TEC | Tecnológico de Costa Rica

Instituto Tecnológico de Costa Rica

Campus Tecnológico Central Cartago

Escuela de Ingeniería en Computación

Bases de Datos II

IC-4302

Resumen 3

Fecha de entrega: 21/03/2023

I-Semestre 2023

Profesor:

Nereo Campos Araya.

Elaborado por:

Diana Sanabria Calvo 2021436548.

--------Apache Spark--------

Podemos ver que la generación diaria de datos aumenta exponencialmente, también lo hace la necesidad de un procesamiento efectivo de esos datos. Los sistemas de procesamiento de datos tradicionales, como los basados en bases de datos relacionales , tampoco son escalables para grandes volúmenes de datos. También se observa que el artículo argumenta que los métodos de procesamiento de datos convencionales son insuficientes para el procesamiento de datos en tiempo real y como resultado, no pueden ofrecer resultados en tiempo real.

Para superar estas dificultades y ofrecer un método escalable y efectivo para procesar grandes volúmenes de datos , se necesita un enfoque novedoso para el procesamiento de grandes conjuntos de datos. Debido a la arquitectura de hardware inherente, la complejidad de los algoritmos empleados y la falta de flexibilidad en el procesamiento de datos en tiempo real, los sistemas de procesamiento de datos tradicionales tienen limitaciones en términos de escalabilidad y rendimiento. Para procesar conjuntos de datos masivos en clústeres de computadoras, Apache Spark es un sistema de procesamiento de datos unificados que ofrece una interfaz de programación sencilla y consistente.

Además, Apache Spark ofrece una biblioteca de operaciones distribuidas para el procesamiento de datos en memoria, lo que acelera el procesamiento de datos al almacenarlos en la memoria del clúster en lugar de en el disco. A lo largo del artículo se afirma que la arquitectura Apache Spark consta de una serie de partes, incluido Spark Core, que proporciona las capacidades informáticas y de escalabilidad fundamentales del sistema; los módulos Spark SQL y DataFrames , que brindan capacidades para procesar datos estructurados y los módulos Spark Streaming y Structured Streaming , que ofrecen capacidades de procesamiento de datos en tiempo real.

El modelo de programación de Apache Spark se basa en la idea de los RDD(conjuntos de datos distribuidos resistentes), que son colecciones inmutables de objetos que se pueden distribuir en un grupo de computadoras y procesar en paralelo y los RDD permiten el procesamiento de datos en paralelo, lo que mejora el rendimiento del sistema. Para manejar conjuntos de datos de cualquier tamaño, desde gigabytes hasta petabytes, Apache Spark está diseñado para ser altamente escalable. Esto es posible mediante la división de datos y cálculos entre grupos de computadoras, además de tener la opción de agregar o eliminar nodos sobre la marcha para adaptarse a las cambiantes demandas de procesamiento de datos. Debido al uso de RAM y la optimización de las operaciones de procesamiento de datos en la memoria, Apache Spark es más rápido que muchos otros sistemas de procesamiento de datos.

La abstracción RDD (Resilient Distributed Datasets) sirve como base para el modelo de programación de Spark, que permite el procesamiento paralelo de colecciones inmutables de objetos distribuidos en un grupo de computadoras. Se explica cómo las transformaciones y las acciones son los dos tipos de operaciones de RDD y como devuelven un resultado al agente de usuario o escriben datos en un almacenamiento externo, como contar o escribir, mientras que las transformaciones son operaciones que crean un nuevo RDD a partir de uno existente, como filtrar o mapear.