Proyecto 2 – Entrega 2

Nicolás Segura – 201716989

David Ruiz Villota – 201620095

Jaime Torres Lopez –

Contenido

Tabla

Descripción generada automáticamente

# Identificación de necesidades analíticas

Como se indica en el enunciado, estos puntos fueron desarrollados en el documento de Excel adjunto.

# Modelado

## Tablas de hechos

* **Fact table:** La granularidad para esta tabla es alta, ya que para cada ubicación (Departamento/municipio) solo contiene las referencias a las dimensiones correspondientes y un periodo(Año/mes).
  + **Código de ubicación:** Esta medida nos permite conocer el valor que hace referencia a un Municipio (Entidad) de un Departamento.
  + **Código de periodo:** Esta medida nos permite conocer el valor que hace referencia al Periodo (Año/Mes) de los datos de las dimensiones.

**\*Nota:** Las medidas anteriormente descritas son de tipo Aditivas ya que se pueden agregar a todas las dimensiones de la tabla de hechos.

## Dimensiones

* **Periodo:** El periodo permite conocer la temporalidad en que se recopiló la información. Este dato nos ayuda a generar un análisis temporal que podemos contrastar con otros datos.
  + **Año:** Corresponde al año en el que se recopiló la información.
  + **Mes:** Corresponde al mes en el que se recopiló la información.

**\*Nota:** En el desarrollo del proyecto, se incluyó el modelado e implementación de la dimensión **periodo**. Esta dimensión controla la temporalidad de a información; por lo que no se requiere hacer manejo de algún tipo de historia de variación lenta (SCD).

## Modelos Dimensionales

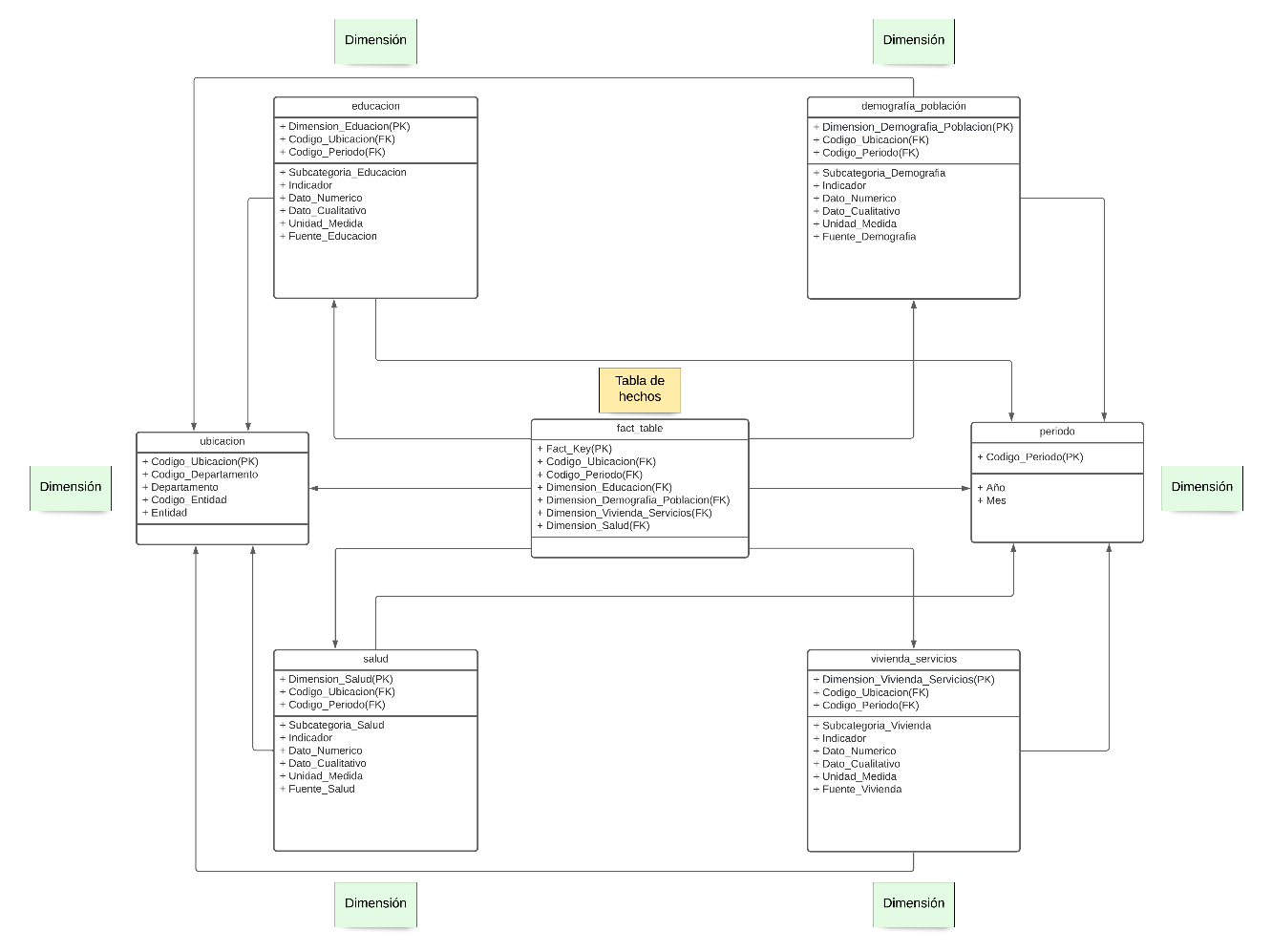


Imagen 1. Modelo dimensional

# Perfilamiento de los datos

## Descripción

Para el perfilamiento de los datos, se tuvieron en cuenta los datos para las dimensiones Demografía y Población, Vivienda y servicios públicos, Educación y Salud. Para todos los datos, se realizó limpieza, normalización e igualmente, se realizaron reportes e informes que nos permitieran visualizar la información relevante de los dataset y de cada una de las variables que lo componían.

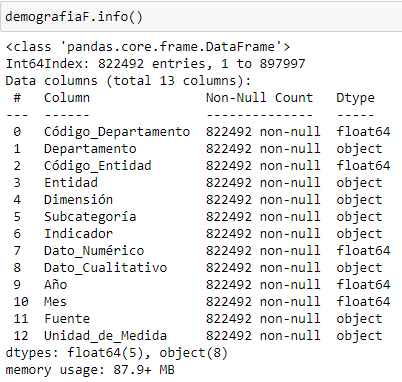


Imagen 2. Informe del dataset Demografía y Población

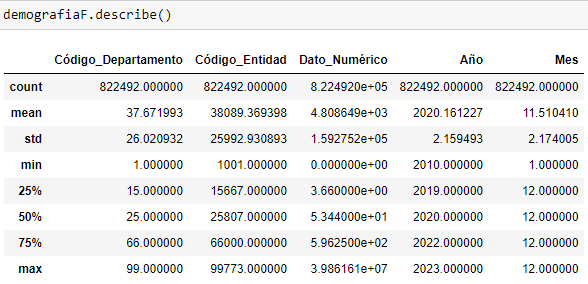


Imagen 3. Reporte del dataset Demografía y Población

Este análisis, permite un mayor acercamiento y entendimiento de la información a tratar, de este modo, se determinó la información relevante para el caso.

El tratamiento y perfilamiento de los datos se encuentra adjunto en el notebook ‘Proyecto2.ipynb’.

## Integración de las fuentes de datos suministradas

Para esta entrega se actualizaron los datos utilizados, pero no se agregaron fuentes de datos externas. Esto se debe a que la información ofrecida por el DNP no es consistente a la obtenida por infraestructura visible, así que para integrar la esta información tendríamos que modificar las tablas para obtener un conjunto de datos consistente. De esta manera, a largo plazo podríamos integrar ambas fuentes y deberíamos contrastar la información de ambos registros, pero con la información en el estado actual, no se podría modelar un esquema que correspondiera a la realidad.

## Diseño del proceso de ETL

Para el proceso de ETL, se estableció el siguiente diseño a implementar

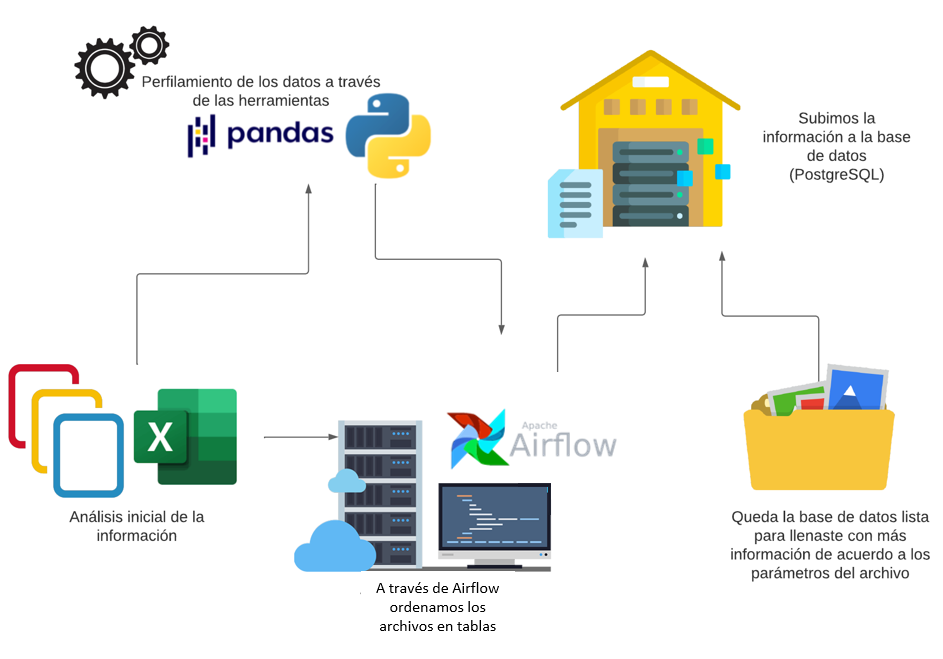


Imagen 4. Diseño proceso ETL

Igualmente, se determinan las siguientes tareas,

* Obtener datos => Recupera y perfila los datos necesarios
* Crear tablas => Crea las tablas de hechos y dimensiones que se usara en la Base de Datos PostgreSQL
* Poblar tablas dimensiones => Puebla las tablas de la Base de Datos con los datos obtenidos anteriormente
* Poblar tabla hechos => Puebla la tabla de hechos creada en la Base de Datos según las tablas de dimensiones y la información con la que se cuente

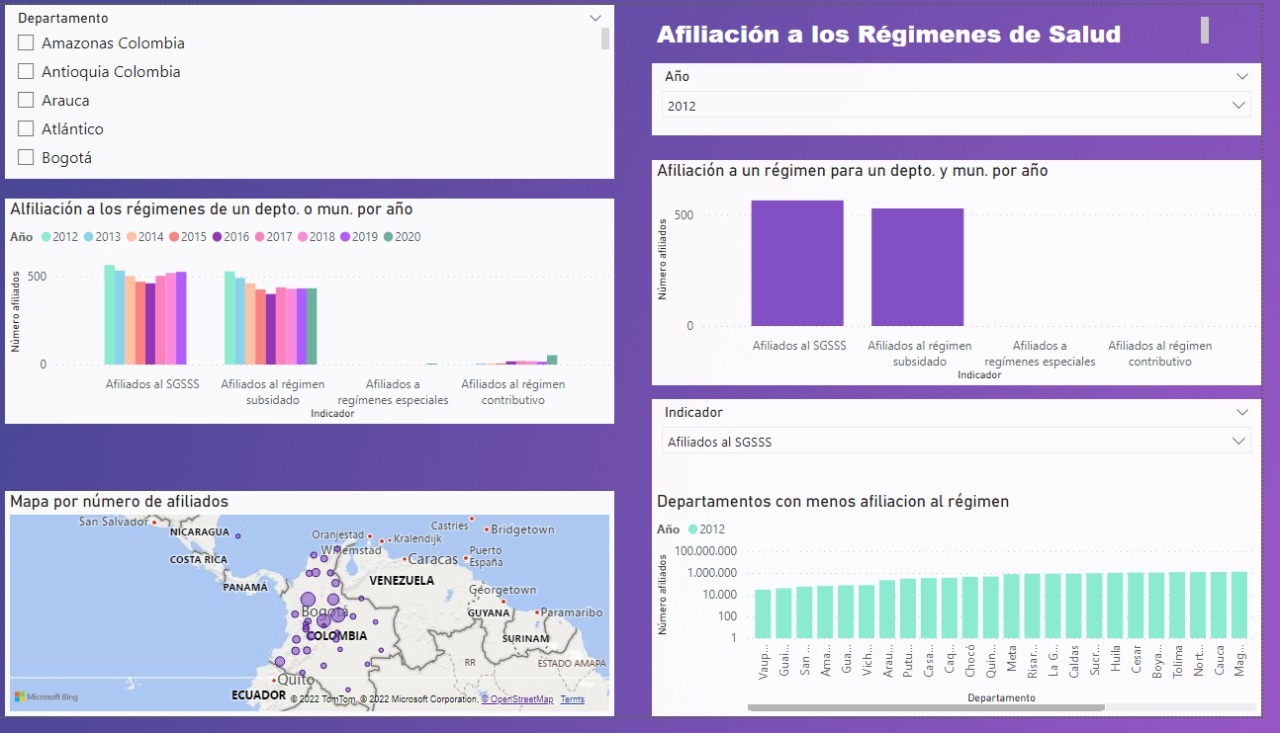
Además del proceso de ETL, sugerimos utilizar el análisis OLAP sobre la información. De este modo podríamos obtener consultas más complejas sin afectar negativamente los sistemas transaccionales. Igualmente, en un futuro sugerimos la integración entre *Apache Airflow* y *Pentaho (PDI),* esto para optimizar la ejecución y agendamiento de tareas y procesos.

Diseño e implementación de dos tableros de control nuevos y una tarea de aprendizaje de máquina

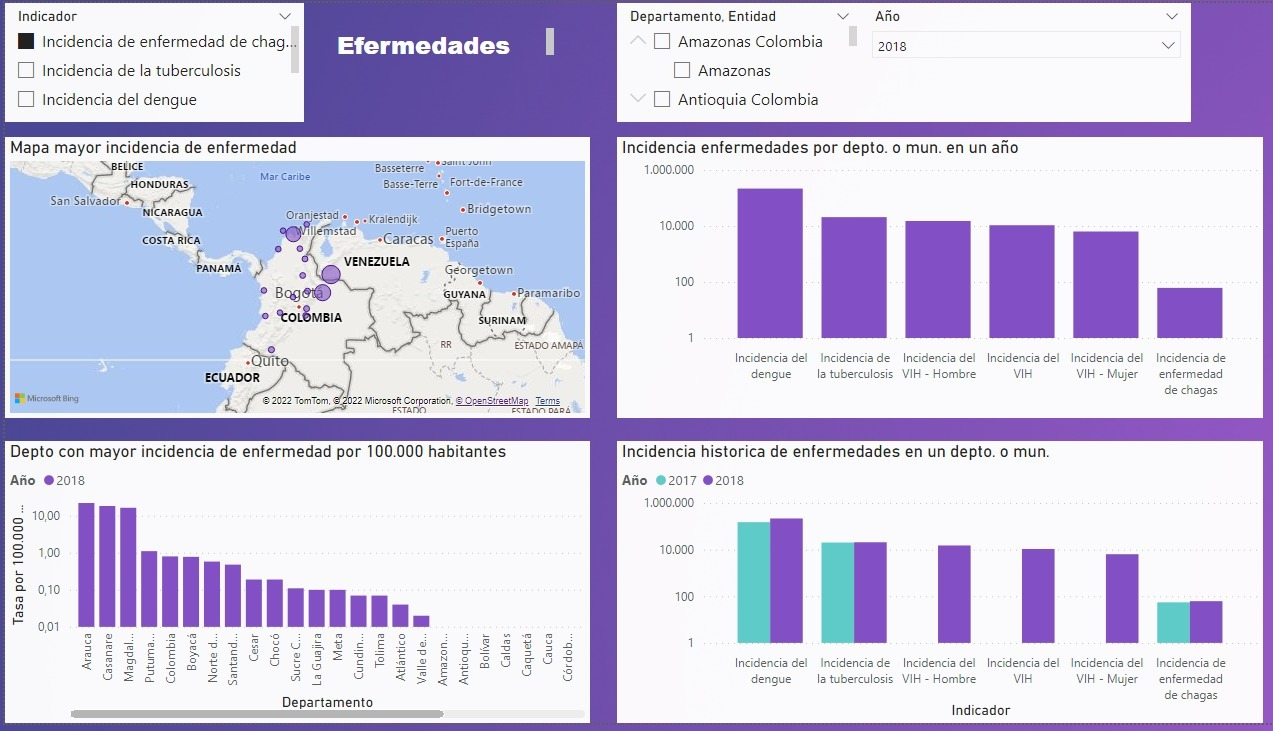
## Tableros de control

Se logró generar los siguientes dashboard.

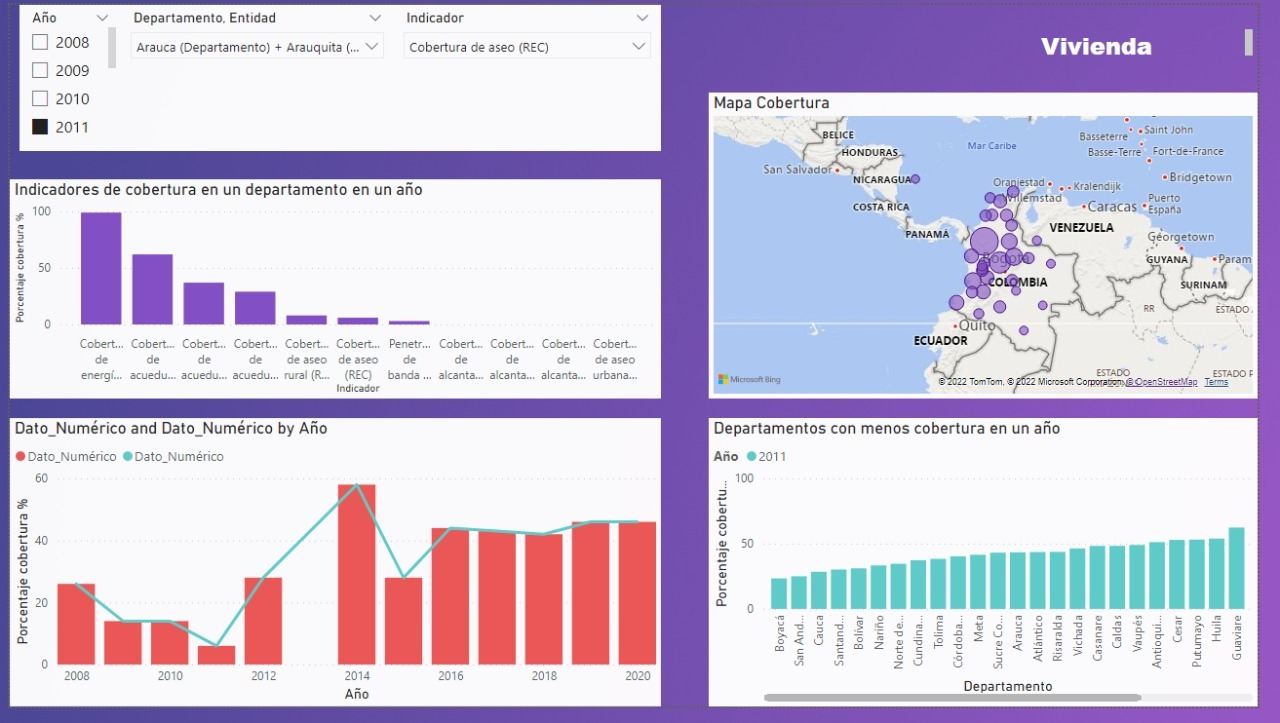
El primero muestra información de afiliados a los Régimenes de salud por departamento y municipio.



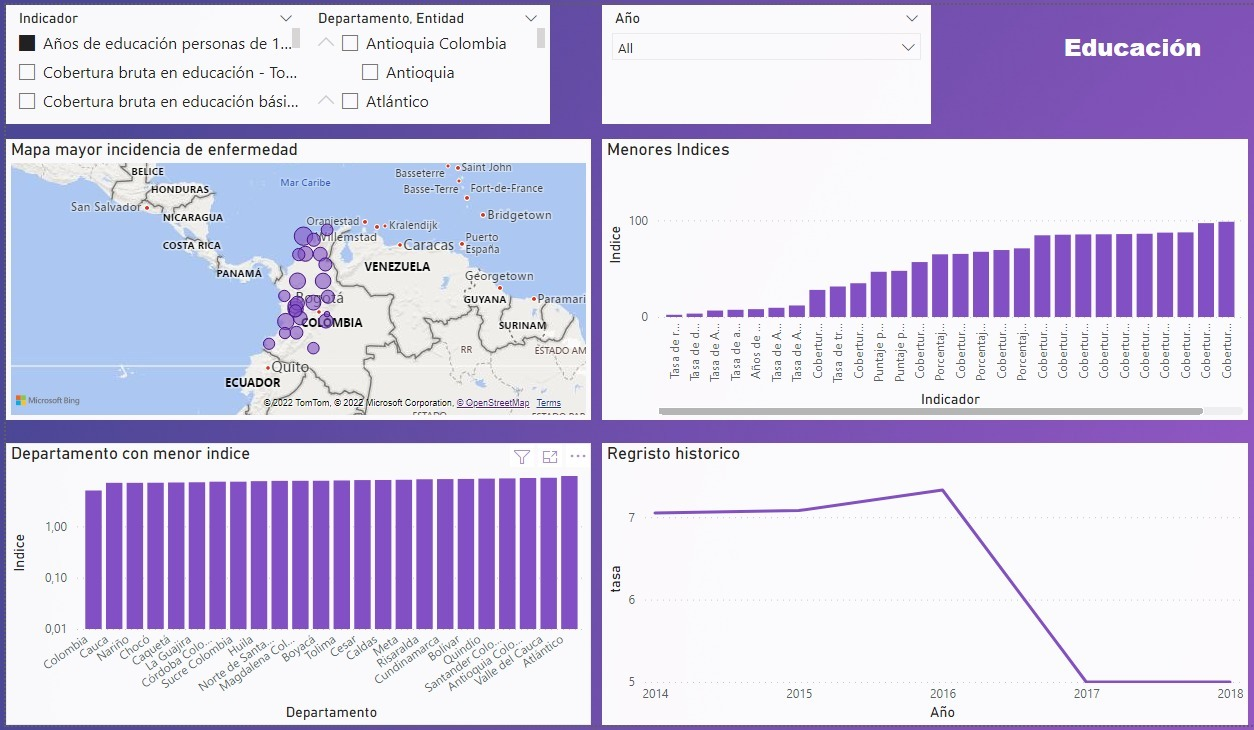
El segundo, muestra la información de enfermedades por departamento y municipio.



El tercero, permite visualizar la información sobre vivienda y servicios públicos en cada departamento y municipio



El cuarto tablero, presenta datos sobre la educación en los departamentos y municipios de Colombia.



## Plan de proyecto

1. Realizar el levantamiento de información y necesidades de BI.
2. Definir conceptualmente las necesidades detectadas.
3. Definir un cronograma para las etapas de desarrollo de solución.
4. Crear un repositorio histórico de datos (Datamart) con la información.
5. Diseñar los procesos de integración de información.
6. Crear los procesos de integración de información.
7. Poblar el Datamart mediante el perfilamiento, extracción y transformación de los datos mediante un ETL.
8. Generar un modelo analítico que nos permitan analizar métricas e indicadores en función de dimensiones.
9. Construir tableros de control con los datos relevantes para el negocio y los indicadores importantes.

## Repartición de tareas

|  |  |
| --- | --- |
| Tarea | Responsable |
| Necesidades Analíticas | Jaime Torres Lopez |
| Modelado | Todos |
| Perfilamiento de datos | David Ruiz Villota |
| Tableros de control | Jaime Torres Lopez |
| Diseño de proceso ETL | David Ruiz Villota, Nicolás Segura Castro |
| Tarea aprendizaje de máquina | Todos |
| Descripción tablas de hechos | Nicolás Segura Castro |
| Documento | Todos |
| Presentación y video | Todos |

|  |  |
| --- | --- |
| **Integrante** | **Puntos** |
| David Ruiz Villota | 33.3 |
| Nicolás Segura | 33.3 |
| Jaime Torres Lopez | 33.3 |