Anteproyecto: Análisis de movilidad urbana con datos de taxis en la Ciudad de Nueva York

Fecha: 11/11/2025
DNI: 38.730.377
Nombre: Braian Alejandro Pucheta
Carrera: Licenciatura en Ciencia de Datos
Cátedra: Análisis de Datos Masivos
Docente: Ing. Gustavo Rivadera
Plan: 2022

1. Introducción

La movilidad urbana en Nueva York genera millones de viajes en taxi diariamente, produciendo un volumen masivo de datos con información valiosa sobre patrones de desplazamiento, zonas de demanda y comportamiento temporal. Con este proyecto propongo analizar los datos públicos de los taxis amarillos (NYC Yellow Taxi Trip Records) para extraer datos significativos que permitan entender y mejorar el servicio de transporte urbano.

El dataset, disponible en formato **Parquet**, contiene millones de registros mensuales con información detallada: ubicaciones de origen y destino, horarios, distancias, tarifas y propinas. Este volumen y complejidad requiere técnicas especializadas de Big Data para su procesamiento efectivo.

La solución se presentará como una aplicación web interactiva desarrollada con Streamlit que mediante visualizaciones intuitivas, permitirá la exploración dinámica de los datos sin requerir conocimientos técnicos avanzados.

Se busca integrar conceptos de análisis de datos masivos vistos en la materia, incluyendo procesamiento distribuido con PySpark, paradigma MapReduce, limpieza de datos, consultas SQL y gráficos interactivos.

2. Definición del problema

¿Cómo aprovechar los datos masivos de taxis amarillos de NYC para entender patrones de movilidad urbana y generar información útil que mejore la toma de decisiones? Esta pregunta me llevó a plantear varios desafíos específicos que decidí analizar paso a paso. El primer desafío es el volumen de datos: archivos mensuales con entre 3 y 5 millones de registros que limitan las capacidades de herramientas como Pandas en memoria local.

El segundo es la calidad de datos: presencia de valores faltantes, outliers extremos (distancias superiores a 100 millas, tarifas negativas), identificadores de zonas fuera del rango válido del sistema TLC, y ruido en mediciones temporales.

El tercer desafío es la complejidad analítica: responder preguntas como identificar zonas de alta demanda por horario, comparar patrones entre días laborables y fines de semana, analizar relación entre distancia, propina y método de pago, y detectar combinaciones de zonas de origen-destino más rentables. Finalmente, existe un desafío de accesibilidad: presentar hallazgos de forma comprensible para usuarios no técnicos mediante interfaces intuitivas.

3. Objetivos del proyecto

El objetivo general del proyecto es desarrollar una solución de análisis de datos masivos que procese, limpie y visualice información de viajes en taxi de NYC, generando información aplicable sobre movilidad urbana mediante una aplicación web interactiva.

Objetivos específicos: Primero, implementar infraestructura de procesamiento distribuido con PySpark para manejar archivos Parquet aplicando el paradigma MapReduce y Spark SQL. Segundo, garantizar la calidad de datos mediante identificación y tratamiento sistemático de inconsistencias, aplicando transformaciones y categorización espacial.

Tercero, realizar análisis exploratorio calculando estadísticas descriptivas, identificando patrones temporales (horas pico, variaciones semanales) y analizando distribuciones espaciales de demanda. Cuarto, desarrollar una aplicación web con Streamlit con gráficos dinámicos, mapas de calor geoespaciales y filtros interactivos generando insights concretos sobre eficiencia operativa, identificando oportunidades de optimización en zonas y horarios específicos.

4. Marco teórico aplicado

Big Data y las tres V: El proyecto trabaja con las características fundamentales de datos masivos: Volumen (millones de registros mensuales), Velocidad (procesamiento distribuido eficiente) y Variedad (campos numéricos, categóricos y temporales). Se aplican conceptos de datificación y cuantificación.

Infraestructura Spark: Con Apache Spark se realizará el procesamiento distribuido de los datos, aprovechando su ejecución en memoria y su modelo basado en DAG para optimizar el rendimiento. El uso de DataFrames y Spark SQL permitirá aplicar transformaciones y consultas eficientes sobre grandes volúmenes de información.

Paradigma MapReduce: Heredado por Spark, este paradigma estructura operaciones en fase Map (filtrar o extraer componentes temporales) y fase Reduce (agregar resultados, como contar viajes por zona o promediar tarifas). La fase Shuffle intermedia redistribuye datos según claves de agregación para procesamiento conjunto.

Formato Parquet: Base de datos analítica orientada a columnas que permite lectura eficiente de campos específicos, compresión optimizada y almacenamiento de metadatos para omitir bloques irrelevantes. Compatible nativamente con ecosistemas distribuidos, facilita procesamiento sin conversiones intermedias.

5. Metodología

Se adoptará un enfoque de desarrollo iterativo que iniciará con el procesamiento de un mes de datos, con el fin de validar el funcionamiento completo del pipeline antes de ampliarlo a períodos mayores. La solución se estructura en tres capas: procesamiento distribuido (backend PySpark con transformaciones y Spark SQL), análisis (consultas complejas y cálculos estadísticos) y presentación (frontend Streamlit con visualizaciones interactivas).

Adquisición de datos: Los archivos Parquet mensuales se descargarán desde NYC TLC con campos esenciales como: timestamps de inicio y fin del viaje, identificadores de zonas de origen y destino, distancia recorrida, duración, tarifas, propinas y método de pago. Se trabajará con 1-3 meses recientes equilibrando volumen representativo y tiempos de procesamiento razonables.

Limpieza de datos: Proceso en cuatro fases: exploración inicial (validar esquema y estadísticas descriptivas), identificación de problemas (cuantificar nulos, outliers e inconsistencias), transformación (filtrar registros inválidos con umbrales contextuales), y enriquecimiento (categorizar franjas horarias, calcular métricas derivadas como velocidad promedio).

Análisis y visualización: El análisis exploratorio abordará tres dimensiones: temporal (distribución por hora/día, patrones laborables vs. fines de semana), espacial (rutas frecuentes) y económico (correlación distancia-tarifa, propinas por método de pago, rentabilidad por segmento). La aplicación Streamlit organizará contenido en bloques con filtros dinámicos, gráficos interactivos Plotly, mapas Folium y optimizaciones mediante cacheo. El pipeline se organizará como una secuencia de tareas que Spark ejecutará solo cuando sea necesario, optimizando así el rendimiento del procesamiento.

6. Solución del problema

Análisis de patrones: Para identificar zonas de alta demanda por horario, los viajes se agruparán según los identificadores de zona de origen y las franjas horarias, generando representaciones visuales de la distribución de la demanda. La variación entre días laborables y fines de semana se mostrará mediante gráficos comparativos de frecuencia horaria, donde se espera observar picos de actividad en horas punta durante los días laborables.

Relaciones económicas: La relación distancia, propina y método de pago se explorará con diagramas de dispersión codificados por color y cálculos de propina

promedio segmentados. Se espera validar que pagos con tarjeta registren propinas mayores por facilidad de cálculo automático. Las rutas más frecuentes y rentables se identificarán combinando pares origen-destino, balanceando métricas de frecuencia e ingreso por milla/minuto.

Insights accionables: La solución generará recomendaciones concretas: optimización de asignación de flota identificando zonas-horarios desatendidos con alta demanda, análisis de eficiencia mediante velocidad promedio por zona revelando corredores congestionados, oportunidades de precio dinámico en zonas con demanda persistente pero escasa oferta, y factores de calidad inferidos mediante análisis de propinas controlando distancia y método de pago.

Presentación web: La aplicación Streamlit contará con una página principal que mostrará métricas clave y visualizaciones interactivas, junto con secciones para el análisis temporal y espacial. Incluirá opciones básicas de filtrado y exploración de datos para facilitar la interpretación de los resultados.

7. Conclusiones

Este anteproyecto plantea una propuesta teórica para aplicar técnicas de Big Data en el análisis de movilidad urbana a partir de datos de viajes en taxi de Nueva York. El objetivo es mostrar cómo el uso de herramientas como PySpark, Spark SQL, paradigma MapReduce y Streamlit puede facilitar el procesamiento, análisis y visualización de grandes volúmenes de información.

La metodología propuesta busca abordar los principales desafíos del problema: manejar el gran volumen de datos mediante procesamiento distribuido, mejorar la calidad de la información mediante limpieza y validación sistemática, y presentar los resultados de forma comprensible a través de una interfaz web interactiva. El uso del formato Parquet se considera clave para optimizar el almacenamiento y la lectura eficiente de los registros.

En conjunto, el proyecto sintetiza conceptos y técnicas estudiadas durante la cursada, demostrando cómo las herramientas open-source permiten desarrollar soluciones escalables y accesibles para el análisis de datos masivos en contextos reales de movilidad urbana.