Big Data Aplicado

UD 6 - Apache Hadoop

Práctica 3 Spark

Usando Datos reales

> 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 reales de Formula 1 de la temporada 2024 utilizando Apache Spark
2. Todos los datos los tienes en este repositorio de kaggle
3. Tienes información completa de todos los campeonatos desde 1950 hasta la última temporada completa.
4. 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.
5. Realiza los siguientes apartados con estos datos facilitados
6. Puedes realizar los ejercicios usando **cualquiera** de las **3 APIs vistas en clase**, y combinarlas como quieras:
    i. DataFrame API
    ii. PySapk SQL

iii. Pandas on Pyspark

Análisis de los Datos

Ejercicio 1: Análisis de los tiempos de los pilotos en las Carreras

***Objetivo**: Analizar los tiempos de vuelta de los pilotos para identificar los mejores rendimientos en cada circuito.*

1. Carga los datos de tiempos por vueltas, circuitos, pilotos y carreras.
2. Explora los DataFrames para entender su estructura y los datos que contienen
3. Prepara los datos. Antes de analizar los tiempos de vuelta, necesitarás unir los tiempos por vuelta con circuitos y pilotos.
4. Identifica el piloto con la vuelta más rápida en cada circuito.
5. Identificar Mejores Rendimientos de cada circuito (Mejores medias de tiempos por vuelta)

Ejercicio 2: Análisis del Impacto de las Paradas en Boxes en los Resultados de las Carreras

***Objetivo**: Investigar cómo las paradas en boxes afectan el rendimiento de los pilotos en las carreras y determinar si ciertas estrategias de paradas en boxes están correlacionadas con mejores resultados.*

1. Carga los datos de tiempos de pit stop y resultados.
2. Explora los DataFrames para entender su estructura y los datos que contienen
3. Prepara los datos. Deberías unir los resultados con los pit_stop.
4. Calcular el impacto de los pit stops: Necesitamos obtener datos clave para evaluar el impacto de los pit stops en el rendimiento, como el tiempo total en pit stops por carrera y el número de paradas, y compáralo con la posición final en la carrera
5. Identificación de patrones o estrategias que pueden ser exitosas. De los datos que has obtenido, que posibles conclusiones o estrategias obtienes? Hay alguna correlación? Podrías ampliarlas o mejorarlas usando más datos? Cuáles y cómo los obtendrías?

Ejercicio 3: Business Intelligence. Analizar el impacto de los tipos de circuito y la posición histórica de los pilotos en cada circuito

***Objetivo**: Descubrir patrones y tendencias que puedan predecir el rendimiento de pilotos y equipos, lo cual es crucial para la planificación estratégica y la toma de decisiones.*

1. Preparación de Datos. Carga los datos de circuitos, resultados, tiempos por vuelta y pilotos.

2. Relacionar las carreras con los circuitos
3. Cálculo de posiciones históricas por circuitos de cada piloto. Podríamos predecir cual podría ser su rendimiento actual que nos influye en la estrategia.
4. Cálculo de tiempos medios de cada piloto por circuitos. Podríamos predecir cual podría ser su rendimiento actual que nos influye en la estrategia.
5. Para un mejor análisis, crea un dataframe que incluya los 2 puntos anteriores
6. Realiza una predicción de la posición del piloto en cada circuito de la temporada actual teniendo en cuenta la media histórica de tiempos, de posición y la experiencia del piloto
7. Visualiza los datos.

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.