**Integrantes**

Andrea Estefania Timaran Buchely

Laura Daniela Espinosa Maca

Carlos Enrique Jaramillo Aros

**Titulo**: Paso a Paso detallado ETL

**Objetivo**: Se busca presentar el detalle de cada paso en la ETL con el fin de explicar la razón y transformaciones.

Paso a Paso

De forma general, la ETL se divide en el tratamiento de la información de acuerdo a las fuentes de información recibidas

Diagrama, Esquemático

Descripción generada automáticamente

1. Gráfico

   Descripción generada automáticamenteFuente de información Archivo CSV

En primera instancia (Copy\_To\_BD) para realizar pruebas, se guardó una copia en la base de datos postgres, para poder cruzar la información y usar este método de validación de información, una vez completada la transformación.

Filtro rows, tiene como condicional revisar que el campo Year\_Remod/Add no sea nul, para después procesar.



En caso de que se cumpla la verdad de esta condicional, se pasa a ordenar y quitar espacios en blanco (TRIM\_nNulls y TRIM\_Nulls) de los campos que serán procesados con sus respectivas transformaciones.

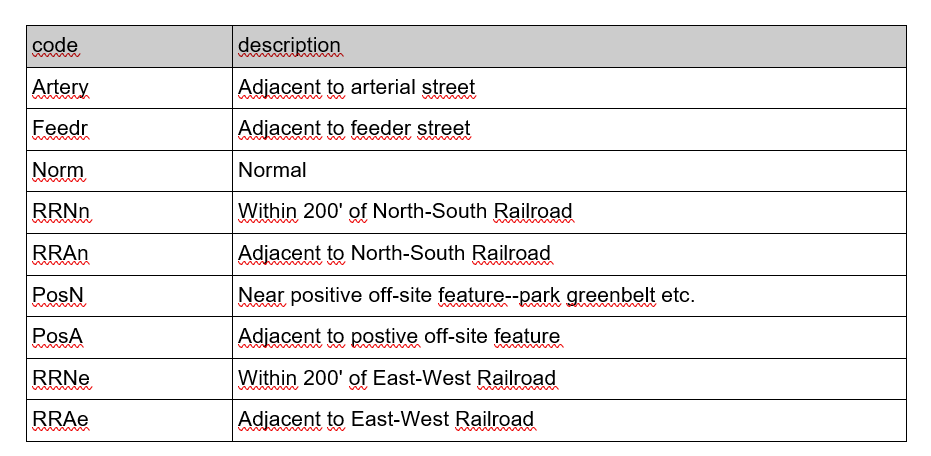
Tabla

Descripción generada automáticamente

Para las columnas, Condition\_1, Condition\_2 y LotShape, se realizaron los mapeos entregados como insumos para el proyecto (value mappers, condition\_1\_mapper, contition\_1\_mapper\_2,LotShapeMapper)

Condition 1 y Condition 2

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

LotShape

Texto

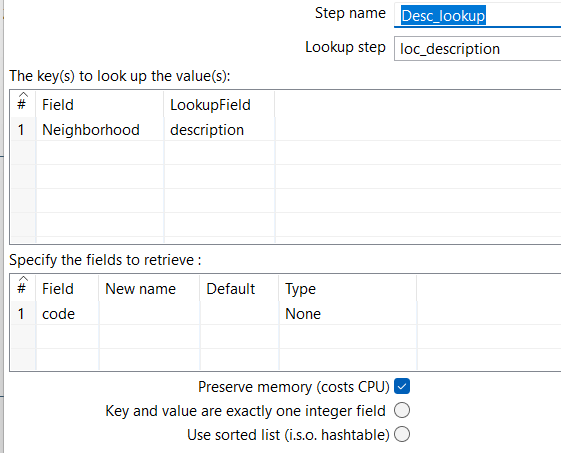
Descripción generada automáticamenteInterfaz de usuario gráfica, Tabla

Descripción generada automáticamente

Para el mapeo de la columna Neighborhood, se realizó la creación de una tabla con en la cual se insertó la información entregada como insumo, esta misma se uso para buscar (Desc\_Lookup) las descripciones y sacar las abreviaciones (Code)

Scripts:





Que fueron renombrados una vez se completó la actividad

Interfaz de usuario gráfica, Aplicación, Tabla

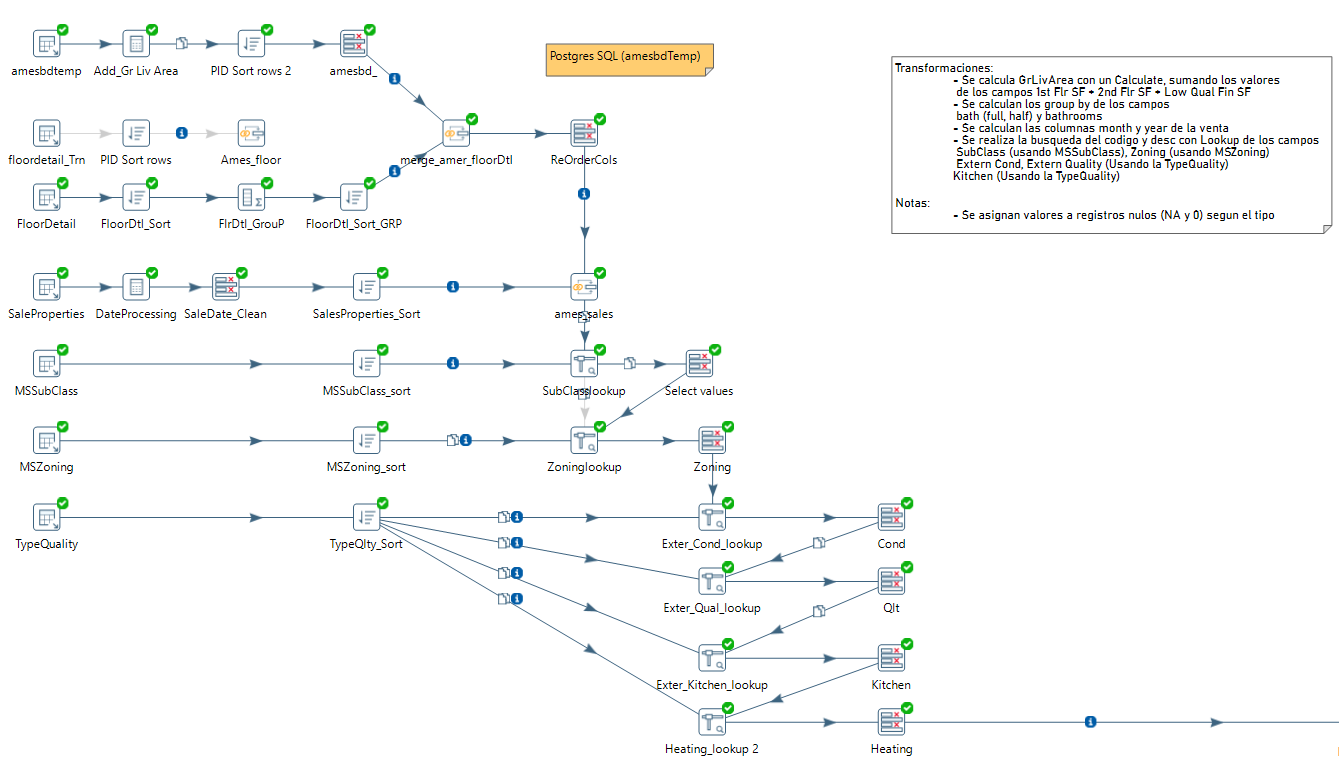
Descripción generada automáticamente

Como último paso de transformación para el archivo plano se validan los valores nulos según su tipo

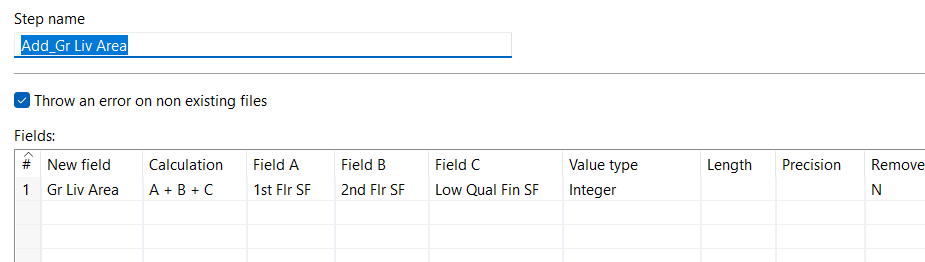
Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

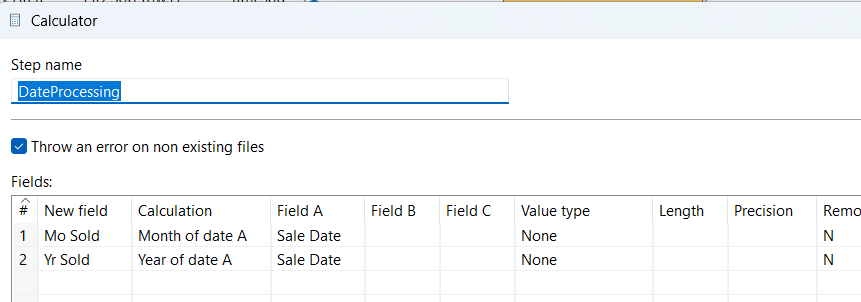
1. Fuente de Información (BD Postgres)



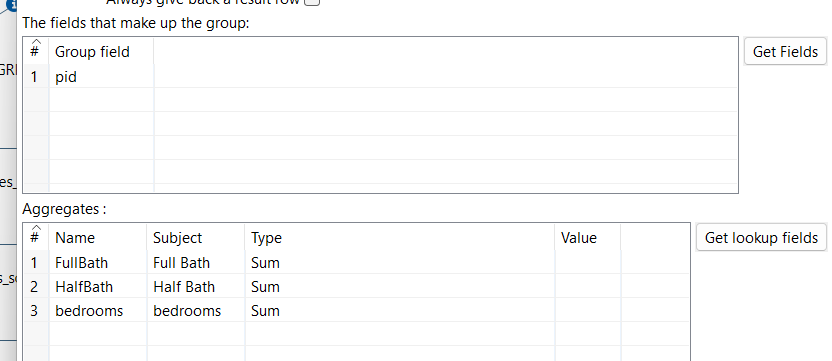
Se calcula el campo Gr Liv Area, sumando las columnas 1st Fir SF, 2nd Fir SF y Low Qual Fin SF



De igual manera, se calculan los campos Mo Sold y Yr Sold desde la columna Sale Date usando las funciones de la calculadora de PDI (Month of date A y Year of date A)



Se realiza la agrupación de baños completos (Full Bath), baños medios (Half Bath) y habitaciones (Bedroom) por cada casa (PID) sin considerar el Piso (Floor)



Todos estos ,se ordenan y se cruzan usando el campo PID

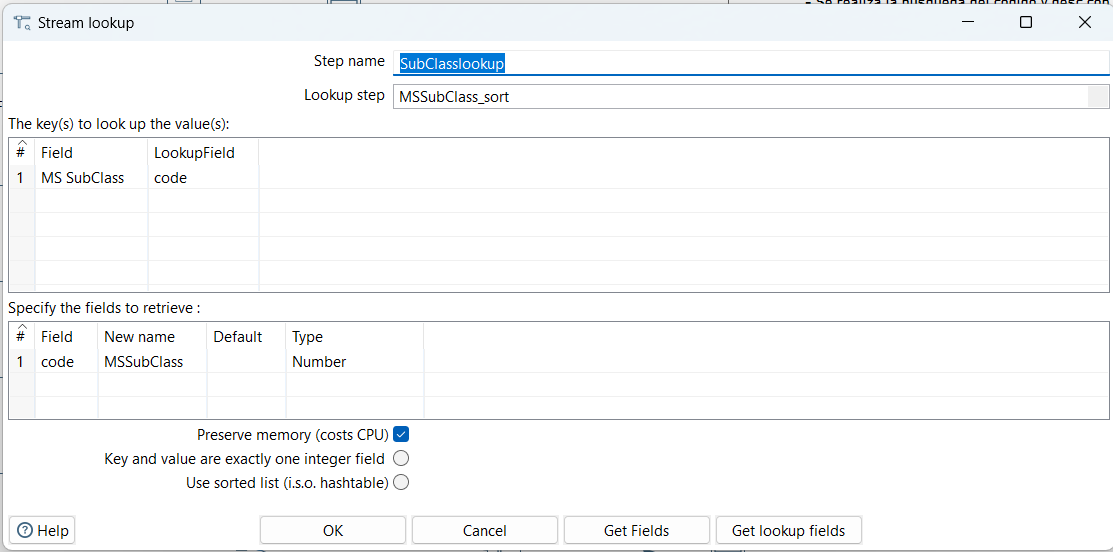
Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

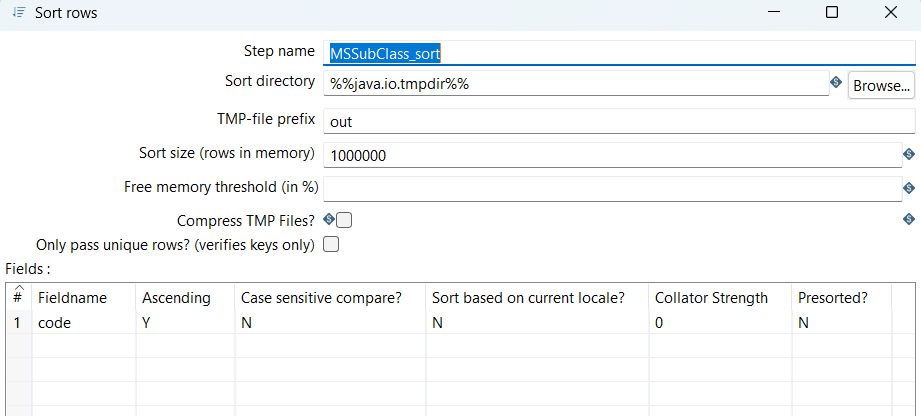
Descripción generada automáticamente

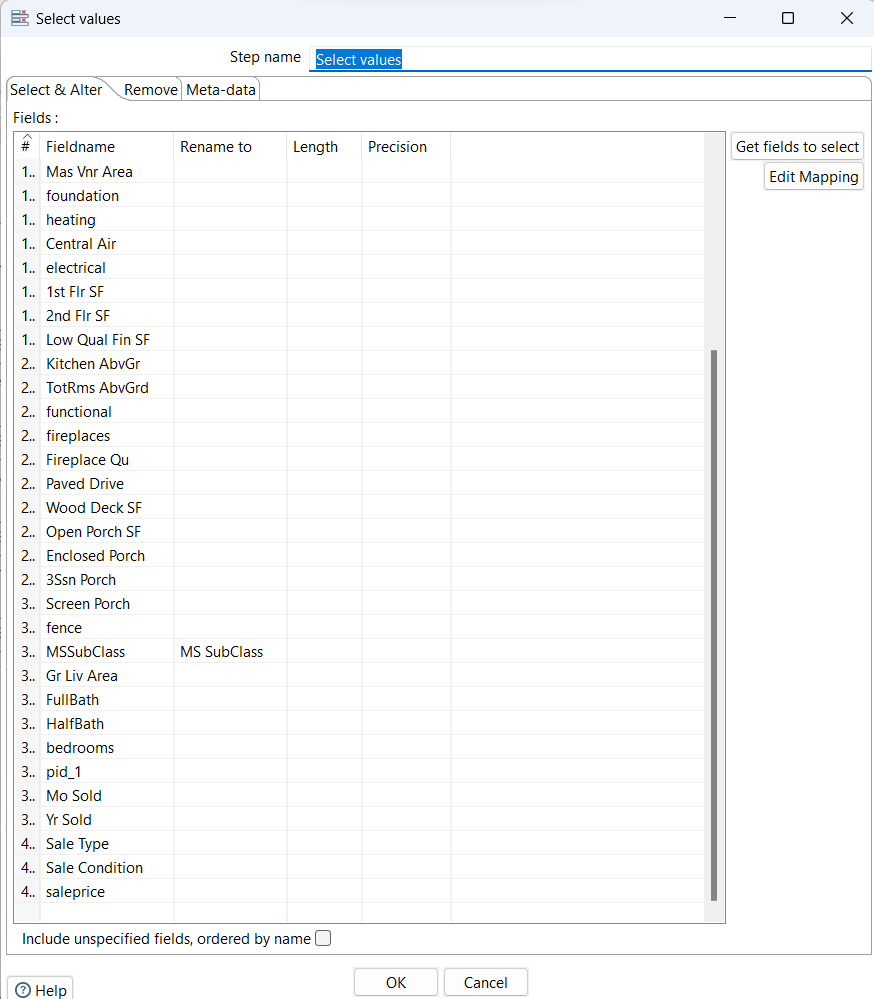
Una vez completo este cruce se pasa a realizar la búsqueda con Lookups de las descripciones y/o códigos de las columnas subclase y Zoning usand la columna que viene desde amesbdtemp contra los códigos en MSSubClass (code) y MSZoning (id), es importante mencionar que antes de realizar estas búsquedas es recomendable realizar sort (MSSubClass Sort y MSZoning Sort) para después realizar selección de variable y renombramiento, usando la columna buscada (LookupField) con el nuevo nombre y eliminando la anterior



Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente





De igual manera para las columnas que hablan de calidad de cocina, y exteriores, se realiza búsqueda contra la tabla TypeQuality en postgres

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

En estos pasos, se realiza búsqueda por las columnas correspondientes según el contexto

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Para Exter cond se usa el campo Exter Cond de la tabla origen (amesbdtemp) y id de typequality

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Para Exter Qual se usa el campo Exter Qual de la tabla origen (amesbdtemp) y id de typequality

Tabla

Descripción generada automáticamente

Para Kitchen Qual se usa el campo Kitchen Qual de la tabla origen (amesbdtemp) y id de typequality

Tabla

Descripción generada automáticamente

Para Heating QC se usa el campo Heating QC de la tabla origen (amesbdtemp) y id de typequality

A cada una de las columnas salientes (Field to retrive) de estas busquedas se da un nombre temporal que será reemplazado por el nombre de la columna inicial en amesbdtemp, tiendo asi la descripcion o código requerido para la entrega de información

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

1. Fuente de información (MongoDB)

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente con confianza media

Para este caso es importante mencionar que no se realiza conexión de base datos, sino que se consume la información recuperando el json que entrega la base de datos NoSQL

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Patrón de fondo

Descripción generada automáticamente con confianza baja

Interfaz de usuario gráfica, Texto, Aplicación

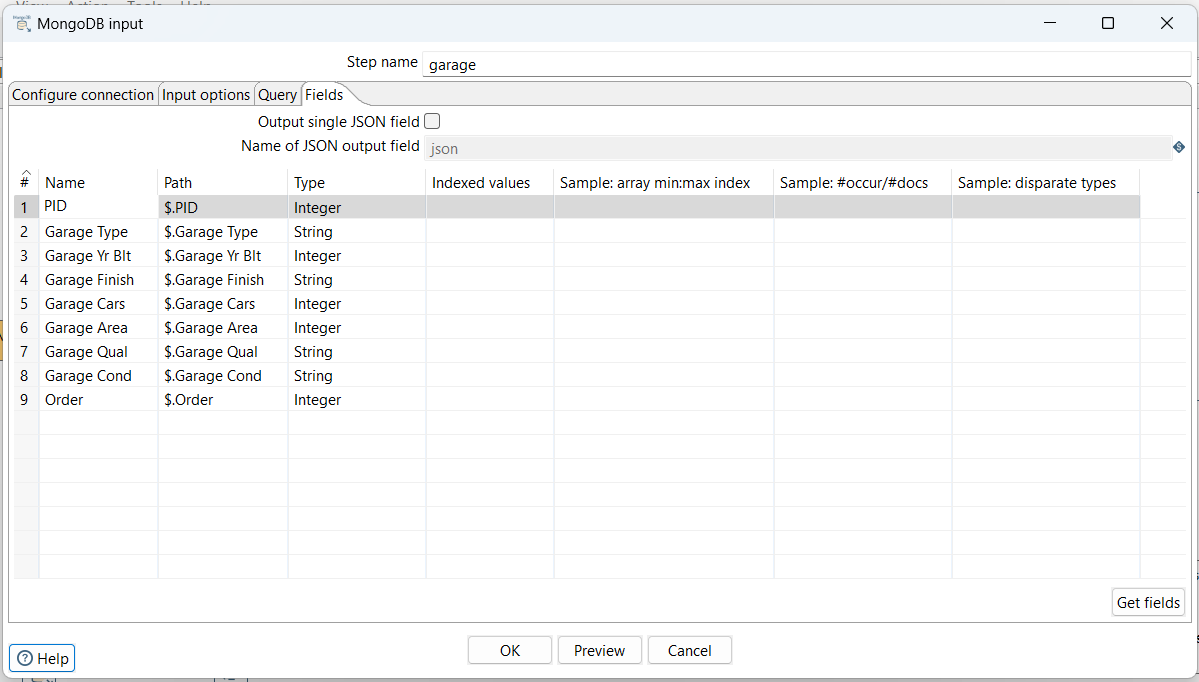
Descripción generada automáticamenteEn caso de haber condiciones para colocar (db.coll.find()) se colocan solo las condiciones, y no los sort o projections

Interfaz de usuario gráfica

Descripción generada automáticamenteComo ultima parte, dependiendo de la configuración deseada, se puede o no permitir la salida (Output singe JSON field) para dar salida en este formato, y asi procesar con un objeto JSON input y posterior procesamiento.

Para el caso de la ruta, se debe leer el roor ($) y después el nombre del campo, al igual que la asignación del tipo (Esto es muy importante)

Estos pasos los realizamos para cada colección de la bd nosql



Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Interfaz de usuario gráfica

Descripción generada automáticamente

Es importante mencionar que todos los cruces a realizar, deberán conservar la integridad de la fuente principal de información (Bsmt), por tanto deben tener left join para evitar perdida de informacion

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Una vez completo esto y seleccionadas las variables, en este paso, evito duplicidad de columnas y de informacion

Interfaz de usuario gráfica, Aplicación, Tabla, Excel

Descripción generada automáticamente

Se realiza la asingacion de valores en caso de ser nulo (NA para categorías y 0 para numéricas)

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

1. Merge de fuentes (USANDO PID)

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Previa a la finalizaicon de la ETL, realizamos el renombramiento de columnas tal como se requieren en el desitno

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

Y guardamos el archivo en el formato y contenido solicitado

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente