

# H2O e Sparkling Water - Tópicos em Estatística 1

2022.1 - UnB - Professor Guilherme Rodrigues

Bruno Gondim, Eraldo Jair, José Felipe, Paulo Henrique e Pedro Aguiar

12/09/2022

# Sumário I

- 1 Introdução
- 2 *H2O - Flow*
- 3 *H2O - R*
- 4 *H2O - Google Cloud*
- 5 *Sparkling Water*
- 6 Fontes

# Section 1

## Introdução

# *H2O* e *H2O.ai*: o que é?

The logo for H2O.ai is displayed within a large yellow rounded square. The text "H2O.ai" is centered in a bold, black, sans-serif font. The "H2O" part is significantly larger and bolder than the ".ai" domain extension.

**H2O.ai**

O *H2O* é uma plataforma de aprendizado de máquina distribuído na memória totalmente de código aberto com escalabilidade linear. É compatível com os algoritmos estatísticos e de aprendizado de máquina mais usados, incluindo máquinas impulsionadas por gradiente, modelos lineares generalizados, aprendizado profundo e muito mais. Também possui uma funcionalidade *AutoML* líder do setor que executa automaticamente todos os algoritmos e seus hiperparâmetros para produzir uma tabela de classificação dos melhores modelos. A plataforma é usada por mais de 18.000 organizações em todo o mundo e é extremamente popular nas comunidades *R* e *Python*, além de ser utilizada por aproximadamente metade das empresas listadas na *Fortune* 500.

A *H2O.ai* é uma empresa de *software* de *machine learning* e inteligência artificial, sediada no Vale do Silício e reconhecida como visionária pelo *Gartner*. Formam um time de “*Makers*” que trouxe para o mercado novas plataformas e tecnologias que impulsionaram o movimento de IA.




É possível executar o *download* do *software* tanto localmente quanto em nuvem.

Ambientes em nuvem suportados:

- Instâncias *Amazon EC2* e *Storage S3* (*RedHat AMI*, *Amazon Linux AMI*, e *Ubuntu AMI*)
- *Amazon AWS*
- *Microsoft Azure*
- *Databricks*
- *IBM Power 9*
- *Nimbix Cloud*
- *Google Cloud*

Em [h2o.ai/download](https://h2o.ai/download), o *download* está disponível.

[H2O.ai](#) [Platform](#) [Solutions](#) [Customers](#) [Partners](#) [Open Source](#) [Community](#) [Company](#)  [Try For Free](#) [Request Demo](#) [Sign In](#) >

[H2O AI Cloud](#) [H2O Wave](#) [Driverless AI](#) [H2O Open Source Platform](#) [Sparkling Water](#) [Enterprise Steam](#) [Enterprise Puddle](#)

## H2O

H2O works with R, Python, Scala on Hadoop/Yarn, Spark or your laptop.

H2O is licensed under the [Apache License, Version 2.0](#)


### Direct Downloads


[Latest Stable Release](#)


[Nightly Bleeding Edge](#)

[Prior Releases](#)

### Cloud Downloads

 Microsoft Azure

 aws

 Google Cloud



## Section 2

### *H2O - Flow*

## H2O - Flow

*H2O Flow* é uma interface de usuário de código aberto para *H2O*. É um ambiente interativo baseado na web que permite combinar execução de código, texto, matemática, gráficos e mídia avançada em um único documento.

*H2O Flow* envia comandos para *H2O* como uma sequência de células executáveis. As células podem ser modificadas, reorganizadas ou salvas em uma biblioteca. Cada célula contém um campo de entrada que permite inserir comandos, definir funções, chamar outras funções e acessar outras células ou objetos na página. Quando você executa a célula, a saída é um objeto gráfico, que pode ser inspecionado para visualizar detalhes adicionais.

Embora o *H2O Flow* suporte *API REST*, *scripts R* e *CoffeeScript*, nenhuma experiência de programação é necessária para executar o *H2O Flow*. Você pode clicar em qualquer operação *H2O* sem nunca escrever uma única linha de código. Você pode até mesmo desabilitar as células de entrada para executar o *H2O Flow* usando apenas a *GUI*. O *H2O Flow* foi projetado para guiá-lo em todas as etapas, fornecendo prompts de entrada, ajuda interativa e fluxos de exemplo.



Version 3.36.1.4

Fast Scalable Machine Learning API  
For Smarter Applications

DOWNLOAD AND RUN

INSTALL IN R

INSTALL IN PYTHON

INSTALL ON HADOOP

USE FROM MAVEN

KUBERNETES

DOWNLOAD H<sub>2</sub>O

### Get started with H<sub>2</sub>O in 3 easy steps

1. Download H<sub>2</sub>O. This is a zip file that contains everything you need to get started.
2. From your terminal, run:

```
cd ~/Downloads
unzip h2o-3.36.1.4.zip
cd h2o-3.36.1.4
java -jar h2o.jar
```



3. Point your browser to <http://localhost:54321>

**Observação:**

ao rodar o código no terminal, o comando “*unzip*” não foi reconhecido pelo *Windows PowerShell*.

Utilizando o comando “*Expand-Archive*” no lugar, funcionou.

# Ambiente de trabalho - H2O Flow

The screenshot displays the H2O Flow web interface. At the top, there is a navigation bar with the H2O logo and a hamburger menu, followed by tabs for Flow, Cell, Data, Model, Score, Admin, and Help. Below this is a toolbar with various icons for file operations and execution. The main workspace is titled 'Untitled Flow' and contains a canvas with a single routine named 'assist'. On the left side, there is an 'Assistance' sidebar listing various routines with their descriptions. On the right side, there is a 'HELP' sidebar with sections for 'Using Flow for the first time?', 'Quickstart Videos', 'Or, view example Flows to explore and learn H2O.', 'STAR H2O ON GITHUB!', 'GENERAL' (with links to Flow Web UI, Importing Data, Building Models, Making Predictions, Using Flows, and Troubleshooting Flow), and 'EXAMPLES' (explaining Flow packs).

**H2O FLOW** Flow Cell Data Model Score Admin Help

Untitled Flow

assist

**? Assistance**

Routine	Description
importFiles	Import file(s) into H <sub>2</sub> O
importSqlTable	Import SQL table into H <sub>2</sub> O
getFrames	Get a list of frames in H <sub>2</sub> O
splitFrame	Split a frame into two or more frames
mergeFrames	Merge two frames into one
getModels	Get a list of models in H <sub>2</sub> O
getGrids	Get a list of grid search results in H <sub>2</sub> O
getPredictions	Get a list of predictions in H <sub>2</sub> O
getJobs	Get a list of jobs running in H <sub>2</sub> O
runAutoML	Automatically train and tune many models
buildModel	Build a model
importModel	Import a saved model
predict	Make a prediction

**HELP**

Using Flow for the first time?

Quickstart Videos

Or, view example Flows to explore and learn H<sub>2</sub>O.

STAR H2O ON GITHUB!

GENERAL

- Flow Web UI ...
- ... Importing Data
- ... Building Models
- ... Making Predictions
- ... Using Flows
- ... Troubleshooting Flow

EXAMPLES

Flow packs are a great way to explore and learn H<sub>2</sub>O. Try out these Flows and run them in your browser.

Connections: 0 H<sub>2</sub>O

Ready

Na próxima vez que você quiser iniciar o *Flow*, rode o seguinte código no terminal:

```
cd ~/Downloads/h2o-3.36.1.4  
java -jar h2o.jar
```

E acesse [localhost](http://localhost) novamente.

Em [tutorial.flow](#), está presente todo o passo a passo de como usar o *H2O Flow*. Alguns pontos importantes da interface são:

- Fluxos de exemplo: *HELP* -> *view example Flows*;
- Começar um novo fluxo: “*nem flow*”;
- Modos de célula: “editar” ou “comandar”
  - Células de comando:
    - MD (*markdown*)
    - CS (código padrão)
    - RAW (comentários de código)
    - H[1-6] (níveis de cabeçalho)
  - *Clips*: salvar células



## Section 3

*H2O - R*

## H2O - R

Código para instalar o pacote “H2O” no R:

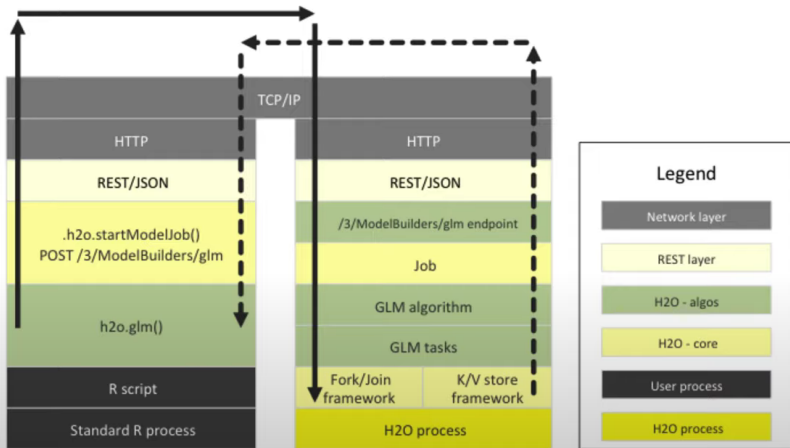
```
# The following two commands remove any previously
# installed H2O packages for R.
if ("package:h2o" %in% search())
{ detach("package:h2o", unload=TRUE) }
if ("h2o" %in% rownames(installed.packages()))
{ remove.packages("h2o") }

# Next, we download packages that H2O depends on.
pkgs <- c("RCurl","jsonlite")
for (pkg in pkgs) {
  if (! (pkg %in% rownames(installed.packages())))
  { install.packages(pkg) }}
```

```
# Now we download, install and initialize  
# the H2O package for R.  
install.packages("h2o", type="source",  
repos=  
"https://h2o-release.s3.amazonaws.com/h2o/rel-zumbo/4/R")  
  
# Finally, let's load H2O and start up an H2O cluster  
library(h2o)  
h2o.init()
```

# Workflor - Arquitetura do H2O.AI no R

## H2O.AI - ARQUITETURA



Resumidamente, Comando  $R \rightarrow H2O \rightarrow$  Ambiente  $R$

Exemplo de comando:

**`h2o_df=h2o.importFile("./path/arquivo.csv")`**

- se transforma internamente num arquivo do tipo *h2oFrame*

## Vantagens:

- Até 100x mais veloz que o *Scikit Learn*.
- Escalabilidade (quanto mais máquinas no *cluster*, melhor.)
- Interface *UI* para monitoramento em tempo real

## Por que usar o *H2O*:

- superar limitações do *R*, tais como:
  - Leitura de arquivos grandes (5GB+)
  - Mau gerenciamento de memória

## Problemas de usar o *H2O* no *R*:

- Modelos em *R* (e ***Python*** também) não lidam tão bem assim com aplicações *Java* e *Scala*

## Section 4

*H2O - Google Cloud*

# H2O - Google Cloud

Recomendação padrão para iniciar o *H2O-3*:

- 4 *CPU's*
- 15 *GB RAM*
- 3 *nodes*



# Instalação

- Faça o *log-in* em [Google Compute Engine Console](#)
- Em [console.cloud.google.com/h2o-3-cluster](https://console.cloud.google.com/h2o-3-cluster), clique em “**abrir**”



## H2O-3 Cluster

[H2O.ai](#)

Automatic Machine Learning for the Enterprise

ABRIR

VER IMPLANTAÇÕES ANTERIORES

- 1 Especifique um nome para esta implantação.
- 2 Selecione uma zona para a implantação.
- 3 Selecione ou personalize um tipo de máquina e quantidade de memória.
- 4 Especifique o número de nós para a máquina virtual.
- 5 Especifique o tipo e o tamanho do disco de inicialização (em GB).
- 6 Especifique os nomes de rede e sub-rede.

Clique em “**Implantar**”. Você será direcionado para a seguinte página:

The screenshot shows the Google Cloud Deployment Manager interface. The left sidebar has a menu with 'Implantações' (Deployments) selected. The main area is divided into two panes. The left pane shows a tree view of the deployment 'h2oai-h2o3-cluster-launcher-1', which is marked as 'EXCLUIR' (Exclude). The tree shows a folder 'h2oai-h2o3-cluster' containing a file 'h2oai-h2o3-cluster.jinja', which in turn contains a folder 'h2o3-cluster' with a file 'vm\_multiple\_instances.py'. This file lists four VM instances: 'h2oai-h2o3-cluster-launcher-1-1', 'h2oai-h2o3-cluster-launcher-1-2', 'h2oai-h2o3-cluster-launcher-1-3', and 'h2oai-h2o3-cluster-launcher-1-firewall'. The right pane shows the details for the 'h2oai-h2o3-cluster' deployment, which is an 'H2O-3 Cluster' solution provided by 'H2O.ai'. It lists configuration details: Instance (us-central1-f), Instance zone (us-central1-f), Instance machine type (e2-standard-4), Connect At (https://34.123.250.7:54321), Username (h2oai), Password (1531168419867869489), and Assigned Network Tags (http-server, https-server, h2o3-server). A note mentions suggested actions for cluster startup. Below the details, there is a section for 'H2O-3 Cluster: noções básicas' (H2O-3 Cluster: basic concepts) and 'Próximas etapas sugeridas' (Suggested next steps), which include 'NOTE CLUSTER STARTUP' and 'NOTE CLUSTER SECURITY'.

Google Cloud My First Project

Pesquisa Produtos, recursos, documentos (/)

Deployment Manager

h2oai-h2o3-cluster-launcher-1 EXCLUIR

h2oai-h2o3-cluster

Implantações

Registro do tipo

h2oai-h2o3-cluster-launcher-1 foi implantado

Overview - h2oai-h2o3-cluster-launcher-1

h2oai-h2o3-cluster h2oai-h2o3-cluster.jinja

h2o3-cluster vm\_multiple\_instances.py

h2oai-h2o3-cluster-launcher-1-1 Instância de VM

h2oai-h2o3-cluster-launcher-1-2 Instância de VM

h2oai-h2o3-cluster-launcher-1-3 Instância de VM

h2oai-h2o3-cluster-launcher-1-firewall firewall

**H2O-3 Cluster**  
Solução fornecida por H2O.ai

Instance

Instance zone us-central1-f

Instance machine type e2-standard-4

Connect At <https://34.123.250.7:54321>

Username h2oai

Password 1531168419867869489

Assigned Network Tags http-server, https-server, h2o3-server

NOTE See suggested actions for notes on cluster startup.

MAIS SOBRE O SOFTWARE

**H2O-3 Cluster: noções básicas**

**Próximas etapas sugeridas**

- NOTE CLUSTER STARTUP**  
Depending on the number of virtual machines requested, you may have to wait a couple minutes before all machines have started up successfully and clustered together. After you are able to see the "Connect At" link, give the cluster a couple minutes to fully start up.
- NOTE CLUSTER SECURITY**  
When connecting to H2O Flow in your browser, you may notice a security alert. We use a self-signed certificate to encrypt traffic to the cluster, which will cause this. See Suggested actions below for next steps.

Acesse o IP em “**Connect At**”. Em seguida, será solicitado o nome do usuário (*h2oai*) e a senha (*1531168419867869489*).

Assim, você estará conectado ao *H2O-3 Cluster*.

## Section 5

### *Sparkling Water*

# Sparkling Water

- O *Sparkling Water* foi um pacote desenvolvido para aprendizado de máquina de forma rápida e escalável.

# $H_2O + Spark = Sparkling\ Water$

- *Sparkling water* vem da junção do *H<sub>2</sub>O* com o *Spark*, onde o termo *Sparkling Water* significa “água com gás”.

# Vantagens do Spark e H2O:

## Vantagens do Spark:

- *ML Pipelines*;
- *ETL* poderoso;
- Algoritmos (NLP);

## Vantagens do H2O:

- Algoritmos avançados;
- Velocidade e acuracia;
- Computação distribuida e paralelizada;
- É possível trabalhar com *R*, *Python* e *Scala*



- *Sparkling Water* contém os mesmos recursos e funcionalidades do *H2O*, mas fornece uma maneira de usar o *H2O* com o *Spark*, uma estrutura de *cluster* de grande escala.
- É ideal para usuários de *H2O* que precisam gerenciar grandes clusters para suas necessidades de processamento de dados e desejam transferir dados do *Spark* para o *H2O* (ou vice-versa).
- Há também uma interface *Python* disponível para permitir o acesso ao *Sparkling Water* diretamente do *PySpark*.

## Section 6

### Fontes

# Fontes

- <https://docs.h2o.ai/h2o-tutorials/latest-stable/index.html>
- <https://spark.rstudio.com/guides/h2o.html>