AWS 決定ガイド

AWS 分析サービスの選択

AWS 分析サービスの選択: AWS 決定ガイド

Table of Contents

AWS 分析サービスの選択

最初のステップを実行する

目的	どの AWS 分析サービスが組織に最適かを判断するのに役立ちます。
最終更新日	2025 年 9 月 24 日
対象サービス	• Amazon Athena • AWS Clean Rooms • Amazon Data Firehose • Amazon DataZone • Amazon EMR • AWS Glue • Amazon Kinesis Data Streams • Amazon Managed Service for Apache Flink • Amazon Managed Streaming for Apache Kafka • Amazon Managed Workflows for Apache Airflow • Amazon OpenSearch Service • QuickSight • Amazon Redshift • Amazon S3 • Amazon SageMaker

序章

データは最新のビジネスの基礎です。人とアプリケーションは、新しい多様なソースからのデータに安全にアクセスして分析する必要があります。データ量も絶えず増加しているため、組織は必要なすべてのデータのキャプチャ、保存、分析に苦労する可能性があります。

これらの課題に対処するには、分析やインサイトのためのすべてのデータサイロを破壊する最新のデータアーキテクチャを構築し、end-to-endのガバナンスにより、組織内のすべてのユーザーが 1 か所でアクセスできるようにします。また、分析システムと機械学習 (ML) システムを接続して予測分析を可能にすることもますます重要です。

この決定ガイドは、 AWS サービスで最新のデータアーキテクチャを構築するための適切な質問をするのに役立ちます。データサイロ (データレイクとデータウェアハウスを接続する) 、システムサイロ (ML と分析を接続する) 、および人材サイロ (組織内の全員の手元にデータを配置する) を分割する方法について説明します。

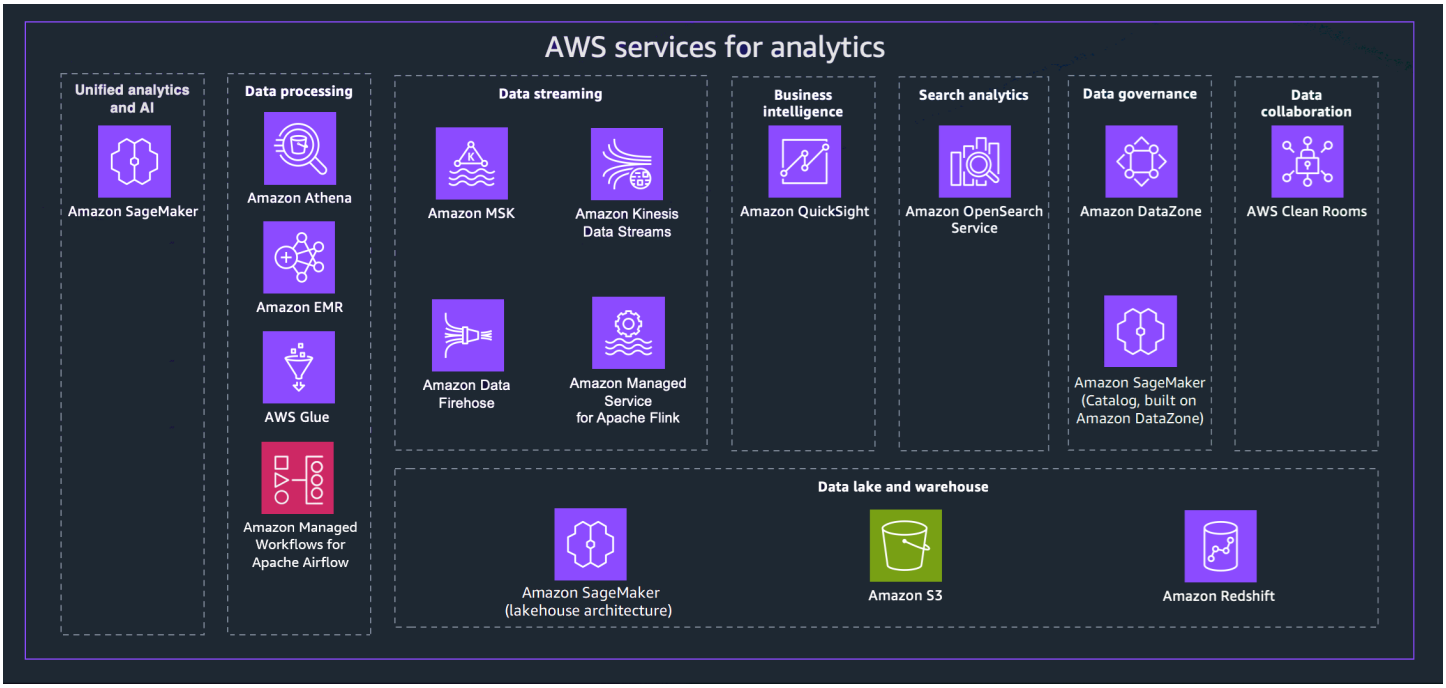
この 8 分間の抜粋は、re:Invent 2024 での Sirish Chandrasekaran と Rick Sears による 1 時間のプレゼンテーションからのものです。架空の会社である Maxdome が、次世代の Amazon SageMaker の一部である SageMaker Unified Studio AI と分析を使用してデータインサイトを引き出す方法の概要を説明します。 Amazon SageMaker

AWS 分析サービスを理解する

最新のデータ戦略は、データの管理、アクセス、分析、および対応に役立つ一連のテクノロジー構成要素を使用して構築されています。また、データソースに接続するための複数のオプションも用意されています。最新のデータ戦略では、チームに次の権限を与える必要があります。

- 任意のツールまたは手法を使用する
- 人工知能 (AI) を使用して、データに関する特定の質問に対する回答を見つけるのに役立ちます。
- 適切なセキュリティとデータガバナンスコントロールを使用して、データにアクセスできるユーザーを管理する
- データサイロを分割して、データレイクと専用データストアの両方を最大限に活用する
- 任意の量のデータを低コストでオープンな標準ベースのデータ形式で保存する
- データレイク、データウェアハウス、運用データベース、アプリケーション、フェデレーティッドデータソースを一貫した全体に接続する

AWS には、最新のデータ戦略の達成に役立つさまざまなサービスが用意されています。次の図は、このガイドで説明する分析 AWS のサービスを示しています。以下のタブには、追加の詳細が表示されます。



Unified analytics and AI

次世代の Amazon SageMaker は、広く採用されている AWS 機械学習 (ML) と分析機能を組み合わせて、分析と AI の統合エクスペリエンスを提供し、すべてのデータへの統一されたアクセスを提供します。Amazon SageMaker Unified Studio を使用すると、モデル開発、生成 AI アプリケーション開発、データ処理、SQL 分析用の使い慣れた AWS ツールを使用して、コラボレーションと構築を高速化できます。これらはすべて、ソフトウェア開発用の生成 AI アシスタントである Amazon Q Developer によって高速化されます。エンタープライズセキュリティ要件を満たすための組み込みガバナンスを使用して、データレイク、データウェアハウス、またはサードパーティーやフェデレーティッドソースからデータにアクセスします。

Data processing

- Amazon Athena は、Amazon S3 に保存されている非構造化、半構造化、構造化データの分析に役立ちます。たとえば、CSV 形式、JSON 形式、列データ形式 (Apache Parquet や Apache ORC など) に対応しています。Athena は ANSI SQL を使用したアドホッククエリの実行に利用でき、データを集約したり、データを Athena にロードしたりする必要はありません。Athena は QuickSight AWS Glue Data Catalogやその他の AWS サービスと統合されます。また、インフラストラクチャを管理することなく Trino を使用してデータを大規模に分析し、Apache Flink と Apache Spark を使用してリアルタイム分析を構築することもできます。
- Amazon EMR は、Apache Hadoop や Apache Spark などのビッグデータフレームワークの実行を簡素化 AWS し、大量のデータを処理および分析するマネージド型クラスタープラットフォームです。これらのフレームワークと、関連するオープンソースプロジェクトを使用する

ことで、分析用のデータやビジネスインテリジェンスワークロードを処理できます。Amazon EMR では、Amazon S3 などの他の AWS データストアやデータベースとの間で大量のデータを変換および移動することもできます。

- を使用するとAWS Glue、100 を超える多様なデータソースを検出して接続し、一元化されたデータカタログでデータを管理できます。ETL パイプラインを視覚的に作成、実行、モニタリングして、データをデータレイクにロードできます。また、Athena、Amazon EMR、Amazon Redshift Spectrum を使用して、カタログ化されたデータをすぐに検索およびクエリできます。
- Amazon Managed Workflows for Apache Airflow (MWAA) は、Apache Airflow のフルマネージド実装であり、クラウド内のデータワークフローの作成、スケジュール、モニタリングを容易にします。MWAA は、ニーズに合わせてワークフロー容量を自動的にスケーリングし、 AWS セキュリティサービスと統合します。MWAA を使用して、データ処理、ETL ジョブ、機械学習パイプラインなど、分析サービス全体のワークフローをオーケストレーションできます。

Data streaming

- Amazon Managed Streaming for Apache Kafka (Amazon MSK) を使用すると、Apache Kafka を使用してストリーミングデータを処理するためのアプリケーションを構築して実行できます。Amazon MSK は、クラスターの作成、更新、削除などに用いられるコントロールプレーンオペレーションを提供します。データの生成と消費などの、Apache Kafka データプレーンオペレーションを使用することができます。
- Amazon Kinesis Data Streams を使用すると、大量のデータレコードストリームをリアルタイムで収集して処理できます。使用されるデータには、IT インフラストラクチャのログデータ、アプリケーションのログ、ソーシャルメディア、マーケットデータフィード、ウェブのクリックストリームデータなどの種類があります。
- Amazon Data Firehose は、Amazon S3、Amazon Redshift、Amazon OpenSearch Service、Splunk、Apache Iceberg Tables などの宛先にリアルタイムのストリーミングデータを配信するためのフルマネージドサービスです。また、Datadog、Dynatrace、LogicMonitor、MongoDB、New Relic、Coralogix、Elastic など、サポートされているサードパーティーサービスプロバイダーが所有するカスタム HTTP エンドポイントまたは HTTP エンドポイントにデータを送信することもできます。
- Amazon Managed Service for Apache Flink を使用すると、Java、Scala、Python、または SQL を使用してストリーミングデータを処理および分析できます。ストリーミングソースと静的ソースに対してコードを作成して実行し、時系列分析、リアルタイムダッシュボードのフィード、メトリクスを実行できます。

Business intelligence

QuickSight は、インタラクティブなビジュアル環境で情報を探索して解釈する機会を意思決定者に提供します。QuickSight は、単一のデータダッシュボードで AWS 、データ、サードパーティーデータ、ビッグデータ、スプレッドシートデータ、SaaS データ、B2B データなどを含めることができます。QuickSight Q を使用すると、自然言語を使用してデータについて質問し、レスポンスを受け取ることができます。たとえば、「カリフォルニアで最も売れているカテゴリは何ですか？」と入力します。

Search analytics

Amazon OpenSearch Service は、OpenSearch クラスターのすべてのリソースをプロビジョニングして、OpenSearch クラスターを起動します。また、障害が発生した OpenSearch Service ノードを自動的に検出して置き換え、セルフマネージドインフラストラクチャに関連するオーバーヘッドを減らします。OpenSearch Service の直接クエリを使用して、Amazon S3 やその他の AWS サービスのデータを分析できます。

Data governance

Amazon DataZone を使用すると、きめ細かなコントロールを使用してデータへのアクセスを管理および管理できます。これらのコントロールは、適切なレベルの権限とコンテキストによるアクセスの確保に役立ちます。Amazon DataZone は、Amazon Redshift、Athena、QuickSight、オンプレミスソース AWS Glue、サードパーティーソースなどのデータ管理サービスを統合することで、アーキテクチャを簡素化します。

Data collaboration

AWS Clean Rooms は、raw データへのアクセスを提供することなく集合データセットを分析できる安全なコラボレーションワークスペースです。コラボレーションするパートナーを選択し、データセットを選択し、それらのパートナーのプライバシー強化コントロールを設定することで、他の企業とコラボレーションできます。クエリを実行すると、 はそのデータの元の場所からデータを AWS Clean Rooms 読み取り、組み込みの分析ルールを適用して、そのデータの制御を維持します。

Data lake and data warehouse

- 次世代の Amazon SageMaker は Apache Iceberg と完全に互換性があり、Amazon Simple Storage Service (Amazon S3) データレイクと Amazon Redshift データウェアハウス間でデータを統合できます。これにより、分析、AI、機械学習 (ML) アプリケーションを 1 つのデータコピーで構築できます。ゼロ ETL 統合により、運用ソースからほぼリアルタイムでデータをストリーミングし、複数のソース間でフェデレーティッドクエリを実行し、Apache Iceberg 互換

ツールを使用してデータにアクセスできます。すべての分析および ML ツールとエンジンに適用されるきめ細かなアクセス許可を定義することで、データを保護できます。

- Amazon S3 は、データレイク基盤に使用できるほぼすべての量と種類のデータを保存および保護できます。Amazon S3 には、特定のビジネス、組織、コンプライアンスの要件を満たすために、データへのアクセスを最適化、整理、設定できる管理機能があります。Amazon S3 Tables は、分析ワークロード用に最適化された S3 ストレージを提供します。標準の SQL ステートメントを使用すると、Athena、Amazon Redshift、Apache Spark などの Iceberg をサポートするクエリエンジンでテーブルをクエリできます。
- Amazon Redshift は、フルマネージド型のペタバイト規模のデータウェアハウスサービスです。Amazon Redshift は Amazon SageMaker のデータレイクハウスに接続できるため、Amazon Redshift データウェアハウスと Amazon S3 データレイク全体の統合データで強力な SQL 分析機能を使用できます。Amazon Redshift で Amazon Q を使用することもできます。これにより、自然言語による SQL の作成が簡素化されます。

AWS 分析サービスの基準を検討する

データ分析を構築するには、さまざまな理由があります AWS。クラウド移行ジャーニーの最初のステップとして、グリーンフィールドプロジェクトまたはパイロットプロジェクトをサポートする必要がある場合があります。または、中断を最小限に抑えながら既存のワークロードを移行する場合もあります。どのような目標であっても、以下の考慮事項が選択に役立ちます。

Assess data sources and data types

利用可能なデータソースとデータ型を分析して、データの多様性、頻度、品質を包括的に把握します。データの処理と分析における潜在的な課題を理解します。この分析は、次の理由から重要です。

- データソースは多様で、さまざまなシステム、アプリケーション、デバイス、外部プラットフォームから取得されます。
- データソースには、データ更新の一意の構造、形式、頻度があります。これらのソースを分析すると、適切なデータ収集方法とテクノロジーを特定するのに役立ちます。
- 構造化データ、半構造化データ、非構造化データなどのデータ型を分析することで、適切なデータ処理とストレージのアプローチが決まります。
- データソースとタイプを分析すると、データ品質評価が容易になり、値の欠如、不整合、不正確さなど、潜在的なデータ品質の問題を予測するのに役立ちます。

Data processing requirements

データの取り込み、変換、クリーンアップ、分析の準備方法に関するデータ処理要件を決定します。主な考慮事項は次のとおりです。

- データ変換： raw データを分析に適したものにするために必要な特定の変換を決定します。これには、データの集約、正規化、フィルタリング、エンリッチメントなどのタスクが含まれます。
- データクレンジング： データ品質を評価し、欠落、不正確、または一貫性のないデータを処理するプロセスを定義します。データクレンジング手法を実装して、信頼性の高いインサイトを得るために高品質のデータを確保します。
- 処理頻度： 分析ニーズに基づいて、リアルタイム、ほぼリアルタイム、またはバッチ処理が必要かどうかを判断します。リアルタイム処理は即時のインサイトを可能にしますが、定期的な分析にはバッチ処理で十分です。
- スケーラビリティとスループット： データボリュームの処理、処理速度、同時データリクエストの数に関するスケーラビリティ要件を評価します。選択した処理アプローチが将来の成長に対応できることを確認します。
- レイテンシー： データ処理に許容されるレイテンシーと、データの取り込みから分析結果までにかかる時間を考慮します。これは、リアルタイム分析や時間的制約のある分析に特に重要です。

Storage requirements

分析パイプライン全体でデータを保存する方法と場所を決定して、ストレージのニーズを判断します。重要な考慮事項は次のとおりです。

- データ量： 生成および収集されるデータ量を評価し、将来のデータ増加を見積もり、十分なストレージ容量を計画します。
- データ保持： 履歴分析またはコンプライアンス目的でデータを保持する期間を定義します。適切なデータ保持ポリシーを決定します。
- データアクセスパターン： データにアクセスしてクエリを実行し、最適なストレージソリューションを選択する方法を理解します。読み取りおよび書き込みオペレーション、データアクセス頻度、データローカリティを検討してください。
- データセキュリティ： 暗号化オプション、アクセスコントロール、データ保護メカニズムを評価してデータセキュリティを優先し、機密情報を保護します。

- コストの最適化： データアクセスパターンと使用状況に基づいて最も費用対効果の高いストレージソリューションを選択して、ストレージコストを最適化します。
- 分析サービスとの統合： 選択したストレージソリューションとパイプライン内のデータ処理および分析ツールをシームレスに統合します。

Types of data

データの収集と取り込みのための分析サービスを決定するときは、組織のニーズと目的に関連するさまざまなタイプのデータを検討してください。考慮すべきデータの一般的なタイプは次のとおりです。

- トランザクションデータ： 顧客の購入、金融取引、オンライン注文、ユーザーアクティビティログなど、個々のインタラクションまたはトランザクションに関する情報が含まれます。
- ファイルベースのデータ： ログファイル、スプレッドシート、ドキュメント、イメージ、オーディオファイル、ビデオファイルなどのファイルに保存される構造化データまたは非構造化データを指します。分析サービスは、さまざまなファイル形式の取り込みをサポートする必要があります。
- イベントデータ： ユーザーアクション、システムイベント、マシンイベント、ビジネスイベントなど、重大な発生またはインシデントをキャプチャします。イベントには、オンストリームまたはダウンストリーム処理用にキャプチャされる高速で到着するデータを含めることができます。

Operational considerations

運用責任はユーザー間で共有され、責任の分担 AWSはモダナイゼーションのさまざまなレベルによって異なります。で分析インフラストラクチャを自己管理 AWS することも、多数のサーバーレス分析サービスを活用してインフラストラクチャ管理の負担を軽減することもできます。

セルフマネージドオプションにより、ユーザーはインフラストラクチャと設定をより詳細に制御できますが、より多くの運用作業が必要になります。

サーバーレスオプションは、運用上の負担の大部分を取り除き、自動スケーラビリティ、高可用性、堅牢なセキュリティ機能を提供します。これにより、ユーザーはインフラストラクチャや運用タスクを管理するのではなく、分析ソリューションの構築とインサイトの促進に集中できます。サーバーレス分析ソリューションには以下の利点があります。

- インフラストラクチャの抽象化： サーバーレスサービスはインフラストラクチャ管理を抽象化し、ユーザーのプロビジョニング、スケーリング、メンテナンスタスクを軽減します。 はこれらの運用面 AWS を処理し、管理オーバーヘッドを削減します。
- 自動スケーリングとパフォーマンス： サーバーレスサービスはワークロードの需要に基づいてリソースを自動的にスケーリングし、手動による介入なしで最適なパフォーマンスを確保します。
- 高可用性とディザスタリカバリ： AWS サーバーレスサービスに高可用性を提供します。 は、データの冗長性、レプリケーション、ディザスタリカバリ AWS を管理し、データの可用性と信頼性を向上させます。
- セキュリティとコンプライアンス： AWS 業界標準とベストプラクティスに従って、サーバーレスサービスのセキュリティ対策、データ暗号化、コンプライアンスを管理します。
- モニタリングとログ記録： AWS サーバーレスサービスの組み込みモニタリング、ログ記録、アラート機能を提供します。ユーザーは Amazon CloudWatch を通じて詳細なメトリクスとログにアクセスできます。

Type of workload

最新の分析パイプラインを構築する場合、さまざまな分析ニーズを効果的に満たすには、サポートするワークロードのタイプを決定することが重要です。ワークロードのタイプごとに考慮すべき主な決定ポイントは次のとおりです。

バッチワークロード

- データ量と頻度： バッチ処理は、定期的な更新を伴う大量のデータに適しています。
- データレイテンシー： バッチ処理では、リアルタイム処理と比較してインサイトの提供に多少の遅延が生じる可能性があります。

インタラクティブ分析

- データクエリの複雑さ： インタラクティブな分析では、迅速なフィードバックのために低レイテンシーのレスポンスが必要です。
- データの視覚化： ビジネスユーザーがデータを視覚的に探索できるように、インタラクティブなデータ視覚化ツールの必要性を評価します。

ストリーミングワークロード

- データ速度とボリューム： ストリーミングワークロードでは、高速データを処理するためにリアルタイム処理が必要です。
- データウィンドウ： ストリーミングデータのデータウィンドウと時間ベースの集計を定義して、関連するインサイトを抽出します。

Type of analysis needed

ビジネス目標と、分析から導き出すことを目的としたインサイトを明確に定義します。さまざまなタイプの分析は、さまざまな目的を果たします。例:

- 記述分析は、履歴の概要を取得するのに最適です
- 診断分析は、過去のイベントの原因を理解するのに役立ちます
- 予測分析は将来の成果を予測する
- 規範的分析は、最適なアクションのレコメンデーションを提供します

ビジネス目標を関連するタイプの分析と一致させます。以下は、適切なタイプの分析を選択するのに役立つ重要な決定基準です。

- データの可用性と品質： 記述分析と診断分析は履歴データに依存しますが、予測分析と規範分析には、正確なモデルを構築するために十分な履歴データと高品質のデータが必要です。
- データ量と複雑さ： 予測分析と規範分析には、大量のデータ処理と計算リソースが必要です。インフラストラクチャとツールがデータ量と複雑さを処理できることを確認します。
- 決定の複雑さ： 決定に複数の変数、制約、目標が含まれる場合、最適なアクションの指針として規範分析が適している場合があります。
- リスク許容度： 規範的分析はレコメンデーションを提供する可能性がありますが、関連する不確実性があります。意思決定者が分析出力に関連するリスクを理解していることを確認します。

Evaluate scalability and performance

アーキテクチャのスケーラビリティとパフォーマンスのニーズを評価します。設計では、増加するデータボリューム、ユーザー需要、分析ワークロードを処理する必要があります。考慮すべき主な決定要因は次のとおりです。

- データ量と増加： 現在のデータ量を評価し、将来の増加を予測します。

- データ速度とリアルタイム要件： データをリアルタイムで処理および分析する必要があるか、ほぼリアルタイムで処理および分析する必要があるかを決定します。
- データ処理の複雑さ： データ処理と分析タスクの複雑さを分析します。計算負荷の高いタスクの場合、Amazon EMR などのサービスは、ビッグデータ処理のためのスケーラブルで管理された環境を提供します。
- 同時実行数とユーザーロード： 同時ユーザー数とシステムのユーザーロードレベルを考慮します。
- 自動スケーリング機能： 自動スケーリング機能を提供するサービスを検討し、リソースが需要に応じて自動的にスケールアップまたはスケールダウンできるようにします。これにより、効率的なリソース使用率とコスト最適化が保証されます。
- 地理的分散： データアーキテクチャを複数のリージョンまたはロケーションに分散する必要がある場合は、グローバルレプリケーションと低レイテンシーのデータアクセスを備えたサービスを検討します。
- コストパフォーマンスのトレードオフ： パフォーマンスのニーズとコストの考慮事項のバランスを取ります。パフォーマンスの高いサービスには、より高いコストがかかる場合があります。
- サービスレベルアグリーメント (SLAs): AWS サービスが提供する SLAs をチェックして、スケーラビリティとパフォーマンスの期待を満たしていることを確認します。

Data governance

データガバナンスは、データアセットの効果的な管理、品質、セキュリティ、コンプライアンスを確保するために実装する必要がある一連のプロセス、ポリシー、コントロールです。考慮すべき主な決定点は次のとおりです。

- データ保持ポリシー： 規制要件とビジネスニーズに基づいてデータ保持ポリシーを定義し、不要になったデータを安全に廃棄するためのプロセスを確立します。
- 監査証跡とログ記録： データアクセスと使用状況をモニタリングするためのログ記録と監査のメカニズムを決定します。包括的な監査証跡を実装して、データの変更、アクセス試行、およびユーザーアクティビティを追跡し、コンプライアンスとセキュリティのモニタリングを行います。
- コンプライアンス要件： 組織に適用される業界固有および地理的なデータコンプライアンス規制を理解します。データアーキテクチャがこれらの規制とガイドラインに沿っていることを確認します。

- データ分類： 機密性に基づいてデータを分類し、データクラスごとに適切なセキュリティコントロールを定義します。
- ディザスタリカバリとビジネス継続性： ディザスタリカバリとビジネス継続性を計画し、予期しないイベントやシステム障害が発生した場合にデータの可用性と回復力を確保します。
- サードパーティーのデータ共有： サードパーティーエンティティとデータを共有する場合は、安全なデータ共有プロトコルと契約を実装して、データの機密性を保護し、データの不正使用を防止します。

Security

分析パイプライン内のデータのセキュリティには、パイプラインのすべての段階でデータを保護し、機密性、完全性、可用性を確保することが含まれます。考慮すべき主な決定点は次のとおりです。

- アクセスコントロールと認可： 堅牢な認証および認可プロトコルを実装して、承認されたユーザーのみが特定のデータリソースにアクセスできるようにします。
- データ暗号化： データベース、データレイク、およびアーキテクチャのさまざまなコンポーネント間のデータ移動中に保存されるデータに適した暗号化方法を選択します。
- データマスキングと匿名化： 特定の分析プロセスを続行しながら、PII や機密データなどの機密データを保護するために、データマスキングや匿名化の必要性を考慮します。
- 安全なデータ統合： 安全なデータ統合プラクティスを確立して、アーキテクチャのさまざまなコンポーネント間でデータを安全にフローし、データ漏洩やデータ移動中の不正アクセスを回避します。
- ネットワーク分離： リソースがパブリックインターネットに公開されないように、Amazon VPC エンドポイントをサポートするサービスを検討してください。

Plan for integration and data flows

分析パイプラインのさまざまなコンポーネント間の統合ポイントとデータフローを定義して、シームレスなデータフローと相互運用性を確保します。考慮すべき主な決定点は次のとおりです。

- データソースの統合： データベース、アプリケーション、ファイル、外部 APIs など、データを収集するデータソースを特定します。データ取り込み方法 (バッチ、リアルタイム、イベントベース) を決定して、データを最小限のレイテンシーで効率的にパイプラインに取り込みます。

- データ変換： 分析用のデータを準備するために必要な変換を決定します。パイプライン内を移動するデータをクリーンアップ、集約、正規化、または強化するツールとプロセスを決定します。
- データ移動アーキテクチャ： パイプラインコンポーネント間のデータ移動に適したアーキテクチャを選択します。リアルタイム要件とデータボリュームに基づいて、バッチ処理、ストリーム処理、またはその両方の組み合わせを検討してください。
- データレプリケーションと同期： データレプリケーションと同期メカニズムを決定して、すべてのコンポーネントでデータをup-to-date保ちます。データの鮮度の要件に応じて、リアルタイムレプリケーションソリューションまたは定期的なデータ同期を検討してください。
- データ品質と検証： データ品質チェックと検証ステップを実装して、パイプラインを通過するデータの整合性を確保します。アラートや エラー処理など、データの検証に失敗したときに実行するアクションを決定します。
- データセキュリティと暗号化： 転送中および保管中のデータを保護する方法を決定します。データの機密性に基づいて必要なセキュリティレベルを考慮して、パイプライン全体で機密データを保護するための暗号化方法を決定します。
- スケーラビリティと耐障害性： データフロー設計により、水平方向のスケーラビリティが可能になり、増加したデータボリュームとトラフィックを処理できることを確認します。

Architect for cost optimization

で分析パイプラインを構築すると AWS 、さまざまなコスト最適化の機会が得られます。コスト効率を確保するために、次の戦略を検討してください。

- リソースのサイズ設定と選択： 実際のワークロード要件に基づいてリソースのサイズを適正化します。過剰プロビジョニングを回避しながら、ワークロードのパフォーマンスニーズに合った AWS サービスとインスタンスタイプを選択します。
- Auto Scaling: さまざまなワークロードが発生するサービスの Auto Scaling を実装します。自動スケーリングは、需要に基づいてインスタンスの数を動的に調整し、トラフィックが少ない時間帯のコストを削減します。
- スポットインスタンス： 重要でない耐障害性のあるワークロードには Amazon EC2 スポットインスタンスを使用します。スポットインスタンスは、オンデマンドインスタンスと比較してコストを大幅に削減できます。
- リザーブドインスタンス： 予測可能な使用量で安定したワークロードのオンデマンド料金よりも大幅なコスト削減を実現するには、リザーブドインスタンスの購入 AWS を検討してください。

- データストレージ階層化： データアクセス頻度に基づいて異なるストレージクラスを使用することで、データストレージコストを最適化します。
- データライフサイクルポリシー： データライフサイクルポリシーを確立して、経過時間と使用パターンに基づいてデータを自動的に移動または削除します。これにより、ストレージコストを管理し、データストレージをその価値に合わせて維持できます。

AWS 分析サービスを選択する

分析ニーズを評価する基準がわかったので、組織のニーズに適した AWS 分析サービスを選択する準備が整いました。次の表は、一連のサービスを共通の機能やビジネス目標に合わせています。

カテゴリ	何に最適化されていますか？	サービス
統合分析と AI	分析と AI 開発 単一の開発環境である Amazon SageMaker Unified Studio を使用して、データ、分析、AI 機能にアクセスするために最適化されています。	Amazon SageMaker
データ処理	インタラクティブ分析 リアルタイムのデータ分析と探索を実行するために最適化されており、ユーザーはデータをインタラクティブにクエリおよび視覚化できます。	Amazon Athena
	ビッグデータ処理 大量のデータの処理、移動、変換に最適化されています。	Amazon EMR
	データカタログ 利用可能なデータ、その構造、特性、関係に関する詳細	AWS Glue

カテゴリ	何に最適化されていますか？	サービス
	情報を提供するように最適化されています。	
	ワークフローオーケストレーション Apache Airflow を使用して分析プロセスと ETL ジョブを調整するデータワークフローを作成、スケジューリング、モニタリングするために最適化されています。	Amazon MWAA
データストリーミング	ストリーミングデータの Apache Kafka 処理 Apache Kafka データプレーンオペレーションを使用し、Apache Kafka のオープンソースバージョンを実行するために最適化されています。	Amazon MSK
	リアルタイム処理 IT インフラストラクチャログデータ、アプリケーションログ、ソーシャルメディア、市場データフィード、ウェブクリックストリームデータなど、迅速かつ継続的なデータの取り込みと集約に最適化されています。	Amazon Kinesis Data Streams

カテゴリ	何に最適化されていますか？	サービス
	リアルタイムのストリーミングデータ配信 Amazon S3、Amazon Redshift、OpenSearch Service、Splunk、Apache Iceberg Tables、およびサポートされているサードパーティーサービスプロバイダーが所有するカスタム HTTP エンドポイントまたは HTTP エンドポイントなどの送信先にリアルタイムのストリーミングデータを配信するように最適化されています。	Amazon Data Firehose
	Apache Flink アプリケーションの構築 Java、Scala、Python、または SQL を使用してストリーミングデータを処理および分析するために最適化されています。	Amazon Managed Service for Apache Flink
ビジネスインテリジェンス	ダッシュボードとビジュアライゼーション 複雑なデータセットを視覚的に表現し、データの自然言語クエリを提供するように最適化されています。	QuickSight

カテゴリ	何に最適化されていますか？	サービス
検索分析	マネージド OpenSearch クラスター ログ分析、リアルタイムアプリケーションモニタリング、クリックストリーム分析用に最適化されています。	Amazon OpenSearch Service
データガバナンス	データアクセスの管理 ライフサイクル全体でデータの適切な管理、可用性、ユーザビリティ、整合性、セキュリティを設定するように最適化されています。	Amazon DataZone
データコラボレーション	安全なデータクリーンルーム 基盤となる生のデータを共有せずに、他の企業とコラボレーションするために最適化されています。	AWS Clean Rooms

カテゴリ	何に最適化されていますか？	サービス
データレイクとウェアハウス	データレイクとデータウェアハウスへの統一されたアクセス Amazon S3 データレイク、Amazon Redshift データウェアハウス、運用データベース、サードパーティーおよびフェデレーティッドデータソース間のデータアクセスを統合するために最適化するためのレイクハウスアーキテクチャ上に構築されています。	Amazon SageMaker
	データレイクのオブジェクトストレージ 実質的に無制限のスケーラビリティと高い耐久性を備えたデータレイク基盤を提供するために最適化されています。	Amazon S3
	データウェアハウス 組織内のさまざまなソースから大量の構造化データ、場合によっては半構造化データを一元的に保存、整理、取得するために最適化されています。	Amazon Redshift

AWS 分析サービスを使用する

これで、ビジネス目標と、データパイプラインの構築を開始するために取り込んで分析するデータの量と速度を明確に理解できたはずです。

を使用して、利用可能な各サービスの詳細について調べる方法を調べるために、各サービスの仕組みを調べるための経路を用意しました。以下のセクションでは、基本的な使用法からより高度な詳細分析を開始するための詳細なドキュメント、実践的なチュートリアル、リソースへのリンクを提供します。

Amazon Athena

- Amazon Athena の開始方法

  Amazon Athena を使用してデータをクエリし、Amazon S3 に保存されているサンプルデータに基づいてテーブルを作成し、テーブルをクエリして、クエリの結果を確認する方法について説明します。

  チュートリアルを始める

- Athena で Apache Spark の使用を開始する

  Athena コンソールのシンプルなノートブックエクスペリエンスを使用して、Python または Athena ノートブック APIs。

  チュートリアルを始める

- Amazon SageMaker レイクハウスアーキテクチャを使用した Athena フェデレーティッドクエリのカタログ化と管理

  Amazon SageMaker のデータレイクハウスを介して、Amazon Redshift、DynamoDB、Snowflake に保存されているデータに対してフェデレーティッドクエリに接続、管理、実行する方法について説明します。

  ブログを読む

- Athena を使用した Amazon S3 でのデータの分析

  事前定義された形式のテキストファイルとして生成された Elastic Load Balancer の口グで Athena を使用する方法について説明します。テーブルの作成、Athena が使用する形式でのデータのパーティション分割、Parquet への変換、クエリパフォーマンスの比較を行う方法について説明します。

ブログ記事を読む

AWS Clean Rooms

- セットアップ AWS Clean Rooms

  アカウント AWS Clean Rooms で AWS を設定する方法について説明します。

  ガイドを読む
- 基盤となるデータを共有 AWS Clean Rooms せずに でエンティティ解決を使用して AWS 、マルチパーティーデータセットのデータインサイトをロック解除する

  準備とマッチングを使用して、共同作業者とのデータマッチングを改善する方法について説明します。

  ブログ記事を読む
- 差分プライバシーが個人レベルでデータを明らかにすることなくインサイトを引き出す方法

  AWS Clean Rooms 差分プライバシーが差分プライバシーの適用を簡素化し、ユーザーのプライバシーを保護する方法について説明します。

  ブログを読む

Amazon Data Firehose

- チュートリアル: コンソールから Firehose ストリームを作成する

  AWS Management Console または AWS SDK を使用して、選択した送信先への Firehose ストリームを作成する方法について説明します。

  ガイドを読む
- Firehose ストリームにデータを送信する

  さまざまなデータソースを使用して Firehose ストリームにデータを送信する方法について説明します。

  ガイドを読む
- Firehose でソースデータを変換する

Lambda 関数を呼び出して受信ソースデータを変換し、変換されたデータを宛先に配信する方法について説明します。

ガイドを読む

Amazon DataZone

- Amazon DataZone の開始方法

  Amazon DataZone ルートドメインを作成し、データポータル URL を取得し、データプロデューサーとデータコンシューマーの基本的な Amazon DataZone ワークフローについて説明します。

  チュートリアルを始める
- 次世代の Amazon SageMaker と Amazon DataZone でのデータ系統の一般提供を発表

  Amazon DataZone が自動系統キャプチャを使用して、 と Amazon Redshift から系統情報を自動的に収集 AWS Glue してマッピングする方法について説明します。

  ブログを読む

Amazon EMR

- Amazon EMR の開始方法

  Spark を使用してサンプルクラスターを起動する方法と、Amazon S3 バケットに保存されているシンプルな PySpark スクリプトを実行する方法について説明します。

  チュートリアルを始める
- Amazon EKS での Amazon EMR の開始方法

  仮想クラスターに Spark アプリケーションをデプロイして、Amazon EMR on Amazon EKS の使用を開始する方法を示します。

  ガイドを見る
- EMR Serverless の使用を開始する

  Amazon EMR Serverless が、最新のオープンソースフレームワークを使用する分析アプリケーションの運用を簡素化するサーバーレスランタイム環境を提供する方法について説明します。

チュートリアルを始める

AWS Glue

- の開始方法 AWS Glue DataBrew

  最初の DataBrew プロジェクトを作成する方法について説明します。サンプルデータセットをロードし、そのデータセットで変換を実行し、それらの変換をキャプチャするレシピを構築し、変換されたデータを Amazon S3 に書き込むジョブを実行します。

  チュートリアルを始める

- を使用したデータの変換 AWS Glue DataBrew

  データアナリストやデータサイエンティストがデータをクリーニングして正規化し AWS Glue DataBrew、分析や機械学習の準備を容易にするビジュアルデータ準備ツールである について説明します。を使用して ETL プロセスを構築する方法について説明します AWS Glue DataBrew。

  ラボの使用を開始する

- AWS Glue DataBrew イマージョンデー

  を使用して AWS Glue DataBrew 、分析と機械学習のためにデータをクリーンアップおよび正規化する方法について説明します。

  ワークショップの開始方法

- の開始方法 AWS Glue Data Catalog

  Amazon S3 バケットをデータソースとして使用する AWS Glue Data Catalog最初の を作成する方法について説明します。

  チュートリアルを始める

- のデータカタログとクローラ AWS Glue

  データカタログの情報を使用して ETL ジョブを作成およびモニタリングする方法について説明します。

  ガイドを見る

Amazon Kinesis Data Streams

- Amazon Kinesis Data Streams の開始方法チュートリアル

  リアルタイムの株式データを処理および分析する方法について説明します。

  チュートリアルを始める
- Amazon Kinesis Data Streams を使用したリアルタイム分析のためのアーキテクチャパターン、パート 1

  時系列データ分析とイベント駆動型マイクロサービスの 2 つのユースケースの一般的なアーキテクチャパターンについて説明します。

  ブログを読む
- Amazon Kinesis Data Streams を使用したリアルタイム分析のためのアーキテクチャパターン、パート 2

  リアルタイム生成ビジネスインテリジェンス、リアルタイムレコメンデーションシステム、Internet of Things データストリーミングと推論の 3 つのシナリオにおける Kinesis Data Streams を使用した AI アプリケーションについて説明します。

  ブログを読む

Amazon Managed Service for Apache Flink

- Amazon Managed Service for Apache Flink とは

  Amazon Managed Service for Apache Flink の基本概念を理解します。

  ガイドを見る
- Amazon Managed Service for Apache Flink ワークショップ

  このワークショップでは、Amazon Managed Service for Apache Flink を使用して Flink アプリケーションをデプロイ、運用、スケーリングする方法について説明します。

  仮想ワークショップに参加する

Amazon MSK

- Amazon MSK の開始方法

メトリクスを使用して、Amazon MSK クラスターの作成、データの生成と消費、クラスターの状態のモニタリングを行う方法について説明します。

ガイドの使用を開始する

- Amazon MSK ワークショップ

  この実践的な Amazon MSK ワークショップで詳しく説明します。

  ワークショップの開始方法

Amazon MWAA

- Amazon MWAA の開始方法

  最初の MWAA 環境を作成し、DAG を Amazon S3 にアップロードして、最初のワークフローを実行する方法について説明します。

  チュートリアルを始める

- Amazon MWAA を使用したデータパイプラインの構築

  Glue、EMR、Redshift などの他の AWS 分析サービスをオーケストレーションするend-to-end のデータパイプラインを構築する方法について説明します。このブログ記事では、MWAA と Cosmos を使用して dbt Core ジョブをオーケストレーションするための、合理化された設定主導のアプローチについて説明します。ジョブは Amazon Redshift で変換を実行します。

  ブログ記事を読む

- Amazon MWAA ワークショップ

  データワークフローのオーケストレーションに Amazon MWAA をデプロイ、設定、使用する方法については、実践的な演習をご覧ください。

  ワークショップの開始方法

- Amazon MWAA のベストプラクティス

  分析ワークフローで Amazon MWAA を使用するためのアーキテクチャパターンとベストプラクティスについて説明します。

  ガイドを読む

OpenSearch Service

- OpenSearch Service の開始方法

  Amazon OpenSearch Service を使用してテストドメインを作成および設定する方法について説明します。

  チュートリアルを始める

- OpenSearch Service と OpenSearch Dashboards を使用したカスタマーサポートコールの視覚化

  次の状況の完全なウォークスルーについて説明します。ある企業がカスタマーサポートの電話を数回受け、分析したいと考えています。各問い合わせの件名は何でしょうか? 肯定的なやり取りの数はいくつでしょうか? 否定的なやり取りの数はいくつでしょうか? マネージャーはこれらの問い合わせのトランスクリプトをどのように検索または確認することができますか?

  チュートリアルを始める

- Amazon OpenSearch Serverless ワークショップの開始方法

  コンソールで AWS 新しい Amazon OpenSearch Serverless ドメインを設定する方法について説明します。利用可能なさまざまなタイプの検索クエリを調べ、目を引くビジュアライゼーションを設計し、割り当てられたユーザー権限に基づいてドメインとドキュメントを保護する方法について説明します。

  ワークショップの開始方法

- コスト最適化ベクトルデータベース: Amazon OpenSearch Service 量子化手法の概要

  OpenSearch Service がスカラーおよび製品量子化手法をサポートしてメモリ使用量を最適化し、運用コストを削減する方法について説明します。

  ブログ記事を読む

QuickSight

- QuickSight データ分析の開始方法

  最初の分析を作成する方法について説明します。サンプルデータを使用して、シンプルな分析またはより高度な分析を作成します。また、独自のデータに接続して分析を作成することも可能です。

ガイドを見る

- QuickSight を使用した視覚化

  ビジネスインテリジェンス (BI) とデータ可視化の技術的な側面について説明します AWS。ダッシュボードをアプリケーションやウェブサイトに埋め込み、アクセスとアクセス許可を安全に管理する方法について説明します。

  コースの開始方法

- QuickSight ワークショップ

  ワークショップで QuickSight ジャーニーをスタートする

  ワークショップの開始方法

Amazon Redshift

- Amazon Redshift Serverless の開始方法

  Amazon Redshift Serverless の基本的なフローを理解して、サーバーレスリソースの作成、Amazon Redshift Serverless への接続、サンプルデータのロード、データに対するクエリの実行を行います。

  ガイドを見る

- Amazon Redshift ディープダイブワークショップ

  Amazon Redshift プラットフォームの使用を開始するのに役立つ一連の演習をご覧ください。

  ワークショップの開始方法

Amazon S3

- Amazon S3 の開始方法

  最初の DataBrew プロジェクトを作成する方法について説明します。サンプルデータセットをロードし、そのデータセットで変換を実行し、それらの変換をキャプチャするレシピを構築し、変換されたデータを Amazon S3 に書き込むジョブを実行します。

  ガイドの使用を開始する

Amazon SageMaker

- Getting started with SageMaker

  プロジェクトの作成、メンバーの追加、サンプル JupyterLab ノートブックを使用して構築を開始する方法について説明します。

  ガイドを読む

- 次世代の Amazon SageMaker の紹介: すべてのデータ、分析、AI の中心

  データ処理、モデル開発、生成 AI アプリ開発を開始する方法について説明します。

  ブログを読む

- SageMaker Unified Studio とは

  SageMaker Unified Studio の機能、および Amazon SageMaker を使用する際のアクセス方法について説明します。

  ガイドを読む

- Amazon SageMaker のレイクハウスアーキテクチャの開始方法

  Amazon SageMaker でプロジェクトを作成し、ビジネスユースケースのデータを参照、アップロード、クエリする方法について説明します。

  ガイドを読む

- Amazon SageMaker のレイクハウスアーキテクチャのデータ接続

  レイクハウスアーキテクチャが、 AWS サービスとエンタープライズアプリケーション間のデータ接続を管理するための統一されたアプローチを提供する方法について説明します。

  ガイドを読む

- SageMaker レイクハウスアーキテクチャを使用した Athena フェデレーティッドクエリのカタログ化と管理

  Amazon SageMaker プロジェクトの Amazon Redshift、DynamoDB、Snowflake に保存されているデータに対してフェデレーティッドクエリに接続、管理、実行する方法について説明します。

  ブログを読む

AWS 分析サービスの使用方法を確認する

Editable architecture diagrams

リファレンスアーキテクチャ図

分析ソリューションの開発、スケーリング、テストに役立つアーキテクチャ図をご覧ください AWS。

分析リファレンスアーキテクチャの詳細

Ready-to-use code

注目のソリューション	AWS ソリューション
での Apache Druid を使用したスケーラブルな分析 AWS Apache Druid のセットアップ、運用、管理に役立つ構築 AWSされたコードをデプロイします。Apache Druid は、費用対効果が高く AWS、可用性が高く、回復力があり、耐障害性のあるホスティング環境です。 このソリューションの詳細	によって構築された、事前設定済みのデプロイ可能なソリューションとその実装ガイドについて説明します AWS。 すべての AWS セキュリティ、アイデンティティ、ガバナンスのソリューションを調べる

Documentation

分析ホワイトペーパー	AWS ビッグデータブログ
ホワイトペーパーを参照して、組織に最適な分析サービスの選択、実装、使用に関する詳細なインサイトとベストプラクティスを確認してください。 分析ホワイトペーパーを見る	特定のビッグデータのユースケースに対応するブログ記事をご覧ください。 AWS ビッグデータブログを詳しく見る

ドキュメント履歴

次の表に、この決定ガイドの重要な変更点を示します。このガイドの更新に関する通知については、RSS フィードをサブスクライブできます。

変更	説明	日付
re:Invent 更新	決定ガイド全体のリンクを更新し、Amazon Managed Workflows for Apache Airflow を追加しました。	2025 年 9 月 24 日
re:Invent 更新	決定ガイド全体で、Amazon SageMaker、Amazon SageMaker Unified Studio (プレビュー対象外) 、Amazon SageMaker Lakehouse への参照を更新しました。	2025 年 9 月 9 日
re:Invent 更新	SageMaker AI Unified Studio と を追加しました AWS Clean Rooms。ドキュメントを AI の新機能で更新しました。	2025 年 2 月 20 日
初版発行	ガイドが最初に公開されました。	2023 年 11 月 17 日