Almacenes de Datos - Proyecto Segundo Bimestre 2021A

Bitácora de Trabajo

Anotaciones

Actividad	*Fecha*	*Comentarios*
Familiarización con las actividades requeridas del proyecto	*11 de septiembre del 2021*	*Socialización de las tareas a realizar con el grupo, y análisis de los Datasets.*
Instalación de Pentaho Data Integration	*12 de septiembre del 2021*	*Se solucionó un problema con las variables de entorno de Java. Se instalo el conector de MySQL para Pentaho.*
Realización de las preguntas y necesidades del negocio que se desean responder con el DWH. Definición de dimensiones y métricas.	*13 de septiembre del 2021*	*Se tuvo problemas con algunas preguntas las cuales podrían dar origen a varios atributos con valores NaN.*
*Realización del diagrama **estrella**. Socialización sobre PDI para la realización de las transformaciones y tareas. Creación del DWH en MySQL*	*14 de septiembre del 2021*	*Se realizaron 6 dimensiones de las cuales se encontró problemas de resumen en una de ellas.*
Creación del DWH en MySQL y pruebas haciendo uso de PDI	*15 de septiembre del 2021*	*Se reinstala MySQL en todos los computadores de los integrantes y se solventan errores que presenta PDI, de la misma manera se verifica que todos tengan completa las instalaciones y sin problemas*
Uso de PDI para el poblado del DWH (Dimensiones) y análisis de los problemas de resumen.	*16 de septiembre del 2021*	*Problema con la agregación de campo extra para la identificación de preguntas o tipos.*
Migración de tabla de hechos	*18 de septiembre del 2021*	*Realizar los distintos joins para los registros de la tabla de hechos*
Realización del Informe	*19 de septiembre del 2021*	

Links empleados
Datasets descargados y extraídos con éxito: Hogares | (ecuadorencifras.gob.ec) https://www.ecuadorencifras.gob.ec/informacion-de-anos-anteriores-hogares/ https://www.youtube.com/watch?v=o7If1a-gkyI&list=PLPgjON4ZM0JBdxxDUAfCS84X79e_2CLNQ Documentation - Hitachi Vantara Lumada and Pentaho Documentation

Bitácora 11/09/2021

Actividades Realizadas:

- Leer y analizar los requisitos del proyecto.
- Socialización de las directivas del proyecto
- Análisis de los Datasets y diccionarios para determinar la información necesaria
- Conversión de codificación de CSV de latín-1 a UTF8 para CSV del año 2018
- Redacción de las Fuentes de datos del negocio en base a los CSVs ya convertidos

Propuestas:

-Análisis de los códigos de las variables para el año 2018-2019 y columnas para el tratamiento de las métricas necesarias

-En base al análisis anterior se procedió a tomar puntos clave para la redacción de las necesidades del negocio como son los problemas ambientales, uso de la electricidad y agua potable.

Tratamiento de los Dimensiones y atributos para reducir el número de columnas del dataset

- Área
- Ciudad
- Clasificación de residuos (s101p11, s101p12a-e)
- Eliminación de residuos (s101p2a-f)
- Tiempo
- Ubicación
- Problema ambiental

Problemas:

Tipo de Codificación del CSV del año 2018
Se encontró que para algunas columnas del dataset

Soluciones:

Para agilitar el proceso se va a investigar sobre herramientas en Pentaho que permitan el preprocesamiento de los datos.
Se realizo la con versión de Latin-1 a UTF8 en Excel

Observaciones

Ninguna

Bitácora 12/09/2021

Actividades Realizadas:

- Se procedió a realizar la descripción de las necesidades del negocio en base a la información del CSV y los tópicos más específicos de análisis
- Se estableció las operaciones y cálculos para obtener detalles de las costumbres de las personas encuestadas
- Identificación de las dimensiones y atributos de las fuentes de datos con sus respectivas jerarquías

Propuestas:

1) Mediante decisión de todos se procedió al filtrado de los tópicos necesarios para el análisis.
2) Para la descripción de las operaciones y cálculos se obtuvo mediante la información de la descripción de las fuentes de datos en base a los CSVs
3)En base al punto anterior se procedió a identificar las dimensiones, atributos y jerarquías de acuerdo con la información que hasta el momento se identificó.

Problemas:

En el archivo CSV y las variables de diccionario se tiene preguntas con distintas opciones las cuales se tomaron las necesarias para interés de análisis.

Soluciones:

Filtrar los campos que no se emplearan en el estudio.

Observaciones

Sin Observaciones

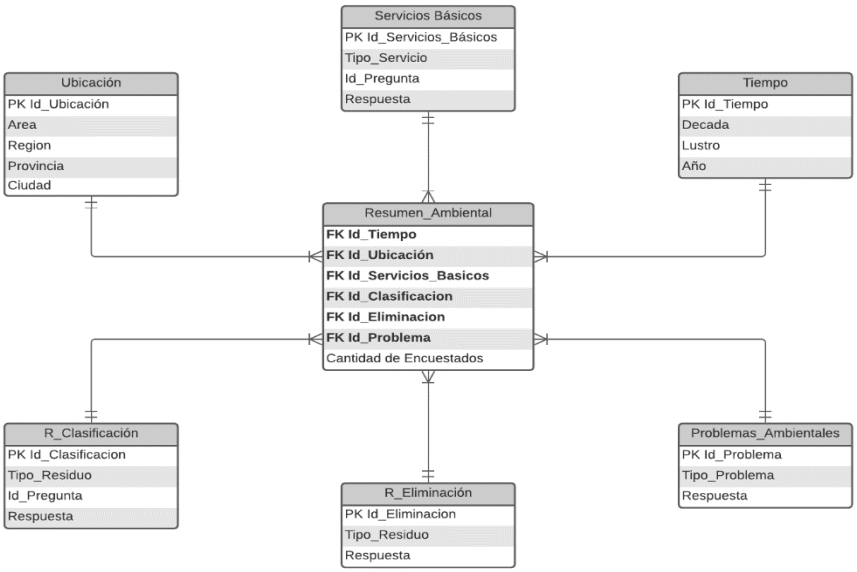
bitácora 13/09/2021

Actividades Realizadas:

- Descripción de las medidas o indicadores, fuente de datos y propiedades de agregación
- Diseño del esquema en nuestro caso estrella en Lucidchart para el DWH con el fin de obtener los requerimientos del negocio planteados
- Se Identificaron problema de resumen en base al desarrollo de nuestro esquema y explicar las respectivas soluciones.

Propuestas:

- En base a las dimensiones se detalló las medidas que se desea obtener para nuestro análisis
- Se diseño el esquema en estrella en base a las dimensiones ya planteadas



- Identificar posibles problemas en cada una de las tablas

Problemas:

- En la tabla Ubicación se tiene un problema debido a que no existe en ocasiones una ciudad especifica, registro guardado como otras en el atributo Ciudad, de lo cual se puede generar un problema de Roll Up al agregarlos ya sea a Provincia o Región.

Soluciones:

- Verificar cambios de preguntas del año 2018 a 2019 antes de desarrollar nuestro esquema en estrella.
- Esto se puede solucionar mediante la definición de un valor por defecto en dichos atributos.

Observaciones
Ninguna

bitácora 15/09/2021

Actividades Realizadas:

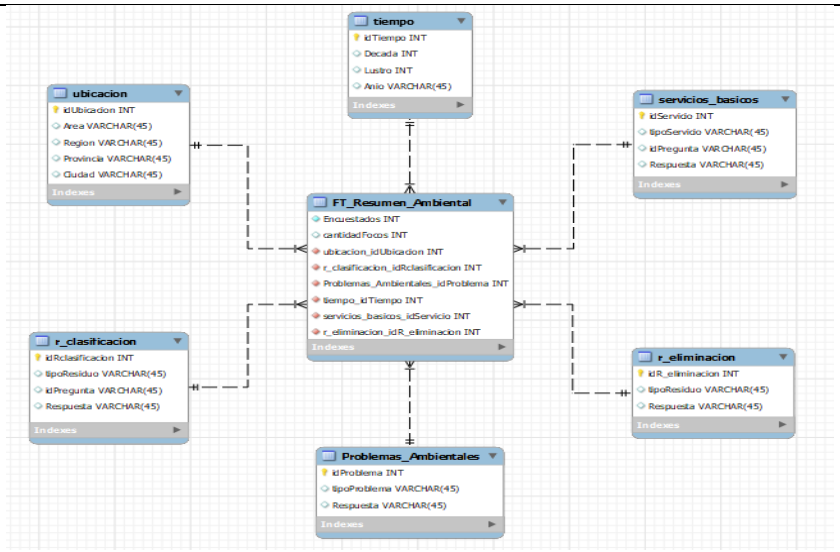
- Solución de Problemas en base a las variables de entorno de Pentaho ya que en varios computadores del grupo la aplicación no se ejecutaba correctamente
- Se instalo el conector a MySQL para poder conectarnos con la base de datos para poder crear nuestro Esquema
- Creación del DWH en MySQL
- Comprobación del funcionamiento de Pentaho y prueba de ciertas funcionalidades

Propuestas:

Revisión de documentación y búsqueda para poder instalar Pentaho de manera correcta

Buscar el conector hacia la base de datos y descargarlo e instalarlo y probar su funcionamiento con MySQL.

Creación del DWH en MySQL y lograr la Conexión de la base de datos para cargar nuestros archivos.



Problemas:

Problemas de complementos para el trabajo búsqueda de la versión de Pentaho que cuenten con los mismos.

Soluciones:

Averiguar el número de versión de Pentaho y descargar e instalar la que cuente con todas las herramientas que se necesita para trabajar

Observaciones
Ninguna

bitácora 16/09/2021

Actividades Realizadas:

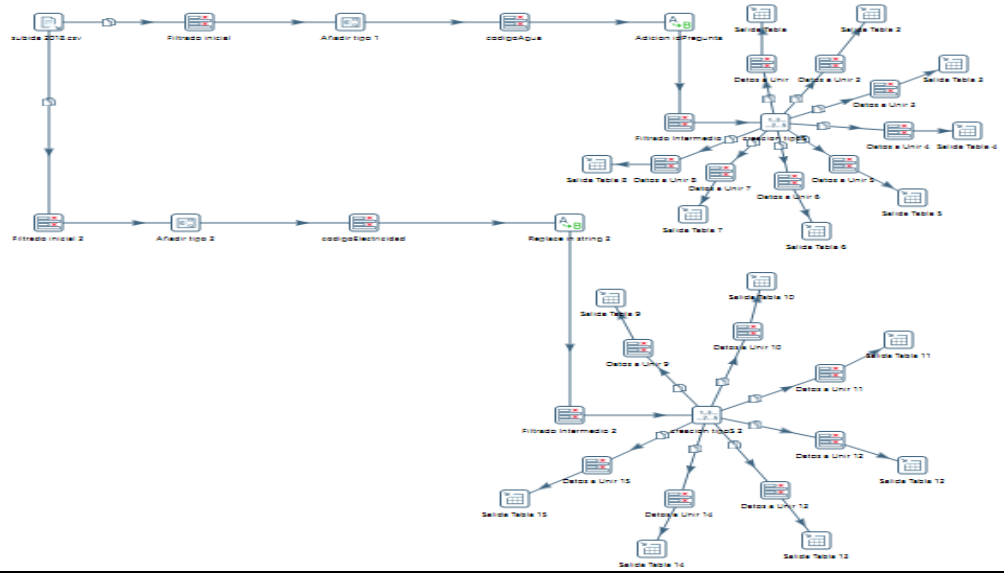
- Se empezó a realizar las transformaciones para poblar la DWH

Propuestas:

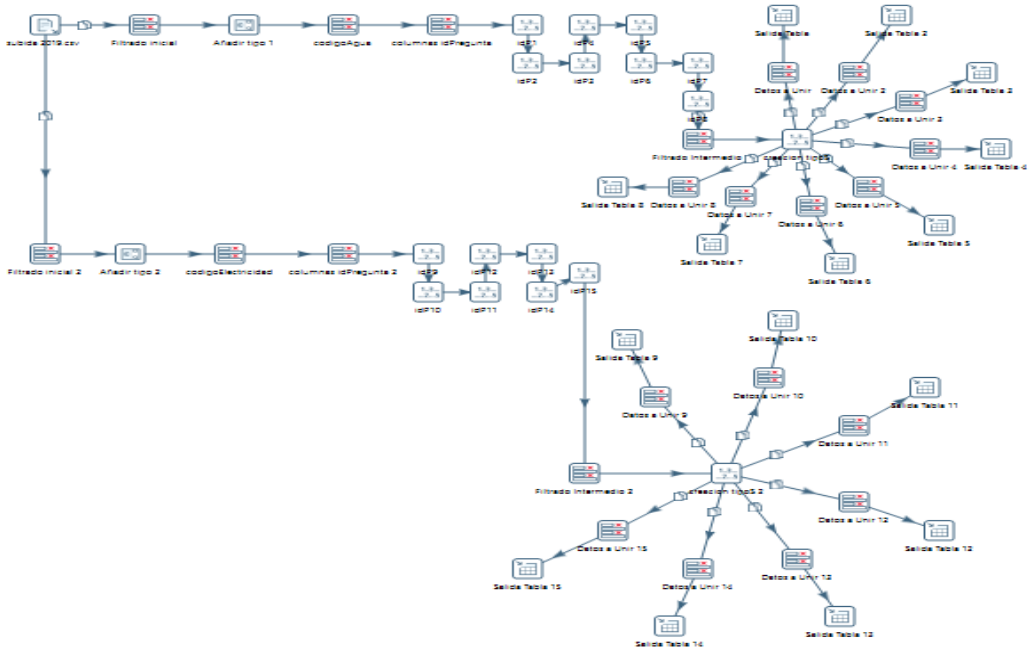
Utilizando spoon de PDI se empezó a utilizar los distintos pasos para la transformación de los datos con la ayuda de la documentación de estos.

Desarrollo:

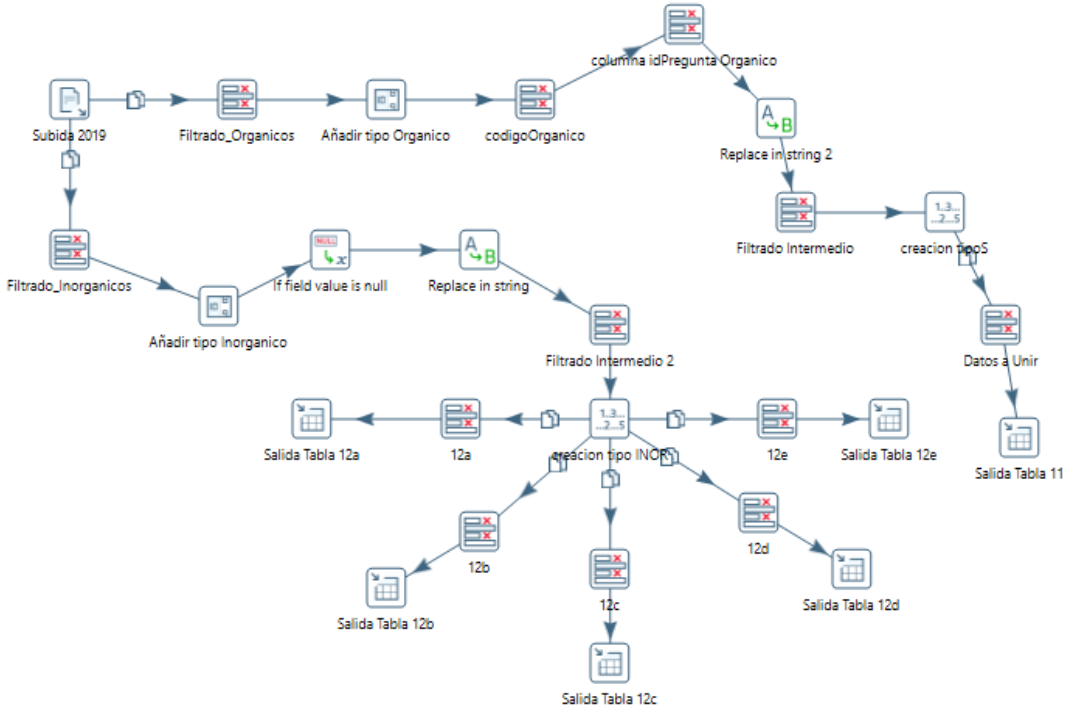
Transformación para la Dimensión Servicios Básicos (2018)



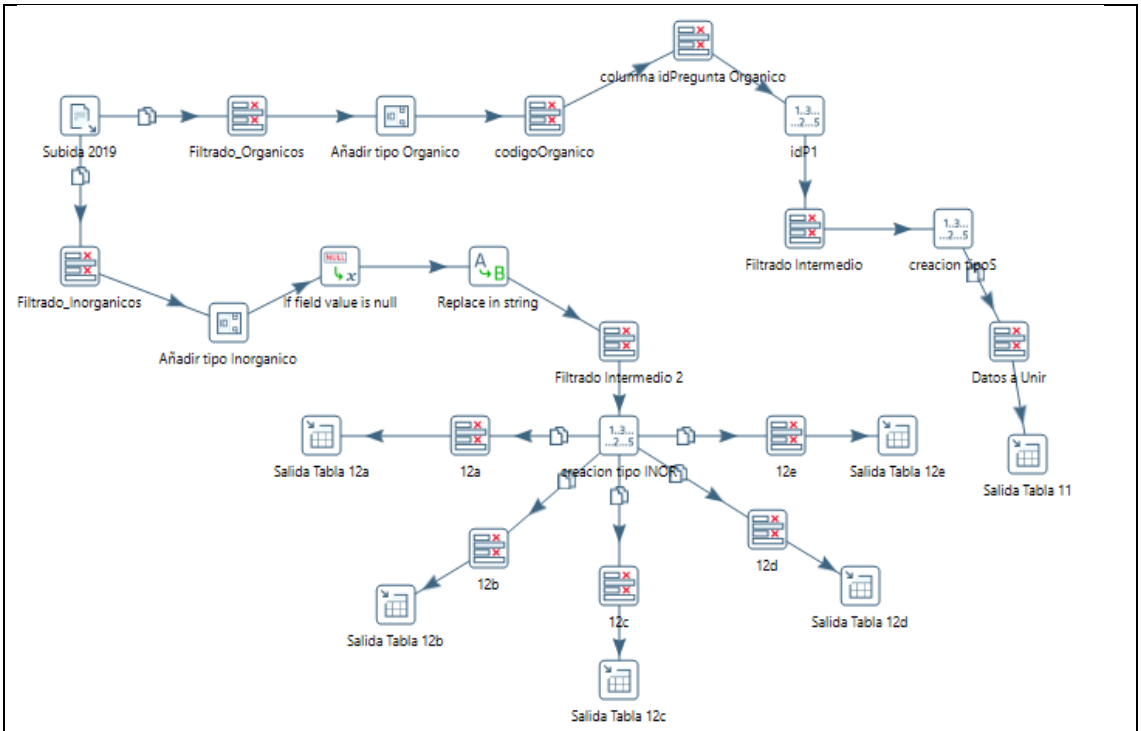
Transformación para la Dimensión Servicios Básicos (2019)



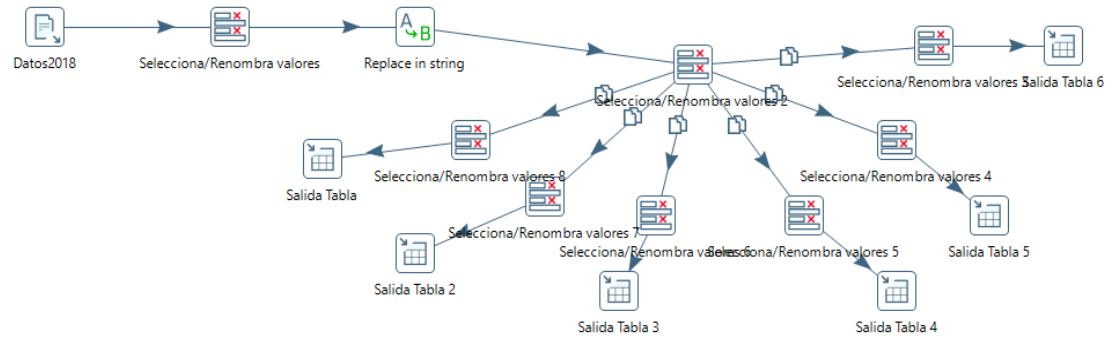
Transformación para la Dimensión Clasificación de Residuos (2018)



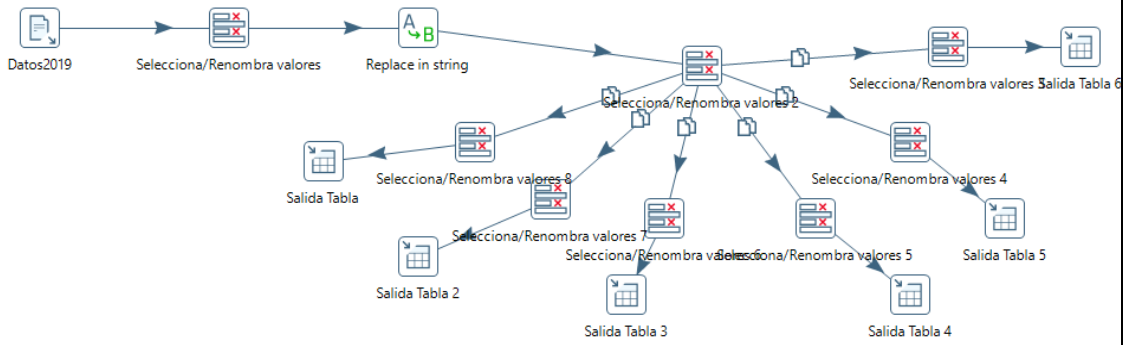
Transformación para la Dimensión Clasificación de Residuos (2019)



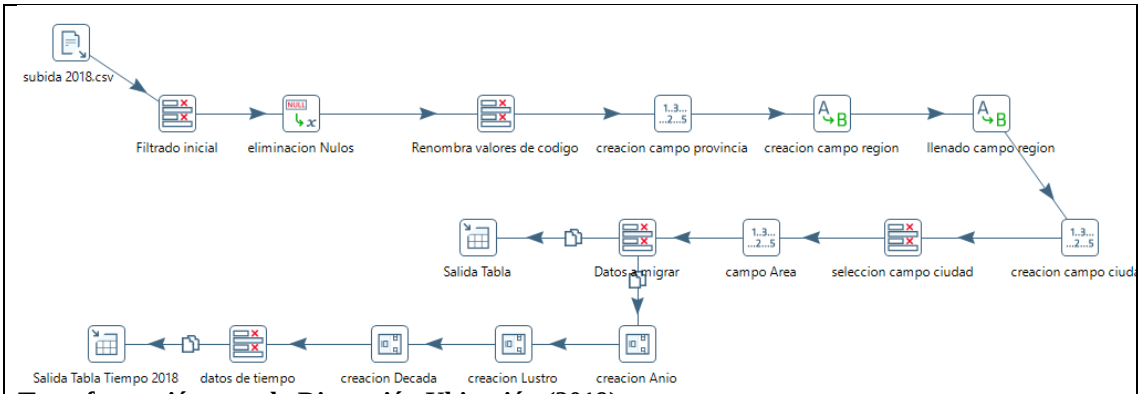
Transformación para la Dimensión Eliminación de Residuos (2018)



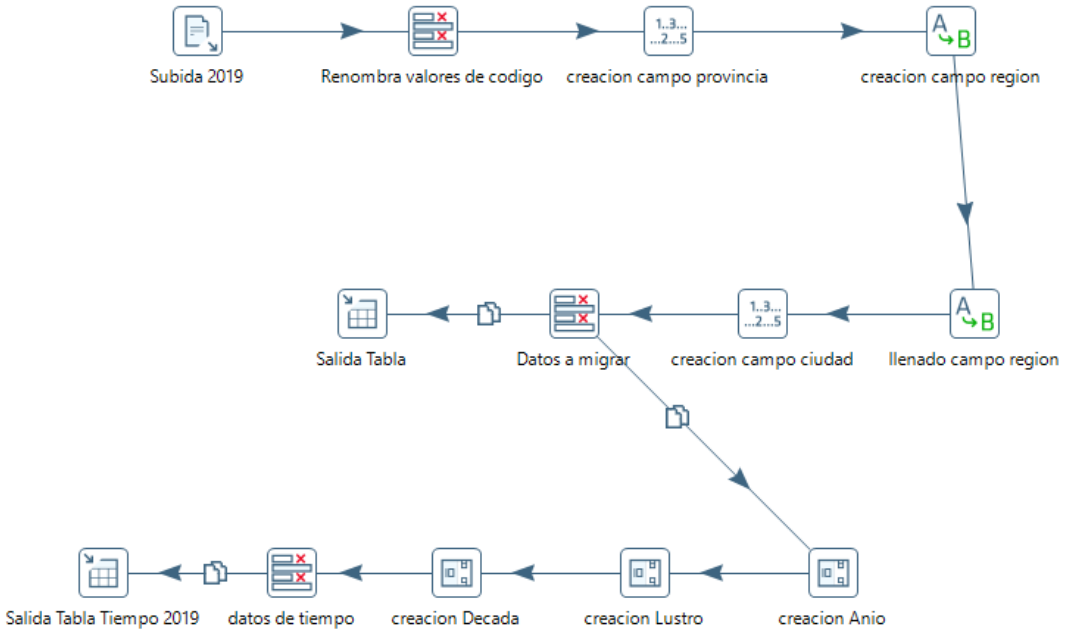
Transformación para la Dimensión Eliminación de Residuos (2019)



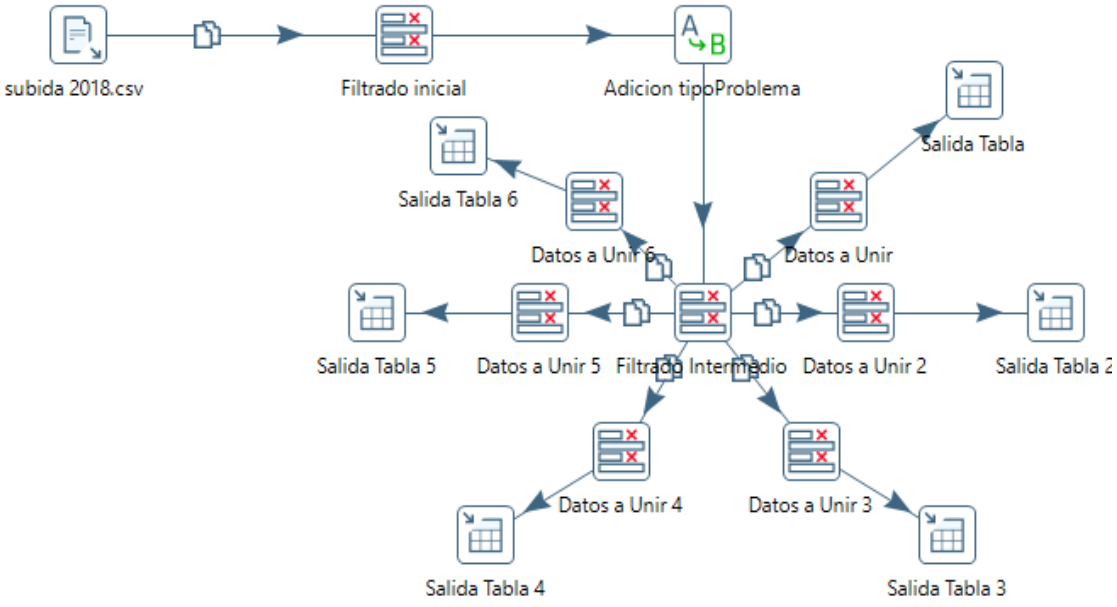
Transformación para la Dimensión Ubicación (2018)



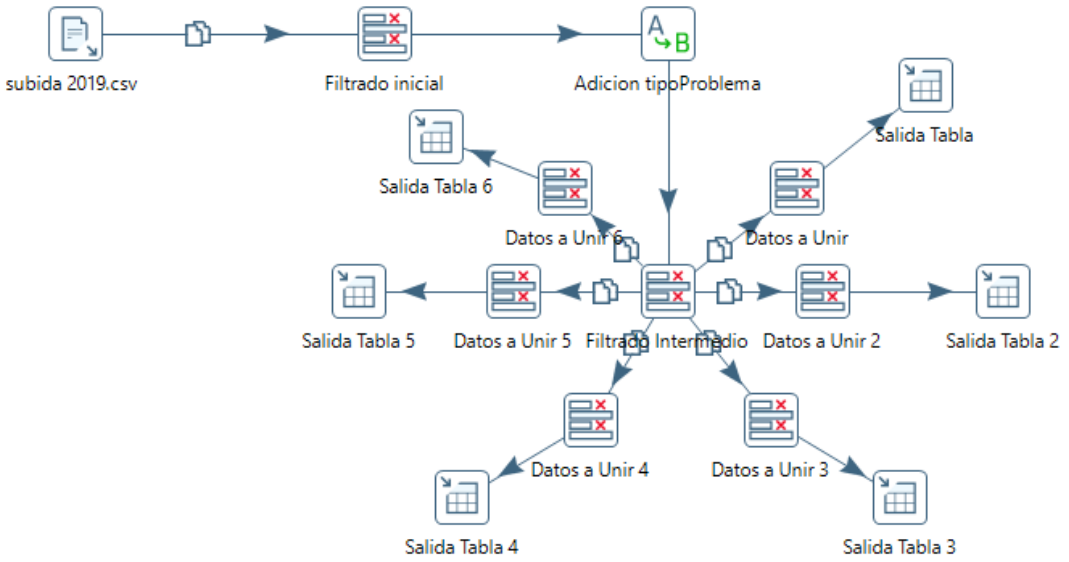
Transformación para la Dimensión Ubicación (2019)



Transformación para la Dimensión Problema Ambiental (2018)



Transformación para la Dimensión Problema Ambiental (2019)



Problemas:
Para poder utilizar el paso reemplazar en String se tuvo que cambiar el tipo de datos de algunas columnas ya que se encontraban en formato de entero y las preguntas se iban a trabajar como cadena de strings.

Soluciones:
Se cambio el tipo de datos en el paso de CSV file input ya que nos muestra todas las cabeceras de los campos con sus respectivos tipos evitando un paso adicional que al inicio se realizó.

Observaciones

Comprobar que las transformaciones al CSV no radique en pérdidas de datos

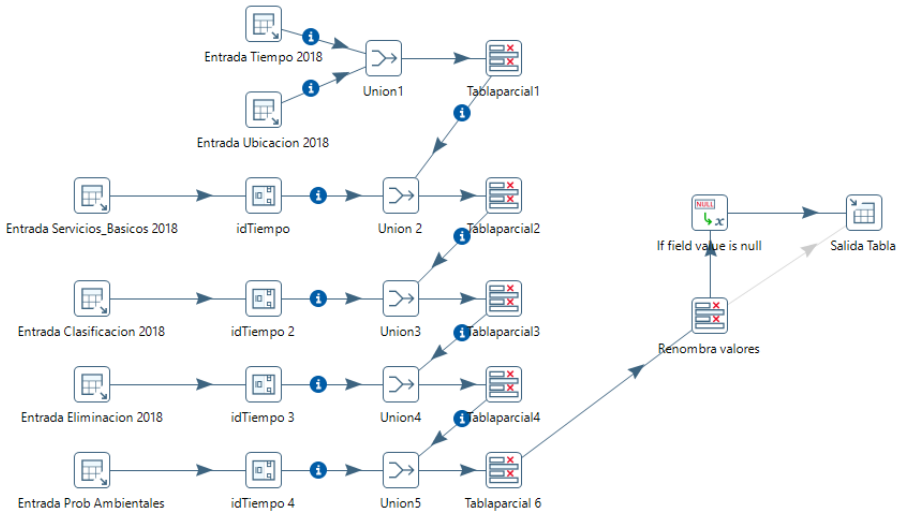
Bitácora 17/07/2021-18/07/2021

Actividades Realizadas:

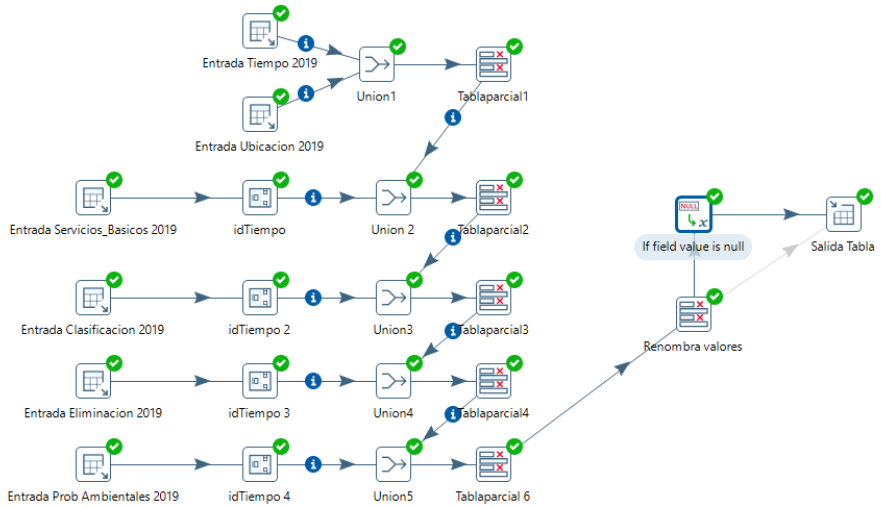
- *Comprobación de que no exista perdidas tras los procesos de transformación en los respectivos CSV*
- *Verificación y poblado de las dimensiones de la base de datos de MySQL*
- *Poblado de la tabla de Hechos.*

Propuestas:

- Se decidió importar desde las tablas de dimensiones los registros para la tabla de hechos.
  Población de la tabla de hechos para registros del 2018



Población de la tabla de hechos para registros del 2019



Problemas:

- El problema corresponde al concatenado de la tabla de hechos por el numero de registros y la cantidad de nulos que se generan debido a esto:

FK_Tiempo	FK_Ubicacion	FK_ServBas	FK_Clasif	FK_Elimina	FK_ProbAmb	Cant_Encuestados
1-12368 (2018)	1-12368 (2018)	1-185520 (2018)	1-74208 (2018)	1-74208 (2018)	1-74208 (2018)	12368
12369-23896 (2019)	12369-23896 (2019)	185521 - 358440 (2019)	74209-143376 (2019)	74209-143376 (2019)	74209-143376 (2019)	11528

idTiempo (2018)	Id de preguntas o campos que generaron mayor numero de registros que la tabla tiempo.
1,2,3…12368	1,2,3…12368 --> s101p1a
	12369…24736--> s101p1b
	24736…37104
	…

Soluciones:

- Se agregaron varias secuencias de datos que simulen identificadores de tiempo para el merging de la tabla de Hechos, no solventado al momento.

Observaciones

La cantidad de nulos dentro de las tablas se resolvió con el ingreso de un valor 1 para el año 2018 y 2 para el año 2019. Los cuales en las filas donde exista 4 valores 1 u 2 corresponde a registros con un solo valor real, rodeado por nulos.

bitácora 19/07/2021

Actividades Realizadas:

- Una vez poblada la base de datos, y realizados los procesos pedidos en la hoja guía del proyecto, se finaliza con la realización del informe

Observaciones

- Ninguna