decimales y codificación de caracteres.

3. **Unificación de esquemas**, homologando nombres de columnas y estructuras hacia un modelo canónico.

4. Validación de la calidad de datos, aplicando reglas de negocio (ejemplo: `cantidad > 0`, `precio_unitario ≥ 0`).

Fase 3: Carga

1. Definición del destino en una base de datos de staging y posteriormente en el Data Warehouse corporativo.

2. Garantizar la disponibilidad de los datos listos para análisis a las 03:00 AM, cumpliendo con el SLA establecido.

Herramienta utilizada

La herramienta seleccionada para implementar el flujo es Apache NiFi, debido a:

- Su interfaz visual e intuitiva para la construcción de flujos.
- Amplia conectividad con múltiples fuentes y destinos.
- Capacidades nativas de trazabilidad, validación y programación batch.
- Flexibilidad para gestionar errores, notificaciones y auditoría.

Buenas prácticas implementadas

- Validación de datos en cada fase, garantizando calidad antes de la carga final.
- Trazabilidad completa mediante almacenamiento de archivos originales y registros de auditoría.
- Ejecución en horario nocturno para evitar impacto en operaciones.
- Manejo de errores y alertas automáticas en caso de fallas durante la ingesta.

Esquema general del flujo

1. Recepción de archivos `.csv` desde sucursales.
2. Validación de estructura y preparación en zona Raw.
3. Transformación: limpieza, conversión, estandarización y validación de reglas de negocio.
4. Consolidación en un dataset único diario.
5. Carga a staging y posterior inserción en el Data Warehouse.
6. Disponibilidad de la información consolidada a las 03:00 AM.

Conclusión

El flujo batch diseñado permitirá consolidar la información de ventas de todas las sucursales, garantizando uniformidad, calidad y disponibilidad de los datos. Con el uso de Apache NiFi y la aplicación de buenas prácticas ETL, la organización contará con una solución eficiente para la generación de reportes unificados y el soporte confiable a la inteligencia de negocios.