BigData

EVALUACIÓN

PROGRAMACIÓN CON SPARK Y LAS API DE ALTO Y BAJO NIVEL

02 AL 05 DE ABRIL

CRITERIOS DE EVALUACIÓN:

- Comprensión de los conceptos básicos de Spark
- Capacidad para manipular y procesar grandes conjuntos de datos utilizando Spark
- Habilidad para implementar algoritmos desarrollados en Spark sobre Hadoop

Los criterios para autoevaluación, coevaluación y heteroevaluación, se define a partir de la rúbrica anexa.

Punto 1: Aplicar los conceptos de RDD en Spark. 35 pts: Se evalúa la manipulación de RDD mediante operaciones de transformación y acción y la optimización de rendimiento utilizando RDD. **También se evalúa el tiempo de ejecución**

Este punto toma en cuentas los dataset de Nasdaq y de companylist. Recuerde el formato de datos es:

Para NASDAQ: exchange, símbolo de la acción, fecha, precio de apertura de la acción, precio máximo de la acción, precio mínimo de la acción, precio de cierre de la acción, volumen de la acción y precio ajustado de cierre de la acción.

Para companylist: Símbolo, Nombre, año oferta pública inicial IPOyear y sector de la industria

1. Calcule, **para cada año** del DataSet dados para el punto 1, **qué sector tuvo el mayor número de operaciones**. La salida debe mencionar el año, el nombre del sector y el valor global de operaciones. El resultado debe ser parecido a:

   Finance,1996,20090342
   Pharma,1996,12312312
   Finance,1997,25612312

   Entregable 1: script de spark donde se use RDD para resolver la problemática, con el nombre ***1_topsectorperyear.py.*** se debe explicar dentro del script las líneas fundamentales de código
   Entregable 2*: Archivo de salida con los resultados, con el nombre 1****_out.txt***

2. Calcule, para cada empresa y sector comercial, qué empresa creció más por año, enumerando también el porcentaje de crecimiento. Los resultados deben estar en un formato similar a:

   Finance,1996,ABCD,46%
   Finance,1997,VFER,64%

   Entregable 3: script de spark donde se use RDD para resolver la problemática, con el nombre **2_*topcompanypersector.py.*** se debe explicar dentro del script las líneas fundamentales de código

Entregable 4*: Archivo de salida con los resultados, con el nombre* ***2_out.txt***

Punto 2: Aplicar los conceptos de API estructurada (30 puntos) se evalúa los conceptos de la API estructurada de Spark, la manipulación de DataFrames mediante operaciones de transformación y acción. Este punto trabaja con los datos de películas que están en **ml-25m.zip. También se evalúa el tiempo de ejecución**

3. Para cada género, encuentre la calificación promedio del género y la cantidad de películas que pertenecen a este género. Si una película pertenece a más de un género, considere el mismo puntaje en cada género. Los resultados deben estar en un formato similar a:

   Crimen, 3.1625, 905
   Romance, 3.156, 1205
   Thriller, 3.148, 1425

   Entregable 5: script de spark donde se use AÏ de alto nivel para resolver la problemática, con el nombre **3_ *genreaveragefilms.py.*** se debe explicar dentro del script las líneas fundamentales de código

   Entregable 6*: Archivo de salida con los resultados, con el nombre* ***3_out.txt***

Punto 3: Aplicar los conceptos de API estructurada y de RDD (35 puntos) . este punto se realiza con el data set que descarga del siguiente link https://www.kaggle.com/datasets/arnabchaki/popular-video-games-1980-2023 El DataSet contiene una lista de videojuegos entre 1980 y 2023, con fechas de lanzamiento, calificación de revisión de usuarios y calificación de revisión de críticos. **También se evalúa el tiempo de ejecución y que el RDD se divida en máximo 2 partes.**

4. ¿Cuál es el videojuego más popular, por cada género y por cada desarrollador Team? Se entregan dos archivos diferentes de resultados
   Entregable 7: script de spark con el nombre 4_ v***ideogames.py.*** se debe explicar dentro del script las líneas fundamentales de código
   Entregable 8*: Archivo de salida con los resultados, con el nombre* ***4_genreout.txt y 4_teamout.txt y todas las partes generadas***

5. ¿por género, cuál es el promedio de puntuación de los video juegos y el promedio del número de caracteres de las opiniones de los usuarios?
   Entregable 9: script de spark con el nombre **5_average*videogames.py.*** se debe explicar dentro del script las líneas fundamentales de código
   Entregable 10*: Archivo de salida con los resultados, con el nombre* ***5_out.txt y todas las partes generadas***