

データ・イン・モーションの原理

エンタープライズ全体にわたるストリーミング
データアーキテクチャーの設計図

データ・イン・モーションへの対応

このソリューションブリーフで紹介するデータ・イン・モーション機能をエンドツーエンドで実装し、成果を上げた Cloudera のお客様の事例を以下にご紹介します。

ある世界的な医療機器メーカーは、メッセージングアーキテクチャーを最新化し、従来製品よりも高い頻度と解像度でより多くのデータを生成する埋め込み型の新しい医療機器製品に対応できるようになりました。

- フロー管理**—医療データはプライバシーに関わるため、データフローが複雑であり、移動中のデータと保存時のデータの両方を暗号化する必要がありました。新しい取り組みでは、NiFi のノーコードのユーザーインターフェースを使用することで、ビジネスに100%集中できるようになり、テクノロジーチームによるサポートが必要最低限になりました。
- ストリームメッセージング**—メッセージの量は、四半期に1度デバイスのステータスレポートを作成していた状態から、健康状態のリアルタイムでのモニタリングへと、飛躍的に増えました。Kafka を採用したことで、それだけの量を複数のオンプレミス環境とクラウド環境にわたって拡張できるようになりました。
- ストリーム処理と分析**—このメーカーでは、データの処理方法をバッチ処理からリアルタイム処理に移行する必要がありました。Flink は、バッチおよびリアルタイム処理だけでなく、近い将来に計画している複雑なイベント処理にも対応します。このため、ストリーム処理/分析エンジンである Flink を導入してサポートするだけでリアルタイム処理への移行を実現できました。

エンドツーエンドのストリーミングアーキテクチャーを失敗に終わらせないために

エンドツーエンドのストリーミングアーキテクチャーでは、最適なメッセージングソリューションを柱に据えるだけでは十分ではありません。これは、お客様のデータの取り込み作業をサポートしてきた経験から Cloudera が学んだことです。つまり、ストリームメッセージング機能に、フロー管理とストリーム処理/分析という2つの要素も統合する必要があります。この3つの要素を適切に統合できれば、持続可能で拡張性と順応性に優れたエンドツーエンドのストリーミングアーキテクチャーを確立できます。しかし、1つでも不十分な要素があれば、三脚椅子のように、構造全体が不完全なものとなります。

このソリューションブリーフでは、Cloudera が考えるデータ・イン・モーション(移動中のデータ)の原理について説明します。エンタープライズ全体のストリーミングデータへのアプローチを評価し、シンプル化するにあたっての設計図として、ビジネスやテクノロジーの意思決定者に役立つ内容となっています。

本書で取り上げるストリーミングアーキテクチャー

以下に挙げる3つの要素を連携させて、エンドツーエンドの統合ストリーミングアーキテクチャーを実現します。

- フロー管理**: 大まかに言えば、多数のプロデューサーとコンシューマーに対してデータを収集、配信、変換することです。
- ストリームメッセージング**: プロデューサーとコンシューマー間でメッセージのプロビジョニングと配信を行います。
- ストリーム処理と分析**: プロデューサーとコンシューマー間でストリーミングされているデータから分析的洞察をリアルタイムで生成する方法です。

ストリーミングアーキテクチャーの鍵となる要素



Cloudera が提唱するデータ・イン・モーションの原理は、フロー管理を実現する Apache NiFi、ストリームメッセージングに対応する Apache Kafka、ストリーム処理と分析を提供する Apache Flink が相互に補完し合うことによってもたらされる機能を基盤としています。

フロー管理に取り組んだお客様の事例

Cloudera は、フロー管理に取り組むお客様を何年も前から支援してきました。

当初はデータレイクへのデータの取り込み方法を改善することを目指していましたが、やがて拡張性に優れたストリーミングイベントアーキテクチャーを実現する必要が生じ、続いてこうしたストリーミングデータを活用して分析を実行できるようにすることが目標となりました。

クラス最高のコンピューティングエンジンは、次のような段階を経て Cloudera のデータ・イン・モーションのビジョンへと進化しました。

1. **NiFi** を最初に導入して、コードなしでデータレイクへのデータの取り込みを可能にしました。その後 NiFi をフェデレーテッドクラスタに実装し、エンタープライズ全体でデータを移動できるようにしました。
2. **Kafka** と NiFi を統合して、拡張性に優れたイベントアーキテクチャーを支える主要なメッセージング基盤を構築しました。
3. **Flink** を第3世代のストリーム処理/分析エンジンとして導入し、リアルタイム処理をバッチ処理と同様に簡単に実行できるようにしました。
4. **Cloudera Data Platform** との統合によって、Shared Data Experience による柔軟性に優れた PaaS (Platform-as-a-Service) を実現しました。

データ・イン・モーションの原理

包括的なエンドツーエンドのデータパイプラインでは、その各要素にクラス最高のコンピューティングエンジンを活用すべきであると Cloudera は考えます。また、プラットフォームを高度に抽象化することによって、エンジンの接続、管理に伴う複雑さにバックグラウンドで対処し、ユーザーがビジネスロジックに集中できるようにする必要があります。Cloudera Data Platform (CDP) では、この2つの原則を次のソリューションで実現します。

1. **Cloudera DataFlow (CDF)** は、ストリーミングデータに関連するあらゆる取り組みを支えるデータ・イン・モーションプラットフォームです。CDF には、以下の3つの要素がすべて統合されています。
 - エッジでのデータの取得とフロー管理
 - 取り込んだデータを Kafka メッセージングバックボーンとの間で相互に直接プロビジョニング
 - ストリーム処理と分析
2. **Shared Data Experience (SDX)** は、統合された一連の共通サービスを提供します(詳しくは、5ページの「Shared Data Experience」を参照)。SDX により、データセンターとクラウド環境全体のセキュリティとガバナンスを統合します。

メッセージングをすべての中心に据えることの限界

エンドツーエンドのアーキテクチャーの中心に、拡張性に優れたメッセージングソリューションを据えることは重要です。しかし、これだけでは拡大を続けるエンタープライズの高度なリアルタイムのユースケースや、技術上、運用上のデータ移動要件に対応するには十分ではありません。

テクノロジーチームは、ビジネスニーズにリアルタイムで対応するために、大規模なモノリシックデータベースアーキテクチャーから、イベントドリブンのアプリケーションやマイクロサービスによる設計へと方向転換しています。また、保存中のデータを分析するだけでなく、リアルタイムのデータストリームから直接意思決定を行うケースが増えています。

Kafka は、大規模な組織のための単一の集中型ストリーミングアーキテクチャーバックボーンとして登場しました。基本的な課題となっていた拡張性の問題に対処し、メッセージの臨機応変なやり取りと継続的なやり取りの両方に高度に最適化されている Kafka の登場は、待ち望まれていたものでした。しかし Kafka では、

さまざまなソースからデータをリアルタイムで取り込んだり、ストリーミング中のデータについてリアルタイムの洞察を生成したりすることに伴う課題に対処できません。こうした課題を解決する最善の方法は、ストリーミングアーキテクチャーにフロー管理とストリーム処理/分析機能を組み込むことです。

フロー管理に関する主な考慮事項

フロー管理の要件を評価するにあたって考慮すべき重要な側面は、ツールの拡張性、使いやすさ、データの出所の3つです。Apache NiFi は、リアルタイムの統合データロジスティクスと、シンプルなイベント処理のためのプラットフォームであり、これら3つの側面すべてに本質的に対応します。

ツールの拡張性

データが発生する場所から消費される場所まで、データのライフサイクルは、広大な地理的境界とセキュリティ境界にまたがります。例えばセンサーデータは、さまざまな理由から、地域のビジネスセンター、取引先、世界本社に逐次的にまたは並行して転送する必要があるかもしれません。そうしたデータの中には、機密性が高く、隔離または除去すべきデータもあると考えられます。組織内の変化によって、データのプロデューサー、コンシューマーがその場その場で追加、削除、修正、再設計されることも、複雑さを生み出すもう1つの要因となります。

データフローのさまざまなシナリオや、組織特有のシナリオに対処するために必要なツールには、優れた柔軟性と拡張性が求められます。NiFi は、エンタープライズ内の多種多様な場所にあり、成熟度も異なるデータソースやターゲットの代わりに機能することで、この要件を満たします。また、データに依存しないため、形式、スキーマ、プロトコル、スピード、サイズが異なる多様な分散ソースに対応できます。

このほかにも NiFi では、機械のセンサー、地理位置情報デバイス、クリックストリーム、ファイル、ソーシャルフィード、ログファイル、動画などをすべて扱えます。さらに、ストリームの統合、データのエンリッチ化やフィルタリングといった簡単なイベント処理も行えます。

ストリーミングメッセージングに取り組んだお客様の事例

Kafka を利用する Cloudera の主なお客様にインタビューを行ったところ、Kafka クラスタ全体を把握できないことが大きな問題であることが明らかになりました。これを Cloudera では「Kafka のブラインドネス」と呼んでいます。

このため Cloudera は、エンタープライズ内の各チームが抱える独自の課題を解決するために、以下をはじめとする統合ツールセットを開発しました(詳しくは、[このページの「ストリーミングメッセージングの完全な把握」](#)を参照)。

- **Kafka プラットフォームの運用チーム** は、クラスタとブローカーを把握できるようになったほか、ブローカーがインフラストラクチャーに及ぼしている影響や、これとは逆にインフラストラクチャーがブローカーに及ぼしている影響も把握できるようになりました。
- **Kafka の DevOps およびアプリケーションチーム** は、プロデューサー、ブローカー、トピック、コンシューマー間のデータフローを完全に把握し、それぞれの主なパフォーマンスメトリクスを監視できるようになりました。
- **ガバナンスおよびセキュリティチーム** は、証拠保全、監査、メタデータ、アクセスコントロール、データリネージの完全な透明性を確保できるようになりました。

Cloudera が「Kafka のブラインドネス」を解消した方法について詳しくは、ホワイトペーパー『[Manage, Monitor and Replicate Apache Kafka Across the Enterprise and Cloud \(エンタープライズおよびクラウド環境での Apache Kafka の管理、監視およびレプリケーション\)](#)』をご覧ください。

データの出所

前述したようなデータの統合に伴う複雑さによって、エンタープライズ全体を移動するデータの出所や属性を理解することも極めて難しくなります。これは、一般に「データの出所」と呼ばれ、CDO や CISO が真っ先に解決しなければならないと認識している重要な問題です。どの時点で、CDO や CISO は、データポイントがシステムによってどのような影響を受けたかを正確に説明できるよう、チームの体制を整えておかなければなりません。

NiFi にはデータの出所に対応できる機能が最初から備わっています。具体的には、実行するすべての処理でデータリネージ情報を非常にきめ細かなレベルで生成し、イベントの前後の変化を記録します。

NiFi を使用するとプロデューサーとコンシューマー間のデータフローを統制できるため、エンタープライズ全体のエンドツーエンドのデータリネージを取得できます。このほかにも、データのガバナンスやセキュリティの確保に活用できるだけでなく、どのシステムがどのように通信し、システム間のレイテンシはどうなっているかといった運用面での認識も高められます。これも非常に有用な点です。

使いやすさ

ビジネスソリューションを主導し、エンドツーエンドのデータフローを理解している特定分野の専門家が、優れたコードの記述方法を知っているとは限りません。このため NiFi は、ノーコードのドラッグ&ドロップインターフェースを使用して、前述したような統合作業を簡単に行えるように作られています。構築したオーケストレーションフローが、実際のデータに影響する実際の関数にリアルタイムで変換されるため、対話形式で作業を進められます。

特定分野の専門家は、何がうまくいき、何がうまくいかないかを実感できるフィードバックを受け取れるため、有益なデータフローを短期間で構築することができます。

さらに、構築済みのプロセッサも300以上揃っており、あらゆるデータソースをあらゆるターゲットに簡単かつシームレスに接続できるため、複雑な作業は不要です。

ストリーミングメッセージングの完全な把握

ストリーミングメッセージングには、高い拡張性と安定性が必要です。この領域でのクラス最高のコンピューティングエンジンは、Apache Kafka です。Kafka は、大規模な組織のための単一の集中型ストリーミングアーキテクチャーバックボーンとして登場し、金融サービス、通信、製造をはじめとする多くの業界にデータ・イン・モーションにおける変革を起こしています。

Cloudera は、Kafka コミュニティに注力し、技術面での密接な関係を継続的かつ積極的に推進しています。この連携により、これまで重要なイノベーションや製品の向上を実現してきました。

Cloudera はデータ・イン・モーションプラットフォームである CDF の一部として、高機能な Kafka 環境をサポートし、拡張可能で包括的なコンポーネントエコシステムを提供しています。以下はその一部です。

- **Cloudera Streams Messaging Manager (SMM)** は、Kafka クラスタ全体のプロデューサー、ブローカー、トピック、コンシューマー間で移動するデータをエンドツーエンドで把握できるようにする、単一の監視/管理ダッシュボードです。
- **Streams Replication Manager (SRM)** は、統合ツールセットの一部として、SMM に直接組み込まれています。SRM は、クラスタ間で Kafka のトピックをレプリケーションするエンタープライズクラスのソリューションであり、ストリーミングアーキテクチャーのビジネス継続性と高可用性を確保します。
- **Schema Registry** を使用すると、スキーマの不一致によって発生する中断を安全な方法で軽減できます。Schema Registry は、Kafka 環境全体のプロデューサースキーマおよびコンシューマースキーマすべての変化を管理、共有、サポートします。
- **Cruise Control** は、大規模な Kafka 環境の負荷分散を行う Kafka コンポーネントです。ユーザーが定義した目標に基づいてパーティションの負荷を自動で調整するだけでなく、異常を検知し、プロアクティブに解決します。

ストリーム処理と分析に取り組むお客様の支援

データフローへの取り組みの第3段階では(詳しくは、2ページの「フロー管理に取り組んだお客様の事例」を参照)、エンタープライズ全体の移動データを容易に管理できる NiFi と、拡張性に優れたイベント/ストリームメッセージングアーキテクチャーである Kafka との統合を基盤として、ストリーミングデータを活用し、リアルタイムで洞察を引き出します。

以前 Cloudera では Apache Storm、Spark Structured Streaming、Kafka Streams をサポートしていましたが、これらではデータの遅延や欠如、複雑なイベント処理といった課題に対処できず、高信頼性、高可用性、データ損失ゼロを保証することもできませんでした。

こうした複雑な分析ニーズに対応する、第3世代のストリーム処理/分析エンジンとして先頃登場したのが Apache Flink です。Cloudera は、Apache Flink をデータ・イン・モーションポートフォリオにさっそく組み込みました。

Flink の機能と、ストリーム処理/分析エンジンの選択に関する重要な技術面、運用面での要素について詳しくは、ホワイトペーパー『Choose the Right Stream Processing Engine for Your Data Needs (データ要件に適したストリーム処理エンジンの選択)』の分析エンジンに関するセクションをご覧ください。

豊富な統制機能とオプションを備えたストリーム処理と分析

3つ目の要素は、ストリームの処理機能と分析機能です。前述の2つの要素は、ストリーミングデータを完全に把握し、データの出所を掌握しながら移動、プロビジョニング、レプリケーションを行う効果的な方法を提供します。しかし、ビジネスの意思決定に役立つ実用的なインテリジェンスを獲得するには、ストリーミングデータをリアルタイムで処理する必要もあります。

Cloudera は、データ・イン・モーションに取り組むお客様の支援を通じて、データパイプラインのあらゆる要件をカバーする、クラス最高のストリーム処理/分析エンジンが必要であると考えようになりました。そこで CDF に組み込んだのが Apache Flink です(詳しくは、4ページの「ストリーム処理と分析に取り組むお客様の支援」を参照)。

Apache Flink は、分散型処理エンジンであり、拡張性に優れたデータ分析フレームワークでもあります。数百万ものデータポイントや複雑なイベントを極めて容易に処理し、予測に基づく洞察をリアルタイムで提供します。Apache Flink をクラス最高と呼べる理由は、一部の高度な分析ユースケースにも対応できる豊富な統制機能を技術面、運用面の両方で備えていることにあります。

Flink は、ストリーミングを(バッチ処理よりも)最優先する手法で大量のストリームデータを大規模に処理しながら、ステートフルストリーミング、1回だけの配信、高度なウィンドウ化技術などの重要な機能もサポートし、耐障害性/高信頼性機能もあらかじめ搭載しています。生成されたデータをリアルタイムで処理できるほか、ストレージファイルシステム、パブリッククラウドのオブジェクトストレージなどの永続性のあるリポジトリ内に格納されたデータも処理できます。

使いやすさの観点では、柔軟性と表現性に優れた Flink の API を使用することで、開発者は複雑なイベント処理などの高度なストリーミングアプリケーションを短時間で作成できます。リアルタイムのデータへのアクセスと処理にこうした API や SQL を使用することで、アプリケーション開発が大幅にシンプル化されます。

何社もの有名大手企業が、リアルタイムのストリーム処理/分析ニーズに対応するために Flink の大規模導入にすでに投資を行っているのには、以下のような理由があります。

- マイクロサービス、バッチ処理、ストリーミングといったあらゆるユースケースに対応できる柔軟性が必要である
- 高スループットが必須である
- 低レイテンシが極めて重要である
- 複雑なイベント処理などの高度な機能が必要である
- 運用面の効率を、技術的な機能と同様に重視する必要がある
- 組織全体での採用を促すために、使いやすさと拡張性が重要である

Flink は Cloudera のデータ・イン・モーションの原理に不可欠なコンポーネントであり、企業のセキュリティフレームワーク、運用プロセス、サポート体制に完全に統合できます。

Shared Data Experience (SDX)

このソリューションブリーフで紹介したクラス最高のコンピューティングエンジンの接続や管理に伴う複雑さをユーザーに意識させない、高度な抽象化プラットフォームを提供することは、Clouderaのデータ・イン・モーションの原理の重要な柱です。

Cloudera SDX は、この柱を支える要素であり、他のプラットフォームプロバイダーとの大きな違いでもあります。SDXを使用すれば、データのセキュリティ、ガバナンス、統制に関するポリシーを1度設定するだけで、データセンター、ハイブリッドクラウド、マルチクラウドといった場所を問わず、すべてのコンポーネントにわたってポリシーを一貫して適用できます。

SDXは、次のソリューションによって、導入の選択肢を増やし、柔軟性を高めます。

- **Apache Ranger** は、きめ細かく一貫性のある集中的なアクセスコントロールによって、エンタープライズ全体のセキュリティを確保します。エンドツーエンドのストリーミングプラットフォーム全体のアセットを一元管理ビューから監視、保護できます。
- **Apache Atlas** は、エンタープライズクラスの監査、リネージ、ガバナンス機能を提供します。エンタープライズ全体のデータを把握し、理解できるようにするための重要なコンポーネントです。
- **Apache Knox** は、エンタープライズのセキュリティポリシーへの準拠を維持しながら、ユーザーがクラスタデータにシームレスかつ安全にアクセスできるようにしたり、ジョブの実行権限を適切に付与したりすることでセキュリティ統制をシンプル化する、ゲートウェイベースのSSOです。

高度な抽象化による統合

Cloudera では、ストリーミングデータのエンドツーエンドでの管理に伴う独自の課題に適切に対処するには、クラス最高のコンピューティングエンジンが必要であると考えています。

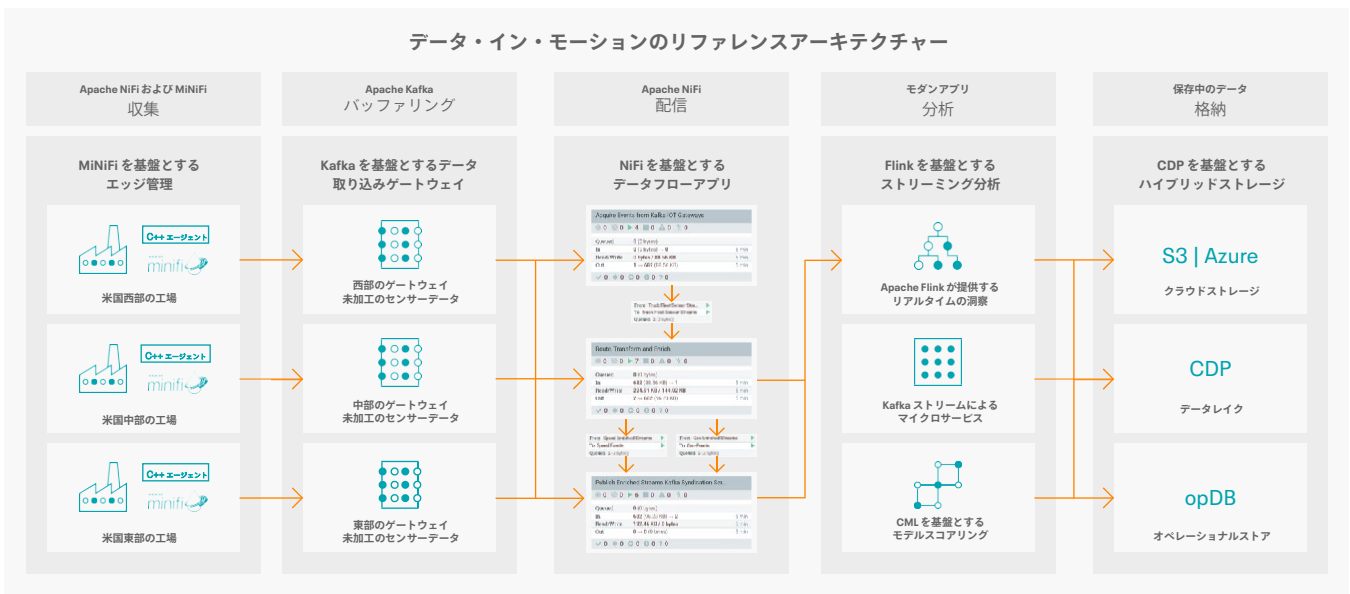
この原理を実際に活用するために、高度な抽象化によってコンピューティングエンジンの接続、管理、統合に伴う複雑さの問題を解決する統合プラットフォームを提供しています。

ユーザーはこれを利用して純粋にビジネスロジックに集中し、エンドツーエンドのデータパイプラインを構築できます。各エンジンにまたがるビジネスロジックのレンダリングは、Cloudera がシームレスに実行するため、ユーザーが複雑さを意識することはありません。すでに触れた例も含め、Cloudera では次のような手法を取っています。

- **ノーコードのユーザーインターフェース。**このインターフェースを使用することで、特定分野の専門家はデータフロー図を実際の関数に変換し、実際のデータにリアルタイムに影響を与えることができます。
- **単一の監視/管理ダッシュボード。**エンドツーエンドでのデータの把握を可能にするとともに、レプリケーションソリューションを提供して、大規模なストリームメッセージング環境全体のビジネス継続性と高可用性を確保します。
- **表現性と柔軟性に優れた API。**こうした API を使用することで、開発者は高度なストリーム処理/分析アプリケーションを容易に構築できます。
- **一連の共通サービスとの緊密な統合。**これにより、エンタープライズのデータセンターとクラウド環境全体で統合されたセキュリティとガバナンスを実現します。

以下のリファレンスアーキテクチャー図は、エンドツーエンドのデータパイプラインの大きな流れの中での NiFi、Kafka、Flink の位置付けと連携を示しています。

データ・イン・モーションのリファレンスアーキテクチャー



Cloudera について

Cloudera は、データの力によって、今日不可能なことも明日には実現できると信じています。Cloudera は、複雑なデータを明確で実践的な洞察に転換する力を人々に与えます。Cloudera は、エッジから AI に至るまで、あらゆる場所のあらゆるデータに対応することが可能なエンタープライズデータクラウドを提供します。Cloudera は、オープンソースコミュニティの絶え間ない革新を原動力に、世界最大規模の企業のデジタルトランスフォーメーションを推進していきます。

Cloudera Data Platform の詳細は
こちらから: cloudera.com/cdp

Cloudera DataFlow の詳細はこちらから:
cloudera.com/cdf

Cloudera の詳細はこちらから:
jp.cloudera.com

ビジョン

Cloudera のデータ・イン・モーションのビジョンでは、一連のサービスが統合されています。こうすることで、エンドツーエンドのデータパイプラインの開発をシンプル化し、組織のセキュリティフレームワーク、運用プロセス、サポート体制と統合し、ビジネスニーズに合わせて拡張、縮小することができます。

Cloudera は、クラス最高のデータストリーミングコンピューティングエンジンが統合されたプラットフォームに、高度な抽象化機能を組み合わせて提供することで、このビジョンを実現しています。このため、お客様はストリーミングデータパイプラインを構築するビジネスロジックに純粋に集中することができます。