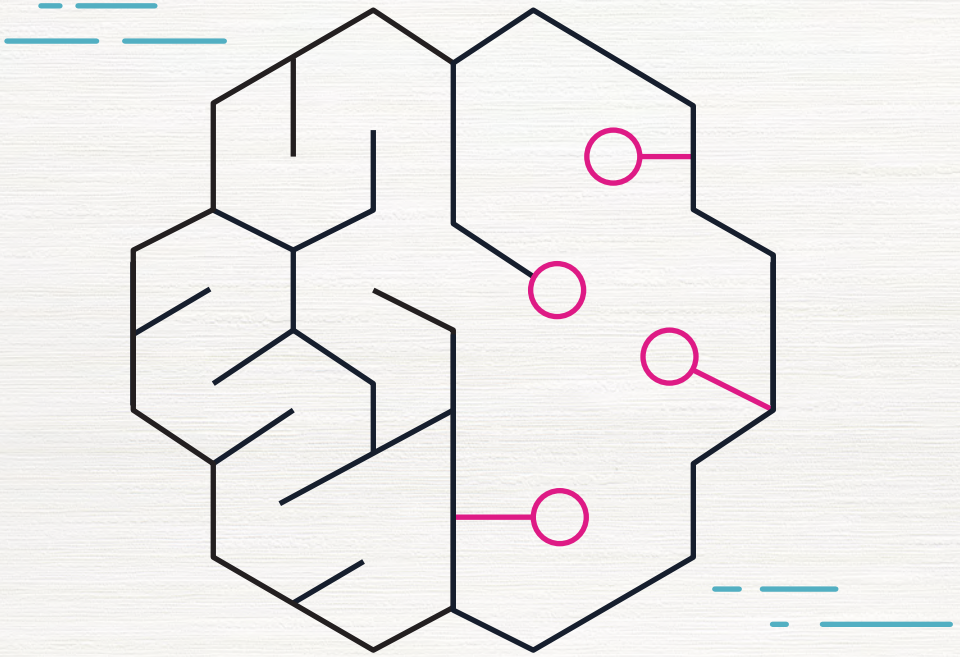
aws startups

機械学習のデータ処理管理

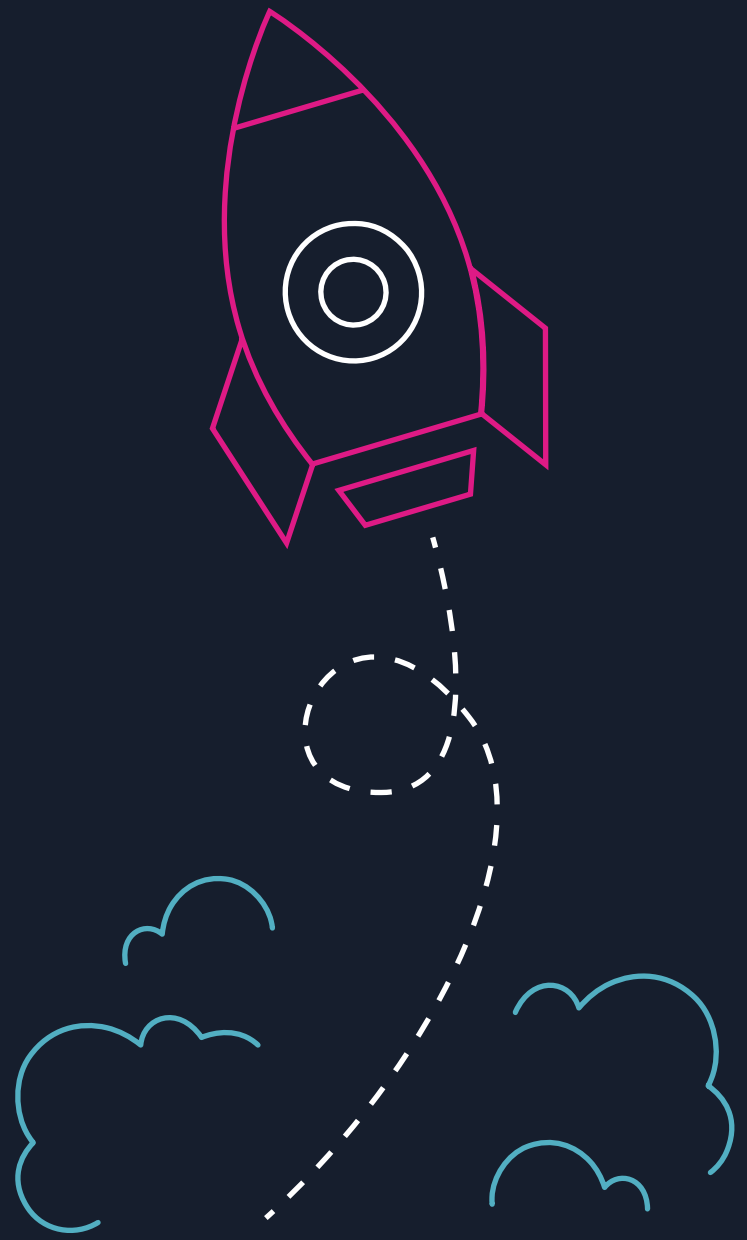


機械学習実践に向けた
データ管理についてのリファレンスガイド

目次

はじめに

この日本語ガイドでは、機械学習を実用化するうえで中核となるデータ処理のインサイトと実践ガイドを提供します。最初に、データの取得、データ品質、データのサイジングなどについて理解を深めます。続いて、AWS が提供する関連のデータ管理サービスおよび処理サービスと共に、機械学習データの管理にこれらのサービスを活用しているお客様の成功事例を紹介します。この日本語ガイドがお客様による機械学習データ処理の基盤となり、今後の段階における機械学習の実用化に継続的に役立つことを願っております。



機械学習データについて

最新の機械学習テクニックでは、モデルの構築は大きくデータに依存しています。機械学習のためのデータのさまざまな側面を理解することは、機械学習の導入を成功させるために不可欠です。この章では、機械学習データに関連するこれらの重要な側面についていくつか説明し、機械学習プロジェクトを開始する際により深い理解が得られるようにします。

機械学習のデータソース

通常、機械学習データセットは組織外または組織内のデータソースから取得します。外部データソースは、さらに無料のオープンデータセットと、データプロバイダーから提供されるプライベートデータセットに分類できます。

外部のパブリック機械学習データソース

これらのデータセットは、国際機関、中央政府または地方自治体、研究機関などの組織によって、過去の運用データ、調査プロジェクト、機械学習の課題などを通じて作成または収集されます。一般的なパブリック機械学習データセットの例をいくつか挙げます (一部を除き英語)。

- AWS Data Exchange
- AWS オープンデータレジストリ
- データ | 世界銀行
- データとマップ — 欧州環境機構 (EEA)
- パブリックデータセット : アマゾン ウェブ サービス
- Google Public Data Explorer
- Competitions – Kaggle
- UCI 機械学習レポジトリ

通常、これらのデータセットは一般公開されており、アクセス、使用、貢献が可能です。機械学習に適したこれらのパブリックデータソースの一覧については、こちらおよびこちらをご覧ください。

外部のプライベート機械学習データソース

これらは、特定の業種や調査分野向けにより独自性の高いデータセットや付加価値の高い可能性のあるデータセットを提供しているベンダーです。通常、これらのベンダーは探索プラットフォーム (Explorium など)、マーケットプレイス (Datarade など)、機械学習 SaaS サービス (Calligo など) を通じてデータセットを提供しています。AWS は、クラウド上のサードパーティーデータを簡単に検索、サブスクライブ、使用できるサービスである、AWS Data Exchange も提供しています。

内部の機械学習データソース

これらは、顧客データ、ビジネス運営データ、またはその知的財産を通じて取得した組織内のデータセットです。これらのデータセットは、データベース、データウェアハウス、ドキュメントストアなどさまざまな内部システム、または CRM や ERP システムなど外部ベンダーの SaaS プラットフォームに保存できます。通常これらは、企業が活用を試みる主な機械学習データソースですが、サイロに分散されていることが多く、管理が困難です。AWS のサービスを活用してこれらのデータセットを統合および管理する方法については、後の章で詳しく説明します。

機械学習データセットのデータ品質

コンピューティングの初期から、「ゴミを入れたら、ゴミが出てくる」と言われてきました。この格言は、AI　の時代においてさらに意味を帯びてきました。AI の重要なブランチである機械学習では、大量のトレーニングデータを使用して機械学習モデルを構築します。機械学習モデルは、その性質上、データ品質の影響を非常に受けやすくなっています。Andrew Ng 博士の「Data Centric AI」リファレンスによれば、AI デベロッパーの 80% の時間がデータの準備に使われており、データの最適化に時間を使えば、アルゴリズムの効率よりもモデルのパフォーマンスが向上する可能性が高いことがわかっています。したがって、機械学習プロジェクトの最初のステップとして、機械学習データの一貫性、正確性、互換性、完全性、適時性、重複または破損したレコードを確認する必要があります。

データ品質の測定と評価

データサイエンスの世界では、データ品質の測定と評価には多くの方法が使用されます。例えば、データサイエンティストは**ベンチマーク**、**コンセンサス**、**Cronbach のアルファテスト**、**レビュー**[1] など、よく使用されるデータ品質測定プロセスに従い、精度、偏向、正確性などの統計的概念を使用してデータ品質を評価しています[2]。より実践的な意味合いにおいては、データ品質に影響を与えるもの (データの測定および収集エラー、ノイズ、外れ値、不足値など)[2] について理解し、それらを測定するために適切なメトリクス (データとエラーの比率、空の値の数など)[3] を定義すると、データ品質の改善に役立ちます。

80%

の AI デベロッパー時間がデータの準備に使われており、データの最適化に時間を使えば、アルゴリズムの効率よりもモデルのパフォーマンスが向上する可能性が高いことがわかっています。

機械学習データとモデル品質の改善

機械学習モデルを改善するための最適な方法は、トレーニング、検証、テストのための品質の高いデータセットから始まります。したがって、データセットの事前処理 (クレンジング、変換、代入) が重要なステップになります。Amazon SageMaker は、SageMaker Studio の SageMaker Data Wrangler や SageMaker Processing など複数の機能を提供し、機械学習データを改善するためにデータの視覚化と準備を支援します。また、SageMaker Clarify は機械学習モデルの偏向を検出し、モデル品質の改善を支援します。これらの機能については、後の章で詳しく説明します。データのクレンジングと変換以外にも、他の考慮事項があります[4]。考慮すべき重要なことのひとつは、(モデルのドリフトを防止するために) 機械学習データとモデルのモニタリングを継続し、更新されたデータセットでモデルを再トレーニングすることです。機械学習のための自動化されたデータ品質管理に向けた調査[5] が、機械学習モデルを継続的に改善する Amazon SageMaker の Data Quality Monitoring および Model Quality Monitoring 機能の開発につながっています。

機械学習モデルをトレーニングするためのデータサンプリングサイズ

機械学習はデータに大きく依存しています。そしてよくある質問の 1 つが、「どれだけデータが必要か ?」ということです。機械学習には、さまざまな問題 (分類、予測、異常検出、クラスタリングなど) を解決するために、異なる学習パラダイム (教師あり学習と教師なし学習など) やアプリケーションドメイン (テキスト分析、自然言語処理、イメージ処理など) があり、非常に複雑です。また、モデルのトレーニングのために使用される機械学習アルゴリズムやフレームワークなど、考慮すべきその他の要素もあります。これらの要素以外にも、モデルの正確性やトレーニング時間に関する具体的な要件があり、必要なデータサイズはコストによっても決定されます。したがって、**この質問に対する回答は 1 つとは限りません。ケースバイケースです。ただし、原則となるガイドをいくつか以下に示します。実験を通じ、ニーズに合わせて調整することができます。**

サンプリングデータサイズまたは機械学習トレーニングデータセットに関するこのトピックでは、多くの調査が行われています[6、7、8]。一般的なガイドラインとなるシンプルな原則として、データセットのサイズはディメンションの約 10 倍以上とし、使用するモデルに依存してはなりません[9、10、11]。その他のよく見られる機械学習ドメインまたは問題の種類については、右の表にサンプルデータサイズの一般的なガイドを示しています。

これらは、初期の機械学習プロジェクト計画の一般的なガイドとして使用します。解決する実際の問題、使用しているフレームワークとアルゴリズム、前述した考慮すべきその他の要件に基づいて、データセットサイズを調整してください。

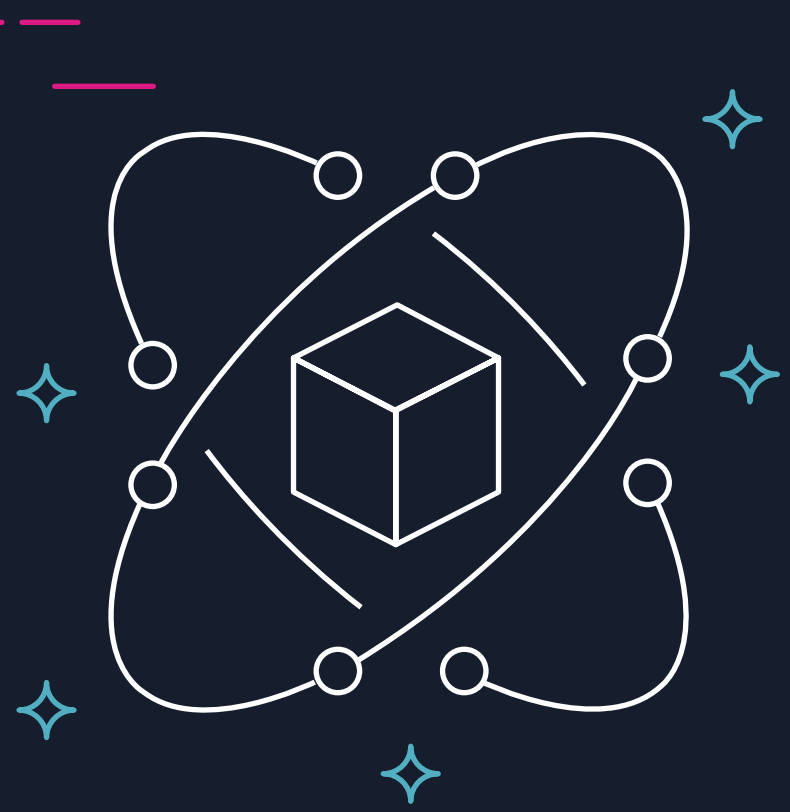
これで機械学習データの価値とさまざまな側面について理解したので、後の章では、機械学習に関連する AWS のサービスが、機械学習データの収集、処理、管理による機械学習プロジェクトの開始にどのように役立つかについて説明します。

	学習パラダイムまたはドメイン	問題の種類	データサイズの原則	メモ	リファレンス
1	教師あり学習	線形回帰、バイナリまたはマルチクラスの分類	10:1 のデータサンプル比 : ディメンション	例 : XGBoost	9、10、11
2	教師あり学習	時系列予測	50 ～ 100 回の観察	例 : ARIMA モデルの使用	12、13
3	コンピュータビジョン	イメージ分類	クラスあたり 1000 イメージ	例 : ImageNet データセットの CNN または DL	14、15、16
4	テキスト分析	ドキュメント分類	カテゴリあたり 100 ドキュメント以上 *	例 : ドキュメント分類	17
5	自然言語処理 (NLP)	感情分析	200 ～ 300 ワード **	例 : Yelp Academic Dataset を使用した感情分析	18

*、** 注意 : ほとんどの場合、これらの数値は参考用の実験データに基づいています。

AWS でのデータの管理

これで機械学習におけるデータの価値と重要性について理解を深めたので、機械学習にデータを効率的に使用する方法について説明します。このセクションでは、機械学習プロセスの一般的なステップと、そのステップを簡単かつ効率的に達成できるよう支援する AWS のサービスについて説明します。



データの収集

前述したように、今日の機械学習はインサイトに満ちた正確なモデルを構築するためにデータに依存しています。そして最初のステップは、機械学習モデルにはどのようなデータが必要か確認し、そのデータを収集してモデルをトレーニングするために利用できるさまざまな方法を評価することです。AWS は、静的なリソースから、またはウェブサイト、モバイルアプリ、およびインターネット接続デバイスなどの新しい動的に生成されたリソースからデータを取り込むための方法を多数提供しています。

Amazon Kinesis Suite: ストリーミングデータは、複数のデータソースによって継続的に生成されるデータです。通常、データレコードで同時に、小さいサイズ (数キロバイト) で送信されます。ストリーミングデータには、アプリケーションログファイル、e コマースによる購入、ソーシャルネットワークからの情報など、さまざまなデータが含まれます。Kinesis は、ストリーミングデータの収集、処理、保存に役立ちます。

- Amazon Kinesis Data Streams は、AWS のスケーラブルなリアルタイムデータストリーミングサービスです。KDS は継続的に数千のソースから 1 秒あたり数 GB のデータを取り込み、収集したデータを数ミリ秒で利用可能にして、リアルタイム分析のユースケースに対応します。

- また、Amazon Kinesis Data Firehose では、ストリーミングデータをデータレイク、データストア、分析サービスに簡単にロードすることができます。Kinesis Data Firehose はフルマネージドで、データのスループットに合わせて自動的にスケールします。また、ロード前にデータストリームをバッチ処理、圧縮、変換、暗号化することもできます。

- Amazon Kinesis Data Analytics では、3 つの簡単なステップでクエリと高度なストリーミングアプリケーションをすばやく簡単に構築できます。それらのステップは、ストリーミングデータソースのセットアップ、クエリまたはストリーミングアプリケーションの記述、処理されたデータの送信先のセットアップです。

例 : ストリーミングソーシャルメディアデータの分析

入力
ソーシャルメディアストリームを Amazon Kinesis Data Streams にロード

Amazon Kinesis Data Streams
リアルタイムの処理のため、ソーシャルメディアフィードを取り込んで保存

Amazon Kinesis Data Analytics
リアルタイムでハッシュタグデータを生成

Amazon Lambda
後処理を実行し、Amazon DynamoDB に書き込み

Amazon DynamoDB
ソーシャルメディアのトレンド結果を保存

出力
マーケティングチームがソーシャルメディアのトレンドをモニタリングし、すばやく対応

例 : クリックストリーム分析

入力
ウェブサイトがクリックストリームデータを Amazon Kinesis Data Firehose に送信

Amazon Kinesis Data Firehose
データを収集し、Amazon Kinesis Data Analytics に送信

Amazon Kinesis Data Analytics
リアルタイムでデータを処理

Amazon Kinesis Data Firehose
処理されたデータを Amazon Redshift にロード

Amazon Redshift
得られた分析モデルを実行し、コンテンツのレコメンデーションを提供

出力
パーソナライズされたコンテンツの提案を表示し、さらにエンゲージメントを実現

クラウドデータの移行 : クラウドにデータを移行する際は、さまざまなユースケースにおけるデータの移動先、移動するデータの種類、利用可能なネットワークリソースなどの検討事項を理解する必要があります。AWS は、あらゆるデータ移行プロジェクトに対して適切なソリューションを提供するデータ転送サービスのポートフォリオを提供しています。これには、ハイブリッドクラウドストレージ、オンラインデータ転送、オフラインデータ転送のニーズが含まれます。詳細については、「クラウドデータ移行」ページを参照してください。

サードパーティーツール : AWS は、すべてのお客様とそのさまざまなユースケースをサポートしたいと考えています。すべてのデータが AWS クラウドネイティブではないことは理解しております。そのため、AWSではサードパーティーのデータプラットフォームとの統合を可能にしています。例えば、人気の高いデータウェアハウスプラットフォームである Snowflake は、大規模にスケール可能なオブジェクトストレージソリューションである Amazon Simple Storage Service (Amazon S3) との統合をサポートしています。これは、S3 のアクセス管理ポリシーを介して S3 へのアクセス権を Snowflake に付与することで実現されます。これについては後で説明します。

導入事例 : データを収集して機械学習モデルをトレーニングするための取り込みパイプラインの開発に成功した、いくつかのスタートアップ企業をご紹介しましょう。最初は、個性的なファッションのマーケットプレイスで新たなショッピング体験を提供している Depop です[19]。Depop は Amazon Kinesis Data Firehose と Amazon Managed Streaming for Kafka を利用して、2500 万点に及ぶアイテムの膨大な在庫と取引をストリーミングしています。Depop はデータを取り込むためのマネージドサービスを活用することで、インフラストラクチャの管理ではなくカスタマーサービスの開発に集中することができます。次は、処方箋と請求データを分析のために取り込み、認可された医療従事者チームの介入のために事例を参照できるようにする企業である axialHealthcare です[20]。axialHealthcare のクラウドベースコンタクトセンターは Amazon Kinesis Data Streams を使用してエージェントのステータスイベントをモニタリングし、Amazon S3 を使用して通話の録音をオブジェクトとして保存しています。

depop | axial Healthcare

データを使用した分析の実行

データの理解は、機械学習モデルの開発にとって重要です。探索的データ解析[21] は、パターンの発見、異常の検出、仮説のテストなどのため、データの初期調査を行うプロセスです。さらに、手持ちのモデルのコンテキストで特徴量を分析することで、どの特徴量が他よりも重要であるかの直感が得られます。特徴量には、モデルパフォーマンスを改善させるもの、まったく改善させないもの、パフォーマンスを低下させるものがあります。AWS は、お客様が収集してクラウドに保存したデータを、追加設定なしですぐに調査することができるサービスを提供しています。

データを理解するための重要な側面のひとつは、パターンの識別です。通常、これらのパターンは、表内のデータを見ているだけではわかりません。適切な可視化ツールを使用すれば、データを短期間でより深く理解できます。表やグラフを作成する前に、何を表示するかを決めなければなりません。例えば、グラフによって、主要業績評価指標 (KPI)、関係性、比較、分布、または構成などの情報を伝達することができます。

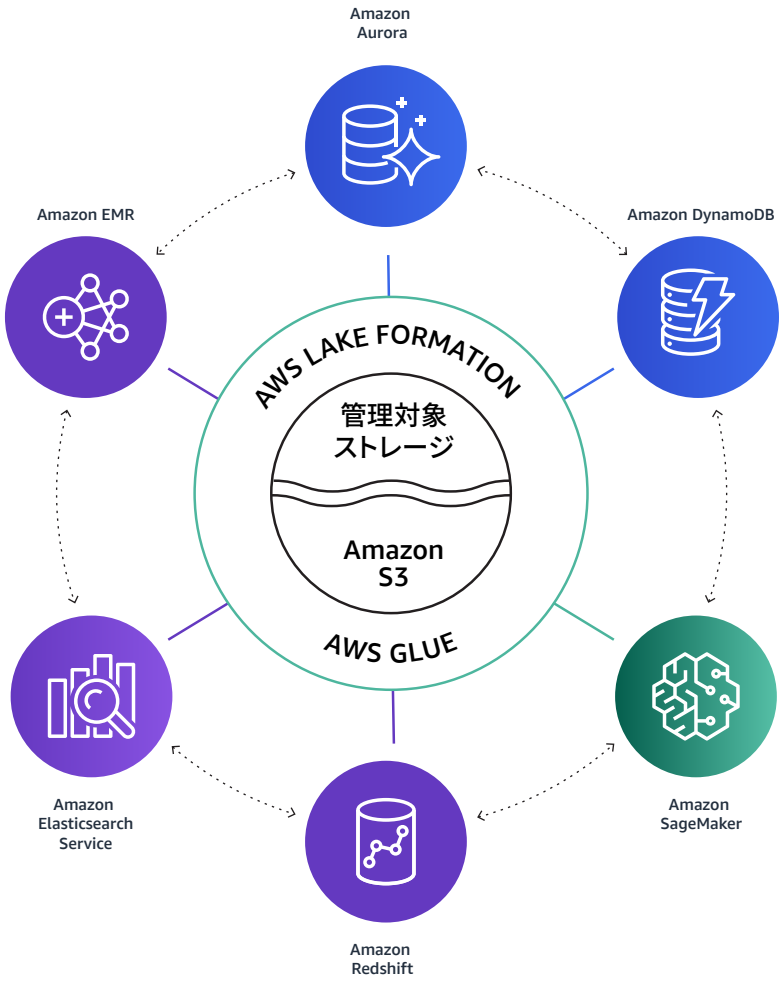
Amazon Athena は、標準 SQL を使用して Amazon S3 のデータの分析を容易にするインタラクティブなクエリサービスです。Athena はサーバーレスなので、インフラストラクチャの管理は不要です。実行したクエリに対してのみ料金が発生します。Athena では、追加の変換なしで、raw 状態で直接データをクエリできます。

Amazon QuickSight は、インタラクティブダッシュボードを簡単に作成して発行できる、機械学習を利用したスケーラブルなビジネスインテリジェンスサービスです。QuickSight は ML Insights も提供しています。これは実績のある AWS の機械学習と自然言語機能を利用して、データに隠れたインサイトやトレンドの発見、主な要因の識別、ビジネスメトリクスの予測に役立ちます。また、Amazon SageMaker に組み込まれた機械学習モデルに QuickSight を接続して、予測型ダッシュボードを作成することもできます。

Amazon SageMaker Studio は単一のウェブベースのビジュアルインターフェイスを提供します。ユーザーは機械学習開発のステップすべてをそこで実行できます。Studio には、すばやくスピンアップできるワンクリックの Jupyter ノートブックが用意されています。基盤となる処理能力は完全に伸縮自在であるため、必要とする適切な処理能力を簡単にスピンアップまたはダウンできます。データセットの分析は、pandas (Python の人気の高いオープンソースのデータ分析および操作ツール) などのツールを使用して直接ノートブック環境で開始できます。データの視覚化では、Matplotlib や Seaborn などのオープンソースライブラリを使用できます。

データの管理 – データレイク

これまで、オペレーションから蓄積されたデータはさまざまなデータサイロに保存され、分析が非常に困難になっていました。データサイロには複数の課題があります。特定のワークロードに必要なデータが複数のサイロ間に分割されてアクセスできない、データが保存されているサイロが特定のワークロードのコスト要件を満たさない、サイロごとに異なる管理、セキュリティ、認可のアプローチが必要で運用コストやリスクが増えるなどです。データレイクを使用すれば、エンタープライズ全体のデータ (構造化データと非構造化データ)をスケーラビリティ、可用性、セキュリティ、柔軟性の高いデータストアに一元的に保存してカタログ化でき、機械学習や分析などのユースケースできわめて大量のデータセットを処理できます。AWS のレイクハウスアーキテクチャは、組織がデータレイクや周辺の目的別分析ストアおよびサービスを構築できるよう支援します。また、データレイクと周辺の目的別分析ストアおよびサービス間のデータの移動 (内部から外部へ、外部から内部へ、外部を迂回) を支援し、データからインサイトを引き出します[22、23]。

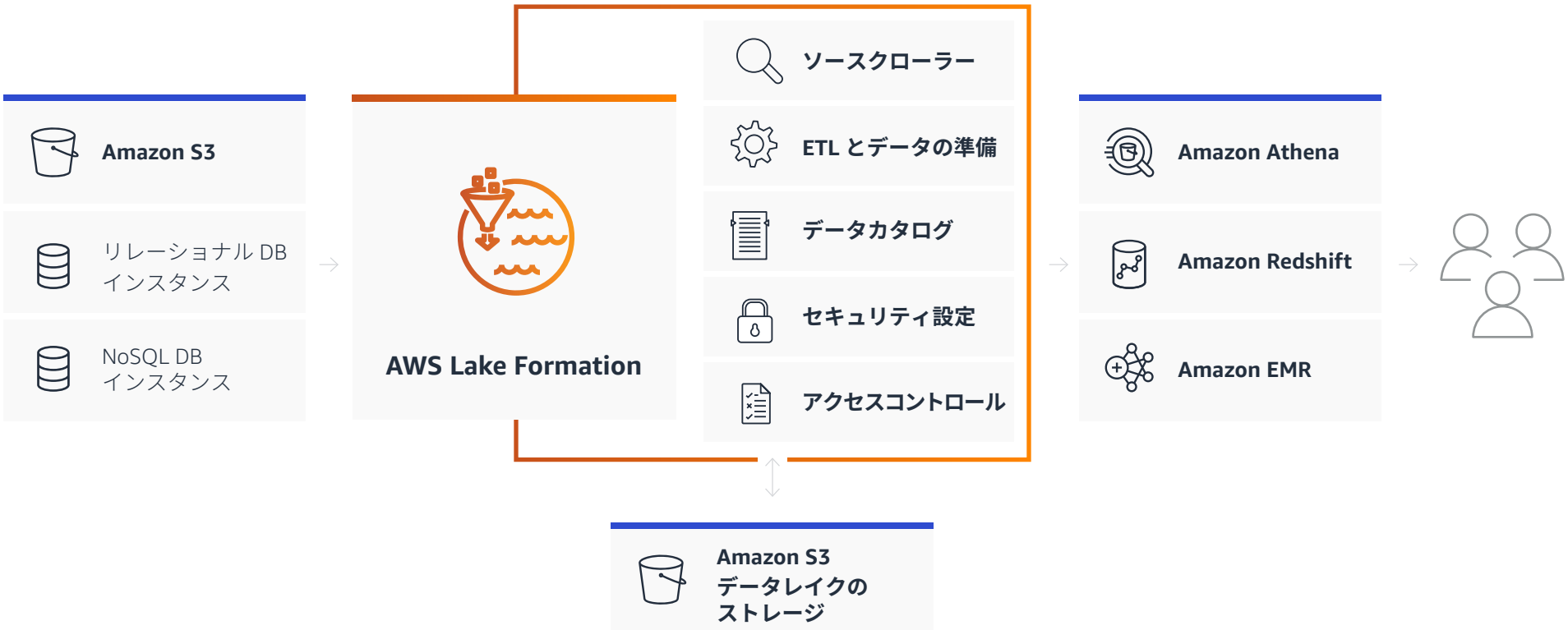


レイクハウスアーキテクチャの中心となるのはデータレイクであり、これは Amazon S3 で始まります。Amazon S3 は、構造化データと非構造化データに対して最大かつ最も高パフォーマンスのオブジェクトストレージサービスであり、データレイクの構築に最適なストレージサービスです[24]。サイズを問わず、スケーラブルでコスト効果の高いデータレイクをセキュアな環境で構築でき、データは 99.999999999% (イレブンナイン) の耐久性で保護されます。Amazon S3 はコスト効果の高い S3 ストレージクラスを幅広く提供しており、該当する料金でさまざまなデータアクセスレベルをサポートしています。S3 Storage Class Analysis を使用して、アクセスパターンに基づきより低いコストのストレージクラスに移動できるデータを発見し、S3 ライフサイクルポリシーを設定して移動を実行できます。また、変化するアクセスパターンや不明なアクセスパターンのデータを S3 Intelligent-Tiering に保存できます。それにより、変化するアクセスパターンに応じてオブジェクトを階層化し、自動的にコスト削減を実現します。クラウドプロバイダーのうちで最も早くデータレイクを構築できるようになったのが Amazon S3 です。AWS は、現在どこよりも多いデータレイクが実行されており、そのデータ分析のすべてにサーバーレスのインタラクティブクエリサービスである Amazon Athena が使用されています。

データレイクに加えて、AWS 目的別データベースと Amazon EMR、Amazon Elasticsearch Service、Amazon Redshift などの分析サービスを組み合わせて使用して、ジョブに合った適切なツールを使用し、可能な限り低コストで高いパフォーマンスとスケールを取得することができます。お客様は、サーバーレスのデータ統合サービスである AWS Glue を使用して、これらのシステム間でデータを移動しています。AWS Lake Formation により、データがデータレイクにあるか、目的別分析ストアにあるかにかかわらず、すべてのデータのセキュリティとガバナンスを管理できます。より具体的には、AWS Lake Formation はすべてのデータ間での 1 つのセキュリティおよびガバナンスコントロールアクセス、きめ細かいアクセスコントロールを使用したデータの機密性の管理、データレイク全体の一元化された監査統制により、データレイクの構築をお客様を支援します。

レイクハウスアーキテクチャに概要を示した目的別データストアの 1 つが、機械学習のユースケースを対象にしています。データジャーニーはデータレイクに取り込まれた raw データで始まり、データ変換や特徴量エンジニアリングプロセスを通じて、モデルトレーニングと推論のための特徴量が作成されます。ここで、次の図に示すとおり、チーム間でさまざまなチーム間でさまざまなモデルが共通の特徴量を共有できるように、作成された特徴量はデータサイエンティスト間でそれぞれのモデルトレーニング/推論用に保存、発見、共有される必要があります。さらに重要なことは、データサイエンティストはモデルを改善するため、時間の経過とともに新しい特徴量を追加しますが、既存の特徴量の再作成や再計算は望みません。これには時間がかかり、モデル予測のレイテンシーが増加するためです。

仕組み



Amazon SageMaker Feature Store は、モデルトレーニングと推論のため、データサイエンティストが他のデータサイエンティストと共通のデータ特徴量を共有できるよう支援します。SageMaker Feature Store は、オフラインの特徴量ストアをトレーニングのためにサポートし、オンラインの特徴量ストアをオンライン推論のためにサポートします[25]。SageMaker Feature Store はモデルトレーニングのために大規模なバッチで特徴量を提供するだけでなく、リアルタイム推論のユースケース用に短い読み取りレイテンシー (数ミリ秒) で特徴量を提供します[26]。Feature Store はモデルの特徴量の一元的なレポジトリを提供するため、一貫した特徴量を提示します。つまり、トレーニングと推論でまったく同じ特徴量が利用できるため、トレーニングと推論間で特徴量が同期しなくなることはありません。SageMaker Studio で特徴量を視覚的に検索して発見することができます。チームのすべてのメンバーは、レポジトリで特徴量を共有して再利用を進め、再処理を減らします。また、Feature Store はチーム間で統一した特徴量の定義のセットを提供するので、チームが共同で作業しやすくなります。

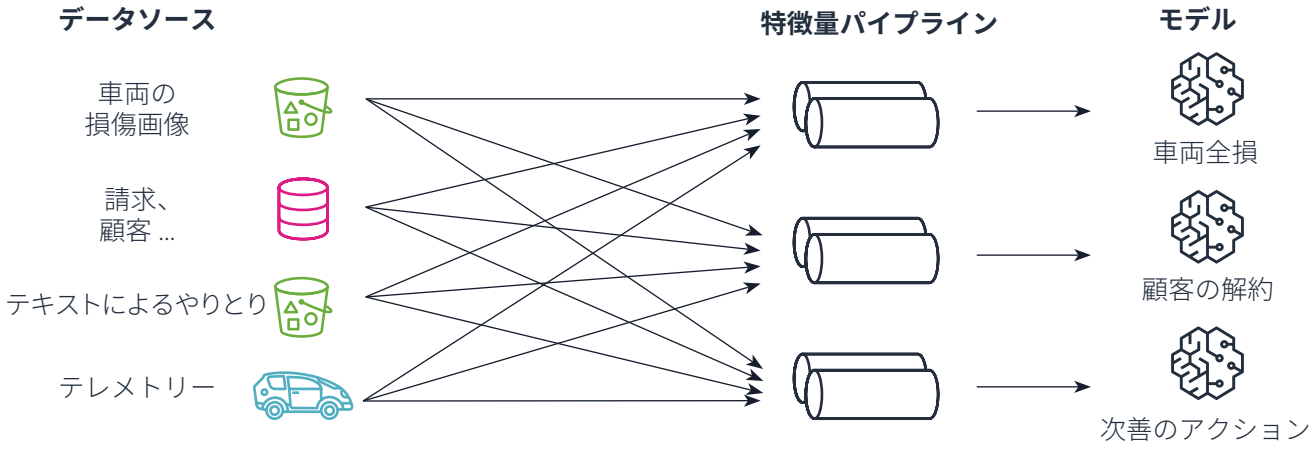
次の図は、専用のスタンドアロン特徴量エンジニアリングを使用したモデルトレーニングと共有特徴量ストアを使用したモデルトレーニングの比較を概念的に示しています。

導入事例：オンライン青果物プラットフォームである HappyFresh は、S3 でデータレイクを開発しました[27]。このスタートアップ企業は Amazon S3 にクリックストリームデータを保存し、AWS Glue を使用してデータを抽出し、顧客の購入パターンの分析のために処理しています。S3 でデータを管理することにより、HappyFresh はパーソナライゼーションサービスに特化し、顧客が商品検索で貴重な時間を失わないようにしています。

happyfresh

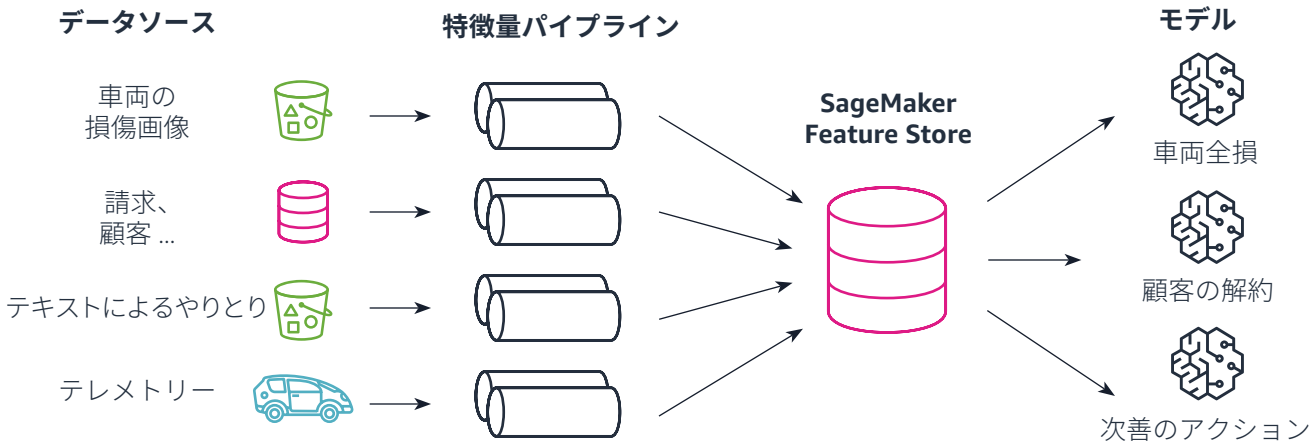
データジャーニー

新しいモデルごとのスタンドアロン特徴量エンジニアリング



- 特徴量の重複
- 市場への投入が遅い
- 不正確な予測

Feature Store を使用して特徴量を 1 回構築すれば、チームやモデル間で再利用可能



- 検索により特徴量グループが発見可能
- 再現性の高い特徴量の変換
- 正確なトレーニングデータセットの抽出
- 推論のための低レイテンシーのルックアップ
- トレーニングと推論で一貫した特徴量

データの処理

データを収集したら、モデルで効果的に使用できる形式に処理する方法を検討する必要があります。処理ステップの例には、機械学習アルゴリズムで予期される入力形式へのデータの変換、列の再スケールと正規化、テキストのクリーニングとトークン化など、数多くがあります。

機械学習モデルの善し悪しは、トレーニングに使用するデータに左右されます。データの収集後は、そのデータの統合、準備、処理などが重要になります。学習と一般化のために最適化されたトレーニングデータが、モデルの成功と正確性にとって鍵となります。データの準備は、小型の統計的に有効なサンプルから始めて、データの整合性を継続的に維持しながら、異なるデータ準備戦略を用いて繰り返し改善される必要があります[28]。AWS では、データの注釈付けと、データの抽出、転送、ロード (ETL) を大規模に行うことができるサービスをいくつか提供しています。

AWS Glue: データを作成したら、一般的に「データ統合」と呼ばれる方法で、分析と機械学習のために準備して組み合わせる必要があります。AWS Glue は、データ統合を簡単にするために、ビジュアルインターフェイスとコードベースのインターフェイスの両方を提供しています。Glue はスケジュールに従って定期的に実行するか、トリガーを通じて実行するように設定できます。例えば、新しいデータが S3 バケットに到達したときに Glue を実行できます。

AWS Batch: バッチコンピューティングを実行します。AWS は、G4 や P4 などの GPU インスタンスファミリーを提供しています。これにより、お客様はスケーラブルな GPU ワークロードを実行できます。AWS Batch により、複数のバッチコンピューティングジョブを AWS で効率的かつ大規模に実行できます。送信したバッチジョブに基づいて、最適な数と種類のコンピューティングリソースが動的にプロビジョニングされます。さらに、AWS Batch により、送信されたジョブが適切なインスタンスにスケジュールおよび配置されるため、ジョブのライフサイクルが管理されます。お客様が提供する AMI の追加により、AWS Batch ユーザーは、GPU を必要とするジョブに対してこの伸縮性と利便性を活用できます[29]。

Amazon EMR: 多くの組織は、データ処理のために Spark を使用しています[30]。この状況において、通常 Spark クラスターは、独自のセットアップ、チューニング、メンテナンスを行う必要性を減らす、Hadoop エコシステムクラスター用のマネージドサービスである Amazon EMR で実行されます。EMR では、必要なコンピューティング、メモリ、ストレージのパラメータでカスタムジョブを実行できます。自動化されたクラスターのセットアップとオートスケーリングを提供し、コスト削減のためスポットインスタンスをサポートします。EMR では、大量のデータに対して迅速かつコスト効果の高い方法で、データセットのインポート、エクスポート、結合などのオペレーションを実行できます。

導入事例 : Guru はナレッジ管理ソフトウェアを提供するスタートアップ企業です[31]。Guru は、Amazon OpenSearch サービスを使用して Elasticsearch クラスターのストレージとスケーリングを管理しています。同社は、検索エンジンの検索結果の関連性を改善するために、Amazon EMR を使用して実験フレームワークを作成することができました。次に、AiCure は、患者の行動をモニタリングし、臨床試験でリモートの患者エンゲージメントを可能にする高度なデータ分析企業です[32]。AiCure は AWS Step Functions と AWS Batch を活用して継続的に AI モデルと推論を大規模に改善し、データを実践的なものにしています。

Amazon SageMaker

データスキーマは、データを整理および操作するための優れた最初のステップです。ただし、スキーマは進化し、コードは古くなり、クエリは低速になることに留意してください。データ中心の AI はデータを改善して、機械学習のエンジニアリングおよびデータサイエンスチームを含むダウンストリームの利用者に高品質のデータを提供することに焦点を当てます。そしてこれは反復的なアプローチです。データの品質により、その場でデータ処理パイプラインが停止する可能性があります。このような問題を事前に把握しないと、誤解を招きやすいレポート、偏向がある　AI/機械学習モデル、その他の予期しないデータ製品が発生する可能性があります。

SageMaker Ground Truth: 正確にラベル付けされたデータは、教師ありモデルの成功にとって不可欠です。不正確なラベルがあると、機械学習モデルは悪い例から学習し、予測が不正確になります。SageMaker Ground Truth は、データの効率的で正確なラベル付けを支援します。自動化されたデータラベリングと人間によるデータラベリングを組み合わせて使用します。カスタムワークフローを構築して、データラベリングジョブのユーザーインターフェイス (UI) を定義することもできます。Amazon SageMaker は、使用開始を支援するため、イメージ、テキスト、オーディオデータラベリングジョブ用のカスタムテンプレートを提供しています — ブログはこちら。

SageMaker Data Wrangler は、機械学習、データ分析、特徴量エンジニアリング、特徴量の重要度分析、偏向検出に特化して設計されています。Data Wrangler は特徴量エンジニアリングと偏向軽減のため、300 を超える組み込みのデータ変換を提供しています。

SageMaker Processing Jobs は、scikit-learn や Apache Spark など使い慣れたオープンソースツールを使用して、任意の Python スクリプトまたはカスタム Docker イメージをフルマネージドで従量制料金の AWS インフラストラクチャで実行できます。このサービスは、クラスター内の多くの SageMaker インスタンスで、カスタムスクリプトまたは Docker イメージを並列化できます。SageMaker Processing では、スクリプトを提供し、インスタンスタイプとクラスターサイズを指定するだけです。

SageMaker Clarify: SageMaker Clarify は、SageMaker Data Wrangler での偏向検出に加えて、モデルトレーニングに最適な列 (特徴量) の選択の支援、トレーニング後のモデルの偏向の検出、モデル予測の説明、モデル予測入出力の統計的ドリフトの検出を行います。

準備 ⟶

SageMaker Ground Truth
機械学習用にトレーニングデータをラベル付け

SageMaker Data Wrangler 新規
機械学習用にデータを集約および準備

SageMaker Processing
組み込みの Python、BYO R/Spark

SageMaker Feature Store 新規
特徴量を保存、更新、取得、共有

SageMaker Clarify 新規
偏向を検出し、モデル予測を理解

構築 ⟶

SageMaker Studio Notebooks
伸縮自在なコンピューティングと共有を備えた Jupyter ノートブック

組み込みのアルゴリズムと持ち込みアルゴリズム
多数の最適化されたアルゴリズムと持ち込みアルゴリズム

ローカルモード
ローカルマシンでのテストとプロトタイプ

SageMaker Autopilot
完全な可視性をもって機械学習モデルを自動的に作成

SageMaker Jumpstate 新規
一般的なユースケース用の事前構築されたソリューション

トレーニングと調整 ⟶

ワンクリックトレーニング
分散インフラストラクチャの管理

SageMaker Experiments
各ステップをキャプチャ、整理、比較

モデルの自動チューニング
ハイパーパラメータの最適化

分散トレーニングライブラリ 新規
大規模データセットとモデル用のトレーニング

SageMaker Debugger 新規
プロファイルの実行をデバッグおよびプロファイル化

マネージドスポットトレーニング
トレーニングコストを 90% 削減

デプロイと管理 ⟶

ワンクリックのデプロイ
フルマネージド、超低レイテンシー、高スループット

Kubernetes および Kubeflow の統合
Kubernetes ベースの機械学習を簡略化

マルチモードエンドポイント
インスタンスあたり複数のモデルをホストしてコストを削減

SageMaker Model Monitor
デプロイされたモデルの正確性を維持

SageMaker Edge Manager 新規
エッジデバイスのモデルを管理およびモニタリング

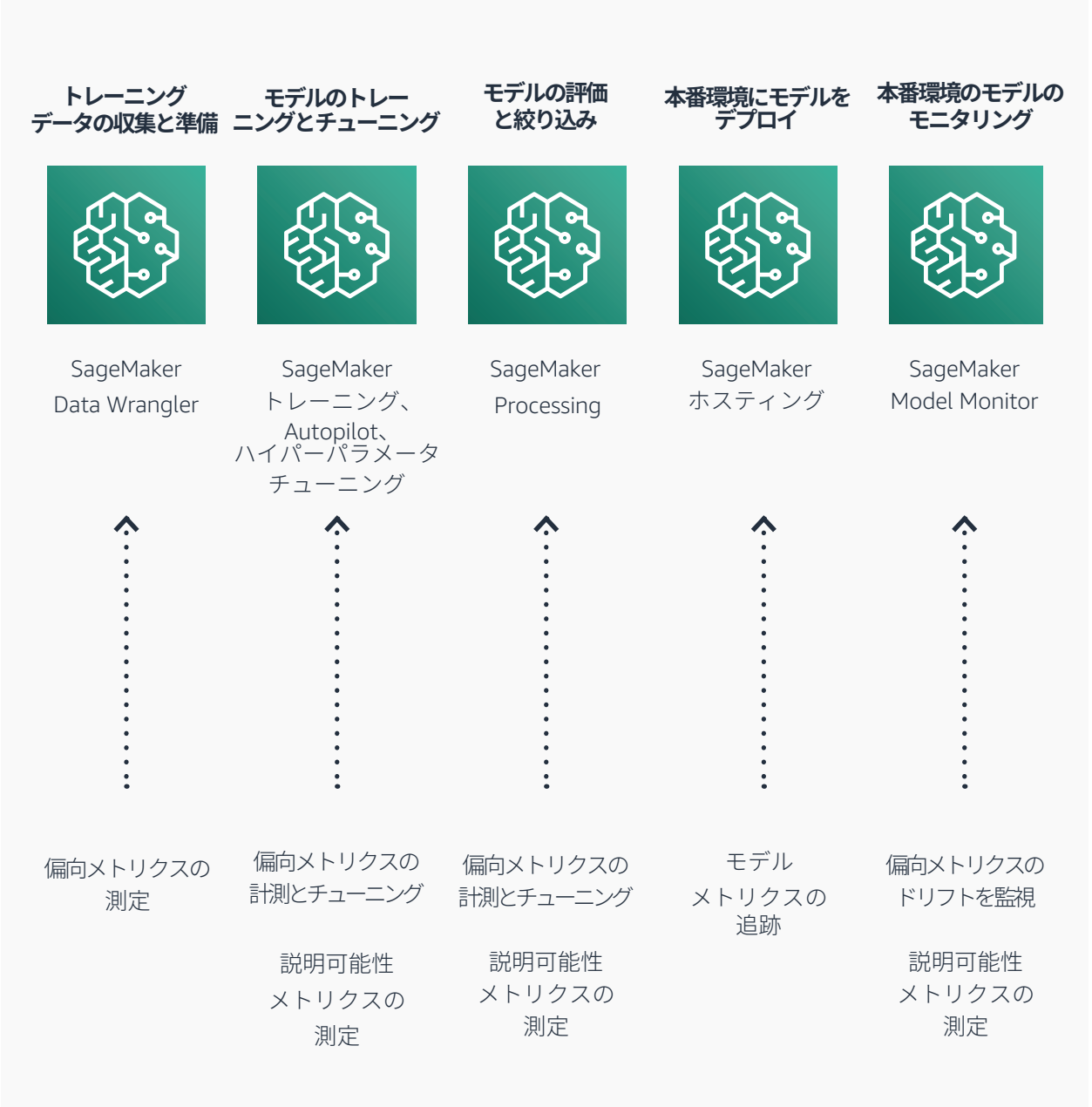
SageMaker Pipelines 新規
ワークフローのオーケストレーションとオートメーション

データの実用化

機械学習を行う際に最大の価値を得るデータの管理方法を学習した後は、モデルトレーニングとチューニングに役立つ AWS ツールについて説明します。

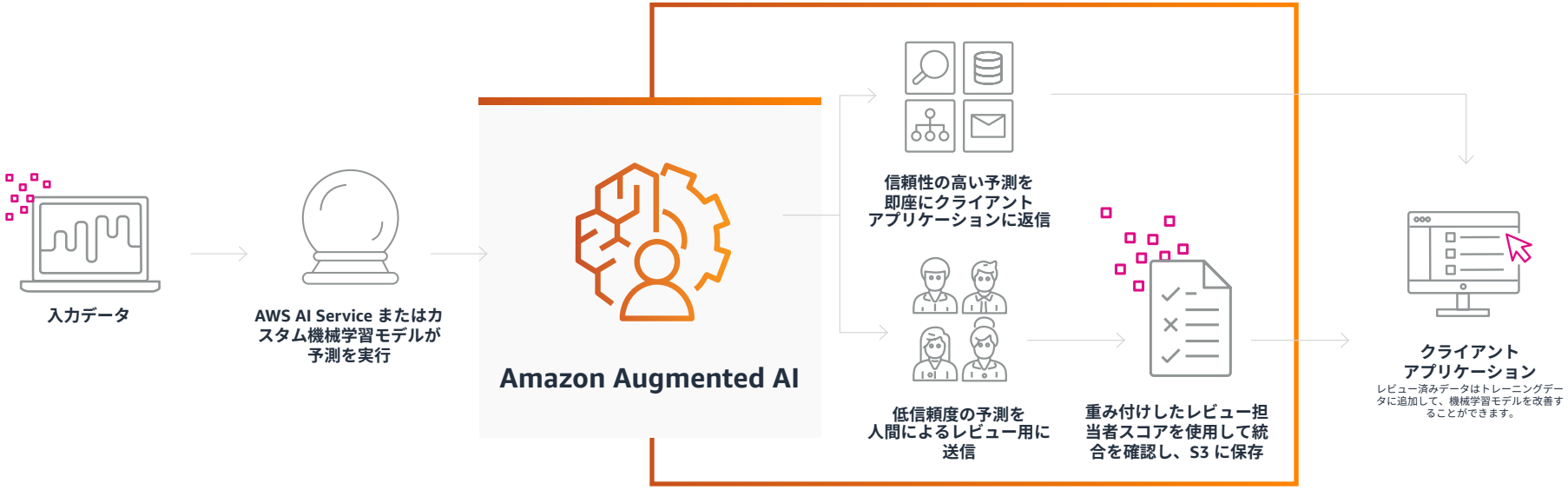
Amazon Augmented AI (Amazon A2I) を使うと、人間による機械学習予測のレビューに必要なワークフローを簡単に構築できます。Amazon A2I では、人間によるレビューシステムの構築や、多数のレビュー担当者の管理に伴う画一的で面倒な作業からデベロッパーを解放し、すべてのデベロッパーが人間によるレビューを実施できるようにします。

多くの機械学習アプリケーションでは、結果が正しいかどうかを確認するため、信頼性の低い予測を人間がレビューする必要があります。例えば、スキャンした住宅ローン申請書から情報を抽出する場合、スキャンや手書きの質が低いと、人間によるレビューが必要になる場合があります。しかし、人間によるレビューシステムの構築には時間と費用がかかります。複雑なプロセスまたは「ワークフロー」の実装、レビュータスクと結果を管理するカスタムソフトウェアの作成、さらに大抵の場合、数多くのレビュー担当者の管理も必要となります。



Amazon A2I では、機械学習アプリケーションの人間によるレビューを簡単に構築、管理できます。Amazon A2I には、コンテンツのモデレーションやドキュメントからのテキスト抽出など、機械学習の一般的なユースケース用の人間によるレビューワークフローが備わっています。これにより、Amazon Rekognition や Amazon Textract からの予測を簡単にレビューできます。Amazon SageMaker やその他のツールで構築した機械学習モデル用に、独自のワークフローを作成することも可能です。Amazon A2I を使うと、モデルで信頼性の高い予測を行えない場合や、その予測を継続的に監査できない場合に、レビュー担当者が介入できるようになります。

- Amazon A2I は、データセットラベルに組み込むこともできます。このブログ (英語) は、Amazon A2I を使用して低信頼度データをレビューおよび強化する方法を詳しく示しています。



Amazon SageMaker Studio: 前述したように、SageMaker Studio は 1 つのウェブベースのビジュアルインターフェイスを提供します。このインターフェイスでは、すべての機械学習開発ステップを実行し、データサイエンスチームの生産性を向上させることができます。ノートブック、実験管理、自動モデル作成、デバッグなどを含む、すべての機械学習開発アクティビティを Studio 内で実行できます。このガイドでは、データ管理の各ステージに合ったさまざまな AWS のサービスについて説明しました。次の表に示すように、Studio はデータ管理に関する SageMaker の機能と機械学習をネイティブに統合します。Studio を使用することで、チームはワークロードを 1 つのビューに統合して管理を容易にし、コラボレーションを高速にできます。

- ワークロード/パフォーマンスがチームにとって重要な場合、Amazon FSx for Lustre を Amazon SageMaker の入力データソースとすることができます。FSx for Lustre を入力データソースとして使用すると、Amazon SageMaker の機械学習トレーニングジョブは、最初の S3 ダウンロードステップを除外して加速されます。SageMaker ジョブは、S3 バケットとリンクされた FSx for Lustre ファイルシステムが作成されたらすぐに開始できます。S3 から完全な機械学習トレーニングデータセットをダウンロードする必要はありません。データは、処理ジョブに対して Amazon S3 から必要に応じて遅延読み込みされます。別の利点として、TCO の削減があります。これは、同じデータセットの反復ジョブに対して、共通オブジェクトの繰り返しのダウンロードを避けることによって実現されます (S3 リクエストコストが削減されます)。

プロジェクト	継続的統合および継続的デリバリー (CI/CD) を使用して、モデルの構築とデプロイパイプラインを自動化
Data Wrangler	機械学習のデータをすばやく集計および準備
Feature Store	機械学習の特徴量を保存、更新、取得、共有するフルマネージドの目的別レポジトリ
Pipelines	エンドツーエンドの機械学習ワークフローを大規模に作成、自動化、管理
Experiment と Trial	機械学習実験を整理、追跡、評価
モデルレジストリ	モデルの系統とメタデータを追跡
エンドポイント	予測エンドポイントを追跡

まとめ

現在の機械学習環境において、モデルの構築を成功させるためにデータは不可欠です。このガイドでは、機械学習モデルで使用する可能性のあるデータソースと、データの品質およびボリュームを測定するための方法について説明しました。次に、データを管理し、機械学習用に準備するために必要な各ステップを通じて、AWS のサービスと機能がどのように役立つかについて説明しました。

データ管理プロセスに精通し、機械学習用にデータの使用開始を検討する際は、ぜひ AWS の機械学習ページにアクセスして機械学習サービスについてご確認いただき、お客様の成功事例や一般的なユースケースをご覧ください。AWS のミッションは明確です。機械学習をすべてのデベロッパーにとって身近なものにすることです。AWS はお客様がデータを理解して活用できるよう支援いたします。

AWS にお問い合わせいただければ、お客様がデータジャーニーと機械学習の導入を加速できるよう支援いたし ます。

クラウド上の機械学習アプリケーションの構築、デプロイ、実行に必要となる無料のオファーやサービスをご用意しています。

今すぐ開始する >>

リファレンス

[1] トレーニングデータの品質を定義して測定する方法
[2] データ品質の評価
[3] データ品質を測定する方法 – データ品質を評価するための 7 つのメトリクス
[4] ビッグデータと機械学習におけるデータ品質の考慮事項 : データのクレンジングと変換を越えて
[5] 機械学習の自動化されたデータ品質管理に向けて
[6、7] 分類のパフォーマンスに必要なサンプルサイズの予測
[8] 大規模なトレーニングセットがどのように必要になるのか？
[9] 分類モデル用のサンプルサイズの計画
[10] データセットサイズとデータディメンション、原則はあるのか？
[11] 必要なトレーニングデータの量は？
[12] 時系列モデルの確認に必要な最小数は？
[13] 経済的な問題や環境上の問題に対するアプリケーションによる介入分析
[14] 深層学習モデル – CNN をトレーニングするために必要な最小サンプルサイズは？
[15] トレーニングデータが十分あることを確認する方法
[16] ニューラルネットワークのトレーニングに必要なイメージの数は？
[17] テキスト分析 101: ドキュメント分類
[18] 感情分析に本当に必要なテキストの量は？
[19] Depop の導入事例
[20] AxialHealthcare の導入事例
[21] 探索的データ解析とは？
[22] レイクハウスアプローチとは？
[23] AWS Lake House からインサイトを引き出す
[24] AWS でのデータレイクストレージ
[25] AWS でのセキュアなエンタープライズ機械学習プラットフォームの構築
[26] Amazon SageMaker Feature Store での機能の作成、保存、共有
[27] HappyRefresh の導入事例
[28] Well-Architected 機械学習レンズ – データの準備
[29] AWS Batch での深層学習
[30] AI/機械学習のデータ処理オプション
[31] GURU の導入事例
[32] AiCure の導入事例


aws
startups