H_2O

DATASHEET

Introducing H2O: Fast. Scalable Machine Learning for Better Predictions

H2O brings clarity to the complexity of large-scale data analysis. H2O is the industry's first scalable machine learning platform trusted by leading businesses, such as PayPal, Nielsen, MarketShare, Trulia, Collective and Cisco. The platform combines cutting-edge algorithms, high-performance in-memory processing, and the freedom of OpenSource to rapidly predict and analyze data at a massive scale.

Product Overview: **Ready for Production**

H2O makes it possible for anyone to easily apply math and predictive analytics to solve today's most challenging business problems. It intelligently combines unique features not currently found in other machine learning platforms including:

➔ **Best of Breed Open Source Technology** – Enjoy the freedom that comes with big data science powered by OpenSource technology. H2O leverages the most popular OpenSource products like Apache™ Hadoop® and Spark™ to give customers the flexibility to solve their most challenging data problems.

➔ **Easy-to-use WebUI and Familiar Interfaces** – Set up and get started quickly using either H2O's intuitive Web-based user interface or familiar programming environments like R, Java, Scala, Python, JSON, and through our powerful APIs.

➔ **Data Agnostic Support for all Common Database and File Types** – Easily explore and model big data from within Microsoft Excel, R Studio, Tableau and more. Connect to data from HDFS, S3, SQL and NoSQL data sources. Install and deploy anywhere – on a desktop, in the cloud, on a Hadoop cluster, or

➔ **Massively Scalable Big Data Analysis** – Train a model on complete data sets, not just small samples, and iterate and develop models in real-time with H2O's rapid in-memory distributed parallel processing.

➔ **Real-time Data Scoring** – Use the Nanofast Scoring Engine to score data against models for accurate predictions in just nanoseconds in any environment. Enjoy 10X faster scoring and predictions than the next nearest technology in the market.

H2O OpenSource Edition

Work with R and Familiar Tools,

- **Community Support** – Documentation, Weekly Meetups, User Forums, and Email Support are all included.

H2O Enterprise Edition

All the features of H_2OStandard Edition, plus:

- **Enterprise Support** – Personalized enterprise support options tailored to a variety of use cases.
- **Model Management** – Web-based application to manage all of your data, models, scoring and deployment into production.

An End-to-End Platform to Solve Your Data Dilemmas

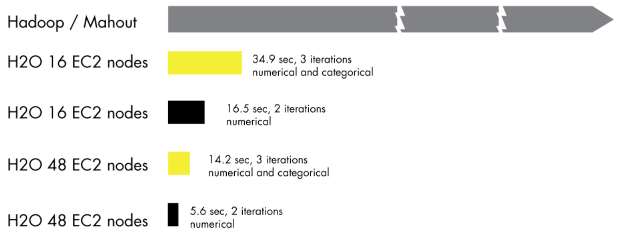
H2O reduces the need for programming and coding to model the data and supports the complete end-to-end analytical workflow:

PARSE	FEATURE	MODEL	SCORE	PREDICT
Support for multiple data sources including Hadoop, Database, and S3. Support for all common file types,	Flexible OpenSource technology supports easy integration with current data tools and platforms, including Hadoop, Spark and SQL	Easy-to-use Web UI delivers intuitive model creation and management and real-time data interaction features	High-performance scoring engine and in-memory processing produces accurate results in nanoseconds	Cutting-edge math algorithms deliver predictions to help you anticipate future needs and solve business challenges

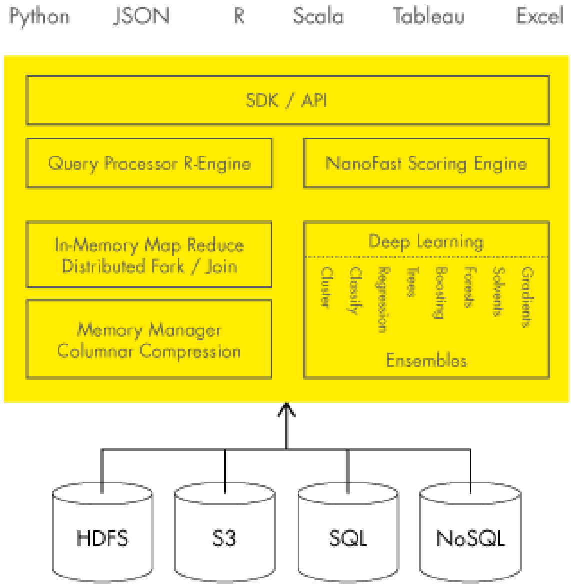
> *"We've tried everything from Mahout, to Vowpal Wabbit, and even R packages like RMR and none work as well as H2O."*
>
> — **Hassan Namarvar,**
>
> **Principal Data Scientist of ShareThis.**

H2O Billion Row Machine Learning Benchmark

GLM Logistic Regression



Compute Hardware: AWS EC2 c3.2xlarge - 8 cores and 15 GB per node, 1 GbE interconnect
Airline Dataset 1987-2013, 42 GB CSV, 1 billion rows, 12 input columns, 1 outcome columnt
9 numerical features, 3 categorical features with cardinaliteis 30, 376 and 380



H2O Customers

Technology Overview:
100% of your data. 100x faster results.

- **Data Compression and Support for All Data Platforms** – H2O compresses all forms of data, typically 2x to 4x better than on disk. The company is confident in its data compression techniques and guarantees that if the data is accessed linearly then H2O can match the performance you see using C or Fortan. The access time is only bound by one's memory bandwidth.

- **In-Memory Parallel Processing** – With H2O one can stop sampling data and start analyzing results on entire data sets to get the complete data picture. H2O in-memory processing engine analyses massive data sets in real-time, enabling customers to analyze over 50% more data that with traditional methods.

- **Cutting Edge Machine Learning Algorithms** – H2O delivers parallel & distributed mathematical models on big data at speeds up to 100X faster than other predictive analytics providers.

- **Native Integration for R, Java, Scala, Python and REST API** – H2O features native support for a number of APIs including R, Java, Scala, Python and REST. For example, using the R interface, one can forward workflows to H2O for big data processing, and work in a familiar interface while running algorithms on data sets that are hundreds of times larger than what would be possible on a desktop.

- **Simple deployment without intermediary transformations** – H2O can collect data from a variety of sources and feed it into elaborate models without the need for customizations, lengthy data transformations, or significant ramp up time. Scalable machine learning made simple.

What's New with H_2O?

Sparkling Water

With the launch of Sparkling Water, H2O now provides a powerful machine learning engine and API for the Spark Platform. By blending H2O's machine learning technology with Spark's fast and intuitive data-munging capabilities, H2O now offers an ideal solution that meets the demands of both the Spark data science and developer communities to build intelligent and smarter applications.

Deep Learning

H2O expands on the company's premier algorithm set to now include deep learning algorithms which model high-level abstractions using multiple non-linear transformations.

Model Management

H2O Model Management is the first Web-based application to visualize your data, models, and score to choose the best champion for your business application.

Try For Yourself

Download the latest version here:
http://www.h2o.ai/download

Data collection is easy. Making predictions is hard.

H2O makes it easy to train large models on your data through faster and better machine learning. Existing big data solutions are batch oriented, using small sample sizes and manual data analysis. H2O delivers real-time and interactive modeling of massive data sets with the latest machine learning technology available in the market. By taking the complexity out of modeling, scoring, and deployment, better predictions become the norm in the enterprise — not the exception.

With H2O, you can:

- **Make better predictions.** Harness sophisticated, ready-to-use algorithms and processing power to analyze bigger data sets, more models, and more variables.
- **Get started with minimal effort and investment.** H2O is an extensible open source platform that offers the most pragmatic way to put big data to work for your business, working with existing languages and tools.

Machine Learning with H2O is as Easy as 1-2-3

1) Prepare Your Data For Modeling	
Data Profiling	Quickly summarize the shape of your dataset to avoid bias or missing information before you start building your model. Missing data, zero values, text, and a visual distribution of the data are visualized automatically upon data ingestion.
Summary Statistics	Visualize your data with summary statistics to get the mean, standard deviation, min, max, cardinality, quantile and a preview of the data set.
Aggregate, Filter, Bin, and Derive Columns	Build unique views with Group functions, Filtering, Binning, and Derived Columns.
Slice, Log Transform, and Anonymize	Normalize, anonymize, and partition to get your data into the right shape for modeling.
Variable Creation	Highly customizable variable value creation to hone in on the key data characteristics to model.
PCA	Principal Component Analysis makes feature selection easy with a simple to use interface and standard input values.
Training and Validation Sampling Plan	Design a random or stratified sampling plan to generate data sets for model training and scoring.
2) Model with State of the Art Machine Learning Algorithms	
Generalized Linear Models (GLM)	A flexible generalization of ordinary linear regression for response variables that have error distribution models other than a normal distribution. GLM unifies various other statistical models, including linear, logistic, Poisson, and more.
Decision Trees	A decision support tool that uses a tree-like graph or model of decisions and their possible consequences.
Gradient Boosting (GBM)	A method to produce a prediction model in the form of an ensemble of weak prediction models. It builds the model in a stage-wise fashion and is generalized by allowing an arbitrary differentiable loss function. It is one of the most powerful methods available today.
K-Means	A method to uncover groups or clusters of data points often used for segmentation. It clusters observations into k certain points with the nearest mean.
Anomaly Detection	Identify the outliers in your data by invoking a powerful pattern recognition model.
Deep Learning	Model high-level abstractions in data by using non-linear transformations in a layer-by-layer method. Deep learning is an example of unsupervised learning and can make use of unlabeled data that other algorithms cannot.
Naïve Bayes	A probabilistic classifier that assumes the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. It is often used in text categorization.
Grid Search	Is the standard way of performing hyper parameter optimization to make model configuration easier. It is measured by cross-validation of an independent data set.
3) Score Models with Confidence	
Predict	Generate outcomes of a data set with any model. Predict with GLM, GBM, Decision Trees or Deep Learning models.
Confusion Matrix	Visualize the performance of an algorithm in a table to understand how a model performs.
AUC	A graphical plot to visualize the performance of a model by its sensitivity, true positive, false positive to select the best model.
HitRatio	A classification matrix to visualize the ratio of the number of correctly classified and incorrectly classified cases.
PCA Score	Determine how well your feature selection is for a particular model.
Multi-Model Scoring	Compare and contrast multiple models on a data set to find the best performer to deploy into production.

About H_2O

H2O is for data scientists and business analysts who need scalable and fast machine learning. H2O is an open source predictive analytics platform. Unlike traditional analytics tools, H2O provides a combination of extraordinary math and high performance parallel processing with unrivaled ease of use. H2O speaks the language of data science with support for R, Python, Scala, Java and a robust REST API. Smart business applications are powered by H2O's NanoFast™ Scoring Engine. Learn more by going to **http://www.h2o.ai** and contact us for more information.