BigQuery

Organizations rely on data warehouses to aggregate data from disparate sources, process it, and make it available for data analysis in support of strategic decision-making.

BigQuery

- BigQuery is the Google Cloud enterprise data warehouse designed to help organizations to run large scale analytics with ease and quickly unlock actionable insights.

- You can ingest data into BigQuery either through batch uploading or by streaming data directly to unlock real-time insights.

BigQuery

- Google BigQuery is a fully managed data warehouse tool.
- As a fully-managed data warehouse, BigQuery takes care of the infrastructure so you can focus on analyzing your data up to petabyte-scale.
- It allows scalable analysis over a petabyte of data, querying using ANSI SQL, integration with various applications, etc.
- To access all these features conveniently, you need to understand BigQuery architecture, maintenance, pricing, and security.

BigQuery

- Google BigQuery is a Cloud Datawarehouse run by Google.
- It is capable of analyzing terabytes of data in seconds.
- If you know how to write SQL Queries, you already know how to query it.
- In fact, there are plenty of interesting public data sets shared in BigQuery, ready to be queried by you

BigQuery

- You can access BigQuery by using the GCP console or the classic web UI, by using a command-line tool, or by making calls to BigQuery Rest API using a variety of Client Libraries such as Java, and .Net, or Python.

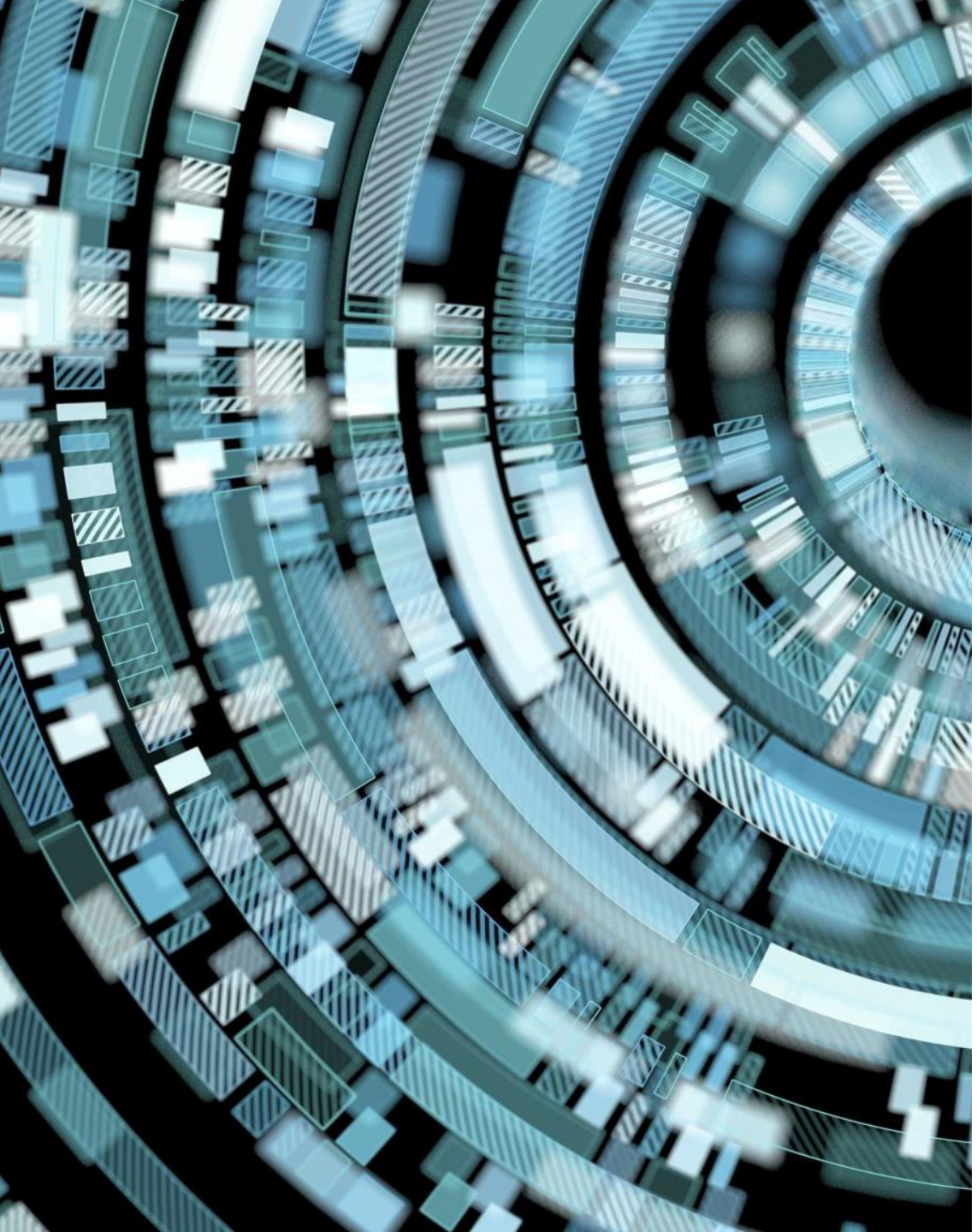
BigQuery unique features

Ease of Implementation: Building your own is expensive, time-consuming, and difficult to scale. With BigQuery, you need to load data first and pay only for what you use.

Speed: Process billions of rows in seconds and handle the real-time analysis of Streaming data.

BI Engine: BigQuery BI Engine is a fast, in-memory analysis service that accelerates many SQL queries in BigQuery by intelligently caching the data you use most frequently.

BI Engine can accelerate SQL queries from any source, including those written by data visualization tools, and can manage cached tables for on-going optimization.



BigQuery unique features cont.,

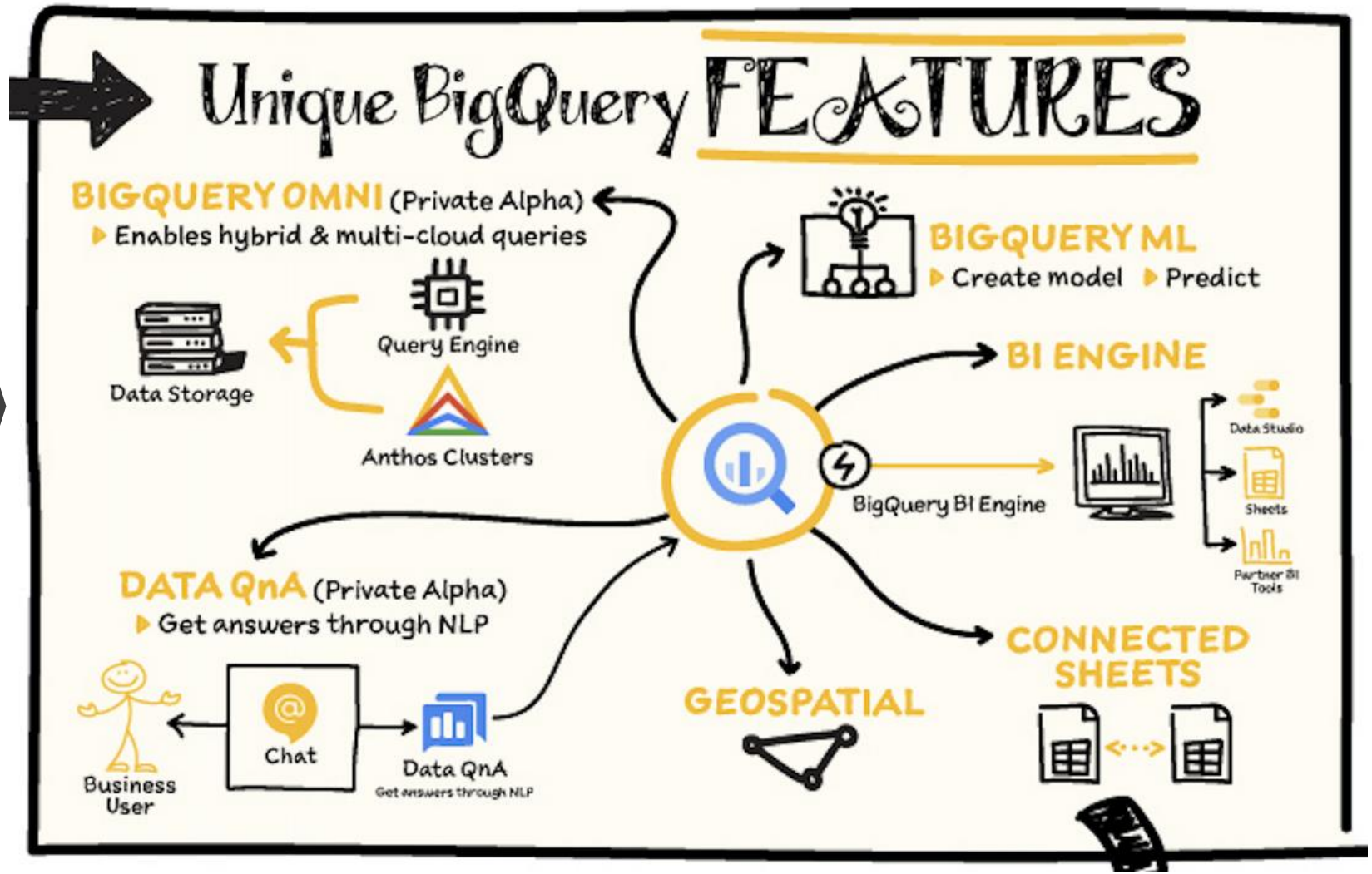
- **BigQuery ML:** BigQuery ML is unlocking machine learning for millions of data analysts.

- **BigQuery Omni** - BigQuery Omni is a flexible, multi-cloud analytics solution powered by Anthos that lets you cost-effectively access and securely analyze data across Google Cloud, Amazon Web Services (AWS), and Azure, without leaving the BigQuery user interface (UI).

- **Data QnA:** Data QnA enables self-service analytics for business users on BigQuery data as well as federated data from Cloud Storage, Bigtable, Cloud SQL, or Google Drive.
- It uses Dialogflow and enables users to formulate free-form text analytical questions, with auto-suggested entities while users type a question.

- **Connected Sheets** -The native integration between Sheets and BigQuery makes it possible for all business stakeholders, who are already quite familiar with spreadsheet tools, to get their own up-to-date insights at any time.

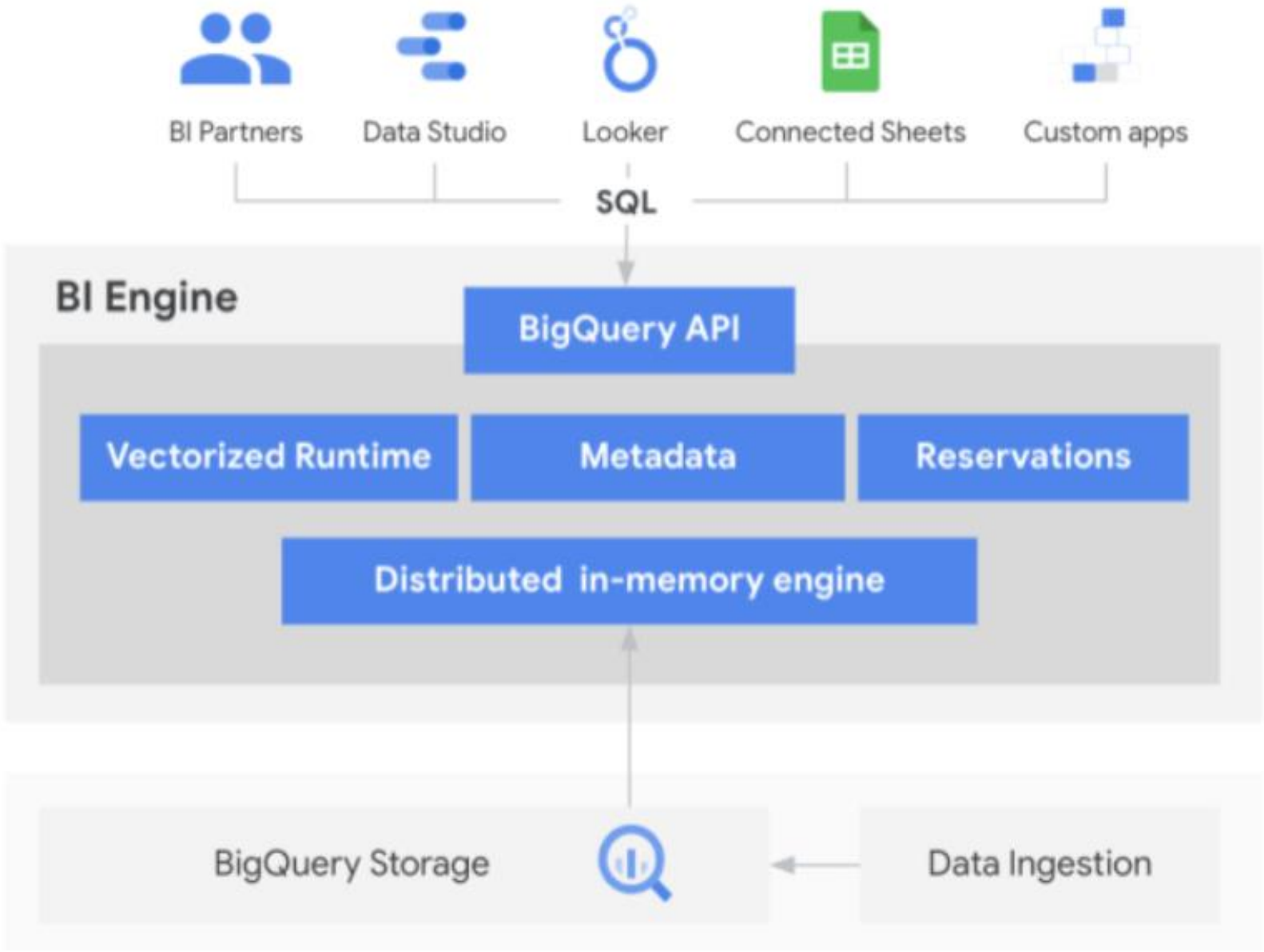
BigQuery unique features



BI Engine provides the following advantages:

- BigQuery API: BI Engine directly integrates with the BigQuery API.
- Any BI solution or custom application that works with the BigQuery API through standard mechanisms such as REST or JDBC and ODBC drivers can use BI Engine without modification.
- Vectorized runtime: With the BI Engine SQL interface, BI Engine introduces a more modern technique called vectorized processing.
- Using vectorized processing in an execution engine makes more efficient use of modern CPU architecture, by operating on batches of data at a time.
- BI Engine also uses advanced data encodings, specifically, dictionary run-length encoding, to further compress the data that's stored in the in-memory layer.
- Seamless integration: BI Engine works with BigQuery features and metadata, including authorized views, column and row level security, and data masking.
- Reservations: BI Engine reservations manage memory allocation at the project location level.
- BI Engine caches specific columns or partitions that are queried, prioritizing those in tables marked as preferred.

BI Engine
SQL interface

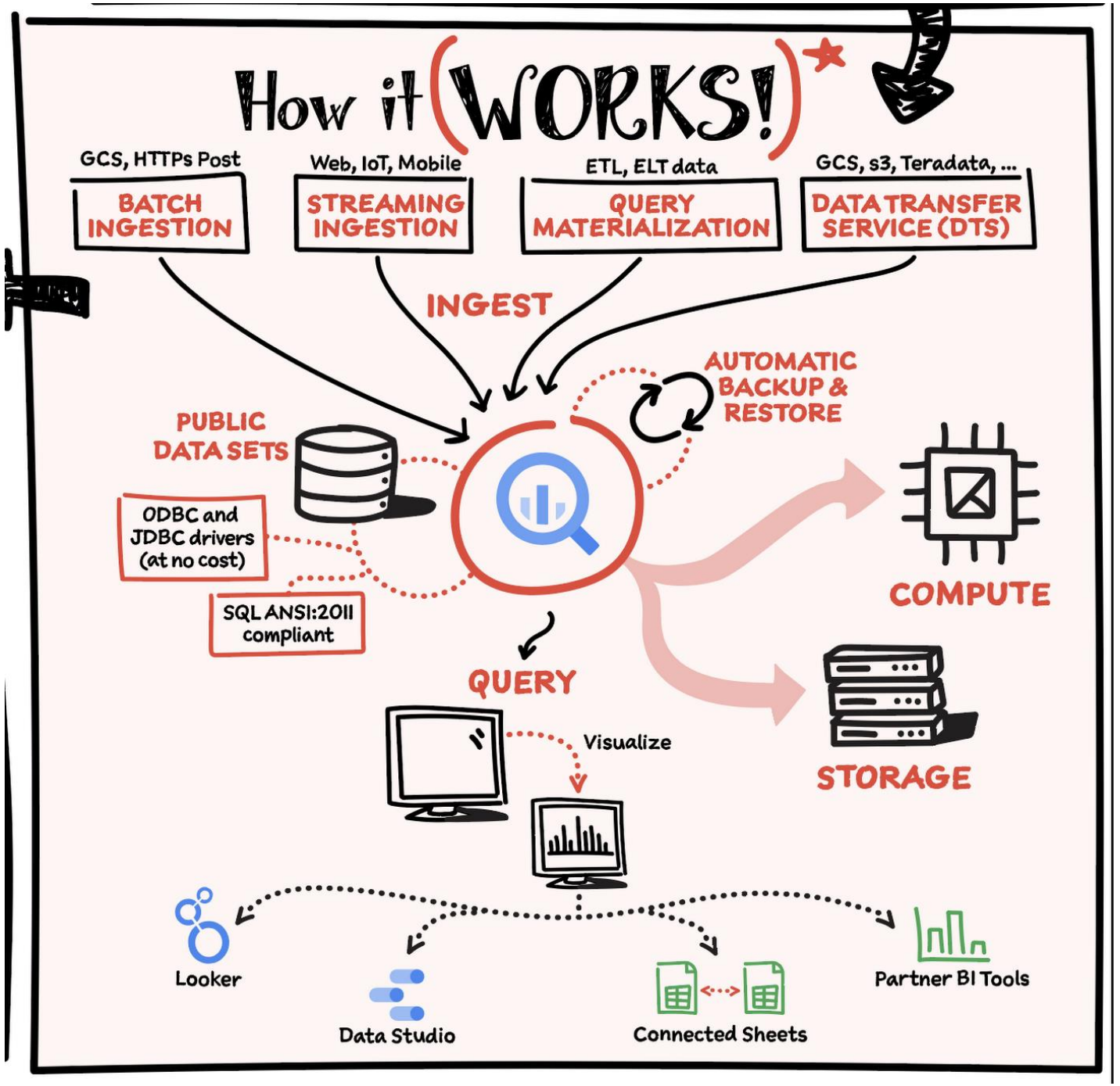


BI Engine SQL interface

- The BI Engine SQL interface expands BI Engine to integrate with other business intelligence (BI) tools such as Looker, Tableau, Power BI, and custom applications to accelerate data exploration and analysis.

- This diagram provides an overview of the BI Engine SQL interface, and the expanded capabilities that it brings to BI Engine.

How does it work?



How does it work?

- The BI Engine SQL interface expands BI Engine to integrate with other business intelligence (BI) tools such as Looker, Tableau, Power BI, and custom applications to accelerate data exploration and analysis.

- This diagram provides an overview of the BI Engine SQL interface, and the expanded capabilities that it brings to BI Engine.

BigQuery Organization

BigQuery is structured as a hierarchy with 4 levels:

- Projects: Top-level containers in the Google Cloud Platform that store the data
- Datasets: Within projects, datasets hold one or more tables of data
- Tables: Within datasets, tables are row-column structures that hold actual data
- Jobs: The tasks you are performing on the data, such as running queries, loading data, and exporting data

Projects

- Projects are the top-level containers that store the data
- Within the project, you can configure settings, permissions, and other metadata that describe your applications
- Each project has a name, ID, and number that you'll use as identifiers
- When billing is enabled, each project is associated with one billing account but multiple projects can be billed to the same account

Datasets

- Datasets allow you to organize and control access to your tables.
- All tables must belong to a dataset. You must create a dataset before loading data into BigQuery.
- You can configure permissions at the organization, project, and dataset level.

Tables

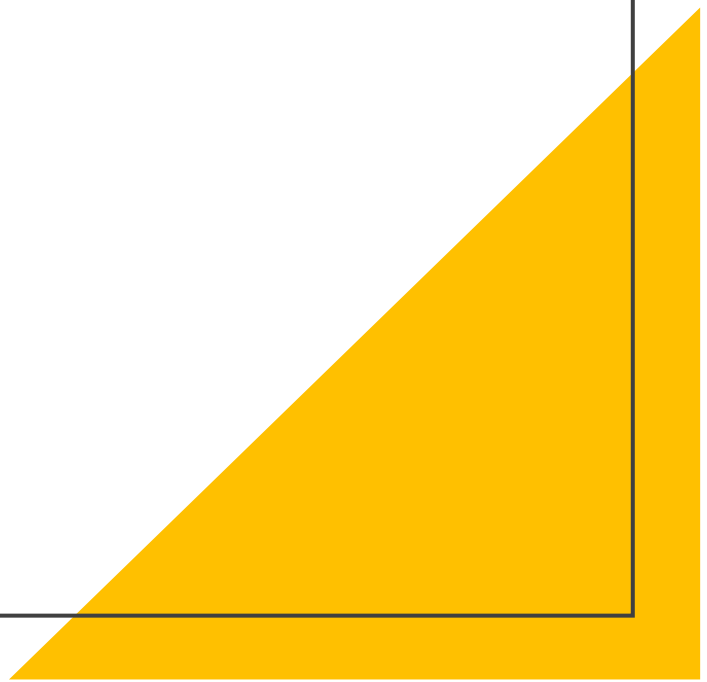
- Tables contain your data in BigQuery .
- Each table has a schema that describes the data contained in the table, including field names, types, and descriptions.
- BigQuery supports the following table types:
  - Native tables: tables backed by native BigQuery storage
  - External tables: tables backed by storage external to BigQuery
  - Views: virtual tables defined by a SQL query

Jobs

- Jobs are objects that manage asynchronous tasks such as running queries, loading data, and exporting data
  - You can run multiple jobs concurrently
  - Completed jobs are listed in the Jobs collection

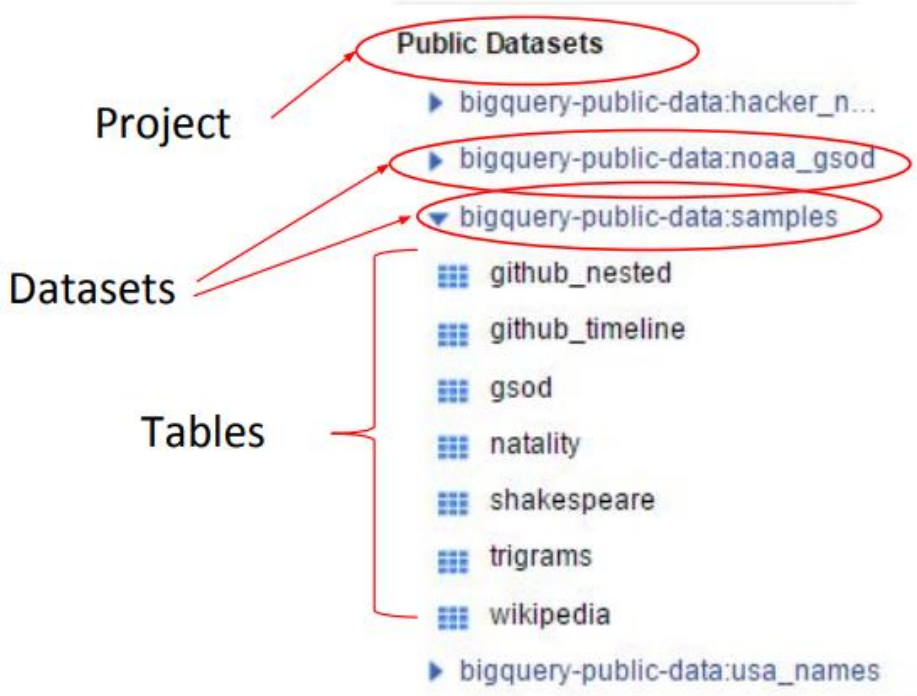
There are four types of jobs:

- Load: load data into a table
- Query: run a query against BigQuery data
- Extract: export a BigQuery table to Google Cloud Storage
- Copy: copy an existing table into another new or existing table



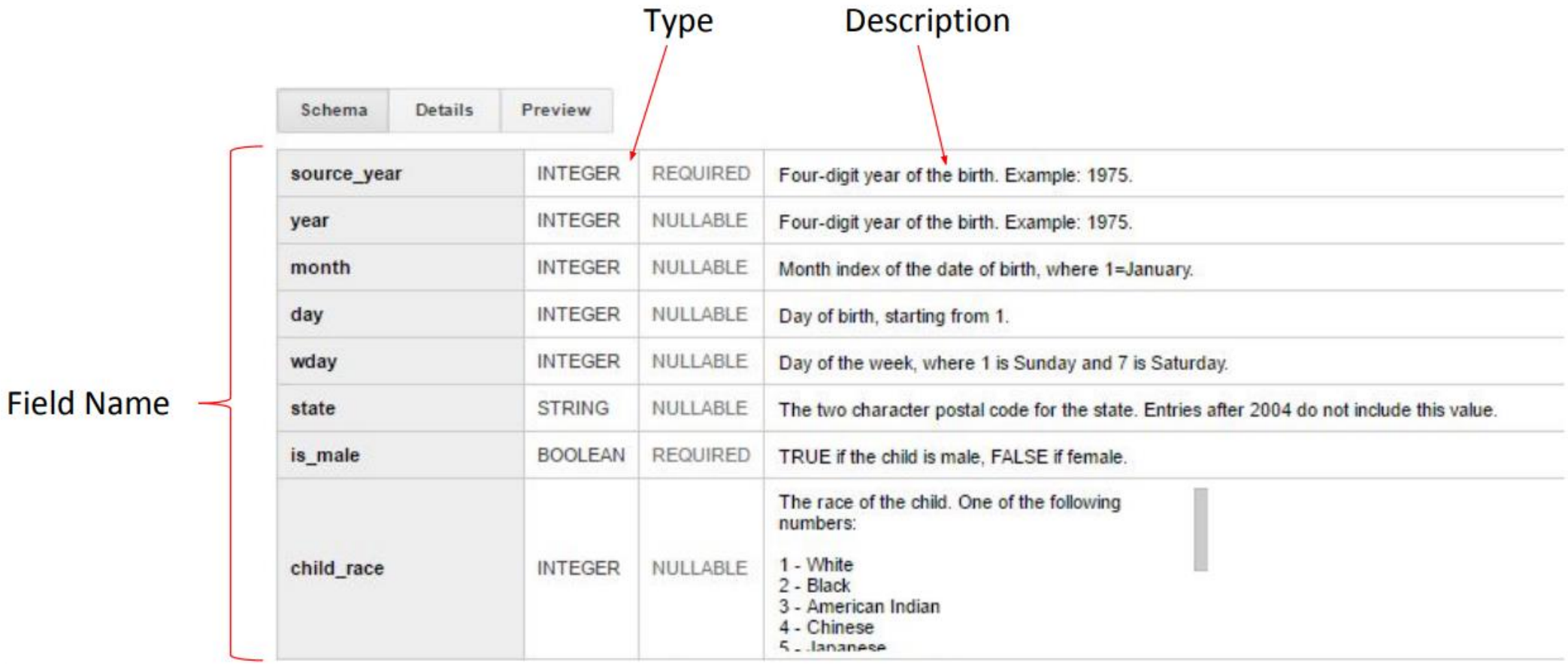
Example: BigQuery, Datasets, and Tables

- Here is an example of the left-pane navigation within BigQuery
- Projects are identified by the project name, e.g. Public Datasets, and ID, e.g. bigquery-public-data
- You can expand projects to see the corresponding datasets, e.g. samples, and tables, e.g. github_nested
- Tables are referenced by their project and dataset as: <project>:<dataset>.<table>
- e.g. bigquery-public-data:samples.natality



Example of Simple Schema

- Schema for table Natality under Sample Datasets



source_year	INTEGER	REQUIRED	Four-digit year of the birth. Example: 1975.
year	INTEGER	NULLABLE	Four-digit year of the birth. Example: 1975.
month	INTEGER	NULLABLE	Month index of the date of birth, where 1=January.
day	INTEGER	NULLABLE	Day of birth, starting from 1.
wday	INTEGER	NULLABLE	Day of the week, where 1 is Sunday and 7 is Saturday.
state	STRING	NULLABLE	The two character postal code for the state. Entries after 2004 do not include this value.
is_male	BOOLEAN	REQUIRED	TRUE if the child is male, FALSE if female.
child_race	INTEGER	NULLABLE	The race of the child. One of the following numbers: 1 - White 2 - Black 3 - American Indian 4 - Chinese 5 - Japanese

Security

- BigQuery offers built-in data protection at scale.
- It provides security and governance tools to efficiently govern data and democratize insights within your organization.
- Within BigQuery, users can assign dataset-level and project-level permissions to help govern data access.
- Secure data sharing ensures you can collaborate and operate your business with trust.
- Data is automatically encrypted both while in transit and at rest, ensuring that your data is protected from intrusions, theft, and attacks.
- Cloud DLP helps you discover and classify sensitive data assets.
- Cloud IAM provides access control and visibility into security policies.
- Data Catalog helps you discover and manage data.

BigQuery

#GCPSketchnote

@PVERGADIA THECLOUDGIRL.DEV

6.30.2021

What is BigQuery?

ENTERPRISE DATA WAREHOUSE

INGEST → STORE → ANALYZE → VISUALIZE

BATCH DATA

STREAMING DATA

99.99% SLA

Google will allocate as many servers as needed to run your query

ON DEMAND INSIGHTS

REAL-TIME INSIGHTS

Unique BigQuery FEATURES

BIGQUERY OMNI (Private Alpha)
- Enables hybrid & multi-cloud queries

Data Storage

Query Engine

Anthos Clusters

BIGQUERY ML
- Create model
- Predict

BI ENGINE

BigQuery BI Engine

Data Studio

Sheets

Partner BI Tools

DATA QnA (Private Alpha)
- Get answers through NLP

Business User

Chat

Data QnA

Get answers through NLP

GEOSPATIAL

CONNECTED SHEETS

BigQuery PRICING

FREE!

Free storage for all public datasets

Query upto 1TB/month for free

BigQuery Sandbox
No credit card required

Batch Load is free

Streaming Inserts

Export is free

Storage

Query

ACTIVE
<90 days

LONG TERM
>90 days

ON DEMAND
$ per query

FLAT RATE
$ for dedicated resources

SECURITY for BigQuery

0 Avoid risk of data loss

1 Identity and Access Management

2 VPC controls

3 Customer managed encryption key (CMEK)

How it WORKS!

GCS, HTTPs Post	Web, IoT, Mobile	ETL, ELT data	GCS, s3, Teradata, ...
BATCH INGESTION	STREAMING INGESTION	QUERY MATERIALIZATION	DATA TRANSFER SERVICE (DTS)

INGEST

AUTOMATIC BACKUP & RESTORE

PUBLIC DATA SETS

ODBC and JDBC drivers (at no cost)

SQL ANSI:2011 compliant

COMPUTE

QUERY

Visualize

STORAGE

Looker

Data Studio

Connected Sheets

Partner BI Tools

BigQuery Use case example

Data Analytics

Trigger	Ingest	Enrich	Analyze	Visualize
Google Cloud Storage: Mobile, Web, Data Store, IoT	Batch Load, Pub/Sub, Streaming Inserts	Dataflow	BigQuery, Vertex AI, Bigtable	Looker, DataStudio, Third-party BI, Google Sheets

BigQuery’s Columnar Database

- Google BigQuery Architecture uses column-based storage or columnar storage structure that helps it achieve faster query processing with fewer resources.

- It is the main reason why Google BigQuery handles large datasets quantities and delivers excellent speed.

MapReduce	BigQuery
1. High Latency.	1. Low Latency.
2. Flexible (complex) batch processing.	2. SQL-like Queries.
3. Unstructured Data.	3. Structured Data.

MapReduce vs Google BigQuery

Data Mesh Architecture with Google Cloud BigQuery

- Data Mesh implementations rely on Google Cloud Platform (GCP) as a common infrastructure, at least for analytical data.
- It is a highly integrated platform.
- Everything is available as developer-friendly self-service.
- The on-demand query performance of BigQuery is remarkable, especially for large data sets.
- BigQuery is the central component for storing analytical data.
- BigQuery is a columnar data store and can perform efficient JOIN operations with large data sets.
- It supports access to files stored in Google Cloud Storage as external tables

Data Mesh Architecture with Google Cloud BigQuery

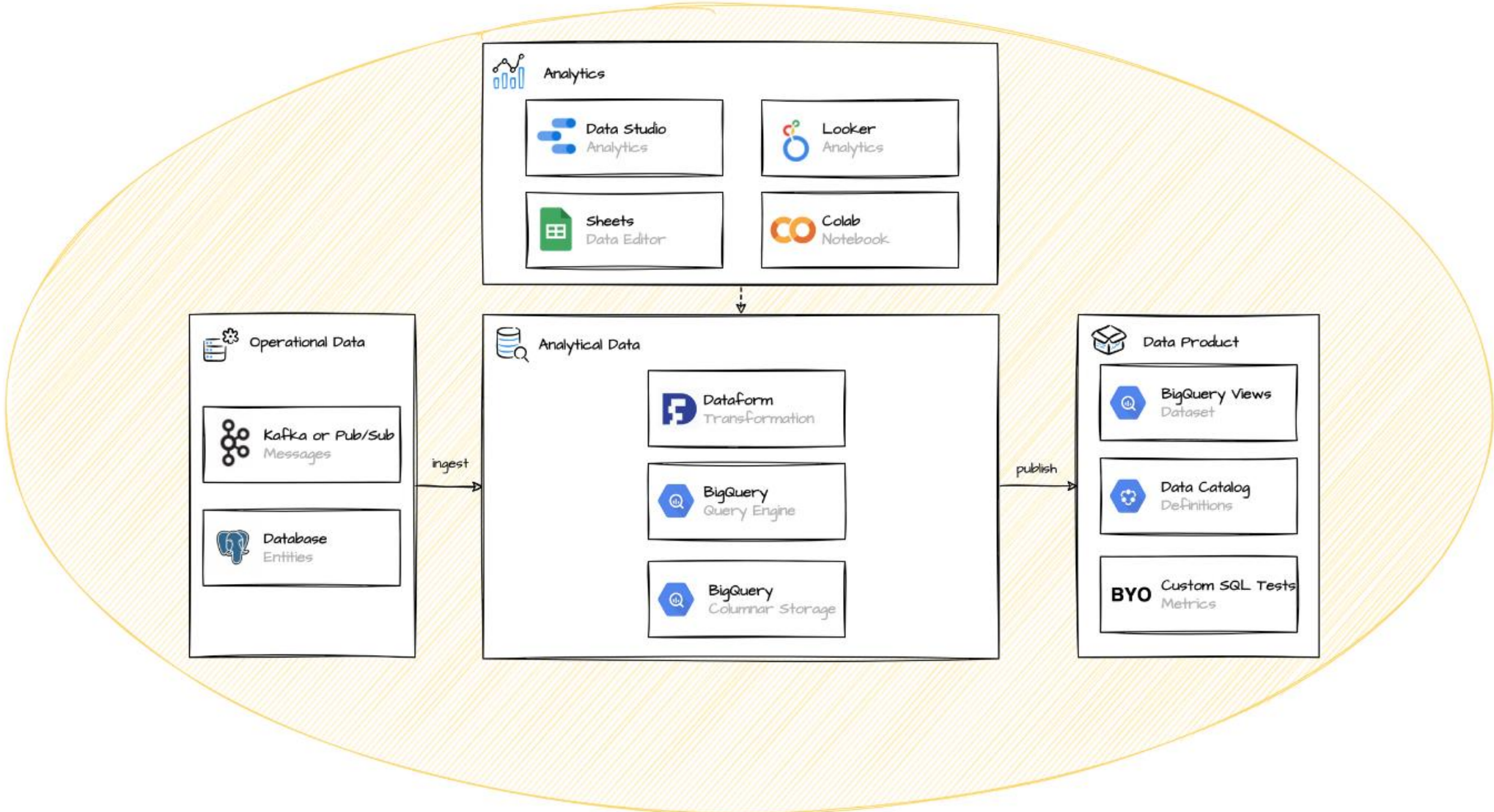


Image source:datamesh-architecture

Reference

- https://cloud.google.com/blog/topics/developers-practitioners/query-big-bigquery-cheat-sheet
- https://hevodata.com/blog/google-bigquery-data-warehouse/
- https://cloud.google.com/bigquery/docs/introduction