Ferramentas para **Big Data**

Apache Spark

Novas Tecnologias em Bd
Aluno: Mario Francisco Ponce Jr

O que é o Apache Spark?

Apache Spark é uma poderosa ferramenta de processamento de big data que oferece diversas funcionalidades para processamento distribuído e análise de dados em larga escala.

- Permite o processamento rápido e eficiente de grandes volumes de dados.
- Oferece suporte a diversas tarefas, como processamento em lote, processamento de streaming, consultas SQL e aprendizado de máquina.
- Possui uma arquitetura flexível e escalável.
- Permite a execução em clusters de computadores para lidar com cargas de trabalho intensivas.
- Fabricante: Apache Software Foundation

Principais Características

Processamento In-Memory

O Apache Spark realiza processamento em memória, o que aumenta significativamente a velocidade de processamento e análise de dados.

Execução de Consulta Adaptativa

é uma funcionalidade do Spark SQL que adapta o plano de execução em tempo de execução, como definir automaticamente o número de reducers e algoritmos de junção.

Estruturadas e dados não estruturados

O Spark SQL trabalha com tabelas estruturadas e dados não estruturados, como JSON ou imagens.

Vantagens e Desvantagens

Vantagens

- Processamento de dados em larga escala.
- Velocidade de execução.
- Múltiplas linguagens de programação.
- Gratuito e de código aberto (open source)

Desvantagens

- Alto consumo de recursos.
- Muito complexo para usuários iniciantes.

Requisitos

- Java 8 ou superior
- Sistema operacional compatível (Linux, macOS, Windows)
- Recursos de hardware adequados para o processamento desejado

Características Relevantes

- Suporte a processamento em memória (processamento mais rápido e eficiente)
- Capacidade de processar diferentes tipos de dados (batch, streaming, SQL)
- Bibliotecas e APIs extensíveis para tarefas específicas (machine learning, processamento de gráficos, etc.)

Biografia

- https://spark.apache.org/docs/latest/