Universidad Católica Del Maule

Facultad Ciencias De La Ingeniería

Escuela de Ingeniería Civil Informática

Curso: INF-522

24-10-2022

ETL en Pentaho

Integrantes: Mario Labbé Sanhueza

Mario González Galdames

Thadli Guerra Ibarra

Luis Inostroza Flores

Docente: Natalia Morales Rosales

Introducción

Como bien sabemos, la inteligencia de negocios es un conjunto de estrategias que resulta hoy en día ser un factor crucial e imprescindible para el funcionamiento de todas las empresas las cuales tienen que manejar grandes volúmenes de datos, debido a que los datos que se van almacenando a lo largo del tiempo pueden ser transformados en información vital y/o valiosa para el funcionamiento o crecimiento de la empresa puesto que esta información puede abarcar la comprensión del funcionamiento actual y también la anticipación de acontecimientos futuros.

A raíz de esto nacen las famosas técnicas ETL (Extract, Transformation, Load), que como su nombre lo indica, permiten extraer datos de diversas fuentes utilizadas por las empresas para depurarlos y prepararlos, y finalmente cargarlos en un almacén de datos donde posteriormente serán analizados.

Para realizar lo descrito anteriormente, existen diversas herramientas de inteligencia las cuales se basan en la utilización de diferentes sistemas de información de inteligencia para brindar soluciones a este tipo de problemáticas, en particular durante el desarrollo de este trabajo se utilizará la herramienta de ETL Pentaho.

Pentaho Data Integration es un software de código abierto que posee herramientas de inteligencia de negocios, que permiten convertir datos brutos en datos netos, descartando la información poco útil y dejando aquella que sí nos puede servir en un formato más limpio. Esta herramienta nos brinda un portal web completo con su repositorio, gestión de usuarios y roles, instalación de plugins para extender funcionalidades, etc.

Durante el desarrollo de este trabajo se utilizará esta herramienta descrita para automatizar un proceso que genera indicadores de gestión para una empresa que da respuestas a clientes sobre demandas laborales, los procesos y recursos serán descritos a lo largo del trabajo.

Descripción de objetos y lógica empleada de cada caso

**Comienzo de trabajo:**

Diagrama

Descripción generada automáticamente

En este paso se crean los objetos necesarios para la lectura de las hojas Excel propuestas. Se crea un nuevo archivo Excel con los feriados correspondientes de los años 2021 – 2022 para poder trabajar de mejor manera las comparaciones posteriores, se crea una fórmula para diferenciar cada día de la semana y luego de ordenar toda la información se une a través de un FULL JOIN.

**1.- ¿Es feriado?: indica con un “NO” o la “fecha registro” si de Hoja Base (fecha**

**registro) corresponde a un día feriado.**

Diagrama

Descripción generada automáticamente

Luego de realizar el “Join” y unir las 2 hojas de Excel, se separan los días hábiles y no hábiles a través de un identificador creado previamente con el objeto “Formula”, para así trabajar con los días por separado, para los días hábiles se crea una nueva columna llamada “¿es feriado?” y se rellenan todas las filas con un “NO”, luego para los días no hábiles se crea de igual forma una columna llamada “¿es feriado?” y se rellena cada fila con la “fecha registro” correspondiente, luego se unen esos 2 resultados al final con un “Dummy”.

**2.- Fecha registro actualizada: Si columna ¿Es feriado? = NO, entonces se registra la fecha de la columna fecha registro; de lo contrario, se debe registrar el primer día hábil siguiente a la fecha registro.**

Diagrama

Descripción generada automáticamente

Este procedimiento se realizó en conjunto con el procedimiento 1, antes de unir los 2 resultados, se crea una nueva columna con el nombre “Fecha registro actualizada” para las filas que tienen “NO” en la comuna “¿es feriado?” y se rellena con la “fecha registro”. Para los días no hábiles se les tiene que registrar el primer día hábil siguiente, así que se separan los sábados y domingos a través de un “Filter Rows” para así crear la fila “Fecha registro actualizada” con 2 días sumados en caso de que sea sábado y 1 día en caso de que sea domingo. Finalmente se unen todos los resultados con un Dummy.

**3.- Fecha Máximo Pago: corresponde a la Fecha registro actualizada más 7 días hábiles. Se debe empezar a contar a partir del día siguiente de la “Fecha registro actualizada”.**

Diagrama, Mapa

Descripción generada automáticamente

Al igual que en el procedimiento anterior, se crea con el objeto “Formula” un identificador para los días de la semana y con la ayuda de un “Filter rows” separamos los días jueves y viernes de los demás, esto es porque solo a los días jueves y viernes es necesario sumarles 11 días mientras que a los demás 9 días para considerar los 7 días hábiles solicitados.

**4.- Plazo máximo ingreso dictamen: Corresponde a la fecha registro más 4 días hábiles. Se debe contar a partir del día siguiente de la fecha registro.**

Mapa de colores

Descripción generada automáticamente con confianza baja

Al igual que en los procedimientos anteriores, este procedimiento consiste en separar los días de la semana con el objeto “filter rows” ya que en este caso habría que sumarles números distintos dependiendo del día. En este caso a los días miércoles, jueves y viernes habría que sumarles 6 días para considerar los 4 días hábiles. Por otro lado al día domingo se le sumarían 4 días, al día sábado se le suman 5 días, al día lunes 4 días y al día martes 6 días. Se crea la comuna “Plazo máximo ingreso dictamen” en cada grupo de días separados y se les suma la cantidad de días correspondientes a la “fecha registro” y se envían a un “Dummy”. Finamente se unen todos los “Dummy” resultantes en un “select values” para eliminar las columnas sobrantes.

**5.- Registro Fuera Plazo: Si fecha dictamen es mayor a Plazo máximo ingreso**

**dictamen, entonces “SI”, de lo contrario “NO”.**

**Diagrama

Descripción generada automáticamente con confianza media**

En este procedimiento se utilizó el objeto “filter rows” con la condición propuesta para separar las filas donde la “fecha dictamen” fuera mayor a “plazo máximo”, luego de separarlas se usa el objeto “calculator” para agregar la columna “registro fuera de plazo” con el respectivo “SI” o “NO”, finalmente se eliminan las columnas innecesarias y se une todo con un “Dummy”.

**6.- Plazo máximo registro 7 días: Corresponde a la fecha registro más 7 días hábiles. Se debe contar a partir del día siguiente de la fecha registro.**

**Diagrama

Descripción generada automáticamente**

Este procedimiento se realiza de igual manera que el procedimiento 3, solo que esta vez se trabaja con la “fecha registro”. En resumen, se le suman 11 días a los días jueves y viernes, a los sábados se le suman 10 días y a los demás 9 días.

**7.- Registro Fuera Plazo 7 días: Si fecha dictamen es mayor a Plazo máximo registro 7 días, entonces “SI”, de lo contrario, “NO”**

**Diagrama

Descripción generada automáticamente**

En este procedimiento se crea el objeto “filter rows” con la condición de que si el registro con su “fecha dictamen” es mayor a “plazo máximo registro 7 dias” se crea la columna “Registro Fuera Plazo 7 Dias” y se rellenan las filas con “SI”, de lo contrario se crea de igual forma la columna y se rellena con “NO”. Finalmente se establecen los formatos pertinentes con el objeto “select values”.

**8.- Registro Anterior: Validar que en la hoja Base Anterior existe el mismo número de registro, si es así, entonces “SI”, de lo contrario, “NO”. Si existen dos registros para el mismo número de registro, considerar solo el que contenga la fecha más reciente. Solo debe considerar un registro.**

**Diagrama, Mapa

Descripción generada automáticamente con confianza media**

Para este procedimiento se utilizan los objetos respectivos para leer la hoja “Base anterior” y ordenar los datos. Una vez hecho esto se usa el objeto “Merge join” para unir la información obtenida con la que teníamos previamente, utilizamos el objeto “filter rows” para validar si existe el mismo número de registros, finalmente con una calculadora se crea la columna respectiva.

**9.- Ultima fecha registro anterior: si columna registro anterior = “SI” entonces copiar la fecha de la columna fecha registro de la hoja Base Anterior, de lo contrario, dejar en blanco.**

**Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente**

Para este procedimiento se utiliza el objeto “filter rows” para separar las filas que poseen la “columna registro anterior = SI” de las que poseen la “columna registro anterior = NO”, luego con una calculadora se crea la columna “ultima fecha registro anterior” para las filas que tienen “columna registro anterior = SI” y se realiza una copia de “fecha registro” de la hoja “Base Anterior” en la columna creada, finalmente se unen los registros con un “Dummy”:

**10.- ¿Tienen la misma fecha?: si columna Ultima fecha registro anterior es igual a fecha registro, entonces “SI”, de lo contrario “NO”**

**Diagrama

Descripción generada automáticamente**

En este procedimiento, se utiliza el objeto “filter rows” para realizar la comparación solicitada, en caso de que las fechas sean iguales, los registros se van a un objeto “calculator”, se crea una nueva columna con el nombre “¿tiene la misma fecha?” y se rellena con “SI”, de lo contrario se rellena con “NO”. Finalmente se usa el objeto “select values” para eliminar las columnas que no se utilizan y ajustar los formatos pertinentes.