データサイエンスのための
クラウドデータプラットフォーム

snowflake

WHITE PAPER

すべてはデータ次第

機械学習（ML）がテクノロジーの主流となってきています。人工知能（AI）とMLの利用に関するTDWIの2019年の調査によると、回答者の92%がML技術を利用していると述べており、ML用のツールを用いて予測モデルを構築しているとの回答も85%存在しました。[1]

データサイエンティストがこのようなMLモデルを構築し、業務を行うには、大量のデータが必要です。AIの時代において、データへのすばやく正確なアクセスは重要な差別化要因となっています。データマネージメント（分析のためのデータの発見、アクセス保護、クレンジング、結合、準備）は、一般的に、モデル構築プロセスのなかでももっとも時間を要する箇所とされています。

効率的なクラウドデータプラットフォームが何より重要

Forbesによると、分析と訓練のためのデータの発見、取り出し、統合、クレンジング、準備に、科学者は最大で業務時間の80%を費やしています。また同調査によると、データサイエンティストは業務時間のほとんどを、データのマイニングやモデリングではなくデータの操作に費やしていることがわかりました。そして、高度なスキルを持つこれらのプロフェッショナルの76%が、データプレパレーションをもっともつまらない作業と考えていることがわかっています。

このホワイトペーパーでは、データサイエンスやMLに関する今日の取り組みを促進するデータ要件を見極める手助けをします。同時に、業界をリードするSnowflakeとパートナーのツールをサポートするクラウドデータプラットフォームにより、それらの要件を満たす方法について解説します。

データサイエンスのための完全無欠なクラウドプラットフォーム

Snowflakeのクラウドデータプラットフォームは、データウェアハウスのパワー、ビッグデータプラットフォームの柔軟性、クラウドの拡張性、およびライブのデータ共有を1つにまとめ、従来のデータプラットフォームソリューションの数分の1のコストで提供します。Snowflakeは、すべてのデータを1カ所でまとめて格納・分析し、社内での利用とデータ共有の両方に対応するための性能、同時実行性、シンプルさをお届けします。クラウドデータプラットフォームは以下の3つの基本ニーズを満たすことから、現在、数千社がこれを基盤として利用しています。

- 単一の統合ソースによりすべてのデータに対応：Snowflakeは、データサイエンティストが1つの一貫したソースで提供される構造化データと半構造化データにアクセスし、組織の持つより多くのデータ資産を容易に発見、統合、クレンジング、利用できるよう支援します。データサイエンスからのアウトプットはSnowflakeにシームレスに再統合され、ビジネスユーザーがそれらにアクセスすることが可能です。
- 効率的かつ高速なデータプレパレーション：Snowflakeは、他のユーザーや部門に影響を与えることなく、SQLを使ってデータの取り込み、変換、クエリー実行、効率性に優れた専用の仮想ウェアハウスを提供します。多くの場合、SnowflakeのSQLは、Sparkなど他のツールと比べてデータプレパレーションを10倍効率的に行えることから、MLタスク間の遅延が緩和されます。
- 充実したパートナーエコシステム：Snowflakeは、データサイエンスに関するあらゆる確立した技術および新興技術との接続性を備えています。これによって、お客様はニーズに合わせて最適なデータサイエンスツールを選ぶことができ、すべてのツールが1つの一貫した統合データプラットフォームにアクセスすることになります。Snowflakeは、AWS S3をはじめとするストレージにシームレスにデータをエクスポートし、あらゆるデータサイエンスツールによるユニバーサルアクセスを可能にします。

SNOWFLAKEクラウドプラットフォームのワークロード

図1：SNOWFLAKEは、データサイエンス以外にも多数のユースケースとワークロードに対応することから、1つのプラットフォームをデータ、アナリティクス、予測分析に活用することが可能です。

[1] "Best Practices Report: Driving Digital Transformation Using AI and Machine Learning" (tdwi.org/bpreports).
[2] Forbes "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says" (bit.ly/38EbXmN).
[3] Forbes "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says" (bit.ly/38EbXmN).

データサイエンスの重要な概念と役割

データサイエンティストは機械学習技術を利用して、データのパターン、関係、相関性、結果、推論を特定します。このようなデータドリブンな発見は、不正検出、保守周期の予測、カスタマーチャーン（顧客離れ）の緩和、販売予測など、多数の先進的なタスクを自動化するモデルに組み込まれます。このプロセスにおける重要な任務・役割としては、以下のようなものがあります。

- データサイエンティストはデータを用いてモデルを構築し、それらの訓練を行います。その際、JupyterやZeppelinといったノートブックと、R、Python、Java、Scalaなどの言語が使われます。
- データアナリスト / 市民データサイエンティストはこれらのモデルを利用して、予測分析と規範的分析を実行します。そして機械学習の実用的理解にもとづき、ビジネス上の意思決定を行います。
- データエンジニアはデータを準備し、MLモデルに継続的にデータをフィードする自動データパイプラインを構築します。

データサイエンスにおける機械学習の役割

データプレパレーション、データディスカバリー、アナリティクス、データモデリングからなる幅広いデータサイエンスの領域のうち、機械学習は主にデータモデリングに対応します。今日のMLツールとデータサイエンスツールは、データ解析、予測モデルと規範的モデルの生成、モデルのプロダクション導入、およびそれらのモデルの経時的な保守の多くの側面に対処できるようになっています。予測分析と規範的分析のアプリは、ウェブブラウジングのパターンを追跡してビジターに製品やサービスをすすめるなど、しばしば人間の介在なしに、自ら意思決定を行うことができます。

データサイエンティストは、アナリティクスツールを利用して仮説を立てたうえで、プログラミング言語とMLライブラリーを使って予測を行います。MLには、線形回帰、ロジスティック回帰、分類、決定木、ディープラーニングなど、いくつもの種類があります。人気のMLライブラリーにはXGBoost、TensorFlow、scikit-learn、PyTorchなどがあります。

データサイエンティストが信頼できる予測を行うモデルの作成と訓練を担当するのに対し、データエンジニアは推測に必要なデータをMLモデルにフィードするデータパイプラインを担当します。このようなML/AIプロセスの結果は、ビジネスユーザーによるデータドリブンな意思決定に利用されます。

機械学習のプロセス

機械学習のイニシアティブの成功は、正しいデータを、正しいタイミングで、正しいモデルに投入できるかどうかにかかっています。これはかならずしも容易ではありません。なぜなら、ほとんどの場合、機械学習のサイクルは、発見、開発から、プロダクションにいたるまで、複数のステージで構成されるからです。MLサイクルの各ステージでデータの追加と準備が複数回行われ、データ要件が異なるといったことも少なくありません。MLでの成功は、正しいデータを、正しい条件で、正しいアナリティクスプラットフォームに投入し、ビジネスで成果をあげることが前提となります。

図2にあるように、データサイエンティストはまず、データを発見、収集、理解、準備するところから始めます（ステップ1 ～ 3）。その際、データをしっかりと理解し、仮説を立てるためにビジネスインテリジェンスツールを利用することが可能です。データサイエンティストは、この反復プロセス全体において、多くのデータセットを試します。データセットの範囲を拡大・拡張するたびに、データサイエンティストはデータエンジニアがデータをロードし、準備を終えるのを待たなければなりません。これが遅れの原因となり、次の反復プロセスまでの間に著しい遅延が生じてしまいます。また、データを正規化された形に「形成」しなければならず、多くのアルゴリズムで微妙に異なるフォーマットが必要になります。

次に、前のステップで準備したモデルを使ってデータの訓練を行い（4）、結果を評価して各モデルの有効性を特定したうえで、特徴エンジニアリング（3）とハイパーパラメーターの調整（4）のサイクルを通して、モデルのさらなる調整を行います。

このようにして訓練されたモデルは、プロダクションに導入され（5）、予測ツールと規範的ツールによってビジネスユーザーをエンパワーします。プロダクション導入されたモデルは、モデルドリフトを特定し、モデルが古くなっていないかどうかを見極めるための評価を継続して受けることになります（6）。新しい訓練データを使って、モデルの再訓練を定期的に行わなければなりません。このようなモデルの更新は、MLサイクルの反復作業がさらにもう1つ増えることを意味しています。つまり、より多くのデータの処理が必要であり、時間がかかり、エラーが生じやすくなります。モデルの再訓練を行う頻度は、数時間おき、数日おき、あるいは数週間おきといったように、ユースケースによって変わってきます。

MLのワークフロー

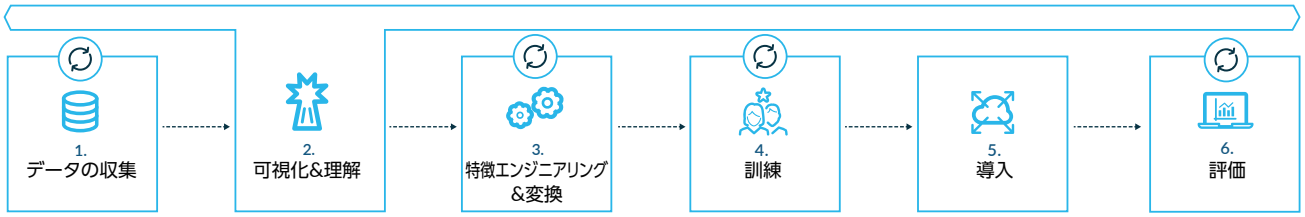

図 2：収集・準備から、訓練、予測、製品化までの ML プロセスをデータにより促進

クラウドデータプラットフォームの役割

機械学習はデータを多用するアクティビティです。各予測モデルの成功は、さまざまな方法で収集、保存、変換、表示される大量の異種データにかかっています。そこには、多くの側面や細部によって特徴づけられ、さまざまな文脈から生じる大量のデータが含まれます。たとえば、カスタマーチャーンを予測する機械学習モデルを構築した場合、販売、サービス、購入、アプリでのやり取りといった顧客行動に関する履歴データとリアルタイムデータの両方を手にすることが考えられます。

Snowflakeのクラウドデータプラットフォームを使えば、データウェアハウス、データマート、データレイクのデータを信頼できる単一の情報源に統合し、これをもとにアナリティクスとデータサイエンスの複数種のアプリケーションを運用することができます。わざわざデータを複製したり、別の場所へ移動したりしなくてもチームメンバー同士で共同作業を行えることから、多様なチームが統制されたデータを社内外で容易に共有できます。JSON、AVRO、XML、ORC、Parquetをネイティブでサポートしており、データサイエンスのワークフローで、ローデータ、構造化データ、半構造化データを容易に見つけて、すぐさまアクセスすることが可能です。

1組のツールを使って構造化データと半構造化データの両方を管理できることから、データの発見と準備のサイクルが短縮されます。さらには、MLアルゴリズムからの出力データはリポジトリーに戻され、ソースデータと共に、ビジネスユーザーがアクセスすることが可能です。つまり、ビジネスユーザー、アナリスト、データサイエンティストのために、すべてのデータが常に最新の状態に保たれ、一貫して保持されることになります。

クラウドデータプラットフォームのメリット

TDWI（The Data Warehousing Institute）では、クラウド用に構築され、機械学習、人工知能、予測アプリケーションの開発からなるデータライフサイクル全体をカバーするモダンデータプラットフォームの取得を推奨しています。このようなプラットフォームで求められるものとは何でしょう？データプレパレーションでは、大量のデータセットをインタラクティブなレスポンスタイムで処理できることが必要です。訓練では、これらのデータセットをインタラクティブに耕す作業が必要です。プロダクションでは、反復と拡張が可能で、信頼できるデータパイプラインが求められます。

データサイエンスの取り組みにおいてSnowflakeクラウドデータプラットフォームの利用が推奨される主な理由を、以下にいくつか紹介します。

- **シンプルさ：**複数のコンピュートプラットフォームを管理でき、統合を継続して行う必要がありません。
- **セキュリティ：**ユーザークレデンシャルを注意深く管理し、すべての伝送を暗号化することで、Snowflake環境で1つのデータコピーを安全に格納します。
- **性能：**クエリー結果がキャッシュされ、MLプロセスとアナリティクスで繰り返しそれらを使用できます。
- **ワークロードの分離：**ユーザーとワークロードがそれぞれ専用のコンピュートリソースを手にすることができます。
- **拡張性：**大量のデータの処理に対応するために、容量をわずか数秒でスケールアップし、完了と同時に容易にリリースできます。秒単位の課金となっていることから、コストが最小限に抑えられます。
- **構造化データと半構造化データに対応：**統合リポジトリー内のあらゆる種類のデータを容易にロード、統合、分析できます。
- **同時実行性：**共有データに対し、同時実行ワークロードを大規模に実行できます。

機械学習とアナリティクスのための データを統合

機械学習アプリケーションのためにデータをプロビジョニングする方法は多数あり、柔軟性が不可欠です。たとえば、一部の組織では、データレイクをデータウェアハウスと組み合わせて使用しています。これによって、大量のローデータをそのままの形態で格納しておき、必要に応じて、さまざまなアナリティクス用途に使用することが可能です。ほとんどのデータサイエンスツールで、データレイクがデータソースとして使用されています。しかし、今日の分析戦略では、ビッグデータプラットフォーム、クラウド、データレイク、データウェアハウスが混在したマルチプラットフォームデータアーキテクチャを用いるケースが増えています。多くの有力組織が、データレイクを完全に取り除き、代わりにすべてのデータをクラウドデータプラットフォームに統合するというやり方を選択しています。このアプローチによって、データレイクを個別に管理することによる複雑さが解消されるだけでなく、データレイクとデータウェアハウスの間のデータ変換パイプラインも不要になります。多目的のクラウドデータプラットフォームをベースとした統合リポジトリーを持つことで、データセットとワークロードごとに適切なストレージ、処理方法、エコノミクスを選択でき、MLとアナリティクスの選択肢が最適化されます。

データを収集し、準備を終えた時点で、アナリティクスツールと予測分析ツールを使って、パターンとインサイトを発見できることが必要です。Snowflakeは一般分析と予測分析の統合を可能にします。そのため、ビジネスインテリジェンスツールとデータサイエンスツールが、同一の統制されたデータについての1つの一貫したビューを持つことになります。すべてのデータサイエンスツールが同じデータ定義を参照することから、クエリー、予測、ダッシュボード、レポートの内容を一貫して再現することが可能です。ローデータとML結果の両方がデータプラットフォーム内に置かれ、容易にアクセスできます。このような統合型のアプローチにより、データサイエンティストは汎用分析のために機械学習アクティビティの結果をデータプラットフォームに戻すことができるだけでなく、これらの結果を意思決定プロセスに組み込むことも可能です。

共通のセマンティクス、データ定義、データモデルがあることで、全員が共通の理解を持つことができます。たとえば、セールスマネージャーがセールスチームのこれまでの業績をまとめたBIレポートを見る場合を考えてみましょう。MLモデルによって、対象アカウントの購買傾向にもとづき、この先の四半期の販売見込みを予測し、計上済みの売上と予測売上の両方を同じレポートによって明らかにするといったことが可能です。

データエンジニアリング、データ統合、データシェイプの自動化

MLで成功するということは、正確かつタイムリーなデータをビジネスユーザーにフィードし、使用するアプリとサービスを追加する、効率性と信頼性に優れたデータパイプラインを構築することを意味します。

データの取り込み

Snowflakeには、データを非同期的にロードし、ただちに利用できるようにするサーバーレスのインジェストサービス「Snowpipe」が含まれます。手動によるデータフラット化タスクが完全に自動化されており、各ターゲットテーブルで必要とされる種類と形態へとデータが変換されます。

標準のコネクターとアダプターを使って、Kafkaをはじめとするメッセージングシステムからイベントストリームを容易に取り込めるいっぽうで、Snowflakeのストリームとタスクにより、SQLジョブのデータロードのスケジュールを簡単に設定できます。

同パイプラインは、自動データインジェストサービスによりMLモデルを「製品化」することで、データ取り込みの複雑なタスクを簡素化します。データサイエンティストはオンデマンドでデータを発見し、準備することができます。次のテストまで数時間から数日待たされるといったことはありません。自動化されたデータパイプラインサービスがプロダクション導入された時点で、データレイクからのETL処理を行わなくてもローデータをただちに利用することができます。データが取り込まれると、自動でモデルを通過し、予測が行われます。また、完全なクラウドベースであることから、データサイエンティストは他のユーザーに影響を与えることなく、専用の仮想ウェアハウスコンピュートリソースを使用できます。

ユニバーサルSQL機能

Snowflakeのお客様は、ETLとELTの堅牢で効率的なワークロードを可能にするユニバーサルSQL機能により、単一の信頼できる中央情報源を利用することができます。Sparkを使用する場合と比べて、SQLを用いたデータ変換がすばやく簡単に行え、コストも少なくてすみます。SQLクエリーの一部としてデータを変換できることから、変換が分析の一部になります。Snowflakeのクラウドアーキテクチャと圧縮機能により、わずかなコストで大量のストリーミングデータをすばやく取り込み、それらを無期限に保管することが可能です。データエンジニアは、Alteryx、Alooma、Matillion、Fivetran、Alation、Informaticaなど、さまざまな統合ツールを使用することができます。

専用のコンピュートリソース

Snowflakeでは、MLのデータ取り込み、データ管理、データプレパレーションのワークロードに対し、ML以外のデータエンジニアリングやアナリティクスのワークロードと競合することのない専用のリソースが提供されます。リソースの競合をなくすことで、ストリーム中にライブデータを取り込んで変換し、すぐさまアナリティクスで使用することができます。ワークロードごとにデータウェアハウスのサイズのカスタマイズを行い、必要に応じてスケールアップし、終了と同時にクラウドサービスをオフにすることが可能です。線形的な拡張性により、予測可能な時間枠でクエリーを実行するのに必要な分だけリソースを要求できます。瞬間的な拡張性と秒単位の課金により、ユーザーとワークグループはそれぞれが使った分だけコンピュートリソースの料金を支払う仕組みになっています。最終的にこのアーキテクチャによって、一貫したデータを提供しつつ、チームのパフォーマンスと効率性を最大限まで高めることが可能です。

堅牢なデータセキュリティ

従来型のデータサイエンスプロジェクトの多くが、分散フレームワークでデータを保管・処理するためのオープンソースフレームワークであるApache Hadoopに依存しています。しかし、Hadoopアーキテクチャは基本的なアクセス制御しか採用しておらず、HIPAA、PCI、GDPRといったデータのセキュリティとプライバシーを統制する重要な業界標準に対応する設計にはなっていません。Amazon S3など、汎用のオブジェクトストアを利用しているデータサイエンスプロジェクトもありますが、こちらもやはり堅牢なデータセキュリティが不足しています。これに対して、Snowflakeクラウドデータプラットフォームは、暗号化、アクセス制御、ネットワークモニタリング、および物理的なセキュリティ対策とあわせて、モニタリング、アラート、サイバーセキュリティの包括的なプラクティスで構成される多層のセキュリティ基盤をベースに構築されています。Snowflakeは、ISO/IEC 27001やSOC 1/SOC 2 Type 2などの業界標準の技術認証を取得しているだけでなく、PCI、HIPAA/Health Information Trust Alliance（HITRUST）、FedRAMP認定といった重要な政府・業界規制にも対応しています。Snowflakeのお客様は、あらゆる種類のデータサイエンスアクティビティで、安全にデータにアクセスしていただけます。

効率的なデータ共有

Snowflakeのプラットフォームは、パブリックおよびプライベートのデータエクスチェンジにより、パートナー、サプライヤー、ベンダー、顧客の間で行われるデータのやり取りを簡素化します。これによって、モデルの有効性を高め、特徴エンジニアリングのさらなる可能性を提供するユニークなデータセットへのアクセスが可能になります。Snowflakeのセキュアデータシェアリングでは、FTPを使ったデータ転送や、アプリケーションを接続するためのAPIの設定が必要ありません。ETLの統合を簡素化し、データプロバイダーとデータコンシューマーの間で「ライブ」データを自動で同期します。ソースデータを複製するのではなく、共有するようになっているため、余分なクラウドストレージが必要ありません。このようなパブリックおよびプライベートのデータエクスチェンジにより、データサイエンティストはローデータや処理済みのデータを共有することで、モデルに対して容易に共同作業を行えます。

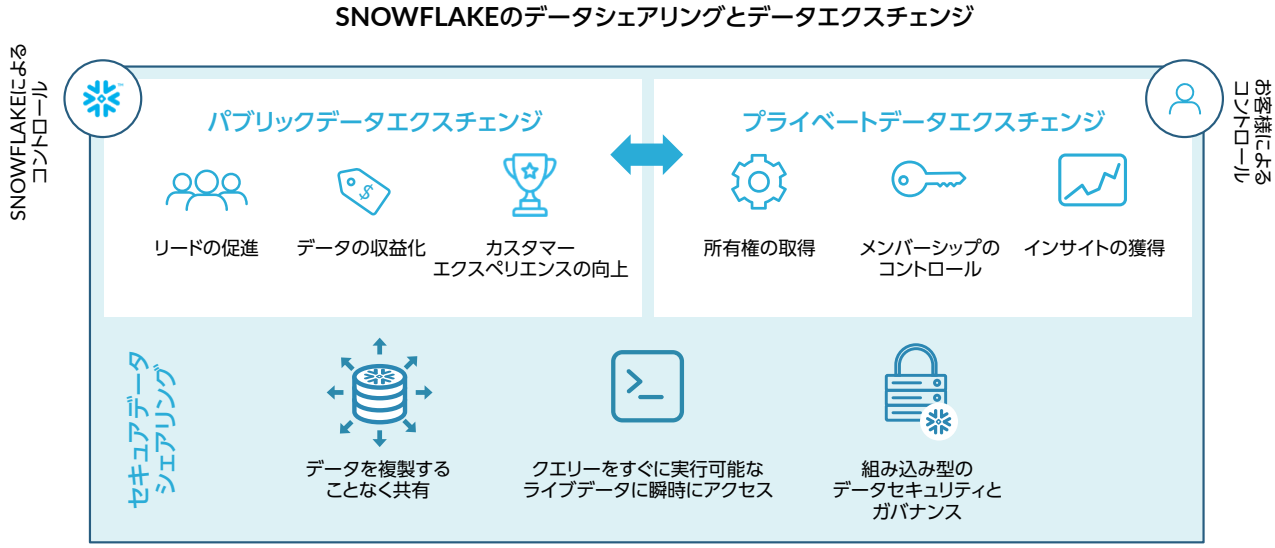

図 3：SNOWFLAKE のデータシェアリングでは、パブリックデータエクスチェンジを介して外部とデータを共有し、顧客、サプライヤー、その他のビジネスパートナーと独自のプライベートデータエクスチェンジを構築することが可能です。

充実したパートナーエコシステム

機械学習分野は目まぐるしく進化しており、毎年新しいツールが追加されています。Snowflakeの充実したパートナーエコシステムを通して、新旧のあらゆるデータサイエンスツール、プラットフォーム、およびPython、R、Java、Scalaなどの言語、PyTorch、XGBoost、TensorFlow、scikit-learnなどのオープンソースライブラリー、JupyterやZeppelinなどのノートブック、DataRobot、Dataiku、H20.ai、Amazon Sagemaker、その他のプラットフォームに直接接続することが可能です。Snowflakeは単一の一貫したデータリポジトリーを提供することから、ツール、言語、ライブラリーを切り替えるたびに基礎データを入れ替える必要がありません。さらには、これらのアクティビティからのアウトプットはSnowflakeに難なくフィードされ、一般のユーザーがアクセスしてビジネスバリューを生み出せるようになっています。

ノートブックベースのMLツール

JupyterやZeppelinといった従来のMLノートブックは、Amazon Sagemaker、Dataiku、Zeplをはじめとする今日の主要なデータサイエンスツールの動力源となっています。このアプローチは、データサイエンティストが選択したフレームワークとアルゴリズムの根本的なコントロール、徹底的な特徴エンジニアリングの実施、ハイパーパラメーターの調整、およびMLモデルの反復的な作成、評価、製品化を可能にします。データサイエンティストは、アルゴリズムを試し、それらの性能を評価して、新しいモデルを選択・改良するプロセスを反復して行うことで、直感を正確な予測へと変えることができます。Amazon SageMakerのユーザーは、Snowflake Python Connectorを利用して、Pandasデータフレームにデータを直接取り込むことが可能です。このような高速接続により、訓練に要する時間が短縮されるだけでなく、ANSI SQLをフルに活用するデータプレパレーションと特徴エンジニアリングのサイクルが最適化されます。

AutoMLツール

もう1つの方法として、RapidMiner、BigSquid、H2o.ai、DataRobotなどのAutoML（自動化された機械学習）ツールを使うことで、アルゴリズムの選択、モデルの訓練、最適なモデルの選定を自動で行えます。これらのツールは高度なアナリティクスへのアクセスを民主化する効果的な方法であり、データアナリストはプログラミングの高度なスキルや数学/統計学の深い知識がなくても、ML機能を実行することができます。これらの2つのアプローチを橋渡しするいくつかのツールにより、データサイエンティストはAutoMLのプロセスをカスタマイズすることが可能です。AutoML分野のリーディングプレイヤーであるDataRobotのプラットフォームには、Snowflakeが最初から統合されており、ユーザーがDataRobotアカウントをSnowflakeにすばやく接続して、データストアとして使用できるようになっています。

アナリティクスパートナーとクラウドパートナー

いずれのMLアプローチを選んだとしても、SnowflakeではTableau、Looker、ThoughtSpot、Sigmaといった他のエコシステムパートナーとの接続を利用することで、ダッシュボード、レポート、ビジネスアナリティクスツールを通して結果を取り込むことができます。さらには、Amazon、Microsoft、Googleの人気のサービスを含め、あらゆる地域のあらゆるクラウドで、データを格納・複製することが可能です。Snowflakeは、Amazon S3、Azure Blob、Google Cloud Storageで保持される外部テーブルにシームレスにデータをエクスポートできることから、あらゆるツールによるユニバーサルアクセスが実現します。たとえば、Snowflakeを利用してAWS上のデータレイクを補完したうえで、Amazon SageMakerと接続して、MLモデルの開発、テスト、導入を大規模に実行するといったことが可能です。同プラットフォームは、データのストレージや処理から、トランザクション管理、セキュリティ、ガバナンス、メタデータ管理まで、あらゆるものを自動化します。

数分で使用を開始

SnowflakeとMLを今すぐ使い始める方法を知りたいという方は、厳選したテクノロジーパートナーとの統合を前もって設定することで導入を簡素化するPartner Connectプログラムの利用をぜひご検討ください。パートナーアプリケーションのプロビジョニングと設定を数分の間に自動で行い、Snowflakeにデータをロードして、ただちに使用することができます。

代表的な事例

ConsumerTrackは、デジタル広告とデジタル出版を手掛ける企業であり、数百のプロバイダーからウェブサイトのパフォーマンスデータを集め、CNNやMSNといったポータルに同時配給しています。同社のML環境では、MySQLとさまざまなオーケストレーションツールを使用することでデータが渋滞を起し、遅延の問題が生じており、データサイエンスチームは頭を抱えていました。

ConsumerTrackはSnowflakeを使って既存のデータレイクを増強し、MLワークフローを自動化するフルマネージド型サービスとしてAmazon SageMakerを選択しました。同サービスはデータの分類と準備を行い、アルゴリズムを選択し、モデルに訓練を施し、そのモデルを導入するための調整と最適化を行い、予測を立てたうえで行動を起こします。

今では、WS LambdaとAWS Glueを使った自動パイプラインを介してデータがデータレイクに送られるようになっています。データはキュレートされたうえで、Snowflakeにロードされ、カスタムアラートによりデータストリームの設定が行われます。Amazon SageMakerは、MLモデルの開発、テスト、構築を簡素化するために、Snowflakeに接続します。

ConsumerTrackは、データの渋滞ポイントをなくし、インサイトの取得にかかる時間を数時間から数分に短縮することに成功しました。Snowflakeは、データの発見と準備にかかる時間を大幅に短縮します。Snowflakeの幅広いエコシステムにより、ConsumerTrackはPython用のネイティブコネクターをはじめ、データサイエンス関連のさまざまなプラットフォームとツールに接続することができています。必要に応じて、同社のデータサイエンスチームはあらゆるブロブストアにデータをエクスポートし、ユニバーサルアクセスに対応することも可能です。

次のステップ

機械学習用のSnowflakeクラウドデータプラットフォームについて、Snowflakeのデータサイエンスページとデータプラットフォームページで詳しく解説しています。

今回取り上げたデータサイエンスパートナー

DataRobot　Amazon SageMaker　dataiku

zepl　H2O.ai　bigsquid

SNOWFLAKEについて

Snowflakeが提供するデータクラウドは、数千もの企業がデータを集積するグローバルなネットワークであり、無限に近い拡張性、同時実行性、性能を実現しています。顧客企業はデータクラウド内で、サイロ化したデータを統合し、管理されたデータを簡単に検索して安全に共有しながら、さまざまな分析ワークロード を実行しています。Snowflakeはデータやユーザーの場所を問わず、さまざまなパブリッククラウドの壁を越えて単一でシームレスなエクスペリエンスを提供します。Snowflakeのプラットフォームはデータクラウドへのアクセスを推進し、利用可能にするエンジンであり、データウェアハウジング、データレイク、データエンジニアリング、データサイエンス、データアプリケーション開発、データ共有におけるソリューションをもたらしています。Snowflake のデータクラウドは、顧客、パートナー、データプロバイダーのビジネスを新たなフロンティアへと導きます。ぜひ、データクラウドにご参加ください。**Snowflake.com**

snowflake®

© 2020 Snowflake. All rights reserved.