Análisis de Afluencia en el Metro de CDMX con PySpark

Aguilar Garcia Juliana
Matematicas Aplicadas y Computacion
Universidad Nacional Autonama de Mexico
Facultad de Estudios Superiores Acatlan
Abril 2025

1. Introducción

Esta práctica tiene como objetivo aplicar técnicas de análisis distribuido con Apache Spark en el entorno Databricks. Se analizaron datos reales sobre la afluencia de usuarios en el sistema de metro de la Ciudad de México, con énfasis en patrones de uso por estación, línea, tipo de pago y fechas específicas.

2. Dataset

El dataset utilizado proviene del portal de Datos Abiertos del Gobierno de la Ciudad de México.

Fuente: https://datos.cdmx.gob.mx/dataset/?organization=secretaria-de-movilidad
Archivo: AfluenciaMetro.csv
Tamaño: ~51 MB
Estructura:
- fecha (date) Día del registro.
- mes (string) Nombre del mes.
- anio (int) Año del registro.
- linea (string) Línea del metro.
- estacion (string) Nombre de la estación.
- tipo_pago (string) Forma de ingreso (Prepago, Boleto, Gratuidad).
- afluencia (int) Número de personas que ingresaron ese día, por tipo de pago.

Figura 1. Diagrama de flujo del proceso de análisis de movilidad.

3. Operaciones realizadas

✔ Lectura de datos: Se utilizó spark.read.csv para cargar el archivo en un DataFrame.
✔ Filtrado de datos: Se filtraron registros con valores nulos y se verificó el correcto tipo de dato en las columnas.
✔ Agrupaciones y resúmenes:
- Afluencia total por estación.
- Afluencia por tipo de pago.
- Estaciones más usadas por línea.
✔ Visualización: Se utilizaron gráficas con matplotlib y pandas para representar los resultados.
✔ Uso de IA: Durante la práctica, se utilizó IA para corregir errores de sintaxis y depurar fragmentos de código.

Estas operaciones permiten detectar patrones de uso y puntos críticos en la movilidad urbana.

4. Resultados principales

- Estaciones más concurridas

estacion	total_afluencia
Pantitlan	200749838
Constitucion de 1917	127996492
Indios Verdes	122995337
Tacubaya	98817240
Cuatro Caminos	94991610

- Tipo de pago mas usado

tipo_pago	total_afluencia
Prepago	3055225264
Boleto	770301660
Gratuidad	581730500

- Líneas del metro con mayor afluencia

linea	total_afluencia
Linea 2	754898822
Linea 3	659785638
Linea B	496690484

5. Fragmentos clave de código con comentarios

```
# Cargar la tabla generada en Databricks
df = spark.sql("SELECT * FROM default.afluencia_metro")

# Agrupar por estación y sumar la afluencia total
estaciones_top = df.groupBy("estacion") \
   .agg(sum("afluencia").alias("total_afluencia")) \
   .orderBy(col("total_afluencia").desc())

# Visualizar las estaciones más concurridas
top_estaciones_pd = estaciones_top.limit(10).toPandas()
plt.bar(top_estaciones_pd['estacion'], top_estaciones_pd['total_afluencia'])
plt.xticks(rotation=45)
plt.title("Top 10 estaciones con más afluencia")
```

```
plt.show()
```

6. Uso IA

La IA se utilize como herramiennta de apoyo puntual para depurar errores de sintaxis, sintetizar informacion relacionada con Databricks despues de visualizer videos en Udemy asi como guia dentro de la misma interfaz.

7. Conclusión

Esta práctica reforzó los conocimientos sobre el manejo de volúmenes de datos mediante PySpark, desde la lectura hasta la visualización. Además, dejo como aprendizaje el uso de herramientas como Databricks e IA como soporte técnico en el desarrollo eficiente de análisis.