



オンプレミスのETLから クラウドドリブンのELTへ

データパイプラインの価値と効率を最大化するためのベストプラクティス



CHAMPION
GUIDES

TABLE OF CONTENTS

- 2** エグゼクティブサマリー
- 3** 基本的な用語とコンセプトの理解
- 4** ELTの台頭
- 6** 汎用的なデータマネジメント戦略の確立
- 7** ELTの実装を検討すべきタイミング
- 7** データパイプラインの選択
- 9** Snowflakeによるデータの処理
- 10** 結論
- 11** Snowflakeについて

エグゼクティブサマリー

これまでのデータパイプラインは、オンプレミスのビジネスアプリケーションから収集される、動きが遅く簡単に分類可能で予測可能なデータに対応するように設計されたものでした。こうしたデータパイプラインは、さまざまなソースからデータを取得し、使える形式に変換してからデータウェアハウスなどのターゲットにロードするという抽出、変換、ロード (ETL) 型プロセスを採用しています。このような旧式のパイプラインは、エンタープライズアプリケーションからの構造化データのソースは問題なく処理できますが、最新のデータ環境の特徴である、多彩なデータタイプや取り込みスタイルには十分に対応することができません。

いっぽう最先端のパイプラインは、まずデータを抽出してロードし、目的の場所に到達してからデータを変換するように設計されています。このサイクルをELTと呼びます。最新のELTシステムは、変換ワークロードをクラウドへ移すことにより、より高い拡張性と伸縮性を実現しています。従来のオンプレミス環境では、ETLジョブは同じインフラストラクチャー上で実行される他のワークロードとのリソース競合が発生しますが、ELTでは、データを未加工の状態で読み込み、データがどのように使われるかははっきりしてから用途に適した様々な方法でデータを変換します。

ELTパイプラインの利用により、さまざまなタイプの未加工データをクラウドデータプラットフォームなどのクラウドベースのリポジトリにロードできます。このようなプラットフォームは、組織全体のデータの取り込み、変換、共有をスピードアップさせま

す。それによってリソースを大量に消費する変換ワークロードをクラウド上で実行できるようになり、スケーラブルなクラウドリソースの処理能力および容量を最大限に活用できます。

次ページ以降でも説明しますが、ELT処理が適しているのは以下のような用途です。

- **膨大なデータ量への対応:**ELTでは、大量の構造化および非構造化データをクラウド上ですばやく処理できます。
- **アナリティクスの実験:**ELT処理により、アナリストやデータサイエンティストはデータの可能性を追求し、特定のプロジェクトの必要に応じたデータ変換オプションを最大化できます。

- **低レイテンシのデータパイプライン:**ELTはデータを即時に転送するため、低レイテンシのアナリティクスやニアリアルタイムのユースケースにおける価値が大きいです。

それぞれの状況やワークロードに適した変換方法を使用することで、データパイプラインの価値を最大限に高めることができます。以下にその方法をご紹介します。



基本的用語とコンセプトの理解

ETLとは、さまざまなソースからデータを抽出し、ステージングサーバーでデータを変換し、データウェアハウス、データレイク、クラウドデータプラットフォームなどのターゲットにロードするソフトウェア統合プロセスです。

従来のデータウェアハウスではデータはリレーショナルデータモデルを前提としてマッピングされており、ターゲットデータベースへのロード前に、クレンジングやエンリッチメント、共通フォーマットへの変換などが必要です。

データの構造化と変換を行うことにより、SQLベースのビジネスインテリジェンス (BI) ツールによる高速かつ効率的な分析が可能となります。しかし、変換プロセスにおいて未加工データの成果物が一部失われるため、データの利用価値は限定的なものとなります。ほとんどのETLワークフローでは、ソースデータベースから取得したデータはデータウェアハウス内でまずス

テージングされます。ステージングサーバーは、フィルタリング、マスキング、エンリッチメント、マッピング、重複排除、複数のソースからのデータ統合といった変換ロジックを実行します。

データエンジニアは、バッチデータアップロードの移動をオーケストレーションし、継続的にデータをストリーミングするデータパイプラインを構築します。こうしたパイプラインを用いてアプリケーション、デバイス、およびイベントストリームからデータを抽出します。基本的なETLワークフローにおいては、ETLパイプラインでまずデータをビジネスに利用できる形式に変換します。ビジネス要件が明確に定義されていればこれで問題ありませんが、機械学習やデータサイエンスなど最近増えている新しいタイプのワークロードでは、データフォーマットの要件を事前に把握できない場合も少なくありません。これとは逆に、例えばまずデータを未加工の状態（あるいはなるべく加工されていない状態）で確保してから、後でさまざまな形式に変換し、さまざまなタイプのモデル、予測エンジン、アナリティクスシナリオに対応する方がデータサイエンティストにとっては望ましいかもしれません。

ELTの台頭

従来のETL操作では、多くの場合専用のコンピュータサーバーで動作する別個の処理エンジンを使用します。ETLタイプのデータベースは、後工程のビジネス要件に従い、データを予め定義されている特定の形式に調整してからロードするようモデリングされています。たとえば、データをダッシュボード上で迅速に表示したり、ロールアップして月次の財務レポートに反映させられるよう、データのソート、要約、パラメーター化などを行ってからロードします。

このようなETL手順は、エンタープライズリソースプランニング(ERP)、サプライチェーン管理(SCM)、カスタマーリレーションシップ管理(CRM)システムなどのエンタープライズアプリケーションによる構造化データソースには適していると言えるでしょう。しかし、こうした旧式のETLパイプラインは、モノのインターネット(IoT)システムからの機械生成データ、ソーシャルメディアネットワークからのストリーミングデータ、インターネットウェブサイトからのウェブログデータ、SaaSアプリからのモバイル利用状況データなど、膨大な量の新しい形式のデータには簡単に対応することができません。構造化データやバッチデータの取り込みには適していますが、スキーマレスデータや半構造化データを収集して取り込むには柔軟性に欠けています。

いっぽう最新のデータパイプラインは、新しい形式の大量なデータに対応してタイムリーなアナリティクスを実行できるよう、先にデータを抽出してロードし、目的地に到達してから変換するよう設計されており、そこではじめてデータの標準化、ク

レンジング、マッピング、および他のソースのデータとの結合を行います。これらの新しいELTデータパイプラインは、膨大な量のデータをコスト効率の高い方法で保存し処理できるクラウドデータウェアハウスやクラウドデータプラットフォームを活用しています。

ELTパイプラインは、構造化されたりレガシーデータだけでなく、非構造化データ、半構造化データ、未加工データを取り込み、そのすべてをクラウドデータプラットフォームまたはデータレイクにロードすることができます。データのステージングは必要ありません。データを未加工のまま維持することで、実験や迅速な反復もスムーズに行うことができます。

ELTアプローチのメリット

パワー:

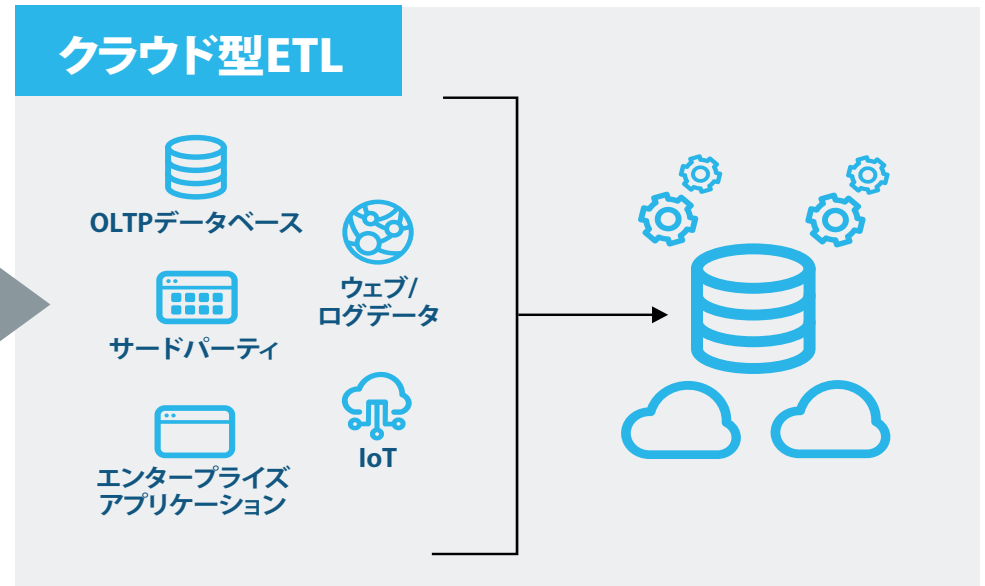
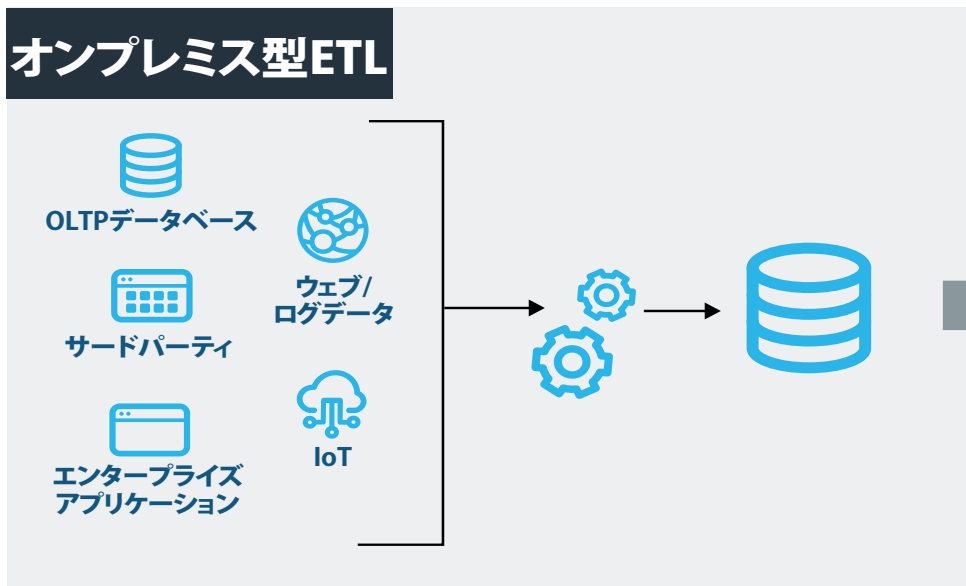
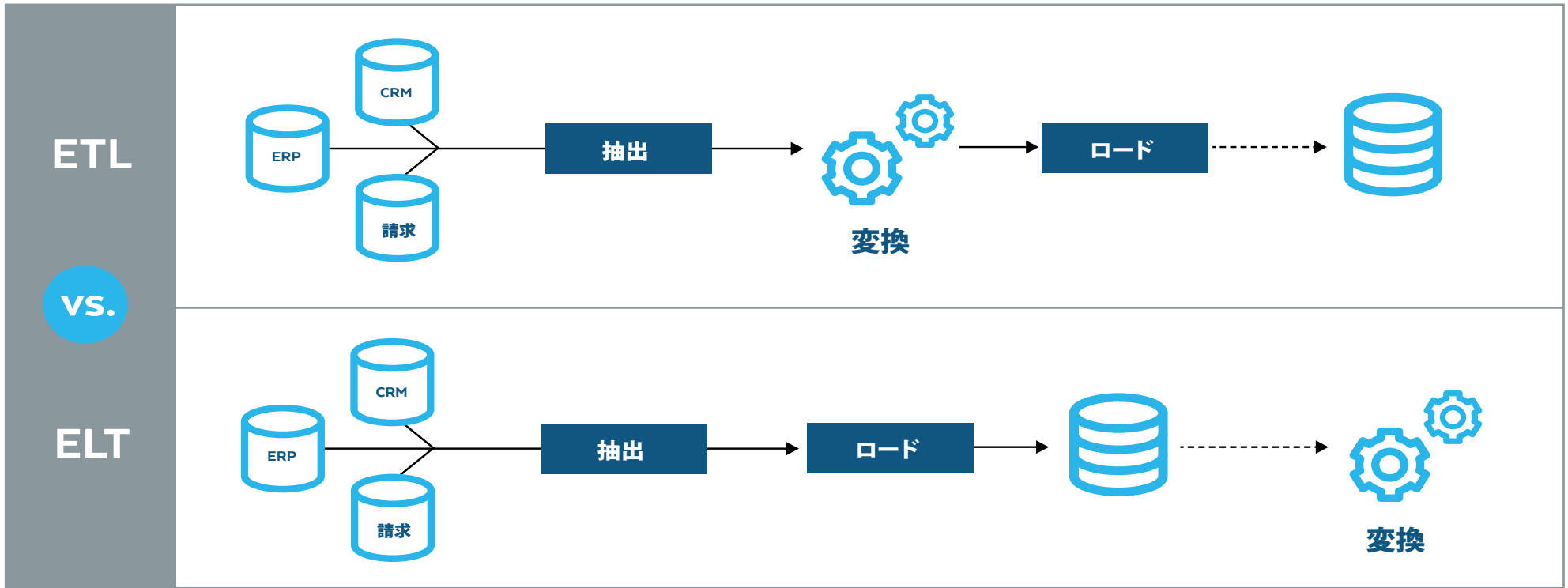
クラウドベースのリポジトリが、拡張性のあるコンピュータサーバーに支えられたほぼ無制限のストレージ機能を提供するため、データ量が増えても対応することができます。

範囲:

ELTパイプラインでは、利用可能なデータはタイプに関わらずそのまま取り込むことができます。データを特定のフォーマットに変換する必要はありません。

柔軟性:

その時々のアナリティクス要件に必要なデータのみを変換することで、複数のチームがレポート、ダッシュボード、データサイエンスモデルなどの様々なタスクに用いるデータを必要に応じて変換できるようになり、データ活用のオプションが最大化されます。



多様性のあるデータ マネジメント戦略の確立

ETLは、データの予測と管理が可能で、更新間隔が定期的である場合には有効なオプションです。一般的にETL型のバッチ取り込みプロセスは、頻繁な更新が不要なアプリケーションデータに対して使用されます。たとえば、日次収益報告向けに小売店のPOSデータを1日の業務終了時にデータウェアハウス内で更新する場合やCRMシステムの顧客データを1時間に1度コールセンターのダッシュボードにアップロードし、その時点の売上およびサービストランザクションを反映する場合などがこれに該当するでしょう。あるいは、従量制課金システムにおいてスマートメーターから収集した電力使用量データを15分ごとに更新するなどのユースケースもあります。

しかし、こうしたやり方には問題が隠れています。ETLシステムや旧式のデータアーキテクチャでは、ばらばらのシステムで生成されるデータがさまざまな場所でサイロ化し、それぞれ個別のニーズに合わせて設計されモデリングされた独自システムに異なるタイプのデータが格納されることとなります。結果として、複数の異なるリポジトリが形成されてしまい、メンテナンスが非常に困難な状況になります。オンプレミスでもクラウド上でも、マーケティングオートメーションシステムのマーケティングデータ、CRMシステムの販売データ、ERPシステムの財務データ、ウェアハウス管理システムの在庫データなどの様々な本番アプリケーションがそれぞれ独自のデータサイロを形成しており、さらに各アプリケーションは本番システムからのデータ収集と分析用途への変換のために、それぞれ専用のETLツールや独自のソフトウェア手順を採用しています。

現在のビジネスおよびアナリティクスのニーズに対応するには、すべてのデータを一つに集約し、さまざまなワークグループ、アプリケーション、ツールからユニバーサルアクセスが可能なシングル・ソース・オブ・トゥールズ（信頼できる唯一の情報源）として機能できるようにするアーキテクチャを構築すべきです。



ELTの実装を検討すべきタイミング

ETL技術は、あるシステムから別のシステムへデータをバッチモードで移動する用途には今後も活発に利用されるでしょう。しかし、ほとんどの旧式のETLソリューションには、扱うことのできないデータタイプが存在します。エンタープライズアプリケーションの構造化データは問題なく扱えますが、例えばIoTシステムの機械生成データ、ソーシャルメディアフィードのストリーミングデータ、JSONイベントデータ、インターネットやモバイルアプリのウェブログデータには適していません。

ETLとELTのどちらを採用するか判断基準としては、以下の基本的ガイドラインを参考にしてください。

- ETLプロセスはテーブル形式の構造を維持する必要があるリレーショナルデータに適しています。
- ELTは、具体的なアナリティクスユースケースが考案されるまで、未加工またはネイティブな状態で維持しなければならない半構造化データに適しています。

また、処理するデータの量や、後工程での分析の準備に必要なスピードなども考慮する必要があります。変換プロセスには、多くのコンピュートサイクルが必要です。ELTでは、自動スケールリングにより、それぞれの動作をサポートするために必要なリソースが即時にプロビジョニングされます。ELT処理により、クラウドの無限のリソースを有効活用して迅速かつ効率的にデータを処理し変換することができます。また、データを独立したサーバーやストレージ機構に移動するのではなく、データが存在する場所で処理できるため、データの移動を最小限に抑えることもできます。

処理エンジンをどこで実行するか、どのようなインフラストラクチャリソースを利用できるか、どのような性能が必要となるかも検討しましょう。サーバー容量の制約など、拡張性あるいは同時実行性についても問題を抱えていませんか？オンライントランザクションプロセス (OLTP) システム、ウェブサイトのやり取り、SaaSアプリケーション、機器のセンサー、ソーシャルメディアのストリームなど、データの生成元がどのようなものであっても、データエンジニアはそのデータを取得し、データリポジトリに取り込み、ビジネスコミュニティがアクセスできるようにするデータパイプラインを開発しなければなりません。多くの場合、データパイプラインの運用は、クラウド内のターゲットデータベースの処理能力を活用することで強化されます。

データパイプラインの選択

ETLが適しているケース：

- 処理するデータの総容量が比較的小さい場合
- ソースデータベースとターゲットデータベースで必要なデータタイプが異なる場合
- 主に処理しているのが構造化データである場合

ELTが適しているケース：

- 処理すべきデータが大量にある場合
- ソースデータベースとターゲットデータベースが同じタイプである場合
- データが半構造化もしくは非構造化データである場合

Snowflakeユーザーのケーススタディ

組織: Paciolanは、チケット販売、資金調達、マーケティング、アナリティクス、技術ソリューションの大手事業者であり、500以上のライブエンタテインメント企業をサポートし年間1億2000万枚以上のチケットを販売しています。

問題点: Paciolanは、独自ETLコードによるデータの構文解析と正規化により半構造化データをリレーショナルデータに変換していました。しかしこれでは、例えば5万件のレコードを変換すると1つのオンプレミスデータウェアハウス内で最大100万行が生成されてしまい、毎日100GB近いデータを扱うETL処理の実行に30分から60分かかっていました。またリソースが限定さ

れているため、アナリストが効果的にデータを要約してロールアップできない状況となっていました。

解決策: 現在、Paciolanは半構造化JSONデータをVariant型としてSnowflakeプラットフォームに格納しています。Snowflakeを、データポルトベースのデータレイク兼データウェアハウスとして利用しています。データポルトは、現代のアジャイルなエンタープライズデータウェアハウスのサポートを目的とした、特定のデータモデル設計パターンを含むアーキテクチャアプローチです。

結果: Snowflakeデータパイプラインの実装により、従来型のデータウェアハウスでは1時間ほどかかっていたETLプロセスをわずか数分で完了できるようになりました。開発者は、シンプルなPythonスクリプトを使用してステートメントを動的に挿入することができます。

メリット

- コンピューティングとストレージの分離により性能が安定し、コストが可視化される
- リアルタイムな伸縮性により、ほぼ無制限のコンピューティングパワーをユーザー数の制限なく提供できる
- Variant型での半構造化データの保存に対応することでより豊かなデータインサイトが得られる

「導入前と導入後の数値を比較したところ、SnowflakeではETLプロセスに使用するコードが90%削減されました。これは私たちにとって大きな収穫です。」

Ashkan Khoshcheshmi
主任ソフトウェアエンジニア
Paciolan

Snowflakeによるデータの処理

Snowflakeプラットフォームには、柔軟性と拡張性を備えたデータパイプライン機能が基本サービスの一環として含まれています。未加工データを直接Snowflakeに取り込むことができるため、データを別のフォーマットに変換するためのパイプラインを作成する必要がありません。Snowflakeではこれらの変換が自動的に実行されるため、ストレージコスト並びにコンピューティングコストが最小限に抑えられます。

また、Snowflakeでは、データサイロを解消することによりデータ管理もシンプル化されます。下流にある複数のアプリケーション用としてデータの複数のコピーを維持する必要はありません。未加工データの元の形状を維持しながら高度に最適化されたストレージ技術を透過的に適用することで、アナリティクスおよびデータ変換の性能が大幅に向上します。

何より重要なのは、Snowflakeがクラウドならではの特性を最大限に活用できるよう設計されているという点です。コンピュートリソースとストレージリソースを分離するマルチクラスターおよび共有データアーキテクチャをベースとすることで、大規模なデータ変換にも対応できます。各タイプのリソースは、それぞれのアプリケーションの特定のニーズに合わせて独立して拡張可能です。

Snowflakeのプラットフォームは、堅牢な処理エンジンを中心に構築されています。この処理エンジンは、データエンジニアリングパイプライン経由でデータを取り込みながら、同時に同じデータを使って機械学習モデルをトレーニングするなど、あらゆるタイプのワークロードを性能を損なわずに処理できるよう

設計されています。拡張性のあるこのパイプラインサービスでは、他のワークロードの処理に影響を与えることなく継続的にデータを取り込むことができます。データエンジニアは、それぞれのデータ取り込みプロセスに割り当てるコンピューティングパワーを決定したり、システムの自動的な拡張を許可したりすることができます。

また、Snowflakeでは、データエンジニアが取り込みストリームを管理するための言語や統合ツールを幅広い選択肢の中から選んだ上で、データパイプラインを構築することもできます。バッチ統合やApache Kafkaによるストリーミング統合など、よく利用されるデータ取り込みのスタイルに幅広く対応しています。さらにデータ処理の共通言語である標準SQLも使用するため、さまざまなタイプのデータを簡単かつ効率的に取り込むことができます。



結論

データの量や種類が増え高速化も進んでいることから、新しいタイプのデータパイプラインと、データを取り込んで利用可能な状態にするより高度なクラウドベースデータ処理エンジンが必要とされるようになっていきます。

ETLプロセスは多くの場合オンプレミスサーバーで処理されますが、容量が固定されており、帯域幅やCPUサイクルにも制限があります。最新のデータ統合ワークロードは、自在に拡張できるクラウドデータベースおよびクラウドデータプラットフォームを活用した処理によりその価値をさらに高めることができます。

このようなクラウドリソースを活用するために、データを抽出してクラウドデータベースにロードし、その後にデータを変換する、いわゆるELTサイクル型データパイプラインを設計する企業が増えています。このアプローチでは、最新のデータ処理エンジンのパワーを活用し、不必要なデータの移動を排除することで、従来のETLプロセスよりも短時間での処理が可能となります。

ELTプロセスでリソースを大量に使用する変換ワークロードをクラウドにプッシュする主な理由は次の2点です。

1. クラウドのほぼ無限のリソースを利用して迅速かつ効率的にデータを処理し変換できるため
2. ビジネス要件を十分に把握できるまでデータを未加工の状態のまま保持できるため

結論として、リソースを大量に消費する変換ジョブは、できるだけETLではなくELTを使用してクラウドベースのターゲットプラットフォームにプッシュすると良いでしょう。このアプローチにより、データパイプラインがシンプル化されデータの移動も最小限に抑えられるほか、データサイロの削減や、データの最終的な使用方法としてのオプションの最大化も実現します。

Snowflakeのデータパイプラインソリューションの詳細については、snowflake.com/workloads/data-engineeringをご覧ください。





Snowflakeについて

Snowflakeは、Snowflakeのデータクラウドによってあらゆる組織が自らのデータを活用できるようにします。顧客企業は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、さまざまな分析ワークロードを実行しています。データやユーザーがどこに存在するかに関係なく、Snowflakeは複数のクラウドと地域にまたがり単一のデータ体験を提供します。多くの業界から何千ものお客様(2022年1月31日時点で、2021年のFortune 500社のうち241社、2021年のForbes Global 2000 (G2K)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。本書で言及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おり、各所有者の商標である可能性があります。Snowflakeが、必ずしもかかる商標所有者と関係を持ち、または出資や支援を受けているわけではありません。