AWS-37

データ活用を推進する AWS の モダンデータ戦略とデータガバナンス

林田 千瑛
技術統括本部 ソリューションアーキテクト, アナリティクス
アマゾン ウェブ サービス ジャパン合同会社

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

自己紹介

林田 千瑛

Chie Hayashida

アマゾン ウェブ サービス ジャパン合同会社

ソリューションアーキテクト

データ分析技術からデータ活用まで
幅広くご支援します！

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

アジェンダ

- データ活用の課題と AWS のモダンデータ戦略
- AWS で実現するデータガバナンス
- 一歩進んだ大規模データ活用基盤の構築

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

データ活用の課題と
AWS のモダンデータ戦略

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

あらゆるユーザーが
データ分析を行う時代へ

データエンジニア/データベース開発者

データアナリスト

データサイエンティスト

BI プロフェッショナル

システム管理者

データ活用の現実

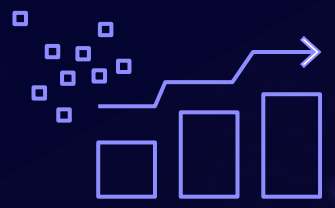
99% データによる意思決定のための積極的な投資を行っている企業の割合

24% データによる意思決定を実行できていると考える企業の数

New Vantage Partners
Big Data and AI Executive Survey 2021

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

データ活用のためのチャレンジ



あらゆるデータに接続

だれでも簡単に分析

個人利用から大規模
データまでスケール

画一的な分析環境によるデータ活用の限界

リアルタイムデータの
利用

コスト最適化

セキュリティ、
データ管理の一元化

AWS の目的別サービス

収集

Amazon Kinesis
ストリーム
データ収集

Amazon AppFlow
SaaS 連携

変換

AWS Glue
マネージドETL

Amazon EMR
マネージド Hadoop

分析

Amazon Redshift
マネージド DWH

Amazon Athena
インタラクティブクエリ

Amazon Opensearch Service
検索/ログ分析

Amazon SageMaker
機械学習

可視化

Amazon QuickSight
可視化

Amazon Managed Service for Grafana
マネージド Grafana

ストレージ

Amazon S3（オブジェクトストレージ）

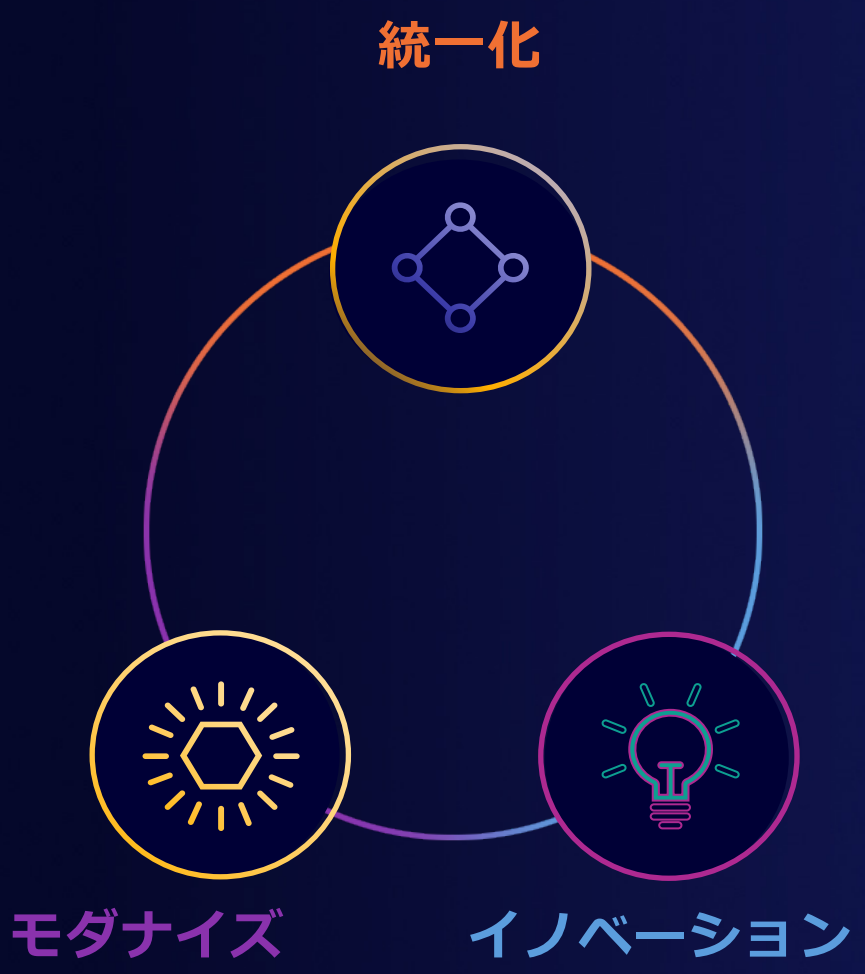
aws

モダンデータ戦略

目的別サービスの協調によりデータ活用を加速

モダンデータ戦略

AWSの新しいデータ戦略



モダナイズ

- モダンなアプリケーションによるオペレーションコスト削減
- 容易にスケール
- ローコード/ノーコードによるスキルギャップ削減

統一化

- データアクセスやセキュリティを統一してリスク削減
- データ管理を一元化してデータ活用を促進
- 得られた洞察からすばやくアクション

イノベーション

- 機械学習による新しいビジネスプロセスの開発
- イノベーションの加速

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

データ活用のステークホルダと利用シーン

マーケティング担当
Excel やダッシュボードでキャンペーンの効果を確認したい

Amazon QuickSight
可視化

Excel,
Spread Sheet

データサイエンティスト
Python や Jupyter Notebook 環境を使って機械学習や統計分析を行いたい

Amazon SageMaker
機械学習

経営層 、マネージャ
KPI を立て、その達成状況をダッシュボードで日々チェックしたい

Amazon QuickSight
可視化

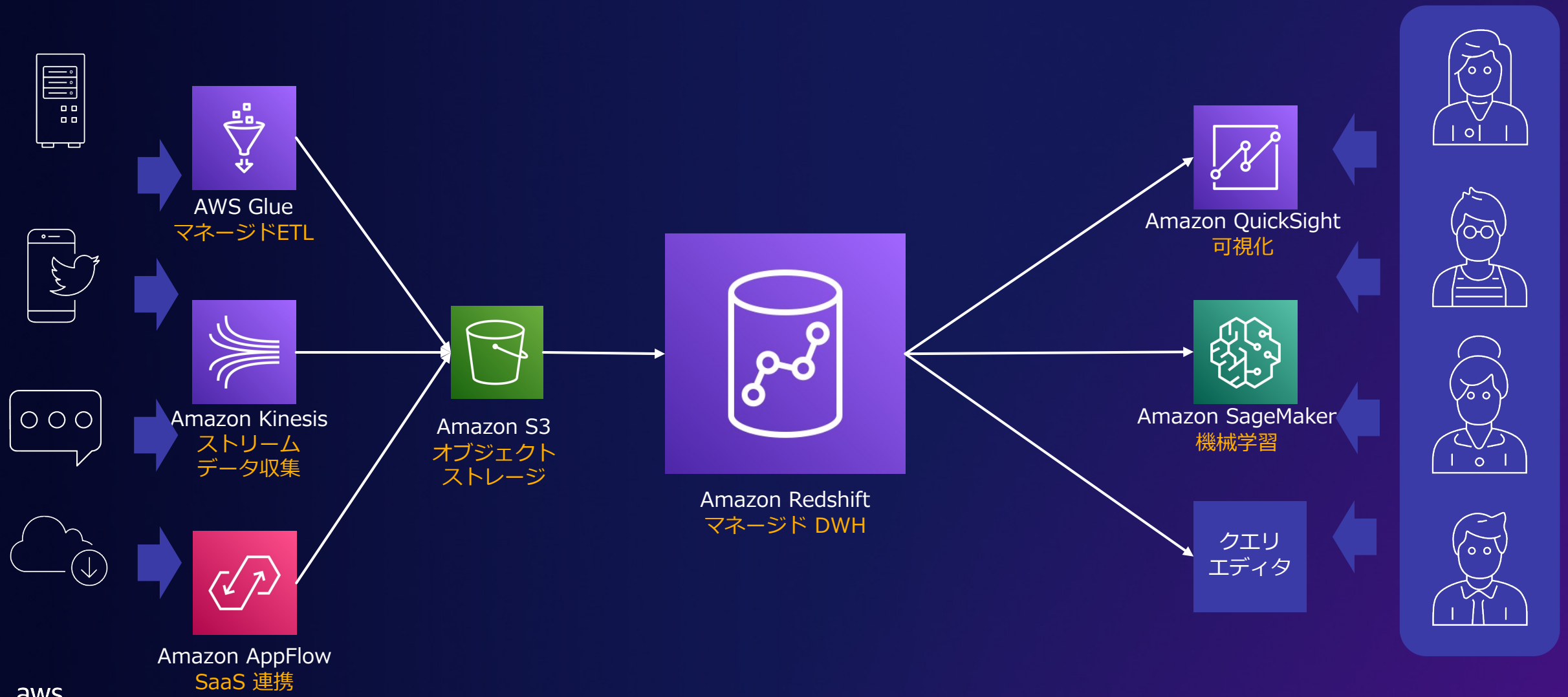
情報システム部門
さまざまなデータを統合し、様々な分析を行いやすいようにデータマートや ETL を整形する

Amazon Redshift
マネージド DWH

AWS Glue
マネージドETL

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ベーシックなデータレイク



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS で実現するデータガバナンス

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

データ活用の広がりとともに起きるデータ管理の課題

データ利用者

- それぞれが好きな環境を使って分析したい
- さまざまなデータを利用したい
- エンジニアじゃなくても気軽に利用できる分析環境がほしい
- 複雑なワークフローを動かしたい

データ管理者

- データを集約して一元的に検索・管理したい
- コンプライアンスを遵守したデータの利用が必要
- データ品質を担保したい
- システムとして安定運用したい

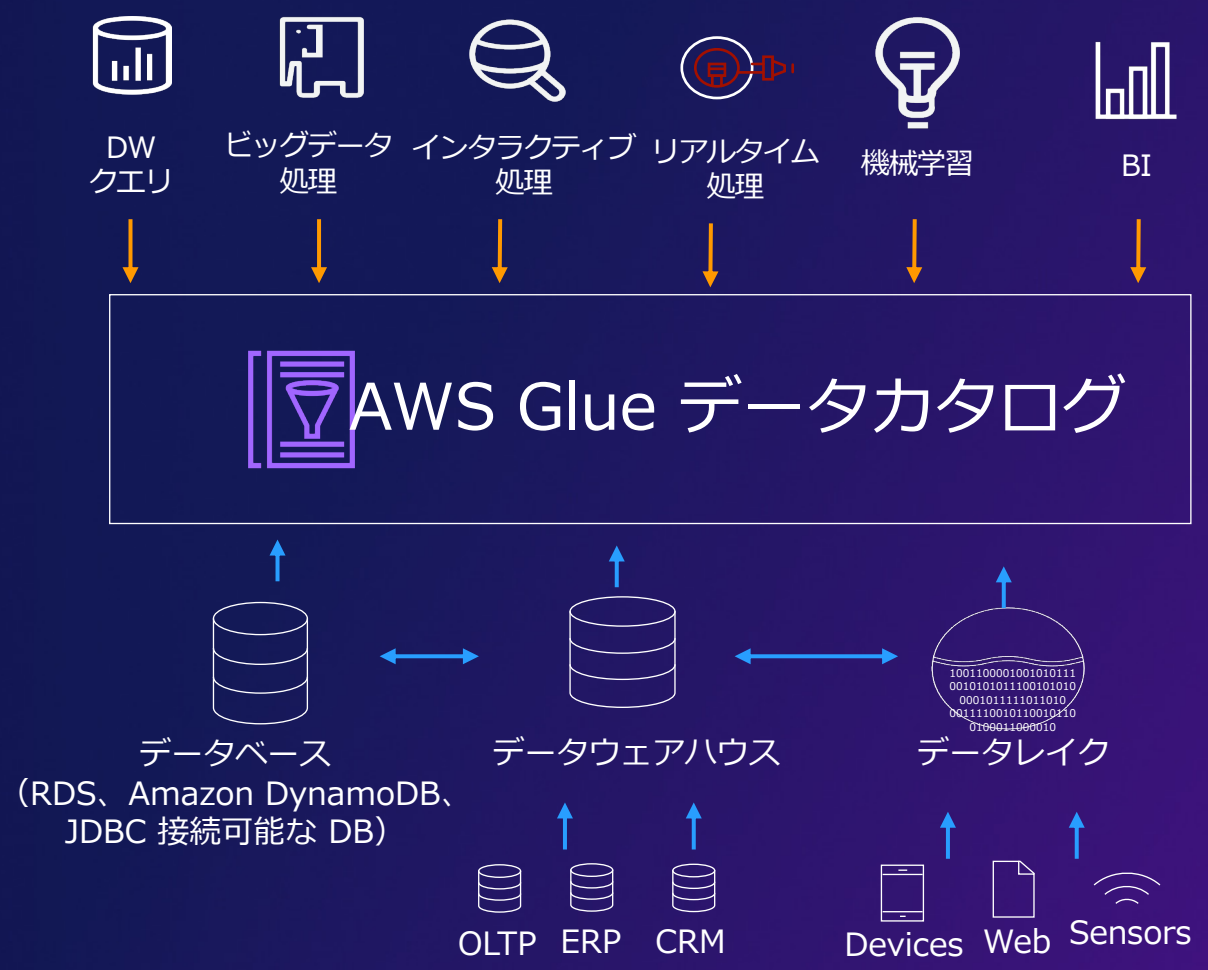
データガバナンス層によるデータ管理の課題解決

すべてのデータをカタログ化して一元管理

メタデータ

バージョニング

- 構造化・半構造化・非構造化データを登録可能
- JSON、CSV、Parquet などのオープンフォーマットに対応
- クローラや API、DDL 文を利用して自動データ登録
- コネクタを利用することでデータをコピーすることなくアクセス可能



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

データカタログが解決するもの

メタデータ

バージョニング

- さまざまなクエリエンジンから同じデータを利用することができる
  - ✓ クエリエンジンごとの ETL 作業が不要
- Single Source Of Truth（信頼できる唯一のデータソース）を実現
  - ✓ 古いデータ、誤ったデータの利用を防ぐ
- 利用者が必要なデータを数分で発見
  - ✓ データ検索や問い合わせの手間を削減

生産性向上、コスト削減

AWS Glue カタログの画面例

メタデータ

バージョニング

Services | Resource Groups | Ben Snively | N. Virginia | Support

AWS Glue

Data catalog
- Databases
  - Tables
  - Connections
- Crawlers
  - Classifiers

ETL
- Jobs
- Triggers
- Dev endpoints

Tutorials
- Add crawler
- Explore table
- Add job

Tables A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Add tables | Action | 検索窓 Filter tables or search by keyword | Save view | Showing: 1 - 13

Name	Database	Location	Classification	Last updated	Deprecated
cloudfront	logs	s3:// -us-east-1/cloudfront/	csv	10 August 2017 10:04 AM UT...	
customer1_public_customer	redshiftnamespace	customer1.public.customer	redshift	10 August 2017 9:49 AM UT...	
customer1_public_lineorder	redshiftnamespace	customer1.public.lineorder	redshift	10 August 2017 9:49 AM UT...	
customer1_public_part	redshiftnamespace	customer1.public.part	redshift	10 August 2017 9:49 AM UT...	
customer1_public_supplier	redshiftnamespace	customer1.public.supplier	redshift	10 August 2017 9:49 AM UT...	
data_store_api_public_agency	spending	data_store_api.public.agency	postgresql	10 August 2017 11:51 AM UT...	
data_store_api_public_awards	spending	data_store_api.public.awards	postgresql	10 August 2017 11:51 AM UT...	
data_store_api_public_budge...	spending	data_store_api.public.budget_authority	postgresql	10 August 2017 11:51 AM UT...	
data_store_api_public_federa...	spending	data_store_api.public.federal_account	postgresql	10 August 2017 11:51 AM UT...	
limo_2fd0ff1c	nytaxianalysis	s3:// /NYC-transportation/limo/	csv	25 July 2017 9:03 PM UTC-4	
ny_pub	nytaxianalysis	s3:// /canonical/NY-Pub/	parquet	25 July 2017 9:49 AM UTC-4	
taxi_303e40bd	nytaxianalysis	s3:// /NYC-transportation/taxi/	csv	25 July 2017 9:03 PM UTC-4	
tweets_27	logs	s3:// /twitter/2017/04/27/	json	10 August 2017 10:33 AM UT...	

テーブル名　DB名　データの場所　フォーマット　更新日時

Feedback | English | © 2008 - 2017, Amazon Web Services, Inc. or its affiliates. All rights reserved. | Privacy Policy | Terms of Use

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

メタデータの構成例

- クエリエンジンからのデータアクセスに必要なテクニカルメタデータは自動登録
- テーブルや列に対するコメントやプロパティによって、データ検索やナレッジシェアも可能
- スキーマバージョン管理により変更内容の比較が可能

メタデータ

バージョニング

テーブル情報

名前	sample_test_out01
説明	
データベース	sample-test
分類	parquet
場所	s3://test-sample-glue-bucket-ijgowerfvwigoev542637/test_out/out01/
接続	
廃止	いいえ
最終更新日	Thu Jun 27 10:35:27 GMT+900 2019
入力形式	org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat
出力形式	org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat
Serde シリアル化ライブラリ	org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe

テーブルプロパティ

Serde パラメータ	serialization.format	1				
テーブルのプロパティ	sizeKey	18183	objectCount	63	UPDATED_BY_CRAWLER	sample-test-data
	CrawlerSchemaSerializerVersion	1.0	recordCount	30	averageRecordSize	143
	CrawlerSchemaDeserializerVersion	1.0	compressionType	none	typeOfData	file

スキーマ情報

スキーマ

表示中: 1 - 7 of 7

	列名	データ型	パーティションキー	Comment
1	deviceid	string		
		string		Web/Mobile
		string		メールアドレス
		string	Partition (0)	JP/US
		string	Partition (1)	YYYY
		string	Partition (2)	mm
		string	Partition (3)	dd

コメント

最終更新日 27 6月 2019 テーブル バージョン (現行バージョン)

表示中: 1 - 2

バージョン	作成日:	作成者:
1	27 6月 2019 10:1...	
0	27 6月 2019 9:57...	

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

プロパティを利用したビジネスメタデータ管理例

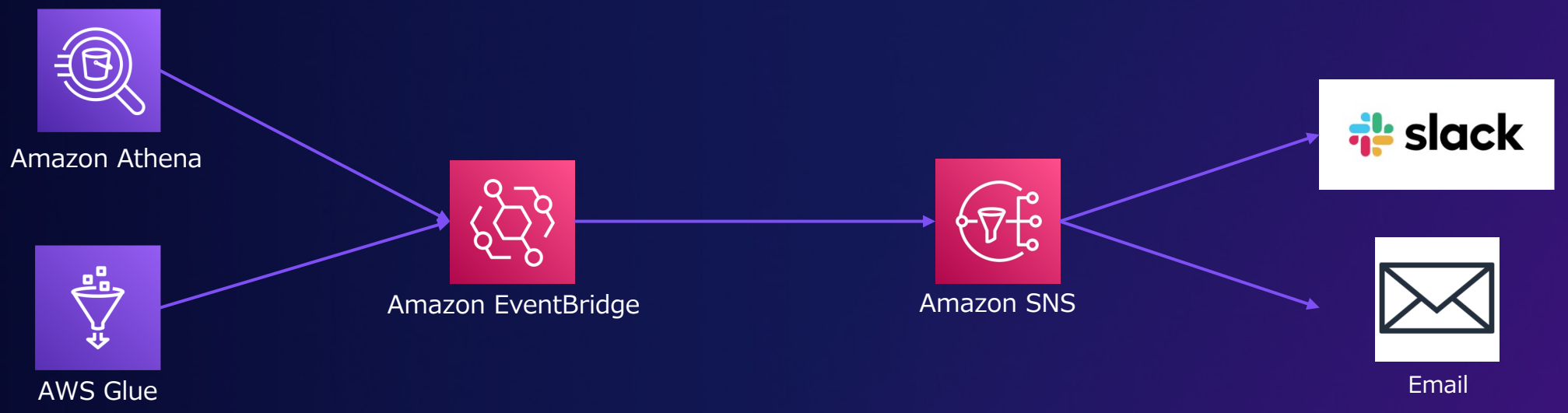
- データオーナ情報
  - ✓ あるデータのオーナにデータの詳細について問い合わせ
  - ✓ ある特定の人が管理するデータを一覧化
- データのトピック情報
  - ✓ ある分析をしたいときに必要なトピックのデータを簡単に発見
- データの有効期限
  - ✓ 期限切れデータの発見・削除
  - ✓ 古いデータの利用による誤ったインサイトを防ぐ
- 生成元データソース
  - ✓ あるデータに異常があったときに影響範囲を確認

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

データの状態監視

モニタリング

- データカタログや定期ジョブのステータスを Amazon EventBridge で監視
  - ✓ カタログの状態変化（スキーマ変更など）
  - ✓ AWS Glue ジョブや Amazon Athena クエリのステータスの変化（ジョブ終了、失敗など）

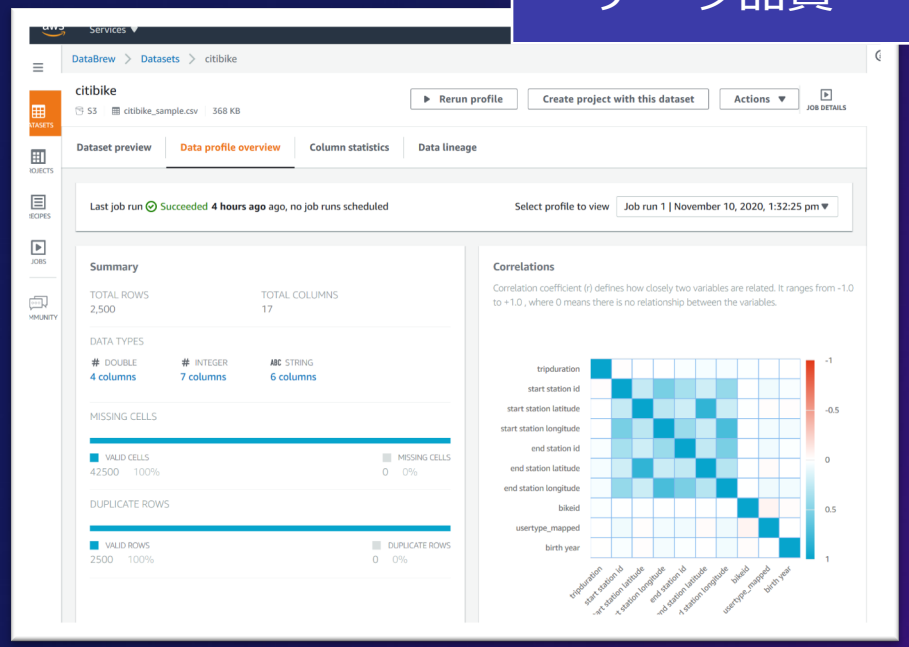


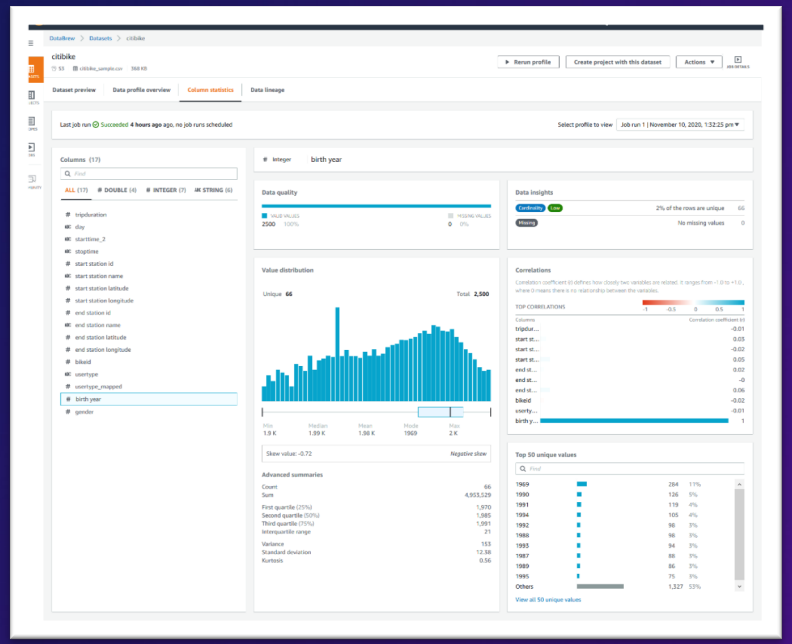
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

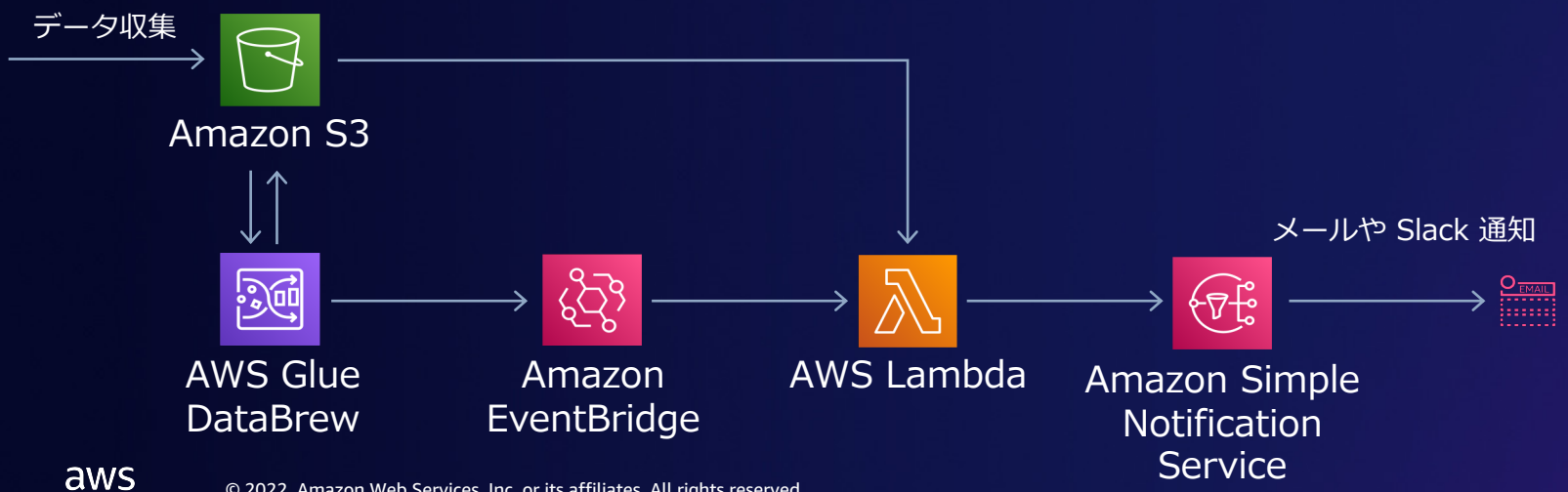
データのプロファイリング

データ品質

- AWS Glue DataBrew のプロファイルジョブを利用すると、テーブルデータのプロファイリングが可能
- カスタマイズ可能なデータ品質ルールを満たすかどうか検査し、結果を通知







© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

データカタログを利用したアクセス権限管理

セキュリティ

データ
プライバシー

- AWS の分析サービス及び 3rd パーティアプリケーションで同じ権限モデルを利用
- ファイル単位でなく DB、テーブル、列、行単位でのきめ細かな権限管理が可能
- アクセス履歴をニアリアルタイムで閲覧
- コンソールと API で操作可能

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

組織におけるデータアクセス制限の要件

セキュリティ

データ
プライバシー

- 部署やロールごとのデータアクセス制限
  - ✓ 組織の機密情報保護
  - ✓ チームごとのアクセスコントロール
- 個人情報（PII）の閲覧制限
  - ✓ ユーザテーブルにおいて住所や名前などの PII 列を閲覧不可に設定
- GDPR、CCPA に準拠するための地域ごとの管理の徹底
  - ✓ US のデータは US のみで利用可など

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

DB 操作と同じスタイルのきめ細かな権限管理

セキュリティ

データ
プライバシー

DB スタイルの GRANT/REVOKE によるアクセスコントロール

GRANTABLE オプションによる権限管理の委譲が可能

Principals	Resources	Metadata permissions	Data Permissions
IAM users and roles, SAML users and groups	Catalog	CREATE_DATABASE	N/A
	Databases	CREATE_TABLE, ALTER, DROP	N/A
	Table	DESCRIBE, ALTER, DROP	INSERT, DELETE, SELECT
	Data filters (row & cells)	DESCRIBE, DROP	SELECT
	Columns	N/A	SELECT

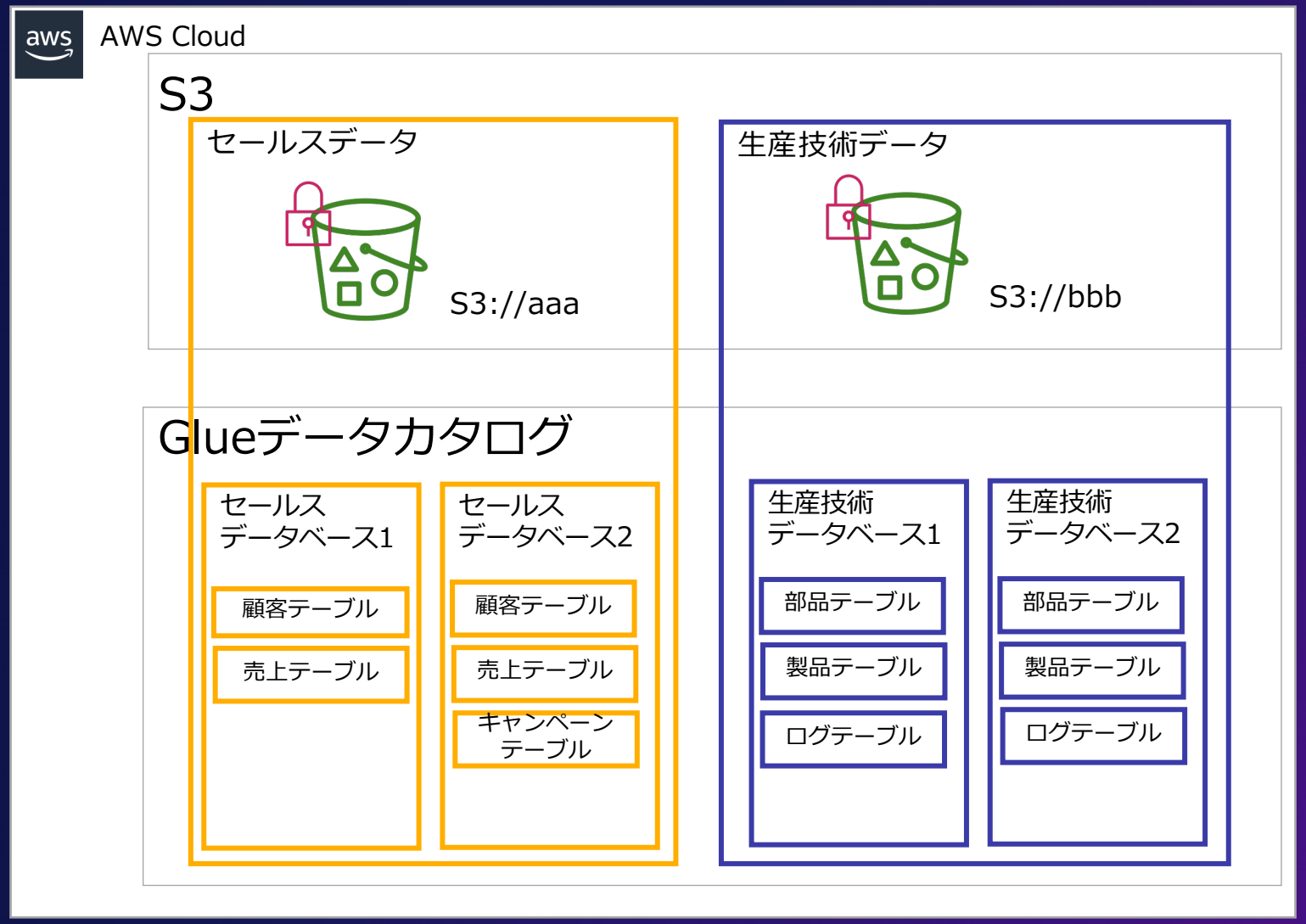
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

セキュリティ

AWS Lake Formation 環境設計の例

- 部署ごとにS3バケットを作成
- 権限を持ったユーザがデータベースやテーブルの作成/編集権限を自由に管理



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Lake Formation 権限付与の例

許可の種類	Lake Formation 全権限	セールスデータの全権限、Pass Role権限	セールスデータの一部の権限（SELECT/DESCRIBE）	生産技術データの全権限、Pass Role権限	生産技術データの一部の権限（SELECT/DESCRIBE）
IAM管理者					
データレイクセキュリティ担当	●				
セールスデータ管理者		●			
セールスデータ閲覧者			●		
生産技術データ管理者				●	
生産技術データ閲覧者					●

セールスデータ管理者が
権限付与/削除

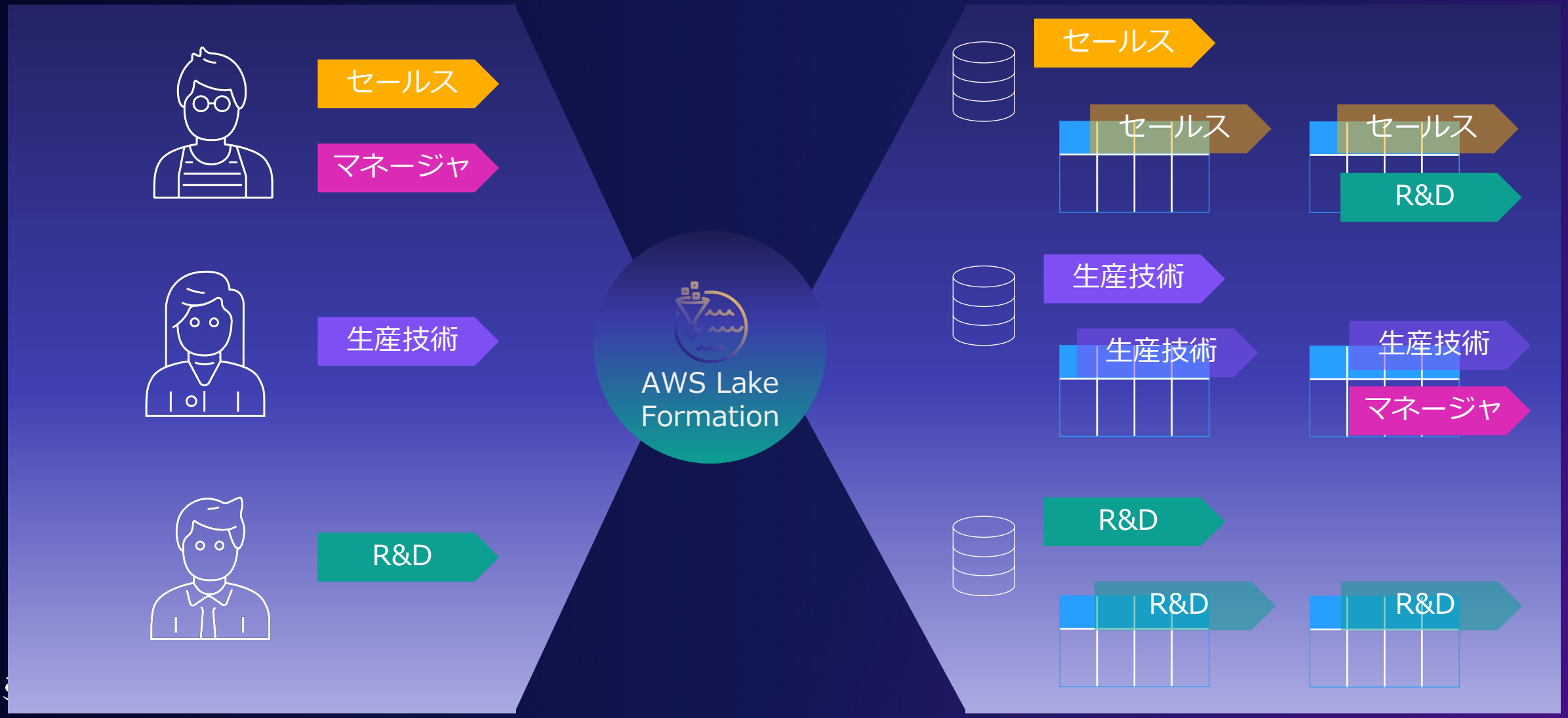
生産技術データ管理者が
権限付与/削除

データレイクセキュリティ
担当者が権限付与/削除

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

組織ごとのアクセスコントロールの例

ユーザとオブジェクトに対する**タグ付け**でアクセスコントロールを実現



PII 検出とデータマスキング

データ
プライバシー

機密情報をデータレイク取り込み時に検出・マスキングして安全にデータ活用

AWS Glue DataBrew

- データを取り込んで
  １クリックで PII の検出が可能
- 該当列中の PII の割合や、
  PII の可能性のある列の提案も行う
- ノーコード ETL 機能（レシピジョブ）
  を利用して PII のマスキングが可能

PII CATEGORIES
PII categories with number of occurrences per identified column.

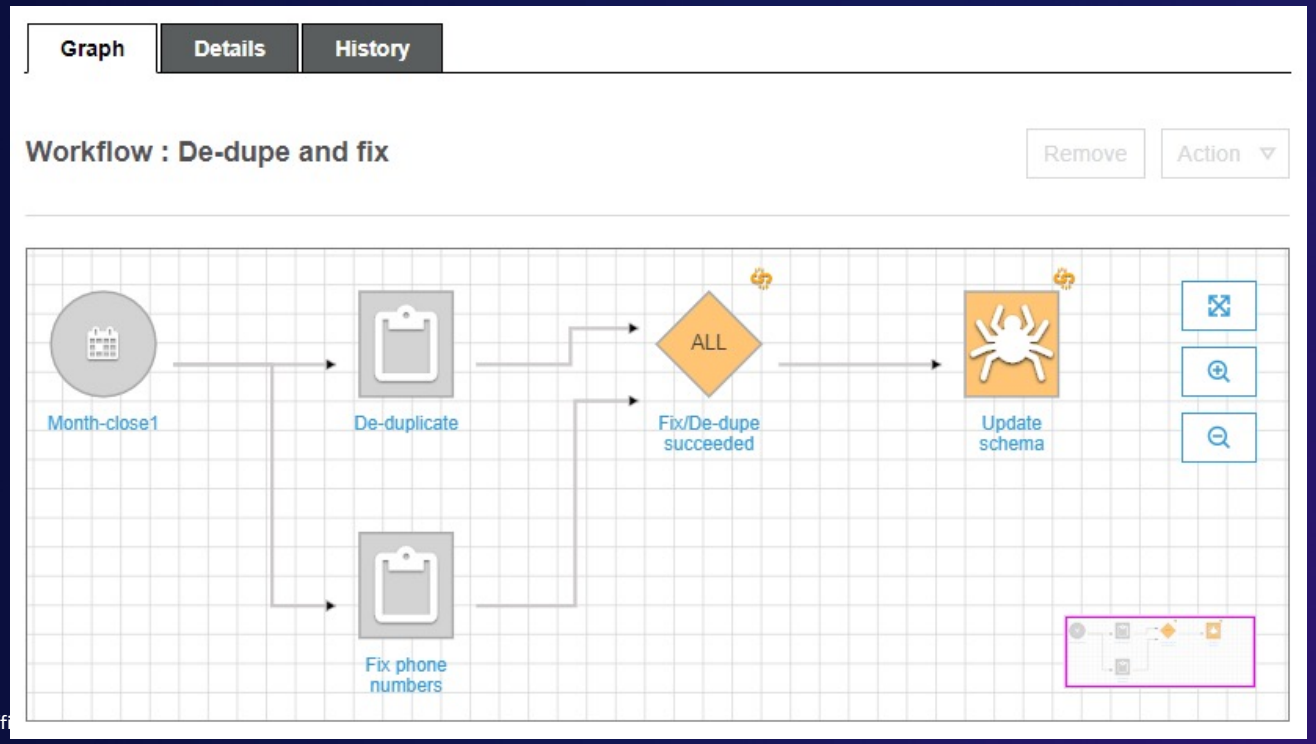
Column name	PII categories	PII column statistics
Id	USA driving license	18% PII ROWS 81% NON PII ROWS
DRIVERS	USA driving license	84% PII ROWS 15% NON PII ROWS
PASSPORT	USA passport number	71% PII ROWS 28% NON PII ROWS
FIRST	Person name	81% PII ROWS 18% NON PII ROWS
LAST	Person name	79% PII ROWS 20% NON PII ROWS
MAIDEN	Person name	22% PII ROWS 77% NON PII ROWS
ADDRESS	Person name	39% PII ROWS 60% NON PII ROWS
CITY	Person name	34% PII ROWS 65% NON PII ROWS

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ワークフロー管理

ワークフロー

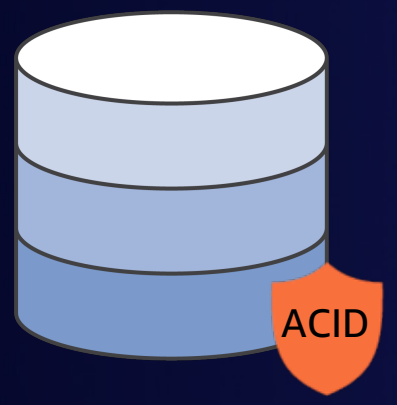
- AWS Glue ワークフローを利用して ETL ジョブのワークフローを一元管理
- 部分リトライや BluePrint を利用したワークフロー生成の自動化も可能

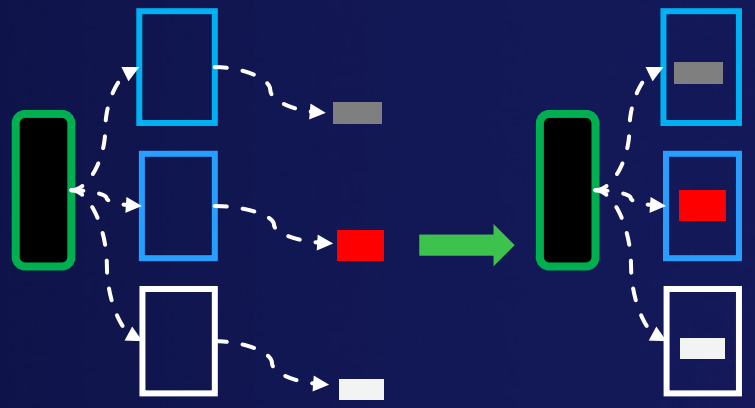


© 2022, Amazon Web Services, Inc. or its affi

AWS Lake Formation Governed Table

データ最適化





ACID トランザクション

データレイクにおける
一貫性を保証した書き込み

信頼性

ストレージ最適化

スモールファイルの自動圧縮、
プッシュダウンフィルタリングによる
スキャン量削減

パフォーマンス

タイムトラベル

過去の状態のデータ
へのクエリ

バージョニング

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

一歩進んだ大規模データ活用基盤の構築

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customer 360：顧客データからビジネス価値を創出

収集 → 統合 → 分析 → 活用

データ統合

- Amazon S3
  データ保存
- AWS Glue
  ETL処理
- AWS Transfer
  ファイル転送
- AWS Data Exchange
  3rd パーティデータ

Web、モバイル、ソーシャルメディアなどからデータ取り込み

顧客情報

- Amazon Cognito
  顧客プロフィールの統合
- Amazon Neptune
  顧客の関係グラフ

プライバシー

- Amazon DynamoDB
  同意管理

セグメンテーション

- Amazon SageMaker
  機械学習による
  セグメンテーション

予測

- Amazon Personalize
  レコメンデーション

活用

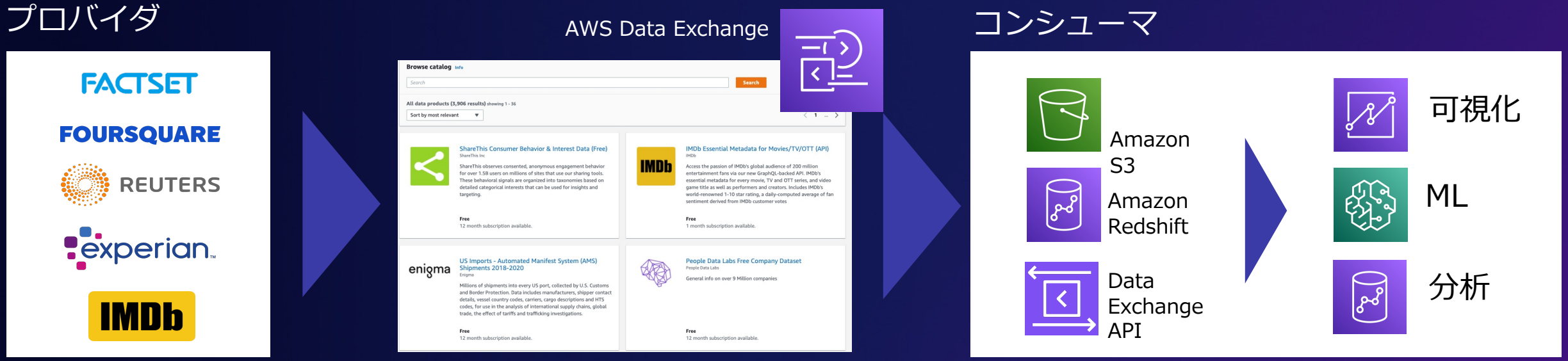
- Amazon Pinpoint
  キャンペーンの実行
- SES, SNS
  メールやプッシュ通知
- Amazon Connect
  コールセンタ自動化
- Amazon DSP、 AMC
  広告最適化
- Amazon Personalize
  商品のレコメンド
- Amazon QuickSight
  顧客動向の可視化
- Amazon RedShift
  アドホック分析
- Amazor Forecast
  需要予測

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Data Exchange で広がるデータ活用

- サードパーティベンダが提供するデータを利用することでより深い洞察を得る

例：
- 天気情報を考慮したレコメンドエンジン
- 地図情報を利用した近くのレストラン情報の検索
- 株価やビットコイン価格データを利用した市場分析



- 250以上のデータプロバイダと3000以上のデータプロダクト（1000以上が無料）
- Amazon S3、Amazon Redshift、API を利用したデータ取得が可能
- 公開またはプライベートでのデータ共有が可能

© 2022, Ama

エンタープライズ企業でのモノリスデータ環境の課題

- コスト管理があいまいになる
- インフラ管理者の権限管理が煩雑
- ETL ワークフローがスパゲッティコード化
- スケールが難しくなる
- １つのエラーで全ての分析機能が止まる

➡ セルフマネージドなデータインフラの利用

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

セルフマネージドなデータインフラ

- ドメイン指向のデータアーキテクチャによってビジネスインサイト導出を加速する

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

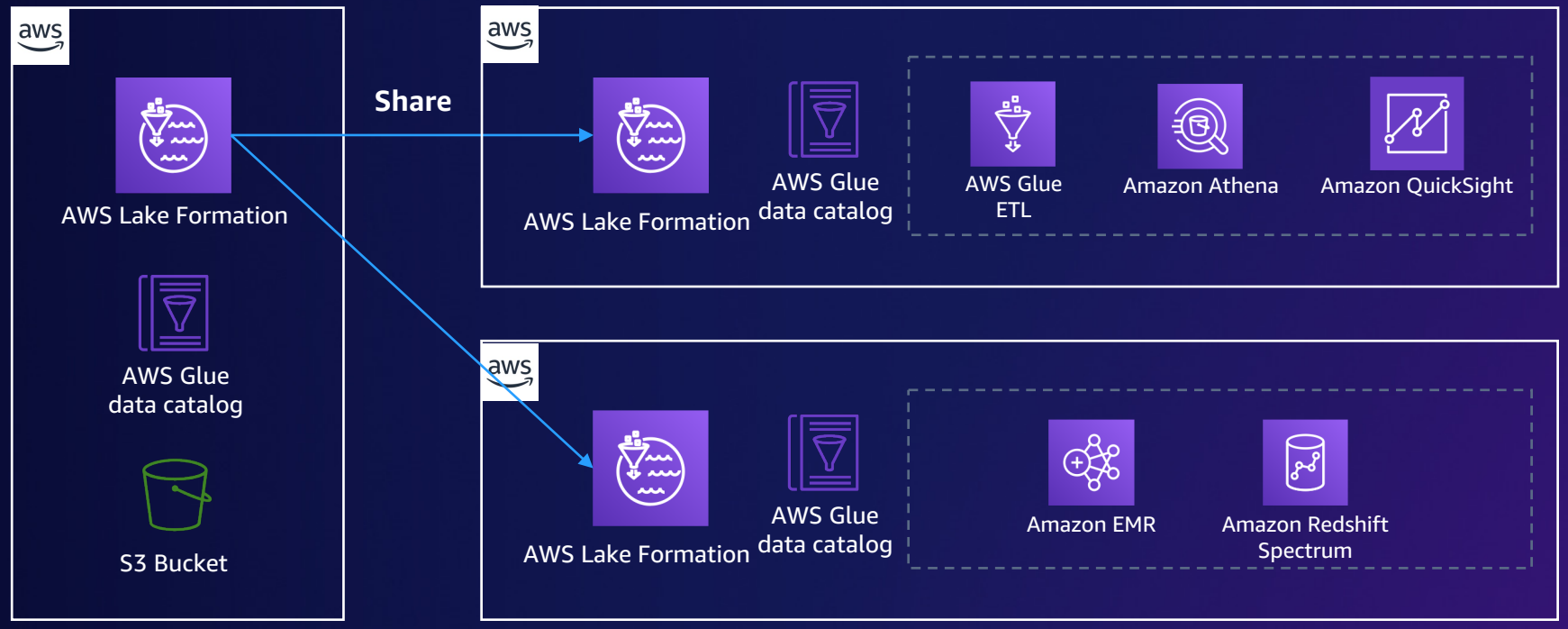
Data Mesh（データメッシュ）

- ドメイン指向による所有権とアーキテクチャの分散
- それぞれのデータを 1 つのプロダクトとして扱う
- 統合的なデータガバナンス環境におけるデータ共有

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Lake Formation で実現する Data Mesh

- データをコピーすることなくプロデューサアカウントのデータをコンシューマアカウントに共有
- クロスアカウントでのDB/テーブル/列/行レベルでのアクセスコントロールが可能

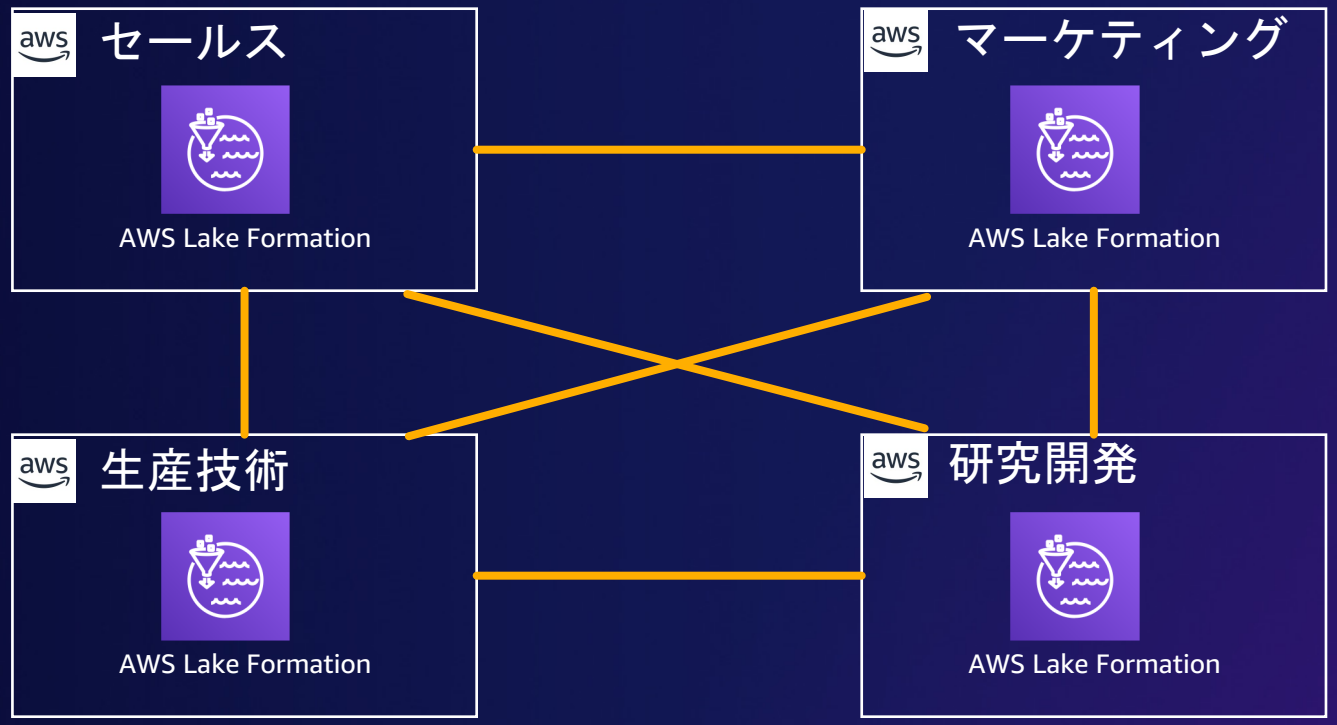


プロデューサアカウント　　コンシューマアカウント

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Lake Formation で実現する Data Mesh

- 各組織がアカウント内でデータプロダクト（DB、テーブル等の Lake Formation オブジェクト）を必要な組織に共有
- メッシュのように複数アカウント間で相互にデータ共有
- 各共有先アカウントでのデータプロダクトへのアクセス履歴をプロデューサアカウントで監査・モニタリング可能



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

データメッシュのメリット

コストと責任の分離

- アカウントを分けることで組織ごとのコスト管理が可能
- 障害時の責任範囲が容易

スケールの効率化

- 組織ごとに自由にクエリエンジンを選択
- 他組織の影響を受けずにリソースチューニング

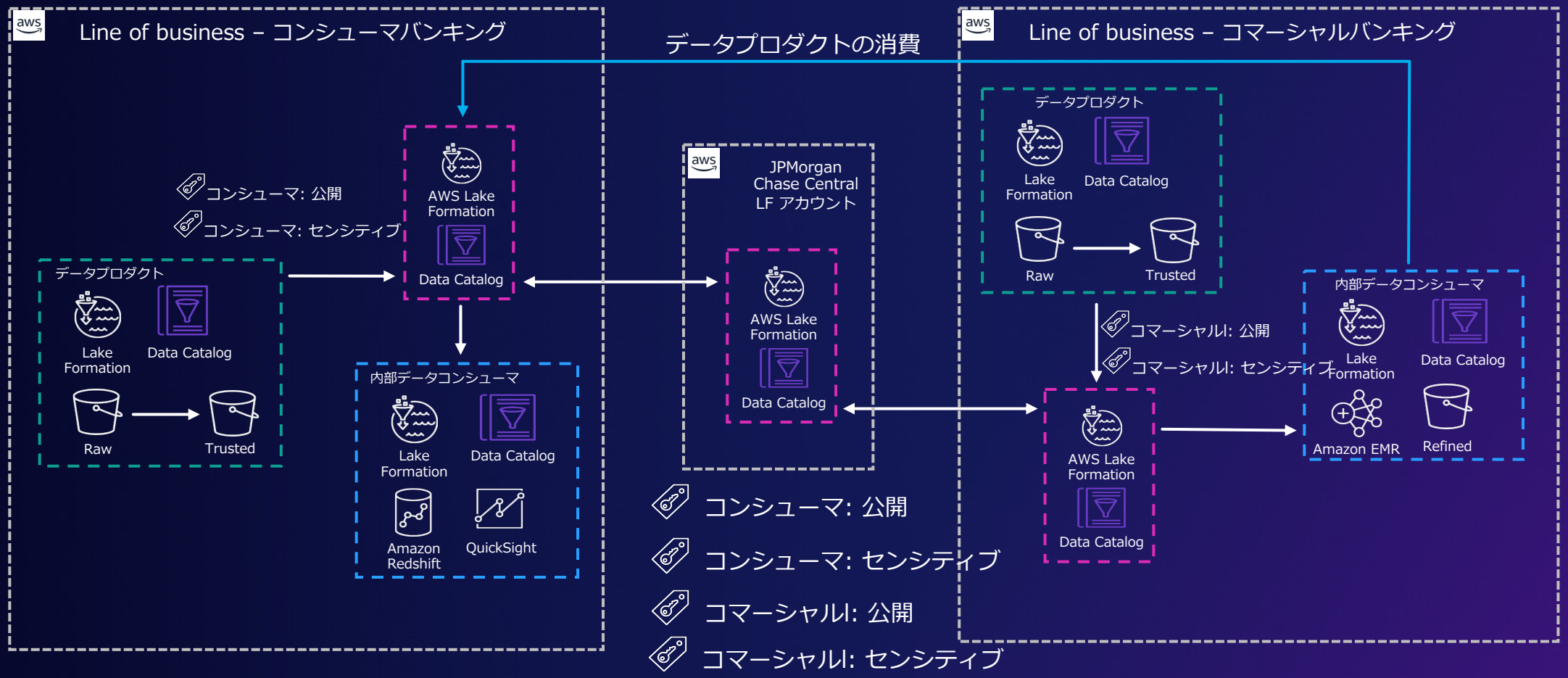
より高速なインサイトの導出

- データ転送を行わずにパイプライン簡素化
- 他組織との調整なくユーザが必要な環境をすぐに提供

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

JP Morgan Chase におけるデータメッシュ事例



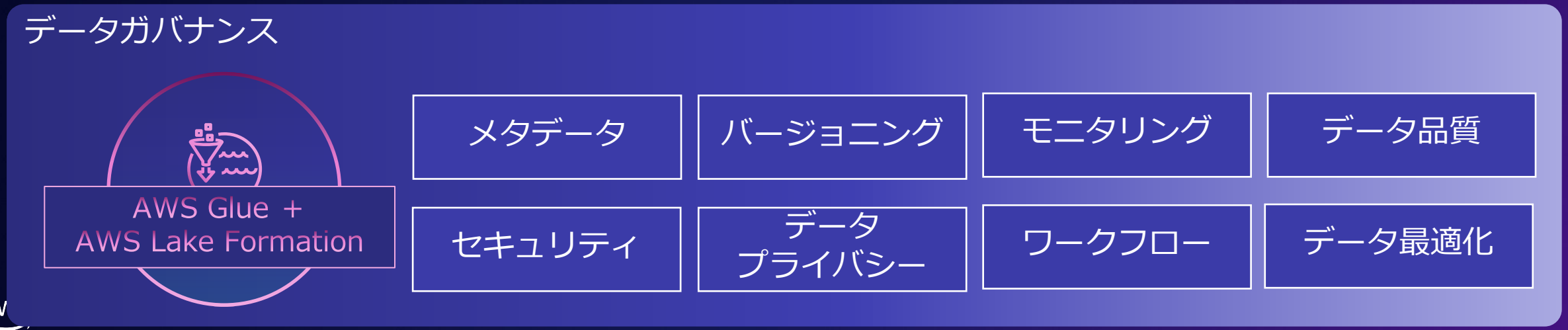
- 各組織がデータプロダクトを所有・管理
- 権限管理と使用状況監査を行い、適切にデータ共有を行う

- JPMorgan Chase 中央カタログ はすべてのデータプロダクトへのアクセスを提供
- データセットへのアクセス権の付与や取り消しは不可

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

まとめ

- モダンデータ戦略では、目的別サービスによりデータ活用を促進する
- データ活用の加速に伴い、データガバナンスの重要度が増している。AWS Glue、AWS Lake Formation を中心としたデータガバナンスにより、データ活用とシステムとしての頑健性を両立する
- データメッシュ環境を構築することで、よりビジネスインサイトに直結した分析環境を提供できる



Thank you!

林田 千瑛（Chie Hayashida）

ソリューションアーキテクト
アマゾンウェブサービスジャパン合同会社

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.