

Introducción

En este análisis se simuló el flujo de clics de usuarios utilizando **Python, PySpark y Pandas** para observar cómo evoluciona la interacción de clics en una plataforma fitness tipo **Gymshark**.

Código utilizado

```
# =====
# * Gym Analytics - Análisis de Flujo
# =====

from pyspark.sql import SparkSession
from pyspark.sql.functions import avg
import pandas as pd
import matplotlib.pyplot as plt
import random
import time

# -----
# [1] Crear dataset simulado (solo la primera vez)
# -----
data = [
    (1, "Camiseta Compresión", 12, 300, "Colombia"),
    (2, "Leggins Mujer", 8, 240, "México"),
    (3, "Short Hombre", 20, 410, "España"),
    (4, "Top Deportivo", 7, 190, "Argentina"),
    (5, "Suplementos", 15, 380, "Chile"),
]

columns = ["usuario_id", "producto", "clicks", "tiempo_sesion", "pais"]
df = pd.DataFrame(data, columns=columns)
df.to_csv("data_simulada.csv", index=False)
print("☑ Dataset 'data_simulada.csv' generado con éxito.\n")

# -----
# [2] Procesamiento por lotes con PySpark
# -----
spark = SparkSession.builder.appName("AnalisisFlujoGym").getOrCreate()
df_spark = spark.read.csv("data_simulada.csv", header=True, inferSchema=True)

metricas = df_spark.groupBy("pais").agg(
    avg("clicks").alias("Promedio_Clicks"),
    avg("tiempo_sesion").alias("Promedio_Tiempo_Sesion")
)

print("📊 Métricas promedio por país:")
metricas.show()

# -----
```

```

# [3] Simulación de flujo (Streaming)
# -----
print("\n⚡ Simulando flujo de clics por usuario...\n")

usuarios = [1, 2, 3, 4, 5]
clicks_totales = {u: 0 for u in usuarios}

for minuto in range(1, 6): # simulación de 5 minutos
    print(f"--- Minuto {minuto} ---")
    for u in usuarios:
        nuevos_clicks = random.randint(1, 10)
        clicks_totales[u] += nuevos_clicks
        print(f"Usuario {u}: +{nuevos_clicks} clics (total: {clicks_totales[u]})")
    print()
    time.sleep(0.5) # pequeña pausa simulando tiempo real

# -----
# [4] Visualización de clics por usuario
# -----
usuarios_list = list(clicks_totales.keys())
clicks_list = list(clicks_totales.values())

plt.figure(figsize=(7, 4))
plt.bar(usuarios_list, clicks_list, color="#af0b0b")
plt.title("Clics Totales por Usuario (Simulación Streaming)", fontsize=12)
plt.xlabel("Usuario ID")
plt.ylabel("Clics Totales")
plt.grid(axis='y', linestyle='--', alpha=0.6)
plt.tight_layout()
plt.savefig("grafico_clicks.png", dpi=120)
plt.close()

print("▣ Gráfico 'grafico_clicks.png' generado correctamente.\n")

# -----
# [5] Reflexión breve
# -----
print("💡 Reflexión:")
print("El procesamiento por lotes analiza datos históricos en bloque,")
print("mientras que el streaming permite analizar datos en tiempo real para")
print("tomar decisiones inmediatas.\n")

spark.stop()

```