Escuela Politécnica Nacional

Almacenes de Datos GR1

Proyecto del segundo bimestre 2021A

Integrantes:

- Alfaro Jefferson
- Gualoto Johnny
- Moreira Erick
- Oña Luis

1. Objetivos

- Diseñar un almacén de datos (DWH) que cumpla con un conjunto requisitos de negocio.
- Analizar y detectar potenciales problemas en las jerarquías de atributos de una dimensión y en las relaciones dimensión-hecho.
- Usar las herramientas ETL de Pentaho Data Integration (PDI) para extraer y transformar datos de diversas fuentes e integrarlo en un almacén de datos.
- Trabajar cooperativamente en un grupo de compañeros de la clase de AD para desplegar un almacén de datos.
- Desarrollar habilidades de comunicación oral y escrita mediante la escritura de un informe del proyecto y su exposición.

2. Fuentes de Datos

Como Fuentes de Datos se tiene dos archivos CSV con datos de muestra que fueron recopilados por el Instituto Nacional de Estadísticas y Censos quien es el responsable de la estadística oficial encargada de planificar, normar y certificar la producción del Sistema Estadístico Nacional, además de producir información estadística pertinente, oportuna y de calidad; e innovar en metodologías, métricas y análisis de información estadística necesaria para el diseño, implementación y evaluación de la planificación nacional. Se tiene información sobre las acciones del buen uso y manejo de agua, energía y residuos sólidos, con el fin de analizar las prácticas de uso de los recursos naturales tanto en niveles individuales y comunitarios en los años 2018 - 2019. Buscando aplicar conceptos de patrones de diseño de base de datos multidimensionales, basados en las fuentes de datos antes mencionadas, cuyo objetivo principal es definir un modelo multidimensional destacando sus dimensiones atributos y medidas, así como la granularidad de la tabla de hechos, como otro objetivo del presente trabajo es crea un almacén de datos (DWH) que integre los datos de nuestras dos fuentes de datos además de identificar posibles problemas de resumen al momento de implementar nuestro esquema y finalmente poblar nuestras tablas del almacén con los registros de los archivos consultados.

A continuación, se describen los archivos csv:

Para empezar, se tiene una descripción generalizada de los dos CSV seleccionados, como se mencionó anteriormente se tiene para la parte del año 2018 -2019.

CODIGO DE LA VARIABLE 2018	CODIGO DE LA VARIABLE 2019	FORMATO DEL DATO

área	área	Numérico
ciudad	ciudad	Numérico
conglomerado	conglomerado	Numérico
vivienda	vivienda	Numérico
hogar	hogar	Numérico
S10P1	s101p11	Categórico
S10P1A	s101p12a	Categórico
S10P1B	s101p12b	Categórico
S10P1C	s101p12c	Categórico
S10P1D	s101p12d	Categórico
S10P1E	s101p12e	Categórico
S10P2A	s101p2a	Categórico
S10P2B	s101p2b	Categórico
S10P2C	s101p2c	Categórico
S10P2D	s101p2d	Categórico
S10P2E	s101p2e	Categórico
S10P2F	s101p2f	Categórico
S10P3	s101p3	Categórico
S10P3A	s101p3a	Numérico
S10P3B	s101p3b	Numérico
S10P4	s101p4	Categórico
S10P4A	s101p4a	Numérico
S10P4B	s101p4b	Numérico
S10P4B1	s101p4b1	Numérico
S10P5A	s101p5a	Categórico
S10P5B	s101p5b	Categórico
S10P5C	s101p5d	Categórico
S10P5D	s101p5e	Categórico
S10P5E	s101p5f	Categórico
S10P5F	s101p5g	Categórico
S10P5G	s101p51a	Categórico
	s101p52a	Numérico
	s101p51b	Categórico
	s101p52b	Numérico
	s101p51c	Categórico
	s101p52c	Numérico
	s101p51d	Categórico
	s101p52d	Numérico
	s101p51e	Categórico
	s101p52e	Numérico
	s101p51f	Categórico
	s101p52f	Numérico
	s101p51g	Categórico

	s101p52g	Numérico
	s101p51h	Categórico
	s101p52h	Numérico
	s101p51i	Categórico
	s101p52i	Numérico
	s101p51j	Categórico
	s101p52j	Numérico
	s101p51k	Categórico
	s101p52k	Numérico
S10P6A	s101p61	Categórico
S10P6B	s101p62	Categórico
S10P6C	s101p63	Categórico
S10P6D	s101p64	Categórico
S10P6E	s101p65	Categórico
S10P6F	s101p66	Categórico
S10P6G	s101p67	Categórico
S10P6H	s101p68	Categórico
S10P7A	s101p71	Categórico
S10P7B	s101p72	Categórico
S10P7C	s101p73	Categórico
S10P7D	s101p74	Categórico
S10P7E	s101p75	Categórico
S10P7F	s101p76	Categórico
S10P7G	s101p77	Categórico
	s101p71a	Categórico
S10P8	s101p8	Categórico
S10P9A	s101p9a	Categórico
S10P9B	s101p9b	Categórico
S10P9C	s101p9c	Categórico
S10P9D	s101p9d	Categórico
S10P10	s101p10	Categórico
S10P11A1	s101p111	Numérico
S10P11A1_1	s101p11111	Categórico
S10P11B2	s101p112	Numérico
S10P11B2_1	s101p11112	Categórico
S10P11C3	s101p113	Numérico
S10P11C3_1	s101p11113	Categórico
S10P11D4	s101p114	Numérico
S10P11D4_1	s101p11114	Categórico
S10P11E5	s101p115	Numérico
S10P11E5_1	s101p11115	Categórico
S10P12A	s101p121	Categórico
S10P12B	s101p122	Categórico

S10P12C	s101p123	Categórico
S10P12D	s101p124	Categórico
S10P12E	s101p125	Categórico
S10P12F	s101p126	Categórico
S10P13	s101p13	Categórico
S10P14A	s101p141	Categórico
S10P14B	s101p142	Categórico
S10P14C	s101p143	Categórico
S10P14D	s101p144	Categórico
S10P17A	s101p171	Categórico
S10P17B	s101p172	Categórico
S10P17C	s101p173	Categórico
S10P17D	s101p174	Categórico
S10P17E	s101p175	Categórico
S10P17F	s101p176	Categórico
S10P18A	s101p181	Categórico
S10P18B	s101p182	Categórico
S10P18C	s101p183	Categórico
S10OBS	s101obs	
	upm	Numérico
	fexp4	Numérico
	estrato4	Numérico
	id_hogar	Numérico

De los CSV seleccionados, se observa que en algunos se encuentran campos numéricos, por lo cual es necesario la visualización del diccionario de variables para identificar en que rangos se encuentran para la transformación de datos en Pentaho.

Las ciudades tendrán la siguiente codificación

Cod_Ciudad	Ciudad
10150	Cuenca
170150	Quito
90150	Guayaquil
70150	Machala
180150	Ambato

De la misma manera es importante la aclaración de ciertos códigos que deben entenderse como tal en el data set, sin embargo, para nuestro estudio únicamente los siguientes campos serán seleccionados

CODIGO DE LA VARIABLE 2018	CODIGO DE LA VARIABLE 2019	FORMATO DEL DATO

área	área	Numérico
ciudad	ciudad	Numérico
S10P1	s101p11	Categórico
S10P1A	s101p12a	Categórico
S10P1B	s101p12b	Categórico
S10P1C	s101p12c	Categórico
S10P1D	s101p12d	Categórico
S10P1E	s101p12e	Categórico
S10P2A	s101p2a	Categórico
S10P2B	s101p2b	Categórico
S10P2C	s101p2c	Categórico
S10P2D	s101p2d	Categórico
S10P2E	s101p2e	Categórico
S10P2F	s101p2f	Categórico
S10P6A	s101p61	Categórico
S10P6B	s101p62	Categórico
S10P6C	s101p63	Categórico
S10P6D	s101p64	Categórico
S10P6E	s101p65	Categórico
S10P6F	s101p66	Categórico
S10P6G	s101p67	Categórico
S10P6H	s101p68	Categórico
S10P7A	s101p71	Categórico
S10P7B	s101p72	Categórico
S10P7C	s101p73	Categórico
S10P7D	s101p74	Categórico
S10P7E	s101p75	Categórico
S10P7F	s101p76	Categórico
S10P7G	s101p77	Categórico
S10P12A	s101p121	Categórico
S10P12B	s101p122	Categórico
S10P12C	s101p123	Categórico
S10P12D	s101p124	Categórico
S10P12E	s101p125	Categórico
S10P12F	s101p126	Categórico

Esto se debe a que estas preguntas son las más representativas en cuanto a los datos que el equipo de trabajo definido como parámetros para el almacén, representan información en cuanto a análisis de problemas ambientales, uso debido de agua y luz, clasificación de residuos y su forma de eliminación.

3. Necesidades del Negocio

El objetivo principal para realizar el almacén de datos es realizar un monitoreo acerca de la información sobre clasificación y eliminación de residuos, Problemas ambientales en

el vecindario, y uso de electricidad y agua potable en distintas áreas y ciudades del Ecuador a través de los años. Las medidas o indicadores que permitirán la valorización son el número de personas encuestadas que cumplen con ciertas prácticas ambientales con relación al manejo de residuos sólidos y prácticas de uso de servicios básicos, además de su percepción de los problemas en su entorno, los métodos más y menos empleados para eliminación de desechos por los encuestados. El almacén de datos debe permitir el análisis continuo de informes. Sumarizando lo más importante **por encuestado anual,** relacionado a **servicios básicos, manejo de residuos y problemas ambientales.**

Las operaciones y cálculos para obtener un mayor detalle acerca de las costumbres de los encuestados sobre los factores ya mencionados son:

- El promedio de personas anual que si clasifican residuos orgánicos.
- La cantidad de personas con buenas prácticas de eliminación de residuos.
- Los problemas ambientales más y menos frecuentes en las ciudades.
- Porcentaje de personas que no clasifican sus residuos sobre la totalidad de encuestados.

4. Problemas

a) Identificar las dimensiones y atributos, mapear los atributos y las fuentes de datos, y, finalmente, especificar la jerarquía de atributos de las dimensiones que conformarán el Almacén de Datos (DWH).

Dimensión: Servicios Básicos

- **Id_Servicios (A):** Atributo generado para el modelado multidimensional.
- **Tipo_Servicio (A):** Atributo generado para el modelado multidimensional.
- **Id_Pregunta (A):** Atributo obtenido del archivo CSV (s101p61, s101p62, s101p63, s101p64, s101p65, s101p66, s101p67, s101p68, s101p71, s101p72, s101p73, s101p74, s101p75, s101p76, s101p77).
- **Respuesta(A):** Atributo obtenido del archivo CSV (s101p61, s101p62, s101p63, s101p64, s101p65, s101p66, s101p67, s101p68, s101p71, s101p72, s101p73, s101p74, s101p75, s101p76, s101p77).

Dimensión: R_Clasificación

- **Id_EncuestadoC (A):** Atributo generado para el modelado multidimensional.
- **Tipo_Residuo (A):** Atributo generado para el modelado multidimensional.
- **Id_Pregunta (A):** Atributo generado para el modelado multidimensional, se genera a partir del atributo (s101p11) del CSV y la combinación de los atributos (s101p12a-e).
- **Respuesta (A):** Atributo obtenido del archivo CSV.

Dimensión: R_Eliminación

- **Id_Eliminacion (A):** Atributo generado para el modelado multidimensional.
- **Tipo_Residuo (A):** Atributo obtenido del archivo CSV (Orgánicos, Papel/Cartón, Plástico, Vidrio, Metal, Tetrapak).
- **Respuesta (A):** Atributo obtenido del archivo CSV (Respuesta numérica del 1 - 8).

Dimensión: Tiempo

- **Id_Tiempo(A):** Atributo generado para el modelado multidimensional.

- **Año (A):** Atributo generado para el modelado multidimensional (por año de cada CSV). Posible Jerarquía a futuro: Década→ Lustro→ Año.

Dimensión: Ubicación

- **Id_Ubicación (A):** Atributo generado para el modelado multidimensional.
- **Área (A):** Atributo obtenido del archivo CSV (Area).
- **Ciudad (A):** Atributo obtenido del archivo CSV (Ciudad). Jerarquía: Región → Provincia → Ciudad.

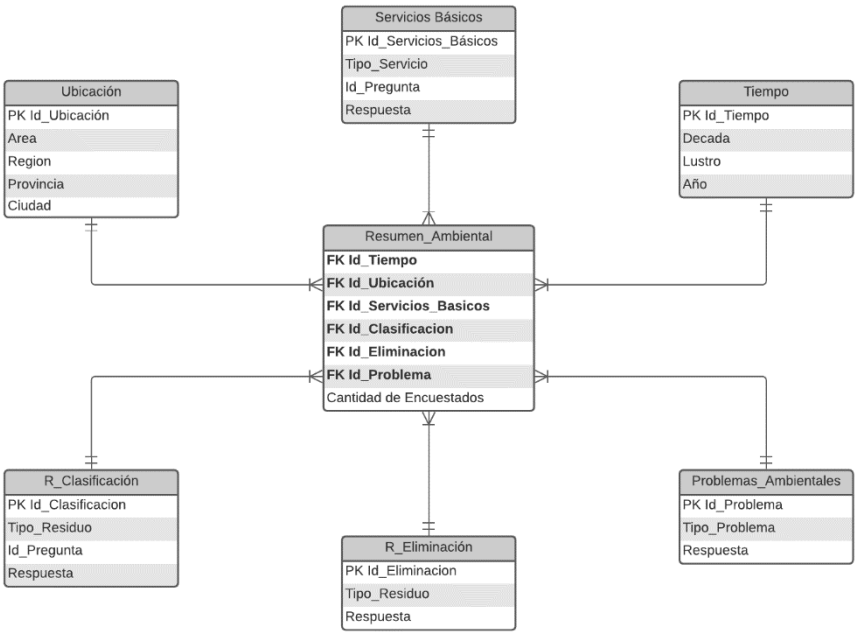
Dimensión: Problema Ambiental.

- **Id_Problema (A):** Atributo generado para el modelado multidimensional.
- **Tipo_Problema (A):** Atributo obtenido del archivo CSV (contaminación visual, Agua contaminada, ruidos excesivos, acumulación de basura, contaminación del aire, presencia de animales callejeros).
- **Respuesta (A):** Atributo obtenido del archivo CSV.

b) Especificar las medidas o indicadores, fuentes de datos y propiedades de la agregación de medidas de la tabla de hechos de DWH.

Cantidad de encuestados (M): Obtenido de CSV (cualquier tipo de pregunta del CSV). Medida Aditiva.

c) Diseñar un esquema en estrella del DWH (o el adecuado) para permitir responder a las preguntas planteadas.



d) Identificar los problemas de resumen en su esquema y explicar las soluciones para resolverlos.

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. Esto se puede solucionar mediante la definición de un valor por defecto en dichos atributos.

5. Pentaho Data Integration

a) Usando Spoon de PDI, implementar las transformaciones y tareas (jobs) necesarias para poblar el DWH (se debe resolver los problemas de resumen identificados).

Se adjuntan graficas de las transformaciones en la bitácora del 16/09/2021

6. Conclusiones y Recomendaciones

a) Conclusiones

- Gracias a Spoon de PDI se pudo realizar la extracción de las fuentes de datos para realizar las respectivas transformaciones sin la necesidad de programar el código y así poblar nuestra DWH mediante pasos sencillos.
- Se evidencio que el diseño correcto de un DWH permite establecer los parámetros y establece las acciones sobre las cuales debe llevarse a cabo los procesos de transformación de un archivo de datos.
- Se comprobó que PDI Spoon es muy robusto y capaz de procesar datos de distintas fuentes para su tratamiento y migración a un DWH sin el uso de IDES para el preprocesamiento de los datos en lenguajes de programación o en varias ocasiones la implementación de lenguaje SQL.

b) Recomendaciones

- Para lograr la conexión al momento de configurar las salidas de las tablas en cada una de las transformaciones como primer paso se debe comprobar si se tiene instalado el conector para MySQL caso contrario no podremos establecer conexión hacia la base de datos
- Para Lograr instalar correctamente el conector en Pentaho se debe verificar la versión con la que nos encontramos trabajando ya que de una versión a otro puede diferir las configuraciones del conector.
- Al momento de configurar las conexiones de las tablas hacia la base de datos se deben tomar en cuenta el nombre de las columnas con las que creamos el diseño en la base de datos.