

Proyecto 2

Santiago Bobadilla – 201820728

Juan José Beltrán – 201819446



30 de MAYO de 2022

Semestre 2022-01

## Introducción

Infraestructura visible es una empresa dedicad a aprovechar y hacer buen uso de información pública disponible con el fin de poder realizar análisis de indicadores socioeconómicos y de diferentes dimensiones. Para este proyecto se van a utilizar los siguientes sets de datos, mejor llamados dimensiones:

1. Demografía y Población
2. Vivienda y Servicios Públicos
3. Educación
4. Salud

Si bien existen otros, se va a optar por primar estos 4 de la manera más precisa y completa posible.

## IDENTIFICACIÓN DE Necesidades Analíticas

Con base en los datos entregados, la entrevista realizada y un estudio del estado del arte se pudo encontrar unos temas claros a abordar. Con el fin de tener una visión de manera más precisa cada uno, así como la finalidad de todos como un todo se usó el análisis de Kimball. Con esto obtuvimos para cada una de las dimensiones una o dos necesidades por abordar. A continuación, se presenta la totalidad de lo que se busca estudiar y luego la justificación general.

Se resalta que, si bien esto fue un primer reconocimiento, los tableros de control y análisis que sea entregado no esta necesariamente ligado a dichas necesidades dado nuevos descubrimientos en los datos, así como propuesta al hablar con el equipo, diferentes a las entrevistas originales.

Tabla

Descripción generada automáticamente

### Justificación

Se propone tanto un tablero de control como una análisis OLAP en las dimensiones más relevantes para un país de desarrollo con altísima desigualdad. Esto se debe a que, si bien es importante ver visualmente la dimensión, vale la pena entender factores que generen influencia en dicho comportamiento. Inicialmente y en este proyecto se va a enseñar el aspecto visual de cada dimensión (la cual, si bien numéricamente depende de si misma, analíticamente depende las otras) con el de encontrar patrones y relaciones que tengan valor.

El valor lo definimos con base a lo comentado en las entrevistas donde se estipulo un grano municipal con el fin de tener la mayor exactitud posible. Además, los temas y dimensiones se eligieron por petición de que se busca tener factores correlacionados entre ellos, donde sea posible buscar patrones.

Del estado del Arte se usaron reportes de Bloomberg de índices de crecimiento de la población en relación con la disponibilidad de recursos naturales con el fin de discernir que datos eran relevantes para Colombia y su influencia estadística sobre otros campos. Ejemplo de esto, una propuesta para proyectos futuros de análisis OLAP de la demografía con los servicios de censos que se tiene.

Cabe resaltar nuevamente el grano es estos tableros y posibles análisis dado que es de suma importancia tener información lo más detallada posible.

## MODELAMIENTO DEL Data MarT

Con el fin de cumplir con los requerimientos de manera detallada, pero sin privar/eliminar información que pueda ser de utilidad para próximos proyectos, se usaron las siguientes dimensiones generales, con el fin de tener un enfoque más profundo de los datos y de la misma manera llegar a hallazgos más interesantes:

* Fecha
* Indicador
* Municipio

Dichas clases fueron llenadas con la información de las 4 categorías estipuladas inicialmente. Así, con esto en mente se paso a generar 4 tablas de hechos, una para cada categoría.

Se va a presentar una representación de una tabla de hechos con sus dimensiones respectivas y se estipula las similitudes y diferencias con respecto a las otras tres. Para este caso, se va a usar Demografía y Población.

Diagrama

Descripción generada automáticamente

La diferencia con la tabla de hechos de salud, educación y vivienda es su valor en la tabla de hechos, dado que las dimensiones siempre están presentes con el fin de poder realizar un inspección detallada de la información. En ese orden de ideas se debe tener presente que las clases estrella que arman las llaves foráneas en la tablas de hechos como se va a explicar en siguientes literales contiene toda la información disponible de las cuatro bases de datos presentadas. Es decir, son aptas para varias tablas de hechos sin importar el requerimiento que se esté cubriendo.

No obstante, a continuación, se presenta el modelo multidimensional completo:

Diagrama

Descripción generada automáticamente

### Tablas de hechos

La granularidad que se uso en las tablas de hecho se basa en dos componentes importantes que se nos comunicaron en las entrevistas:

1. Nivel geográfico en el que se quiere presentar los datos.
2. Especificidad del tema de la información.

En primera instancia se busca mostrar la información de manera municipal en un momento específico del tiempo, es decir, se quiere que en cada una de las tablas de hechos el grano sea una “foto” del municipio respecto a los indicadores en algún momento del tiempo. Si bien hay una categoría más amplia que es departamento, se nos informó que no hay realmente mucho que aportar diferente a estudios del arte en dicha granularidad. Segundo, se revisó y detalló que existen estudios e informes por categoría ya detallados; pero, que la información interna dividida por indicadores no tiene mayor relevancia. Así, para nosotros, se toma como relevante e importante de mostrar.

Como segundo hecho importante para resaltar, se indica que se tiene una sola medida usualmente llamada ‘valor\_nombreDelHecho’ en cada table de hechos y es semi aditiva. Si bien esto no es muy claro desde el modelamiento se debe tener en cuenta un ejemplo de dicha información para entender cuando es aditiva y cuando no.

Tabla

Descripción generada automáticamente

Si observamos de manera detallada el ejemplo de la tabla de hechos presentado para educación, se puede ver que el valor es aditivo para escalar de granularidad. Es decir, reportar valores departamentales y no municipales, o escalar en la temporalidad, es decir no una año-mes-día, sino reportar por meses, o años. Si bien esto es simple dado que tenemos la menor granularidad posible, se debe tener presente que no hay posibilidad de adhesión si estamos trabajando con diferentes indicadores. Esto seria erróneo dado que pueden existir meticas diferentes, así como temas diferentes cuando se plantea una adhesión. Esto se da de manera generalizada para las 4 tablas de hechos usadas en este modelo multidimensional.

### Justificación del modelo

Para el manejo de historias se analizó si alguna de las clases requería de variaciones temporales u históricas de algún tipo. Dicho análisis como se puede ver con el diagrama muestra que inicialmente se habla de información estática estandarizada que difícilmente presenta cambios en temporalidades bajas. Por consiguiente, cada clase se cataloga como tipo 1, es decir, sin memoria histórica y solo con información presente. Esto por lo que se mencionaba en la sección inmediatamente anterior, y es que se tienen “fotos” del indicador del municipio en un momento específico del tiempo, por lo cual, si se toma otro dato en otro momento del tiempo, será un nuevo registro y no se necesitará modificar el valor directamente, a menos de que haya sido erróneo; y si esto último pasa, entonces no será necesario guardar los datos que quedaron mal almacenados o calculados.

## Datos y ETL

Los datos tuvieron las siguientes fases de exploración y modificación con el de poder construir un ETL completo.

1. Exploración de rangos lógicos de ciertos datos, así como su correlación e inferencia.
2. Construcción de la información necesaria y correcta para el ETL.

### Exploración

En primera medida se usó la siguiente información:

* Tabla de demografía y población: Específicamente en la población menor/mayor de edad. Subcategoría: "Población menor y mayor de edad" e indicador: "Porcentaje de población menores de 15 años".
* Tabla de salud. Subcategoría: "Tasas y coberturas" e indicador: "Tasa de mortalidad infantil en menores de 1 año (x cada 1.000 nacidos vivos)", y "Tasa de mortalidad infantil en menores de 5 años".
* Tabla de educación. Subcategoría: "Acceso a la educación" e indicador: "Cobertura bruta en educación primaria" y "Cobertura bruta en educación básica".
* Tabla de vivienda: Subcategoría: "Vivienda y acceso a servicios públicos" e indicadores: "TODOS LOS CENSOS".

En segunda medida, el proceso para cada tabla fue:

1. Filtrar por máscaras de verdad.
2. Generar una columna de tipo Date donde se asume que el día siempre es 01 dado la falta de información. Esto se hace para facilidad al momento de guardar la información en la base de datos.
3. Se corrige todo valor numero a un formato simple e internacional estadounidense. Es decir, el decimal es el ‘.’ y la coma es separador decimales, más solo se usa para belleza al momento de presentar datos.
4. Se corrige rangos lógicos de los datos con base en su naturaleza.
5. Se eliminan todas las columnas que no generen información relevante.
6. Se eliminan todos los registros con información nula.
7. Se revisa la desviación de los valores con el fin de tener un mayor conocimiento.

Se puede ver con mayor detalle este proceso en el Notebook anexado, sin embargo, se muestra un ejemplo de la clase demografía donde se observa la dispersión/varianza de los datos, así como los casos atípicos numéricamente hablando.

Tabla

Descripción generada automáticamente con confianza media

Esto nos permite realizar inferencias importantes de cuales son los datos considerados normales para dicho indicador y cuales fueron atípicos. Así, tener ideas iniciales de que datos valen la pena que se vean en los tableros de control.

### Tablas Data Mart

Para construir el ETL se requería de la información presente en el Data Mart. Así, se buscó construir en primera instancia las dimensiones que conforman las clases estrella, para posteriormente ver como quedarían las tablas de hechos. En ese orden de ideas se realizó un proceso genérico similar para estas clases:

1. Generación de funciones auxiliares en caso de que fueran necesarias si las tablas necesitaban una transformación exacta. Ejemplo, pasar el mes de numero a texto.
2. Aclaración de supuestos en caso de que fueran necesarios. Entre estos el más importante es:
   1. No hay registros con información repetida.
   2. Se asume que el día en la fecha es 01.
   3. Se diferencia por valor conjunto. Es decir, hay diferencia al pensar en departamento-municipio y no solo por uno de los dos.
3. Creación de tabla con información no repetida.

El resultado de dicho proceso queda de la siguiente manera:

Tabla

Descripción generada automáticamente

Es importante notar las dimensiones de la tabla, donde se observa que, de miles de datos, cuales son realmente únicos y usables para las diferentes tablas de hechos al momento de registrar el proceso correspondiente.

Las demás tablas y el detalle de la limpieza y transformación se pueden encontrar en el Notebook anexado.

Por último, las tablas de hechos son la unión (JOIN/MERGE) de las diferentes tablas con valores y sus dimensiones estrellas removiendo registros con valores nulos, o columnas repetidas o sin valor real. Anteriormente ya se presentó un ejemplo de esta afirmación.

### ETL

Se muestra el resultado del proceso ETL luego de los pasos de autenticación y conexión con la base de datos PostgreSQL; así como el grafo relacionado al proceso.

Diagrama

Descripción generada automáticamenteTabla

Descripción generada automáticamente

Si bien existieron errores en el proceso, el resultado y la carga de la información fue exitosa. Solamente como muestra de la confirmación, se pueden ver a continuación consultas dentro de la base de datos PostgreSQL, en donde se ve que la información quedó cargada tal como se planeó:

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación, Tabla

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Para más información, les pedimos acceder a la máquina virtual o descargar el repositorio que viene adjunto.

## Tableros de Control

### ARQUITECTURA DE SOLUCIÓN:

La arquitectura es bastante sencilla si se tiene en cuenta que la información está en bases de datos distribuidas, lo cual hace que la solución del problema sea bastante escalable. A continuación, se muestra la arquitectura, en donde se ve que una máquina puede soportar todo el proceso, no obstante, la máquina virtual dada se quedó un poco corta para los requerimientos de la empresa:

Diagrama

Descripción generada automáticamente

Posteriormente, en el archivo “Tableros Proyecto 2.pbix”, se podrán ver los tableros de control, teniendo en cuenta que en el video se mostrará a detalle sus funcionalidades. Finalmente, cabe mencionar que los datos que se importaron al tablero de control están dados por consultas SQL que unen la información de todas las tablas, de tal forma que se obtenga la información de forma escalable y distribuida.

Enlace al video: <https://youtu.be/90IdzvXrBCQ>

## Actividades Realizadas

Las actividades realizadas fueron:

* Juan José Beltrán: Creador de los tableros de control y ETL.
  + Tiempo promedio total: 11.4 horas
* Santiago Bobadilla: Exploración de datos y creación de Data Mart.
  + Tiempo promedio total: 9.9 horas

Los puntos totales se van a repartir 50-50 entre los dos miembros.

## Anexos

1. Jupyter Notebook
2. Tablero de Control con los Resultados Específicos
3. Video
4. Datos modificados usados para la determinación de resultados