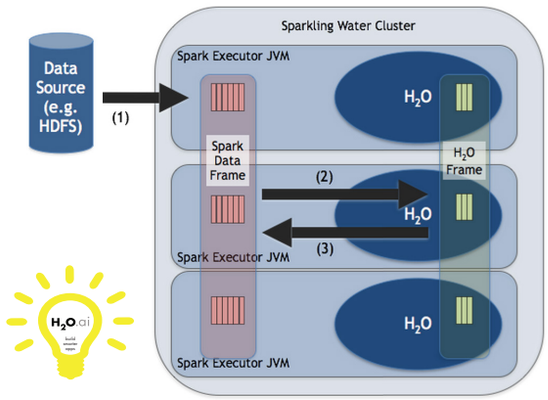
H2O.ai

DATASHEET

Fast, Scalable Machine & Deep Learning for Smarter Applications

H2O.ai is an open-source enterprise software analytics company located in Mountain View, California. Drawing on the expertise of leading data scientists and software engineers, we build high-performance machine learning software. With a rapidly growing community of more than 25,000 users worldwide, we are building a grassroots movement of data science and software developers building smarter applications. H2O.ai is a Gartner Cool Vendor in Data Science for 2015.

Spark + H2O SPARKLING WATER

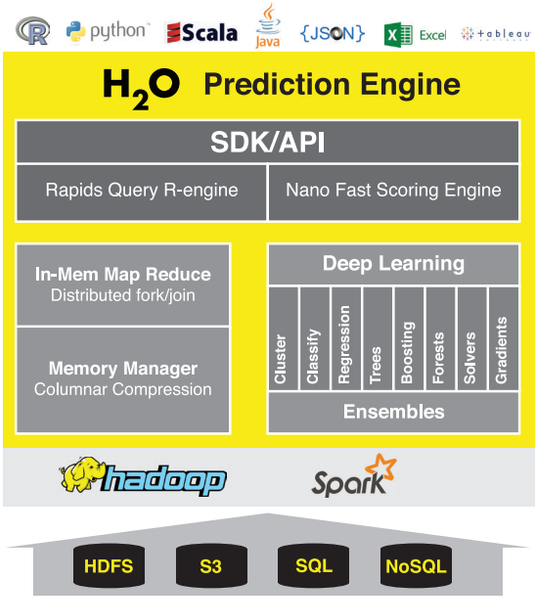


CREATE SMARTER SPARK APPS WITH H2O LIBRARY

Sparkling Water blends data science workflows into developers' applications using H2O's machine learning technology and Spark's fast data munging capabilities. Sparkling Water provides the API calls to transform an H2O Frame to a Spark Data Frame, allowing access to Spark's SQL engine and Sparkling Water conversely transforms Data Frames to H2O Frames for access to H2O's algorithms.

In-Memory	*Columnar Distributed Data Frames*
Data Engineering	*SQL and Data Frame Filters and Maps*
Machine Learning	*H2O's Library and MLlib*
Extensibility	*Input: HDFS, Databases, S3, Hive* *Output: POJO*
API	*Scala, Python, R, REST API*
Real-Time	*Spark Streaming, Nano-Fast Scoring*

CAPABILITIES



DATA MUNGING

- Quantiles, Summaries, Histograms
- Filtering, Splicing, Binding
- Lightning Fast Group-bys

MODEL SELECTION AND VALIDATION

- Grid Search: Hyperparameter optimization to programmatically tune parameters and find best models
- Cross Validation: Validify models for similar behavior across independent subsets.
- Variable Importance: Visualize Machine Learning models and interpret results in human readable fashion.

SCORING

- Export big models as binary serialized object to quickly resume workflows between sessions
- Export models and workflows as plain old java objects (POJOs) to quickly plug model into an application or product.

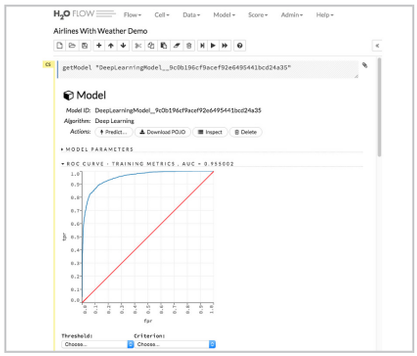
ALGORITHMS

- **Deep Learning:** World record 99.1% accuracy on MNIST data, nonlinear solutions to dimensionality reduction and outlier detection, online learning: continually update old models with new data
- **Generalized Linear Modeling:** Elastic-net regularization, L-BFGS solver to handle million columns wide data, scales almost strongly linearly (billion row logistic regression in 5 seconds)
- **Gradient Boosting Machine:** One of the first distributed and parallelized implementations, produces an ensemble of shallow decision trees with increasing refined approximations improving predictions
- **Distributed Random Forest:** Build hundreds of random trees in parallel
- **Ensembles:** Super learner algorithm that learns the optimal combination of base learners, combines best of Deep Learning, GLM, DRF, and GBM.
- **K-Means:** Cluster like-elements for unsupervised learning use cases
- **Generalized Low Rank Models:** Latest algorithm to come from Stanford- convex optimization, matrix factorization, imputation of missing data, dimensionality reduction, recommendation engine
- **Naive Bayes:** Highly scalable linear learning for fast textual classification

H2O FLOW

H2O provides an interactive web interface called Flow that seamlessly blends a command line, text-based shell with a modern GUI. With Flow, H2O users can combine:

- Coffeescript code
- R/Spark/Python code
- Text as Markdown
- Mathematical expressions
- Charts and plots
- Video content



There is a minimal learning curve for front end users who want to create Flows, iPython Notebooks, or R Scripts. Packages and modules for R and Python are built atop a uniform REST API.

R

```
> library(h2o)
> h2o.init()
> df <- h2o.importFiles(path = ...)
```

Python

```
>>> import h2o
>>> h2o.init()
>>> df = h2o.import_file(...)
```

Sparkling Water

```
scala> import org.apache.spark.h2o._
scala> val h2oContext = new H2OContext(sc).start()
scala> import h2oContext._
scala> val df = new H2OFrame(new java.io.File(...).toURI)
```

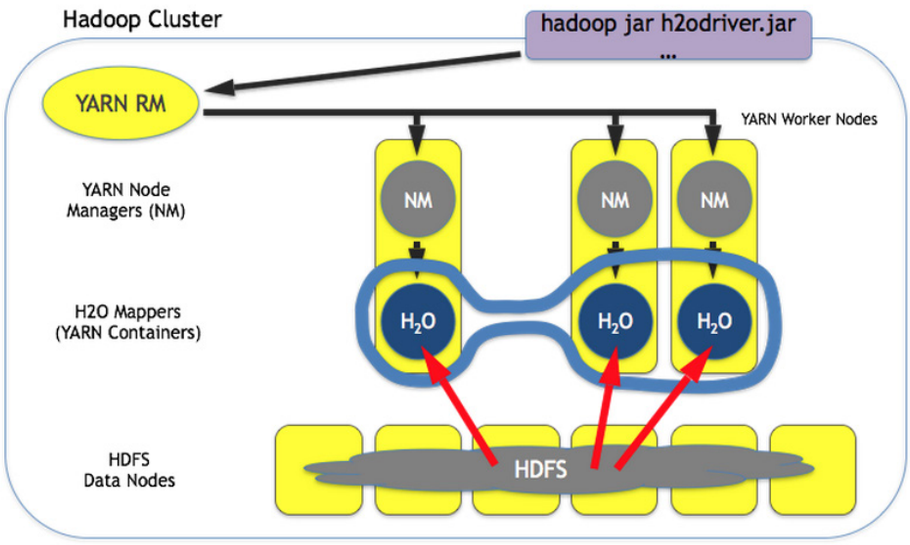
Tel: +1.650.227.4572 | sales@h2o.ai

http://www.github.com/h2oai

@h2oai

H2O.ai
DATASHEET

LAUNCH ON HADOOP



COMMUNITY AND OPEN SOURCE

H2O is the fastest growing machine learning and data science project on Github with over 20,000 commits and 10,000+ installations. H2O is Apache v2 Open Source and allows extensibility and customization by users. With 150+ meetups over 2 years, H2O is a word-of-mouth movement that brings mathematicians, engineers, and analysts together to learn and improve the product.

Join us on the H2O Open Tour 2016!

CHICAGO, IL MAY 3RD
NEW YORK, NY JULY 19TH
SAN FRANCISCO, CA SEPT 13-14TH
DALLAS, TX OCT 26TH

CUSTOMERS

Capital One, Cisco, Nielsen Catalina Solutions, nc, Zurich, MarketShare a Neustar Solution, Transamerica, PayPal

KEY BENEFITS

BETTER PREDICTIONS -
Ready-to-use, powerful algorithms that use all data. Parallel distribution on data enables accurate computations across one or many nodes by moving the code to the data.

SPEED -
In-memory parallel processing provides real-time responsiveness, increases efficiency, and enables users to run more models, no sampling required.

EASE OF USE -
Easy to setup and adopt with intuitive H2O Flow webUI. Support for existing languages R, Java, Scala, and Python through H2O's REST API.

EXTENSIBILITY -
Seamless Hadoop integration with distributed data ingestion from HDFS and S3. Models are built and exportable in plain Java code.

SCALABILITY -
Easy to iterate, develop, and train models on all data sets.

REAL-TIME SCORING -
Predict and score more accurately and 10x faster than the next best technology on the market.

TRY H2O FOR YOURSELF! Download the latest version at http://www.h2o.ai/download

JOIN OUR COMMUNITY! http://h2o.ai/events | http://github.com/h2oai

OUR PARTNERS

R, databricks, MAPR, cloudera, tableau, Hortonworks

Tel: +1.650.227.4572 | sales@h2o.ai
http://www.github.com/h2oai
@h2oai