Cloud OnAir

最新アップデート
Google Cloud データ関連ソリューション

2021 年 2 月 5 日 放送

Agenda

1. Smart Analytics ソリューションの方向性
2. Smart Analytics 最新アップデート情報のご紹介

Cloud OnAir

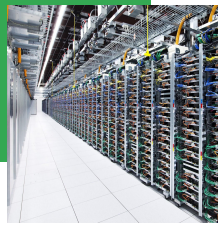
Smart Analytics ソリューションの方向性

Speaker

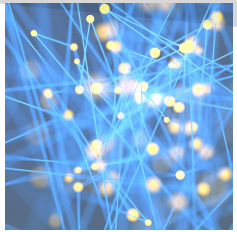
北原 裕士

グーグル・クラウド・ジャパン合同会社
データアナリティクス スペシャリスト

Smart Analytics ソリューション

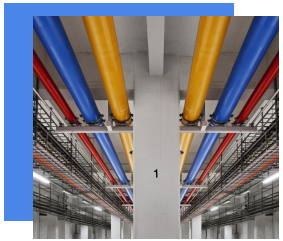


BigQuery を用いて
データ分析プラットフォーム
をモダナイズ



Dataproc などによる
オープンで柔軟な
マネージドデータレイク構築

Dataflow＆PubSub を利用し
俊敏なビジネスを実現



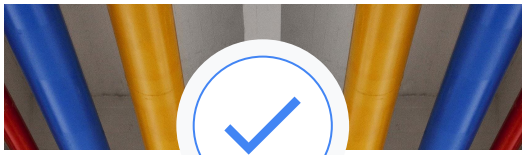
エンタープライズ対応の
データ連携サービス によってデータ
のサイロ化を解消

Looker で実現する
データドリブンビジネス
エンタープライズ BI と
データ アプリケーション

Smart Analytics のビジョン

オープン /
マルチクラウド

インテリジェント



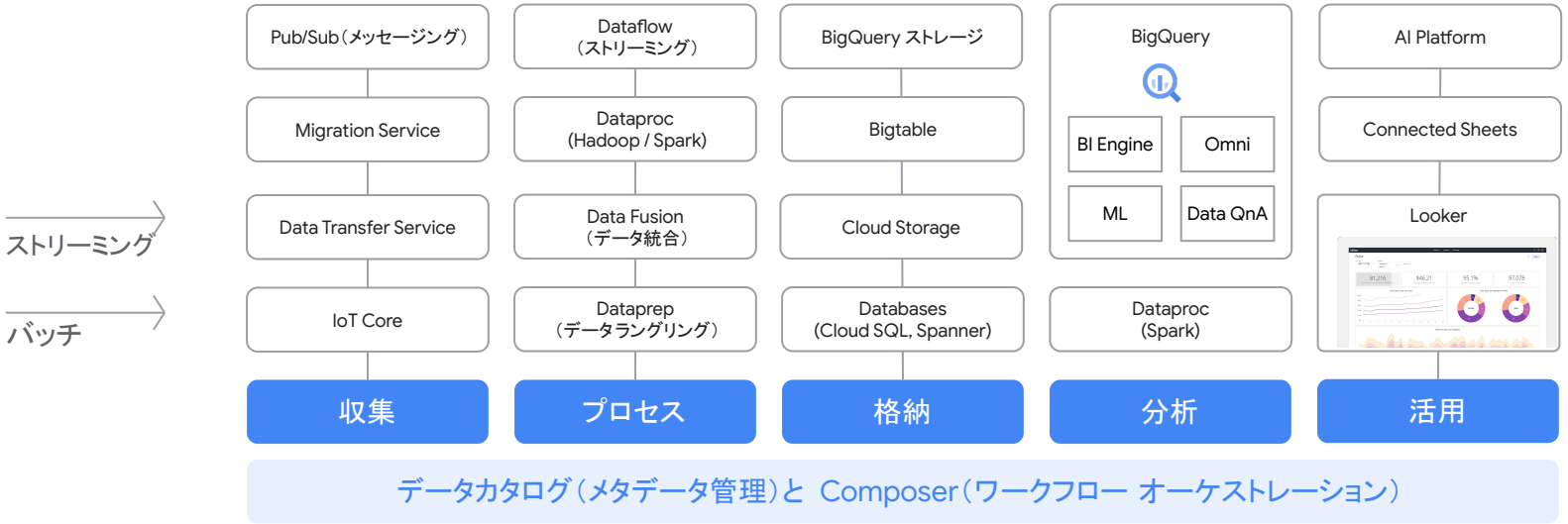
フレキシブル

ミッションクリティカルなワークロードに対しても**実証済の信頼性**を持つ

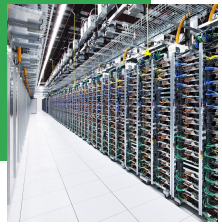
オープン ｜ インテリジェント ｜ フレキシブル な

アナリティクス基盤を提供すること

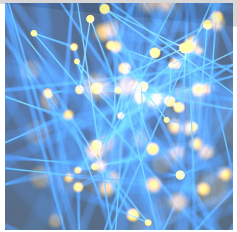
Smart Analytics プラットフォーム



Smart Analytics ソリューション

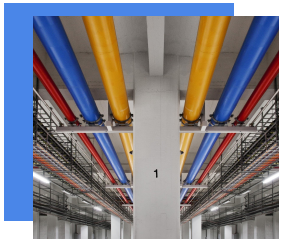


BigQuery を用いて
データ分析プラットフォーム
をモダナイズ



Dataproc などによる
オープンで柔軟な
マネージドデータレイク構築

Dataflow＆PubSub を利用し
俊敏なビジネスを実現



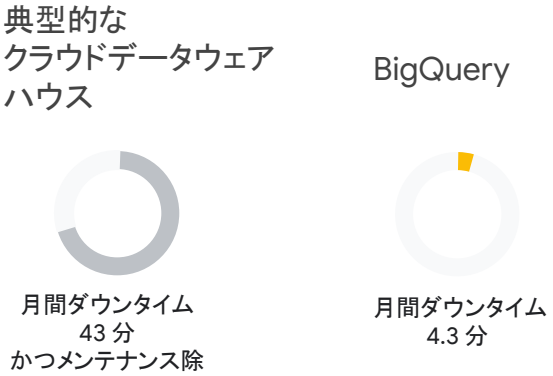
エンタープライズ対応の
データ連携サービス によってデータ
のサイロ化を解消

Looker で実現する
データドリブンビジネス
エンタープライズ BI と
データ アプリケーション

BigQuery でデータ分析プラットフォームをモダナイズ

BigQuery

- 99.99 % の高い可用性を提供
- 複数ゾーンでの自動冗長化・レプリケーションによる高可用性と高耐久性
- 東京・大阪リージョンにおけるデータセットコピーサポート
- VPN や 専用線による接続サポートとデータ持ち出し制限



クラウド データベース評価

今年から開設された 2020 年 Gartner マジック クアドラントのクラウド データベース管理システム（DBMS）部門のリーダーに Google が選出されたことをご報告します。この評価は Google Cloud のデータ分析およびデータベースのビジョンと戦略によるものであり、Google Cloud を最適なデータ プラットフォームとして選んだあらゆる業界と地域のお客様の成長に表れています。

Gartner のマジック クアドラントにおいてビジョンの完全性の面で特に優れたリーダーとして、Google が 3 社のベンダーのうちの 1 社に選出されました。Google はマルチクラウドとハイブリッドの約束を果たし、あらゆる地域と業界におけるさまざまな顧客ベースでの導入を実現しています。また、強力な財務ガバナンス機能を備えた柔軟な料金設定の新しい基準を打ち立て、多様なエコシステム全体で各企業と提携しています。また、ビジョンを実現し、BigQuery Omni でマルチクラウド データ ウェアハウスを提供する最初のハイパースケール プロバイダとして自社の取り組みに誇りを持っています。

Gartner、Google を 2020 年マジック クアドラントのクラウド データベース管理システム部門のリーダーに選出

BigQuery でデータ分析プラットフォームをモダナイズ

フルマネージド データウェアハウス
高スケーラビリティの分析プラットフォーム
サーバーレスでスケールと信頼性を提供

インテリジェント データウェアハウス
将来を予測するデータ分析
AI/ML による予測分析

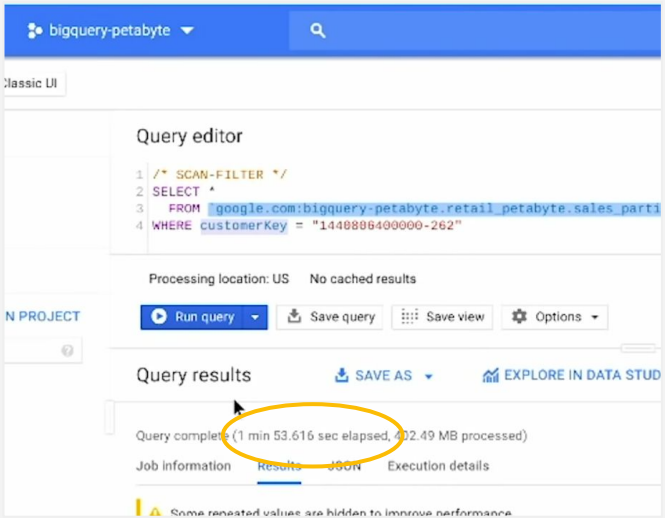
セルフサービス データウェアハウス
全てのユーザーに対応するデータ分析
セルフサービスでデータを利用

マルチクラウド データウェアハウスプラットフォーム横断でデータを分析
マルチクラウドのデータ分析

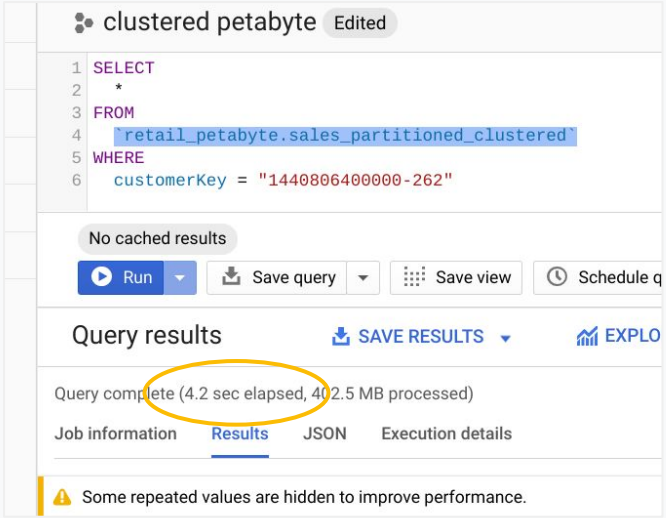
リアルタイム データウェアハウス
データをリアルタイムで分析
常に高速、常に最新

継続的な改善

2018



2020



管理機能のアップデート

Reservations

エンタープライズ ワークロード管理

予測可能な料金

アイドル スロットの共有

Flex slots

60 秒から必要最小時間のスロット確保

秒単位の課金

いつでもキャンセル可能

データベース管理

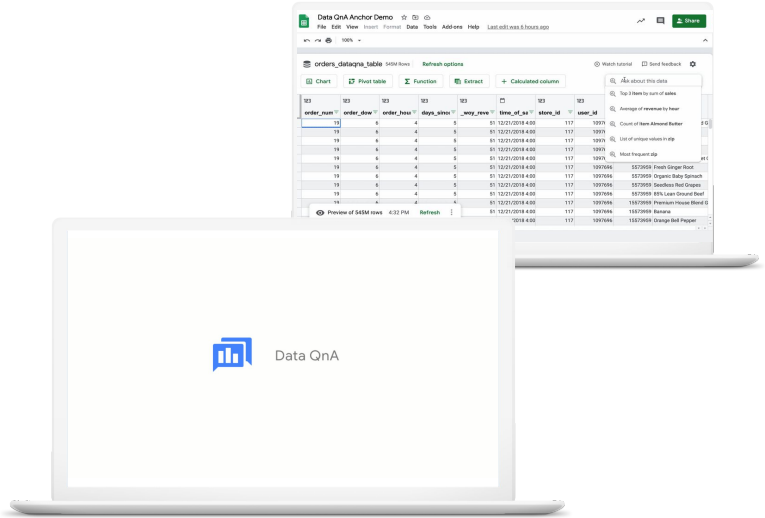
Information Schema

データ型の拡張

SQL 対応構文拡大

Data QnA: BigQuery の自然言語インターフェース

1. 自然言語を使ったセルフサービス分析を通じてインサイトを民主化
2. アドホック レポートを不要にすることで BI チームの生産性が向上
3. Google スプレッドシート、BigQuery、Chatbots、カスタム UI（API 経由）、Looker、Google Voice など、さまざまなインターフェースを介してアクセス



リアルタイム データウェアハウス

Streaming V2

標準で数百万 QPS（クエリ / 秒）

Exactly once セマンティクス

クエリ パフォーマンスの低下なし

マテリアライズド ビュー

効率的でシームレスなメンテナンス

常時整合

スマートなクエリのリルーティング

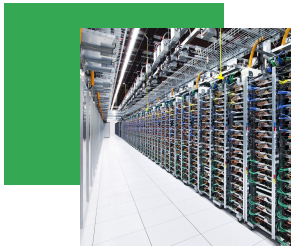
BI Engine

インメモリ実行エンジン

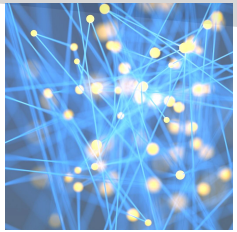
1 秒未満のクエリレスポンス

スマート チューニング

Smart Analytics ソリューション

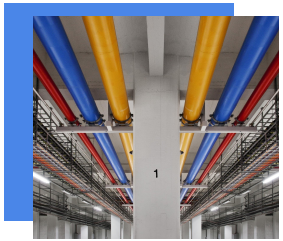


BigQuery を用いて
データ分析プラットフォーム
をモダナイズ



Dataproc などによる
オープンで柔軟な
マネージドデータレイク構築

Dataflow & PubSub を利用し
俊敏なビジネスを実現



エンタープライズ対応の
データ連携サービス によってデータ
のサイロ化を解消

Looker で実現する
データドリブンビジネス
エンタープライズ BI と
データ アプリケーション

オープンで柔軟なマネージド データレイク構築

Dataproc

オープン

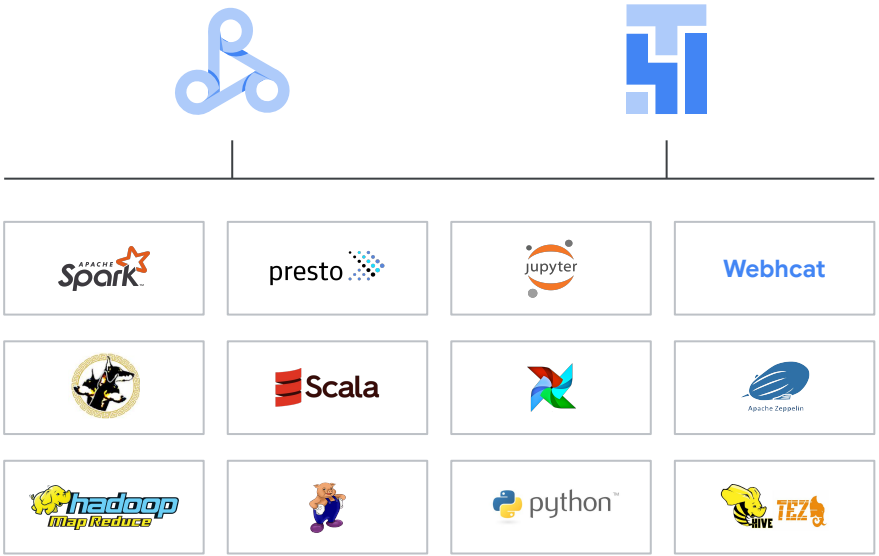
OSS データアナリティクス
オープンでセキュアなマネージド環境
マネージド Hive メタストア
Dataproc on Kubernetes

フレキシブル

スピードと柔軟性
素早いデプロイと柔軟なスケーラビリティ

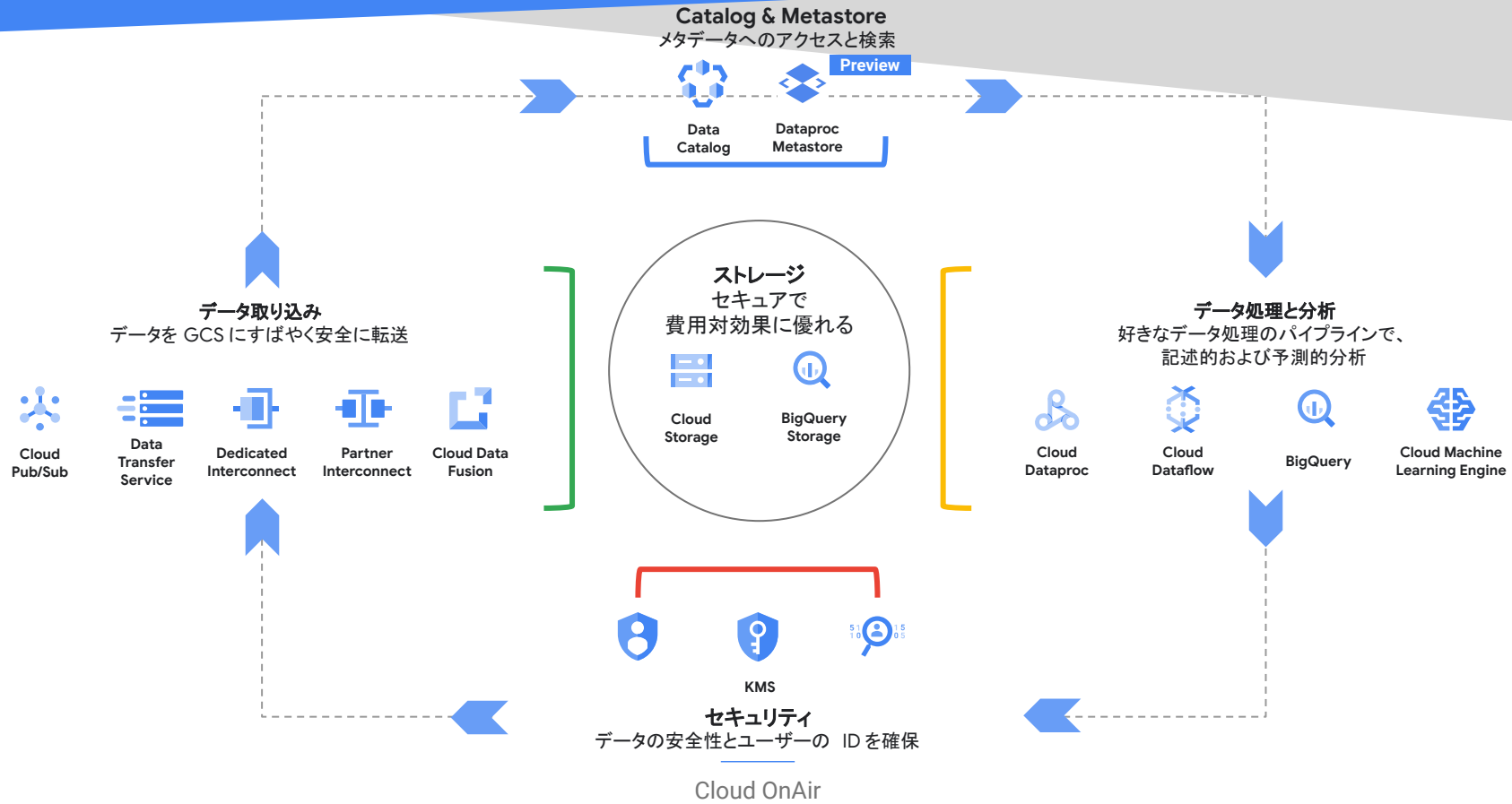
インテリジェント

エンタープライズ対応ノートブック
ノートブックを中心とした分析環境



... 他にも様々なオプション コンポーネントに対応

オープンで柔軟なマネージド データレイク構築

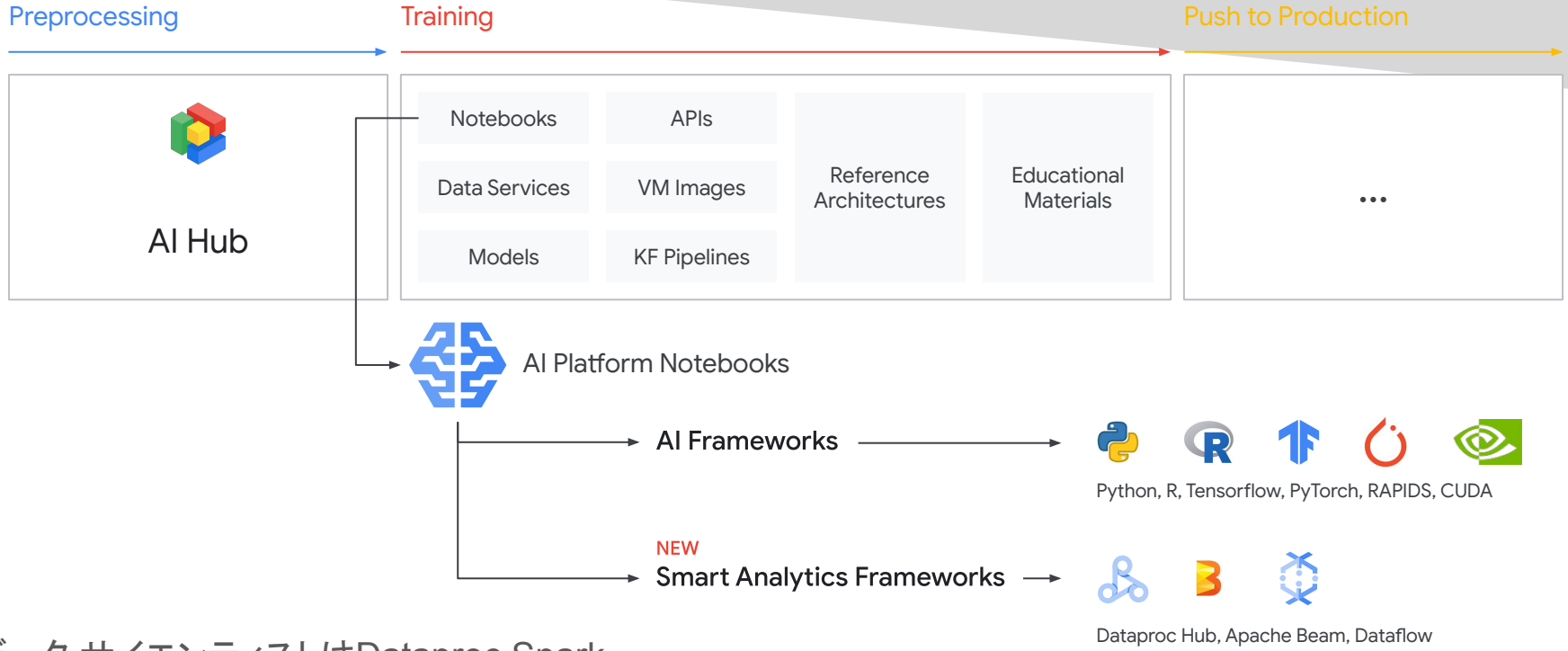


Cloud OnAir

エンタープライズ対応ノートブック
Dataproc Hub + AI Platform Notebooks

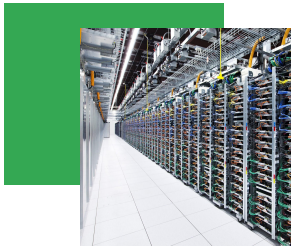
- マネージド Jupyter Notebook 環境からDataproc クラスタのApache Spark などへ簡単に接続
- 管理者
  - ノートブック インスタンスとDataproc クラスタを事前構成
  - ノートブック利用管理と利用状況のモニタリング
- データ サイエンティスト
  - セルフサービスでノートブック インスタンスとSpark クラスタを利用
  - PySpark, SparkML などのオープソース ツールを利用
  - GPU を接続したクラスタへのノートブックの接続も可能

エンタープライズ対応ノートブック
Dataproc Hub + AI Platform Notebooks

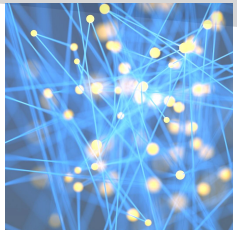


データ サイエンティストはDataproc Spark,
Apache Beam, Dataflow もセルフサービスで利用可能

Smart Analytics ソリューション

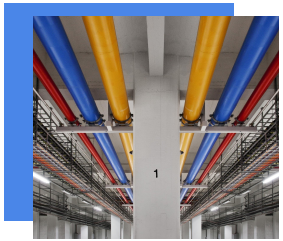


BigQuery を用いて
データ分析プラットフォーム
をモダナイズ



Dataproc などによる
オープンで柔軟な
マネージドデータレイク構築

Dataflow＆PubSub を利用し
俊敏なビジネスを実現



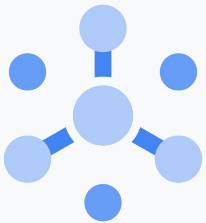
エンタープライズ対応の
データ連携サービス によってデータ
のサイロ化を解消

Looker で実現する
データドリブンビジネス
エンタープライズ BI と
データ アプリケーション

ストリーミング処理で俊敏なビジネスを実現

1 **取込**

信頼性の高い
データ取込と配布



Pub/Sub

2 **変換**

素早くシンプルに、高速で正確な処理を実行

Dataflow

3 **分析 / サービング**

機械学習と
データウェアハウス

AI Platform

BigQuery

Bigtable

Cloud SQL

ストリーミングデータ取込

1 取込

Google Cloud

Pub/Sub

リアルタイム分析の　ためのメッセージングとイベント取り込み

Pub/Sub Lite

最適なコストでの イベント取り込みとデータ メッセージング

パートナー
ソリューション

Confluent Cloud

フルマネージド Kafka によるオンプレミスからクラウドへの簡単な移行

変更データキャプチャ

Debezium ベースの コネクタを使用した CDC 向けの Dataflow サンプル ソリューション

データ変換機能

2 変換

Dataflow SQL

BigQuery のウェブ UI 内で直接 SQL を使用して、ストリーミング Dataflow パイプラインを開発

Jupyter Notebook の統合

インタラクティブな Jupyter 環境でパイプラインを反復的に構築してプロトタイプを作成

フレックステンプレートによる共有とスケーリング

すべての Dataflow パイプラインで開発、共有、スケーラブル導入が容易に

リコメンデーションで作業を迅速に

コストの削減と最適化に向けたガイダンスに従って Dataflow を操作

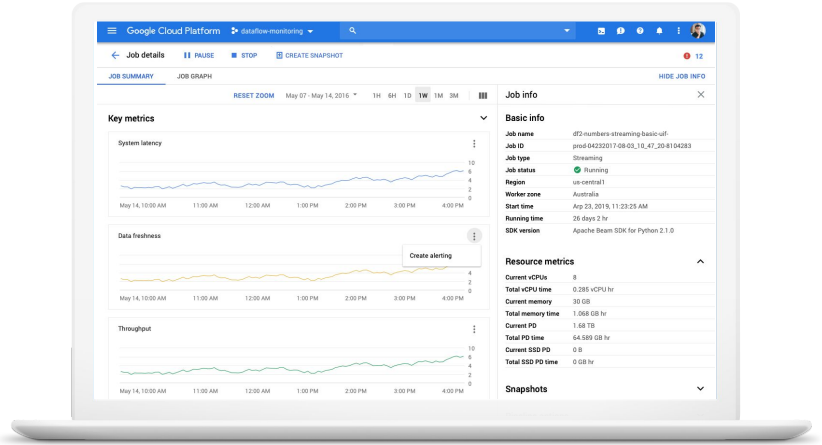
可視性に優れたオペレーション

パイプライン オブザーバビリティ ダッシュボード

すべての重要なパイプライン指標が時間と関連付けられた 1つのダッシュボードに表示されるため、パイプラインオペレーターは以下のことが可能に:

パフォーマンスやコストをさらに改善するために、 Dataflow の自動スケーリングの決定の仕組みを理解する

パイプラインのレイテンシとスループットを最適化する



可視性に優れたオペレーション

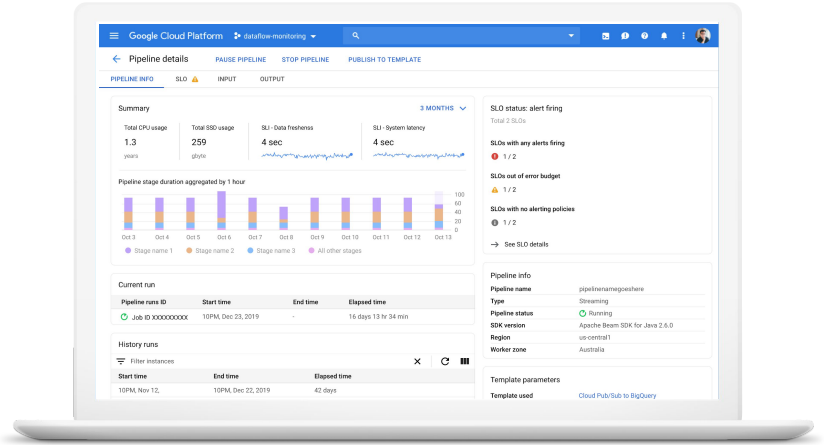
マネージド パイプライン

何百ものパイプラインや繰り返しバッチジョブを実装する際の　オペレーションが明確になり、オペレーターは以下のことが　可能に:

複数のジョブを実行する場合のリソース割り当てを理解する

データ更新 SLO を定義して管理する

個別のパイプラインステージにドリルダウンして修正　最適化する



ストリーミング AI/ML

3 分析 / サービング

Google Cloud AI やワーカー ローカルモデルによるオンライン予測を、Dataflow ストリーミング パイプラインに統合

パターン認識

異常検出

予測

Dataflow と Cloud AI Platform を統合するための新しい Apache Beam トランスフォーム

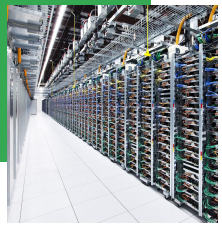
ストリーミング予測用の Apache Beam の新しいトランスフォーム tfx_bsl/runInference

LSTM および BoostedTrees を使用した異常検出ソリューション

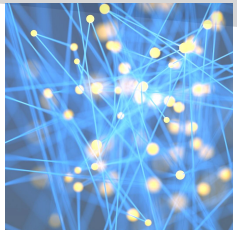
ビデオやイメージのパターン検出のソリューション

cloud.google.com/dataflow/ にアクセス

Smart Analytics ソリューション

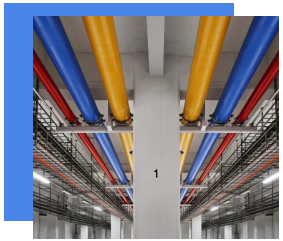


BigQuery を用いて
データ分析プラットフォーム
をモダナイズ



Dataproc などによる
オープンで柔軟な
マネージドデータレイク構築

Dataflow＆PubSub を利用し
俊敏なビジネスを実現



エンタープライズ対応の
データ連携サービス によってデータ
のサイロ化を解消

Looker で実現する
データドリブンビジネス
エンタープライズ BI と
データ アプリケーション

データ連携サービスでデータサイロを解消

Data Fusion

フルマネージド、クラウドネイティブなデータ統合ソリューション

- コーディング不要なGUI 上でのデータ変換プロセスの開発
- 100 以上のプラグイン- コネクター、変換、アクション
- テストとデバッグ機能を備えた、1000 以上の変換プロセス実行
- あらかじめ用意されたパイプライン
- データセットや列レベルでのデータリネージュ管理

Data Catalog

フルマネージドでスケーラブルなメタデータ管理と検索

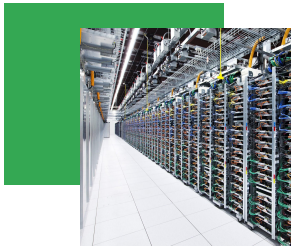
- シンプルなメタデータ検索
- データガバナンス機能組み込み
- メタデーター元管理

Composer

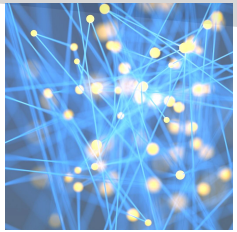
Apache Airflow で構築された、フルマネージドのワークフローオーケストレーションサービス

- ハイブリッドおよびマルチクラウド環境にまたがるパイプラインを作成、スケジューリング、モニタリング
- GCP サービスとの統合
- 特定のベンダーに依存する必要がなくなり、使用も簡単

Smart Analytics ソリューション

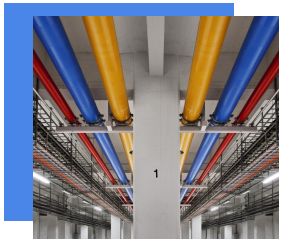


BigQuery を用いて
データ分析プラットフォーム
をモダナイズ



Dataproc などによる
オープンで柔軟な
マネージドデータレイク構築

Dataflow＆PubSub を利用し
俊敏なビジネスを実現

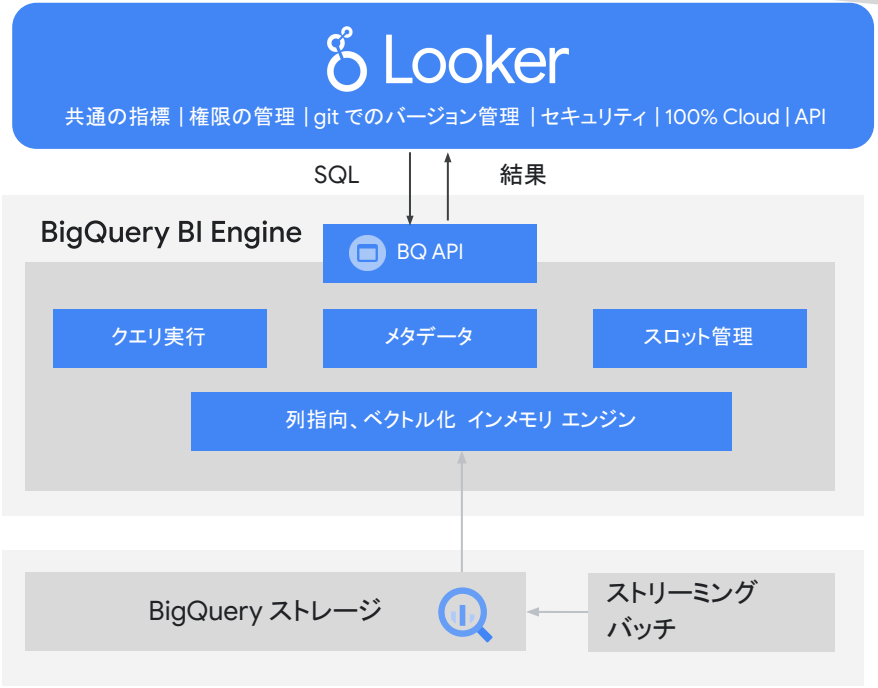


エンタープライズ対応の
データ連携サービス によってデータ
のサイロ化を解消

Looker で実現する
データドリブンビジネス
エンタープライズ BI と
データ アプリケーション

Looker と BigQuery BI Engine の連携

1. Looker は BI Engine とシームレスに連携
2. お客様は何も変更せず BI Engine によるパフォーマンス向上の恩恵を受けることが可能
3. OLAP キューブ構築、運用の手間が不要
4. 秒以下の遅延、スタースキーマ サポート



Smart Analytics
最新アップデート情報のご紹介

BigQuery アップデートまとめ（2020 年後半～）

BigQuery

SQL 関連

- GA DDL の拡張
- GA 認可済 UDF
- GA ユニコード テーブル名
- GA 動的 SQL ステートメント
- GA 関数追加
- GA ジョブ キャンセル
- GA 日付算術演算子
- Preview BigNumeric

UI

- Preview 新しい UI
- Preview 検索と自動補完

データ転送

- GA 大阪リージョン対応
- GA VPC SC 対応

ML

- GA Matrix Factorization モデル
- GA 時系列モデル
- GA 勾配ブースト モデル
- GA DNN モデル
- GA モデル エクスポート

性能最適化

- GA 時間パーティション単位選択

ワークロード管理

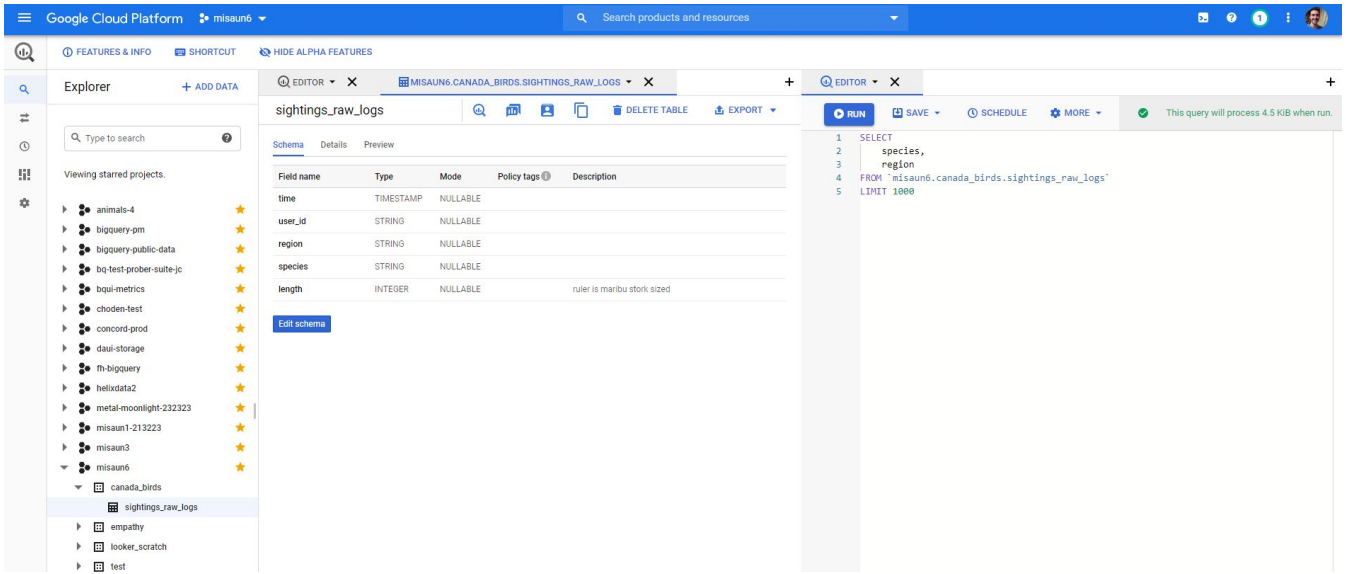
- GA スロット購入 100 単位
- GA INFORMATION_SCHEMA

セキュリティ

- GA 列レベル セキュリティ

新しいユーザー インターフェース

- マルチタブ編集
- クエリエディタ
  - 入力補完
  - 折り畳み
  - ショートカット
- リソースパネル
  - 動的にロード
  - 検索機能の向上
  - シングルクリックでピン留め



DDL 拡張

- テーブルオペレーションの新しいコマンド
  - ADD COLUMN（ALTER TABLE コマンド）
  - TRUNCATE TABLE
    - 無料のオペレーション
  - Unicode テーブルの命名
- 外部ストレージの読み取りと書き込み
  - CREATE OR REPLACE EXTERNAL TABLE
  - DROP EXTERNAL TABLE
  - EXPORT DATA

関数追加

- 新しいDATE 関数
  - DATE の加減演算子 (“+”, ”-”)
  - LAST_DAY
- 新しい文字列関数
  - 連結演算子 (“||”)
  - LEFT, RIGHT
  - INSTR
  - INITCAP
  - TRANSLATE
  - ASCII
  - UNICODE
  - CHR
  - SOUNDEX
  - REGXP_SUBSTR
  - REGXP_EXTRACT
  - REGXP_INSTR
  - OCTET_LENGTH

INFORMATION SCHEMA の拡張

- テーブルの情報スキーマ
  - TABLES
  - TABLE_OPTIONS
  - COLUMNS
  - COLUMN_FIELD_PATHS
- ビュー情報スキーマ
  - VIEWS
- ルーティンの情報スキーマ
  - ROUTINES
  - ROUTINE_OPTIONS
  - PARAMETERS
- データセットの情報スキーマ
  - SCHEMATA
  - SCHEMATA_OPTIONS

Data Lake アップデートまとめ（2020 年後半～）

Dataproc

Hadoop / Spark

GA
2.0 イメージ
Apache Hadoop 3
Apache Spark 3

クラスタ 管理

Preview
高度な柔軟性モード

Preview
クラスタ起動と停止

Preview
永続履歴サーバー

ジョブ 管理

Preview GA
ワークフロータイムアウト

GA
再実行可能なジョブ

セキュリティ

GA
個人用クラスタ認証

Preview
サービスアカウントによる マルチテナンシー

メタデータ 管理

Preview
Dataproc Metastore

オプション コンポーネント

GA
Docker

GA
Flink

GA
Ranger

GA
Solr

コンピュート ノード

GA
単一テナントノードサポート

GA
バランス永続ディスクサポート

GA
Shielded VM サポート

Cloud OnAir

Dataproc 2.0

デフォルト
コンポーネント

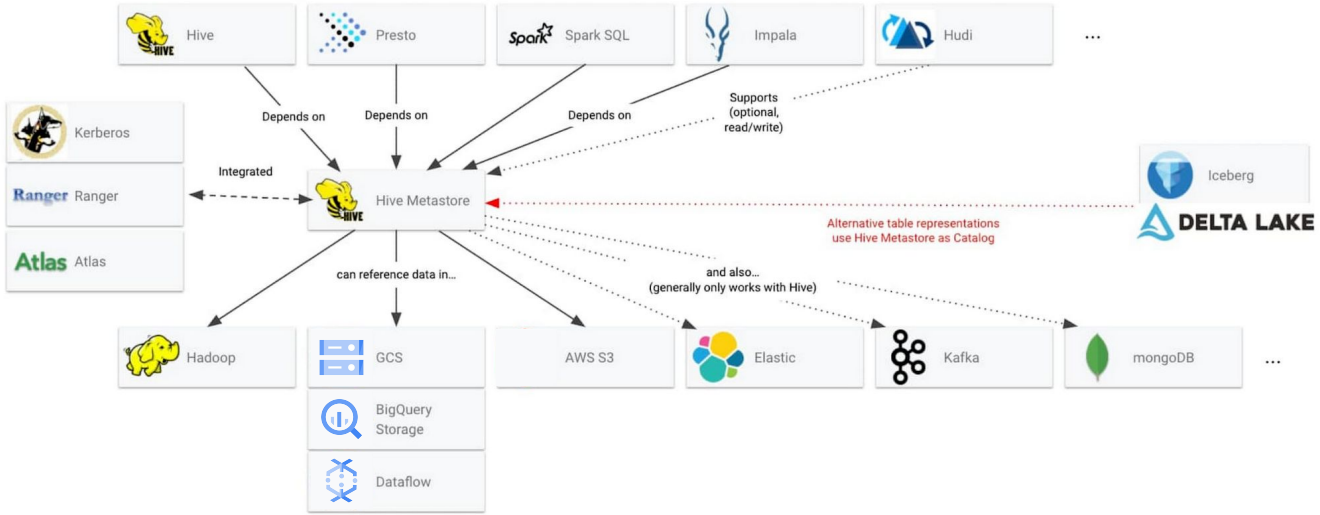
Apache **Hadoop**	Apache **Spark**	Apache **Hive**	**Java**
3.2	3.1	3.1	11
Apache **Iceberg**	Apache **Pig**	Delta Lake	**Python**
0.10.0	0.18.0	0.7.0	3.8

オプション
コンポーネント

Apache **Flink**	Apache **HBase**	Apache **Druid**	**Presto SQL**
1.12	2.2	0.20	340
Apache **Ranger**	Apache **Atlas**	Apache **Knox**	**JuypterLab**
2.0	2.0	1.4	3.0

Dataproc Metastore

高可用性と自動修復機能を備えたオープンソースのフルマネージド Apache Hive メタストアサービス



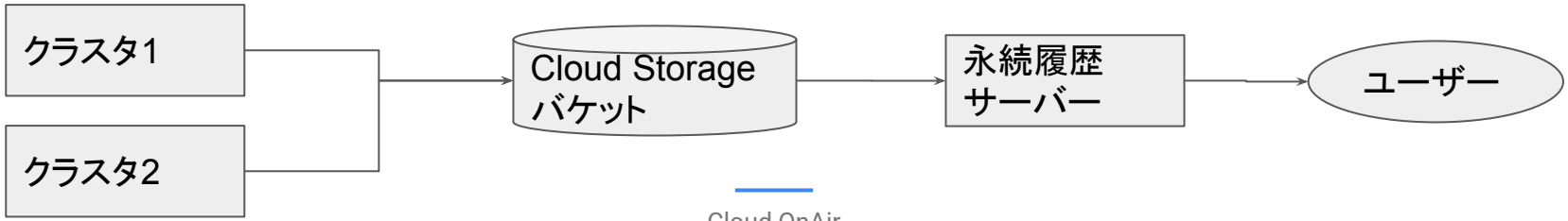
永続履歴サーバー

概要

- Dataproc クラスタを削除後もジョブ履歴を確認できるよう、ジョブ履歴表示用のUIを提供

特徴

- ジョブ実行するサーバーではCloud Storage にログを蓄積するように構成
- 単一ノードのDataproc クラスタ上で実行される永続履歴サーバーから、Cloud Storage のジョブ履歴ファイルへアクセス



サービスアカウントによるマルチテナンシー

概要

- セキュアにDataproc クラスタを複数ユーザーで共有する仕組み

特徴

- サービスアカウントにマップされた複数のユーザーでクラスタを共有
- ユーザーはそれぞれのワークロードをそれぞれの別のユーザーとしてクラスタで実行する
  - ジョブはKerberos プリンシパルとして特定のOS ユーザーで実行
  - GCS などのGoogle Cloud リソースには、マップされたサービスアカウントのクレデンシャルを用いてアクセス

制約

- Kerberos 認証を有効にしてクラスタを構成
- サービスアカウントにマップされていないユーザーはジョブを実行できない
- コンポーネント ゲートウェイは使用できない
- ワークフローの利用は不可

Streaming アップデートまとめ（2020 年後半～）

Dataflow

- GA — ワーカーVM へのネットワークタグ
- GA — フレックステンプレート
- GA — Java 11 ランタイム
- Preview — カスタム コンテナ
- Preview — GPU サポート
- GA — ノートブック使用インタラクティブ開発
- Preview — Beam DataFrames

Pub/Sub

- Preview GA — メッセージ順序指定
- GA — Pub/Sub Lite
- GA — メッセージフィルタリング

データ統合ソリューション アップデートまとめ（2020 年後半～）

Data Fusion

- GA バージョン 6.1.4 / 6.2.2 / 6.3.0
- GA Dataproc クラスタ自動スケーリング
- GA インスタンス作成時のバージョン指定
- GA Dataproc 実行時サービスアカウント指定
- GA インスタンス バージョンアップ
- GA BigQuery ビュー / マテリアライズドビューのサポート

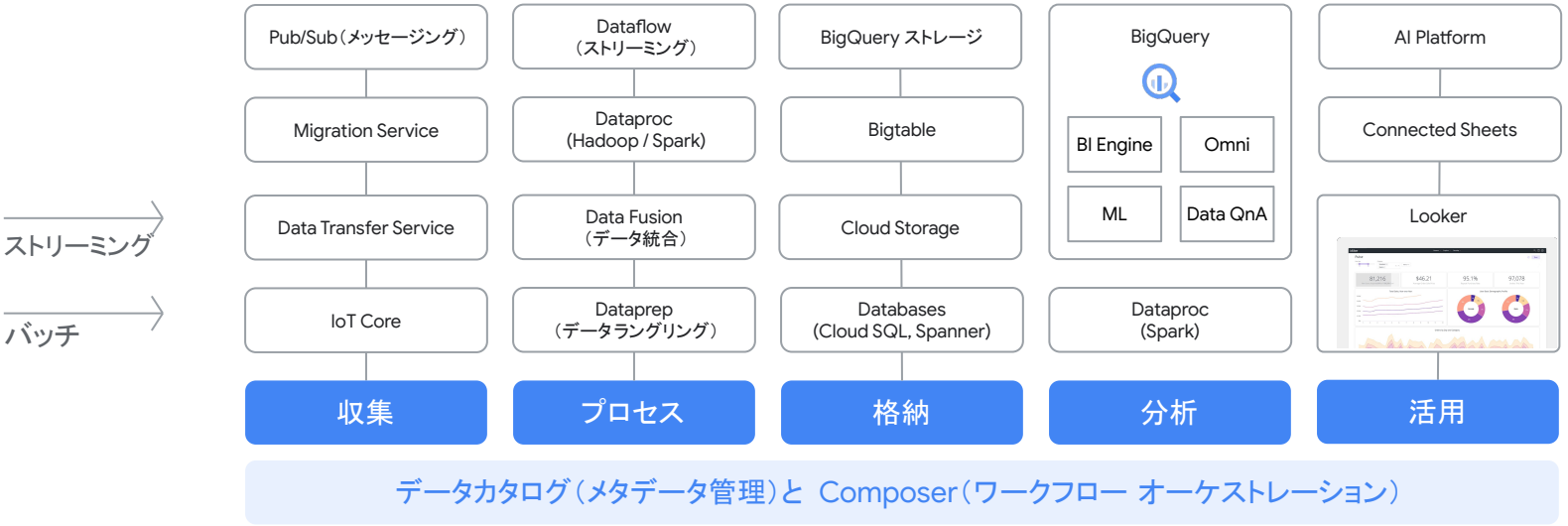
Composer

- GA Apache Airflow 1.10.10 / 1.10.12
- GA VPC SC 対応
- GA シークレットマネージャー
- Preview 新しい Logs タブ
- GA Composer Log 出力項目追加
- GA Airflow web server ネットワーク ACL
- GA Cloud Monitoring メトリクス追加
- Preview 顧客管理のセキュリティキー（CMEK）

参考情報

Smart Analytics 各プロダクト リリースノート

https://cloud.google.com/release-notes/all#data-analytics



Thank you

Cloud OnAir