Tecnicatura Universitaria en Procesamiento y Explotación de Datos

Bases de Datos Multidimensionales

Trabajo Integrador Final

Alumno:Rodrigo Zamora

Profesores

Elias Walter, Fernandez Maximiliano

Introducción

Mediante una base de datos que proviene del Sistema Único de Boleto Electrónico (SUBE), que se implementó en el Área Metropolitana de Paraná. Se realizará un datawarehouse con el fin de poder almacenar toda esta información, generando métricas que permitan obtener nuevos conocimientos a partir de los datos.

Situación problemática

Los servicios de colectivos son un medio de transporte público que se encarga de brindar servicios de transporte a pasajeros.

Existe el problema de las "horas pico" en el transporte público, es una preocupación común en las ciudades. Se refiere a los momentos del día en los que se produce una mayor concentración de pasajeros en los colectivos, lo que puede tener varias implicaciones y desafíos. Los colectivos a menudo experimentan una sobrecarga de pasajeros. La capacidad de los vehículos puede superarse, lo que resulta en un viaje incómodo y a veces inseguro para ellos. La sobrecarga y los retrasos durante estas horas pueden afectar la eficiencia del sistema de transporte público en su conjunto.

La cantidad de movimiento de personas en los colectivos, implica el manejo de importantes cifras de dinero a diario, por esa razón es necesario gestionar eficazmente las finanzas relacionadas con el transporte público.

A partir de estos problemas se planteó la siguiente pregunta:¿Qué patrones y tendencias se pueden identificar en la base de datos de sube en el de abril en relación con la frecuencia, tipo de contrato y número de colectivo de los viajes para mejorar el sistema de transporte de Paraná en las zonas urbanas?

Objetivo

El presente proyecto tiene como objetivo demostrar la importancia de una base de datos multidimensional en la gestión y planificación del transporte público, utilizando datos de alta calidad del sistema SUBE. Este enfoque puede proporcionar soluciones prácticas y valiosas para mejorar la calidad de vida de los ciudadanos y optimizar la operación de los servicios de colectivos en el Área Metropolitana de Paraná.

Estado del arte

El presente estado del arte proporciona una visión general de las investigaciones y desarrollos recientes en el campo de la optimización y sostenibilidad de la movilidad en el sistema de transporte público (TP) de Paraná. Este análisis se basa en el artículo titulado "Evaluación de la Relación Oferta-Demanda en el Sistema de Transporte Público de Paraná: Enfoque en la Optimización y Sostenibilidad de la Movilidad" de Juan Francisco Jaurena, Rafael Díaz Arias, Walter Elías y Joaquín Lambarri de la Facultad de Ingeniería de la Universidad Nacional de Entre Ríos. El artículo investiga la relación entre la oferta y la demanda del TP en Paraná y destaca la importancia de equilibrar estas dos variables para lograr una operación eficiente y sostenible del sistema.
El artículo aborda la relevancia de optimizar el sistema de transporte público como un medio para reducir la congestión vehicular, disminuir la contaminación ambiental y mejorar la satisfacción de los usuarios. Además, señala que la optimización puede contribuir a la política de reducción de vehículos particulares en favor del TP.

Se destaca que uno de los principales desafíos en la operación del TP es gestionar las horas pico, cuando la demanda es significativamente mayor. El artículo subraya que la falta de equilibrio entre la oferta y la demanda en estas horas puede dar lugar a aglomeraciones y disminución de la calidad del servicio.

Patrones de Uso en Horas Pico
El análisis revela patrones de uso en las horas pico, con una explicación detallada sobre cuántas tarjetas están asociadas con estas franjas horarias de alta demanda y la implicación de esta información para la operación eficiente del sistema de TP.

Marco teórico

Un data warehouse, o almacén de datos, es una estructura de almacenamiento de información diseñada para la consolidación, organización y gestión de datos empresariales con el propósito de facilitar el análisis y la toma de decisiones. Es una parte fundamental de la inteligencia empresarial (BI, por sus siglas en inglés) y se utiliza para reunir datos de diversas fuentes, transformarlos en un formato consistente y almacenarlos de manera eficiente para su posterior análisis.

Conceptos importantes:

Consolidación de datos: recopilan datos de múltiples fuentes, como sistemas transaccionales, hojas de cálculo, bases de datos, registros de ventas, y más. Estos datos pueden ser heterogéneos y estar distribuidos en diferentes formatos y ubicaciones.
Organización y estructura: Los datos se organizan de manera coherente y se almacenan en una estructura diseñada para facilitar su consulta y análisis. Los datos se transforman en un formato que es útil para los analistas y los tomadores de decisiones.
Acceso y consulta eficientes: Los sistemas de gestión de bases de datos utilizados en los data warehouses están optimizados para consultas complejas y análisis de datos. Esto garantiza

que los usuarios puedan acceder a la información de manera eficiente, incluso cuando se trata de grandes volúmenes de datos.
Apoyo a la toma de decisiones: Los data warehouses permiten a los usuarios realizar análisis de datos, generar informes y visualizaciones, y obtener información valiosa para la toma de decisiones estratégicas en la empresa.
Integración con herramientas de BI: Los data warehouses a menudo se integran con herramientas de inteligencia empresarial, como Tableau, Power BI o QlikView, para facilitar la creación de informes y paneles interactivos.
No Volatilidad: La información almacenada no se modifica ni se elimina; una vez que se almacena, se convierte en información de solo lectura y se mantiene para futuras consultas.

En el diseño de un Data Warehouse, se distinguen tres etapas: conceptual, lógica y física. El diseño conceptual se centra en la estructura abstracta de datos, incluyendo hechos y dimensiones. El diseño lógico se traduce en un modelo de datos relacional con tablas y relaciones. El diseño físico se enfoca en la implementación concreta en hardware y software, considerando la distribución de datos y el rendimiento.

Elección de modelo kimball

Adaptabilidad a necesidades específicas: Su enfoque en la creación de data marts brinda la flexibilidad necesaria para ajustar la estructura de los datos a las características únicas de mi proyecto y a los requisitos específicos de análisis. Esto me permite diseñar data marts específicos que se adapten a las necesidades precisas de cada departamento o área de estudio involucrada en mi investigación.
Flexibilidad en el diseño de data marts: El modelo Kimball ofrece una flexibilidad excepcional en el diseño de data marts. Puedo definir dimensiones y medidas de acuerdo con las necesidades cambiantes de mi proyecto y ajustarlas a medida que avanzo en mi investigación. Esta flexibilidad me facilita la exploración de datos y la adaptación a las particularidades de mi proyecto sin estar limitado por una estructura de datos centralizada rígida.
Rapidez en la obtención de resultados: La necesidad de obtener resultados de manera rápida es fundamental en mi proyecto de investigación. El modelo Kimball destaca en la entrega rápida de soluciones, lo que me permite implementar data marts de manera ágil y comenzar a analizar datos sin demoras innecesarias.

Diseño e implementación

La base de datos es un registro de las transacciones realizadas con las tarjeta sube, tiene un total de 1698204 filas y 44 columnas de las cuales voy a utilizar:
IDLINEA: proporciona el número del colectivo, por ejemplo si es un 6, 10 o 15 etc.
CODIGOCONTRATO: tiene el tipo de contrato que tiene la tarjeta, jubilado, estudiante entre otros.
MONTO: el pago que realiza la tarjeta al subir al colectivo.
FECHATRX: la fecha que indica año, mes, dia, hora, minutos y segundos en la que se realiza la transacción.
DESCUENTO: el descuento que se realiza.

Con las columnas seleccionadas del dataset se utilizaron para dimensiones las variables FECHATRX, CODIGOCONTRATO y IDLINEA. Las otras columnas, DESCUENTO y MONTO son utilizadas para las métricas.
Con las columnas seleccionadas del dataset se realizarán las siguientes dimensiones

Tiempo
- id_tiempo PK
- fecha
- mes
- dia_semana
- hora

Contrato
- id_contrato PK
- tipo

Línea
- id_linea PK
- número

por último se realizará la tabla de hechos
Hechos
- id_hechos PK
- id_Tiempo FK
- id_Contrato FK
-
- id_Línea FK
- cantidad (métrica)
- monto (métrica)
- descuento (métrica)

con los anteriores datos se van a realizar las siguientes métricas:

1. Cantidad de transacciones
2. cantidad de dinero(monto)
3. cantidad de descuento

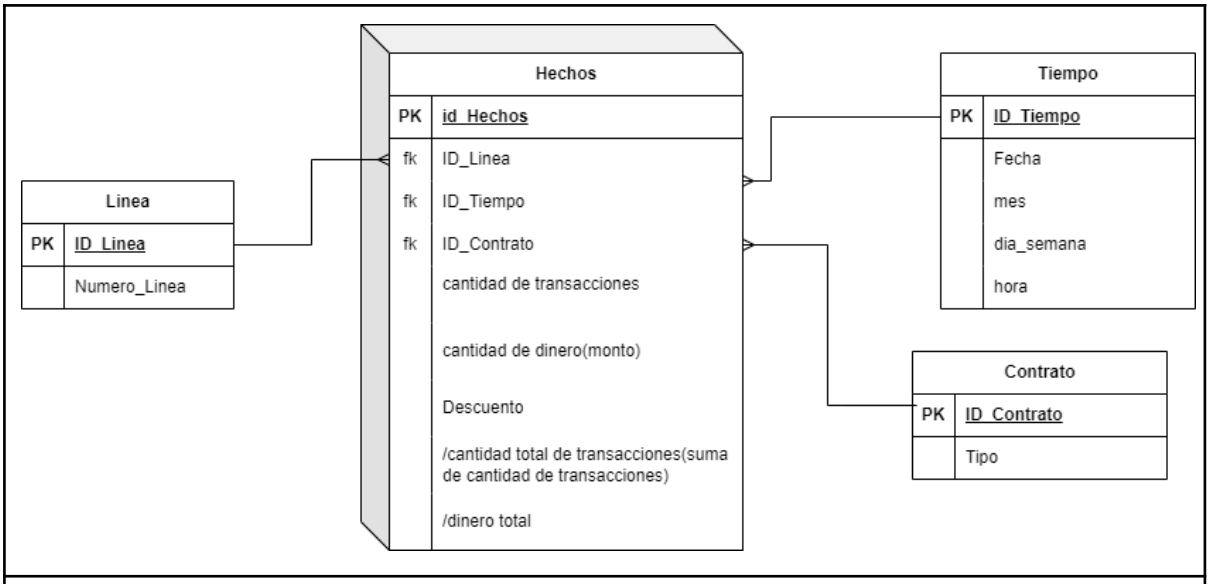


Diagrama conceptual del modelo multidimensional del DW

A partir del diagrama conceptual del modelo multidimensional del DW se genera el DDL para SQL el cual da como resultado el diagrama de tabla que se presenta a continuación

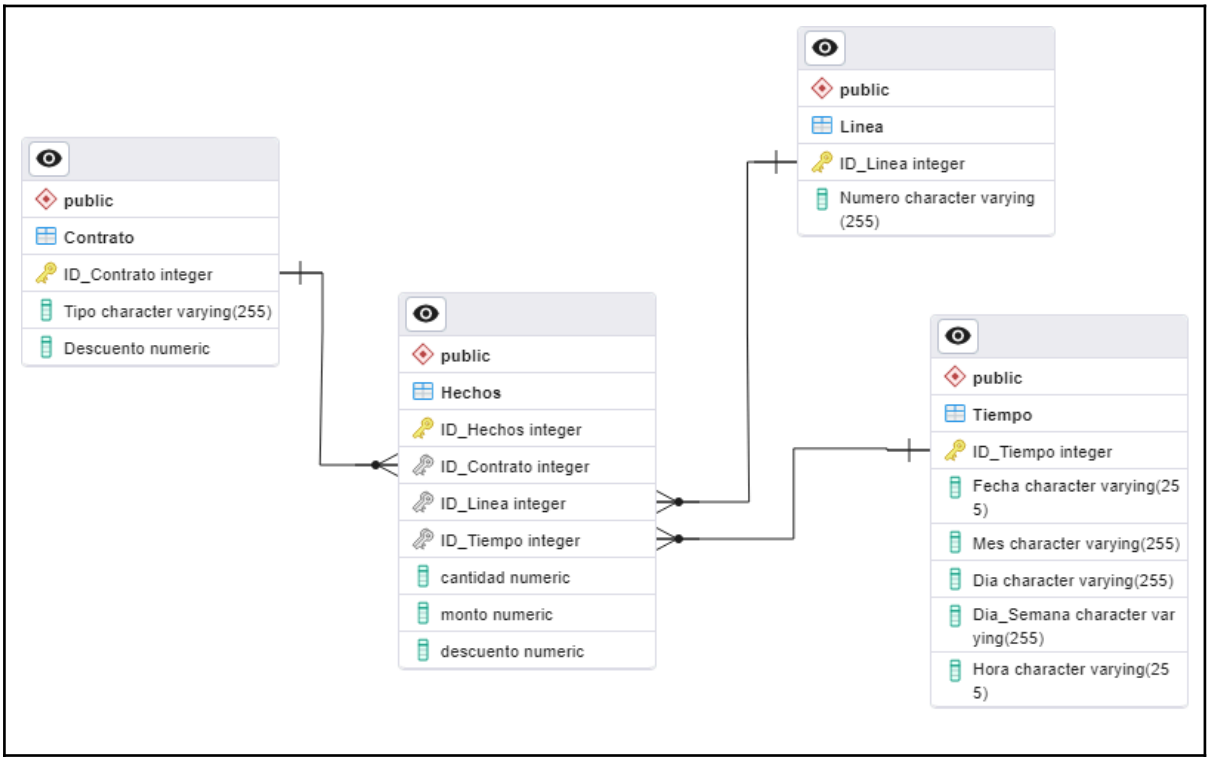


Diagrama de tablas de la implementación en SQL del DW

Lo que se puede ver en el diagrama es que el datawarehouse es de tipo estrella, tomé esta decisión ya que las relaciones entre las tablas de dimensiones y la tabla de hechos son directas, lo que facilita las consultas y el acceso a los datos, también debido a su estructura simple y denormalizada, las consultas en un modelo de estrella tienden a ser más rápidas y eficientes.

Cabe aclarar que en la dimensión tiempo, Fecha tiene los datos de la columna FECHATRX de la bases de datos SUBE en cambio mes, dia, dia_semana y hora son creadas a partir de dividir FECHATRX por mes/dia/hora

ETL

ETL es un acrónimo que significa "Extract, Transform, Load", y se refiere a un proceso comúnmente utilizado en la gestión de datos y la informática para mover y transformar datos desde una fuente de origen hasta un destino donde puedan ser almacenados, analizados y utilizados de manera efectiva.
Primero realice la extracción de los datos que eran importantes de la base de datos SUBE, las cuales fueron las columnas IDLINEA, CODIGOCONTRATO, MONTO, FECHATRX y DESCUENTO.
Segundo hice la transformación de los datos, en esta parte tuve que modificar CODIGOCONTRATO ya que los diferentes tipos de contratos estaban en formato numérico, imposibilitando entender a qué tipo de contrato hacían referencia, gracias al diccionario de datos proporcionado con la bases de datos se pudieron transformar para saber qué número hacía referencia a un tipo de contrato específico, el mismo trabajo realice con IDLINEA.
Diccionario de datos utilizado para la transformación de las variables IDLINEA y CODIGOCONTRATO.

522	jubilado
523	obrero
524	emp.publico
525	universitario
526	secundario
527	primario
528	secundario AM
529	primario AM
530	gentileza
531	ord. 9238
621	tarifa social
602	tarifaplana

Diccionario de datos de IDLINEA

ID Cifrado	ID Verdadero
1203	1
1221	2
1222	3
1223	4
1224	5
1225	6
1226	7
1227	8
1228	9
1229	10
1230	11-21
1231	12
1232	14
1233	15
1234	20

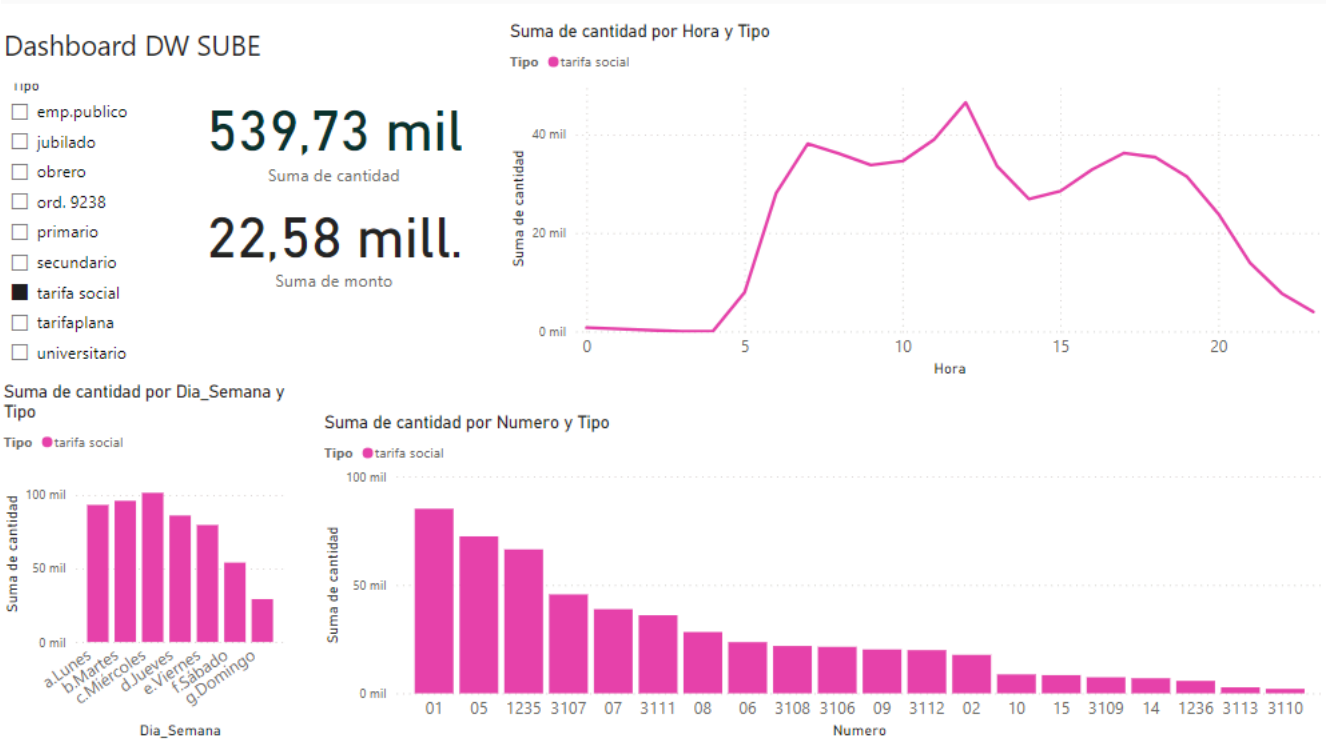
Diccionario de datos de CODIGOCONTRATO

Con la columna FECHATRX genere más variables, estas son nombre_dia(indica si es lunes, martes, miércoles etc) las otras son Dia, Hora y Mes.
En la etapa cargado. primero conecto desde python donde tengo la bases datos a Postgres para cargarlos, comienzo a cargar las dimensiones, para esto primero creo un Data Frame para cada una y con sus datos correspondientes,al principio tuve problemas con esto ya que los nombres de las columnas del dataframe tienen que ser iguales al de las dimensiones, también el tipo de dato tiene que ser el mismo en las dimensiones y en el dataframe. Los datos fueron cargados utilizando una función dada por la cátedra que actualiza una tabla de dimensión de un DW con los datos nuevos.
En la dimensión Línea se cargan los valores únicos de la columna IDLINEA y se le asigna un id a cada uno, lo mismo con la dimensión Contrato donde se cargan los valores únicos de la columna CODIGOCONTRATO.

Lo que se carga a la tabla de Hechos son los datos que incluyen información de contratos, líneas y fechas de transacciones, también la cantidad de dinero involucrada en esas transacciones. El código primero agrupa estos datos por contrato, línea y fecha, y luego suma el dinero para cada grupo y también cuenta la cantidad de transacciones para cada grupo. Por lo tanto, para cada combinación única de contrato, línea y fecha, se obtiene la suma total del dinero involucrado y la cantidad de transacciones.

Para finalizar realice un dashboard en Power BI de los datos, en la cual podemos ver que en la parte superior izquierda aparecen los tipos de contratos, esto funciona como filtro que dependiendo el tipo de contrato que seleccionemos ese va a ser el que se muestre en los gráficos. Arriba a la derecha podemos ver un gráfico de línea que muestra la cantidad de transacciones del contrato seleccionado en las distintas horas del día, abajo a la derecha el gráficos de barras muestra la cantidad de transacciones del contrato en las diferentes líneas

de colectivo, por último el otro gráfico de barras tiene la cantidad de transacciones del contrato por dia de la semana.



conclusión

Este proyecto de Data Warehouse basado en el modelo Kimball ha abordado con éxito la gestión de datos relacionados con el transporte público en el Área Metropolitana de Paraná. Ha demostrado cómo la implementación de un Data Warehouse puede ayudar a resolver problemas relacionados con las horas pico y la eficiencia en la gestión del transporte público. La visualización de datos en Power BI proporciona una herramienta valiosa para la toma de decisiones basada en información concreta.