



Analytics - Visualização, Relatórios e Tomada de Decisões com Big Data



Big Data Analytics com Azure HDInsight – Parte 2

Analytics, Visualização, Relatórios e Tomada de Decisões com Big Data





Big Data Analytics com Azure HDInsight

Parte 2

- HortonWorks Sandbox (localmente e no Azure)
- Apache Spark
- Conexões de ferramentas de BI com o Cluster
- Conexões de linguagens de programação com o Cluster
- H2O, Sparkling Water, H2OFlow
- Cluster R



Apache Spark

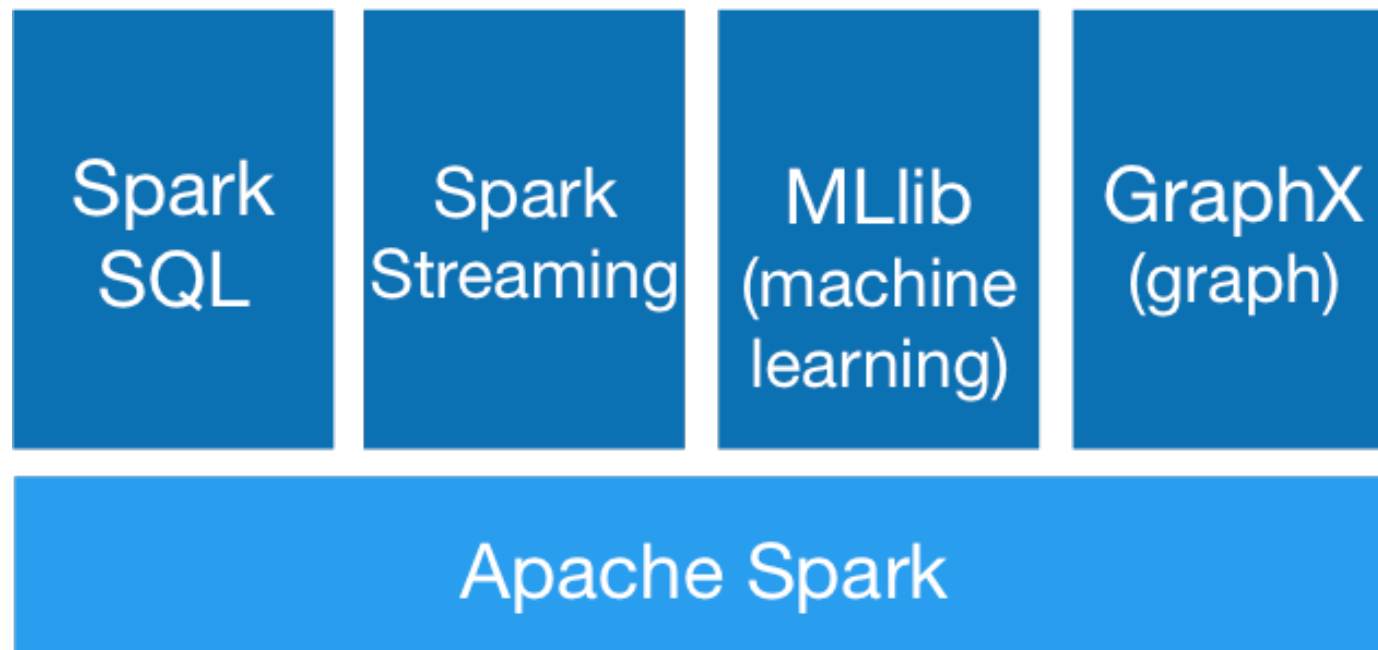
- Framework para processamento de Big Data com foco em velocidade e facilidade de uso
- Desenvolvido na Universidade de Berkeley, Califórnia em 2009. Em 2010, seu código foi aberto como projeto da Fundação Apache
- Permite que aplicações em clusters Hadoop executem até 100 vezes mais rápido em memória e até 10 vezes mais rápido em disco
- Suporta Java, Scala e Python, Clojure e R





Apache Spark

- Suporta operações Map/Reduce, consultas SQL, Streaming de Dados, Machine Learning e Processamento de Grafos





H2O.ai

- Empresa do Vale do Silício, fundada em 2011, focada no desenvolvimento de produtos para Big Data e Machine Learning
- O H2O usa compressão de dados em memória permitindo a manipulação de datasets com bilhões de linhas em clusters relativamente pequenos
- Plataforma H2O inclui interfaces para R, Python, Scala, Java, CoffeeScript/Javascript
- Plataforma web H2O-Flow possibilita a criação de workflows analíticos por profissionais não-técnicos
- H2O é projetado para executar dentro de um Cluster Hadoop/Spark (mesmo em modo standalone)

The H2O.ai logo, consisting of the text "H2O.ai" in a bold, sans-serif font, with the "2" as a subscript, set against a solid yellow square background.

H₂O.ai



H2O.ai

- H2O inclui diversos algoritmos de Machine Learning, supervisionados e não-supervisionados, desde os mais simples como Regressão Logística, Naive Bayes, K-Means até Random Forests, Gradient Boosting e Deep Learning
- A plataforma facilita a criação e comparação entre milhares de modelos

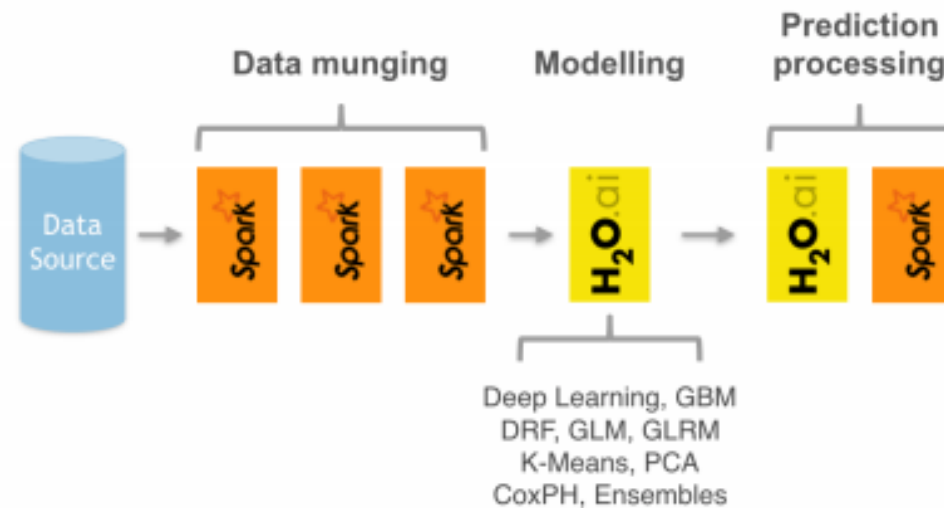
The H2O.ai logo, consisting of the text "H2O.ai" in a bold, black, sans-serif font, set against a solid yellow rectangular background.

H₂O.ai



Sparkling Water (H2O + Spark)

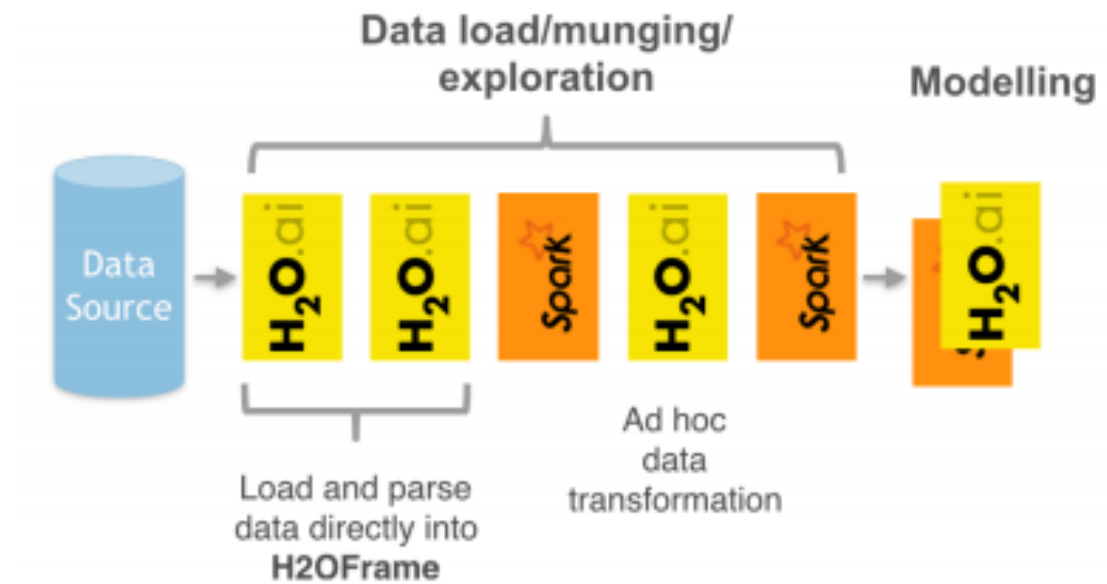
- Permite a combinação entre algoritmos de ML escaláveis do H2O com as capacidades do Spark
- Ideal para usuários que queiram usar o H2O em grandes clusters de computadores
- Exemplo: Você pode usar o Spark SQL para alimentar modelos de ML no H2O e usar os resultados obtidos do modelo novamente no Spark





Sparkling Water

- Transformação de Dados





Sparkling Water

Fontes de Dados

- Spark RDDs: API permite carregar dados em H2OFrames
- Sistemas de Arquivos Locais
- HDFS
- S3
- HTTP/HTTPS

Formato de Dados

- CSV
- SVMLight
- ARFF
- Parquet



FIM

Analytics, Visualização, Relatórios e Tomada de Decisões com Big Data

