DELL Technologies

intel

ホワイトペーパー

DataRobot を Dell EMC インフラストラクチャで使用して AI 主導のエンタープライズを実現

オンプレミスで利用できるデータ サイエンス プラットフォームのリファレンス アーキテクチャ

要旨

Dell Technologies、DataRobot®、Intel®の 3 社は、信頼できるインフラストラクチャ上で組織が AI トランスフォーメーション プロジェクトを実装できるように、オンプレミスのエンタープライズ人工知能（AI）ソリューション スタックを提供しています。

2020 年 2 月

目次

概要 3

ソリューションの概要 3

Dell EMC インフラストラクチャで強化される DataRobot 機械学習ソフトウェア 3

リファレンス アーキテクチャと実装 6

Dell EMC インフラストラクチャ上の DataRobot エンタープライズ ML プラットフォームの価値 14

ガバナンスと透明性 14

実稼働環境へのモデルの配置 14

まとめ 16

詳細情報 16

概要

機械学習（ML）と AI への関心は高まる一方ですが、これはアルゴリズムが空前の進歩を遂げたことが要因となっています。AI を採用したいという願望が組織内で高まり、それが実際のビジネス価値へと形を変えるにつれて、データ サイエンスの人材が不足してきています。そのため、DataRobot のエンタープライズ AI プラットフォームなどの、自動化されたソリューションの導入が増加しています。エンタープライズ/ビジネス AI プラットフォームを使用することで、社内にいる各領域の専門家を補強し、既存のビジネス プロセスとデータ パイプラインの中で AI プロジェクトを作成、導入、管理することができます。

本書では、組織がオンプレミスのエンタープライズ AI プラットフォームを実装するために活用できる DataRobot を取り上げ、Dell EMC インフラストラクチャにおけるアーキテクチャと実装について説明します。このエンタープライズ AI プラットフォームは、既存の Big Data やデータ レイクのプラットフォームと統合することも、スタンドアロンのマルチ ユーザー環境として単独で実行して複数のソースからデータを取り込むこともできます。組織の既存のデータ ソースとのこのような緊密な統合により、組織内の複数のチームが、より迅速に ML を実装して、効率的に使用できるようになります。

datarobot.com

ソリューションの概要

DataRobot は、エンタープライズ AI のリーダーであり、今日のインテリジェンス革命でしのぎを削っているグローバル企業に信頼できる AI ソフトウェアと AI 対応サービスを提供しています。DataRobot のエンタープライズ AI ソフトウェアにより、ML モデルの構築、導入、管理がエンドツー エンドで自動化されるため、専門知識がなくてもデータ サイエンスを利用できるようになります。DataRobot、Dell Technologies、Intel は共同で、AI を導入したい、または AI の使用を加速してビジネス価値とプロセスを改善したいと考えている組織のニーズに応えるために、拡張性の高いリファレンス アーキテクチャを設計してきました。DataRobot を Dell EMC インフラストラクチャに装備することによって、一般のデータ サイエンティストが、特定のアルゴリズムをいつどのように適用するかについて理解したりコーディングを習得したりすることなく、高度な機械学習モデルを作成する能力を得られます。また、モデル構築プロセスでの反復手順が自動化されているため、データ サイエンティストは生産性を向上し、独自の専門知識をモデルの選択や微調整に使用できるようにもなります。

Dell EMC インフラストラクチャで強化される DataRobot 機械学習ソフトウェア

DataRobot ソフトウェアは、AI を大規模な環境で提供し、時間が経過してもパフォーマンスを継続的に最適化することによって、**ビジネス価値を最大限に高めます。**DataRobot の最先端のソフトウェアと世界レベルの AI の実装、トレーニングとサポート サービスの実績のある組み合わせによって、規模、業種、リソースを問わずあらゆる組織が、AI によるビジネス成果の向上を推進できるようになります。

Dell EMC PowerEdge サーバーはデータ センターの基盤です。PowerEdge サーバーのポートフォリオでは、アプリケーションのパフォーマンスを最適化するための柔軟な設計が利用できます。シングル ソケット サーバーのポートフォリオでは、将来的な成長を視野に入れ、バランスのとれたパフォーマンスとストレージ容量を用意しています。2 ソケット サーバーのポートフォリオでは、コンピューティングとメモリーの最適なバランスにより、パフォーマンスを最大限に高め、将来の需要に応じて拡張でき、ほぼすべてのワークロードに適応できるようにするための機能が組み合わされています。Dell EMC の 4 ソケット サーバーのポートフォリオでは、最上位のサーバーでアプリケーションに対して最高のパフォーマンスと大規模な拡張性を実現、インメモリー データベースのワークロードやハイ パフォーマンス コンピューティング（HPC）から、データ分析、AI、GPU データベース アクセラレーションまで広範囲に対応します。

インテル® からは、インテル Xeon スケーラブル プロセッサー向けに最適化されたライブラリーとフレームワークが提供されます。TensorFlow™、MXNet、PaddlePaddle、Caffe、PyTorch® などが含まれます。それらを使用したソフトウェア最適化によって、DL のパフォーマンスが向上します。たとえば、Python 用のインテル ディストリビューションでは、インテル MKL などの統合インテル パフォーマンス ライブラリーで NumPy、SciPy、scikit-learn などの AI 関連の Python ライブラリーが高速化されるため、AI のトレーニングと推論が加速されます。DataRobot はこれらのライブラリーとフレームワークを統合して、信頼性が高くパワフルで包括的な AI 開発ソフトウェア プラットフォームを企業が迅速に導入できるようにします。

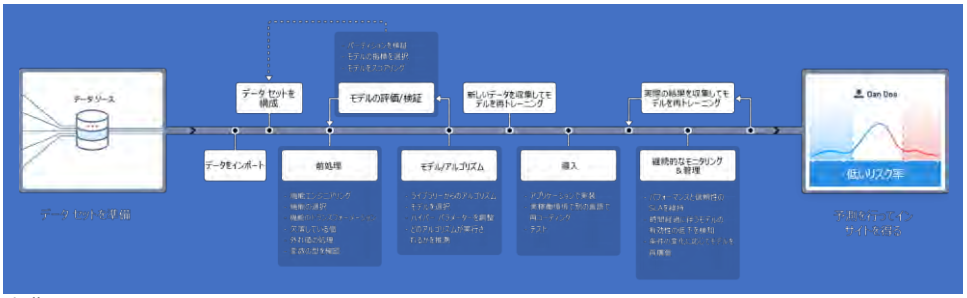
出典：datarobot.com

DataRobot は、包括的なソリューションとして、ML モデルの開発と導入の重要なフェーズ全体にわたって付加価値を提供します。

- **データを特定します。**DataRobot 内の AI カタログは、データ エンジニア、データ スチュワード、データ サイエンティスト、アナリストが、信頼できる AI 資産にセルフサービスでアクセスできる一元化された情報源として機能します。
- **データを取り込みます。**DataRobot は、構造化データと非構造化データを特定の形式に変換します。この形式は最適なパフォーマンスに各アルゴリズムで必要とされます。また、データのパーティション分割の際にはベスト プラクティスに従います。
- **機能を設計します。**DataRobot は、既存の数値、カテゴリー、テキストの機能から新しい機能を設計します。DataRobot は、追加の機能設計によるメリットが得られるアルゴリズムと、そうでないアルゴリズムを把握していて、データの特性を考慮して理にかなう機能のみを生成します。
- **アルゴリズムを検討します。**あらゆるビジネス上の問題やデータ セットを 1 つのアルゴリズムで解決できるわけではありません。DataRobot では、数百もの多様なアルゴリズムと適切な前処理を利用できます。ユーザーは、これらを利用して、AI の課題に最適なアルゴリズムを見つけるためにデータに対してテストを実施できます。
- **アルゴリズムを選択します。**DataRobot は、ユーザーがデータに適したアルゴリズムを選択するのに役立ちます。
- **ML モデルをトレーニングし、調整します。**DataRobot は、ユーザーのデータに基づいてモデルをトレーニングし、スマート ハイパー パラメーター チューニングを使用して、各アルゴリズムの最も重要なハイパー パラメーターを調整します。
- **アルゴリズムの最適な組み合わせを見つけます。**アンサンブル モデルまたはブレンダー モデルは、一般的には個々のアルゴリズムよりパフォーマンスが優れています。DataRobot は、ブレンドする最適なアルゴリズムを見つけて、各アンサンブル モデル内のアルゴリズムの重み付けを調整します。
- **モデルを直接比較します。**DataRobot は、数十のモデルを構築してトレーニングし、その結果を比較します。モデルで使用されているプログラミング言語や機械学習ライブラリーに関係なく、正確さ、スピード、正確さとスピードの最も効率的な組み合わせで、モデルをランク付けします。ユーザーは、直感的なグラフィカル ユーザー インターフェイス（GUI）を使用して各モデルを調べて、どのモデルで進めるかを選択することができます。
- **信頼を築きます。**DataRobot は透明性を確保できるように、人間が解釈できる方法でモデルの決定を説明し、各モデルの精度に最も大きな影響を与える機能と、各機能に適合するパターンを示します。また、DataRobot は、特定の予測が行われた主な理由を示すために、予測に関する説明も提供します。
- **実稼働対応モデルを導入します。**DataRobot は、リアルタイムの予測、バッチ導入、Apache® Hadoop®でのスコアなどの方法にかかわらず、ユーザーが迅速に運用化してエンタープライズ アプリケーションと統合できる、実稼働対応モデルを生成します。また、ユーザーは、R、Python®、Apache Spark®、MLlib、H2O®などのツールを使用して独自のモデルを開発し、DataRobot ライブラリーを呼び出してそのモデルをアクティブにすることもできます。

- **監視と管理を行います。**導入後に DataRobot を使用すると、予測を実際の結果と比較して最新のデータで新しいモデルをトレーニングすることが容易になります。DataRobot は、時間が経過するとモデルのパフォーマンスが低下する場合は、先を見越してハイライト表示します。

DataRobot が登場する前は、専門のデータ サイエンティストが時間のかかる面倒な多くのデータ サイエンス プロセスを実行する必要がありました。



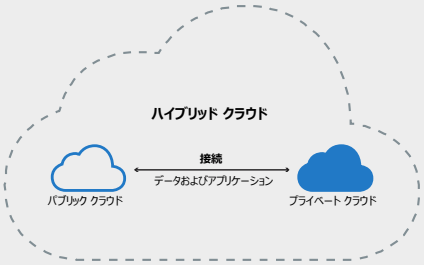
出典：datarobot.com

DataRobotを使用すれば、データ サイエンス向けのガードレールと、その他の自動化ベスト プラクティスによって、一般のデータ サイエンティストでも信頼できるAIを構築できるため、専門のデータ サイエンティストは生産性を向上させることができます。

出典：datarobot.com

DataRobot は、すでに企業内にあるテクノロジーのエコシステム内で簡単に統合できます。セキュリティやデータ プライバシーのテクノロジー、データ統合や可視化のツール、Hadoop や SQL データベースなどのインフラストラクチャ プラットフォームと統合できます。構造化データと非構造化データは、データ レイク、テーブル、他の企業ソースから取り込むことができます。ユーザーは、グラフィカル インターフェイスまたはプログラム インターフェイスを使用してシステムと対話します。

AI のワークロードは、運用するデータに依存し、組織内外のさまざまなソースからの多くのデータを必要とします。通常は、アプリケーションをデータ セットと同じ場所に配置しておく方が理にかなっています。



2 世代インテル Xeon スケーラブル プロセッサー

第 2 世代インテル 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 スケーラブル プロセッサーより動作周波数が高くなり、アーキテクチャが改良され、AI とディープ ラーニング（DL）の推論の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 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。

DataRobot ソフトウェアには、次の 3 つの独立した完全統合製品が含まれています。

- **Automated Machine Learning** には、単純線形回帰から、統計的な従来の回帰モデル、さらには勾配ブースティングやニューラル ネットワークなどのより複雑な手法まで、さまざまな回帰手法が組み込まれています。また、このソフトウェアでは、単純 2 項分類問題と、最大 100 カテゴリーまでの複雑なマルチクラス分類問題も解決できます。

- **Automated Time Series** は、データ系列の将来の価値を履歴と傾向に基づいて予測するための高度なモデルを開発する複雑なプロセスを自動化します。このプラットフォームは、時系列機能エンジニアリングを統合して予測信号を検出します。基本と高度の両方の時系列モデルを使用して予測精度を最適化し、時間経過に伴うインサイトを可視化するための多くの方法や、実稼働環境にモデルを導入するための多くの方法を提供します。

- **MLOps** は、実稼働環境で機械学習モデルを導入、管理、制御して、データ サイエンス チームへの投資を最大化し、リスクと法令順守を管理するための、一元化されたハブを提供します。その対象には、DataRobot の自動機械学習機能を使用して構築されたモデルだけでなく、Python や R などのプラットフォームやフレームワークを使用して専門のデータ サイエンティストのチームによって構築されたモデルも含まれます。

リファレンス アーキテクチャと実装

多くの組織では、既存のデータ セットにアクセスするために、AI と分析のインフラストラクチャをオンプレミスに置く必要があります。これらのデータ セットは容量が大きく、複数の組織内に散らばっているため、クラウドとの間で使い勝手がよい方法やコスト効率に優れた方法で転送することはできません。また、クラウド プロバイダーが提供するよりもはるかに低料金でコンピューティング リソースを取得し運用できる組織もあります。

AI と HPC 向けのクラウド ホスティングのメリットと制約に関して Moor Insights and Strategy が発表した最近のレポート。このレポートでは、利用可能なサービスと必要なコンピューティング リソースの取得が容易であるため、「AI を試したいと考えている組織にとっては、クラウドから始めることは非常に理にかなっていることがあります」と述べられています。[1]

「しかし、多くの組織は、アプリケーションが規模を拡大して実行し始めるため、最終的には AI と HPC 向けに相当の規模のコンピューティング インフラストラクチャを必要とするようになります。これに加え、データ転送やスループットのコストを考慮すると、組織が AI に関して成熟するにつれて、オンプレミスのインフラストラクチャを構築する方がコスト的に有利になり始めます」と続けられています。

コストに加えて、組織が考慮する必要のある重要な要因は、「データ重力」とセキュリティ/プライバシーに関する懸念です。AI のワークロードは、運用するデータに依存し、組織内外のさまざまなソースからのかなりの量のデータを必要とします。通常は、アプリケーションをデータ セットと同じ場所に配置しておく方が理にかなっています。セキュリティに関しては、所有権のある大量のデータ セットを取り扱いながら、安全性を確保する必要がある AI プロジェクトの場合、使い勝手が良くなる可能性は無視されることがあります。特に金融や医療の市場においては顕著です。

DataRobot のソフトウェア プラットフォームでは、さまざまなパブリック クラウド プロバイダーのマネージド サービスに加えて、単一ノードの Linux®システムから大規模な Hadoop 環境まで、複数のオンプレミス環境に対応しています。このような環境では、DataRobot のサービスは Docker®コンテナーとして導入されます。

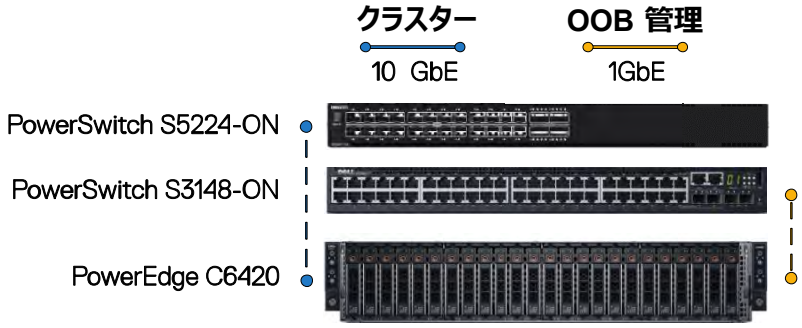
DataRobot のすべてのサービスを実行する単一 Linux サーバーから、弾性に優れたマルチノードの仮想化環境まで（各サービスで最適なパフォーマンスと予測可能なレスポンス タイムを実現するリソース要件をカスタムに導き出せる柔軟性を備える）、複数のインストールの検証を行いました。

[1] Moor Insights and Strategy『AI and HPC: Cloud or On-Premises Hosting』（2019 年 2 月）。

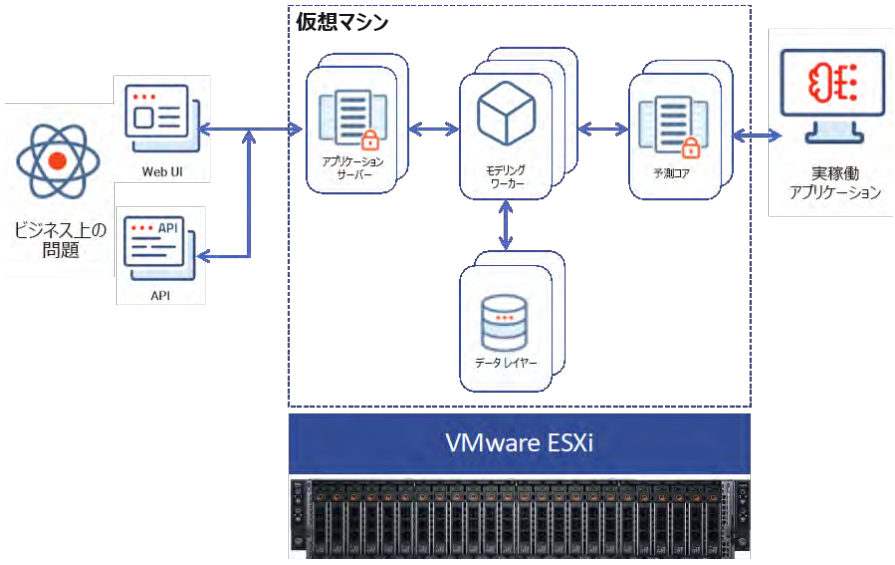
別のインスタンスでは、コンピューティング ワークロードとオブジェクト ストレージ サービスを Hadoop クラスター全体に分散させることができます。Hadoop の導入により、プロビジョニング済みの Hadoop クラスターに DataRobot をインストールする柔軟性がもたらされ、ハードウェアコストの削減とデータへの直接パスが可能になります。

以下のリファレンス アーキテクチャは、スタンドアロン Linux 導入を使用して複数の同時ユーザーをサポートするためのハードウェア インフラストラクチャを示しています。VMware® ESX®環境に複数の Linux 仮想マシン（VM）を展開し、それぞれの仮想マシンに DataRobot サービスを 1 つずつ別々に展開しました。

同時ユーザーをサポートするには、DataRobot ソフトウェア プラットフォームのアプリケーション、コンピューティング、ストレージのニーズを満たすために、合計 4 台の Dell EMC PowerEdge サーバーを使用する必要があります。Dell Technologies のポートフォリオには、このワークロードによく適しているサーバーが複数ありますが、このホワイトペーパーでは、2U シャーシに 4 台の 2 ソケット サーバーを搭載している Dell EMC PowerEdge C6420 サーバーを取り上げて検証しています。CPU コア数とメモリー容量が本書に記載している推奨事項と一致していれば、他の 2 ソケット サーバー（PowerEdge R640 や R740XD など）で行うことも可能です。



DataRobot 環境での主要なサービスは、アプリケーション サーバー、データ レイヤー、モデリング ワーカー、予測サーバーです。次の図は大まかなフローを示しています。



出典：datarobot.com

アプリケーション サーバーには、主要な管理コンポーネントすべてが収容されています。このサーバーは、認証、プロジェクト管理、ユーザー管理に対処し、DataRobot のパブリック API に対するエンドポイントを提供します。また、さまざまなプロジェクトによって作成されるモデリング リクエストのキューも管理します。それらのリクエストは、モデリング ノードで実行されている**モデリング ワーカー**によって実行されます。1 つのモデリング ノードでは複数のモデリング ワーカー（タスクを実行するための処理能力の一部）をホスティングでき、1 つの DataRobot システムに複数のモデリング ノードを持たせることができます。

DataRobot を購入すると、所定の数の並列モデリング ワーカーを利用できます。一般的に、1 つのモデリング ワーカーで 1 つのモデルのトレーニング、または 1 つの追加インサイト オプションの生成（機能の影響と予測の説明など）を行うことができます。

モデリング ワーカーは並行して実行されるため、ユーザーは複数のモデルの作成や、追加インサイトの生成を同時に行うことができます。これにより、特定のプロジェクトに関連づけられているすべてのモデルをトレーニングするのに要する時間を最小限に抑えることができます。

DataRobot ソフトウェアは、ユーザーがモデルのトレーニングに使用できるデータ量は、割り当てられている物理ハードウェア リソースによってのみ制限されるように設計されています。モデリング ワーカーでは、タスクを正常に操作して完了することができるように、マシン リソースの量が可変である必要があります。**ベース モデリング ワーカー**には、特定の CPU とメモリーの制限（通常は 4 つの CPU コアと 30GB の RAM）が割り当てられていて、サイズが（ディスク上で非圧縮状態で）1.5GB までのトレーニング データ セットと、10 クラスまでのマルチクラス分類アプリケーションを使用するモデルを構築することができます。

一方、**フレキシブル モデリング ワーカー**では、CPU コア数が 4 から 20 まで拡張でき、より大容量のメモリー割り当てを使用して、より大規模なトレーニング データ セットと、100 クラスまでのマルチクラス分類アプリケーションを処理することができます。DataRobot は、データ サイズや問題の複雑さが増すにつれて、コンピューティング能力とメモリーを動的に追加します。フレキシブル モデリング ワーカーが拡張できる容量は、モデリング ノード上で利用可能なリソース量によって制限されます。

また、モデリング ワーカーはステートレスであるため、必要に応じて環境に参加したり離脱したりするように構成することもできます。これにより、仮想プライベート クラウド（VPC）を使用して構成している場合に、ハードウェア コストを節約できます。Hadoop クラスター内では、これらのワーカーは YARN コンテナーです。トレーニング済みのモデルは**データ レイヤー**に書き戻され、それらの精度はアプリケーション サーバーを介してモデルのスコアボードに反映されます。

予測は、**Web UI** を使用してデータをアップロードするか、または **API** エンドポイントを使用することによって、バッチで生成できます。どちらの場合も、予測を生成するために一時的にモデリング ワーカーが使用されます。低レイテンシーで大容量の予測環境を実現するには、**予測コア**を使用することが推奨されます。予測コアは、モデリング アクティビティとの競合を回避するために予測用に予約されています。

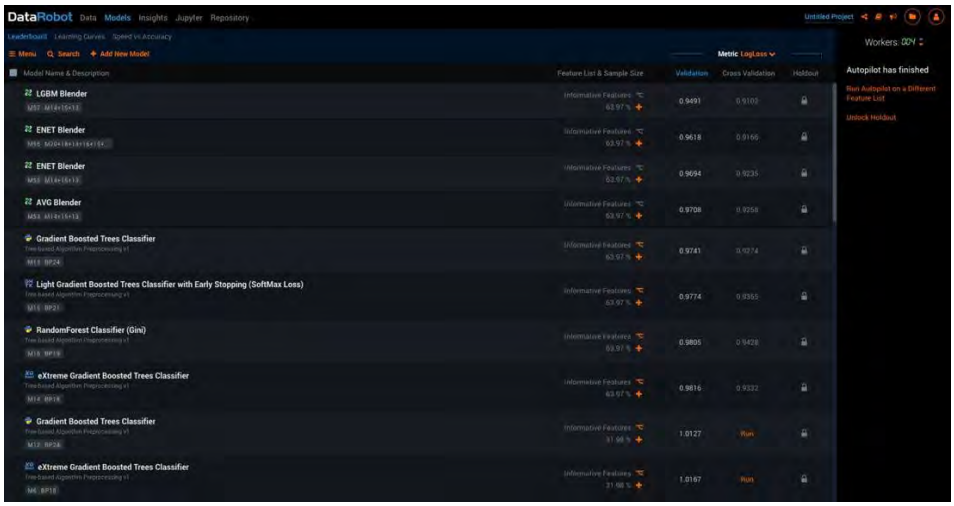
予測コアを予約するシステムによって、特定のユーザーが予測を生成する必要があるときに予測リソースを効果的に使用できるようなり、モデリングするユーザーは、モデリング リソースが予測のために使用されている間、待たなくて済むようになります。さらに、それらの予測と提供されたデータに関する主要な統計情報は、アプリケーション サーバーに戻されて、データのドリフトと正確さの分析などの、モデルの正常性をモニタリングできるようにユーザーに表示されます。

予測コアを DataRobot から切り離された環境に導入することもできるため、企業は隔離されたネットワークにモデルを導入できます。モデリング ワーカー数が 30、20、12 のシステムをサポートするための推奨構成を以下の表に示します。

	30 のモデリング ワーカー（大規模）	20 のモデリング ワーカー（中規模）	12 のモデリング ワーカー（基本）
サーバー	PowerEdge C6420 x 4	PowerEdge C6420 x 3	PowerEdge C6420 x 2
プロセッサー	インテル® Xeon® Gold 6248		
メモリー	384GB DDR-2933		
ストレージ	480GB BOSS（ブート用） x 2、Dell Express Flash NVMe P4610 1.6TB SFF x 6		
ネットワーキング	インテル Ethernet 10G 4P X710 SFP+ rNDC		

このリファレンス アーキテクチャは Dell EMC PowerEdge のテスト データ センターでセットアップされており、多様なワークロードを実行する場合のインフラストラクチャ要件を決定するために、複数のユースケースがテストされています。自動化された機械学習機能と DataRobot プラットフォームの機能を使用して、さまざまなユースケースを実行しました。ユースケースの例として 5 つの公開データ セットを選択し、DataRobot を活用して、データを取り込み、機能を設計し、選択したデータに対してアルゴリズムを調査して選択し、複数の ML モデルを調整し、トレーニングしました。DataRobot は、モデル トレーニング フェーズの一環として、スマート ハイパー パラメーター チューニングを使用して、各アルゴリズムの最も重要なハイパー パラメーターを調整します。これらのフェーズは、計算負荷が高いため、複数のモデリング ワーカーを並行して活用するスケールアウト手法と、アルゴリズムのインテリジェントな選択を使用して、トレーニングの時間を最小限に抑えています。

DataRobot はパフォーマンスが上位のモデルを示すスコアボードを提示するので、正確さのスコア、検証の正確性、モデルの追加詳細（使用される入力変数など）を比較して、それらのモデルを検討することができます。導入の適性を判断するための他の考慮事項（予測速度や解釈可能性など）があるので、実稼働環境に最も適したモデルは最も正確なモデルではないことがあります。DataRobot はモデルを比較して評価するためのツール、レポート、プロットを提供します。これにより、ユーザーがどのモデルをスコアボードから選択するかを判断することが簡単になります。



出典：datarobot.com

インテル DC SSD（ソリッド ステート ドライブ）

インテル DC SSD は、インテル Xeon スケーラブル プロセッサー向けに最適化されており、さまざまなメディア タイプと容量が提供されています。たとえば、インテル 3D NAND SSD（インテル SSD DC P4500、P4501、P4600 シリーズなど）は、クラウド インフラストラクチャ向けに最適化されており、卓越した品質、信頼性、高度な管理機能、保守機能を提供しているため、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 Optane™ SSD DC P4800X では、高速キャッシュと高速ストレージによってアプリケーションが高速化されて、サーバーあたりの拡張性が向上し、レイテンシーの影響を受けやすいワークロードのトランザクション コストが削減されます。さらに、インテル Optane SSD DC P4800X を使用すると、データ センターでより多くの低コストのデータ セットを導入して、大容量のメモリー プールから新たなインサイトを得ることができるようになります。

DataRobot で評価されたユースケース

不正検出。消費者詐欺（クレジット カード詐欺、ローン詐欺、銀行の破綻など）を見抜くことは、旧式のルールやシステムでは難しい場合があります。クレジット カード詐欺の場合は、DataRobotを使用して以前のクレジット カード取引をモデル化することによって、詐欺を検出し、防止することができます。このモデルは、最近のトランザクションが詐欺的であるかどうかを識別します。その目標は、詐欺的取引を 100%検出して、不正への間違った分類を減らすことです。

データ セット：Predicting Fraud in Financial Payment Services

顧客の解約予測。顧客の解約は、組織の成長にとって大きな妨げとなります。DataRobot は、顧客の解約を予測して、顧客維持（銀行部門）とマーケティング部門に提示できる実用的なアラートを作成するための予測モデルを開発することができます。

データ セット：Bank Customer Churn Prediction

保険会社におけるリスク評価。保険会社には、死亡率、罹患率、壊滅的なリスクなどの複数のリスクが関連しています。DataRobot を使用すると、これらのリスクを評価して格付けすることができます。このプラットフォームは、保険会社が危険度を測定して推奨保険料を算出するのに役立ちます。

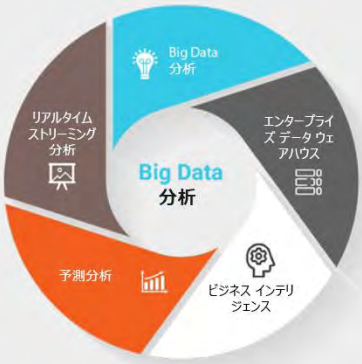
データ セット：Prudential Life Insurance Assessment

ヒッグス粒子検出。ヒッグス粒子は機械学習での最新の改良のおかげで発見されたものであり、現在は、データ セット内の非常に多くのパラメーターをシステムがどの程度処理できるかを判断するための共通のベンチマーク データ セットとして使用されています。

データ セット：HIGGS Data Set

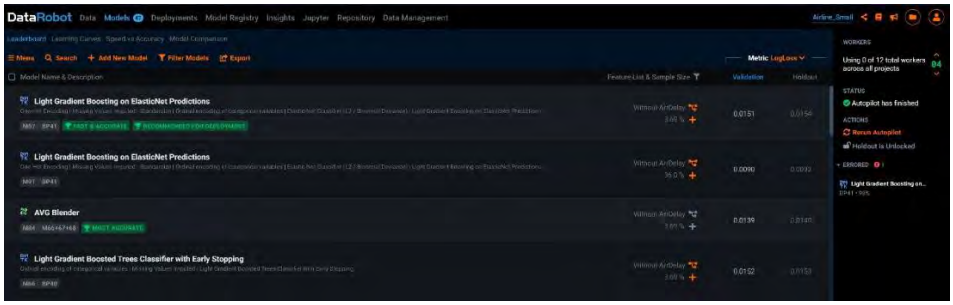
航空路線の遅延予測。航空路線の遅延を予測することは、旅行者やオペレーターにとってメリットがあります。DataRobot を使用すると、顧客は別のフライト経路を選択できるか、または遅延するリスクが大きい場合は、少なくともキャンセルするか、そもそも航空券を購入しないかを選択できるようになります。オペレーターは、遅延に直面したときに、自分の組織の履行と競合他社の履行を分析することができます。

データ セット：Year 2008



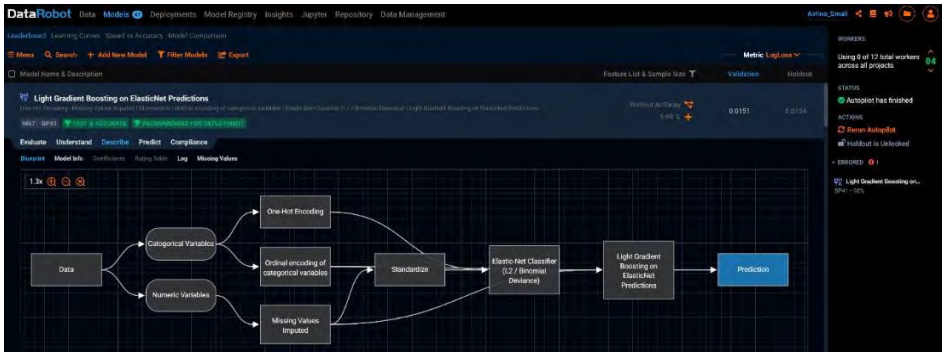
出典：progressive.in

航空路線の遅延を予測するようにトレーニング済みの複数のモデルを示すスコアボード ビューを以下に示します。



出典：datarobot.com

導入が推奨されるモデルの詳細を確認します。この場合は、ブループリントを見ると LGB モデルが推奨されています。



出典：datarobot.com

Big Data の Hadoop 環境との統合

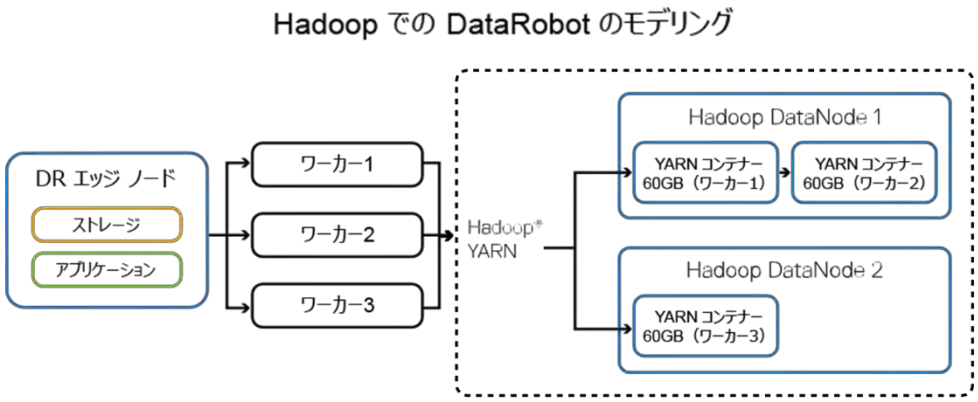
DataRobot は、Hadoop ベースのデータ分析環境を増強します。Cloudera® Hadoop 環境に導入すると、DataRobot に組み込まれているソフトウェア統合により、導入、モニタリング、管理がシンプルになります。以下のメリットがあります。[2]

- **パーセルを使用してインストールする。** DataRobot の新規リリースとアップデート リリースを、Cloudera Manager を使用してダウンタイムなしで導入できます。
- **CSD が統合されている。** Cloudera Manager は、DataRobot で使用されているリソースをモニタリングできます。
- **Kerberos。** LDAP/AD（認証）がサポートされている
- **Sentry が有効になっている。** 承認/ガバナンス/コンプライアンス
- **YARN が有効になっている。** マルチ テナント環境用のリソース マネージャー
- **Spark/MLlib を使用。** モデリング タスク向けの柔軟なインメモリー データ処理

これらの統合により、DataRobot のユーザーは、Hadoop 分散ファイル システム（HDFS）に保存されている過去のデータを表す、既存のデータ レイクを活用でき、それらを自動モデル トレーニングに利用できるようになります。モデル トレーニングと推論向けの Apache Spark の統合により、データ運用チームが数年に渡って管理してきたのと同じ Apache Spark と Hadoop の環境で、DataRobot ベースの分析ワークロードをスケーリングできるようになります。

[2] DataRobot「DataRobot Brings Automated Data Science to Hadoop」（2016 年 3 月）。

DataRobot は、エッジ ノードとして Hadoop のクラスターと統合されています。この構成では、YARNがリソース管理を担当し、Apache Sparkのエグゼキューターがモデリング タスクに活用されます。ワーカー ノードが DataRobot からモデル トレーニング操作を受け取ると、必要なメモリーが YARN によって Hadoop DataNode 上に割り当てられて、モデル トレーニングが開始されます。

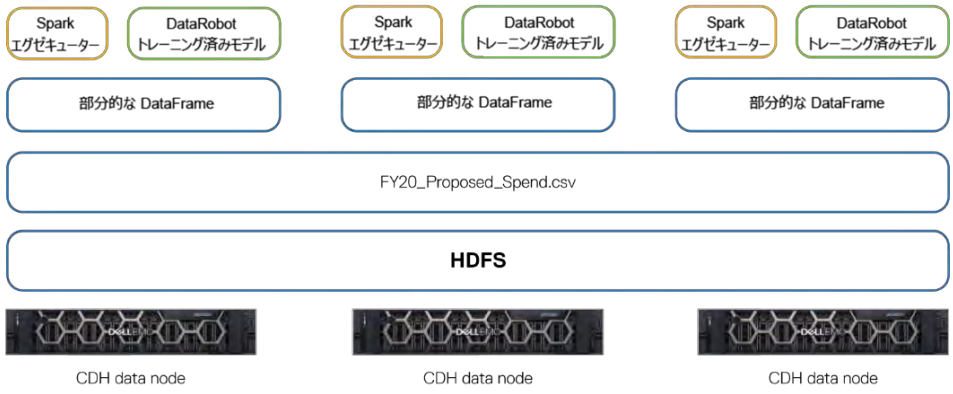


・YARN は、ワーカーがモデルをトレーニングするときに、データ ノードでメモリを割り当てます。
・各モデルは、利用可能なデータ ノード上のメモリでトレーニングされます。

出典：Transforming Insurance Analytics with Big Data and Automated Machine Learning

このアーキテクチャでは、各 YARNコンテナーは、1つのモデル候補のトレーニングとDataRobotを担当し、Spark Driverとして機能して、選択プロセスをオーケストレーションします。これにより、DataRobotは、分散トレーニング向けに設計されていないか、または Spark 互換形式で使用できる、アルゴリズムと機械学習ライブラリーを使用できるようになります。各 Spark エグゼキューターは、計算のためにローカル ノード リソースに対して実行されます。単一モデルの分散トレーニング操作のように、他のエグゼキューターとデータを共有する必要はありません。[3]

Apache Spark を使用してモデリング タスクを実行すると、推論またはモデルのスコアリング操作の動作が異なります。トレーニング時に行われるモデル候補の選択プロセスとは異なり、インメモリー分散型 Apache Spark データ構造は推測操作に大きなメリットをもたらします。推論時に、各ワーカーは、トレーニング済みモデルの同一コピーを使用します。各ワーカーは、データセットの分散表現である DataFrame のパーティションをスコアリングします。各パーティションの結果は、レポート作成用に DataRobot に送信されます。



[3] Dell Technologies「Training an AI Radiologist with Distributed Deep Learning」2018 年 8 月）。

DataRobot のテストでは、Cloudera Hadoop 向け Dell EMC Ready Solution のアーキテクチャ ガイドに記載されている設計とびベスト プラクティスに従って、以下の Dell EMC PowerEdge C6420 サーバー構成を使用しました。[4]

システム要素	DataRobot アプリケーション ノード	Cloudera ワーカー ノード	Cloudera 管理ノード
サーバー モデル	Dell EMC PowerEdge C6420 x 1	Dell EMC PowerEdge C6420 x 2	Dell EMC PowerEdge C6420 x 1
CPU	インテル Xeon Gold 6230 プロセッサー	インテル Xeon Gold 6240 プロセッサー	インテル Xeon Gold 6230 プロセッサー
DRAM メモリー	192GB	384GB	192GB
容量 ストレージ	3.8TB SATA SSD	3.8TB SATA SSD	3.8TB SATA SSD
ネットワーク	1 – 埋め込み型インテル ギガビット I350-t LOM 2 – インテル Ethernet 10G 2P X520 アダプター	1 – 埋め込み型インテル ギガビット I350-t LOM 2 – インテル Ethernet 10G 2P X520 アダプター	1 – 埋め込み型インテル ギガビット I350-t LOM 2 – インテル Ethernet 10G 2P X520 アダプター
ソフトウェア- （動作条件）	CentOS® Linux 7.6.1810 Docker サービス	Cloudera Hadoop 5.16.1 HDFS DataNode YARN ノード マネージャー Spark ゲートウェイ Hive ゲートウェイ	Cloudera Manager 5.16.1 YARN リソース マネージャー HDFS NameNode Zookeeper Spark 履歴サーバー
DataRobot ソフトウェア	DataRobot 5.2.2 リリース アプリケーション		DataRobot 5.2.2 リリース パーセルとサービス マスター サービス ETL コントローラー ETL デフォルト サービスと ETL クイック ワーカー サービス

役割
DataRobot アプリケーション ノード – このノードは、DataRobot アプリケーションのインストールに使用され、Docker のコア サービス（Nginx、Mongo、Redis、RabbitMQ など）と Hadoop 構成同期を維持します。
Cloudera 管理ノード – 管理ノードは、Cloudera Manager、HDFS NameNode、YARN リソース マネージャーをインストールするために使用されます。
Cloudera ワーカー ノード – ワーカー ノードは、YARN スケジューラーによって管理されているモデリング タスクをオフロードするために使用されます。

[4] 『Dell Ready Bundle for Cloudera Hadoop Architecture Guide version 5.10』

Dell EMC インフラストラクチャ上の DataRobot エンタープライズ ML プラットフォームの価値

DataRobot エンタープライズ AI ソフトウェアは、専門知識がなくてもデータ サイエンスを利用できるようにし、AI を適切な規模で構築、導入、メンテナンスするためのエンド ツー エンドのプロセスを自動化します。最新のオープンソース アルゴリズムを利用し、クラウド、オンプレミス、またはフル マネージド AI サービスとして利用可能な DataRobot を使用すると、ユーザーは AI のパワーを活用して、より優れたビジネス成果を生み出すことができます。

- **ROI に優れた AI。** DataRobot は、AI 主導でありたいと考えている組織に価値と成功をもたらすことに徹底的に重点を置いて開発されています。データ サイエンティストの生産性を向上し、データ サイエンティスト以外のユーザーが従来のデータ サイエンスの手法を習得することなく、AI の構築、導入、保守を行うことができるようになることにより、AI のユースケースのスループットが加速されます。手作りのコードによる少数の機械学習モデルの開発とテストに数週間または数か月かけるのではなく、既存のチーム（データ サイエンスの専門知識を問わず）が数百ものモデルを構築し、パフォーマンスが最も優れているモデルを数時間で導入することができます。

- **信頼できる AI。** DataRobot は、常に計画どおりに機能する信頼性の高い AI を提供することに取り組んでいるため、ユーザーは重要なビジネス上の意思決定を自信を持って下すことができます。専用の開発チームが直感的な AI エクスペリエンスを提供することに重点を置いているため、ユーザーは DataRobot の予測と推測を簡単に理解でき、データ サイエンスのスキル レベルに関係なく他の人に説明することができます。組み込みのガードレール、自動化されたモデルのドキュメント作成とその他の機能により、組織特有の価値と倫理を矛盾なく反映した、人間中心の AI が実現されます。

- **所有する AI。** AI インフラストラクチャは、お客様の最も戦略的な資産になる可能性を秘めているため、そのあらゆる側面をお客様が所有している必要があります。AI を構築し管理するための複数のオプションが用意されているので、DataRobot によって、組織はプラットフォームをどこででもどのようにでも望みどおりに使用できる柔軟性がもたらされます。お客様の AI の知的財産は、ベンダーロックインに陥ることなく、お客様が所有したままです。

ガバナンスと透明性

AI と機械学習の多くのアプローチは、「ブラック ボックス」で行われるため、モデルがどのようにトレーニングされて、どうしてその予測が行われたかについての詳細は部分的にしか知ることができません。DataRobot は違います。モデルの構築、導入、管理のプロセス全体にわたって透明性に特に重点を置いています。

AI の作成者は、データがどのように処理されたか、どのような機能が設計されたか、どのようなアルゴリズムが使用されたかなど、モデル構築プロセスのすべてのステップについての詳細を確認できます。AI のオペレーターはモデルのドキュメントを自動的に生成して、モデルがどのようにトレーニングされたかを理解でき、モデルの結果に最も大きな影響を与えているデータを簡単に識別することもできます。このプロセスによって、ビジネス ルールとロジックをモデルに適用して、モデルの整合性を損なう可能性のあるバイアスやその他の問題を特定することもできます。また、AI の利用者は、モデルによって行われたすべての予測に対して、その意思決定に寄与した主な要因などについて、人間が分かりやすい説明を得ることができます。

実稼働環境へのモデルの配置

AI と機械学習のプロジェクトは、お客様のビジネスの未来を推進するものです。しかし、多くの企業では、データ サイエンスへの投資を最大限に活用できていません。そのような組織には、実稼働環境で機械学習を導入、管理、制御するためのスキルとリソースが不足しています。さらに、適切なモニタリングと制御を行わずに AI モデルを使用した結果、収益の低下や、経営陣、投資家、顧客からの信頼を失うなどの、壊滅的な事態に陥る可能性があります。

DataRobot に組み込まれているすべてのモデルは、すぐに導入できる状態になっています。

- **アップロード** – 新しいデータ セットを DataRobot にアップロードし、バッチでスコアリングしてダウンロードします。
- **作成** – アプリケーションからデータを直接スコアリングできるように、REST API のエンドポイントを作成します。低レイテンシーで高スループットの予測要件に適合するように、独立した予測サーバーを使用することができます。
- **エクスポート** – Hadoop でのインプレース スコアリングのためにモデルをエクスポートします。
- **ダウンロード** – 編集可能なソース コードまたは自己完結型実行ファイルとしてスコアリング コードをダウンロードし、アプリケーションに直接埋め込むことによって、計算負荷の高い操作を高速化します。

導入後は、お客様のビジネスを推進する機械学習モデルで、市況が変化しても正確かつ一貫性のある結果を得られることが DataRobot の MLOps によって保証されます。実稼働環境にあるすべてのモデルからの指標の概要（予測リクエストの数、主要な正常性統計情報など）を一目で確認できます。

- ［**Service Health**］では、運用またはエンジニアリングの観点からのコア パフォーマンス指標（レイテンシー、スループット、エラー、使用状況）が表示されます。
- ［**Data Drift**］では、モデルの信頼性に影響を与える可能性のある傾向があるかどうかをユーザーに知らせるために、時間経過に伴うデータ特性の変化をプロアクティブに探しています。
- ［**Accuracy**］では、予測に対応する実際の値（つまりグラウンド トゥルース）を比較しているので、標準の機械学習指標を使用してモデルのパフォーマンスを評価することができます。

DataRobot は、モデルを頻繁にアップデートする機能を提供しています。競争力のある新しいモデルをテストし、ビジネス アプリケーションへのサービス提供を継続したままアプリケーションをオンザフライで変更することもできます。ML の導入に関連するガバナンス ポリシーを適用し、強力なガバナンス手法に必要なデータ（モデルを公開したユーザー、変更が行われた理由、時間経過に伴ってどのモデルが実施されていたかなど）を取得します。

Dell Technologies、DataRobot、Intel のコラボレーションによるメリット

- デル カスタマー リューション センターで DataRobot の AI プラットフォームにアクセスしてテスト実行し、既存の IT 環境への導入や統合に関するガイダンスを提供できるソリューション エキスパートに相談してください。
- シンプルになった、Dell EMC インフラストラクチャ上の DataRobot の注文プロセスをご利用ください。
- デルと DataRobot による共同サポートを受けることができます。

まとめ

デルは、組織がデータ センターにエンタープライズ AI プラットフォームを実装するために活用できる、Dell EMC インフラストラクチャ上の DataRobot プラットフォームのオンプレミスでの実装を発表しました。このエンタープライズ AI プラットフォームは、既存の Big Data やデータ レイクのプラットフォームと統合することも、スタンドアロンのマルチ ユーザー環境として単独で実行することもできます。

データ サイエンス チームが必要とするさまざまなモデリング機能と予測機能に対応できる、3 つの異なるスタンドアロン構成が検証済みです。DataRobot を Hadoop と統合することにより、Hadoop のコンピューティング リソースを DataRobot の機械学習に活用でき、HDFS に格納されている既存データを活用できます。

このホワイトペーパーとリファレンス アーキテクチャは、DataRobot、Intel、Dell Technologies の共同作業であり、組織で以下のことができるようになります。

- 自社で独自に行う（最適な構成やサイジングなどを決定する）場合よりも迅速に、オンプレミスのソリューションを立ち上げることができます。
- 安定性と相互運用性をエンジニアがテスト済みであるため、自信を持って導入できます。
- さまざまなユースケースとトレーニング/推論機能向けに適正なサイズのハードウェア インフラストラクチャが提供されているため、コストの削減を実現できます。

Dell EMC Ready Architecture for Hadoop を使用している組織は、DataRobot を統合して、Big Data 分析機能を拡張することができます。

詳細情報

- DataRobot
- Delltechnologies.com/ai
- Delltechnologies.com /referencearchitectures
- Delltechnologies.com/hpc
- Hpcatdell.com
- Delltechnologies.com/servers
- @dellemcservers
- Intel and AI インテル SSD データセンター ファミリー
- Intel Deep Learning Boost
- Intel Framework Optimizations
- Intel Deep Learning Reference Stack
- Intel Builders
- #IntelBuilders

DataRobot

DELL Technologies

intel