Big Data Aplicado

UD 6 - Apache Hadoop

Práctica 2 Spark

Entendiendo las diferencias entre APIs

> Usando cualquiera de las opciones disponibles Spark (cluster propio, docker o Databricks), realiza la siguiente práctica

1. Imagina que eres un científico/a de datos que tiene que analizar un conjunto de datos de Formula 1 de la temporada 2023 utilizando Apache Spark
2. Los datos incluyen tiempos de vuelta de pilotos en todos los circuitos, información de pilotos y equipos, y detalles de los circuitos.
3. La práctica tiene como objetivo extraer insights sobre el rendimiento de pilotos y equipos, así como entender cómo diversos factores influyen en los resultados de las carreras.
    i. Lista de pilotos con sus respectivos equipos: pilotos_df
    ii. Lista de circuitos con detalles adicionales: circuitos_df
    iii. Tiempos por vuelta de todos los pilotos en cada circuito de la temporada: vueltas_df
4. Realiza los siguientes apartados con estos datos facilitados

5. Debes realizar los ejercicios usando las **3 APIs vistas en clase**, para observar así las diferencias entre ellos:

    i. DataFrame API
    ii. PySapk SQL
    iii. Pandas on Pyspark

1. Calcula el tiempo medio por vuelta de cada piloto en toda la temporada.
2. Identifica el piloto con la vuelta más rápida en cada circuito.
3. Determina la vuelta más rápida de la temporada y el piloto.
4. Analiza la consistencia de los pilotos mediante la desviación estándar de sus tiempos de vuelta.
5. Comparación de Equipos: Evalúa el rendimiento general de los equipos comparando los tiempos promedio de vuelta de sus pilotos.
6. Análisis de Circuitos: Determinar cuáles circuitos presentan mayores desafíos para los pilotos, basándose en la variabilidad de los tiempos de vuelta.

Entrega:

La práctica debe ser entregada como un notebook de Jupyter o un script de Python que incluya comentarios explicativos sobre cada paso del análisis, asegurando que el código sea comprensible y bien organizado.