

*Ejercicio de Laboratorio 7: ETL – P2*

Instituto Politécnico Nacional.

Escuela Superior de Cómputo.

Licenciatura en ciencia de datos.

Bases de Datos Avanzadas

Emiliano López Méndez.

Hinostroza Loera Leonardo

Antonio Aguilar Velazquez

Escudero Robles Arantxa

Lugo Acosta Paulina

**Introduccion**

En el presente laboratorio, abordaremos un flujo completo de proceso ETL (Extracción, Transformación y Carga), utilizando herramientas avanzadas para el manejo y limpieza de datos. Este ejercicio práctico se enfoca en asegurar la calidad de los datos, eliminando valores faltantes y resolviendo códigos postales incompletos mediante la integración de información adicional. La finalidad de esta actividad es fortalecer nuestras competencias en la manipulación y transformación de datos, con el objetivo de prepararlos para análisis profundos o su almacenamiento en bases de datos, procesos fundamentales en el campo de la ciencia de datos..

**Desarrollo de la Actividad:**

**Extracción de Datos**: Se inicia el proceso con la carga de datos desde un archivo CSV. En esta fase, se usa un paso de entrada de archivo de texto para acceder al archivo sales\_data.csv, visualizar su estructura y preparar los datos para los pasos siguientes.

**Filtrado de Registros con Códigos Postales Faltantes**: Se aplica un filtro para retener solo aquellos registros completos en los que se disponga de un código postal, evitando cargar datos incompletos en la base de datos.

**Resolución de Códigos Faltantes**: Para los registros sin código postal, se integran datos adicionales de un segundo archivo que contiene información de ciudades, estados y códigos postales. Esto permite enriquecer los datos originales y completar los campos vacíos.

**Limpieza y Normalización de Datos**: Una vez que los datos cuentan con la información necesaria, se efectúan ajustes adicionales. En el campo país, se unifican los valores, reemplazando términos como "Estados Unidos" por "EE. UU." y categorizando las ventas en pequeñas, medianas y grandes.

**Ejecución y Orquestación**: Finalmente, se utiliza la herramienta Pentaho para coordinar y ejecutar los trabajos de transformación de manera programada. Este proceso asegura la consistencia y la actualización periódica de los datos, simulando un escenario en el cual un sistema externo coloca el archivo de entrada en una ubicación específica, y el trabajo de ETL se ejecuta automáticamente.

**Parte 1 Extract and load data**

En el Paso 1, recuperará datos de un archivo plano CSV y utilizará el paso Entrada de archivo de texto para conectarse a un repositorio, ver el esquema del archivo y recuperar el contenido de los datos.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Texto

Descripción generada automáticamente Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente Interfaz de usuario gráfica, Texto

Descripción generada automáticamente Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente Interfaz de usuario gráfica, Texto, Aplicación, Word

Descripción generada automáticamente

**Parte 2 Filter for missing codes**

Después de completar el Paso 1: Extraer y cargar datos, estará listo para agregar un componente de transformación a su canalización de datos. El archivo fuente contiene varios registros a los que les faltan códigos postales. Esta sección del tutorial filtra aquellos registros a los que les faltan códigos postales, donde el POSTALCODE no es nulo (la condición verdadera) y garantiza que solo se carguen registros completos en la tabla de la base de datos.

Interfaz de usuario gráfica, Tabla

Descripción generada automáticamente Diagrama

Descripción generada automáticamente Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**Parte 3 Resolve missing codes**

Después de completar el Paso 2: Filtrar los códigos faltantes, estará listo para resolver los códigos postales faltantes. En esta sección, aprenderá a utilizar un segundo archivo de texto que contiene una lista de ciudades, estados y códigos postales, para buscar los códigos postales de aquellos registros en los que faltan campos, que es la rama falsa de su Filtro. paso de filas.

Primero, utilizará un paso de entrada de archivo de texto para leer el archivo fuente. Luego, utilizará un paso de búsqueda de secuencia para incorporar los códigos postales resueltos a la secuencia. Por último, utilizará el paso Seleccionar valores para cambiar el nombre de los campos en la secuencia, eliminar campos innecesarios y más.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente Tabla

Descripción generada automáticamente Diagrama

Descripción generada automáticamente Diagrama

Descripción generada automáticamente con confianza media Interfaz de usuario gráfica, Texto, Aplicación, Tabla

Descripción generada automáticamente Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente Texto

Descripción generada automáticamente Interfaz de usuario gráfica, Tabla

Descripción generada automáticamente con confianza media Escala de tiempo

Descripción generada automáticamente

**Parte 4 Clean the data**

Después de completar el Paso 3: Resolver los datos faltantes, puede limpiar y categorizar aún más los datos en depósitos antes de cargarlos en una base de datos relacional. En esta sección, limpiará los datos del campo PAÍS asignando Estados Unidos a EE. UU. mediante el paso del asignador de valores. La limpieza de los datos garantiza que solo haya una versión de EE. UU.

Además, aprenderá a utilizar depósitos para categorizar los datos de VENTAS en categorías pequeñas, medianas y grandes mediante el paso Rango de números. Aprenderá cómo insertar estas funciones de limpieza y categorización en su transformación justo antes del paso Escribir en la base de datos en el lienzo.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

**Parte 5 Run the transformation**

Pentaho Data Integration proporciona varias opciones de implementación. La sección Ejecutar una transformación en el documento Integración de datos de Pentaho explica estas y otras opciones disponibles para su ejecución. En esta sección del tutorial, creará una transformación utilizando la opción Ejecución local.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza baja Tabla

Descripción generada automáticamente Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**Parte 6 Orchestrate with jobs**

Los trabajos se utilizan para coordinar actividades de ETL como:

Definir el flujo y las dependencias que controlan el orden lineal para que se ejecuten las transformaciones.

Preparación para la ejecución comprobando condiciones como "¿Está disponible mi archivo fuente?" o "¿Existe una tabla?"

Realizar operaciones de carga masiva de bases de datos.

Ayudar a la gestión de archivos, como publicar o recuperar archivos mediante FTP, copiar archivos y eliminarlos.

Envío de notificaciones de éxito o fracaso a través de correo electrónico.

Para esta parte del tutorial, imagine que un sistema externo es responsable de colocar su entrada sales\_data.csv en su ubicación de origen todos los sábados por la noche a las 9 p. m. Quiere crear un trabajo que verifique que el archivo haya llegado y luego ejecute la transformación para cargar los registros en la base de datos. En un ejercicio posterior, programará el trabajo para que se ejecute todos los domingos por la mañana a las 9 a. m.

Imagen que contiene Diagrama

Descripción generada automáticamente Diagrama

Descripción generada automáticamente Texto

Descripción generada automáticamente

**Conclusión**

La práctica realizada permitió afianzar nuestras habilidades en el proceso ETL, destacando la importancia de la calidad y completitud de los datos para evitar errores en análisis posteriores. Al resolver valores nulos y estandarizar campos, se logró un dataset coherente y listo para ser almacenado o utilizado en análisis futuros. Este ejercicio también reforzó el uso de herramientas como Pentaho para la automatización de tareas en proyectos de ciencia de datos, contribuyendo a la eficiencia y organización en el manejo de grandes volúmenes de información.