Por que Spark?

Por que Machine Learning com Spark?

- Spark ML
  - Processamento distribuído
  - Eficiente para conjuntos de dados gigantes, 100x mais rápido
  - Modelos de ML mais simples

- Bibliotecas de ML do Python (ex. Scikit-Learn)
  - Processamento em memória, não distribuído (único nó)
  - Excelente performance enquanto “cabe” na memória

Porque Pyspark?

ML

- Spark Possui 2 principais bibliotecas de Machine Learning:
  - Mllib: baseado no formato RDD: descontinuada (apenas manutenção)
  - ML (também Spark ML, MLlib DataFrame-based API) baseado totalmente no formato SQL DataFrame

O Que é Spark

- Ferramenta de Processamento de Dados (Não é Data Storage)
- Distribuído em um Cluster
- Em memória
- Veloz
- Escalável
- Dados em HDFS ou Cloud
- Particionamento

Spark

Universidade da Califórnia iniciou projeto Spark em 2009

Versão 1.0 lançada em Maio de 2014 pela Fundação Apache