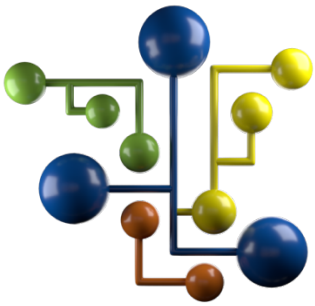

Data Science Academy

www.datascienceacademy.com.br

Engenharia de Dados com Hadoop e Spark

Apache Spark MLLib x Apache Mahout

Apache Mahout é o framework de Machine Learning do Hadoop e MLLib é o framework de Machine Learning do Spark. Eles são concorrentes ou complementares? Qual é o melhor e quando cada um deve ser utilizado? É o que vamos responder agora.

A Fundação Apache tem introduzido muitos frameworks de Machine Learning e um dos mais utilizados em ambientes de larga escala é o Apache Mahout. Já existem muitas empresas usando o Mahout para criar sistemas de recomendação ou construir modelos preditivos sobre grandes conjuntos de dados. A Amazon talvez seja um dos exemplos mais emblemáticos. A empresa utiliza o Mahout em seus sistemas de recomendação e segundo a empresa obteve um crescimento de 35% nas vendas desde a implementação do framework.

O Mahout demonstrou ter algumas funcionalidades que o Spark MLLib ainda se quer implementou. Mas o Mahout tem um pequeno problema. Ele é executado sobre o Hadoop MapReduce, o que restringe e muito sua performance. Algoritmos de Machine Learning geralmente utilizam muitas iterações, o que pode tornar o Mahout lento. Em contraste, o MLLib foi construído sobre o Spark, que é muito mais veloz que o Hadoop MapReduce.

A principal diferença entre Mahout e MLLib recai sobre os frameworks em que eles são executados, Hadoop MapReduce ou Apache Spark. A questão é que o Spark é muito mais veloz e por conta disso, o MLLib processa os algoritmos de Machine Learning de forma muito mais rápida. Entretanto, o MLLib não implementa alguns dos algoritmos implementados no Mahout. As atualizações do Mahout também são menos frequentes que as atualizações do MLLib. A equipe do Mahout está desenvolvendo uma nova biblioteca para álgebra linear chamada Samsara, que promete mudar completamente o funcionamento do Mahout. Os dois frameworks suportam processamento paralelo e distribuído, suportam linguagem Java e Python, permitem processar grandes conjuntos de dados e são open-source.

Algumas pesquisas recentes indicaram que em um único job MapReduce, o Spark é significantemente mais veloz que o Mahout.

Ok, então qual framework de Machine Learning devo utilizar?

Se você estiver começando em Machine Learning para computação em larga escala, o Spark MLLib seria uma escolha mais segura. O Apache Mahout deve ser usado apenas em casos bem específicos com volumes de dados na casa de Petabytes e para utilizar algoritmos que não estejam implementados no MLLib. Já existem esforços para migrar o Apache Mahout para ser executado sobre o Spark, o que vai torná-lo muito parecido ao MLLib. Portanto, se tiver que criar modelos preditivos para grandes conjuntos de dados, sua melhor escolha seria o Spark MLLib.

Mas saiba que já existe um novo framework que promete superar a velocidade do MLLib e executar algoritmos de Machine Learning bem mais poderosos e voltados para Inteligência Artificial. É o H2O e esse é o site, h2o.ai. Mas isso é assunto para outro curso.