alteryx

Google Cloud Platform でのセルフサービス分析：

成功を確実にするデータ準備の 5 つの教訓

ビジネス分析の課題に取り組み、生産性と革新性を高めるうえで、データは不可欠な存在です。

そして、今日の動的で複雑なデータ環境において、従来のデータウェアハウスやビジネスインテリジェンスソリューションによるアプローチはもはや有効ではありません。Tableau や Qlik といったソリューションで従来のデータウェアハウスを活用しようとしても、結局は表計算ソフトでの作業や、IT 部門のサポートに頼らざるを得ません。

こうした問題を解決するのが、セルフサービスの分析ソリューションです。

本書では、Alteryx が企業におけるセルフサービス分析の導入を支援する際に学んだ 5 つの教訓について、アナリストがあらゆる形式のデータを、あらゆる規模やスピードで自在に活用できる包括的なスマート分析スイートである Google Cloud Platform を交えながらご紹介します。

レッスン 1
クラウドへの移行時のリフト & シフトを廃止

多くの企業が、分析をクラウドに移行する際に分析アプリケーションをクラウド上の仮想マシンにインストールして利用する「リフト＆シフト」方式を採用しています。こうしたアプローチでは、インフラのメンテナンスは容易になりますが、データ分析の提供方法の改善、ビジネスメリットの向上、運用コストの削減にはつながりません。また、プロセスやツール自体も以前と変わらないままとなることから、ユーザーがセルフサービスによる付加価値を享受できません。

真のセルフサービスを実現するには、クラウドネイティブな分析ソリューションを利用して、データの取り込み、保存、準備、レポート作成のプロセスを、クラウド用に構築されたシステムとネイティブに統合できるように設計する必要があります。

また、動的で弾力性があり、スケーラブルで、コンテナ化が進むクラウド環境において、今後想定されるマイクロサービスの提供をサポートできるものでなければなりません。今日成功を収めている企業では、オンプレミスの融通の利かないデータプラットフォームを再構築する代わりに、クラウドの俊敏性を活かしながら、Kubernetes や Docker などのオープンソースシステムを使用してコンテナとデータフローをオーケストレーションしています。

Google Cloud Platform は、ネイティブのサーバーレススマート分析スイートを備えており、セルフサービス分析にを簡単に実行できます。各コンポーネントを使用状況に応じて簡単にアクティブ化できるようになっているため、動的にスケーリングしてリソースの割り当てを減らすことができます。また、メンテナンスの作業を事前に計画する必要はありません。企業は、あらゆる分析コンポーネントを自在に活用して、リソースを柔軟に使用し、コストを管理できるため、データと、データがビジネスにもたらす価値に集中することができます。

レッスン 2
ETL とセルフサービスは相容れない

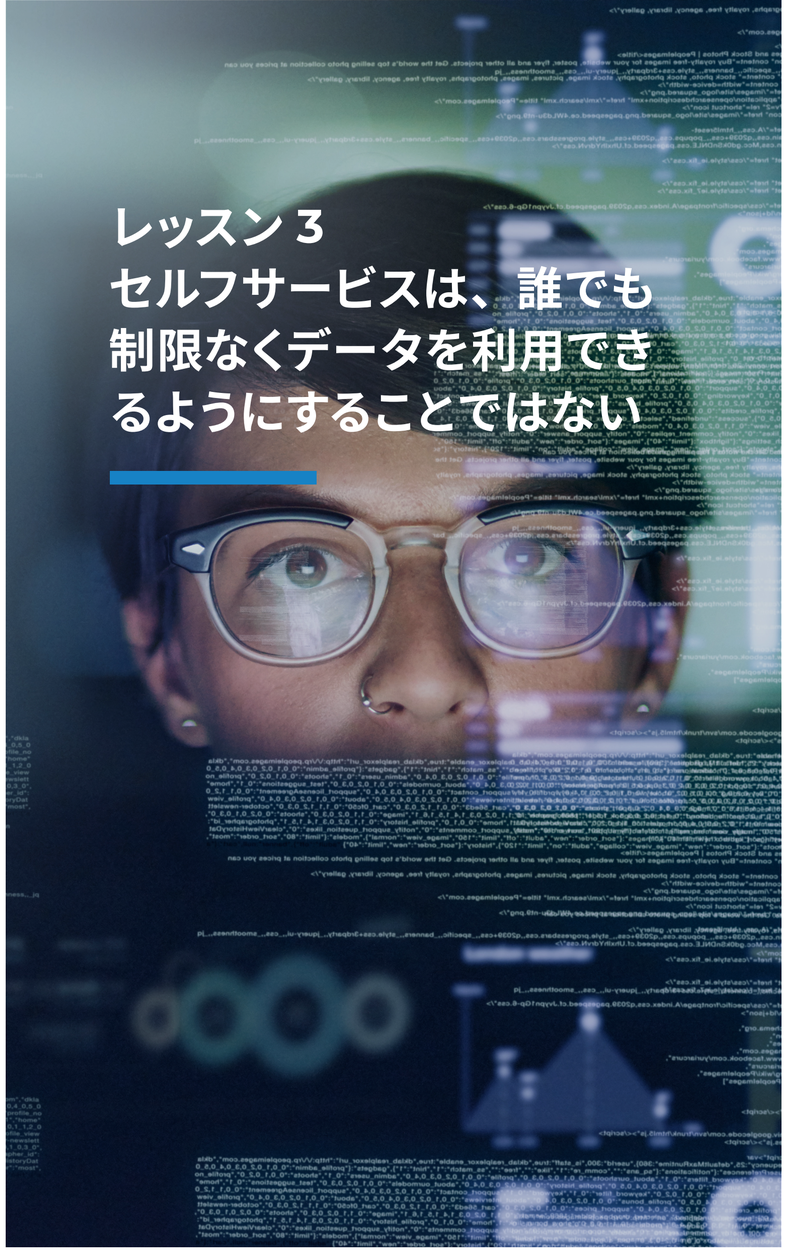
ETL（抽出・変換・書き込み）などの従来のデータ統合テクノロジーを利用して、オンプレミスや他のクラウドに保存されているデータを Google Cloud Platform に移行するためのパイプラインを構築することは可能ですが、変わりに StreamSets、Fivetran、Google Cloud Data Fusion などの新しいクラウド向けデータ統合 / ストリーミングテクノロジーを使用すれば、効率的にシステム間でデータをやり取りし、Google Cloud Platform で分析可能な状態にすることができます。

データレイクやデータウェアハウスに移行した後も ETL を使い続けたいと思うかもしれませんが、ETL はセルフサービス分析には向いていません。

ETL は、大規模なデータ移行に向け、定義されたリアルタイム変換やバッ変換を定期的に行うデータエンジニア向けの、非常に技術的で複雑な技術となっており、ビジネスデータに精通した専門家がデータにアクセスし、分析に利用するためには、特別なデータ準備が必要になります。

Google Cloud Platform では、スマート分析スイートの一部として、データ準備ソリューションである、Cloud Dataprep by Trifacta を提供しています。Cloud Dataprep by Trifacta では、データの品質を評価し、データを改良、標準化、クレンジングし、データ同士を結合してさまざまなデータの計算を処理することができます。

Trifacta の Cloud Dataprep は、データエンジニア、データアナリスト、ビジネスアナリストなど、データを扱うあらゆる専門家がデータレイクやデータウェアハウスを活用して、データの内容を反復的に改良し、下流のビジネス主導の分析で利用できるように設計されています。



レッスン 3
セルフサービスは、誰でも制限なくデータを利用できるようにすることではない

セルフサービス分析を利用できる人が増えれば増えるほど、自分自身で分析をする人が増えますが、セルフサービスとは、誰でも制限なくデータを利用できるようにすることではありません。より多くのユーザーがセルフサービスを利用できるようにするためには、以下のような慎重なガバナンスが不可欠です。

- 制御不能なデータ拡散の防止
- 規制要件の遵守
- ビジネス上の意思決定に使用するデータの信頼性を維持

データアセットを保護することと、ユーザーが適宜連携してデータから価値を引き出せるようにすることの、適切なバランスを保つことが重要です。

組織は、ガバナンスやセキュリティの観点からデータアセットを保護することと、ユーザーがコラボレーションしてデータから価値を引き出せるようにすることの間で、適切なバランスを取る必要があります。こうしたバランスをとるためには、以下の 3 つのアプローチがあります。

サイロの削減

データのストレージと処理を、無限の拡張性を備えたクラウドで一元化し、エンドユーザーがクラウド上で各自のデータを管理できるようにすることで、データサイロ化を回避できます。ユーザーは、データを抽出し、表計算ソフトで複製する代わりに、抽出されたデータを収集して、独自の準備ルーチンを実行し、クラウド内で、またはクラウドからレポートを作成できます。

データカタログを使用

データ定義、メタデータ、データ系列に関する知識を管理する共有リソース（一元管理されたカタログや用語集など）により、ユーザーが素早くデータを検索し、組織がデータソースを管理し、そのライフサイクルを監視することが可能になります。また、機械学習（ML）および人工知能（AI）ソリューションを取り入れることで、データに関するメタデータおよび関連知識の収集と管理を自動化できるようになります。

データ系列の追跡と文書化

データ系列、すなわち、データが誰によってどのように使用され変換されてきたかというとは、規制機関への報告と監査において欠かせない情です。また、分析、視覚化、処方的分析における推奨事項のもととなるデータの履歴や、分析を行うためのデータパイプラインに対する新しい要件の影響などを理解する必要がある意思決定者にとっても重要です。

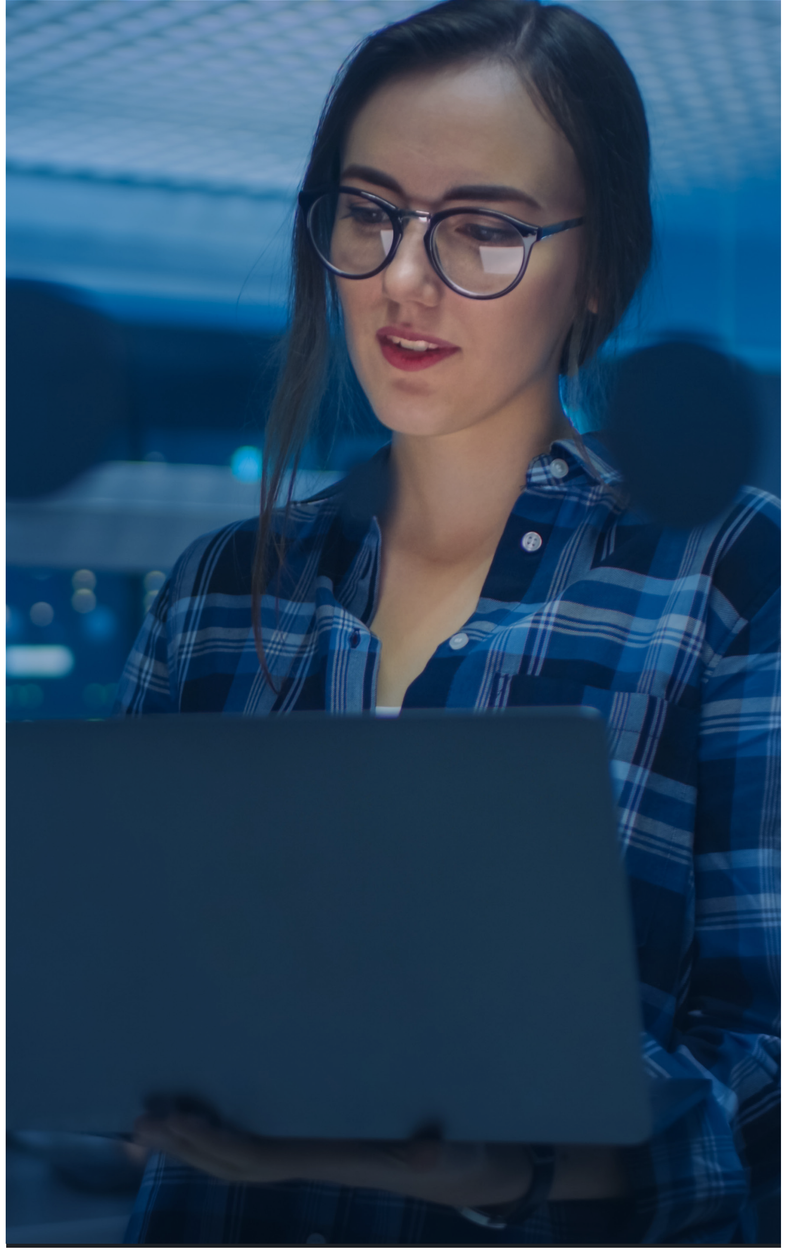
レッスン 4
役割と責任を確認する

セルフサービスによるデータ準備により、企業がデータ分析のバリューチェーンにおける主要ステークホルダーの役割と責任を特定し、厳格なデータパイプラインプロセスをより柔軟で俊敏なフローとして見直すことが可能になります。これにより、各ステークホルダーが、ニーズに合わせてさまざまな方法でセルフサービスでのデータ準備を行い、デザイン、コミュニケーション、コラボレーションという共通のフレームワークを用いて分析を提供することができるようになります。

セルフサービスでのデータ準備は、データエンジニア、データアナリスト、データサイエンティスト、データアーキテクト、そして分析を必要とする経営者や管理者が互いに協力し合うための新たなアプローチを提供します。

データアナリストは通常、ビジネスアナリスト、プロジェクトマネージャー、その他関連する役職に加えて、ビジネスユーザーとも緊密に連絡を取りながら、生データを探索し、ビジネスに関する問題の解決に役立つ可能性があるデータを見つける必要があります。また、できるだけ早く簡単に解決策を見つけることを目指しています。

Cloud Dataprep by Trifacta は、導入が容易で、コーディングの技術的な専門知識を必要としないため、データアナリストにとって理想的なツールとなっており、データ準備プロセスの初期段階から関与し、ビジネスニーズに合わせてデータを探索し、プロトタイプを作成できます。また、自動化によって持続可能性や再現性を実現したい場合にも、データエンジニアと容易に連携し、エンドツーエンドのデータパイプラインをオーケストレーションできます。



データ処理とデータアーキテクチャを設計、構築、管理し、分析とデータサイエンスをサポートする**データエンジニア**は、生データの探索やプロファイリングなど、データの変換プロセスに深く関わっています。データエンジニアの主な目標は、データ関連プロセスを合理化および自動化し、より多くのデータを管理できるようにすることです。

Cloud Dataprep by Trifacta は、データエンジニアやデータアナリストが設計するさまざまなデータフローを運用化し、監視できるため、データエンジニアにとっても理想的なツールです。Cloud Dataprep by Trifacta を使用すれば、データエンジニアがあらゆるステークホルダーと容易に連携して、データインフラストラクチャの要件を理解し、ユーザーがデータを探索、分析、モデル化、活用する方法を改善するためのガイダンスを提供できるようになります。

データサイエンティストは、機械学習や人工知能を活用したアルゴリズムの設計やモデリングにおいて、専門知識やスキルを発揮します。しかし、業務時間の最大 80% を日常的なデータ準備作業に奪われてしまい、イノベーションに費やす時間はほとんど残っていません。

Cloud Dataprep by Trifacta は、クラウドネイティブなデータ準備ソリューションであり、日常的なデータ準備作業を簡素化し、それらの作業を迅速かつ低コストで完了させることができるため、データサイエンティストにとって理想的なツールです。

レッスン 5
データ品質の確保は全員の責任

すべてのステークホルダーは、分析、可視化、アルゴリズムのもととなるデータが信頼できるものであることを把握する必要があります。そして、予測モデルの結果や分析によるインサイトの品質は、そのもととなるデータの良し悪しによって決まります。

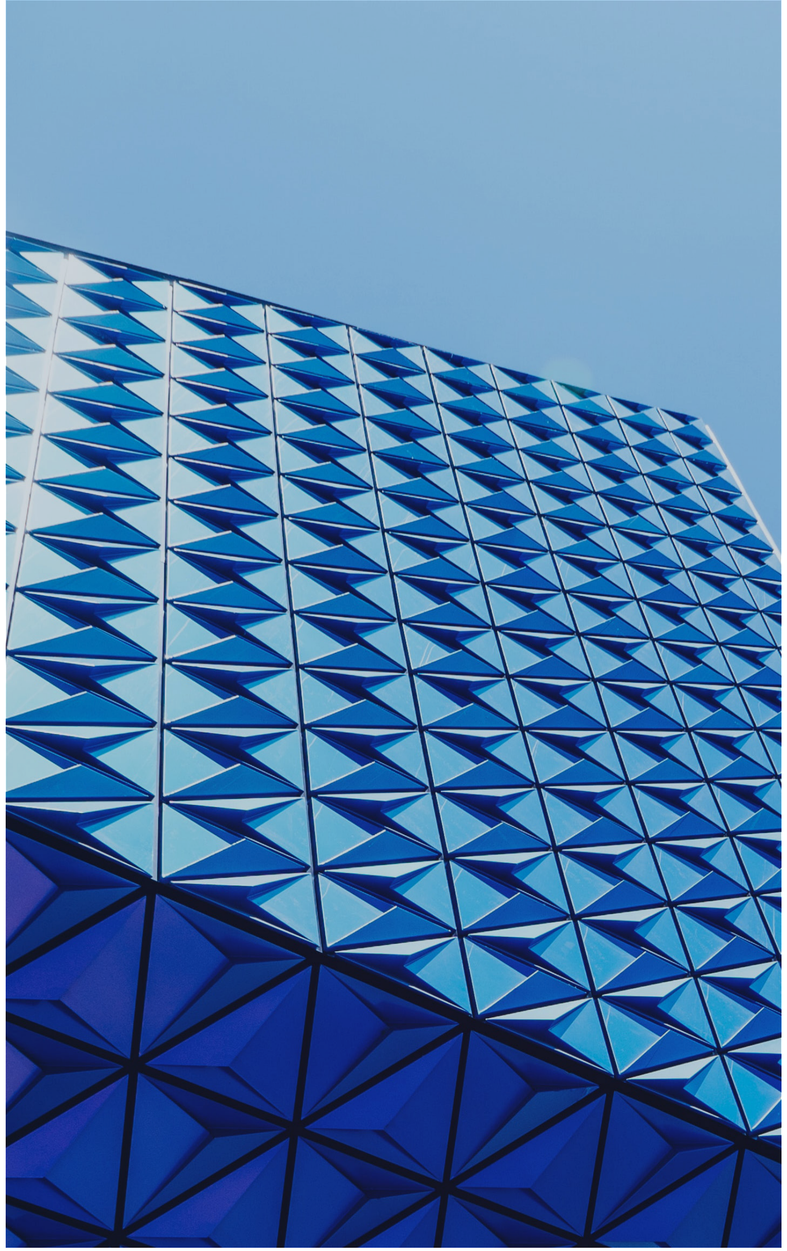
そのため、いかなる組織も、クリーンなデータを利用できるように、優先事項として取り組む必要があります。

全社的でデータ品質の維持に取り組むことができれば理想的ですが、それは現実的な話ではありません。データ量が膨大なうえに、自社やサードパーティからの多種多様なデータソース、データタイプ、コンテキストが混在しているため、データ品質の評価、問題の修正、新しいデータ品質問題の監視が困難になっています。また、クラウドデータレイクや最新のデータウェアハウスなどの新しいデータリポジトリが、データ品質の課題をさらに複雑なものにしています。

こうした課題に対処するためには、データ準備の段階でデータ品質の維持に取り組むことが重要です。Cloud Dataprep by Trifacta は、ML と AI を適用してデータクレンジング処理を自動化することにより、データの正確性、一貫性、完全性を向上させます。自動化によって、大規模なデータリポジトリにも対応し、不正確、無効、欠落、不一致の可能性が高いデータ値を迅速に特定できるようになるります。また、より踏み込んで精査すべき異常値には自動的にフラグが立てられます。

データプロファイリングやクレンジングの自動化により、クラウドデータストレージに統合されたソース間の不整合を検出し、潜在的なデータの重複を特定したり、コードフリーの自動変換によってデータ品質の問題を修正する方法を視覚的に確認したりできるようになります。

また、新たな構造化、非構造化、半構造化データが取り込まれ、クラウドに統合されると、データ品質が継続的に検証されるようになります。こうした継続的な検証により、検証作業の終了を待たずに結果の確認やテストが行えるようになるため、今日のアジャイル開発手法の大敵である遅延を回避できるようになります。



総括

以下に Alteryx が、多くの企業におけるクラウドでのセルフサービス分析の導入支援を通じて得た 5 つの重要な教訓をご紹介します。

1. オンプレミスでの分析からクラウドへと移行する際に、シンプルな「リフト & シフト」のアプローチは有効ではありません。

包括的なネイティブのサーバーレススマート分析スイートを備えた Google Cloud Platform など、クラウド専用に設計された分析コンポーネントを備えたクラウドプラットフォームを選択する必要があります。

2. ETL テクノロジーを用いれば、オンプレミスのデータを円滑にクラウドに移行できますが、ETL は極めて専門的で柔軟性に欠けるため、セルフサービス分析には適していません。Cloud Dataprep by Trifacta などのセルフサービスのデータ準備ソリューションを導入すれば、あらゆるビジネスユーザーや技術ユーザーが、分析に必要なデータを活用できるようになります。

3. セルフサービスは、誰でも制限なくデータを利用できる無法地帯ではありません。セルフサービス分析を成功させるには、慎重なガバナンスとオープン性の適正なバランスを保つ必要があります。

そのためには、データ系列を理解したり、ビジネスデータの用語集を活用したりすることが重要です。

4. データを分析に活用できる高度な資産にするためには、すべてのステークホルダーとのコラボレーションを、誰もが理解できる言葉で進められるアジャイルプロセスの導入が不可欠です。

5. データ品質の確保は全員の責任です。Cloud Dataprep by Trifacta は、自動のデータ品質評価と継続的な検証および解決を実現し、セルフサービス分析に利用するデータの信頼性を大幅に向上させます。

alteryx

ALTERYX について

分析自動化プラットフォームのグローバルリーダーである Alteryx は、分析、データサイエンス、ビジネスプロセス自動化を統合した単一のエンドツーエンドプラットフォームにより、デジタルトランスフォーメーションの加速と分析自動化の未来の実現を支援しています。世界中のあらゆる規模の企業が Alteryx の分析自動化プラットフォームを活用し、インパクトのあるビジネス成果と現代に即した人材の迅速なスキルアップを実現しています。詳しくは、www.alteryx.co.jpをご覧ください。

Alteryx Analytics Cloud

Alteryx Analytics Cloud の一部である Alteryx では、自動化されたデータパイプラインを通じて、迅速かつ効率的にデータへの接続、プロファイル、変換、配信を行うことができ、それらすべてをノーコードまたはローコードで実現できます。データ品質を視覚的に調査し、分析用のデータ準備を加速させ、データパイプラインをわずか数秒で構築・導入することが可能です。

詳細を見る >

Alteryx Machine Learning Platform

自動機械学習（AutoML）と特徴量エンジニアリングにより、全社規模でのデータサイエンスの拡張を実現。ビジネス分野のスペシャリストからデータサイエンティストに至るまでの誰もがインサイトを加速できるようになります。

詳細を見る >

Alteryx Designer

データの準備、ブレンディング、レポーティング、予測分析、データサイエンスなど、分析のすべてのステップを自動化します。どんなデータソース、ファイル、アプリケーション、データ型にもアクセスでき、300 以上の自動化ビルディングブロックを備えた使いやすくパワフルなセルフサービスプラットフォームを介して、誰もが自在に活用可能な分析結果をすばやく得ることができます。

詳細を見る >

Alteryx Intelligence Suite

初心者もエキスパートも、革新的な機械学習モデルをわずか数分で作成できます。Alteryx Intelligence Suite は、データに潜むあらゆるインサイトを解き放ちます。構造化データの分析から、テキストやドキュメントに埋もれたインサイトの発掘まで、Intelligence Suite を使用すれば、ビジネスのどんな難題も解決できるようになります。

詳細を見る >

Alteryx Server

Alteryx Server では、分析ワークフローのプロセス、モデル、データを組織全体で簡単に拡張、共有、管理することができます。Alteryx Designer で作成した分析ワークフローを Alteryx Server に公開することで、レポート作成や結果の出力をスケジュール設定して自動実行できるようになります。データガバナンス、一元管理のセキュリティ、高可用性が兼ね備わった Alteryx Server は、部門や組織全体での分析の拡張をスムーズに促します。

詳細を見る >