

DATAMASS
GDAŃSK SUMMIT 2017

Machine learning in R using H2O framework

Gdańsk, 30.09.2017

Who am I?



Mariusz Liksza

Solutions Developer at Acxiom since July, 2016

Lecturer in postgraduates studies data science since March, 2017

Previous work:

- Data Analyst at Logisfera FI
- Junior Data Analyst at BEST
- Intern Data Scientist at Pin Your Client

Education:

- M. S. Financial Mathematics, Gdańsk University of Technology
- B. S. Methods of data analysis, University of Gdańsk

Background: Mathematics, Statistics, Informatics

Tools: R, Python, SAS, MATLAB, Hadoop Ecosystem, H2O

Hobby: volleyball, tennis, chess, big data



**POLITECHNIKA
GDAŃSKA**

WYDZIAŁ FIZYKI TECHNICZNEJ
I MATEMATYKI STOSOWANEJ

Machine Learning with H2O

H2O Overview:

- About H2O.ai
- What is H2O?
- Architecture
- Algorithms

H2O Web UI

H2O on R

H2O on Hadoop Cluster

H2O on Spark - Sparkling Water

H₂O.ai

H2O.ai (formerly oxdata)

Founded in 2012, Mountain View, CA

H2O.ai is the maker behind H2O, the leading open source deep learning platform for smarter applications and data products.

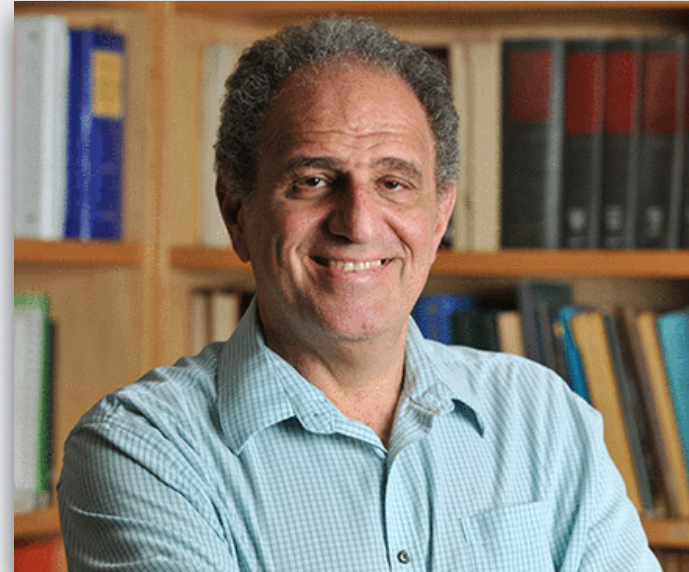


H2O.ai was co-founded by SriSatish Ambati - CEO (on the right) and Cliff Click - CTO (on the left).

Scientific Advisory Council



Professor of Statistics Trevor Hastie, a collaborator of John Chambers on S , is an expert on generalized additive models and statistical learning theory.



Professor in the Departments of Statistics Robert Tibshirani, a collaborator with Bradley Efron on bootstrapping, is an expert on generalized additive models and statistical learning theory

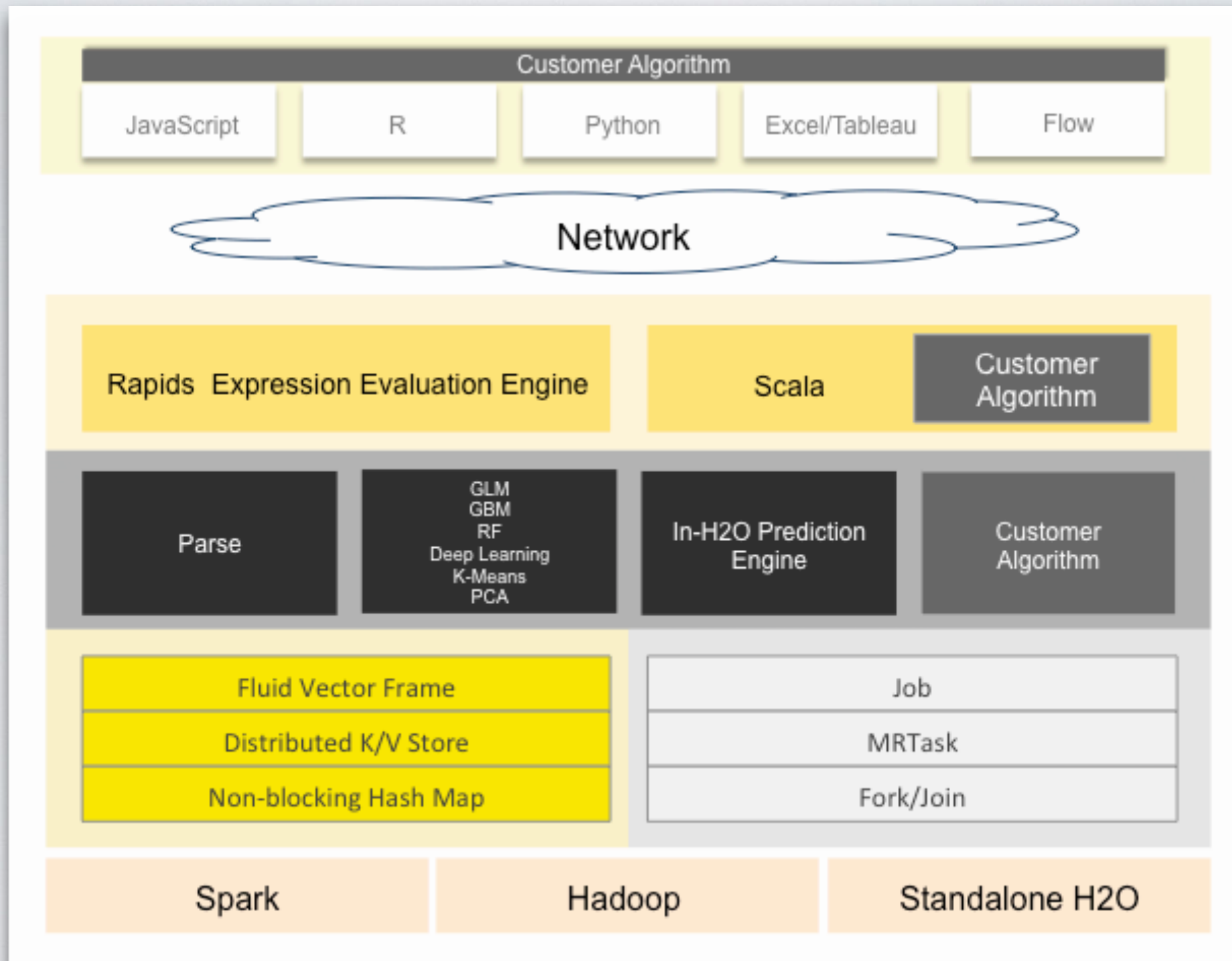


Professor Stephen P. Boyd is an expert in convex minimization and applications in statistics and electrical engineering.

What is H2O?

- ☑ H2O is the world's leading open source deep learning platform
- ☑ H2O is used by over 80,000 data scientists and more than 9,000 organizations around the world
- ☑ H2O has easy to use Web Interface
- ☑ Written in Java
- ☑ REST API (JSON) drives H2O from R, Python, Excel, Tableau, Scala & JSON
- ☑ Distributed Algorithms Scale to Big Data
- ☑ <https://github.com/h2oai>

H2O Architecture



REST API Clients

All REST API clients communicate with H2O over a socket connection.

JavaScript

The embedded H2O Web UI is written in JavaScript, and uses the standard REST API.

R

R scripts can use the H2O R package [`library(h2o)`]. Users can write their own R functions that run on H2O with `'apply'` or `'ddply'`.

Python

Python scripts currently must use the REST API directly. An H2O client API for python is planned.

Excel

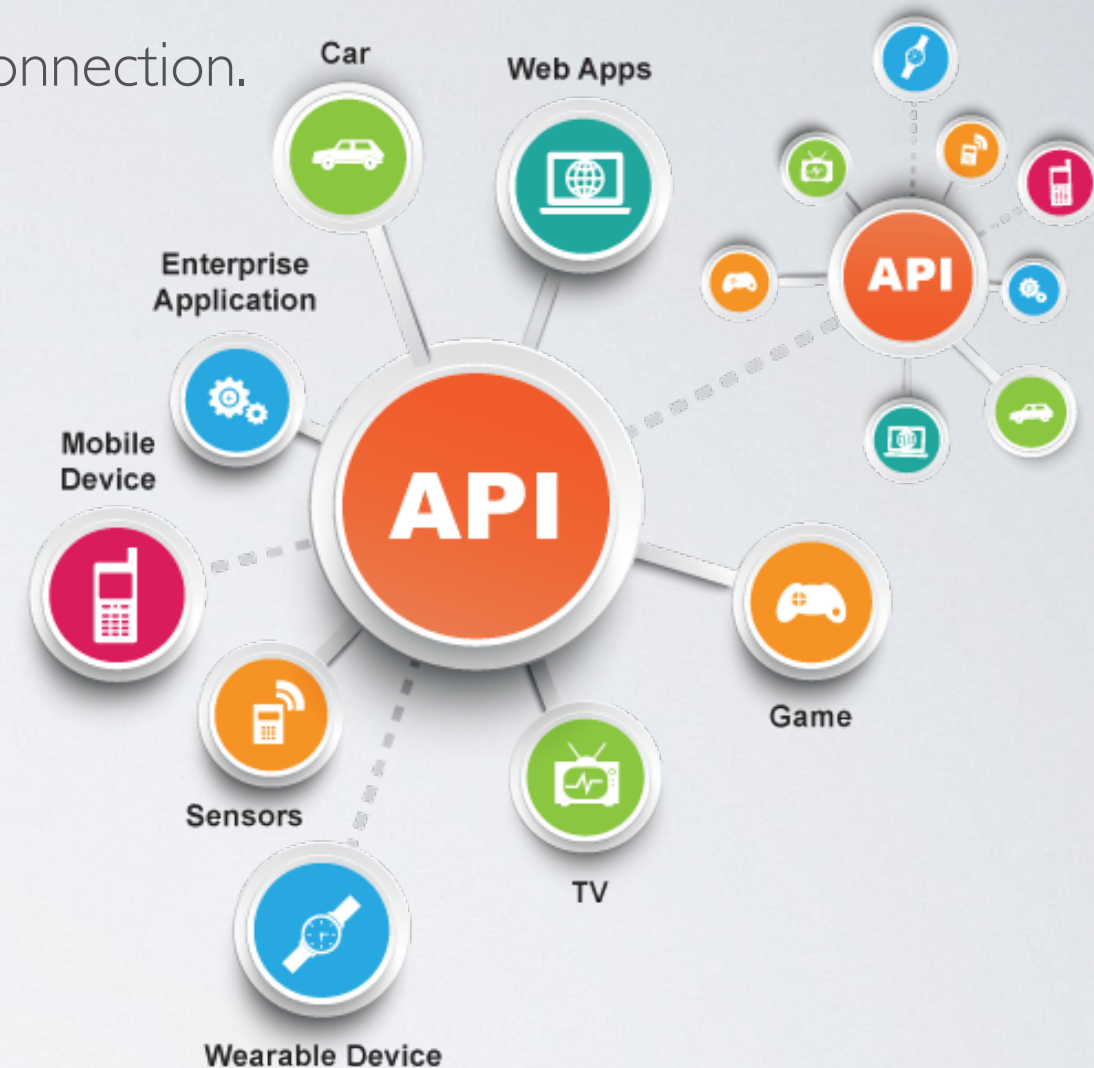
An H2O worksheet for Microsoft Excel is available. It allows you to import big datasets into H2O and run algorithms like GLM directly from Excel.

Tableau

Users can pull results from H2O for visualization in Tableau.

Flow

H2O Flow is the notebook style Web UI for H2O.



Algorithms in H2O

Supervised learning:

Statistical Analysis:

- GLM with Regularization
- Naive Bayes

Ensembles

- Distributed Random Forest
- Gradient Boosting Machine

Neural Networks:

- Deep Learning

Unsupervised learning:

Clustering:

- K-means

Dimension Reduction

- PCA
- Generalized Low Rank Models

Anomaly Detection:

- Autoencoders (anomaly detection)

H2O Web UI - Using Flow

H2O Flow is an open-source user interface for H2O. It is a web-based interactive environment that allows you to combine code execution, text, mathematics, plots, and rich media in a single document.

GLM_Example



```
getModel "glm-97176b71-c4dc-4d3a-bf36-53b5a4e4ab03"
```

180ms

Model

Model ID: glm-97176b71-c4dc-4d3a-bf36-53b5a4e4ab03

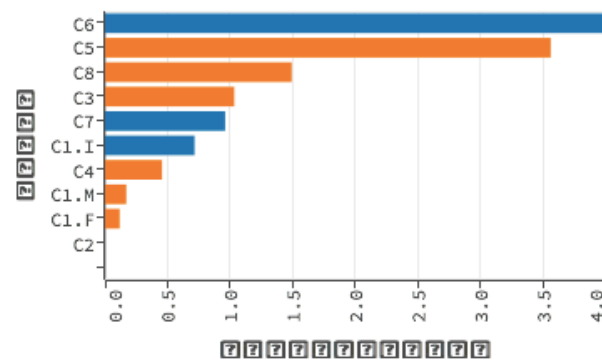
Algorithm: Generalized Linear Modeling

Actions: Refresh Predict... Download POJO Download Model Deployment Package Export Inspect Delete

MODEL PARAMETERS

SCORING HISTORY

STANDARDIZED COEFFICIENT MAGNITUDES



OUTPUT

OUTPUT - GLM MODEL (SUMMARY)

OUTPUT - SCORING HISTORY

OUTPUT - TRAINING METRICS

OUTLINE FLOWS CLIPS **HELP**

Help

PACK

examples


- [GBM_Example.flow](#)
- [DeepLearning_MNIST.flow](#)
- [GLM_Example.flow](#)
- [DRF_Example.flow](#)
- [K-Means_Example.flow](#)
- [Million_Songs.flow](#)
- [KDDCup2009_Churn.flow](#)
- [QuickStartVideos.flow](#)
- [Airlines_Delay.flow](#)
- [GBM_Airlines_Classification.flow](#)
- [GBM_GridSearch.flow](#)
- [RandomData_Benchmark_Small.flow](#)
- [GBM_TuningGuide.flow](#)


With H2O Flow, you can capture, rerun, annotate, present, and share your workflow. H2O Flow allows you to use H2O interactively to import files, build models, and iteratively improve them. Based on your models, you can make predictions and add rich text to create vignettes of your work – all within Flow's browser-based environment.

Start working with H2O - its easy peasy!

1. Go to <https://www.h2o.ai> and click *Download* button
2. Choose *Download H2O Latest Stable Version*
3. *Just follow 3 simple steps located there (see below screen)*

[DOWNLOAD AND RUN](#)[INSTALL IN R](#)[INSTALL IN PYTHON](#)[INSTALL ON HADOOP](#)[USE FROM MAVEN](#)


 **DOWNLOAD H₂O**



Get started with H₂O in 3 easy steps

1. Download H₂O. This is a zip file that contains everything you need to get started.
2. From your terminal, run:

```
cd ~/Downloads
unzip h2o-3.10.4.3.zip
cd h2o-3.10.4.3
java -jar h2o.jar
```


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

DEMO - H₂O Flow

The logo for H2O.ai, featuring the text "H2O.ai" in black and yellow on a yellow rectangular background.

H₂O.ai

H2O on R



Requirements:

R \geq 3.1.0 & Java \geq 7

Tested on many versions of Linux,
OS X and Windows

You can install it from CRAN!

All computations are performed in the H2O cluster
and initiated by REST calls from R so you don't have
to worry anymore about R memory :)

Step 1: The R user calls the importFile function



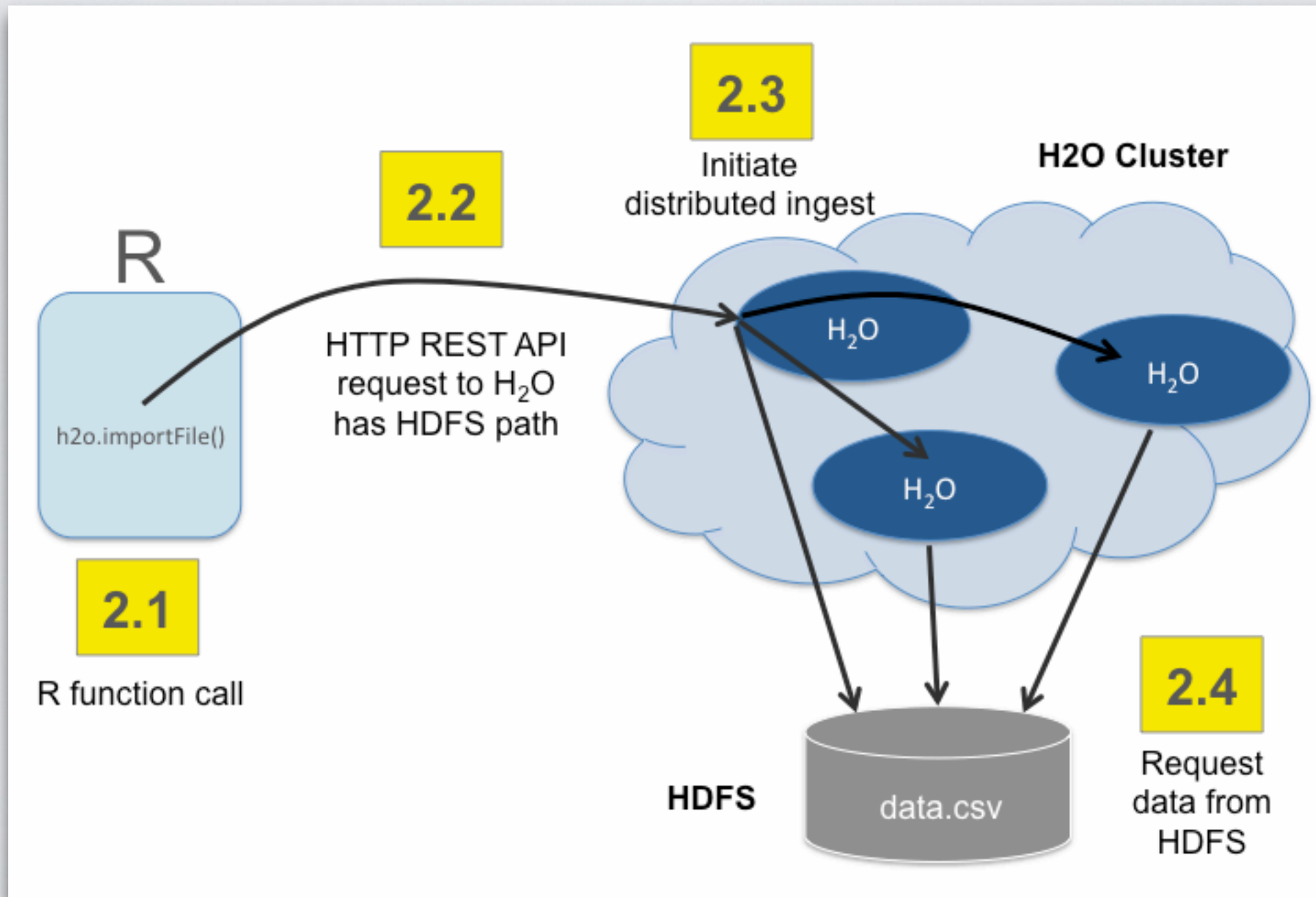
R user



```
h2o_df = h2o.importFile("hdfs://path/to/data.csv")
```

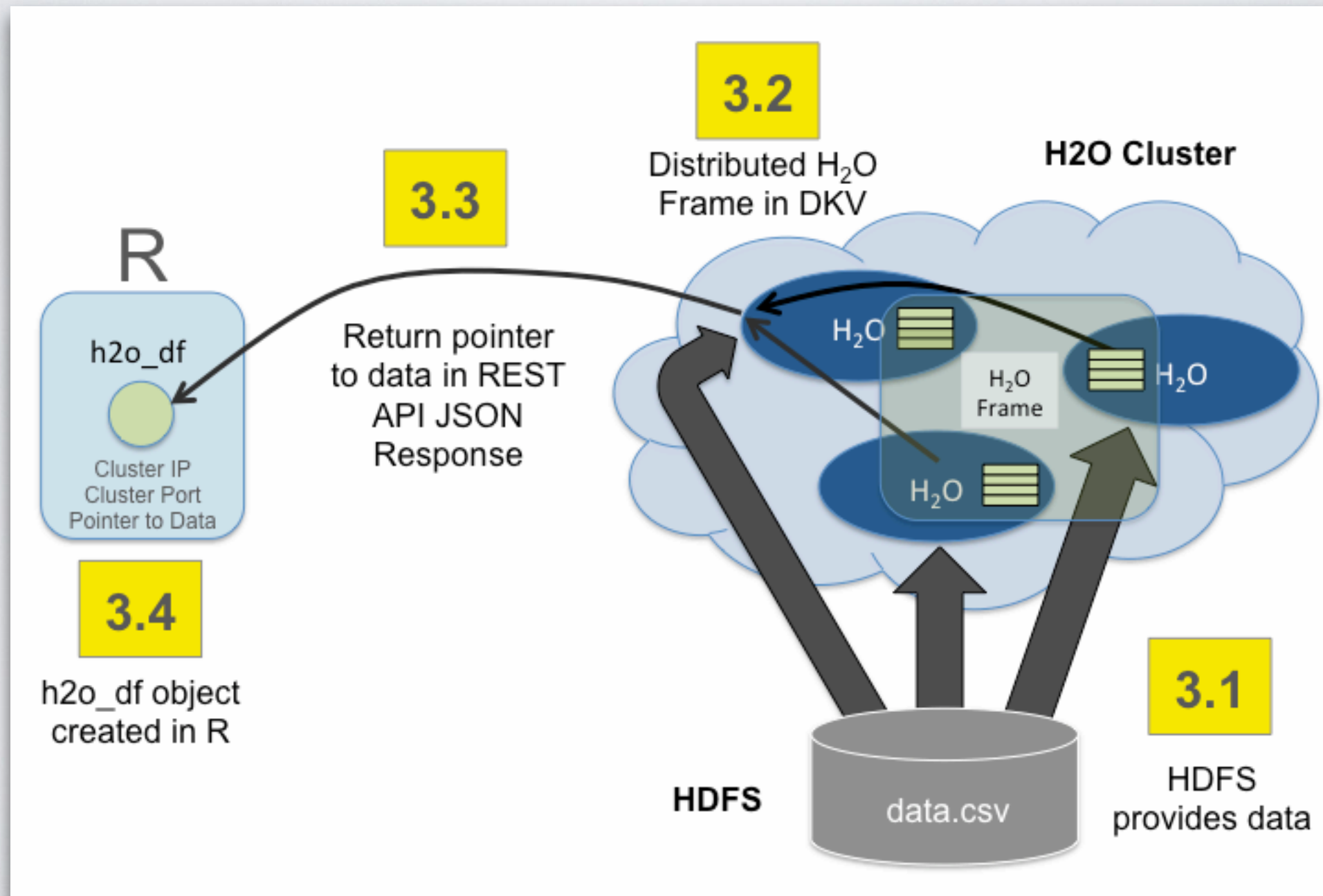
Step 2: The R client tells the cluster to read the data

The thin arrows show control information.



Step 3: The data is returned from HDFS into a distributed H2O Frame

The thin arrows show control information. The thick arrows show data being returned from HDFS. The blocks of data live in the distributed H2O Frame cluster memory.



DEMO - H₂O in R

H2O on Hadoop

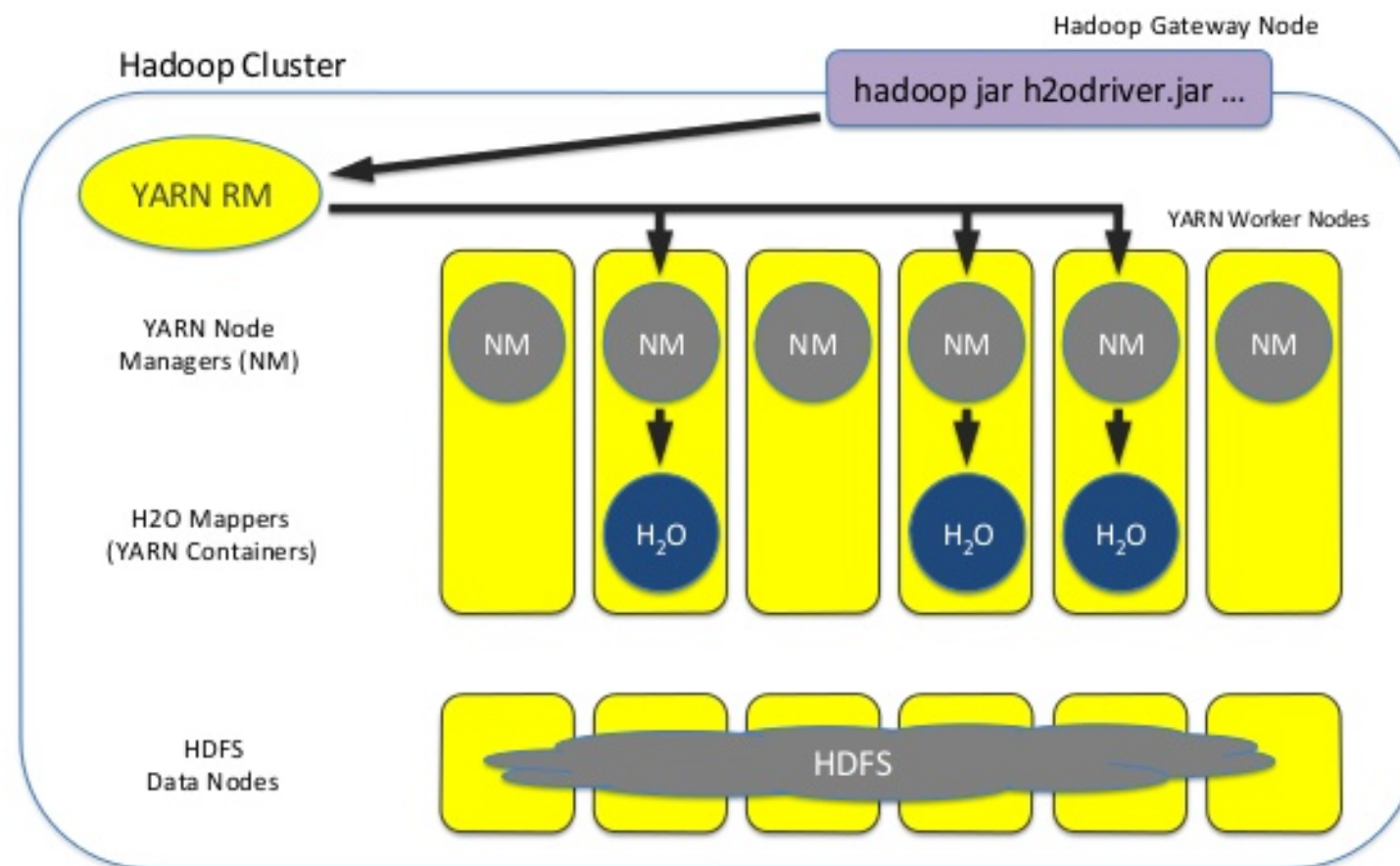
You can launch H2O directly on Hadoop:

```
hadoop jar h2odriver.jar -nodes 6 -mapperXmx 6g -output /outputDir
```

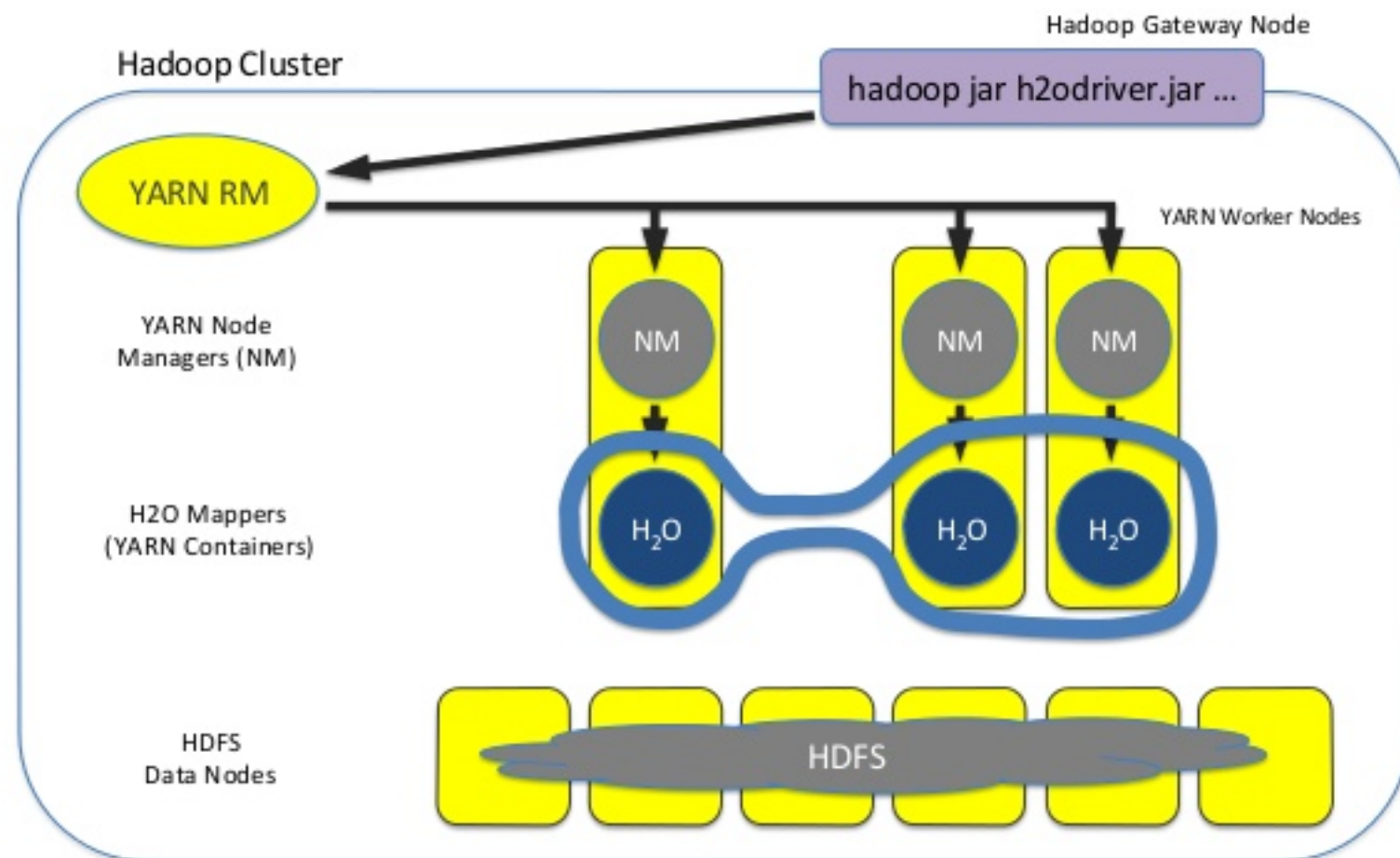
Check all options here (section Hadoop Launch Parameters):

<http://docs.h2o.ai/h2o/latest-stable/h2o-docs/hadoop.html>

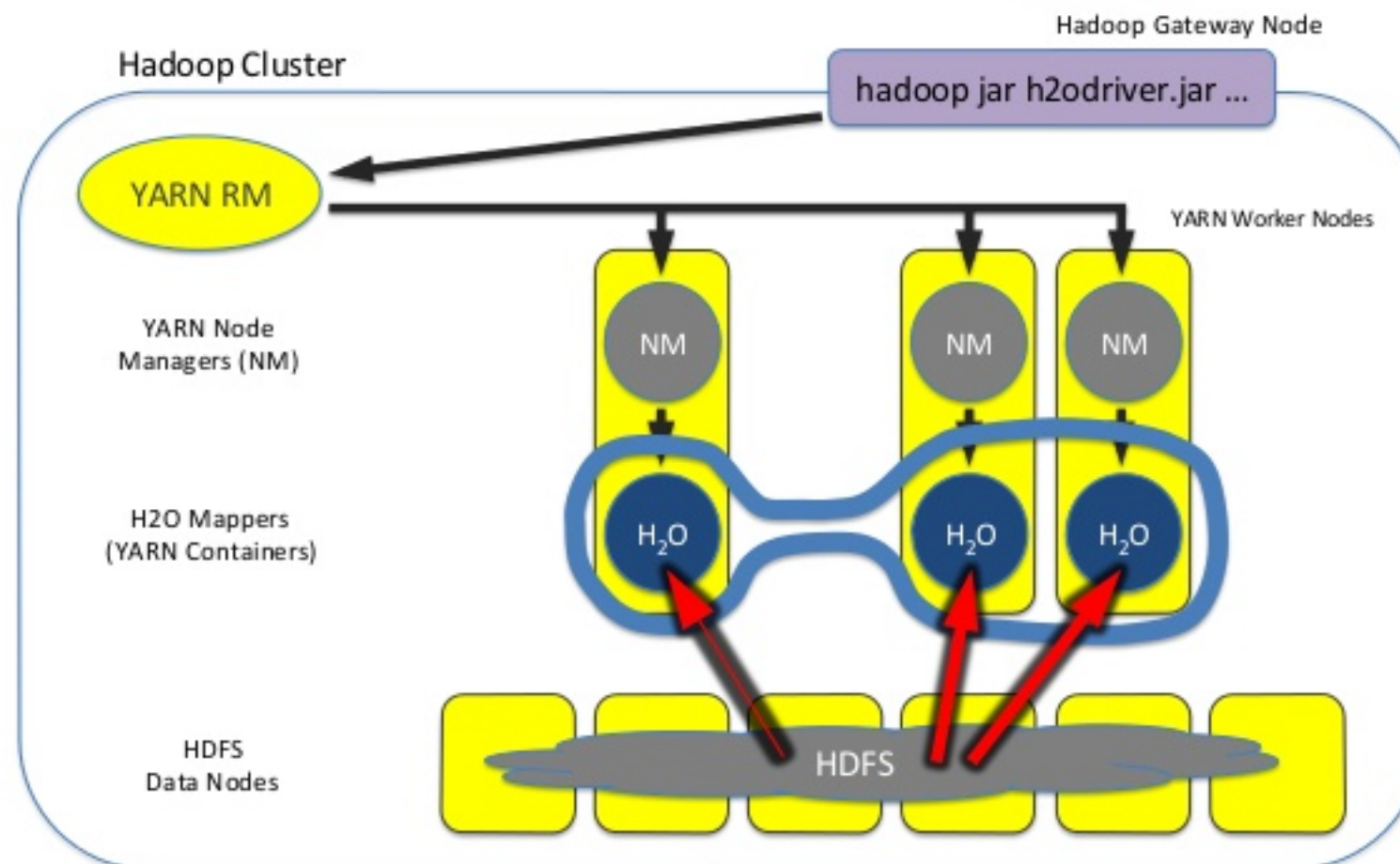
H2O on YARN Deployment



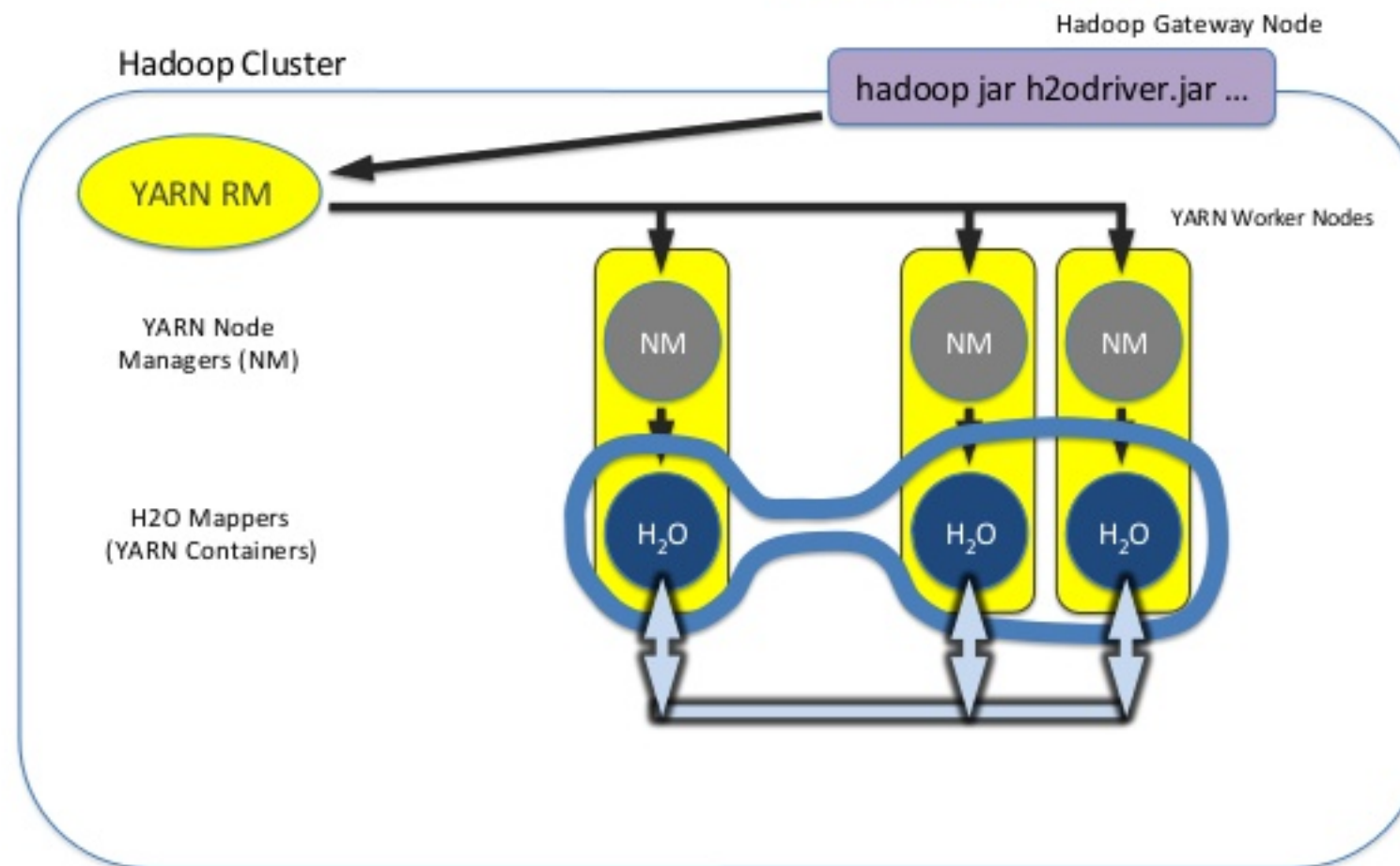
Now You Have an H2O Cluster



Read Data from HDFS *Once*



Build Models *in-Memory*



DEMO - run H2O on Hadoop Cluster

H₂O on Spark - Sparkling Water

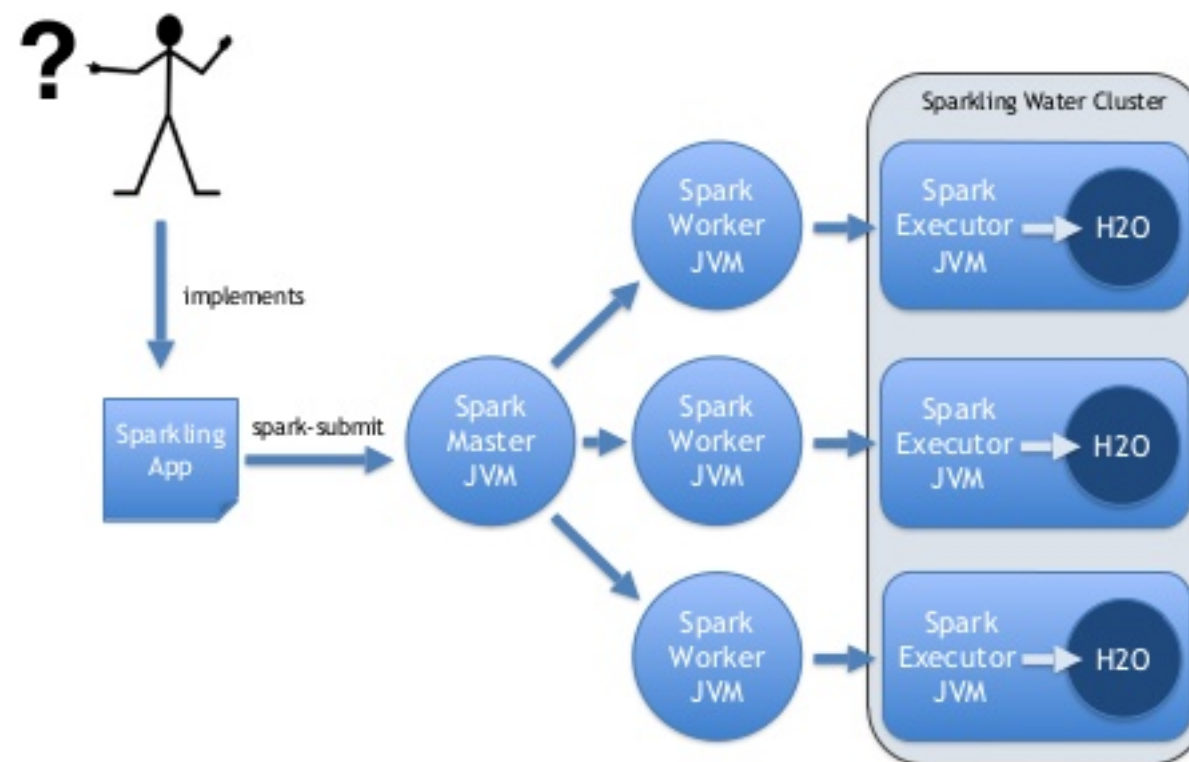
Spark  + H₂O

SPARKLING
WATER

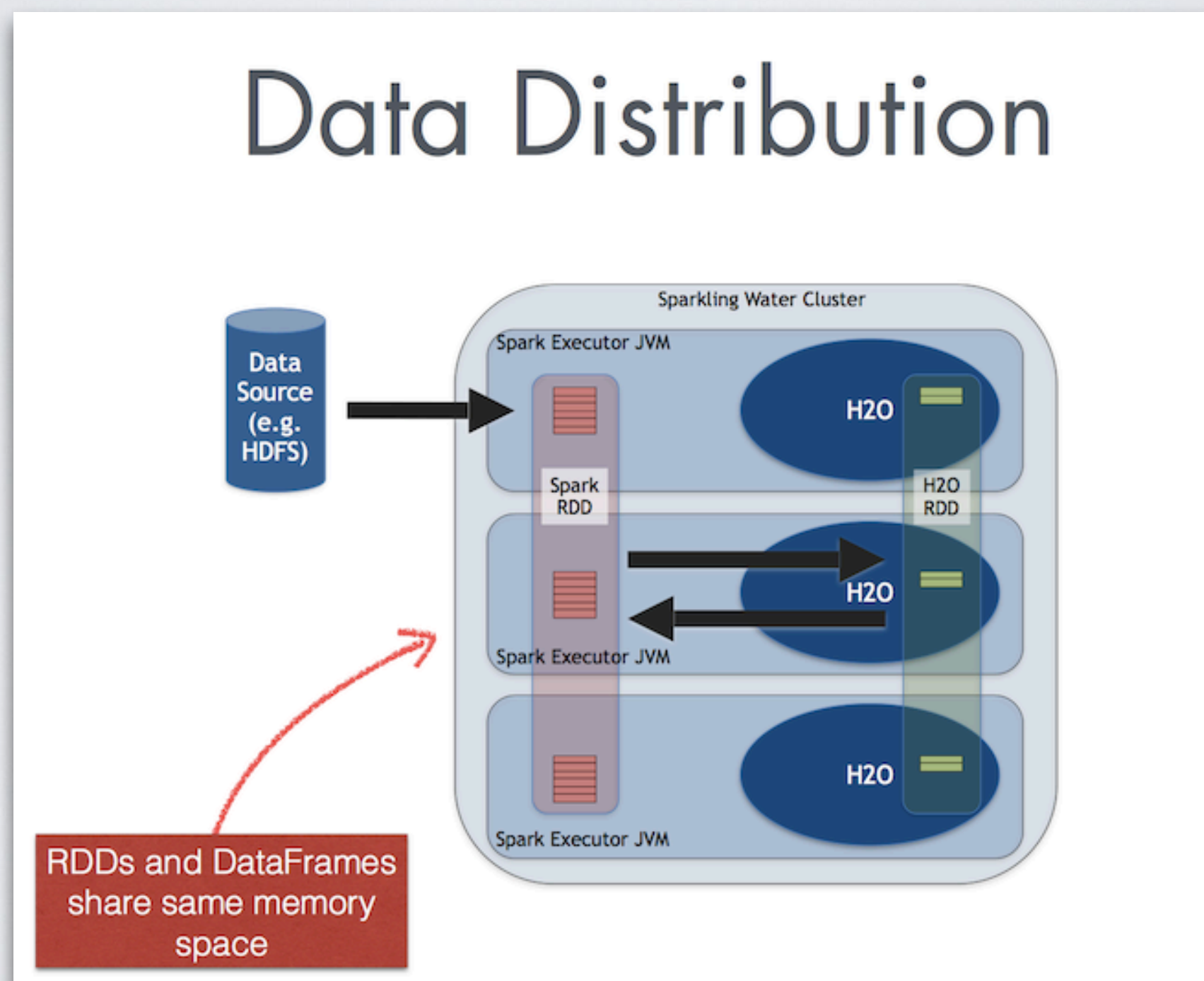


Sparkling Water allows users to combine the fast, scalable machine learning algorithms of H2O with the capabilities of Spark. With Sparkling Water, users can drive computation from Scala/R/Python and utilize the H2O Flow UI, providing an ideal machine learning platform for application developers.

Sparkling Water Design



Sparkling Water enables transformation between different types of Spark RDD and H2O's H2ORFrame, and vice versa.



source: <https://www.slideshare.net/0xdata/2015-02-19stratataalk>