

JAPAN | 2024

aws SUMMIT



AWS-10

AWS で実現するデータガバナンス

大園 純平

アマゾン ウェブ サービス ジャパン合同会社
シニアアナリティクススペシャリストソリューションアーキテクト



Speaker

大園 純平 (おおぞの じゅんぺい)

シニアアナリティクススペシャリスト
ソリューションアーキテクト

Data & AI の事業開発組織に所属

お客様のデータガバナンスを含めた
データの取り組みを技術的に支援



Agenda

データガバナンスとその課題

AWS で実現するデータガバナンス

データガバナンスの進め方のポイント

まとめ

データガバナンスとその課題

データは戦略的資産である

97.0%

データに関する施策への
投資を増やしている企業

53.0%

データ戦略を策定している企業



出典: NewVantage Partners, DATA AND AI LEADERSHIP EXECUTIVE SURVEY 2022

2025 年までに
デジタルビジネスの拡大を目指す組織の

80%

が、データガバナンスに対する
最新のアプローチを採用していない
ために失敗する



出典: Gartner, Choose Adaptive Data Governance over One-Size-Fits-All for Greater Flexibility

データガバナンスはデータ戦略の最優先事項

データガバナンス

データ/分析/AI を基盤とする新しいビジネスイニシアチブの実現

データドリブンなカルチャーの促進

組織のための分析またはAIベースのインサイトの生成

リスクおよび規制コンプライアンス

データ/分析/AI 倫理

データの収益化

データ漏えいとハッキング防止

その他（具体的にお答えください）

不正防止



出典: Amazon Web Services, 2024 CDO Insights: Data & Generative AI

データガバナンスとは何か

漏えいや誤用を防ぐため

企業内のデータを

限られたアクセス範囲に
閉じ込めて守ること



ビジネス上の取り組みや運用を
サポートするため

企業内のデータを

適切な状態にしておくこと



出典: Amazon Web Services, What is Data Governance?
https://aws.amazon.com/what-is/data-governance/?nc1=h_ls

データガバナンスは イノベーションを加速するためのガードレール



適切なデータの
発見, アクセス, 共有



データを
保護してセキュア
に保つ



適切な
監査とコントロール
を可能にする

データガバナンスにおけるチャレンジ



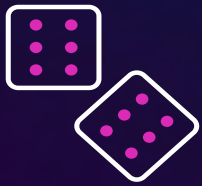
MORE DATA

データサイズと種類の増加
データ品質の定常的なモニタリングが必要



MORE SOURCES

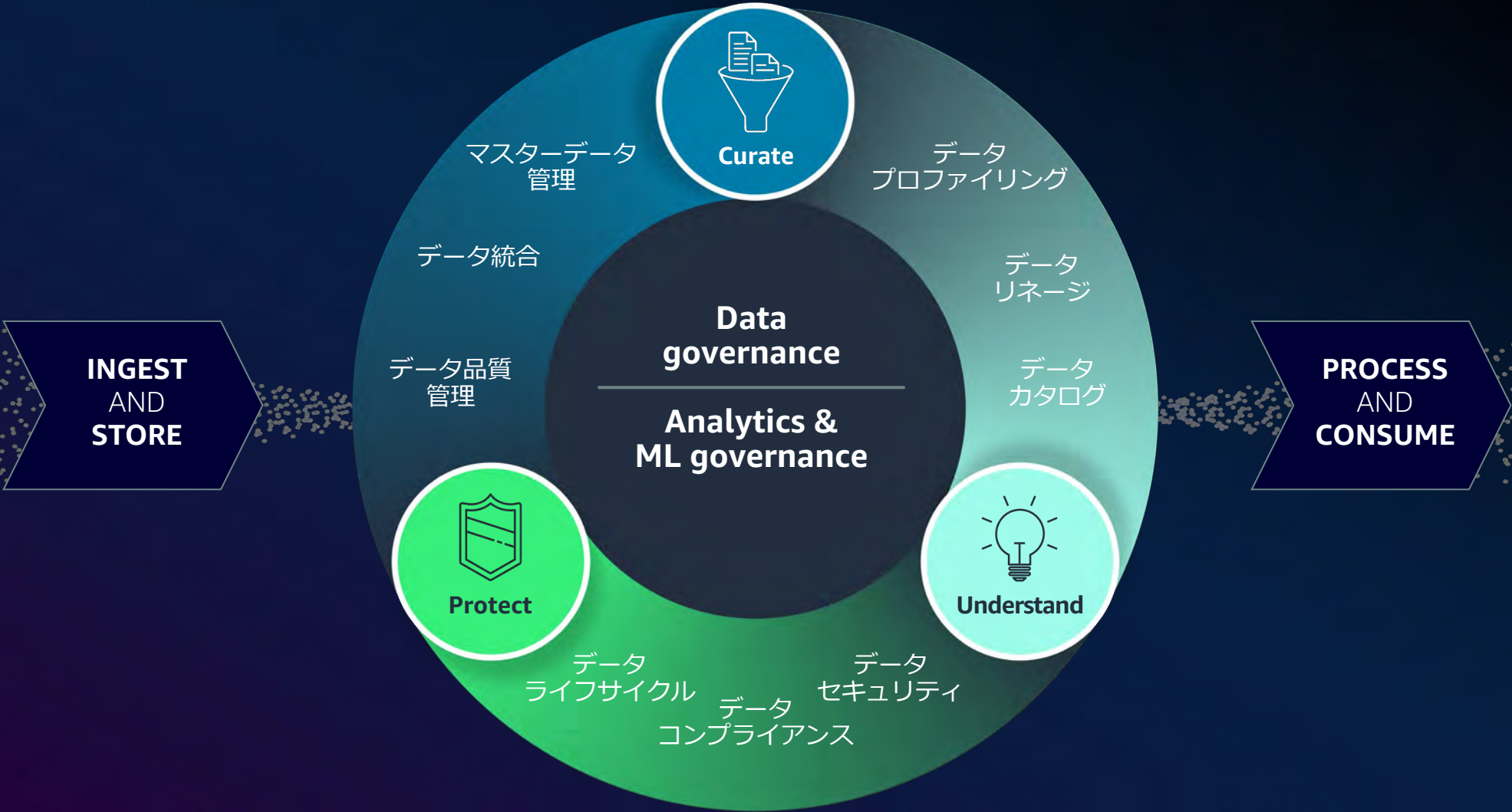
データが複数のシステムに分散され、アクセスに時間がかかる
ユーザーがすべてのデータを理解することは困難



MORE BUSINESS UNITS

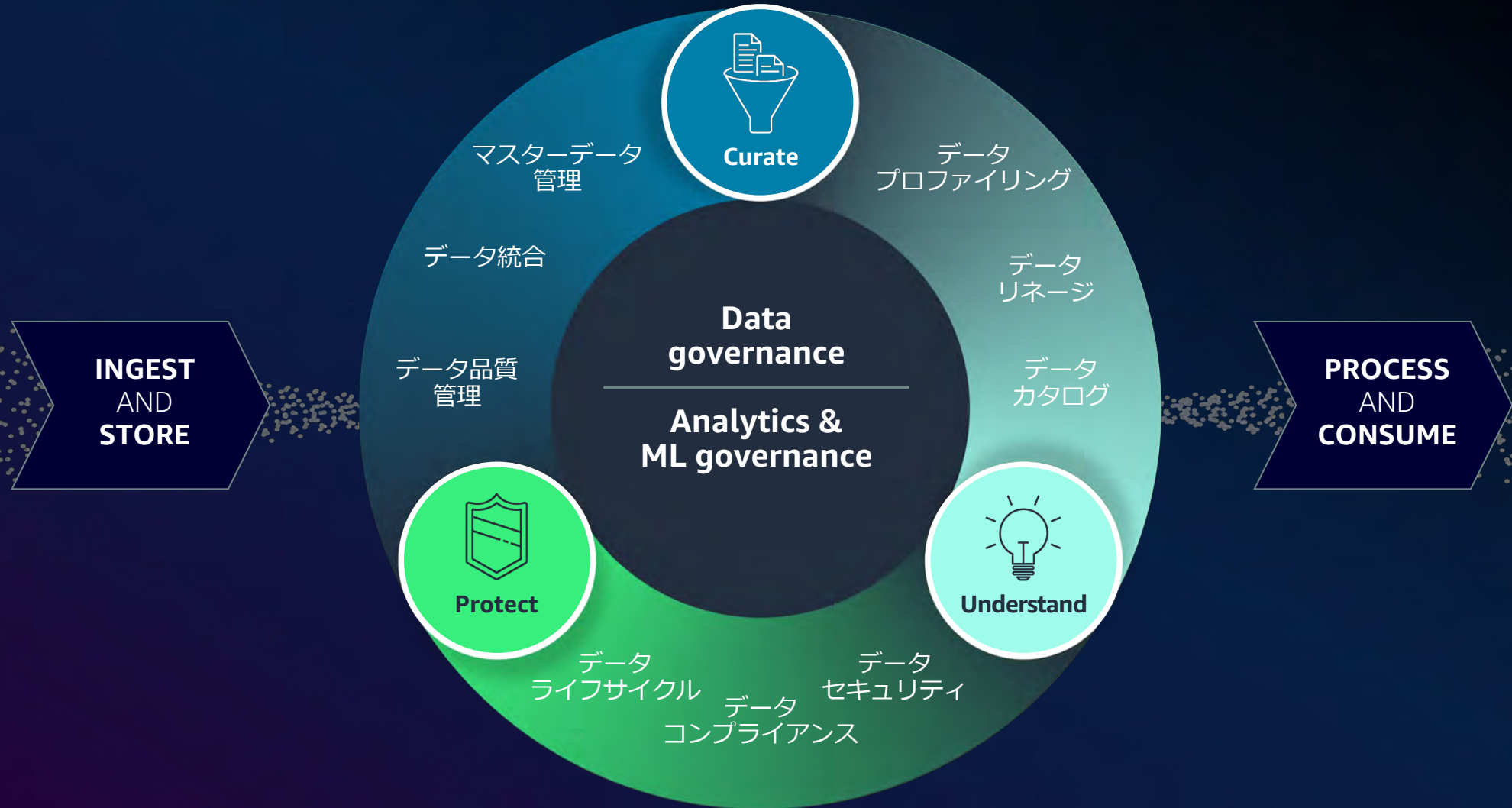
より多くのユーザーがより多くのユースケースでデータを使用
多くのメンバーはデータや利用ポリシーに関する知識を持たない

AWS におけるデータガバナンスフレームワーク









AWS で実現するデータガバナンス







AWS におけるデータガバナンスフレームワーク



AWS におけるデータガバナンスフレームワーク

3つの柱	キーワード	AWS サービス
 <p>Curate</p> <p>データを大規模に分類・整理し 管理および活用しやすい形にする</p>	<p>機密情報の特定と対処</p> <p>データ品質管理</p>	 <p>AWS Glue</p>
 <p>Understand</p> <p>ユーザーがコンテキストに基づいて データを発見, 理解, アクセス可能にする</p>	<p>メタデータ管理</p>	 <p>Amazon DataZone</p>
 <p>Protect</p> <p>データを適切に保護し安全に共有する</p>	<p>データセキュリティ管理</p>	 <p>Amazon DataZone</p>

AWS におけるデータガバナンスフレームワーク

3つの柱	キーワード	AWS サービス
 <p>データを大規模に分類・整理し 管理および活用しやすい形にする</p>	機密情報の特定と対処 データ品質管理	 <p>AWS Glue</p>
 <p>ユーザーがコンテキストに基づいて データを発見, 理解, アクセス可能にする</p>	メタデータ管理	 <p>Amazon DataZone</p>
 <p>データを適切に保護し安全に共有する</p>	データセキュリティ管理	 <p>Amazon DataZone</p>

データを分類・整理する目的

利用者が自信をもって
データをビジネスに
活用できるようにする



機密情報への対処
(個人情報, 企業機密事項 etc)



データ品質管理
(一貫性, 精度, 完全性, 整合性)

品質の悪いデータは深刻な結果をもたらす



Multicenter Study > [Pediatr Emerg Care](#). 2019 Sep;35(9):637-642.

doi: 10.1097/PEC.0000000000001277.

Pediatric Weight Errors and Resultant Medication Dosing Errors in the Emergency Department

Kristin M Hirata¹, Ann H Kang, Gina V Ramirez, Chieko Kimata, Loren G Yamamoto

Affiliations + expand

PMID: 28976456 DOI: 10.1097/PEC.0000000000001277

Abstract

Background: An accurate weight is critical for dosing medications in children. Weight errors can lead to medication-dosing errors.

Objectives: This study examined the frequency and consequences of weight errors occurring at 1 children's hospital and 2 general hospitals.

Methods: Using an electronic medical record database, 79,000 emergency department encounters of children younger than 5 years were analyzed. Extreme weights were first identified using weight percentiles. Encounters with potential weight errors were further evaluated using a retrospective chart review to determine whether a weight error and medication-dosing error occurred.

Results: The percentage of weight errors of total encounters at all 3 institutions was low (0.63% on average), but a large proportion of weight errors led to subsequent medication-dosing errors (34% on average). The children's hospital did not have clinically significantly lower occurrences of weight errors or weight-based medication errors. Common weight errors included the weight in pounds being substituted for the weight in kilograms and decimal placement errors.

Conclusions: Weight errors were uncommon at the 3 emergency departments that we studied, but they led to weight-based medication-dosing errors that had the potential to cause harm.

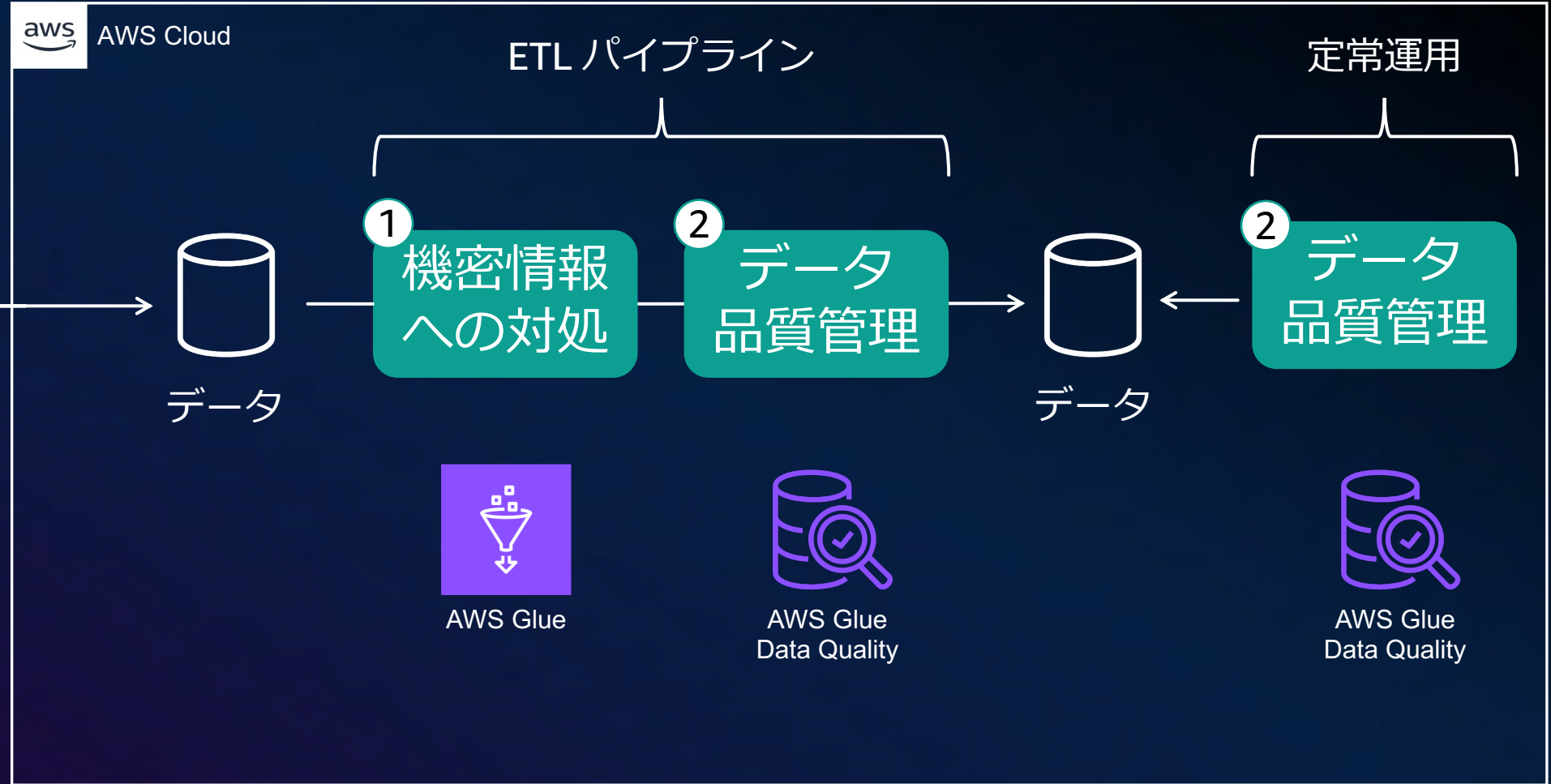
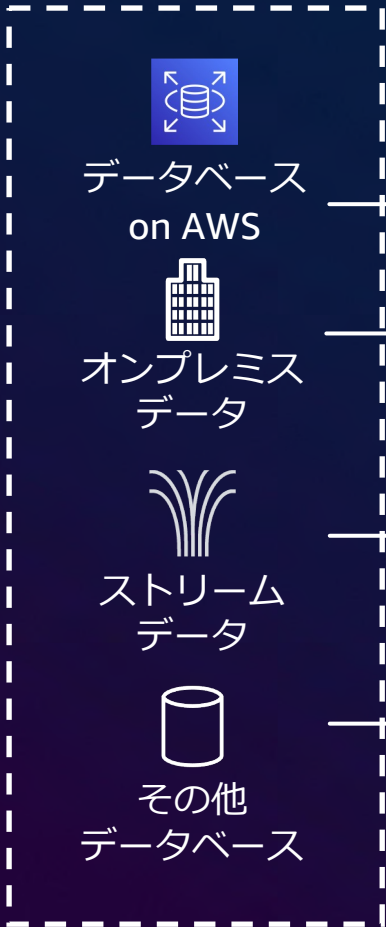
- 患者の体重記録の誤りや欠損は投薬量の誤りにつながる可能性がある
- 5歳未満の小児を対象とした79,000件の救急外来で発生した体重記録ミスの頻度とその結果について調査
- 体重記録ミスの発生は比較的まれ(0.63%)であったが、体重の記録ミスの多くはその後の投薬ミスにつながった(34%)

出展: Hirata, Kang, Kimata, Ramirez, and Yamamoto, [NIH National Library of Medicine](#)

機密情報への対処とデータ品質管理



データソース



AWS Glue の特徴

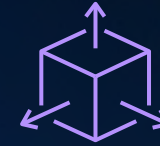
サーバーレスデータインテグレーションサービス



AWS Glue



データカタログ
(テクニカルメタデータ)



サーバーレス ETL エンジン



スキーマの自動収集



機密データ検知など
豊富なビルトイン変換機能



他システム連携の
豊富なコネクタ



データ品質管理

参考セッション

AWS-11: 生成 AI が変える、データアナリティクス
AWS-41: 実践! AWS Glue の ETL 開発運用ベストプラクティス

① AWS Glue における機密データ検知

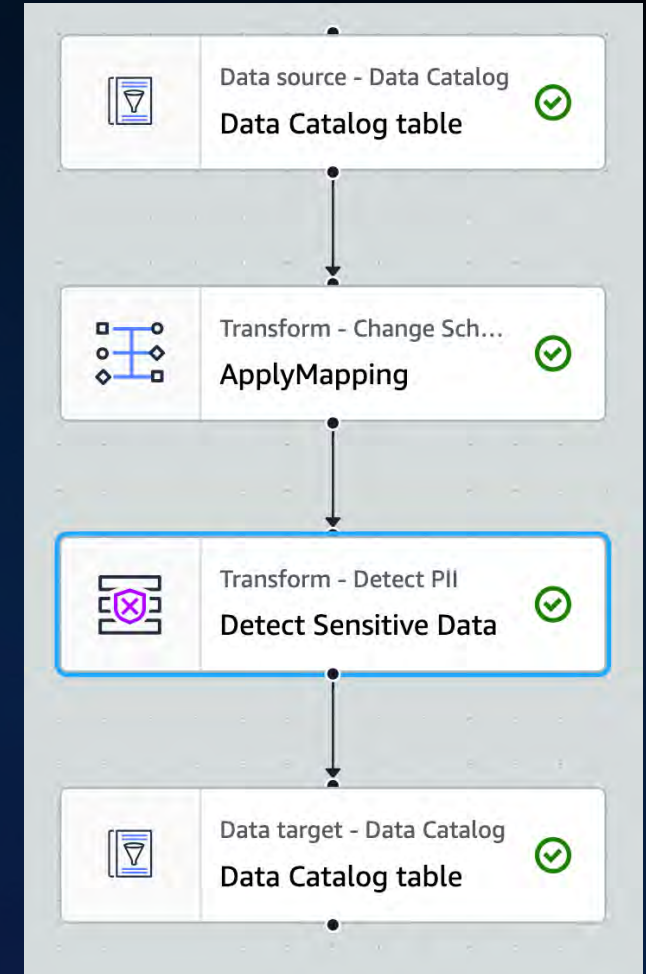
データソースをスキャンし、個人識別情報 (PII) を自動識別

- 識別された PII は自動削除または **** など固定文字列へ変換/マスク可能
- 検出可能なビルトインの PII パターン例

日本独自のデータタイプ

- JAPAN_BANK_ACCOUNT (銀行口座)
- JAPAN_DRIVING_LICENSE (運転免許番号)
- JAPAN_MY_NUMBER (マイナンバー)
- JAPAN_PASSPORT_NUMBER (パスポート番号)

- ビルトインの検出パターンの他正規表現でカスタムパターンも定義可能



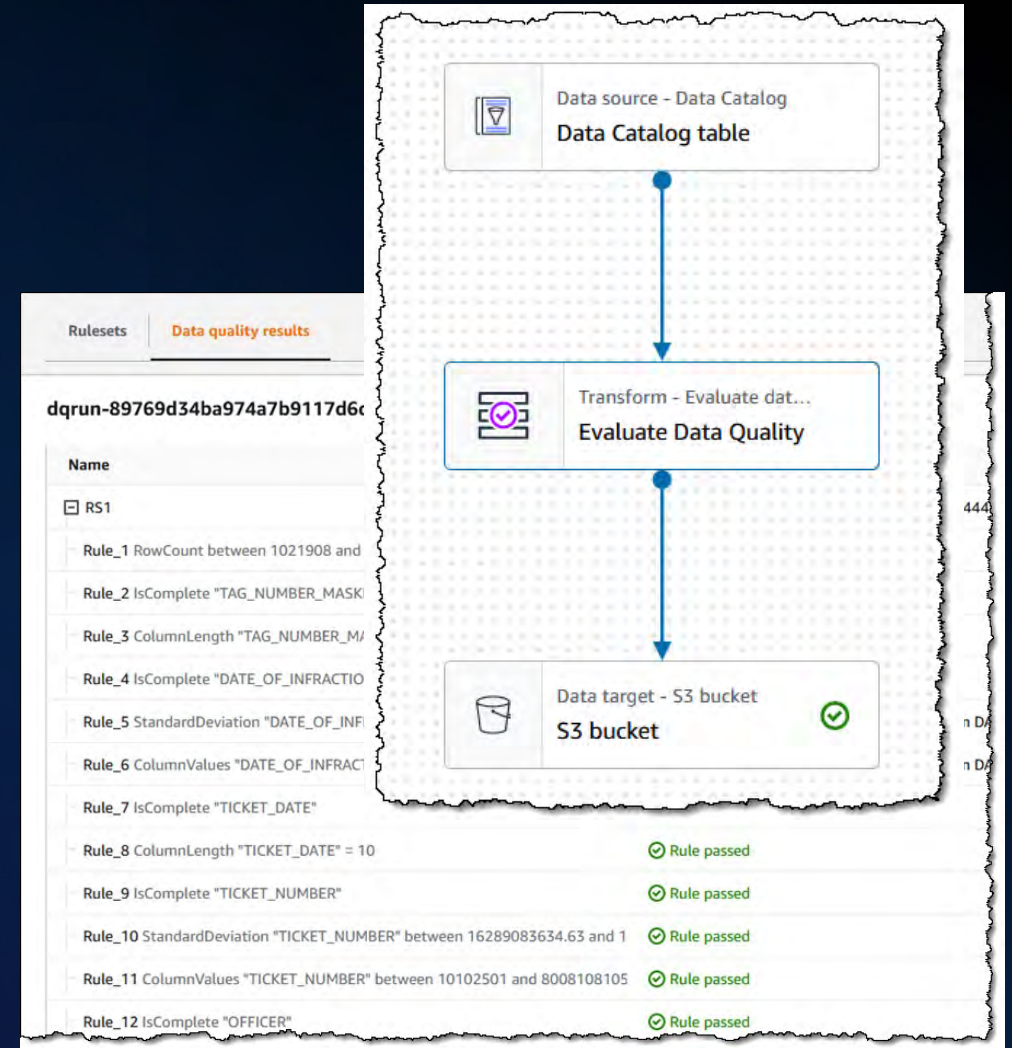
参考: <https://docs.aws.amazon.com/glue/latest/dg/sensitive-data-managed-data-types.html>

② Glue Data Quality によるデータ品質管理



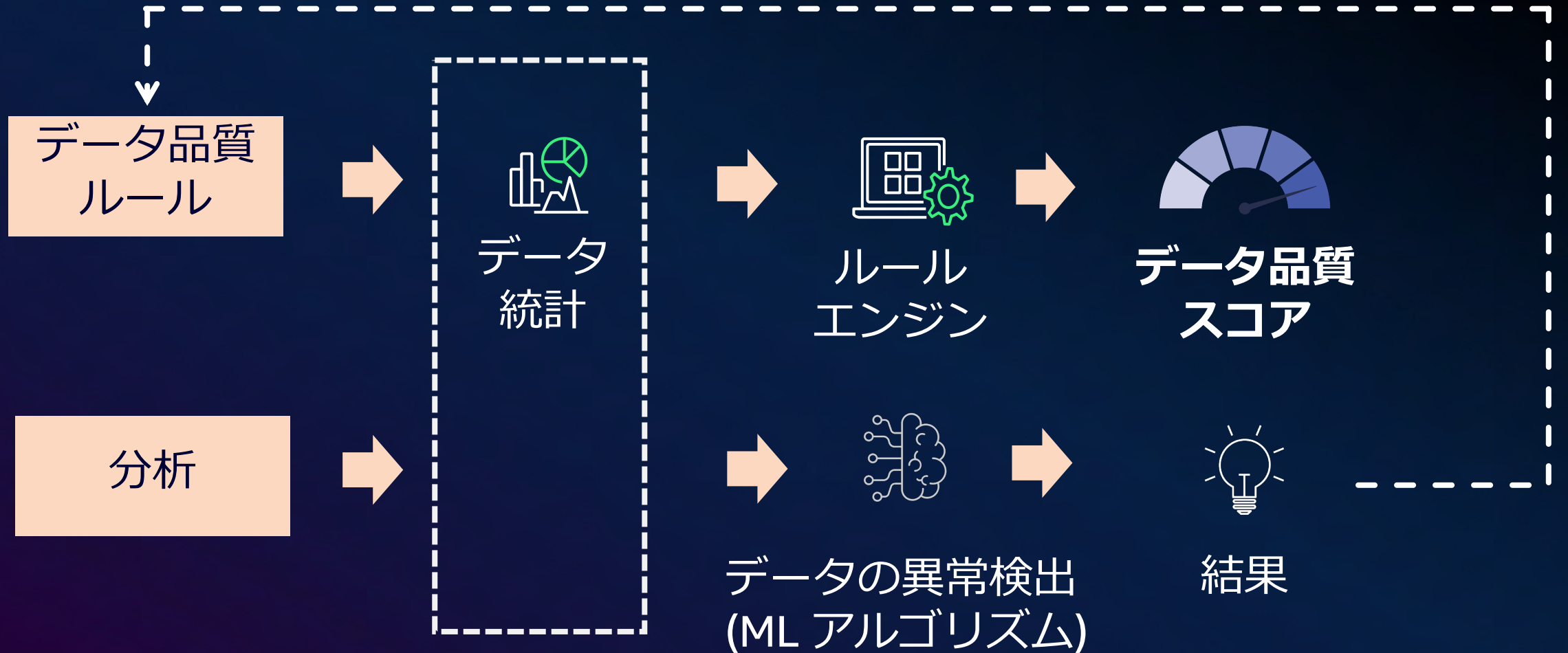
データの品質を自動的に測定・モニタリング・管理

- データを自動的に分析しデータに関する統計を取得し, データ品質ルールを推薦
- 推薦されたルールまたは手動定義したルールに基づきデータ品質をモニタリング
- ETL パイプラインの各ステップでルール定義と警告を設定可能
- データカタログ上のテーブルに対してデータ品質ルールを定義することで定常的にデータの状態を監視することも可能



参考: <https://aws.amazon.com/jp/glue/features/data-quality/>

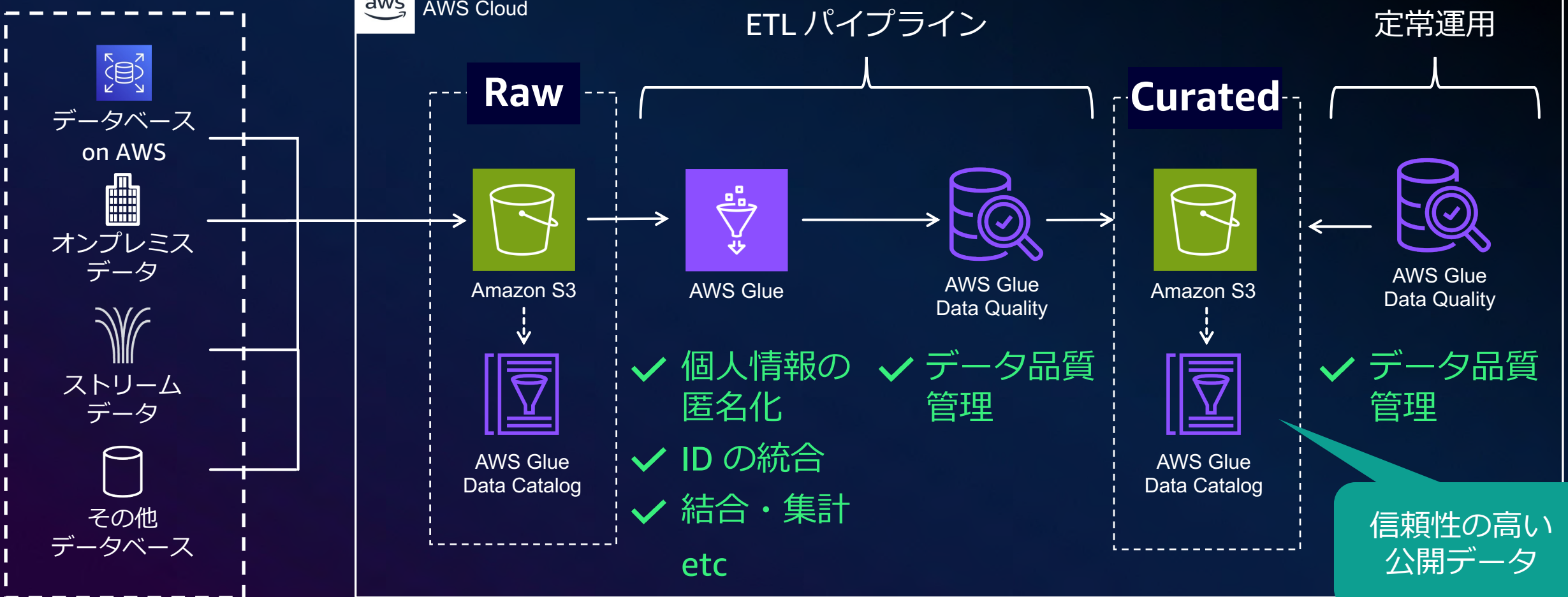
機械学習でデータ品質ルールを段階的に強化









公開に適した高品質なデータを整備する



データソース



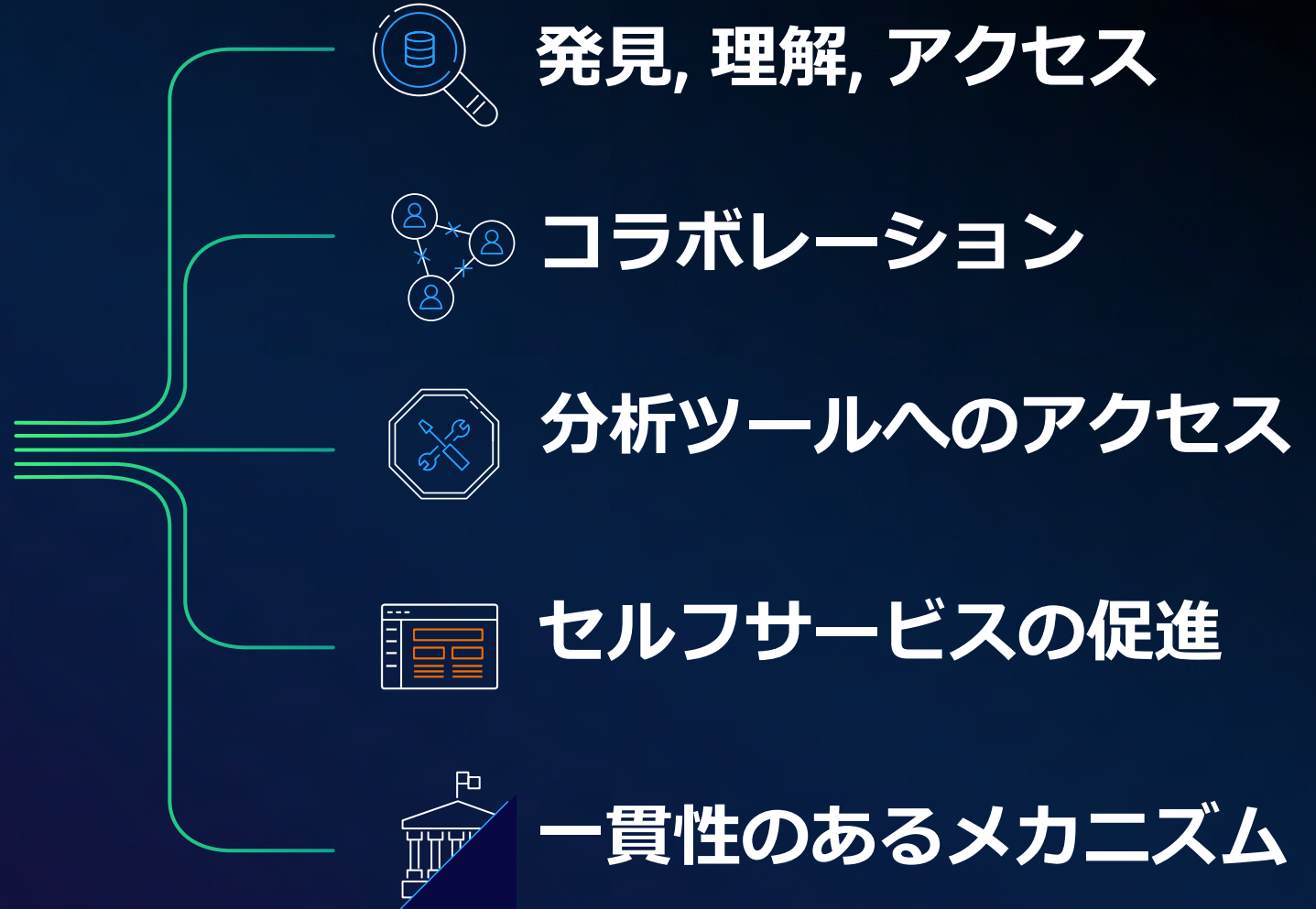
AWS におけるデータガバナンスフレームワーク

3つの柱	キーワード	AWS サービス
 <p>データを大規模に分類・整理し 管理および活用しやすい形にする</p>	機密情報の特定と対処 データ品質管理	 <p>AWS Glue</p>
 <p>ユーザーがコンテキストに基づいて データを発見, 理解, アクセス可能にする</p>	メタデータ管理	 <p>Amazon DataZone</p>
 <p>データを適切に保護し安全に共有する</p>	データセキュリティ管理	 <p>Amazon DataZone</p>

必要なデータは整った, では次は?



求められている
データガバナンスの
ソリューション





Amazon DataZone

ビルトインのガバナンスにより組織の境界を越えてデータを活用する



組織の境界を越えた
データアクセス管理



データとツールを
共有して関係者を
つなぎ、ビジネス
インサイトを促進

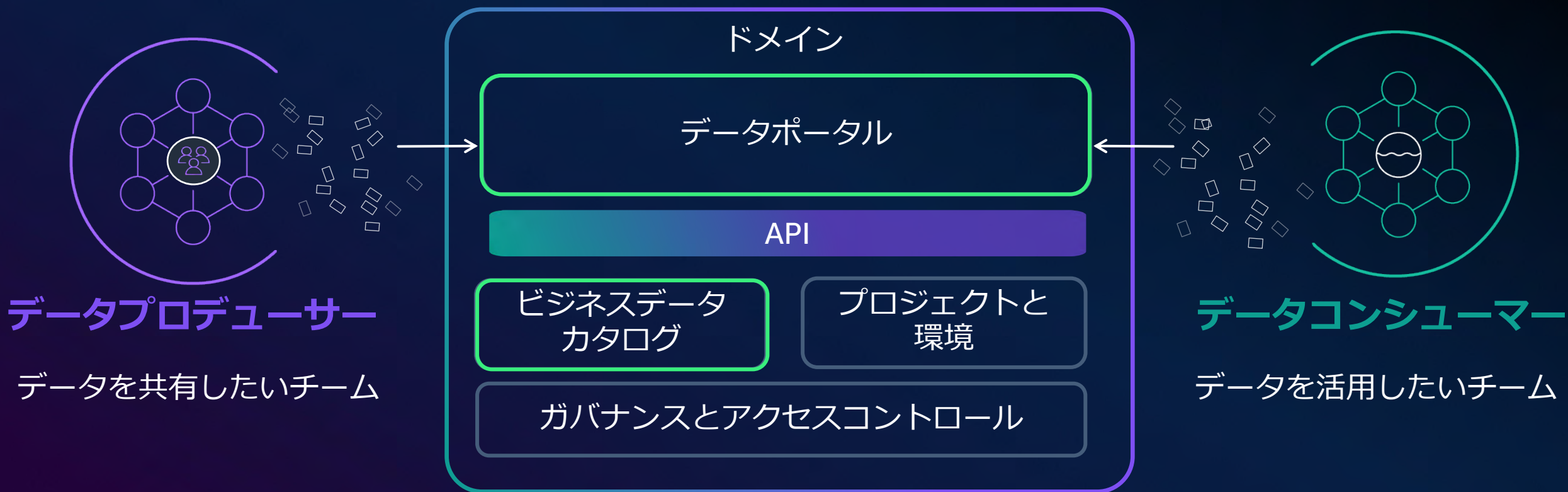


生成 AI でデータの発見
とカタログ作成を自動化

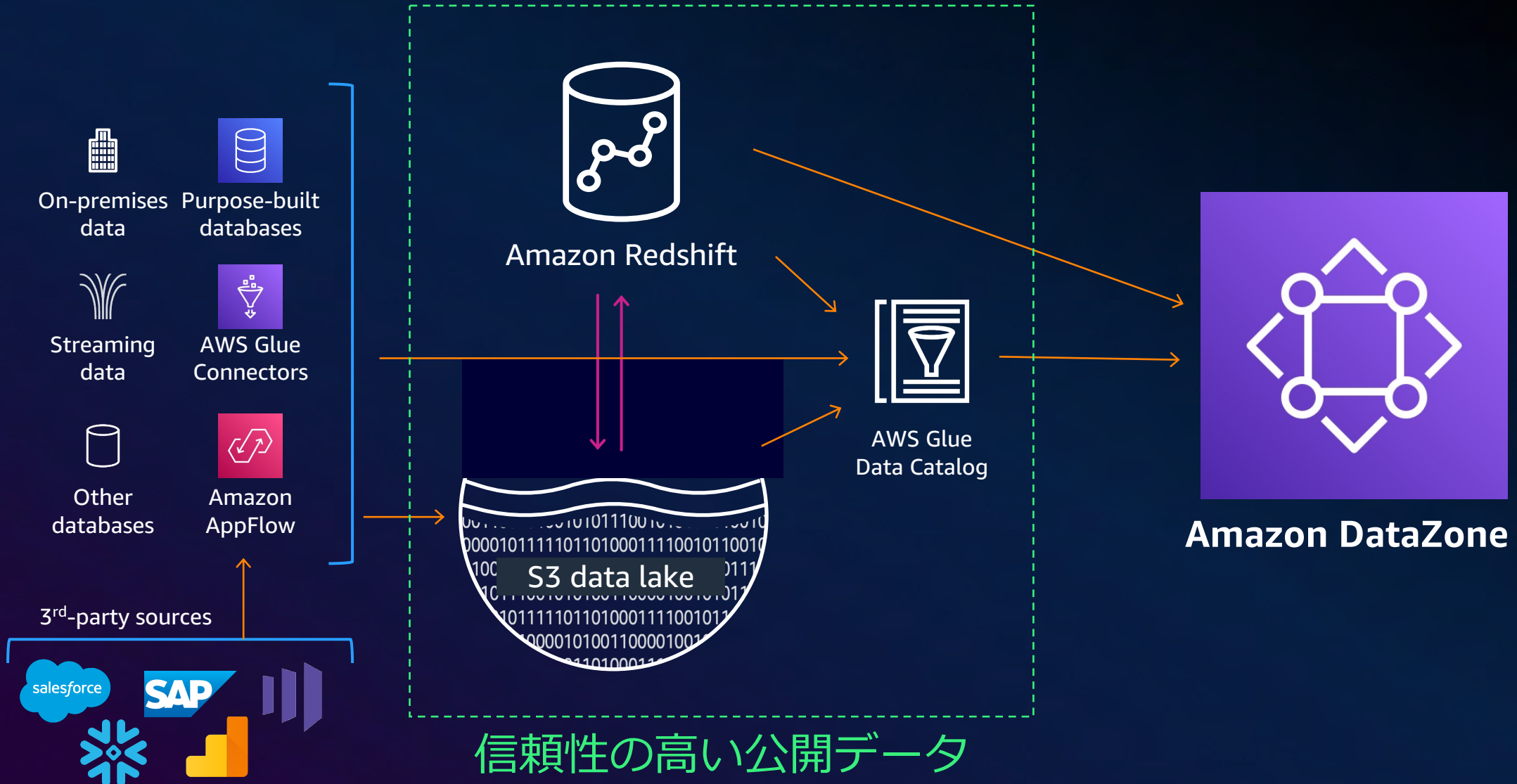
Amazon DataZoneのコアコンポーネント



Amazon DataZone



すべてのデータ資産をメタデータとして登録





メタデータにおける代表的な要件

要件

データがどこにあるか分からないので検索したい

データの意味とそれがどのように分析に活用できるか知りたい

データが信頼できる品質のデータなのかを知りたい

Amazon DataZone が提供する機能・特徴

データポータル

ビジネスメタデータ機能

データ品質連携機能



メタデータにおける代表的な要件

要件

データがどこにあるか分からないので検索したい

データの意味とそれがどのように分析に活用できるか知りたい

データが信頼できる品質のデータなのかを知りたい

Amazon DataZone が提供する機能・特徴

データポータル

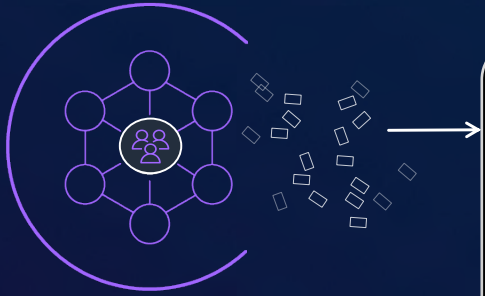
ビジネスメタデータ機能

データ品質連携機能

データポータル機能



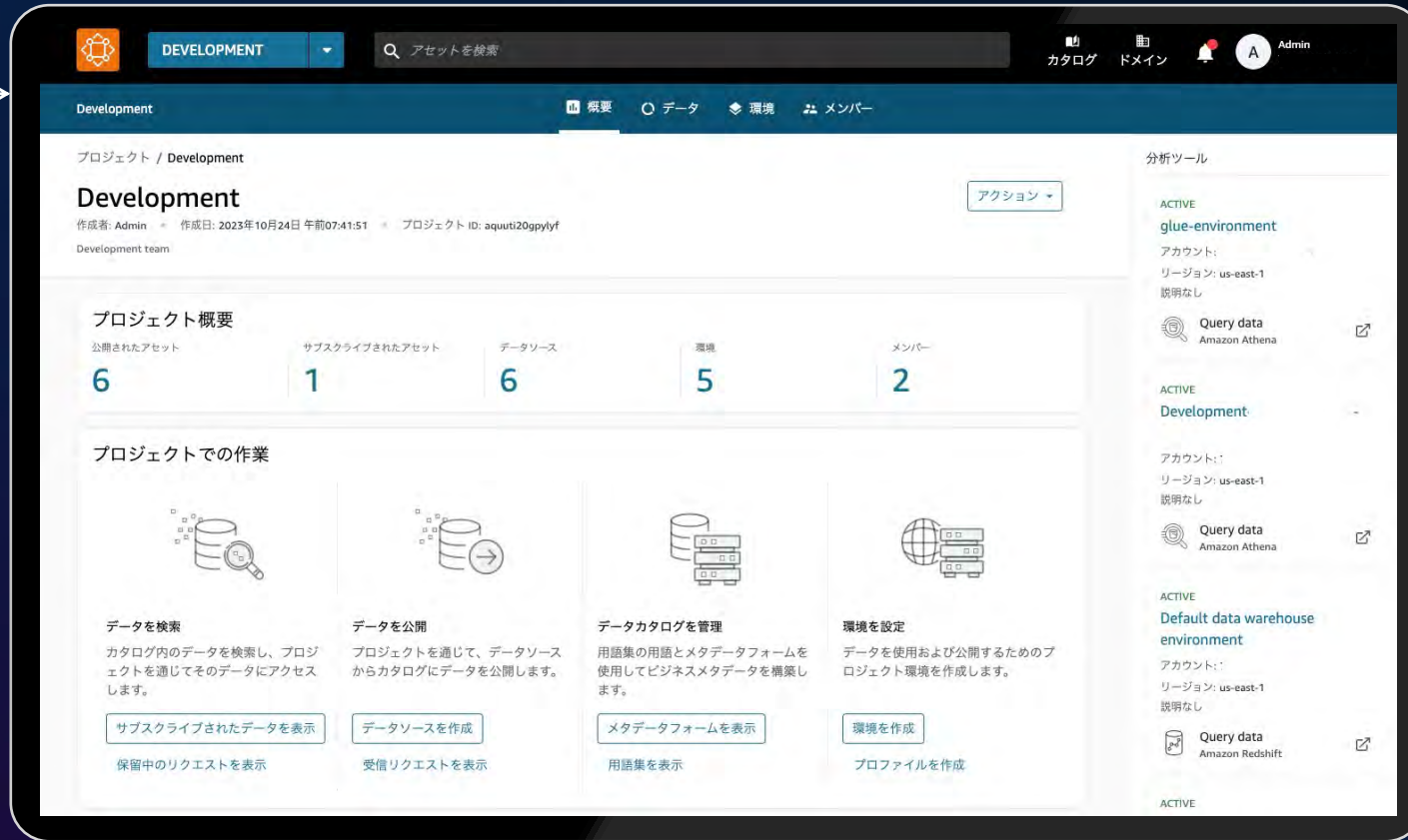
データ活用に関わるアクション・操作を行うための統合環境を提供



データプロデューサー

データを共有したいチーム

既存のテーブル
などアセットを
メタデータ
として登録



データコンシューマー

データを活用したいチーム

メタデータを検索,
発見し,アクセス
申請, データ分析
までを実行可能

メタデータにおける代表的な要件



要件

Amazon DataZone が提供する機能・特徴

データがどこにあるか分からないので検索したい

データポータル

データの意味とそれがどのように分析に活用できるか知りたい

ビジネスメタデータ機能

データが信頼できる品質のデータなのかを知りたい

データ品質連携機能

ビジネスメタデータ機能



必要なビジネスメタデータを、データ資産に付与することが可能



テーブルやカラムのビジネスラベル、説明、利用上の注意 (README) など



データソースから自動的に取得される
テクニカルメタデータ



データ資産に付加するメタデータである
ビジネス用語集とメタデータフォーム

The screenshot shows the AWS Glue console interface. At the top, there's a navigation bar with the Glue logo, a 'DEVELOPMENT' environment selector, and a search bar labeled 'アセットを検索'. Below this, the page title is 'Development' with a '概要' (Summary) button. The main content area shows a breadcrumb '← 戻る' and the title 'Amazon Reviews Table Named Resource'. Below the title, it lists technical details: '技術名: amazon_reviews_table_named_resource' and 'アセットタイプ: Glue Table'. A description reads 'Amazon レビューに関するテーブル'. There are several tabs: 'ビジネスメタデータ' (Business Metadata), 'スキーマ', 'データ品質', 'サブスクリプションリクエスト', and 'マイサブスクリプション'. The 'ビジネスメタデータ' tab is selected. Under this tab, there are sections for '概要' (Summary) and 'README'. The '概要' section contains the text 'このテーブルは Amazon レビューテーブルです。'. The 'README' section is titled '利用上の注意' (Usage Notes) and contains the text 'このデータを使って分析をする際には、情報の取り扱い区分に沿って適切なご利用をお願いします。'.

ビジネスメタデータ機能



必要なビジネスメタデータを、データ資産に付与することが可能



テーブルやカラムのビジネスラベル、説明、利用上の注意 (README) など



データソースから自動的に取得される
テクニカルメタデータ



データ資産に付加するメタデータである
ビジネス用語集とメタデータフォーム

The screenshot shows the AWS Glue console interface. At the top, there's a navigation bar with the AWS logo, a 'DEVELOPMENT' dropdown menu, and a search bar containing 'アセットを検索'. Below this, the page title is 'Development' with a '概要' (Summary) button. The main content area shows a breadcrumb trail '←戻る' and the title 'Amazon Reviews Table Named Resource'. Below the title, it lists technical details: '技術名: amazon_reviews_table_named_resource' and 'アセットタイプ: Glue Table'. A description reads 'Amazon レビューに関するテーブル'. A horizontal menu below the description includes 'ビジネスメタデータ', 'スキーマ', 'データ品質', 'サブスクリプションリクエスト', and 'マイサブスクリ'. Two sections are highlighted with orange boxes: '概要' (Summary) with the text 'このテーブルは Amazon レビューテーブルです。' and 'README' with the heading '利用上の注意' (Usage Notes) and the text 'このデータを使って分析をする際には、情報の取り扱い区分に沿って適切なご利用をお願いします。'.

ビジネスメタデータ機能

必要なビジネスメタデータを、データ資産に付与することが可能



テーブルやカラムのビジネスラベル、説明、利用上の注意 (README) など



データソースから自動的に取得される
テクニカルメタデータ



データ資産に付加するメタデータである
ビジネス用語集とメタデータフォーム

用語集の用語 [これは何ですか?](#)

極秘 - 社内



メタデータフォーム [これは何ですか?](#)

AWS Glue テーブル

Glue データカタログ ID	[REDACTED]
場所	s3://[REDACTED]
リージョン	us-east-1
分類	parquet
テーブル ARN	arn:aws:glue:us-east-1:[REDACTED]
テーブル名	amazon_reviews_table_named_r
プライマリキー	product_category

Amazon Review 用メタデータフォーム

フィールド名	値
情報の取り扱い区分	極秘 - 社内
国外ユーザーからの参照可否	はい

ビジネスメタデータ機能

必要なビジネスメタデータを、データ資産に付与することが可能



テーブルやカラムのビジネスラベル、説明、利用上の注意 (README) など



データソースから自動的に取得される
テクニカルメタデータ



データ資産に付加するメタデータである
ビジネス用語集とメタデータフォーム



用語集の用語 [これは何ですか?](#)

極秘 - 社内

メタデータフォーム [これは何ですか?](#)

AWS Glue テーブル

Glue データカタログ ID	[REDACTED]
場所	s3://[REDACTED]
リージョン	us-east-1
分類	parquet
テーブル ARN	arn:aws:glue:us-east-1:[REDACTED]
テーブル名	amazon_reviews_table_named_r
プライマリキー	product_category

Amazon Review 用メタデータフォーム

フィールド名	値
情報の取り扱い区分	極秘 - 社内
国外ユーザーからの参照可否	はい

ビジネスメタデータ生成の自動化



手動分類・整理



分類・整理に時間がかかり
手作業による負荷が高いと
不完全なメタデータに
なってしまふことが多い

生成 AI ベースの 自動分類・整理



メタデータの自動生成により
作業負担を抑え、より迅速に
充実したメタデータを整備できる

データ資産のサマリー, 活用事例の自動生成



ボトルネックになりがちなメタデータの整備を生成AIの力で大きく促進

- バックグラウンドで Amazon Bedrock の LLM (大規模言語モデル) を利用し, わかりやすいテーブルとカラムの名前と説明を自動的に生成

Amazon Reviews Table Named Resource

技術名: amazon_reviews_table_named_resource • アセットタイプ: Glue Table

説明なし

ビジネスメタデータ • スキーマ データ品質 サブスクリプションリクエスト マイサブスクリプション 履歴

🔍 フィルター

名前	データタイプ	説明
Marketplace marketplace	string	The marketplace where the review was written
Customer ID customer_id	string	The identifier for the customer who wrote the review
Review ID review_id	string	The unique ID of the review



Amazon Bedrock

データ資産のサマリー, 活用事例の自動生成



ボトルネックになりがちなメタデータの整備を生成AIの力で大きく促進

- バックグラウンドで Amazon Bedrock の LLM (大規模言語モデル) を利用し, わかりやすいテーブルとカラムの名前と説明を自動的に生成
- さらに, データ資産の概要や活用例 (ユースケース) といった人手の登録が困難な情報も生成 AI の力を活用し, 自動的に生成

Amazon Reviews Table Named Resource

技術名: amazon_reviews_table_named_resource • アセットタイプ: Glue Table

説明なし

[ビジネスメタデータ](#) • [スキーマ](#) [データ品質](#) [サブスクリプションリクエスト](#) [マイサブスクリプション](#) [履歴](#)

概要

The amazon_reviews_table_named_resource table contains Amazon customer review data. Each row represents a single customer review product information, and customer details.

The review_id column contains the unique ID for each review. Details on the actual review are contained in columns like star_rating, review_text, and total_votes. The product being reviewed is identified by the product_id and product_title columns. Additional product details are in the product_information columns.

Each review is connected to the customer that wrote it via the customer_id column. The marketplace and year provide context on when a review was made. The verified_purchase and vine columns indicate review authenticity details.

Overall, this table enables analysis of Amazon review data to identify insights into products, reviews, customers, voting patterns, and how they change over time. The detailed review content and metadata create a rich dataset for mining consumer feedback and sentiments.

Use Case

The amazon_reviews_table_named_resource contains rich customer review data that can enable several valuable use cases for different users.

Product Teams: Analyzing reviews by product_id, star_rating, helpful_votes, and review content in columns like review_headline and review_text can help identify product strengths and areas for improvement to guide development. Review trends over time (year column) can reveal changing consumer sentiments.

Marketing Teams: The detailed review data can assist marketing professionals in better positioning and messaging products based on what customers value. Identifying influential reviewers via total_votes and helpful_votes can also be useful.



メタデータにおける代表的な要件

要件

Amazon DataZone が提供する機能・特徴

データがどこにあるか分からないので検索したい

データの意味とそれがどのように分析に活用できるか知りたい

データが信頼できる品質のデータなのかを知りたい

データポータル

ビジネスメタデータ機能

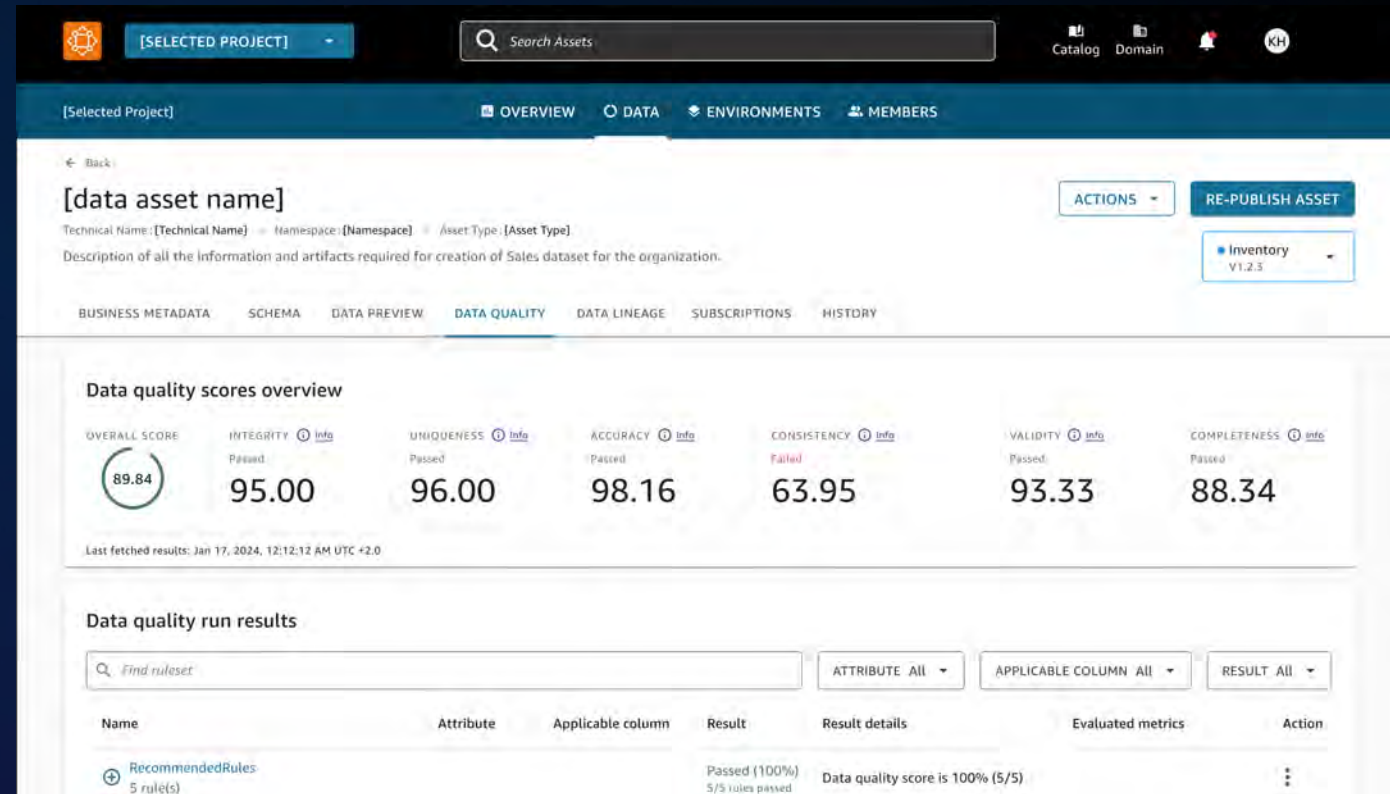
データ品質連携機能

データ品質連携機能









データ資産のデータ品質を高めて活用を加速

- 公開されたデータ資産の詳細に表示されるデータ品質メトリクス
- AWS Glue Data Quality や 3rd Party のデータ品質ツールからのメトリクスの表示を自動化
- データ利用者はデータポータルで最新のデータ品質スコアを確認しデータ活用判断に利用できる



AWS におけるデータガバナンスフレームワーク

3つの柱	キーワード	AWS サービス
 <p>Curate</p> <p>データを大規模に分類・整理し 管理および活用しやすい形にする</p>	<p>機密情報の特定と対処 データ品質管理</p>	 <p>AWS Glue</p>
 <p>Understand</p> <p>ユーザーがコンテキストに基づいて データを発見, 理解, アクセス可能にする</p>	<p>メタデータ管理</p>	 <p>Amazon DataZone</p>
 <p>Protect</p> <p>データを適切に保護し安全に共有する</p>	<p>データセキュリティ管理</p>	 <p>Amazon DataZone</p>

Amazon DataZoneのコアコンポーネント



Amazon DataZone



データプロデューサー
データを共有したいチーム

データコンシューマー
データを活用したいチーム

組織間を越えたデータセキュリティ管理の難しさ



“分析に必要なセールスとクリックストリームのデータはどこで入手できるのか？”



“誰がこの財務データにアクセスさせてくれるのか？”



“販売データにアクセスするためのポリシーに従っているか確認したい”



“このデータをデータ分析ツールに接続するには？”



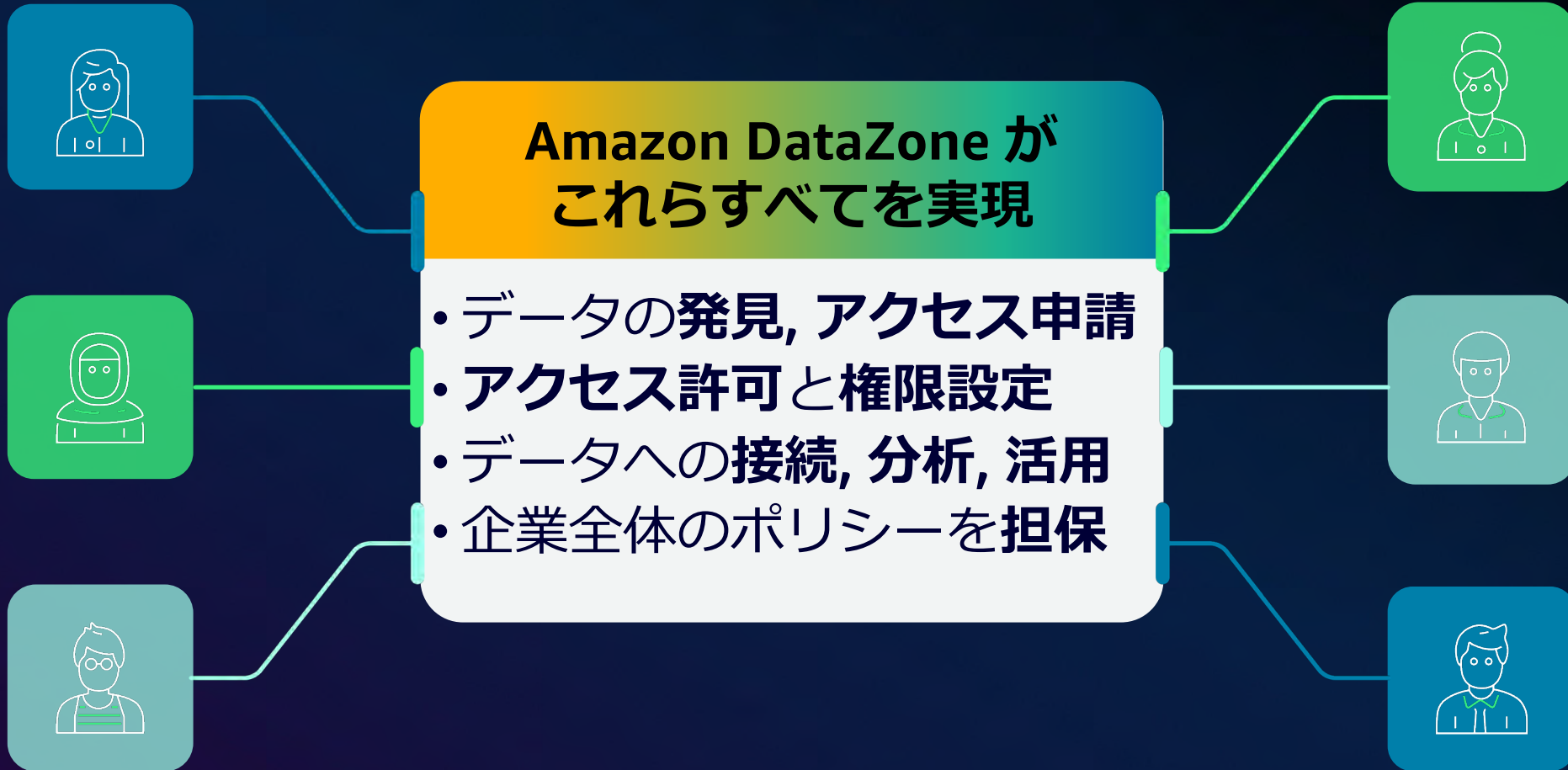
“誰がこの販売データにアクセスできるのか？”



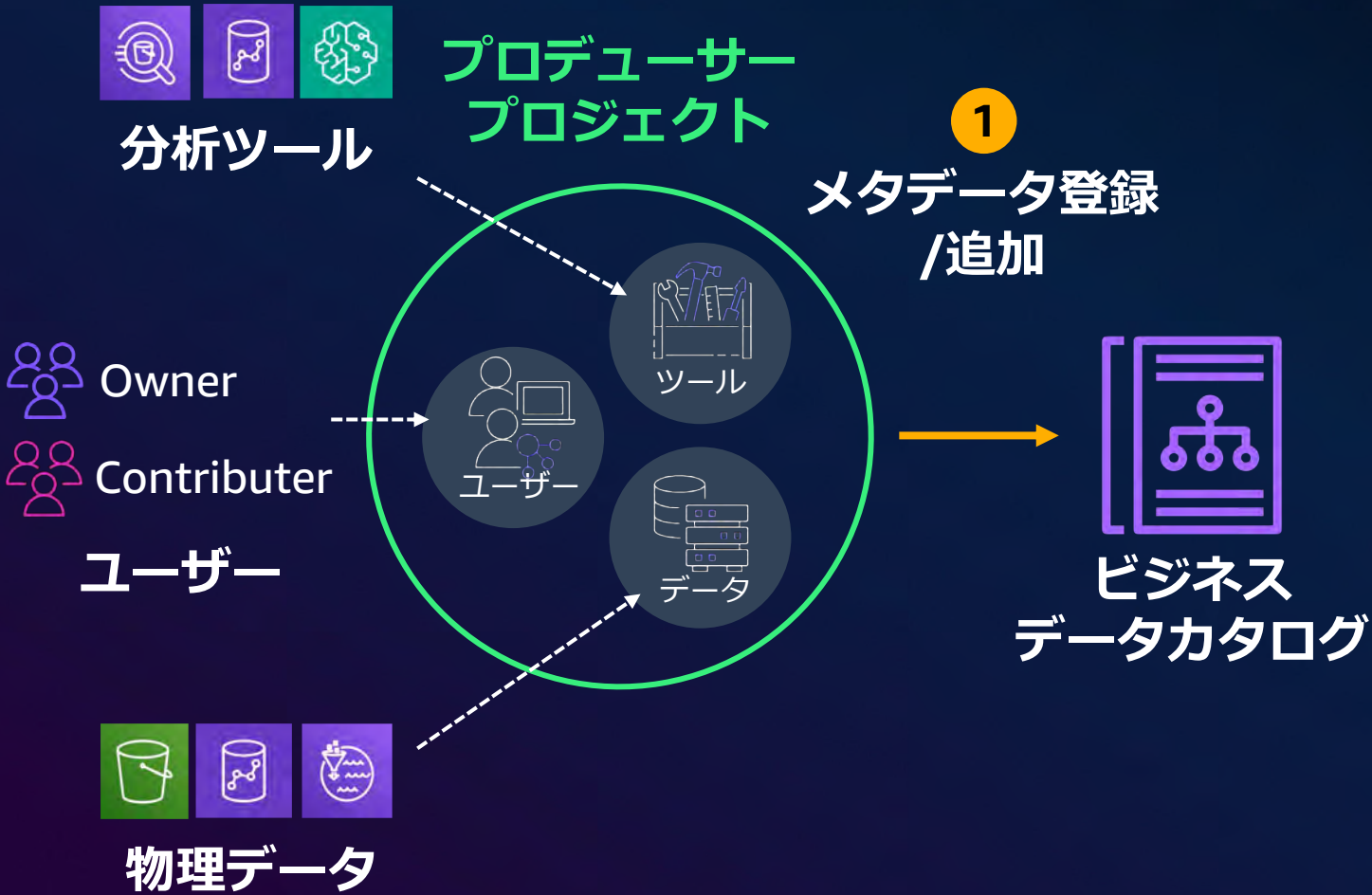
“この分析はどこでチームと共有できるのか？”



イノベーションを加速するアクセスコントロールモデル









メタデータ登録のワークフロー (プロデューサー)



アクセス申請のワークフロー (コンシューマー)



AWS におけるデータガバナンスフレームワーク

3つの柱	キーワード	AWS サービス
 <p>Curate</p> <p>データを大規模に分類・整理し 管理および活用しやすい形にする</p>	<p>機密情報の特定と対処 データ品質管理</p>	 <p>AWS Glue</p>
 <p>Understand</p> <p>ユーザーがコンテキストに基づいて データを発見, 理解, アクセス可能にする</p>	<p>メタデータ管理</p>	 <p>Amazon DataZone</p>
 <p>Protect</p> <p>データを適切に保護し安全に共有する</p>	<p>データセキュリティ管理</p>	 <p>Amazon DataZone</p>

データガバナンスの進め方のポイント

フレームワークやテクノロジーだけでは
データガバナンスを浸透させていくことは難しい

データガバナンスを実現するための主要な役割



データスチュワード

ビジネスニーズに合わせてデータを管理・活用推進していく役割

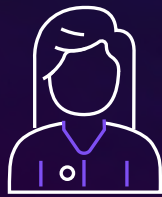
- データスチュワードは作るものではなく、組織の中から見つける
- 正式に任命し、役割を与えることが重要



データオーナー

データを所有する管理者

- データに関するポリシーの策定などを行う
- データスチュワードと緊密に連携する

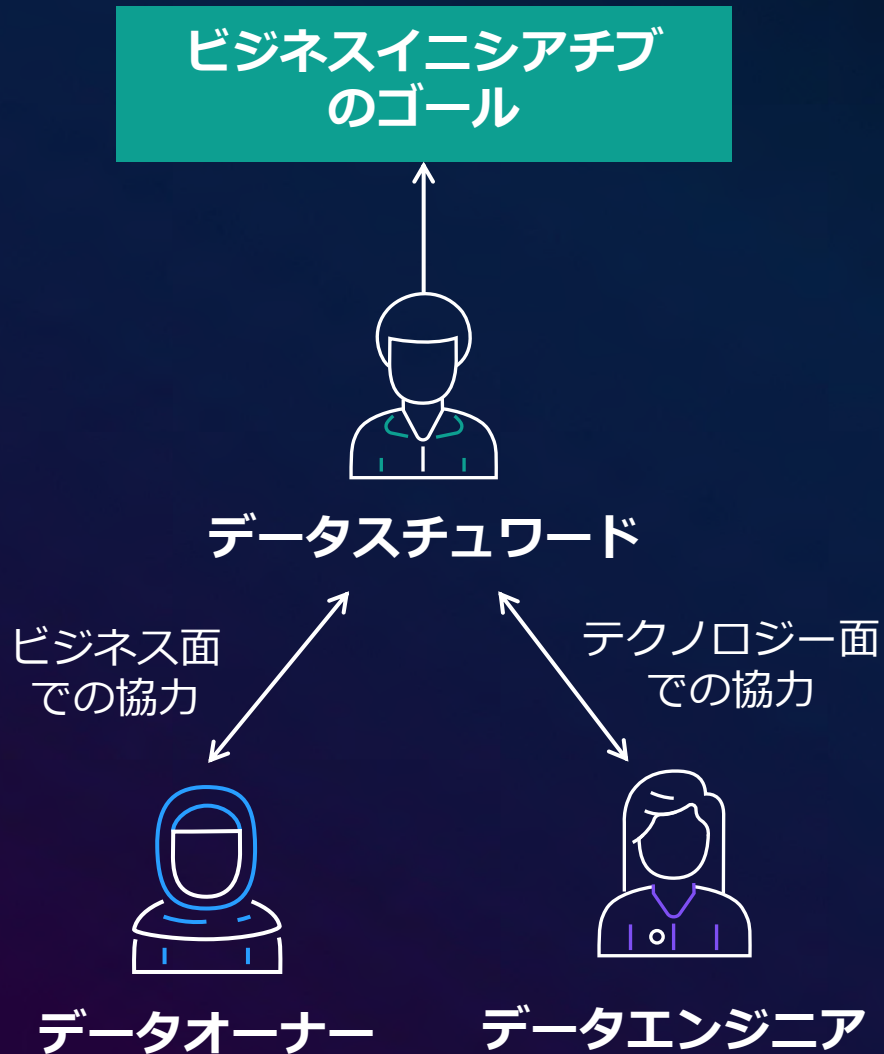


データエンジニア

データの変換と活用をサポート

- データスチュワードをテクノロジー観点でサポートする

ビジネスイニシアチブのゴールから逆算して始める

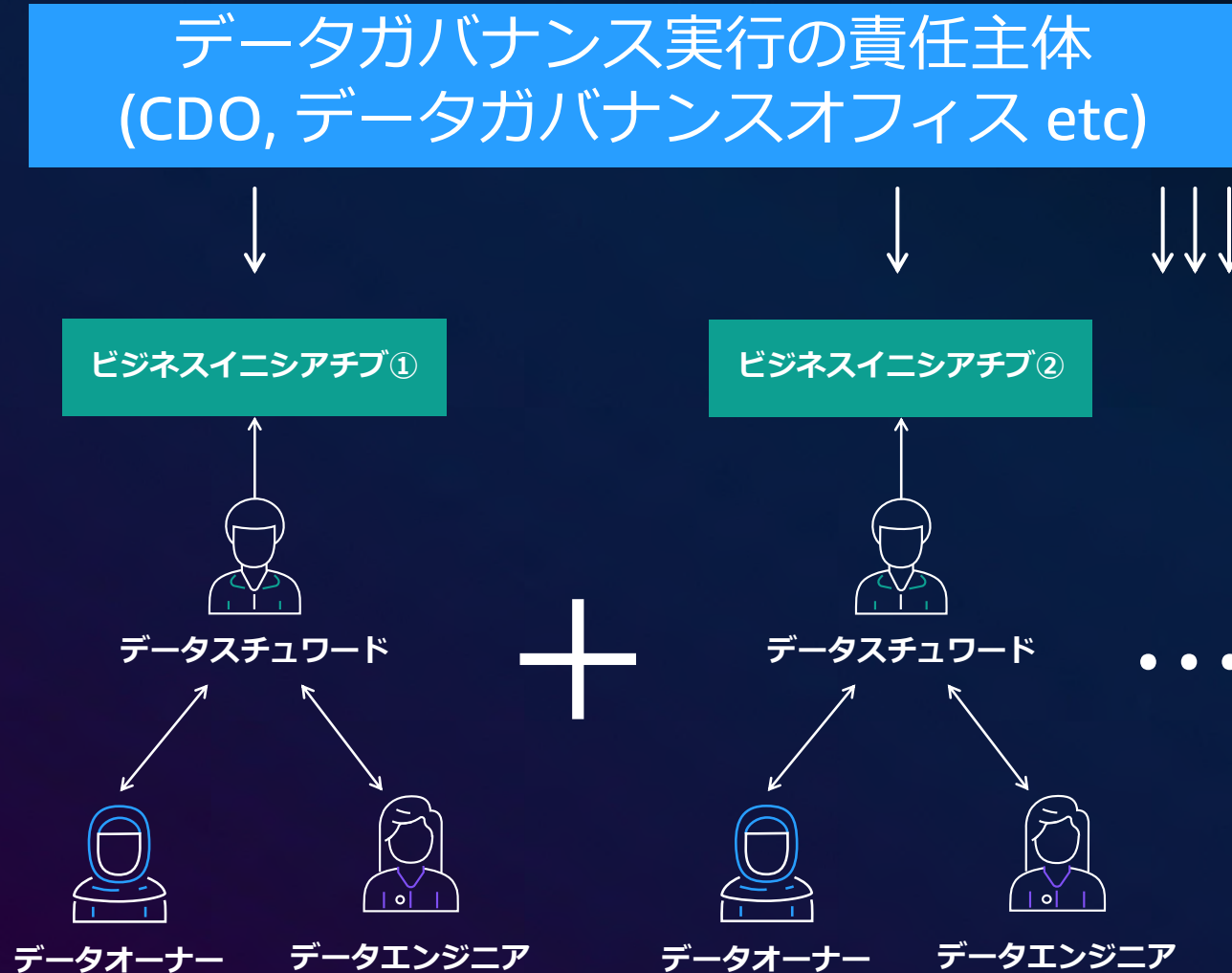


○ 特定のビジネスイニシアチブをサポートする形で始める



× "データガバナンスの推進"を目的としたイニシアチブとして始めない

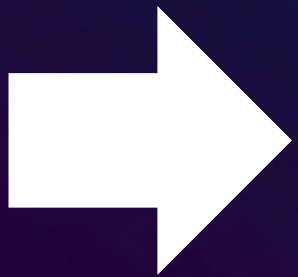
全体にデータガバナンスをスケール



- 特定イニシアチブの短期的な成功を優先しつつも、データガバナンスの適用範囲を狭くしすぎないようにする
- 徐々に他のビジネスイニシアチブ全体にデータガバナンスをスケールさせる

マインドセット変革

「誰もがコンシューマーになりたがるが
プロデューサーにはなりたがらない」



- ビジネスイニシアチブを起点とした取り組みと成功事例を増やす
- マネジメントによるコミットメント
- 適切なインセンティブ設計

まとめ

まとめ

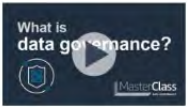








- データガバナンスとは、ビジネス上の取り組みや運用をサポートするために企業内のデータを適切な状態にしておくことであり、データを保護しながら活用を促進するためのもの
- はじめから大きな取り組みとするよりも、ビジネスを起点にその要件に合わせて、データガバナンスフレームワークの要素のうち優先度の高いところから取り組みを始める
- Amazon DataZone, AWS Glue など AWS の関連サービスを有効活用することで、データガバナンスの実現を加速できる

ぜひ、はじめましょう

AWS Data governance master class



Watch our videos

<p>What is data governance?</p>  <p>Learn why business and IT need to work together, and how Amazon's working backwards method can help you.</p>	<p>How do we align with business initiatives?</p>  <p>Align to funded business initiatives. Determine what data is needed for those business initiatives.</p>	<p>What is a good model for data governance?</p>  <p>Use data governance capabilities that help you curate the best data, understand the context of that data, and protect it across the lifecycle.</p>
<p>How do we plan a data governance roadmap?</p>  <p>Identify funded business initiatives and how their data requirements will overlap. Don't forget to establish fit-for-purpose guardrails.</p>	<p>How can we distribute data governance responsibilities?</p>  <p>Balance centralized capabilities with decentralized responsibilities to increase speed and agility.</p>	<p>How can we use the enterprise operating model?</p>  <p>Leverage enterprise capabilities for strategic planning to promote your successes and plan future projects.</p>
<p>Does self-service analytics relate to data governance?</p>  <p>Enable self-service while establishing coherent, trustworthy analytics and reducing wasteful reinvention.</p>	<p>Does ML governance relate to data governance?</p>  <p>Use good quality data in your ML models. Plan and manage various aspects of model building, deployment, monitoring, and more.</p>	<p>How can AWS help with data governance?</p>  <p>Choose tools that support your data governance model: curate, understand, and protect. AWS has services that support your strategy.</p>



AWS Data Governance Workshop

AWS でのデータガバナンスワークショップ

はじめに

AWS でのデータガバナンスワークショップへようこそ。このワークショップの目的は、データガバナンスのさまざまな課題の対応方法を実際に体験することです。

データガバナンスの実践的な定義: ビジネスの取り組みやオペレーションのニーズをできるだけ効率的、効果的、安全に満たすような状態にデータを維持します。



Thank you!

大園 純平

アマゾン ウェブ サービス ジャパン合同会社

シニアアナリティクススペシャリストソリューションアーキテクト

