Trabajo Práctico: Procesamiento y Carga de Datos en una Base de Datos

Materia: Base de Datos

Fecha de Entrega: **junio 14, 2024**

*Texto realizado con el soporte de **ChatGPT***

***Objetivo:***

El objetivo de este trabajo práctico es desarrollar una aplicación que realice el procesamiento y la carga de un conjunto de datos en una base de datos. Este proceso, conocido como ETL (Extract, Transform, Load), incluye varios pasos intermedios para limpiar y preparar los datos, resolviendo potenciales problemas antes de su inserción en la base de datos.

***Descripción General:***

Los estudiantes deberán procesar un conjunto de datos a través de un flujo de trabajo (Workflow) que incluirá varios pasos intermedios dependiendo del conjunto de datos elegido. La meta es garantizar que los datos estén limpios y listos para su carga en una base de datos SQLite, la cual será utilizada para facilitar el trabajo local sin inconvenientes.

***Instrucciones Detalladas:***

1. Selección del Conjunto de Datos:

   Los estudiantes pueden buscar y seleccionar un conjunto de datos de fuentes públicas como Kaggle o cualquier otro repositorio de datos públicos. La cantidad de datos es indistinta; lo importante es poder construir el Workflow y ejecutar todos los pasos necesarios para su procesamiento y carga.

2. Análisis de Preprocesamiento:

   Identificar y describir los posibles problemas presentes en los datos. Por ejemplo:

   - Valores nulos o faltantes.
   - Datos fuera de rango.
   - Inconsistencias en el formato de los datos.
   - Duplicados.

   Proponer soluciones para cada uno de estos problemas.

3. Preprocesamiento de los Datos:

Implementar los pasos necesarios para limpiar y transformar los datos. Algunos ejemplos incluyen:

- Eliminación o imputación de valores nulos.
- Normalización de datos para asegurar que todos los valores estén dentro del rango adecuado.
- Conversión de tipos de datos (por ejemplo, convertir cadenas a fechas).
- Eliminación de duplicados.
- Validación de integridad y consistencia de los datos.
- Documentar cada paso realizado durante el preprocesamiento.
- Carga de Datos en la Base de Datos:
- Crear una base de datos SQLite y definir las tablas necesarias para almacenar los datos procesados.

4. Cargar los datos limpios en la base de datos.

   Verificar que los datos han sido insertados correctamente y que la base de datos está en un estado consistente.

***Entrega:***

Los estudiantes deben entregar:

Un informe detallado que incluya el análisis de preprocesamiento, las soluciones propuestas, y una descripción de cada paso realizado durante el preprocesamiento.

El código fuente de la aplicación desarrollada.

Un archivo de la base de datos SQLite con los datos cargados.

***Evaluación:***

Se considerará al momento de la evaluación los siguientes puntos

- La correcta identificación y análisis de problemas en los datos.
- La implementación efectiva de los pasos de preprocesamiento.
- La precisión y claridad en la documentación del proceso.
- La funcionalidad de la aplicación para cargar los datos en la base de datos.
- La calidad del código y su adherencia a buenas prácticas de programación.

Fecha de Entrega: **junio 14, 2024**

***Recursos Adicionales:***

**Kaggle**

Web: https://www.kaggle.com/

Blog: https://www.kaggle.com/datasets/kaggle/kaggle-blog-winners-posts

Foro: https://www.kaggle.com/general

Documentación: https://www.kaggle.com/docs/api

**SQLLite**

Wiki: https://en.wikipedia.org/wiki/SQLite

Foro: https://sqlite.org/forum/

Tutorial: https://www.youtube.com/watch?v=3oJCP7bwJ0Q

Subreddit: https://www.reddit.com/r/sqlite/

Libros: https://www.sqlite.org/books.html

YouTube: https://www.udemy.com/topic/sqlite/

*Herramientas y utilidades:*

DB Browser for SQLite: https://sqlitebrowser.org/dl/

SQLiteStudio: https://sqlitestudio.pl/

***Definiciones realizadas con el soporte de Gemini AI de Google***

**Kaggle**

Plataforma web que reúne a la comunidad de ciencia de datos más grande del mundo, con más de 5 millones de miembros activos en 194 países. Ofrece una amplia gama de herramientas y recursos para aprender, practicar y competir en el campo de la ciencia de datos y el aprendizaje automático.

¿Qué puedes hacer en Kaggle?

Participar en concursos: Kaggle organiza regularmente competiciones donde los participantes compiten por crear los mejores modelos de aprendizaje automático para resolver problemas del mundo real. Estos concursos ofrecen premios en efectivo y son una excelente manera de poner a prueba tus habilidades y aprender de otros expertos

**Workflow**

Un Workflow, o flujo de trabajo, es un modelo digital de un proceso que se racionaliza y divide en diferentes tareas para optimizar el rendimiento y el uso de recursos. Un Workflow es un conjunto de tareas que siguen un orden y reglas específicas para obtener un resultado de manera eficiente. Las tareas pueden ser manuales o automatizadas y pueden estar dispuestas secuencialmente o en paralelo

**Problemas comunes en el procesamiento de datos una descripción general**

Los datos son el activo más importante de muchas organizaciones en la actualidad. Sin embargo, la información contenida en los datos puede verse comprometida por diversos problemas que surgen durante su recolección, almacenamiento y procesamiento. Estos problemas comunes en el procesamiento de datos pueden afectar negativamente la calidad de los análisis, la toma de decisiones y el rendimiento general de las empresas.

A continuación, se presenta una descripción general de los cuatro problemas comunes mencionados:

**1. Valores nulos o faltantes:**

Los valores nulos o faltantes se producen cuando no hay información registrada en un campo o variable particular de un conjunto de datos. Esto puede deberse a diversas razones, como errores de entrada, sistemas incompletos o la naturaleza inherente de los datos. La presencia de valores nulos puede generar sesgos en los análisis y dificultar la interpretación de los resultados.

**2. Datos fuera de rango:**

Los datos fuera de rango se refieren a valores que no se ajustan a los límites esperados o definidos para una variable específica. Esto puede ocurrir por errores de entrada, mediciones incorrectas o inconsistencias en las definiciones de rango. Los datos fuera de rango pueden afectar la precisión de los cálculos y generar resultados erróneos.

**3. Inconsistencias en el formato de los datos:**

Las inconsistencias en el formato de los datos se presentan cuando la forma en que se representan los datos no es uniforme o coherente dentro de un conjunto de datos. Esto puede incluir variaciones en el formato de fechas, horas, números, texto o unidades de medida. Las inconsistencias de formato dificultan la lectura, el análisis y la manipulación de los datos.

**4. Duplicados:**

Los datos duplicados se refieren a registros o filas que contienen la misma información varias veces dentro de un conjunto de datos. Esto puede deberse a errores de entrada, procesos de integración de datos deficientes o la naturaleza de la fuente de datos. Los duplicados pueden inflar artificialmente el tamaño del conjunto de datos y afectar la precisión de los análisis.

**Impacto de los problemas comunes en el procesamiento de datos:**

Los problemas comunes en el procesamiento de datos pueden tener un impacto significativo en las organizaciones de diversas maneras:

- **Reducción de la calidad de los datos**: Los datos de baja calidad pueden generar resultados inexactos y análisis erróneos, lo que lleva a decisiones equivocadas y estrategias ineficaces.
- **Dificultades en la integración de datos**: Las inconsistencias en el formato y la presencia de duplicados pueden dificultar la integración de datos de diferentes fuentes, lo que limita la visión general de la información.
- **Aumento de los costos**: La limpieza y corrección de datos de baja calidad puede requerir un tiempo y recursos considerables, lo que aumenta los costos operativos.
- **Daño a la reputación**: La toma de decisiones basada en datos erróneos puede dañar la reputación de una organización y generar desconfianza entre los clientes o stakeholders.

**Prevención y mitigación de problemas comunes en el procesamiento de datos:**

Para prevenir y mitigar los problemas comunes en el procesamiento de datos, las organizaciones pueden implementar diversas estrategias:

- **Establecer estándares de calidad de datos**: Definir estándares claros para la calidad de los datos, incluyendo la precisión, integridad, consistencia y completitud.
- **Implementar procesos de validación de datos**: Implementar mecanismos para validar la calidad de los datos en el momento de la entrada, durante el procesamiento y antes del análisis.
- **Utilizar herramientas de limpieza de datos**: Emplear herramientas especializadas para identificar, corregir y eliminar errores, inconsistencias y duplicados en los conjuntos de datos.

- **Capacitar al personal**: Brindar capacitación al personal sobre la importancia de la calidad de los datos, las técnicas de limpieza de datos y las mejores prácticas para la gestión de información.

Al abordar proactivamente los problemas comunes en el procesamiento de datos y adoptar un enfoque preventivo, las organizaciones pueden garantizar la calidad y confiabilidad de sus datos, lo que se traduce en mejores decisiones, mayor eficiencia y un mayor retorno de la inversión en sus iniciativas de análisis de datos.

**Estado del arte en el desarrollo de Workflow basado en Althor**

Desarrollo de una Aplicación ETL para Procesamiento y Carga de Datos en una Base de Datos

***1. Introducción y Contexto***

Bueno, para la presentación de informes se pueden incluir ejemplos ilustrativos como gráficos y resúmenes tabulares. Sin embargo, las prestaciones de estos sistemas están muy limitadas por las condiciones de hardware y software de que disponemos. Para ello, se pretende realizar la implementación de un datawarehouse y/o la implementación de un sistema ETL que permitan tapar las carencias de estos sistemas. Sin embargo, hay que tener en cuenta que la implantación de un sistema ETL conlleva una importante carga de trabajo en los departamentos de informática, siendo necesario definir tareas bien planificadas tanto en la identificación de las necesidades de cara a la carga de datos (ETL) como en el diseño técnico a nivel software de la gestión de dicha información (Datawarehouse) con el objetivo de racionalizar los costes de implantación en base a las inversiones tecnológicas realizadas.

El trabajo de fin de máster que se presenta a continuación está centrado en el desarrollo de una aplicación ETL (Extracción, Transformación y Carga) que permita la integración de la información disponible en los diferentes sistemas de ATOS con el objetivo de obtener una vista única de la información en tiempo real. El uso de un sistema ETL tiene como ventajas que posibilita la obtención de la información necesaria para alimentar el Datawarehouse para el año 2019 o de cara a los trabajadores o colaboradores habilitados. La auditoría completa del tratamiento realizado sobre los datos, para que se conozcan aspectos como el detalle de los errores que puedan originarse o incluso el valor de las sumas y otros totales recuperados en un momento determinado del proceso. Permitiría un mayor y mejor conocimiento del dato y la información sobre su evolución. Resulta importante para el Área de Tecnología e Información - ATI de la Fundación bancaria vinculada a la entidad, puesto que posibilita una mayor transparencia

respecto a las funciones y actividades realizadas sobre los datos y sus propiedades, ya que conlleva una mayor trazabilidad del dato.

### *2. Conceptos Fundamentales del ETL*

A lo largo del proceso ETL, los datos son sometidos a las primeras fases de calidad, es decir, a transformaciones y verificaciones que mejoran su idoneidad, para ser utilizados posteriormente por usuarios finales. Dado que la fuente de los datos (generalmente la base de datos de la organización y la base de datos de producción, es interno y externo, respectivamente) suele ser diferente a los usuarios que los utilizan (internos y externos), los usuarios finales deben realizar piezas de software (llamadas aplicaciones de reporting) que posibiliten el análisis de los datos, presupuestas de los datos y las representaciones gráficas de los mismos, para finalmente obtener informes relacionales. Reports es informes derivados: Se trata de informes generados por usuarios, con un nivel de granularidad determinado. La necesidad de obtener regularmente determinados reports obliga muchas veces a planificar y realizar de forma automática la modalidad habitual mediante la programación de un proceso.

Ambos tipos de cargas son necesarios, teniendo en cuenta que no todas las aplicaciones de reporting únicamente trabajan con cubos de datos, sino que normalmente necesitan trabajar también con los datos subyacentes de éstos. Los procesos de carga del medio OLTP permiten de forma automática identificar los datos que han cambiado desde la última carga y almacenar exclusivamente dichos cambios en dichos históricos. Dicho motor implementa las siguientes capacidades fundamentales. Motor ETL: de operaciones putativas de cargado, que soporta la ejecución distribuida y concurrente de muchos flujos, garantizando la consistencia de los datos y aplicando ciertas optimizaciones para reducir el tiempo de cargado.

### *3. Diseño y Arquitectura de la Aplicación ETL*

El diseño y desarrollo de la arquitectura de la herramienta ETL fue un desafío importantísimo. Existen dos tipos de diseño hardware, el de servidor y cliente. Se decidió desplegar solo un servidor con las funcionalidades de procesamiento y ejecución de procesos, lo que trajo beneficios. El proceso máximo de procesos aumentó en un 90%, el procesamiento paralelo disminuyó en un 76%, ayudando a disminuir el índice de clientes si se mantiene un cliente activo cuando este se cierre.

En la figura se muestra el diseño de la arquitectura hardware del sistema. Los usuarios acceden a la aplicación y esta a la Base de Datos. La aplicación ejecuta las cargas de datos, para esto requiere un servidor de directorio (fileserver), el cual contiene los ficheros planos, posible fuente

de datos que sean compartidos por diferentes aplicativos. Seguidamente se lista el diseño de la arquitectura proveniente del diseño de la aplicación EPM / ETL. Microsoft Windows Server es un sistema operativo de servicios muy robusto y confiable que se ofrece para los servidores de empresas, el cual se selecciona con el fin de que presente facilidades en cuanto a la virtualización de máquinas y espacio para instalar diferentes servidores. Es el servidor encargado del manejo de los recursos generales del sistema, entre estos, la administración de la Base de Datos y las interfaces de comunicación con los usuarios. Mediante consola de administración, se encarga de manejar ficheros y directorios, configuración de seguridad, transferencia de información entre diferentes servidores. Es el motor para la ejecución de trabajos en lotes, ofreciendo mayor potencia de proceso y flexibilidad al permitir dividir los trabajos en diferentes trabajos.

***4. Selección y Preparación de los Datos***

Selección y preparación de los datos. Dada la necesidad de que la aplicación ETL obtenga la información necesaria para realizar los procesos de transformación y carga de datos, será necesario implementar un proceso previo a la etapa de ejecución normal del flujo de datos (E y T por separado o los dos juntos): un procedimiento que seleccione y recolecte la información desde sus diferentes fuentes, prepare y depure los datos que serán utilizados por las correspondientes capas para que la siguiente etapa de Transformación y Carga E&D corresponda con aquel otro procedimiento que transforme y cargue los datos en la oportuna Base de Datos. Este proceso se compone de dos partes: la selección y la depuración de los datos.

Puesto que la arquitectura planteada clasifica los datos en función del dominio de estos, será necesario definir cuál ha de ser la capa encargada de recopilarlos y prepararlos para los trabajos de las otras capas. En función del análisis de los Sistemas de información llevado a cabo en el apartado 4 (Selección y preparación de los datos), la realización de este proceso desde el servidor de aplicación de la capa de Dominio para los servicios de Infraestructura y AdSis puede agilizar de manera oportuna el desarrollo de la aplicación, planteando que sea la misma capa de Dominio la que desarrolle las funciones de esta capa. En cambio, si la aplicación no pudiese acoplar los datos desde Infraestructura, necesitará definir un nuevo procedimiento que recupere los datos del SGBD de Infraestructura generándolo en la capa de Dominio.

***5. Procesamiento y Transformación de los Datos***

A continuación, se muestra un esquema de estadísticos que recoge una lista de campos junto con su descripción, incluyendo los siguientes campos categóricos: tipo de variable, operación estadística, descripción y registro/valor (más información de las estadísticas descriptivas).

Variable = Nombre del campo

Tipo Variable: Indica el tipo de variable que se va a seleccionar:

- Numérica: Los valores para esta variable son numéricos

- Categórica (nominal): El campo será seleccionado como categórico (Ver más información sobre las opciones disponibles al seleccionar una categórica)

- Fecha/Hora: El campo contiene fechas

Operación: Operación estadística a aplicar

- Ninguna: Ninguna operación se refiere a que no se mostraría el campo en la tabla resultante

- Medida: Calcula la medida indicada en función de los parámetros seleccionados al cualificar la operación

- Cuenta: Cuenta el número de registros

- Distinto: Cuenta el número de registros distintos con respecto a un campo

- Porcentaje: Calcula el porcentaje respecto al total de las filas, en función del ámbito seleccionado, con respecto a una medida

- Suma Acumulativa: Suma acumulativa de una medida en función de la jerarquía seleccionada

- Rango: Número de filas desde el inicio de la tabla con respecto a una operación de agrupación (ASCENDENTE/DESCENDENTE)

- Disallow: Permit indica si se aplica o no la restricción de valores únicos al campo

Tamaño (p.e. VARCHAR) = Tamaño de caracteres que puede tener el campo

Alias = Nombre que tendrá el campo a seleccionar en la tabla resultante.

***6. Carga de Datos en una Base de Datos SQLite***

Ya que se implementó el módulo de extracción y transformación de datos, entonces, una última etapa es realizar el almacenamiento de los datos ya procesados en el formato que el usuario considere pertinente para la consulta y utilización en otras herramientas o plataformas. Usualmente, una forma de presentar la información es almacenarla en una base de datos relacionales. Para almacenar los datos transformados en un motor de base de datos se debe

especificar el motor al que la ETL destinará los datos y luego se deben modelar las tablas y definir las relaciones que conformarán la estructura de la base de datos. En este trabajo, se seleccionó como motor de base de datos SQLite el cual es un motor muy utilizado por su ligereza y que permite incluir la base de datos en el paquete de la aplicación.

La capa de acceso a datos es la que realiza la tarea de interactuar directamente con el motor de base de datos seleccionado. Algo muy positivo de usar SQLite como motor de base de datos es que es compatible con varias plataformas, con independencia del lenguaje de programación empleado. Por este motivo, Android gestiona las bases de datos SQLite mediante la API contenida en el paquete android.database.sqlite. Aunque SQLite proporciona un motor de base de datos ligero, rápido y versátil, es muy limitado en cuanto a tipos de datos, puesto que solo soporta cinco tipos: texto (TEXT/BLOB), enteros (INTEGER), decimales reales (REAL) y fechas (TEXT). Además, el motor SQLite no dispone de componentes para la verificación y el desarrollo de bases de datos, es decir, no permite el uso de CONSTRAINT para la definición de restricciones referenciables y tampoco integridad referencial. En cuanto a la cláusula CHECK, hay que tener en cuenta que hasta la versión 2.0.1, la base de datos de Android no tiene en cuenta dicha cláusula.

### *7. Pruebas y Validación del Proceso ETL*

Todo proceso de desarrollo de software necesita pruebas que garanticen que lo diseñado cumple con los requisitos establecidos. La prueba consistirá en verificar que los datos de entrada son válidos, que los datos cumplen los requisitos mínimos, visualizar los resultados para garantizar que los datos son correctos o realizar alguna comparación de los datos cargados con los datos fuente. El tipo de pruebas a realizar serán pruebas de caja negra o funcionales, en donde se presenta un conjunto de pruebas que típicamente involucra la verificación de los resultados de la carga de datos usando el proceso de ETL, sin hacer referencia a cómo el proceso transforma los datos, sino a que los datos son correctos. Dentro de las pruebas tenemos las pruebas unitarias y pruebas de integración. En el caso de las pruebas unitarias consiste en verificar que el proceso sea capaz de leer el archivo fuente, transformar y cargar la información al sistema; para ello se utilizaron pequeños archivos con información acotada de 10 a 20 registros.

Una vez establecidas las pruebas unitarias, se pasó a la etapa de pruebas de integración de los procesos ETL con la aplicación; se usó un archivo de personas de buen tamaño para verificar que el proceso de escritura de registros a la base de datos sea eficiente y con ello establecer parámetros para los procesos restantes. Luego, previo a la ejecución de los procesos completos, se realizaron pruebas a fondo con los archivos de personas ya trabajados (censo 2002 y registro

civil). Se verificó el desempeño y si se cumple con los requisitos mínimos establecidos. Durante la ejecución del ETL se generan archivos de reporte que muestran información detallada de las excepciones que se produjeron, detallando el registro que causó el error; estos archivos se generaron en cada uno de los procesos para dar información según la necesidad solicitada. En el caso de la base de datos, se generaron un par de vistas que permiten una primera visualización de los datos cargados.

### *8. Conclusiones y Futuras Mejoras*

La implementación de una aplicación ETL toma en cuenta un conjunto de herramientas que facilitan la recolección, transformación y carga de datos hacia una base de datos. El desarrollador debe tener estos conceptos implantados, pero los sets que necesita para la construcción e implementación eran poco claros durante el desarrollo. Para tal rama de pruebas, el conjunto de herramientas disperso que ofrece SAS se integró en un método descriptivo. Esto incluye una revisión especial de los sets de datos permanentes, intermedios y finales para entender bien cómo cada procedimiento manipulaba estos sets. Al final, se requiere una revisión y afianzamiento de todos los conceptos vistos en la primera rama. Por lo cual, es altamente factible que en la mayoría de las grandes aplicaciones de ETL existan subaplicaciones, ejecutables, shell y scripts implementados para limpiar y/o preparar archivos de datos específicos antes de que estos sean cargados al almacén de datos, o para validar la información resultante y escribir los archivos de salida.

Además, las aplicaciones ETL se benefician de contar con herramientas de auditoría, tales como perfiles de control y estadísticas, que facilitan la representación en dos niveles de las desviaciones con respecto a la información anterior y posterior. La aplicación SAS DI Studio permite ejecutar de forma automática las transformaciones utilizando el estatus de ejecución del job, ya que si falla alguna tabla por las restricciones, se cuelga la operación y se deben recargar los datos. Otras herramientas ETL no cuentan con esto, generando un resultado distinto al final del proceso.