Laboratorio 8.
Transformaciones con Spark

INSTRUCCIONES:

En esta guía se detallan las instrucciones para poder completar los ejercicios que complementan el contenido visto en la semana. El objetivo de la misma es que el estudiante se familiarice y conozca cómo utilizar los diferentes métodos de transformación de DataFrames dentro del nmódulo del Spark SQL. Esta es la transformación principal de datos utilizando este framework.

Realizar de forma **grupal**, si surge alguna duda siéntase en la comodidad y libertad de hacer preguntas al catedrático. Para este único laboratorio, deberá utilizar su ambiente dentro de **Databricks**.

PREPARACIÓN DEL AMBIENTE

Descargue en sus equipos personales **la base de datos de los principales resultados de la PNC** para los accidentes de tránsito hasta 2024 siguiendo **este enlace**. Dentro del material proporcionado, copie el archivo que descargó dentro de una carpeta en su ambiente de trabajo (workspace) de su ambiente personal de Databricks. Esto lo puede lograr de la siguiente manera: seleccione la opción de New → Add or upload data desde el menú del lado izquierdo. En el apartado de **Files** escoja la opción Upload files to a volume. Importe el archivo desde su equipo, y en la parte inferior de los catálogos disponibles, escoja un lugar en donde desea almacenar estos datos. Por ejemplo, usando su workspace siga la ruta de All catalogs → workspace → default → <coloque un nombre de volumen a su elección>. Esto le dará como resultado una ruta de tipo **/Volumes/workspace/default/<my volume>** que podrá referenciar luego en su Notebook.

EJERCICIOS

La fuente de datos corresponde a los accidentes de tránsito en Guatemala (2020–2024), divididos en tres grandes bases:

- Hechos de tránsito: número de accidentes clasificados por año, mes, departamento, día de la semana, hora, tipo de accidente, y en algunos casos por zona (en el municipio de Guatemala).

- Vehículos involucrados: cantidad y características de vehículos (tipo, color, modelo), además de información sobre los conductores (sexo, edad, condición).

- Fallecidos y lesionados: víctimas por accidentes, con desagregación por edad, sexo, tipo de accidente, hora, día, mes y departamento.

Un detalle importante es que los datos están en distintas hojas de un mismo Excel, y no existe un ID único que relacione directamente las tres bases. Para unirlas habrá que apoyarse en columnas comunes como: año, mes, día, hora, departamento, zona, tipo de accidente, etc.

CARGA DE DATOS Y ANÁLISIS EXPLORATORIO (20 PTS)

Como primer paso, deberá procesar dicho archivo de excel. Cada hoja del Excel contiene una tabla diferente (hechos, vehículos, fallecidos/lesionados), puede consultar el índice del mismo para mayor detalle. Utilizando puramente python con las librerías que ya conoce, deberá convertir las hojas que sean de su interés en archivos que pueda leer con spark posteriormente.

Deberá cargar los datos que vaya a utilizar utilizando la función **spark.read** y también deberá apoyarse de las definiciones del *header* y el *inferSchema* de dicho método para manejar las columnas según corresponda. A continuación, responda las siguientes preguntas:

1. (5 pts) Mostrar cuántos registros hay en cada tabla (hechos, vehículos, fallecidos, lesionados). Muestre algunos resultados con la función **.show()**. Genere un describe y summary para aquellas columnas que considere importantes según cada archivo.
2. (5 pts) Identificar los años disponibles en cada tabla y validar si coinciden.
3. (5 pts) Mostrar los valores distintos de tipo de accidente.
4. (5 pts) Calcular cuántos departamentos únicos aparecen en las bases.

PREGUNTAS DE ANÁLISIS (80 PTS)

5. (5 pts) ¿Cuál es el total de accidentes por año y departamento? Apóyese de la función **groupBy**. Investigue la función **display** que tiene Databricks y muestre su resultado en formato de gráfico de barras.
6. (5 pts) ¿Qué día de la semana registra más accidentes en 2024? Graficar con display en un gráfico de columnas.
7. (5 pts) Mostrar la distribución de accidentes por hora del día en el municipio de Guatemala. Graficar en un histograma.
8. (10 pts) Unir la tabla de hechos de tránsito con la de vehículos usando una llave compuesta por año, mes, departamento y tipo de accidente. ¿Cuántos registros combinados se logran?
9. (10 pts) De la unión anterior, calcular el promedio de vehículos por accidente en cada departamento. Guardar este resultado en formato Parquet. Luego, vuelva a cargarlo y grafique los 10 departamentos con más vehículos/accidente.
10. (5 pts) Encontrar el top 5 de colores de vehículos más involucrados en accidentes.
11. (5 pts) Calcular cuántos lesionados por atropello hubo en 2024, por mes. Graficar en serie temporal (línea).

12. (10 pts) Relacionar accidentes con fallecidos usando llaves (año, mes, departamento, tipo de accidente). Calcular el total de fallecidos por cada tipo de accidente. Graficar en barras horizontales.
13. (5 pts) Usar **withColumn** para clasificar accidentes en franjas horarias: Mañana [6-12), Tarde [12-18), Noche [18-24), Madrugada [0-6). Mostrar cuántos accidentes ocurren en cada franja.
14. (5 pts) Calcular el ratio de fallecidos por accidente en cada departamento (fallecidos / accidentes). Guardar el resultado en Parquet.
15. (5 pts) Identificar los grupos de edad más afectados en fallecidos y lesionados. Graficar en barras que permitan comparar a ambos grupos.
16. (7 pts) Calcular, para el municipio de Guatemala, cuántos accidentes hay por zona y cuántos fallecidos se reportan en cada una. Generar un gráfico de barras con ambos indicadores.
17. (8 pts) Crear un DataFrame que muestre el porcentaje de accidentes donde el conductor era hombre vs mujer (tabla vehículos). Guardar como Parquet. Finalmente, vuélvalo a cargar y grafique con display en gráfico de pie.

EVALUACIÓN

- **(20 puntos)** Haber completado con exactitud la sección de Análisis Exploratorio.
- **(80 puntos)** Haber respondido de manera correcta evidenciando el paso a paso de su solución para todas las preguntas de análisis.

NOTA: para poder obtener derecho a nota, cada estudiante del grupo deberá haber entregado el ejercicio del día Lunes 29 de Septiembre: Ejercicio - Spark SQL + DataFrames.

MATERIAL A ENTREGAR

- Script de Python (.ipynb) que utilizó con la discusión generada utilizando markdown.

FECHAS DE ENTREGA

- **AVANCE:** Jueves 2 de octubre de 2025 17:20hrs: Actividades de la **1 a la 7** de la sección de ejercicios.
- **DOCUMENTO FINAL COMPLETO:** Domingo 5 de octubre de 2025 a las 23:59.

Sugerencia: El segundo día de clase de la semana tendrá un tiempo de aclaración de dudas con el profesor, se le sugiere que avance en la resolución del laboratorio en los pasos del contenido teórico visto en la clase presencial para que aclare todas sus dudas al respecto en dicho espacio.