Google Cloud

Google Cloud UPDATES:
2023 年 Data Analytics / ML編

2023-02-20
Google Cloud

Google Cloud Updates について

- このイベントでは、Google Cloud に関する四半期分のアップデートの振り返りを行っていきます
- 既存ユーザ様を主な対象としているため、基本的には、プロダクトの概要レベルの説明は行いません
- 出入りは自由ですが、退出時にアンケートにご協力下さい
- 質問は、Chat か、こちらのフォームをご活用下さい
- 今回は、2022-10-01 ~ 2023-01-30 の DataAnalytics / AI,ML 系のプロダクトのアップデートの振り返りを行っていきます

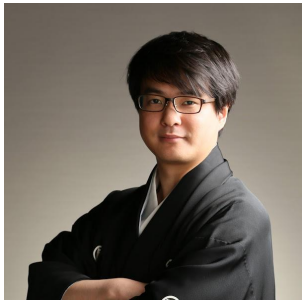
本日のスピーカー

Yu Yamada

BigQuery,
ガバナンス

Yutaro Saito

ETL & Workflow



Shingo Furuyama

BQML, Vertex AI,
その他

01

Data Analytics - BigQuery ガバナンス

GA 2022-10-10

Multi-statement transactions

/ 特徴

- 複数の SQL 文を 1 つのトランザクションとして実行
- DML、SELECT、一時表、一時関数の作成、削除をサポート
- INFORMATION_SCHEMA.JOBS_BY_* でトランザクション ID やトランザクションの成否が確認可能

/ ユースケース

- 複数の処理を束ねてアトミックに扱いたい場合
- マルチ ステートメントトランザクションを使っている他社 DB からのバッチ処理移行

```
BEGIN

BEGIN TRANSACTION;

CREATE TEMP TABLE tmp
  AS SELECT * FROM mydataset.NewArrivals WHERE warehouse =
'warehouse #1';

DELETE mydataset.NewArrivals WHERE warehouse = 'warehouse #1';

MERGE mydataset.Inventory AS I
USING tmp AS T
ON I.product = T.product
WHEN NOT MATCHED THEN
 INSERT(product, quantity, supply_constrained)
 VALUES(product, quantity, false)
WHEN MATCHED THEN
 UPDATE SET quantity = I.quantity + T.quantity;

DROP TABLE tmp;

COMMIT TRANSACTION;

EXCEPTION WHEN ERROR THEN
  -- Roll back the transaction inside the exception handler.
  SELECT @@error.message;
  ROLLBACK TRANSACTION;
END;
```

Multi-statement transactions

Dataplex - Data Exploration Workbench

GA 2022-10-20

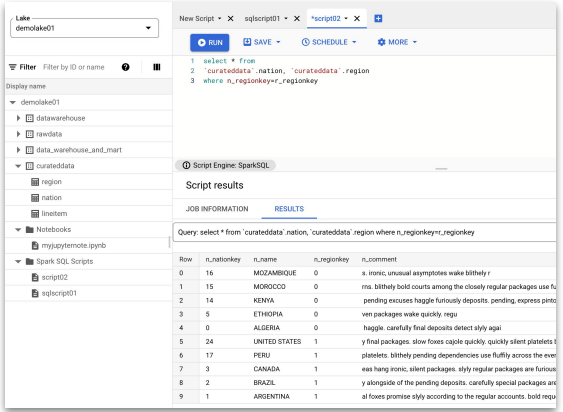
/ 特徴

- Cloud Storage / BigQuery にあるデータをSpark SQL / notebook で探索ができる
- スケジュールも可能
- 共有してコラボレーションが可能
- サーバレスなインフラで実行

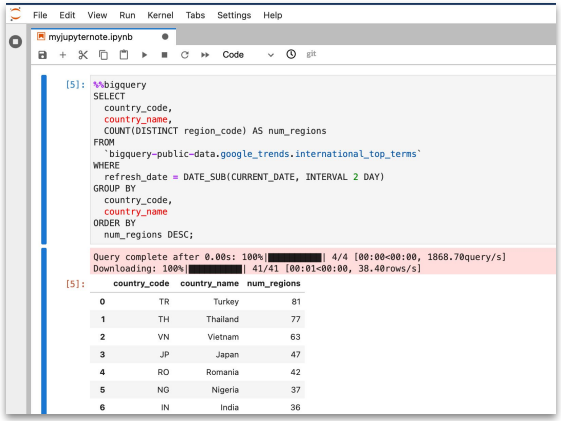
/ ユースケース

- Dataplex 内で管理されているデータに対してすぐに分析や探索がしたい

Spark SQL



Notebook

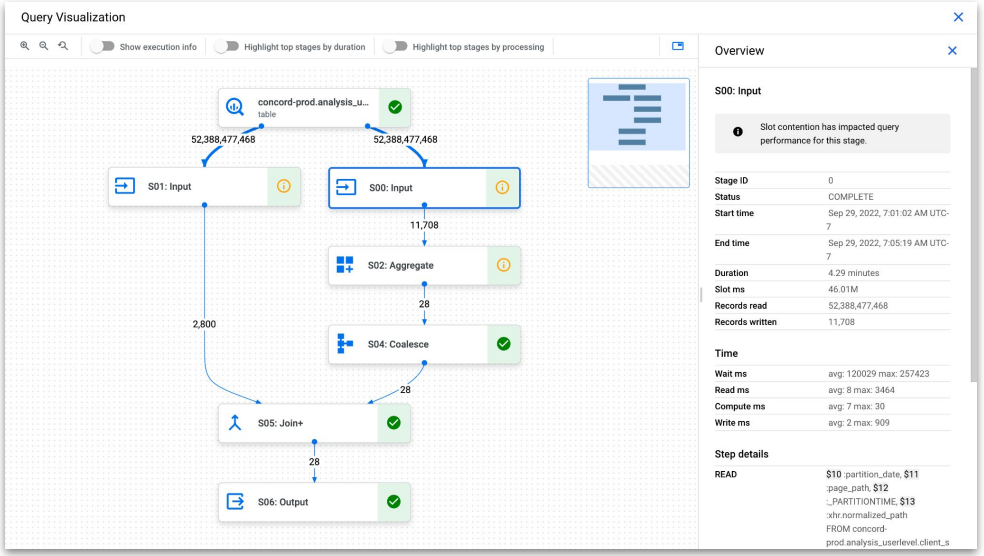


BigQuery - Query Execution Graph

GA 2022-11-03

/ 特徴

- クエリ性能診断に必要な各種情報を可視化
  - クエリの実行計画
  - 各ステージの詳細の表示（スロット利用時間、読み取り件数、書き込み件数）
  - 処理量が多いステージのハイライト
  - 実行時間が長いステージのハイライト
  - パフォーマンス分析情報



/ ユースケース

- 遅いクエリの分析、原因の解析

Preview 2022-12-01

BigQuery - BigLake for Iceberg tables

/ 特徴

- オープンソーステーブルフォーマットであるIceberg を BigLake でサポート
- Spark で作成した Iceberg table が自動的に BigQuery テーブルとして認識される
- Row, column level security, masking 等 BigLake のセキュリティ機能をIceberg テーブルに適用
- BigLake Metastore (Iceberg catalog) を参照することにより、object.list オペレーションが不要となり、Pruning により必要なファイルにのみアクセス。パフォーマンスメリットが得られる。

/ ユースケース

- OSS フォーマットをベースとしたLakehouseアーキテクチャを BigQuery / BigLake で実現
- Spark ETL プロセスで Iceberg テーブルを作成し、BigQuery からクエリ実行。

Preview 2022-12-12

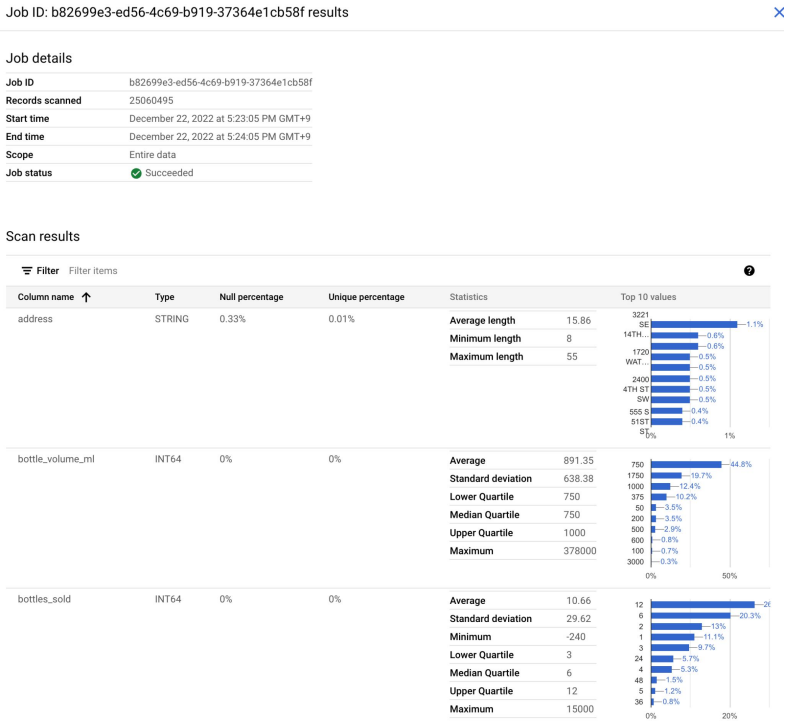
Dataplex - Data Profiling

/ 特徴

- Dataplex の Asset として登録されている BigQuery テーブルに対して、自動プロファイリングにより以下の統計が取得できる
  - Null 率
  - ユニーク率
  - 統計情報 (Type により異なる)
  - Top 10 values
- Preview 期間は無料

/ Use case

- スキーマ情報が不明なデータの検査
- データ品質チェックの値確認



Preview 2022-12-12

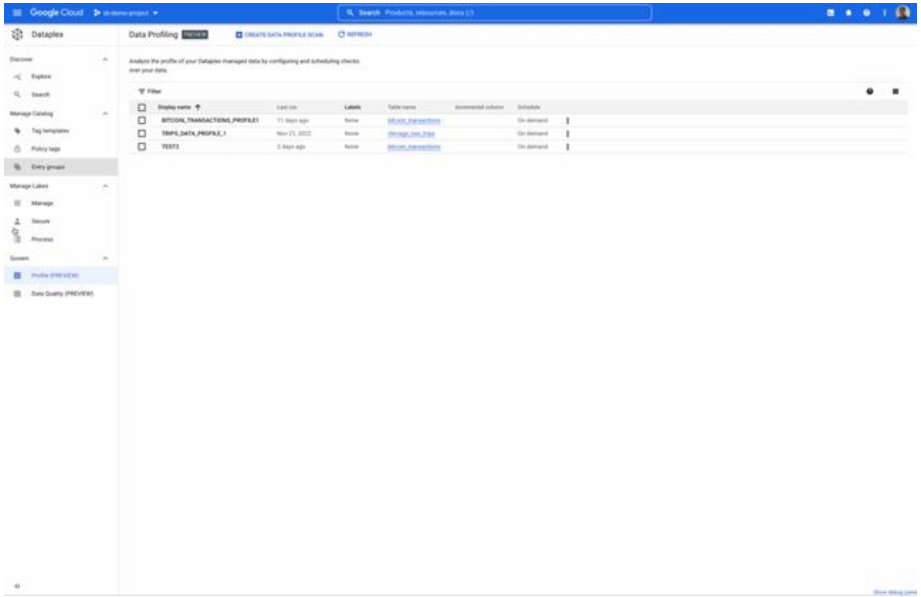
Dataplex - Auto Data Quality (AutoDQ)

/ 特徴

- テンプレート化されたルールをもとに、データ品質チェックをクイックに行う
- Preview 期間は無料

/ Use case

- 自社のデータガバナンスが届かない相手先 (外部システム) からのデータ連携などで、データが要件を満たすか (or クレンジングされているか) チェック
- 通常の Dataplex DQ Task との使い分け
  - 簡易・クイックに => AutoDQ
  - しっかり・カスタム => 通常 DQ Task



Preview 2022-12-12

Dataplex - Auto Data Quality (AutoDQ)

事前定義ルール

- 範囲チェック(ex : min 1 - max 100)
- Not Null チェック
- 特定の値チェック(ex : USD, JPY, EUR のみ)
- 正規表現チェック(ex : 先頭 customer* で始まる)
- ユニーク性チェック
- Statistic Range Expectation (ex : 金額の平均がxxx より高い)

Rule type (Name in Google Cloud console)	Row-level or aggregate rule	Description	Supported column types	Rule-specific parameters
`RangeExpectation` (Range Check)	Row-level	Check if the value is between min and max.	All numeric, date, and timestamp type columns.	**Required:** • `Passing threshold percentage` • `mean`, `min`, or `max` values: Specify at least one value. **Optional:** • Enable `strict min`: If enabled, the rule check uses ">" instead of ">=". • Enable `strict max`: If enabled, the rule check uses "<" instead of "<=". • Enable `ignore null`: If enabled, null values are ignored in the rule check.
`NonNullExpectation` (Null check)	Row-level	Validate that column values are not NULL.	All supported column types.	**Required:** Passing threshold percentage.
`SetExpectation` (Set check)	Row-level	Check if the values in a column are one of the specified values in a set.	All supported column types, except `Record` and `Struct`.	**Required:** • Set of string values to check against. • Passing threshold percentage. **Optional:** • Enable `ignore null`: If enabled, null values are ignored in the rule check. • Enable `invert` condition: If enabled, rule checks if the values are not from the set.
`RegexExpectation` (Regex check)	Row-level	Check the values again a specified regular expression.	String	**Required:** • Regex pattern used to check. • Passing threshold percentage. • Note: Google Standard SQL provides regular expression support using the re2 library; See that documentation for its regular expression syntax.
ın			All supported column types, except `Record` and `Struct`.	**Required:** Column and dimension from the supported parameters.
n ure ge			All supported numeric column types.	**Required:** `mean`, `min`, or `max` values: Specify at least one value. **Optional:** • Enable `strict min`: If enabled, the rule check uses ">" instead of ">=". • Enable `strict max`: If enabled, the rule check uses "<" instead of "<=". • Enable `ignore null`: If enabled, null values are ignored in the rule check.

GA 2023-01-09

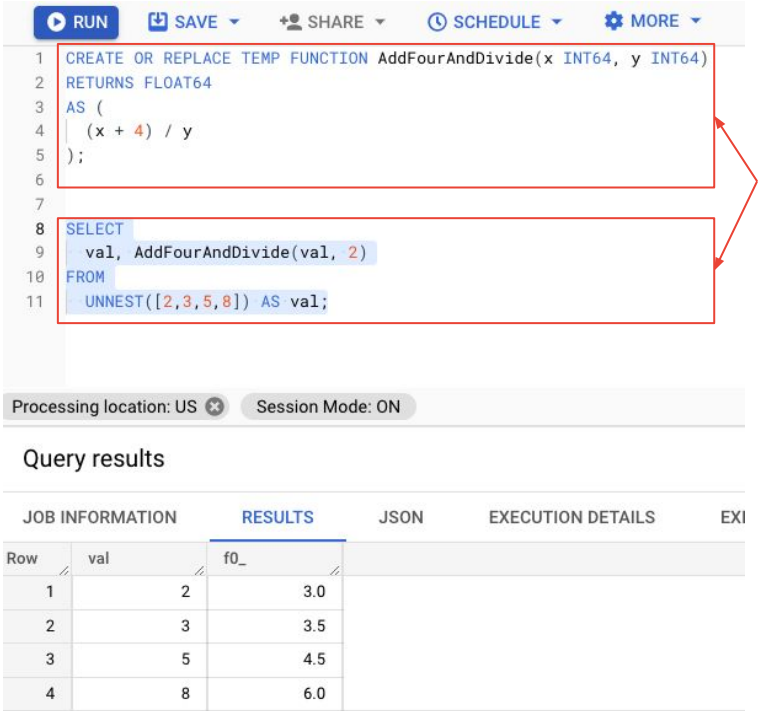
BigQuery - Session の機能追加

/ 特徴

- Temp Function が Session 内で利用可能に
  - Session 内のみで有効
  - Session 終了後、自動的に削除
- Temp キーワードを含む SQL に OR REPLACE, IF NOT EXISTS が利用可能に
  - Temp function
  - Temp table

/ Use case

- Temp function / Temp table を使った処理のデバッグ



GA 2023-01-10

BigQuery - Reservation の DDL の追加、改善

/ 特徴

- 購入済みのコミットメントの変更
  - プランの変更
  - リニューアルプランの変更
- 作成済みの予約の変更
  - アイドルスロットの利用の設定
  - スロット量
  - 同時実行数（Query Queue 利用時）
- CREATE 時にオプションの指定が可能に
  - コミットメント、予約、割当

/ Use case

- プランの変更を SQLで実行
- BigQuery Script など SQL のみで BigQuery Reservation 関連の操作を実行する
  - ピーク時に Flex スロットを購入しオフピークにはもとに戻す

```
/*  追加でFlex Slotを100購入 */
CREATE CAPACITY `admin_project.region-us.my-commitment`
OPTIONS (
 slot_count = 100,
 plan = 'FLEX');

/* 購入したスロットを既存の予約（100スロット）に追加して200にする
ALTER RESERVATION `admin_project.region-us.my-reservation`
SET OPTIONS (
 slot_capacity = 200
);
```

Preview 2023-01-30

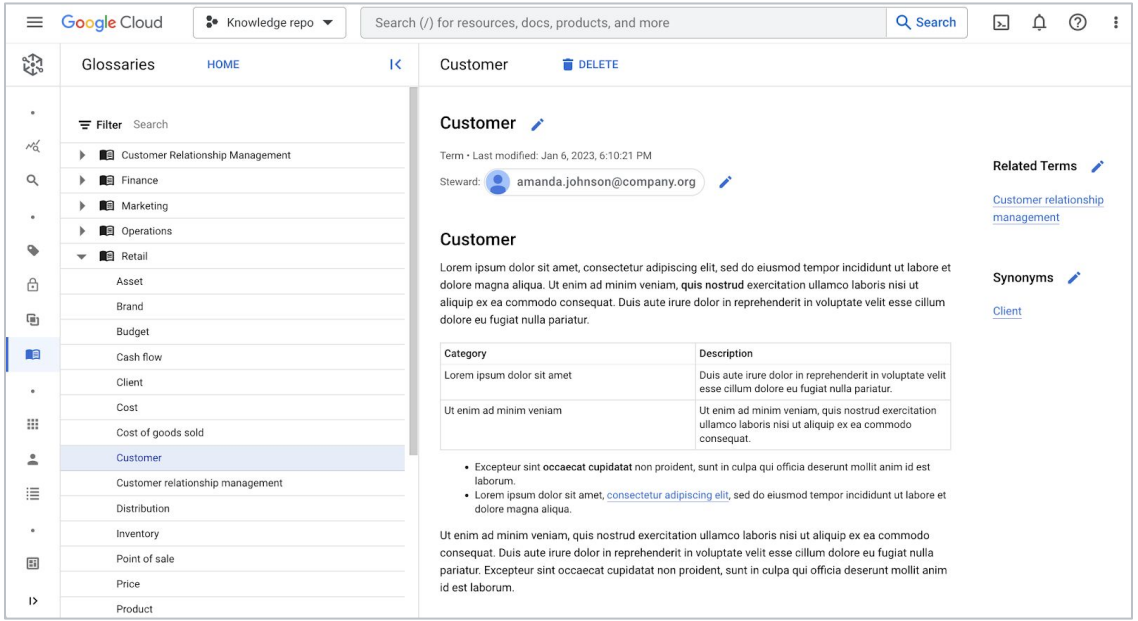
Dataplex - Business Glossary

/ 特徴

- ビジネス用語集を作成・管理
- 用語を BigQuery のカラムに関連付けて参照することが可能

/ Use case

- そのビジネスドメインに不案内な分析ユーザのデータ理解
- 用語が特殊であったり同じ用語でも組織内でコンテキストの違いで異なる意味になる場合



Preview 2023-01-30

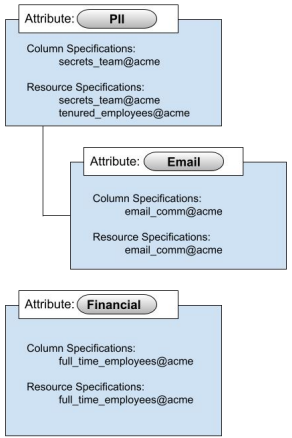
Dataplex - Attribute store

/ 特徴

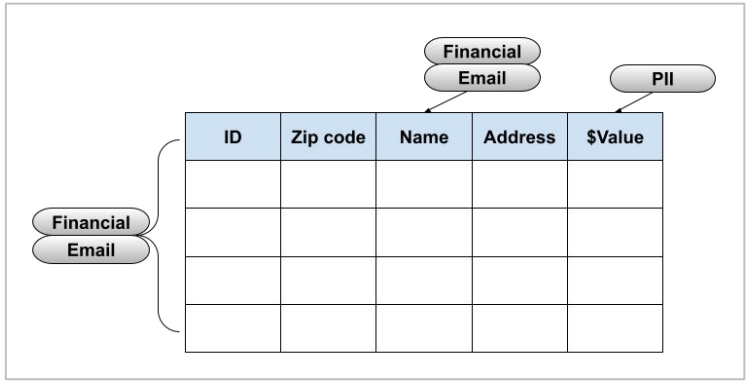
- Attribute を階層化してアクセスポリシー設定
- テーブルと列の ACL を管理
- Dataplex の管理アセットに対して実施可能

/ Use case

- Dataplex でリソースを管理しているユーザで BigQuery のポリシータグのように階層的・一元的にテーブル・列レベルでのアクセス制御を管理したい



- Employees in `secrets_team@acme`, `tenured_employees@acme`, `full_time_employees@acme`, and `email_comm@acme` can access the table. This is because Dataplex merges the resource specifications of the attributes `Financial` and `Email`, and the attribute `Email` inherits the specifications from the attribute `PII`.
- Employees in `secrets_team@acme`, `email_comm@acme`, `full_time_employees@acme` can access the column **Name**. This is because Dataplex merges the column specifications of the attributes `Financial` and `Email`.
- Only employees in `secrets_team@acme` can query the column **$Value**.



02

Data Analytics - ETL&Workflow

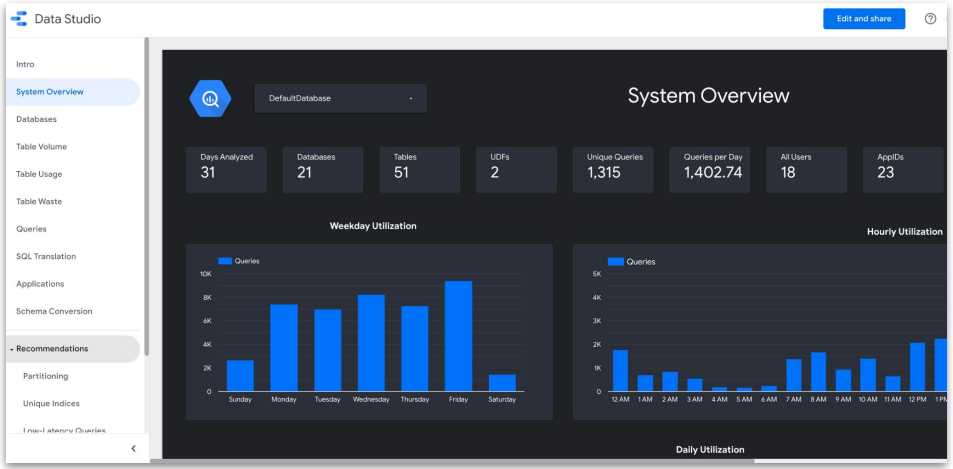
Preview 2022-11-01

BigQuery - Redshift Migration アセスメント

/ 特徴

Redshift に関する情報を分析し移行の複雑度、移行方法を評価

- 既存システム
  - データベース、スキーマ、テーブル等の情報
- BigQuery の安定状態
  - 自動変換可能なクエリ数、データ保存の費用、テーブルの最適化の提案
- 移行パス
  - 移行作業自体に関する情報



/ ユースケース

- Redshift 移行時の移行支援
- PoC ワークロードの抽出

GA 2022-11-03

BigQuery - ラップされた鍵を管理する SQL 関数

/ 特徴

- ラップされた鍵を管理するSQL 関数
  - KEYS.NEW_WRAPPED_KEYSET
  - KEYS.ROTATE_WRAPPED_KEYSET
  - KEYS.REWRAP_KEYSET

/ ユースケース

- ラップされた鍵で列レベルの暗号化
- AEAD 暗号化をよりセキュアに実施したい

```
SET kms_resource_name =
'gcp-kms://projects/my-project/locations/us/keyRings/my-key-ring/cryptoKeys/my-
crypto-key';
SET wrapped_keyset = (SELECT KEYS.NEW_WRAPPED_KEYSET(kms_resource_name,
'AEAD_AES_GCM_256'));

SELECT AEAD.ENCRYPT(
   KEYS.KEYSET_CHAIN(
     kms_resource_name,
     wrapped_keyset),
     'plaintext',
     '') AS encrypted_content
```

BigQuery - Load Data for S3 and Azure Blob

GA 2022-11-04

/ 特徴

- S3 、Azure Blob 上にあるファイルを直接 BigQuery へロードが可能に(要 BigQury Omni のスロット)
- US multi-region と US-EAST-4 region に対応

/ ユースケース

- S3 上にあるファイルをセキュアに BigQuery にロードして分析したい

```
LOAD DATA INTO `mydataset.testparquet`
FROM FILES (uris = ['s3://test-bucket/sample.parquet'],
format = 'PARQUET')
WITH CONNECTION `aws-us-east-1.test-connection`

;
```

Cross-cloud transfer from Amazon S3
Cross-cloud transfer from Azure Blob Storage

BigQuery - データの追加の改善

GA 2022-11-07

/ 特徴

- BigQuery にデータを追加する方法がよりわかりやすくなった。
- 一般的に利用されるローカルファイル、GCS、外部データソース
- Analytics Hub の Listing からの追加
- その他、様々なソースからの追加も可能に

Preview 2022-11-08

Data Fusion - Private Instance 使用時における DNS での名前解決

/ 特徴

- プライベートインスタンスでソースとシンクにホスト名、ドメイン名を指定して設定可能
- IP アドレス直指定ではないので、ソースやシンクのフェイルオーバー等によるパイプラインの変更が不要
- Data Fusion 6.7.0 以降で利用可能

/ ユースケース

- オンプレや Private network 内のデータソースやシンクにホスト名でアクセスしたい

Resolve domain names or hostnames

Preview 2022-11-11

Composer - Airflow トリガーと Deferrable operators

/ 特徴

- ワーカーのアイドル期間中にワーカースロットを解放して別のタスクで利用することが可能
- ワーカーを効率的に利用可能
- 簡単に利用可能
- Composer 2、且つ、Airflow 2.2.5、2.3.3以降 で利用可能

/ ユースケース

- 処理を外部サービス(BigQuery / Dataproc)で実行していて且つ実行時間が長い

```
insert_query_job = BigQueryInsertJobOperator(
    task_id="insert_query_job",
    configuration={
        "query": {
            "query": INSERT_ROWS_QUERY,
            "useLegacySql": False,
        }
    },
    location=LOCATION,
    deferrable=True,
)
```

Preview 2022-11-11

Pub/Sub - BigQuery Subscriptions JSON Type のサポート

/ 特徴

- 全ての String の項目、data と attributes に　適用可能
- JSON Data Type の恩恵を受けられる
  - 格納効率（データサイズ）
  - JSON 関数の高速化

/ ユースケース

- JSON データを展開せずにStreaming で BigQuery に入れる

Field name	Type
subscription_name	STRING
message_id	STRING
publish_time	TIMESTAMP
data	JSON
attributes	JSON

subscription_name	message_id	publish_time	data	attributes
projects/839210930825/subsc...	6335549661394078	2022-11-24 09:07:25.235000 U...	{"col1": "hello", "col2": "world"}	{"googclient_schemaname": "projects/tenishimdemo/schemas/sample","key1": "value1","key2": "value2","googclient_schemae
projects/839210930825/subsc...	6335583816702330	2022-11-24 09:17:55.651000 U...	{"col1": "abc", "col2": "def"}	{"googclient_schemaname": "projects/tenishimdemo/schemas/sample","googclient_schemaencoding": "JSON"}
projects/839210930825/subsc...	6335710543489158	2022-11-24 09:33:27.009000 U...	{"col1": "hello", "col2": "world"}	{"googclient_schemaencoding": "JSON","key1": "value1","googclient_schemaname": "projects/tenishimdemo/sche

Preview 2022-11-17

BigQuery - Metadata Caching

/ 特徴

- メタデータをキャッシュしてクエリのパフォーマンスを向上
- BigLake テーブルと Object テーブルに対して有効
- Cloud Storage のリスティングを防ぐので大量のファイルをもつ BigLake テーブルに有効
- 期限を指定可能（30 分～7 日間）
- リフレッシュは手動と自動が選択可能

/ ユースケース

- Hive Partitioned table など大量のファイルをもつテーブルを BigLake テーブルとして BigQuery から高速にアクセスしたい

```
CREATE EXTERNAL TABLE `my_dataset.object_table`
WITH CONNECTION `us.my-connection`
OPTIONS(
 object_metadata = 'SIMPLE',
 uris = ['gs://mybucket/*'],
 max_staleness = INTERVAL 1 DAY,
 metadata_cache_mode = 'AUTOMATIC'
);
```

Metadata Caching Performance

Preview 2022-12-05

Dataproc Metastore - Administrator Interface

/ 特徴

- (Dataproc Metastore をアタッチした) Dataproc や Hive のインスタンスがなくてもDataproc Metastore に保管されたメタデータの参照や管理ができるように
- Read-only operation でできること(roles/metastore.metadataQueryAdmin)
  - メタデータのクエリ
- Read and write operation (roles/metastore.metadataMutateAdmin)
  - データベース、テーブルやパーティションのロケーション変更
  - テーブルを別のデータベースへ移動

Dataproc Metastore administrator interface

GA 2022-12-05

Cloud Composer - Composer Local Development CLI tool

/ 特徴

- ローカルで実行されるApache Airflow 環境で DAG の開発ができるように

/ ユースケース

- DAG のテスト
- PyPI パッケージのテスト
- Airflow configuration option のテスト

GA 2022-12-06

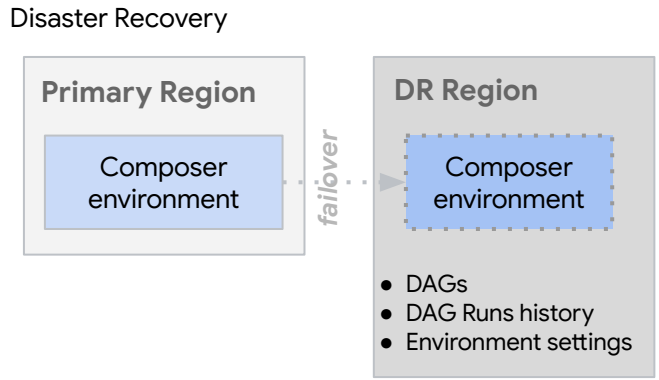
Cloud Composer - Environment snapshot / Scheduled snapshot

/ 特徴

- Composer 環境の設定と状態のバックアップを GCS バケットへ保管

/ ユースケース

- オペレーションエラーからの復旧
- 新しいバージョンで作成したComposer 環境にスナップショットを適用することで、side-by-side アップデートを実現
- DR 環境へのフェイルオーバー



Preview 2022-12-06

Cloud Data Fusion - Cloud Data Fusion 6.8.0

CDAP 6.8.0 リリース

/ 特徴

- Dataplex Batch Source と Dataplex Sink プラグインが GA
- Datastream を利用した Oracle から BigQuery への Replication が GA
  - GCS バケット経由の Replication の模様
- BigQuery batch source pushdown サポート
- AND trigger のサポート
- Replication job のバージョンアップグレードはできないため、利用している場合要注意

BREAKING Upgrading the Cloud Data Fusion version for Replication jobs is broken. Upgrading Replication jobs to Cloud Data Fusion version 6.8.0 isn't recommended.

https://cdap.atlassian.net/wiki/spaces/DOCS/pages/1845067785/CDAP+Release+6.8.0

Preview

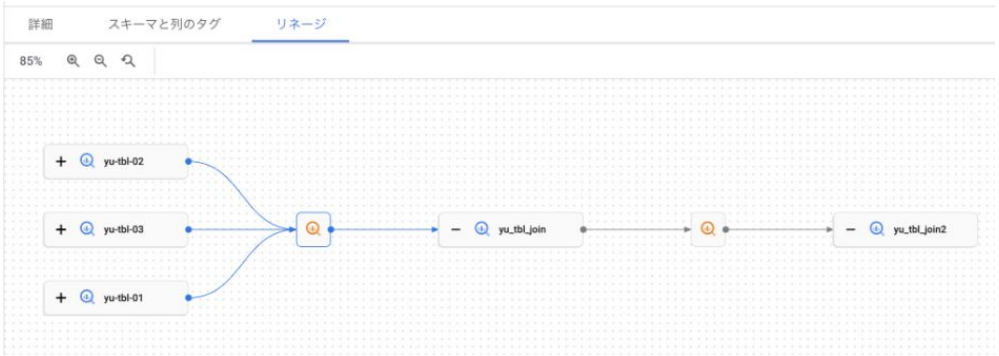
Dataplex - Data Lineage

/ 特徴

- Table の依存関係を可視化
- Datamart 作成の SQL も確認可能
- BQ の UI からも確認可能

/ Use case

- 生成元データの確認
- データガバナンス整備



Preview 2022-12-13

Cloud Composer - Data lineage

/ 特徴

- Composer 環境で Dataplex data lineage integration を有効にすることで、lineage 情報が Data Lineage API 経由で連携される
- 対応 version : Composer 2 (version 2.1.2+) (Airflow version 2.2.5+) のみ
- 対応 operators : BigQuery, GCS 関連 (links)
- Airflow タスク実行の終了時に Lineage 情報が連携される (1-2 sec) 既存タスクのパフォーマンスには影響しない

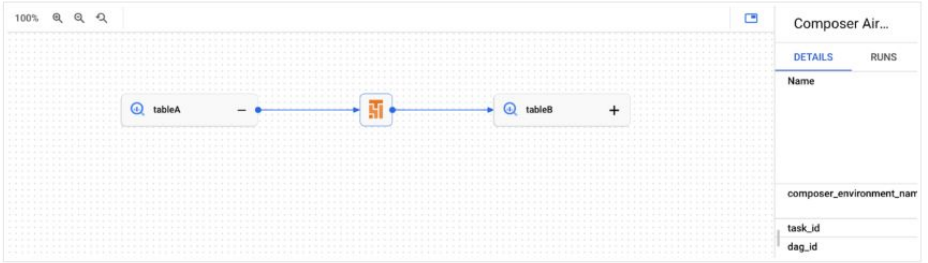
/ Use case

- Composer, Data Fusion, BigQuery, GCS, などを横断してパイプラインを構成しており、全体の Lineage 情報を統合管理したい

For example, running the following task:

```
task = BigQueryInsertJobOperator(
    task_id='snapshot_task',
    dag=dag,
    location='<dataset-location>',
    configuration={
        'query': {
            'query': 'SELECT * FROM dataset.tableA',
            'useLegacySql': False,
            'destinationTable': {
                'project_id': GCP_PROJECT,
                'dataset_id': 'dataset',
                'table_id': 'tableB',
            },
        }
    },
)
```

Results in creating the following lineage graph in Dataplex UI:



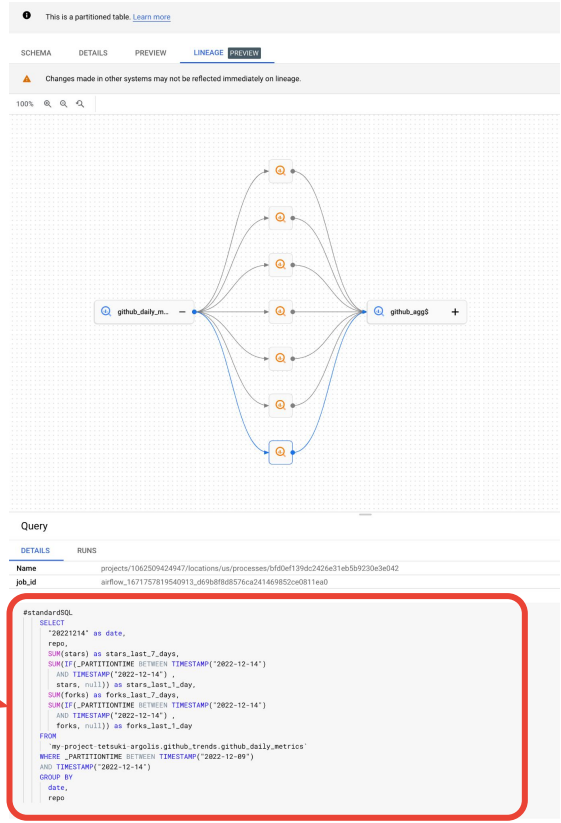
Preview 2022-12-13

Cloud Composer - Data lineage

/ Tips

- Operator が実行する SQL がそのまま表示される
  INSERT などの部分が省略されているため注意が必要

```
# Task 3
t3 = BigQueryOperator(
    task_id='t3_bq_write_to_github_agg',
    use_legacy_sql=False,
    write_disposition='WRITE_TRUNCATE',
    allow_large_results=True,
    sql='''
    #standardSQL
    SELECT
      "{{ yesterday_ds_nodash }}" as date,
      repo,
      SUM(stars) as stars_last_7_days,
      SUM(IF(_PARTITIONTIME BETWEEN TIMESTAMP("{{ yesterday_ds }}")
        AND TIMESTAMP("{{ yesterday_ds }}") ,
        stars, null)) as stars_last_1_day,
      SUM(forks) as forks_last_7_days,
      SUM(IF(_PARTITIONTIME BETWEEN TIMESTAMP("{{ yesterday_ds }}")
        AND TIMESTAMP("{{ yesterday_ds }}") ,
        forks, null)) as forks_last_1_day
    FROM
      `_PROJECT_ID_.github_trends.github_daily_metrics`
    WHERE _PARTITIONTIME BETWEEN TIMESTAMP("{{ macros.ds_add(ds, -6) }}")
    AND TIMESTAMP("{{ yesterday_ds }}")
    GROUP BY
      date,
      repo
    '''.replace('_PROJECT_ID_', PROJECT_ID),
    destination_dataset_table='_PROJECT_ID_.github_trends.github_agg${{ yesterday_ds_nodash }}'.replace(
        '_PROJECT_ID_', PROJECT_ID),
    dag=dag)
```



GA 2022-12-15

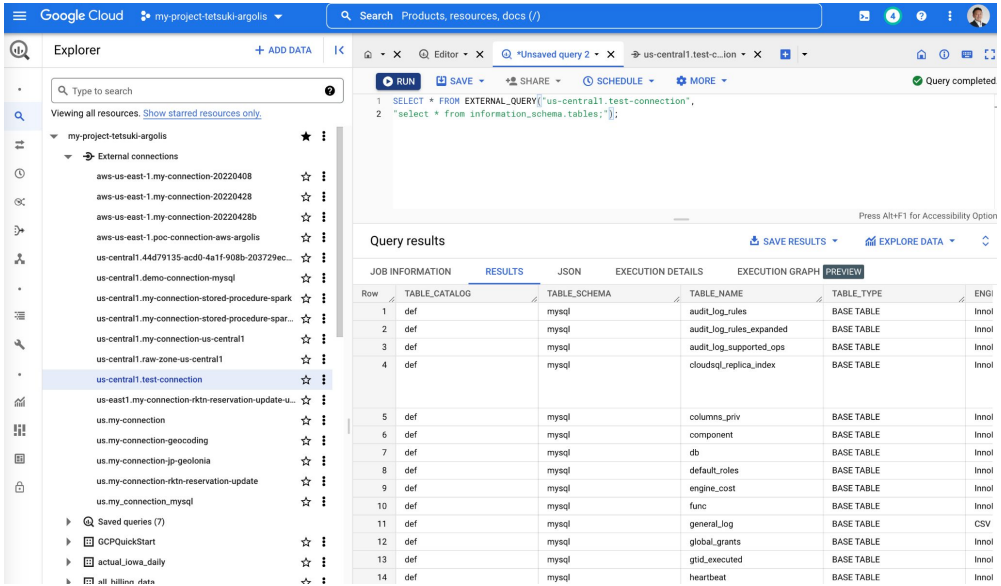
BigQuery - query Cloud SQL with Private connection

/ 特徴

- 従来は Public connection のみだった Cloud SQL に対する Federated Query が Private Connection にも対応

/ Use case

- Cloud SQL を 閉域網内のみで利用 (Private Connection モード) されているお客様で、BigQuery を用いてデータ分析を行いたい



03

AI, ML - BQML,  Vertex AI, その他

Preview 2022-10-05

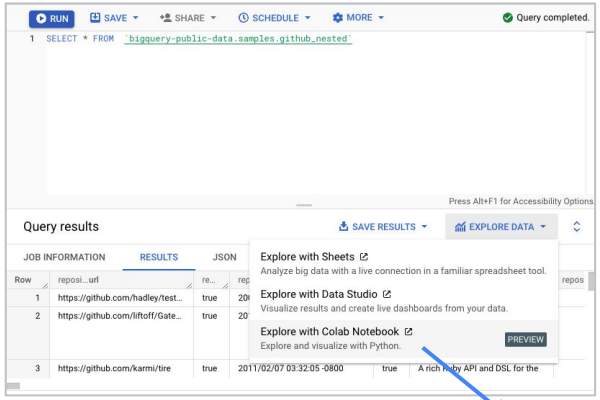
Explore query results in Colab

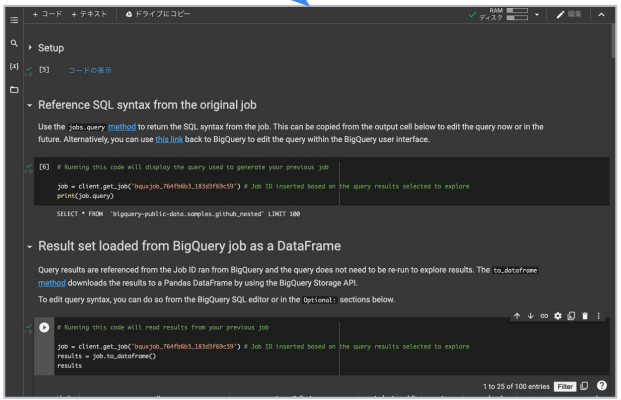
/ 特徴

- BigQuery の SQL 実行結果を Colab Notebooks で分析できる

/ ユースケース

- BigQuery にデータが有り、Colab でデータサイエンスのタスクを実行している





[Explore data in Colab]

Preview 2022-10-11

Stored procedure for Apache Spark

/ 特徴

- BigQuery から SQL で Apache Spark の pyspark のコードがストアドプロシージャとして実行可能
- 実行環境には Dataproc を指定できる

/ ユースケース

- PySpark で実装されたワークロードがあり、BigQuery のワークロードと統合して容易に運用

/ 注意点

- Scala は未対応

[Apache Spark のストアド プロシージャを操作する]

[Connect to Apache Spark]

PySpark コードを直接書いた場合のストアドの作成

```
CREATE OR REPLACE PROCEDURE my_bq_project.my_dataset.spark_proc()
WITH CONNECTION `my-project-id.us.my-connection`
OPTIONS(engine="SPARK")
LANGUAGE PYTHON AS R"""
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("spark-bigquery-demo").getOrCreate()

# Load data from BigQuery.
words = spark.read.format("bigquery") \
  .option("table", "bigquery-public-data:samples.shakespeare") \
  .load()
words.createOrReplaceTempView("words")

# Perform word count.
word_count = words.select('word', 'word_count').groupBy('word').sum('word_count')
word_count.show()
word_count.printSchema()

# Saving the data to BigQuery
word_count.write.format("bigquery") \
  .option("writeMethod", "direct") \
  .save("wordcount_dataset.wordcount_output")
"""
```

GCS にある PySpark コードを参照するパターン

```
CREATE PROCEDURE my_bq_project.my_dataset.spark_proc()
WITH CONNECTION `my-project-id.us.my-connection`
OPTIONS(engine="SPARK", main_file_uri="gs://my-bucket/my-pyspark-main.py")
LANGUAGE PYTHON
```

GA 2022-10-11

Analytics Hub

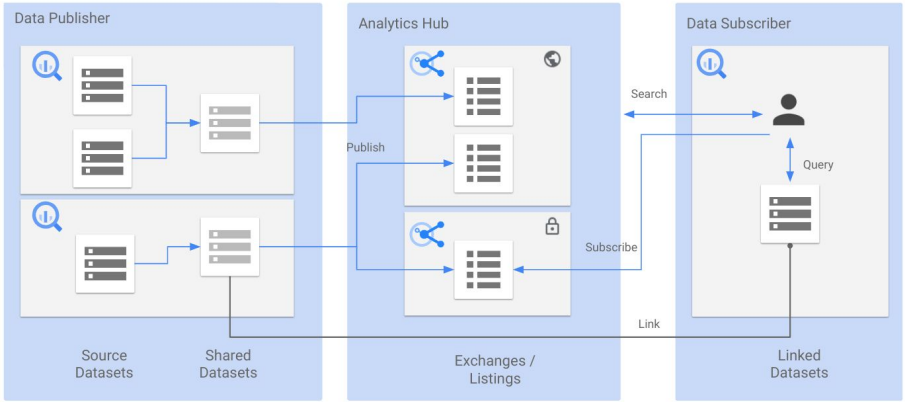
/ 特徴

- 組織の境界を越えてデータと分析情報を共有するデータ交換プラットフォーム
- セルフサービスで、データを複製せず、セキュアに共有可能
- Google の Public Dataset とかけあわせて分析もできる
- Dataplex(Data Catalog) 統合で検索可能に
- 公開されたデータのサブスクリプションの一覧と削除が可能に

/ ユースケース

- 自社のデータを他社に提供し収益化
- 組織内のデータ共有のセルフサービス化

Analytics Hub



Analytics Hub の概要

Preview 2022-10-24

BigQuery - BI Engine の新しいダッシュボード メトリックス

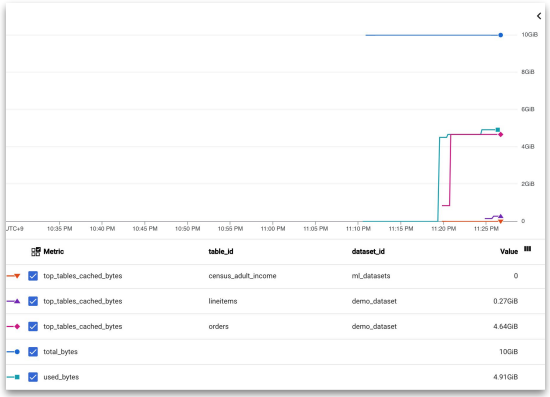
/ 特徴

- BI Engine Top Tables Cached Bytes
  - 上位のテーブルのキャッシュ上のバイト数
- BI Engine Fallback Count
  - BI Engine を利用しなかったクエリの割合(クエリ数/秒)
  - 原因ごとに表示も可能

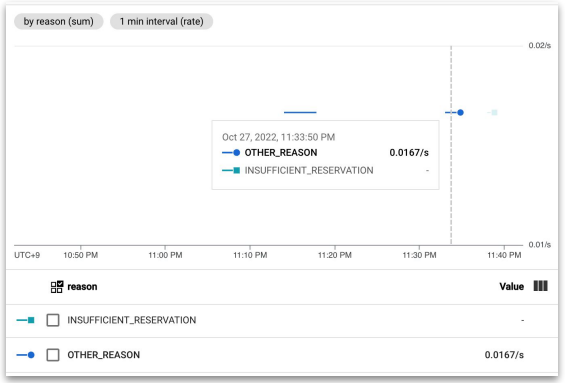
/ ユースケース

- BI Engine が有効に利用されているかの確認
- 目的のテーブルがキャッシュされているか
- BI Engine の効果を測る

BI Engine Top Tables Cached Bytes



BI Engine Fallback Count



BigQuery - Search Index

GA 2022-10-27

/ 特徴

- テキストインデックスによりテーブル内から高速に特定の文字列を検索
- 列名が不明でもテーブル全体から高速に検索
- STRING と Native JSON データに対応
- 専用の SEARCH 関数を提供
- 索引のメンテナンスは BigQuery が自動的に実施
- スロットの利用量とクエリコストを削減
- **Tokyo リージョンでも使用可能に**

/ ユースケース

- ログ分析
- セキュリティ監査
- 選択性の高い文字列検索

/ 制限

- あいまい検索やタイポ等のドキュメント系の検索には未対応

Search Indexの作成

```
CREATE SEARCH INDEX my_index ON Logs(ALL COLUMNS);
```

Search Indexを利用した検索

```
SELECT * FROM Logs WHERE SEARCH(Logs, 'bar');
```

Introduction to Search In BigQuery

BigQuery - リモート関数

GA 2022-10-18

/ 特徴

- SQL から Cloud Functions / Cloud Run を実行可能
- BQ のユーザ定義関数で実装できない処理が実装可能（外部APIの呼び出し）
- 複数の言語で実装可能（Node.js, Python, Go, Java, .Net, Ruby, PHP）
- VPC-SC のサポート

/ ユースケース

- SQLで
  - Cloud DLP で匿名化、再識別化
  - 画像、動画、自然言語 API の利用
- ELT でできる範囲が広くなりアーキテクチャのシンプル化

Cloud Functionsで関数の定義

```
import json
_MAX_LOSSLESS=9007199254740992

def remote_add(request):
 try:
   return_value = []
   request_json = request.get_json()
   calls = request_json['calls']
   for call in calls:
     return_value.append(sum([int(x) if isinstance(x, str) else x for x in call if x
is not None]))
   replies = [str(x) if x > _MAX_LOSSLESS or x < -_MAX_LOSSLESS else x for x in
return_value]
   return_json = json.dumps( { "replies" :  replies} )
   return return_json
 except Exception as inst:
   return json.dumps( { "errorMessage": 'something unexpected in input' } ), 400
```

BigQueryでリモート関数の作成

```
CREATE OR REPLACE FUNCTION demo_dataset.remote_add(x INT64, y INT64)
RETURNS INT64
REMOTE WITH CONNECTION `tetsu-project01.us.remote_add_connection`
OPTIONS (endpoint =
'https://us-central1-tetsu-project01.cloudfunctions.net/remote_add');
```

SQLから実行

```
SELECT val, demo_dataset.remote_add(val, 2)
FROM UNNEST([NULL,2,3,5,8]) AS val;
```

リモート関数の操作

Preview 2022-11-17

BigQuery - Object Tables

/ 特徴

- BigQuery で Cloud Storage 上の非構造化データを SQL で分析可能に
- ファイルのメタデータ情報の保持
- Data という疑似列も参照可能でBQML と一緒に利用可能
- Signed Object URL を生成し、Remote functionへ渡すことも可能
- BigQuery のセキュリティも適用可能

/ ユースケース

- Cloud Storage 上の画像、動画等のファイルをSQLを使って機械学習にかけて分析する
- 画像、動画等のファイルをBigQuery 上のその他のデータとかけあわせて分析する

Field name	Type	Mode
uri	STRING	NULLABLE
generation	INTEGER	NULLABLE
content_type	STRING	NULLABLE
size	INTEGER	NULLABLE
md5_hash	STRING	NULLABLE
updated	TIMESTAMP	NULLABLE
metadata	RECORD	REPEATED
name	STRING	NULLABLE
value	STRING	NULLABLE

```
CREATE TABLE my_dataset.my_inference_results AS
SELECT uri, content_type, vision_feature
FROM ML.PREDICT(
 MODEL my_dataset.vision_model,
 SELECT ML.DECODE_IMAGE(data) AS vision_input
 FROM my_dataset.object_table
);
```

Introduction to object tables

GA 2023-01-03

BigQuery ML - Vertex AI Model Registry との統合

/ 特徴

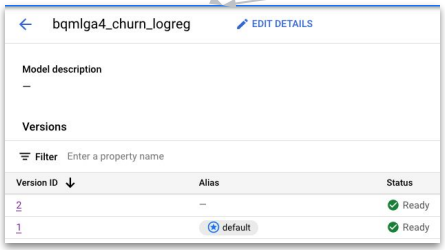
- Vertex AI Model Registry に BigQuery ML モデルを登録して管理、バージョニング、モニタリングできる
- 登録したモデルをVertex AI Endpoints　へデプロイ
- 評価指標の比較、追跡ができる

/ Use case

- BigQuery で作成したモデルを使い Vertex AIでオンライン予測をする
- BigQuery ML を含めた ML Ops
- Dataform (preview) でデータとモデルをビルドしてステージングする

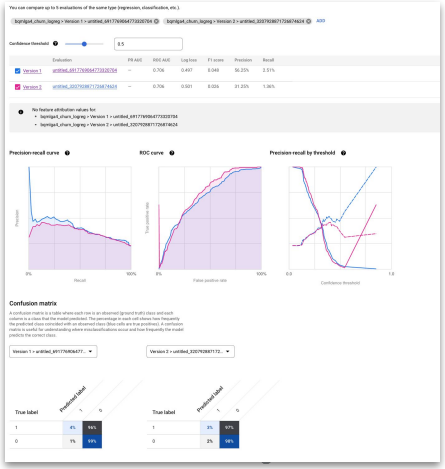
```
CREATE MODEL bqmlga4.churn_logreg
OPTIONS(
 MODEL_REGISTRY="vertex_ai",
 vertex_ai_model_id="bqmlga4_churn_logreg",
 MODEL_TYPE="LOGISTIC_REG",
 INPUT_LABEL_COLS=["churned"]
) AS
SELECT （省略）
```

```
CREATE MODEL bqmlga4.churn_logreg2
OPTIONS(
 MODEL_REGISTRY="vertex_ai",
 vertex_ai_model_id="bqmlga4_churn_logreg",
 MODEL_TYPE="LOGISTIC_REG",
 INPUT_LABEL_COLS=["churned"]
) AS
SELECT （省略）
```



Model Registry に登録

評価指標の比較



Preview 2023-01-04

BigQuery ML - Vision model の画像分析のサポート

/ 特徴

- Vision model を import として BigQuery ML でオブジェクトの検出やOCR 等が可能
- BigQuery の Object テーブルでGCS 上のファイルにアクセス
- Tensorflow モデル
  - ImageNet
  - ResNet 50

/ Use case

- BigQuery ML で画像の推論や文字の認識をする

```
CREATE MODEL my_dataset.my_vision_model
OPTIONS(
 model_type = 'TENSORFLOW',
 model_path =
'gs://my_bucket/my_model_folder/*');
```

```
SELECT * FROM
ML.PREDICT(
 MODEL my_dataset.vision_model,
 (SELECT uri, ML.DECODE_IMAGE(data) AS input
 FROM my_dataset.object_table)
);
```

Preview 利用申し込み form

Preview 2023-01-17

BigQuery ML - Sparse Input のサポート

/ 特徴

- 値がほとんどゼロまたは空であるデータに対するトレーニングにおける空間効率が改善される（1 hot-encodingなどによって発生する多数の 0 の無駄を省ける）
- Array [STRUCT < int, numerical>] で Sparse なカテゴリ変数を効率的に保存

/ Use case

- 多数のカテゴリ変数をもつデータでモデルをトレーニングする

	bigquery	features	hello	now	sparse	supports	world
Hello World	0	0	1	0	0	0	1
BigQuery supports sparse features now	1	1	0	1	1	1	0

ARRAY<STRUCT<k INT64, v INT64>> で表すと
Hello World は、[(2, 1), (6, 1)]
BigQuery supports …. は、[[0,1],[1,1],[3,1],[4,1],[5,1]]

Support sparse input for feature columns
Sparse Features Support in BigQuery

Preview 2023-01-17

BigQuery ML - ARIMA_PLUS_XREG モデル

/ 特徴

- 多変量時系列予測が可能に
- 数値、カテゴリ、配列の自動特徴量エンジニアリング
- 季節の傾向や休日の検出などARIM_PLUSモデルの機能も使用できる
- Reservation のみで利用可能

/ Use case

- 多変量解析による時系列予測

```
CREATE OR REPLACE MODEL
   `bqml_tutorial.seattle_pm25_xreg_model`
 OPTIONS (
   MODEL_TYPE = 'ARIMA_PLUS_XREG',
   time_series_timestamp_col = 'date',
   time_series_data_col = 'pm25')
AS
SELECT
 date,
 pm25,
 temperature,
 wind_speed
FROM
 `bqml_tutorial.seattle_air_quality_daily`
WHERE
 date
 BETWEEN DATE('2012-01-01')
 AND DATE('2020-12-31')
```

シアトルの大気質データからの多変量時系列予測
The CREATE MODEL statement for the ARIMA_PLUS_XREG model

GA 2023-01-19

BigQuery ML - Vertex AI Pipelines での BigQuery ML 用のコンポーネント

/ 特徴

- 20 以上のコンポーネントが利用可能
- パイプラインのメタデータのトラッキングが可能
- Vertex AI のオンライン予測などシームレスな連携が可能に

/ Use case

- Vertex AI Pipeline で BigQuery ML を含めたML のパイプラインを作成する

主要なコンポーネント

- BigqueryQueryJobOp
- BigqueryCreateModelJobOp
- BigqueryExportModelJobOp
- BigqueryPredictModelJobOp
- BigqueryEvaluateModelJobOp
- BigqueryDropModelJobOp
- BigqueryEvaluateModelJobOp
- BigqueryExplainForecastModelJobOp
- BigqueryExplainPredictModelJobOp
- BigqueryForecastModelJobOp

google_cloud_pipeline_components.v1.bigquery module

Appendix

その他のアップデート

その他

Cloud Composer

October 06, 2022

ANNOUNCEMENT Starting from January 2023, the default version for new Cloud Composer environments changes from Cloud Composer 1 to Cloud Composer 2. New environments will use the latest Cloud Composer 2 version with the default Airflow 2 version (`composer-2-airflow-2`). Currently, the default version is `composer-1-airflow-1.10.15` .

Dataproc

October 03, 2022

ANNOUNCEMENT Preemptible SPOT VMs can be used as secondary workers in a Dataproc cluster. Unlike legacy preemptible VMs with a 24-hour maximum lifetime, Spot VMs have no maximum lifetime.

その他

BigQuery

2022/11/14

FEATURE The slot estimator helps you manage slot capacity based on historical performance metrics. This feature is now generally available (GA).

Dataproc

2022/11/14

FEATURE If a Dataproc Metastore service uses the gRPC endpoint protocol, a Dataproc or self-managed cluster located in any region can attach to the service.

2022/11/17

FEATURE Dataproc Serverless for Spark supports Spark and System metrics. These metrics are enabled by default. Spark driver and executor metrics can be customised using overrides.

その他

Pub/Sub

2022/12/1

その他

BigQuery

2022/12/8

FEATURE The demo query guide helps you query a public dataset from Google Trends and is now in preview.

Data Fusion

2022/11/30

DEPRECATED Cloud Data Fusion version 6.4 is no longer supported as of December 1, 2022. You should upgrade your instances to run in a supported version. For instructions, see Upgrading your Cloud Data Fusion environment.

2022/12/01

FEATURE Google Cloud Platform Plugins version 0.20.4 is generally available (GA) in Cloud Data Fusion versions 6.7.1 and 6.7.2. This version includes Dataplex Source and Sink plugins in GA. For more information, see the CDAP Hub release log.

FEATURE Google Cloud Platform Plugins version 0.19.3 is generally available (GA) in Cloud Data Fusion version 6.6.0. This version includes Dataplex Source and Sink plugins in GA. For more information, see the CDAP Hub release log.

その他

BigQuery

December 22, 2022

FEATURE The Lineage tab in the table properties page lets you track how your data moves and transforms through BigQuery. This feature is now in preview.

CHANGED BigQuery now blocks saving query results to Google Drive from projects inside a VPC Service Controls protected perimeter.

PubSub Lite

December 19, 2022

FEATURE Pub/Sub Lite now supports export subscriptions. You can use an export subscription to export Pub/Sub Lite messages to a destination Pub/Sub topic. This feature is generally available (GA).

その他

BigQuery

FEATURE The following geography functions are now generally available (GA):

- `ST_ISCLOSED` : Returns `TRUE` for a non-empty geography, where each element in the geography has an empty boundary.
- `ST_ISRING` : Checks if a geography is a linestring and if the linestring is both closed and simple.

Dataproc

FEATURE Dataproc Serverless for Spark now supports `spark.dataproc.diagnostics.enabled` property that enables auto diagnostics on Batch failure. Note that enabling auto diagnostics will hold compute and storage quota after Batch is complete and until diagnostics is finished.

その他

BigQuery ML Jan 03

CHANGED Customers can use BigQuery ML to train and run models on BigLake in Cloud Storage. See Data Cloud Blog and End to end unstructured data use cases demo.

Dataflow Jan 03

FEATURE Starting in version 2023-01-03-RC00, the Google-provided Dataflow templates support ES6 syntax for JavaScript user-defined functions (UDFs). This change is backwards-compatible. ES5 syntax and existing user-defined functions are still supported.

When you run Google-provided templates using the latest version, your jobs are upgraded automatically on restart. If you want to keep running an earlier version of a template, when you run the template, specify version `2022-12-15-00_RC00` or earlier.

その他

Cloud Data Fusion Jan05

FEATURE

The SAP SuccessFactors Batch Source plugin is available in Preview. You can connect your data pipeline to an SAP SuccessFactors Source and a BigQuery Sink with this plugin in Cloud Data Fusion versions 6.5.1 and later.

その他

Cloud Composer

January 25, 2023

FEATURE **Airflow 2.4.3** is available in Cloud Composer images.

その他

February 01, 2023

FEATURE The BigQuery Data Transfer Service can now transfer data from Azure Blob Storage into BigQuery. This feature is now in preview.

January 31, 2023

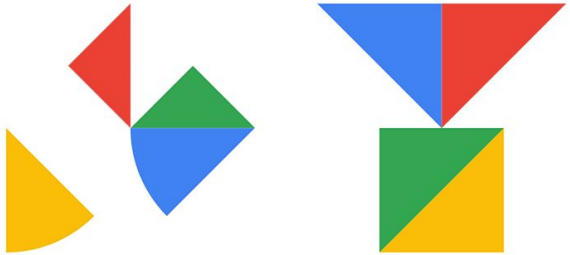
FEATURE Azure workload identity federation is now generally available (GA) for BigQuery Omni connections. You can now create a connection for federated identity using Google Cloud console.

CHANGED **Cloud console updates**: When you create datasets, select locations to run specific queries, or create exchanges in Analytics Hub, you now see separate options for multi-region and specific regions. Based on your selection, you see a list with more options.

January 30, 2023

FEATURE You can search for BigQuery partners in the BigQuery Partner Center. This feature is in Preview.

Google Cloud Day ’23 Tour



東京	（オンライン）	5 月 23 日（火）～ 25 日（木）
大阪	（ハイブリッド）	6 月 2 日（金）
名古屋	（ハイブリッド）	6 月 22 日（木）
福岡	（ハイブリッド）	6 月 30 日（金）

企業の DX を加速する、そのヒントを 4 都市からお届けします。

今すぐ登録　goo.gle/gcd23_1p