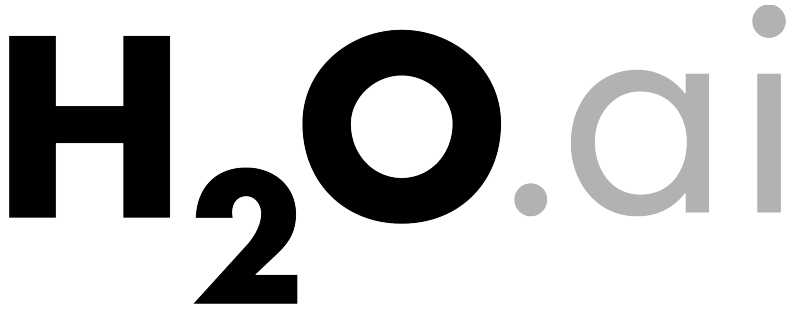
H_2O Training (20th Feb, 2018)

- Agenda
  - Sign in / Setup Equipment
  - Introduction
  - Machine Learning Basics
  - Hands-on (Web/R/Python)
    - Regression
    - Classification
    - Clustering
  - Driverless AI demo
  - Q&A / Open Discussion

- Lunch Break (12:00 – 13:00)



Jo-fai (Joe) Chow

Data Scientist

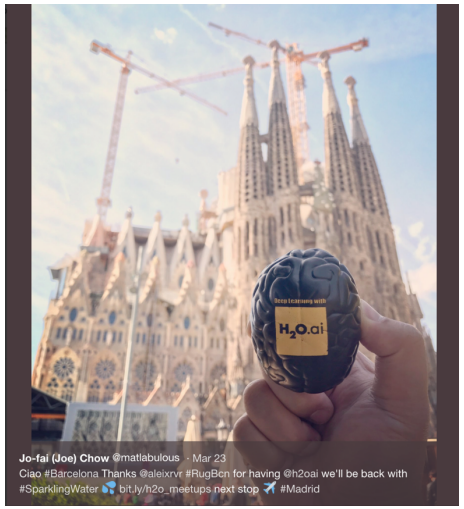
joe@h2o.ai

@matlabulous

About Me

- Civil (Water) Engineer
  - 2010 – 2015
    - Consultant (UK)
      - Utilities
      - Asset Management
      - Constrained Optimization
    - EngD (Industrial PhD) (UK)
      - Infrastructure Design Optimization
      - Machine Learning + Water Engineering
      - Discovered H_2O in 2014
- Data Scientist
  - 2015 – 2016
    - Virgin Media (UK)
    - Domino Data Lab (Silicon Valley)
  - 2016 – Present
    - H_2O.ai (Silicon Valley)
  - How?
    - bit.ly/joe_kaggle_story

What is Joe's role at H_2O.ai?



- Data Scientist
  (the job title …)
- Sales Engineer /
  Conference & Meetup Speaker /
  Community Manager
  (hard truth about tech startup ...)
- Unofficial Photographer of H_2O.ai SWAG
  (the travelling data scientist)
- H_2O.ai SWAG EMEA Distributor
  (please help yourself)

H_2O.ai

About H_2O.ai

Company Overview

Founded	2011 Venture-backed, debuted in 2012
Products	• **H_2O Open Source In-Memory AI Prediction Engine** • Sparkling Water (H_2O + Spark) • Enterprise Steam • Driverless AI
Mission	Operationalize Data Science, and provide a platform for users to build beautiful data products
Team	85+ employees • Distributed Systems Engineers doing Machine Learning • World-class visualization designers
Headquarters	Mountain View, CA





H2O.ai

THE H_2O.ai TEAM

Scientific Advisory Council

Dr. Trevor Hastie

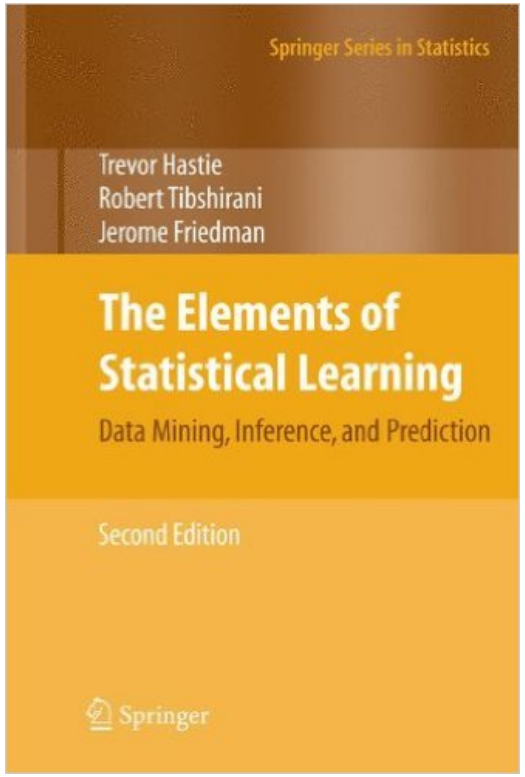
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

Dr. Robert Tibshirani

- Professor of Statistics and Health Research and Policy, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Author, *Regression Shrinkage and Selection via the Lasso*
- Co-author, *An Introduction to the Bootstrap*

Dr. Steven Boyd

- Professor of Electrical Engineering and Computer Science, Stanford University
- PhD in Electrical Engineering and Computer Science, UC Berkeley
- Co-author, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*
- Co-author, *Linear Matrix Inequalities in System and Control Theory*
- Co-author, *Convex Optimization*



H2O.ai

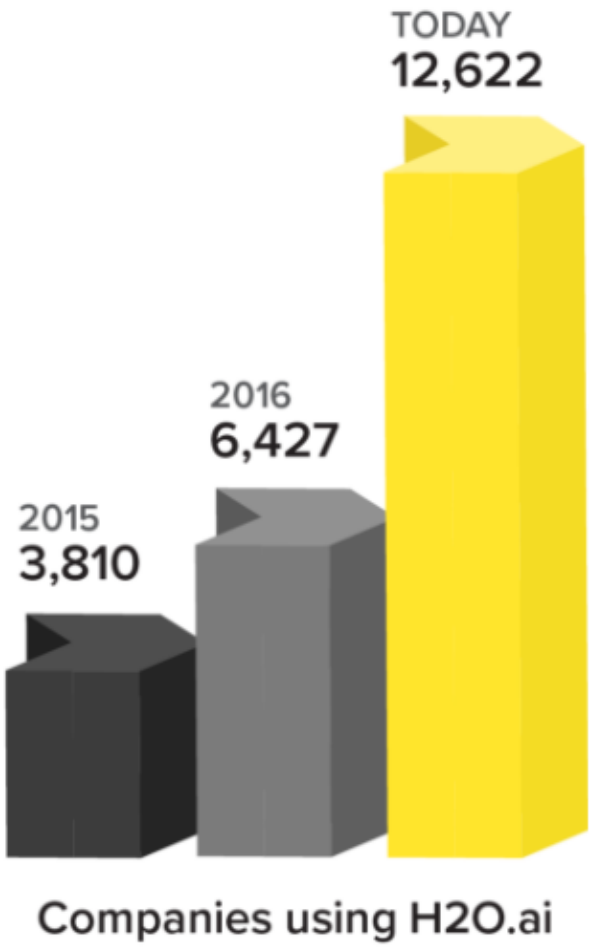
wenphan
@wenphan

Following

So much brain power in one place: @ArnoCandel and Stanford profs. Boyd, Tibs, and Hastie. Hacking algos at @h2oai HQ

H2O.ai

12,000+ Companies use H2O — World Wide Community Adoption

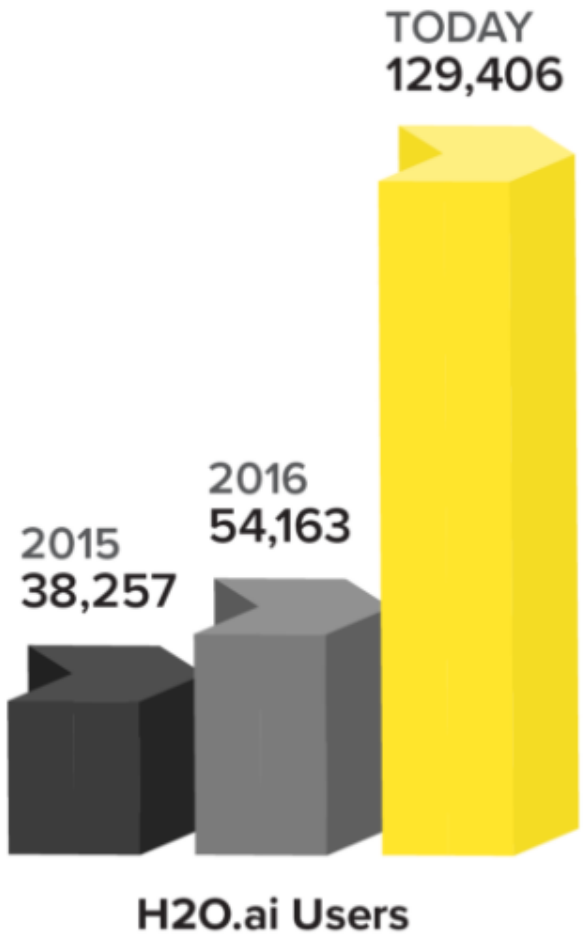


222 OF THE FORTUNE 500 ♥ H_2O

8 OF TOP 10 BANKS

7 OF TOP 10 INSURANCE COMPANIES

4 OF TOP 10 HEALTHCARE COMPANIES



H_2O.ai

H2O.ai Select Paying Customers

"Overall customer satisfaction is very high." - Gartner

H2O.ai

Award-Winning Application (PwC + H_2O.ai)

pwc | Global

PwC Global › About us › › Stories about our people, clients and communities

Harnessing the power of AI to transform the detection of fraud and error

Setting the scene

PwC has invested significantly in pioneering the use of artificial intelligence for the audit and has partnered with H2O.ai, a leading Silicon Valley-based AI company.

Following 18 months of development, the first outcome of this partnership is PwC's GL.ai, the first module of PwC's Audit.ai - a revolutionary bot that does what humans can't. Its AI analyses billions of different data points in seconds and applies judgement to detect anomalies in general ledger transactions.

"The reason this is such a brilliant tool is the ability to look at different risks in context at the same time. For example, it would be uneconomical for an auditor to look at every single user's pattern of activity and decide what was unusual. With GL.ai, the algorithms do it for us."

Laura Needham partner, PwC UK

PwC
@PwC
Follow

Exciting night at this year's @WAI_News Awards: PwC wins 2017 Audit Innovation of the Year! pwc.to/Glai17 #taandiab17

10:15 PM - 4 Oct 2017

http://www.pwc.com/gx/en/about/stories-from-across-the-world/harnessing-the-power-of-ai-to-transform-the-detection-of-fraud-and-error.html

H2O In Action

Check out our website h2o.ai

ADP

Various data leaders discuss the transformative impact of H2O AI for ADP.

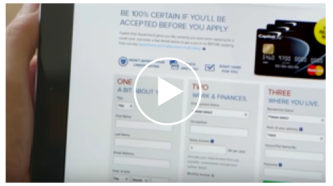
Insurance

What data products mean and why H2O keeps this industry leader relevant.

Progressive

See how Progressive uses H2O predictive analytics for User-based Insurance (UBI).

Capital One



Capital One uses H2O machine learning for various use cases.

MarketShare

H2O predictive analytics helps boost the impact and results of digital marketing.

Kaiser

Kaiser uses H2O machine learning to save lives.

Zurich Insurance

Zurich turned to H2O as a strategic differentiator for commercial insurance.

Comcast

Comcast uses H2O to improve customer experience.

McKesson

McKesson discusses the adoption of artificial intelligence in healthcare.

Macy's

Macy's uses H2O for personalized site recommendations.

Transamerica

Transamerica turns to H2O to develop an insurance recommendation platform.

Paypal

Paypal turned to H2O Deep Learning for fraud detection and customer churn.

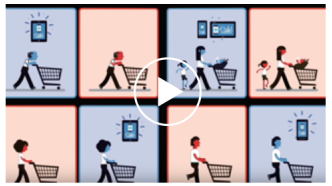
eBay

eBay chose H2O for open source machine learning.

Cisco

Cisco uses H2O to build a scalable model factory to improve sales and marketing.

Nielsen Catalina Solutions



H2O helps the country's largest TV behavior analytics company optimize ad performance.

H2O.ai

Partners

NVIDIA's (NASDAQ: NVDA) invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined modern computer graphics, and revolutionized parallel computing. More recently, GPU deep learning ignited modern AI -- the next era of computing -- with the GPU acting as the brain of computers, robots, and self-driving cars that can perceive and understand the world.
Website

Founded in 1975, Microsoft (Nasdaq "MSFT") is the worldwide leader in software, services, devices and solutions that help people and businesses realize their full potential. You can launch Sparkling Water on Azure HDInsights with just a few clicks to build your data science pipeline on the cloud.
Website | Documentation

Watson Data Platform enables AI-powered decision making by simplifying and automating the development and operationalization of new insights. It enables unprecedented levels of collaboration amongst data savvy professionals.

Kensu's mission is to lift Data Science to the Enterprise level focusing on the production environment and the maintenance across time.
Website

SigOpt is the optimization platform that amplifies your research. SigOpt takes any research pipeline and tunes it, right in place.
Website

Nimbix is the leading provider of purpose-built cloud computing for machine learning, AI and HPC applications. Powered by JARVICE™, the Nimbix Cloud provides high-performance software as a service, dramatically speeding up data processing for Energy, Life Sciences, Manufacturing, Media and Analytics applications.
Website

Cloudera delivers a modern platform for data management and analytics, helping businesses solve their most challenging problems with data.

Anaconda is the leading open data science platform powered by Python. Continuum Analytics is the creator and driving force behind Anaconda. We put superpowers into the hands of people who are changing the world.

The MapR Converged Data Platform integrates enormous power of Hadoop and Spark with global event streaming, real-time database capabilities, and enterprise storage.

Splunk Inc. provides the leading platform for Operational Intelligence. Customers use Splunk to search, monitor, analyze and visualize machine data.
Website

DataScience.com pairs data expertise with powerful tools to help businesses unlock the value in their data. Enabling data science for every business.
Website

MapD makes queries faster and creates a fluid and immersive data exploration experience that removes the disconnect between an analyst and their data. Making extracting insight from data effortless and lightning fast.
Website

Databricks provides a just-in-time data platform, to simplify data integration, real-time experimentation, and robust deployment of production applications.

Hortonworks drives actionable intelligence with Connected Data Platforms that maximize the value of all data—data-in-motion and data-at-rest.

Minio is an object storage server built for cloud application developers and devops.

Community Expansion

15 Meetups a month

66,843	33	50	45	18
members	interested	Meetups	cities	countries

Find out more: **www.h2o.ai/community/**

22 FEB

Thursday, February 22, 2018

External Registration Required: Amsterdam AI & Deep Learning Meetup at ING

Hosted by Jo-fai Chow

From Amsterdam Artificial Intelligence & Deep Learning

You're going

Share | Tweet | Invite

Details

Many thanks to ING for hosting our next event in Amsterdam.

Note: EXTERNAL REGISTRATION (Eventbrite) is required. Please use the following link to register. You must provide your real name and email address. Please also bring your ID for security check (just in case).

https://www.eventbrite.com/e/amsterdam-ai-deep-learning-meetup-at-ing-tickets-42783160585

Agenda:

17:50 - 18:10 - Doors open, pizzas, drinks and networking

18:10 - 18:20 - Introduction

18:20 - 18:45 - Talk 1: Anomaly Detection in Finance using Isolation Forest by Andreea Bejinaru

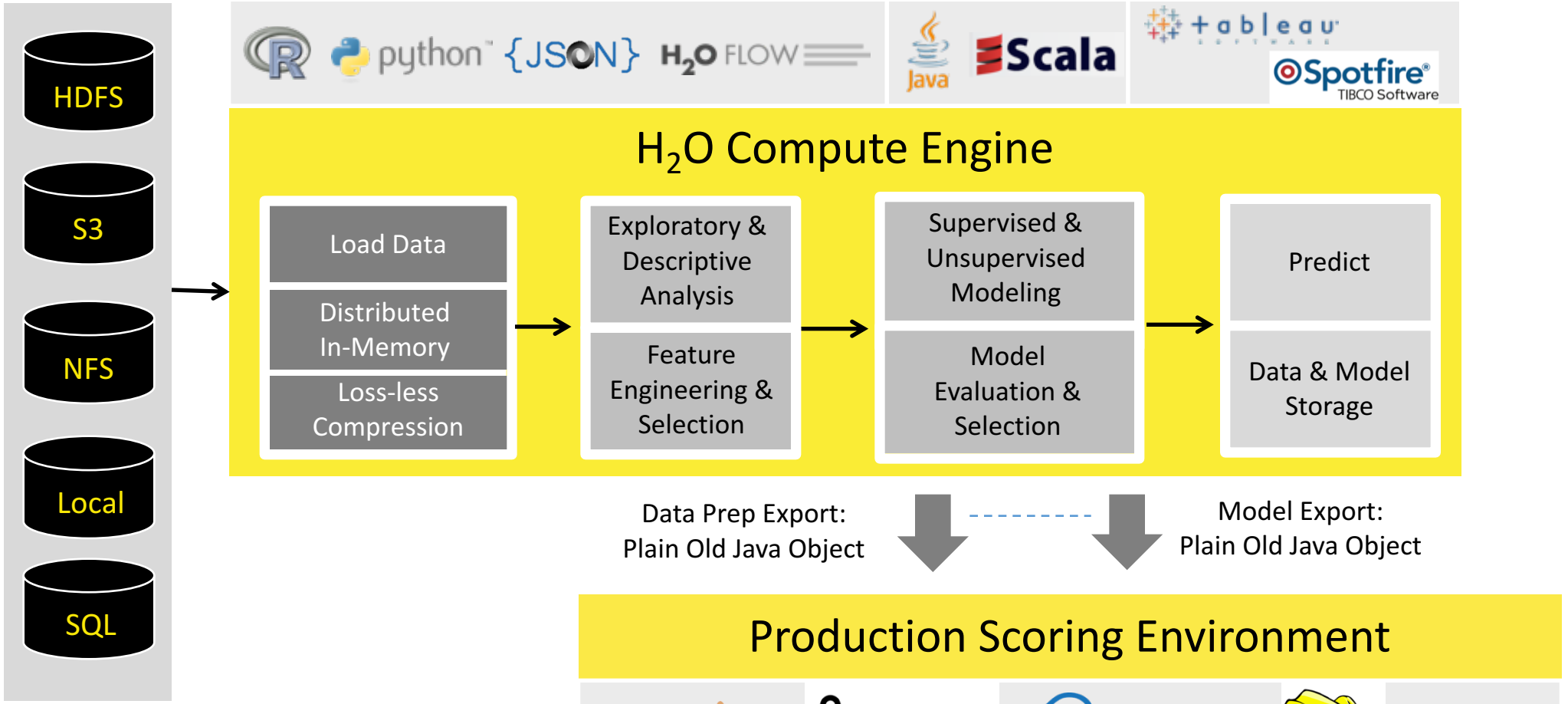
Organizer tools

Thursday, February 22, 2018
5:45 PM to 8:15 PM
Add to calendar

Needs a location

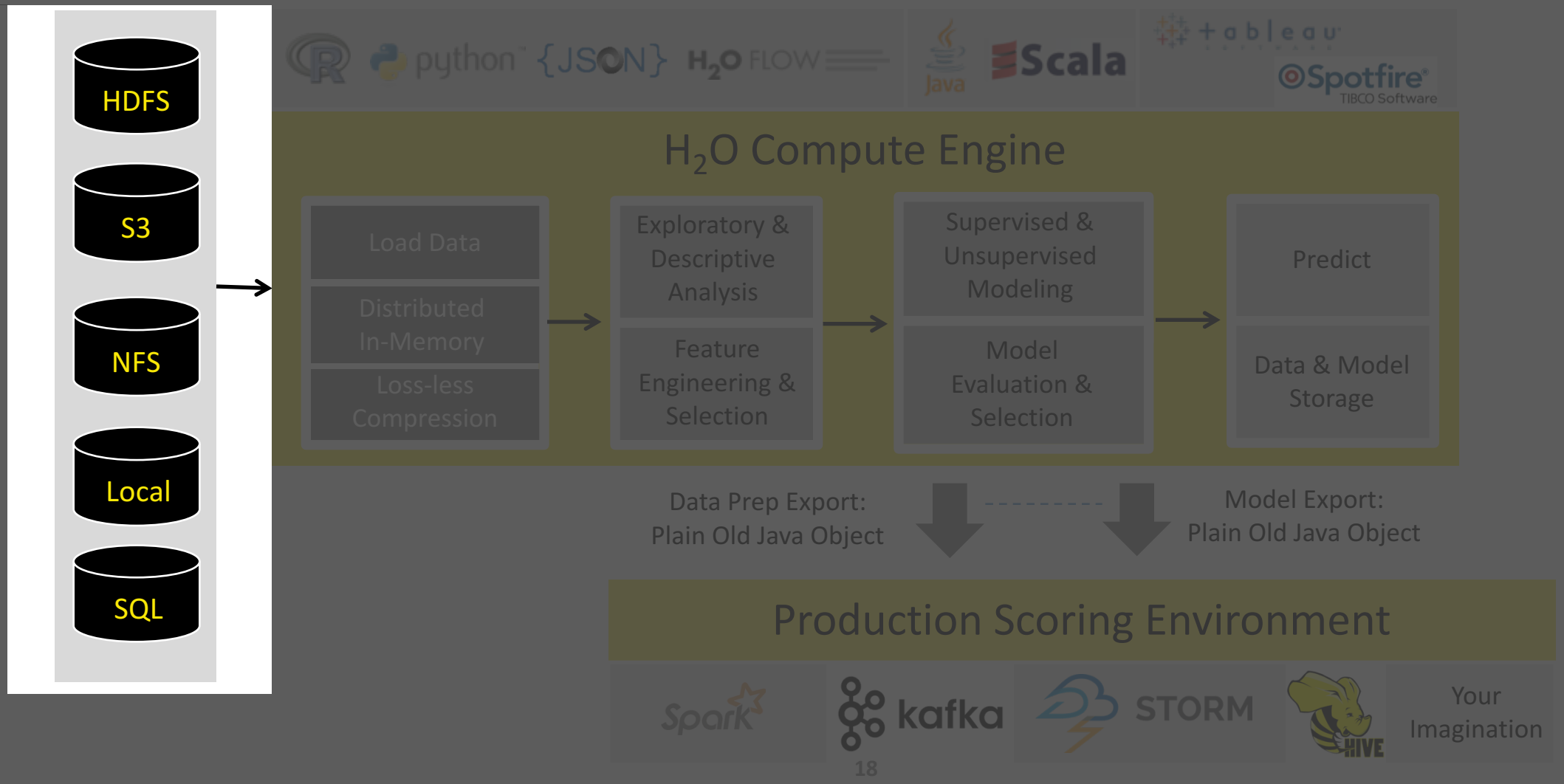
H_2O Machine Learning Platform

High Level Architecture



High Level Architecture

Import Data from Multiple Sources

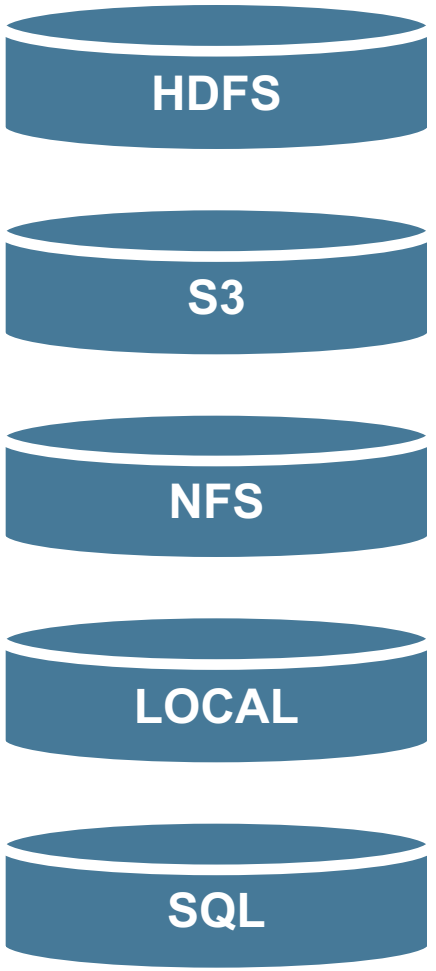


Supported Formats & Data Sources

9 Formats

File type or Folder of Files

- CSV
- XLS
- XLSX
- ORC*
- Hive*
- SVMLight
- ARFF
- Parquet
- Avro 1.8.0*



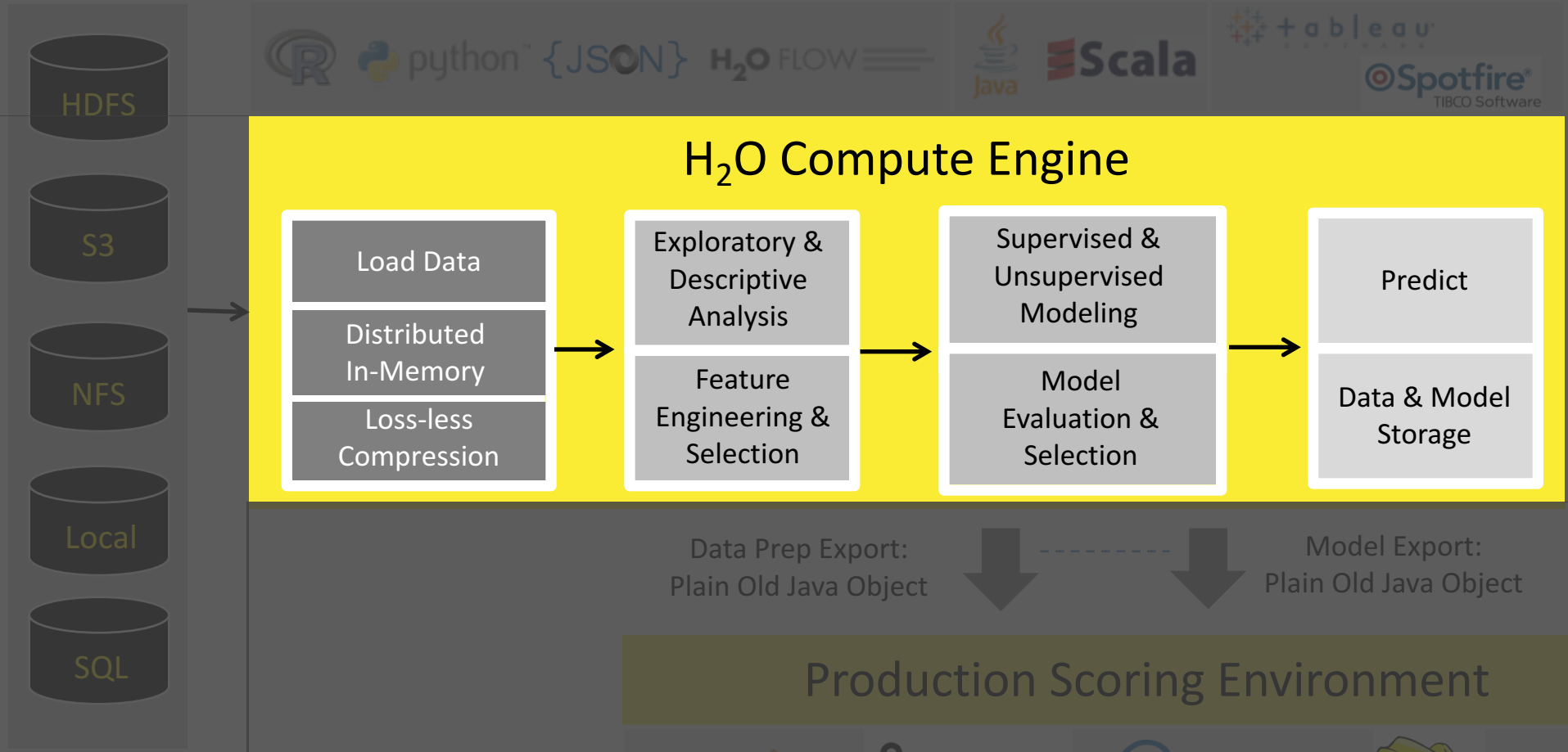
5 Sources

* 1. only if H2O is running as a Hadoop job
* 2. Hive files that are saved in ORC format
* 3. without multi-file parsing or column type modification

H2O.ai

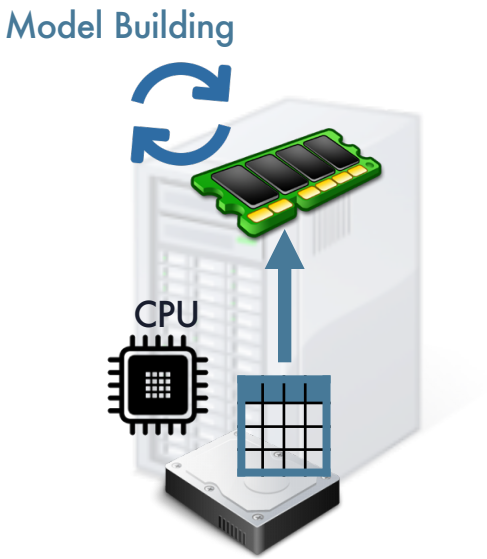
High Level Architecture

Fast, Scalable & Distributed Compute Engine Written in Java

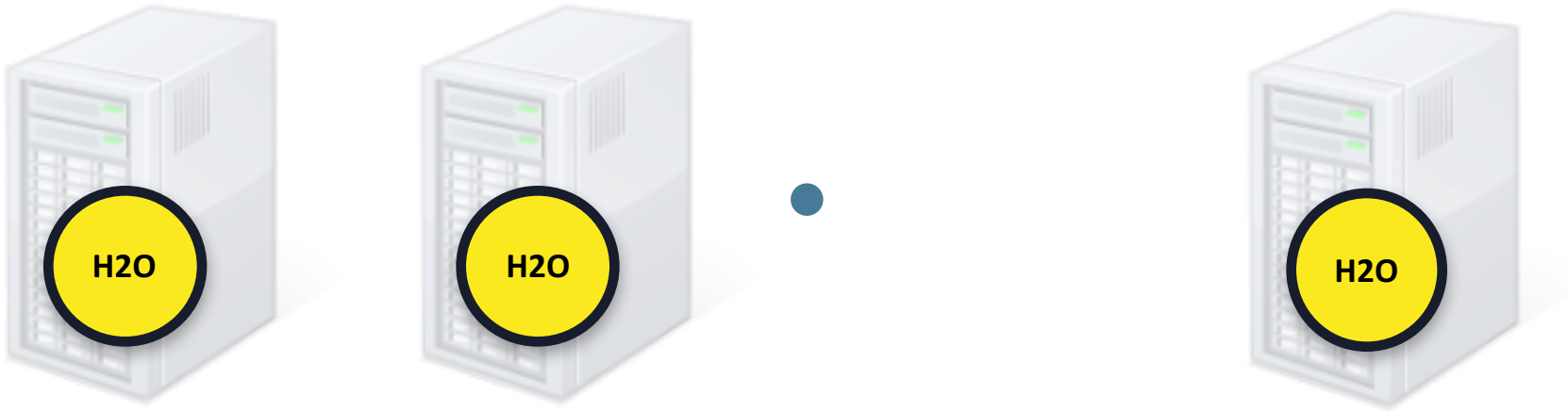


H_2O Core

H_2O Core

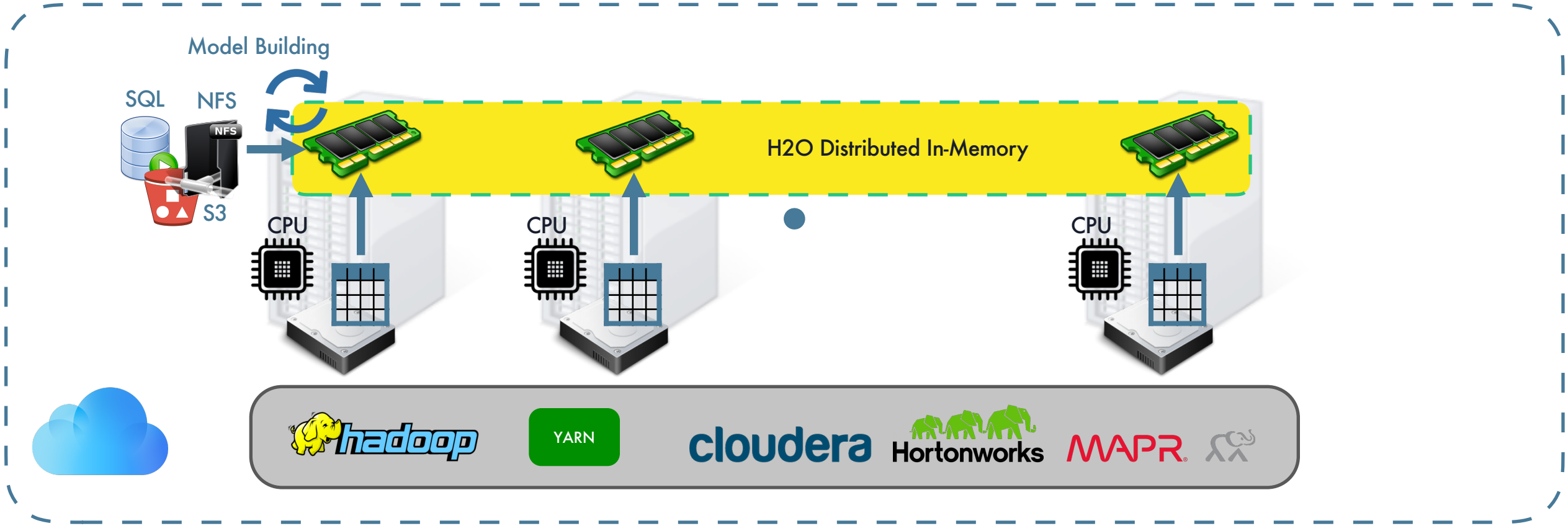


H_2O Core



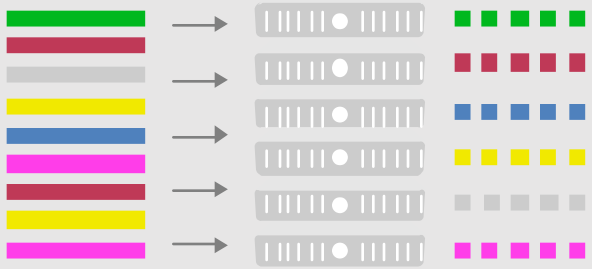
H_2O Core

H_2O Core

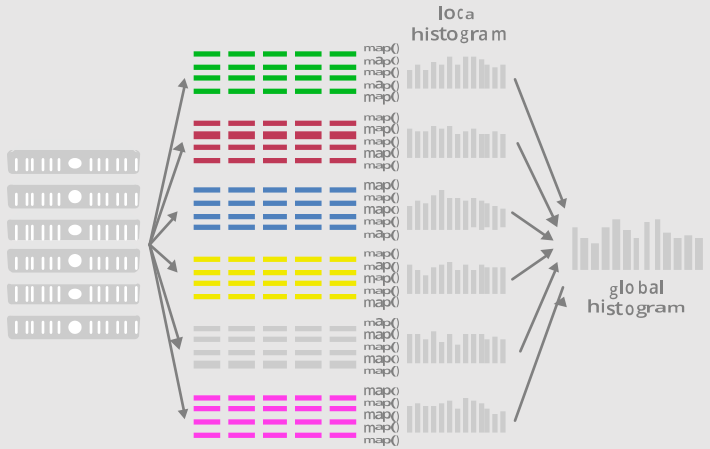

Model Building
SQL
NFS
NFS
S3
H2O Distributed In-Memory
CPU
CPU
CPU
hadoop
YARN
cloudera
Hortonworks
MAPR

Distributed Algorithms

Foundation for Distributed Algorithms



Parallel Parse into **Distributed Rows**



Fine Grain Map Reduce Illustration: Scalable Distributed Histogram Calculation for GBM

Advantageous Foundation

- Foundation for In-Memory Distributed Algorithm Calculation - **Distributed Data Frames** and **columnar compression**
- All algorithms are distributed in H_2O: GBM, GLM, DRF, Deep Learning and more. Fine-grained map-reduce iterations.
- **Only enterprise-grade, open-source distributed algorithms in the market**

User Benefits

- "Out-of-box" functionalities for all algorithms **(NO MORE SCRIPTING)** and uniform interface across all languages: R, Python, Java
- **Designed for all sizes of data sets, especially large data**
- **Highly optimized Java code for model exports**
- **In-house expertise for all algorithms**

Algorithms Overview

Supervised Learning

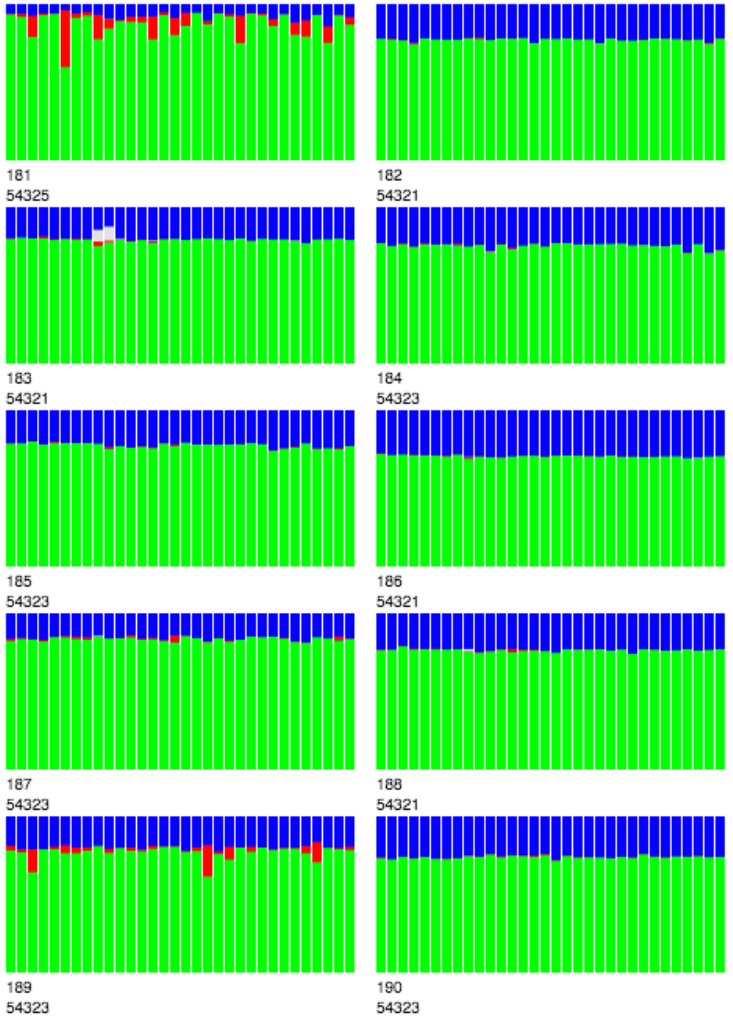
Statistical Analysis	• **Generalized Linear Models**: Binomial, Gaussian, Gamma, Poisson and Tweedie • **Naïve Bayes**
Ensembles	• **Distributed Random Forest**: Classification or regression models • **Gradient Boosting Machine**: Produces an ensemble of decision trees with increasing refined approximations
Deep Neural Networks	• **Deep learning**: Create multi-layer feed forward neural networks starting with an input layer followed by multiple layers of nonlinear transformations

Unsupervised Learning

Clustering	• **K-means**: Partitions observations into k clusters/groups of the same spatial size. Automatically detect optimal k
Dimensionality Reduction	• **Principal Component Analysis**: Linearly transforms correlated variables to independent components • **Generalized Low Rank Models:** extend the idea of PCA to handle arbitrary data consisting of numerical, Boolean, categorical, and missing data
Anomaly Detection	• **Autoencoders**: Find outliers using a nonlinear dimensionality reduction using deep learning

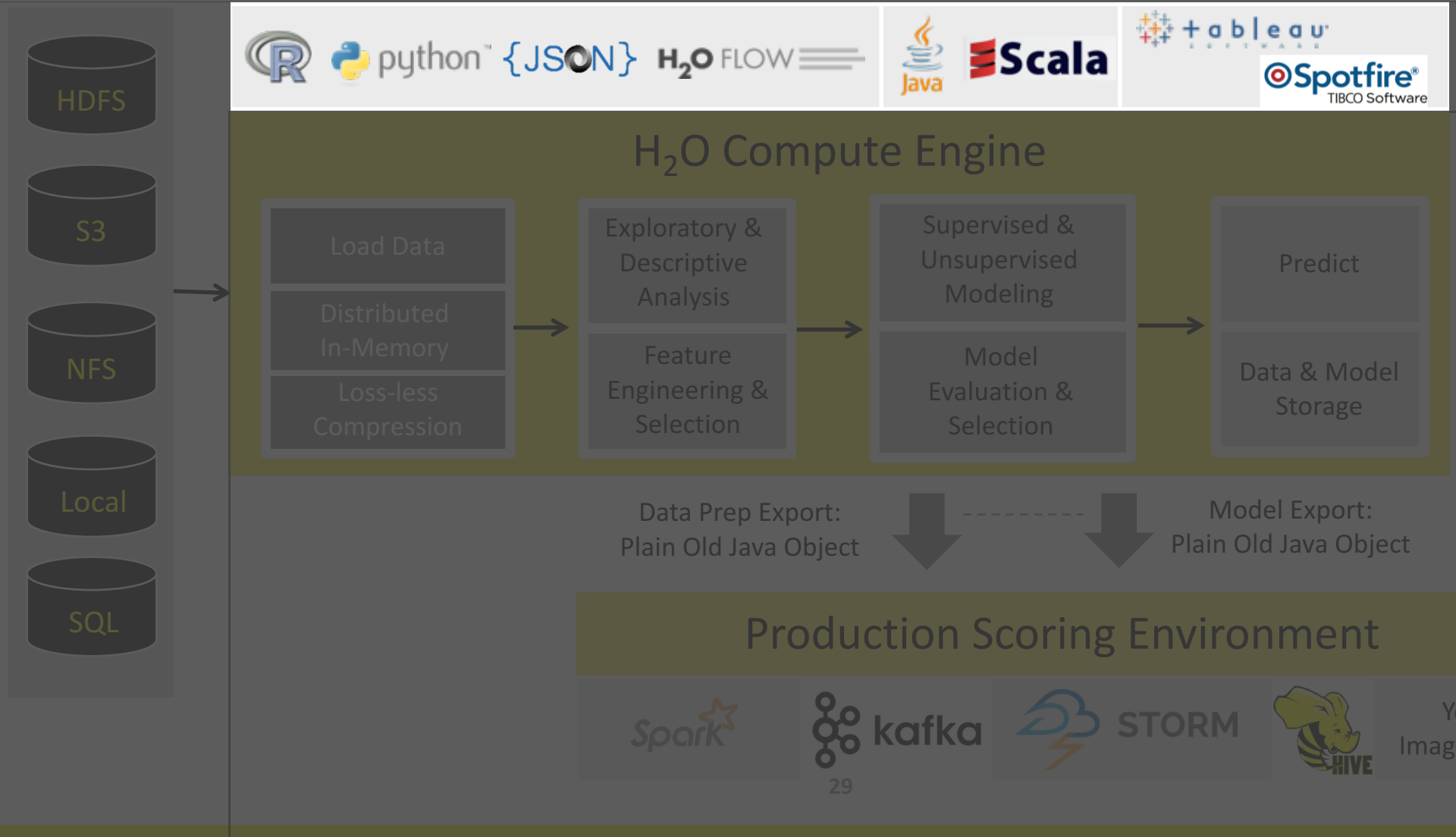
H_2O Water Meter (CPU Monitor)

10 x 32 = 320 Cores

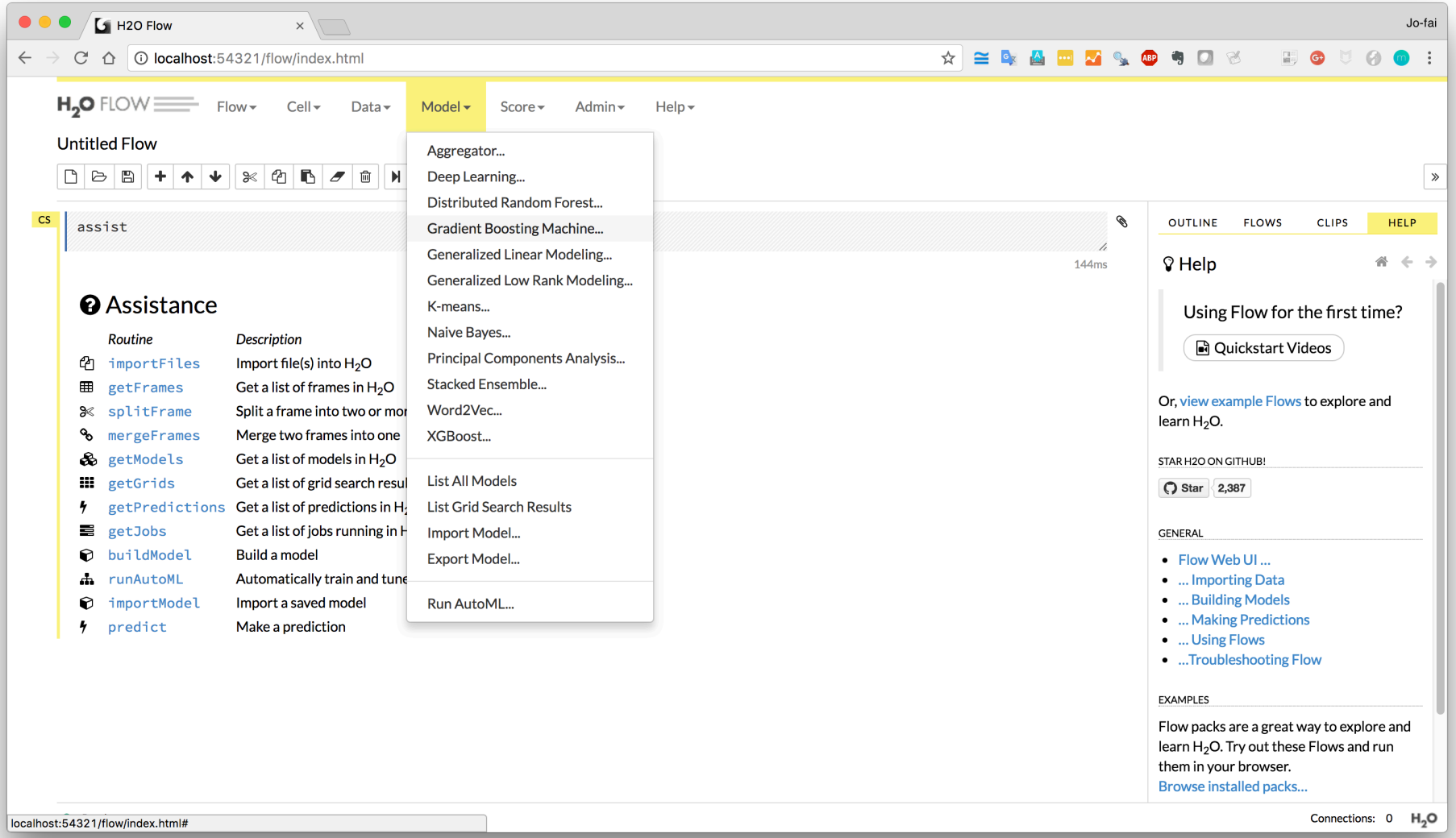


High Level Architecture

Multiple Interfaces

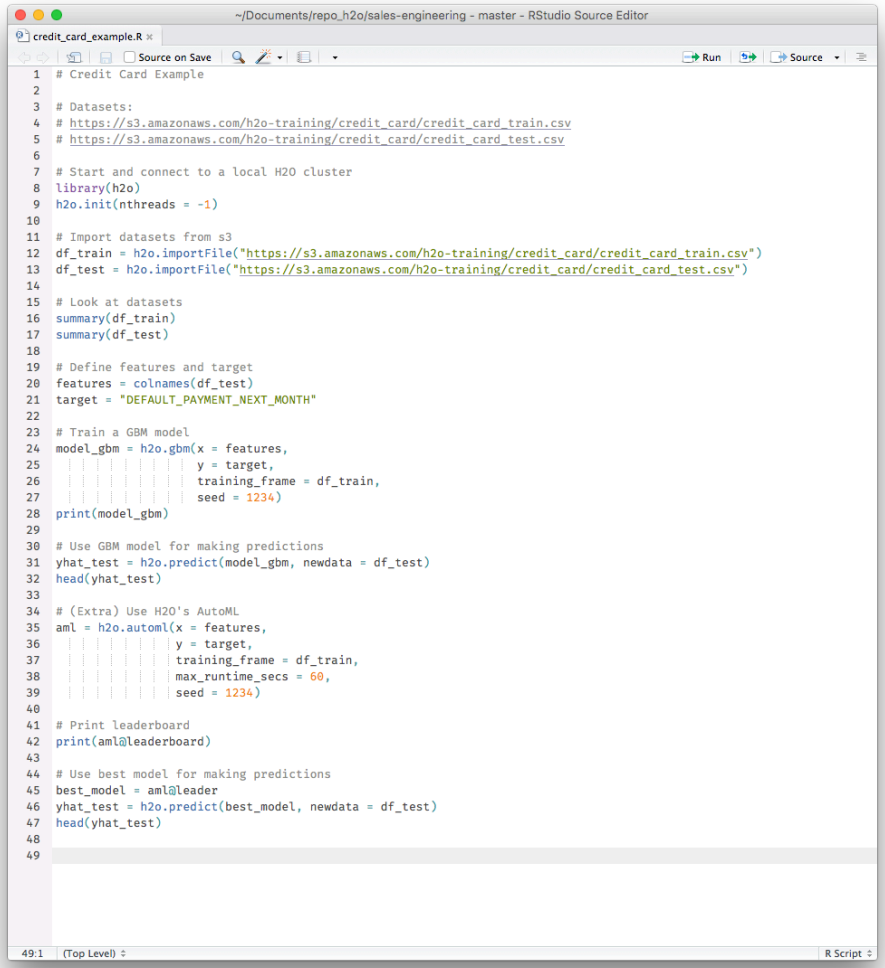


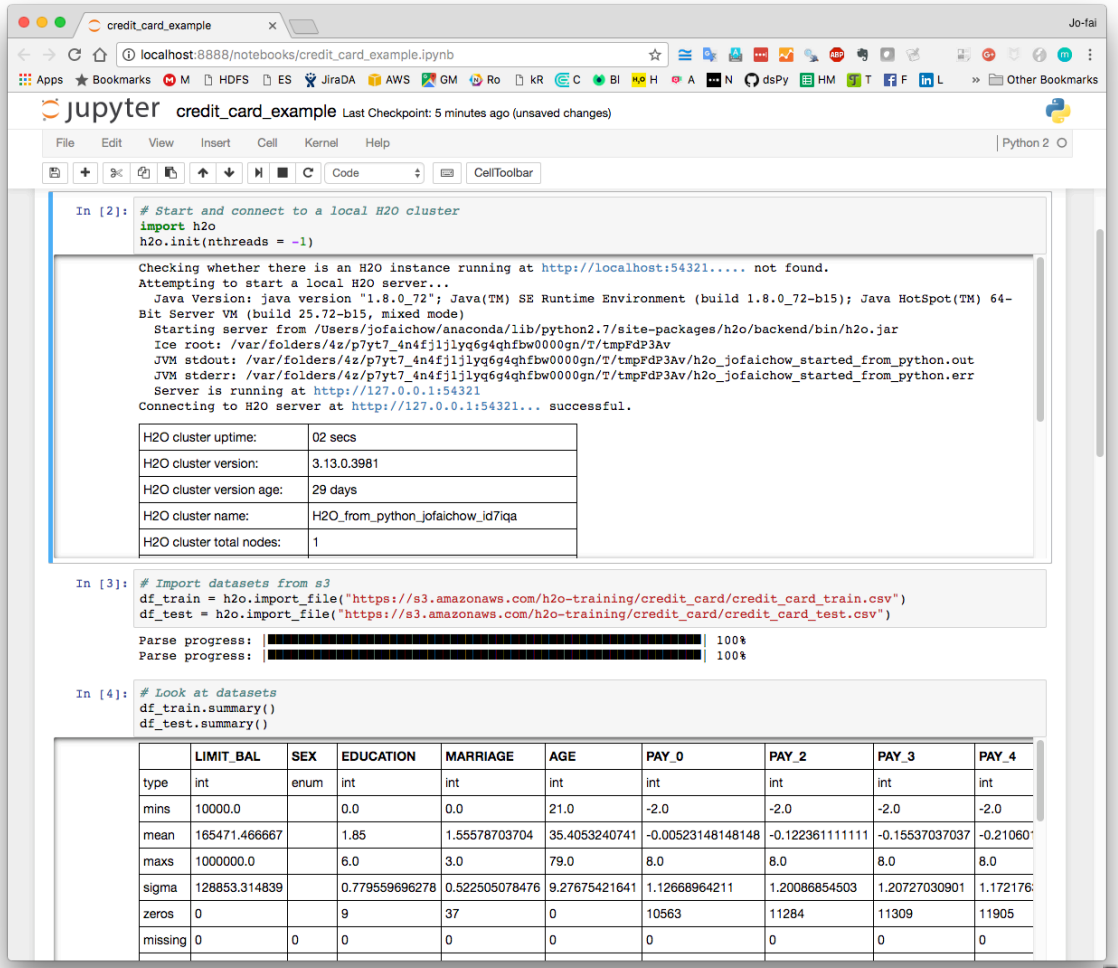
H_2O Flow (Web)



H2O.ai

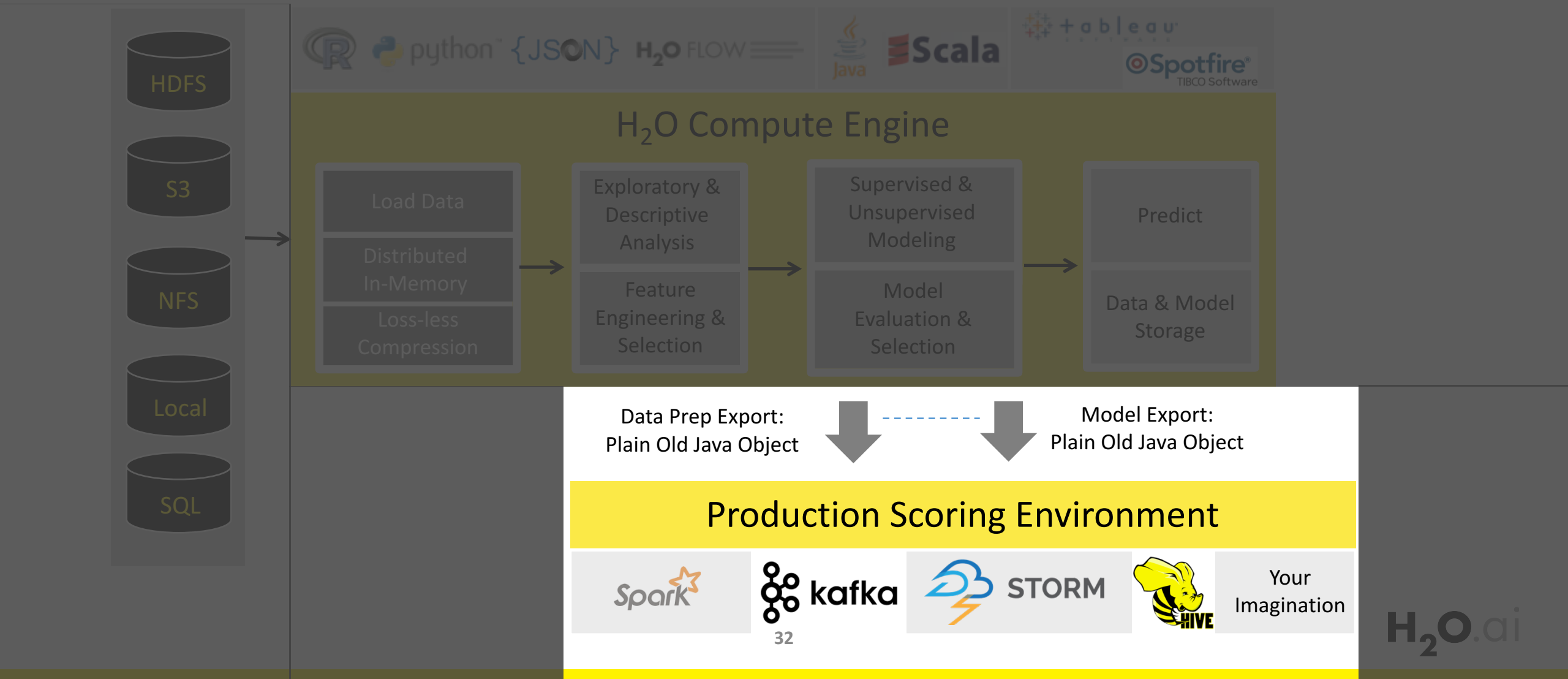
Interface – R and Python

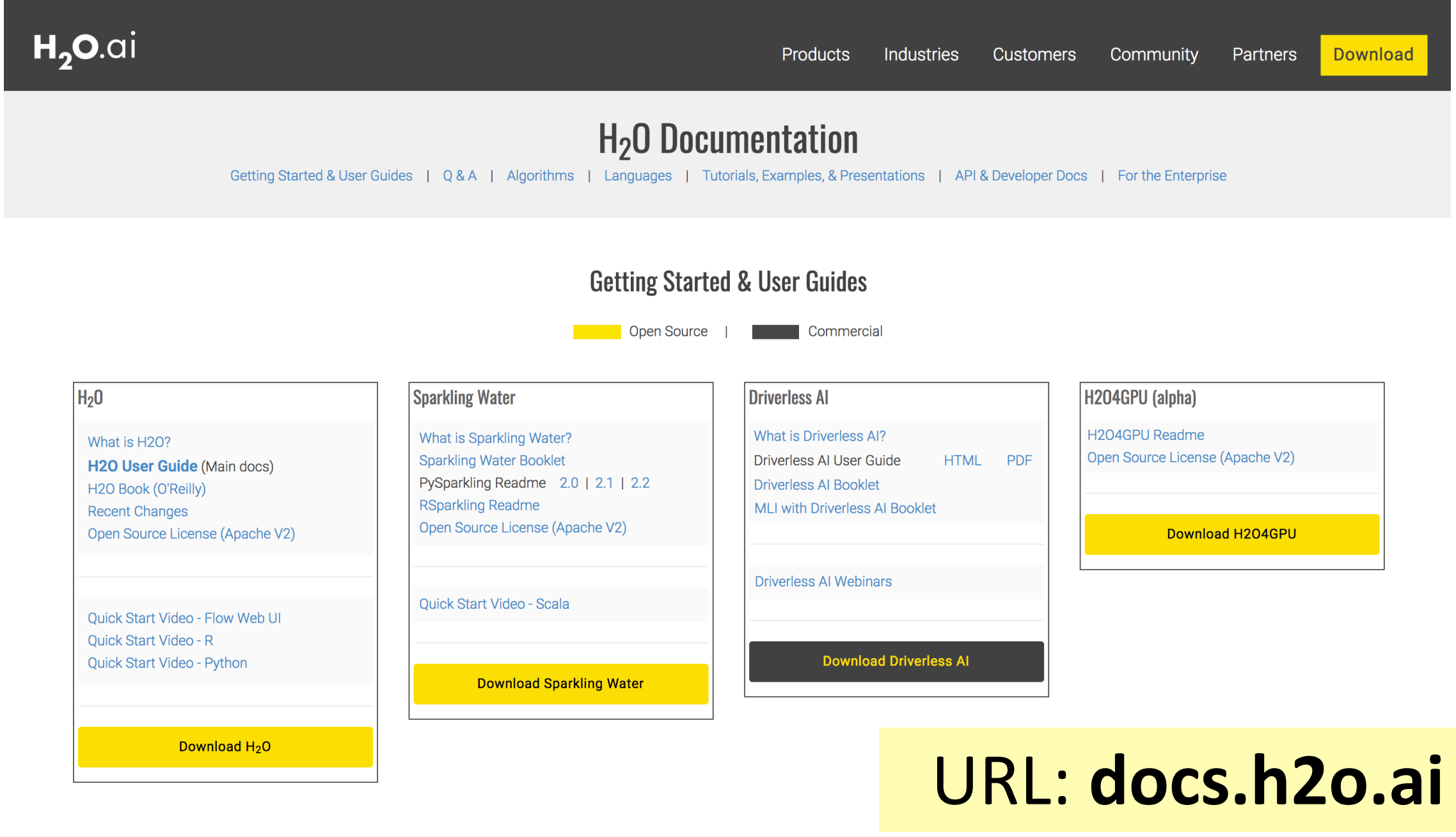




High Level Architecture

Export Standalone Models for Production





URL: **docs.h2o.ai**

H2O.ai