




MACHINE LEARNING COM H2O E SPARK

Big Data



H2O

- Plataforma Open Source de análise preditiva e de Machine Learning
- Distribuído e escalável
- Fornece recursos para construir modelos de ML em Big Data com recursos para publicação em ambientes corporativos.




H2O

- Desenvolvido principalmente em Java.
- Sua principal estrutura é um Key-Value Store Distribuído (DKV), usado para acessar e referenciar dados, modelos, objetos, etc., em todos os nós / máquinas
- Algoritmos são implementados no paradigma Map Reduce utilizando o framework Java Fork / Join.

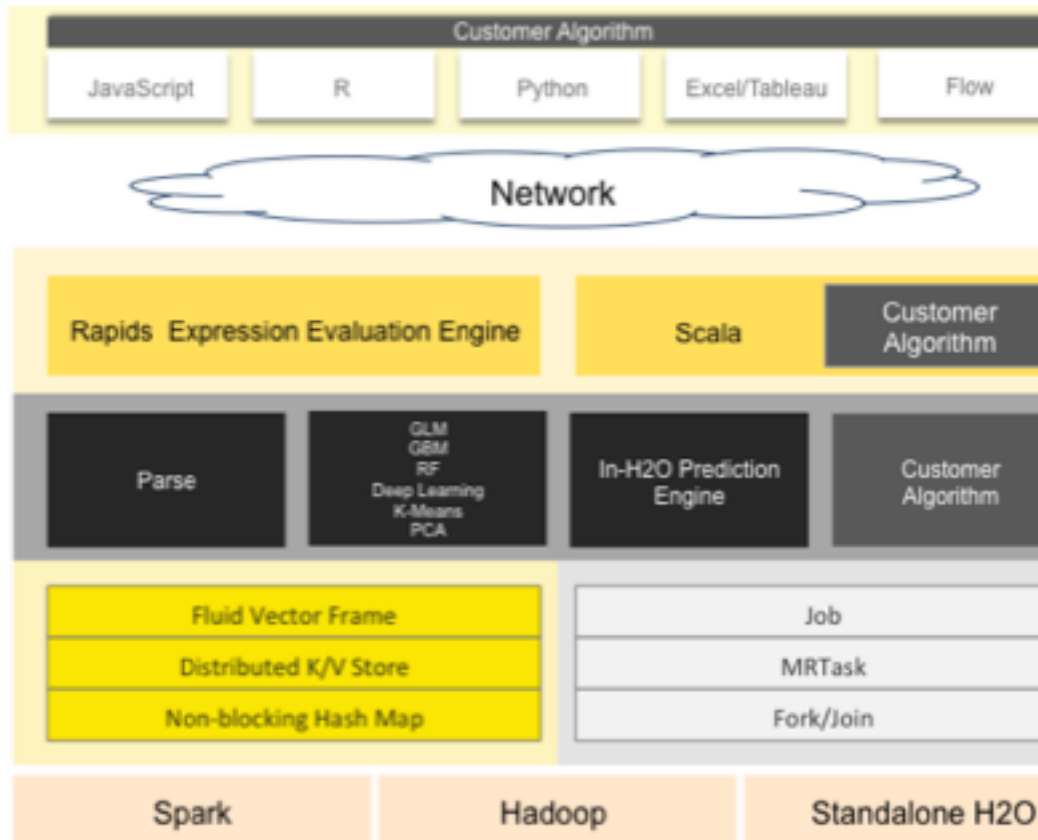
H2O

- Os dados são lidos em paralelo e são distribuídos pelo cluster, armazenados na memória em um formato colunar de forma compacta.
- Possui uma API REST para acesso a todos os recursos a partir de um programa ou script externo, via JSON sobre HTTP.
 - A API é usada pela interface Web do H2O (Flow UI), pelo R (H2O-R) e pelo Python (H2O-Python).



H2O

- Disponibiliza algoritmos para Deep Learning, Gradient Boosted Trees, Random Forest, Decision Trees, GLM, K-means, PCA, etc



H2O



SPARKLING WATER

- Integração do H2O no ecossistema Spark, facilita o uso do H2O em workflows do Spark.
- Projetado como um aplicativo Spark comum
 - fornece uma maneira de iniciar os serviços do H2O em cada nó do cluster Spark e acessar dados armazenados nas estruturas de dados do Spark e do H2O.

SPARKLING WATER

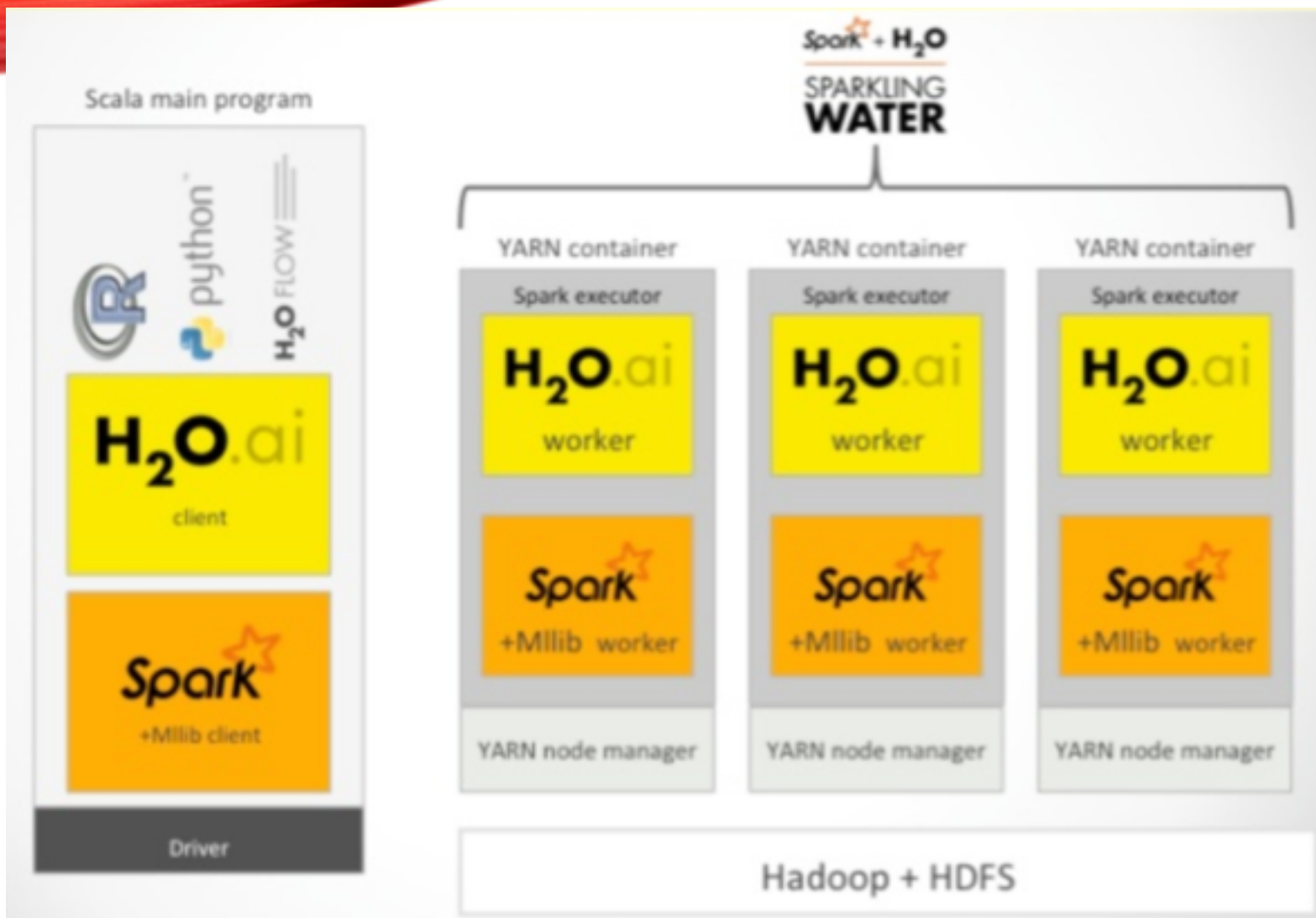
- O Diver cria um SparkContext (sc) que, por sua vez, é usado para criar um H2OContext (hc) que é usado para iniciar os serviços de H2O nos Executors.
- H2OContext é uma conexão ao cluster de H2O e também facilita a comunicação entre o H2O e o Spark.
- Quando um H2OCluster é iniciado, ele possui a mesma topologia que o cluster do Spark e os nós de H2O compartilham as mesmas JVMs que os Executors do Spark.

SPARKLING WATER

- Os dados no cluster do Spark, armazenados como um RDD, precisam ser convertidos em um H2OFrame (Dataframes distribuídos do H2O).
 - Isso requer uma cópia de dados devido à diferença no layout de dados no Spark e no H2O.
 - Sobrecarga é baixa devido à compressão de dados do H2O.

SPARKLING WATER

- Ao converter um H2OFrame em RDD, o Sparkling Water cria um wrapper em torno do H2OFrame para fornecer uma API semelhante à RDD.
 - Nesse caso, nenhum dado é duplicado e os dados são exibidos diretamente do H2OFrame subjacente.
 - Como o H2O é executado nas mesmas JVMs que os Executors do Spark, a movimentação de dados do Spark para o H2O ou vice-versa é feita em memória, no mesmo processo Java.



SPARKLING WATER

OPERAÇÕES COM H2OFRAME

- <http://docs.h2o.ai/h2o/latest-stable/h2o-pv/docs/frame.html>
- <http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-munging.html>
- <http://docs.h2o.ai/h2o/latest-stable/h2o-pv/docs/modeling.html>



H2O GLM

- <https://www.slideshare.net/0xdata/distributed-alm-tk20150127atl>

NO NOSSO CLUSTER

```
cd sparkling-water-2.3.8
```

```
PYSPARK_DRIVER_PYTHON="ipython" bin/pysparkling --conf  
spark.scheduler.minRegisteredResourcesRatio=1 --conf  
spark.dynamicAllocation.enabled=false --conf  
spark.executor.instances=5
```