Diciembre 2022
Universidad de Los Andes
Jesús David Barrios 201921887
Sergio Peñuela 201922873
Jhoan Diaz 201819861

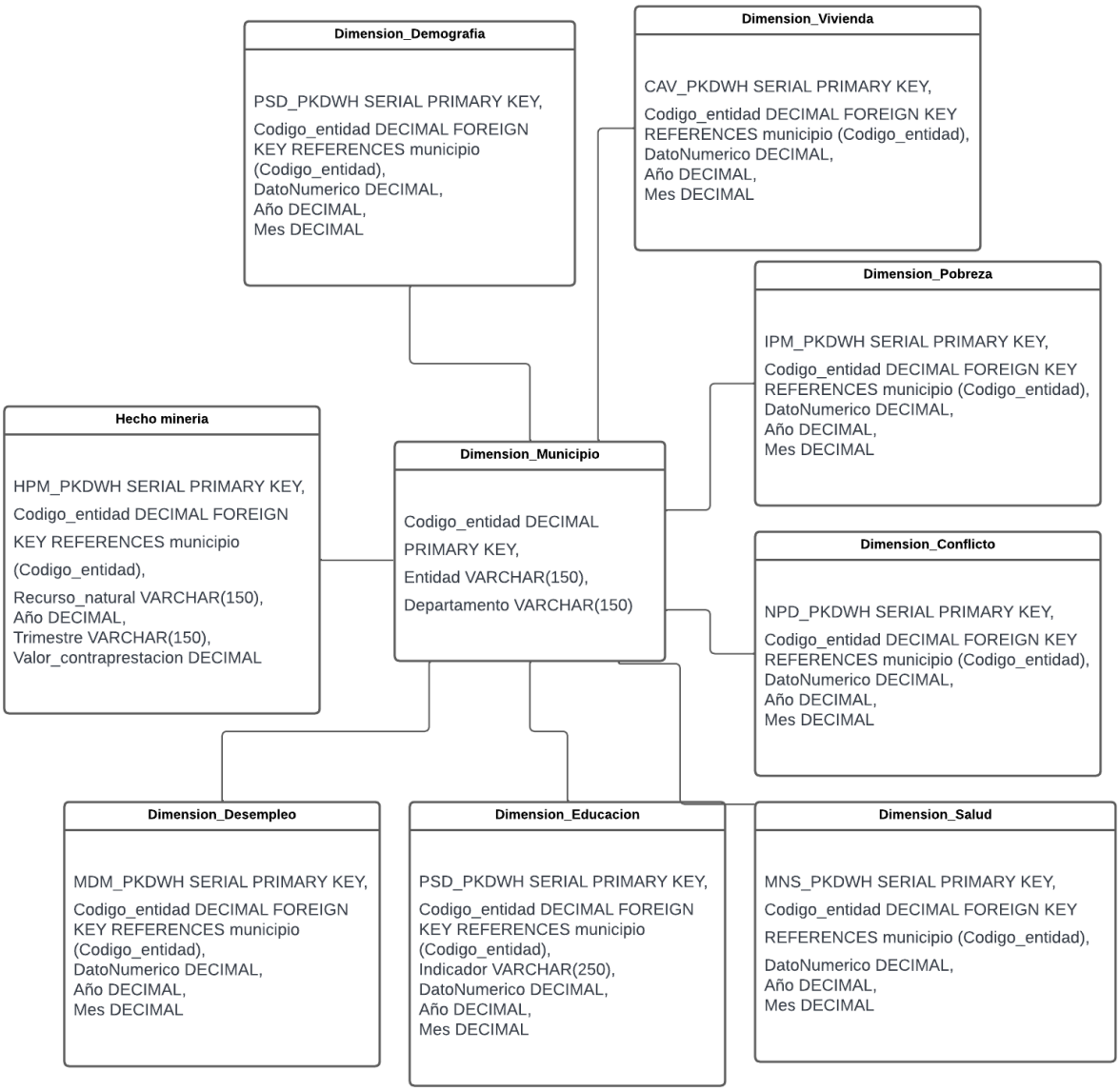
Proyecto 1 Inteligencia de negocios – Etapa 2

## Necesidades analíticas

| Tema analítico | Análisis requeridos o inferidos | Categoría del análisis - Tablero de control, análisis OLAP, Minería de datos | Procesos de negocio | Fuentes de datos y datos |
|---|---|---|---|---|
| **Impacto de la mineria en el nivel académico de los habitantes del departamento/municipío** | Desempeño general en puntajes icfes de los habitantes del departamento | análisis OLAP | Proyectos de innovación | ICFES |
| | Contraprestaciones de proyectos mineros en el departamento | | | Ministerio de minas |
| | | | | |
| **Impacto de la mineria en la salud de los habitantes** | Muertes por departamento (anual, mensual, etc) | Tablero de control | Mediciones e impacto de la mineria en Colombia | DANE |
| | Cantidad homicidios por departamento | | | DANE |
| **Como la demografia cambia dependiendo de la cantidad de proyectos mineros que haya en el departamento/municipio** | Puntaje de SISBEN promedio por departamento | OLAP | Mediciones e impacto de la mineria en Colombia | DANE |
| | Población total por departamento | | | DANE |
| | | | | |
| **Impacto de la mineria en la vivienda de los colombianos** | Cobertura de energía eléctrica | OLAP | Mediciones y proyectos | DNP |
| | Cobertura de alcantarillado | | | |
| **Impacto de la mineria en la pobreza** | IPM por departamento | OLAP | Mediciones y proyectos | DANE |
| | Población en condición de miseria | | | DANE |

## Modelo propuesto para Data Marts

Con el objetivo de dar respuesta a los objetivos de negocio, se plantea un modelo que permita fácil acceso y uso de los datos para el tablero de control. Se decidió incluir en el modelo todas las dimensiones de datos entregadas, y se escogió un indicador para cada una de ellas. Así, en las tablas de cada una de las dimensiones, se tienen solamente los valores de los indicadores escogidos. A continuación, se presenta el modelo dimensional utilizado para el proyecto:




En este caso la granularidad de la tabla de hechos es bastante baja, pues nos estamos enfocando en un proceso muy específico, ya que queremos ver si se logra encontrar alguna relación entre la producción minera y algún indicador de alguna de las dimensiones propuestas. Con esto, se busca ver qué patrones se identifican y poder utilizar esta información para los objetivos de negocio. En este caso, la tabla de hechos solo tiene la medida aditiva equivalente al valor de la contraprestación del hecho y una fecha de la medición.

En cuanto a la dimensión municipio, no se tienen medidas sino solamente atributos que permiten identificar el municipio por nombre, departamento y código. Esta dimensión, es referenciada por la tabla de hechos y por el resto de las dimensiones.

En cuanto a las otras dimensiones, tenemos una composición similar a la de la tabla de hechos. Estas están compuestas por una medida aditiva que establece el valor del indicador escogido y una fecha. La dimensión educación, tiene un atributo adicional dado el indicador que se escogió: puntaje del ICFES. Este atributo, permite saber si el valor es de matemáticas o de lectura crítica.

En cuanto al manejo de historia, para todas las dimensiones se utilizó el tipo 2. Esto implica, que, para cada municipio y dimensión, cada que se obtiene un nuevo dato para un indicador, se crea una nueva fila.

Se vuelve a resaltar, que en el modelo se priorizó la facilidad de acceso para el tablero de control.

**Entendimiento de los datos**

Para llevar a cabo el entendimiento de los datos cargamos los datos en un Jupyter Notebook para poder ver los datos que venían en cada excel y ver como estaban organizados, ver que columnas era de nuestro interés y ajustar los valores de los datos.

De igual manera, en términos de calidad de datos se verificó su completitud y duplicidad. En términos generales, los datos que recibimos no tuvieron problema en cuanto a métricas de calidad. Para los valores nulos que se encontraron, en caso de ser en columnas categóricas, se cambió el valor por 'N/A o sin datos', y para los numéricos, dada la baja cantidad de nulos, estos se remplazaron por la media o se eliminaron.

Para cada dimensión se eligió solamente el indicador con el que se decidió trabajar, por lo que, para reducir el tamaño de los datos procesados, todas aquellas columnas y registros que no tuvieran relación con el indicador y no fueran necesarias para su entendimiento fueron eliminadas y de esta manera reducir el tiempo que toma llevar a cabo el procesamiento de todos los datos. Además, se renombraron las columnas de los diferentes dataframes para mantener una homogeneidad entre las tablas y una congruencia con el modelo planteado.

**Diseño e implementación del proceso de ETL**

El diseño del ETL se encuentra en el Excel publicado en el repositorio y adjuntado a este documento. El enlace al Excel en el repositorio es el siguiente: https://github.com/JESUSDAVIDBARRIOS/repo_labs_BI/blob/main/proyecto2/Dise%CC%83o%20ETL.xlsx

**Arquitectura de solución**

Con el objetivo de dar respuesta a los objetivos de negocio, se plantea una arquitectura que permita fácil acceso y uso de los datos. Para esto, inicialmente, se propone crear un ETL que permita preparar los datos y organizarlos en una base de datos de fácil acceso. Para la base de datos se utiliza PostgreSQL la cual tiene múltiples ventajas de desempeño, y además de accesibilidad y costos al ser de código abierto. Por su parte, para el ETL, se utiliza Apache AirFlow, el cual permite una fácil conexión con la base de datos y la creación de los flujos de trabajo de manera intuitiva.

Con esto implementado y desplegado, se puede acceder a los datos desde el tablero de control. Cabe resaltar, que para este proyecto no se realizó el despliegue de la

base de datos, por lo que, para este ejercicio, se utilizaron los CSV generados por el ETL para crear el tablero de control. Esto, dado que no se va a mantener un servidor que ejecute la aplicación sin haberlo presentado primero al cliente.

Para el tablero de control, se utilizó Power BI, dado que es una herramienta de fácil conexión y difusión que puede ser utilizada por el cliente. Con el objetivo de mostrar la mayor cantidad de información en el tablero de control y la relación de las dimensiones con la minería, se diseñó un tablero de control con una página para cada dimensión. En cada una de las páginas, se presentan los valores de las contraprestaciones por municipio y departamento, acompañados de los valores del indicador escogido para cada dimensión. Además, se incluye la opción de filtrar por años y por departamento. Para el caso de la dimensión de educación, también se incluye un filtro para escoger entre matemáticas y lectura crítica dado que se escogió el puntaje del ICFES como indicador

Para cada una de las páginas, se presentan dos gráficos: (1) un mapa que permite ver el valor de la contraprestación y el valor del indicador de la dimensión para cada entidad territorial; y (2) una gráfica de barras que agrupa los datos por departamento y enfrenta el valor de la dimensión con el de minería. Se puede acceder al tablero de control implementado en el siguiente enlace: https://app.powerbi.com/links/UNWhcAOt5n?ctid=fabd047c-ff48-492a-8bbb-8f98b9fb9cca&pbi_source=linkShare&bookmarkGuid=5f3193bd-fad1-4c7d-8871-b361a51a40c6

**Actividades realizadas**

En cuanto a la repartición de 100 puntos por el trabajo, se considera que estos pueden ser repartidos de manera equitativa entre los miembros del grupo.

- Jesús David Barrios

  El estudiante tuvo el rol de líder de datos y líder de negocio, por lo que se encargó de estar alineado con el problema de negocio a lo largo del proyecto y de gestionar la preparación y manejo de datos. En cuanto a actividades, el estudiante realizó el entendimiento y procesamiento de datos inicial y participó en la implementación del tablero de control. Para esto, se tuvo una dedicación de alrededor de 7 horas y media.

- Sergio Peñuela

  El estudiante tuvo el rol de líder de proyecto. En este caso la mayoría se cuadró por medio del grupo de whatsapp de los integrantes del grupo y se tuvo una reunión de trabajo intensa el fin de semana. En el caso de las actividades el estudiante realizó la implementación del ETL y participó en la implementación del tablero de control. Para esto, se dedicaron alrededor de 7 horas.

- Jhoan Diaz

Jhoan tuvo el rol del líder de analítica. En este caso se encargó de diseñar el ETL del proyecto y participó en la implementación del tablero de control. Para esto, se tuvo una dedicación de alrededor de 7 horas y media.

El modelo multidimensional se diseñó en conjunto.