Delta Lake シリーズ

導入事例

Delta Lake でイノベーションを加速

databricks

このeBookの概要

Databricks の Delta Lake の eBook シリーズは、データを扱う方が Delta Lake のフル機能を理解して利活用するための支援を目的として提供されています。

この eBook「Delta Lake シリーズ：導入事例」では、実際のお客様が Delta Lake を活用してデータにまつわる難題をどのように解決しているのか、導入事例を通じてご紹介します。

学べる内容

さまざまなビジネスの課題を解決するソリューションの実例を通じて、Delta Lake のケイパビリティの活用法を理解できます。

databricks

目次

序章 Delta Lake とは

01 **事例 1 ： ヘルスダイレクト・オーストラリア**
Databricks の活用により、セキュアでパーソナライズされたオンライン患者ケアを提供

02 **事例 2 ： コムキャスト**
Delta Lake と MLflow で視聴者エクスペリエンスを変革

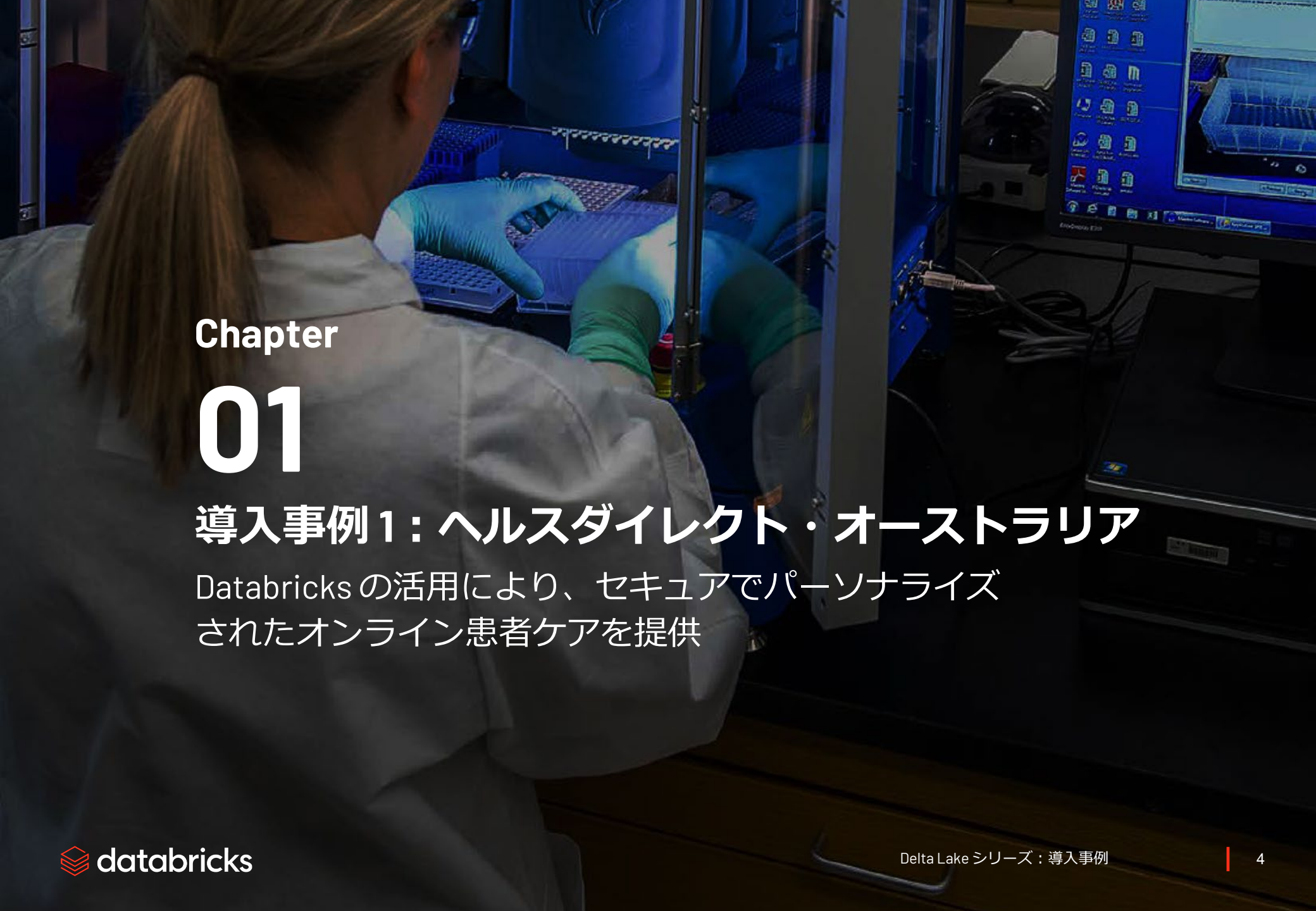
03 **事例 3 ： バイアコム 18**
Hadoop から Databricks への移行により視聴者エンゲージメントを強化

Delta Lake とは

Delta Lake は、データの信頼性を高め、迅速な分析をクラウドのデータレイクにもたらす統合データ管理システムです。既存のデータレイク上で動作し、Apache Spark™ API と完全な互換性があります。

Databricks では、Delta Lake がデータレイクにもたらす信頼性、性能、ライフサイクル管理を実証してきました。Databricks を導入したお客様は、Delta Lake を活用することで、不正なデータ取り込み、コンプライアンスのためのデータ削除の煩雑さ、データ収集時のデータ欠落など、さまざまな課題を解決しています。

Delta Lake は、データレイクへの高品質なデータの取り込みを高速化し、セキュアでスケーラブルなクラウドサービスを通じてデータチームによるコラボレーションを可能にします。



Chapter

01

導入事例1：ヘルスダイレクト・オーストラリア

Databricks の活用により、セキュアでパーソナライズされたオンライン患者ケアを提供

01

導入事例 1:
ヘルスダイレクト・オーストラリア

Databricks の活用により、セキュアでパーソナライズされたオンライン患者ケアを提供

オーストラリアの公衆衛生情報サービス （National Health Services Directory：NHSD）を管轄するヘルスダイレクト・オーストラリアは、テラバイト級のデータを活用して時間主導型活動基準のヘルスケアトランザクションをカバーし、ヘルスケアサービス、情報提供、サポートの改善に注力しています。ヘルスダイレクト・オーストラリアは、厳格なガバナンス要件、チームのサイロ化、従来のシステムのスケーラビリティの欠如といった課題を解決するために Databricks への移行を実施しました。その結果、ダウンストリームの機械学習のためのデータ処理の高速化および、HIPAA 要件を満たすデータセキュリティの確保に成功しています。

事例：ヘルスダイレクト・オーストラリア
業種：医療・ヘルスケア、ライフサイエンス

6 倍

データ処理が 6 倍高速化

2,000 万

データの読み込みを高速化、2,000 万件のレコードを 20 分で取り込み

データ品質、ガバナンス、サイロ化、スケーラビリティの問題を解決

ヘルスダイレクト・オーストラリアは、高まる規制に対応すべく、包括的なデータ品質の向上と、ガバナンスの強化を目指していましたが、データの保存やアクセスに関する問題に直面しました。ダウンストリーム分析用のデータを効率的に準備するうえで、乱立するデータサイロが障壁となり、スタック内のさまざまなシステムにデータソースが分散し、それらのデータが同期されていないことでデータ読み取りの一貫性に影響を及ぼしていました。また、品質の低いデータによるエラー率の上昇や効率性の低下という問題もありました。ばらばらなアーキテクチャが運用上のオーバーヘッドを生み、患者を十分に理解するという重要なミッションの妨げとなっていました。

さらに、予約方法、料金設定、e-ヘルスのトランザクションなどにおける顧客・患者ニーズの変化に伴い、 10 億以上のデータポイントを取り込む必要があり、データ量は 1 TB 以上になることが予測されました。ヘルスダイレクト・オーストラリアのチーフアーキテクトであるピーター・ジェームズ（Peter James）氏は、次のように述べています。

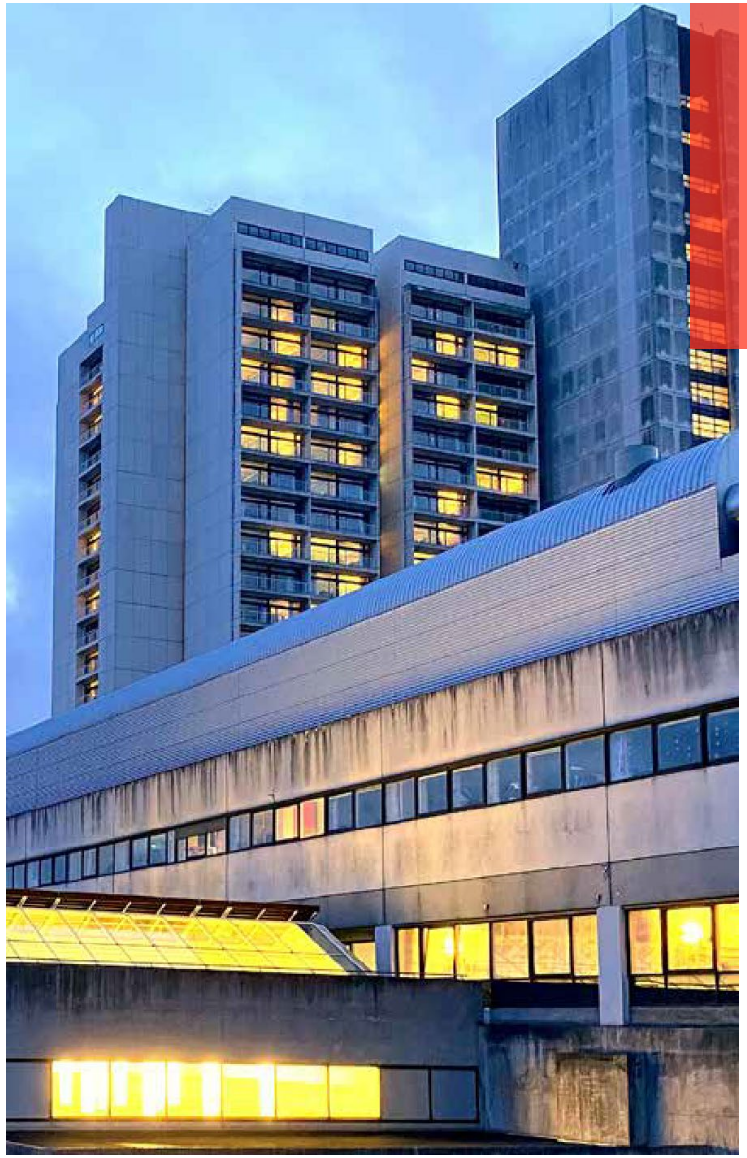
「私たちはデータに関して多くの課題を抱えていました。まずデータ処理が不効率でした。バッチのオーバーランも発生していました。24時間の時間枠を取っていては満足な医療データやサービスを提供できなくなることに気づきました。」

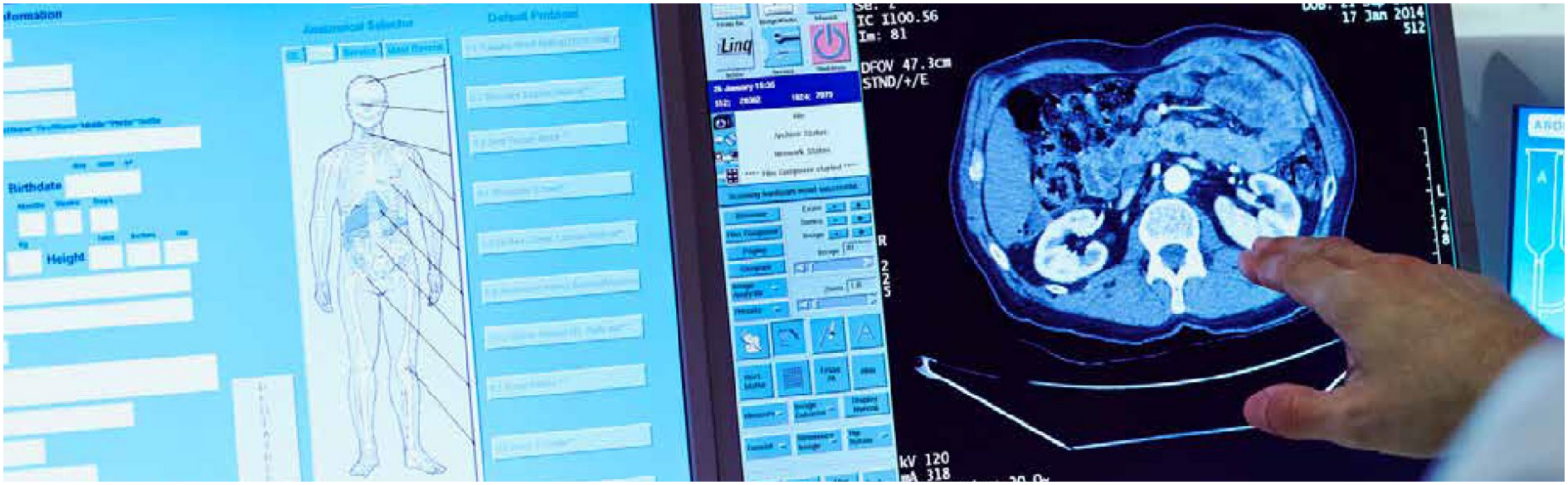
ヘルスダイレクト・オーストラリアは、業務を適切にサポートするためには、エンドツーエンドのプロセスと技術スタックを刷新する必要があることを認識しました。

Databricks と Delta Lake で分析を刷新

ヘルスダイレクト・オーストラリアは、データエンジニアリングをシンプルにし、データサイエンスのイノベーションを加速させる統合データ分析プラットフォームを導入することを決断しました。Databricks の Notebook 環境を利用することで、特殊な用途のジョブを毎回実行する必要がなく、コントロールされた方法でコンテンツを変更できるようになりました。

「Databricks は、データ運用だけでなくチーム運営にも良い影響を及ぼしています。アナリストとデータオペレーションチームが直接やりとりするようになり、全体の工数が半減しました。彼らが協力し合うことでサービスの提供スピードが大幅に向上しています。」（ピーター・ジェームズ氏）





ヘルスダイレクト・オーストラリアは、Delta Lake を利用して、Landing、Raw、Staging、Gold といった論理的なデータゾーンを作成しています。これらのゾーン内では、構造化データも非構造化データも、そのまま Delta Lake のテーブルに保存されます。そこから、メタデータ駆動型のスキーマを使用し、そのテーブル内のネストされた構造にデータを保持します。これにより、あらゆるソースからのデータを一貫して処理することができ、データを利用するさまざまなアプリケーションへのデータのマッピングが簡素化されます。

また、構造化ストリーミングにより、全ての ETL バッチジョブをストリーミング ETL ジョブに変換し、複数のアプリケーションに一貫したサービスを供給できるようになりました。ヘルスダイレクト・オーストラリアは、Spark 構造化ストリーミング、Delta Lake、Databricks 統合データ分析プラットフォームの導入により、アーキテクチャが大幅に改善され、性能の向上、運用上のオーバーヘッドの削減、プロセスの効率化に成功しています。

データパイプラインの高速化が可能にする患者中心のヘルスケア

ヘルスダイレクト・オーストラリアは、Databricks による性能の改善と Delta Lake によるデータの信頼性向上の結果、名前のファジーマッチングアルゴリズムの精度を 95% まで向上させ、手作業を排除しました。Databricks 導入以前は手作業による確認に依存しており、80% に満たない精度でした。

また、Delta Lakeと構造化ストリーミングによる運用効率の向上により、月に 3 万件以上の自動更新を処理できるようになりました。Databricks 導入以前は、信頼性の低いバッチジョブを手作業で行っており、同数の更新を処理するために 6 か月を要しており、データ処理の効率が 6 倍向上したことになります。

「Databricks の導入により市場投入までの時間を短縮できました。分析や運用管理が効率化し、医療部門の新たなニーズに対応できるようになっています。」
（ピーター・ジェームズ氏）

さらに、データの読み込み速度を 1 分間に 100 万レコードまで高めることができ、2,000 万レコードにおよぶデータセット全体を 20 分で読み込めるようになりました。Databricks 導入以前は、同じ 100 万件のトランザクションに 24 時間以上を要しており、アナリストの迅速な意思決定が妨げられ、効果的なデータ活用が行えませんでした。

これらに加え、コンプライアンス要件を満たすうえで重要であったデータセキュリティの大幅な改善を実現しました。Databricks は、HIPAA などの標準的なセキュリティ基準に対応しているため、ヘルスダイレクト・オーストラリアは、同国が定めるセキュリティ要件も満たすことができました。大幅なコスト削減を実現しつつ、役割変更に伴うアクセス権限の更新、メタデータレベルのセキュリティの変更、データ侵害などのインシデントを監視・検知することで、データの確実性を維持できるようになりました。

「Databricks の導入により市場投入までの時間を短縮できました。分析や運用管理が効率化し、医療部門の新たなニーズに対応できるようになっています。」（ピーター・ジェームズ氏）

ヘルスダイレクト・オーストラリアの未来は明るいものになるでしょう。Databricks の導入により、データと分析の価値を証明し、それらが組織運営とビジョンに与える影響の大きさを示しました。文書化されたデータリネージや品質が保証され、データへのアクセスが透明化されたことで、さまざまなビジネスグループやアナリストグループがより簡単かつ迅速にデータから価値を引き出せるようになりました。ヘルスダイレクト・オーストラリアは、全ての人の医療・ヘルスケアを向上させるという目的を達成すべく、積極的なデータの利活用に取り組んでいます。

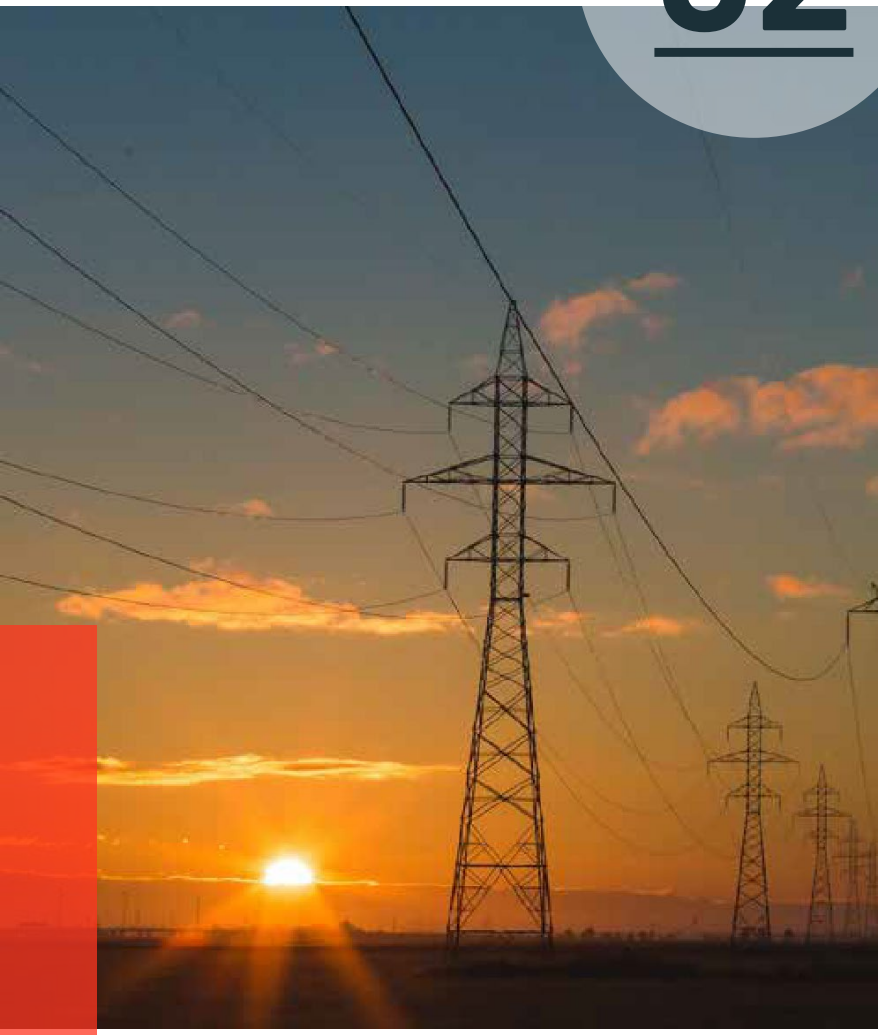
Chapter

02

導入事例 2：コムキャスト

Delta Lake と MLflow で視聴者エクスペリエンスを変革

02



導入事例 2：コムキャスト社

Delta Lake と MLflow で視聴者エクスペリエンスを変革

事例：コムキャスト
業種：メディア／エンターテインメント

10 倍
データ処理コストを 10 倍削減

90%
インフラ管理工数を 90% 削減

時間短縮
モデル展開に要する時間を数週間から数分に短縮

米メディア大手のコムキャスト社は、テクノロジーを活用して数 100 万の視聴者に対し、パーソナライズされたエクスペリエンスを提供することを目指していました。しかし、データパイプラインの処理能力が不足していること、データサイエンスに関わる部門間のコラボレーションが困難であることが、目標達成の障壁となっていました。コムキャスト社は、問題の解決策として、Delta Lake や MLflow が統合されている Databricks を導入。ペタバイト規模のデータのための高性能なデータパイプラインを構築し、機械学習モデル 100 種類以上のライフサイクルの管理を簡素化しました。その結果、エミー賞受賞にもつながる、革新的でパーソナライズされた視聴者エクスペリエンスを実現しました。

データや機械学習のニーズに対応できないインフラ

特定の番組に対する視聴者の声によるリクエストに素早く対応すること、また、数 10 億件におよぶ視聴者とのやり取りを実用的な洞察に変えることが、コムキャスト社の IT インフラストラクチャ、データ分析およびデータサイエンス部門にとって大きな課題でした。 コムキャスト社ではさらに、クラウド、オンプレミス、また、場合によってはデバイスへの直接接続など、異なる環境にモデルを展開する必要がありました。Databricks 導入前のコムキャスト社は次のような課題を抱えていました。

膨大なデータ

エンターテインメントシステムから数 10 億件のイベント、音声対応リモコンから 2000 万件以上のデータが生成され、分析のためにセッション化が必要なデータはペタバイト規模になっていた。

脆弱なパイプライン

データパイプラインは複雑で、失敗を繰り返し、回復作業が困難だった。 さらに、多数の小規模なファイルの管理に手間がかかり、ダウンストリームの機械学習タスクのためのデータ取り組みに遅延が生じていた。

不十分なコラボレーション

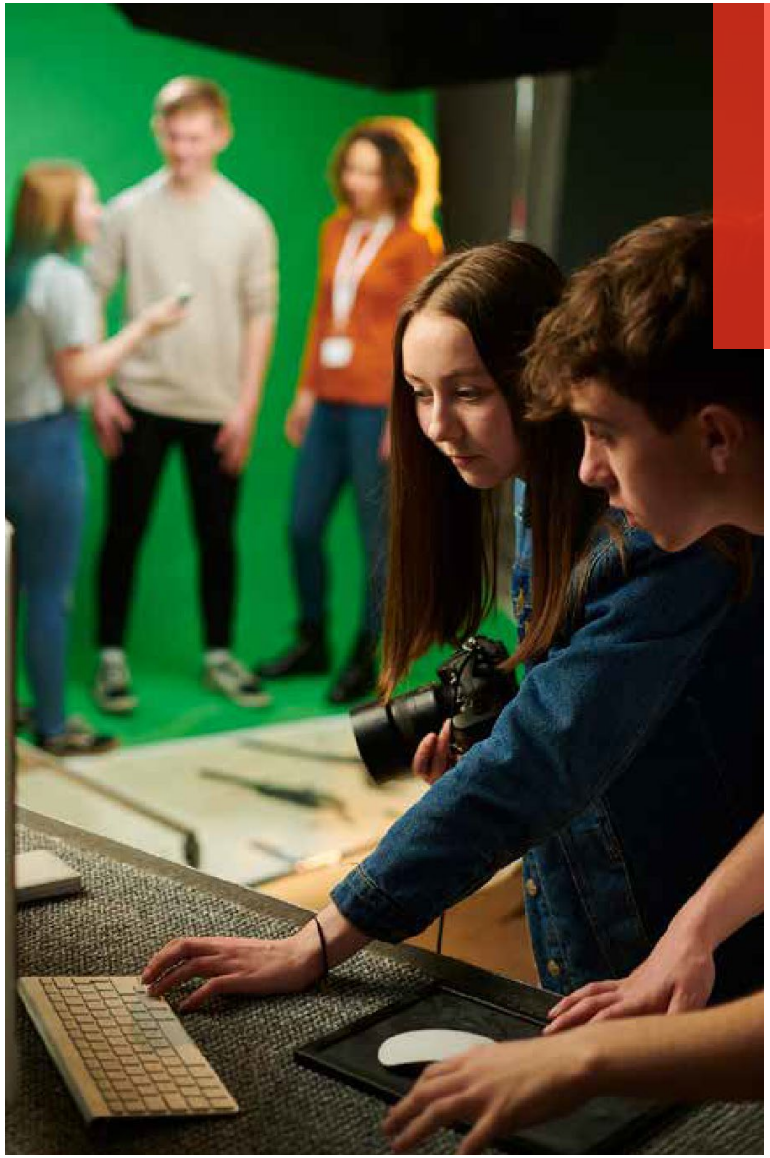
世界中に分散するデータサイエンティストが、それぞれ異なるスクリプト言語を使用しており、コードの共有と再利用が困難であった。

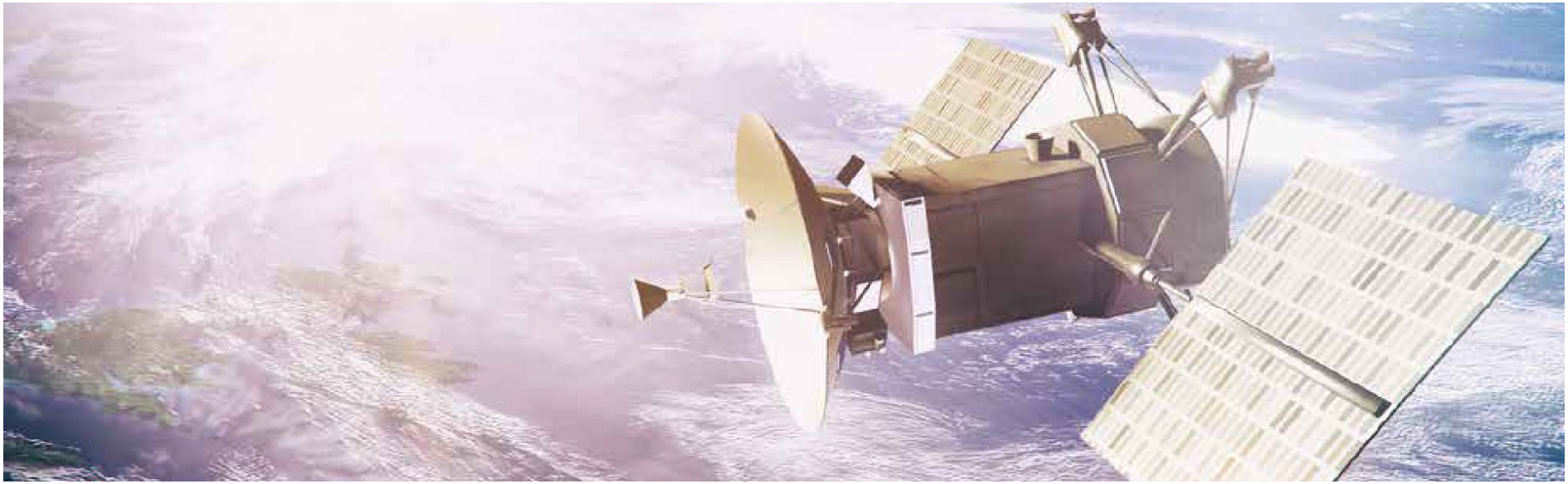
機械学習モデルの不効率な管理

数百種類の機械学習モデルの開発、トレーニング、展開が主に手動で行われており、遅く、複製が困難で、拡張性に欠けていた。

部門間の見解の相違

最新のツールとモデルの使用を推進する開発部門と、実績のあるインフラストラクチャ上での展開を希望した運用部門に、見解の相違が生じていた。





Delta Lake によるインフラの自動化とデータパイプラインの高速化

コムキャスト社は、視聴者を喜ばせるための新しい施策を打ち出すためには、データ取り込みから機械学習モデルの展開まで、データ分析プロセスの刷新が必要であるとの結論に達し、Databricks の統合データプラットフォームの導入に至りました。その結果、リッチデータセットの構築、大規模な機械学習の最適化、複数部門が共有するワークフローの合理化とコラボレーションの促進、インフラの簡素化、優れた視聴者エクスペリエンスの提供が可能になりました。

インフラ管理の簡素化
クラスタの自動管理、および、自動スケーリングやスポットインスタンスなどのコスト管理機能による運用コストの削減。

高性能データパイプライン
データの取り込み、エンリッチメント、動画や音声アプリケーションやデバイスからのテレメトリー生データの初期処理が可能になった。

大量の小規模ファイルの確実な処理
Delta Lake による、膨大な数の小規模ファイルの迅速かつ確実な取り込みを可能にする最適化。

コラボレーションワークスペース
インタラクティブなノートブックにより、チーム間の共同作業とデータサイエンスの創造性が向上。これにより、モデルのプロトタイピングが迅速になり、高速なイテレーションが可能になった。

機械学習ライフサイクルの簡素化
マネージド MLflow によって、Kubeflow 環境における機械学習ライフサイクルとモデルの提供が簡素化され、数百種類以上のモデルの追跡・管理を容易にした。

大規模な ETL の信頼性向上
Delta Lake が効率的な大規模分析パイプラインを可能にし、履歴データとストリーミングデータの統合を確実に行い、より豊かな洞察を抽出できるようになった。

機械学習を活用したパーソナライゼーション

競争の激しいエンターテインメント業界では、立ち止まることは後退を意味します。コムキャスト社は、分析のための統合プラットフォームの導入によって AI を活用した未来型エンターテインメントを先取りし、視聴者エクスペリエンスをより魅力的なものにすることでエンゲージメントを維持し、競争優位性を高めています。

エミー賞に輝く視聴者エクスペリエンス

Databricks の導入により、インテリジェントな音声コマンドを使った革新的な視聴者エクスペリエンスを実現。エンゲージメントを高めることに成功し、エミー賞を受賞。

コンピューティングのコストを 1/10 に削減

Delta Lake の利用により、データの取り込みを最適化し、性能を向上させると同時にマシンの台数を 640 から 64 に削減。 インフラ管理が容易になり、データの分析に注力できるようになった。

DevOps 工数を削減

ユーザー 200 人のオンボーディングに要する DevOps のリソースを、5 名から 0.5 名に削減。

データサイエンスの生産性向上

インタラクティブな単一のワークスペースで複数の言語をサポートすることで、グローバルに分散するデータサイエンティスト間のコラボレーションを促進。さらに、Delta Lake により、データ部門はデータパイプライン上のデータにいつでもアクセスできるようになり、迅速なモデルの構築とトレーニングが可能になった。

モデル展開の高速化

異なるプラットフォームでのモデル展開が可能になり、展開時間が数週間から数分に短縮。

Chapter

03

導入事例 3 ：バイアコム 18（Viacom18）

Hadoop から Databricks への移行により
視聴者エンゲージメントを強化

03

導入事例 3 : バイアコム 18 （Viacom18）

Hadoop から Databricks への移行により 視聴者エンゲージメントを強化

10 年間で 40 倍の成長を遂げたインドのエンターテインメントネットワーク企業バイアコム 18 社（Viacom18 Media Pvt. Ltd.）は、月間 6 億人以上の視聴者に対して、マルチプラットフォーム、多世代、多文化のブランドエクスペリエンスを提供しています。

バイアコム 18 社は、数百万人におよぶ視聴者に対してより魅力的なエクスペリエンスを提供すべく、Hadoop 環境から Databricks に移行しました。大規模なデータを効率的に処理できないことが主な理由でした。

Databricks の導入により、インフラ管理の合理化、データパイプラインの高速化、データ部門の生産性向上が実現しました。現在は、視聴者エクスペリエンスのパーソナライズ、ビジネスの最適化および、ROI の向上に取り組んでいます。

事例：バイアコム 18
業種：メディア／エンターテインメント

26 %

運用効率の改善により総コストを 26% 削減

視聴者数とデータ量の増加が Hadoop の限界を超える

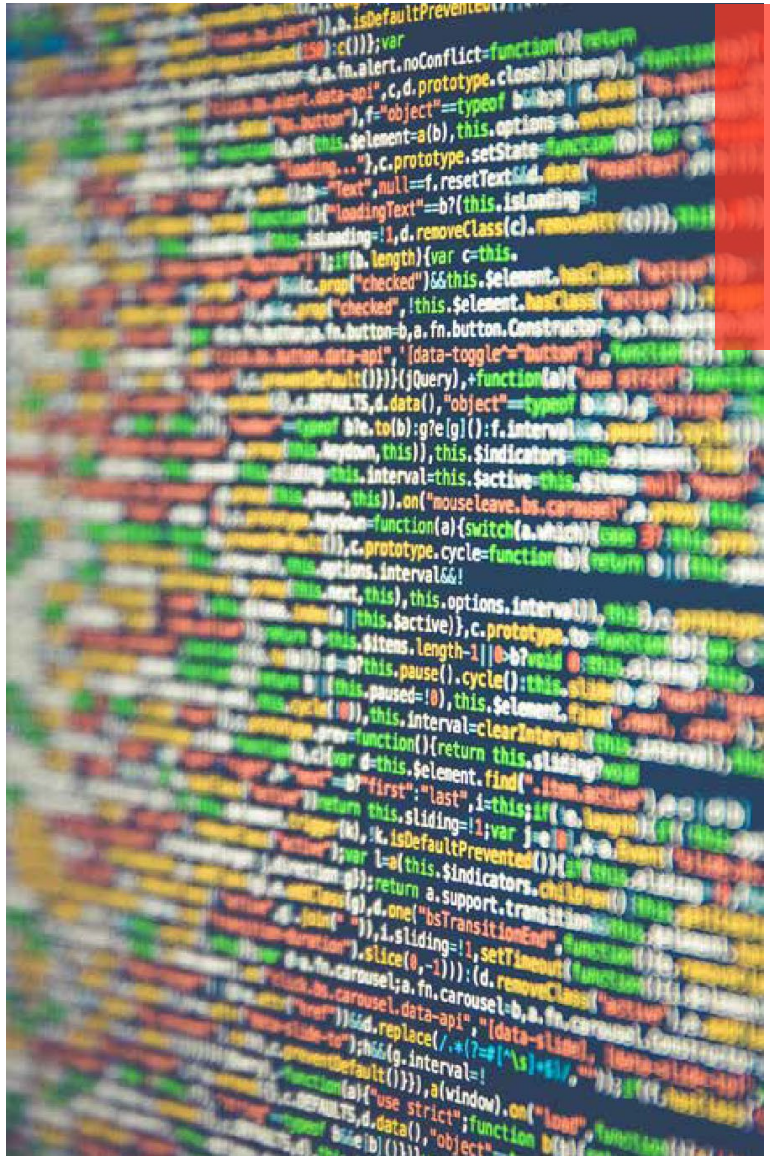
ネットワーク 18 社（Network18）とバイアコム CBS 社（ViacomCBS）のジョイントベンチャーであるバイアコム 18 社は、高度にパーソナライズされた視聴者エクスペリエンスの提供を重視しています。この戦略の核となるのは、毎日蓄積される視聴データをもとにした強力な顧客分析を可能にするエンタープライズデータアーキテクチャの導入です。しかし、インド全土におよぶ何百万人もの消費者をサポートするためには膨大な量のデータを処理しなければなりません。例えば、同社が提供するオンデマンドビデオ配信プラットフォームである VOOT は、日々 45,000 時間以上に該当するコンテンツを取り込んで処理する必要があるため、1 日に 700 GB から 1 TB のデータを生成することになります。

バイアコム 18 社のデジタル変革・技術部門アシスタント VP であるパリヤット・デイ（Parijat Dey）氏は、次のように述べています。

「コンテンツは私たちの業務の中核をなすものです。視聴履歴や好みに基づくパーソナライズされたコンテンツを視聴者に提案することで、視聴率や顧客ロイヤルティの向上を図っています。」

バイアコム 18 社は、データレイクの運用にオンプレミスの Hadoop を活用していましたが、90 日分のローリングデータを適切に処理することができず、同社が定める SLA を満たせていませんでした。また、分析のニーズにも応えることができず、顧客エクスペリエンスだけでなく全体のコストにも影響が及んでいました。

この課題に正面から取り組むためには、1 日単位のスナップショットに頼るのではなく、より長期間のデータトレンドの分析が可能なモダンデータウェアハウスの導入が必要でした。また、インフラストラクチャをシンプルにするプラットフォーム、すなわち、自動スケーリングなどの機能によりクラスタのプロビジョニングを容易にしてコンピューティングコストの削減を可能にするプラットフォームが必要でした。



Databricks で分析と機械学習のデータ処理を高速化

バイアコム 18 社は、処理能力とデータサイエンスのケイパビリティを強化するために、Salesforce、データ分析、ビッグデータの大手コンサルティング企業であるセレバル社（Celebal Technologies）と提携しました。セレバル社は、Azure Databricks を活用した統合データ分析プラットフォームにより、バイアコム 18 社のデータウェアハウス機能のモダナイズと大規模なデータ処理の高速化を実現しました。

Delta Lake 上のデータキャッシング機能により、特に重要であったクエリの高速化が実現したほか、自動スケーリング機能を備えたクラスタ管理や、ストレージとコンピューティングの分離によってインフラ管理がシンプルになり、運用コストの最適化が実現しました。

「Delta Lake でデータパイプラインの管理がシンプルになりました。運用コストも低減し、ダウンストリームの分析とデータサイエンスによる気づきの発見がスピードアップしています。」（パリヤット・デイ氏）

さらに、Databricks の Notebook 機能がバイアコム 18 社に想定外の好効果をもたらしました。データ部門が共通のワークスペースを活用し、モデルトレーニング、アドホック分析、ダッシュボードやレポートの作成など、PowerBI でさまざまな工程を共同で行うようになったことで、生産性が向上しています。

データ活用による視聴者エクスペリエンスのパーソナライズ

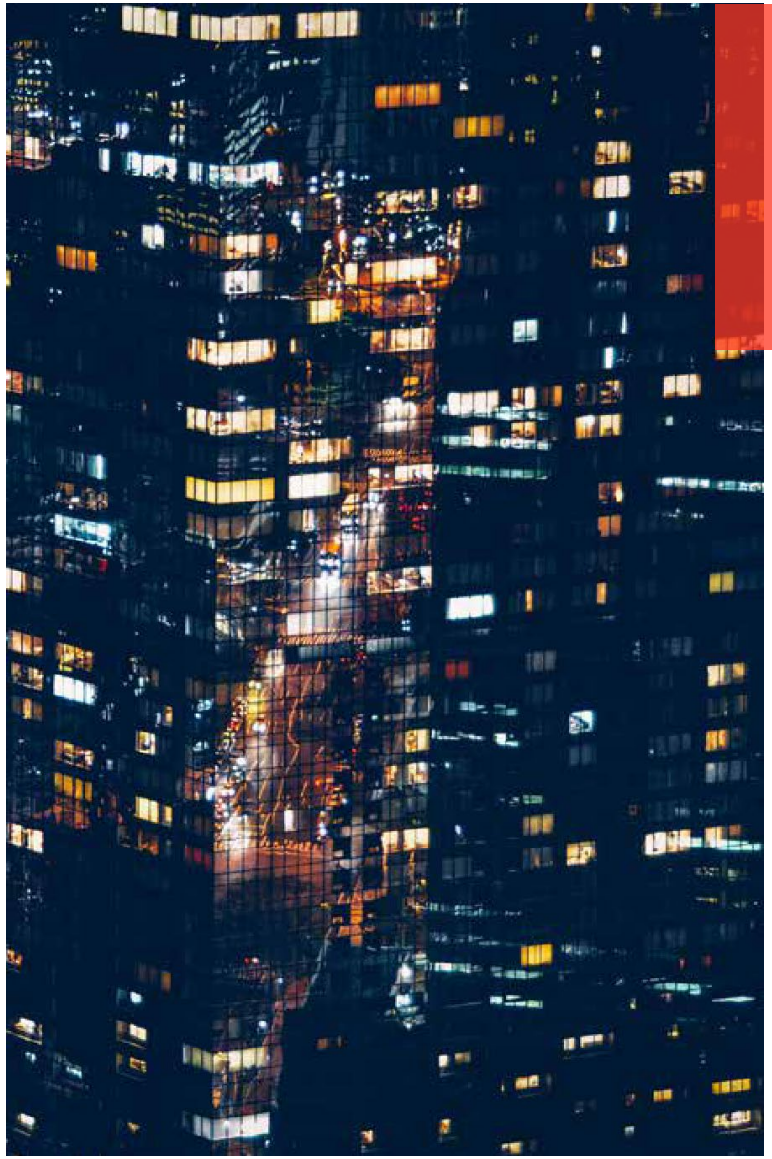
Databricks は、セレバル社と協力してバイアコム 18 社の部門間のコラボレーションと生産性を向上させ、革新的な顧客ソリューションの提供と気づきの取得を可能にしました。データ部門は Databricks を活用してシームレスにデータをナビゲートできるようになり、顧客サービスの向上へとつながっています。

「Databricks の導入により、膨大なデータのきめ細かい分析が可能になり、顧客の行動やエンゲージメントに関する知見をアナリストやデータサイエンティストに提供できるようになっています。」（パリヤット・デイ氏）

性能向上に加えてクエリ時間が短縮されたことで、データ量が日々増加しているにもかかわらず、総所有コストは低減しました。

「Azure Databricks によってプロセスが大幅に合理化され、生産性が約 26% 向上しました。」（パリヤット・デイ氏）

デイ氏は、Hadoop から Databricks への移行によって大きなビジネス価値を得られたと述べています。バイアコム 18 社は、障害発生時のコスト削減、大規模な処理速度の向上を実現し、アドホック分析の簡素化によってデータ探索を容易にし、魅力的な顧客エクスペリエンスを提供するイノベーションを実現しました。



次のステップ

この eBook では、Delta Lake とその機能が性能を向上させる仕組みについて解説しました。このシリーズの他の eBook では、Delta Lake のリソースを詳しくご紹介しています。

この eBook の後続シリーズ

- Delta Lake シリーズ：基礎と性能
- Delta Lake シリーズ：機能
- Delta Lake シリーズ：レイクハウス
- Delta Lake シリーズ：ストリーミング

Delta Lake をさらに詳しく

- Databricks の Web サイト
- Databricks の無料トライアル
- 技術トークシリーズ：Delta Lake の基礎（英語）
- 技術トークシリーズ：Delta Lake を深掘り（英語）
- Web セミナー：Delta Lake でデータレイクにオープンソースの信頼性を（英語）

databricks