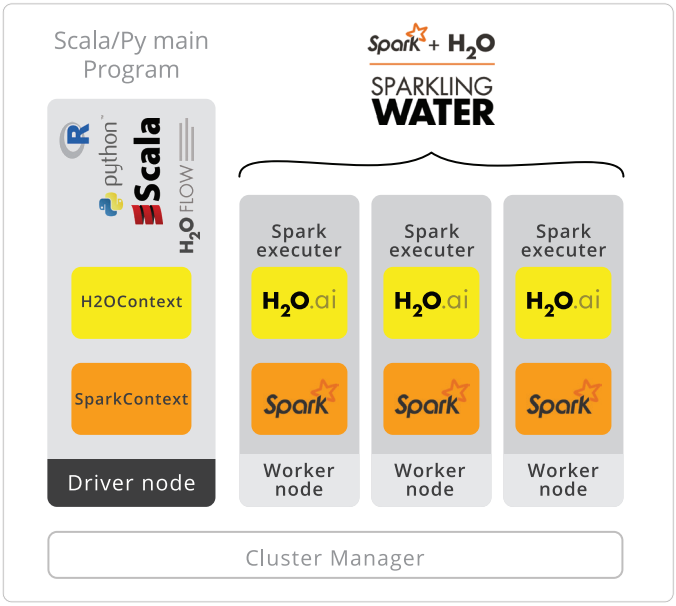
Sparkling Water

Sparkling Water blends data science workflows into developers' applications using H2O's machine algorithims and Spark's fast data munging capabilites. Sparkling Water enables usage of H2O algorithms with Spark Data Frames by providing a transparent API to exchange data between H2O Frames and Spark Data Frames.

Why Sparkling Water?



Sparkling Water was designed to allow users to get the best of Apache Spark - its elegant APIs, SQL-based data munging, machine learning pipelines - along with H2O's computation speed of fully-featured machine learning algorithms. Sparkling Water **also allows for greater flexibility when it comes to finding** the best algorithm for a given use case. Apache Spark's **MLib offers a library of popular algorithms directly** built using Spark. Sparkling Water empowers enterprise customers to use H2O algorithms in conjunction with, or instead of, MLlib algorithms on Apache Spark.

1. **Parallelized data processing:** H2O is designed to quickly process huge amounts of data in a distributed and fully parallelized fashion.

Benefits

- Seamlessly transition back and forth between Spark and H2O
- Use Scala, Python or R to build models
- Power of Spark SQL-based data munging combined with the speed of H2O
- All the features of H2O included (Flow - UI, model export)

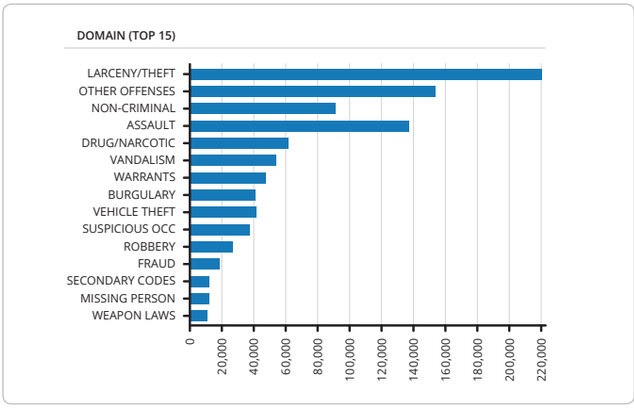
Highlights

- **Accuracy:** AutoML, Ensembles, GBM, GLM, DRF, Deep Learning
- **Speed:** In Memory, Distributed Computation
- **Interface:** R, Python, Flow
- **Developers:** Spark API, PySpark, Sparklyr
- **Community:** Expert Data Scientists, Developers, Data Engineers
- **Cloud:** Databricks Cloud, AWS, Azure

Features

- Seamless integration with Spark API.
- Run Scala code in Flow.
- H2O algorithms are exposed as Spark estimator enabling transparent integration with Spark machine learning pipeline
- Bringing H2O's Visual Intelligence to MLlib algorithms.
- Support of Driverless AI MOJO pipelines

2. **Streamline model training, evaluation & comparison and scoring:** H2O operationalizes this process by:
   a. Providing a library of ML algorithms **supporting advanced, algorithm-specific features.** Moreover,



H2O allows combining models into ensembles (super-learners) or finding the best model with AutoML.

b. Performing **fast exploration of hyper-space of parameters** (a.k.a. grid search).

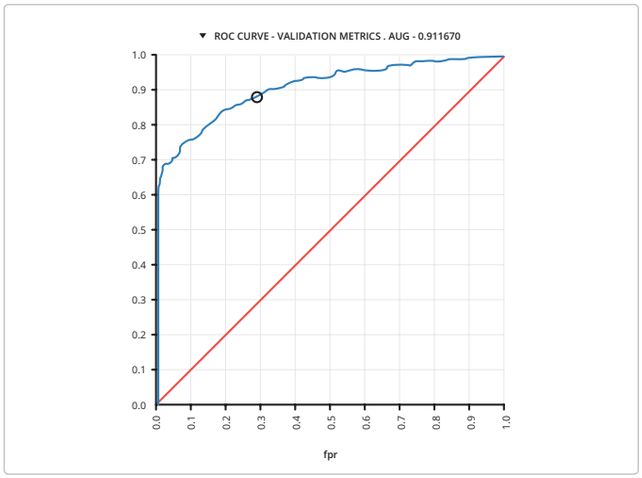
c. Providing the ability to **specify various criteria** that identify and select the best model, e.g. accuracy, building time, scoring time, etc.

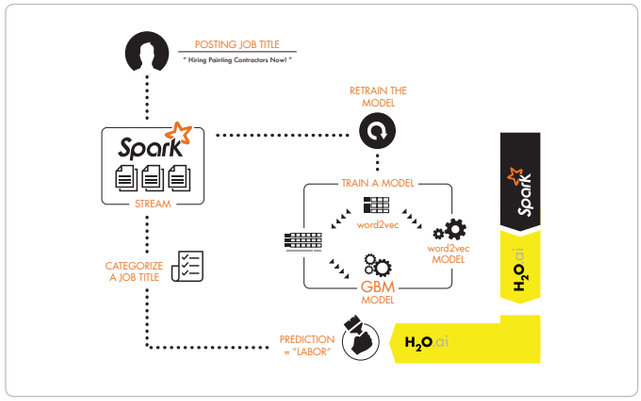
d. **Ability to continue model training** with modified parameters and additional relevant input data.

e. Continuous modeling feedback: **Visualization of various model characteristics** on-the-fly during training as well as of the final model.

3. **Deployment of optimized models:** Model deployment is one of the most critical elements of the machine learning process. H2O and Driverless AI allows for the export of trained models as an optimized code for deployment into target systems (i.e., web services, applications, etc.) The exported models can be also used as part of Spark machine learning pipelines.

4. **Sparkling Water deployment:** Easy use of Sparkling Water with existing Spark distribution with help of published Sparkling Water package. Moreover, Sparkling Water provides two operation modes (internal and external) which reflect demand of





Sparkling Water use-case example architecture

different execution environments and allow to manage H2O cluster as part of Spark cluster or separatelly.

About H2O.ai

H2O.ai is the open source leader in AI and automatic machine learning with a mission to democratize AI for everyone. H2O.ai is transforming the use of AI to empower every company to be an AI company in financial services, insurance, healthcare, telco, retail, pharmaceuticals and marketing. H2O.ai is driving an open AI movement with H2O, which is used by more than 18,000 companies and hundreds of thousands of data scientists. H2O Driverless AI, an award winning and industry leading automatic machine learning platform for the enterprise, is helping data scientists across the world in every industry be more productive and deploy models in a faster, easier and cheaper way. H2O.ai partners with leading technology companies such as NVIDIA, IBM, AWS, Intel, Microsoft Azure and Google Cloud Platform and is proud of its growing customer base which includes Capital One, Nationwide Insurance, Walgreens and MarketAxess. H2O.ai believes in AI4Good with support for wildlife conservation and AI for academics. Learn more at www.H2O.ai

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