**Nombres:** Nicolás Andrés Tobo Urrutia (201817465), Vilma Marcela Tirado Gómez (201632317), Juan Felipe Rubio Perdomo (201718384)

**Fecha:** 20 octubre, 2020

**Proyecto 2 – Proceso ETL**

Proceso de ETL

A partir de todo el proceso de perfilamiento realizado en el documento del anterior laboratorio, vamos a desarrollar en este laboratorio los procesos realizados para poder extraer, transformar y cargar los datos en un data warehouse a partir de los archivos .xlsx entregados.

El proceso ETL inicia con la extracción de datos de los archivos xlsx dados en el enunciado del proyecto (*peajes2019(2).xlsx* y *Tarifas de peaje a cargo del INVIAS - 2019.xlsx*) los cuales indican las tarifas de cada peaje en Colombia y información relevante de cada uno de los peajes. Como primer paso se pasaron estos archivos a formato .csv, acto seguido fueron analizados con la herramienta *panda\_profiling* con el fin de identificar estadísticas descriptivas de los datos, números de registros y cantidad de registros faltantes. De este proceso se procedieron a tomar las siguientes decisiones respecto a la extracción de datos:

* Se eliminaron las categorías “especiales”de vehículos dado que en su mayoría más del 90% de los registros de los peajes, no tienen este tipo de clasificaciones en su catálogo de tarifas.
* La información de contacto y administración es eliminada puesto que no presenta información relevante para el negocio, dado que la orientación del Análisis OLAP es orientada mas hacia las rutas.
* Mantener uno de los dos valores que caracterizan al departamento ya que dos columnas se refieren a la misma información.

A partir de los datos podemos a plantear las siguientes dimensiones con el fin de modelar cuales serían los criterios principales por los cuales realizar búsquedas dentro del data mart: **Departamento, Peaje, Fecha** y **Ruta, además** planteamos las siguientes medidas que pueden ser calculadas: **Valor de categorías a pagar en COP, valor a pagar por contar con ejes extras, recaudos realizados por peajes de una ruta en especifico y Pago.**

Para procesar los resultados se decidió usar la herramienta de spoon que nos permite cargar los csv explicados en la entrega previa. Mediante esta herramienta se pueden transformar todos los datos que tengan incongruencias es decir se transformaran los datos para sean apropiados para la carga. Si se tiene tiempo posteriormente se cargarán los resultados en Google Cloud.

La tabla de hechos representa la transacción de pagar un peaje, en ella se pondrán las llaves foráneas:

* Identificador del departamento
* Identificador del peaje
* Identificador de la ruta
* Identificador del departamento.
* Identificador de pago
* Costo valor pagado por el vehículo dependiendo de su tipo.

Mediante las cuales se podrán identificar los diferentes tipos de recaudos generados por los peajes a nivel nacional.

El negocio requiere información sobre 2 procesos, el recaudo del peaje y el tráfico en las vías. Para esto se plantean 2 tablas de hechos una para cada proceso. Estas tablas tendrían las siguientes dimensiones

* Peaje
* Fecha
* Hora

La información sobre la via en la cual se encuentra el peaje y el departamento de este se manejaría como niveles de la dimensión peaje. La hora se tiene en cuenta como dimensión para consultar el trafico por hora.

Duda: ¿Es necesario tener 2 tablas de hecho? ¿No seria suficiente tener solo una y medir el trafico en cada via mediante el conteo de los recaudos hechos en cada peaje?