Pipeline de datos ETL para Análisis de Viajes de Uber en NYC (2014)

Resumen Ejecutivo

Este proyecto establece una robusta canalización de datos ETL (Extracción, Transformación, Carga) para analizar los datos de viajes de Uber en la ciudad de Nueva York durante el período de abril a septiembre de 2014. Utilizando una arquitectura Medallón (Bronze, Silver, Gold) implementada en Databricks Free Edition, la solución ingesta datos crudos, los limpia y enriquece, y finalmente los agrega en un formato optimizado para el análisis de negocio. El objetivo principal es proporcionar insights accionables sobre patrones de demanda, rendimiento de bases y distribución geográfica de los viajes, permitiendo a los equipos de ciencia de datos y analistas de negocio tomar decisiones informadas.

1. Contexto del Negocio y Objetivos

1.1. Problema de Negocio

La gestión eficiente de una plataforma de transporte como Uber requiere una comprensión profunda de los patrones de demanda y oferta. Sin un análisis de datos estructurado, Uber NYC podría enfrentar desafíos como:

- **Ineficiencia Operacional:** Desconocimiento de las horas y ubicaciones de mayor demanda, llevando a una asignación subóptima de conductores.
- **Experiencia del Usuario Subóptima:** Tiempos de espera prolongados o falta de disponibilidad en horas pico.
- **Planificación Estratégica Deficiente:** Dificultad para identificar oportunidades de crecimiento o áreas de mejora en el servicio.

1.2. Preguntas de Negocio Hipotéticas a Responder

Este proyecto busca responder preguntas clave para la operación y estrategia de Uber NYC, tales como:

- ¿Cuáles son las horas pico de demanda de Uber en NYC y cómo varían por día de la semana?
- ¿Cómo se distribuye la demanda de viajes geográficamente en NYC? ¿Existen "hotspots" de recogida?
- ¿Cuál es el rendimiento de cada base de Uber en términos de volumen de viajes y actividad?
- ¿Cómo ha evolucionado la demanda de viajes mes a mes durante el período analizado (abril-septiembre de 2014)?
- ¿Existen patrones de demanda distintivos entre días de semana y fines de semana?

1.3. Objetivos del Proyecto

- **Construir una Canalización de Datos Confiable:** Desarrollar una pipeline ETL automatizada y escalable para procesar datos de viajes de Uber.
- **Proporcionar Datos de Alta Calidad:** Asegurar que los datos transformados sean limpios, consistentes y enriquecidos para el análisis.
- **Generar Insights de Negocio:** Crear un modelo de datos que facilite la extracción de métricas clave y la identificación de patrones de demanda.
- **Demostrar Habilidades Técnicas:** Servir como un portafolio profesional que muestre experiencia en ingeniería de datos, especialmente con Databricks y Delta Lake.

1.4. Descripción del Conjunto de Datos

El proyecto se basa en el conjunto de datos 'Uber Pickups in New York City' de Kaggle, que se centra específicamente en las recogidas de Uber en la ciudad de Nueva York durante el período de abril a septiembre de 2014.

Descripción General

- **Periodo de los Datos:** Abril a septiembre de 2014.
- **Volumen de Datos:** Más de 4.5 millones de registros de recogidas de Uber.
- **Formato de los Archivos:** Cada mes está contenido en un archivo CSV separado.
- **Ubicación:** Ciudad de Nueva York.

Diccionario de Datos

Los archivos mensuales contienen las siguientes columnas:

Columna	Descripción
Date/Time	La fecha y hora exactas en las que se realizó la recogida del viaje de Uber.
Lat	La latitud geográfica del punto de recogida.
Lon	La longitud geográfica del punto de recogida.
Base	El código de la base de la empresa afiliada a la recogida de Uber, según la Comisión de Taxis y Limusinas de Nueva York (TLC).

Archivos

Los nombres de los archivos correspondientes a cada mes son:

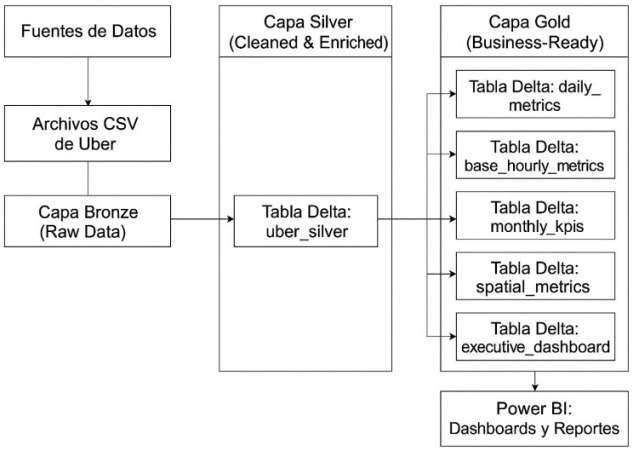
- uber-raw-data-apr14.csv
- uber-raw-data-aug14.csv
- uber-raw-data-jul14.csv
- uber-raw-data-jun14.csv
- uber-raw-data-may14.csv
- uber-raw-data-sep14.csv

Fuente: https://www.kaggle.com/datasets/fivethirtyeight/uber-pickups-in-new-york-city

2. Arquitectura de la Solución: Enfoque Medallón

La solución implementa una arquitectura de lago de datos con el enfoque Medallón, que organiza los datos en tres capas distintas: Bronze (Crudo), Silver (Curado) y Gold (Consumible). Esta estructura promueve la calidad, la gobernanza y la facilidad de consumo de los datos.

2.1. Diagrama de Arquitectura



2.2. Descripción de las Capas

- **Capa Bronze (Raw Data):**
  - **Propósito:** Ingesta de datos crudos directamente desde las fuentes, preservando su formato original. Actúa como un archivo histórico inmutable.
  - **Contenido:** Los datos de los archivos CSV de Uber tal cual, con la adición de metadatos de auditoría (timestamp de ingesta, nombre del archivo fuente, versión de la capa).
  - **Formato:** Delta Lake.
- **Capa Silver (Cleaned & Enriched):**
  - **Propósito:** Limpieza, validación y enriquecimiento de los datos de la capa Bronze. Los datos aquí están listos para un análisis más profundo y la creación de características.
  - **Contenido:** Datos limpios, con tipos de datos corregidos, valores nulos manejados, duplicados eliminados y nuevas características generadas (ej. hora de recogida, día de la semana, tipo de día, categoría de tiempo).
  - **Formato:** Delta Lake.
- **Capa Gold (Business-Ready):**
  - **Propósito:** Agregación y modelado de datos para casos de uso específicos de negocio. Las tablas en esta capa están altamente optimizadas para el rendimiento de consultas y el consumo por herramientas de BI.
  - **Contenido:** KPIs ejecutivos, métricas temporales (diarias, horarias, mensuales), rendimiento por base, y datos geoespacial.
  - **Formato:** Delta Lake.

3. Fases del Pipeline ETL

El pipeline ETL se implementa a través de una serie de notebooks de Databricks, cada uno encargado de una fase específica.

3.1. Extracción (E): Capa Bronze (01_Uber_Bronze_Layer.ipynb)

- **Objetivo:** Ingestar los datos crudos de los archivos CSV en una tabla Delta en la capa Bronze.
- **Proceso Paso a Paso:**
  1. **Configuración de Rutas:** Se definen las rutas de origen (archivos CSV) y destino (tabla Delta Bronze).
  2. **Validación de Archivos Fuente:** Se listan y verifican los seis archivos CSV (uber-raw-data-apr14.csv a uber-raw-data-sep14.csv) en el directorio de origen.
  3. **Definición de Esquema:** Se define un esquema explícito para los archivos CSV (Date/Time como StringType, Lat y Lon como DoubleType, Base como StringType) para asegurar la consistencia.
  4. **Lectura de Datos:** Se leen todos los archivos CSV en un único DataFrame de Spark, infiriendo el esquema y tratando la primera fila como encabezado.

5. **Añadir Metadatos de Auditoría:** Se añaden columnas adicionales para trazabilidad:
    - ingestion_timestamp: Marca de tiempo de cuándo se ingirió el registro.
    - source_file: Nombre del archivo CSV de origen.
    - bronze_layer_version: Versión de la capa Bronze (ej. "1.0").
6. **Carga a Delta Lake:** El DataFrame resultante se escribe en formato Delta Lake en la ruta de la capa Bronze (/Volumes/workspace/default/uber_etl_azure/bronze/uber_bronze), utilizando el modo overwrite para reingestas o append para ingestas incrementales.

- **Validación:** Se verifica el recuento de registros y se muestra una muestra de los datos para confirmar la ingesta.

3.2. Transformación (T): Capa Silver (02_Uber_Silver_Layer.ipynb)

- **Objetivo:** Limpiar, validar y enriquecer los datos de la capa Bronze, preparándolos para el análisis.
- **Proceso Paso a Paso:**
    1. **Configuración y Lectura:** Se leen los datos de la tabla uber_bronze de la capa.
    2. **Análisis de Calidad de Datos (Inicial):** Se realiza un análisis exploratorio para identificar problemas como:
        - Esquema actual de los datos.
        - Conteo de valores nulos por columna.
        - Conteo de registros duplicados.
        - Muestra de datos crudos.
    3. **Limpieza de Datos:**
        - **Renombrado de Columnas:** Se renombran columnas a un formato más amigable y consistente (ej. Date/Time a pickup_datetime, Lat a latitude, Lon a longitude, Base a base_code).
        - **Conversión de Tipos de Datos:** Se convierte pickup_datetime a tipo timestamp y se asegura que latitude y longitude sean Double.
        - **Validación de Rangos Geográficos:** Se filtran los registros cuyas coordenadas de latitud y longitud caen fuera de los límites aproximados de la ciudad de Nueva York.
        - **Manejo de Nulos y Duplicados:** Se eliminan los registros con valores nulos en columnas críticas y se eliminan los registros duplicados.
    4. **Ingeniería de Características:** Se crean nuevas columnas derivadas para enriquecer los datos y facilitar el análisis:
        - pickup_hour: Hora de la recogida.
        - pickup_day_of_week: Nombre del día de la semana (ej. "Monday").
        - pickup_day_of_week_num: Número del día de la semana (1=Domingo, 7=Sábado).
        - pickup_month: Mes de la recogida.
        - pickup_year: Año de la recogida.
        - pickup_date: Fecha en formato yyyy-MM-dd.

    - pickup_week_of_year: Semana del año.
    - time_category: Categorización de la hora del día (ej. "Morning Rush", "Midday", "Evening Rush", "Night", "Late Night/Early Morning").
    - day_type: Categorización del día (ej. "Weekday", "Weekend").
    - trip_id: Un ID único para cada viaje (ej. usando un UUID o una combinación de columnas).
  5. **Añadir Metadatos de Auditoría (Silver):** Se añaden silver_processing_timestamp y silver_layer_version.
  6. **Carga a Delta Lake (Silver):** El DataFrame transformado se guarda como una tabla Delta (uber_silver) en la capa Silver (/Volumes/workspace/default/uber_etl_azure/silver/).
- **Validación:** Se verifica el esquema final, el recuento de registros y se muestra una muestra de los datos limpios y enriquecidos.

3.3. Carga (L): Capa Gold (03_Uber_Gold_Layer.ipynb)

- **Objetivo:** Crear agregaciones de negocio, KPIs y métricas analíticas desde la capa Silver, optimizadas para el consumo en Power BI.
- **Proceso Paso a Paso:**
  1. **Configuración y Lectura:** Se leen los datos de la tabla uber_silver de la capa Silver.
  2. **Creación de Tablas Gold Especializadas:** Se generan cinco tablas Gold, cada una para un caso de uso de negocio específico:
    - **daily_metrics (Agregaciones Diarias):**
      - Agregaciones: total_trips, active_bases, avg_latitude, avg_longitude, morning_rush_trips, evening_rush_trips, midday_trips, night_trips, late_night_trips.
      - Dimensiones: pickup_date, pickup_day_of_week, day_type.
    - **base_hourly_metrics (Métricas Horarias por Base):**
      - Agregaciones: hourly_trips, avg_latitude, avg_longitude, weekend_percentage.
      - Dimensiones: base_code, pickup_hour, time_category.
    - **monthly_kpis (KPIs Mensuales):**
      - Agregaciones: total_monthly_trips, mom_growth (crecimiento mes a mes), rush_hours_percentage, night_activity_percentage.
      - Dimensiones: pickup_month, pickup_year.
    - **spatial_metrics (Métricas Geoespaciales):**
      - Agregaciones: total_trips, avg_latitude, avg_longitude, trip_density.
      - Dimensiones: latitude_bin, longitude_bin (para agrupar por áreas geográficas).
    - **executive_dashboard (Resumen Ejecutivo Global):**
      - Agregaciones: total_trips_period, total_active_bases, total_active_days, avg_daily_trips, avg_trips_per_base, rush_hours_share, night_activity_share, weekend_vs_weekday_ratio, center_latitude, center_longitude, geographic_coverage_lat, geographic_coverage_lon.

        - Esta tabla contendrá un único registro con KPIs de alto nivel para todo el período.
    3. **Carga a Delta Lake (Gold):** Cada DataFrame agregado se escribe como una tabla Delta separada en la capa Gold (/Volumes/workspace/default/uber_etl_azure/gold/).
- **Validación:** Se verifica el recuento de registros y se muestra una muestra de los datos de cada tabla Gold para confirmar las agregaciones.

3.4. Dashboard y Visualización (04_Uber_Dashboard_Databricks.ipynb)

- **Objetivo:** Cargar las tablas Gold y realizar visualizaciones preliminares dentro de Databricks, simulando la preparación para Power BI.
- **Proceso Paso a Paso:**
    1. **Configuración y Carga de Tablas Gold:** Se cargan las cinco tablas Gold generadas en la fase anterior.
    2. **Generación de KPIs Ejecutivos:** Se muestran los KPIs principales del executive_dashboard (ej. Total de Viajes, Bases Activas, Promedio Viajes/Día).
    3. **Análisis Temporal:**
        - Tendencias mensuales de viajes y crecimiento mes a mes.
        - Distribución de viajes por día de la semana y tipo de día (Weekday/Weekend).
        - Patrones de demanda horaria general.
    4. **Performance por Bases:**
        - Ranking de las bases más activas.
        - Distribución horaria de viajes por base.
    5. **Análisis Geoespacial:**
        - Visualización de hotspots de recogida (ej. mapa de calor).
    6. **Insights de Negocio:** Se extraen conclusiones clave de las visualizaciones generadas.
- **Nota:** Aunque este notebook realiza visualizaciones básicas, la integración final con Power BI implica conectar Power BI directamente a las tablas Delta de la capa Gold.

4. Tecnologías y Herramientas

4.1. Databricks Free Edition

- **Justificación:**
    - **Plataforma Unificada:** Databricks ofrece un entorno unificado para el procesamiento de datos, machine learning y colaboración, lo que simplifica el desarrollo y la gestión del pipeline ETL.
    - **Spark Optimizado:** Aprovecha Apache Spark con optimizaciones de rendimiento, ideal para procesar grandes volúmenes de datos como el conjunto de datos de Uber.
    - **Delta Lake Integrado:** Soporte nativo para Delta Lake, permitiendo transacciones ACID, versionado de datos y esquemas evolucionables.
    - **Colaboración:** Los notebooks facilitan el trabajo en equipo y la documentación del código.

- **Free Edition:** Permite demostrar las capacidades de la plataforma sin costos iniciales, ideal para un proyecto de portafolio.

4.2. Delta Lake

- **Justificación:**
  - **Fiabilidad (ACID Transactions):** Garantiza la atomicidad, consistencia, aislamiento y durabilidad de las operaciones de datos, crucial para pipelines de datos robustos.
  - **Esquema Evolucionable:** Permite modificar el esquema de las tablas sin romper los pipelines existentes, adaptándose a cambios en los datos.
  - **Manejo de Datos a Escala:** Optimizado para grandes volúmenes de datos, permitiendo un rendimiento eficiente en la ingesta y consulta.
  - **Control de Versiones:** Facilita la auditoría y la capacidad de "viajar en el tiempo" a versiones anteriores de los datos, lo que es invaluable para la depuración y la reproducibilidad.
  - **Integración con Spark:** Funciona de forma nativa con Apache Spark, lo que simplifica el desarrollo de transformaciones de datos.

4.3. Power BI

- **Justificación:**
  - **Visualización Interactiva:** Permite crear dashboards y reportes interactivos y dinámicos, facilitando la exploración de datos por parte de analistas de negocio.
  - **Conectividad Robusta:** Puede conectarse directamente a Delta Lake (a través de Databricks SQL Analytics Endpoint), lo que permite consumir los datos de la capa Gold de manera eficiente.
  - **Facilidad de Uso:** Interfaz intuitiva para la creación de visualizaciones, incluso para usuarios no técnicos.
  - **Compartibilidad:** Facilita la distribución de informes y dashboards a diferentes stakeholders dentro de la organización.

5. Modelo de Datos

El modelo de datos sigue la arquitectura Medallón, con esquemas que evolucionan en cada capa para mejorar la calidad y la utilidad de los datos.

5.1. Capa Bronze (uber_bronze)

- **Propósito:** Almacenar los datos crudos tal como se ingieren desde los CSV.
- **Esquema:**
  - Date/Time: StringType (Fecha y hora de la recogida)
  - Lat: DoubleType (Latitud del punto de recogida)
  - Lon: DoubleType (Longitud del punto de recogida)
  - Base: StringType (Código de la base de Uber)
  - ingestion_timestamp: TimestampType (Marca de tiempo de la ingesta)

- source_file: StringType (Nombre del archivo CSV de origen)
- bronze_layer_version: StringType (Versión de la capa Bronze)

5.2. Capa Silver (uber_silver)

- **Propósito:** Datos limpios, validados y enriquecidos, listos para análisis.
- **Esquema:**
  - pickup_datetime: TimestampType (Fecha y hora de la recogida)
  - latitude: DoubleType (Latitud del punto de recogida)
  - longitude: DoubleType (Longitud del punto de recogida)
  - base_code: StringType (Código de la base de Uber)
  - ingestion_timestamp: TimestampType (Marca de tiempo de la ingesta)
  - source_file: StringType (Nombre del archivo CSV de origen)
  - bronze_layer_version: StringType (Versión de la capa Bronze)
  - pickup_hour: IntegerType (Hora de la recogida)
  - pickup_day_of_week: StringType (Día de la semana de la recogida)
  - pickup_day_of_week_num: IntegerType (Número del día de la semana)
  - pickup_month: IntegerType (Mes de la recogida)
  - pickup_year: IntegerType (Año de la recogida)
  - pickup_date: StringType (Fecha de la recogida en formato YYYY-MM-DD)
  - pickup_week_of_year: IntegerType (Semana del año de la recogida)
  - time_category: StringType (Categoría horaria: "Morning Rush", "Midday", "Evening Rush", "Night", "Late Night/Early Morning")
  - day_type: StringType (Tipo de día: "Weekday", "Weekend")
  - trip_id: StringType (ID único del viaje)
  - silver_processing_timestamp: TimestampType (Marca de tiempo del procesamiento en Silver)
  - silver_layer_version: StringType (Versión de la capa Silver)

5.3. Capa Gold (Tablas Agregadas)

- **Propósito:** Tablas optimizadas para el consumo directo por herramientas de BI y análisis de negocio.

5.3.1. daily_metrics

- **Esquema:**
  - pickup_date: DateType
  - pickup_month: IntegerType
  - pickup_year: IntegerType
  - pickup_day_of_week: StringType
  - pickup_day_of_week_num: IntegerType
  - day_type: StringType
  - total_trips: LongType (Total de viajes en el día)

- active_bases: IntegerType (Número de bases activas en el día)
- active_hours: IntegerType (Número de horas con actividad en el día)
- avg_latitude: DoubleType (Latitud promedio de las recogidas)
- avg_longitude: DoubleType (Longitud promedio de las recogidas)
- stddev_latitude: DoubleType (Desviación estándar de la latitud)
- stddev_longitude: DoubleType (Desviación estándar de la longitud)
- morning_rush_trips: LongType (Viajes en hora pico de la mañana)
- evening_rush_trips: LongType (Viajes en hora pico de la tarde)
- midday_trips: LongType (Viajes en horario de mediodía)
- night_trips: LongType (Viajes en horario nocturno)
- late_night_trips: LongType (Viajes en horario de madrugada)
- first_trip_hour: IntegerType (Primera hora con viajes)
- last_trip_hour: IntegerType (Última hora con viajes)

5.3.2. base_hourly_metrics

- **Propósito:** Análisis del rendimiento de cada base por hora.
- **Esquema:**
  - base_code: StringType
  - pickup_hour: IntegerType
  - time_category: StringType
  - hourly_trips: LongType (Total de viajes por base y hora)
  - avg_latitude: DoubleType
  - avg_longitude: DoubleType
  - weekday_trips: LongType (Viajes en días de semana)
  - weekend_trips: LongType (Viajes en fines de semana)
  - weekday_percentage: DoubleType (Porcentaje de viajes en días de semana)
  - weekend_percentage: DoubleType (Porcentaje de viajes en fines de semana)

5.3.3. monthly_kpis

- **Propósito:** Seguimiento de KPIs a nivel mensual.
- **Esquema:**
  - pickup_month: IntegerType
  - pickup_year: IntegerType
  - total_monthly_trips: LongType
  - mom_growth: DoubleType (Crecimiento porcentual mes a mes)
  - rush_hours_percentage: DoubleType (Porcentaje de viajes en horas pico)
  - night_activity_percentage: DoubleType (Porcentaje de viajes nocturnos)

5.3.4. spatial_metrics

- **Propósito:** Entender la distribución geográfica de la demanda.
- **Esquema:**

- latitude_bin: DoubleType (Latitud binned)
- longitude_bin: DoubleType (Longitud binned)
- total_trips: LongType (Total de viajes en el bin geográfico)
- trip_density: DoubleType (Densidad de viajes)
- avg_pickup_hour: DoubleType (Hora promedio de recogida en el bin)
- most_active_base: StringType (Base más activa en el bin)

5.3.5. executive_dashboard

- **Propósito:** Resumen de alto nivel para la dirección.
- **Esquema:**
  - total_trips_period: LongType
  - total_active_bases: IntegerType
  - total_active_days: IntegerType
  - avg_daily_trips: DoubleType
  - avg_trips_per_base: DoubleType
  - rush_hours_share: DoubleType
  - night_activity_share: DoubleType
  - weekend_vs_weekday_ratio: DoubleType
  - center_latitude: DoubleType
  - center_longitude: DoubleType
  - geographic_coverage_lat: DoubleType
  - geographic_coverage_lon: DoubleType

6. Visualización y Hallazgos (Ideas para Power BI)

La capa Gold está diseñada para ser directamente consumible por Power BI. A continuación, se presentan ideas para dashboards y visualizaciones clave, junto con los posibles hallazgos.

6.1. Dashboard de KPIs Ejecutivos

- **Visualizaciones:**
  - Tarjetas de KPI: Total de viajes, Promedio de viajes por día, Número de bases activas.
  - Gráfico de barras: Distribución de viajes por time_category (Morning Rush, Midday, etc.).
  - Gráfico de anillos/pastel: Proporción de viajes de Weekday vs. Weekend.
- **Hallazgos Potenciales:**
  - Identificar rápidamente el volumen total de operaciones y la eficiencia promedio.
  - Confirmar si las horas pico (mañana/tarde) representan la mayor parte de la demanda.
  - Determinar la importancia relativa de los fines de semana en comparación con los días laborales.

6.2. Análisis Temporal

- **Visualizaciones:**

    - Gráfico de líneas: Tendencia mensual de total_monthly_trips y mom_growth.
    - Gráfico de barras: Viajes por pickup_day_of_week_num (ordenado por día de la semana).
    - Gráfico de área apilado: Distribución horaria (pickup_hour) de viajes, segmentado por day_type.
- **Hallazgos Potenciales:**
    - Observar el crecimiento constante de la demanda de Uber de abril a septiembre de 2014.
    - Identificar los días de la semana con mayor y menor actividad.
    - Visualizar patrones de demanda diarios, como el doble pico en días de semana (mañana y tarde) y una demanda más distribuida en fines de semana.

6.3. Performance por Bases

- **Visualizaciones:**
    - Gráfico de barras: Top N base_code por total_trips.
    - Gráfico de líneas múltiples: hourly_trips para las bases principales, mostrando su patrón de demanda a lo largo del día.
    - Tabla: Detalles de base_hourly_metrics con filtros por base_code y time_category.
- **Hallazgos Potenciales:**
    - Determinar qué bases son las más activas y contribuyen más al volumen total de viajes.
    - Identificar si algunas bases tienen patrones de demanda horaria únicos (ej. una base cerca de un aeropuerto con picos tempranos).

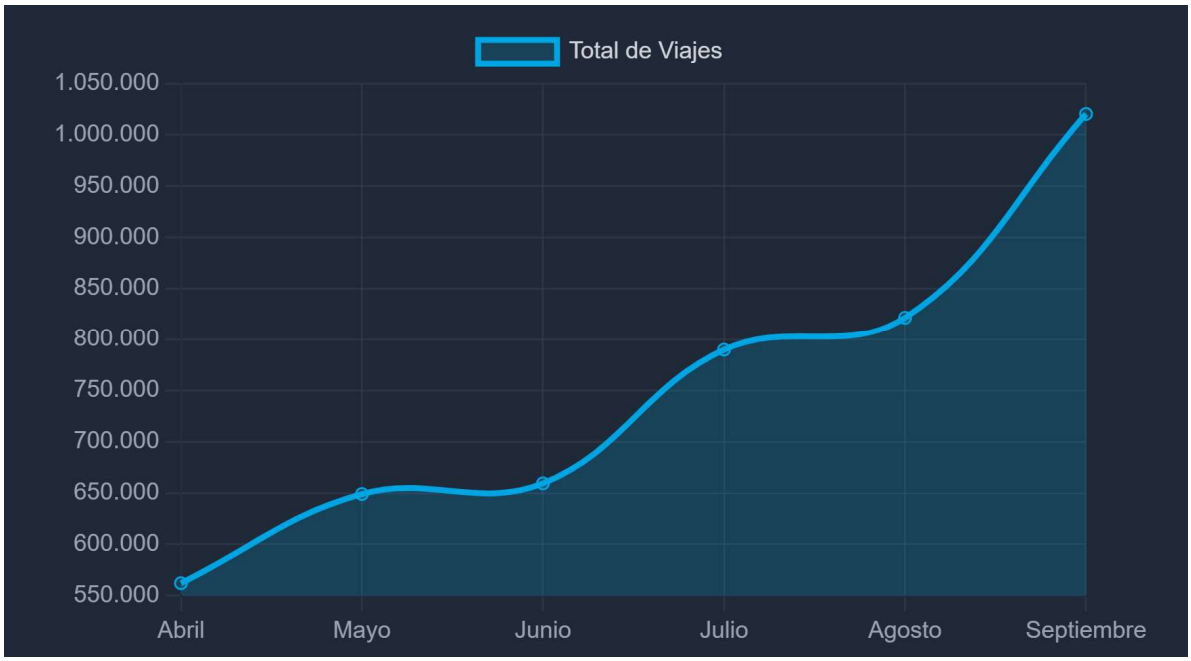
6.4. Análisis Geoespacial

- **Visualizaciones:**
    - Mapa de calor (Heatmap) en un mapa de NYC: Utilizando latitude_bin y longitude_bin con trip_density para mostrar los "hotspots" de recogida.
    - Puntos en el mapa: Representando los avg_latitude y avg_longitude de los bins, con tamaño proporcional a total_trips.
- **Hallazgos Potenciales:**
    - Identificar las áreas geográficas de mayor concentración de recogidas en NYC.
    - Observar cómo la densidad de viajes varía en diferentes partes de la ciudad.
    - Detectar si hay áreas con alta demanda pero baja cobertura de bases (oportunidades de expansión).

6.5. Ilustraciones Clave del Dashboard

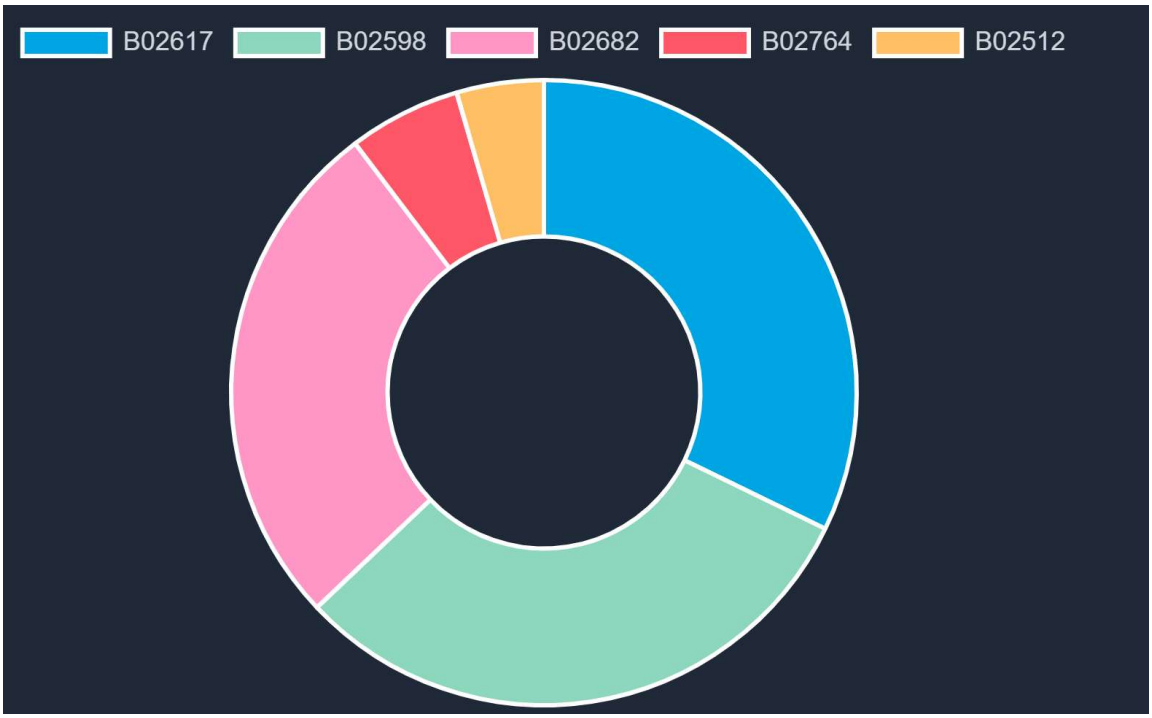
A continuación, se presentan descripciones de algunas visualizaciones clave que se podrían construir en Power BI, junto con los insights que ofrecen:

- Crecimiento Mensual de Viajes
  El análisis revela un crecimiento sostenido y significativo en el volumen de viajes mes a mes, culminando en un impresionante aumento del 24% en septiembre. Esto indica una rápida adopción y expansión del servicio de Uber en NYC durante 2014.



- Patrones de Demanda por Hora del Día
  La demanda horaria muestra un claro patrón de "horas pico" durante la tarde y noche, especialmente entre las 17:00 y las 21:00. Este insight es crucial para la gestión de la flota y la implementación de tarifas dinámicas para equilibrar la oferta y la demanda.

- Rendimiento por Base de Operaciones
  Las cinco bases de Uber tienen una participación de mercado distinta. Las bases B02617, B02598 y B02682 dominan el volumen de viajes, manejando conjuntamente la gran mayoría de las recogidas en la ciudad y siendo pilares de la operación.



- Hotspots Geográficos de Demanda
  El análisis geoespacial identifica zonas de alta concentración de recogidas. Áreas como Midtown Manhattan, el Distrito Financiero y el East Village emergen como los principales "hotspots", indicando dónde se debe concentrar la oferta de vehículos. Las coordenadas en este gráfico son representaciones conceptuales de las ubicaciones para ilustrar la densidad.

7. Conclusiones y Futuras Mejoras

7.1. Conclusiones del Trabajo Realizado

Este proyecto ha demostrado la capacidad de construir una canalización de datos ETL de extremo a extremo robusta y escalable para el análisis de viajes de Uber en NYC. La implementación de la arquitectura Medallón en Databricks con Delta Lake ha permitido:

- **Ingesta Eficiente:** Procesar millones de registros CSV en la capa Bronze, preservando la fidelidad de los datos.
- **Transformación de Calidad:** Limpiar, validar y enriquecer los datos en la capa Silver, resolviendo problemas de formato, nulos y duplicados, y añadiendo características valiosas para el análisis.
- **Modelado para el Negocio:** Crear tablas Gold optimizadas para consultas de negocio, facilitando la extracción de KPIs y el análisis de tendencias.
- **Habilitación de Insights:** Proporcionar una base de datos confiable y estructurada que permite a los analistas de negocio y científicos de datos responder preguntas críticas sobre la demanda de Uber en NYC.

7.2. Futuras Mejoras y Próximos Pasos

Para seguir evolucionando este proyecto y extraer aún más valor de los datos, se proponen las siguientes mejoras:

- **Integración de Datos Adicionales:**
  - **Datos Meteorológicos:** Incorporar datos de clima de NYC para analizar cómo las condiciones climáticas (lluvia, nieve, temperatura) afectan la demanda de viajes.
  - **Eventos Especiales:** Integrar un calendario de eventos (conciertos, partidos deportivos, ferias) para correlacionar picos de demanda con acontecimientos específicos.
  - **Datos de Tráfico:** Añadir información sobre el tráfico para entender su impacto en los tiempos de viaje y la eficiencia.
  - **Análisis de Sentimiento de Redes Sociales:** Integrar datos de redes sociales (ej. Twitter) para capturar el sentimiento público sobre Uber y eventos, lo que podría influir en la demanda.
- **Análisis Predictivo:**
  - **Pronóstico de Demanda:** Desarrollar modelos de Machine Learning (ML) para predecir la demanda de viajes por hora y por ubicación, lo que permitiría a Uber optimizar la asignación de conductores.
  - **Detección de Anomalías:** Implementar algoritmos para identificar patrones de viaje inusuales que podrían indicar fraudes o problemas operativos.
- **Optimización de Costos y Rendimiento:**
  - **Particionamiento Inteligente:** Optimizar el particionamiento de las tablas Delta en las capas Silver y Gold basándose en patrones de consulta para mejorar el rendimiento y reducir costos.

    - **Z-Ordering:** Aplicar Z-Ordering en Delta Lake para mejorar la lectura de datos en consultas que involucran múltiples columnas.
    - **Streaming de Datos:** Evaluar la posibilidad de migrar a una ingesta de datos en tiempo real (streaming) para un análisis casi instantáneo de la demanda.
- **Gobernanza y Monitoreo:**
    - **Data Quality Checks Automatizados:** Implementar herramientas de monitoreo de calidad de datos para alertar sobre inconsistencias o anomalías en el pipeline.
    - **Catálogo de Datos:** Crear un catálogo de datos para documentar los metadatos de las tablas, facilitando su descubrimiento y uso por parte de los equipos de negocio.
- **Expansión a la Plataforma Azure:**
    - **Ingesta y Orquestación con Azure Data Factory:** Utilizar Azure Data Factory para orquestar la pipeline ETL, desde la ingesta de datos de origen (Azure Storage Account) hasta la carga en las capas Bronze, Silver y Gold.
    - **Procesamiento de Datos con Azure Databricks:** Aprovechar Azure Databricks como el motor de procesamiento principal para las transformaciones de datos en las capas Silver y Gold, manteniendo la flexibilidad y escalabilidad de Spark.
    - **Almacenamiento de Datos Curados en Azure SQL Database:** Cargar las tablas finales de la capa Gold en Azure SQL Database para un consumo eficiente por parte de aplicaciones de negocio y herramientas de BI.
    - **Visualización de Resultados en Power BI:** Conectar Power BI directamente a Azure SQL Database para crear dashboards y reportes interactivos, permitiendo a los analistas de negocio explorar los insights.
    - **Gestión Segura de Credenciales con Azure Key Vault:** Centralizar y proteger las credenciales y secretos de conexión (por ejemplo, para Azure SQL Database) utilizando Azure Key Vault, garantizando la seguridad del pipeline.
    - **Automatización y Despliegue Continuo con Azure DevOps:** Implementar Azure DevOps para la integración continua (CI) y el despliegue continuo (CD) del pipeline de datos, automatizando las pruebas y el despliegue de los notebooks de Databricks y la infraestructura de Azure.