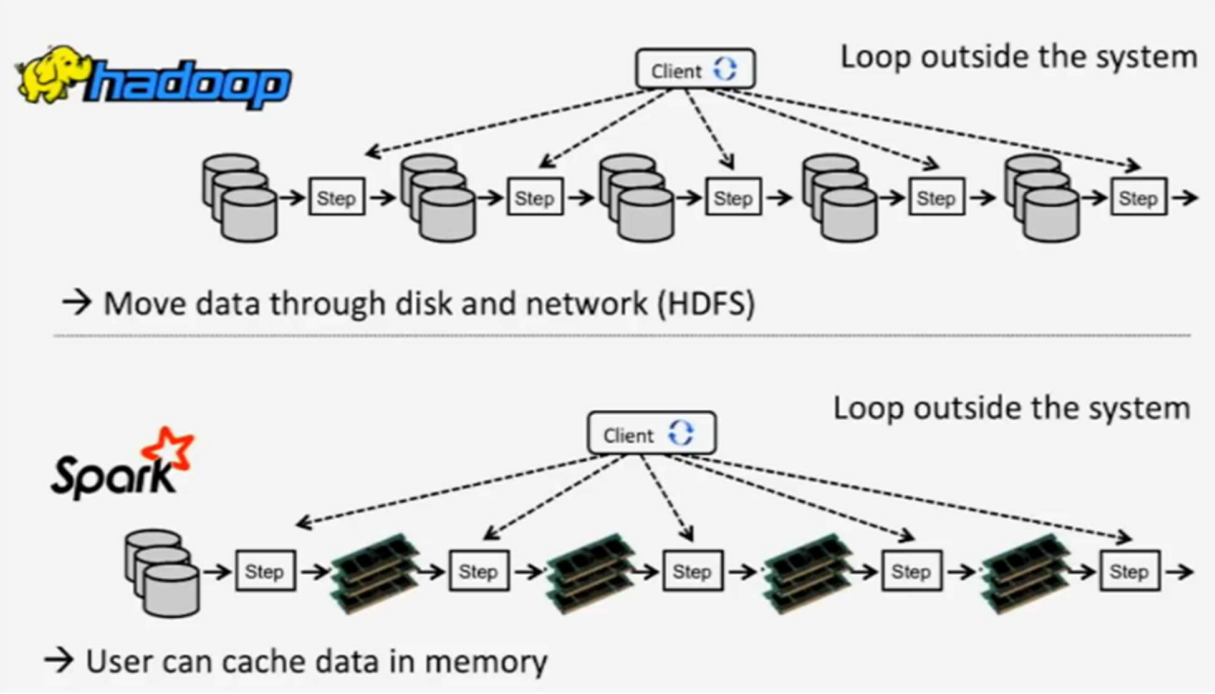
Machine Learning e IA em Ambientes Distribuídos Versão 2.0

Apache Spark e Big Data Analytics

Apache Spark é um dos assuntos mais quentes do momento em tecnologias de Big Data Analytics. A quantidade de dados gerados em todo o mundo aumenta de forma exponencial e o Spark é claramente a solução computacional expressamente concebida para lidar com este nível de crescimento. Primeiramente criado como parte de um projeto de pesquisa na Universidade de Berkeley nos EUA, Spark é um projeto open source no universo do Big Data, construído para análises sofisticadas, velocidade de processamento e facilidade de uso. Ele unifica capacidades críticas de análise de dados, como SQL, análise avançada em Machine Learning e streaming de dados, tudo isso em uma única estrutura.

O Spark tem muitas vantagens se comparado as outras tecnologias de Big Data e do paradigma MapReduce, como o Hadoop MapReduce e o Storm. Inicialmente, o Spark oferece um framework unificado e de fácil compreensão para gerenciar e processar Big Data com uma variedade de conjuntos de dados de diversas naturezas (por exemplo: texto, grafos, etc), bem como de diferentes origens (carga de dados em batch ou streaming de dados em tempo real).

O Spark permite que aplicações executem até 100 vezes mais rápido em memória e até 10 vezes mais rápido em disco, do que usando Hadoop MapReduce. Permite o desenvolvimento rápido de aplicações em Java, Scala ou Python, além de linguagem R. Além disso, vem com um conjunto integrado de mais de 80 operadores de alto nível e pode ser usado de forma interativa para consultar dados diretamente do console. Além das operações de Map/Reduce, suporta consultas SQL, streaming de dados, aprendizado de máquina e processamento de grafos. Desenvolvedores podem usar esses recursos no modo stand-alone ou combiná-los em um único pipeline. Neste capítulo, veremos o que é o Spark, como ele se compara com uma solução típica com MapReduce e disponibiliza um conjunto completo de ferramentas para processamento de Big Data.



Entre as principais características do Spark, podemos citar:

- Spark realiza operações de MapReduce
- Spark pode utilizar o HDFS
- Spark permite construir um workflow de Analytics
- Spark utiliza a memória do computador de forma diferente e eficiente
- Spark é veloz
- Spark é flexível
- Spark é gratuito

E por que aprender a usar o Apache Spark? Por diversas razões: é atualmente uma das tecnologias mais quentes em Big Data Analytics, devido sua velocidade de processamento. Mais e mais empresas estão adotando infraestrutura de Big Data que tem o Spark como um dos componentes principais. Existe cada vez mais suporte de outras empresas e existe alta demanda por profissionais que conheçam processamento de dados em tempo real. Portanto, existem diversas razões pelas quais você deveria aprender a usar o Spark.

O Spark é um tema avançado, que requer conhecimentos em ciência da computação para que seja bem compreendido. Ele é poderoso e isso tem um preço. Praticamente todas as operações em Spark são feitas via linha de comando e não via interface gráfica. Não teremos agora o recurso de arrastar e soltar. Teremos que construir todas as operações que serão executadas. Precisamos mais do que nunca da sua dedicação. Exatamente por se tratar de um tema avançado, são poucos os profissionais que dominam o Spark a ponto de criar soluções de análise de dados em tempo real. Nós vamos guiá-lo pelo processo de aprendizagem, mas sua dedicação é fundamental.

A melhor fonte de informação sobre o Spark é a documentação oficial (em inglês):

https://spark.apache.org/documentation.html