DELL Technologies

intel

ホワイトペーパー

H2O.ai 向けリファレンス アーキテクチャ

機械学習向けに最適化された Dell EMC インフラストラクチャで H2O.ai を使用して AI のパワーを解き放つ

要旨

このホワイトペーパーでは、Dell Technologies、Intel®、H2O.ai によって共同開発されたオンプレミスのエンタープライズ AI プラットフォームのリファレンス アーキテクチャに関する技術的な考慮事項とサイジングのガイダンスについて説明します。最新のインテル® Xeon® スケーラブル プロセッサーと NVMe ストレージを搭載し、最適化された Dell EMC インフラストラクチャで H2O.ai ソフトウェアを実行することによって、組織は AI を使用してカスタマー エクスペリエンスを向上させ、ビジネス プロセスを効率化し、不正や無駄を減らすことができます。

2020 年 2 月

H2O.ai

目次

概要

IoT（Internet of Things）、モバイル テクノロジー、モバイル アプリケーションなどの新たなテクノロジーによって、さまざまな業種の組織において想像もしていなかったスピードとボリュームでデータが作り出されています。多くの組織は、人工知能（AI）を使用して、そのようなすべてのデータを、より適切で迅速な意思決定を可能にする原動力に変えることを検討しています。AI が実現するインサイトによって、組織はリスクを軽減し、価値を創出し、新たな機会を見いだすことができます。ただし、AI ワークロードを実行可能なシステムは、導入が複雑であり、ハードウェアとソフトウェアの広範な統合とテストを必要とします。

これらの課題を解決するために、Dell Technologies は、Intel、H2O.ai と共同で、H2O® Driverless AI エンタープライズ プラットフォームと、H2O.ai オープンソース プラットフォームである H2O と H2O Sparkling Water 向けのリファレンス アーキテクチャを開発し、ベンチマークを行っています。数千ものグローバル企業が H2O.ai のソフトウェア ソリューションに信頼を置いています。これは同社のソリューションが、勾配ブースティング マシン、一般化線形モデル、ディープ ラーニング（DL）など、最も広く使用されている統計アルゴリズムと機械学習（ML）アルゴリズムに対応しているためです。ベンチマーク テストでは、これらの Dell Technologies のリファレンス アーキテクチャで、高い精度を維持しつつパフォーマンスが大幅に向上することが示されています。

これらのリファレンス アーキテクチャを使用することで、IT チームは、Dell EMC のエンタープライズ クラスのサーバー、ストレージ、ネットワーキングと H2O ソフトウェアの組み合わせを安心して活用することができ、予測分析機能を利用して自動化と ML タスクを最適化することができます。この組み合わせは、エンタープライズ クラスの AI システムを導入する際のコストと複雑さを軽減し、専門知識なしでも AI を使えるようにし、より多くの組織が AI を活用して、カスタマー エクスペリエンスを向上させ、プロセスを効率化し、不正や無駄を減らすことができるように設計されています。

Dell Technologies、Intel、H2O.ai は、データ サイエンティストを支援し、AI の導入を促進するために、エンジニアリング検証済みのリファレンス アーキテクチャを共同開発しています。

ソリューションの概要

このホワイトペーパーでは、Dell Technologies と H2O.ai のコラボレーションによって提供される、さまざまなデータ サイエンス プラットフォームと ML のリファレンス アーキテクチャについて説明しています。

検証済みソフトウェア

検証済みソフトウェアとして、H2O Driverless AI エンタープライズ プラットフォームと、H2O と H2O Sparkling Water のオープンソース ソフトウェア プラットフォームがあります。これらの H2O のソフトウェア プラットフォームは、高度に最適化されたインテルのライブラリーと ML フレームワークを活用して、Dell EMC インフラストラクチャ上でのパフォーマンスが強化されています。

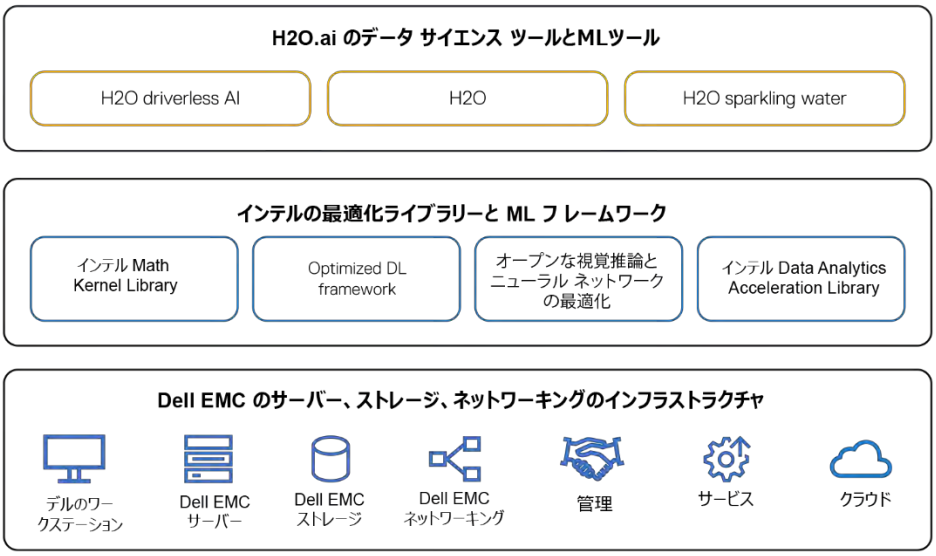


図 1： H2O.ai、Intel、デルによる AI と ML のプラットフォーム

IT を変革する革新的な設計

Dell EMC PowerEdge サーバーは適応性を備えた完全な IT ソリューションの基盤であり、これを導入することで、優れたパフォーマンスと容易な管理機能により、お客様の成功に必要となるビジネス アプリケーションをより効率的かつ効果的に実行できます。

AI 向けインテル Xeon スケーラブル プロセッサー

インテル Xeon プロセッサーは、優れたパフォーマンス、柔軟性、拡張性、TCO の削減を実現します。ハードウェア ベースの AI アクセラレーションの最近の進歩と、AI フレームワークと AI 特化ライブラリーに対するソフトウェアの最適化により、インテル Xeon スケーラブル プロセッサーを搭載した Dell EMC PowerEdge サーバーは、企業にとって利用実績が豊富で信頼を置いている CPU プラットフォーム上で優れたパフォーマンスと拡張性を提供します。

インテル DAAL は、分析のすべてのステージで役立ちます。

- 前処理：解凍、フィルタリング、正規化
- トランスフォーメーション：集約と次元縮小
- 分析：要約統計とクラスタリング
- モデリング：トレーニング、パラメーター推定、シミュレーション
- 検証：仮説検証とモデル誤差検出
- 意思決定：予測と決定木

検証済みハードウェア

Dell EMC PowerEdge R740xd は、2 ソケットの 2U ラック サーバーであり、メモリー、I/O 容量、ネットワークの拡張性に優れたオプションを使用して、複雑なワークロードを実行できるように設計されています。膨大なストレージ容量に対応したオプションが提供されているため、大量のストレージを必要とするデータ集約型アプリケーションに最適であり、I/O パフォーマンスが犠牲になることもありません。

インテル DC P4600 NVMe SSD は、インテル 3D NAND SSD であり、卓越した品質、信頼性、高度な管理機能と保守機能性が提供されるため、サービスの中断を最小限に抑えることができます。

インテル Xeon スケーラブル プロセッサーは、要求の厳しいデータ センターのワークロード向けに最適化されています。このプロセッサー ファミリーでは、前世代のインテル Xeon プロセッサーよりも周波数が高くなり、アーキテクチャが改良され、AI と DL の推論ワークロードが強化されています。

第 2 世代インテル Xeon スケーラブル プロセッサーでは、インテル Deep Learning（DL）Boost によって AI パフォーマンスが次のレベルに引き上げられており、ベクトル ニューラル ネットワーク命令（VNNI）が追加されてインテル アドバンスト ベクトル エクステンション 512（インテル AVX-512）命令セットが拡張されています。インテル DL Boost によって、VNNI を使用するように最適化された DL ワークロードの推論パフォーマンスが大幅に高速化されており、前世代のインテル Xeon スケーラブル プロセッサーと比較して 30 倍もの速度向上を達成するケースもあります。

最適化された機械学習ライブラリー

インテル Data Analytics Acceleration Library（インテル DAAL）は、アプリケーションによる予測を高速化し、コンピューティング リソースを増加させることなく大規模なデータ セットを分析できるようにする、使い勝手のよいライブラリーです。ハイ パフォーマンスを実現できるように、データの取り込みとアルゴリズムのコンピューティングの両方が最適化されます。また、アプリケーションのさまざまなニーズに対応できるように、オフライン、ストリーミング、分散型の使用モデルもサポートされています。

ハイ パフォーマンスのロジスティック回帰、拡張 GBM 機能、ユーザー定義のデータ変更プロシージャなどの新機能があります。

デルの H2O Driverless AI 向けリファレンス アーキテクチャ

H2O Driverless AI は、最先端の予測分析モデルの自動開発と迅速な導入を可能にする、ハイ パフォーマンスの単一ノードのエンタープライズ プラットフォームです。選択されたデータ セットに対する最適な予測分析モデルを自動的に構築し選択することによって、データ サイエンス プロジェクトをシンプルにします。

Driverless AI によって、ユーザーはグラフィカル ユーザー インターフェイス（GUI）を使用して数回クリックするだけで、モデリング パイプラインをトレーニングして導入することができます。上級ユーザーは、Python®や Java®などのさまざまな言語を使用して、クライアント/サーバーAPI を活用することができます。また、Driverless AI では、機械学習解釈可能性（MLI）機能セットの一部として、理由コードと説明を使用した、統計的に厳密な自動データ可視化とインタラクティブなモデル解釈を行うことができます。

H2O Driverless AI は、Apache® Hadoop®分散ファイル システム（HDFS）、Apache Spark®、Amazon® Simple Storage Service（S3）、Microsoft® Azure® Data Lake、または他のデータ ソースから、インメモリー分散型キー バリュー ストアにデータを直接取り込むことができます。さらに、最適なモデルの生成を支援するために、アルゴリズムとそのハイパー パラメーターを使用して AutoML 機能が自動的に実行されます。

Dell Technologies は、ビジネス クリティカルなワークロードの導入に使用される、業界をリードするサーバーとデータ ストレージにより、長年にわたってデータ分析のパイオニアであり続けています。デルは、信頼性とセキュリティを備えたデータ主導型の組織を支援できるように最適化され高速化された AI ソリューションの必要性を理解しています。そのため、デルのエンジニアリング チームは、Intel と H2O.ai との共同作業により、企業における AI 機能を強化するように調整されている、H2O Driverless AI 向けのリファレンス アーキテクチャを作成しました。

図 2 に、以下のステップがある自動化ワークフローで H2O Driverless AI がどのように動作するかを示します。

1. **データのドラッグ アンド ドロップ：**さまざまなクラウドやデスクトップのデータ ソースからのプレーン テキスト ソースを表形式データに変換します。HDFS、SQL Server®、Amazon S3、Snowflake、Google® BigQuery™、Azure Blog Storage、ローカル データ ストレージにあるデータを取り込むことができます。
2. **モデリングと可視化：**データ セットを作成し、モデル構築プロセスを始める前にユーザーがデータを素早く理解できるように、最も関連度の高いデータに基づいて可視化します。
3. **自動モデル最適化：**ベスト プラクティス モデルとハイ パフォーマンス コンピューティング（HPC）を融合させることによって、そのデータに最適なモデルを決定します。
4. **自動スコアリング パイプライン：**Python や Java によるスコアリング パイプラインなどの低レイテンシーのパイプラインを使用して、機能変換と検証、チューニング、選択、導入のためのモデルなどの実験を行います。

自動機能エンジニアリング、機械学習、解釈可能性

DRIVERLESSAI

出典：h2o.ai

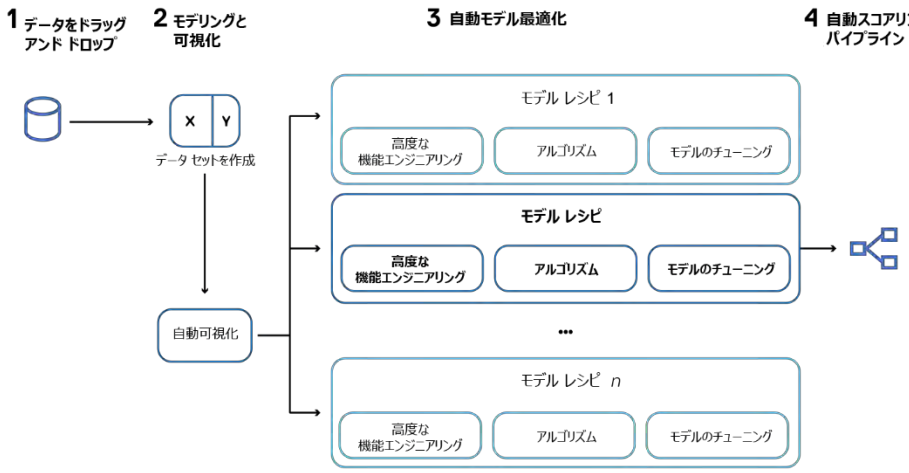


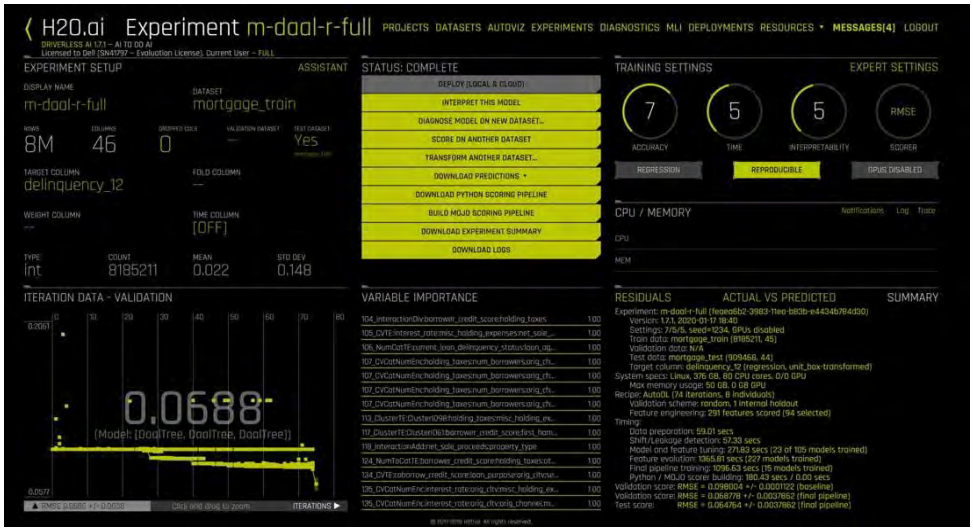
図 2：H2O Driverless AI のワークフロー

H2O Driverless AI は、ビジネス アナリスト、データ サイエンティスト、AI を活用したいと考えているドメイン ユーザー向けに設計されており、エンド ツー エンドのデータ サイエンス用の GUI から操作します（図 3 を参照）。

H2O Driverless AI は、堅牢でハイ パフォーマンスで革新的で検証済みの、以下のような機能により、時間、コスト、信頼性の課題を解決することに成功しています。

- **自動機能エンジニアリング：**データ サイエンティストが複数のアルゴリズムから最も正確なデータを取得できるようにします。アルゴリズムと機能変換のライブラリーにより、このソフトウェア プラットフォームでは、指定されたデータ セットに対して機能をうまく構築し設計することができます。
- **データと導入の柔軟性：**Hadoop HDFS、Amazon S3 など、さまざまなデータ ソースがサポートされています。これは、Microsoft Azure、Amazon Web Services®（AWS）、Google Cloud Platform™などのクラウドに導入することができます。
- **自動データ可視化（AutoVis）：**最も関連度の高いデータ統計情報に基づいてデータ プロットを自動的に選択して、ユーザーが非常に大規模なデータ セットの構成を理解し、傾向と、モデリング結果に影響を与える可能性のある問題を見つけるのを支援します。

- **自動モデル ドキュメント作成（AutoDoc）：**データ サイエンティストやデータ エンジニアが関与することなく、各実験のレポートを生成します。AutoReport には、使用されたデータ、選択された検証スキーマ、モデルと機能のチューニング、作成された最終的なモデルに関する詳細が含まれています。



出典：h2o.ai

図 3：H2O Driverless AI のユーザー インターフェイス

- 時系列のレシピ：ほぼすべての予測期間に対して最適化するための時系列の予測機能を提供します。構造化文字データを提供して、時系列データや他の欠落している値のギャップを埋めます。
- **TensorFlow™による自然言語処理（NLP）：**テキスト文字列を機能に自動的に変換します。TensorFlow を使用して、より大きなテキスト ブロックを処理し、利用可能なすべてのデータを使用してセンチメント分析、文書分類、コンテンツのタグ付けなどのビジネス上の問題を解決するためのモデルを構築します。
- **自動スコアリング パイプライン：**完了済みの実験に対して、Python のスコアリング パイプラインと、新しい超低レイテンシーの自動スコアリング パイプラインの両方を生成します。高度に最適化された低レイテンシーの実稼働対応の、どこにでも導入できる Java コードに、機能エンジニアリングと最適な ML モデルを導入します。
- **ML 解釈可能性（MLI）：**成果の信用と認証が形成されるように、AI 主導のビジネス上の意思決定とモデリング結果に関する正確な説明を可読形式で提供します。さまざまな手法と方法論を使用して解釈を行います。
- **自動理由コード：**顧客に影響を与える重要な意思決定を企業が説明できるように、モデルのスコアリングの決定における主要な肯定的要因と否定的要因が分かりやすい言葉で示されます。
- **カスタム レシピのサポート：**ML アルゴリズム、機能エンジニアリング、スコアラー、構成のカスタム レシピをインポートして、個別に使用することも、組み込みレシピと組み合わせて使用することも、組み込みレシピの代わりに使用することもできます。自動 ML パイプラインと最適化の選択肢に大きな影響を与えます。
- **GPU サポート：**オプションの GPU アクセラレーションを活用して、自動 ML を高速化します。XGBoost、GBM、GLM、K-Means などのマルチ GPU アルゴリズムがサポートされています。

Driverless AI は、企業向けの高速で正確で解釈可能な AI を提供します。

自動化された機械学習（AutoML）は、機械学習を現実世界の問題に適用するエンド ツー エンドのプロセスを自動化するプロセスです。

リファレンス アーキテクチャと実装

このセクションで説明しているリファレンス アーキテクチャは、代表的な H2O Driverless AI ソリューションの基本構成を表しています。

ハードウェアとソフトウェアの構成の概要

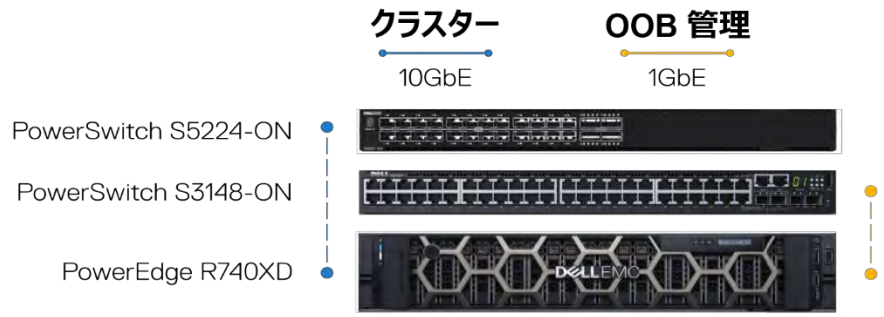


図 4：ハードウェアとソフトウェアの設定

H2O Driverless AI のコンピューティング ノード	
サーバー	Dell EMC PowerEdge R740xd x 1
プロセッサー	インテル Xeon Gold スケーラブル 6248 x 2
メモリー	384GB DDR4 @ 2667MHz
ドライブ	オペレーティング システム：480GB SSD x 2 を搭載した BOSS カード データ：Dell Express Flash NVMe 4610 1.6TB SFF x 12
ネットワーキング	インテル Ethernet 10G 4P x710 SFP+ rNDC
トップオブラック（ToR）スイッチ	管理：Dell Networking PowerSwitch 3148-ON（1GbE） クラスター：Dell Networking PowerSwitch S5224-ON（10/25GbE）
ソフトウェア	**バージョン**
オペレーティング システム	Red Hat® CentOS® Linux® 7.6 または Red Hat Enterprise Linux（RHEL）7
H2O ソフトウェア	H2O Driverless AI 1.7.1
ライブラリー	インテル DAAL バージョン 2019.5

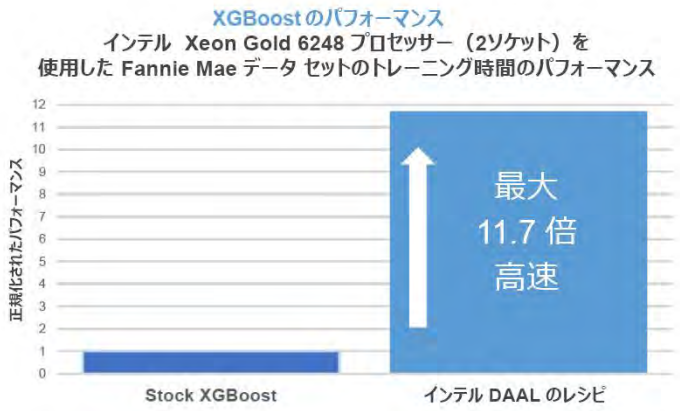
インテル DAAL のレシピを使用した CPU アクセラレーション モデルの構築

デルと H2O は、第 2 世代インテル Xeon スケーラブル プロセッサーの最新のパフォーマンス機能を活用するために、ML モデルの構築にインテル DAAL のレシピを使用することを推奨しています。Intel は、企業が迅速かつ大規模に ML を実現できるように、DAAL のレシピを作成して、H2O.ai のオープンソースのレシピ リポジトリーで公開しました。オープンソースの DAAL のレシピを使用すると、Stock アルゴリズムに比べてパフォーマンスが大幅に向上します。

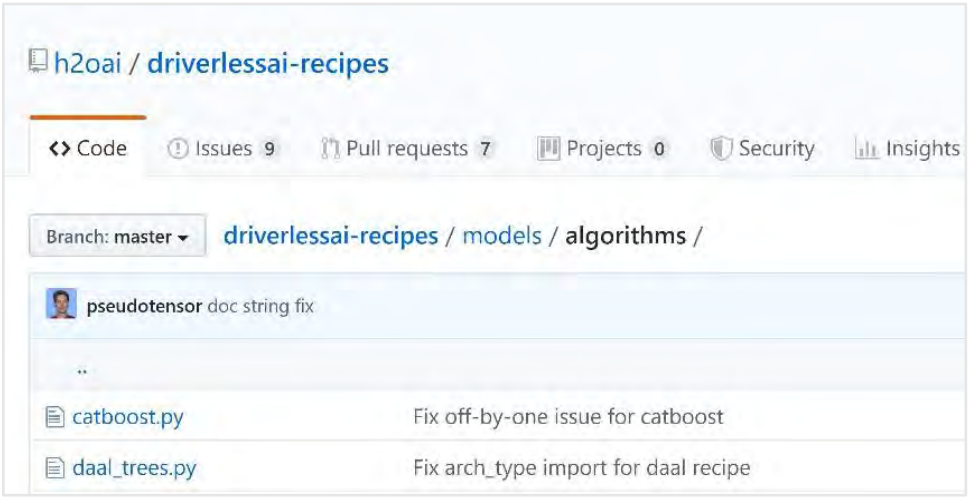
DAAL を使用してパフォーマンス向上を測定するために、H2O Driverless AI を使用して、Fannie Mae®の一戸建て住宅のローン運用実績データ セットを使用してモデルを構築し、借り手の債務不履行状況を予測しました。モデルは Driverless AI で利用可能な Stock XGBoost アルゴリズムを使用してトレーニングされていて、H2O.ai の GitHub から入手可能な DAAL のレシピを使用して構築されたモデルと比較しました（図 5 を参照）。

インテル DAAL では、1 時間以内にトレーニングが完了し、Fannie Mae の 800 万レコードのデータ セットでの債務不履行状況を予測するモデルが生成されました。Stock XGBoost アルゴリズムでは、モデルのトレーニングに約 10 時間かかりました。観察の結果、インテル DAAL では、正確さを維持しながら、トレーニング プロセスが Stock XGBoost モデルに比べて 11.7 倍も高速化されることがわかりました。

H2O Driverless AI の自動化により、データ サイエンティスト、データ エンジニア、ドメイン サイエンティストは、より迅速かつ効率的にプロジェクトに取り組むことができるようになります。



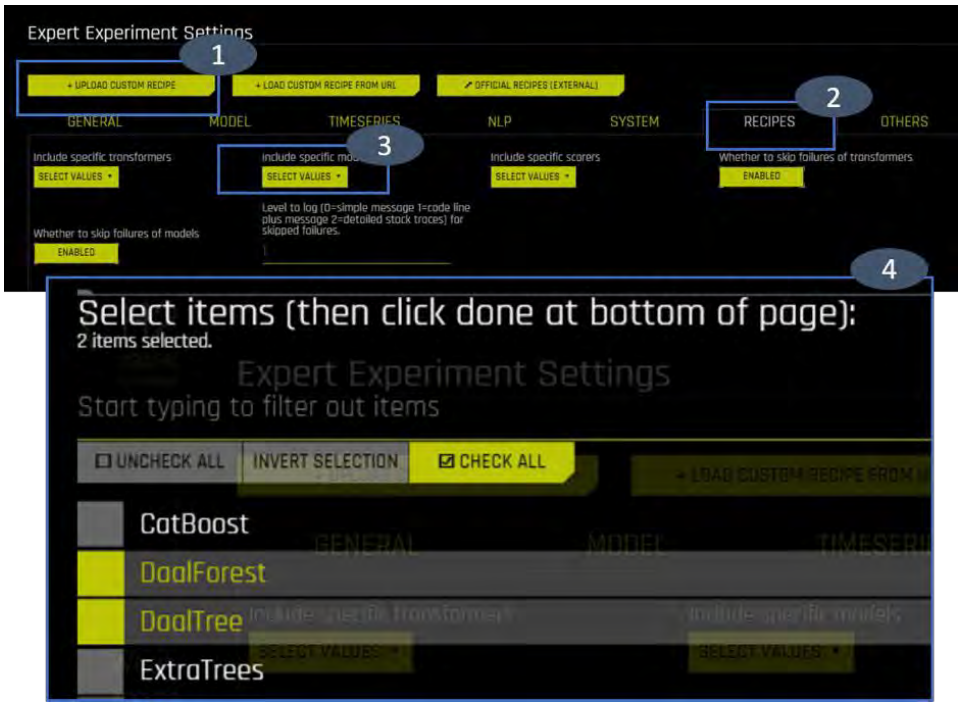
インテル DAAL のレシピは、Driverless AI の GitHub のカスタム レシピのリポジトリーから入手できます（図 5 を参照）。



出典：h2o.ai

図 5：H2O.ai の GitHub リポジトリーにある Driverless AI のレシピ

Driverless AI のユーザー インターフェイスでマウスを数回クリックするだけで、Driverless AI でインテル DAAL を利用できます。DAAL のレシピをインポートするための 4 つの手順を図 6 に示しています。



出典：h2o.ai

図 6：インテル DAAL のレシピを Driverless AI にインポートする手順

H2O Driverless AI は自動化を使用することにより、データ サイエンティスト、データ エンジニア、ドメイン サイエンティストが、より迅速かつ効率的にプロジェクトに取り組むことができるようにします。以前は数か月かかったタスクが、H2O Driverless AI を使用することによって、数時間または数分に短縮できます。主な機能として、自動機能エンジニアリング、モデルの検証、モデルのチューニング、モデルの選択と導入、MLI、時系列、NLP、モデル スコアリング用の自動パイプライン生成があります。ユーザー独自のレシピを持ち込む機能によって、ユーザーはインテル DAAL を活用してトレーニング時間を短縮でき、AI プロジェクトの実施と実装にかかる時間を短縮することができます。

デルの H2O オープンソース プラットフォーム向けリファレンス アーキテクチャ

H2O.ai のオープンソース プラットフォームは、Apache v2 ライセンスに基づいており、エンタープライズ サポート サブスクリプション付きで提供されています。データ サイエンティスト向けに設計されており、R と Python がサポートされ、H2O Flow と呼ばれるインタラクティブな GUI がサポートされています。H2O は、既存の Big Data インフラストラクチャと併用できる、インメモリーの分散型 ML アルゴリズムをまとめたものです。ベアメタル、Apache Hadoop、Apache Spark クラスターが含まれることがあります。HDFS、Spark、Amazon S3、Azure Data Lake などのデータ ソースからデータをインメモリーの分散型キー バリュー ストアに直接取り込むことができます。H2O は、次の表に示している 3 つの方法で導入することができます。

H2O オープンソース プラットフォームの導入の選択肢

導入方法	詳細
H2O	H2O は計算の分配とノード間通信を担当しています。
H2O Sparkling Water	H2O は Apache Spark と統合されています。H2O ジョブは、Spark マスターに送信されてから、Spark エグゼキューター内で実行されます。
H2O on Hadoop	YARN（Yet Another Resource Negotiator）は、H2O ジョブを Hadoop 上の MapReduce タスクとしてスケジュールします。

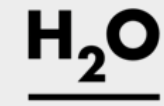
Dell Technologies のエンジニアリング チームは、H2O と H2O Sparkling Water 向けのリファレンス アーキテクチャを作成して、ベンチマークを行いました。

H2O の主な機能は次のとおりです。

- **アルゴリズムの信頼性：**分散コンピューティング環境は、新規に開発されたアルゴリズムに依存しています。H2O では、ランダム フォレスト、GLM、XGBoost、GBM、DL、GLRM（Generalized Low Rank Models）、Word2Vec、その他多数のアルゴリズムがサポートされています。
- **言語への適合性：**H2O では、R、Python など、モデルを構築するために広く使用されているプログラミング言語がサポートされています。これにより、開発者の作業が容易になります。また、コーディングを必要としない直感的な GUI である H2O Flow を利用することもできます。
- **ワークフローの自動化：**H2O の AutoML 機能により、ML ワークフローが自動化されて、指定したタイムライン内で多数のモデルの自動トレーニングと自動チューニングが可能になります。H2O の Stacked Ensemble 機能により、パフォーマンスが上位のモデルを特定できます。
- **パフォーマンスの強化：**H2O は、データの正確性を損なうことなく、膨大なデータ セットを管理できます。これは、ノードとクラスターをシリアル化するインメモリー処理によって行われます。Big Data の分散処理により、スピードと効率性が向上します。
- **簡単な導入：**H2O は、どのような環境でも高速かつ正確なスコアリングを実現できるように、導入が簡単な Java モデルを利用しています。

H2O Sparkling Water の主な機能は次のとおりです。

- **アルゴリズムへの容易なアプローチ：**H2O Sparkling Water を使用すると、開発者は分散コンピューティング環境向けの多様な H2O アルゴリズムを利用できるようになります。教師ありと教師なしの両方のアプローチで、ランダム フォレスト、GLM、GBM、XGBoost、GLRM、Word2Vec、その他多数のアルゴリズムを利用できます。
- **適応性：**H2O Sparkling Water を使用すると、アプリケーション開発者に理想的な ML プラットフォームが提供されて、Scala®、R、Python からのシームレスな計算が可能になります。また、オープンソースのユーザー インターフェイスである H2O Flow を利用することもできます。
- **簡単な導入と正確なスコアリング：**Sparkling Water では、導入が簡単で高精度の Java スコアリング モデルが利用されています。



H2O Flow の GUI を使用したインメモリーの分散型機械学習アルゴリズム

出典：h2o.ai

H2O AI オープンソース エンジンの Spark との統合

出典：h2o.ai

H2O オープンソース プラットフォーム向けのハードウェアとソフトウェアの構成

Dell Technologies の H2O オープンソース プラットフォーム向けリファレンス アーキテクチャでは、システム設計への柔軟なビルディング ブロック アプローチが採用されているため、ユーザーは個々の構成要素を組み合わせて、ユーザー特有のワークロードやユースケースに最適化されたシステムを構築することができます。このソリューションは、Sparkling Water を実行する H2O クラスターまたは Spark クラスターを高可用性環境に導入できるよう、3 台以上の Dell EMC PowerEdge R740xd サーバーで構成されます。Dell Technologies のエンジニアリング チームは、5 ノードまでの構成でテストとベンチマークを行いました（図 7 を参照）。

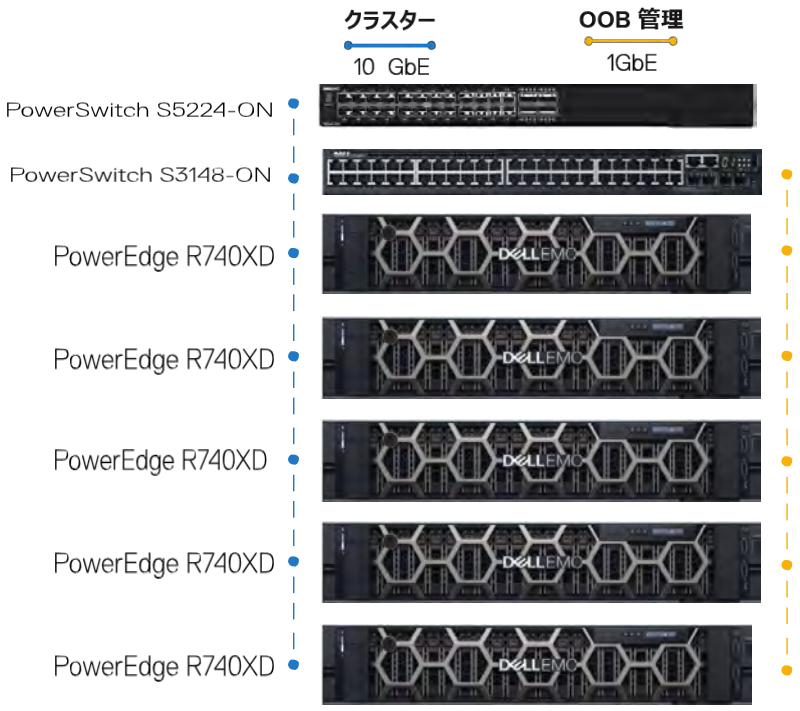


図 7：H2O と H2O Sparkling Water 向けのハードウェアとソフトウェアの構成

H2O.ai オープンソース プラットフォームのコンピューティング ノード	
サーバー	Dell EMC PowerEdge R740xd x 3～5
プロセッサー	インテル Xeon Gold スケーラブル 6248 x 2
メモリー	384GB DDR4 @ 2667MHz
ドライブ	オペレーティング システム：480GB SSD x 2 を搭載した BOSS カード データ：Dell Express Flash NVMe P4610 1.6TB SFF x 12
ネットワーキング	インテル Ethernet 10G 4P X710 SFP+ rNDC
ToR スイッチ	管理：Dell Networking PowerSwitch S3148-ON（1GbE）、クラスター：Dell Networking PowerSwitch S5224-ON（10/25GbE）
ソフトウェア	**バージョン**
オペレーティング システム	CentOS Linux Release 7.6 または RHEL 7
H2O ソフトウェア	H2O Flow（3.26.0.3）、JDK12、Sparkling Water 2.3.1
Cloudera®クラスター	CDH 5.16.2 Spark 2.3.0
ライブラリー	インテル DAAL バージョン 2019.5

H2O オープンソース プラットフォームでのトレーニングと推論のパフォーマンス

H2O Sparkling Water クラスターのパフォーマンスは、3 ノードと 5 ノードのクラスターで評価されました。このモードでは、H2O は Spark ワーカーを介して起動し、Spark がジョブ スケジューリングとノード間の通信を管理します。インテル Xeon Gold 6248 プロセッサーを搭載した Dell EMC PowerEdge R740xd サーバー3 台と同サーバー5 台を使用し、Fannie Mae の一戸建て住宅のローン運用実績データ セットから得られた住宅ローンのデータ セットを使用して、XGBoost と GBM のモデルをトレーニングしました。

H2O の XGBoost モデルと GBM モデルは、回帰と分類の問題によく使用されており、追加 CPU コアとシステム メモリーを拡張して利用するように最適化されています。追加の Spark ワーカー ノードがモデルのトレーニングに使用されている場合、予測モデルを構築するためのトレーニング時間が長くなります。

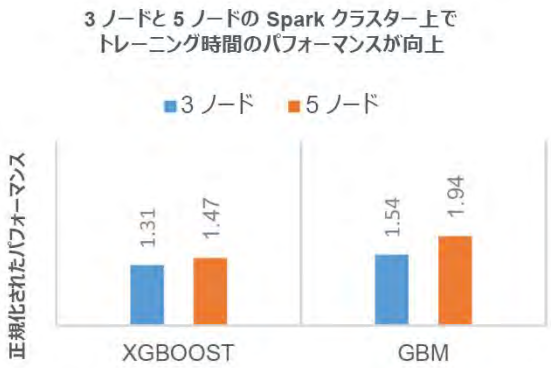


図 8： Sparkling Water クラスター上のデータ セットのトレーニング時間の改善

H2O Sparkling Water には、指定されたタイムライン内で多数のモデルの自動トレーニングと自動チューニングを行うための AutoML 機能も含まれています。この機能を使用するために、複数のモデルに対して評価とチューニングを行ってパフォーマンスが上位のモデルを特定するタイムラインが指定されています。

この機能を使用して、1 時間と 2 時間のタイム ウィンドウ内で XGBoost モデルの自動チューニングを実行しました。このモデルは、データ サイエンティストがハイパー パラメーターを使用してトレーニングしたモデルと比較されました。ハイパー パラメーターは、データ サイエンティストが実験の繰り返しと手動での最適化を経て選択したものです。通常、これらのタスクを手動で実行する場合、トレーニングされるモデルとデータ セットによっては数日から数週間かかることがあります。H2O の AutoML 機能は、図に示しているように、XGBoost モデルのチューニング パラメーターを特定するのに有効であることが確認されています。

トレーニング済みモデルの精度の比較を図 9 に示しています。H2O オープンソース プラットフォームの AutoML 機能によって、専門家ではないユーザーが使用できて使い勝手のよい ML ソフトウェアに分類されています。H2O AutoML のインターフェイスはパラメーター数ができる限り少なくなるように設計されているため、ユーザーは H2O の Flow UI でデータ セットを指定し、応答列を特定し、必要に応じて、時間的制約またはトレーニングされるモデルの総数の上限を指定するだけで済みます。

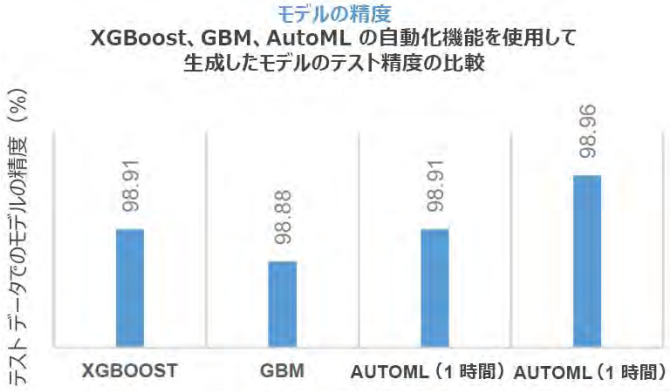


図 9： ワークフロー自動化の手法を使用したモデルの正確さの比較

Sparkling Water で予測モデルがトレーニングされたら、Sparkling Water の推論機能を使用して推論を実行し、トレーニング済みのモデルを評価することができます。トレーニングを実行する前にデータ セットが分割されているため、そのモデルにとって新しいデータに対して推測を実行することができます。このモデルは、住宅ローンのデータ セットの 80%を使用してトレーニングされ、残りの 20%に対して推論が実施されました。1 つの Spark ワーカー ノードでの推定速度は、1 秒間あたり 170 万レコードであることが観測されました。

図 10：単一ノードの Sparkling Water を使用して測定された推定スループット

Sparkling Water だけでなく、スタンドアロンの H2O クラスターを使用した場合のパフォーマンス向上も評価しました。このシナリオでは、H2O ソフトウェアは、モデルが複数のクラスター ノードを使用してトレーニングされている場合に、計算の分配と通信の処理を担当します。単一ノードで GBM モデルをトレーニングした場合を基準として、3 ノード クラスターで GBM モデルをトレーニングした場合に得られるスピードアップを図 11 に示しています。

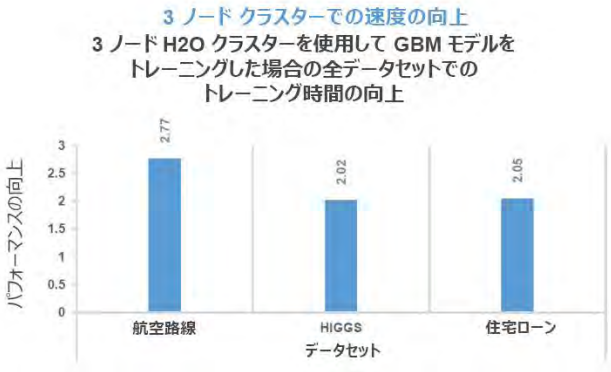


図 11：H2O クラスターでの GBM のトレーニング時間の改善

住宅ローンのデータ セットを使用してローンの債務不履行を予測できるモデルを構築するベンチマークに加えて、航空路線のデータ セット [1] を使用して、フライトの遅延またはキャンセルの可能性を予測するための GBM モデルを構築し、HIGGS データ セット [2] を使用して、ヒッグス粒子を生成する信号処理を特定するモデルを構築して、ベンチマークを行いました。

[1] 航空路線のデータ セットに関する情報は次のページで入手できます。 https://github.com/h2oai/h2o-2/wiki/Hacking-Airline-DataSet-with-H2O.

[2] HIGGS データ セットは https://archive.ics.uci.edu/ml/datasets/HIGGS からダウンロードできます。

H2O.ai について

H2O.ai は、オープンソースの H2O ソフトウェア プラットフォームから、Sparkling Water による Apache Spark との統合や受賞歴のある H2O Driverless AI プラットフォームまで、AI とデータ サイエンスのさまざまなプラットフォームを、困難な状況にいる専門家のデータ サイエンティストに提供しています。

詳細情報：h2o.ai

デルのインフラストラクチャでの H2O.ai の機械学習ソリューションの導入

Dell Technologies は、お客様が AI トランスフォーメーションを加速できるように、H2O.ai を使用したリファレンス アーキテクチャを開発しました。H2O Driverless AI は、インテル Xeon スケーラブル プロセッサーの多数のコアとアーキテクチャを活用する最先端の予測分析モデルの自動開発と迅速な導入を可能にする、ハイ パフォーマンスのコンピューティング ソフトウェア プラットフォームです。

Sparkling Water は、より大規模なクラスターを処理ニーズに合わせて活用し、データを Spark と H2O の間でシームレスに転送する必要があるお客様に最適です。Dell Technologies のエンジニアリング チームは、Dell EMC PowerEdge サーバー インフラストラクチャを使用して 3 ノードと 5 ノードのクラスターをテストしました。その結果では、大規模なデータ セットを使用して複雑な予測モデルをトレーニングする場合にパフォーマンスが拡張され、トレーニング時間が短縮されることが示されています。また、テストで H2O Sparkling Water の新しい AutoML 機能を評価して、多様なユースケースで高いレベルの正確さを実現することができました。

H2O Driverless AI は H2O の商用製品であり、自動化によって AI プロジェクトをより迅速かつ効率的に完了させることができます。H2O Driverless AI を使用すると、自動機能エンジニアリング、モデルの検証、モデルのチューニング、モデルの選択、導入を数時間または数分で行うことができます。さらに、MLI、時系列、NLP、およびモデルのスコアリング用の自動パイプライン生成は、Driverless AI で利用できる有用な機能です。カスタム レシピを使用すると、Driverless AI の自動化されたアプローチを、データ サイエンティストの専門分野の知識によって強化することができます。インテル DAAL のレシピを利用すると、XGBoost モデルのトレーニング時間を大幅に短縮できることを確認しました。

堅牢な機能を備えている H2O はパワフルな AI ツールです。Dell Technologies の H2O Driverless AI と H2O オープンソース プラットフォーム向けリファレンス アーキテクチャでは、最新のインテル プロセッサーと組み合わされており、コスト効率に優れたソリューションと最適化されたデータ管理を提供することによって、AI 導入の実現可能性が高まります。

必要なときに支援を提供

Dell Technologies は、戦略から実装、そして継続的な最適化まで、データ分析と AI のサービスを提供します。望みどおりのビジネス成果を迅速かつ適正な規模で達成するために必要となる、人材とプロセス テクノロジーの橋渡しを支援します。その支援には、AI テクノロジーの実装と事業化、データ エンジニアリング能力の加速の支援などがあります。

詳細情報

Delltechnologies.com/referencearchitectures

Delltechnologies.com/ai

インテル SSD データ センター ファミリー

Intel AI Builders

DELL Technologies

intel