Jakub Háva
jakub@h2o.ai

Introduction to Sparkling Water

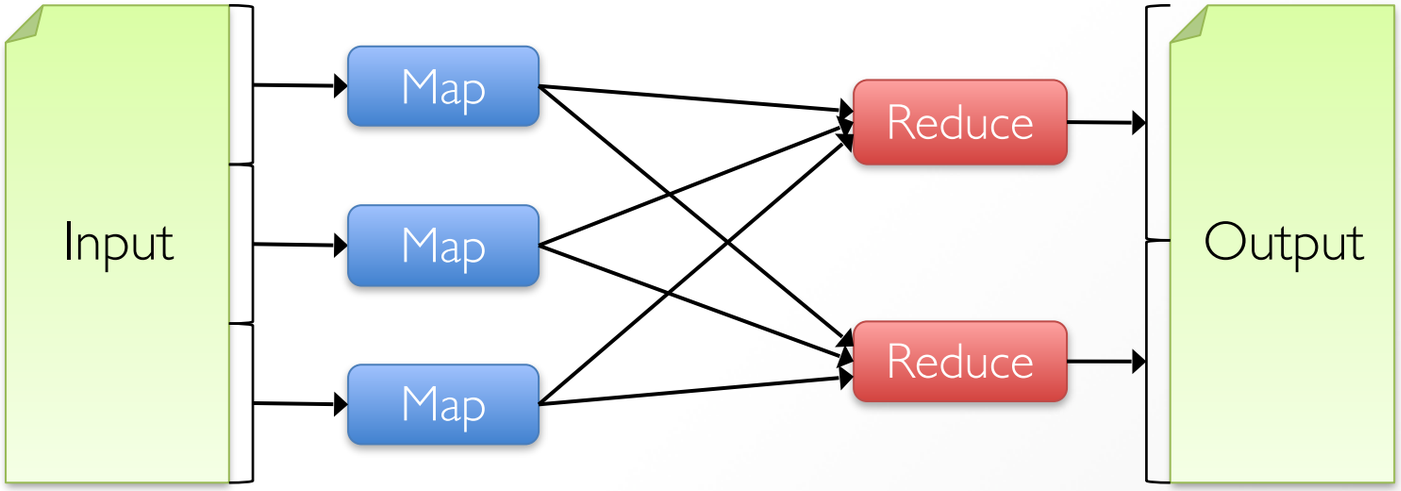
Firemní semináře @ MFF UK, Praha
October 24, 2018

Who are we?

- **Kuba**
  - Senior Software engineer at H2O.ai - Code Owner - Sparkling Water
  - Master's at Charles University (CZ)
    - Implemented high-performance cluster monitoring tool for JVM based languages (JNI, JVMTI, instrumentation)
- **Michal**
  - VP of Engineering at H2O.ai
  - Author of Sparkling Water
  - Ph.D at Charles University (CZ), PostDoc at Purdue (US)

Data Flow in typical Map-Reduce

- Acyclic data flow, from storage A to storage B
- Benefit is that runtime can decide where to run tasks and can automatically recover in case of their failures



Inefficiencies of acyclic Data Flow

- Inefficient for application that repeatedly reuse a *working set* of data
  - Iterative algorithms - machine learning
  - Interactive data mining tools (R, Scala interpreter)
- With typical MapReduce, the data are reloaded for each query

Solution:

RDD

RDD: Resilient Distributed Dataset

- Allow apps to keep working sets in memory for efficient reuse
- Whilst keep the benefits of MapReduce
  - Fault tolerance, data locality, scalability
- General approach

Spark Programming Model

H_2O.ai
Machine Intelligence

Spark Runs on

- YARN, Mesos, Standalone cluster
- 1 driver node
- Several executor nodes
  - can be started dynamically when needed based on provided resources by resource manager

RDDs

- Immutable, partitioned collections of objects
- Created through parallel transformations
  - **map**, **filter**, **groupBy**, **join**…
  - Can be cached for efficient use on various levels
- Transformations does not start any computations - lazy behaviour
- Actions start the computations
  - **count**, **reduce**, **collect**, **save**
- RDDs are split across multiple partitions on multiple nodes

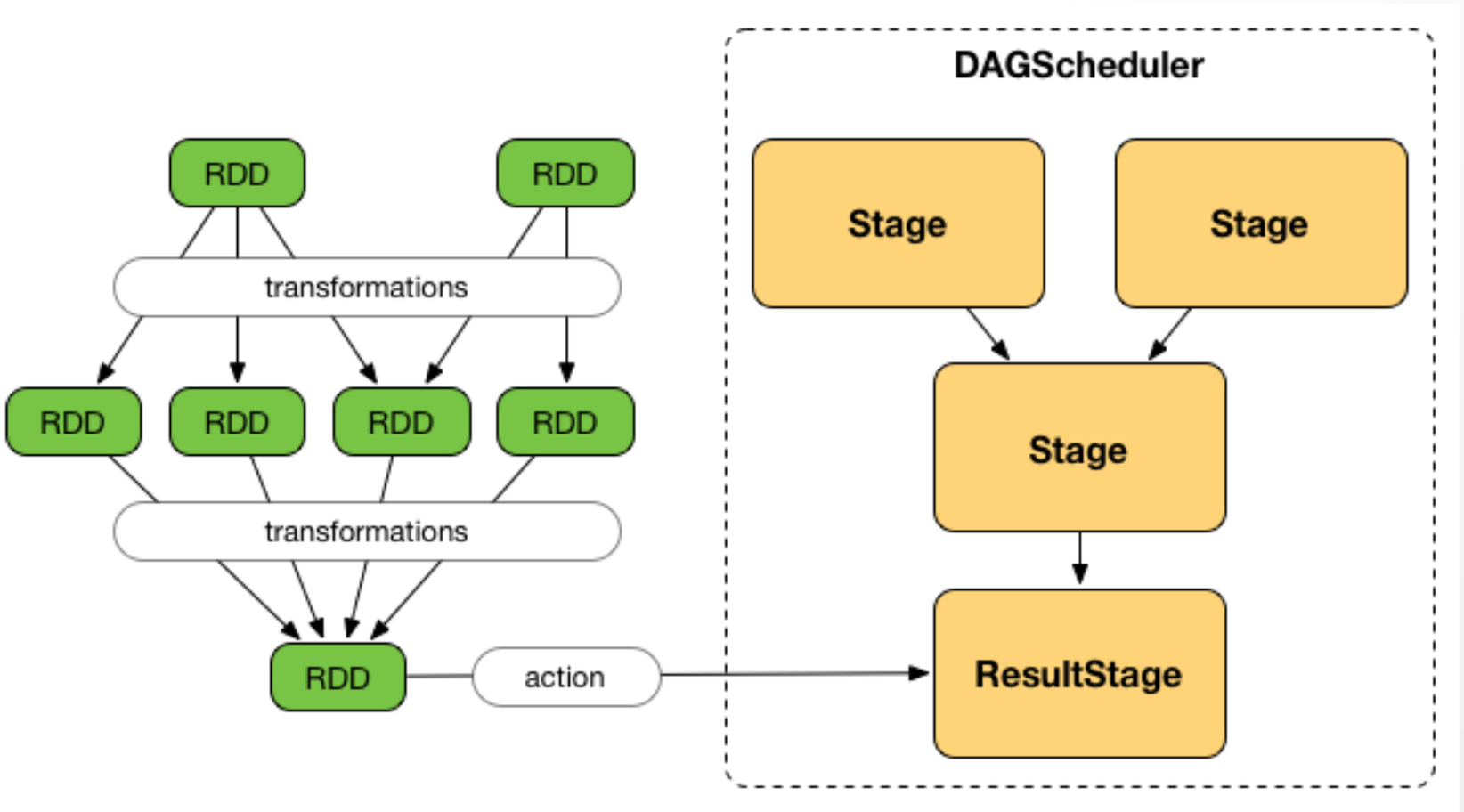
RDD lineage

- Result of applying transformations on the RDD
- val rdd = sc.textFile(…)
- val filtered = rdd.filter(…)
- **toDebugString** method on RDD - see the lineage
- Plan describing the computation

Spark Job

- Starts with an action called on RDD
- Job is split into multiple stages (depends on type of transformations - **stage boundaries**)
- Stage is executed by DAG Scheduler
- Each stage is split into several tasks (usually #partitions in the stage)

DAG Scheduler



Abstractions on top of RDD

RDD

- Low-level transformations
- Data is unstructured
- Good when we want to manipulate with data using functional programming constructs
- Don’t have schema of the data
- On heap, thus affected by GC and are using Java Serialisation

RDD Example

```
rdd.filter(_.age > 21)                // transformation
   .map(_.last)                       // transformation
   .saveAsObjectFile("under21.bin") // action
```

```
rdd.filter(_.age > 21)
```

Dataframes

- Build on top of RDDs => immutable distributed collection of data
- Data organised into columns like a table in relational database
- Have a schema
- DSL and SQL like specific language
- Better performance
  - Can use the schema and a lot of optimilizations - Spark Query Optimiser
- Can be stored off-heap, Spark specific serialisation

Datarame example

```
df.filter("age > 21")
```

```
df.filter(df.col("age").gt(21))
```

Dataset

- First appeared in Preview of Spark 1.6
- Attempt to bring the best from RDDs and DataFrames
  - Compile type safety as in RDD
  - Performance as on DataFrames
- Off-heap storage (outside the JVM memory)
- Usage of encoders to translate between JVM representation and Spark internal binary format, no need to de-serialise the object in order to read it

Dataset example

```
val sc = new SparkContext(conf)
val sqlContext = new SQLContext(sc)
import sqlContext.implicits._
val sampleData: Seq[ScalaPerson] = ScalaData.sampleData()
val dataset = sqlContext.createDataset(sampleData)
```

```
dataset.filter(_.age < 21);
```

Introduction to Scalable & Automatic Machine Learning with H_2O

H_2O.ai

Agenda

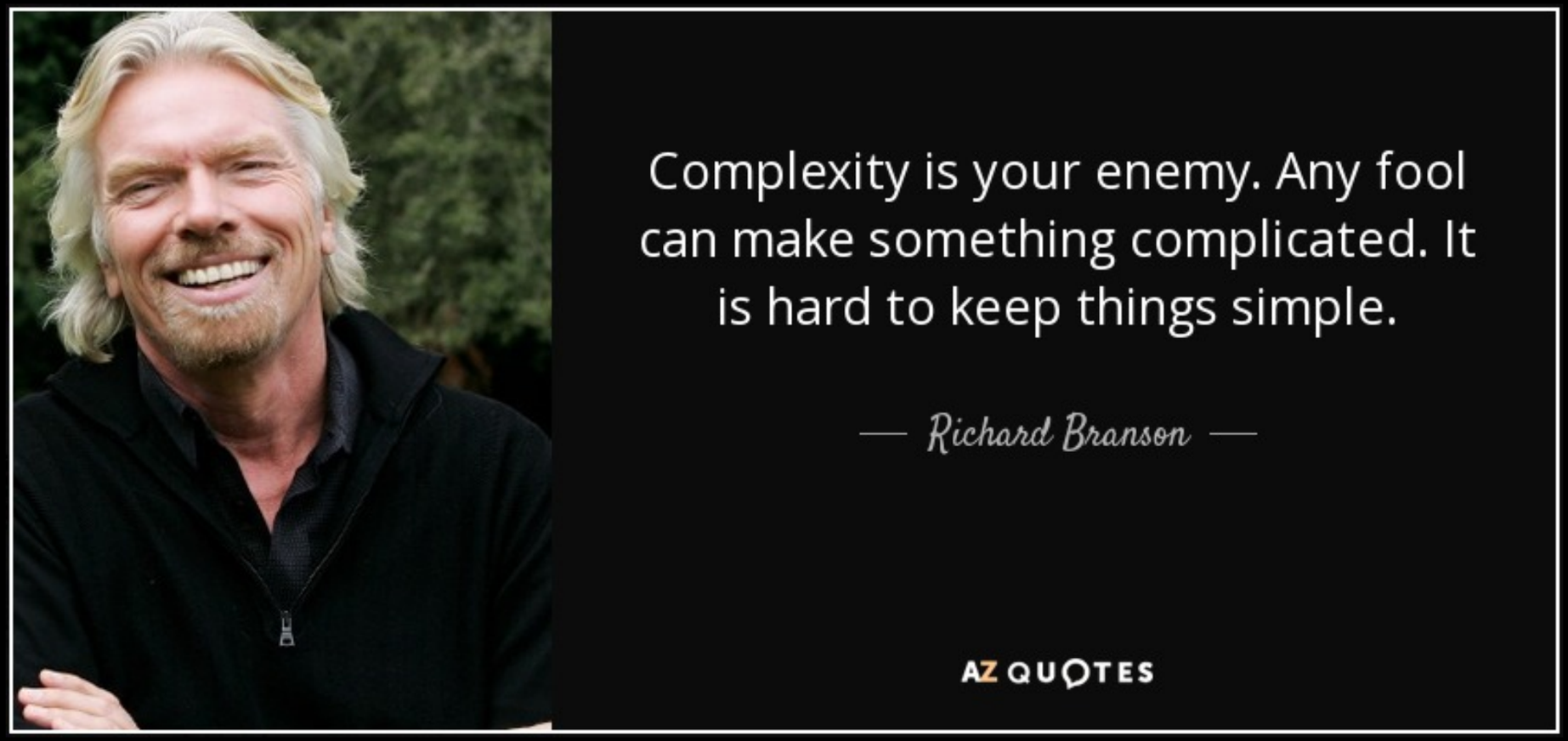
- Talk 1: Introduction to H_2O
  - Company and People
  - H_2O Open Source ML Platform
  - Live Demos
- Talk 2: The Next Generation of Machine Learning on Apache Spark
  - Introduction to Sparkling Water
  - Live Demos
- Talk 3: Introduction to Driverless AI
  - Introduction to Driverless AI

Company Overview

Founded	2012, Series C in Nov, 2017
Products	• Driverless AI – Automated Machine Learning • H_2O - Open Source Machine Learning Platform • H2O4GPU - Lightning Fast Machine Learning on GPUs • Sparkling Water - Integration of H_2O and Apache Spark
Mission	Democratize AI. Do Good.
Team	~100 employees • Distributed Systems Engineers doing Machine Learning • World-class visualization designers
Offices	Mountain View, London, Prague

H_2O.ai

Our Mission:
Make Machine Learning Accessible to Everyone



Scientific Advisory Council

Dr. Trevor Hastie

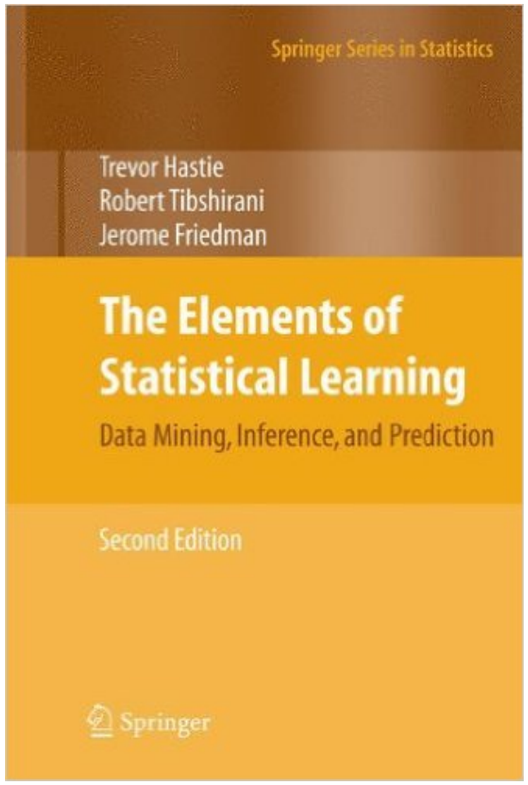
- John A. Overdeck Professor of Mathematics, Stanford University
- PhD in Statistics, Stanford University
- Co-author, *The Elements of Statistical Learning: Prediction, Inference and Data Mining*
- Co-author with John Chambers, *Statistical Models in S*
- Co-author, *Generalized Additive Models*

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*




Amy Vu-Tran


Angela Bartz


Anmol Bal


Apoorv Chaudhary


Arno Candel


Ashrith Barthur


Leland Wilkinson


Magnus Stensmo


Maral Mandjian


Marios Michailidis


Mark Chan


Mark Landry


Avni Wadhwa


Benjamin Campbell


Branden Murray


Cari Andrews


Chandan Maruthi


Cheri Poff


Mateusz Dymczyk


Mathias Müller


Matt Dowle


Megan Kurka


Michal Kurka


Michal Malohlava


Das Narayanan


David Crawford


Dmitry Larko


Eric Gudgion


Erin LeDell


Gregory Kanevsky


Michal Raška


Monika Mullerova


Navdeep Gill


Nidhi Mehta


Nikhil Shekhar


Nishant Kalonia


Ian Gomez


Jacqueline Scott


Jag Bining


Jakub Hava

Jan Gamec

Jeff Fohl

Ondrej Bilek

Pasha Stetsenko

Patrick Aboyoun

Patrick Hall

Patrick Rice

Prithvi Prabhu

Jeff Gambera

Jo-Fai Chow

John Schiess

Jon Mckinney

Jon Oslowski

Jorge Luis Hernandez Villapol

Priya Ravindhran

Raji

Ravi Purushotma

Rosalie Bartlett

Sanjay Joshi

Sirisha Bavireddy

Josephine Wang

Justin Loyola

Karen Hayrapetyan

Kathy Lee

Kristina Arabo

Lauren DiPerna

SriSatish Ambati

Terry Tang

Tom Kraljevic

Venkat Swaminathan

Venkatesh Yadav

Vinod Iyengar

Wen Phan

Wendy Wong

H2O Team

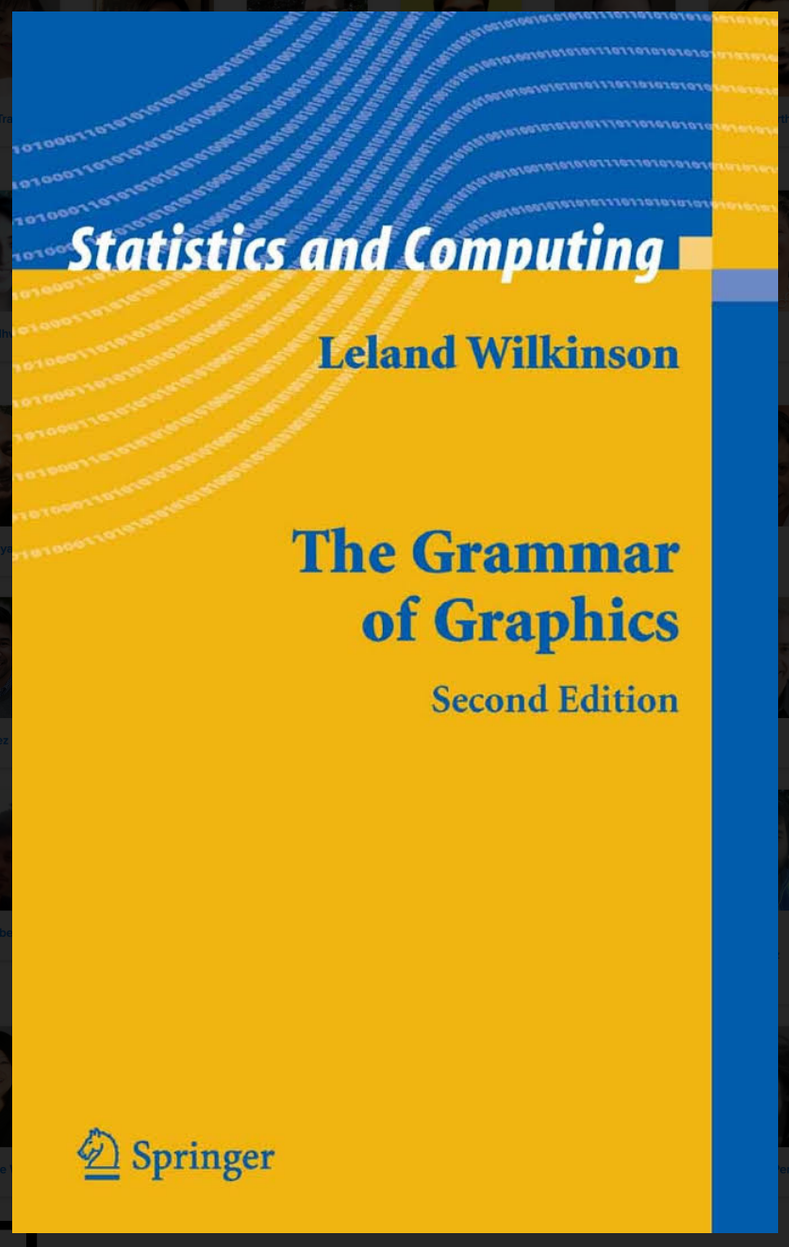
H2O.ai

Arno Candel, CTO
Fortune's 2014 Big Data All-Star

Sri Ambati, Co-founder & CEO

H_2O Team

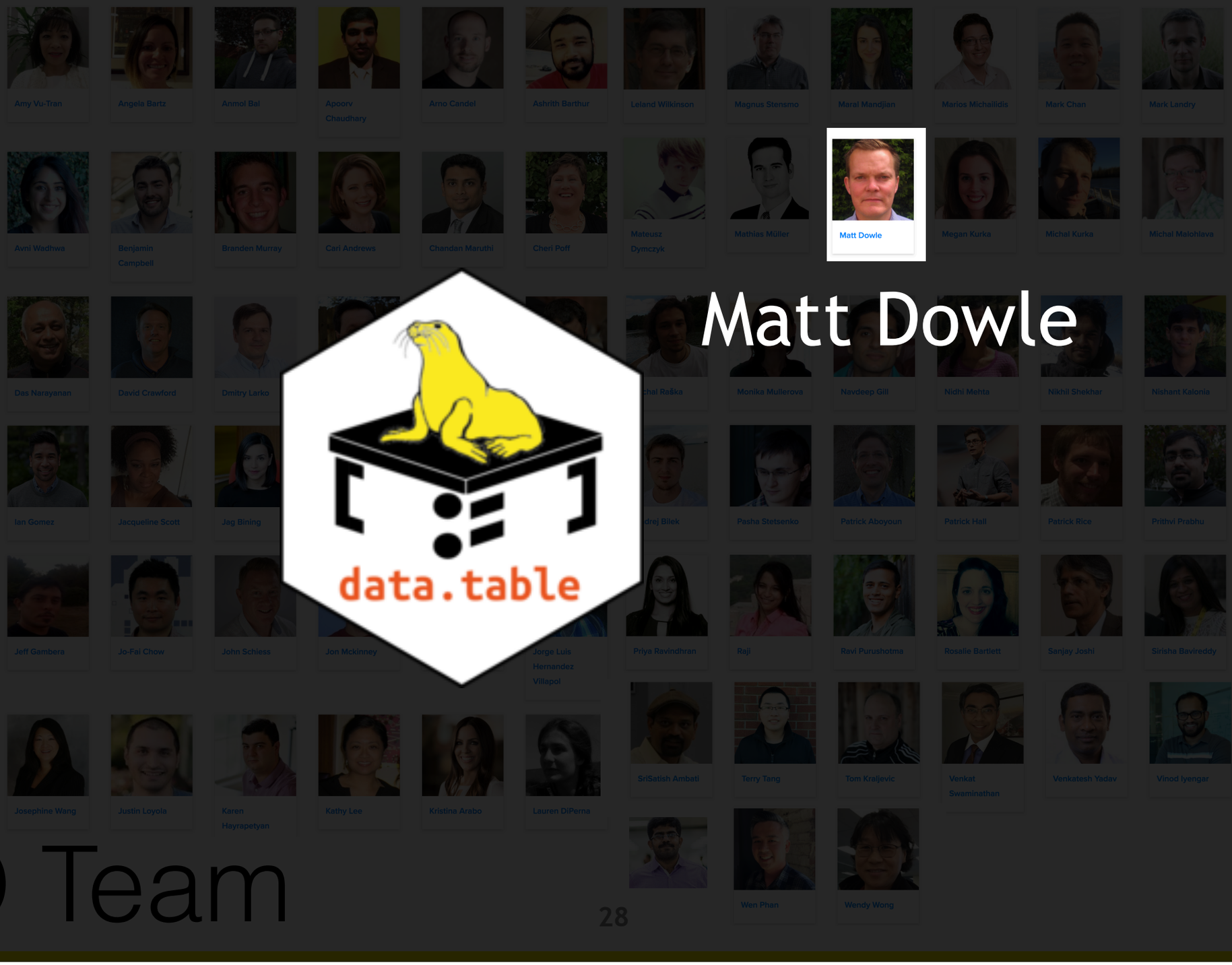
H_2O.ai



Leland Wilkinson

Origin of R Package `ggplot2`

H_2O Team



Matt Dowle

H_2O Team

H_2O.ai

H_2O Products

In-Memory, Distributed Machine Learning Algorithms with H2O Flow GUI

H2O AI Open Source Engine Integration with Spark

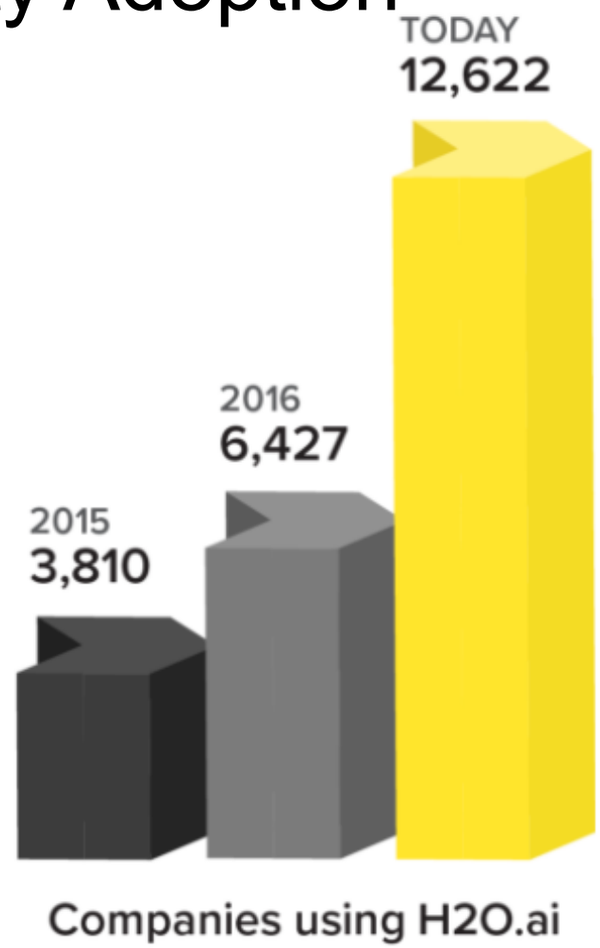
Lightning Fast machine learning on GPUs

Automatic feature engineering, machine learning and interpretability

Steam

Secure multi-tenant H2O clusters

Worldwide Community Adoption



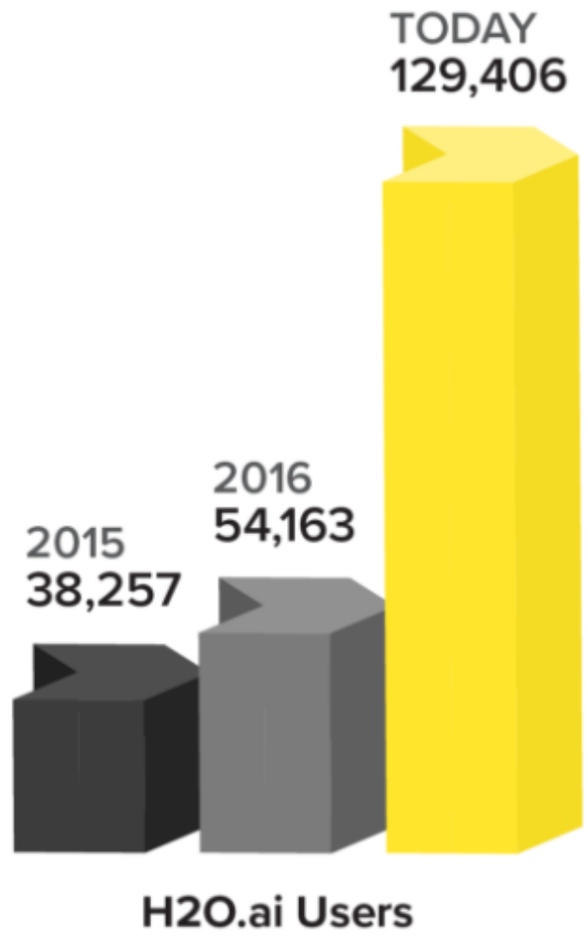
222 OF THE FORTUNE 500

H_2O

8 OF TOP 10 BANKS

7 OF TOP 10 INSURANCE COMPANIES

4 OF TOP 10 HEALTHCARE COMPANIES



* DATA FROM GOOGLE ANALYTICS EMBEDDED IN THE END USER PRODUCT

H2O.ai Solution Leadership Across Verticals

H2O.ai

Gartner names H2O as **Leader** with the **most completeness of vision**

- H2O.ai recognized as a **technology leader with most completeness of vision**
- H2O.ai was recognized for the mindshare, partner network and status as a **quasi-industry standard** for machine learning and AI.
- **H2O customers gave the highest overall score** among all the vendors for sales relationship and account management, customer support (onboarding, troubleshooting, etc.) and overall service and support.

Figure 1. Magic Quadrant for Data Science and Machine-Learning Platforms

Source: Gartner (February 2018)

H2O.ai

Platforms with H_2O integration

H_2O + KNIME Talk at KNIME Summit Mar 2017

Figure 1. Magic Quadrant for Data Science and Machine-Learning Platforms

Source: Gartner (February 2018)

H_2O.ai

Community Expansion

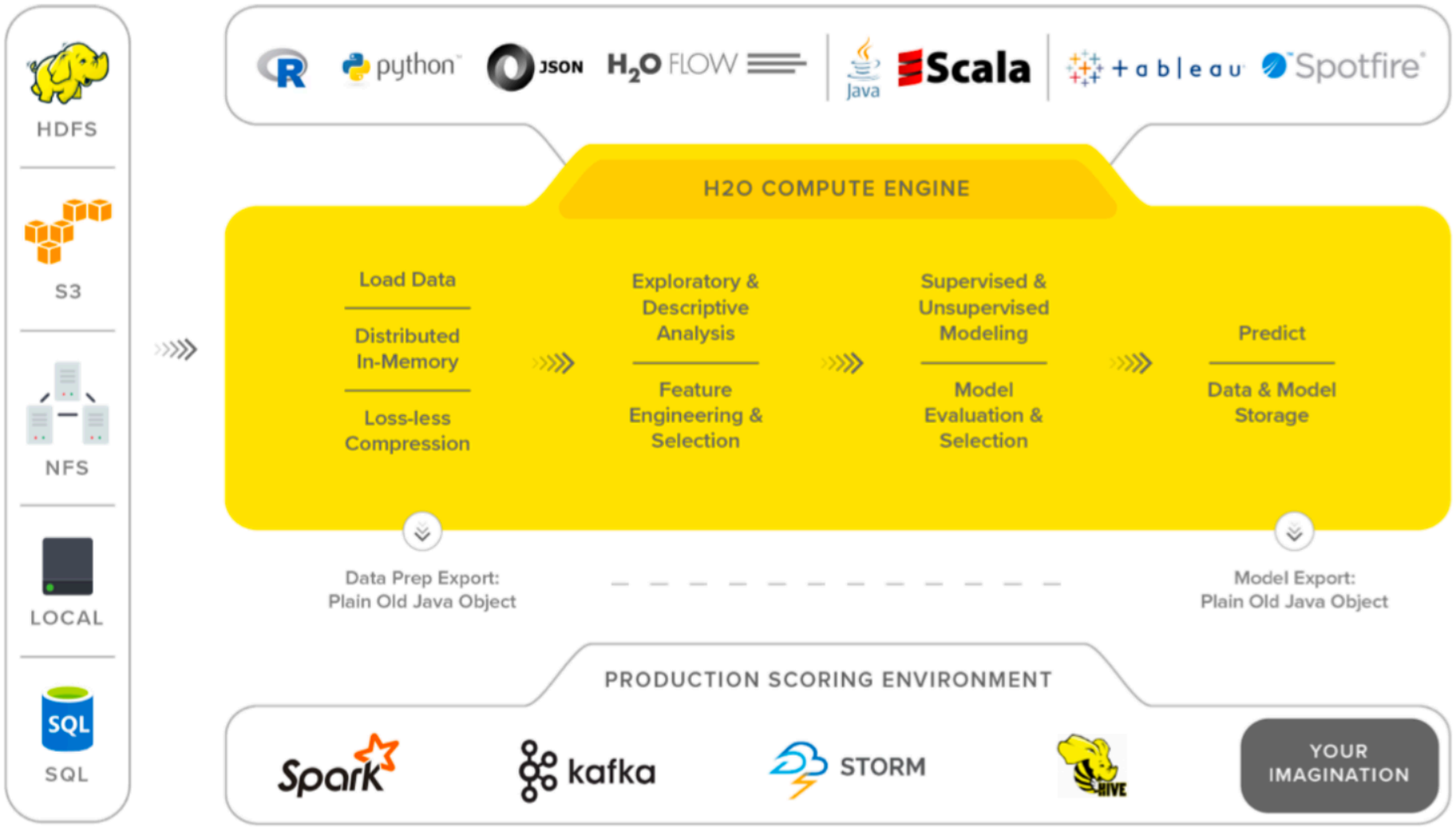
Meetup

88,286	43	52	48	18
members	interested	Meetups	cities	countries

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

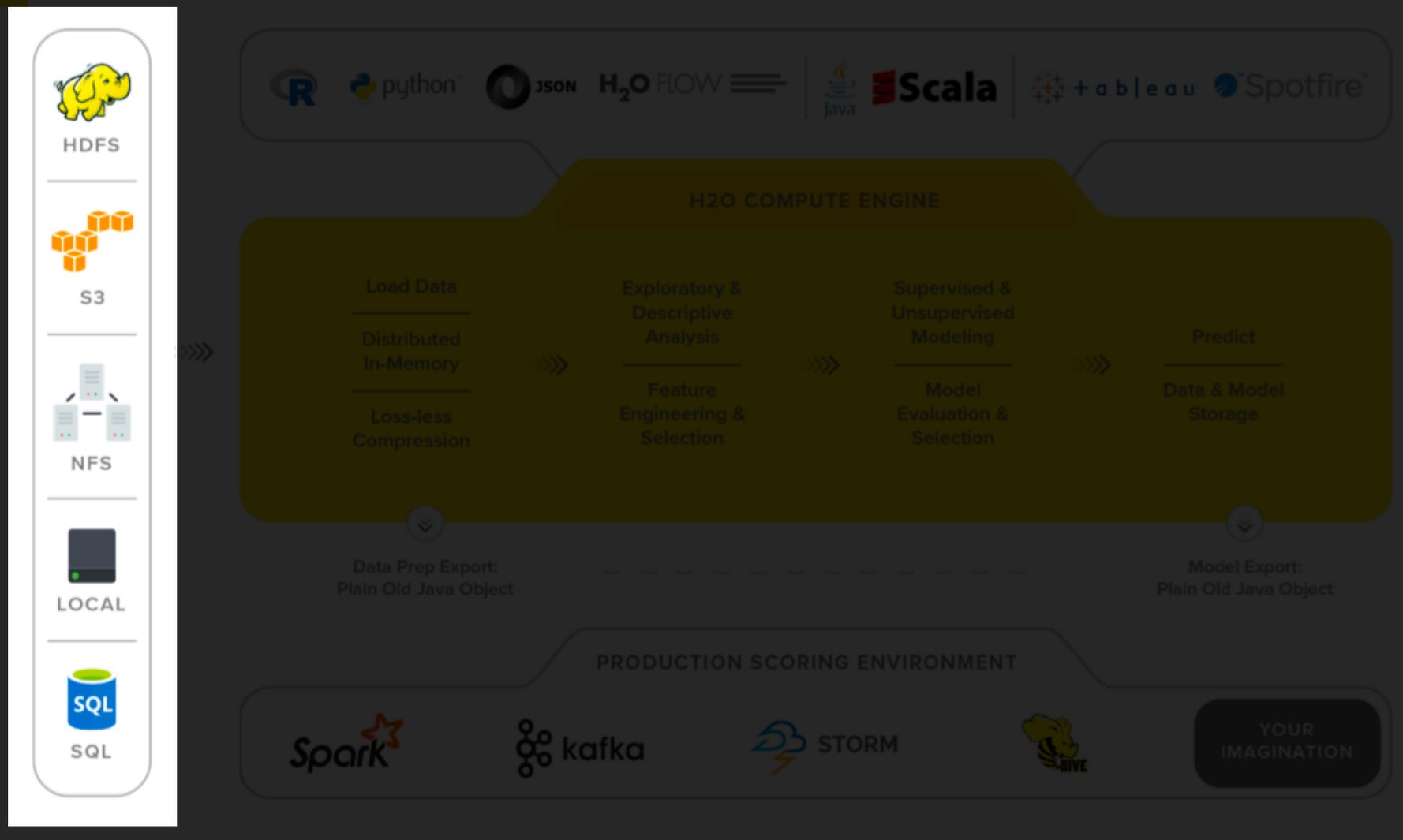
H_2O Machine Learning Platform

High Level Architecture



High Level Archite

Import Data from Multiple Sources

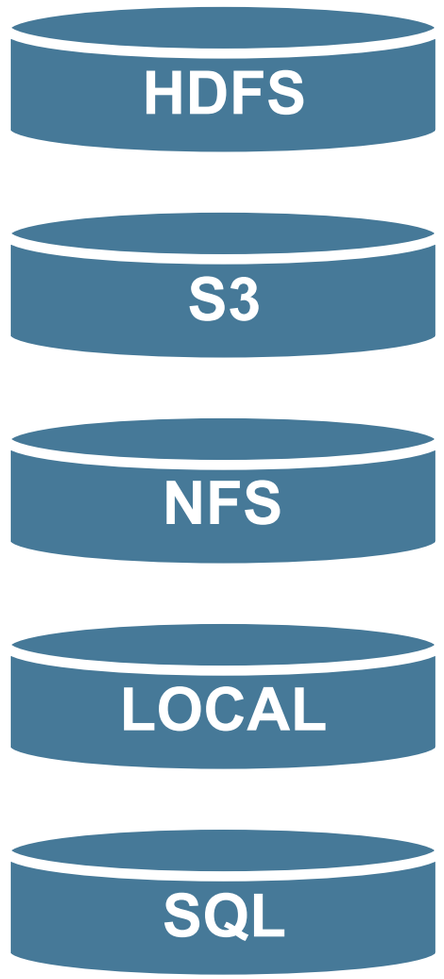


Supported Formats & Data Sources

9 Formats

File type or Folder of Files

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

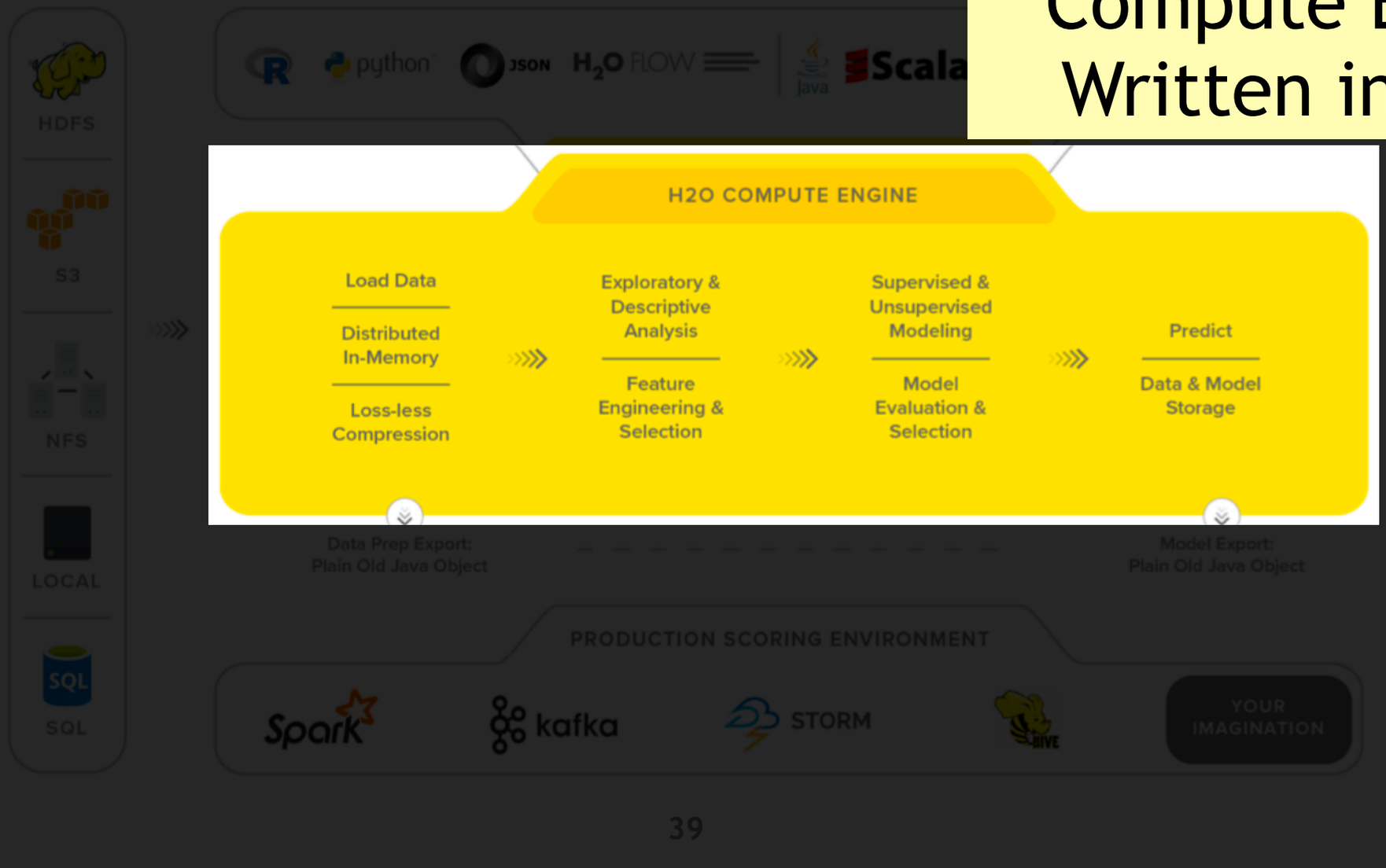


5 Sources

* 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

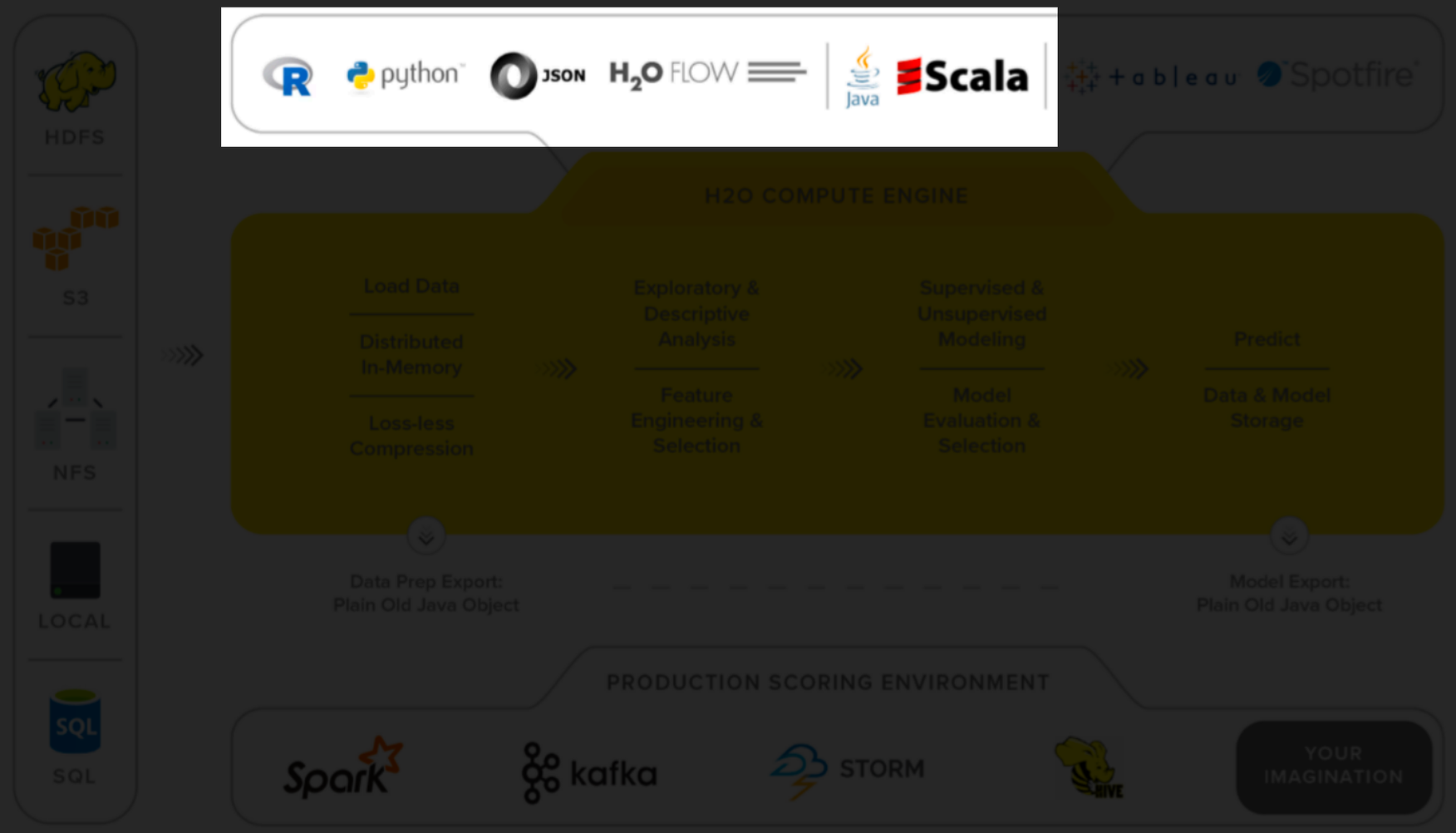
High Level Archit

Fast, Scalable & Distributed Compute Engine Written in Java



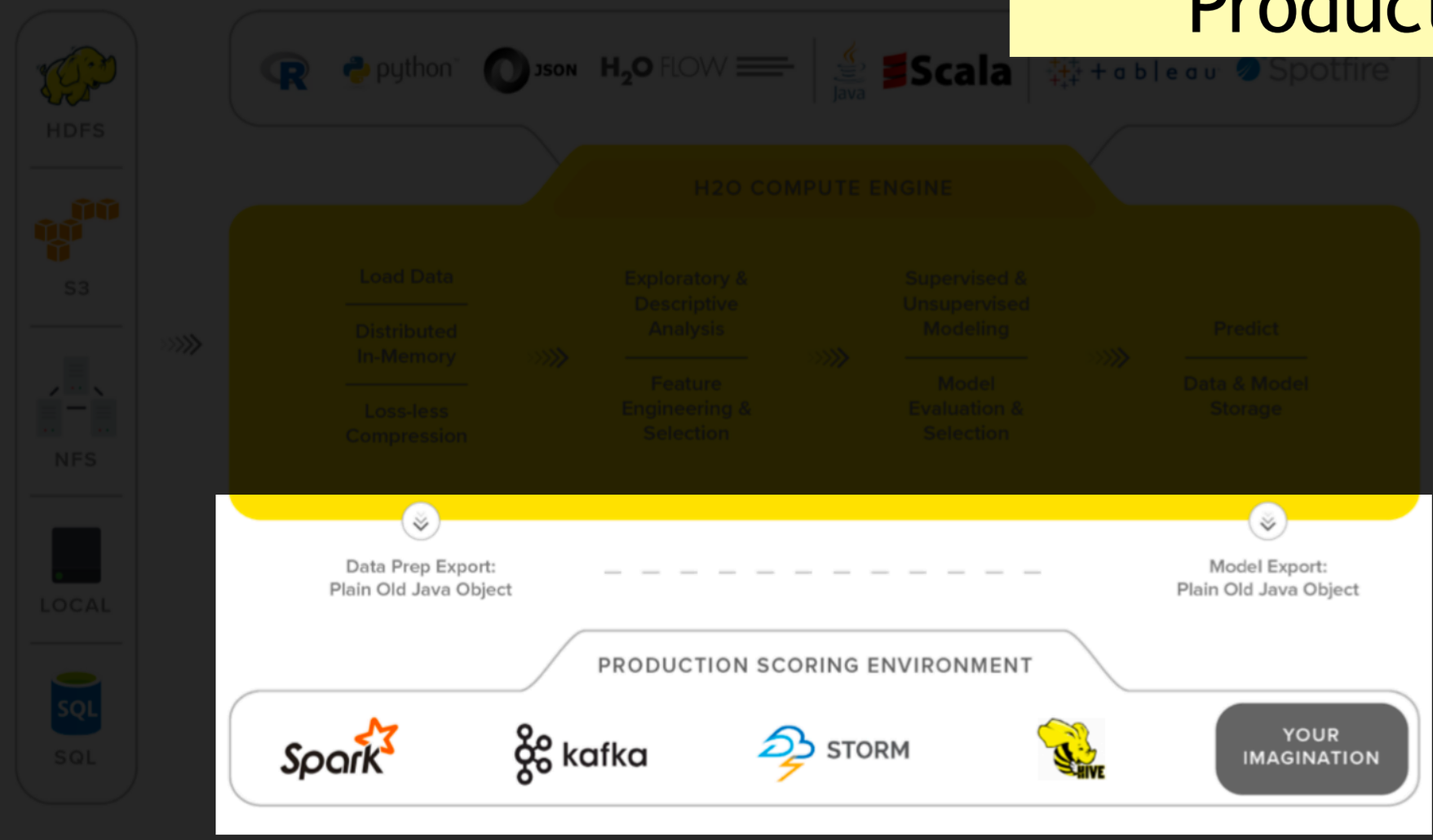
High Level Architecture

Multiple Interfaces



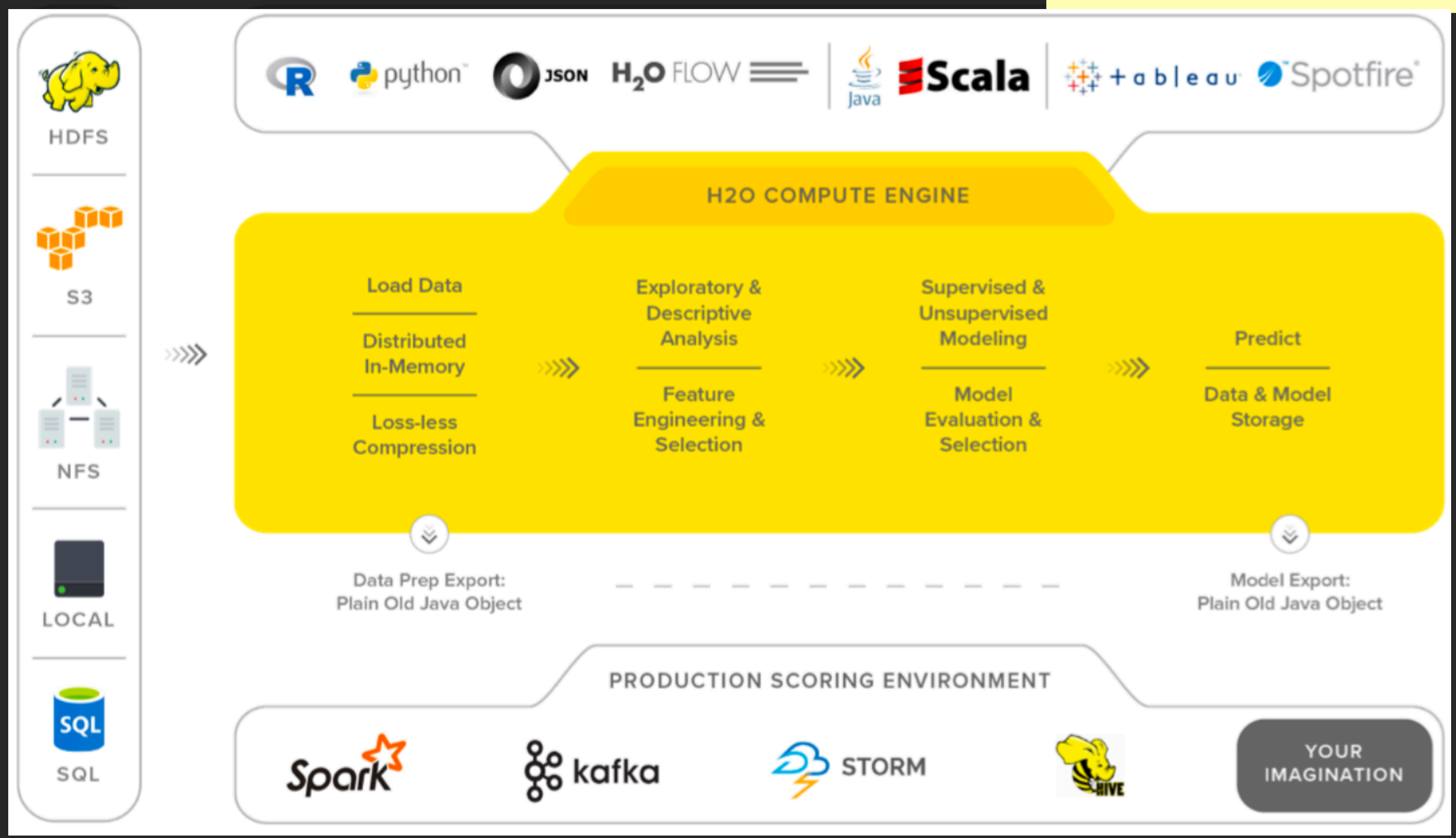
High Level Archit

Export Standalone Models for Production



High Level Archit

H_2O - Open Source Machine Learning Platform



Distributed Algorithms

Foundation for Distributed Algorithms

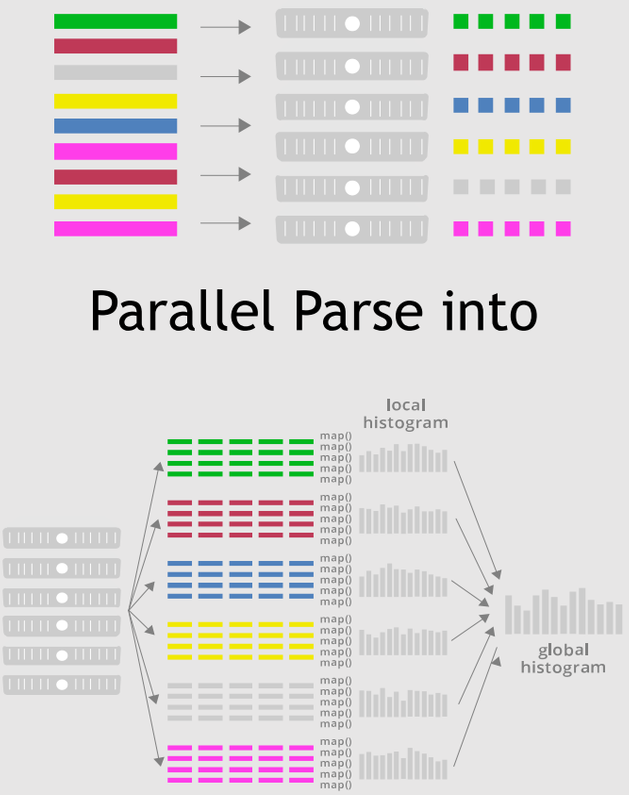


Illustration: Scalable

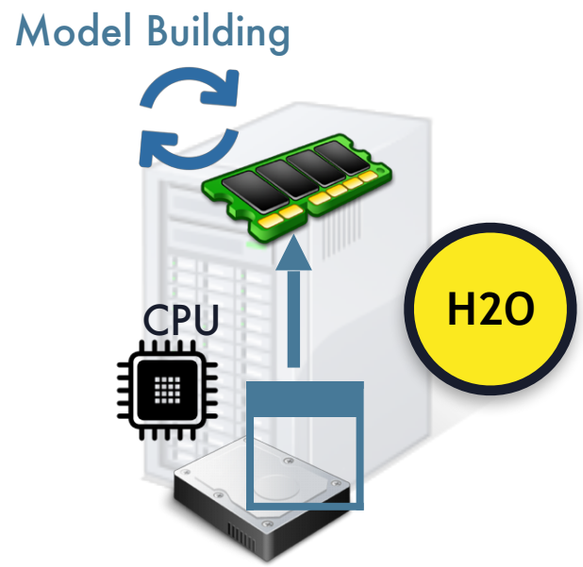
Advantageous Foundation

- 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

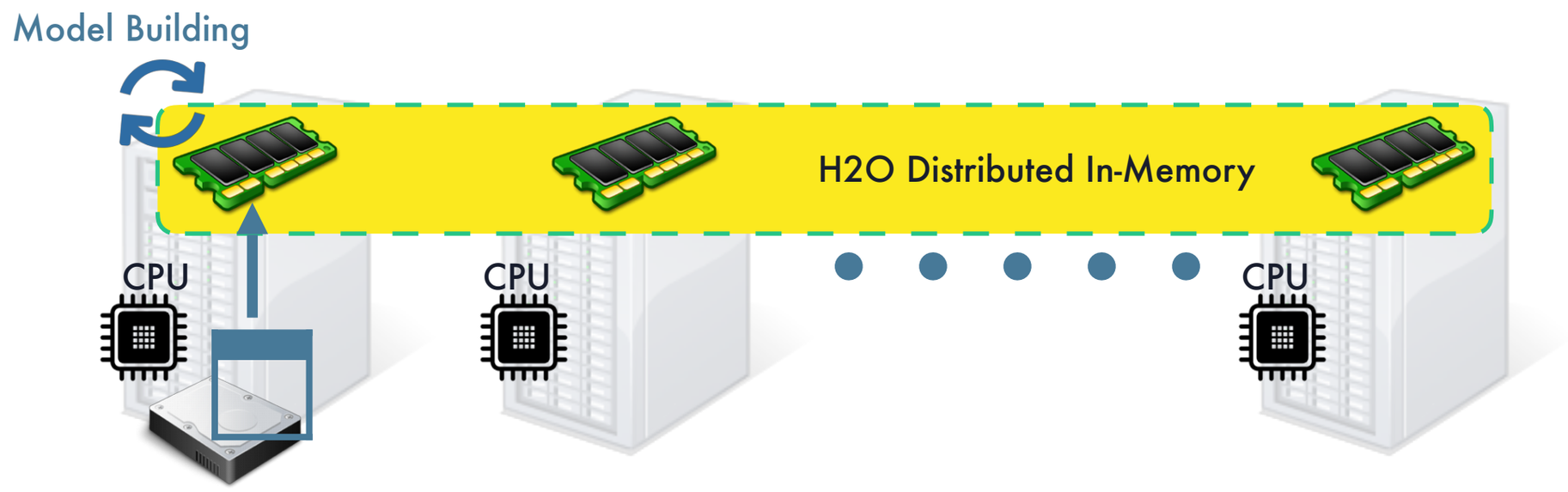
- "Out-of-box" functionalities for all algorithms **(NO MORE SCRIPTING)** and uniform interface across all languages: R, Python, Java
- **Designed for all sizes of data sets, especially large data**
- **Highly optimized Java code for model exports**
- **In-house expertise for all algorithms**

H_2O Core

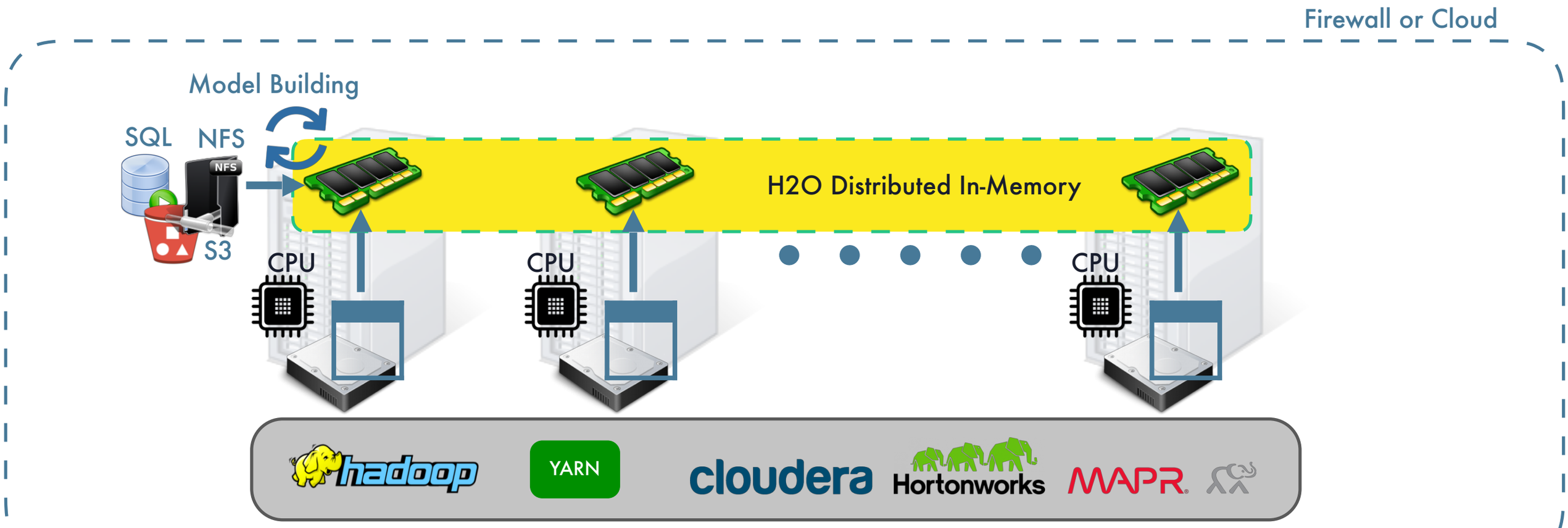


H_2O Core

H_2O Core



H_2O Core


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

H_2O-3 Algorithms Overview

Supervised Learning

Statistical Analysis

- **Generalized Linear Models**: Binomial, Gaussian, Gamma, Poisson and Tweedie
- **Naïve Bayes**
- **Distributed Random Forest**: Classification or regression models

Ensembles

- **Gradient Boosting Machine**: Produces an ensemble of decision trees with increasing refined approximations

Deep Neural Networks

- **Deep earning**: 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

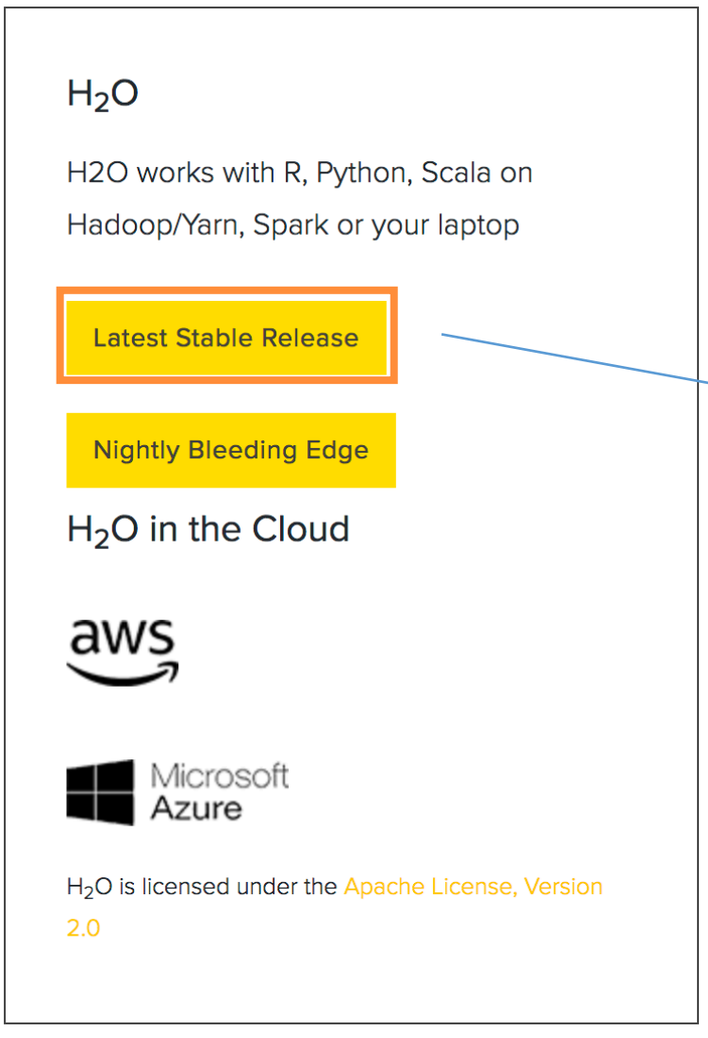
- **Principal Component Analysis**: Linearly transforms correlated variables to independent components
- **Generalized Low Rank Models**: extend the idea of PCA to handle arbitrary data consisting of numerical, Boolean, categorical, and missing data

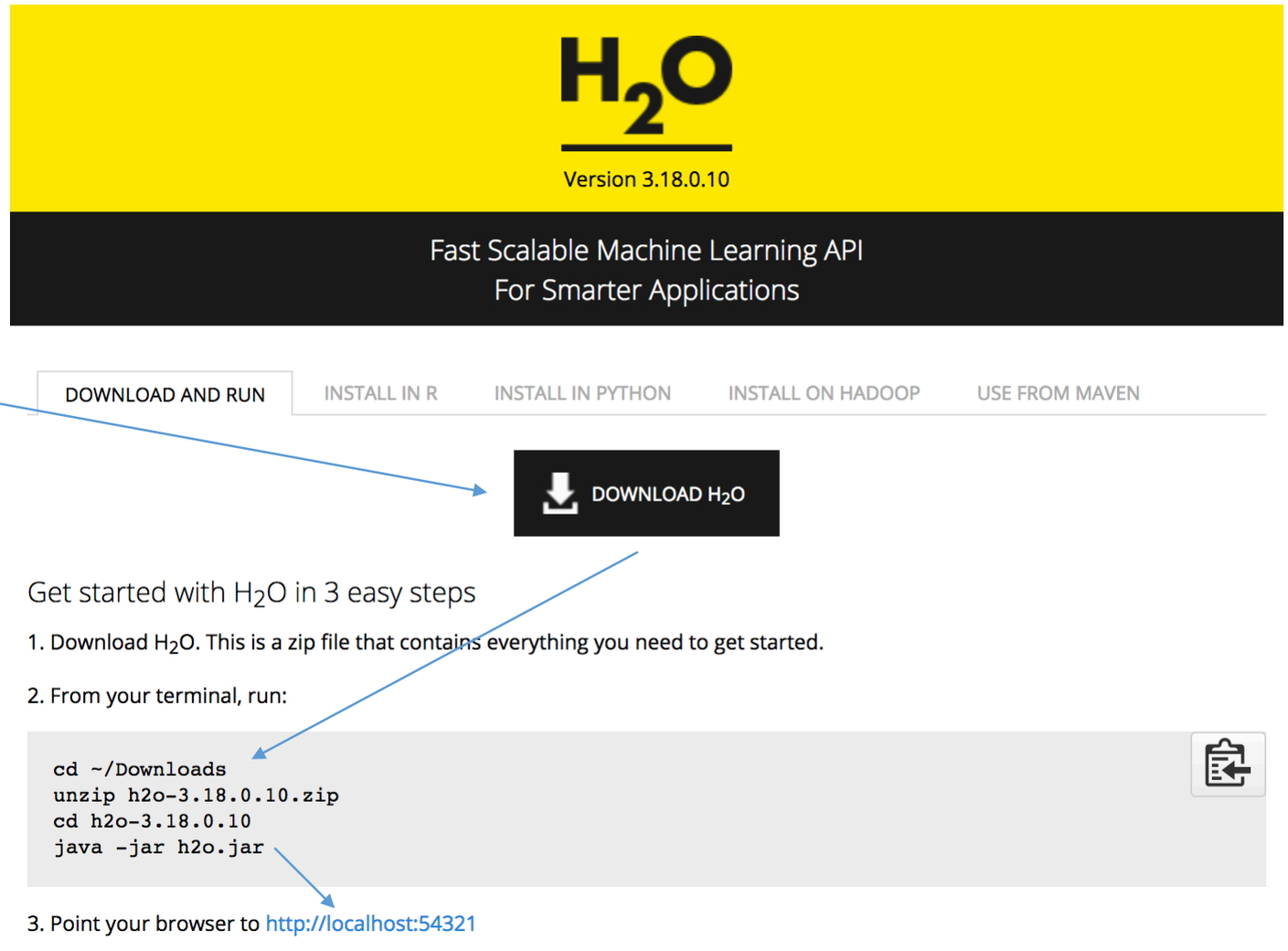
Anomaly Detection

- **Autoencoders**: Find outliers using a nonlinear dimensionality reduction using deep learning

H_2O.ai

Downloading H_2O

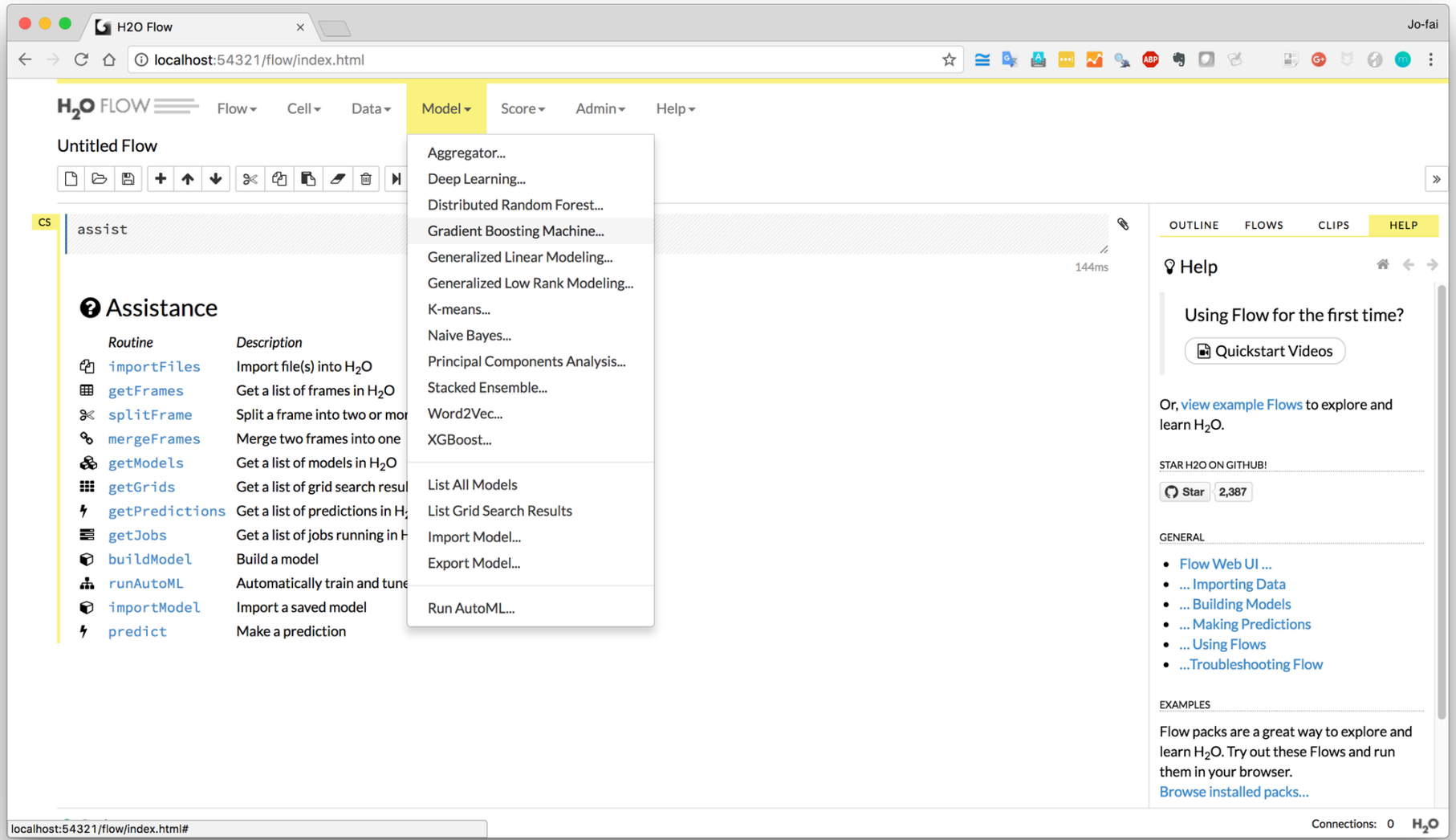




H2O.ai

h2o.ai/download/

H_2O Flow



H_2O.ai

H2O.ai Products Industries Customers Community Partners Download

H2O Documentation

Getting Started & User Guides | Q & A | Algorithms | Languages | Tutorials, Examples, & Presentations | API & Developer Docs | For the Enterprise

Getting Started & User Guides

Open Source | Commercial

H2O

What is H2O?
H2O User Guide (Main docs)
H2O Book (O'Reilly)
Recent Changes
Open Source License (Apache V2)

Quick Start Video - Flow Web UI
Quick Start Video - R
Quick Start Video - Python

Download H2O

Sparkling Water

What is Sparkling Water?
Sparkling Water User Guide 2.3 2.2 2.1
Sparkling Water Booklet
RSparkling Readme
PySparkling User Guide 2.3 2.2 2.1
Recent Changes 2.3 2.2 2.1
Open Source License (Apache V2)

Quick Start Video - Scala

Download Sparkling Water

Driverless AI

What is Driverless AI?
Driverless AI User Guide HTML PDF
Recent Changes
Driverless AI Booklet
MLI with Driverless AI Booklet

Quick Start Video - Downloading Driverless AI
Quick Start Video - Launching an Experiment
Driverless AI Webinars

Download Driverless AI

H2O4GPU (alpha)

H2O4GPU Readme
Open Source License (Apache V2)

Download H2O4GPU

docs.h2o.ai

Jakub Háva
jakub.hava@h2o.ai

The Next Generation of Machine Learning on Apache Spark

Firemní semináře @ MFF UK, Praha
October 24, 2018

H_2O Products

H_2O.ai

In-Memory, Distributed Machine Learning Algorithms with H2O Flow GUI

Spark + H_2O SPARKLING WATER

H2O AI Open Source Engine Integration with Spark

H_2O4GPU

Lightning Fast machine learning on GPUs

DRIVERLESSAI

Automatic feature engineering, machine learning and interpretability

Steam

Secure multi-tenant H2O clusters

H2O+Spark = Sparkling Water

H_2O.ai

Machine Intelligence

Sparkling Water

- Transparent integration of H2O with Spark ecosystem - MLlib and H2O side-by-side
- Transparent use of H2O data structures and algorithms with Spark API
- Excels in existing Spark workflows requiring advanced Machine Learning algorithms
- Deployment tool for Driverless AI MOJOs

Functionality missing in H2O can be replaced by Spark and vice versa

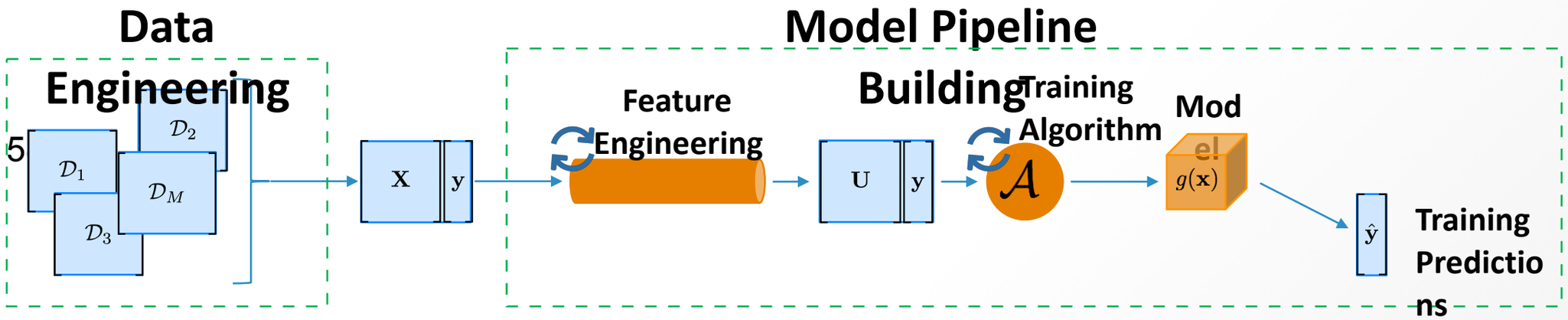
Benefits

- Additional algorithms
  - NLP
- Powerful data munging
- ML Pipelines

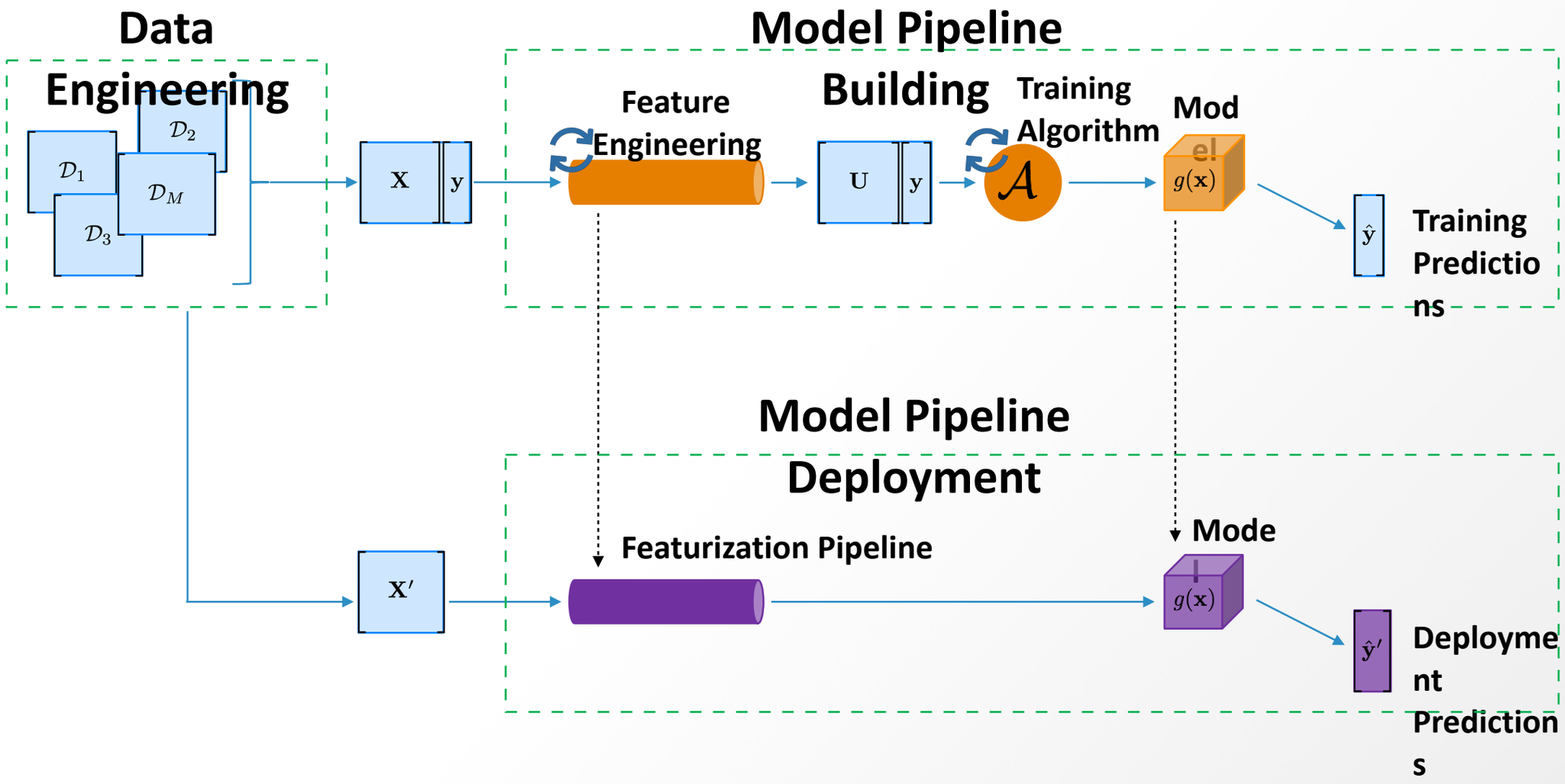
- Advanced algorithms
  - speed v. accuracy
  - advanced parameters
- Fully distributed and parallelised
- Graphical environment
- R/Python interface

How it fits to ML Life-Cycle ?

Basic ML Lifecycle



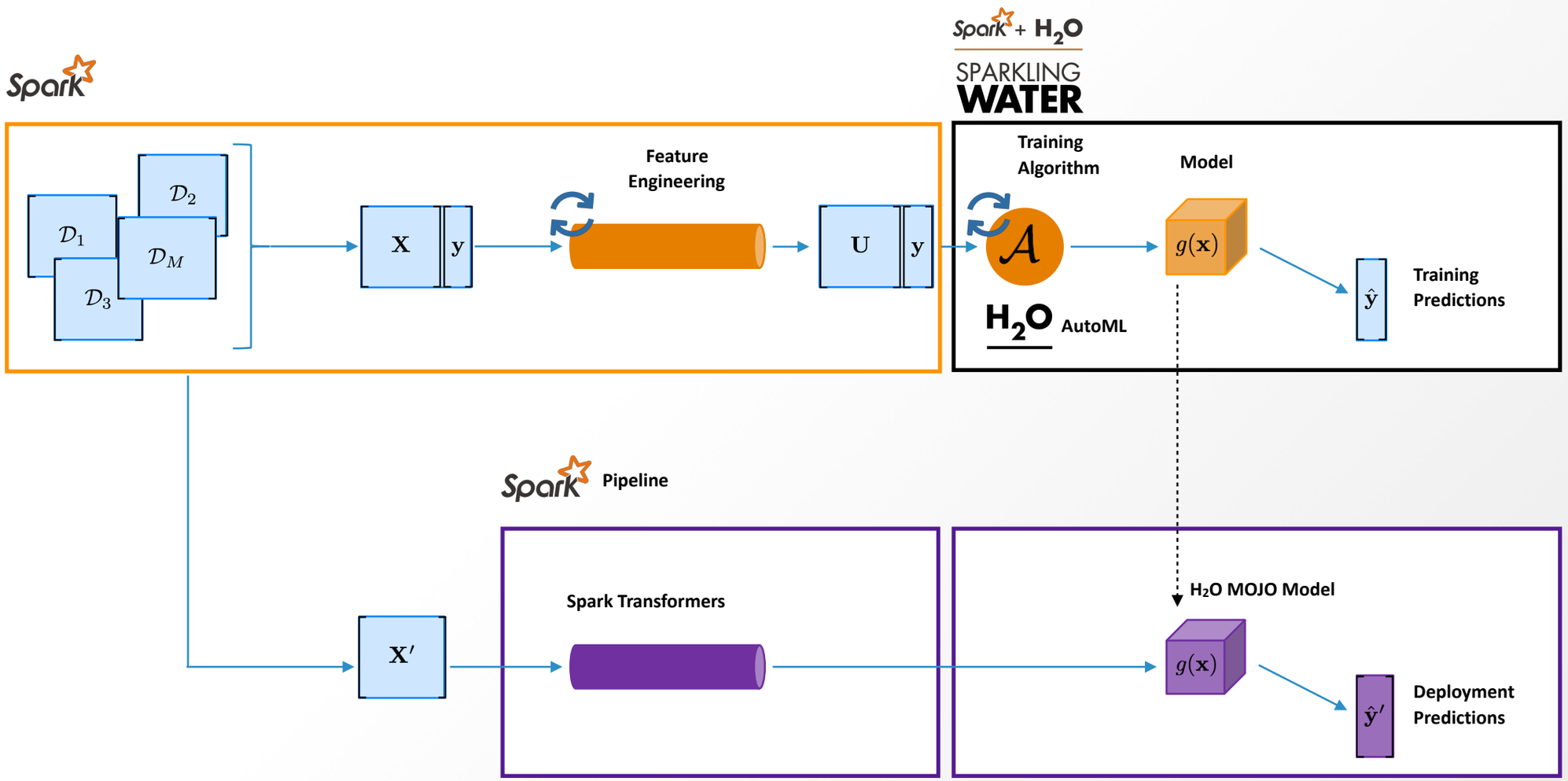
Basic ML Lifecycle



Example Implementation

Model Building			Model Deployment	
Data Engineering	**Feature Engineering**	**Training Algorithm**	**Deployment Pipeline**	**Model**
Spark		**H2O**	**Spark**	**H2O MOJO**
Spark	H2O Driverless AI		Spark	H2O Driverless AI MOJO

Basic ML Lifecycle



How to use Sparkling Water?

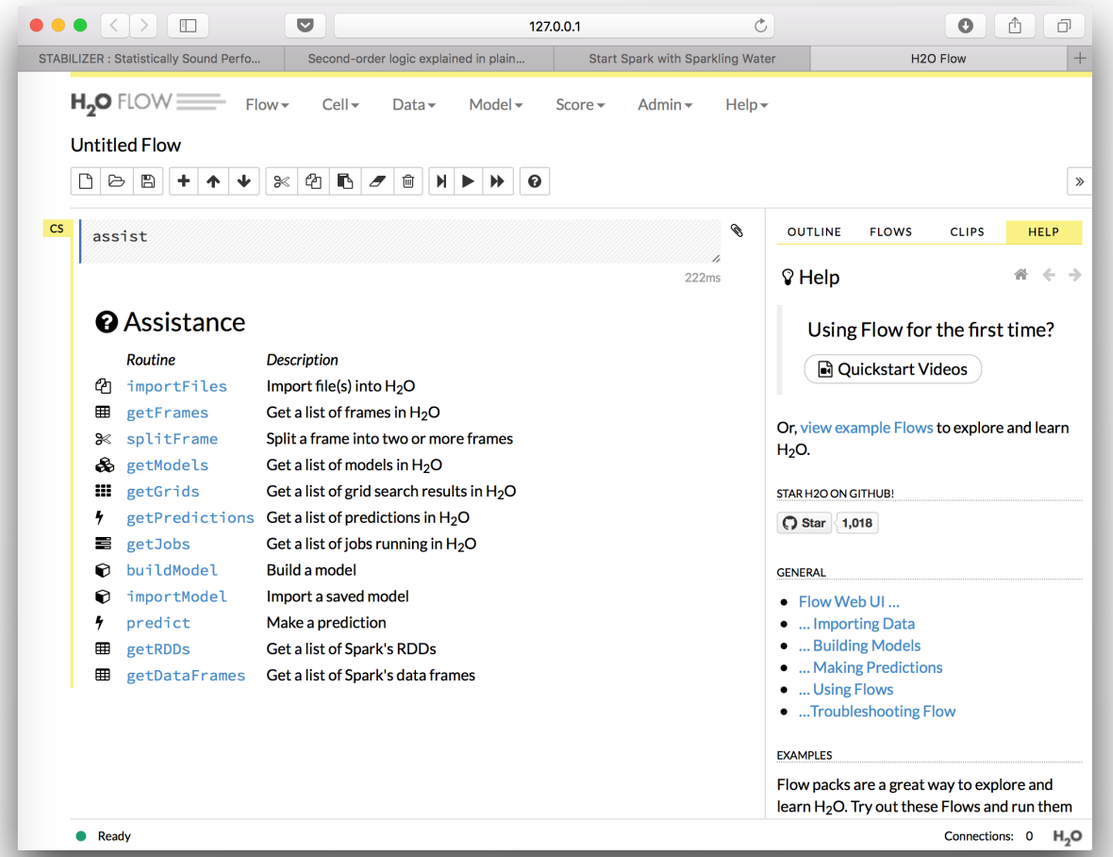
H_2O.ai

Machine Intelligence

Start spark with Sparkling Water

start.sh — Raw

```
$SPARK_HOME/bin/spark-submit \
--class water.SparklingWaterDriver \
--packages ai.h2o:sparkling-water-examples_2.10:1.6.3 \
--executor-memory=6g \
--driver-class-path scalastyle.jar /dev/null
```

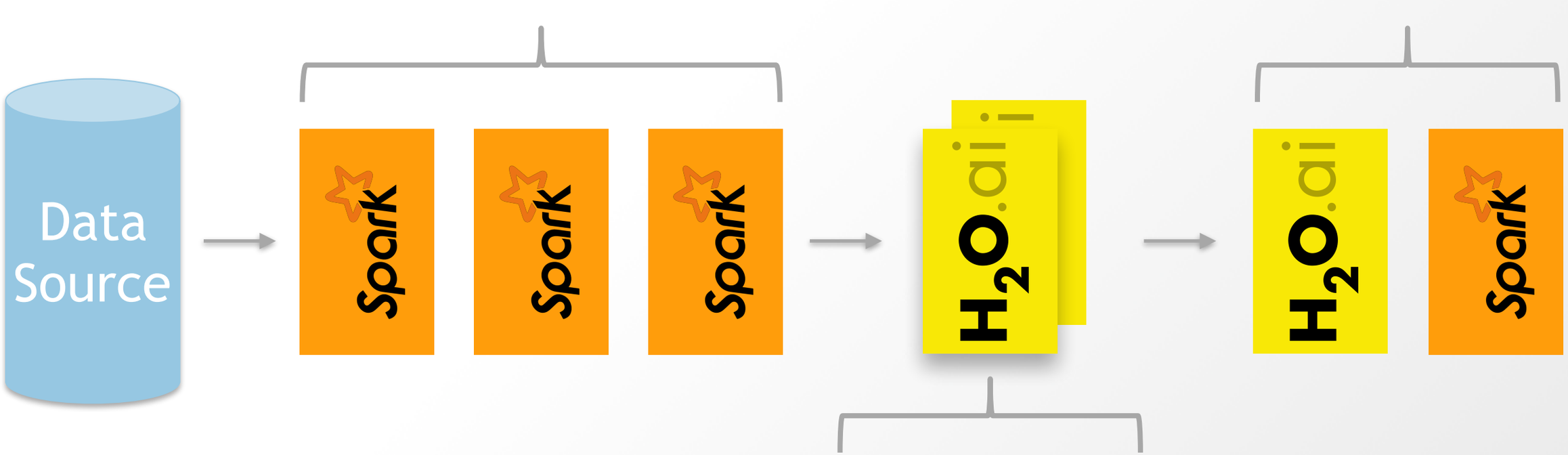


Model Building

Data munging

Modelling

Prediction processing

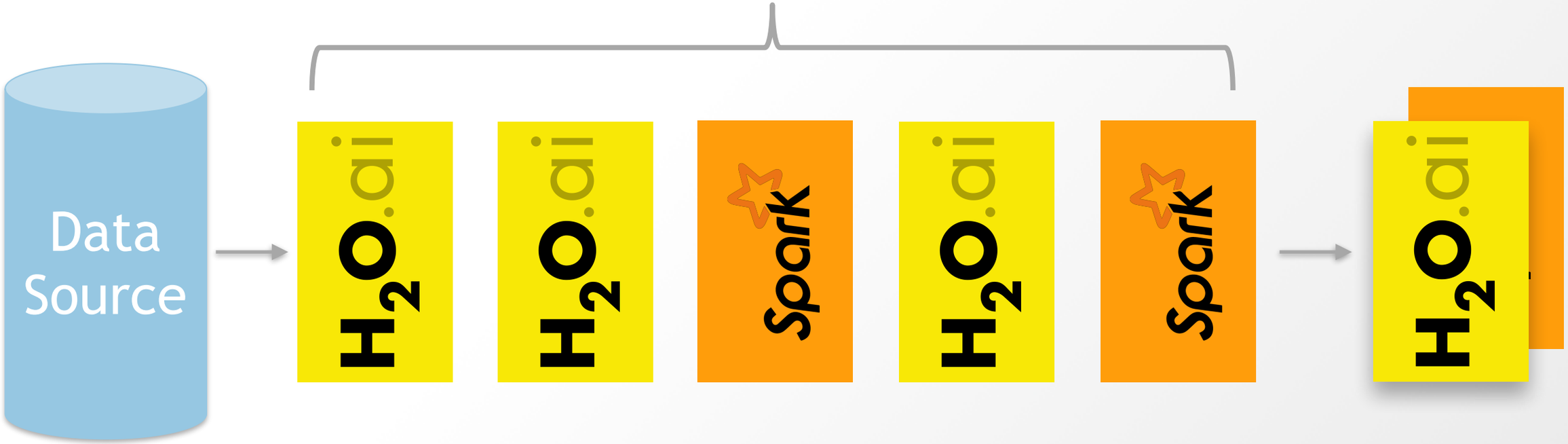


Deep Learning, GBM
DRF, GLM, GLRM
K-Means, PCA
CoxPH, Ensembles

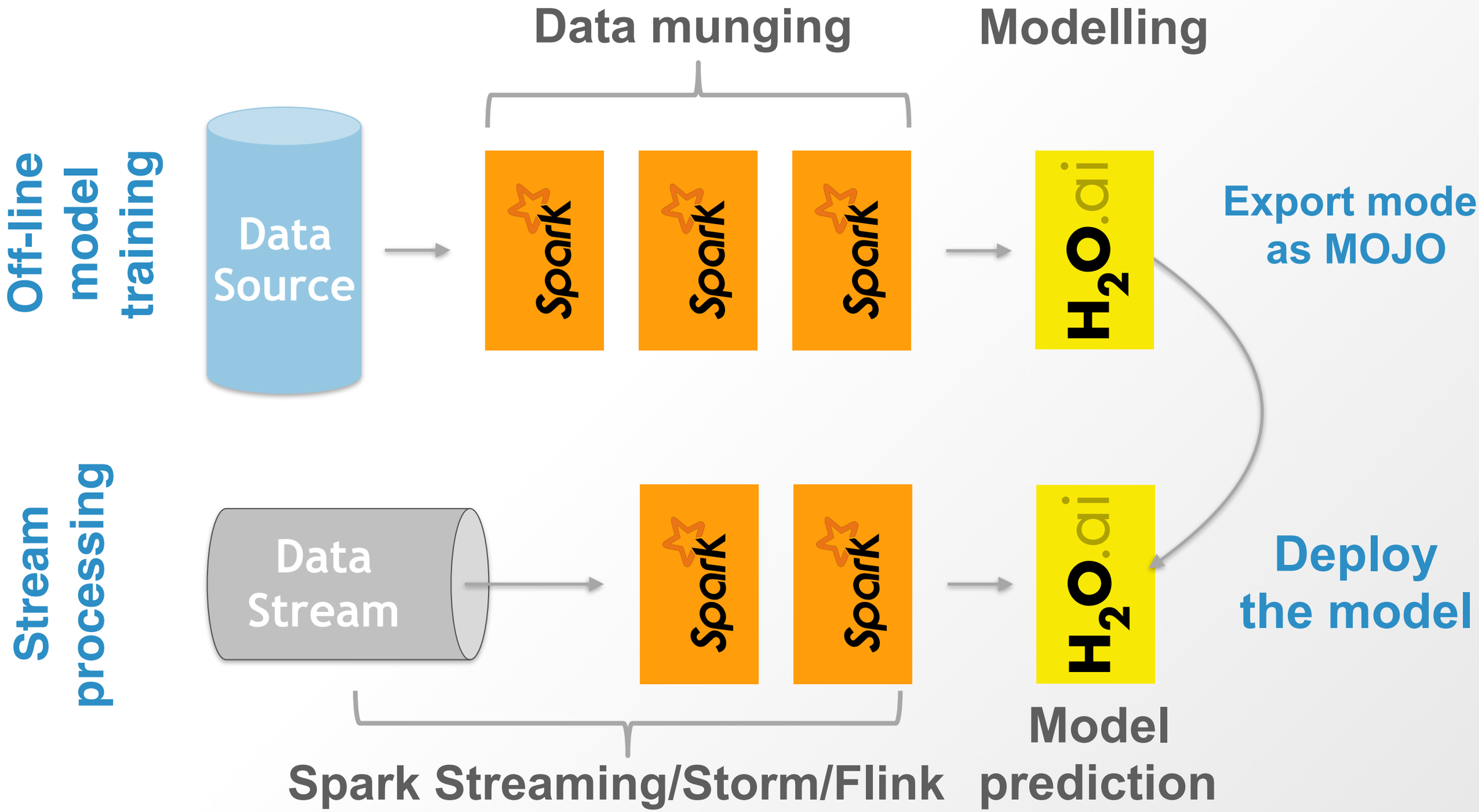
Data Munging

Data load/munging/
exploration

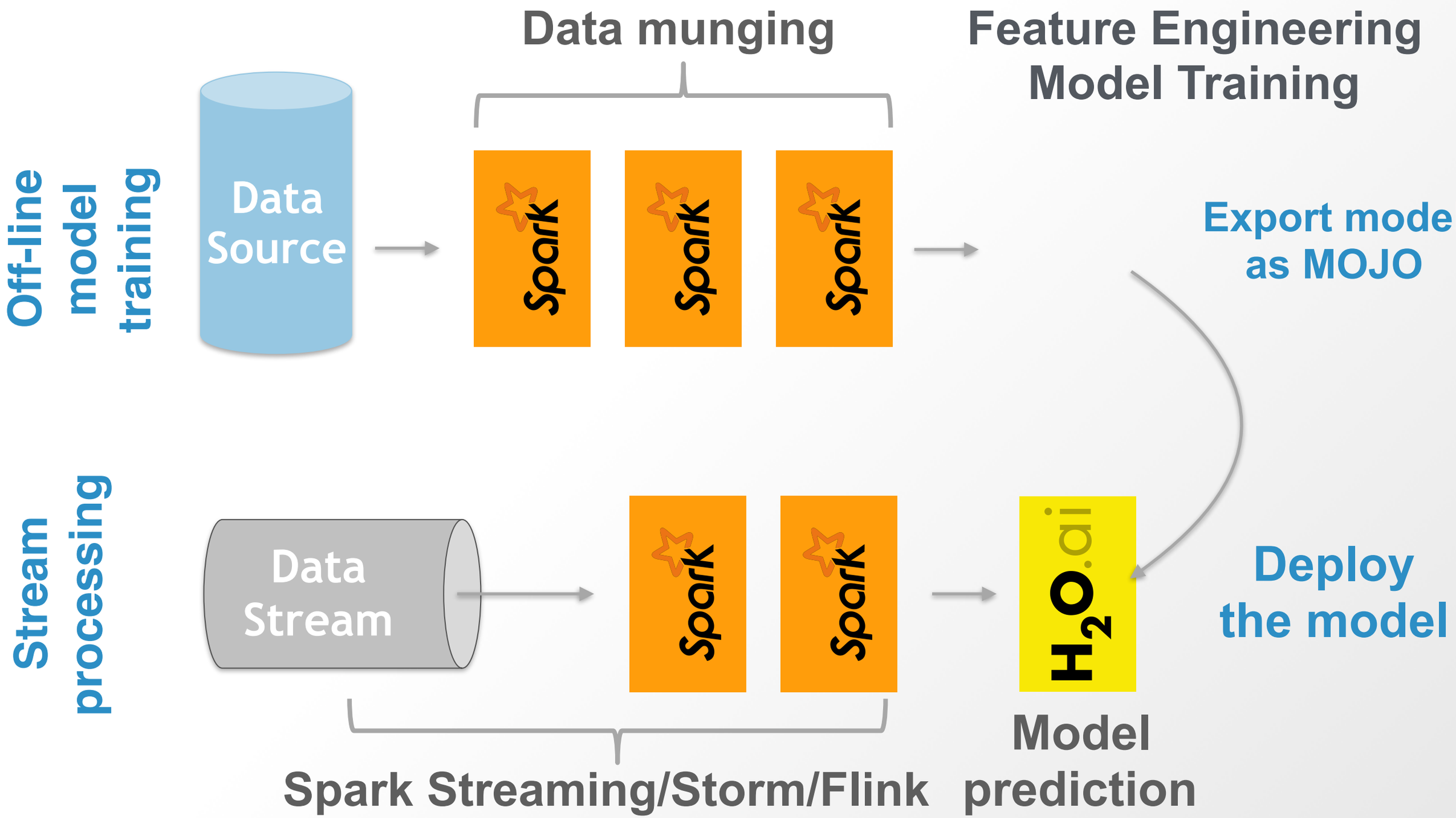
Modelling



Stream Processing



Stream Processing 2



Scoring

- POJO
  - Plain Old Java Object
- MOJO
  - Model Object Optimised
- MOJO Pipeline
  - MOJO from DAI with additional future transformations
- No runtime dependency on H2O/Driverless AI frameworks

Features

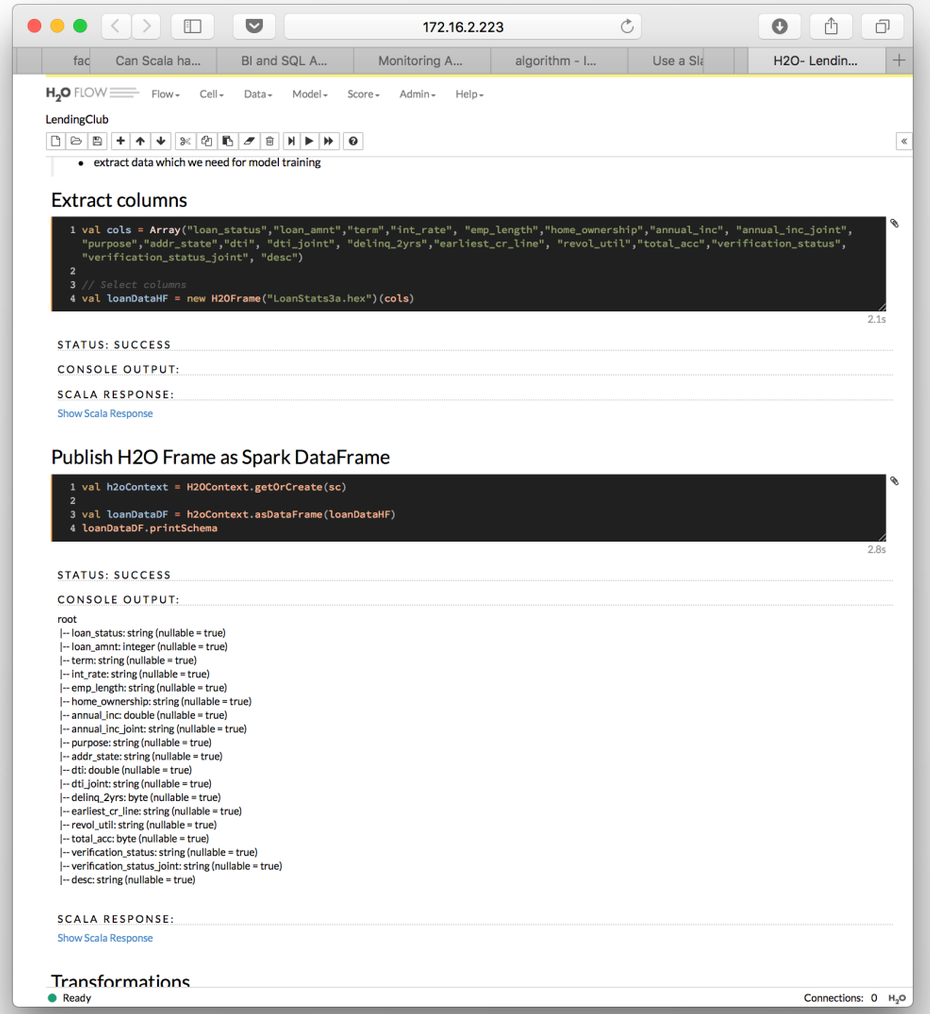
Overview

H_2O.ai

Machine Intelligence

Scala code in H2O Flow

- New type of cell
- Access Spark from Flow UI
- Experimenting made easy



H2O Frame as Spark's Datasource

- Use native Spark API to load and save data
- Spark can optimise the queries when loading data from H2O Frame
- Use of Spark query optimiser

Machine learning pipelines

- Wrap our algorithms as Transformers and Estimators
- Support for embedding them into Spark ML Pipelines
- Can serialise fitted/unfitted pipelines
- Unified API => Arguments are set in the same way for Spark and H2O Models
- Integration with Mojo pipelines

PySparkling

- PySparkling is on PyPi
- Contains all H2O and Sparkling Water dependencies, no need to worry about them
- Just add in on your Python path and that's it
- Correctly specifies dependency on PySpark

RSparkling

- Sparkling Water for R
- Based on SparklyR package
- Independent on Spark and Sparkling Water version

And others!

- Support for Datasets
- Zeppelin notebook support
- XGBoost Support in Distributed Mode
- Support for high-cardinality fast joins
- Secure Communication - SSL
- Support for Sparse Data conversions
- External Cluster Stabilisation
- Enterprise security support - LDAP, Kerberos
- Driverless AI Pipeline deployment availability

Coming features

- Integration with Steam
  - Deploy on remises
  - Deploy to Cloud

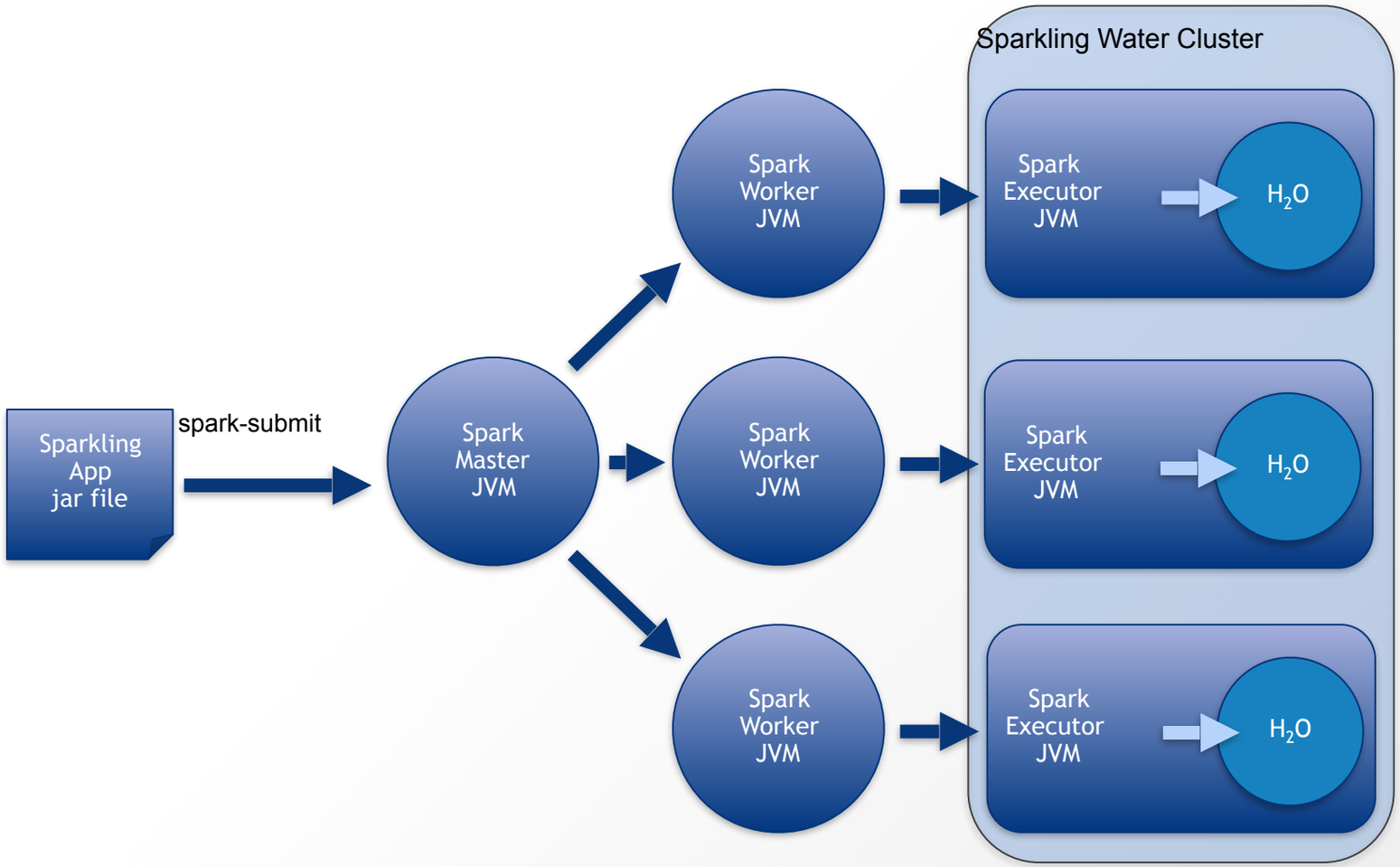
What is inside?

Two Backends

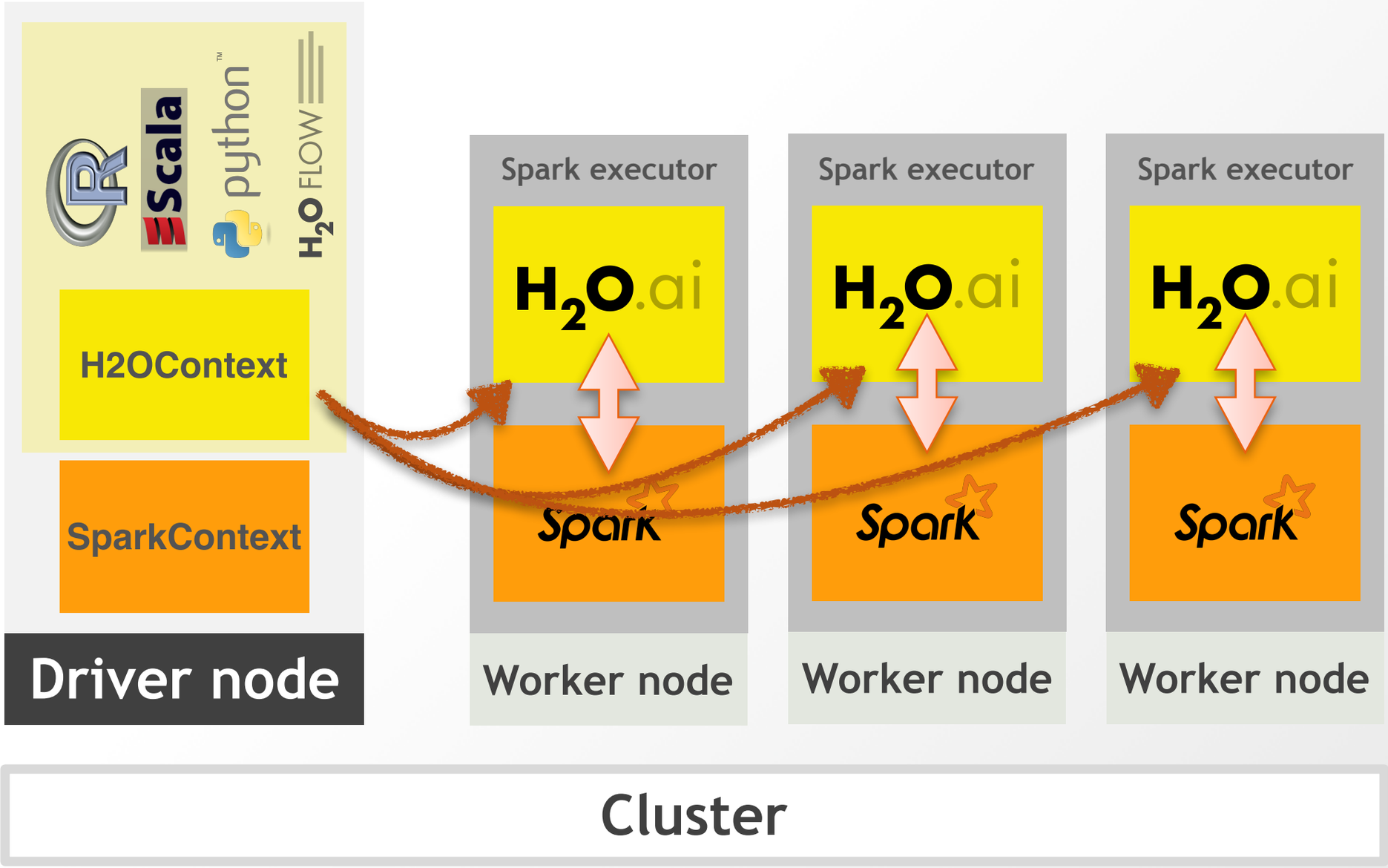
- Sparkling Water has two backends
  - External
  - Internal
- The backend determines where H2O cluster is located
- Each backend is good for different use-cases

Internal Backend

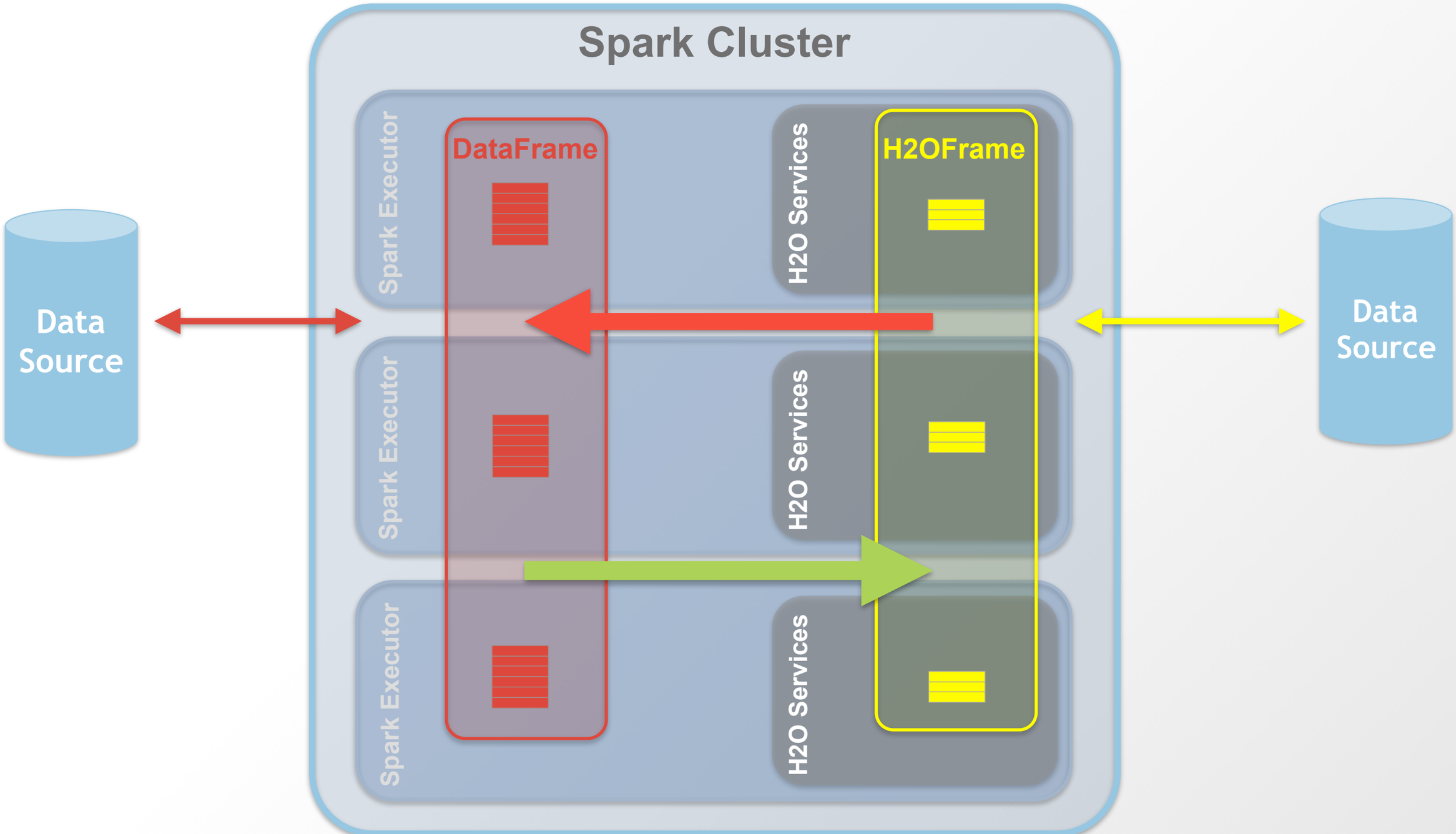
Sparkling Water Internal Backend



Internal Backend



Data Transfers



Pros & Cons

- Advantages
  - Easy to configure
  - Faster (no need to send data to different cluster)
- Disadvantages
  - Spark kills or joins new executor => H2O goes down
  - No way how to discover all executors

External Backend

Overview

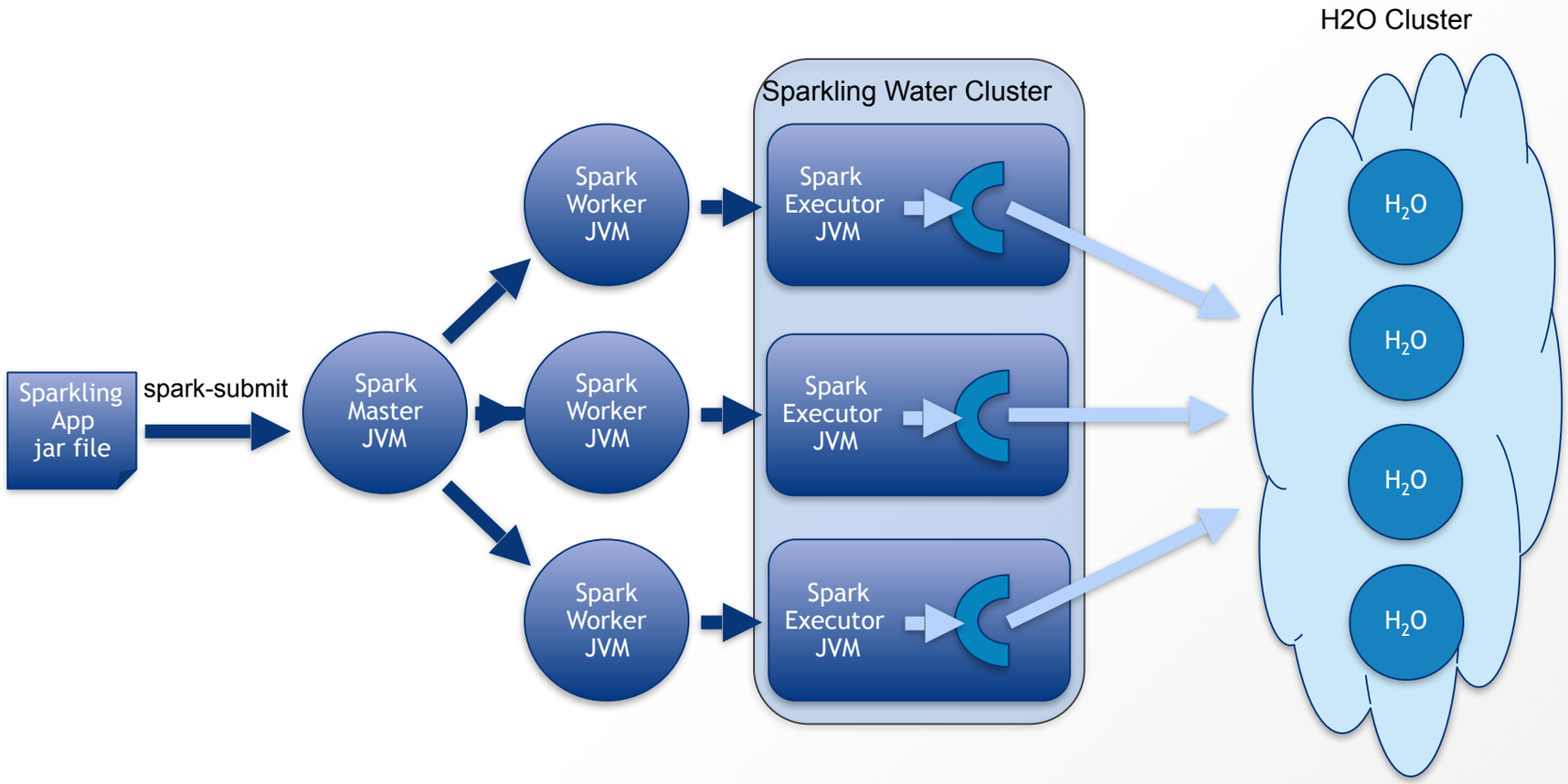
- Sparkling Water is using external H2O cluster instead of starting H2O in each executor
- Spark executors can come and go and H2O won't be affected
- Start H2O cluster on YARN automatically

Separation Approach

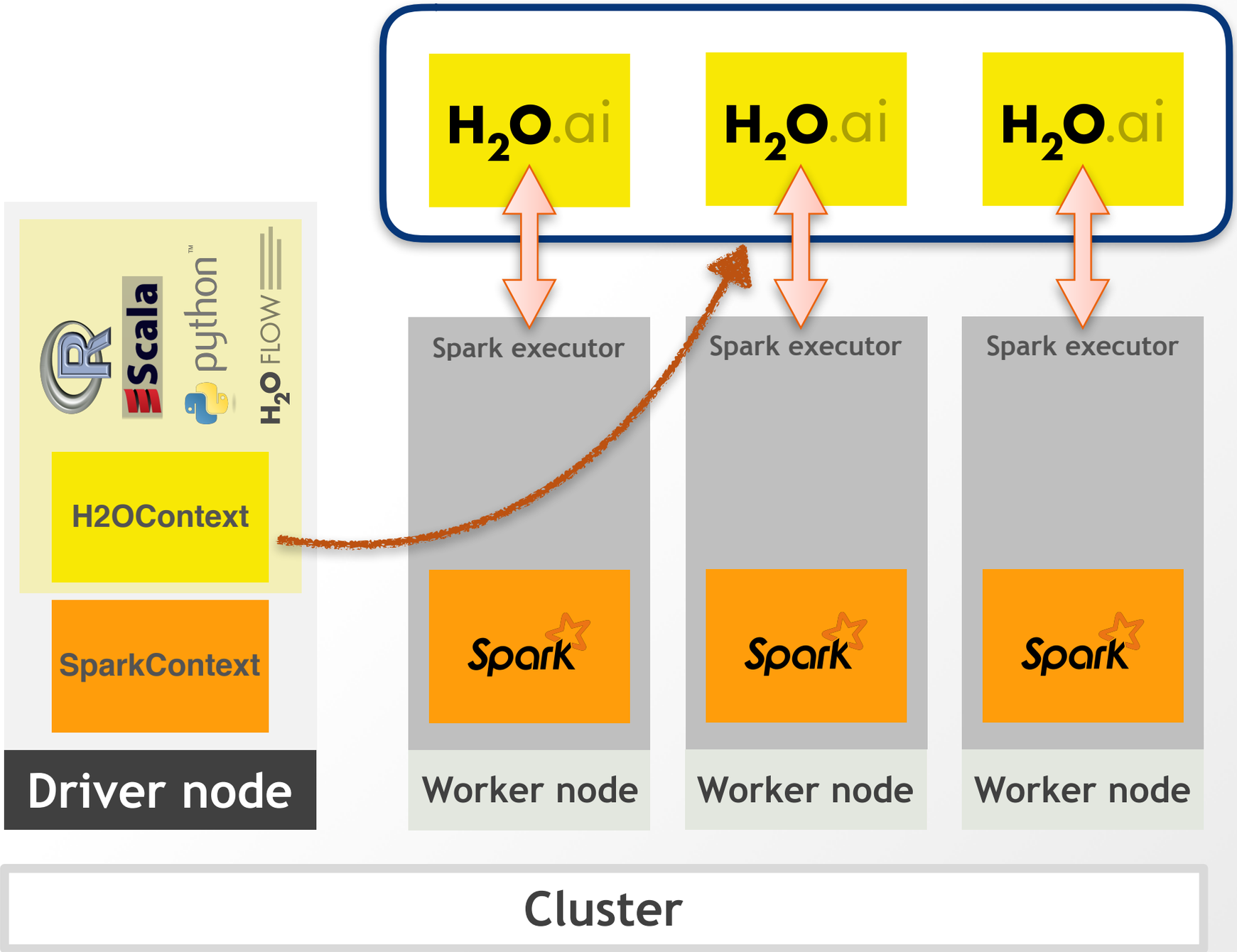
- **Separating Spark and H2O**
  - But preserving same API:
    val h2oContext = H2OContext.getOrCreate("http://h2o:54321/)

- Spark and H2O can be submitted as Yarn job and controlled in separation
  - But H2O still needs non-elastic environment (H2O itself does not implement HA yet)

Sparkling Water External Backend



External Backend



Pros & Cons

- **Advantages**
  - H2O does not crash when Spark executor goes down
  - Better resource management since resources can be planned per tool

- **Disadvantages**
  - Transfer overhead between Spark and H2O processes
    - under measurement with cooperation of a customer

Modes

- **Auto Start mode**
  - Start h2o cluster automatically on YARN
- **Manual Start Mode**
  - User is responsible for starting the cluster manually

More Info

- Documentation: http://docs.h2o.ai
- Tutorials: https://github.com/h2oai/h2o-tutorials
- Slidedecks: https://github.com/h2oai/h2o-meetups
- Videos: https://www.youtube.com/user/0xdata
- Events & Meetups: http://h2o.ai/events
- Stack Overflow: https://stackoverflow.com/tags/sparkling-water
- Google Group: https://tinyurl.com/h2ostream
- Gitter: http://gitter.im/h2oai/sparkling-water

H2O.ai

Thank you!

Sparkling Water is open-source ML application platform combining power of Spark and H2O

Learn more at h2o.ai
Follow us at @h2oai
PS: We are hiring!


MAKERS
GONNA
MAKE
XML
JAKER


2307
MAKERS GONNA MAKE
H2O
THIRSTY?

H2O AI World London
29 & 30 October

- **H2O.ai** is a **Leader** in 2018 Gartner Magic Quadrant for Data Science & Machine Learning Platform.
- **H2O AI World** is our **Community Event**.
  - 29th October: Hands-on H_2O Training
  - 30th October: Tech Talks, Real-World Use Cases, Kaggle Grandmasters Panel & More

Register at: **world.h2o.ai**
Promotional Code: **LASTCHANCE**