MODULO 08 - EJERCICIO 04-A

ALEXIS YURI M.

Diseñe un flujo de ingesta de datos en modalidad batch, utilizando herramientas ETL, que permita consolidar registros de clientes desde múltiples fuentes, transformarlos y cargarlos a un repositorio unificado.

1. Objetivos del Flujo

Problema que resuelve: unificar registros de clientes fragmentados y duplicados provenientes de tiendas físicas (POS), sitio web y aplicación móvil en una base única, consolidada y actualizada diariamente para análisis (comportamiento, fidelización, segmentación).

Qué datos deben consolidarse: identificadores fuente (id\_pos, id\_web, id\_app), nombre y variantes, emails, teléfonos, direcciones, fecha de nacimiento, canales preferidos, historial de compras, estado de fidelidad y metadatos de origen .

2. Descripción del Proceso ETL

Extracción:

- Orígenes de datos:

Archivos .csv de los sistemas de punto de venta de las tiendas físicas.

Tablas de base de datos (SQL o NoSQL) del sitio web.

API de la aplicación móvil.

- Frecuencia: La extracción se realizará diariamente, idealmente durante las horas de menor actividad (noche).

- Formato de entrada: Los datos se obtendrán en sus formatos nativos: .csv, formato de base de datos relacional y JSON de la API.

Transformación:

- Limpieza de datos: Se eliminarán registros con valores nulos o inconsistentes en campos críticos como rut y nombres.

- Eliminación duplicados: Se identificarán y fusionarán los registros duplicados de un mismo cliente utilizando un identificador único (por ejemplo, rut) como clave de negocio. Si un cliente tiene datos en las tres fuentes, se consolidará la información en un solo registro.

- Normalización de campos: Se estandarizarán los campos de datos. Por ejemplo, se convertirá todo a mayúsculas o minúsculas, se eliminarán espacios extra y se estandarizarán los formatos de fechas y direcciones.

- Enriquecimiento: Se puede agregar un campo que indique el origen de cada registro (tienda, web, app) para análisis posteriores.

Carga:

- Destino: Los datos transformados y unificados se cargarán en un Data Warehouse centralizado, como por ejemplo, en una base de datos en la nube. Este repositorio estará optimizado para consultas analíticas.

- Frecuencia: La carga se realizará una vez al día, después de que la etapa de transformación se haya completado con éxito.

3. Selección de una Herramienta ETL

Considerando el escenario de consolidación de datos multi-origen y el procesamiento batch, Talend sería una opción adecuada ya que ofrece una interfaz visual que facilita la creación de flujos de trabajo ETL sin necesidad de escribir mucho código, es versátil y tiene una gran cantidad de conectores pre-construidos que permiten interactuar fácilmente con los distintos orígenes de datos mencionados (archivos CSV, bases de datos, APIs). Además, su capacidad para manejar la limpieza y la eliminación de datos duplicados la hace ideal para este caso de uso.

Otras opciones, como AWS Glue, también son viables si la infraestructura de la empresa ya se encuentra en la nube de AWS, lo que permitiría una integración nativa y simplificada.

4. Incorporación de buenas prácticas.

Ventana de ejecución: El flujo se programará para ejecutarse entre las 2:00 y las 4:00 AM, cuando la actividad de los sistemas de registro es mínima. Esto asegura que el proceso no afecte el rendimiento operativo ni las transacciones en tiempo real.

Gestión de fallos:

- Validación de nodos: Se incluirán nodos de validación en el flujo para verificar la integridad de los datos extraídos antes de la transformación y para asegurar que el número de registros cargados sea coherente con la entrada.

- Logs y alertas: Se configurará un sistema de logeo detallado para registrar cada paso del proceso, incluyendo el número de registros procesados, los errores de limpieza y los registros duplicados. Se implementarán alertas automáticas (por correo electrónico o Slack) que se disparen en caso de fallos críticos, como una conexión de origen fallida o una carga incompleta.

5. Representación Esquemática.

Extracción:

Extracción concurrente: El flujo inicia extrayendo simultáneamente los datos de las tiendas físicas (.csv), el sitio web (DB) y la aplicación móvil (API).

Validación de la fuente: Cada flujo de datos extraído se somete a un primer paso de validación para asegurar que el archivo o la conexión sea correcta y el formato sea el esperado.

Transformación:

Normalización y limpieza: Los datos de cada origen se limpian y normalizan de manera individual. Esto incluye la estandarización de campos (nombre, email, teléfono).

Consolidación y eliminación de duplicados: Los datos de los tres orígenes se combinan en una única tabla de trabajo. Se aplica la lógica de deduplicación para identificar y fusionar los registros del mismo cliente, creando un registro unificado y maestro.

Carga:

El conjunto de datos final, limpio y consolidado, se carga en el Data Warehouse centralizado. Esta carga se puede realizar en modo truncate-and-load (eliminar y volver a cargar) o upsert (insertar nuevos registros y actualizar los existentes).

Monitoreo y reporte:

Una vez finalizada la carga, se genera un reporte final con un resumen del proceso (número de registros extraídos, transformados y cargados) y se envían notificaciones sobre el éxito o fracaso de la ejecución.

