DELL Technologies

White paper

Reference Architectures for H2O.ai

Unleash the power of AI using H2O.ai on Dell EMC infrastructure optimized for machine learning

Abstract

This white paper outlines technical considerations and sizing guidance for on-premises enterprise AI platform reference architectures jointly developed by Dell Technologies, Intel® and H2O.ai. Running H2O.ai software on optimized Dell EMC infrastructure with the latest Intel® Xeon® Scalable processors and NVMe storage, enables organizations to use AI to improve customer experiences, streamline business processes, and decrease waste and fraud.

February 2020

Contents

Executive summary

Emerging technologies, such as Internet of Things (IoT) and mobile technologies and applications, are creating data at speeds and volumes never imagined for organizations across a wide range of industries. Many are exploring the use of artificial intelligence (AI) for converting all this data into the fuel that enables better decisions, faster. Insights enabled by AI can help organizations reduce risk, create value and uncover new opportunities. However, deploying systems capable of running AI workloads can be complex, requiring extensive integration and testing of the hardware and software.

To overcome these challenges, Dell Technologies is collaborating with Intel and H2O.ai to develop and benchmark reference architectures for the H2O® Driverless AI enterprise platform and the H2O.ai open-source platforms, H2O and H2O Sparkling Water. Thousands of global enterprises trust H2O.ai software solutions because they work with some of the most widely used statistical and machine learning (ML) algorithms including gradient boosting machines, generalized linear models, deep learning (DL) and more. Benchmarking tests showed significant performance improvements with high levels of accuracy for these Dell Technologies reference architectures.

Using these reference architectures, IT teams can confidently leverage a combination of Dell EMC enterprise-grade servers, storage and networking together with H2O software, utilizing predictive analytics capabilities to optimize automation and ML tasks. This combination is designed to lower the cost and complexity of adopting enterprise-grade AI systems, democratizing AI and enabling more organizations to use it to improve customer experiences, streamline processes, decrease waste and fraud.

Dell Technologies, Intel and H2O.ai collaborate to develop engineering-validated reference architectures to empower data scientists and accelerate AI adoption.

Solution overview

This white paper covers a range of data science platforms and ML reference architectures that came out of the Dell Technologies and H2O.ai collaboration.

Validated software

Validated software includes the H2O Driverless AI enterprise platform and the H2O and H2O Sparkling Water open-source software platforms. These software platforms from H2O leverage highly optimized Intel libraries and ML frameworks to deliver enhanced performance on Dell EMC infrastructure.

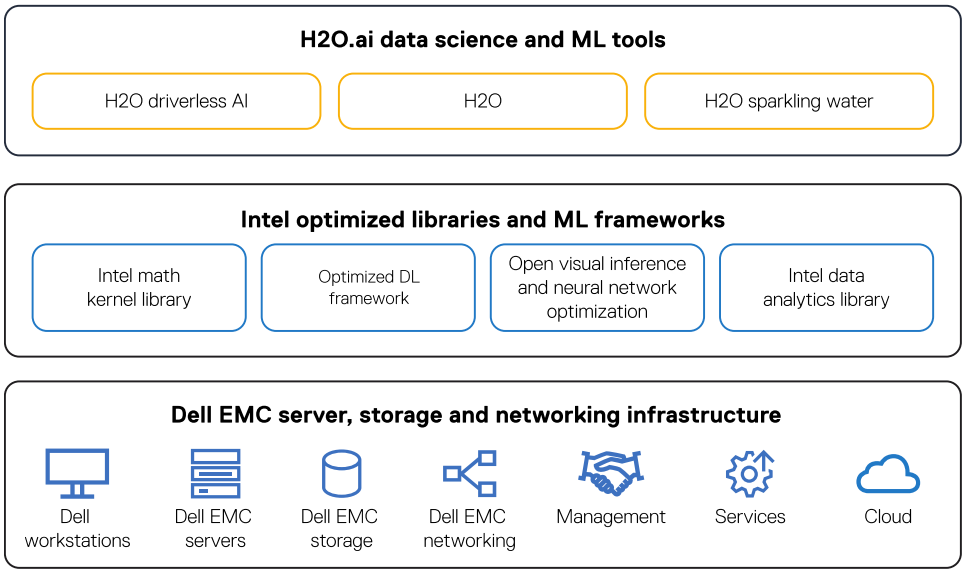


Figure 1. *AI and ML platform with H2O.ai, Intel and Dell*

Intel Xeon Scalable processors for AI

Intel Xeon processors deliver enhanced performance, flexibility, scalability and lower TCO. With recent advancements in hardware-based AI acceleration, software optimizations to AI frameworks and specialized AI libraries, Dell EMC PowerEdge servers with Intel Xeon Scalable processors provide great performance and scalability on the CPU platform enterprises know and trust.

Validated hardware

The Dell EMC PowerEdge R740xd is a two-socket, 2U rack server designed to run complex workloads using highly scalable memory, I/O capacity and network options. It offers extraordinary storage capacity options, making it well-suited for data-intensive applications that require greater storage, while not sacrificing I/O performance.

Intel DC P4600 NVMe SSDs are Intel 3D NAND SSDs, offering outstanding quality, reliability, advanced manageability and serviceability to minimize service disruptions.

Intel Xeon Scalable processors are optimized for demanding data center workloads. This processor family features higher frequencies than previous-generation Intel Xeon processors, along with architecture improvements, AI and DL inference workload enhancements.

The second-generation Intel Xeon Scalable processors take AI performance to the next level with Intel Deep Learning (DL) Boost, which extends the Intel Advanced Vector Extensions 512 (Intel AVX-512) instruction set with Vector Neural Network Instructions (VNNI). Intel DL Boost significantly accelerates inference performance for DL workloads optimized to use VNNI—sometimes by as much as 30X compared to a previous-generation Intel Xeon Scalable processor.

Optimized Machine Learning library

The Intel Data Analytics Acceleration Library (Intel DAAL) is an easy-to-use library that helps applications deliver predictions more quickly and analyze large data sets without increasing compute resources. It optimizes data ingestion and algorithmic compute together for high performance. It also supports offline, streaming and distributed usage models to meet a range of application needs.

New features include high-performance logistic regression, extended GBM functionality and user-defined data modification procedures.

Intel DAAL can help with all stages of analytics

- **Pre-processing:** Decompression, filtering and normalization
- **Transformation:** Aggregation and dimension reduction
- **Analysis:** Summary statistics and clustering
- **Modeling:** Training, parameter estimation and simulation
- **Validation:** Hypothesis testing and model error detection
- **Decision-making:** Forecasting and decision trees

Dell reference architecture for H2O Driverless AI

H2O Driverless AI is a high-performance, single-node enterprise platform for automatic development and rapid deployment of state-of-the-art predictive analytics models. It simplifies data science projects by using automation to build and select the best predictive analytics models for the chosen data set.

With Driverless AI, users can train and deploy modeling pipelines from the graphical user interface (GUI) with a few clicks. Advanced users can leverage the client/server API through a variety of languages such as Python® or Java®. Driverless AI also allows for statistically rigorous automatic data visualization and interactive model interpretation with reason codes and explanations as part of its machine learning interpretability (MLI) feature set.

H2O Driverless AI can ingest data directly from Apache® Hadoop® Distributed File System (HDFS), Apache Spark®, Amazon® Simple Storage Service (S3), Microsoft® Azure® Data Lake, or any other data source into its in-memory distributed key value store. In addition, its AutoML function automatically runs through algorithms and their hyper-parameters to help produce the best models.

Dell Technologies has long been a pioneer in data analytics, with industry-leading server and data storage used to deploy business-critical workloads. Dell understands the need for optimized and accelerated AI solutions to empower data-driven organizations with reliability and security. That's why Dell engineering teams are collaborating with Intel and H2O.ai to create reference architectures for H2O Driverless AI, tuned to enhance AI capabilities in enterprises.

Figure 2 shows how H2O Driverless AI operates in an automated workflow, with the following steps.

1. **Drag-and-drop data:** Converts plain text sources, including a variety of cloud and desktop data sources, into a tabular data. It's capable of absorbing HDFS, SQL Server®, Amazon S3, Snowflake, Google® BigQuery™, Azure Blog Storage and Local Data Storage.
2. **Modeling and visualization:** Creates data sets and generates visualizations based on the most relevant data to help users get a quick understanding of data prior to starting the model-building process.
3. **Automatic model optimization:** Amalgamates best-practice models and high-performance computing (HPC) to determine the optimum model for the data.
4. **Automatic scoring pipeline:** Uses low-latency pipelines such as Python or Java scoring pipelines for experiments that include feature transformations and models for validation, tuning, selection and deployment.

Automatic feature engineering, machine learning and interpretability

DRIVERLESSAI

Source: h2o.ai

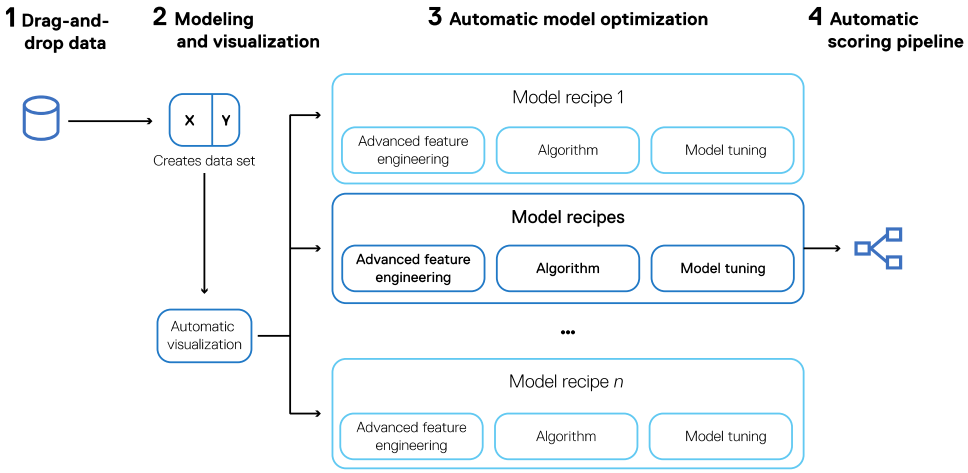


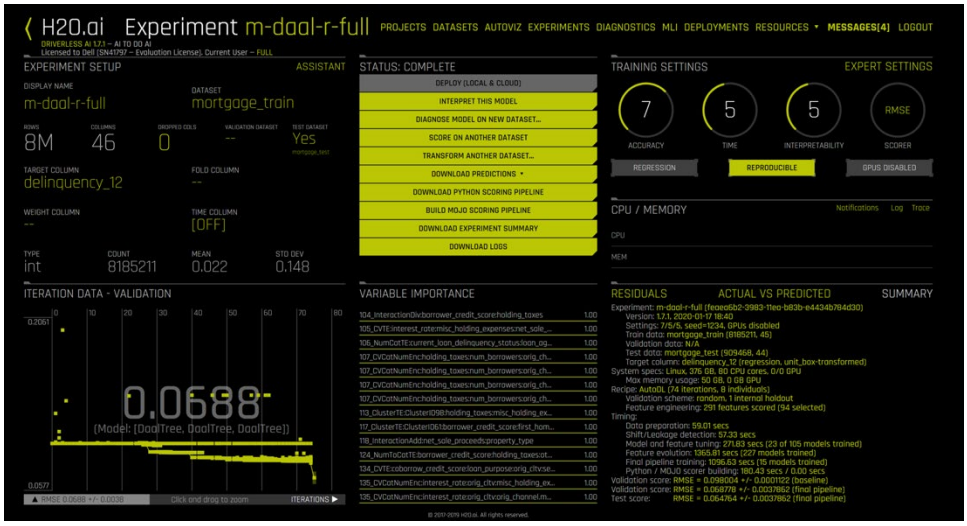
Figure 2. *The H2O Driverless AI workflow*

Built for domain users wanting to leverage AI, business analysts and data scientists, H2O Driverless AI is operated from a GUI for end-to-end data science, as shown in Figure 3.

H2O Driverless AI is successful in resolving the challenges of time, cost and trust with its robust, high-performance, innovative and validated features, such as:

- **Automatic feature engineering:** Enables data scientists to retrieve the most accurate data from algorithms. A library of algorithms and feature transformations enables the software platform to skillfully build and design features for a given data set.
- **Flexibility of data and deployment:** Supports a variety of data sources including Hadoop HDFS, Amazon S3 and more. It can be deployed in the cloud, including Microsoft Azure, Amazon Web Services® (AWS) and Google Cloud Platform™.
- **Automatic data visualization (AutoVis):** Automatically selects data plots based on the most relevant data statistics to help users understand the composition of very large data sets, discover trends and possible issues that could impact modeling results.

- **Automatic model documentation (Auto Doc):** Generates reports for each experiment without the involvement of data scientists and data engineers. The AutoReport includes details about the data used, the validation schema selected, model and feature tuning, and the final model created.



Source: h2o.ai

Figure 3. *H2O Driverless AI user interface*

- **Time-series recipes:** Delivers time-series forecasting capabilities to optimize for nearly any prediction time window. Delivers structured character data and fills the gaps in time-series data and other missing values.
- **Natural language processing (NLP) with TensorFlow™:** Automatically converts text strings into features. Uses TensorFlow to process larger text blocks and build models using all available data to solve business problems like sentiment analysis, document classification and content tagging.
- **Automatic scoring pipelines:** Generates both Python scoring pipelines and new ultra-low-latency automatic scoring pipelines for completed experiments. Deploys feature engineering and the winning ML model in a highly optimized, low-latency, production-ready Java code that can be deployed anywhere.
- **ML interpretability (MLI):** Provides accurate explanations of AI-driven business decisions and modeling results in a readable format to generate trust and authentication for outcomes. Uses a host of different techniques and methodologies for interpretation.
- **Automatic reason codes:** Shows the key positive and negative factors in a model's scoring decision in plain language, enabling the business to explain significant decisions that impact customers.
- **Custom recipe support:** Imports custom recipes for ML algorithms, feature engineering, scorers and configuration that can be used individually or in combination with, or instead of, built-in recipes. Grants greater influence over the automatic ML pipeline and the optimization choices made.
- **GPU support:** Leverages optional GPU acceleration to speed up automatic ML. This includes multi-GPU algorithms for XGBoost, GBM, GLM, K-Means and more.

Driverless AI delivers fast, accurate, interpretable AI for enterprise

Automated machine learning (AutoML) is the process of automating the end-to-end process of applying machine learning to real-world problems.

Reference architecture and implementation

The reference architecture described in this section represents a base configuration for a typical H2O Driverless AI solution.

Hardware and software configuration overview

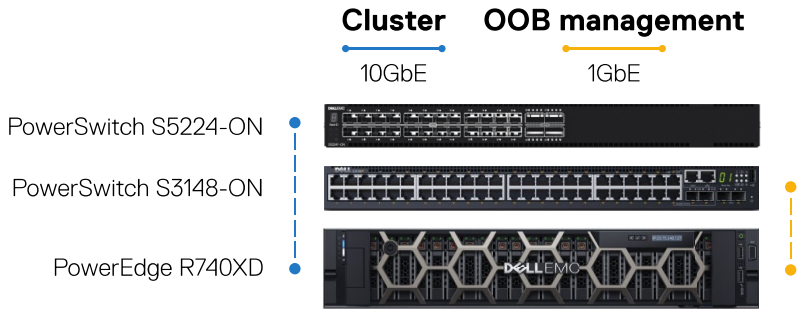


Figure 4. *Hardware and software configuration*

H2O Driverless AI compute node	
Server	1x Dell EMC PowerEdge R740xd
Processor	2x Intel Xeon Gold Scalable 6248
Memory	384GB DDR4 @ 2667 MHz
Drives	Operating system: BOSS card with 2x 480GB SSD Data: 12x Dell Express Flash NVMe 4610 1.6TB SFF
Networking	Intel Ethernet 10G 4P x710 SFP+ rNDC
Top of rack (ToR) switch	Management: Dell Networking PowerSwitch 3148-ON (1GbE) Cluster: Dell Networking PowerSwitch S5224-ON (10/25GbE)
Software	**Version**
Operating system	Red Hat® CentOS® Linux® 7.6 or Red Hat Enterprise Linux (RHEL) 7
H2O software	H2O Driverless AI 1.7.1
Library	Intel DAAL version 2019.5

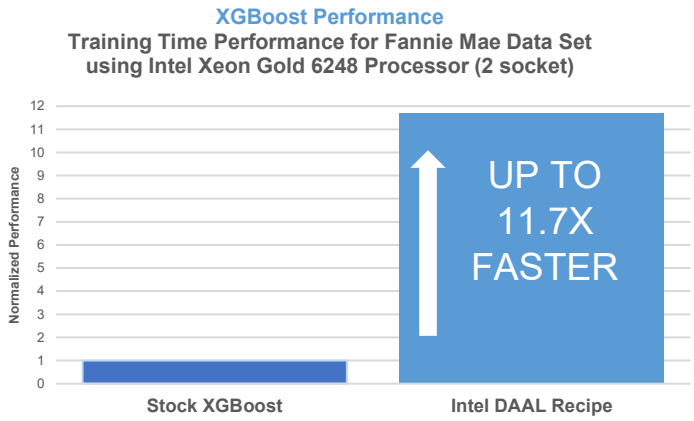
Build CPU-accelerated models using Intel DAAL recipes

Dell and H2O recommend using Intel DAAL recipes for building ML models, to leverage the latest performance features of second-generation Intel Xeon Scalable processors. Intel created and unveiled the DAAL recipe on the H2O.ai open-source recipe repository to enable enterprises to achieve ML at speed and scale. Using the open-source DAAL recipes can provide significant performance gains over stock algorithms.

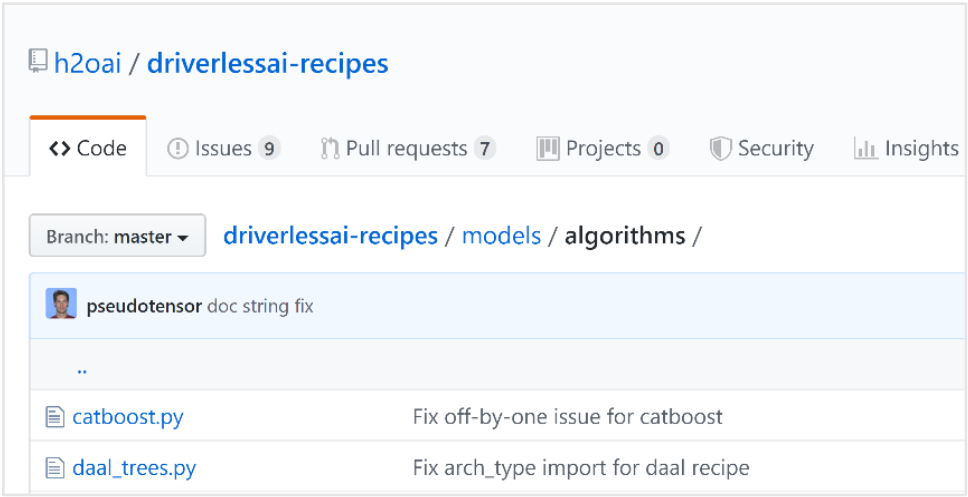
To measure the performance gain using DAAL, we used H2O Driverless AI to build a model using the Fannie Mae® Single-Family Loan Performance data set to predict the delinquency status of the borrower. A model was trained using the stock XGBoost algorithm available in Driverless AI and compared against one built using the DAAL recipe available at the H2O.ai GitHub, as shown in Figure 5.

Intel DAAL completed training in under one hour and generated a model to predict the delinquency status on the Fannie Mae data set with eight million records. The stock XGBoost algorithm took approximately 10 hours to train the model. Intel DAAL was observed to speed up the training process by an impressive margin of 11.7 times over the stock XGBoost model while maintaining accuracy.

H2O Driverless AI automation empowers data scientists, data engineers and domain scientists to work on projects faster and more efficiently.



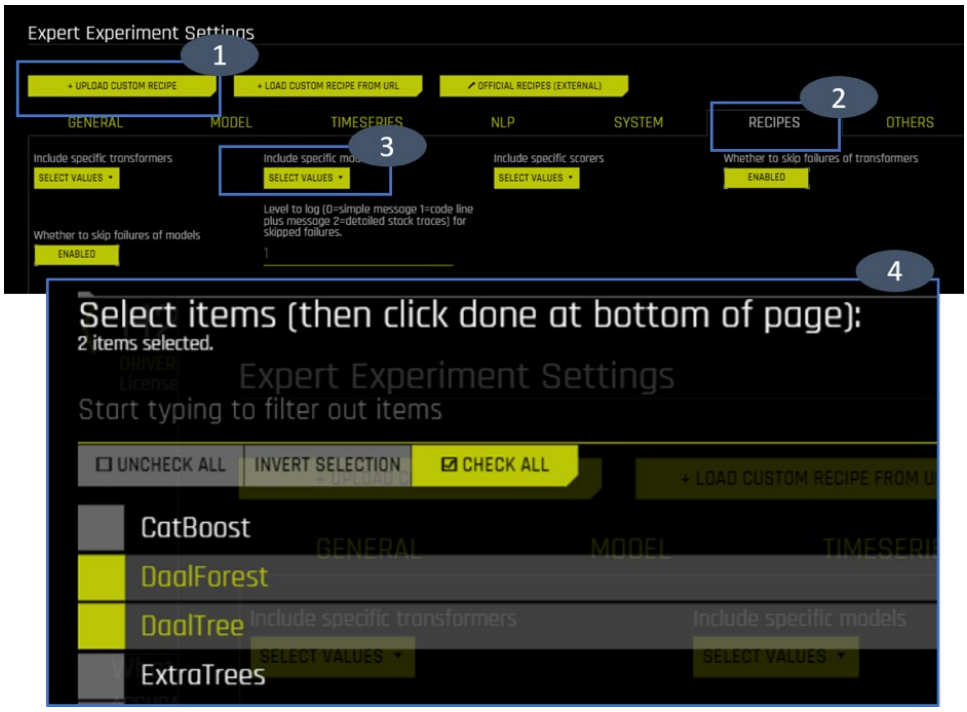
The Intel DAAL recipe is available in the Driverless AI GitHub repository for custom recipes as shown in Figure 5.



Source: h2o.ai

Figure 5. *Driverless AI recipes on the H2O.ai GitHub repository*

Leveraging Intel DAAL in Driverless AI can be accomplished using a few mouse clicks in the Driverless AI user interface. The four steps to import the DAAL recipe are shown in Figure 6.



Source: h2o.ai

Figure 6. *Steps to import Intel DAAL recipe into Driverless AI*

H2O Driverless AI empowers data scientists, data engineers and domain scientists to work on projects faster and more efficiently by using automation. Tasks that previously took months can be reduced to hours or minutes by using H2O Driverless AI. Automatic feature engineering, model validation, model tuning, model selection and deployment, MLI, time-series, NLP and automatic pipeline generation for model scoring are key features. The ability to bring your own recipes enables users to leverage Intel DAAL to reduce training time and shorten the time it takes to execute and implement AI projects.

Dell reference architecture for H2O open-source platforms

H2O.ai open-source platforms are Apache v2 licensed and offered with enterprise support subscriptions. They're built for data scientists, with support for R and Python and an interactive GUI called H2O Flow. H2O is a collection of in-memory, distributed ML algorithms that can be used with existing big data infrastructures. These can include bare metal, Apache Hadoop and Apache Spark clusters. It can ingest data directly from HDFS, Spark, Amazon S3, Azure Data Lake, or other data sources, into its in-memory distributed key value store. H2O can be deployed in three ways, as described in the following table.

H2O open-source platform deployment choices

Deployment method	Details
H2O	H2O is responsible for distribution and inter-node communication.
H2O Sparkling Water	H2O is integrated with Apache Spark. H2O jobs are submitted to the Spark master and then executed within Spark executors.
H2O on Hadoop	YARN (yet another resource negotiator) schedules H2O jobs as MapReduce tasks on Hadoop.

Dell Technologies engineering teams have created and benchmarked reference architectures for H2O and H2O Sparkling Water.

Key capabilities of H2O include:

- **Algorithm reliability:** Distributed computing environments depend on algorithms that are developed from the ground up. H2O supports such algorithms, including Random Forest, GLM, XGBoost, GBM, DL, Generalized Low Rank Models (GLRM), Word2Vec and numerous others.
- **Language compliance:** H2O supports prevalent programming languages such as R, Python and others to build models. This makes developers' tasks easier. They can also make use of H2O Flow, an intuitive GUI that does not require coding.
- **Workflow automation:** The AutoML feature of H2O automates ML workflows for automatic training and tuning of many models within a specified timeline. H2O's Stacked Ensembles feature identifies top-performing models.
- **Enhanced performance:** H2O can manage huge data sets without hampering data accuracy. It does this with its in-memory processing which serializes nodes and clusters. Distributed processing on big data improves speed and efficiency.
- **Easy deployment**: H2O makes use of easy-to-deploy Java models for fast, accurate scoring in any environment.

In-memory, distributed machine learning algorithms with H2O Flow GUI

Source: h2o.ai

Key capabilities of H2O Sparkling Water include:

- **Easy approach to algorithms:** With the help of H2O Sparkling Water, developers are empowered to make use of broad range of H2O algorithms for distributed computing environments. These algorithms, for both supervised and unsupervised approaches, include Random Forest, GLM, GBM, XGBoost, GLRM, Word2Vec and numerous others.
- **Adaptability:** H2O Sparkling Water enables seamless computation from Scala®, R or Python by providing an ideal ML platform for application developers. They can also make use of H2O Flow, an open-source user interface.
- **Easy deployment and accurate scoring:** Sparkling Water makes use of easy-to-deploy and highly accurate Java scoring models.

H2O AI open source engine integration with Spark

Source: h2o.ai

Hardware and software configuration for H2O open-source platforms

The Dell Technologies reference architecture for H2O open-source platforms uses a flexible building block approach to system design, where individual building blocks can be combined to build a system that is optimized specifically for your unique workloads and use cases. The solution includes at least three Dell EMC PowerEdge R740xd Servers for a high-availability deployment of H2O cluster or Spark cluster running Sparkling Water. Dell engineering has tested and benchmarked up to five nodes, as shown in Figure 7.

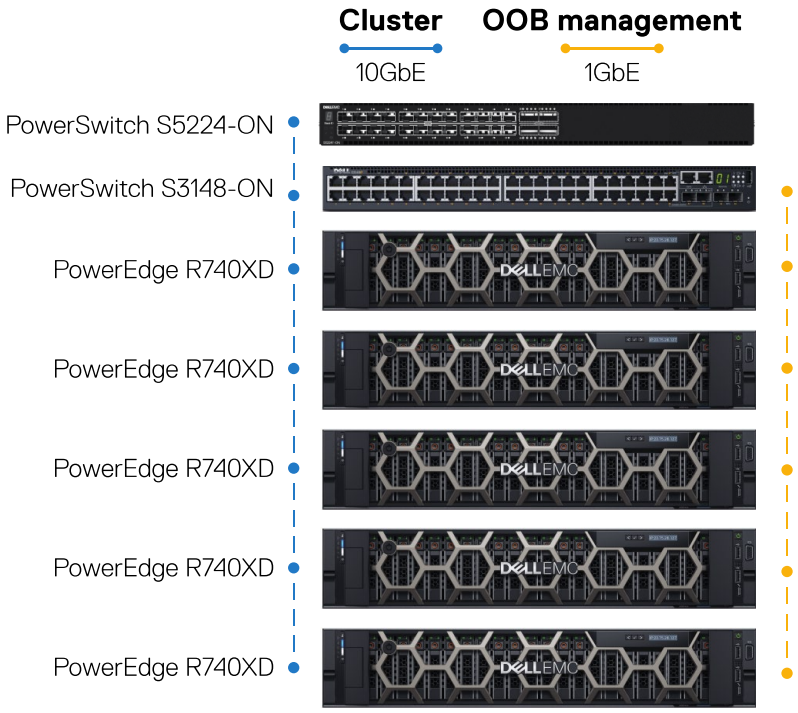


Figure 7. *Hardware and software configuration for H2O and H2O Sparkling Water*

H2O.ai open-source platform compute node	
Server	3x to 5x Dell EMC PowerEdge R740xd
Processor	2x Intel Xeon Gold Scalable 6248
Memory	384GB DDR4 at 2667 MHz
Drives	Operating system: BOSS card with 2x 480GB SSD Data: 12x Dell Express Flash NVMe P4610 1.6TB SFF
Networking	Intel Ethernet 10G 4P X710 SFP+ rNDC
ToR switch	Management: Dell Networking PowerSwitch S3148-ON (1GbE) Cluster: Dell Networking PowerSwitch S5224-ON (10/25GbE)
Software	**Version**
Operating system	CentOS Linux Release 7.6 or RHEL 7
H2O software	H2O Flow (3.26.0.3), JDK12, Sparkling Water 2.3.1
Cloudera® cluster	CDH 5.16.2 Spark 2.3.0
Library	Intel DAAL version 2019.5

Training and inference performance on H2O open-source platforms

H2O Sparkling Water cluster performance was evaluated on three- and five-node clusters. In this mode, H2O launches through Spark workers and Spark manages the job scheduling and communications between the nodes. Three and five Dell EMC PowerEdge R740xd Servers with Intel Xeon Gold 6248 processors were used to train XGBoost and GBM models using the mortgage data set derived from the Fannie Mae Single-Family Loan Performance data set.

XGBoost and GBM models from H2O are commonly used for regression and classification problems and are optimized to scale and leverage additional CPU cores and system memory. The training times to build the predictive model is increased when additional Spark worker nodes are used to train the model.

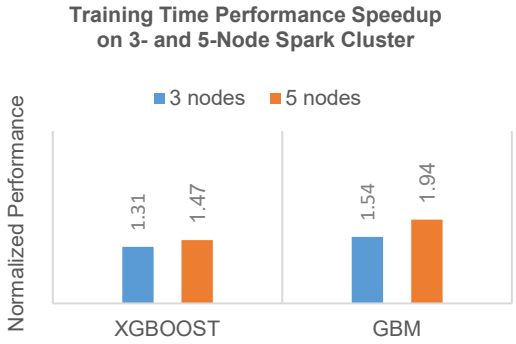


Figure 8. *Training time improvement for a data set on a Sparkling Water cluster*

H2O Sparkling Water also includes AutoML capabilities for automatic training and tuning of many models within a specified timeline. To use this capability, a timeline is specified when multiple models are evaluated and tuned to identify the top-performing model.

We used this feature to perform automatic tuning of the XGBoost model within a window of one and two hours. This is compared against a model that was trained using hyper-parameters that were selected by the data scientist after iterative experimentation and manual optimization. Typically, these tasks, when done manually, can take days to weeks depending on the model and data set being trained. The H2O AutoML feature was observed to be effective in identifying the tuning parameters for the XGBoost model, as illustrated.

The accuracy of the trained models is compared in Figure 9. The AutoML capability of the H2O open-source platform puts it in the class of a user-friendly ML software that can be used by non-experts. The H2O AutoML interface is designed to have as few parameters as possible so the user can just point to the data set in the H2O Flow UI, identify the response column and, optionally, specify a time constraint or limit on the number of total models trained.

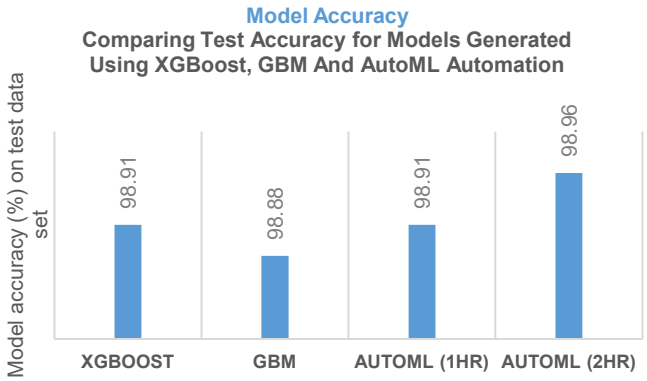


Figure 9. *Comparing model accuracy using workflow automation techniques*

Once the predictive model is trained in Sparkling Water, inferencing can be performed using the inference capability of Sparkling Water to evaluate the trained mode. The data set is split prior to performing training so inferencing can be performed on data that is new to the model. The model was trained on 80% of the mortgage data set and inference was performed on the remaining 20%. The inference speed on a single Spark worker node was observed to be 1.7 million records per second.

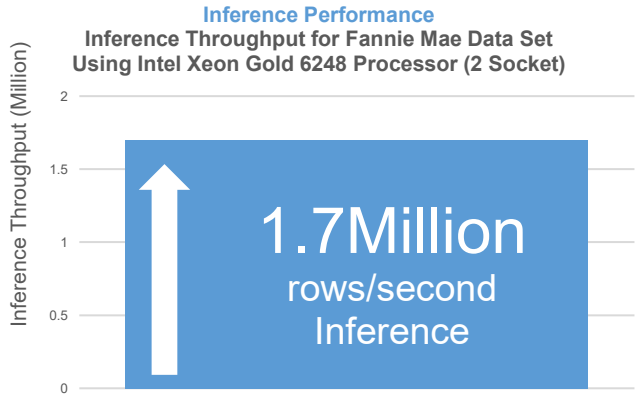


Figure 10. *Measured inference throughput using Sparkling Water on a single node*

In addition to Sparkling Water, we also evaluated the performance improvement when utilizing a stand-alone H2O cluster. In this scenario, H2O software is responsible for distributing the computation and handling communication when the model is trained using multiple cluster nodes. The speedup obtained when training a GBM model on a three-node cluster relative to using a single node for training is shown in Figure 11.

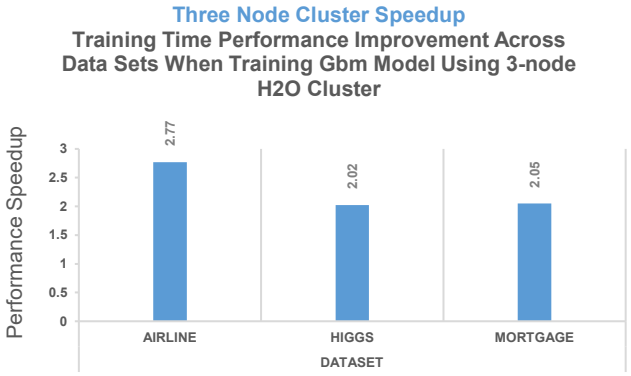


Figure 11: *GBM training time improvements on an H2O cluster*

In addition to the benchmarking, the mortgage data set to build a model that can predict delinquent loans, we used the Airlines DataSet[1] to build GBM models to predict the probability of a flight being delayed or cancelled, and the HIGGS data set[2] to identify signal processes that produce the HIGGS boson particle.

[1] Information on the Airline DataSet can be obtained at: https://github.com/h2oai/h2o-2/wiki/Hacking-Airline-DataSet-with-H2O.

[2] The HIGGS data set can be downloaded from: https://archive.ics.uci.edu/ml/datasets/HIGGS.

About H2O.ai

H2O.ai offers a range of AI and data science platforms from the open-source H2O software platform to integrations for Apache Spark with Sparkling Water and the award-winning H2O Driverless AI platform, delivering an expert data scientist in a box.

Learn more at: h20.ai

Deploy H2O.ai machine learning solutions on Dell infrastructure

Dell Technologies has developed reference architectures with H2O.ai to help you accelerate your AI transformation. H2O Driverless AI is a high-performance, computing software platform for automatic development and rapid deployment of state-of-the-art predictive analytics models which leverages the high core count and architecture of Intel Xeon Scalable processors.

Sparkling Water is ideal for those who need to leverage larger clusters for their processing needs and want to transfer data from Spark to H2O — or vice versa — seamlessly. Dell engineering teams tested a three- and five-node cluster using Dell EMC PowerEdge server infrastructure that shows performance scaling and reduced training times for training complex predictive models using large data sets. We also evaluated the new AutoML capabilities of H2O Sparkling Water in our testing and were able to achieve high levels of accuracy for diverse use cases.

Driverless AI is a commercial offering from H2O that can be used to complete AI projects faster and more efficiently via automation. Driverless AI can be used to perform automatic feature engineering, model validation, model tuning, model selection and deployment in hours or minutes. In addition, MLI, time-series, NLP and automatic pipeline generation for model scoring are valuable features that are available in Driverless AI. Custom recipe allows Driverless AI's automated approach to be enhanced by the data scientist's expert domain knowledge. We observed that the Intel DAAL recipe could be leveraged to reduce training times substantially for XGBoost models.

With its robust features and capabilities, H2O is a powerful AI tool. The Dell reference architectures for H2O Driverless AI and H2O open-source platforms, combined with the latest Intel processors has made incorporating AI more feasible by presenting cost-effective solutions and optimized data management.

Assistance when you need it

Dell Technologies provides data analytics and AI services, from strategy through to implementation and ongoing optimization. It helps to bridge the people, process technology needed to achieve desired business outcomes with speed and scale. This includes implementing and operationalizing AI technologies and helping accelerate data engineering capabilities.

Learn more

Delltechnologies.com/referencearchitectures

Delltechnologies.com/ai

Intel SSD Data Center Family

Intel AI Builders

DELL Technologies

intel