MODULO 08 - EJERCICIO 03-A

ALEXIS YURI M.

Diseñe un flujo de ingesta de datos en modalidad batch que resuelva esta problemática utilizando herramientas ETL. El flujo debe contemplar extracción, transformación y carga, especificando tareas técnicas y buenas prácticas.

1. Definición del Objetivo del Flujo de Ingesta Batch.

El objetivo principal es consolidar la información de ventas de todas las sucursales del país en una única fuente de datos centralizada. Con esto se busca:

- Crear una fuente única de la verdad: Eliminar los silos de información para que el área de Business Intelligence pueda generar reportes fiables y consistentes.

- Habilitar el análisis de negocio: Permitir a la gerencia obtener una visión completa del rendimiento de las ventas, identificar tendencias por región, comparar el desempeño de las sucursales y optimizar estrategias comerciales.

- Mejorar la eficiencia: Automatizar un proceso de consolidación que, de otro modo, se haría de forma manual, lo que reduce los errores y libera tiempo para el análisis.

2. Descripción de la estructura del Flujo ETL

El flujo se llevaría a cabo cada noche en un proceso ETL:

Extracción:

- Tipo de archivos: .csv.

- Frecuencia: Diaria.

- Origen: Servidores locales de las sucursales o un directorio centralizado donde se suben los archivos.

Transformación:

- Limpieza: Eliminar filas vacías, valores nulos o caracteres especiales no deseados.

- Conversión de formatos: Estandarizar formatos de fechas (ej. DD-MM-AAAA), números (ej. punto decimal) y textos (ej. mayúsculas).

- Estandarización: Unificar los nombres de las columnas, códigos de productos y categorías de venta para garantizar la consistencia.

- Validación de datos: Aplicar reglas para asegurar que los datos estén dentro de rangos válidos (ej. un monto de venta no puede ser negativo).

Carga:

- Destino final: Un Data Warehouse en la nube para el análisis de grandes volúmenes de datos históricos.

3. Elección de la herramienta ETL.

Se seleccionaría Talendc como la herramienta ETL para este proyecto, ya que ofrece una plataforma visual de diseño de flujos de datos. Su interfaz de arrastrar y soltar permite a los equipos crear rápidamente los pipelines de integración sin necesidad de escribir código complejo. Además, cuenta con conectores pre-construidos para archivos CSV y bases de datos, lo que agiliza la extracción y la carga.

4. Buenas Prácticas para el Proceso Batch.

Se incorporarían las siguientes 2 buenas prácticas:

- Validación de datos: En la etapa de transformación, se implementarían reglas para rechazar o marcar registros con errores de formato o valores atípicos. Esto asegura que solo los datos limpios y confiables lleguen al destino final, mejorando la calidad del análisis.

- Programación: El flujo se programaría para que se ejecute en horarios de baja actividad del sistema (ej. durante la noche), para no afectar el rendimiento de los sistemas transaccionales de las sucursales.

5. Representación del Flujo

A continuación se muestra el flujo ETL como un esquema técnico paso a paso:

- Recolección: Un proceso programado se conecta a los directorios de archivos .csv de las sucursales.

- Extracción: Talend extrae los archivos uno por uno y los lee como flujos de datos.

- Filtrado y Limpieza: Los datos son procesados para eliminar registros defectuosos o nulos.

- Transformación: Los datos se unifican, se estandarizan los formatos y se validan con reglas del negocio.

- Carga: Los datos limpios se cargan en el Data Warehouse centralizado.

- Monitoreo: El proceso se registra para verificar la cantidad de registros procesados, los errores y el tiempo de ejecución.

Notificación: Se envía una notificación al equipo de BI cuando el proceso haya finalizado.