**Tarea: ETL**

**Introducción**

**Objetivo**

Realizar un proceso ETL básico en Pyspark

**¿Para qué?**

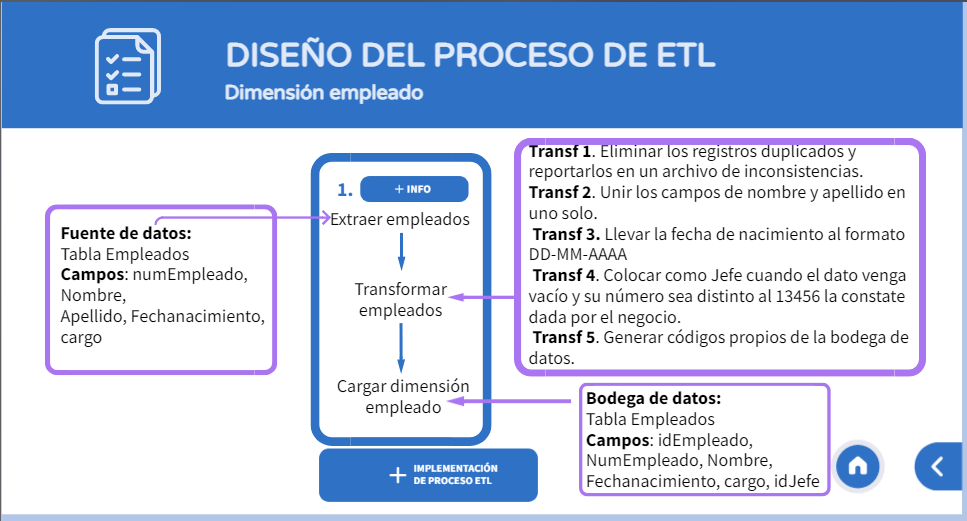
Practicar lo aprendido en el tutorial de ETL

**¿Qué necesita?**

1. Modelo multidimensional asociado al proceso de movimientos
2. Notebook para trabajar: puede usar la seccion 3 “Espacio para desarrollar la tarea” al final del notebook del tutorial para realizar esta actividad
3. Servidor SQL con base de datos relacional “WWImportersTransactional” y base de datos relacional que corresponde a la bodega de WWI de cada estudiante “Estudiante\_i”

**Enunciado**

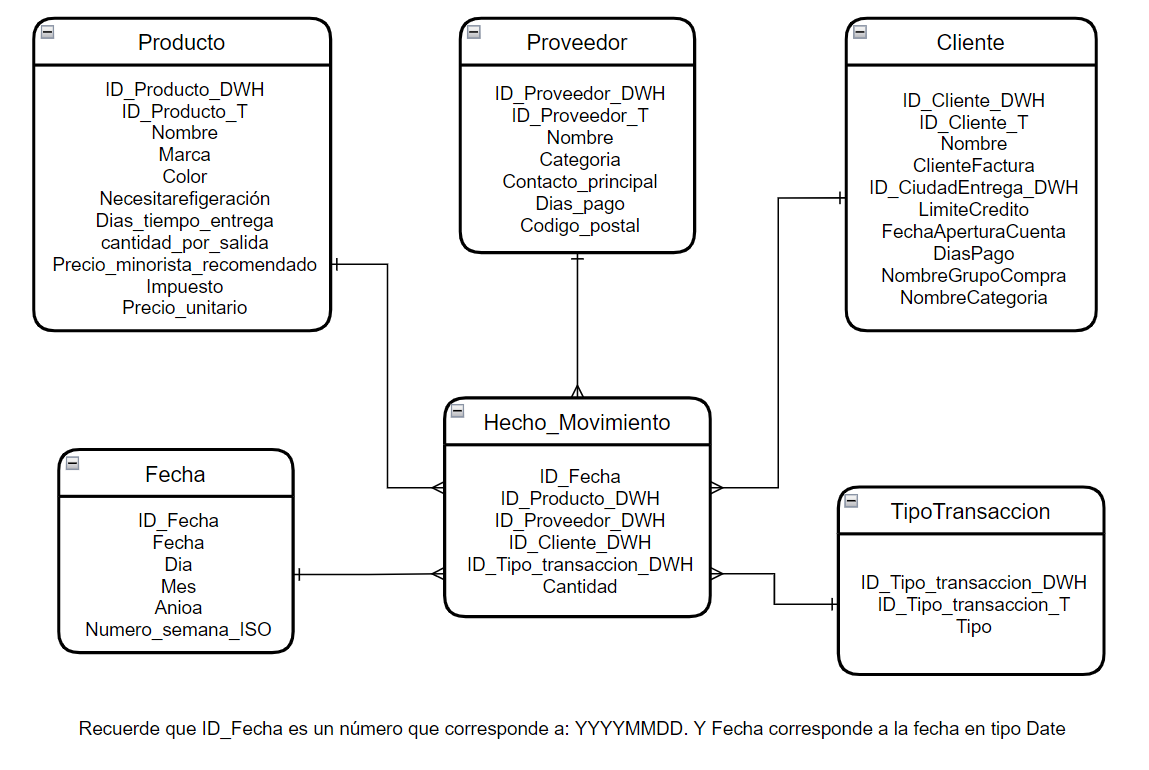
Ahora que sabe cómo realizar un proceso ETL, dado el modelo multidimensional del proceso de negocio de movimientos de inventario realice las siguientes actividades:

1. **Entregable 1 - Diseño del ETL:** Diseñe el ETL para las dimensiones proveedor, tipoTransaccion, fecha y para la tabla de hechos. A nivel de la tabla de proveedores incluya la tabla categoriasProveedores donde encuentra información de las categorías. El diseño del ETL es un diagrama como lo encuentra en la infografía de proceso ETL y puede observar en la siguiente figura. En el entregable utilice la siguiente [plantilla para presentar el diseño de ETL](https://miad-modelo-datos.github.io/guias_v1/Tareas/PlantillaDise%C3%B1oETL.xlsx) e incluya una descripción general que permita comprender detalles que pueden ser confusos.  
   
2. **Entregable 2 - Implementación del ETL:** implementación del proceso ETL para las dimensiones Proveedor, TipoTransaccion, Fecha y para la tabla de hechos. En el entregable incluya una descripción general que permita comprender el proceso de implementación del ETL.

Note que para este proceso de negocio, las dimensiones Producto y Cliente son iguales a las del hecho Orden, este caso se conoce como dimensiones compartidas. Usted debe concentrarse en las dimensiones Proveedor, TipoTransaccion, Fecha y la tabla de hechos que no existen en la bodega de datos actualmente.

Es importante que mantenga las tablas creadas para que los tutores puedan validar las tablas en el momento de calificar su tarea.

WWImporters le comparte el modelo multidimensional que ha decido utilizar. Este modelo representa los movimientos(transacciones) que se hacen sobre el inventario de WWImporters. En particular, se observa que se tiene información de los tipos de transacciones, realizadas por un proveedor, relacionado con un cliente y un producto específico en una fecha determinada. En el modelo se dejan explícitos los dos tipos de identificadores que se crean a nivel de la bodega. el propio de la Bodega, (con el sufijo DWH) y el que viene de la fuente (con el sufijo T).



Sobre los resultados del entendimiento de datos, Wide World Importers les comenta lo siguiente:

* Cada fila representa una transacción o movimiento de productos en el inventario
* Los días de pago no pueden ser negativos no tiene sentido para nuestro negocio. Por favor corregir multiplicando los datos negativos por -1.
* Sobre la regla de negocio dada para la actividad de entendimiento de datos “La cantidad máxima de productos movidos es 50 millones por transaccion”, el negocio revisó y encontró que efectivamente gracias a los avances en su operación, ya puede mover más que la cantidad de 50 millones por transacción, por lo cual elimina esa regla de negocio.
* La falta de datos antes del 2014 es un error de extracción de datos. Los nuevos datos incluyen este año.
* Nuestro análisis concluye que la información que se ha duplicado totalmente no es útil. Por favor no tenerlos en cuenta.
* “El formato de fechas manejado es YYYY-MM-DD HH:MM:SS si tienen hora, minutos y segundos. De lo contrario el formato es YYYY-MM-DD”: En cuanto a formatos de fechas estamos de acuerdo con que los estandarizemos y el formato sea el especificado en la regla de negocio.
* Existen proveedores que tienen 2 filas una con un nombre y otra con el mismo nombre mas un “Inc” o “Ltd”. Unimos estos a un solo proveedor dado que se trató de un error de digitación.
* El código postal igual para todos nuestros proveedores es un error que también fue corregido.
* Cantidades negativas significan salidas de productos del inventario
* El negocio indica que las tablas de categoriasProveedores y TiposTransaccion fueron analizadas previamente, por su grupo de consultores.

**Recuerde**

* El diseño es fundamental y en particular el detalle de las transformaciones
* Dejar evidencia de las tablas creadas y pobladas en la base de datos que le fue asignada en el curso
* Entregar el notebook únicamente con la información solicitada en la tarea