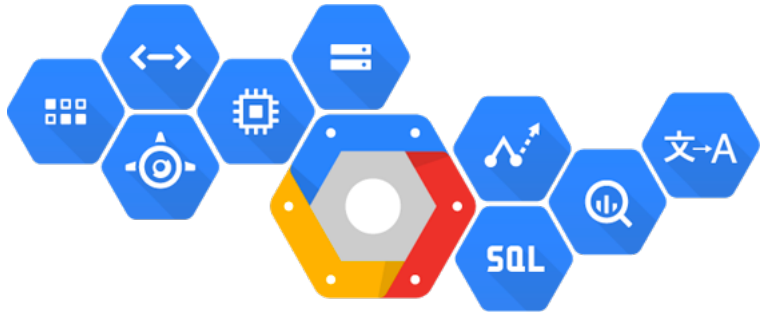
Architecture: Big Data / Data Analytics services



Data Analytics – Core Services

- Pub/Sub
- Dataflow
- Dataproc
- BigQuery

Data Management & ETL

- Cloud Composer
- Cloud Data Fusion
- Dataprep
- Data Catalog

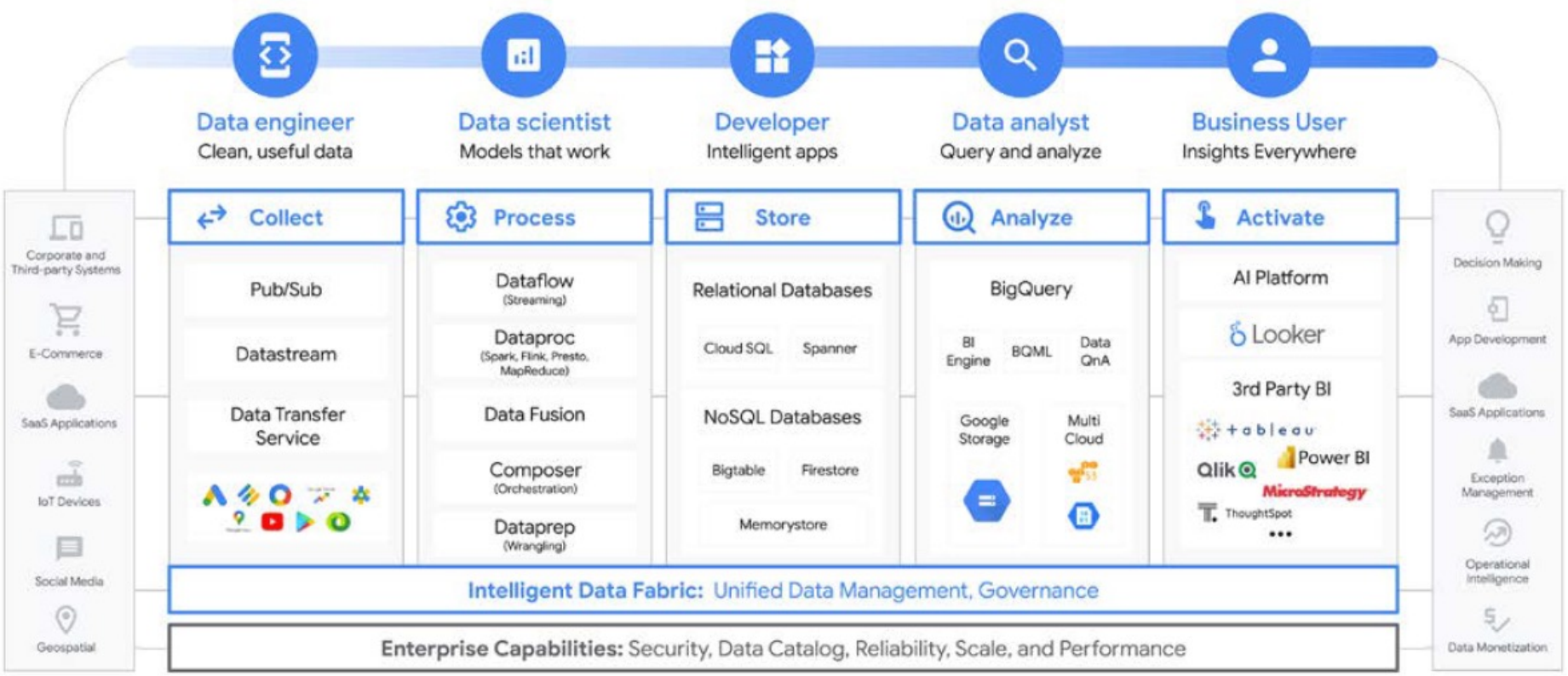
Analytics

- Looker
- Analytics Hub
- Dataplex

Source: https://cloud.google.com/docs, https://cloud.google.com/products

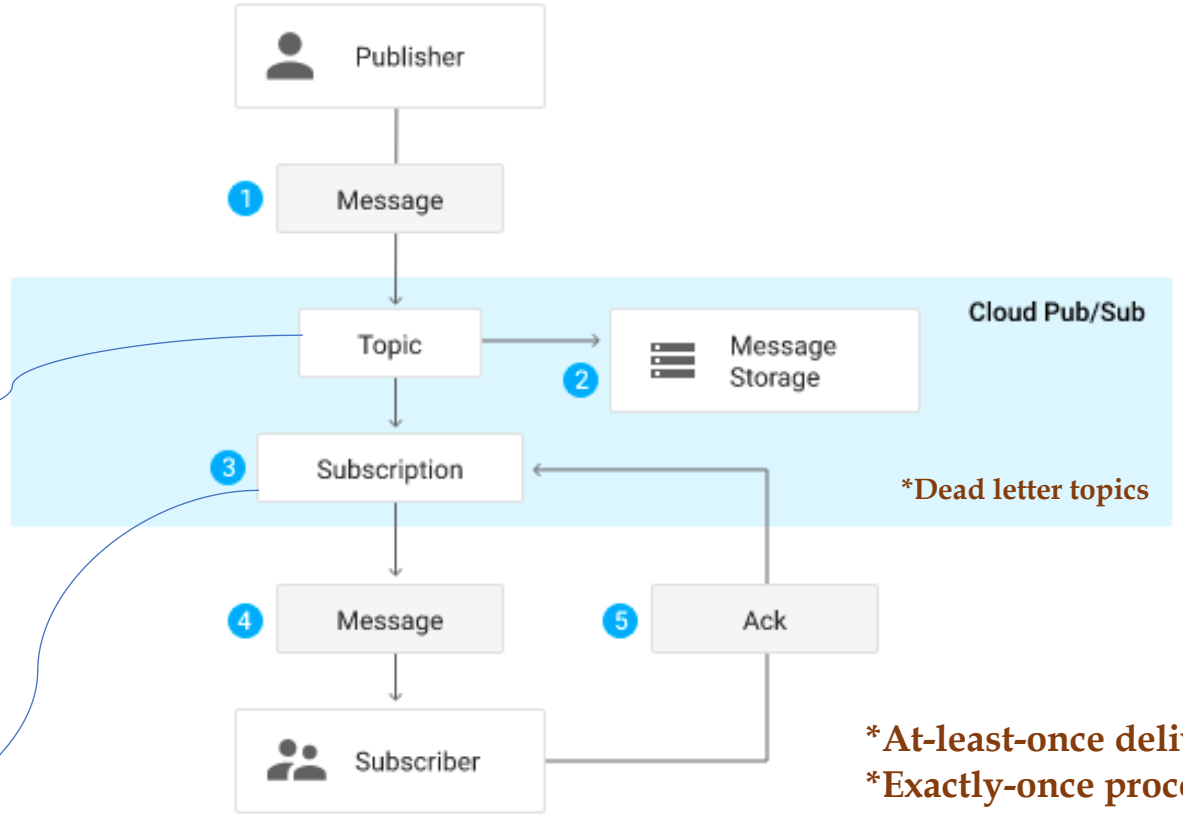
GCP: Data Analytics Lifecycle

The key differentiators of an analytics data platform built on GCP are that it is open, intelligent, flexible and tightly integrated. It Solves for every stage of the data analytics lifecycle, from ingestion to transformation and analysis, to business intelligence and AI.



Pub/Sub is an asynchronous messaging service that decouples services that produce and process events.

Pub/Sub message flow:



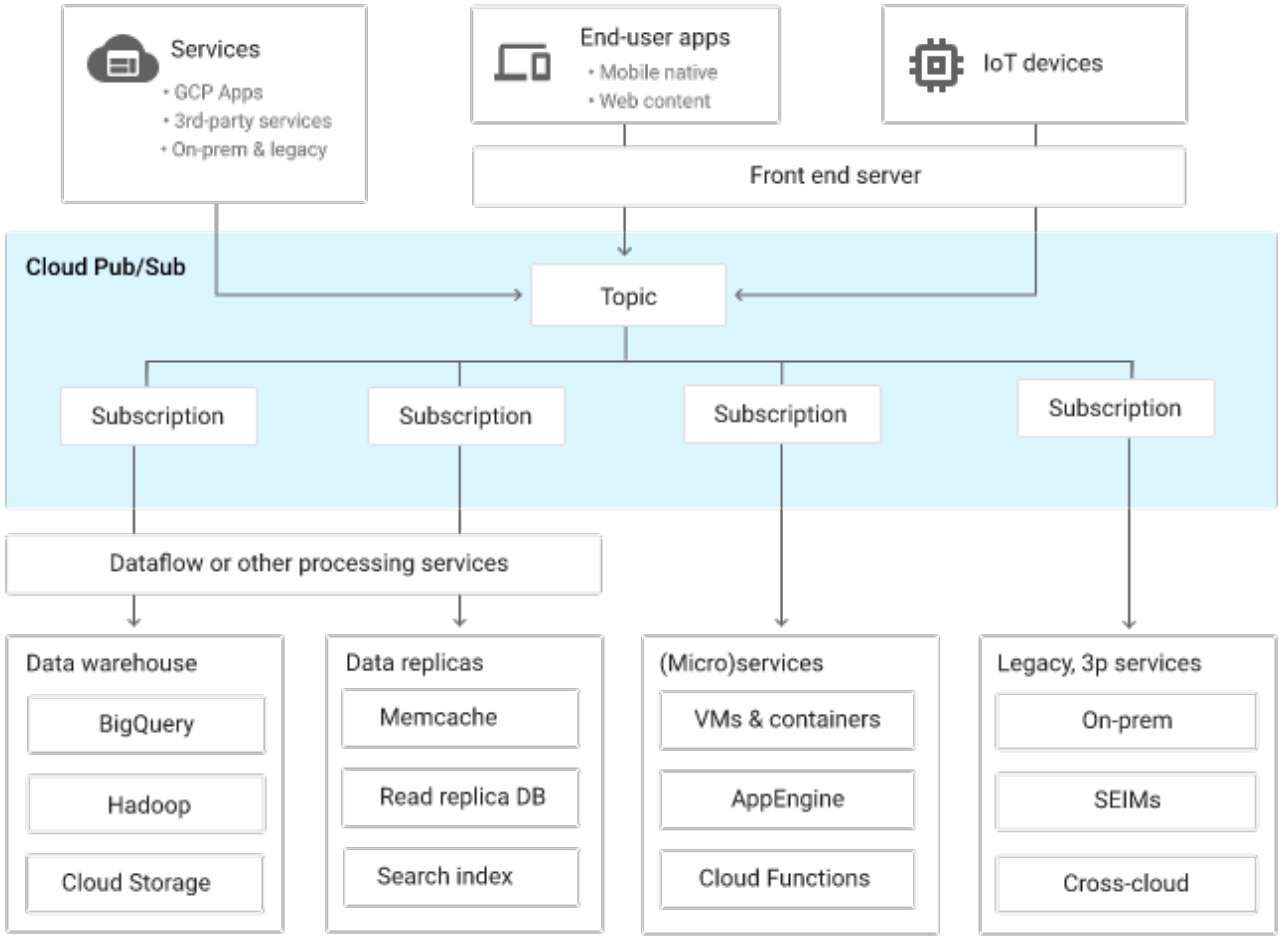
***Topic**: A named resource to which messages are sent by publishers.*

***Subscription**: A named resource representing the stream of messages from a single, specific topic, to be delivered to the subscribing application.*

***At-least-once delivery**
***Exactly-once processing (using Dataflow)**

Source: https://cloud.google.com/docs, https://cloud.google.com/products

Publisher and subscriber endpoints



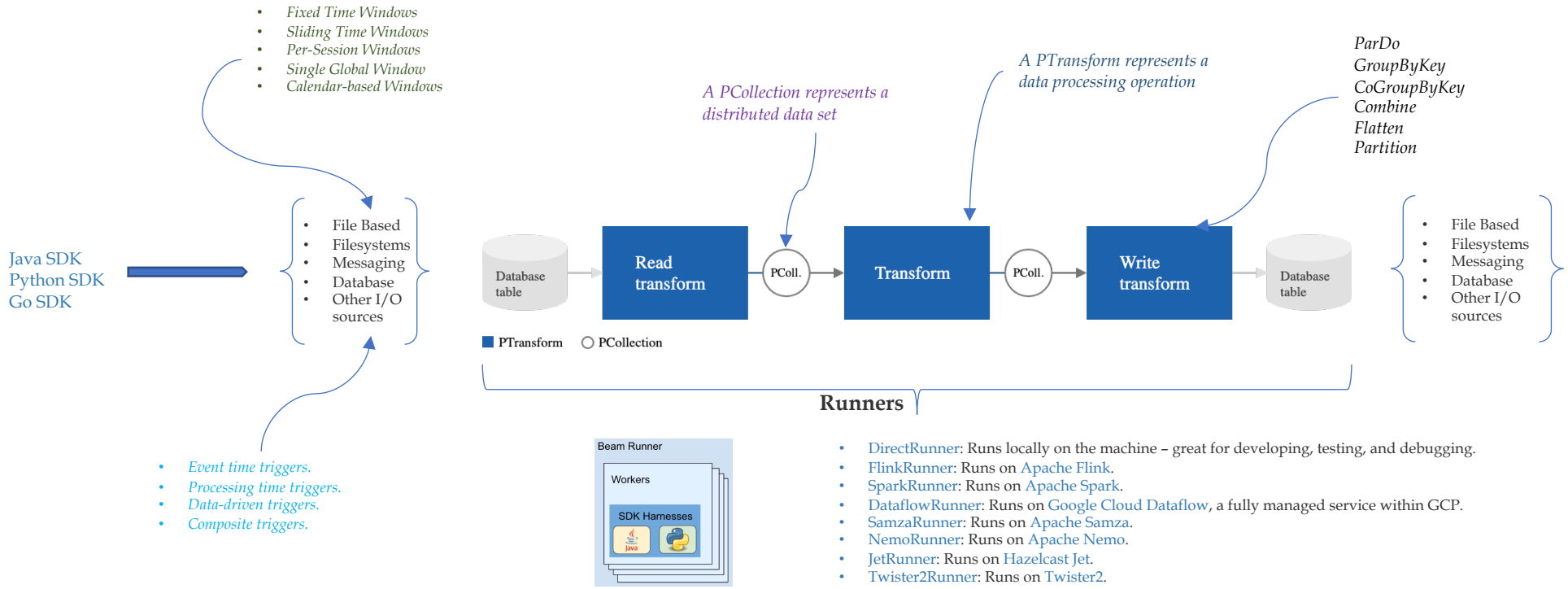
Common use cases

- **Enterprise event bus**
- **Real-time event distribution**
- **Parallel processing and workflows**
- **Data streaming from IoT devices**

❖ Publishers and Pull subscribers can be any application that can make HTTPS requests to pubsub.googleapis.com

Source: https://cloud.google.com/docs, https://cloud.google.com/products

Dataflow is a managed service for executing unified stream and batch data processing that's serverless and is developed using the Apache Beam SDK.
(*An open-source unified programming model to implement batch and streaming data processing pipeline that runs on any execution engine*).



Creating a pipeline

```
import apache_beam as beam
from apache_beam.options.pipeline_options import PipelineOptions

with beam.Pipeline(options=PipelineOptions()) as p:
    pass # build your pipeline here
```

Creating a PCollection

```
lines = p | 'ReadMyFile' >> beam.io.ReadFromText(
      'gs://some/inputData.txt')
```

Applying transforms

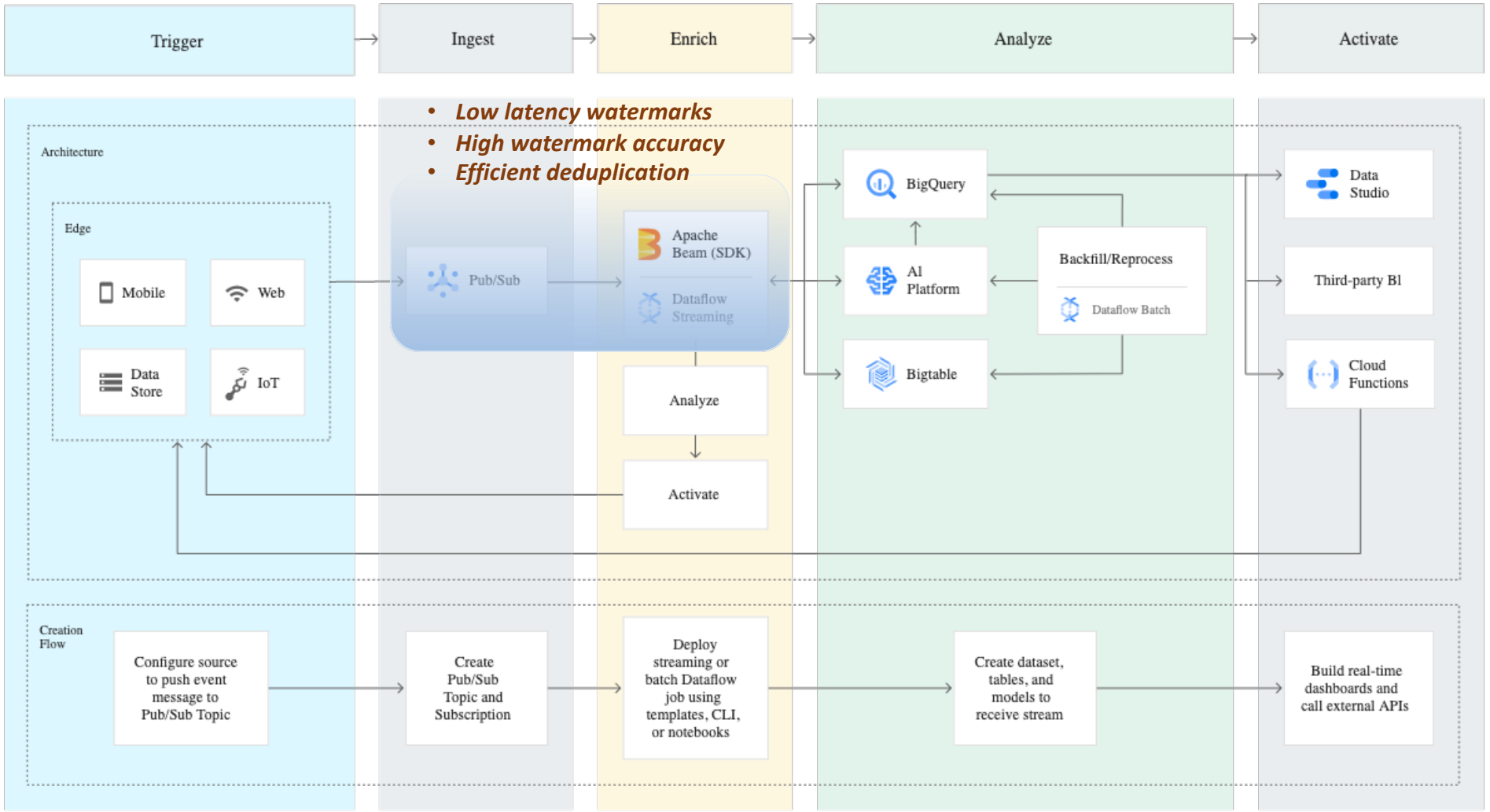
```
[Output PCollection] = [Input PCollection] | [Transform]
```

Applying ParDo

```
class ComputeWordLengthFn(beam.DoFn):
def process(self, element):
   return [len(element)]
word_lengths = words | beam.ParDo(ComputeWordLengthFn())
```

Source: https://cloud.google.com/docs, https://cloud.google.com/products

Stream analytics Use Case:



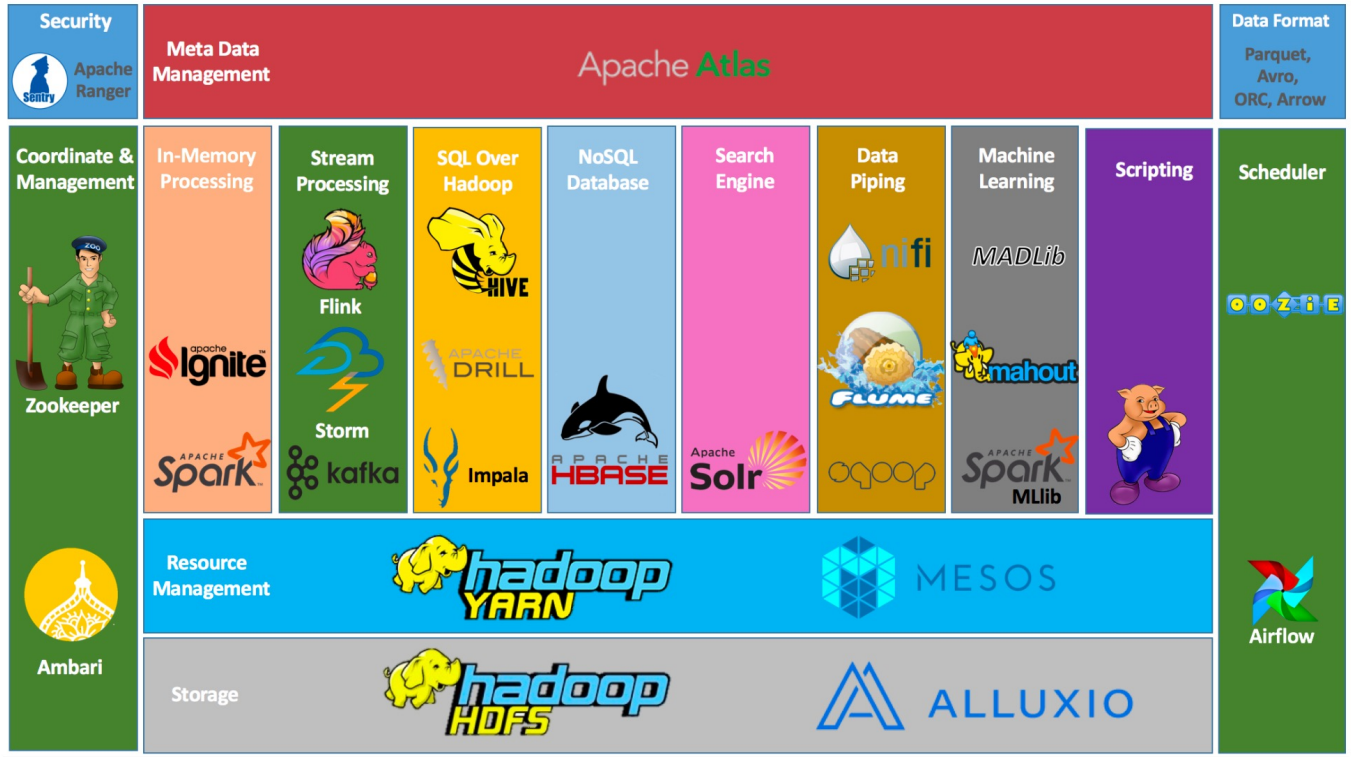
Source: https://cloud.google.com/docs, https://cloud.google.com/products

Big Data: Dataproc

Dataproc is a managed Apache Spark and Apache Hadoop service which takes advantage of open-source data tools for batch processing, querying, streaming, and machine learning.

Apache Hadoop Ecosystem

The Hadoop Distributed File System (HDFS) is a distributed file system based on master/slave architecture. An HDFS cluster consists of NameNode and DataNodes, data is replicated across datanodes (default 3) for highly fault-tolerant system.



Why Dataproc?

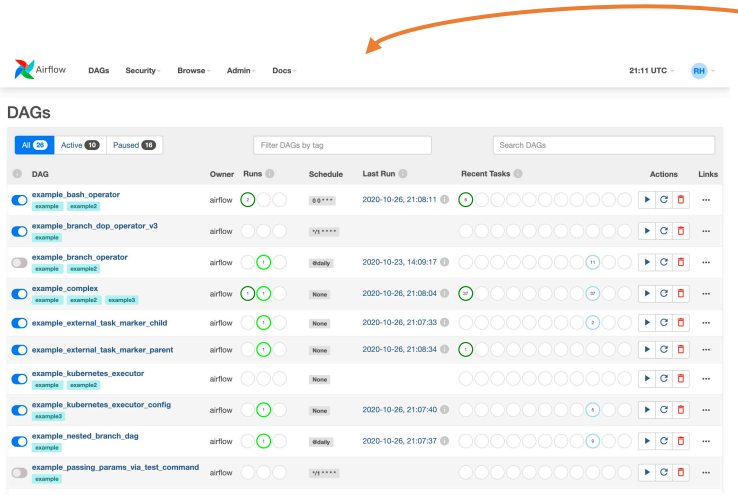
- **Low cost**
- **Super fast**
- **Integrated**
- **Enterprise Security**
- **Containerize OSS jobs**

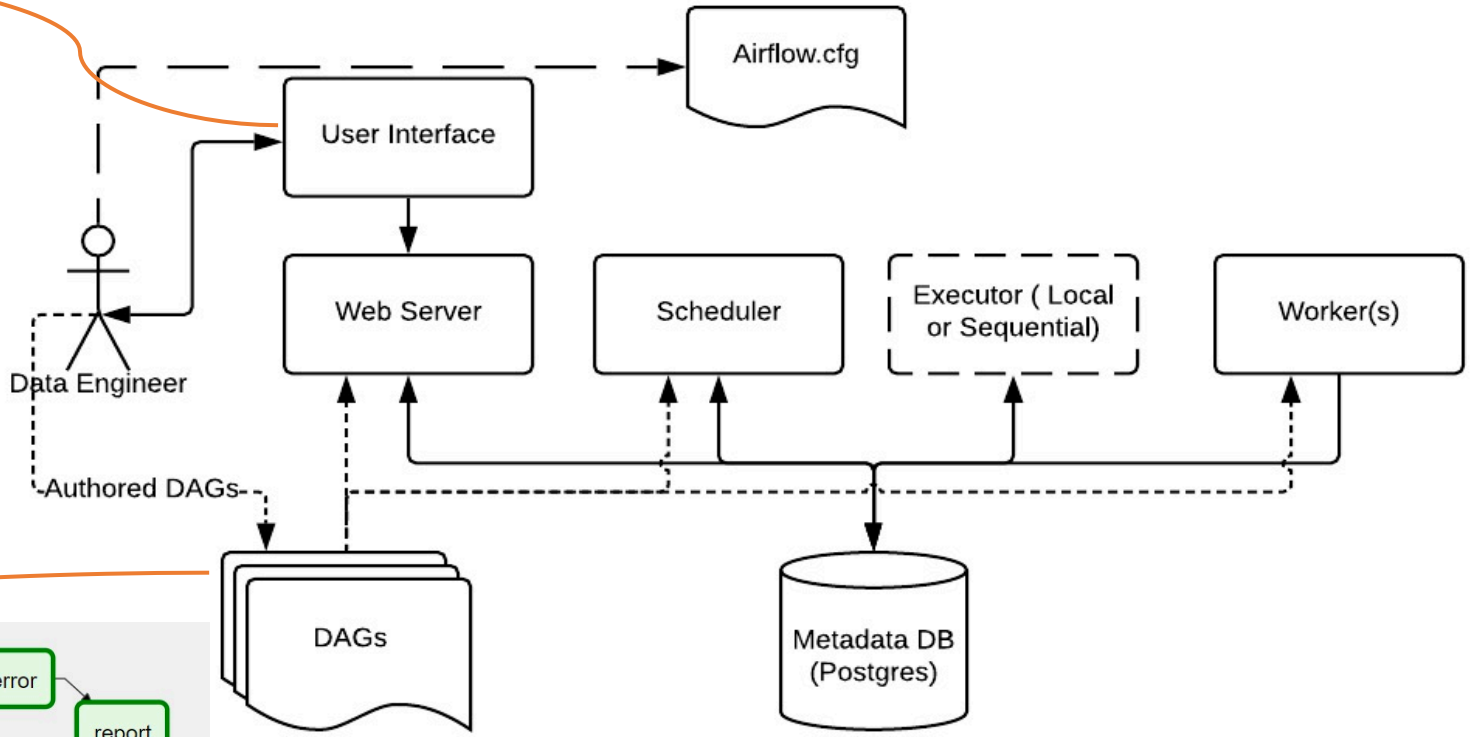
Source: https://cloud.google.com/docs, https://cloud.google.com/products

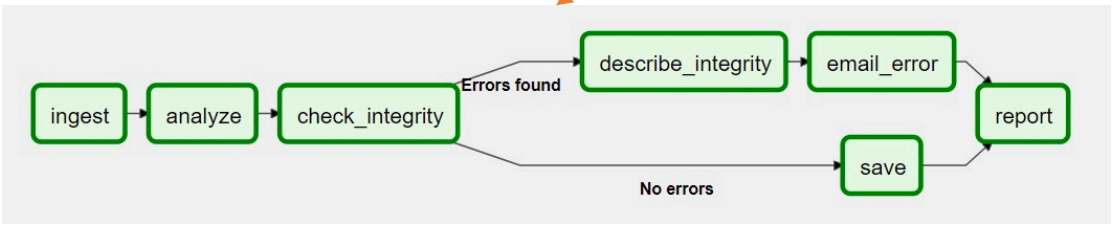
Big Data: Composer

A fully managed workflow orchestration service to author, schedule, and monitor pipelines that span across hybrid and multi-cloud environments. It's built on the Apache Airflow open-source project and operated using Python.

Basic Airflow architecture







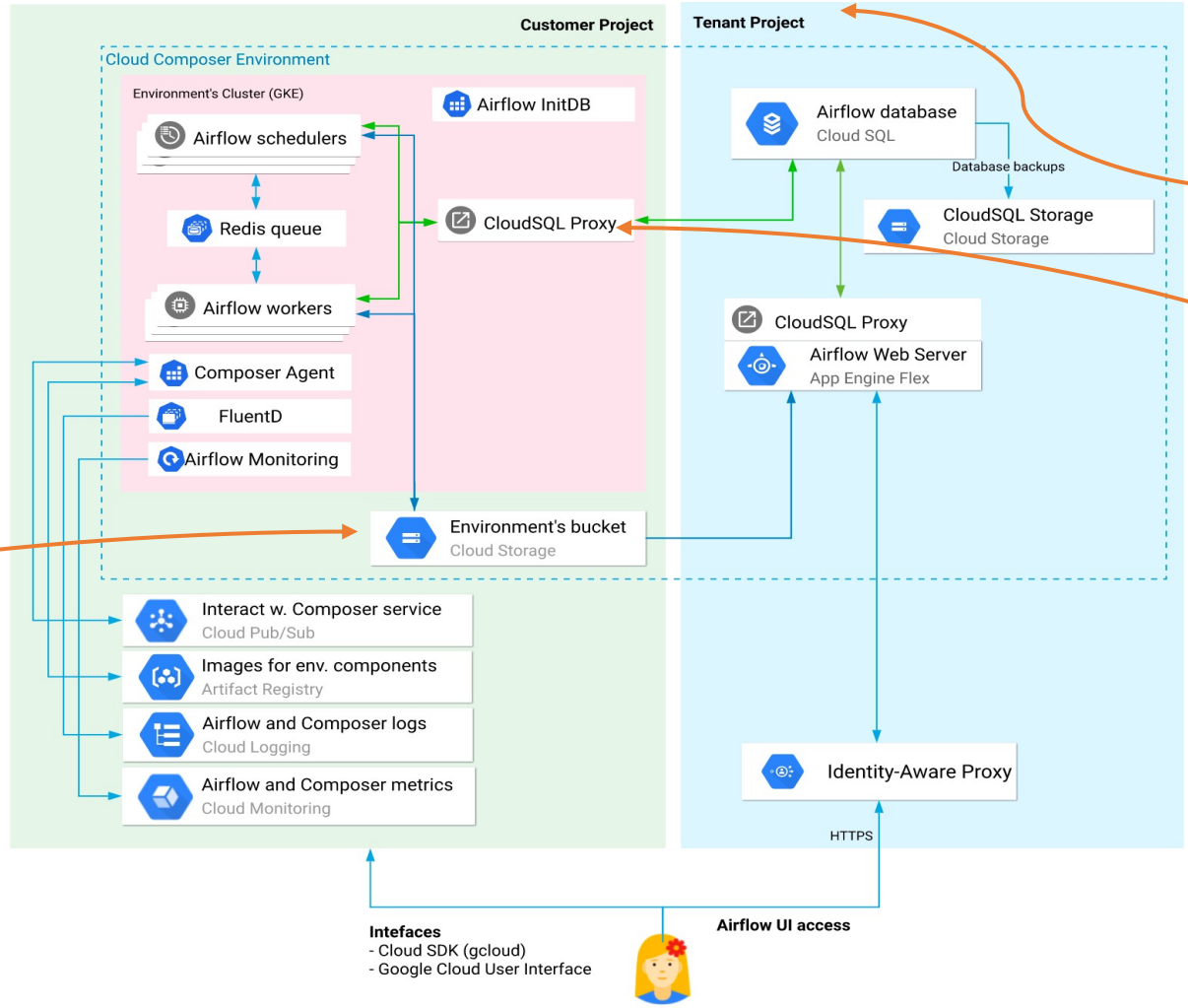
DAG (a Directed Acyclic Graph)

Source: https://cloud.google.com/docs, https://cloud.google.com/products

Cloud Composer environments are self-contained Airflow deployments based on Google Kubernetes Engine.

Architecture configurations:

- Public IP architecture
- Private IP architecture
- Private IP with Domain restricted sharing (DRS) architecture



The tenant project hosts a Cloud SQL instance, Cloud SQL storage, and a App Engine Flex instance that runs the Airflow web server.

Airflow schedulers and workers in the customer project communicate with the Airflow database through a Cloud SQL proxy instances located in the customer project.

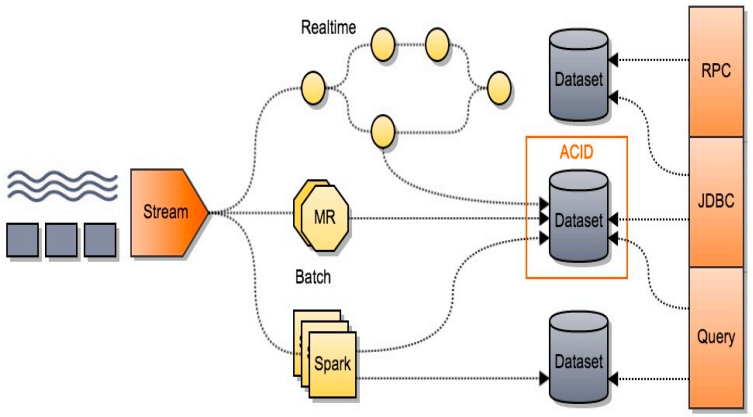
Environment's bucket is a Cloud Storage bucket that stores DAGs, plugins, data dependencies, and Airflow logs.

Source: https://cloud.google.com/docs, https://cloud.google.com/products

Big Data: Data Fusion

Cloud Data Fusion is a fully managed, cloud-native, enterprise data integration service for quickly building and managing data pipelines. It is powered by the open source project CDAP(Cask Data Application Platform).

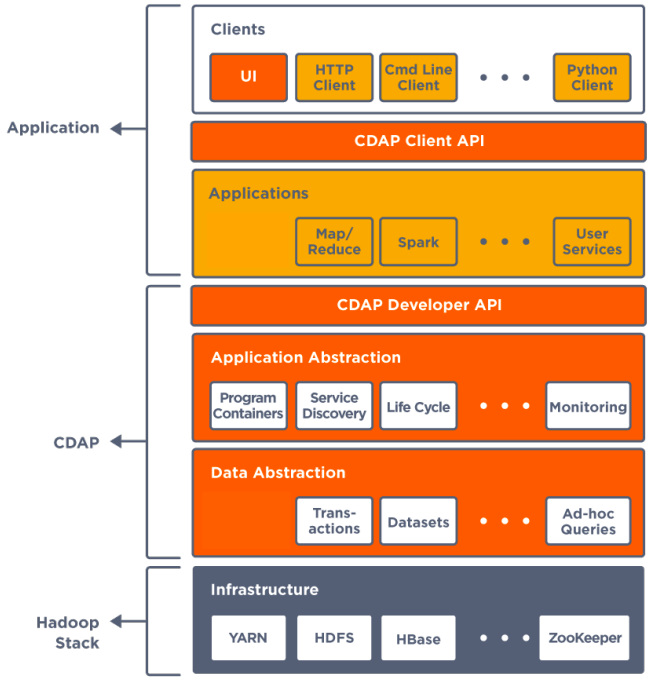
Anatomy of Big Data Applications



As an application developer building a Big Data application, following five are the primarily concerns:

- ✓ Data Collection
- ✓ Data Exploration
- ✓ Data Processing
- ✓ Data Storage
- ✓ Data Serving

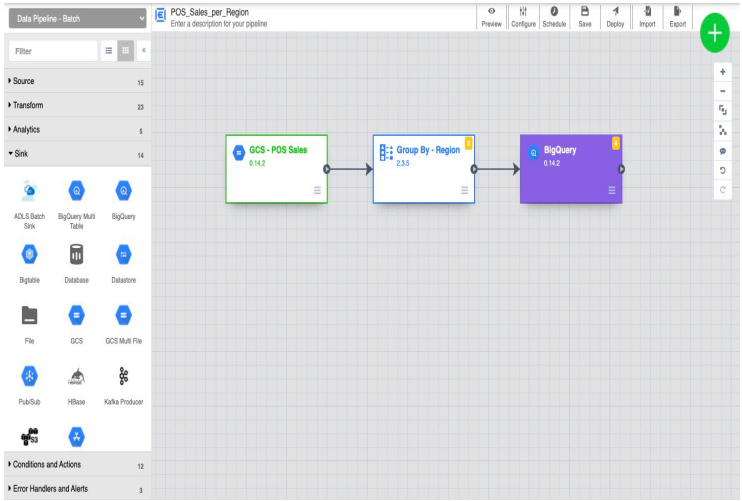
CDAP Abstraction



CDAP can be run in three different runtime modes:

- **Distributed CDAP** for staging and production.
- **CDAP Sandbox** for testing and development on a developer's laptop.
- **In-Memory CDAP** for unit testing and continuous integration pipelines.

Code-free self-service UI driven hybrid and multi-cloud data integration tool

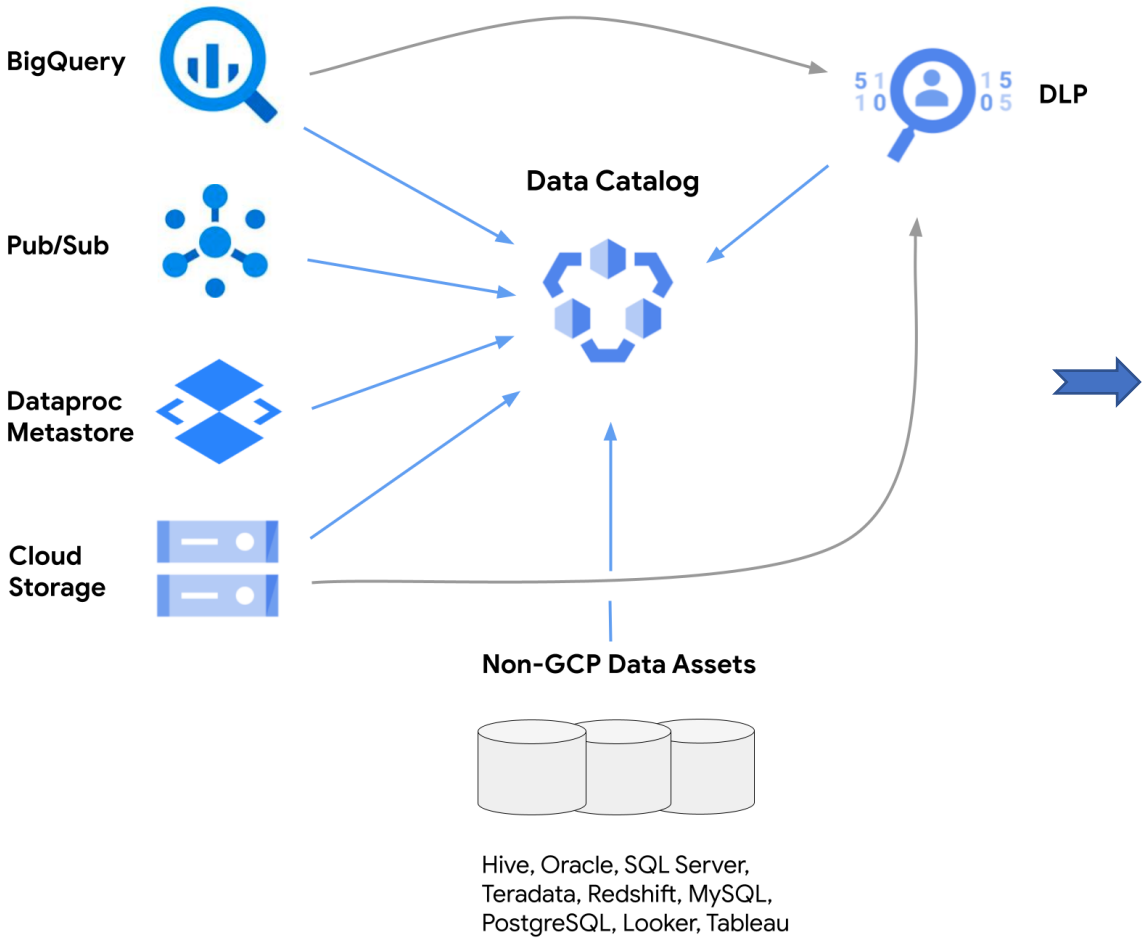


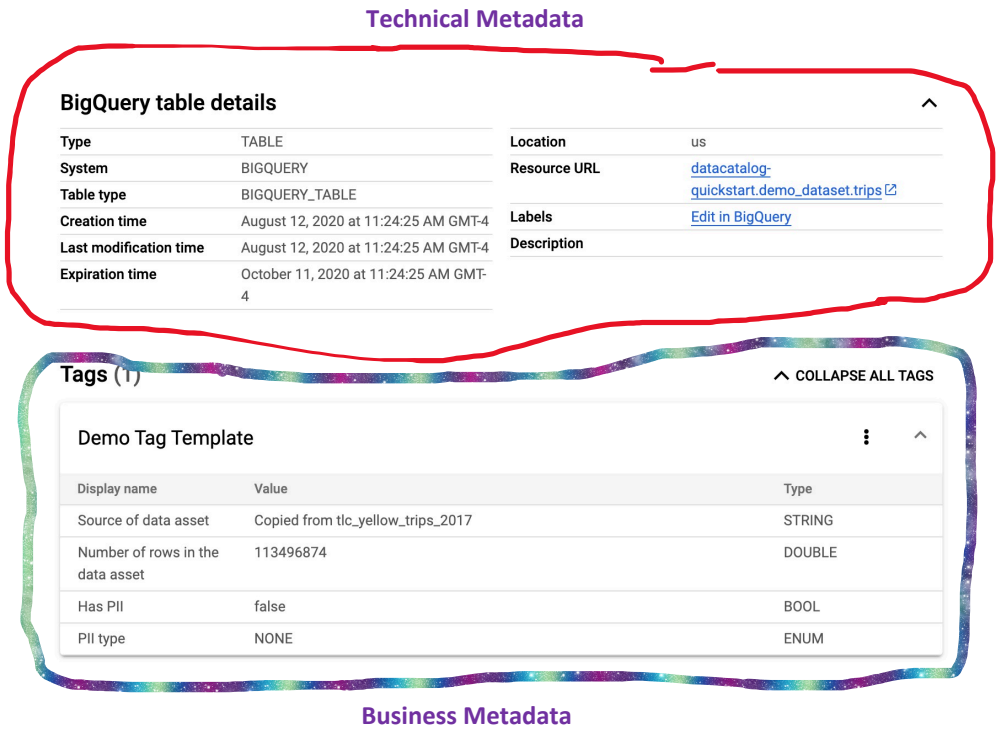
- ✓ Integrated with Google's big-data tools (Storage, Dataproc, BigQuery, Spanner etc)
- ✓ Enterprise Grade Security
- ✓ Metadata and Lineage integration
- ✓ Built-in connectors for modern and legacy systems

Source: https://cloud.google.com/docs, https://cloud.google.com/products

Big Data: Data Catalog

A fully managed and highly scalable data discovery and metadata management service. It can pinpoint data with a simple but powerful faceted-search interface and sync technical metadata automatically and create schematized tags for business metadata.



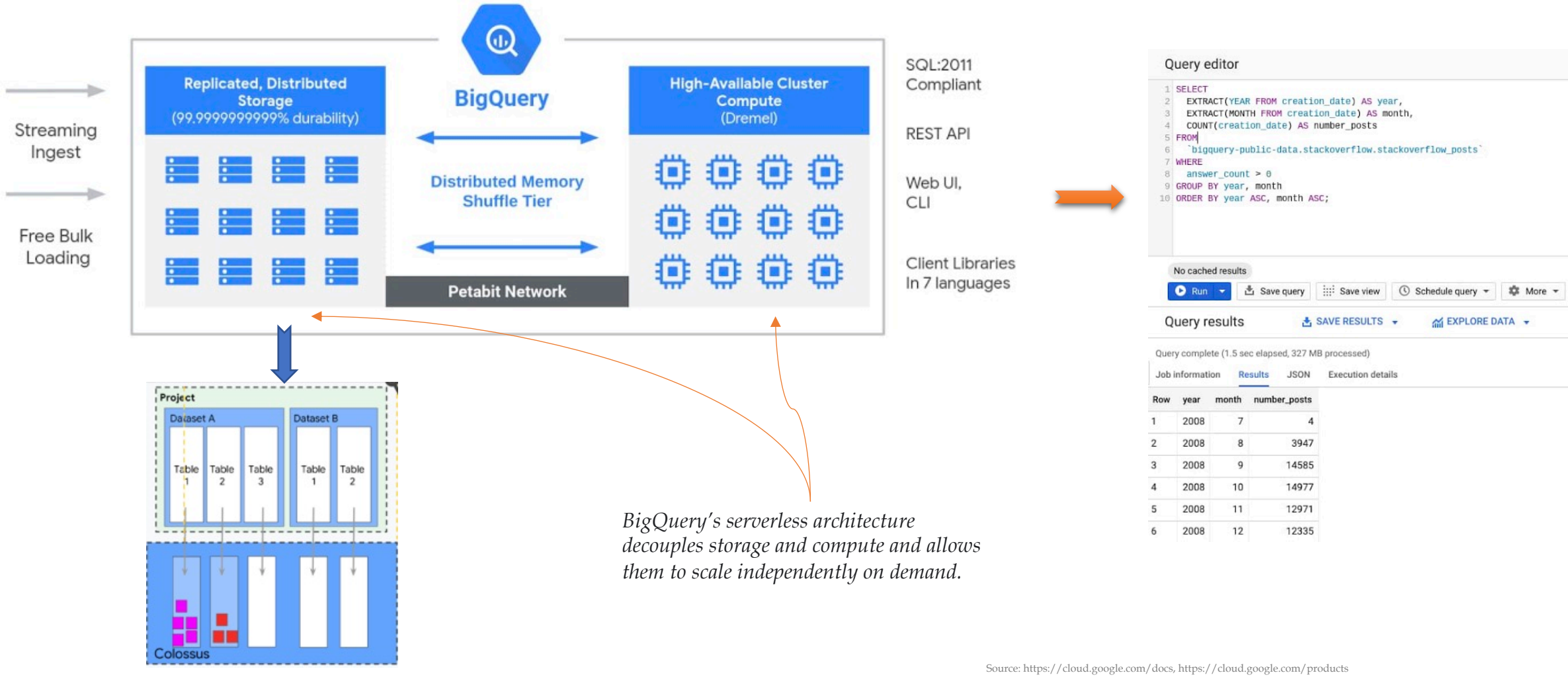


Source: https://cloud.google.com/docs, https://cloud.google.com/products

BigQuery: Architecture

BigQuery is Google Cloud's fully managed, petabyte-scale, and cost-effective analytics data warehouse that helps you manage and analyze data with built-in features like machine learning, geospatial analysis, and business intelligence.

BigQuery Architecture:



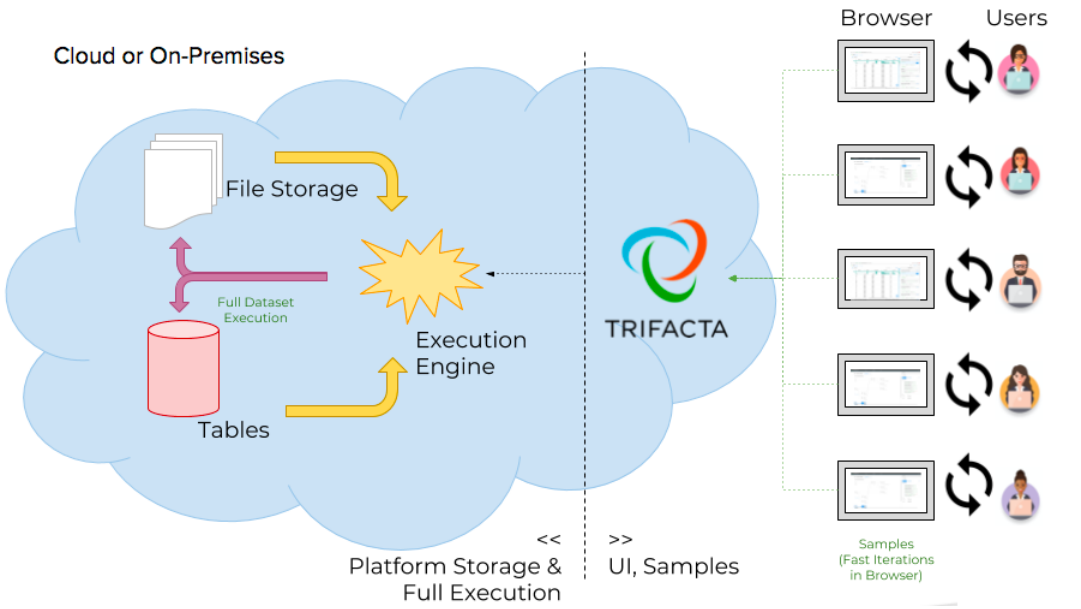
BigQuery's serverless architecture decouples storage and compute and allows them to scale independently on demand.

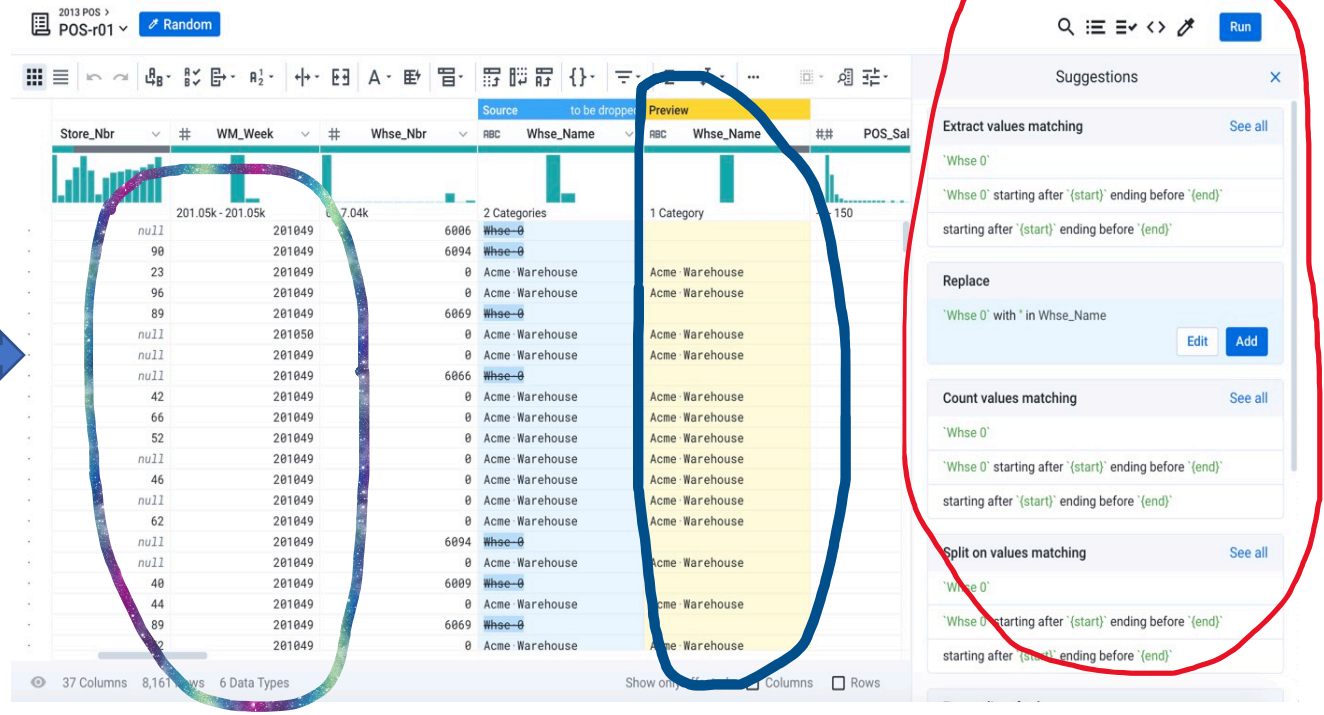
Source: https://cloud.google.com/docs, https://cloud.google.com/products

Big Data: Cloud Dataprep

Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis, reporting, and machine learning.

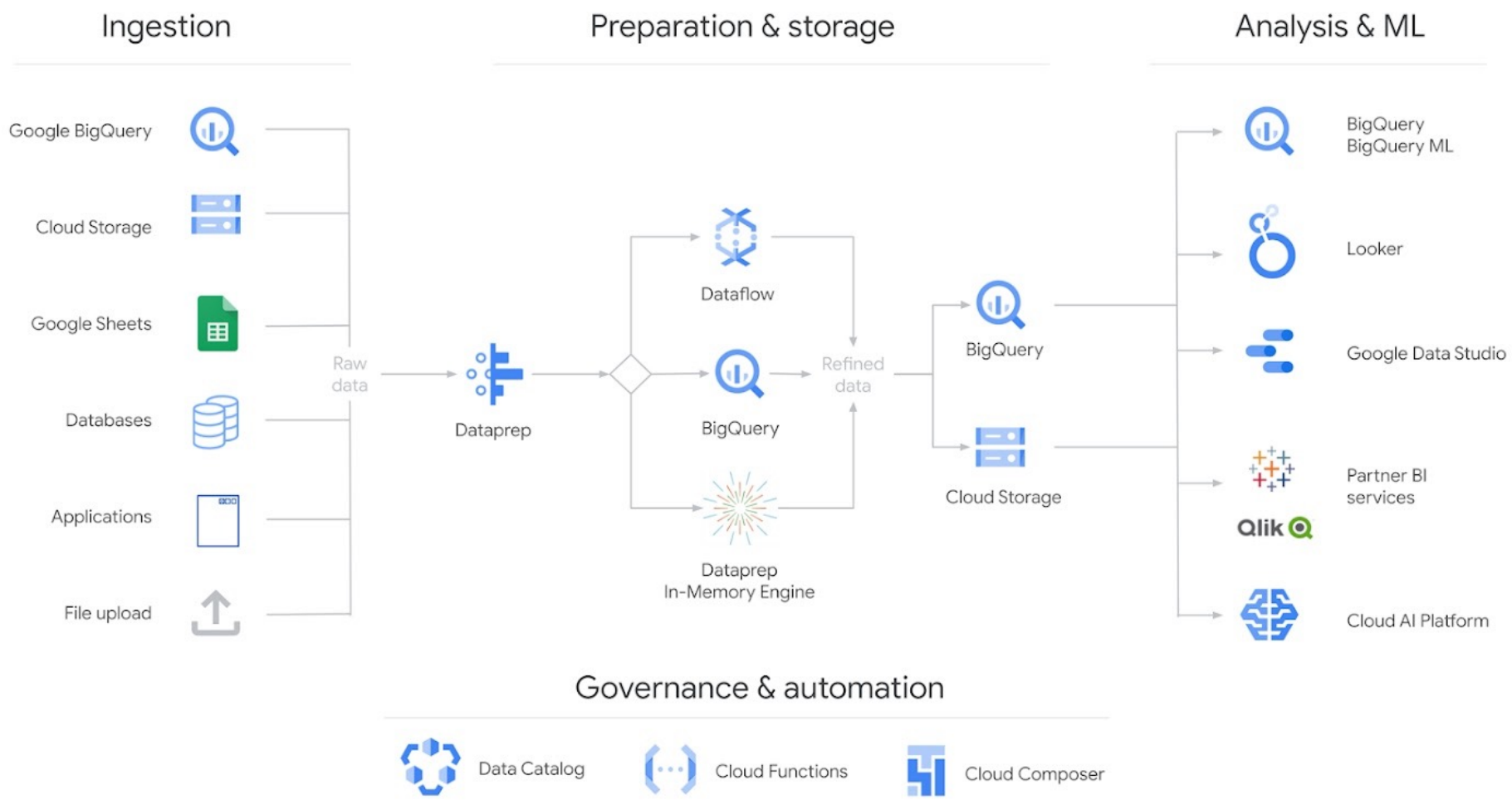
How does it work?





Source: https://cloud.google.com/docs, https://cloud.google.com/products

Cloud Dataprep ELT pipeline architecture:

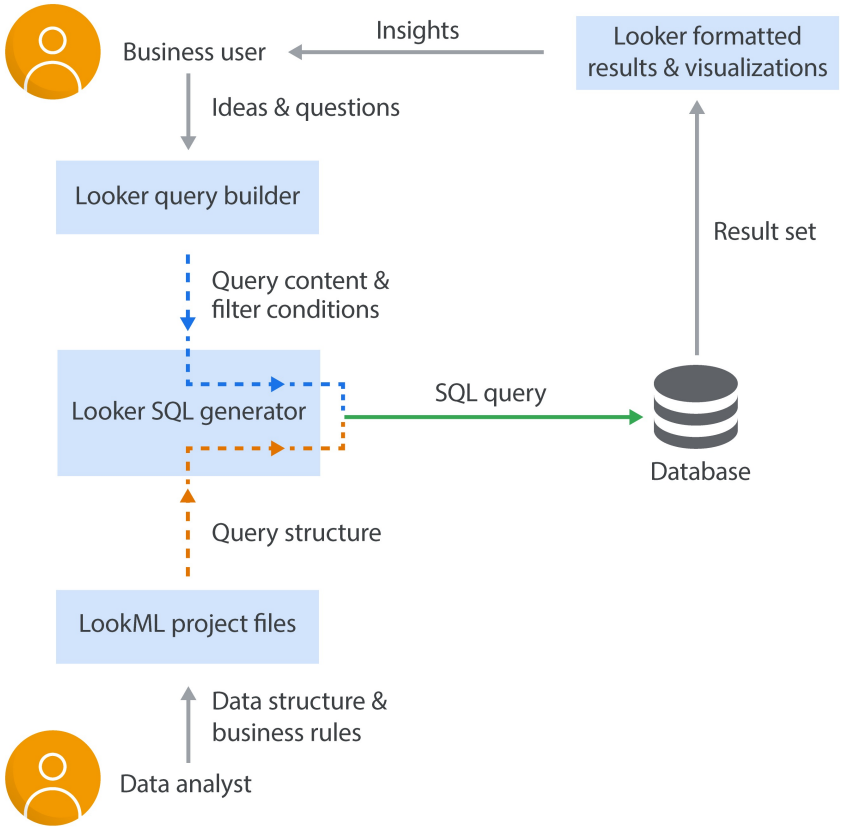


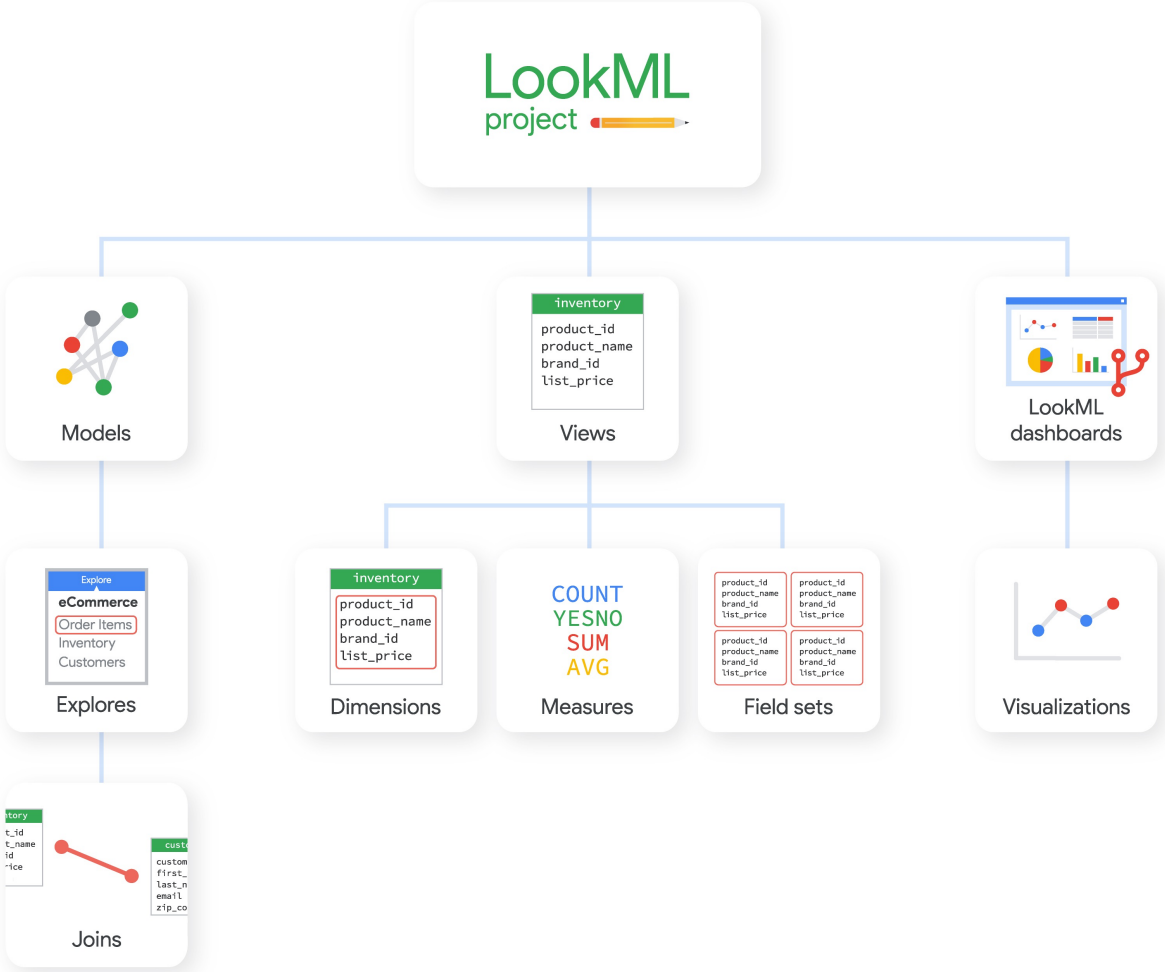
Source: https://cloud.google.com/docs, https://cloud.google.com/products

Big Data: Looker

An enterprise platform for business intelligence, data applications, and embedded analytics. It's an integrated end-to-end multi cloud platform to Connect, analyse, and visualize data across Google Cloud, Azure, AWS, on-premises databases, or ISV SaaS applications.

What is LookML?

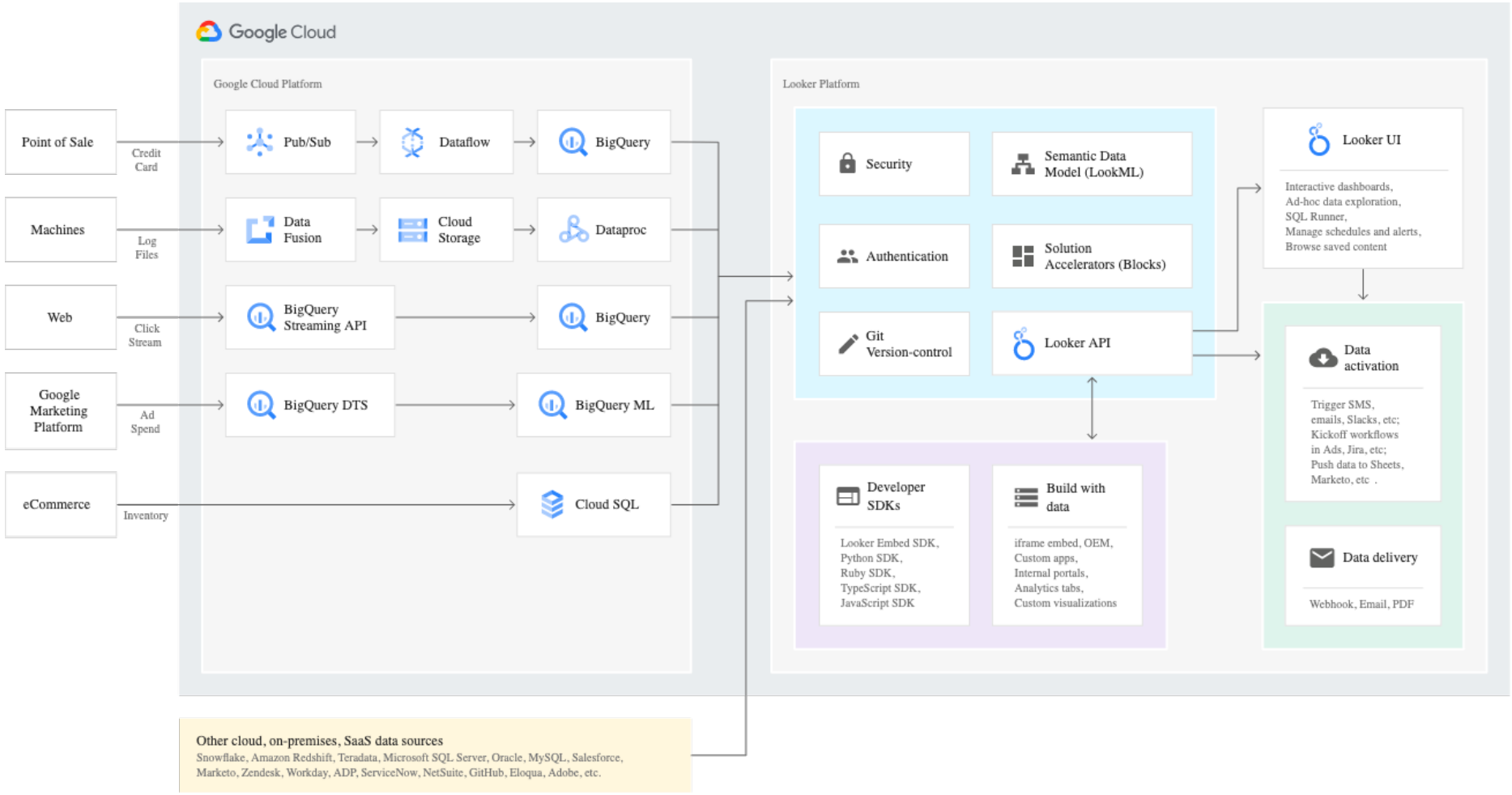




Source: https://cloud.google.com/docs, https://cloud.google.com/products

Big Data: Looker

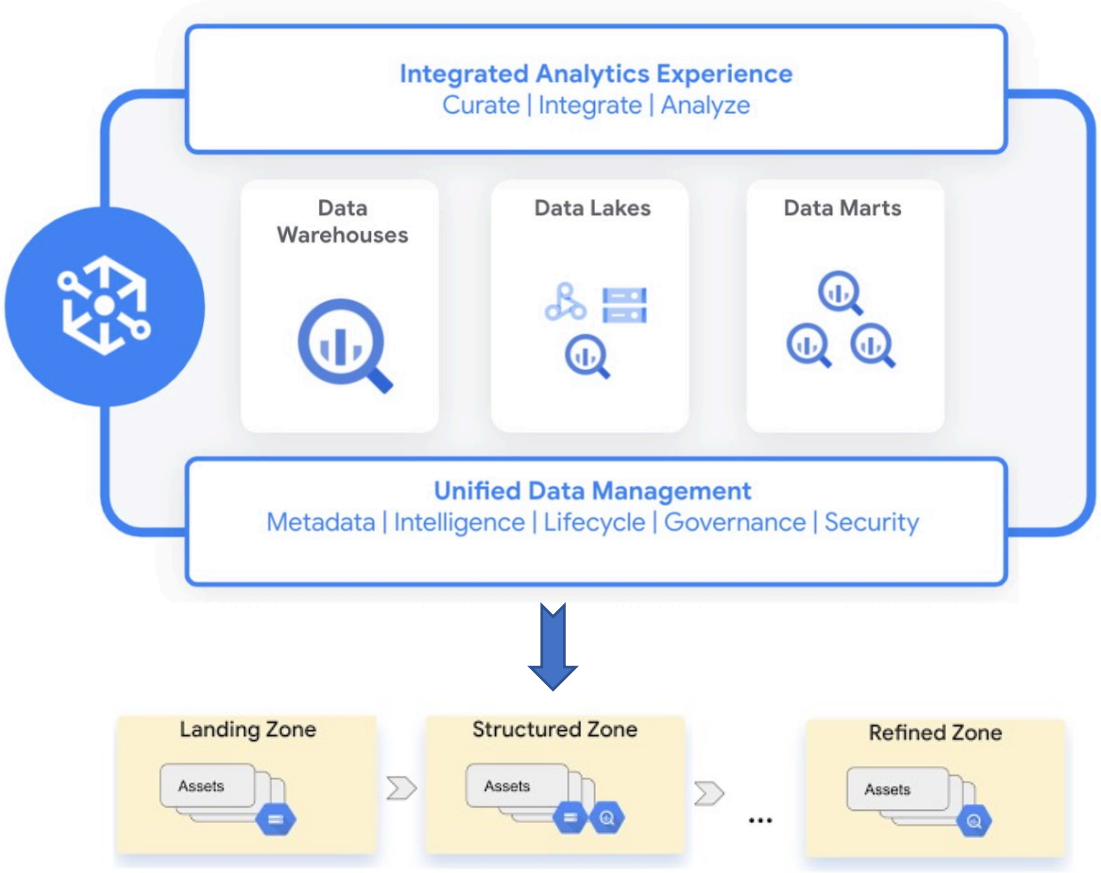
An enterprise platform for business intelligence, data applications, and embedded analytics.

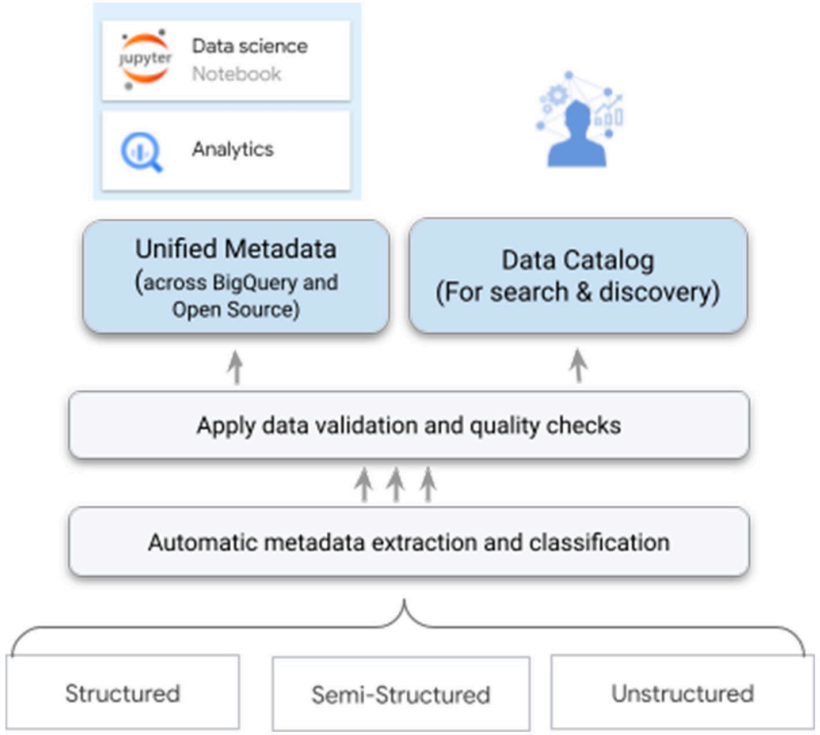


Source: https://cloud.google.com/docs, https://cloud.google.com/products

Dataplex

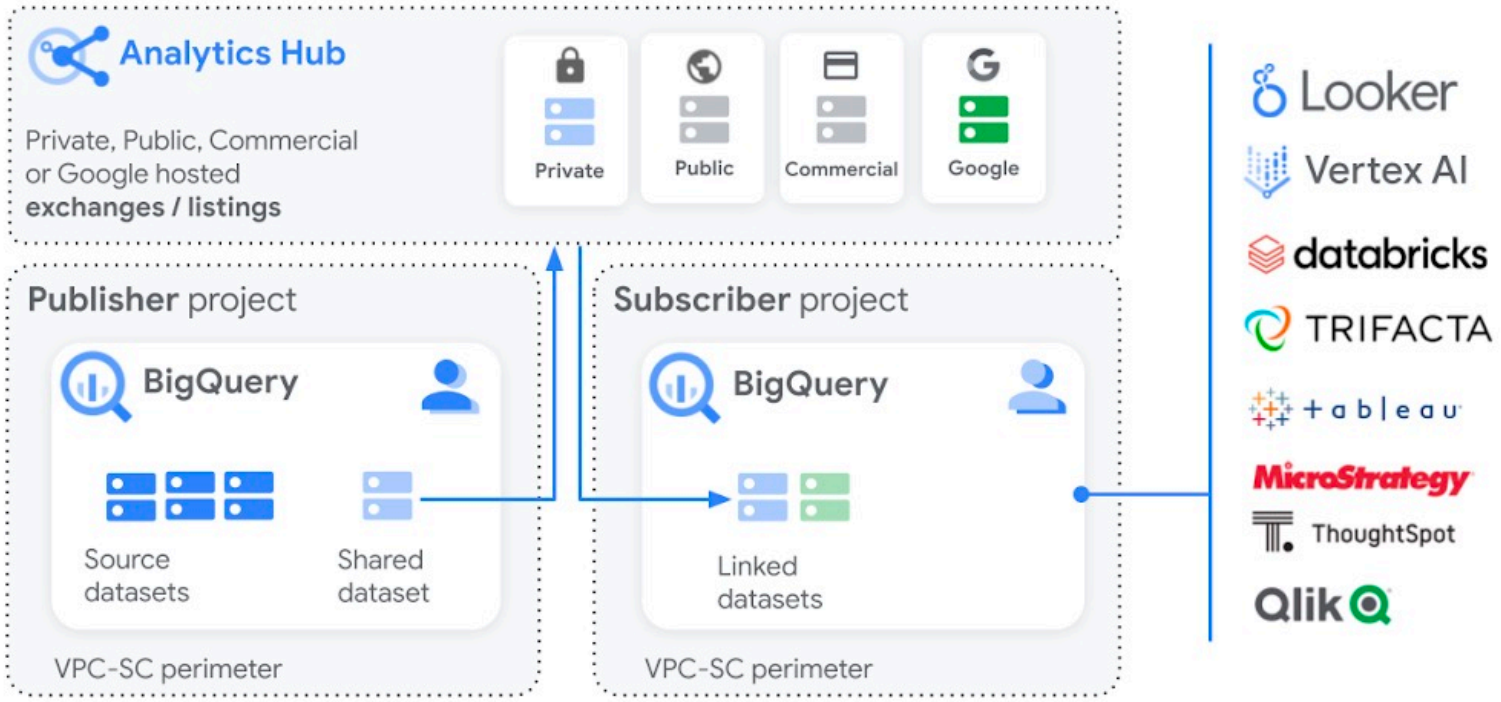
Dataplex is an intelligent data fabric that provides a way to centrally manage, monitor, and govern data across data lakes, data warehouses and data marts, and make this data securely accessible to a variety of analytics and data science tools.





Analytics Hub

Analytics Hub is a new fully managed service, available in preview(Q3), that helps unlock the value of data sharing, leading to new insights and increased business value.



A sharing model for scalability, security, and flexibility

Data Analytics: Use Cases

Data Warehouse Modernization

- BigQuery
- BigQuery ML
- BigQuery GIS

Business Intelligence


Looker


BigQuery BI Engine


Connected Sheets


Data Studio

Workflow Orchestration


Cloud Composer


Cloud Functions

Data Lake Modernization


BigQuery


Dataproc


Cloud Storage

Marketing Analytics


Google Marketing Platform


Data Studio


BigQuery Data Transfer Service

Data Security & Governance


Data Catalog


Cloud DLP


Cloud IAM

Stream Analytics


Dataflow


Pub/Sub


BigQuery

Data Integration


Cloud Data Fusion


Dataflow


Pub/Sub

Source: https://cloud.google.com/docs, https://cloud.google.com/products