



2022年 データサイエンスおよびアナリティクスの 6大トレンド

データクラウドが機械学習を加速させる



CHAMPION
GUIDES

EBOOK

TABLE OF CONTENTS

- 3** はじめに
- 4** 予測分析および処方的分析へ向かう動きは今後も進展
- 5** トレンド#1: 使いやすいMLツールがデータアナリストとデータサイエンティストの業務をパワーアップ
AutoML
API経由で提供されるAIサービス
- 6** トレンド#2: 統合プラットフォームによりアナリティクスとML間のギャップを解消
- 7** トレンド#3: Snowflakeデータクラウドが新しいデータへのアクセスを拡大
- 9** トレンド#4: 特徴量ストアを利用した大規模なML特徴量の管理と展開
- 10** トレンド#5: 新世代分散型トレーニングフレームワークがSparkに代わる有望な代替手段に
- 11** トレンド#6: 続々とリリースされる新製品がMLライブラリ、ツール、フレームワークの新たな選択肢を提供
- 12** 2022年、あなたの機械学習が加速する
- 13** Snowflakeについて

はじめに

過去10年間でデータサイエンスは大きな発展を遂げましたが、多くの企業がデータサイエンスや機械学習 (ML) への積極的な投資を続ける一方で、高度なアナリティクスの真の成果や競争面でのアドバンテージを十分に享受している企業は極めて少数です。いったいなぜでしょうか？それは、大規模なMLの運用に必要なツールの多くが非常に複雑で、それらを使いこなすのに必要なスキルセットが不足しているためです。しかし、そんな状況にもいま変化が訪れようとしています。近年の技術進歩は、データサイエンティストやデータアナリストの業務のあり方を大きく変えようとしています。2022年、MLをさらに進展させ、企業を記述的・診断的アナリティクスから予測的・処方的アナリティクスへと進ませる6つの大きなトレンドが浮上しています。記述的・診断的アナリティクスは「過去に何が起きたか、そしてその理由は何か」を解明する技術ですが、これに対し予測的・処方的アナリティクスとは「これから何が起きるか」を予測し、「その未来を変えるためにはどうすれば良いか」について重要な指針を提供する技術です。

本eBookでは以下について学習します。

- データサイエンティストがフレキシブルな言語プログラミングを活用でき、さらにデータアナリストが使いやすいMLツールや統合データプラットフォームのメリットを享受できれば、MLとアナリティクスの間のギャップは解消されます。
- Snowflakeのデータクラウドは、セキュアなエコシステムを介してすぐに利用可能なサードパーティデータへのアクセスを提供することにより、データシェアリング、および非構造化データも含めたさまざまなデータタイプの利用を拡大します>(*非構造化データのサポートは現在プレビュー提供中です)
- 特徴量ストアによる特徴量の再現性、発見可能性(見つけやすさ)、スケーラビリティの実現により、データサイエンティストはML特徴量を大規模に管理・展開できます。
- 新しい分散型トレーニングフレームワークはパフォーマンスを最大2,000倍高速化し、Sparkより優れた代替手段となりつつあります。
- MLライブラリ、ツール、フレームワークの急速な発達に伴い、データサイエンスやML分野の投資をフューチャープルーフ化するための新たなソリューションのニーズが浮上しています。

予測分析および処方的分析へ 向かう動きは今後も進展

2022年、データサイエンスはようやく、多くの企業が長年期待してきたレベルに到達しようとしています。過去10年間にわたり、データサイエンスとMLはこれからの企業のビジネスのあり方を全く新しいものに変えるであろうとの期待のもと、データサイエンスおよびML分野に大規模な投資が行われてきました。しかし多くの企業は依然としてアナリティクスの真の効果を楽しんでいないとは言い難く、MIT Sloan Management ReviewとBoston Consulting Groupの発表したレポートによれば、AIへの投資により有意な経済的メリットが得られていると回答したのは全企業のわずか10%に留まっています。¹

企業がデータサイエンスに投資するのは競争面でのアドバンテージが得られることを期待するからですが、MLの規模拡大に必要なツールやスキルセットの多くはどこにも見つけられないか、あっても不十分です。データサイエンティストは依然として需要が大きく高コストな人的リソースですが、そのように貴重なデータサイエンティストの業務時間の多くがデータの選定やデータ準備などの煩雑な作業に費やされています。一方、多くの企業は充実したデータアナリストチームを擁しており、彼らには業務上の課題にダイレクトに対処する能力がある一方で、アナリティクスからデータサイエンスのレベルに進んで自らMLモデルを構築するには必要な技術的バックグラウンドが不足しているのが現状です。

2021年中の進展から続く動きとして、今年2022年、MLおよびデータサイエンスの分野では6つの興味深いトレンドが浮上しています。毎月続々と新しいツールや技術がリリースされており、データサイエンティストの業務をスピードアップすると共にデータアナリストが記述的アナリティクスの先へ進んでライトなデータサイエンスやMLに取り組むことを可能にしています。

こうした前に進む動きを支えるもの、それはクラウドです。データサイエンティスト、データアナリストのいずれもが、仮想空間ではほぼ無制限なコンピューリソースを提供するクラウド技術のメリットを享受しています。またクラウドはデータレイク、データウェアハウス、データマートの統合によりデータのサイロ化を解消し、ユーザーはスピーディーかつセキュアで使いやすいデータシェアリングやアナリティクスを1か所から利用できます。

クラウドの普及によりデータは実効性のあるアセットとなりつつあり、これを活用する新たなツールの登場がMLをさらに前進させています。その結果企業はいよいよ、データをモビライズして、未来の予測だけでなく処方的アナリティクスにより特定の望ましいシナリオの確率を高める能力を手に入れようとしています。

2022年、データサイエンスのあり方を形づくり、アナリティクスをMLの領域へと近づける6つのトレンドは何かを次ページからご説明します。



トレンド#1: 使いやすいMLツールが データアナリストとデータサイエンティストの 業務をパワーアップ

多くの企業では、充実したデータアナリストチームを擁する一方でデータサイエンティストの数は十分ではありません。その主な理由は、データサイエンティストの人材が不足しており採用には大きなコストがかかるからです。データアナリストはMLモデルの構築のために必要なデータサイエンススキルが不十分なため、ML運用規模拡大において典型的な課題となっているのがデータサイエンティストの確保です。

しかし、データサイエンスのテクニカルな要素を自動化することで、改善された新しいMLツールが次々と登場してMLへの入口を大きく開こうとしています。こうしたツールの利用により、データアナリストは手動での構築作業なしにパワフルなMLモデルにアクセスできるようになり、一方データサイエンティストは複数の処理ステップを自動化して生産性を向上できるようになりました。いま、自動化された機械学習つまりAutoMLと、API経由で提供される各種AIサービスが、手動でのデータ準備やモデルの構築・トレーニング作業を不要にしつつあります。

AUTOML

AutoMLツールとは、まさにその名の通り「MLを自動化するツール」です。AutoMLツールは、これまでデータサイエンティストが手動で行ってきたMLモデルの開発や展開に必要なさまざまなタスクを自動化します。AutoMLは、データサイエンティストにとってもデータアナリストにとってもこれまでの業務のあり方を一変させるテクノロジーです。なぜなら、AutoMLツールの利用により、データユーザーはデータ準備やモデルのトレーニング・選定などMLワークフローの特定の部分を自動化することができるようになるからです。

しかもAutoMLはアナリストのためだけのツールではありません。AutoMLの優れた機能はこれまでデータサイエンティストの業務時間の最大80%を占めていたデータのロード、選定、準備、クリーニングなどの煩雑な作業を業務時間の45%に圧縮する効果があり（Anacondaが実施しDatanamiが報告するデータサイエンティスト調査結果による）、データサイエンティストの業務のあり方を大きく変えています。² これらの煩雑で時間のかかる作業を排除することで、データサイエンティストの生産性は向上し、分析作業により多くの時間を掛けることが可能になります。また、手動でのモデリング作業における人的エラーも減少し、モデル精度の向上につながっています。

従来AutoMLについて指摘されてきた点としてツール自体がブラックボックス化しているという課題がありましたが、これについてはほぼ解消されています。現在のAutoMLサービスはモデルについて十分な透明性や説明性を提供しており、監査やバイアス検出において非常に重要なメリットとなっています。データサイエンティストにとってのAutoMLの大きなメリットは、複数のモデルを並行して素早く構築しテストできる点です。

DataRobot、Dataiku、H2Oなどのプロバイダーが提供するAutoMLツールはここ2年ほどで大きく進化し、一方でAmazon SageMaker Autopilotなどのソリューションも導入と普及が進んでいます。2021年には、モデルの展開とモニタリングの自動化が進んだことでMLモデルの実用化が大きく進展しました。

API経由で提供されるAIサービス

データ分野で人気が高まっているもう1つのアプローチがAIサービスで、現在ではすぐに使えるモデルがAPI経由で提供されています。これにより企業は、ごく一般的なアクティビティの

ためにモデルを1から構築してトレーニングする必要がなくなり、特定のタスクを実行するトレーニング済みのモデルに簡単にアクセスできます。必要なタスクが自然言語処理（NLP）であっても、自動音声認識（ASR）であっても、あるいは画像認識であっても、APIを介してアプリケーションにAIサービスをプラグ&プレイするだけでタスクが実行され、データサイエンティストの関与は一切必要としません。

Amazonは、Amazon Lex、Polly、Rekognition、Forecast、Translateなど一連の完全管理型AIサービスを提供しています。³ その優れた能力の一例として、例えばRekognitionはアプリからAPIを経由してAmazonに画像を送信すると、AIサービスが受信した画像の分類と記述を返信してそれが何の画像なのかを教えてください。これらのユーティリティは単に時間や手間を節約するだけではなく、データサイエンティストを一般的なサービスの再構築といった作業から解放し、それぞれのビジネスに高度にカスタマイズされたモデルの構築やトレーニングに注力できるようになっています。

AutoMLツールやAIサービスはML導入へのハードルを下げ、学識経験がなくても誰でもデータサイエンスにアクセスすることを可能にします。しかし、これらのツールがその真の価値を発揮するのは、ユーザーの持つ既存技術とシームレスに統合された時でしょう。**Snowflake Partner Connect**を導入すれば、企業はSnowflake製品と技術パートナーの製品の間に事前構築されたインテグレーションを通してデータからより迅速にインサイトを得られるようになります。Snowflake Partner Connectでは、さまざまな最新MLツールやサービスに気軽にトライして御社のビジネスニーズに最も適した製品を選んでもいただけます。

トレンド#2: 統合プラットフォームにより アナリティクスとML間のギャップを解消

1つの組織の内部や組織間などデータサイロがあらゆる所に存在するのは良く知られていますが、見逃されがちなのは、組織上のデータサイロだけでなく、特にデータサイエンティストとデータアナリストの間にいわゆる「アナリティクスサイロ」が存在することです。こうしたアナリティクスサイロは、データサイエンティストとデータアナリストが用いる業務手法やスキルセットの相違により発生します。また、データサイロは、データサイエンティストとデータアナリストとを隔てる障壁の1つに過ぎません。データサイエンティストとデータアナリストは、取り扱うデータの種類(生データ vs 処理済みデータ)も、データソース(データレイク vs データウェアハウス、データマート)、言語(Python、Java vs SQL)、ツール(ML vs BI)もそれぞれ異なります。

組織上のサイロと同様に、アナリティクスサイロはデータサイエンティストとデータアナリスト間のスムーズなコラボレーションや統合を阻害します。その結果、企業は本来「1+1」を越えた相乗効果が期待できるはずのデータサイエンティストチームとデータアナリストチームの連携により得られるべき成果を取り逃がしています。

例えばデータアナリストはデータを利用して主要なビジネス指標を提供し「**は何故起きたのか?」の疑問に答えを出そうとします。Sisuによればデータアナリストの真の価値はそのスピードにあり、データセットを素早く分析し同時にステークホルダーとやり取りしながらデータに秘められたインサイトを発掘します。⁴ データアナリストが目指すのは企業が市場の機会を収益化するのを支援することですが、データアナリストは予測MLモデルの構築に必要なデータサイエンススキルを持たないため、その取り組む業務の大部分は過去の事象に関するものとなります。そのためデータアナリストはBIツールに依存して分析を行います。そのダッシュボードには基本的に制限要素が伴います。ダッシュボードはデータに基づいて「これまでに何が起きたのか」を知るには有効ですが、データアナリストにとって、データをさらに能動的に深く掘り下げて「これから何が起るのか」「どうやってそれに影響を及ぼすことができるのか」を解き明かすのは容易ではありません。

一方データサイエンティストには、未来を予測するだけでなく今後のビジネスの成果に影響を与えるMLモデルを構築する能力があります。しかし、データアナリストに比べてダイナミックで目まぐるしく変化するビジネス環境に対応する能力は不足しがちです。Sisuはデータサイエンティストを「深く狭い働き手」

と表現しています。データサイエンティストは未知の課題に取り組むことにより万が一投入した時間やエネルギーが無駄になることを嫌う傾向があり、そのため企業によるデータサイエンスの取り組みは既知の課題(データアナリストが発見した課題であることも多い)の解明に向きがちです。⁵

Snowflakeデータクラウドが提供するツールは、より意義のある成果やスケールの実現を可能に似します。Snowflakeはアナリティクスサイロを解消し、共有特徴量ストアの提供やデータエンジニアリングパイプラインの再利用により、アナリティクスもMLもガバナンスと一貫性を備えた共通の指標やデータを用いて実行することができます。データサイエンスから得られるインサイトをSnowflakeプラットフォーム上で共有し、データアナリストがこれらのデータにアクセスしダッシュボードと分析に取り込むことで、データサイエンスチームが構築するモデルの価値を最大限に高めることができます。

また**Snowpark**(現在パブリックプレビュー提供中)はデータクラウドにデータのプログラマビリティとフレキシビリティを付与する開発者フレームワークです。Snowparkは、言語サポートをSQL以外にも拡大することでデータアナリストとデータサイエンティストを隔てる言語の壁を解消します。データサイエンティストはそれぞれ自分の使い慣れたプログラミング言語(Python(プレビュー提供中)、Java、Scala)で作成し、さらにデータを移動させることなくMLワークフローの一部としてデータ処理を完了できます。つまり、データサイエンティストとデータアナリストがデータクラウド上で同じデータを介してコラボレーションできるのです。

トレンド#3: SNOWFLAKEデータクラウドが新しいデータへのアクセスを拡大

IDCによれば、2020年中に64.2ゼタバイトのデータが新規作成もしくは複製されました。これは事前の予測を上回っており、世界的なコロナウィルスの流行がデジタルデータの利用に及ぼした影響の結果と考えられます。⁶ Analytics Insightの報告するIDCの予測によれば、2025年には世界のデータの80%が非構造データとなる見込みで、現状非構造化データの0.5%しか分析されていないことを考えると企業にとっては憂慮すべき状況と言えます。⁷

このように非構造化データの増加が予測されることから、データサイエンティストは構造化・半構造化データの分析と並んで非構造化データの分析需要に対処を迫られることになるでしょう。非構造化データは、画像、動画、音声、PDF、およびその他多様な業界固有のファイル形式を含む複雑なデータを格納するデジタルファイルの集まりで構成されています。こうした複雑なデータは他のデータタイプとの統合が極めて難しく、すべてのデータタイプに対応する一元化されたデータソースがなければ非構造化データがサイロ化することは必至です。その結果、データサイエンティストにとっては非構造化データの検索、分析、クエリを行うことが困難となり、複数のシステムから個別にデータ収集せざるを得なくなります。

こうしたデータ管理上の課題に加え、企業にとって、自らがビジネストレンドや競争トレンドの解明に必要とするすべてのデータを漏れなく独自に生成もしくは取得することはほぼ不可能です。データから最大限の価値を引き出すには、組織内および組織間でのデータセットの共有や統合が最良のアプローチであるという認識がますます広がっています。そのためデータサイエンティストやデータアナリストは、外部データの付加によりMLモデルや分析を補完して精度を上げるために絶えず新しいデータを探しています。

Snowflakeでは、データサイエンティストやデータアナリストは一元化されたグローバルなシステムにアクセスして非構造化データを含むすべてのデータタイプを管理することができます（非構造化データのサポートは現在プレビュー提供中です）。今後もより多くの非構造化データが生成されていく中で、Snowflakeのデータクラウドは統合され単一のデータソースを提供することで、データサイエンティストとデータアナリストによるデータのモデリングをスピードアップしあらゆるデータからより多くの価値を引き出すことを可能にするでしょう。

こうした多彩なデータタイプのサポートに加え、Snowflakeはサードパーティデータの利用向けに、セキュリティとガバナンスを備え、しかもフレキシブルでシームレスなアクセスを提供します。Snowflakeの提供するサードパーティへのアクセスは以下の3タイプです。



- 1 **Snowflakeデータクラウド**では、Snowflakeのカスタマー、パートナー、データプロバイダー、データサービスプロバイダーがそれぞれ自らのデータに接続して他のユーザーとシームレスに共有し、他のユーザーの提供するデータやデータサービスを利用することができます。Snowflakeプラットフォームに支えられたデータクラウドは、データサイロの壁を打ち破り、すべての企業が統合された1つのデータに接続できるようにします。またデータクラウドは、急速に増加する商品化されたデータセットに適切なガバナンス下でしかもスピーディーかつ簡単にアクセスすることを可能にするシームレスなデータ利用ソリューションでもあります。
- 2 Snowflakeデータクラウドの基盤要素の1つが、従来のデータ移行に伴う障壁を解消する**Snowflakeセキュアデータシェアリング**技術です。Snowflakeでは原則的に、データをコピーしたり送信することは一切なく、ユーザーはデータが元の場所からそのままライブデータを共有できます。適切なガバナンスとセキュリティの下でデータを参照するというこのシンプルなアクセス方式により、レイテンシーや同時実行ユーザー間の競合に悩まされることなくデータを利用できます。データの変更は1つのバージョンにのみ行われるため、データはすべてのユーザーに対して最新状態に維持され、その結果常に最新データに基づくデータモデルの稼働が保証されます。

- 3 Snowflakeセキュアデータシェアリングは、すぐにクエリ可能なライブデータへのアクセスを提供する一元化されたロケーションである**Snowflakeデータマーケットプレイス**の基盤となる技術です。セキュアでガバナンスされたデータをエコシステムを構成するビジネスパートナー、サプライヤー、カスタマー間でやり取りし、さらにはサードパーティデータプロバイダーやデータサービスプロバイダーの提供するデータも利用可能です。Snowflakeデータマーケットプレイスは、欲しいデータセットを見つけたり、ベンダーとの契約を締結したり、あるいは内部データとの相互運用性確保のためにデータを管理するなどの煩わしい業務なしにすぐにデータを利用できるソリューションです。データサイエンティストやデータアナリストはこれまでより簡単に新しいデータを取得できるようになります。Snowflakeマーケットプレイスでのデータ共有以外にも、企業はSnowflakeセキュアデータシェアリングを利用して信頼できる特定のパートナー、サプライヤー、ベンダー、カスタマーとデータを共有することができます。

すべてのデータクラウドユーザーは、ほんの数クリックで外部データにアクセスして利用できます。データクラウド内に格納され次第、データは直ちに共有と利用が可能になり、CSVファイルを送信したり手動でのバージョン管理を行う必要はありません。データサイエンティストは、あらゆるトピックについてリアルタイムデータや変遷する状況を含めたほぼ無限大のデータにシームレスにアクセスしてモデルのエンリッチメントを行うことができます。



トレンド#4: 特徴量ストアを利用した大規模なML特徴量の管理と展開

データサイエンティストが新しいMLモデルを構築する際に必要となる時間のかかる作業が、データの準備と特徴量の作成です。特徴量の作成は、特定の形式のデータ列を見つけて準備し機械学習モデルに投入することにより行います。1つのモデルについて特徴量を作成できたら、データサイエンティストはさらに次のモデルのために同じ特徴量を書き換えるか、あるいは既存の特徴量を検索して場所を特定するという作業に取り組まなければなりません。

幸い2021年中にML特徴量の貯蔵庫である特徴量ストアの普及が急速に進み、特徴量の検索性、コラボレーション性、スケーラビリティが大幅に向上しました。データサイエンティストは特徴量ストアから必要な特徴量をすぐに見つけて変換し利用可能な状態にできるため、結果としてテストと生産化までにかかる時間は短縮されます。

特徴量ストアの大きなメリットの1つは、他のデータサイエンティストが作った特徴量を再利用することでチーム間のコラボレーションが促進されることです。また特徴量ストアの利用により、データサイエンティストは多くの場合すでに存在するデータパイプラインを再定義する必要がなくなるため、トレーニング済みモデルの展開に必要な時間と労力が軽減されます。

特徴量ストアは、各チームがそれぞれのモデルに必要な強化済み・リファインメント済みデータに簡単にアクセスできることから、MLモデル強化に向けた最良のアプローチとして広く認識されつつあります。しかし、特徴量ストアの構築はそれほど簡単では有りません。特徴量をすぐ稼働可能な状態で提供するには、特徴量ストアに再現性、発見可能性(見つけやすさ)、スケーラビリティを予め組み込んでおく必要があります。

- 1 モデルの再現性の確保**には、特徴量を1つのロケーションに集約させ、データや特徴量のバージョン管理を適切に行う必要があります。データサイエンティストは、時間を遡ってモデルのトレーニングに使用した特徴量やデータを見つけることができなければなりません。また一度定義した特徴量はその後のユースケースにも利用可能である必要があるため、特徴量ストア自体の定期的更新と適切なバージョン管理が必要となります。
- 2 発見可能性(見つけやすさ)**の確保には、特徴量の作成を個別のノートブックインスタンスではなく統合リポジトリで一元管理する必要があります。コラボレーションを可能にし特徴量ストアの効率的な再利用性をサポートするには、特徴量をスムーズに検索し見つけることを可能とするカタログの提供が必要となります。
- 3 特徴量ストアのスケーラビリティ**とは、MLユースケースの増大に合わせて格納する特徴量も拡張させられることを意味します。特徴量ストアは数百から数千の規模の特徴量に対応する必要があり、トレーニングと推論の両ワークフローを効率的にサポートし続けなければなりません。

Snowflakeは特徴量ストアの構築のための2種類のアプローチを提供しており、いずれのアプローチでも新しいシステムの構築は不要で、またデータサイエンティストとデータアナリスト間でデータがサイロ化することはありません。

1つ目のアプローチはSnowflake上に直接特徴量ストアを構築するやり方です。これにより特徴量は1つのデータプラットフォーム上に維持され、既存のインジェスチョン、ELT、カタログツールによりサポートされます。データサイエンティストはこの一元化されたスケーラブルなロケーションで必要なデータを見つけてアクセスし、機械学習モデルのトレーニング、推論、スコアリングを素早く簡単に行うことができます。

2つ目のアプローチは、Snowflakeパートナーを利用して特徴量ストアを構築するやり方です。カスタマーデータは生データもしくはモデリングされたデータ形式でSnowflakeデータクラウド内に維持され、そのデータの上にSnowflakeパートナーアプリケーションを利用してセマンティック層を構築し、特徴量の管理、モニタリング、提示を行います。このアプローチに利用できるパートナーは、Tecton、Rasgo、AtScale、Iguazio、Hopworks等です。

トレンド#5:新世代分散型 トレーニングフレームワークが SPARKに代わる有望な代替手段に

データサイエンティストは、常にモデルのトレーニングや展開作業を効率化できる戦略はないか模索しています。最近、新世代の分散型トレーニングワークフレームがいくつか登場し、Apache Sparkを大きく凌ぐ驚異のスピードとパフォーマンスの提供によりデータサイエンティストの効率化ニーズに応えるソリューションとして関心を集めています。

特に注目を集めているソリューションの1つが、Pythonで構築される分散型トレーニングフレームワークであるDaskです。⁸ Daskはデータサイエンティストによるモデル精度向上のスピードアップを目的に設計されています。データサイエンティストはPython上であらゆるタスクをエンドツーエンドに遂行できることから、Sparkのように実行のためにコードを変換する必要がありません。これにより、複雑性が緩和され効率が向上します。

もう1つのPythonベースのオープンソースフレームワークがRAPIDSで、Daskの上に構築されるソリューションです。⁹ RAPIDSは、データパイプラインの提供に加えてデータサイエンスコードをCPUでなく完全にGPU上で実行することによりコンピュータ時間と速度を最適化します。Saturn Cloudは最近RAPIDSとSparkの比較を実施しました。それによると、RAPIDSを用いたモデルトレーニングは20ノードGPU クラスタ上で1秒で完了し、一方Sparkでは同様の価格の20ノードCPUクラスタを用いて完了までに37分を要しました。これに基づきSaturn Cloudは、GPUを用いたRAPIDSの処理速度はSparkの2,000倍早く、コストもSparkの数分の1であると結論づけています。¹⁰

これらの分散型トレーニングフレームワークは現実世界ですでに成果を上げ始めています。Walmartは、RAPIDSをDaskおよびXGBoost (MLアルゴリズムの一種) を組み合わせてデータアナリティクスおよびMLに導入しています。NVIDIAのレポートによれば、Walmartは「GPUサーバー1台で、これまで20ノードCPUサーバーでかかっていた処理時間のわずか4%で同じ予測モデルを実行することができるようになった」と報告しています。¹¹ これによりWalmartでは、従来複数のCPUを用いて数週間を要していたモデルの実行が今では4時間で完了しています。

いずれの企業もトレーニングフレームワークについては戦略的に検討に取り組んでいる一方で、過去にはさまざまな壁に直面することもありました。現在では、新しい技術の登場によりはるかに多くのことが可能になり、Python上で直接すべてを稼働させることであらゆるタスクをより早く実行できることが実証されています。Sparkへのモデルの変換の必要性排除することで、企業は複雑性の緩和と効率性の向上を実現しています。Snowflakeプラットフォームでは、複数の異なる分散型トレーニングフレームワークをテストした上で御社に最適なソリューションを選ぶことができます。

トレンド#6: 続々とリリースされる新製品がMLライブラリ、ツール、フレームワークの新たな選択肢を提供

データサイエンスの分野は急速に進化を続けています。毎月のように新開発のMLやAIソリューションがリリースされるだけでなく、新たなスタートアップ企業やツール、ソリューションが続々と登場しています。データサイエンス業界がこのように急速なイノベーションを遂げている現在、1つのツールのみには縛られてはいけません。



そのためには、特定のベンダーやフレームワーク、アルゴリズムに紐付けられない汎用性のあるプラットフォームを選定することが重要です。いまフューチャーブルーなプラットフォームを選定しておくことで、今後登場するであろう新たなMLツールをシームレスにプラットフォームに統合させることができるでしょう。次世代のツールを導入したいがためにプラットフォームをまるごと更新することは何としても避けたいものです。

Snowflakeプラットフォームの他にはない独自の強みは、その最新のアーキテクチャにあります。個別のしかし論理的には連動したコンピュートとストレージを提供することで、Snowflakeでは他のシステムでは別々のレイヤーを連動させるために必要となっている手動のクラスタ構築作業が不要となっており、これによりほぼ無限のスケーラビリティ、リアルタイムな伸縮性、および極めて高い同時実行性によりデータクラウドを支える**マルチクラスタ型共有アーキテクチャ**を構成しています。

この基本となるアーキテクチャに加え、Snowflakeは多様なアプローチでデータサイエンスをサポートします。

- Snowflakeの**外部関数**はあらゆるタイプのサードパーティ型、ホスティング型、およびカスタムMLサービスをサポートしており、SQLからの簡単なアクセスを可能にします。
- チームによってはSQL以外の言語を好むケースも多いことを考慮して、**Snowpark** (パブリックプレビュー提供中) は言語サポートをJava、Scala、Python (プレビュー提供中) にも拡大しています。Snowparkは、データサイエンティストが、データフレームなど自分の使い慣れたプログラミング方法で希望の言語を使ってコードを記述し、さらにデータ準備とワークロードをSnowflake上で直接実行することを可能にします。

- **Javaユーザー定義関数(UDF)** (パブリックプレビュー提供中) のサポートにより、Snowflake内でのトレーニング済みモデルの実行が可能です。つまり、MLパートナーの技術を利用して構築・トレーニングしたモデルをSnowflakeの伸縮性のある高性能エンジン内に持ち込んで直接実行することができるのです。
- 非構造化データの増加に伴い、非構造化データを管理・処理するさまざまな手法が並行して開発されています。例えば一連の新しいラベリングサービスは、画像などの非構造化データの手動でのタグ付けを可能にします。**Snowflakeセキュアデータシェアリング**を利用すれば、非構造化データにタグ付けしたうえでデータを移動させることなくプロバイダーと共有できます (非構造化データのサポートは現在パブリックプレビュー提供中です)。さらに、Snowflake上で非構造化データ分析ツールを用いることにより、Hugging Faceなどの提供するNLPサービスやAWS RekognitionなどのAIサービスを利用した非構造化データの強化も可能です。

2022年、あなたの機械学習が加速する

データサイエンスはいま驚くべきスピードで世界の主流となりつつあります。過去10年の間に、企業の関心はレポートや履歴分析から高度な数学モデルやMLを利用したデータサイエンスの応用へとシフトしてきました。クラウドがすべてを一変させました。増え続けるデータを低コストに収集・格納する能力が確立されたいま、MLを利用したデータモデル構築のニーズが高まっています。

データを素早くスケーラブルに、かつ適切なセキュリティとガバナンスの下で分析し共有したいなら、いま最新のデータプラットフォームの実装は不可欠です。Snowflakeは、データ統合と効率的なデータ準備を実現するアーキテクチャと広範なパートナーエコシステムの利用を可能にするソリューションです。データをモビライズして、データサイエンスとML分野で動き始めた新たなトレンドのメリットをいち早く享受しましょう。

Snowflakeは、データサイエンスにおけるすべての制約を解消するソリューションです。あなたも機械学習を加速させてみませんか？





SNOWFLAKEについて

組織はSnowflakeのデータクラウドを利用してサイロ化されたデータを統合し、データを検索して安全に共有しながら、複数のクラウドおよび地域にまたがってさまざまな分析ワークロードを実行しています。数多くの組織(2022年1月31日時点で、2021年のForbes Global 2000のうち488社を含む)が、Snowflakeデータクラウドを全社で幅広いビジネスに活用しています。詳しくは、[snowflake.com](https://www.snowflake.com)をご覧ください。



© 2022 Snowflake Inc. All rights reserved. Snowflake、Snowflakeのロゴ、および本書に記載されているその他すべてのSnowflakeの製品、機能、サービス名は、米国およびその他の国におけるSnowflake Inc.の登録商標または商標です。本書で記載または使用されているその他すべてのブランド名またはロゴは、識別目的でのみ使用されているものであり、それぞれの所有者の商標である可能性があります。Snowflakeとそれらの所有者との間にはスポンサーシップやエンドースメントなどの関係は存在しません。

引用元

- ¹ on.bcg.com/2Kh3grW
- ² bit.ly/3lg5EZp
- ³ amzn.to/3sv5qFx
- ⁴ bit.ly/2XJOa1u

- ⁵ bit.ly/2XJOa1u
- ⁶ bit.ly/3D7u3wh
- ⁷ bit.ly/3lkt1qx
- ⁸ dask.org

- ⁹ rapids.ai
- ¹⁰ bit.ly/3srNbRv
- ¹¹ bit.ly/3FZozW8