Proyecto 2: Infraestructura Visible

F. Bedoya, C. Salinas, N. Orjuela

Grupo #20
Sección #1
Inteligencia de Negocios
Departamento de Ingeniería de Sistemas y Computación
Universidad de Los Andes
Bogotá, Colombia

Repositorio con el código del proyecto

1. Requerimientos analíticos

Tema analítico	Análisis requeridos o inferidos	Categoría del análisis - Tablero de control, análisis OLAP	Procesos de negocio	Fuentes de datos y datos
Comportamiento de la población demográfica por género y nivel de urbanización en Colombia	Extracción del porcentaje de hombres y mujeres por departamento por año	Análisis OLAP	Gran Encuesta Integrada de Hogares realizada por el DANE	TerriData - Demografia Y Poblacion.xslx
	Extracción de la cantidad de población rural y urbana por departamento por año	Análisis OLAP	Gran Encuesta Integrada de Hogares realizada por el DANE	TerriData - Demografia Y Poblacion.xslx
	Visualización del porcentaje de hombres y mujeres por año utilizando un mapa de Colombia	Tablero de control	Gran Encuesta Integrada de Hogares realizada por el DANE	TerriData - Demografia Y Poblacion.xslx
	Visualización de la cantidad de población rural y urbana utilizando un mapa de Colombia	Tablero de control	Gran Encuesta Integrada de Hogares realizada por el DANE	TerriData - Demografia Y Poblacion.xslx

Tema analítico	Análisis requeridos o inferidos	Categoría del análisis - Tablero de control, análisis OLAP	Procesos de negocio	Fuentes de datos y datos
Tasas de analfabetismo y cobertura de educación en Colombia	Extracción de tasa de analfabetismo por departamento por año y por nivel de urbanización (urbano o rural)	Análisis OLAP	Gran Encuesta Integrada de Hogares realizada por el DANE	TerriData - Educacion.xslx
	Extracción del indicador de cobertura neta total por departamento y año	Análisis OLAP	Gran Encuesta Integrada de Hogares realizada por el DANE	TerriData - Educacion.xslx
	Visualización de la tasa de analfabetismo cruzado con la cobertura neta total en la educación utilizando un mapa de Colombia	Tablero de control	Gran Encuesta Integrada de Hogares realizada por el DANE	TerriData - Educacion.xslx

Tema analítico	Análisis requeridos o inferidos	Categoría del análisis - Tablero de	Procesos de negocio	Fuentes de datos y datos
Tasas de deserción intra-anual y repitencia del sector oficial en educación básica y media en Colombia	Extracción de la tasa de deserción intra-anual por departamento por año	Análisis OLAP	Gran Encuesta Integrada de Hogares realizada por el DANE	TerriData - Educacion.xslx
	Extracción de la tasa de repitencia por departamento por año	Análisis OLAP	Gran Encuesta Integrada de Hogares realizada por el DANE	TerriData - Educacion.xslx
	Visualización de la tasa de deserción intra-anual cruzado con la tasa de repitencia utilizando un mapa de Colombia	Tablero de control	Gran Encuesta Integrada de Hogares realizada por el DANE	TerriData - Educacion.xslx

Tema analítico	Análisis requeridos o inferidos	Categoría del análisis - Tablero de	Procesos de negocio	Fuentes de datos y datos
Correlación entre cobertura de vacunacion (DPT, Triple Viral, y pentavalente menores de 1 año) con el registro especial de prestadores y sedes de servicios de salud en Colombia	Extracción y geolocalizacion de los prestadores y sedes de servicios de salud.	Análisis OLAP	Encuesta interna de datos Ministerio de Salud y Protección Social - Dirección de Prestación de Servicios y Atención Primaria	datos.gov.co (https://www.datos.gov.co/Salud y-Protecci-n-Social/Registro-Especial-de-Prestadores-y-Sedes de-Servic/c36g-9fc2/) datos locales en: datos/Registro_Especial_de_Pre stadores_y_Sedes_de_Servicios_ de_Salud.csv
	Extracción tasa de vacunación (DPT,Tripe Viral, y Pentavalente) por departamento	Análisis OLAP	Tabulación datos vacunación nacionales por parte del Ministerio de Salud y Protección Social	TerriData - Salud.xslx
	Visualización de los centros de salud en el mapa colombiano	Tablero de control	Encuesta interna de datos Ministerio de Salud y Protección Social - Dirección de Prestación de Servicios y Atención Primaria	datos.gov.co (https://www.datos.gov.co/Salud y-Protecci-n-Social/Registro-Especial-de-Prestadores-y-Sedes de-Servic/c36g-9fc2/) datos locales en: datos/Registro_Especial_de_Pre stadores_y_Sedes_de_Servicios_ de_Salud.csv
	Visualización de las tasas por departamento de vacunación (DPT, Triple Viral y Pentavalente) en el mapa colombiano	Tablero de control	Tabulación datos vacunación nacionales por parte del Ministerio de Salud y Protección Social	TerriData - Salud.xslx
	Visualización del cruzado de tasas de vacunación por departamento vs cantidad de sedes de servicios de salud en Colombia	Tablero de control	Encuesta interna de datos Ministerio de Salud y Protección Social - Dirección de Prestación de Servicios y Atención Primaria / Tabulación datos vacunación nacionales por parte del Ministerio de Salud y Protección Social	TerriData - Salud.xslx y datos.gov.co (https://www.datos.gov.co/Salud y-Protecci-n-Social/Registro-Especial-de-Prestadores-y-Sedes de-Servic/c36g-9fc2/) datos locales en: datos/Registro_Especial_de_Pre stadores_y_Sedes_de_Servicios_ de_Salud.csv

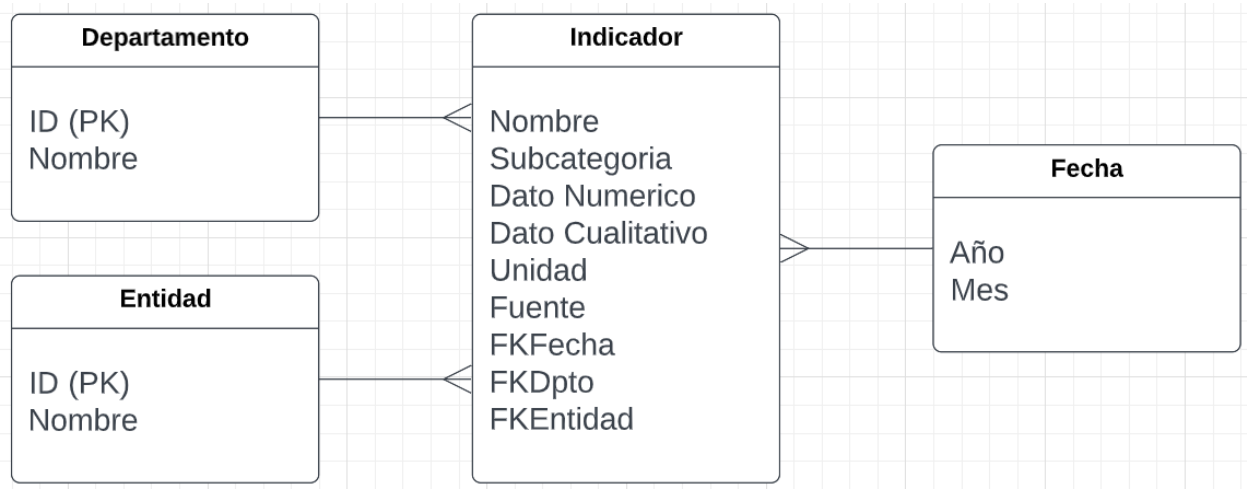
Tema analítico	Análisis requeridos o inferidos	Categoría del análisis - Tablero de	Procesos de negocio	Fuentes de datos y datos
Cobertura de alcantarillado y acueducto en Colombia	Extracción de la cobertura de alcantarillado por departamento por año	Análisis OLAP	Gran Encuesta Integrada de Hogares realizada por el DANE	TerriData - Vivienda y Servicios Publicos.xslx
	Extracción de la cobertura de acueducto por departamento por año	Análisis OLAP	Gran Encuesta Integrada de Hogares realizada por el DANE	TerriData - Vivienda y Servicios Publicos.xslx
	Visualización de la cobertura de alcantarillado cruzado con la cobertura de acueducto utilizando un mapa de Colombia	Tablero de control	Gran Encuesta Integrada de Hogares realizada por el DANE	TerriData - Vivienda y Servicios Publicos.xslx

En el contexto del negocio es importante entender que el correcto manejo y administración de los recursos es algo que debe ser transparente, así como también lo debe ser los resultados de las soluciones de infraestructura propuestas y elaboradas por el gobierno nacional. Después de haber revisado la organización de Infraestructura visible notamos que los proyectos actuales que se llevan a cabo en relación con analítica de datos refieren específicamente a propuestas de infraestructura que afectan directa o indirectamente indicadores socioeconómicos y demográficos, como, por ejemplo, la red nacional de bibliotecas públicas y como estas afectan la formación de las personas y apoyan planes y proyectos a nivel nacional de acuerdo con políticas públicas de educación. La base de datos de Infraestructura visible evidencia un objetivo de completitud de la información de instalaciones gubernamentales, pero también una deficiencia en la visualización y entendimiento a profundidad de los datos presentados,

comparándola específicamente con su homónimo de la oficina federal de prisiones de Estados Unidos.

2. Modelado de Data Marts

a. Modelos dimensionales



Al analizar los datos que se encuentran en los archivos .xslx nos dimos cuenta que estos tienen la misma estructura, por lo que proponemos este modelo para utilizarlo en las dimensiones escogidas, específicamente aquellas que vamos a implementar para este proyecto, es decir, la dimensión de educación.

b. Justificación de los modelos dimensionales

La granularidad de la tabla de hechos se establece mediante sus dimensiones. Es importante entender el indicador geográficamente y por eso se incluyen las dimensiones de departamento y entidad (mejor entendida como municipio) y la fecha en la que este indicador se tomó. La fecha tiene una granularidad bastante gruesa, solo contando con año y mes, pero es importante considerar que para indicadores de tan grande aspecto y poca variabilidad diaria o semanal, los reportes son comúnmente mensuales y los cambios que se observan se reportan mensual o anualmente.

El hecho se compone del nombre del indicador, la subcategoría del indicador para ayudar a filtrar indicadores parecidos, el resultado numérico o cualitativo reportado por el indicador, la unidad del resultado y la fuente de la que fue extraído. De aquí no hay ninguna medida aditiva, además de las que son cualitativas, el único valor numérico es el Dato Numérico para el cual en su gran mayoría son índices o promedios que no tendrían sentido sumar especialmente a través de múltiples localizaciones geográficas.

Es importante recalcar que el manejo de la historia del hecho, necesario para ver la variabilidad anual de los indicadores se maneja en el hecho mismo a través de la dimensión fecha y no se utiliza ninguno de los tipos de manejo histórico desde las dimensiones. Esto se hizo porque ninguna dimensión representa una potencial variabilidad y el nuevo registro o cambio de un indicador se debe ver reflejado sobre sus resultados y fecha.

3. Entendimiento de los datos, creación del Data Mart y proceso ETL

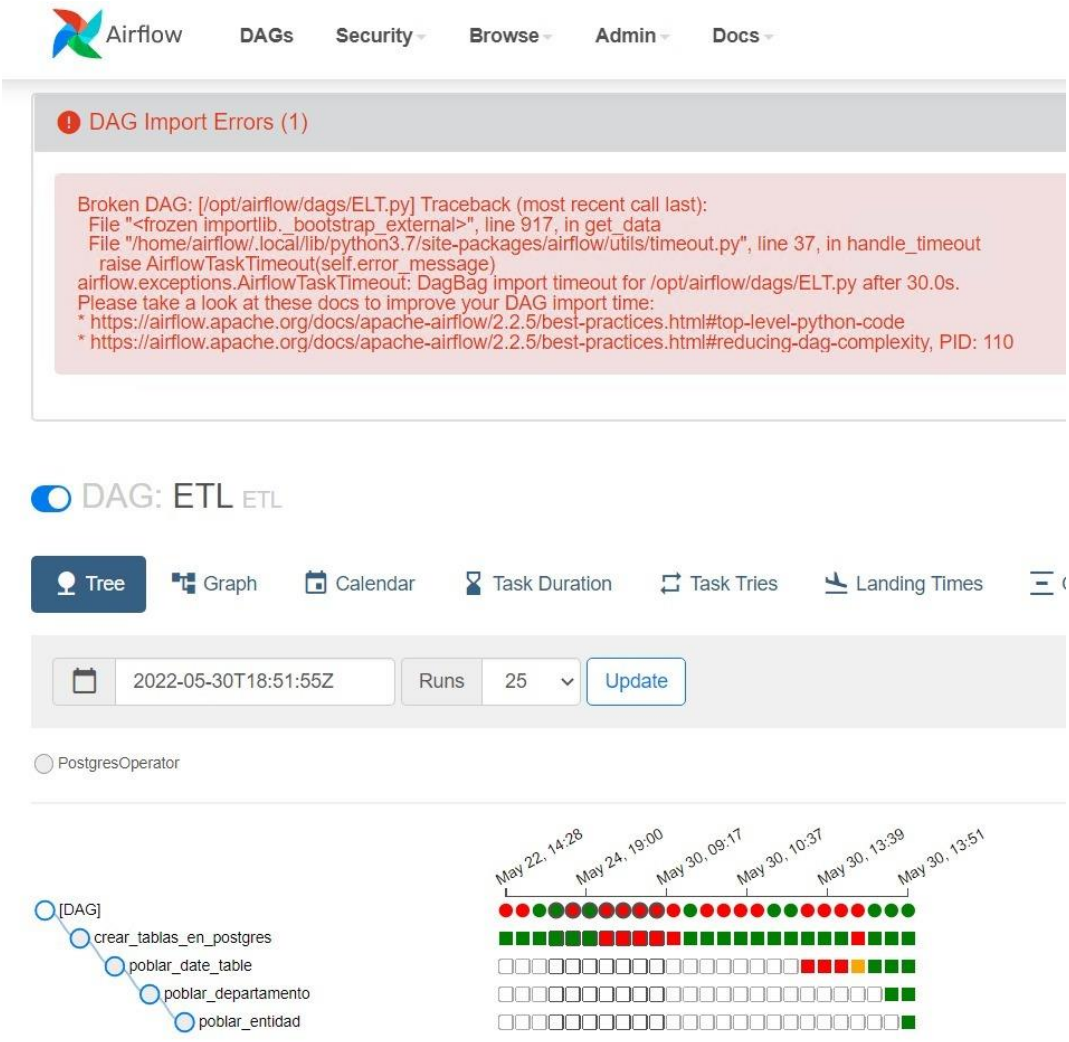
a. Entendimiento de los datos

Los datos están muy limpios y son muy trabajables. Lo que mas nos gustaría recalcar es que son consistentes. Un código de departamento o de entidad se mantiene para el mismo departamento o entidad a través de todos los datos. Aun así, son muchos datos y esto tiene un efecto negativo en la infraestructura utilizada para el despliegue. Los datos no presentan ningún nulo por fuera de lo esperado para el negocio, es decir un valor nulo en alguno de los resultados del indicador ya sea cualitativo o cuantitativo por su naturaleza mutuamente exclusiva. De los casi 200,000 datos (de la fuente de datos de educación) que se tienen todos son completos y se rigen con un tipo de dato constante, en su mayoría un String, exceptuando las fechas, códigos y resultados que son datos numéricos.

b. Diseño e implementación del proceso de ETL

Para la implementación del modelo ETL se obtienen los datos de manera dinámica desde la fuente de TerriData utilizando un Task dentro de Airflow y se descomprime en la máquina de virtual de Docker. Después, otro Task que utiliza el archivo obtenido para generar los CSV de todas las dimensiones y de la tabla de hechos. Es en este mismo Task donde partiendo el archivo principal se manejan los nulos obtenidos y los tipos de datos de las columnas.

Finalmente, en un Task por dimensión y tabla de hechos se insertan a la base de datos PostgreSQL.



Aquí podemos ver como los Tasks consecutivos se van ejecutando y como la población de las tablas se realiza de manera correcta. Aun así, podemos ver un error del DAG cuando se incluye el Task de poblar la tabla de hechos debido a que toma mucho

tiempo en la importación (aunque corra después sin errores). Esto hace que los Workers de Airflow tengan errores de sincronización. Para resolver esta problemática se intentó cambiar la variable de ambiente que indica el Timeout del DagBag Import pero no se pudo lograr una solución estable en el contenedor de Docker.

4. Propuesta de una arquitectura de solución

a. Propuesta de una arquitectura de solución

Queremos mencionar que algunas de las limitaciones que se dieron en el proyecto se debieron a la máquina virtual que se provee y la falta de almacenamiento y memoria principal que esta trae. Debido a esto se recomienda un despliegue en la nube en Microsoft Azure por medio de un contenedor y de su provisión de una base de datos PostgreSQL.

En temas de procesamiento, lo más pertinente en el futuro puede ser usar PySpark, obteniendo los datos dinámicamente a un Bucket S3 o Amazon Redshift y procesando con el mencionado PySpark en Amazon EMR. Esto, aunque posiblemente tenga un costo elevado, ayudaría a reducir problemas de tiempo y falta de recursos y seria una infraestructura extremadamente robusta. A partir de aquí se pueden consumir los datos usando PowerBI conectado a una base de datos remota en la nube.

b. Implementación de los tableros de control

Los tableros de control que se presentan fueron desarrollados a través de Power BI y como hubo problemas al momento de realizar el proceso ETL en la máquina virtual, los datos se obtienen de la copia local que guardan los archivos .pbix dentro de sí mismos.

- Tablero de control de las tasas de analfabetismo y cobertura neta de educación en Colombia
- Tablero de control de la tasa de deserción intra-anual y repitencia del sector oficial en educación básica y media en Colombia

5. Video

El video se encuentra en este enlace.

6. **Actividades realizadas**

- **Redacción de requerimientos analíticos (1 hora):** Felipe Bedoya, Camilo Salinas, Nicolás Orjuela
- **Modelado de Data Marts (1 hora):** Felipe Bedoya, Nicolás Orjuela
- **Entendimiento de las fuentes de datos (2 horas):** Felipe Bedoya
- **Diseño e implementación del proceso de ETL (6 horas):** Felipe Bedoya
- **Implementación de los tableros de control (5 horas):** Camilo Salinas, Nicolás Orjuela

Repartición de puntos (100):

- Felipe Bedoya: 33,3 puntos
- Camilos Salinas: 33,3 puntos
- Nicolás Orjuela: 33,3 puntos