Trabajo final

Infrati

2023-2

Integrantes

Yasmin Johanna García

Javier Ricardo Muñoz

Yesid Montaño Cuero

Introducción

La integración de datos provenientes de diversas fuentes es un desafío fundamental en la gestión de información empresarial. El proceso de Extract, Transform, Load (ETL) se convierte en un componente esencial para consolidar datos de bases de datos heterogéneas. En este contexto, la combinación de fuentes como bases de datos relacionales (utilizando lenguaje SQL), bases de datos NoSQL como MongoDB y sistemas de almacenamiento como Elephant, demanda una estrategia ETL meticulosa y adaptable.

El primer paso, la extracción (Extract), involucra la recopilación de datos desde las fuentes respectivas. En el caso de bases de datos relacionales, se puede emplear consultas SQL para extraer datos tabulares, mientras que, en MongoDB Compass, la extracción implica la captura de documentos no estructurados. En paralelo, para una base de datos en Elephant, el proceso implica la lectura y transferencia de grandes conjuntos de datos.

Una vez obtenidos los datos, el siguiente paso es la transformación (Transform). Este proceso implica la limpieza, normalización y estructuración de los datos para garantizar la coherencia y homogeneidad. En el caso de datos provenientes de diferentes bases de datos, se deben aplicar transformaciones específicas para alinear esquemas y formatos. Aquí, la interoperabilidad entre las estructuras de datos relacionales y NoSQL debe ser cuidadosamente gestionada.

Finalmente, el paso de carga (Load) consiste en insertar los datos transformados en la base de datos de destino, ya sea otra base de datos relacional, NoSQL o un sistema como Elephant. Este proceso garantiza que los datos consolidados sean cohesivos y estén listos para su análisis o consulta posterior. La comprensión detallada de las peculiaridades de cada fuente de datos y la implementación efectiva de procesos ETL son cruciales para una integración de datos exitosa en entornos diversificados.

Análisis y ejecución del proceso

En esta perspectiva, resultó importante iniciar el proceso con un reconocimiento exhaustivo de las diversas fuentes. Como primera medida, se llevó a cabo un análisis detallado de las fuentes relacionales disponibles, evaluando su vinculación con el archivo de salida esperado del proceso ETL.

Simultáneamente, se abordó la comprensión de las fuentes no relacionales disponibles y su correspondencia con el formato anticipado del archivo resultante del proceso ETL. Posteriormente, la secuencia de pasos continuó con la ejecución del proceso de extracción y transformación de datos desde la base de datos relacional, seguido por la realización del mismo procedimiento para las fuentes no relacionales. Este enfoque permitió integrar de manera efectiva la información proveniente de diversas fuentes, asegurando la coherencia entre los datos extraídos y el formato final deseado.

En el proceso fue importante reconocer las llaves primarias que permitían hacer Múltiples Joins además de ordenar (Sort Rows) desde la clave PID. El documento CSV comprendía una serie de características asociadas a propiedades que debian concatenarse para al final cumplir con el objetivó de, bajo dichas características, verificar su precio final

Interfaz

Este proceso integral se desarrolló en la plataforma Pentaho, cuyo enfoque principal fue la ejecución de tareas de Extracción, Transformación y Carga (ETL). No obstante, Pentaho ID destaca por ofrecer funcionalidades más robustas que abarcan diversas áreas, como:

Funcionalidades Clave de Pentaho:

- **ETL (Extracción, Transformación y Carga):** Pentaho Data Integration, también conocido como Kettle, facilita la integración de datos desde diversas fuentes, siendo una herramienta esencial para el proceso ETL.
- **Generación de Informes:** Pentaho Reporting permite la creación de informes detallados y visualmente atractivos, con capacidades flexibles de diseño y distribución.
- **Análisis OLAP:** Pentaho Analysis Services posibilita la creación de cubos OLAP para un análisis multidimensional eficiente.
- **Paneles de Control (Dashboards):** Pentaho Dashboard Designer facilita la creación de paneles de control interactivos, proporcionando una visualización intuitiva de los datos.
- **Minería de Datos:** Pentaho Data Mining ofrece herramientas para descubrir patrones y tendencias en grandes conjuntos de datos.

Pentaho es empleado por profesionales de inteligencia empresarial, analistas de datos y desarrolladores para implementar soluciones de análisis y gestión de datos. La plataforma se destaca por su alta personalización y compatibilidad con diversas fuentes de datos, convirtiéndola en una herramienta versátil para abordar los desafíos de integración y análisis de datos en entornos empresariales.

Conclusiones

En nuestro caso, usamos diferentes versiones del ETL; algunas generaban errores y otras transformaciones obtenían algunos, pero no todos los atributos, finalmente obteniendo como producto final

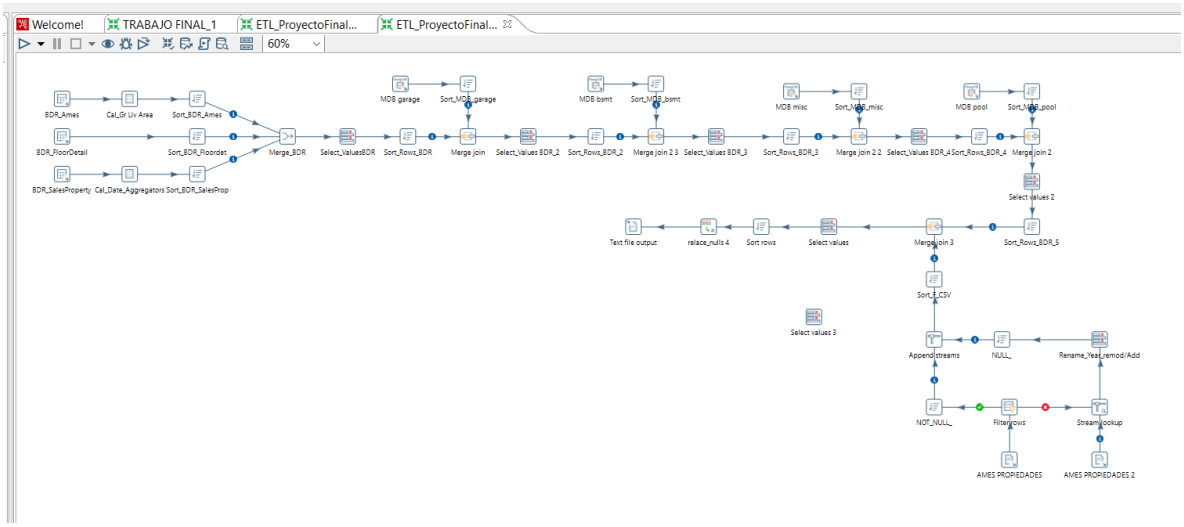

Figura 1: Trabajo Final, vista Pentaho

Una de las cuestiones importantes en este proceso fue el identificar el número final de atributos; esto nos permitió ajustar cada paso y que no se perdiera información en el proceso.

Al final se obtuvo un documento de 81 atributos y 2930 datos.

Lecciones aprendidas

En el proceso de Extracción, Transformación y Carga (ETL) y en la gestión de bases de datos tanto SQL como NoSQL, además el uso de MongoDB Compass, ha consolidado un conocimiento profundo sobre la imperiosa necesidad de integrar datos de manera eficaz. Este viaje de aprendizaje ha destacado la complejidad inherente al abordar diversas fuentes de datos, desde las relacionales hasta las NoSQL, subrayando la importancia crítica de realizar procesos de limpieza y transformación para garantizar la coherencia y la calidad de los datos. La capacidad de ejecutar consultas SQL para extraer información de bases de datos relacionales se ha revelado como una habilidad fundamental, al igual que la apreciación de los modelos NoSQL, especialmente en el caso de MongoDB.

En paralelo, la aplicación de Pentaho como una herramienta integral ha añadido un nivel de sofisticación a la gestión de datos. La eficacia de Pentaho Data Integration en la orquestación de flujos ETL, la versatilidad de Pentaho Reporting para la generación de informes detallados y la potencia de Pentaho Analysis Services para análisis multidimensionales, han enriquecido sustancialmente el conjunto de habilidades del individuo. La capacidad de crear paneles de control interactivos mediante Pentaho Dashboard Designer ha demostrado ser esencial para la visualización intuitiva de datos, proporcionando una interfaz que facilita la interpretación y toma de decisiones informadas.

El aprendizaje ETL, bases de datos SQL y NoSQL, junto con el uso de Pentaho, ha ampliado significativamente la comprensión y habilidades para gestionar y analizar datos de manera efectiva

en entornos empresariales dinámicos. Esta experiencia ha subrayado la versatilidad y la integración como elementos esenciales para afrontar los desafíos actuales en la inteligencia empresarial y la gestión de datos. Este continuo proceso de aprendizaje ha fortalecido la apreciación por la importancia estratégica de la integración de datos y la analítica para la toma de decisiones fundamentadas en el ámbito empresarial.

Bibliografía:

- Hall, S., Shavor, J., Bouman, R., & Dong, Z. (2006). *Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL*. Wiley.
- Matt Casters, Roland Bouman, Jos van Dongen. (2009). *Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration*. Wiley.
- Gómez, A., & Mejía, S. (2015). *Pentaho 5.0 Reporting by Example: Beginner's Guide*. Packt Publishing
- *Tecnología y Negocios. (13 de Diciembre de 2023). Obtenido de Ipod: https://www.itop.es/blog/item/que-es-pentaho-y-cuales-son-sus-beneficios.html*