PySpark e Apache Kafka Para Processamento de Dados em Batch e Streaming

O Que é Processamento Distribuído?

O processamento distribuído refere-se a um sistema no qual componentes de software localizados em computadores interconectados comunicam-se e interagem para alcançar um objetivo comum. Nesse tipo de sistema, as tarefas são divididas em partes menores e distribuídas por diferentes máquinas, permitindo que o trabalho seja realizado em paralelo, otimizando o desempenho e a eficiência.

Existem vários motivos e vantagens para adotar o processamento distribuído:

Escala: Permite que sistemas manipulem grandes volumes de dados e tráfego, distribuindo a carga entre vários computadores ou servidores.

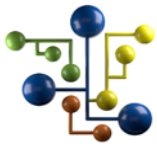
Resiliência: Como as tarefas são distribuídas por várias máquinas, se uma máquina falhar, o sistema como um todo pode continuar a funcionar, dependendo de como é projetado.

Flexibilidade: Os sistemas distribuídos podem ser expandidos adicionando-se mais máquinas ao sistema sem a necessidade de interromper ou reconfigurar todo o sistema.

Otimização de Recursos: A capacidade de processamento, memória e armazenamento de várias máquinas pode ser combinada, permitindo que tarefas complexas sejam realizadas mais rapidamente do que em uma única máquina.

É importante notar que, enquanto os sistemas distribuídos oferecem muitas vantagens, eles também apresentam desafios. A coordenação entre os nós, a garantia de consistência de dados, a recuperação de falhas e a segurança são questões complexas que devem ser cuidadosamente gerenciadas em ambientes distribuídos.

E esse é um dos temas estudados neste curso!



Equipe DSA

Muito Obrigado!
Continue Trilhando Uma Excelente Jornada de Aprendizagem.