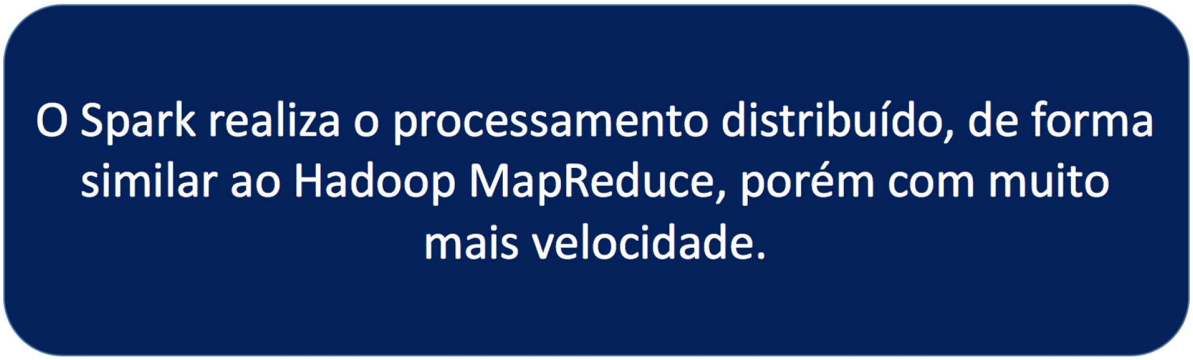
Machine Learning e IA em Ambientes Distribuídos Versão 2.0

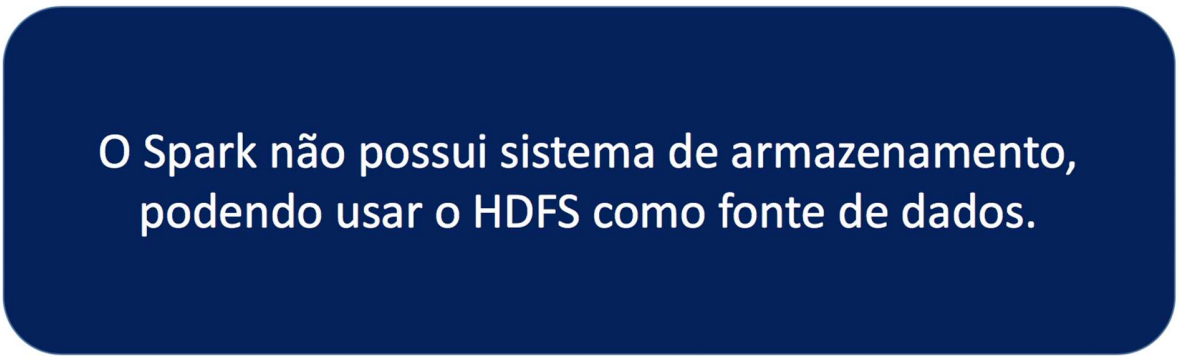
Hadoop MapReduce x Apache Spark

Hadoop MapReduce e Apache Spark são os dois frameworks mais populares para computação em cluster e análise de dados de larga escala (Big Data). Estes dois frameworks escondem a complexidade existente no tratamento de dados com relação a paralelismo entre tarefas e tolerância a falha por meio da exposição de uma simples API com informações para os usuários. Vamos ver aqui as principais diferenças entre as duas tecnologias.

O Hadoop já existe ha mais de 10 anos e tem provado ser a melhor solução para o processamento de grandes conjuntos de dados. O Hadoop possui 2 componentes principais, o HDFS e o MapReduce, sendo o HDFS para armazenamento distribuído, através de clusters e o MapReduce para processamento distribuído, também através de clusters. O MapReduce é uma ótima solução para cálculos de único processamento, mas não muito eficiente para os casos de uso que requerem cálculos e algoritmos com várias execuções, processamento típico de Machine Learning. Isso porque cada etapa no fluxo de processamento tem apenas uma fase Map e uma fase Reduce e os resultados intermediários são armazenados em disco.

O MapReduce é um motor de computação distribuída fornecido pelo Hadoop. Enquanto HDFS fornece um sistema de arquivos distribuído para o armazenamento de grandes conjuntos de dados, MapReduce fornece uma estrutura de computação para o processamento de grandes conjuntos de dados em paralelo em um cluster de computadores. Ele abstrai computação em cluster e fornece construções de alto nível para escrever aplicações de processamento de dados distribuídos.





É importante ter atenção quando se compara Hadoop e Spark. O Spark realiza processamento distribuído, similar ao Hadoop MapReduce. O Spark usa a infraestrutura do Hadoop Distributed File System (HDFS), mas melhora suas funcionalidades e fornece ferramentas adicionais. Devemos olhar para o Spark como uma alternativa para o MapReduce do Hadoop em vez de um simples substituto, como uma solução abrangente e unificada para gerenciar diferentes casos de uso de Big Data. Soluções como o Microsoft Azure suportam Hadoop e Spark juntos para construir uma robusta solução de Big Data, por exemplo.