Google Cloud

データ分析における データ ガバナンス

2022 / 06 / 07

Google Cloud
Data Analytics Specialist
Hiroshi Kitahara

01

データ ガバナンスとは

データガバナンスとは

安全で信頼できるデータの活用を可能とするためのデータマネジメントの活動で、
以下のような目標を持って実行される

- データを民主化し意思決定に組み込む一方で、セキュリティを確保し不正使用を防ぐ
- データの取得から利用、廃棄に至るまでのライフサイクル全体を一貫した原則で管理する
- 組織に蓄積されたデータに一元的なアクセスを提供し、高いデータ品質を確保する
- 規制や基準へ適合し、コンプライアンスを確保する

データ ガバナンスが必要とされる理由

データ ガバナンスへ取り組み始めるきっかけ

データ量増大	データ種別の 多様化、複雑化	データ利用者増 データ民主化	データ活用 ユースケース 拡大	情報セキュリティ プライバシー

データ ガバナンスによって得られる効果

堅牢なデータ ガバナンスによるビジネス メリット

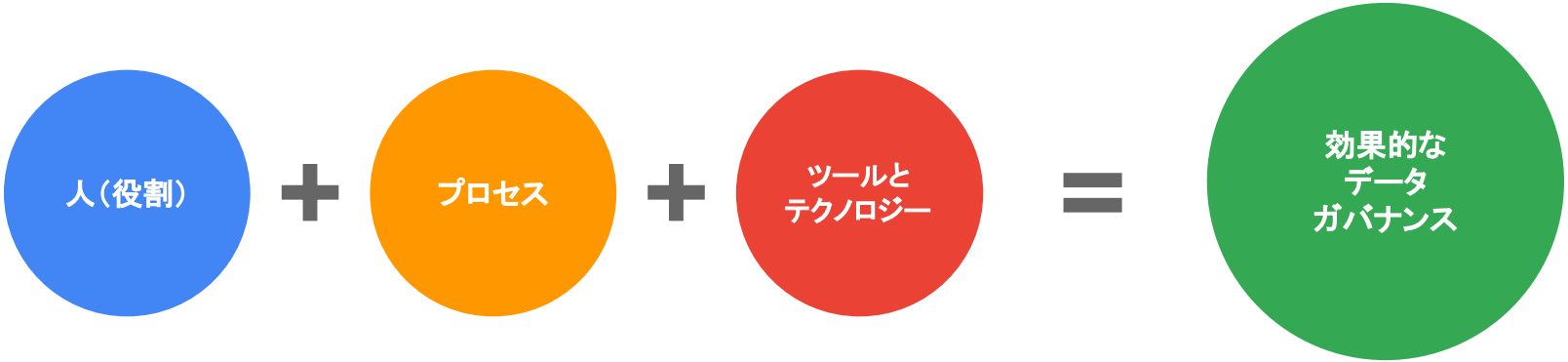
- オペレーション効率の向上
- より多様な分析、より的確な意思決定
- リスク管理
- 規制への対応（GDPR、個人情報保護法 等）
- 顧客サービスの向上、収益拡大

02

データ ガバナンスの推進

効果的なデータ ガバナンス

効果的なデータ ガバナンス戦略を進めるには、人（役割）、プロセス、
ツールとテクノロジーの各要素をバランスよく推進することが必要。



データ ガバナンスを担う人々とロールの例

データ ガバナンス委員会（Data governance council）

- 各ビジネス部門を代表するリーダーで構成
- ハイレベルなガバナンス原則を決定
- データドメインを特定し、データオーナーとスチュワードの役割を割り当て

データオーナーとスチュワード（ Data owner & Data steward）

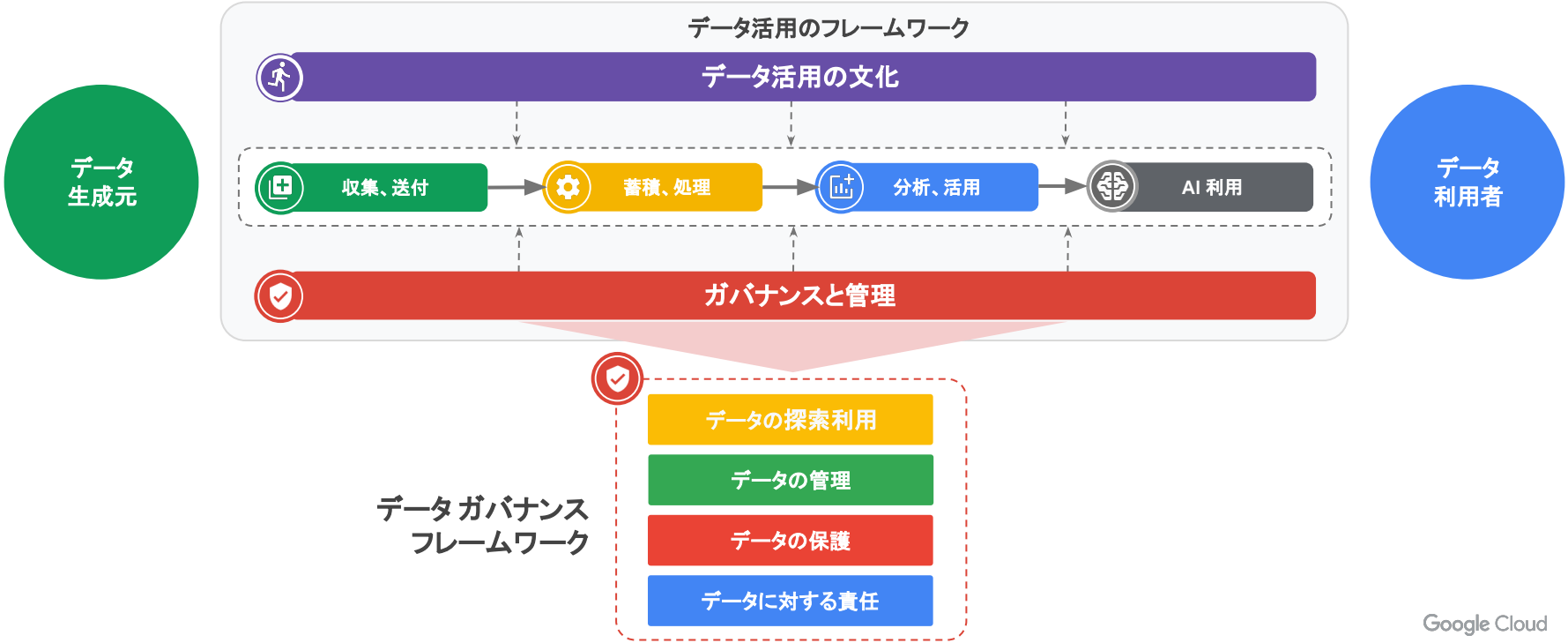
- ビジネス部門の責任者が担当（例 顧客データ: マーケティングや営業部門、
  財務データ: 財務部門、人事データ: 人事部門など）
- データの内容や品質を担保し、データ ガバナンス委員会によって定められた目標を推進

データ管理者（Data Custodian）

- IT 部門が担当
- データオーナーが指定したポリシーに従って、データの取得、保護、保存、共有を実施

データ利活用におけるデータ ガバナンスの位置付け

データの収集から利用にわたるステップの全ての段階にデータ　ガバナンスと管理のフレームワークを組み合わせることで、データ活用を安心して促進することができるようになる。



データ ガバナンスのフレームワーク

データの探索利用	データの管理	データの保護	データに対する責任
データ分類	データ ライフサイクル ＆ 情報資産管理	データマスキング	データオーナーシップ
データリネージュ	参照データ（コード）管理	データ暗号化	データポリシー ＆ 基準
メタデータ ＆ データカタログ	マスターデータ管理	アクセス管理	データドメイン単位のガバナンス
データ品質	データ SRE	監査 ＆ コンプライアンス	データ倫理
		データ所在管理	
		データ復旧	
正確で完全なデータが利用可能であること	データが管理下にあって誰もが理解できること	データの漏洩や失われることを防ぐ	ポリシーと説明責任を定める

データ ガバナンスのフレームワーク

- データマネジメントのフレームワークや規制の例
  - Enterprise Data Management Council (EDMC)
    - Data Management Capability Assessment Model (DCAM)
    - Cloud Data Management Capability (CDMC)
  - GDPR
  - 個人情報保護法
  - HIPAA
  - PCI-DSS

データ ガバナンスを実現するステップ

データガバナンスはじめの一歩

- **データの探索とアセスメント**
- **機微情報の特定と分類**
- **データカタログの整備**
- **データ品質に対する期待を文書化**
- **個人、グループ、ロールを整理しアクセス権限を割り当て**
- **定期的な監査**
- **データ保護のための追加的な対応**

小さく始めて**検証、学習、反復**するアプローチ

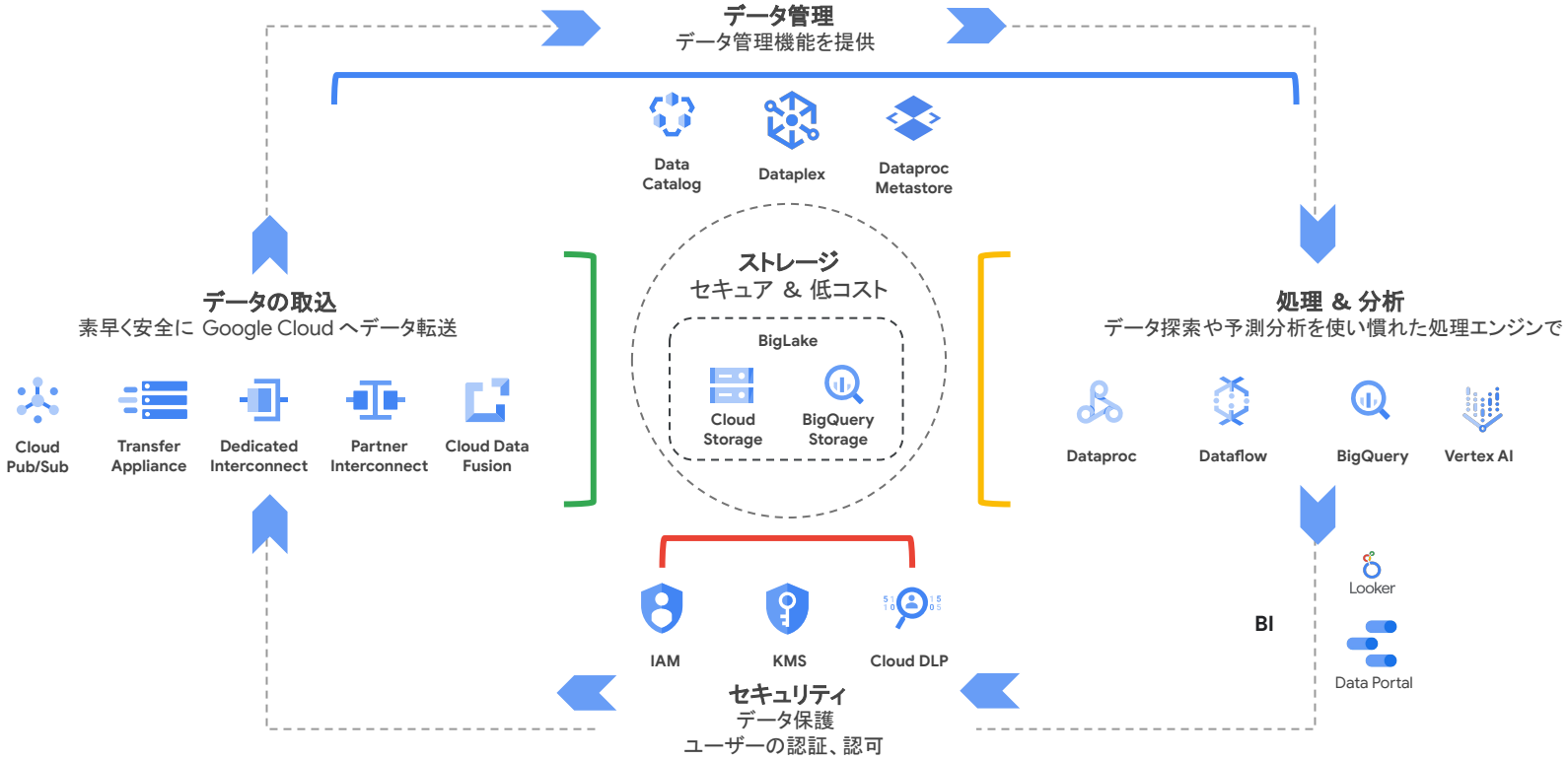
データ ガバナンス プログラムを計画、開始、そして継続するために重要なステップ：

- ビジネスケースの構築
- 指針の文書化
- マネジメントの支援
- オペレーティングモデルの構築
- 報告のフレームワーク
- 分類体系や概念・用語の共通理解
- 適切なテクノロジースタックの利用
- 教育とトレーニング

03

Google Cloud で実現する データ ガバナンス

Google Cloud で構築するデータ基盤とガバナンス関連機能

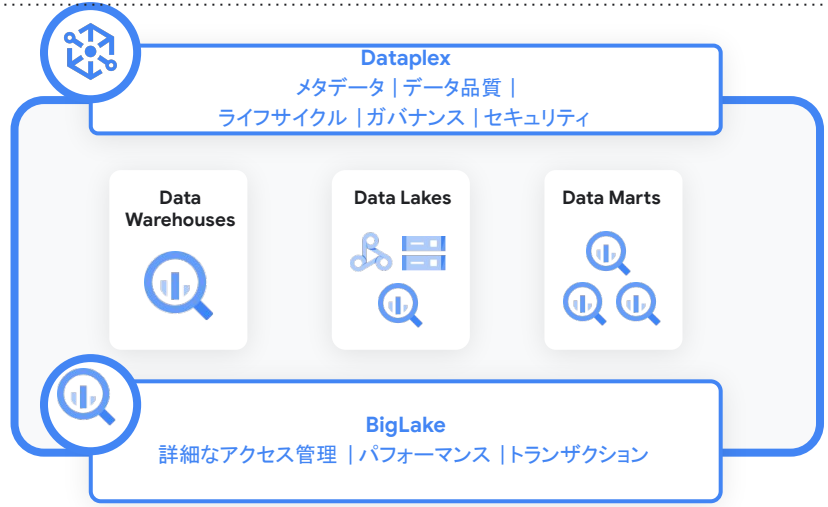


データ ガバナンスのコントロール プレーン

Dataplex

Dataplex

分散したデータを一元化して、**データ管理を自動化** し、**より強力な大規模な分析** を可能にするインテリジェントなデータ　ファブリック



Dataplex が解決する課題

サイロ化されたデータの一元管理

様々なプロジェクトやデータストア、さらにマルチクラウド基盤に分散したデータをグループ化し、ユーザへのアクセス管理やガバナンスを実施。
また BigLake との連携によって、さまざまなフォーマットに対応するパフォーマンスに優れたデータ分析基盤を実現。

インテリジェントなデータ管理

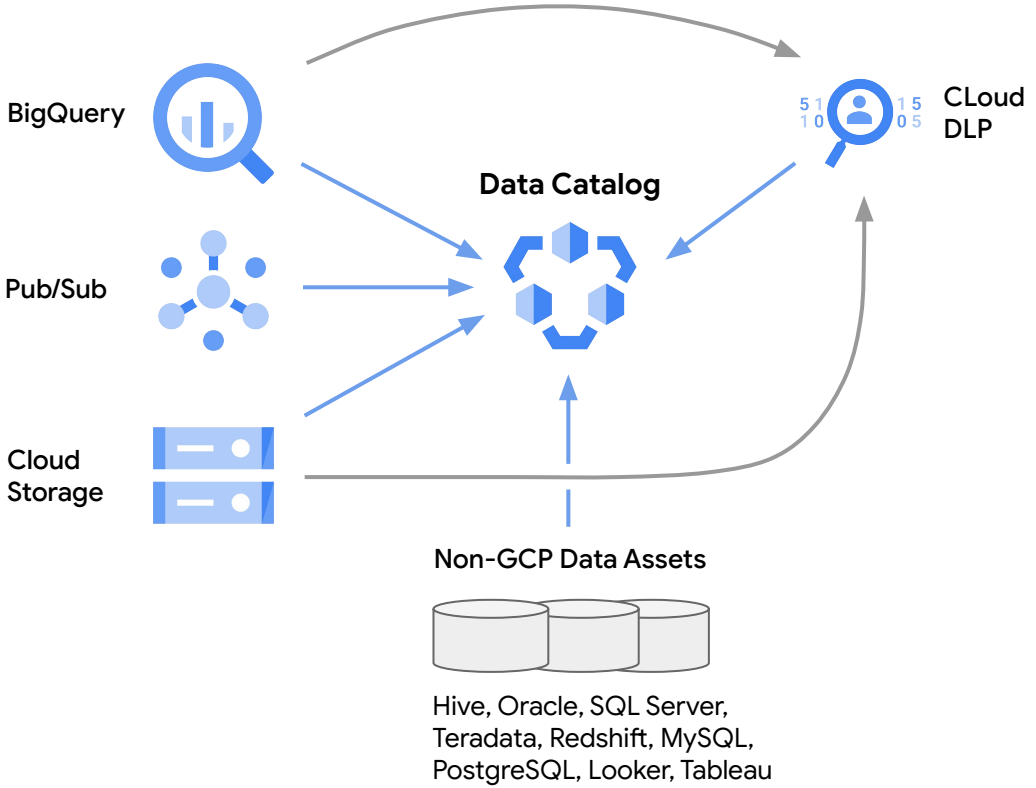
メタデータ検出によるデータの発見・探索。
ユーザ定義や自動化によるデータ品質の管理・運用。

データの見つけやすさを実現する Google Cloud サービス

データの探索利用	データの管理	データの保護	データに対する責任
データ分類	データライフサイクル & レコード管理	データマスキング	データオーナーシップ
データリネージュ	参照データ(コード)管理	データ暗号化	データポリシー & 基準
メタデータ & データカタログ	マスターデータ管理	アクセス管理	データドメイン単位のガバナンス
データ品質	データ SRE	監査 & コンプライアンス	データ倫理
		データ所在管理	
		データ復旧	

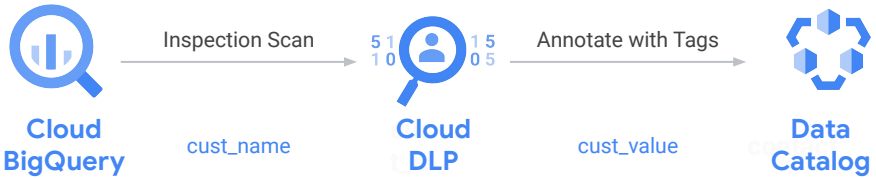
メタデータ & データカタログ
Data Catalog

1. Google Cloud 上のデータアセットの
テクニカル メタデータを
ニア リアルタイムで同期
2. Cloud DLP とのインテグレーションにより、
個人情報データに自動でタグ付け
3. OSS コネクタ経由で Google Cloud
以外のデータアセットをサポート
4. タグテンプレートによるビジネスメタデータの
管理



データ分類

BigQuery と Cloud DLP（Data Loss Prevention）の連携で個人情報関連のタグを自動付与



column:credit_card_number: DLP Result Scan Template (travel-bookings-236421.dlp_result_scan_template)

Attribute	Display name	Value
scan_date_time	Date Time of Scan	04/04/2019 04:44
result	Result	CREDIT_CARD_NUMBER
has_pii	Has PII Information	true
num_rows_scan	Number of Rows to Scan in DLP	1000
result_percentual	Result percentual	1
all_result	all result	|CREDIT_CARD_NUMBER(LIKELY : 100.0%);|
num_rows_table	Number of Rows in the table scanned	13000010
result_likelihood	Results likelihood	LIKELY
num_match	Number of Results	1000

Cloud DLP により自動
タグ付けされたメタデータ タグ

データ品質

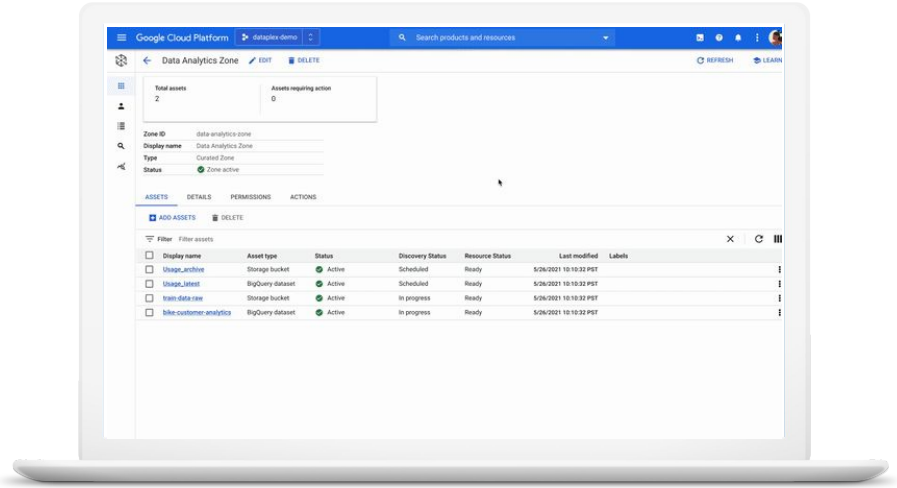
Dataplex 自動データ検出とデータ品質タスク

自動データ検出と分類

- 動的なスキーマ検出と型のマッピング
- スキーマドリフトの検知

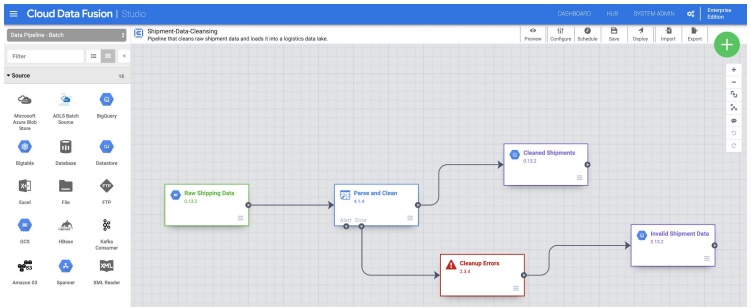
データ品質タスク

- ユーザ定義ルールによるメタデータおよびデータのチェック

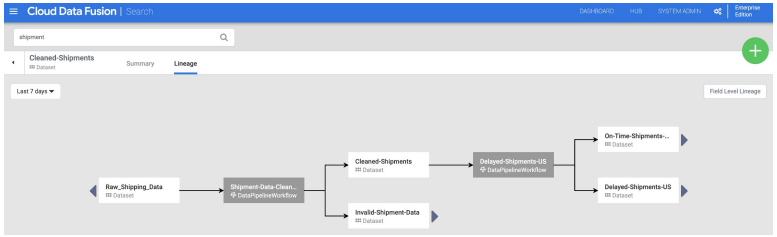


データリネージュ

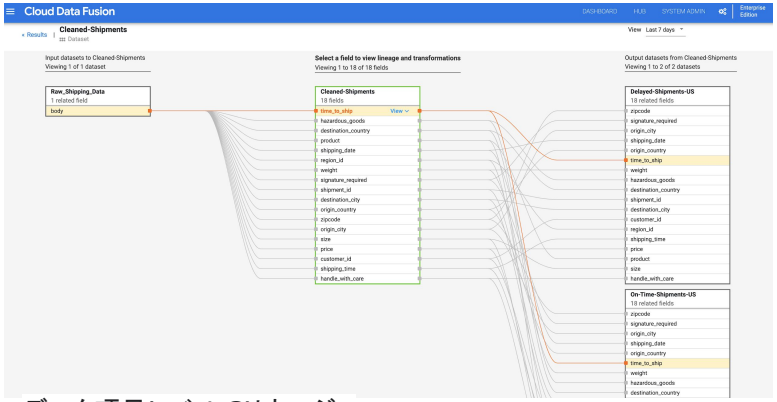
Cloud Data Fusion によるデータリネージュの追跡



データ パイプライン定義



データセットレベルのリネージュ



データ項目レベルのリネージュ

データの管理を実現する Google Cloud サービス


Dataplex


Data Catalog


BigQuery


Looker

データの探索利用	データの管理	データの保護	データに対する責任
データ分類	データライフサイクル & レコード管理	データマスキング	データオーナーシップ
データリネージュ	参照データ(コード)管理	データ暗号化	データポリシー & 基準
メタデータ & データカタログ	マスターデータ管理	アクセス管理	データドメイン単位のガバナンス
データ品質	データ SRE	監査 & コンプライアンス	データ倫理
		データ所在管理	
		データ復旧	

データ ライフサイクル管理

Dataplex でのデータの取込み、整理、キュレート、保護、アーカイブタスク

ワンクリックのテンプレート

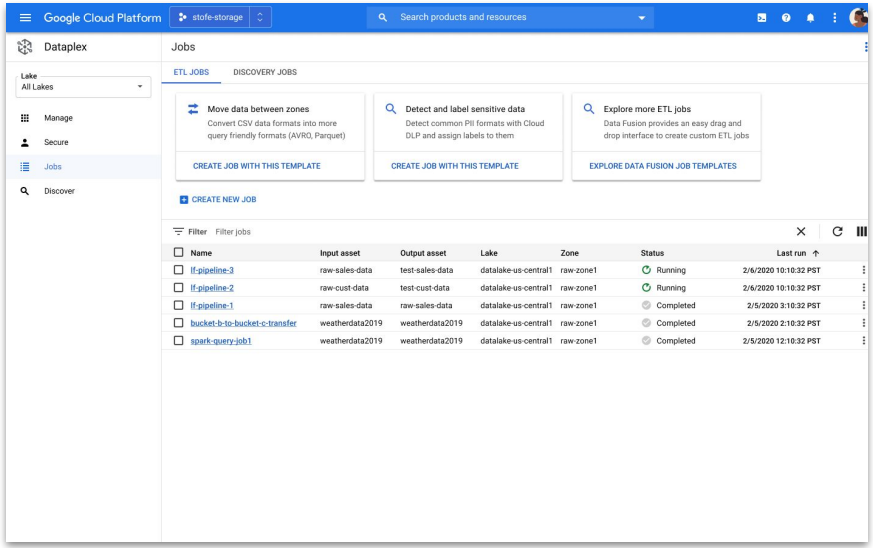
- データの移動、階層化
- インフラ管理不要
- Dataflow、Data Fusion と統合

統一された監視

- テンプレート、検出ジョブ、スケジュールされた ノートブック、その他の ETL ジョブ全体の パイプラインを監視

拡張可能なプラットフォーム

- カスタム変換を構築および監視するためのカスタムジョブ



データ保護を実現する Google Cloud サービス


Dataplex


Data Catalog


BigQuery


Cloud DLP


Cloud Key Management

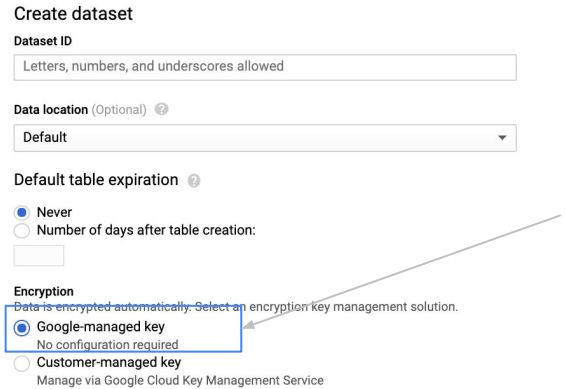

IAM

データの探索利用	データの管理	データの保護	データに対する責任
データ分類	データライフサイクル & レコード管理	データマスキング	データオーナーシップ
データリネージュ	参照データ(コード)管理	データ暗号化	データポリシー & 基準
メタデータ & データカタログ	マスターデータ管理	アクセス管理	データドメイン単位のガバナンス
データ品質	データ SRE	監査 & コンプライアンス	データ倫理
		データ所在管理	
		データ復旧	

データ暗号化

BigQuery データ保管時の暗号化

- デフォルトで**透過的に暗号化**されて保存
- 読み取りの際に自動的に復号化
- Google 管理または顧客管理の暗号鍵（CMEK）を利用可能

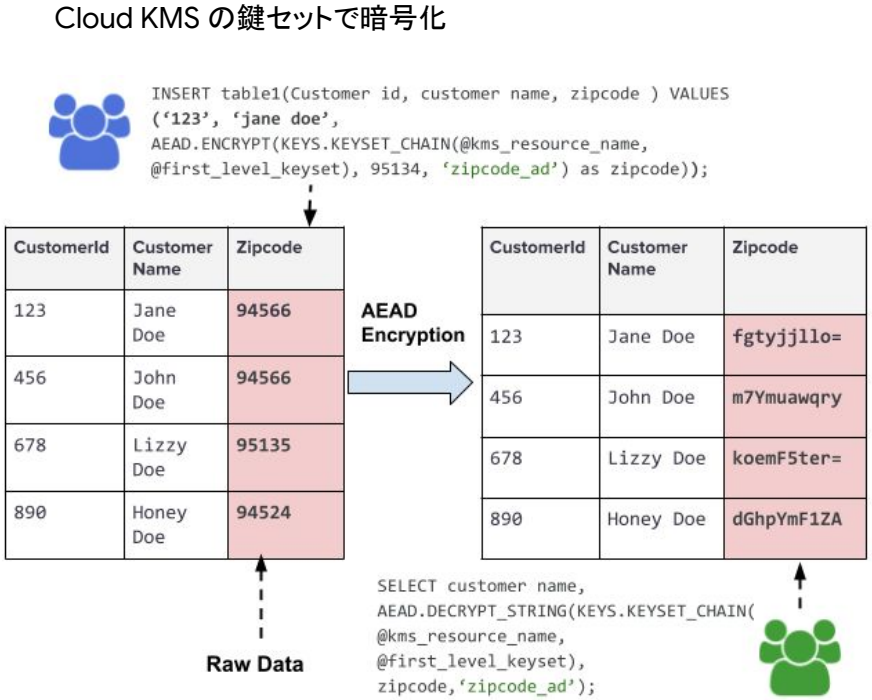


Google 管理、顧客管理の暗号鍵の二択のどちらか
=> 必ず暗号化される

データ暗号化

BigQuery テーブルの値の暗号化

- 個人情報など機微な情報を
  暗号化することができる
  - SQL 関数として AEAD 暗号化関数を利用可能
  - AES GCM 256 bit
- 鍵セット
  - BigQuery 内で作成したもの
  - Cloud KMS で作成したもの



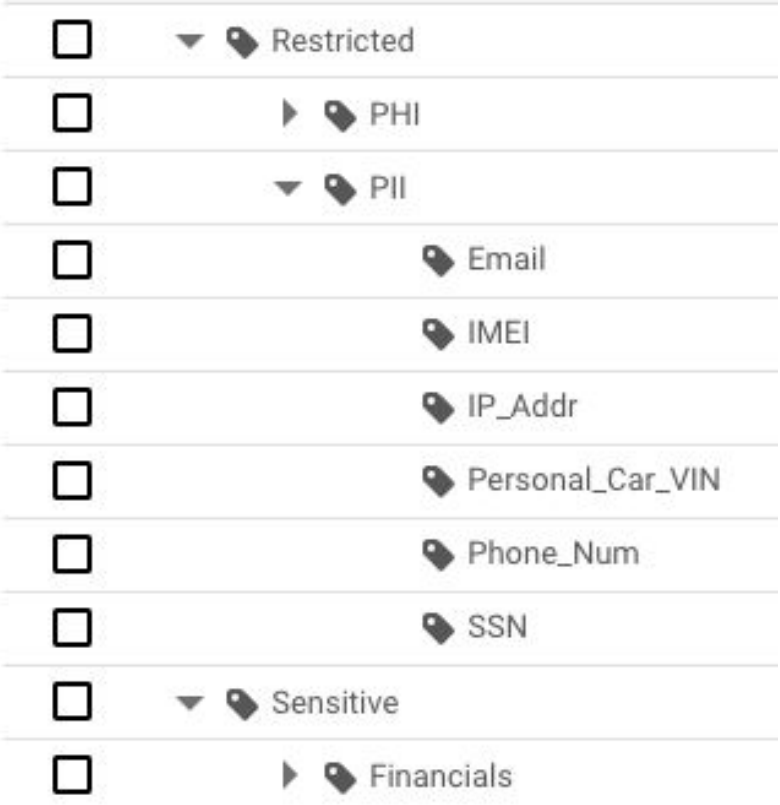
データ分類とアクセス管理

Data Catalog ポリシータグ

データクラス単位でのアクセス管理

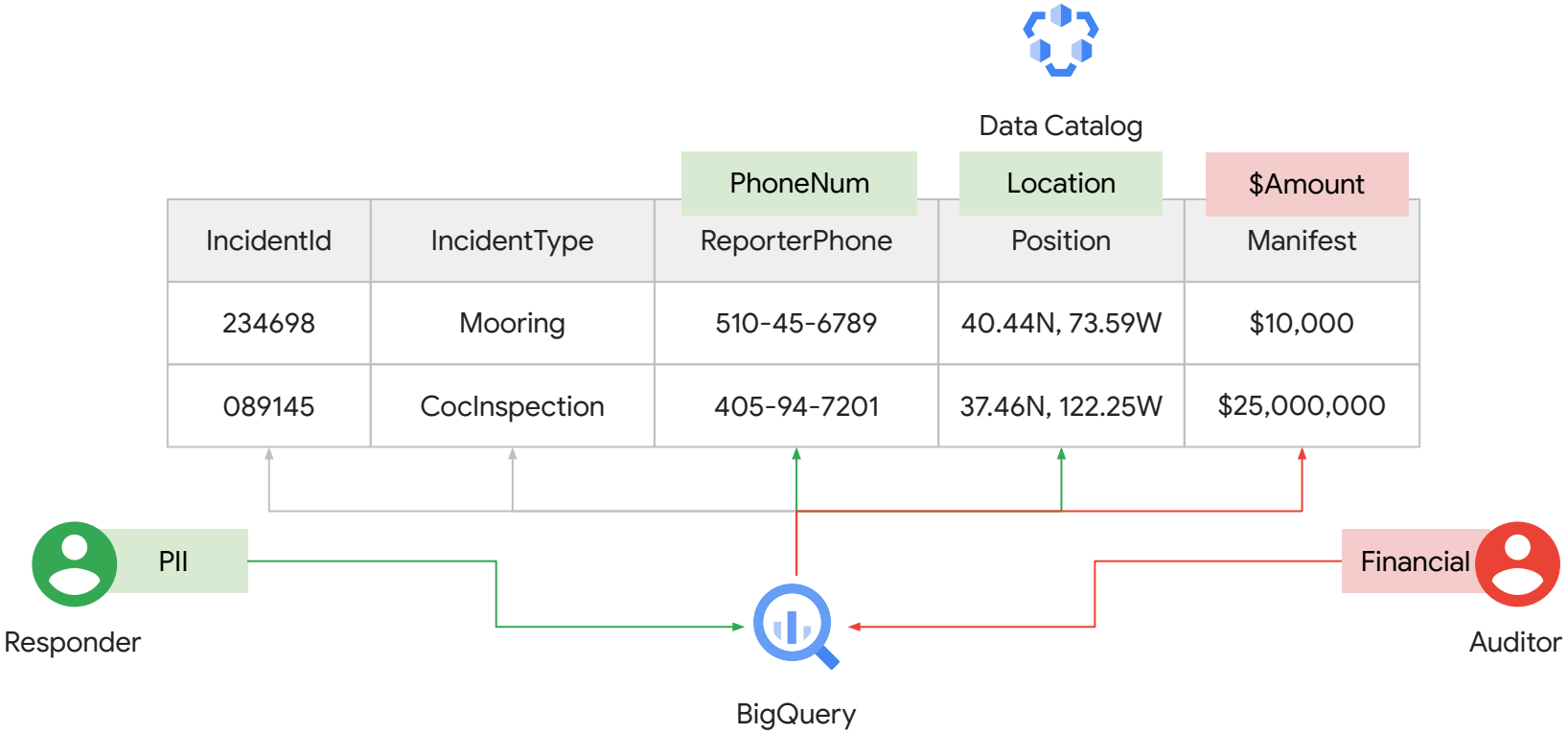
BigQuery の列レベルセキュリティは、Data Catalog で管理するデータクラスの階層に従って制御される

- Data Catalog がデータクラスのポリシーを管理
- 監査ログもデータクラスを考慮



アクセス管理

Data Catalog と BigQuery を組み合わせ、列レベルでのアクセス管理



データマスキングとコンプライアンス

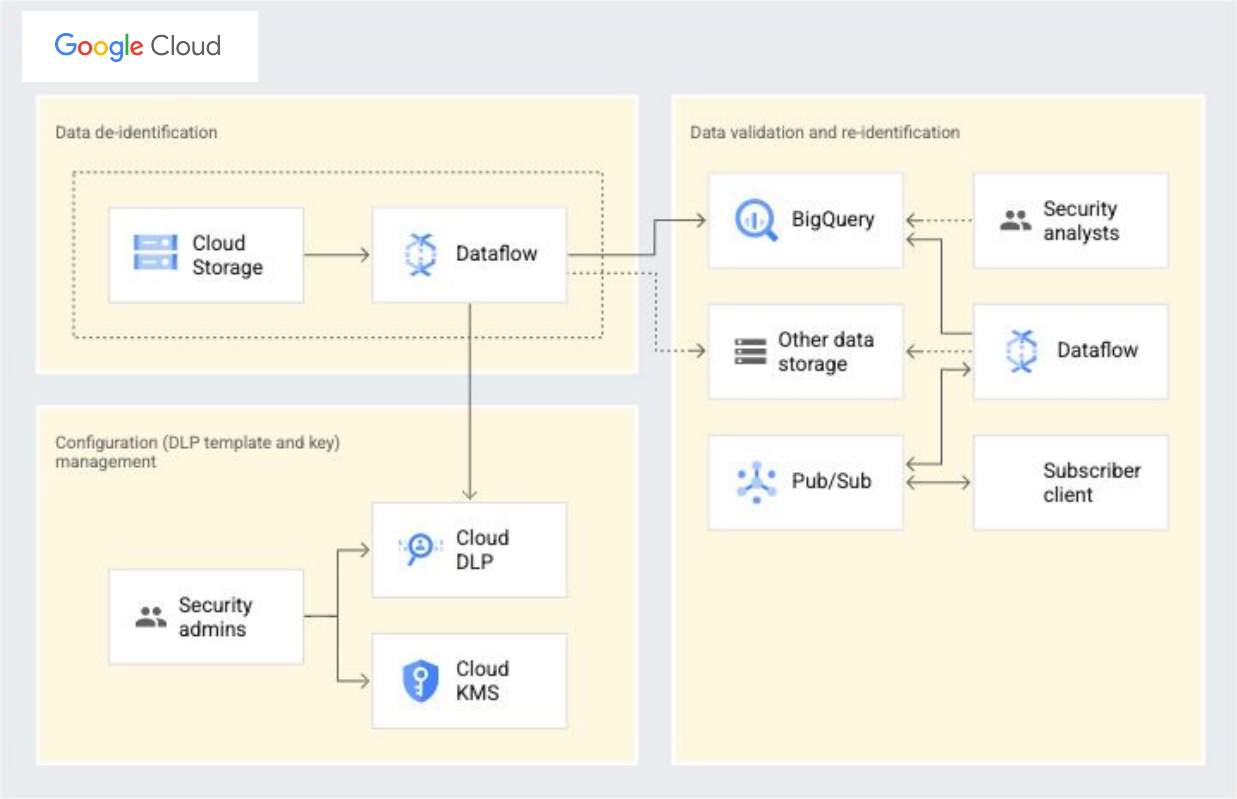
Cloud Data Loss Prevention (DLP)

- データを分類し、個人情報等の機微情報を検知
- データマスキングや、フォーマットを維持した暗号化変換
- k-匿名性など再識別リスクの分析

ID	Job Title	Phone	Comments
359740	Senior Engineer	307-964-0673	Please email them at jane@imadethisup.com
981587	VP, Engineer	713-910-6787	none
394091	Lawyer	692-398-4146	Updated phone to: 692-398-4146
986941	Senior Ops Manager	294-967-5508	none
490456	Junior Ops Manager	791-954-3281	Tried to verify account with their SSN 222-44-5555

データマスキング

Cloud DLP, Cloud KMS と Dataflow の連携による匿名化パイプライン



データに対する責任を実現する Google Cloud サービス


Dataplex


Data Catalog


Looker

データの探索利用	データの管理	データの保護	データに対する責任
データ分類	データライフサイクル & レコード管理	データマスキング	データオーナーシップ
データリネージュ	参照データ(コード)管理	データ暗号化	データポリシー & 基準
メタデータ & データカタログ	マスターデータ管理	アクセス管理	データドメイン単位のガバナンス
データ品質	データ SRE	監査 & コンプライアンス	データ倫理
		データ所在管理	
		データ復旧	

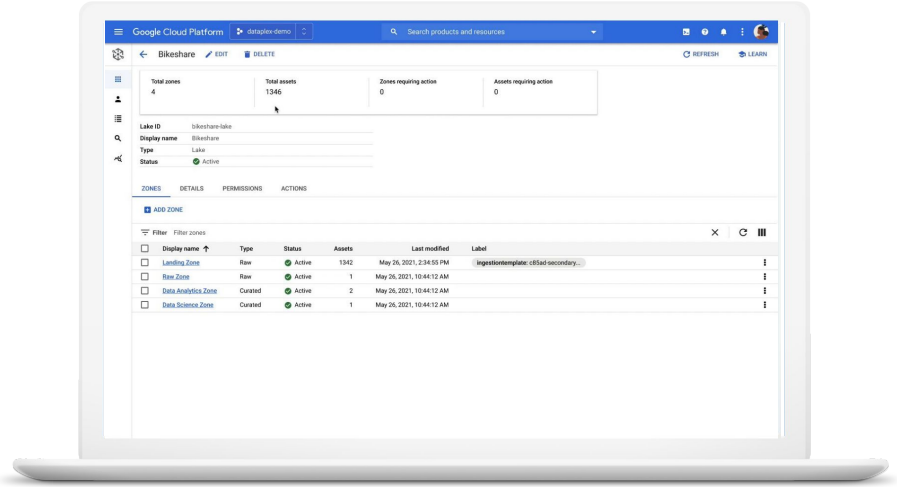
データドメイン単位のガバナンス

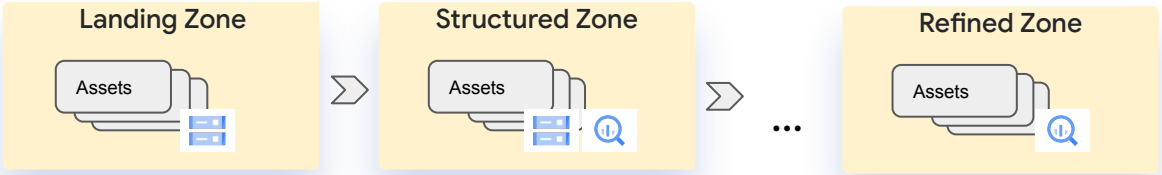
Dataplex による論理的なデータ構成とセキュリティ管理

- ビジネスのユースケースと LOB のニーズに基づいてデータを Lake と Zone に整理

  データを移動せずに、異なるデータストアやプロジェクトからのデータを同じ Zone 内に結びつける

- 論理構造を下記の基盤として利用 :
  - データのアクセス性
  - セキュリティのコントロール





04

まとめ

まとめ

- データ ガバナンスを確立することにより、安全で信頼できるデータの活用が可能に
- データ ガバナンスはツール・テクノロジーだけでなく、人（役割）、プロセスの側面と組み合わせて推進することが必要
- 小さく始めて反復を繰り返しながら拡大するのがベストプラクティス
- Google Cloud のさまざまなサービスによって、データ取得からデータ利用にまたがるライフサイクル全般に対して、データ ガバナンスの機能を実現
- Google Cloud のデータ ガバナンス関連機能に関する詳細
  クラウドにおけるデータ ガバナンス - パート 1 - 個人とプロセス | Google Cloud Blog
  クラウドにおけるデータ ガバナンス - パート 2 - ツール | Google Cloud Blog
  Overview of data security and governance | BigQuery | Google Cloud

Thank you.

Google Cloud