Tarea: Analítica de Datos con PySpark

Introducción

En esta actividad, los estudiantes aplicarán los conceptos de analítica de datos utilizando PySpark, un entorno distribuido que permite procesar grandes volúmenes de información de manera eficiente. La práctica se centrará en el manejo de archivos CSV, realizando operaciones de lectura, limpieza, análisis exploratorio y escritura de resultados en nuevos formatos. El propósito es desarrollar habilidades para el tratamiento de datos estructurados empleando un enfoque analítico y reproducible.

Objetivo

Desarrollar un flujo completo de análisis de datos utilizando PySpark sobre un conjunto de datos en formato CSV, incluyendo los pasos de carga, limpieza, transformación, análisis y exportación de resultados.

Metodología

1. Descargue el archivo de datos proporcionado en formato CSV (por ejemplo, 'ventas.csv').
2. Cargue el archivo en un DataFrame de PySpark utilizando el método `spark.read.csv()` con la opción `header=True`.
3. Realice un análisis exploratorio de los datos mostrando el esquema (`printSchema()`) y las primeras filas (`show()`).
4. Limpie los datos eliminando registros nulos o inconsistentes utilizando `dropna()` o `fillna()`.
5. Ejecute operaciones analíticas como:
   - Agrupamiento (`groupBy`) por categoría o región.
   - Cálculo de promedios, máximos y mínimos (`agg`, `avg`, `max`, `min`).
   - Filtrado de registros mediante condiciones lógicas.
6. Guarde los resultados en un nuevo archivo CSV usando `write.csv()` dentro de una carpeta llamada 'output_csv'.
7. Documente su proceso dentro del notebook con comentarios claros que describan cada paso.

Rúbrica de Evaluación

Criterio	Excelente (90-100%)	Satisfactorio (70-89%)	Insuficiente (<70%)
Lectura y exploración del CSV	Carga correcta del archivo CSV y muestra del	Carga parcial o con errores menores.	No se logra cargar el archivo o se omite la exploración inicial.

	esquema con explicación detallada.		
Limpieza y transformación de datos	Aplica correctamente limpieza, filtrado y transformación de columnas con justificación.	Aplica algunas transformaciones pero con errores o sin explicación.	No realiza limpieza o las transformaciones son incorrectas.
Análisis y operaciones agregadas	Realiza cálculos analíticos relevantes con funciones de agrupamiento y agregación.	Realiza operaciones simples pero incompletas o con errores de interpretación.	No aplica funciones analíticas o las aplica incorrectamente.
Exportación y documentación	Guarda correctamente el resultado en CSV y documenta cada paso del proceso.	Exporta resultados parcialmente o con errores de formato.	No exporta resultados ni documenta el proceso.

Valor total: 100 puntos.