で成長とスケーリングのためのデータレイクの設計 AWS クラウド

AWS 規範ガイダンス

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

AWS 規範ガイダンス: で成長とスケーリングのためのデータレイクの設計 AWS クラウド

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon の商標およびトレードドレスは Amazon 以外の製品およびサービスに使用することはできません。また、お客様に誤解を与える可能性がある形式で、または Amazon の信用を損なう形式で使用することもできません。Amazon が所有していないその他のすべての商標は Amazon との提携、関連、支援関係の有無にかかわらず、それら該当する所有者の資産です。

Table of Contents

序章 1
ターゲットを絞ったビジネス成果 2
一般的なスケーリングの課題 3
データレイクの初期デプロイ 3
データコンシューマーの増加 4
データプロデューサーの増加 5
リファレンスアーキテクチャ 7
リファレンスアーキテクチャコンポーネント 9
データプロデューサー 10
一元化されたカタログ 12
AWS Lake Formation 13
AWS Glue 14
データコンシューマー 15
オンボーディングとアクセス許可の付与 18
データプロデューサーのオンボーディング 18
データコンシューマーのオンボーディング 19
データコンシューマーアカウントの Select アクセスを付与する 20
よくある質問 22
このデータレイクリファレンスアーキテクチャはエンタープライズ組織により適していますか? 22
組織にデータプロデューサーが 1 つしかない場合でも、このリファレンスアーキテクチャを使用できますか? 22
データレイクは、1 つのデータプロデューサーを複数のデータコンシューマーに直接接続します。このガイドのデータレイクリファレンスアーキテクチャは引き続き重要ですか? 22
組織は、変更を加えずにオンボーディングとアクセス許可のワークフローに従うべきですか? 23
リソース 24
ドキュメント履歴 25
..... xxvi

で成長とスケーリングのためのデータレイクの設計 AWS クラウド

Wei Shao、Amazon Web Services (AWS)

2021 年 10 月 (ドキュメント履歴)

組織は、データモダナイゼーション戦略の一環として、アマゾン ウェブ サービス (AWS) クラウド上のデータレイクの設計と構築を進めています。データレイクは、構造化データと非構造化データをあらゆる規模で保存し、さまざまな内部ユーザーと外部ユーザーが利用できるリポジトリです。

ただし、増大するビジネスユースケースに対応するデータレイクにデータを取り込むには、かなりの時間と労力がかかる場合があります。コストを削減し、データから生成される価値を最大化するために、多くの組織はデータレイクへのデータの 1 回限りの取り込みを計画し、このデータを複数回消費します。データレイクが大きくなったときにビジネスステークホルダーに価値を提供するには、データの生産、共有、消費に合わせてスケールするデータレイクアーキテクチャを設計することが重要です。

スケーラブルなデータレイクアーキテクチャは、データレイクからより多くのデータを取り込むと同時に価値を得るための強固な基盤を組織に提供します。スケーラビリティの制約により、速度が低下したり中断されたりすることなくデータインサイトを継続的に取得することで、スケーラブルなデータレイクは組織の競争力を維持するのにも役立ちます。

通常、データレイクにはデータプロデューサーとデータコンシューマーがあります。データプロデューサーは、データドメインからデータを収集、処理、保存することでデータアセットを作成します。これらの集合データアセットは、データレイクのコンテンツを形成します。データプロデューサーは、データアセットをデータレイクのデータコンシューマーと選択的に共有することを選択できます。

データコンシューマーは、ビジネスユースケースを達成するためにデータプロデューサーからのデータを必要とし、このデータを独自のデータと組み合わせることもあります。データプロデューサーとデータコンシューマーは通常、組織の一部ですが、必ずしもそうではありません。重要なのは、データプロデューサーまたはデータコンシューマーの両方を同時に指定できることです。

スケーラブルなデータレイクアーキテクチャは、次の成果を達成するのに役立ちます。

- データプロデューサーは、データ共有プロセス全体を維持する必要なく、大規模にオンボードします。これにより、データプロデューサーはデータをデータレイクにオンボードし、データドメインからのデータの収集、処理、保存に集中できます。

- データコンシューマーは、全体的なコストと管理オーバーヘッドを増やすことなく、複数のデータプロデューサーのデータにアクセスできます。

このガイドでは、組織がデータレイクを成長させるときに発生する可能性がある一般的なスケーリングの課題について説明し、データレイクリファレンスアーキテクチャを提供し、データプロデューサーとデータコンシューマーをオンボーディングしてアクセス権を付与するためのアプローチの概要を説明します。このガイドのデータレイクリファレンスアーキテクチャは、 が提供するさまざまな機能を活用していますAWS Lake Formation。このガイドは、エンタープライズデータアーキテクト AWS クラウド、データプラットフォームアーキテクト、デザイナー、データドメインリードなど、のデータレイクの設計を担当するチームを対象としています。

ターゲットを絞ったビジネス成果

で成長とスケールのためのデータレイクを設計した後、次の 3 つの結果が期待されます AWS クラウド。

- 組織内の複数の事業部門におけるデータ共有とデータ消費のオーバーヘッドを軽減します。
- 組織が外部データプロデューサーを含め、データレイク内でデータを共有するのに役立つ、安全で一貫性のあるアプローチ。
- スケーラビリティの制約により、速度が低下したり中断されたりすることなく、データインサイトを継続的に取得します。

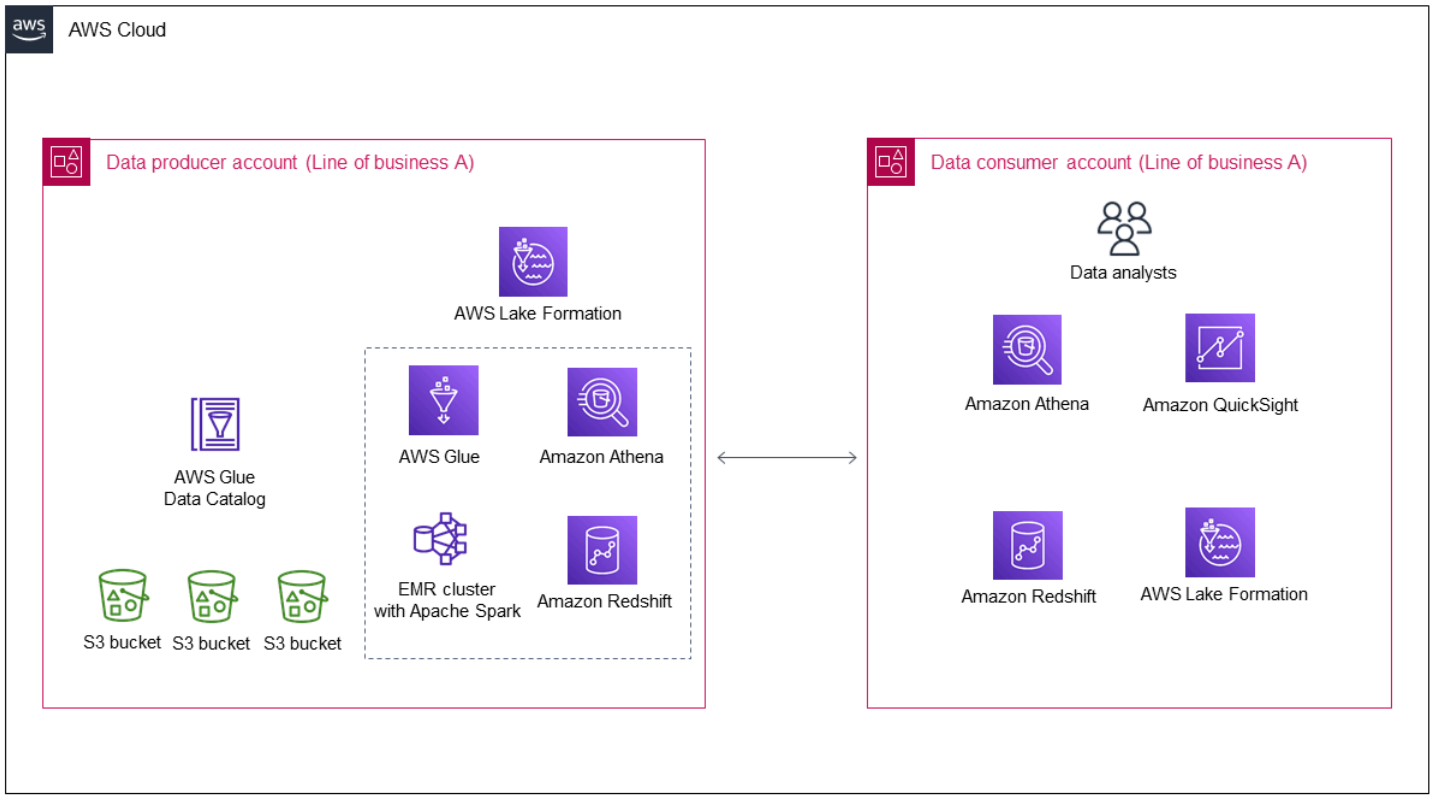
一般的なスケーリングの課題

データレイクは、最初のデプロイ後にデータが大きくなると、いくつかの段階を経ます。スケーラブルアーキテクチャを使用してデータレイクを設計しなかった場合、組織は課題に遭遇し、データレイクの成長によって不利になる可能性があります。

以下のセクションでは、一般的なデータレイクの成長がスケーリングの課題を引き起こす方法について説明します。

データレイクの初期デプロイ

次の図は、基幹業務 A による最初のデプロイ後のデータレイクのアーキテクチャを示しています。



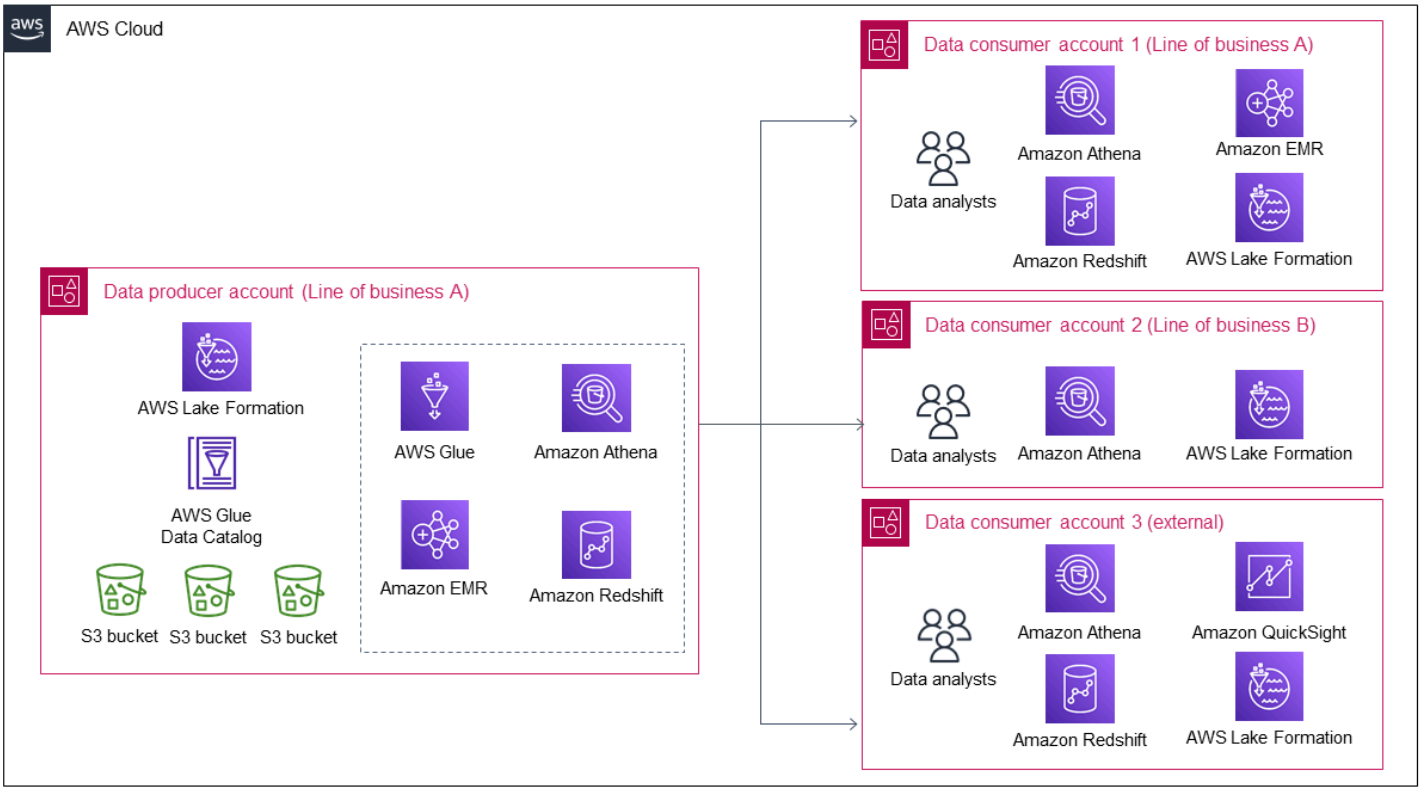
この図は、次のコンポーネントを示しています。

- データプロデューサーアカウントは、データを収集して処理し、処理されたデータを保存して、消費用に準備します。
- データプロデューサーアカウントのデータは、複数のデータレイヤーを持つことができる Amazon Simple Storage Service (Amazon S3) バケットに保存されます。
- データ処理に AWS サービス (AWS Glueや Amazon EMR など) を使用できます。

- データプロデューサーは、データレイクでデータを生成して保存するだけでなく、データコンシューマーと共有するデータとその共有方法を決定する必要があります。 は、データプロデューサーからデータコンシューマーへのクロスアカウントデータ共有を管理するだけでなく、データプロデューサーアカウントでデータレイク AWS Lake Formation を管理します。
- データコンシューマーアカウントは、特定のビジネスユースケースのためにデータプロデューサーアカウントの共有データを消費します。

データコンシューマーの増加

次の図は、基幹業務 A のデータが増えると、より多くのデータがデータレイクに取り込まれることを示しています。その後、データレイクはより多くのデータコンシューマーを惹きつけ、データを活用してデータから価値を得ます。



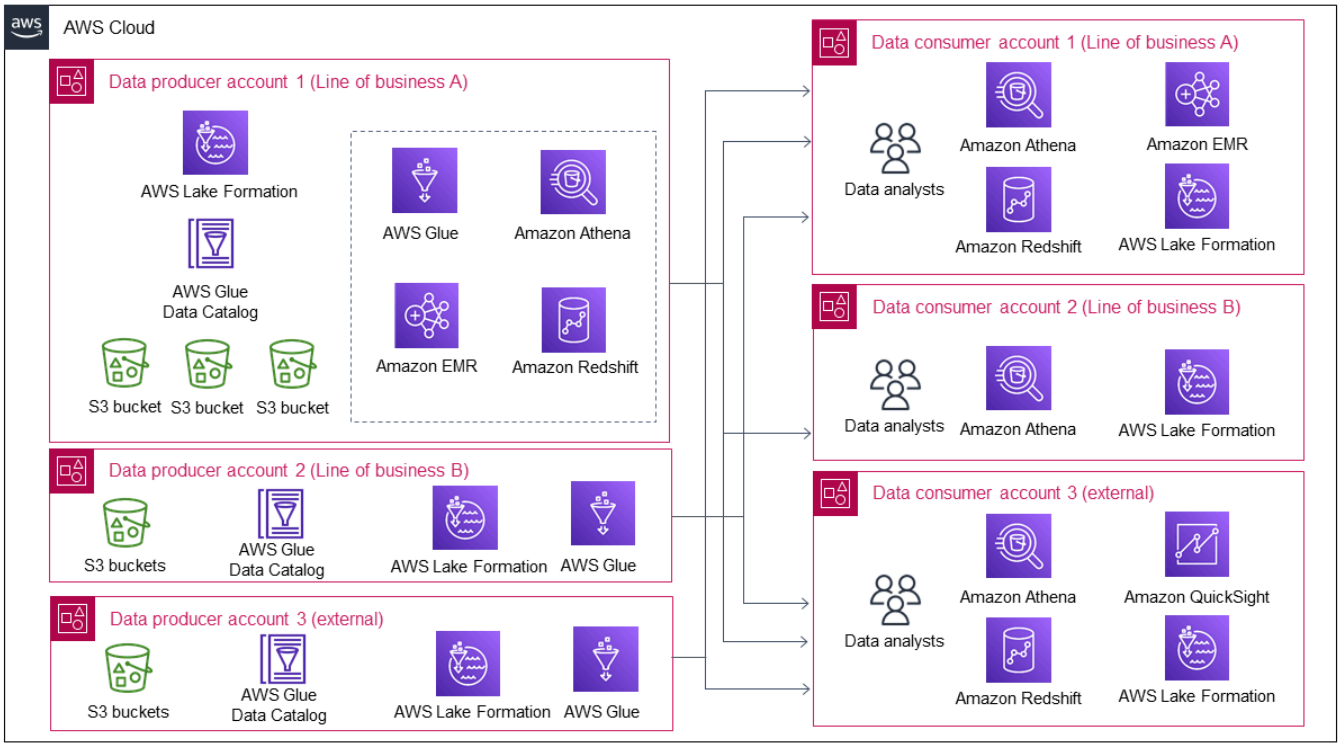
この図は、組織が既存のデータアセットからほぼ継続的な値を生成する方法と、より多くのデータコンシューマーを引き付ける方法を示しています。ただし、データコンシューマーが増加すると、データプロデューサーにはこの増加に対応するための次の 2 つのオプションしかありません。

- 個々のデータコンシューマーによるデータ共有とアクセスを手動で管理します。これはスケーラブルなアプローチではありません。

- データ共有とデータアクセスの管理のための自動または半自動プロセスを開発します。これはスケーラブルなオプションかもしれませんが、内部データコンシューマーと外部データコンシューマーのセキュリティ管理要件が異なるため、設計と構築には多大な時間と労力が必要です。今後、ソリューションの改善には追加の時間と労力も必要になります。

データプロデューサーの増加

次の図は、複数の事業部門がデータプロデューサーとして参加する場合のデータレイクアーキテクチャを示しています。



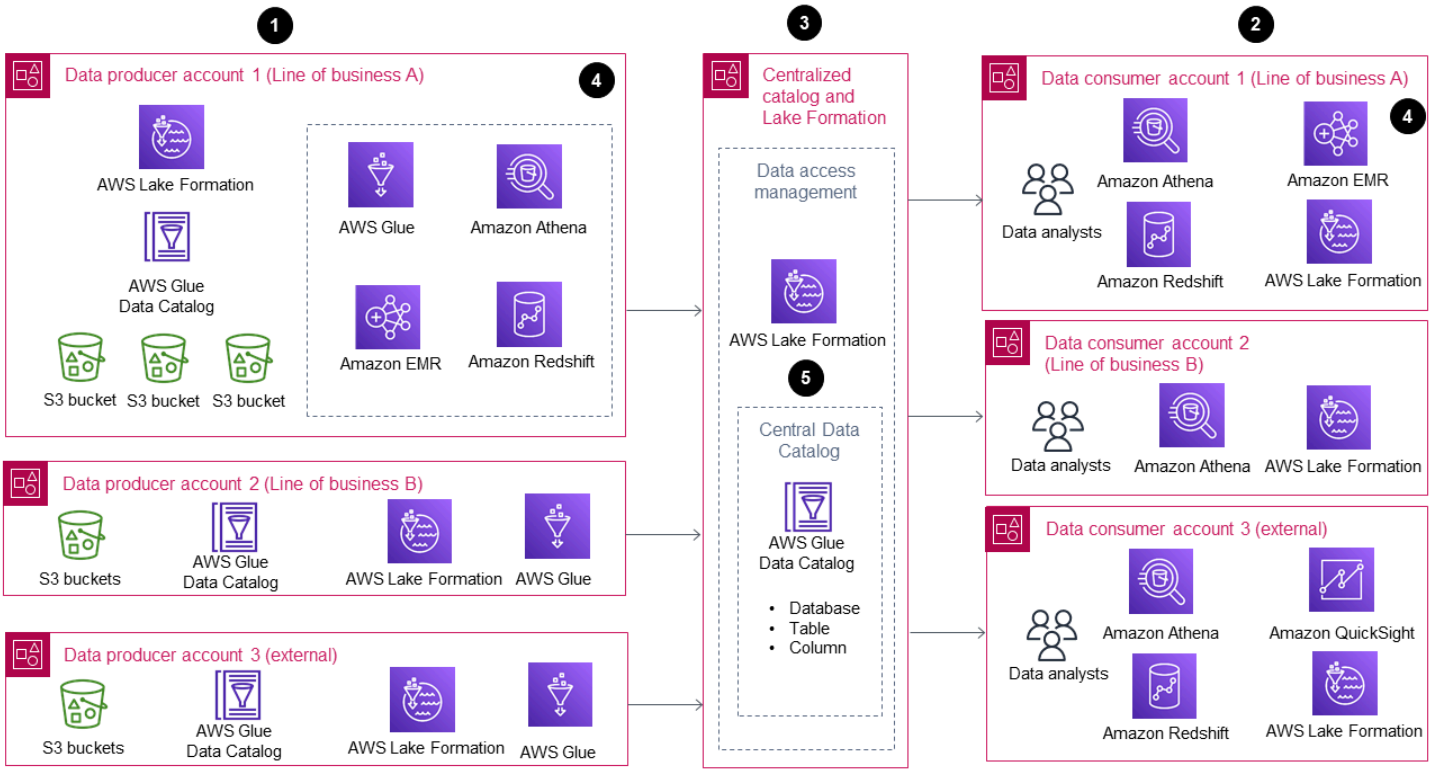
データレイクのアーキテクチャは、3 つのデータプロデューサーと 3 つのデータコンシューマーだけでも、ますます複雑になります。

各データプロデューサーは、複数のデータコンシューマーのデータ共有とデータアクセス管理を処理する必要があります。すべてのデータプロデューサーがデータ共有とデータアクセス管理のための自動または半自動プロセスを開発することは現実的ではありません。一部のデータプロデューサーは、データを共有しないことを選択するため、コストがかからない管理オーバーヘッドを回避できます。同様に、各データコンシューマーは、さまざまなデータ消費プロセスを理解するために、複数のデータプロデューサーとやり取りする必要があります。つまり、個々のデータコンシューマーは、さまざまなデータ共有パターンを処理するための管理オーバーヘッドが増加しています。

多くの組織では、このデータレイクはボトルネックを引き起こし、成長やスケールができません。これは、ボトルネックを取り除くために組織がデータレイクを再設計および再構築する必要があることを意味します。これにより、多大な時間、リソース、コストがかかる可能性があります。

リファレンスアーキテクチャ

次の図は、 でデータレイクを成長およびスケーリングするためのこのガイドのリファレンスアーキテクチャを示しています AWS クラウド。



この図は、次のコンポーネントを示しています。

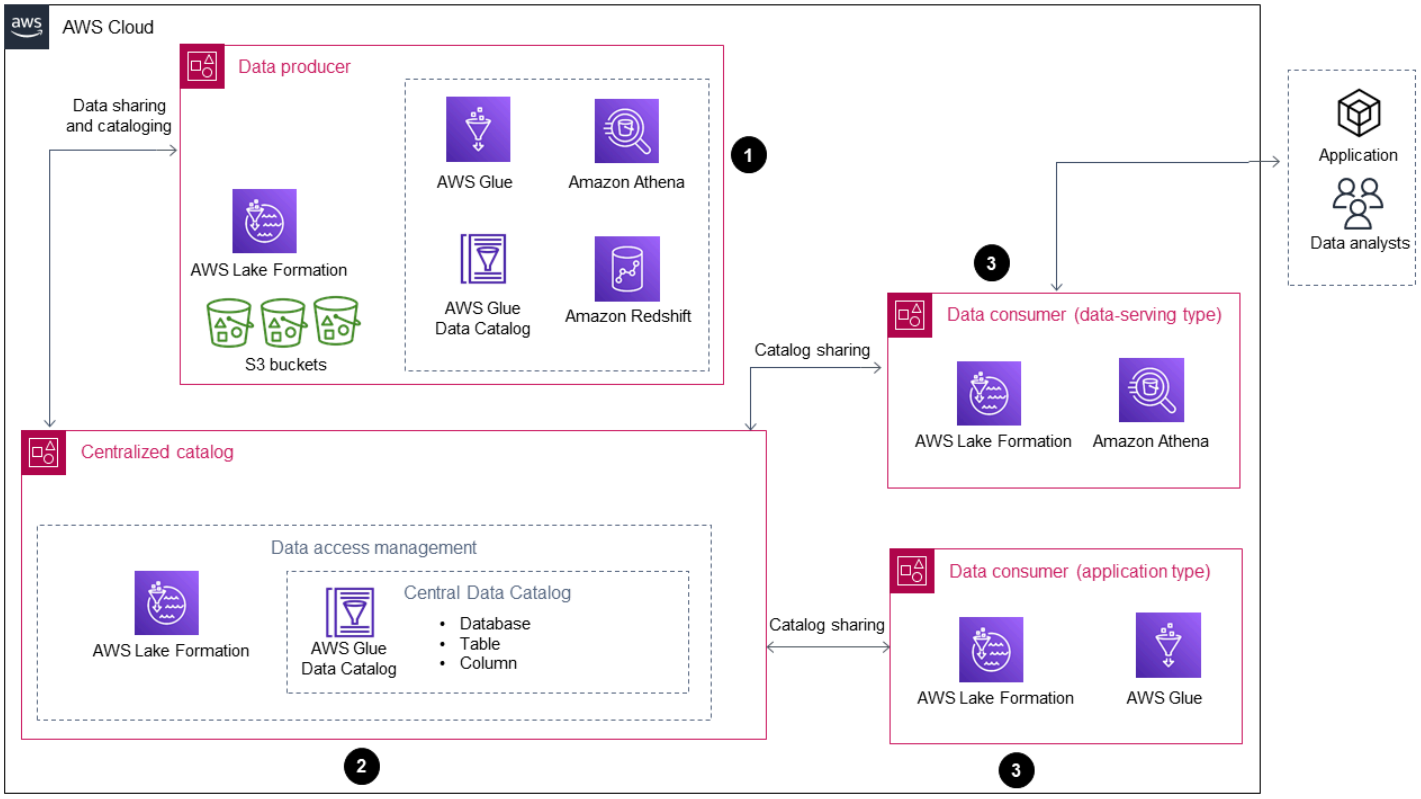
1. 別の のデータプロデューサーレイヤー AWS アカウント。
2. 異なる のデータコンシューマーレイヤー AWS アカウント。
3. 内の一元化されたカタログ AWS アカウント。
4. 各事業部門には 1 つのデータプロデューサーと 1 つのデータコンシューマーしかありませんが、ガイドのリファレンスアーキテクチャは、事業部門ごとに複数のデータプロデューサーとデータコンシューマーをサポートしています。一般的に、1 つのデータプロデューサーに、データ供給タイプとアプリケーションタイプの両方を含む 1 つ以上のデータコンシューマーをオンボードします。詳細については、このガイドの リファレンスアーキテクチャコンポーネント セクションを参照してください。
5. 一元化されたカタログは、データプロデューサーとデータコンシューマーがデータを共有および消費するために使用するインターフェイスです。

リファレンスアーキテクチャのアプローチにより、データ共有と消費を標準化し、管理オーバーヘッドを増やすことなく、データプロデューサーとデータコンシューマーを個別にスケーリングできます。リファレンスアーキテクチャは、さまざまなデータプロデューサーにまたがるデータの本番稼働と配布も可能にします。すべてのデータプロデューサーは、データレイクの一部となり、データを共有し、データレイクが提供する全体的な価値に貢献できます。

このアプローチにより、組織は 1 つのパイプラインでデータ収集と処理を制限することでボトルネックを引き起こすことなく、事業部門と外部データ所有者全体でデータ価値を収集できます。

リファレンスアーキテクチャコンポーネント

次の図は、このガイドのリファレンスアーキテクチャの 3 つのコンポーネントであるデータプロデューサー、データコンシューマー、集中型カタログを示しています。



この図は、次のコンポーネントを示しています。

1. データプロデューサーアカウント。
2. 一元化されたカタログ。
3. データサービスタイプとアプリケーションタイプを含む 2 つのデータコンシューマーアカウント。

> **Note**
>
> この図は 1 つのデータプロデューサーと 2 つのデータコンシューマーのみを示していますが、このガイドのリファレンスアーキテクチャは、データレイクの拡大時に複数のデータプロデューサーとデータコンシューマーをサポートしています。

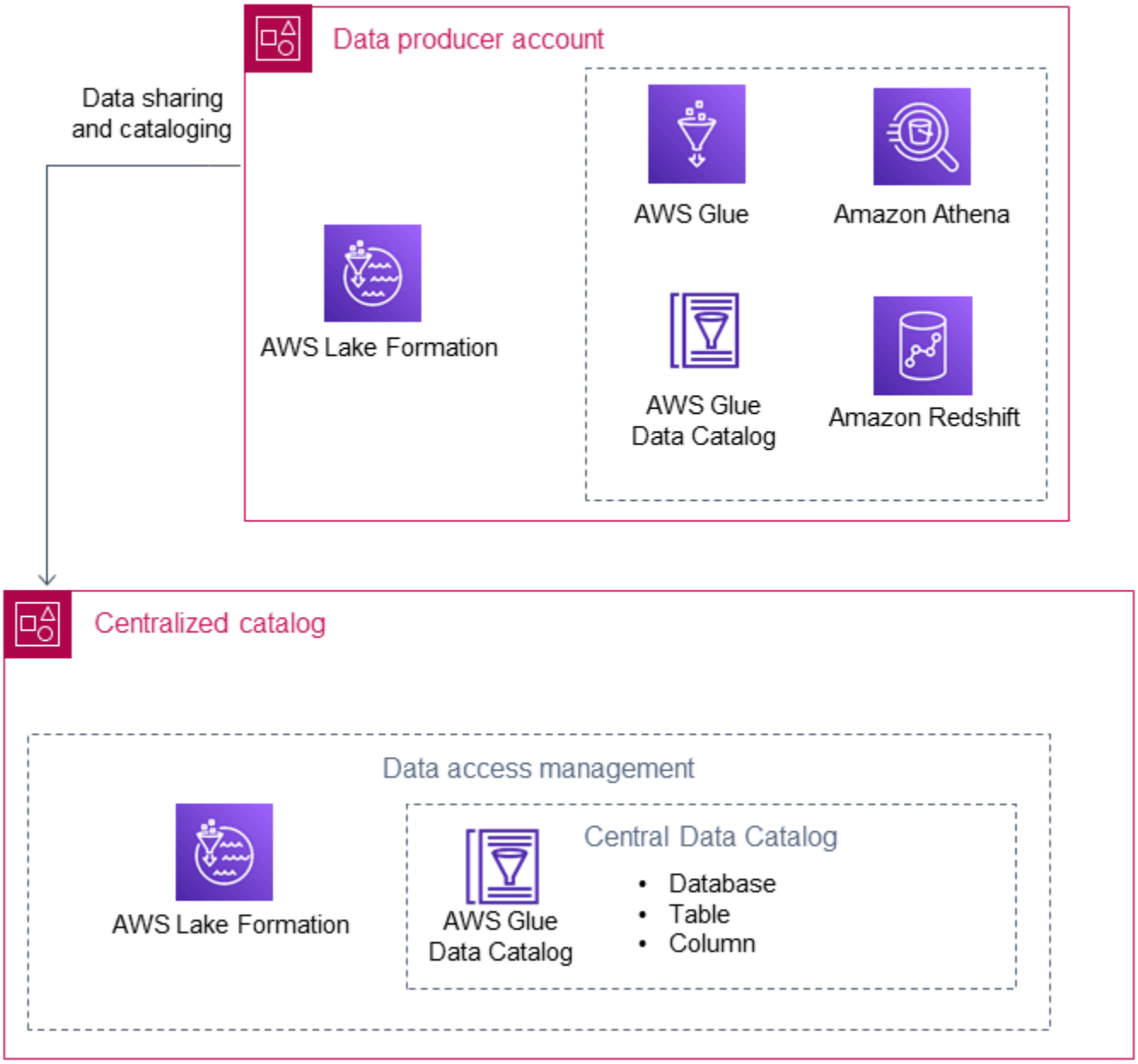
各コンポーネントを理解することで、組織は新しいデータプロデューサーとデータコンシューマーを独立して一貫してオンボーディングすることで、再利用可能なパターンを実装し、データレイクを拡張できます。

以下のセクションでは、リファレンスアーキテクチャのコンポーネントについて説明します。

- データプロデューサー
- 一元化されたカタログ
- データコンシューマー

データプロデューサー

データプロデューサーは、データアセットの品質のモニタリングと確保に加えて、データドメインからデータを収集、処理、保存します。次の図は、このガイドのリファレンスアーキテクチャのコンポーネントとしての データプロデューサーアカウントを示しています。



各データプロデューサーには、内部データ処理で使用されるプライベートデータカタログが AWS Lake Formation AWS アカウント で管理されています。データプロデューサーは、一元化されたカタログにデータに対する選択的なアクセス許可を提供します。つまり、一元化されたカタログアカウントの Lake Formation は、データプロデューサーが共有したいデータにアクセスできます。

つまり、データプロデューサーはデータコンシューマーと直接やり取りしません。代わりに、データプロデューサーアカウントとそのデータストレージの場所は、データコンシューマーから完全に抽象化され、非表示になります。このアプローチは、データコンシューマーの増加を経験するデータプロデューサーの不要なオーバーヘッドを排除することでコストを削減します。

新しいデータの場所が一元化されたカタログによって登録されている場合、データプロデューサーのデータの場所を変更しても、データコンシューマーには影響しません。データプロデューサーが特定のデータの共有を停止する場合は、一元化されたカタログのアクセス許可を削除できます。これにより、データコンシューマーはデータにアクセスできなくなり、各データコンシューマーのアクセスを手動で取り消す必要がなくなります。

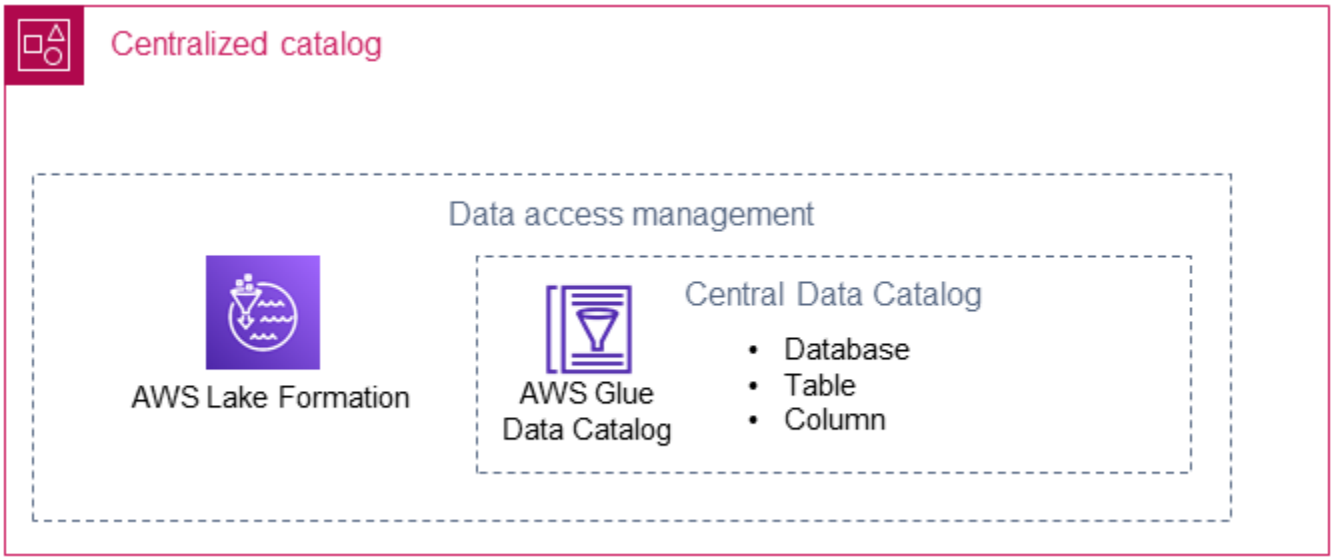
パブリックデータカタログとプライベートデータカタログを使用することで、データプロデューサーは、プライベートデータカタログを介して内部データアクセスを個別に管理しながら、データコンシューマーと共有するものを選択できます。

次の表は、データプロデューサーが一元化されたカタログとデータを共有するために使用する 2 つの AWS サービスを示しています。

Amazon Simple Storage Service (Amazon S3)	S3 バケットのバケットポリシーを調整して、一元化されたカタログ内の AWS Identity and Access Management (IAM) ロールへのデータアクセスを提供します。 データプロデューサーは、Amazon S3 を中間データレイヤーとして使用して、他のデータアプリケーションまたはサービスに保存されているデータを共有することもできます。
AWS Key Management Service (AWS KMS)	集中型カタログ内の IAM ロールと、データプロデューサーアカウント内の共有 Amazon S3 データを暗号化する AWS KMS keys ために使用される に、 AWS マネージドキーのアクセス許可を付与します。

一元化されたカタログ

次の図は、一元化されたカタログがデータレイク内のデータプロデューサーとデータコンシューマーを接続する方法を示しています。



一元化されたカタログは、データプロデューサーアカウントの共有データカタログを保存および管理します。一元化されたカタログは、共有データの技術メタデータ (テーブル名やスキーマなど) もホストし、データコンシューマーがデータにアクセスする場所でもあります。

データコンシューマーは、一元化されたカタログ内の複数のデータプロデューサーのデータにアクセスし、このデータを独自のデータと組み合わせてさらに処理できます。一元化されたカタログを使用すると、データコンシューマーがさまざまなデータプロデューサーに直接接続する必要がなくなり、運用上のオーバーヘッドが軽減されます。

一元化されたカタログは、データプロデューサーとコンシューマーによるデータ共有とデータ消費を可視化できるため、一元化されたデータガバナンス機能 (アクセス監査など) を適用するのに最適な場所です。

以下のセクションでは、一元化されたカタログが AWS Lake Formation と を使用する方法について説明します AWS Glue。

AWS Lake Formation

AWS Lake Formation は、 AWS Glue データレイク内の複数のデータプロデューサーの場所を指すデータベースを データカタログに作成するために役立ちます。Lake Formation の AWS Identity and Access Management (IAM) ロールは、一元化されたカタログに作成されます。Lake Formation を使用すると、一元化されたカタログはデータリソース (データベース、テーブル、列など) をデータコンシューマーと選択的に共有できます。Lake Formation マネージドリソースは、次の 2 つの方法のいずれかを使用してデータコンシューマーと共有されます。

- 名前付きリソースメソッド – このメソッドは、アカウント間でマネージドリソースを共有します。データベース、テーブル、または列名を指定する必要があり、リソースを組織、組織単位 (OU)、または と共有できます AWS アカウント。共有と管理のオーバーヘッドを減らすには、可能な限り高いレベル (たとえば、 ではなく組織または OU AWS アカウント) でリソースを共有することをお勧めします。ただし、このアプローチが組織のデータセキュリティコントロール要件を満たしていることを確認する必要があります。
  - 注: この方法は、 AWS サービスがデータプロデューサーからデータを消費するアプリケーションタイプのデータコンシューマーに適しています。このタイプのデータコンシューマーからのデータアクセス要件は、アプリケーション駆動型、規範的、比較的静的です。
- Lake Formation タグベースのアクセスコントロール (LF-TBAC) メソッド – LF-TBAC は、データ供給タイプのデータコンシューマーに特に役立ちます。ただし、Lake Formation のタグ付けされたリソースは現在、組織 AWS アカウント レベルまたは OU レベルでのみ共有できます。

AWS Glue

一元化されたカタログ内のデータプロデューサー AWS Glue ごとに、 にデータベースを作成する必要があります。一元化されたカタログは AWS Glue を使用してすべてのデータプロデューサーからデータベースをホストするため、データベース名がすべてのデータプロデューサーで一意であり、データプロデューサーとそのデータタイプを反映していることを確認する必要があります。たとえば、次のデータベース命名構造を使用できます。 `<Data_Producer>-<Environment>-<Data_Group>`

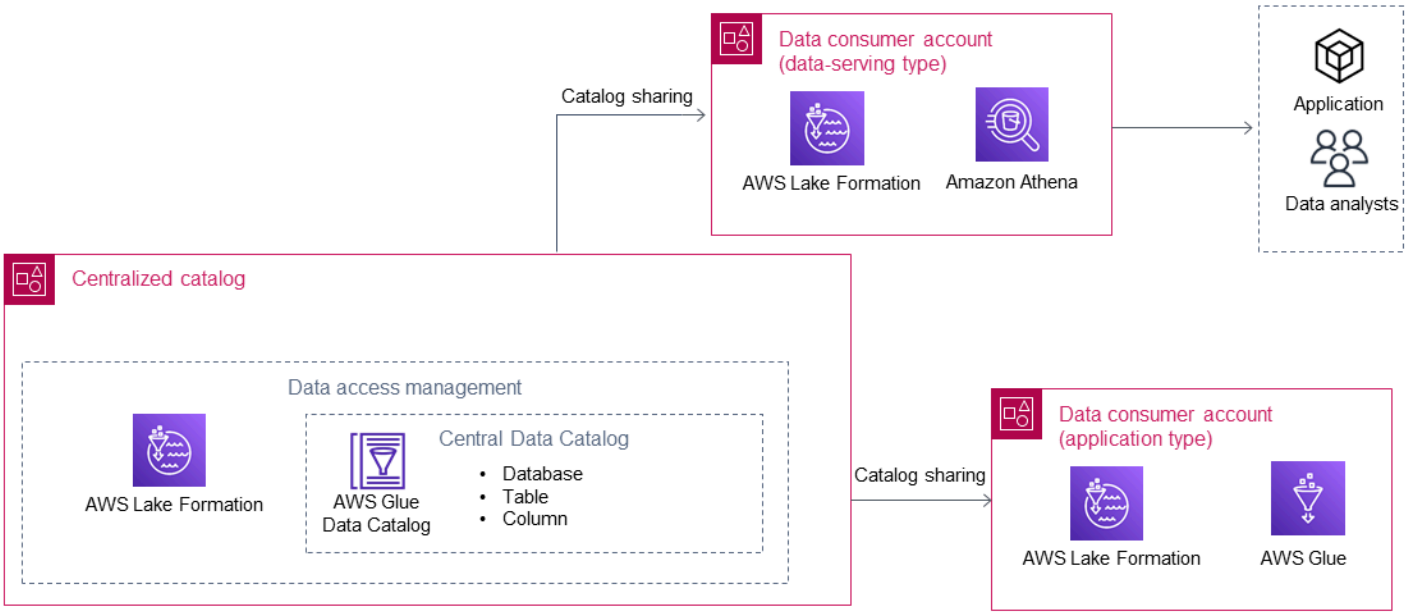
- `<Data_Producer>` – データプロデューサーの名前。
- `<Environment>` – `dev`開発環境、`sit`システム統合テスト環境、`prod`本番環境などのデータレイク環境。
- `<Data_Group>` – データプロデューサーから論理グループにデータを分離するために使用されるデータグループの名前。ソースシステム名、ID、または略語を名前として使用できます。データベースの説明は、データベースの内容と目的を記述するのに役立ちます。

データプロデューサーのデータで AWS Glue クローラを使用して、一元化されたカタログのデータベースにスキーマを維持できます。データプロデューサーが同じ頻度で定期的にデータを作成する場合は、1 AWS Glue つのクローラを使用できます。それ以外の場合は、複数の AWS Glue クローラを使用して、異なるクローリング頻度に対応する必要があります。ビジネスユースケースに応じて、クローラは事前定義された頻度でスケジュールすることも、イベントによって開始することもできます。

API を呼び出し AWS Glue てスキーマを作成または更新 AWS Glue することで、 でテーブルスキーマを維持することもできます。これにより柔軟性が得られますが、コードの開発とメンテナンスには追加の労力が必要です。ユースケースとビジネス価値を評価し、要件を満たし、オーバーヘッドが最も少ないオプションを選択します。

データコンシューマー

データコンシューマーは、一元化されたカタログがデータプロデューサーを使用して共有した後、データプロデューサーからデータを消費します AWS Lake Formation。次の図は、データレイク内の 2 つのデータコンシューマーを示しています。



データコンシューマーには、アプリケーションとデータ供給の 2 種類があります。次の表に、これら 2 つのタイプを示します。

アプリケーションタイプ	アプリケーションデータコンシューマーは、独自のアプリケーションを実行します AWS アカウント。アプリケーションは AWS Identity and Access Management (IAM) ロールを使用してデータプロデューサーから共有データにアクセスし、ロジックに従って処理します。

	通常、このタイプのデータコンシューマーには、アプリケーションのニーズを満たすための規範的なデータ要件があります。
データ提供タイプ	データ供給データコンシューマーは通常、個人 (データアナリストやデータサイエンティストなど) と、独自のものを持たないアプリケーション (ビジネスインテリジェンスアプリケーションなど) を対象としています AWS アカウント。 1 つの組織のデータレイクに複数のデータ供給データコンシューマーが存在する可能性があります。たとえば、さまざまな事業部門が、ユーザーがデータレイクからデータを消費できるように、独自のデータ供給データコンシューマーを設定することを選択する場合があります。これらのデータコンシューマーには、 AWS サービス AWS アカウント (Amazon Athena などAWS IAM アイデンティティセンター) を介して共有データにアクセスするためにデータコンシューマーアカウントのエンドユーザーが使用する独自の IAM ロールプリンシパル (に関連付けられた IAM ロールなど) が設定されています。 通常、このタイプのデータコンシューマーには幅広いデータ要件があり、継続的に増加しています。

AWS Lake Formation は、クロスアカウントデータ共有と集中型カタログへのアクセスのためにデータコンシューマーが使用する最も重要な AWS サービスです。データベースが一元化されたカタログによって共有されると、共有リソースはデータコンシューマーアカウントの Lake Formation で使用できます。その後、必要に応じて、データプロデューサーからのアクセス許可を使用して、データコンシューマーアカウントのローカル IAM プリンシパルにデータアクセスを許可できます。その後、共有データは Lake Formation と統合された AWS サービス (Amazon Athena や など) で使用できま

す AWS Glue。次の AWS サービスを使用して、データコンシューマーアカウントの共有データにアクセスできます。

- Amazon Athena は、標準 SQL を使用して Amazon Simple Storage Service (Amazon S3) 内のデータを直接分析するのに役立つインタラクティブなクエリサービスです。Athena と Lake Formation の詳細については、Amazon Athena Athena ドキュメントの「Athena が Lake Formation に登録されたデータにアクセスする方法」を参照してください。
- Amazon Redshift Spectrum を使用すると、データを Amazon S3 のファイルから構造化データと半構造化データを効率的にクエリおよび取得できます。Redshift Spectrum と Lake Formation の詳細については、Amazon Redshift ドキュメントの「Lake Formation での Redshift Spectrum の使用」を参照してください。
- AWS Glue は、完全マネージド型の抽出、変換、ロード (ETL) サービスであり、データを分類し、クリーンアップし、強化し、さまざまなデータストアとデータストリーム間で確実に移動するシンプルで費用対効果の高いサービスです。 AWS Glue ETL ジョブに関連付けられた IAM ロールは、必要なアクセス許可がある場合、Lake Formation によって管理されるデータレイクデータにアクセスできます。
- Amazon EMR は、ビッグデータフレームワーク (Apache Hadoop や Apache Spark など) を実行して、大量のデータを処理および分析するのに役立ちます。Amazon EMR と Lake Formation の詳細については、Amazon EMR ドキュメントの「Amazon EMR と Lake Formation の統合」を参照してください。
- Amazon Quick は、スケーラブルでサーバーレス、埋め込み可能、機械学習 (ML) を活用したビジネスインテリジェンスサービスで、データレイクからのデータを分析および視覚化するために使用できます。Quick と Lake Formation の詳細については、Quick ドキュメントの「Lake Formation を介した接続の承認」を参照してください。
- Amazon SageMaker AI Data Wrangler (Data Wrangler) は、ML のデータの集約と準備にかかる時間を短縮します。Data Wrangler と Lake Formation の詳細については、Amazon SageMaker AI ドキュメントの「Amazon SageMaker AI Data Wrangler で ML データを準備するAmazon SageMaker 」を参照してください。

オンボーディングとアクセス許可の付与

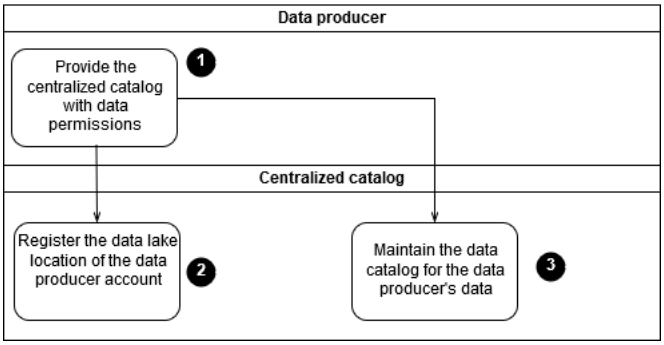
このガイドのデータレイクリファレンスアーキテクチャは、データプロデューサーとデータコンシューマーを個別にスケールするだけでなく、それらのデータコンシューマーをオンボーディングしてアクセス権を付与するための一貫したプロセスを定義して確立するのに役立ちます。

以下のセクションでは、データプロデューサーとデータコンシューマーのオンボーディングプロセスと、データコンシューマーアカウントでアクセス権を付与する方法について説明します。このガイドでは、一元化されたカタログとデータコンシューマーの間で名前付きリソースメソッドを使用します。LF-TBAC メソッドのプロセスは似ていますが、少し異なります。組織のデータガバナンスのプラクティスとポリシーを満たすために、これらのアプローチを評価して設定することをお勧めします。

これらの 2 つの方法の詳細については、このガイドの一元化されたカタログ「」セクションを参照してください。

データプロデューサーのオンボーディング

次の図は、新しいデータプロデューサーをデータレイクにオンボードする方法を示しています。



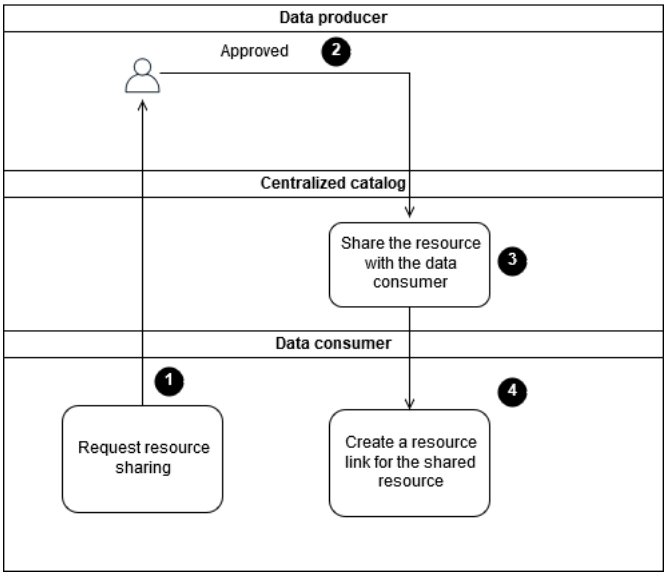
この図は、次のオンボーディングプロセスを示しています。

1. データプロデューサーは、一元化されたカタログにデータへのアクセス (Amazon Simple Storage Service (Amazon S3) バケットや など) を選択的に提供します AWS KMS key。アクセスは、一元化されたカタログの AWS Identity and Access Management (IAM) プリンシパルに提供され、データプロデューサーのデータレイクの場所を に登録 AWS Lake Formation し、データプロデューサーのカタログの維持に使用される IAM プリンシパルを登録します。
2. 一元化されたカタログの Lake Formation を使用するデータプロデューサーのデータレイクの場所 (S3 バケットなど) を登録します。

3. AWS Glue データカタログのデータプロデューサーから新しいデータのデータベース、テーブル、テーブルスキーマを作成します。

データコンシューマーのオンボーディング

次の図は、新しいデータコンシューマーをデータレイクにオンボードする方法を示しています。



この図は、次のオンボーディングプロセスを示しています。

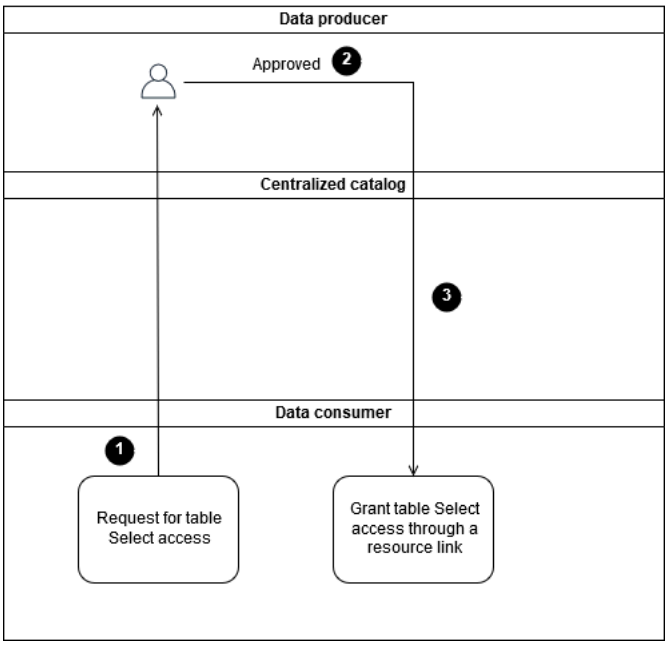
1. データコンシューマーは、データプロデューサーのデータを表示するための承認をリクエストし、アクセスする必要があるデータを指定します。
2. データプロデューサーのデータスチュワードは、データコンシューマーからのリクエストを確認し、以下を行うかどうかを評価します。
   - リクエストされたデータベースの一部またはすべてのテーブルを共有します。すべてのテーブルをデータコンシューマーと共有することによるデータセキュリティへの影響がない場合は、データベースレベルの共有をお勧めします。これにより、テーブルレベルの共有の管理オーバーヘッドを回避できます。
   - データコンシューマーの組織、OU、またはアカウントレベルで共有します。
3. データプロデューサーによって承認されると、必要な Data Catalog リソースが一元化されたカタログのデータコンシューマーと共有されます。
4. リソースリンクは、Lake Formation を使用してデータコンシューマーのアカウントで作成し、一元化されたカタログ内の共有データカタログリソースを指すことができます。

オンボーディングプロセスが完了すると、データコンシューマーの Lake Formation 管理者は、一元化されたカタログとリソースリンクからデータベースカタログリソースを表示できます。この段階では、データコンシューマーのアカウントの他のユーザーはデータプロデューサーのデータにアクセスできません。

データコンシューマーアカウントの Select アクセスを付与する

次の図は、データコンシューマーアカウントのローカル IAM プリンシパルと共有データリソース`Select`へのアクセスを許可するプロセスを示しています。ローカル IAM プリンシパルは、個々のユーザーの IAM ロールでも、特定の AWS サービスによって消費される IAM ロールでもかまいません。

> **Note**
>
> 共有されるデータが機密性が低い場合は、データプロデューサーの承認を必要とせずに、データコンシューマー自体にアクセス許可を委任できます。これは、信頼と共有がすでに確立されているためです。



図表に示す内容は以下のステップです。

1. データコンシューマーアカウントの個々の IAM プリンシパルは、データコンシューマーアカウントの IAM プリンシパルからリソースリンク`Select`へのアクセスをリクエストします。

2. データプロデューサーのデータスチュワードは、データコンシューマーからのリクエストを確認し、すべての要件が満たされた場合に承認を提供します。
3. Select アクセスが許可され、IAM プリンシパルがリクエストされたデータを消費できるようになります。

よくある質問

このセクションでは、 で拡張およびスケールするためのデータレイクの設計に関する一般的な質問に対する回答を提供します AWS クラウド。

このデータレイクリファレンスアーキテクチャはエンタープライズ組織により適していますか?

このガイドのデータレイクリファレンスアーキテクチャは、あらゆる規模の組織に属するデータレイクに適用できます。リファレンスアーキテクチャは、データ交換インターフェイスを標準化し、データレイクを維持および拡大するためのオーバーヘッドとコストを削減し、組織のデータレイクが拡大するあらゆる規模に適用できます。

組織にデータプロデューサーが 1 つしかない場合でも、このリファレンスアーキテクチャを使用できますか?

このガイドのデータレイクリファレンスアーキテクチャは、組織内に 1 つのデータプロデューサーしかない場合でも、依然として関連性があり有益です。一元化されたカタログがない場合、データプロデューサーはデータコンシューマーの増加を処理する必要があるため、複雑さとオーバーヘッドが増加します。データレイクも組織の長期的なアセットであり、通常、組織はより多くのデータプロデューサーを追加します。例えば、コンプライアンス上の理由から、または組織が独自のデータプロデューサーを持つ別のビジネスユニットを取得したために、機密データを保存するために追加のデータプロデューサーが必要になる場合があります。

データレイクは、1 つのデータプロデューサーを複数のデータコンシューマーに直接接続します。このガイドのデータレイクリファレンスアーキテクチャは引き続き重要ですか?

データレイクリファレンスアーキテクチャは、長期的には組織に利益をもたらします。2 ステップのアプローチを使用して、新しいデータコンシューマー向けの一元化されたカタログを構築することから始めることができます。その後、既存のデータコンシューマーを一元化されたカタログに接続できます。

組織は、変更を加えずにオンボーディングとアクセス許可のワークフローに従うべきですか?

いいえ。このセクションの主な目的は、オンボーディングプロセスに必要な論理アクティビティブロックを説明することです。すべての組織はプロセスをカスタマイズする必要があり、データの機密性によっては複数のプロセスがある場合もあります。

もう 1 つの考慮事項は、プロセスフローが でリソースベースの共有アプローチを使用することです AWS Lake Formation。タグベースの共有など、Lake Formation でサポートされている他のデータ共有方法もあります。このプロセスの違いは、特定の共有方法に合わせて調整できます。

リソース

- AWS Glue デベロッパーガイド
- AWS Lake Formation デベロッパーガイド
- Lake Formation でのクロスアカウントアクセス
- Lake Formation でのデータカタログ許可の付与と取り消し
- AWS Glue データカタログの入力

ドキュメント履歴

以下の表は、本ガイドの重要な変更点について説明したものです。今後の更新に関する通知を受け取る場合は、RSS フィード をサブスクライブできます。

変更	説明	日付
—	初版発行	2021 年 10 月 1 日

翻訳は機械翻訳により提供されています。提供された翻訳内容と英語版の間で齟齬、不一致または矛盾がある場合、英語版が優先します。