# Segundo Proyecto

## Infraestructura Visible: Entrega 1

Autores:
Miguel Ángel Acosta  (201914976)
Andrés Felipe Rincón (201914118)
Ángela Liliana Jiménez (201912941)

# Tabla de Contenidos



# Modelado de Data Marts

El **objetivo de negocio** escogido es almacenar la información de las muertes registradas para un grupo demográfico específico (edad, sexo) en una locación geografica particular (departamento) dado un causante de muerte (evento)
La **granularidad** escogida en este modelo es anual, se va a registrar un conjunto de muertes de forma anual.
Las **dimensiones** a manejar serán:

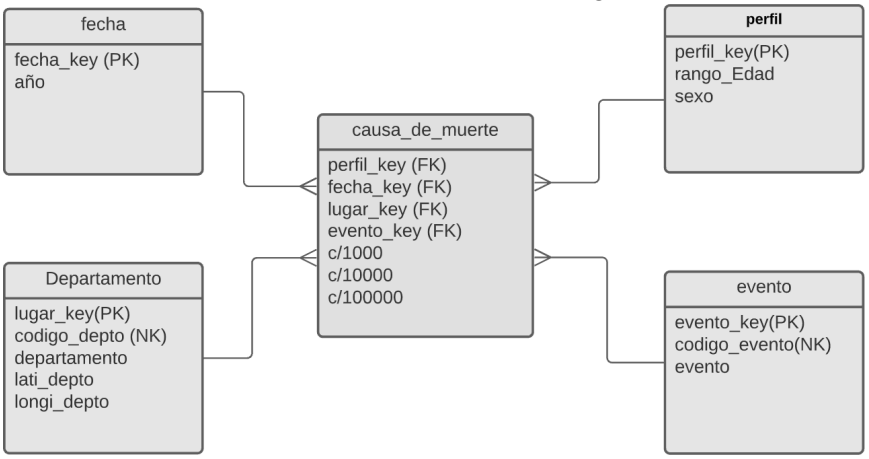
1. Fecha:
   a. Año: Representa el año en el que ocurrió un hecho, permite analizar a lo largo del tiempo el estado de las muertes, por ejemplo ver si hubo un alza en la mortalidad ya sea general o por ejemplo de un departamento.
      Es una dimensión cambiante per se, manejada como una mini dimensión (manejo II) y que va cambiando año tras año, ya que nos interesa ver el cambio año tras año sin reemplazar el valor (descartamos manejo I), se sabe que el año actual es el vigente (descargamos Manejo II) y es un valor que puede aumentar indefinidamente (descartamos manejo III), luego el manejo IV es suficiente y no requiere de manejos más complejos.
2. Lugar:
   a. Codigo_depto: Es la llave natural del atributo departamento, es útil para comparar con otras bases de datos que usen este número.
   b. Nombre_depto: Nombre común del departamento, permite analizar a nivel geográfico las muertes.
   c. Lati_depto: Latitud del departamento, útil para ubicarlo geográficamente.
   d. Longi_depto: Longitud del departamento, útil para ubicarlo geográficamente.
3. Perfil:
   a. Edad: Es el grupo de edad en el que se encuentra un conjunto de muertes, permite analizar a nivel de estructura de población las muertes.
   b. Sexo: Es el sexo relacionado a un grupo de muertos, permite análizar a nivel de sexo las muertes.
4. Evento:
   a. Codigo_evento: Es la llave natural del atributo evento, es útil para comparar con otras bases de datos que usen este número.
   b. Evento: Es el nombre común del evento causante de muerte, permite análizar a nivel de causas las muertes.

Los hechos/medidas a manejar serán:

1. c/1000: Representa la cantidad de casos encontrados para un hecho por cada 1000 habitantes.
2. c/10000: Representa la cantidad de casos encontrados para un hecho por cada 10000 habitantes.
3. c/10000: Representa la cantidad de casos encontrados para un hecho por cada 100000 habitantes.

Todos los **hechos/medidas** son no aditivas, esto porque estamos manejando ratios entre casos presentados por un hecho específico y una cantidad particular de habitantes. Los hechos se promedian para darles significado respectivo a alguna dimensión.

El **modelo dimensional propuesto** que se diseñó es el siguiente:




# Perfilamiento de Datos

Tenemos 3 fuentes de datos:

1. La fuente de datos de muertes: data_2010_2017.csv [1]
   Se eliminaron los datos referentes a totales, para poder manejar cada registro como un hecho único. También se eliminarón muertes en el extranjero, muertes de sexo indeterminado y fuertes de edad desconocida, lo anterior porque no contamos con la cantidad de población determinada para algún grupo
   A continuación se presentan algunos datos extraídos de este conjunto de datos:

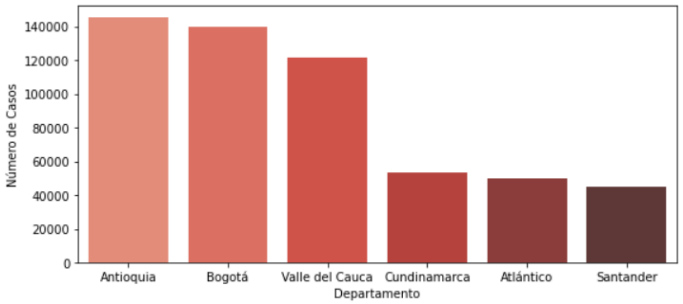



*Figura 1: Conteo de casos por departamento*

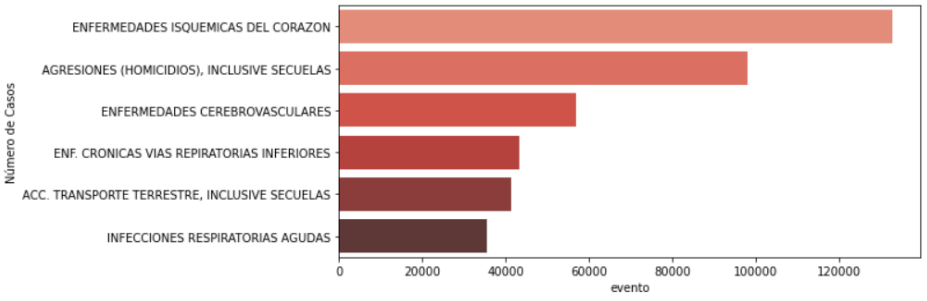


*Figura 2: Conteo de casos por evento*

Podemos ver en los dos ejemplos presentados como tenemos dos gráficas que únicamente representan el conteo de muertes, este conteo por sí solo no es relevante, es necesario normalizar, es decir, pasar de contadores a ratios, que nos indiquen esas muertes a qué porcentaje de su población pertenecen, para esto es que usaremos las otras fuentes de datos.

2. Fuente de datos geográficos: colombia_depto.csv [2]
   Esta fuente simplemente nos permitirá unir los departamentos con sus respectivas referencias geográficas, a continuación podemos ver los puntos geográficos.

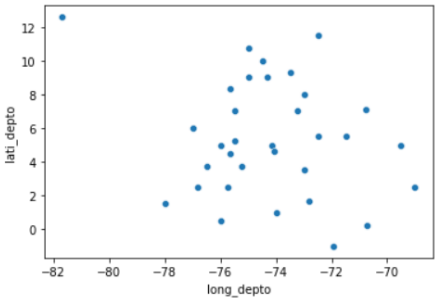



*Figura 3: Ubicaciones geográficas*

Podemos ver que los puntos son coherentes al representar Colombia (el punto a la izquierda es San Andrés)

3. Fuente de datos población geográfica: colombia_depto.csv [3]
   Simplemente contiene la población por departamento en colombia en 2018

| | departamento | población |
|---|---|---|
| **0** | Amazonas | 79020 |
| **1** | Antioquia | 6677930 |
| **2** | Arauca | 391020 |

*Figura 4: Población por departamento*

4. Fuente de datos población demográfica: poblacion_datos.csv [4]
   De esta fuente de datos pudimos obtener la población para un determinado grupo demográfico, pudimos hallar el número de personas en un año, entre un rango de edad y de un género determinado, a continuación una muestra de los datos:

| | anio | edad | sexo | poblacion |
|---|---|---|---|---|
| **244** | 2017 | 20-24 años | Hombres | 2222756 |
| **38** | 2011 | 15-19 años | Hombres | 2221205 |
| **4** | 2010 | 15-19 años | Hombres | 2217321 |

*Figura 5: Grupos demográficos*

Por tomar un ejemplo podemos ver como en el año 2017 para la edad de 20-24 años hubo 2222756 hombres.

El proceso para juntar todo el conjunto de datos que tenemos y entregar una versión definitiva de los datos fue:

1. Usamos la información del conjunto de datos de la *figura 4* para hallar el total de personas en colombia en 2018
2. Calculamos con los datos de la *figura 4* y lo hallado en el paso anterior qué porcentaje de personas representa cada departamento
3. Correlacionamos la fuente de datos de la *figura 5* con la fuente de datos de la *figura 1*, de tal forma que hallamos la población de un año determinado, para un género determinado y para un grupo de edad determinado.
4. Asumimos que cada departamento aporta la misma proporción de población a colombia todos los años por lo que correlacionando lo hallado en el paso 3 con los porcentajes del paso 2 hallamos la población de un año determinado en un departamento determinado y para un género determinado en un grupo de edad determinado

5. Con el número de casos y la población obtenida, realizamos la razón de estos dos datos y la multiplicamos por el número de personas sobre el cúal realizaremos la comparación común:

| | sexo | anio | edad | codigo_depto | nombre_depto | codigo_evento | evento | num_casos | poblacion | c/1000 | c/10000 | c/100000 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **74220** | Mujeres | 2017 | 45-64 años | 99 | Vichada | 303 | ENFERMEDADES ISQUEMICAS DEL CORAZON | 1 | 11921 | 0.08389 | 0.83886 | 8.38856 |
| **74221** | Mujeres | 2017 | 45-64 años | 99 | Vichada | 612 | ENFERMEDADES SISTEMA URINARIO | 1 | 11921 | 0.08389 | 0.83886 | 8.38856 |
| **74222** | Mujeres | 2017 | 5-14 años | 99 | Vichada | 506 | AHOGAMIENTO Y SUMERSION ACCIDENTALES | 1 | 8538 | 0.11712 | 1.17123 | 11.71234 |

*Figura 6: Datos completamente procesados*

Nuestras estimaciones plantean que por ejemplo el año 2017 murió una mujer en Vichada entre 5-14 años, sobre una población de ese mismo tipo de 8538 mujeres, lo que en las columnas subsecuentes se observa como información normalizada sobre 1000, 10000 o 100000 personas, es decir por ejemplo para el mismo caso se podría leer que 0,11 mujeres de cada 1000 para ese mismo grupo murieron.

# Proceso ETL

Partiendo de los datos completamente procesados presentados en la *figura 6* Se realizó un proceso para crear las dimensiones y la tabla de hechos, quedando varios csv's para cada uno de estos, podemos ver aquí una muestra de la tabla de hechos y de una dimensión:

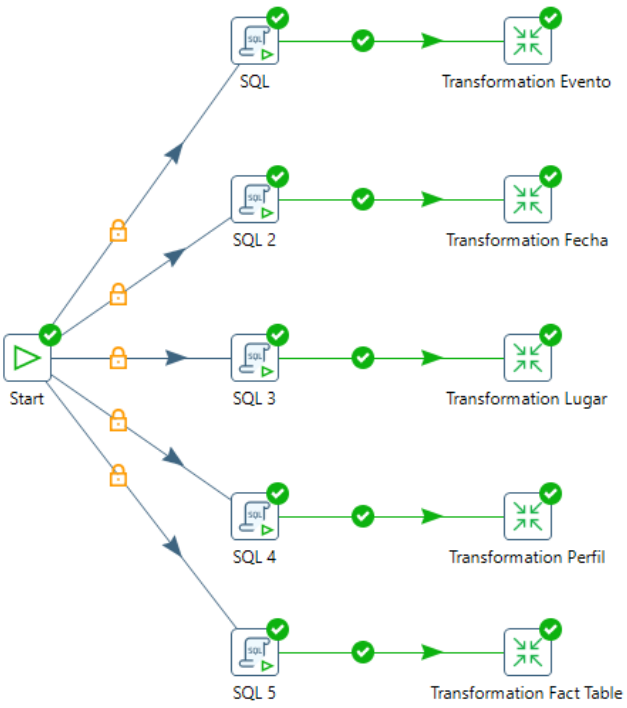
| | perfil_key | edad | sexo |
|---|---|---|---|
| **0** | 1 | 0-4 años | Hombres |
| **1** | 2 | 0-4 años | Mujeres |
| **2** | 3 | 15-44 años | Hombres |
| **3** | 4 | 15-44 años | Mujeres |
| **4** | 5 | 45-64 años | Hombres |

*Figura 7: Muestra Dimensión Perfil*

| | c/1000 | c/10000 | c/100000 | fecha_key | perfil_key | lugar_key | evento_key |
|---|---|---|---|---|---|---|---|
| **0** | 0,00386 | 0,03859 | 0,38594 | 1 | 1 | 1 | 1 |
| **1** | 0,02316 | 0,23156 | 2,31562 | 1 | 1 | 1 | 1 |
| **2** | 0,01581 | 0,1581 | 1,581 | 2 | 1 | 1 | 1 |
| **3** | 0,02767 | 0,27667 | 2,76674 | 2 | 1 | 1 | 1 |

*Figura 8: Muestra Fact Table*

Luego con esos CSV's se empieza un proceso de carga a la base de datos en spoon, tenemos un job principal que se encarga de cargar los datos a una base de datos postgresql, a continuación el job principal:




# Carga de Datos

Finalmente podemos ver como se cargaron los datos a la base de datos y una consulta simple para poder ver la tabla de hechos en la base de datos

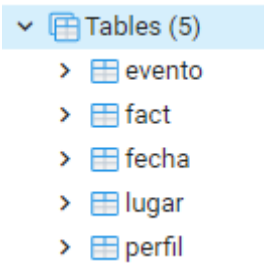



*Figura 9: Tablas Creadas*

```
7   select * from fact
```

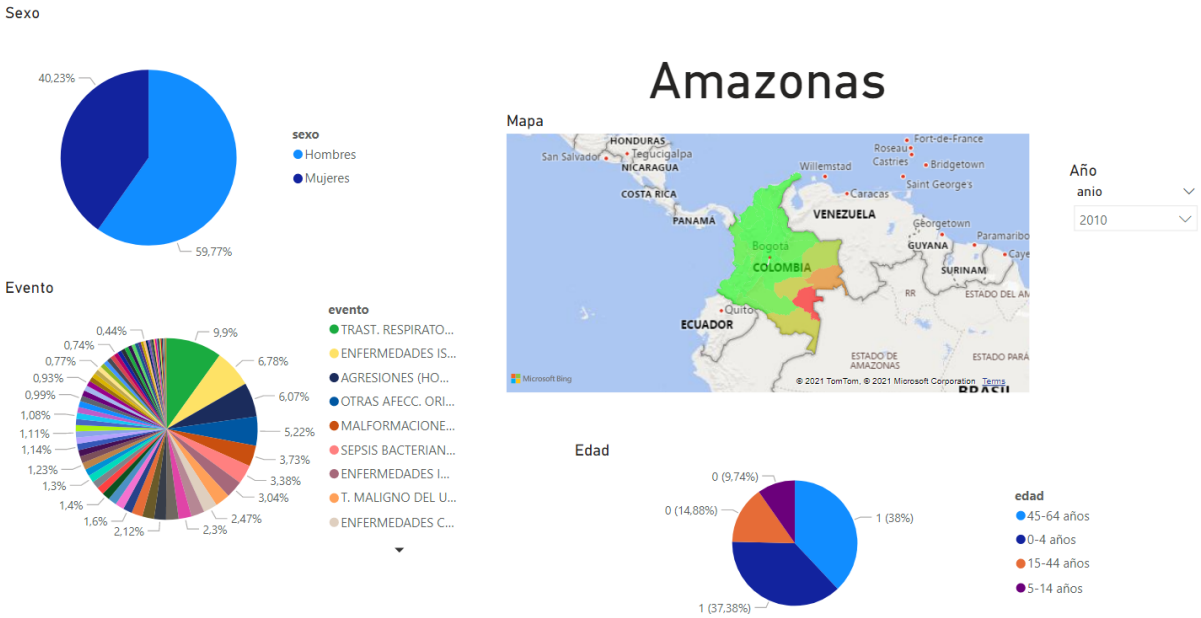
Data Output | Explain | Messages | Notifications

| | fact_key [PK] integer | fecha_key integer | perfil_key integer | lugar_key integer | evento_key integer | c_over_1000 numeric | c_over_10000 numeric | c_over_100000 numeric |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 1 | 1 | 0.00386 | 0.03859 | 0.38594 |
| 2 | 1 | 1 | 1 | 1 | 1 | 0.02316 | 0.23156 | 2.31562 |
| 3 | 2 | 2 | 1 | 1 | 1 | 0.01581 | 0.1581 | 1.581 |
| 4 | 3 | 2 | 1 | 1 | 1 | 0.02767 | 0.27667 | 2.76674 |
| 5 | 4 | 3 | 1 | 1 | 1 | 0.00401 | 0.04008 | 0.40084 |
| 6 | 5 | 3 | 1 | 1 | 1 | 0.01203 | 0.12025 | 1.20253 |

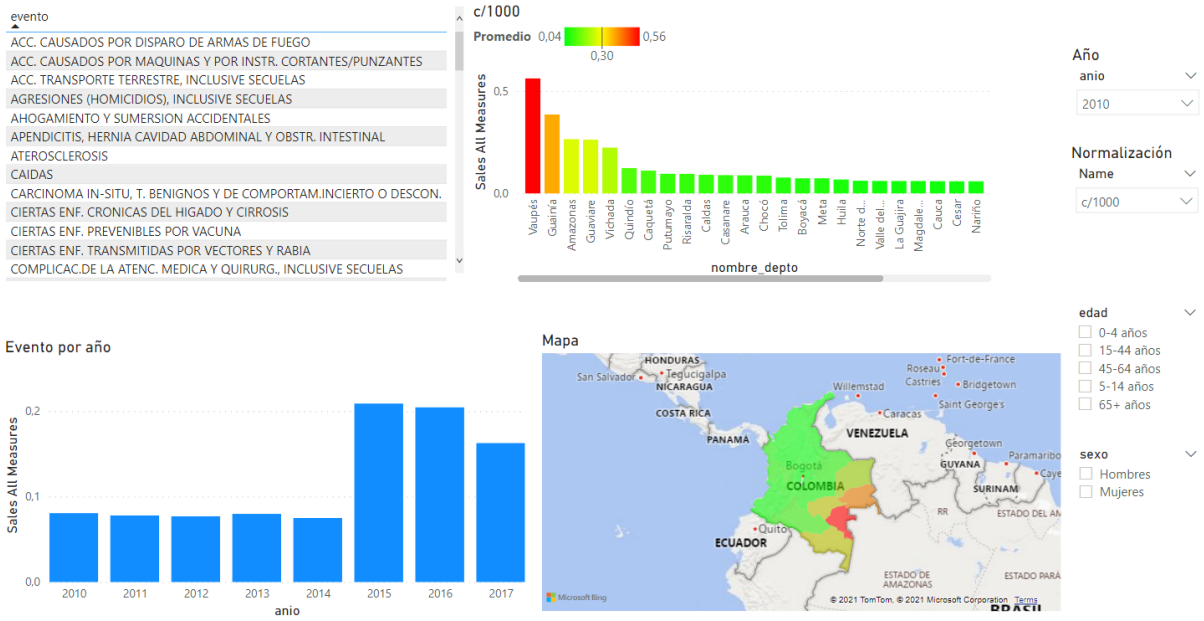
*Figura 10: Fact Table*

# Arquitectura de la Solución

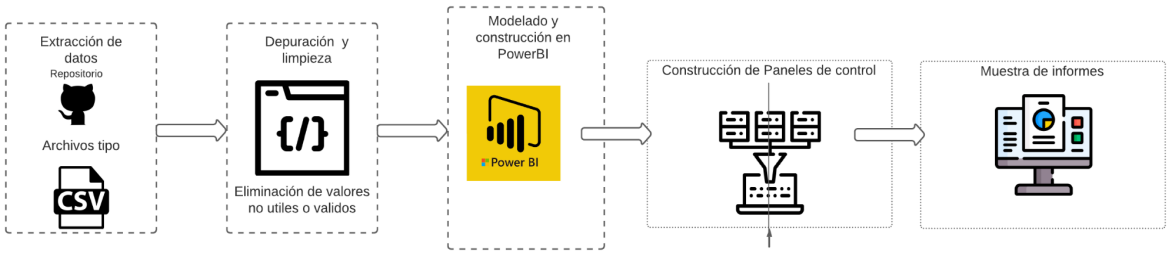
**Diseño de los tableros de control:**
Para los tableros de control planeamos tener 2 marcos diferentes: Por lugar y por evento. Para el lugar desplegamos de acuerdo a la zona seleccionada la diferente información de dicho lugar, esto quiere decir que por cada lugar se visualizarán los porcentajes de sexo, tipo de evento y rangos de edad en un año determinado.
Por otro lado para el tablero de los eventos, se podrá filtrar según un evento específico y evidenciar en el  mapa las diferentes zonas afectadas y la proporción.
Así mismo se verá la cantidad de casos normalizados (según cantidad de habitantes) en diferentes proporciones tanto por año como por lugar.
**Implementación de los tableros de control:**
**Por lugar:**




**Por evento:**




**Proponer la arquitectura de solución:**
La arquitectura es la siguiente:




Se espera que los paneles permitan llegar a conclusiones respecto a los diferentes datos y su relación.

# Video

https://youtu.be/SJKzXqbS4Yo

# Trabajo en Grupo

Miguel Angel Acosta - Scrum Master

Andres Felipe Rincón - Líder de Planeación

Angela Liliana Jiménez - Líder de Diseño

Todos los miembros del grupo formaron parte de las distintas etapas del proyecto, incluyendo el perfilamiento de los datos, el proceso ETL y la construcción de los tableros de control en Power BI. Por esto, consideramos que todos los miembros del grupo deben tener una repartición de 33 puntos por igual.