Informe sobre el Diseño y Modelado de una Base de Datos Multidimensional para Análisis de Tráfico en Paradas de Colectivo

Este tiene como finalidad un informe detallado del diseño de una base de datos multidimensional diseñada para la recolección y análisis de datos provenientes de cámaras sensoras ubicadas en paradas de colectivos. Las cámaras tienen la capacidad de detectar y contar personas esperando permitiendo calcular métricas relacionadas con el tiempo de espera y la afluencia de personas en las paradas. La base de datos multidimensional está estructurada alrededor de una tabla de hechos y 5 dimensiones, facilitandonos la generación de indicadores útiles para la toma de decisiones en la gestión del transporte público.

Métricas Relevantes

Para el análisis de datos de tráfico en paradas de colectivos, se han definido las siguientes métricas clave:

1. **Cantidad de Personas en Parada:** Un conteo del número de personas detectadas por las cámaras en cada parada de colectivos, lo cual nos permite evaluar la demanda de usuarios en distintos horarios y lugares.
2. **Tiempo Promedio de Espera (Avg Tiempo de Espera):** El tiempo promedio que un usuario pasa esperando en una parada de colectivos antes de abordar un colectivo. Esta métrica se calcula restando la hora de llegada de la persona y la hora de partida.

Tabla de Hechos: `Personas_esperando_en_parada`
La tabla de hechos centraliza los datos cuantitativos que se desea analizar. Está compuesta por las siguientes métricas y los identificadores de cada dimensión.

Campo	Descripción
`cantidad_personas`	Número de personas en la parada.
`avg_tiempo_espera`	Tiempo promedio de espera de usuarios.

id_parada	Identificador de la parada de colectivo
id_cliente	Identificador de tipo de cliente (usuario).
id_clima	Identificador de condición climática.
id_colectivo	Identificador del colectivo abordado.
hora, dia, mes, año	Datos temporales detallados del evento.

Dimensiones

Para enriquecer el análisis y dar contexto a los datos de la tabla de hechos, se han diseñado cinco dimensiones: Parada, Fecha, Cliente, Clima y Colectivo.

1. **Dimensión Parada:** Proporciona información sobre la ubicación de cada parada.
    - Campos: Nombre_Parada, Barrio, Localidad, Provincia.
2. **Dimensión Fecha:** Desglosa los datos temporales para permitir análisis detallados según la hora del día, día, mes y año.
    - Campos: Hora, Día, Mes, Año.
3. **Dimensión Cliente:** Segmenta los datos según atributos de los usuarios de colectivos.
    - Campos: Genero, Tipo_de_Cliente.
4. **Dimensión Clima:** Indica las condiciones climáticas presentes durante el evento de espera en la parada.
    - Campo: Condición (registrado a través de un proceso ETL con datos meteorológicos históricos).
5. **Dimensión Colectivo:** Contiene información relevante del vehículo (colectivo) utilizado.
    - Campos: Marca, Antigüedad, Matrícula.

Diagrama Entidad-Relación (ER) y Diagrama de tablas

A continuación, se presenta el diseño del modelo ER y del diagrama de tablas que representa la estructura de la base de datos multidimensional. Este diagrama ilustra las relaciones entre la tabla de

hechos Personas_esperando_en_parada y las dimensiones asociadas (Parada, Fecha, Cliente, Clima, Colectivo).

Figura 1. Diagrama Entidad-Relación (ER)

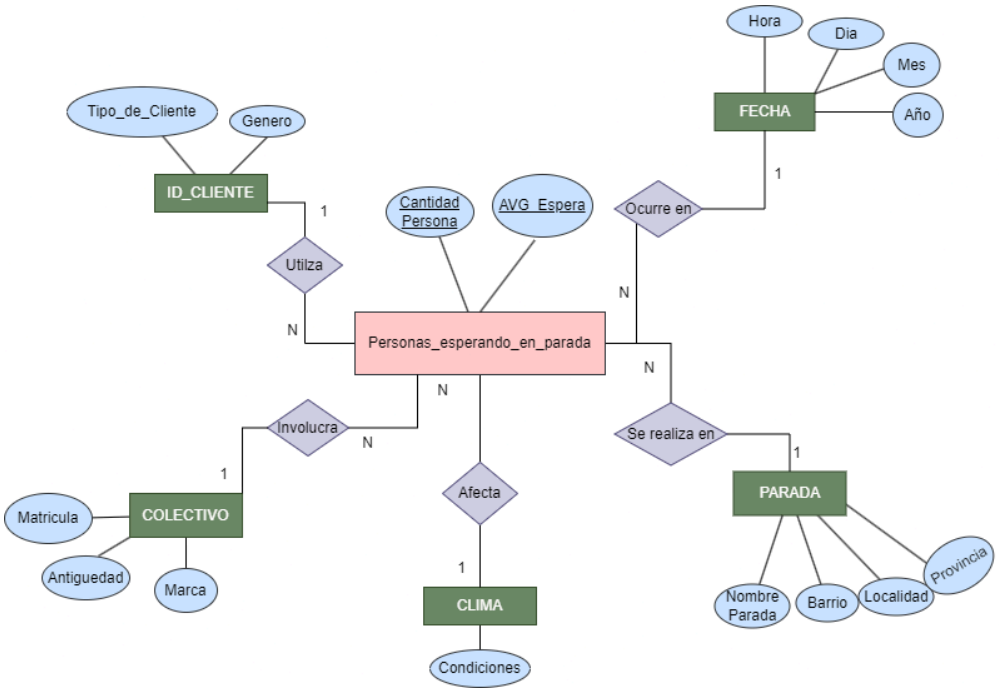
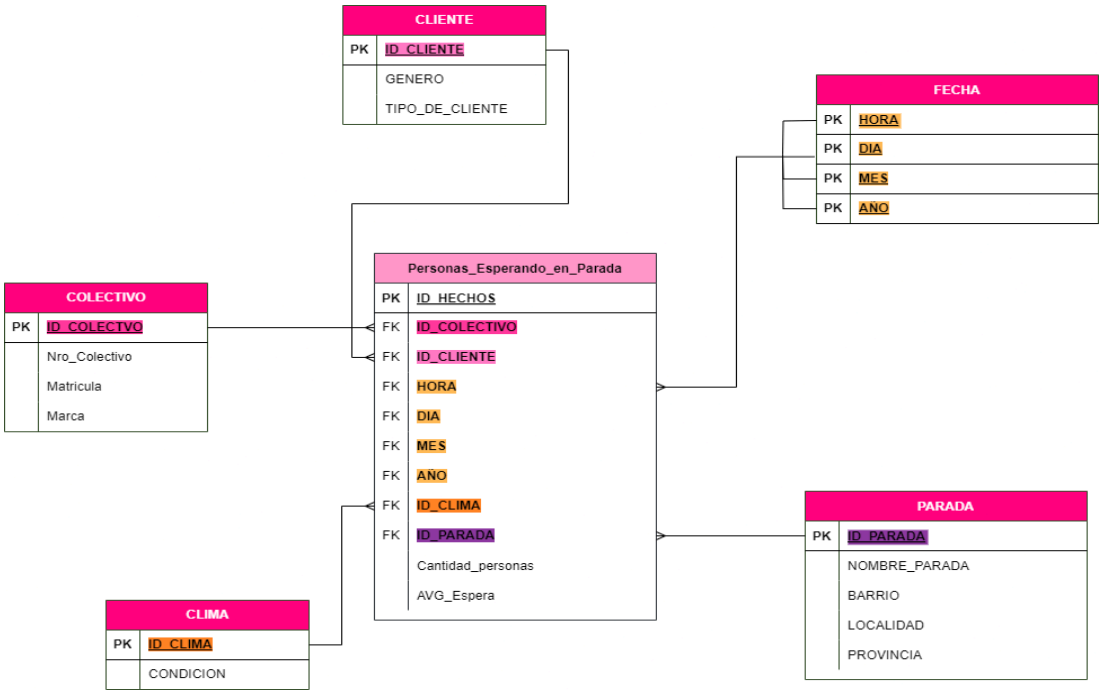


Figura 2. Diagrama de tablas



DDL (Data Definition Language)

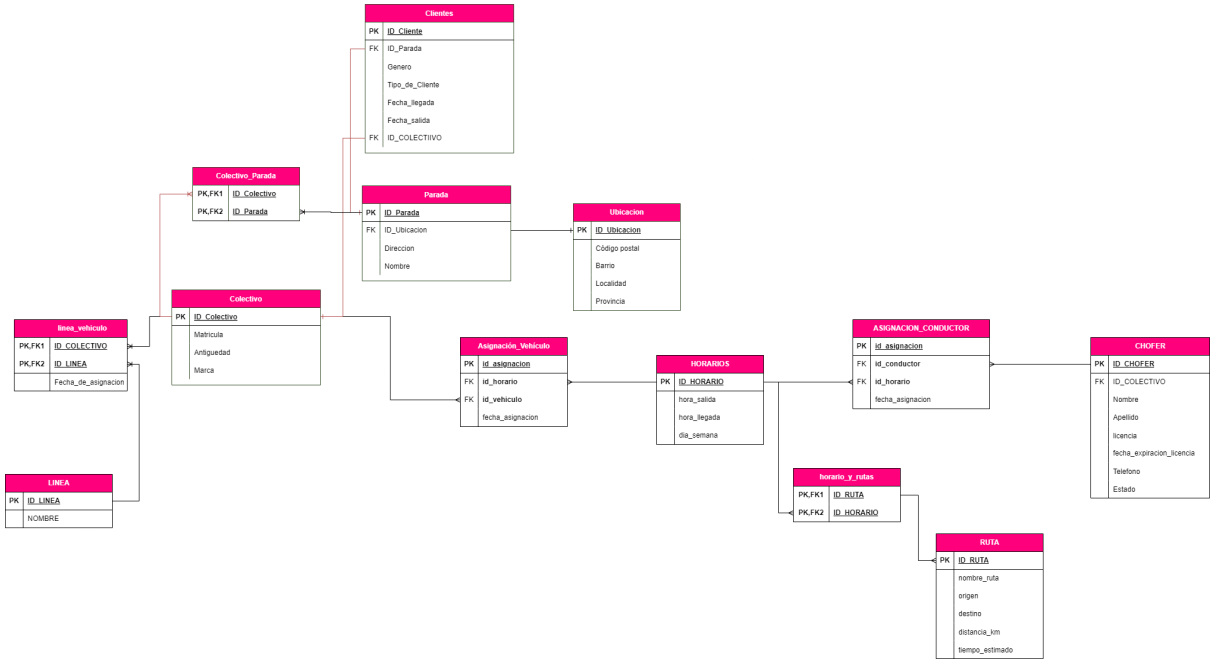
El siguiente script DDL define las estructuras de las tablas en el modelo multidimensional:

https://drive.google.com/file/d/1Vi48gh-3sGFF1e1O6epW8z2Nw2R9GfaE/view?usp=drive_link

Diagrama de tablas de la base de datos transaccional (Fuente de datos)

A continuación, se presenta el diagrama de tablas que representa la estructura de la base de datos transaccional.

Figura 3. Diagrama de tablas



DDL (Data Definition Language)

El siguiente script DDL define las estructuras de las tablas de la base de datos transaccional:

https://drive.google.com/file/d/1n11XzzpYfXn6hS1I9IyqAo15irO_tXrS/view?usp=sharing

Consultas de Inserción

Las siguientes consultas SQL permiten la carga de datos desde una base de datos transaccional hacia las tablas de dimensiones. Estas consultas seleccionan y transfieren datos únicos para evitar redundancias y asegurar la calidad de la información.

```
INSERT INTO "CuboMultidimensional".dim_cliente (tipo_cliente)
SELECT DISTINCT Tipo_de_Cliente
FROM transaccional.Clientes;

INSERT INTO "CuboMultidimensional".dim_parada (parada, barrio, localidad, provincia)
SELECT DISTINCT
p.Nombre AS parada,
u.Barrio AS barrio,
u.Localidad AS localidad,
u.Provincia AS provincia
FROM transaccional.Parada p
JOIN transaccional.Ubicacion u ON p.ID_Ubicacion = u.ID_Ubicacion;
INSERT INTO "CuboMultidimensional".dim_colectivo (numero_colectivo, matricula)
SELECT DISTINCT CAST(li.nombre AS VARCHAR), c.Matricula
FROM transaccional.Colectivo c
JOIN transaccional.linea_vehiculo l ON c.id_colectivo = l.vehiculo_id
JOIN transaccional.linea li ON l.linea_id = li.id_linea;

INSERT INTO "CuboMultidimensional".dim_fecha (hora, dia, mes, ano)
SELECT DISTINCT
  EXTRACT(HOUR FROM Fecha_llegada)::integer AS hora,
  EXTRACT(DAY FROM Fecha_llegada) AS dia,
  EXTRACT(MONTH FROM Fecha_llegada) AS mes,
  EXTRACT(YEAR FROM Fecha_llegada) AS ano
FROM transaccional.Clientes;
```

La dimensión clima no es mostrada porque es generada a través del proceso ETL.
Link hacia el proceso ETL.
https://drive.google.com/file/d/1lKRWEsRQ1pei-PorvOm0x74B8xaWkjVl/view?usp=sharing

Carga de Datos en la Tabla de Hechos

Para completar la carga de datos en la tabla de hechos, se utiliza una consulta que obtiene las métricas calculadas a partir de las tablas transaccionales y las dimensiones correspondientes:

```
SELECT
AVG(EXTRACT(EPOCH FROM (c.Fecha_salida - c.Fecha_llegada)) / 60) AS avg_tiempo_espera,
  COUNT(c.ID_Cliente) AS cantidad_personas_en_parada,
  NULL AS id_clima,
  dc.id_cliente,
  co.id_colectivo,
  dp.id_parada,
  dp.localidad,
  df.hora, df.dia, df.mes, df.ano
FROM transaccional.Clientes c
JOIN "CuboMultidimensional".dim_cliente dc ON dc.tipo_cliente = c.Tipo_de_Cliente
JOIN transaccional.Parada p ON p.ID_Parada = c.ID_Parada
JOIN "CuboMultidimensional".dim_parada dp ON dp.parada = p.nombre
JOIN transaccional.Colectivo co ON co.ID_Colectivo = c.colectivo_id
JOIN "CuboMultidimensional".dim_fecha df
  ON df.dia = EXTRACT(DAY FROM c.Fecha_llegada)
  AND df.mes = EXTRACT(MONTH FROM c.Fecha_llegada)
  AND df.ano = EXTRACT(YEAR FROM c.Fecha_llegada)
  AND df.hora = EXTRACT(HOUR FROM c.Fecha_llegada)
JOIN transaccional.linea_vehiculo lv ON c.colectivo_id = lv.vehiculo_id
-- Aquí se usa una subconsulta correlacionada para obtener la fecha_de_asignacion más cercana
WHERE lv.fecha_de_asignacion = (
  SELECT fecha_de_asignacion
  FROM transaccional.linea_vehiculo lv_sub
  WHERE lv_sub.vehiculo_id = c.colectivo_id
        ORDER BY ABS(EXTRACT(EPOCH FROM (c.fecha_salida - lv_sub.fecha_de_asignacion)))
  LIMIT 1
)
GROUP BY co.id_colectivo, dp.id_parada, dc.id_cliente, df.hora, df.dia, df.mes, df.ano;
```

Código ETL en Python para Inserción en la Tabla de Hechos

En el siguiente fragmento de código ETL en Python, se utiliza un `DataFrame` para almacenar los datos de métricas procesadas y luego insertarlas en la tabla de hechos:

```
for _, row in df_clientess.iterrows():
    cursor.execute("""
        INSERT INTO "CuboMultidimensional".fact_transporte (
            avg_tiempo_espera,
            cantidad_personas_en_parada,
            id_cliente,
            id_clima,
            id_colectivo,
            id_parada,
            hora,
            dia,
            mes,
            ano
        ) VALUES (%s, %s, %s, %s, %s, %s, %s, %s, %s, %s);
    """, (
        row['avg_tiempo_espera'],
        row['cantidad_personas_en_parada'],
        row['id_cliente'],
        row['tipo_de_clima'],
        row['id_colectivo'],
        row['id_parada'],
        row['hora'],
        row['dia'],
        row['mes'],
        row['ano']
    ))
```