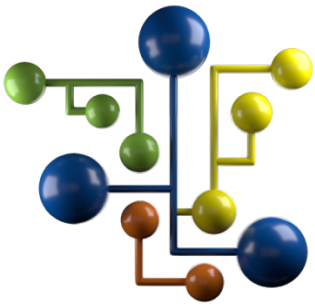


Data Science Academy

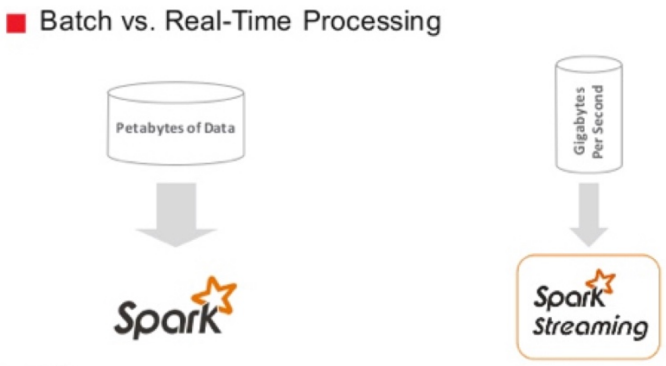
www.datascienceacademy.com.br

Engenharia de Dados com Hadoop e Spark

Apache Spark Streaming

Muito do processamento realizado em procedimentos analíticos são feitos em dados devidamente armazenados e disponíveis, seja em bancos de dados ou arquivos csv. Carregamos os dados, fazemos alguns filtros e aplicamos técnicas de análise de dados. Esse procedimento funciona bem para resolver alguns problemas de negócio. Mas e quando precisamos realizar análises que não podem esperar todo o processo de carga, filtro e manipulação dos dados? Por exemplo: detecção de fraudes com cartões de crédito. Faz sentido aguardar 5 horas, para processar dados e então obter uma visão sobre o que está acontecendo? Precisamos detectar a fraude no momento em que ela ocorre e para isso precisamos analisar dados em tempo real. Esta é a proposta do Spark Streaming.

A vida não acontece em batches (lotes), sendo na verdade um fluxo contínuo de acontecimentos. Muitos dos sistemas que desejamos monitorar e entender, acontecem como um fluxo contínuo de eventos - batimentos cardíacos, correntes oceânicas, métricas de máquinas, ou sinais de GPS. A lista, assim como os eventos, é essencialmente infinita. É natural, então, que possamos recolher e analisar informações a partir desses eventos, como um fluxo de dados. Mesmo análise de eventos esporádicos, como o tráfego de web sites pode se beneficiar de uma abordagem de streaming de dados.



Com Spark podemos manipular os dados de acordo com a forma como eles são gerados. Se tivermos um grande volume de dados por exemplo de transações comerciais de uma rede de varejo ao longo de um ano, podemos processar isso em batch. Carregamos os dados uma única

vez, processamos e analisamos os dados. Mas podemos também, coletar dados à medida que eles são gerados, processar e analisar. O Spark suporta as duas abordagens.

Há muitas vantagens potenciais de manipulação de dados como fluxos, mas até recentemente esse era um trabalho difícil. Atualmente Streaming de dados e análises em tempo real estão se tornando cada vez mais o padrão do mercado. E por que existe agora uma explosão de interesse em streaming de dados? A resposta a essa pergunta é que as novas tecnologias agora estão disponíveis para lidar com streaming de dados em níveis de alto desempenho, grande escala e de forma muito fácil - o que está levando mais empresas a lidar com dados como um stream.

O Apache Spark Streaming é um sistema de processamento de streaming, tolerante a falhas e escalável, o que significa que rapidamente podemos aumentar a quantidade de nodes em um cluster e assim expandir sua capacidade.

Por ser parte do framework, o Spark Streaming se integra com o MLLib, o Spark SQL e o Graphx. A partir da versão 2.0, o Spark Streaming suporta streaming estruturado de dados com o Streaming DataFrame. O Spark Streaming pode receber dados de diversas fontes e produzir resultados que podem ser usados por diversas soluções do mercado.

O Spark Streaming funciona com o conceito de microbatching para simular análise de fluxo em tempo real. Isso significa que um programa em batch é executado em intervalos frequentes para processar todos os dados ao longo do streaming. Embora esta abordagem seja inadequada para aplicações de baixa latência (aquelas que realmente requerem "real, real-time"), é uma maneira inteligente de utilizar pequenos processos em lote (microbatching) para se aproximar de uma atividade em tempo real e funciona bem para muitas situações.

Quer dizer que o Spark Streaming não é ”real” real-time?

Sim, isso mesmo. Na prática, o que esse módulo faz é gerar micro-batches a partir do fluxo de dados capturados e com isso simular um processamento em tempo real. Não deixa de ser genial, mas na prática, o Spark Streaming não é o que chamamos de real, real-time. Mas como os micro-batches são velozes e processados em memória, temos a impressão de estarmos trabalhando com dados em tempo real. Isso será muito bom em algumas situações, mas não será em outras. Lembre-se: não existe tecnologia perfeita.

O Spark Streaming pode receber dados de diversas fontes e para cada uma dessas fontes haverá um receiver. Aqui as fontes de dados suportadas pelo Spark Streaming:

- Flat Files (à medida que são criados)
- TCP/IP
- Apache Flume
- Apache Kafka
- Amazon Kinesis
- Mídias Sociais (Facebook, Twitter, etc...)
- Bancos NoSQL
- Bancos Relacionais

Uma importante vantagem de usar o Spark para Big Data Analytics é a possibilidade de combinar processamento em batch e processamento de streaming em um único sistema.

Streaming de dados é considerado uma das tecnologias mais promissoras e o Apache Streaming, além de ser fácil de utilizar, é totalmente integrado as demais bibliotecas do Spark, permitindo a criação de soluções bem robustas.

Com o Apache Streaming podemos: coletar e analisar dados direto da fonte e à medida que são gerados, transformar e sumarizar os dados, aplicar modelos de Machine Learning e fazer previsões em tempo real.