24/11/2023

Spark

Introducción

Sistema de computación basado en map reduce, motor de procesoado de datos (no los alamacena. Trabaja en **paralelo**

Ventajas (respecto a Hadoop):

* El procesamiento es en memoria, va mucho mas rápido que en disco. Menos operaciones de escritura/lcutra
* Framework de procesameitno unificado (se puede ejecutar Spark en Java, R, Python…)
* Se puede utilizar sparkSQL. Manejar una gran cantidad de datos utilizado SQL
* Interactuar con múltiples sistemas de almacenamiento

Spark puede leer de Kafka, Postgre, HDFS, elastic, Warehouse, S3, Cassandra, etc…

Cores de Spark

* Spark Core: conjunto de librerías
* Spark SQL: procesamiento de datos estructurado. Pensado para tratamiento de datos
* Saprk Streaming: Ingesta en tiempo real.
* Spark MLLib: librería con algoritmos de Machine Learning
* Spark Graph:
* Spark R: permite ejecutar Spark en R

Arquitectura

* Program Driver: el programa que contiene el método principal (punto de partida). Reside en el nodo maestro.
* Woker: instancias donde los ejectores se alojan para ejecutar el codigo
* Executor: RAM del nodo esclavo donde reside el bloque de datos
* Cluster Manager: responsable de asignar los recursos

Spark RDD

Resilient Distributed Dataset (RDD): colección distribuida inmutable de objetos. Cada conjunto de datos se puede calcular en distintos nodos (procesar libros por ejemplo)

Utilian avaluación perezosa. Mantiene las tranformaciones en grafos. Es una estrategia de evalución que retrasa el calcúlo hasta que es necesario para operaciones posteriores. Sus beneficios son:

* Incremento del rendimiento al evitar cálculos innecesarios.
* Construir estructuras de datos potencialmente infinitas
* Definir estructuras de control como abstracciones (en lugar de operaciones primitivas)

Operaciones RDD (Acciones)

Devolver valores

Operaciones RDD (Transformaciones)

Devuelde un RDD

Spark SQL

Procesamiento de datos estructurados. Tipos de datos

* Dataframe: conjunto de datos organizados
* Dataset: conjunto de datos

Spark Streaming

Procesamiento en tiempo real. Características princpilaes:

* Tecnología popular y madura
* Gran comunidad de desarrolladores
* APIs para programar Java, Scalara
* LIbrerias de Machine Learning
* No es streaming real. No es ideal cuando se buscan latencias de menos de 0.5 segundos
* Tolerante a fallos
* Muchos parámetros de configuración

Stream discreto que se parte en batch. Solo se puede trabajar con RDD

Operadores en DSTREAM

Operadores en ventana