Universidad del Valle de Guatemala

CC3084 – Data Science

Laboratorio 8: Spark SQL y DataFrames

Diederich Solís

Sara Guzman

Guatemala, 2025

Índice

1. Introducción

El presente laboratorio tiene como objetivo aplicar los conceptos de procesamiento distribuido mediante **Apache Spark**, utilizando el modelo de *RDDs* y *DataFrames* para realizar análisis exploratorios y consultas analíticas sobre datos de accidentes de tránsito en Guatemala.

Se utilizó **PySpark** en un entorno de **Google Colab** con integración a Google Drive, lo que permitió desarrollar el laboratorio de forma local sin necesidad de Databricks. El dataset contiene información histórica de hechos de tránsito, vehículos involucrados y personas fallecidas o lesionadas, organizadas en tres tablas principales:

- Hechos de tránsito
- Vehículos involucrados
- Fallecidos

El propósito fue combinar, limpiar, agrupar y visualizar la información empleando funciones nativas de Spark SQL y DataFrames, además de generar salidas optimizadas en formato **Parquet** para su análisis posterior.

2. Metodología

El desarrollo se realizó completamente en PySpark siguiendo las siguientes etapas:

1. **Lectura y normalización de datos:** Los archivos CSV fueron cargados desde Google Drive, aplicando limpieza, estandarización de nombres de columnas y filtrado de años válidos (2013–2023).
2. **Transformaciones principales:** Se utilizaron funciones como `withColumn`, `groupBy`, `agg`, `join`, `filter` y `orderBy` para realizar agregaciones, combinaciones y clasificaciones. Además, se crearon nuevas columnas con `when` y `otherwise`.
3. **Visualización:** Se empleó `matplotlib` para graficar resultados en barras, columnas, líneas y pastel, equivalentes a la función `display()` de Databricks.
4. **Almacenamiento:** Los resultados finales se guardaron en formato `Parquet` (incisos 5, 9, 14 y 17) para optimizar su lectura y procesamiento.

3. Resultados y Análisis

Inciso 5: Total de accidentes por año y departamento

Mediante la función `groupBy(.ªnio", "departamento").count()` se determinó el total de accidentes por año y departamento. El año con mayor número de accidentes fue **2023**, con el departamento de **Guatemala** registrando 3,457 casos.

Archivo generado: `accidentes_por_ano_depto.parquet`

Inciso 6: Día con más accidentes en 2023

El análisis mostró que el **domingo** fue el día de la semana con más accidentes, lo cual coincide con un mayor movimiento vehicular recreativo los fines de semana.

Inciso 7: Distribución por hora del día (Municipio de Guatemala)

El histograma evidenció que las horas de **12:00 a 18:00 horas (tarde)** concentran la mayor cantidad de accidentes, coincidiendo con las horas pico de tráfico.

Inciso 8: Unión de hechos y vehículos

La unión mediante una llave compuesta (`anio, mes, dia, hora, departamento, municipio, zona, tipo_accidente`) permitió obtener más de **90,000 registros combinados**, evidenciando la correspondencia entre hechos y vehículos involucrados.

Inciso 9: Promedio de vehículos por accidente

El promedio general fue de **2 vehículos por accidente**. Los departamentos con mayores promedios fueron Guatemala, Santa Rosa y Alta Verapaz.

Archivo generado: `prom_vehiculos_por_accidente.parquet`

Inciso 10: Top 5 colores de vehículos

Los colores más frecuentes en accidentes fueron:

1. Gris
2. Negro

3. Blanco
4. Azul
5. Rojo

Inciso 11: Lesionados por atropello (2023)

Los meses con más lesionados fueron **enero y julio**. La serie temporal presentó una distribución uniforme a lo largo del año sin picos extremos.

Inciso 12: Relación accidentes-fallecidos

El tipo de accidente con mayor número de fallecidos fue la **colisión**, seguido por choques y derrames, mostrando que la mayoría de muertes se asocian a impactos de alta velocidad.

Inciso 13: Franjas horarias

Los accidentes se distribuyeron principalmente en:

- Tarde (12–18 h)
- Noche (18–24 h)
- Mañana y madrugada (menor incidencia)

Inciso 14: Ratio de fallecidos por accidente

Los departamentos con ratios más altos fueron **Huehuetenango** y **Alta Verapaz**, debido a un menor número de accidentes pero una mayor mortalidad por evento.

Archivo generado: `ratio_fallecidos_por_accidente.parquet`

Inciso 15: Grupos de edad más afectados

El grupo de edad **15–29 años** fue el más afectado, tanto en fallecidos como lesionados, indicando una alta vulnerabilidad de los conductores jóvenes.

Inciso 16: Accidentes y fallecidos por zona (Municipio de Guatemala)

Las zonas **6, 12 y 18** presentaron los mayores índices de accidentes y fallecidos, asociadas a alta densidad vehicular y tráfico pesado.

Inciso 17: Porcentaje de accidentes por sexo del conductor

El 86.5 % de los accidentes involucraron a **conductores hombres**, frente al 13.5 % de mujeres.

Archivo generado: `porcentaje_accidentes_por_sexo.parquet`

4. Conclusiones

1. Apache Spark permitió realizar análisis distribuidos de datos masivos de forma eficiente y reproducible.
2. Los resultados confirman que los **jóvenes hombres** representan el grupo más vulnerable en accidentes viales.
3. Las horas de la **tarde y noche** concentran la mayor incidencia de accidentes, especialmente los fines de semana.
4. El departamento de **Guatemala** es el más crítico en número total de accidentes.
5. El uso del formato **Parquet** optimizó la consulta y almacenamiento de resultados.

5. Archivos Generados

Archivo	Contenido
`accidentes_por_ano_depto.parquet`	Total de accidentes por año y departamento
`prom_vehiculos_por_accidente.parquet`	Promedio de vehículos por accidente
`ratio_fallecidos_por_accidente.parquet`	Relación entre fallecidos y accidentes
`porcentaje_accidentes_por_sexo.parquet`	Distribución de accidentes por género