- CICLO DE VIDA ANALÍTICO DEL DATO -

Uso de PIG y HIVE

1. Introducción

En esta práctica vamos a trabajar con el lenguaje HiveQL, PigLatin y Hadoop. Dicha práctica la realizaremos sobre la shell de Hive y Pig, llamadas hive shell y grunt respectivamente. Desde esta shell iremos escribiendo las instrucciones sin necesidad de crear un script.

En un año de Olimpiadas, ¡qué mejor que trabajar con un conjunto de datos de resultados de 120 años!. Para esta práctica vamos a utilizar un fichero de datos que contiene los atletas y resultados de los 120 años de historia de las Olimpiadas: desde Atenas 1896 a Río 2016.

Dicho conjunto de datos ha sido obtenido de: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results donde tenéis información sobre el contenido del mismo: campos, estructura, etc.

A continuación se enumeran las diferentes consultas a realizar tanto en Pig como en Hive. La solución a los ejercicios debéis enviarlos a través de la plataforma en la tarea que se ha creado para dicho propósito.

2. Ejercicios obligatorios

Estos ejercicios deberéis realizarlos tanto en PIG como HIVE.

2.1 Carga de dataset (1 punto)

NOTA: Para mostrar las evidencias de cada punto, incluye capturas de pantalla, así como la instrucción ejecutada.

1. **(0,25 puntos)** Crea una carpeta en HDFS en la ruta **/hadoop/dataset**.
2. Descarga el fichero *athleteEvents.zip* desde la plataforma virtual y cópialo en la ruta que acabas de crear.
3. Para comprobar que el archivo se ha copiado bien:
    a. **(0,25 puntos)** Lista el contenido de la carpeta, incluyendo el tamaño del archivo.
    b. **(0,25 puntos)** Muestra las primeras líneas del fichero, leyendo directamente de HDFS (sin copiar al sistema de ficheros local) en línea de comandos.
4. **(0,25 puntos)** Crea la carpeta **/hadoop/out** en HDFS para alojar los resultados de los ejercicios.

2.2 Consultas (9 puntos)

Rellena esta tabla con el tiempo (segundos) que tarda en realizarse cada consulta **(0,5 puntos)**.

Consultas	Pig Local	Pig MapReduce	Hive
C1			
C2			
C3			
C4			

Analiza los resultados obtenidos e intenta justificar porqué se obtienen esos resultados. ¿Cuál es más rápido Pig o Hive? ¿Por qué? **(1 punto)**

2.2.1. Pig (4 puntos)

Abre la shell de Pig en modo local y ejecuta las siguientes operaciones sobre el dataset.

1. **(C1. 0,5 puntos)** Lista el deportista (o deportistas) más jóvenes en conseguir una medalla de oro. Muestra el nombre, juegos, año, deporte y evento.
2. **(C2. 0,5 puntos)** Lista todos los deportistas españoles que han participado en las Olimpiadas a lo largo de la historia. Muestra todos los campos, ordenando el listado de manera alfabética por nombre.

3. **(C3. 1 punto)** ¿Cuál es el deportista (o deportistas) mejor/es en la disciplina "*Sailing*" en los 120 años? Lista qué número de Olimpiadas se han realizado por continente, listando el año, la ciudad y si han sido Juegos Olímpicos de verano o invierno.
4. **(C4. 1 punto)** ¿Cuál es el deportista (o deportistas) que más medallas ha conseguido?. Muestra el nombre y el número de medallas de oro, plata y bronce conseguidas. ¿Coincide con el atleta que más medallas de oro tiene en la historia de los Juegos Olímpicos?

Ahora cierra la shell de Pig y vuelve a arrancarlo en modo **mapreduce**. Ejecuta las consultas anteriores comprobando cuántos trabajos de Hadoop son necesarios para cada una **(1 punto)**.

2.2.2. Hive (3,5 puntos)

Abre la shell de Hive y ejecuta las siguientes operaciones.

1. **(0,5 puntos)** Realiza lo siguiente:
    a. Crea una base de datos que se llame **practicas** y actívala.
    b. Crea una tabla para alojar los datos del dataset. Debes tener en cuenta el formato del fichero y los campos que tiene.
    c. Carga el fichero en la tabla y haz una primera selección de todos los campos con un máximo de 10 registros.

Vamos ahora a repetir cada consulta que hemos realizado con Pig para así comprobar la expresividad de ambos lenguajes. Entra en la consola de Hive y obtén:

1. **(C1. 0,5 puntos)** Lista el deportista (o deportistas) más jóvenes en conseguir una medalla de oro. Muestra el nombre, juegos, año, deporte y evento.
2. **(C2. 0,5 puntos)** Lista todos los deportistas españoles que han participado en las Olimpiadas a lo largo de la historia. Muestra todos los campos, ordenando el listado de manera alfabética por nombre.
3. **(C3. 1 punto)** ¿Cuál es el deportista (o deportistas) mejor/es en la disciplina "*Sailing*" en los 120 años? Lista qué número de Olimpiadas se han realizado por continente, listando el año, la ciudad y si han sido Juegos Olímpicos de verano o invierno.
4. **(C4. 1 punto)** ¿Cuál es el deportista (o deportistas) que más medallas ha conseguido?. Muestra el nombre y el número de medallas de oro, plata y bronce conseguidas. ¿Coincide con el atleta que más medallas de oro tiene en la historia de los Juegos Olímpicos?

3. Material a entregar

Subid a Moodle un fichero único en formato zip que contenga:

- Memoria en formato PDF que incluya las respuestas a cada cuestión planteada, incluyendo una captura de pantalla que evidencie la instrucción o comando ejecutados así como el resultado obtenido.
- Adjuntar un fichero TXT con los comandos ejecutados para resolver cada cuestión.

FECHA MÁXIMA DE ENTREGA: 17 DE OCTUBRE DE 2021