

10 分で分かる

Google Cloud データ分析関連サービス

.....

Google Cloud

カスタマーエンジニア

星 美鈴

Google Cloud

# データ活用の流れ

1

データを収  
集する

2

データを  
処理する

3

データを  
蓄積する

4

データを  
分析する

5

データを活  
用する

# Google Cloud から提供されるデータ分析関連プロダクト

## 収集

スケーラブルな  
データ収集



Cloud Pub/Sub



Data Transfer  
Service



Cloud IoT  
Core



Storage Transfer  
Service

## 処理

信頼性のある  
ストリームデータ  
パイプライン



Cloud  
Dataflow



Cloud  
Dataproc



Cloud Dataprep

## 蓄積

データレイク  
データ ウェアハウス



Cloud  
Storage



BigQuery  
storage

## 分析

データ ウェアハウス



BigQuery  
analysis engine

## 活用

先進的な分析



Cloud AI  
Services



Google  
Data Portal



Looker



Connected  
Sheets

## 全体



Cloud Composer



Cloud Data Fusion



Data Catalog

# Google Cloud から提供されるデータ分析関連プロダクト

## 収集

スケーラブルな  
データ収集



Cloud Pub/Sub



Data Transfer  
Service



Cloud IoT  
Core



Storage Transfer  
Service

## 処理

信頼性のある  
ストリームデータ  
パイプライン



Cloud  
Dataflow



Cloud  
Dataproc



Cloud Dataprep

## 蓄積

データレイク  
データ ウェアハウス



Cloud  
Storage



BigQuery  
storage

## 分析

データ ウェアハウス



BigQuery  
analysis engine

## 活用

先進的な分析



Cloud AI  
Services



Google  
Data Portal



Looker



Connected  
Sheets

## 全体



Cloud Composer

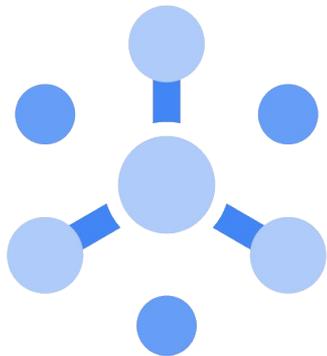


Cloud Data Fusion



Data Catalog

# Cloud Pub/Sub



システム間の疎結合化/長時間接続を提供する  
マネージドメッセージングサービス

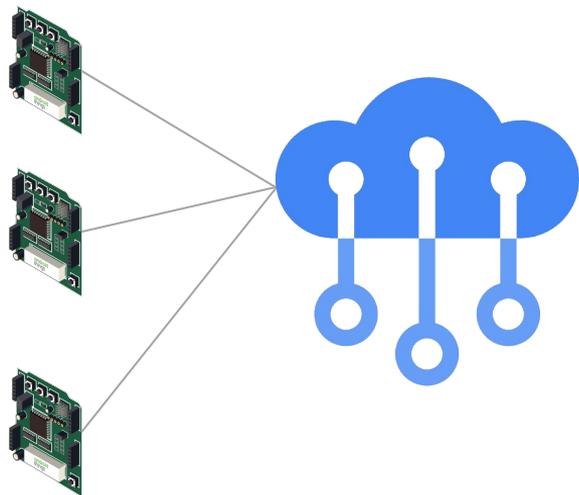


0~数百 GB/sに自動スケール



多様な配信パターンに対応  
(一対一 / 一対多 / 多対多、Push / Pull)

# Cloud IoT Core



多数のデバイス管理が可能な  
マネージドサービス



セキュアに双方向の通信が可能

# BigQuery Data Transfer Service

## サポートされるデータソース

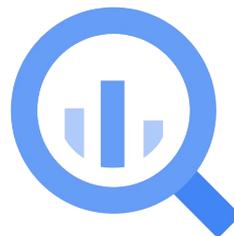
- Amazon S3
- キャンペーンマネージャ
- Cloud Storage
- Google 広告
- Googleアドマネージャー
- Google Merchant Center
- Google Play
- 検索広告 360
- YouTube チャンネル
- YouTube コンテンツ所有者
- Teradata
- Amazon Redshift
- その他サードパーティ

## BigQuery Data Transfer Service

簡単にセット

セキュア転送

定期、自動更新

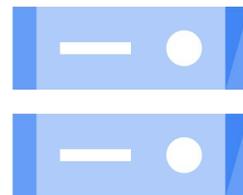


BigQuery  
storage

# Storage Transfer Service

サポートされるデータソース

- Amazon S3
- Azure Blob Storage
- バケットの異なるCloud Storage
- ウェブアドレスで参照できるデータ
- オンプレミスデータ



Cloud Storage

# Google Cloud から提供されるデータ分析関連プロダクト

## 収集

スケーラブルな  
データ収集



Cloud Pub/Sub



Data Transfer  
Service



Cloud IoT  
Core



Storage Transfer  
Service

## 処理

信頼性のある  
ストリームデータ  
パイプライン



Cloud  
Dataflow



Cloud  
Dataproc



Cloud Dataprep

## 蓄積

データレイク  
データ ウェアハウス



Cloud  
Storage



BigQuery  
storage

## 分析

データ ウェアハウス



BigQuery  
analysis engine

## 活用

先進的な分析



Cloud AI  
Services



Google  
Data Portal



Looker



Connected  
Sheets

## 全体



Cloud Composer

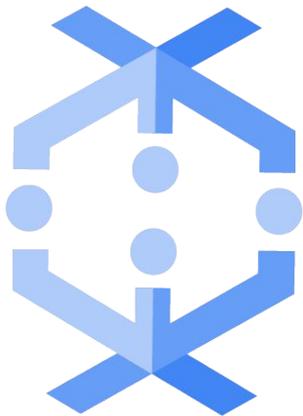


Cloud Data Fusion



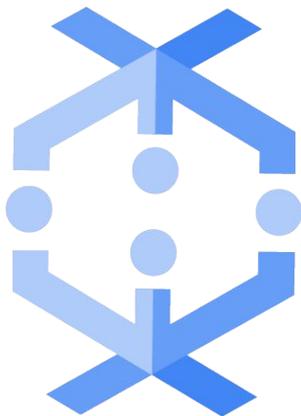
Data Catalog

# Cloud Dataflow



- フルマネージドのデータ処理サービス
- オープン (Apache Beam)
- 分散処理、オートスケール
- ストリームおよびバッチの両方に対応

# Cloud Dataflow



Google Cloud

A screenshot of the Google Cloud Platform interface showing a Dataflow job details page. The page is titled "Job details" and displays a pipeline graph with several stages. The stages are: PubsuIO.Read (Running, 1 hr 34 min 13 sec), Apply 1 Min Window (302 elements/s, 3 min 46 sec), AvroIO.Write (302 elements/s, 10 hr 45 min 25 sec), BigQueryIO.Write (302 elements/s, 44 days 17 hr 25 min 11 sec), Get Sale Lines (302 elements/s, 5 min 51 sec), Apply 10 Mins Window (302 elements/s, 3 min 44 sec), Count Store Sales (302 elements/s, 3 hr 44 min 12 sec), and Create Data Point (Running, 5 min 28 sec). The right sidebar shows a "Job summary" section with two line charts: "System latency (seconds) last 6h" and "Data freshness (seconds) last 6h". Below the charts, there is a table of job details including Job name, Job ID, Region, Job status, SDK version, Job type, Start time, Elapsed time, and Encryption type.

Google Cloud Platform

Job details

Job

Job summary

System latency (seconds) last 6h

Oct 6, 2019 11:24 AM

5 min interval (mean)

Data freshness (seconds) last 6h

Oct 6, 2019 11:24 AM

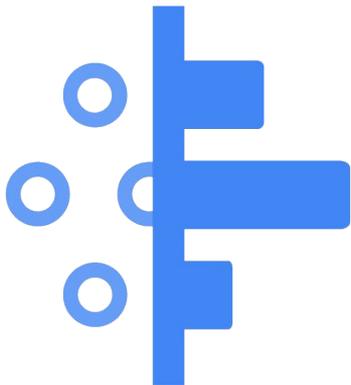
5 min interval (mean)

System latency: Create alerting

Data freshness: Create alerting

|                 |                                    |
|-----------------|------------------------------------|
| Job name        | process-orders-pipeline-23         |
| Job ID          | 2019-10-01_00_36_49-10999945955308 |
| Region          | us-central1                        |
| Job status      | Running                            |
| SDK version     | Apache Beam SDK for Java 2.11.0    |
| Job type        | Streaming                          |
| Start time      | Oct 1, 2019, 4:36:50 PM            |
| Elapsed time    | 4 days 21 hr                       |
| Encryption type | Google-managed key                 |

# Cloud Dataprep



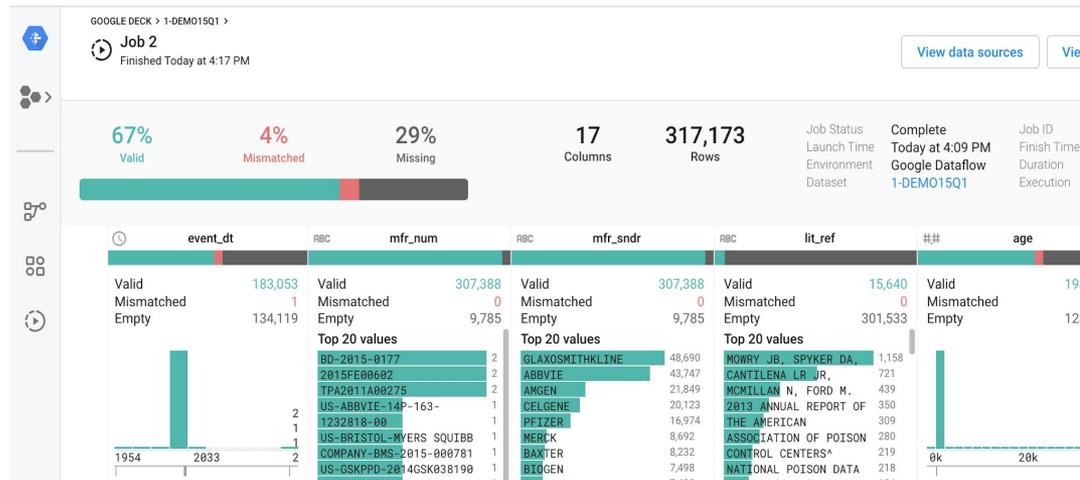
Google Cloud



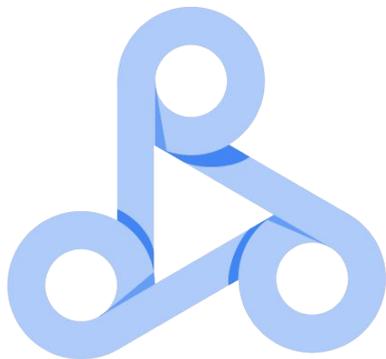
フルマネージドのデータ処理サービス



GUI でデータを整形、加工可能



# Cloud Dataproc



- Apache Hadoop / Spark のマネージドサービス
- ワンクリックまたはワンコマンドにより約 90 秒で Hadoop クラスタを構築
- プリエンプティブル VM の利用や不要時の無効化によりコストを最適化

# Google Cloud から提供されるデータ分析関連プロダクト

## 収集

スケーラブルな  
データ収集



Cloud Pub/Sub



Data Transfer  
Service



Cloud IoT  
Core



Storage Transfer  
Service

## 処理

信頼性のある  
ストリームデータ  
パイプライン



Cloud  
Dataflow



Cloud  
Dataproc



Cloud Dataprep

## 蓄積

データレイク  
データ ウェアハウス



Cloud  
Storage



BigQuery  
storage

## 分析

データ ウェアハウス



BigQuery  
analysis engine

## 活用

先進的な分析



Cloud AI  
Services



Google  
Data Portal



Looker



Connected  
Sheets

## 全体



Cloud Composer

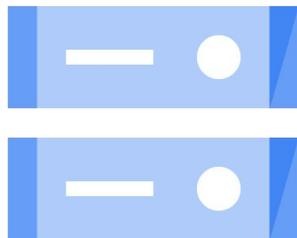


Cloud Data Fusion



Data Catalog

# Cloud Storage



容量無制限のオブジェクトストレージ



高い耐久性(年間 99.999999999 %)

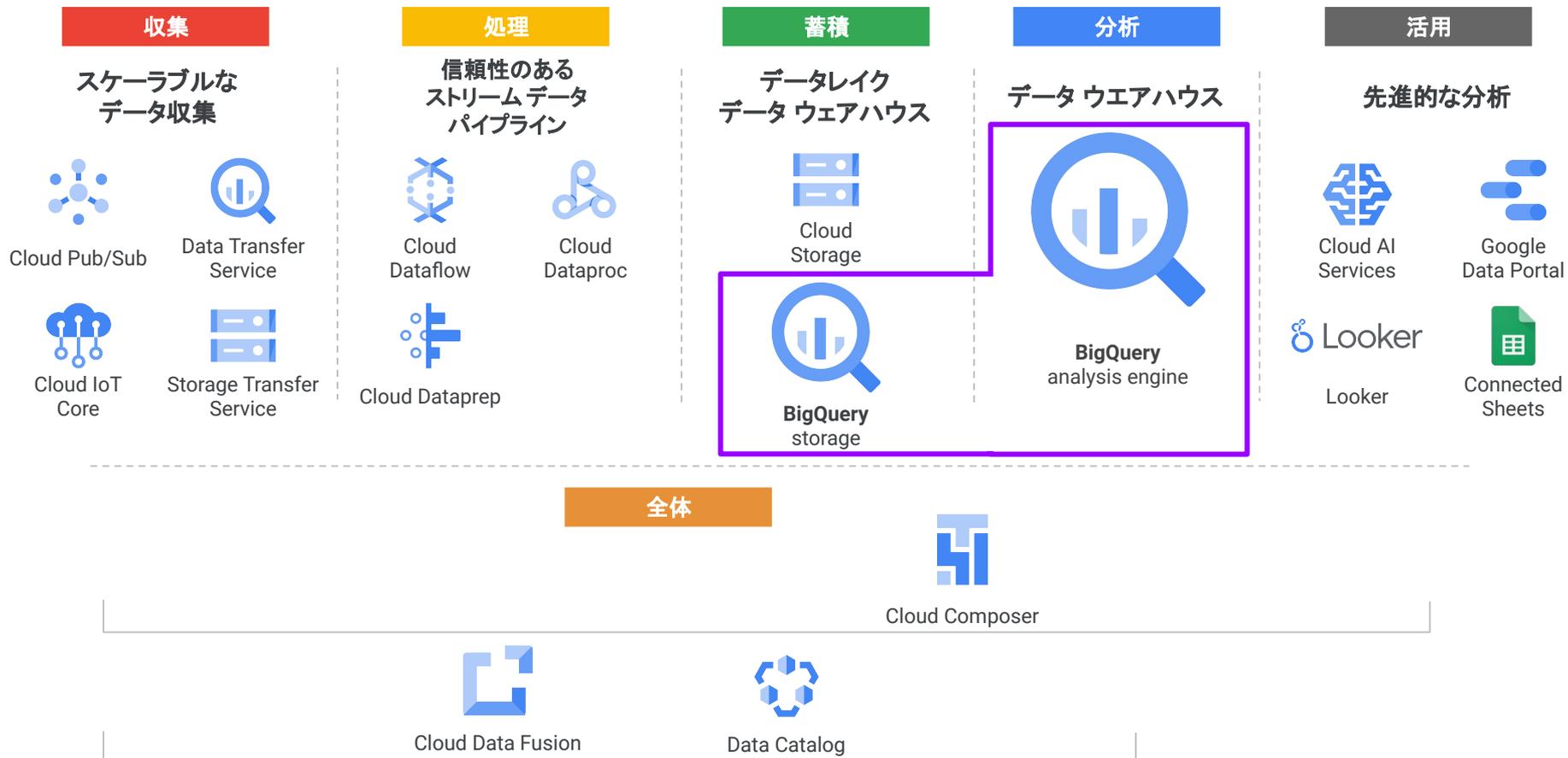


アクセス頻度に応じたストレージクラス  
(Standard / Nearline / Coldline / Archive)



指定条件に応じ低料金クラスへ容易に移行

# Google Cloud から提供されるデータ分析関連プロダクト



# Google BigQuery

エンタープライズ向け  
データウェアハウス

無制限のストレージと  
ペタバイト級クエリ

強固なセキュリティと  
高可用性



## 特徴

フルマネージド、サーバーレス

リアルタイム分析

AI の基盤 / BigQuery ML

BI の基盤 / BI Engine

# Google Cloud から提供されるデータ分析関連プロダクト

## 収集

スケーラブルな  
データ収集



Cloud Pub/Sub



Data Transfer  
Service



Cloud IoT  
Core



Storage Transfer  
Service

## 処理

信頼性のある  
ストリームデータ  
パイプライン



Cloud  
Dataflow



Cloud  
Dataproc



Cloud Dataprep

## 蓄積

データレイク  
データ ウェアハウス



Cloud  
Storage



BigQuery  
storage

## 分析

データ ウェアハウス



BigQuery  
analysis engine

## 活用

先進的な分析



Cloud AI  
Services



Google  
Data Portal



Looker



Connected  
Sheets

## 全体



Cloud Composer

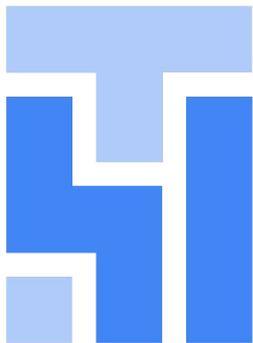


Cloud Data Fusion



Data Catalog

# Cloud Composer



マネージドのデータ系ジョブ管理サービス



オープン (Apache Airflow)



ハイブリッド、マルチクラウドに跨がるパイプラインを作成可能

# Cloud Composer



Google Cloud

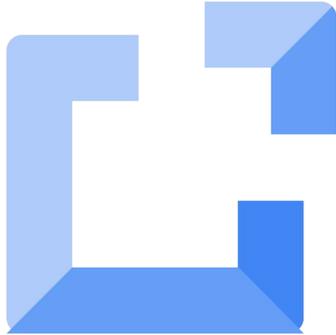
The screenshot shows the Airflow web interface for a DAG named 'bigquery\_github\_trends\_v5'. The interface includes a navigation bar with 'Airflow' and various menu items like 'DAGs', 'Data Profiling', 'Browse', 'Admin', 'Docs', and 'About'. Below the navigation bar, there are view options: 'Graph View', 'Tree View' (selected), 'Task Duration', 'Task Tries', 'Landing Times', 'Gantt', 'Details', and 'Go'. A 'Base date' field is set to '2018-01-10 21:00:00' and the 'Number of runs' is set to '25'. A legend indicates that green circles represent 'success' and red circles represent 'running'. The DAG graph shows a sequence of tasks: '[DAG]' -> 'bq\_check\_hackernews\_github\_agg' -> 'bq\_write\_to\_hackernews\_github\_agg' -> 'bq\_write\_to\_github\_agg' -> 'bq\_write\_to\_github\_daily\_metrics' -> 'bq\_check\_githubarchive\_day' -> 'bq\_write\_to\_hackernews\_agg' -> 'bq\_check\_hackernews\_full'. To the right of the graph is a Gantt chart showing the execution timeline of these tasks from December 24 to January 07, with green bars indicating successful runs and red bars indicating running or failed runs.

# Cloud Data Fusion



- マネージドのデータ統合管理サービス
- オープン(CDAP)
- GUI 操作でデータ処理、メタデータ管理、ハイブリッド/マルチクラウドに跨がるパイプラインの管理が可能

# Cloud Data Fusion



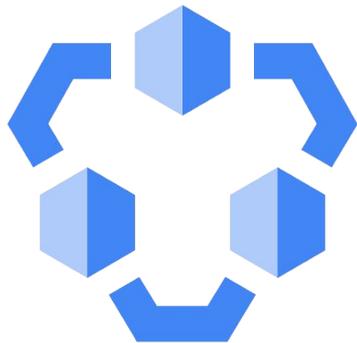
Google Cloud

The screenshot displays the Cloud Data Fusion interface for a pipeline named "complex\_data\_pipeline". The top navigation bar includes "Cloud Data Fusion | Pipeline", "DASHBOARD", "HUB", and "SYSTEM AD". Below the navigation bar, the pipeline name "complex\_data\_pipeline" is shown with a "cdap-data-pipeline" tag. Action buttons for "Configure", "Schedule", "Stop", "Run", and "Summary" are visible. A summary bar indicates "Run 1 of 1" with a "Succeeded" status, a start time of "09-17-2019 11:33:43 PM", a duration of "8 mins 23 secs", 7 warnings, and 0 errors. The main area shows a flow diagram with the following components:

- Inbound triggers (0)**: A vertical bar on the left.
- Database - m\_atm** (2.2.1): A green box with "Out 99 / Errors 0".
- Google Cloud Storage** (0.12.5): A green box with "Out 1 / Errors 0".
- Wrangler** (4.0.1): A blue box with "Out 99 / Errors 0" and "Alert Error".
- ParseJSON** (2.2.1): A blue box with "Out 1 / Errors 0".
- Joiner** (2.2.1): A blue box with "Out 1 / Errors 0".
- Save** (0.12.5): A blue box with "In 1 / Errors 0".

The flow diagram shows data from the Database and Google Cloud Storage sources feeding into Wrangler and ParseJSON respectively. The outputs of Wrangler and ParseJSON feed into the Joiner, which then feeds into the Save component.

# Data Catalog



- マネージドのメタデータ管理サービス
- 検索技術によるデータ資産検出の効率化
- タグ付けなどによるデータ管理の簡易化

# Google Cloud から提供されるデータ分析関連プロダクト

## 収集

スケーラブルな  
データ収集



Cloud Pub/Sub



Data Transfer  
Service



Cloud IoT  
Core



Storage Transfer  
Service

## 処理

信頼性のある  
ストリームデータ  
パイプライン



Cloud  
Dataflow



Cloud  
Dataproc



Cloud Dataprep

## 蓄積

データレイク  
データ ウェアハウス



Cloud  
Storage



BigQuery  
storage

## 分析

データ ウェアハウス



BigQuery  
analysis engine

## 活用

先進的な分析



Cloud AI  
Services



Google  
Data Portal



Looker

Looker



Connected  
Sheets

## 全体



Cloud Composer

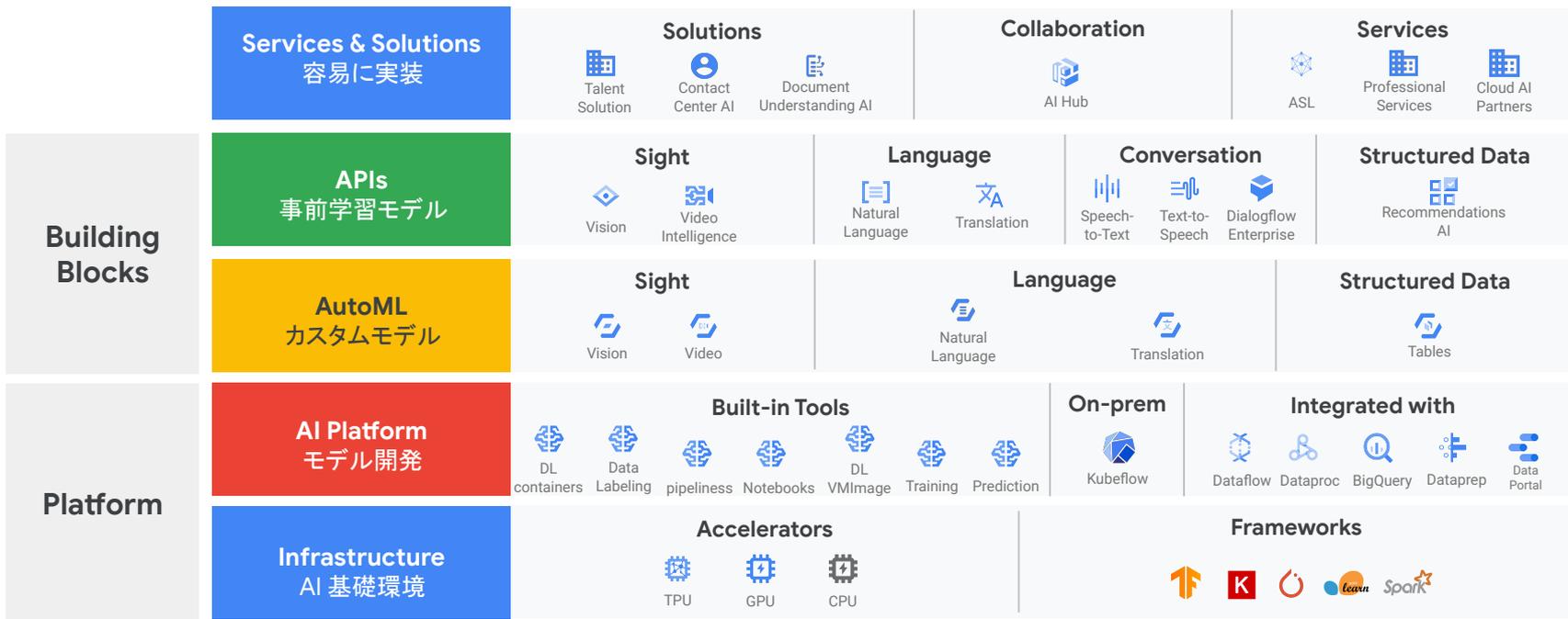


Cloud Data Fusion

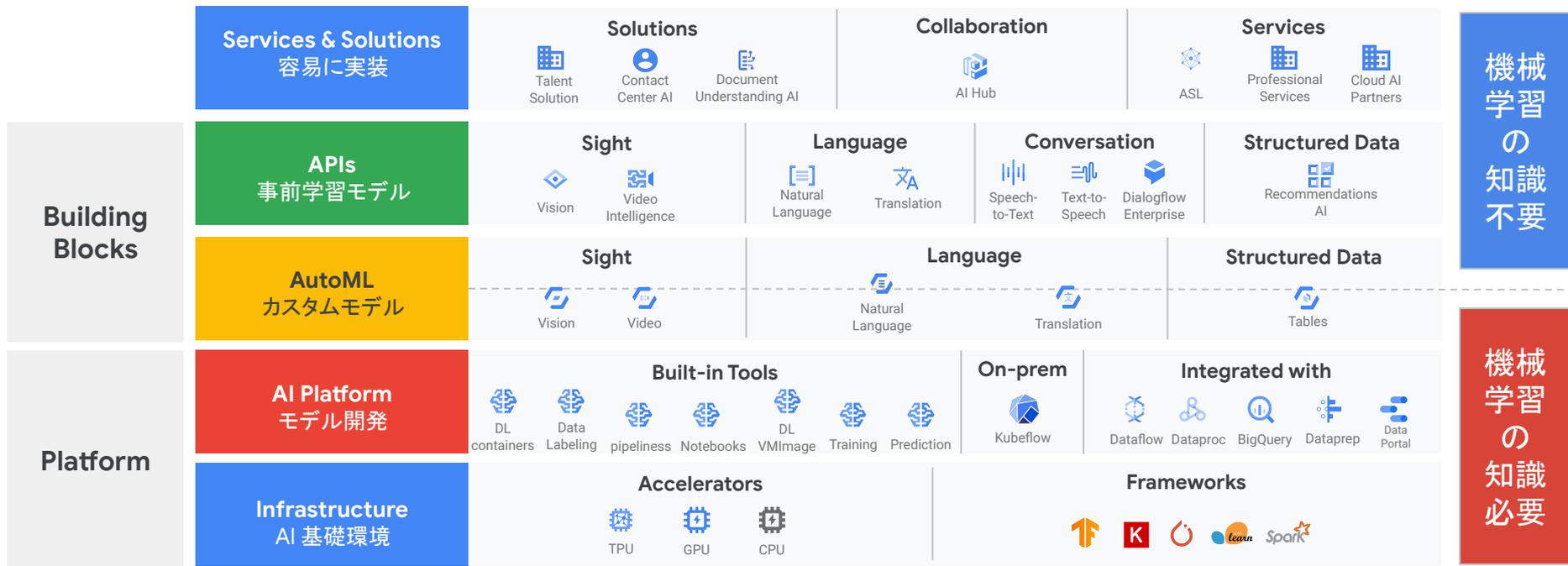


Data Catalog

# Cloud AI サービス群



# Cloud AI サービス群



# Data Portal



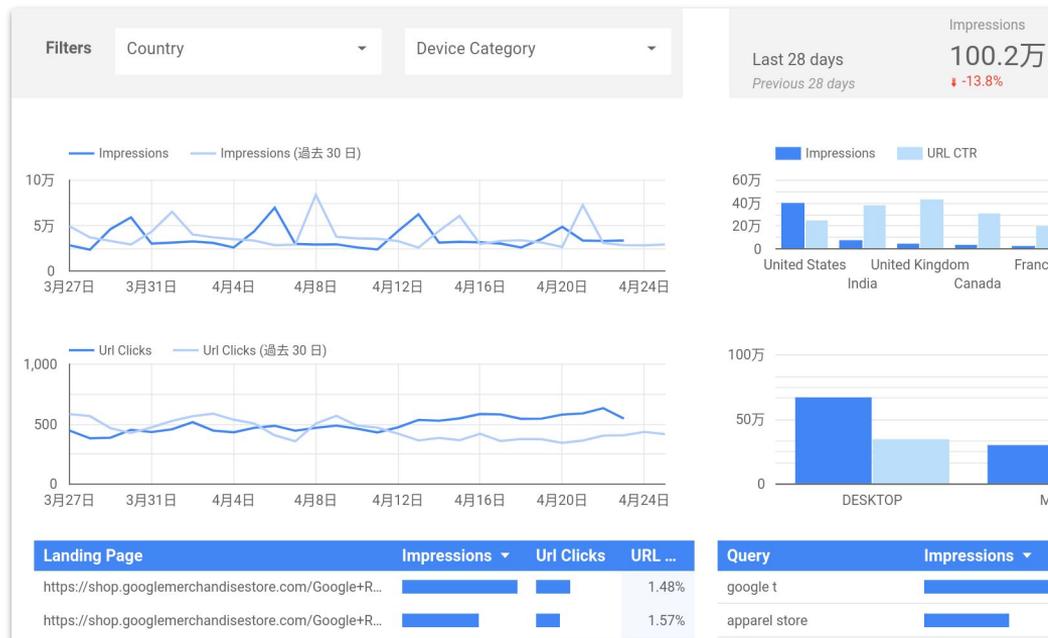
無料のデータ可視化ツール



アドホックな可視化や  
レポートに最適



作成レポートを他ユーザ、  
グループと共有可能



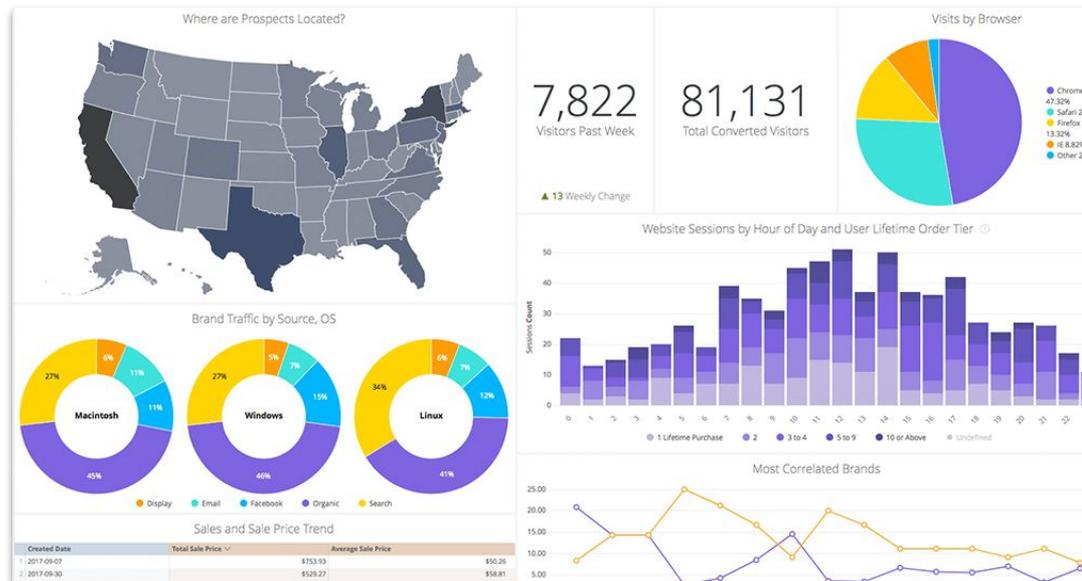
# Looker



エンタープライズ向け  
BI 分析プラットフォーム

セキュアで、企業全体の  
ガバナンスを維持可能

ビジネスツールと連携し  
インサイトを迅速に行動へ



Google Cloud

# Connected Sheets



BigQuery の何十億行のデータを Sheets で分析

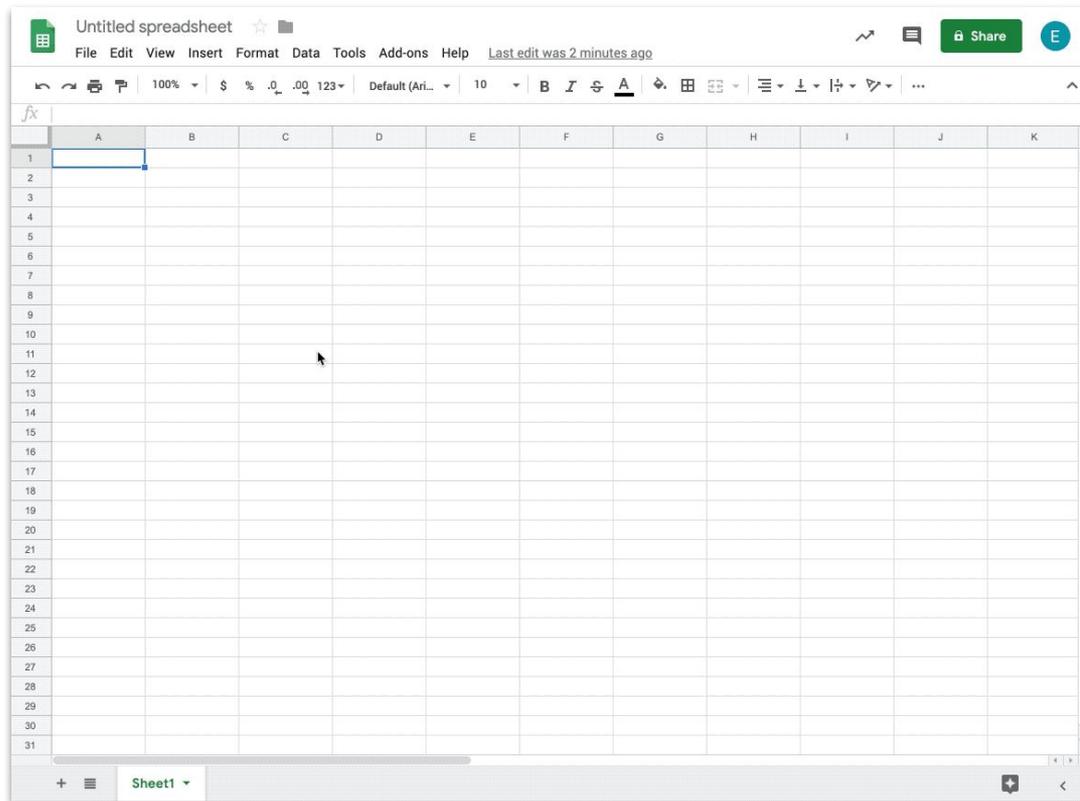


データは常に最新を維持



慣れ親しんだ  
スプレッドシートで  
データ探索が可能

Google Cloud



※ Google Workspace Enterprise の契約が必要

# Google Cloud データ分析関連プロダクトをひと言で表現

## 収集

Cloud Pub/Sub  
Cloud IoT Core  
Data Transfer Service  
Storage Transfer Service

リアルタイムにデータを処理するマネージドメッセージキュー  
IoT デバイスの管理やデータ取り込みのためのマネージドサービス  
BigQuery にデータを取り込むサービス  
Google Cloud Storage にデータを取り込むサービス

## 処理

Cloud Dataflow  
Cloud Dataprep  
Cloud Dataproc

ストリーム / バッチでの ETL 処理が可能なマネージド Apache Beam  
GUI で ETL 処理が可能なマネージドサービス  
マネージド Spark / Hadoop

## 蓄積

Cloud Storage

オブジェクトストレージ

## 分析

BigQuery

ビッグデータ用分析システム、超並列分散データウェアハウス

## 全体

Cloud Composer  
Cloud Data Fusion  
Data Catalog

ワークフローオーケストレーションのためのマネージド Apache Airflow  
GUI でデータパイプラインを管理するマネージド CDAP  
メタデータ管理のためのマネージドサービス

## 活用

Cloud AI  
Google Data Portal  
Looker  
Connected Sheets

ユーザの専門知識レベルに幅広く対応した AI サービス群  
データを簡単に可視化するダッシュボード  
エンタープライズ向け BI プラットフォーム  
Google スプレッドシートから BigQuery へアクセスし分析する機能

データ基盤全体をカバーする  
書籍が 2/20 に出版されました！

是非皆様のデータ活用  
にお役立てください！

Google Cloud

業務で使えるデータ基盤構築

# Google Cloud

ではじめる

実践

# データ エンジニアリング 入門

下田倫大、寛野雄太、  
饗庭秀一郎、吉田啓二 [著]



データ分析・活用・管理のためのデータ基盤の要件と  
Google Cloud の各種サービスをこの1冊で体系的に学ぶ

データウェアハウス ● データレイク ● ETL・ELT 処理  
データパイプラインマネジメント ● データ統合 ● セキュリティ ● コスト管理  
BI ● データの可視化 ● 地理情報分析 ● 機械学習 ● リアルタイム分析

技術評論社

**Thank you**