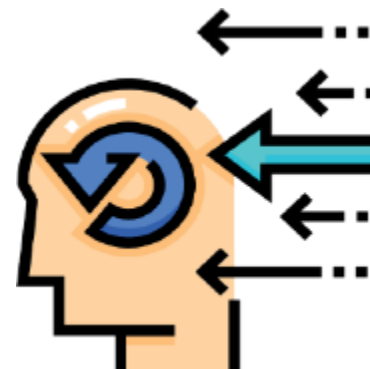




Módulo 7

Sesión N° 2



ACTIVIDAD:



Análisis de Datos con PySpark: Transformaciones, Filtrado y Consultas SQL

- Objetivo: Aplicar transformaciones y consultas en PySpark mediante la creación y manipulación de RDDs y DataFrames, para procesar y estructurar datos en contextos de Big Data.



Instrucciones:

1. Requerimientos específicos:

- Instalar PySpark
- Los estudiantes trabajarán con una lista de datos que simula transacciones de ventas.

```
4s !pip install pyspark

Requirement already satisfied: pyspark in /usr/local/lib/python3.11/dist-packages (3.5.5)
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.11/dist-packages (from pyspark) (0.10.9.7)

0s [2] transacciones = [
    ("Chile", "Electrónica", 1200),
    ("Perú", "Ropa", 800),
    ("México", "Electrónica", 1500),
    ("Chile", "Alimentos", 700),
    ("Perú", "Electrónica", 1000),
    ("México", "Ropa", 600)
]
```

2. Instrucciones:

- Parte 1: Crea y Manipula un RDD
 - Crear un RDD a partir de la lista transacciones usando parallelize().
 - Aplicar un filter() para seleccionar solo transacciones del país "Chile".
 - Convertir el RDD filtrado en mayúsculas utilizando map().
- Parte 2: Crea y consulta un DataFrame
 - Convertir la lista transacciones en un DataFrame con columnas: País, Categoría, Monto.
 - Filtrar las transacciones donde el monto sea mayor a 1000.
 - Consultar las ventas totales por país utilizando Spark SQL.





- Parte 3: Consultas SQL sobre el DataFrame:
 - Registrar el DataFrame como tabla temporal.
 - Ejecutar una consulta SQL para obtener el total de ventas por categoría.
 - Mostrar los resultados finales.
- Parte 4: Evaluación de la Actividad:
 - Creación de RDD y filtrado correcto.
 - Conversión a DataFrame y ejecución de consultas SQL.
 - Interpretación correcta de los resultados obtenidos.

Código Sugerido:

<https://colab.research.google.com/drive/1OZEFyb8a5nVenQUY35yDLuaibAISHcg0?usp=sharing>

