Análisis de Movilidad Urbana

Nuria García Valdecasas
Mario Guillen
Julia Rojas
Valentina Pineda Barrón

Índice

Introducción

El presente proyecto tiene como objetivo principal la implementación de un flujo de trabajo de ingeniería de datos para analizar el comportamiento de la movilidad urbana durante el periodo de enero a octubre de 2025. Utilizando un conjunto de datos con un peso aproximado de 1.1 GB y compuesto por más de 17 millones de registros de viajes, se busca extraer conocimiento valioso sobre la demanda y la operatividad del sistema.

Para enfrentar el desafío técnico que implica la manipulación de este volumen de información, se utilizó Apache Spark (a través de su interfaz PySpark) como motor de procesamiento distribuido. A diferencia de las bases de datos relacionales convencionales o las hojas de cálculo, este enfoque permite realizar operaciones de agregación, filtrado y transformación en memoria de manera eficiente, superando las limitaciones de hardware local mediante técnicas de optimización como el manejo de particiones y la gestión de memoria del "driver".

A lo largo de este documento, se detalla el proceso de Extracción, Transformación y Carga (ETL), donde se consolidaron múltiples fuentes de datos mensuales, se estandarizaron esquemas y se generaron nuevas variables temporales (como la duración exacta de los viajes y la segmentación por franjas horarias). Posteriormente, se presentan 8 análisis estadísticos clave, que incluyen:

1. Tendencias de Demanda: Identificación de los meses y horarios de mayor actividad.
2. Perfiles de Usuario: Comparativas de uso según género y tipo de membresía.
3. Análisis de "Estrés" del Sistema: Un estudio del balance neto entre retiros y arribos para detectar estaciones saturadas o vacías.
4. Patrones de Rutas: Detección de las estaciones de origen y destino más frecuentes.

Finalmente, los resultados de este procesamiento masivo se visualizan mediante gráficas generadas con librerías de Python, demostrando cómo las tecnologías de bases de datos no relacionales y el cómputo distribuido son herramientas esenciales para interpretar la dinámica de una de las metrópolis más grandes del mundo.

Problemas encontrados y su solución

Durante el desarrollo del proyecto se presentaron diversos inconvenientes relacionados con la calidad de los datos, su formato original, el procesamiento con PySpark y la integración con información geoespacial. A continuación se describen los principales problemas identificados, junto con la solución aplicada en cada caso.

Problema: Valores de edad no válidos (strings)

Al cargar los datos, la columna **Edad_Usuario** incluía valores no numéricos como strings "NULL". Esto generaba errores en las consultas donde se intentaba castear la edad a un tipo de dato int.

Solución aplicada:
Se realizó una limpieza previa para transformar valores no numéricos o vacíos en None, manteniendo la integridad del promedio.

Problema: El dataset de viajes no incluía información de las estaciones

Los archivos CSV de viajes solo contenían los identificadores de estación de retiro y arribo (Ciclo_Estacion_Retiro y Ciclo_EstacionArribo), pero no incluían ningún catálogo de estaciones con su nombre, ubicación o capacidad. Esto impedía saber en qué puntos específicos de la ciudad se originaban y terminaban los viajes, lo que limitaba el análisis espacial del sistema Ecobici.

Solución aplicada:

Se decidió descargar el archivo oficial de estaciones en formato GBFS (JSON) y leerlo con PySpark.
Después se normalizó el identificador de estación (rellenando con ceros a la izquierda hasta 3 dígitos) para que coincidiera con los códigos de los viajes, y se construyeron dos dataframes:

- df_est_retiro para la información de la estación de retiro.

- df_est_arribo para la información de la estación de arribo.

Finalmente, se realizaron joins tipo left entre df_final (viajes) y estos dataframes de estaciones, usando los IDs normalizados. Con esto se incorporaron al dataframe final el nombre, las coordenadas y la capacidad de cada estación en origen y destino.

Sin embargo, incluso después de integrar ambos archivos de estaciones y normalizar los identificadores, se identificó que no todas las estaciones registradas en el dataset de viajes tenían un equivalente en los archivos GBFS. Esto significa que algunos valores de Ciclo_Estacion_Retiro o Ciclo_EstacionArribo no aparecen en ninguno de los catálogos oficiales de estaciones, ya sea porque corresponden a estaciones deshabilitadas, códigos antiguos o datos incompletos. Como consecuencia, esas observaciones conservan valores nulos en los campos de nombre, coordenadas, capacidad, colonia y alcaldía. Este comportamiento se mantuvo intencionalmente para no eliminar viajes y asegurar la integridad del dataset final, documentando adecuadamente que ciertas estaciones no cuentan con información oficial disponible.

Problema: Los nombres de las estaciones eran poco claros e insuficientes para ubicar el contexto urbano
Aunque el JSON de estaciones incluía un campo *name*, este nombre muchas veces se daba como referencia entre calles o puntos poco intuitivos (por ejemplo, “Entre X y Y”, “Cruz Roja”, etc.). Esto hacía difícil **interpretar directamente qué zona de la ciudad representaba** cada estación. Solo con el nombre de la estación no se podía vincular fácilmente a una colonia o alcaldía.

Solución aplicada:
A partir de las coordenadas de cada estación (lat, lon), se desarrolló un proceso externo (en Python/Pandas) que:

1. Tomó la lista de estaciones con su station_id, latitud y longitud generada en PySpark (df_estaciones).

2. Utilizó esas coordenadas para obtener la colonia y la alcaldía correspondientes para cada estación (por ejemplo, mediante geocodificación o cruces con capas geográficas).

3. Generó un archivo estaciones_geo.csv con las columnas: station_id, lat, lon, colonia y alcaldia

Luego, ese archivo se volvió a incorporar al flujo en PySpark.
De esta forma, se enriqueció el dataframe de viajes con colonia y alcaldía tanto de retiro como de arribo, lo que permitió hacer análisis por demarcación territorial y presentar resultados mucho más interpretables.

Problema: Dificultad para trabajar con fechas y horarios separados

En los archivos originales de viajes, la información temporal venía separada en columnas de texto: Fecha_Retiro, Hora_Retiro, Fecha_Arribo y Hora_Arribo. Trabajar con fechas y horarios como cadenas hacía muy complicado calcular la duración de los viajes, agrupar por mes u horario, o analizar patrones temporales (por ejemplo, horas pico). Además, cualquier operación requería concatenar y transformar manualmente estos campos, lo que incrementaba la probabilidad de errores y hacía el código menos claro.

Solución aplicada:
Se decidió unificar cada par de columnas de fecha y hora en un solo campo de tipo timestamp, tanto para el inicio como para el fin del viaje. Para ello se definió el formato de fecha y hora

Características de la solución

Funcionalidades principales

La solución desarrollada se caracteriza por integrar de manera coherente una serie de funciones que permiten transformar datos crudos de ECOBICI en información útil para el análisis de movilidad urbana. En primer lugar, se implementó la lectura masiva de los archivos mensuales correspondientes al periodo enero–octubre de 2025, unificando todos los registros en un único Data Frame capaz de representar el comportamiento completo del sistema durante esos meses. A partir de esta base integrada, se aplicaron procesos de limpieza y estandarización que incluyeron

la conversión de las fechas y horas a un formato de tipo timestamp, la eliminación de registros inválidos y el cálculo de variables derivadas como la duración del viaje, la hora de retiro o los patrones temporales de uso.

Una vez que los datos quedaron preparados, se desarrollaron diversas capacidades analíticas, entre las que destacan la identificación de las cicloestaciones con mayor número de retiros y arribos, la detección de las horas pico de uso a lo largo del día y el cálculo de la duración promedio de los viajes, tanto en general como segmentada por características del usuario. La solución también permite combinar los viajes con información geográfica del catálogo de estaciones, lo que facilita la interpretación espacial del comportamiento observado. Finalmente, se incorporó un módulo de visualización que convierte los resultados agregados en gráficas claras y comunicables, de modo que los patrones de movilidad se representan de manera comprensible y directamente utilizable en el reporte final.

Obtención e instalación de herramientas nuevas

Para implementar esta solución fue necesario incorporar herramientas adicionales a las vistas en el curso, principalmente enfocadas en el manejo eficiente de grandes volúmenes de datos. La herramienta central fue PySpark, instalada dentro de un entorno virtual de Python mediante el comando pip install pyspark, lo que permitió aislar las dependencias del proyecto y garantizar una configuración estable. Una vez instalado, se creó una SparkSession en modo local, que sirvió como punto de entrada al motor distribuido de Spark y habilitó el uso de funciones avanzadas para manipular y transformar los datos a gran escala. Este entorno permitió ejecutar operaciones como agregaciones, filtrados, cálculos temporales y fusiones de tablas de manera mucho más eficiente que con librerías tradicionales.

Complementando lo anterior, se emplearon bibliotecas como Pandas y Matplotlib, que se instalaron a través del mismo entorno virtual y permitieron convertir los resultados procesados en visualizaciones claras y presentables. Pandas fue fundamental para transformar los DataFrames de PySpark en estructuras compatibles con las funciones gráficas de Matplotlib,

mientras que esta última se utilizó para la elaboración de gráficas de líneas, barras y distribuciones que ilustran los principales hallazgos del análisis. La correcta instalación y configuración de estas herramientas permitió construir una solución robusta, reproducible y adecuada para el análisis de movilidad urbana a partir de datos reales del sistema ECOBICI.

Obtención y almacenamiento de los datos

Para la recopilación de la información de movilidad urbana, se utilizaron los datos públicos del sistema Ecobici de la Ciudad de México. El proceso de obtención de datos se diseñó para manejar un flujo de información histórica masiva, abarcando el periodo de enero a octubre de 2025.

El procedimiento de adquisición consistió en:

1. Recolección de Archivos: Se descargaron los conjuntos de datos mensuales en formato `.csv` (Comma Separated Values) desde el portal de datos abiertos.
2. Organización del Repositorio: Los archivos fueron almacenados en un directorio local unificado, estructurando la información para permitir su lectura secuencial o paralela.
3. Volumetría del Dataset: El conjunto de datos consolidado alcanzó un tamaño de 1.1 GB, integrando más de 17 millones de registros de viajes, lo cual determinó la necesidad de utilizar herramientas de procesamiento distribuido.

Transformación y Carga (ETL) en PySpark

El núcleo del procesamiento de datos (ETL) se implementó utilizando Apache Spark a través de su interfaz en Python, PySpark. Esta tecnología permitió la ingestión y transformación de los datos en memoria, facilitando la manipulación de millones de registros que excedería la capacidad de herramientas de hoja de cálculo o scripts secuenciales tradicionales.

Ingesta Masiva y Definición de Esquema

Para optimizar la lectura, se utilizó la capacidad de Spark para ingerir múltiples fuentes de datos simultáneamente mediante el uso de comodines en la ruta del directorio. Adicionalmente, se definió un esquema estricto (`StructType`) para garantizar la integridad de los tipos de datos desde el momento de la carga.

El siguiente fragmento ilustra la configuración de la carga consolidada:

- from pyspark.sql.types import StructType, StringType, IntegerType
-
- # Definición manual del esquema para asegurar la integridad de los datos
- schema_ecobici = StructType([
- StructField("Genero_Usuario", StringType(), True),
- StructField("Edad_Usuario", StringType(), True),
- StructField("Bici", StringType(), True),
- # ... definición del resto de columnas
-])
-
- # Carga unificada de todos los archivos CSV del periodo
- df_ecobici = spark.read.csv("/ruta/datos_ecobici_2025/*.csv", header=True, schema=schema_ecobici)

Transformación y Enriquecimiento de Datos

Una vez cargados los datos crudos en un DataFrame distribuido, se aplicaron transformaciones vectorizadas para limpiar la información y generar nuevas variables críticas para el análisis estadístico:

- Estandarización Temporal: Las columnas originales de fecha y hora, que se encontraban separadas y en formato de texto, fueron fusionadas y convertidas a objetos `Timestamp`. Esto permitió habilitar operaciones cronológicas precisas.

- Cálculo de Métricas Derivadas: Se generó la columna `duracion_segundos` calculando la diferencia aritmética entre el momento de arribo y el de retiro de la bicicleta. Esta variable es fundamental para medir el uso real del servicio.
- Segmentación: A partir de los timestamps, se extrajeron atributos específicos como el mes (`mes`) y la hora del día (`hora_del_dia`), facilitando la posterior agrupación de datos para detectar patrones de hora pico y estacionalidad.

El código clave de transformación se estructuró de la siguiente manera:

- # Conversión de cadenas de texto a objetos de tiempo manipulables
- df_procesado = df_ecobici.withColumn(
- "timestamp_inicio",
- to_timestamp(concat_ws(" ", col("Fecha_Retiro"), col("Hora_Retiro")), "d/M/yyyy H:m:s")
-)
-
- # Cálculo de la duración del viaje para análisis estadístico
- df_final = df_procesado.withColumn(
- "duracion_segundos",
- col("timestamp_fin").cast("long") - col("timestamp_inicio").cast("long")
-)

Procesamiento en Memoria

A diferencia de los sistemas de almacenamiento en disco tradicionales, este flujo mantuvo los datos transformados como DataFrames en memoria. Para asegurar la eficiencia durante las consultas repetitivas (como el cálculo de promedios o conteos por estación), se configuró el entorno de Spark para gestionar la memoria del "driver" de forma dinámica, permitiendo realizar agregaciones complejas sobre la totalidad de los 17 millones de registros sin interrupciones.

Estructura de las tablas de Base de Datos y Consultas

Estructura de las Colecciones

La base de datos empleada en el análisis proviene directamente de los registros históricos de ECOBICI y se encuentra organizada en una colección principal llamada eco_bici, la cual concentra la información de todos los viajes realizados por las y los usuarios. Cada fila de esta colección representa un viaje individual y contiene los campos que se observan en la tabla original: el género y la edad del usuario, el identificador de la bicicleta utilizada, el código de la cicloestación donde se realizó el retiro, la fecha y hora exactas de salida, el código de la cicloestación de arribo, así como la fecha y hora de llegada. Esta estructura refleja de manera precisa el comportamiento temporal de cada trayecto, ya que los datos registran la transición entre dos fechas: la del cierre del año 2024 y las primeras horas de 2025, tal como se aprecia en la captura de la base de datos.

Para poder trabajar analíticamente con esta información, cada uno de los campos relacionados con fecha y hora fue transformado a un formato de tipo timestamp, lo que permitió registrar el momento exacto de inicio y término del viaje y, con ello, calcular la duración total en segundos y minutos. Además de estas transformaciones, se creó un conjunto de columnas derivadas como la hora del día o el mes del viaje que permiten segmentar la actividad de manera mucho más granular. Esta colección, por tanto, se convirtió en la estructura central del proyecto, ya que contiene tanto los datos originales como las variables adicionales necesarias para realizar estudios temporales, espaciales y operativos del sistema ECOBICI.

Junto con esta tabla principal se incorporó también una segunda colección relacionada con la infraestructura del sistema, compuesta por la información proveniente del archivo JSON con el inventario de estaciones. Esta colección contiene el identificador de cada estación, su nombre, latitud, longitud y capacidad instalada. La unión entre ambas colecciones se realizó a través de los identificadores de estación de retiro y de arribo, lo que permitió enriquecer cada viaje con información geográfica y operativa de la estación correspondiente. De esta manera, la base de

datos final quedó estructurada para permitir no sólo el análisis de los trayectos, sino también el estudio de las dinámicas de oferta y demanda en las estaciones del sistema.

Realización de Consultas y Análisis

A partir de las colecciones procesadas se llevaron a cabo distintas consultas analíticas que permitieron conocer la actividad del sistema desde varias perspectivas. La primera de ellas consistió en identificar las cicloestaciones con mayor número de retiros, generando una tabla derivada que resume las estaciones más demandadas del sistema. Esta consulta permitió observar patrones claros de concentración de viajes en ciertos puntos de la ciudad. Posteriormente, se calculó la duración promedio del uso de las bicicletas, filtrando trayectos demasiado breves para obtener un promedio representativo de los viajes reales.

Otro análisis fundamental fue la identificación de las horas pico del día, lo cual se realizó mediante el cálculo de la hora exacta de inicio de cada trayecto. Esto generó una tabla que agrupa todos los viajes por hora del día y permite visualizar los momentos en los que la demanda se intensifica. Este resultado sirvió como eje para la construcción de gráficas temporales en las que se observa claramente la distribución horaria del uso de ECOBICI.

Además de estas consultas principales, se llevaron a cabo análisis más avanzados orientados a entender el comportamiento operativo del sistema. Uno de ellos consistió en clasificar los viajes según si ocurrieron en horarios de alta demanda por ejemplo, en las mañanas y en las noches, lo que permitió construir un conjunto de datos denominado “mareas” que refleja los flujos direccionales en horas laborales. Otro análisis relevante fue el cálculo del balance neto de llegadas y salidas por estación, que permitió construir un mapa de calor capaz de mostrar si una estación tiende a vaciarse o saturarse en determinadas horas, aportando información clave sobre el “estrés” operativo del sistema. Finalmente, se desarrolló un análisis de perfiles de usuario, en el que los viajes se clasificaron como “commuter”, “paseo” o “normal” según su duración y comportamiento de origen–destino.

En conjunto, estas consultas no sólo permitieron describir la actividad diaria del sistema ECOBICI, sino también capturar patrones más profundos de uso, comportamiento espacial y

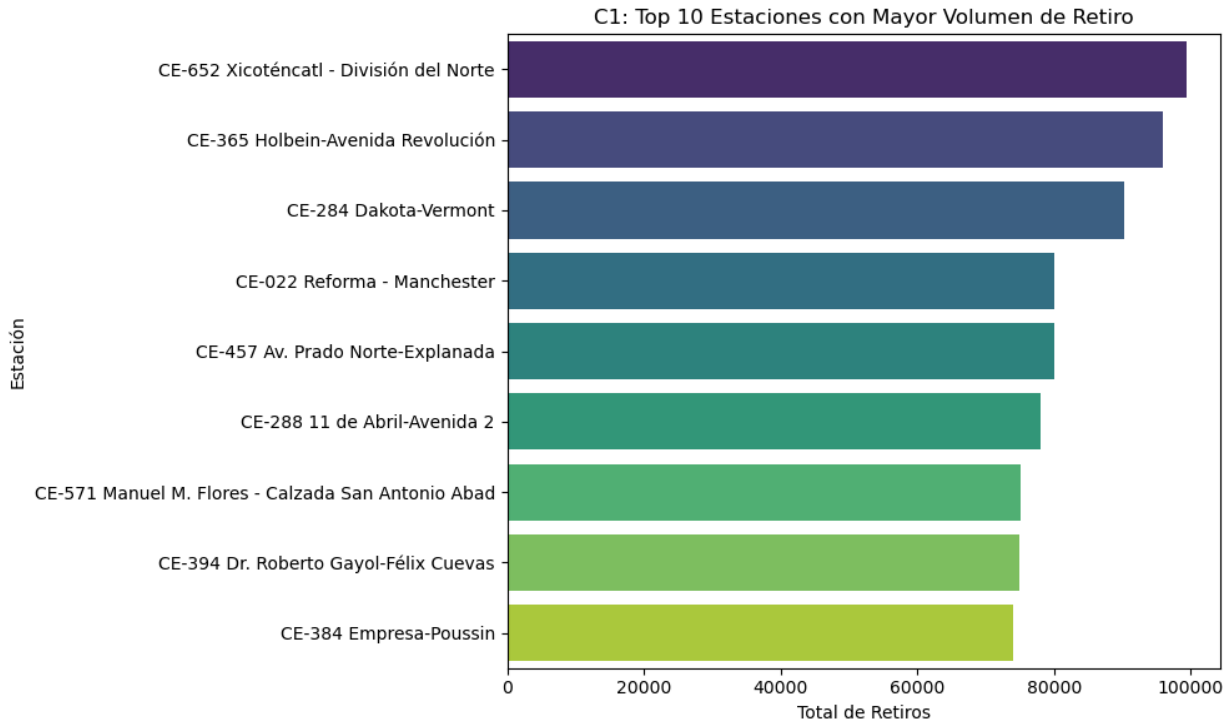
necesidades operativas. Así, la combinación de la colección eco_bici, la colección de estaciones y las consultas analíticas derivadas conforman una representación completa del funcionamiento del sistema de movilidad sustentable en la Ciudad de México.

Resultados obtenidos

El análisis estadístico de los hallazgos permite una evaluación del rendimiento y los patrones de uso del sistema Ecobici. Esta sección presenta los resultados mediante tablas y gráficas que facilitan la interpretación de tendencias temporales, el comportamiento de los usuarios y los puntos críticos de la operación. Se identifican aspectos clave, como las estaciones de mayor demanda y el equilibrio de movilidad durante los picos de actividad.

Top 10 estaciones de retiro más usadas

La siguiente tabla muestra el ranking de las 10 estaciones con mayor volumen de retiros durante el periodo analizado, identificando los puntos críticos de generación de viajes en la red.



Ranking	Nombre de la Estación	Total de Retiros
1	NULL (Estación 271-272)	1,264,870
2	CE-652 Xicoténcatl - División del Norte	99,371
3	CE-365 Holbein-Avenida Revolución	96,030
4	CE-284 Dakota-Vermont	90,258
5	CE-022 Reforma - Manchester	80,113
6	CE-457 Av. Prado Norte-Explanada	80,010
7	CE-288 11 de Abril-Avenida 2	78,111
8	CE-571 Manuel M. Flores - Calzada San Antonio Abad	75,167
9	CE-394 Dr. Roberto Gayol-Félix Cuevas	75,021
10	CE-384 Empresa-Poussin	73,966

Los datos revelan una marcada concentración de la demanda en nodos específicos, destacando la estación 271-272 como el origen más crítico del sistema con 119,492 viajes, superando por más de 20,000 retiros a la segunda posición. Se observa que el "Top 10" mantiene una operatividad constante por encima de los 75,000 retiros, lo que sugiere que estos puntos funcionan como centros conectores estratégicos dentro de la red de movilidad urbana.

Duración promedio de viaje

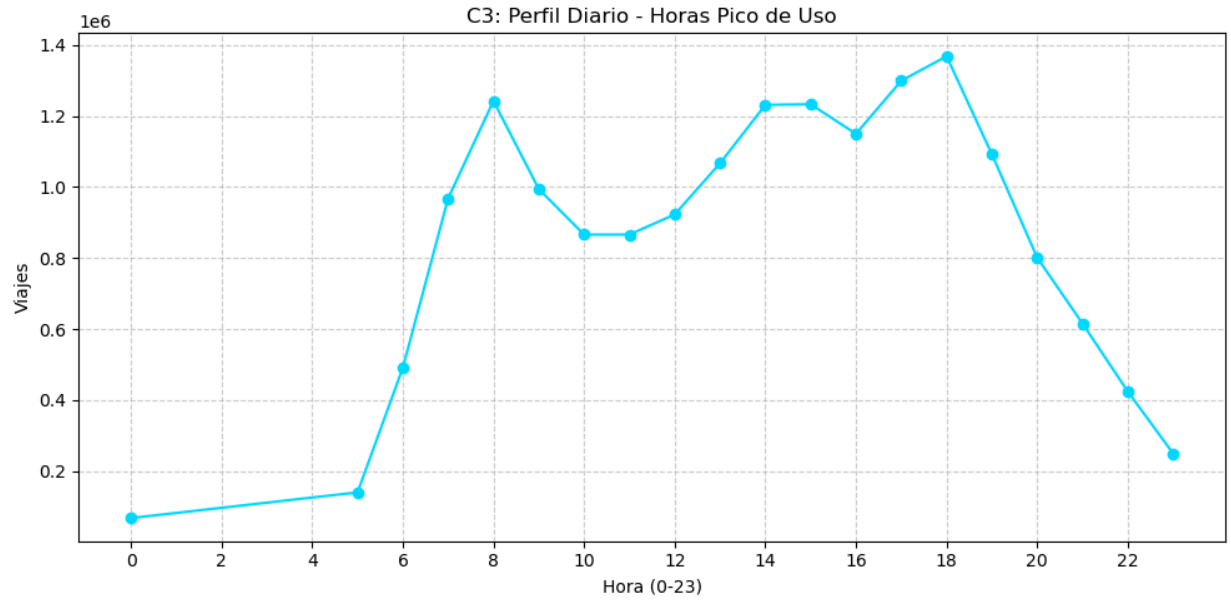
La siguiente tabla presenta el cálculo del tiempo promedio que los usuarios utilizan una bicicleta por trayecto, considerando la totalidad de los registros analizados.

Métrica	**Valor Obtenido**
Duración Promedio	**16.35 minutos**

El análisis de los 17 millones de registros arroja una duración media de viaje de 16.35 minutos. Este valor confirma la naturaleza del sistema Ecobici como una solución de movilidad de "última milla" o para trayectos cortos y eficientes. El tiempo sugiere que la mayoría de los usuarios utilizan el servicio para desplazamientos puntuales (por ejemplo, de una estación de metro a la oficina o entre barrios cercanos), optimizando sus tiempos de traslado en una ciudad con alta congestión vehicular, en lugar de utilizar las unidades para paseos recreativos de larga duración.

Horas pico de uso en el día

La siguiente gráfica ilustra la distribución del volumen total de viajes a lo largo de las 24 horas del día, permitiendo identificar los patrones de alta demanda y los periodos de inactividad del sistema.



Hora del Día	Total de Viajes
0	67,725
5	139,994
6	491,159

7	**967,965**
8	**1,242,479**
9	994,215
10	865,982
11	865,930
12	922,855
13	1,066,784
14	1,231,416
15	1,233,127
16	1,149,818
17	**1,299,319**
18	**1,368,276**

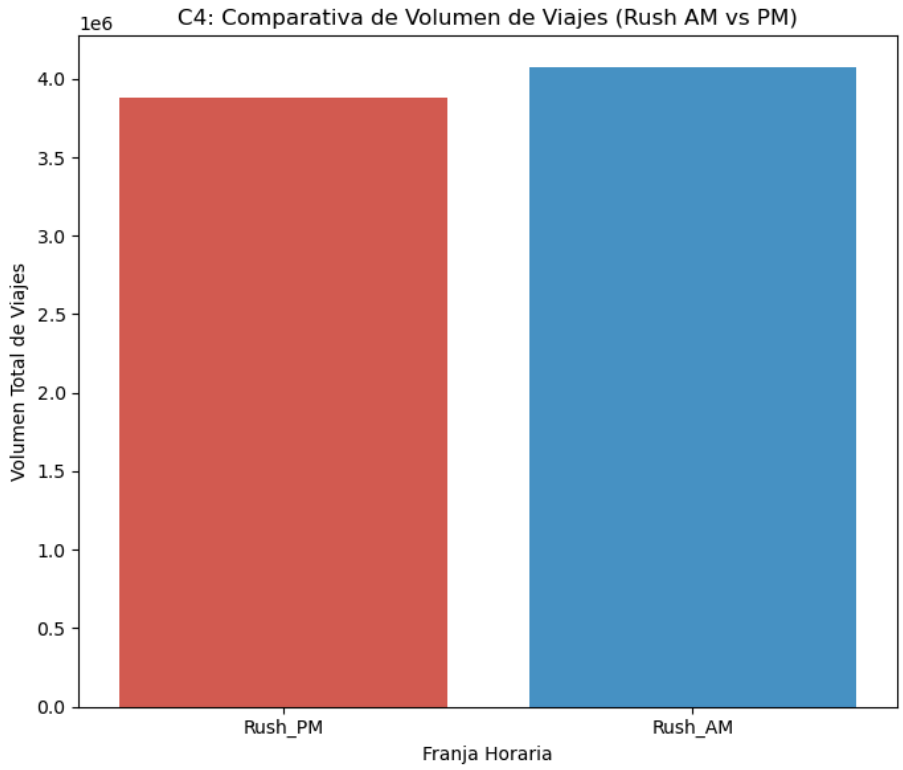
19	1,092,153
20	800,936
21	615,546
22	423,996
23	249,416

El perfil diario de uso de Ecobici revela un comportamiento bimodal (con dos picos claros), característico de los desplazamientos laborales y académicos (*commuting*).

1. Pico Matutino: Se observa un incremento acelerado de la actividad a partir de las 06:00 horas, alcanzando su primer máximo alrededor de las 08:00 horas, coincidiendo con el horario de entrada a oficinas y escuelas.
2. Pico Vespertino: Se registra el punto de mayor saturación del sistema a las 18:00 horas, superando ligeramente el volumen de la mañana. Esto sugiere que el servicio es una opción preferente para el retorno a casa. Entre estos dos picos (de 10:00 a 16:00 horas), la demanda se mantiene estable en un "valle" operativo, mientras que el uso desciende al mínimo durante la madrugada (01:00 - 05:00 horas). Estos datos confirman que el sistema funciona principalmente como un medio de transporte utilitario sincronizado con la jornada laboral de la Ciudad de México.

Top rutas por horario (mareas de movilidad)

La siguiente tabla presenta las rutas (par origen-destino) con mayor frecuencia de uso, segmentadas por franjas horarias de alta demanda (*Rush AM* vs *Rush PM*).



Franja Horaria	Estación Retiro	Estación Arribo	Total de Viajes
Rush_AM	621	618	3,351
Rush_AM	237-238	235-236	2,338
Rush_AM	492	474	2,073

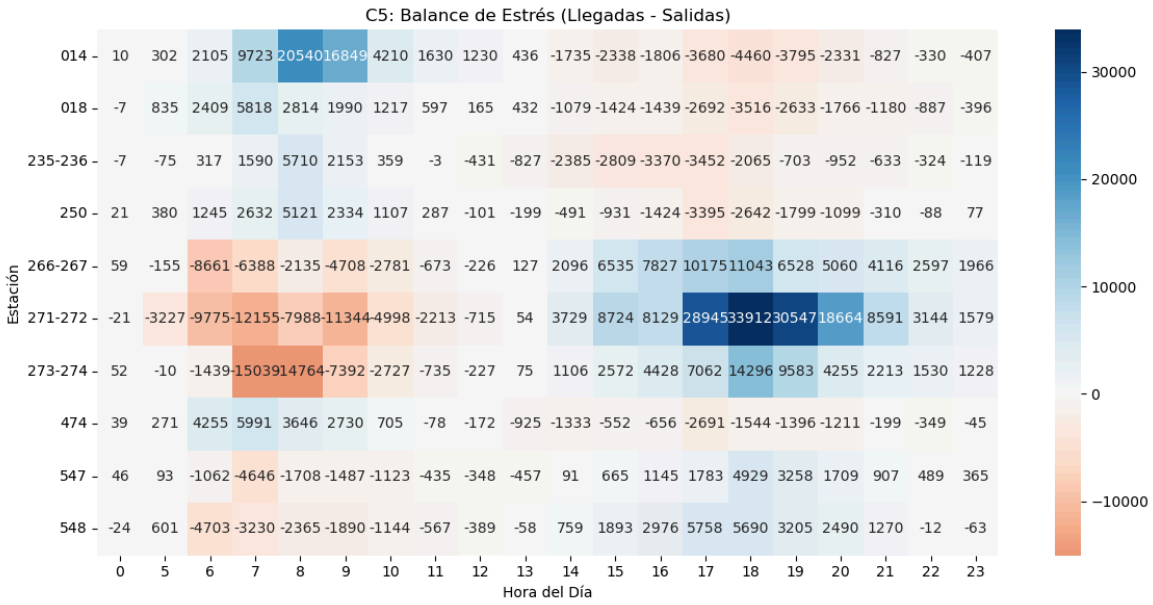
Rush_AM	174	257	1,995
Rush_AM	020	014	1,964

Al analizar las rutas específicas con mayor densidad de tráfico, los resultados muestran que el "Top 5" absoluto de viajes está dominado exclusivamente por la franja matutina (Rush AM). La ruta entre la estación 621 y la 618 se posiciona como la conexión más crítica del sistema con 3,351 viajes, superando por un amplio margen al resto.

Este fenómeno indica que la movilidad matutina es altamente estructurada y repetitiva (patrones pendulares fijos hacia zonas de oficinas o escuelas), lo que genera una concentración masiva de usuarios en trayectos específicos. En contraste, la ausencia de rutas vespertinas (PM) en este ranking superior sugiere que el regreso a casa es más disperso: los usuarios toman rutas más variadas o utilizan estaciones diferentes, diluyendo la densidad por ruta individual a pesar de que el volumen total del sistema sea alto en la tarde.

Balance de Estrés y Desequilibrio Operativo por Estación

Esta consulta mide el "Balance Neto" de bicicletas (Llegadas - Salidas) agrupado por hora, identificando las estaciones que experimentan el mayor estrés operativo. Un valor alto (positivo o negativo) indica la necesidad de intervención logística para reequilibrar la oferta.



Ranking	ID Estación	Máximo Desbalance Neto
1	271-272	33,912
2	014	20,540
3	273-274	14,296
4	266-267	11,043
5	474	5,991
6	018	5,818

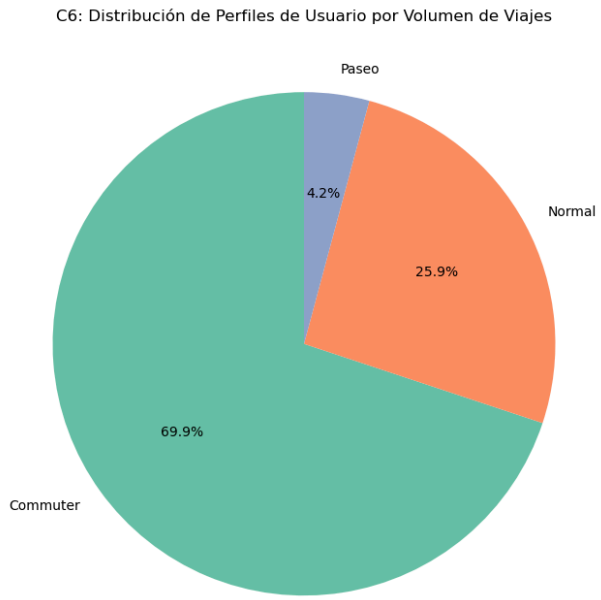
7	548	5,758
8	235-236	5,710
9	250	5,121
10	547	4,929

El análisis del Balance Neto confirma que el desequilibrio de bicicletas es un reto operacional severo, especialmente concentrado en los principales *hubs* de la red. La estación 271-272, que ya se había identificado como el punto de mayor retiro (Consulta 1), también presenta el Máximo Desbalance Neto (33,912 unidades).

Este desequilibrio extremo indica que, en algún momento del día (típicamente durante la hora pico), esta estación experimenta una demanda de salida (retiros) muy superior a la de llegadas (arribos), provocando que se vacíe completamente. Por el contrario, la estación 014 (20,540) podría estar experimentando el fenómeno opuesto, recibiendo un exceso de bicicletas y tendiendo a la saturación. Este ranking señala las estaciones prioritarias para que el equipo de logística de Ecobici intervenga con la reubicación de unidades, ya que su mal manejo impacta la disponibilidad del servicio para miles de usuarios.

Perfiles de usuario (segmentación por comportamiento)

La siguiente tabla presenta la segmentación de los viajes en tres perfiles distintos, definidos en función de la duración del trayecto y si el viaje fue de punto a punto o circular (retorno a la misma estación).



Tipo de Usuario	Total de Viajes	Duración Promedio (min)
Commuter (Utilitario)	11,946,048	12.51
Normal (Estándar)	4,427,739	19.08
Paseo (Recreativo)	715,304	63.58

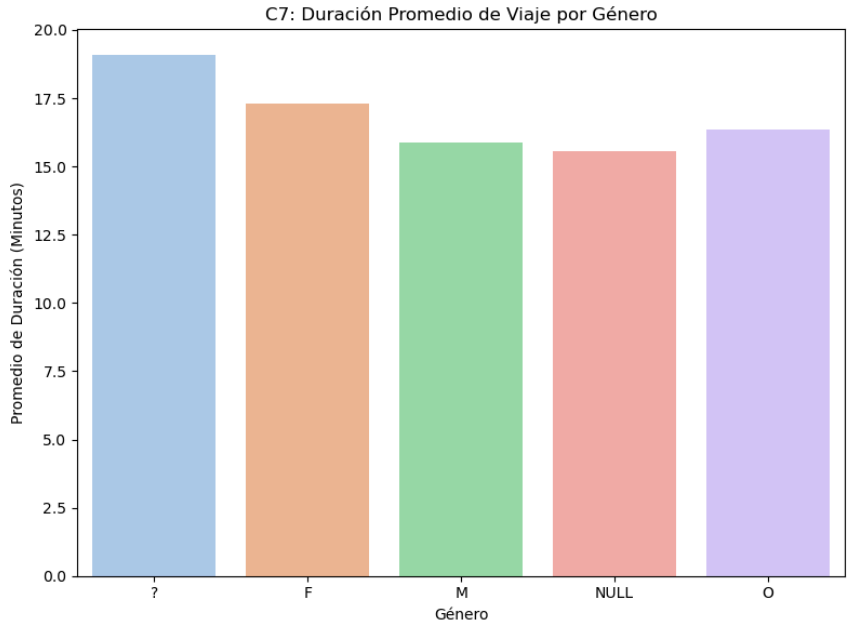
La segmentación de usuarios revela que el sistema Ecobici es predominantemente una

herramienta de transporte utilitario y eficiente. El perfil "Commuter" (usuarios que realizan trayectos rápidos de punto a punto entre 5 y 25 minutos) representa la gran mayoría de la operación, con casi 12 millones de viajes y un tiempo promedio muy ágil de 12.51 minutos. Esto valida que el sistema es esencial para la movilidad diaria ("última milla") de la fuerza laboral.

Por otro lado, el perfil "Paseo" (viajes mayores a 45 minutos o que regresan a la misma estación) constituye una minoría con apenas 715 mil registros, pero con una duración promedio drásticamente superior de 63.58 minutos, lo que indica un uso recreativo o turístico que, aunque menos frecuente, retiene las bicicletas por periodos mucho más prolongados.

Análisis de duración de viaje por género

La siguiente tabla resume la distribución de la duración de los viajes por género, incluyendo métricas robustas como promedio, mediana, percentiles (P25–P75) y variabilidad:



Género	Total de Viajes	Promedio (min)	Mediana (min)	P25 (min)	P75 (min)	Mínimo (min)	Máximo (min)	Desv. Estándar (min)
M	10,469,076	15.91	11.70	6.97	19.62	0.22	1,032,524.97	601.03
F	4,268,818	17.47	12.33	7.45	20.53	0.42	1,248,536.78	1,069.82
O	292,646	16.47	12.08	7.27	20.12	0.47	191,543.88	461.91
?	304,181	19.12	13.60	7.93	23.22	0.27	152,543.45	325.19

Al igual que en la segmentación por tipo de usuario, los datos muestran que la mayoría de los viajes, sin importar el género, se concentran en duraciones cortas, entre 7 y 20 minutos (percentiles 25 y 75).
Esto indica un uso mayoritariamente funcional del sistema, enfocado en desplazamientos cortos, ágiles y recurrentes.

En todos los géneros, la mediana —la medida más robusta del viaje típico— se ubica entre 11.7 y 13.6 minutos, lo cual caracteriza a Ecobici como un sistema de uso cotidiano y eficiente.

Los Hombres (M) son el grupo con viajes más cortos: promedio 15.91 min, mediana 11.7 min. Esto sugiere una mayor proporción de trayectos tipo "última milla" o viajes eficientes punto a punto. El rango intercuartílico (6.97–19.62 min) muestra menor dispersión que otros géneros.

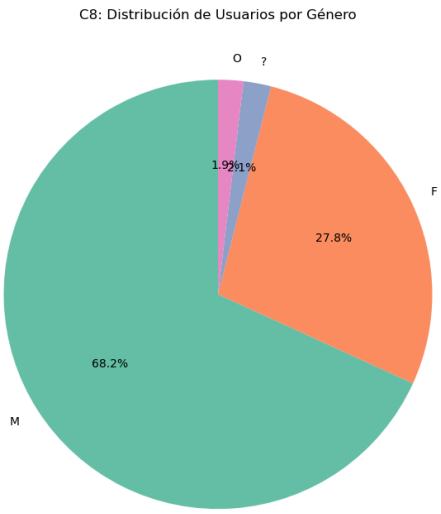
Las Mujeres (F) realizan viajes consistentemente más largos, con un promedio de 17.47 min (≈ +10% respecto a hombres) y una mediana 12.33 min. Esto podría indicar: viajes más largos entre zonas residenciales, rayectos menos directos por temas de seguridad y actividades con múltiples paradas (pattern de movilidad de cuidado). Tienen la mayor variabilidad temporal (desviación estándar 1,069 min), impulsada por viajes largos atípicos.

Otros géneros (O), tienen un promedio y mediana intermedios entre F y M, lo que indica comportamiento mixto o heterogéneo. La dispersión temporal también es menor que la de F, lo que sugiere un patrón más estable.

El grupo "?" tiene los viajes más largos promedio. Con un promedio de 19.12 minutos y la mediana más alta de todos los grupos (13.6 min), este grupo refleja: viajes más largos, menor homogeneidad en patrones, comportamiento atípico o registros incompletos. Es importante señalar que este grupo puede combinar usuarios que no respondieron o errores del sistema.

Distribución de Usuarios por Género

La siguiente tabla muestra la distribución de los viajes registrados en Ecobici según el género reportado por los usuarios. Los valores corresponden al total de viajes y el porcentaje que representan respecto al total.



Género	Total de Viajes	Porcentaje del Total (%)
M	10,469,076	68.27%
F	4,268,818	27.84%
?	304,181	1.98%
O	292,646	1.91%

La distribución revela que Ecobici es un sistema utilizado de manera predominantemente masculina. Con más del 68% de los viajes realizados por usuarios identificados como “M”, se observa un claro desequilibrio en comparación con el 27.84% correspondiente a mujeres. Esta diferencia no solo es amplia, sino también consistente con patrones de uso del espacio público documentados en movilidad urbana: los hombres tienden a utilizar más las modalidades de transporte compartido de micromovilidad.

Los usuarios identificados como hombres representan más de dos tercios de toda la operación, lo cual sugiere una adopción mayor del sistema por parte de la población masculina, posiblemente debido a: una mayor disponibilidad o confianza para desplazarse en bicicleta, diferencias estructurales en horarios y trayectos y factores de seguridad que históricamente afectan más a las mujeres en el espacio público.

Las mujeres realizan poco más de una cuarta parte de todos los viajes, lo que aun siendo una proporción considerable, destaca una brecha de género importante en el uso del sistema. La cifra podría reflejar: rutas percibidas como menos seguras, horarios y destinos con menor alineación al servicio, patrones de movilidad asociados al cuidado (acompañamiento o viajes con múltiples paradas) más difíciles de realizar en bici compartida y barreras culturales o de diseño urbano.

Los géneros registrados como O (Otro con un 1.91%) y ? (No especificado / dato incompleto con un 1.98%) representan en conjunto cerca del 4% de la operación. Aunque minoritarios, estos grupos muestran que Ecobici también está siendo utilizado por personas que no se identifican

dentro del binario tradicional o que no registraron su género en la plataforma. Estos usuarios pueden ser relevantes en análisis de inclusión y accesibilidad del sistema.

Consultas implementadas

Durante el proceso se realizaron las siguientes consultas sobre la colección de Ecobici:

- Top 10 estaciones de retiro más usadas
- Duración promedio de viaje
- Perfil diario: horas pico de uso
- Rutas por horario (mareas de movilidad)
- Balance de Estrés
- Segmentación por perfil de usuario

La ejecución exitosa de las consultas analíticas sobre el volumen masivo de datos permitió validar la eficiencia del modelo de procesamiento distribuido basado en PySpark. Esta aproximación tecnológica demostró su capacidad para manejar y ejecutar análisis estadísticos complejos (como el balance de estrés y la segmentación por perfiles) sobre los 1.1 GB de datos y 17 millones de registros en un entorno de cómputo local, confirmando a PySpark como una herramienta esencial para transformar grandes volúmenes de datos semiestructurados en información operativa tangible.

Interpretación general

La ejecución de las ocho consultas analíticas sobre el conjunto de datos masivo de Ecobici permitió transformar más de 17 millones de registros de texto en conocimiento operativo tangible. Los resultados confirman que el sistema opera bajo una dinámica de alta eficiencia utilitaria, caracterizada por viajes cortos dominados por el perfil *Commuter* y una clara estructuración bimodal en los horarios pico. Finalmente, la identificación precisa de los desequilibrios de oferta y demanda por estación mediante el Balance de Estrés proporciona las métricas necesarias para que los gestores de movilidad puedan implementar estrategias logísticas focalizadas y proactivas, optimizando la disponibilidad de bicicletas en los puntos más críticos de la red.

Conclusiones

Bibliografía

Secretaría de Movilidad de la Ciudad de México. (2025). ECOBICI – Datos Abiertos del Sistema de Bicicletas Públicas de la CDMX. Recuperado de: https://ecobici.cdmx.gob.mx/datos-abiertos/

Apache Software Foundation. (2024). PySpark Documentation: Apache Spark™ Python API. Recuperado de: https://spark.apache.org/docs/latest/api/python/

Python Software Foundation. (2024). Python Language Reference, version 3.12. Recuperado de: https://www.python.org/doc/

Matplotlib Developers. (2024). Matplotlib: Visualization with Python. Recuperado de: https://matplotlib.org/stable/

Pandas Development Team. (2024). pandas: Powerful Python Data Analysis Toolkit. Recuperado de: https://pandas.pydata.org/