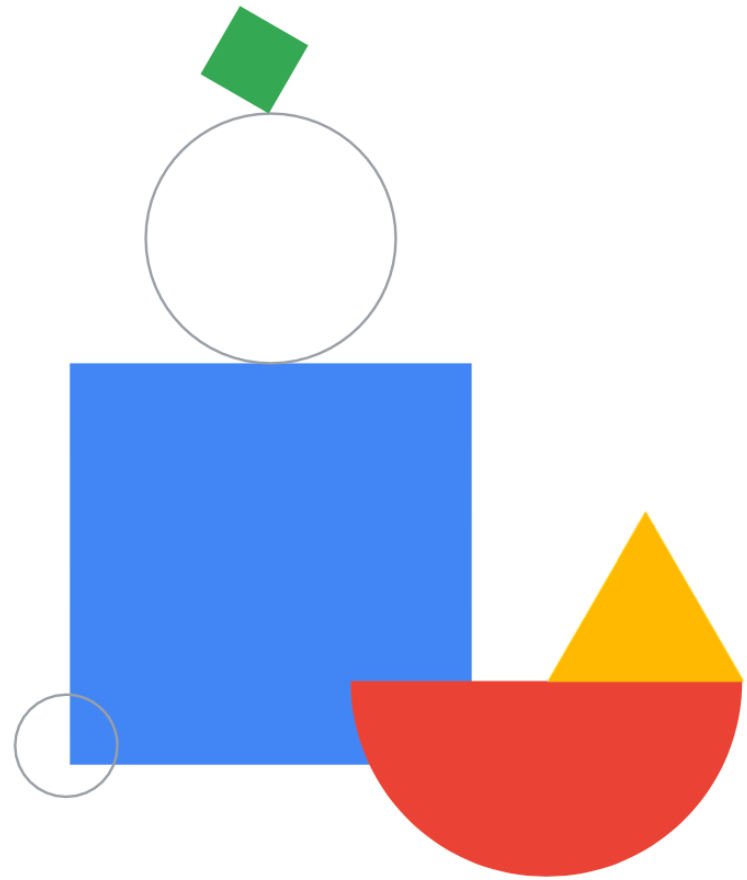
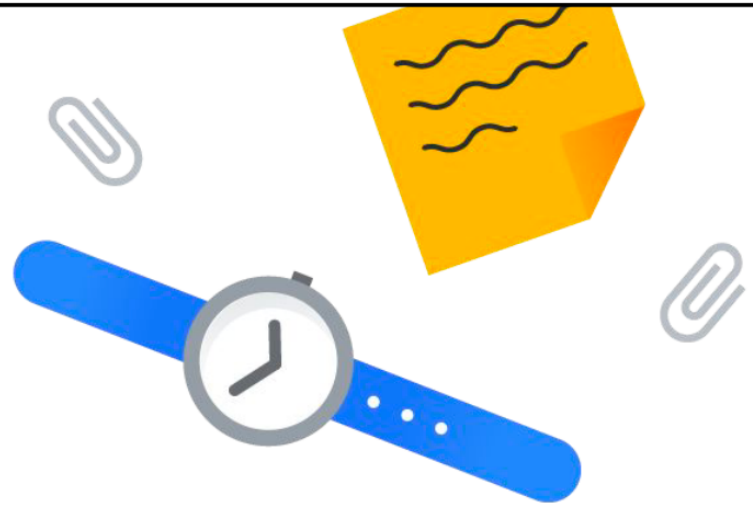
Google Cloud

Introduction to Building Batch Data Pipelines



Proprietary + Confidential



Module agenda

01 EL, ELT, ETL

02 Quality Considerations

03 How to Carry out Operations in BigQuery

04 Shortcomings

05 ETL to Solve Data Quality Issues

Proprietary + Confidential

EL, ELT, ETL

Proprietary + Confidential

The method you use to load data depends on how much transformation is needed

EL

Extract and Load

ELT

Extract, Load, and Transform

ETL

Extract, Transform, and Load

Proprietary + Confidential

When would you use EL?

Architecture

Extract data from files on Cloud Storage

Load it into BigQuery's native storage

You can trigger this from Cloud Composer, Cloud Functions, or scheduled queries

When you'd do it

Batch load of historical data

Scheduled periodic loads of log files (e.g. once a day)

But only if the data is already clean and correct!

Proprietary + Confidential

When would you use ELT?

Architecture

Extract data from files in Cloud Storage into BigQuery.

Transform the data on the fly using BigQuery views, or store into new tables.

When you'd do it

Experimental datasets where you are not yet sure what kinds of transformations are needed to make the data useable.

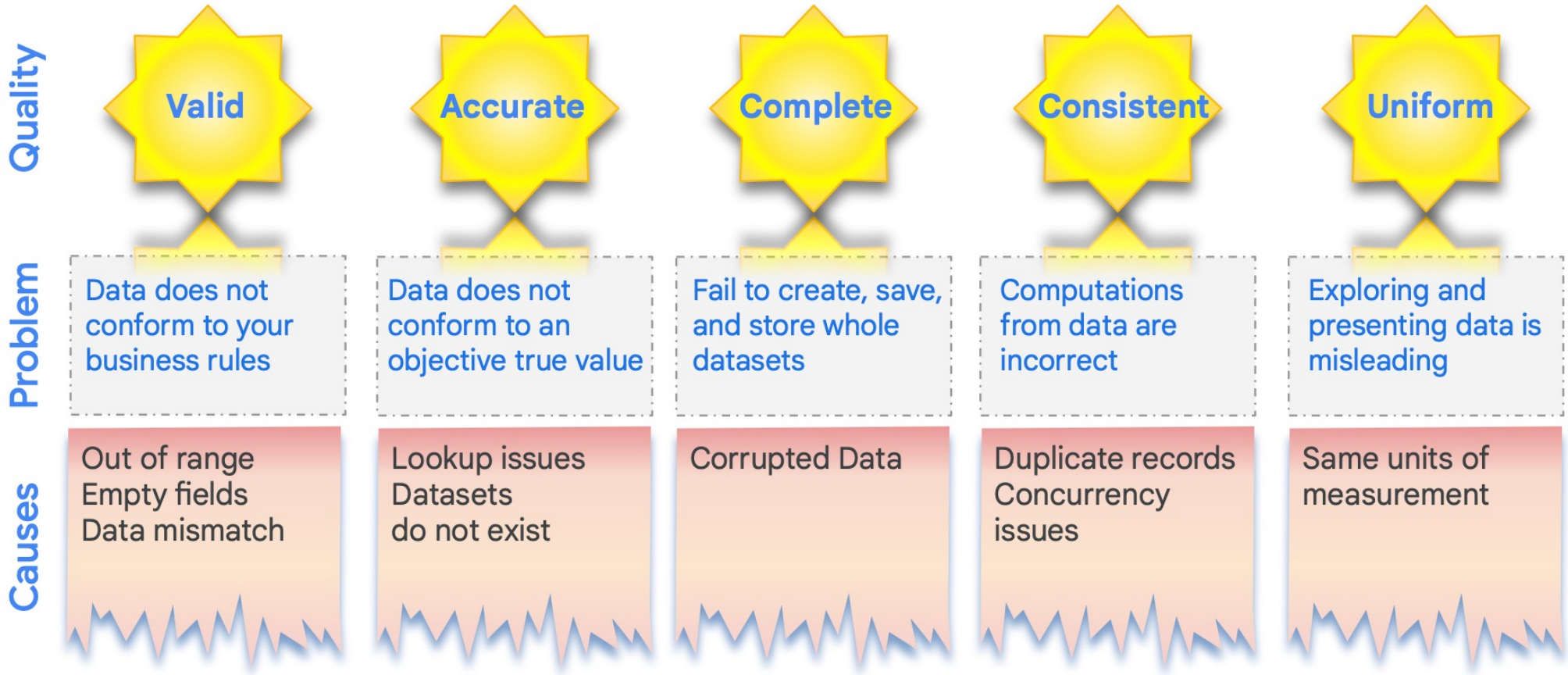
Any production dataset where the transformation can be expressed in SQL.

Proprietary + Confidential

Quality Considerations

Proprietary + Confidential

What are the purposes of Data Quality processing?



Proprietary + Confidential

BigQuery can fix many data quality issues using SQL and Views

Proprietary + Confidential

How to Carry out Operations in BigQuery

Proprietary + Confidential

Filter to identify and isolate invalid data

Setup Field Data Type Constraints

Specify fields as NULLABLE or REQUIRED

```
SQL: NULLABLE or REQUIRED
```

Proactively check for NULL values

```
SQL: NULL
```

Check and Filter for Allowable Range values SQL Conditionals:

```
SQL: CASE WHEN, IF ( )
```

Require Primary Keys / Relational Constraints in upstream source systems (remember, BigQuery is an analytics warehouse not your primary operational database)

Filter rows

```
WHERE(condition)
```

Filter aggregations

```
HAVING(condition)
```

Filters NULLs but leave blanks

```
WHERE field IS NOT NULL
```

Filters NULLs and blanks

```
WHERE field IS NOT NULL AND field <> “”
```

A NULL is the absence of data. A BLANK is a value of data.
Consider if you are trying to filter out both NULLS and BLANKS.

Proprietary + Confidential

Detect duplication, enforce uniqueness for consistency

Store one fact in one location and use IDs for lookup

Use String Functions to clean data:

```
PARSE_DATE( )
SUBSTR( )
REPLACE( )
```

A difference means there are duplicates

```
COUNT(DISTINCT field)
```

```
COUNT(field)
```

>1 indicates duplicates

```
COUNT(field)
GROUP BY(field)
```

Proprietary + Confidential

Test data against known good values for accuracy

Create test cases or calculated fields to check values

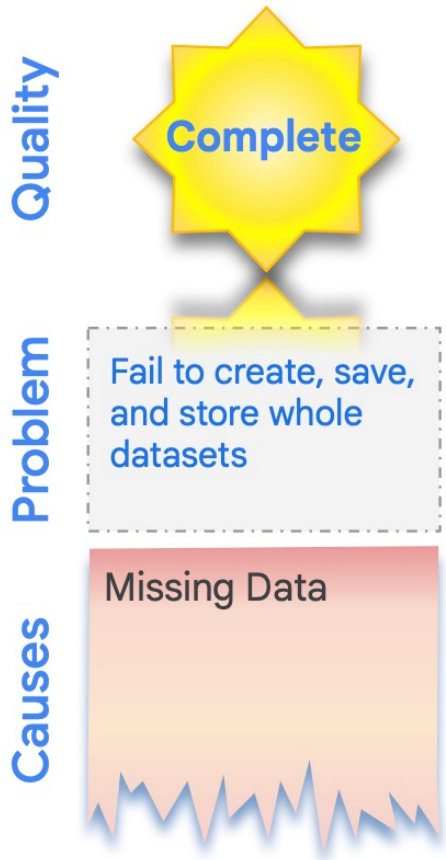
```
SQL: (quantity_ordered * item_price) AS sub_total
```

Lookup values against an objective reference dataset

```
SQL: IN( ) with a subquery or JOIN
```

Proprietary + Confidential

Identify and fill in missing values for completeness



Thoroughly explore the existing dataset shape and skew and look for missing values

```
SQL: NULLIF( ) , IFNULL( ),  COALESCE( )
```

Enrich the existing dataset with others using UNIONs and JOINs

```
SQL: UNION, JOIN
```

Example: Multiple years of historical data are available for analysis

Verify file integrity with checksum values (hash, MD5)

The automatic process of detecting data drops and requesting data items to fill in the gaps is called "backfilling". It is a feature of some data transfer services.

Proprietary + Confidential

Make data types and formats explicit for uniformity

Document and comment your approach

Use FORMAT ( ) to clearly indicate units

```
SQL: FORMAT( )
```

CAST( ) data types to the same format and digits

```
SQL: CAST( )
```

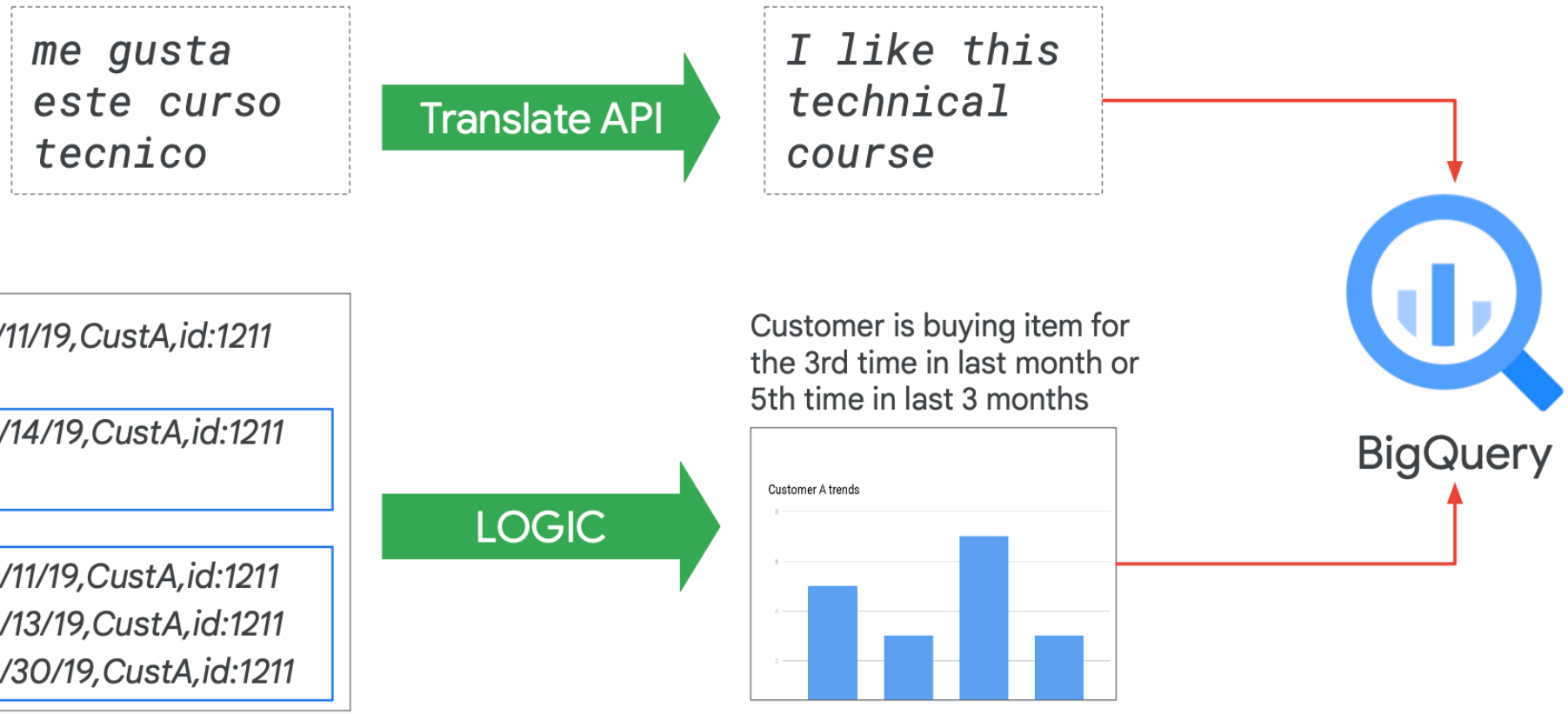
Label all visualizations appropriately

Proprietary + Confidential

Shortcomings

Proprietary + Confidential

What if the transformations cannot be expressed in SQL? Or are too complex to do in SQL?



Proprietary + Confidential

Build ETL pipelines in Dataflow and land the data in BigQuery

Architecture

Extract data from Pub/Sub, Cloud Storage, Cloud Spanner, Cloud SQL, etc.

Transform the data using Dataflow.

Have Dataflow pipeline write to BigQuery.

When you'd do it

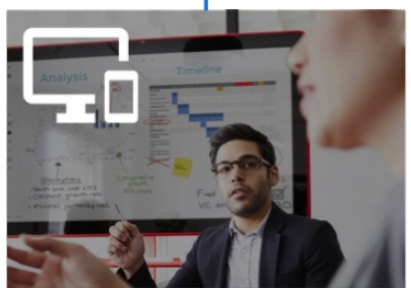
When the raw data needs to be quality-controlled, transformed, or enriched before being loaded into BigQuery.

When the data loading has to happen continuously, i.e. if the use case requires streaming.

When you want to integrate with continuous integration / continuous delivery (CI/CD) systems and perform unit testing on all components.

Proprietary + Confidential

Google Cloud offers a range of ETL tools

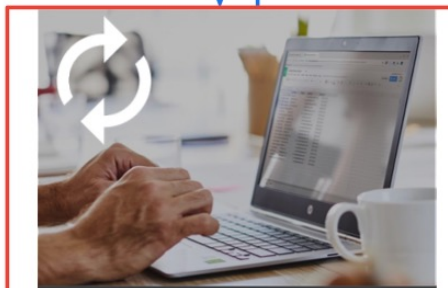


Data ingestion at any scale

Pub/Sub

Data Transfer Services

Cloud IoT Core

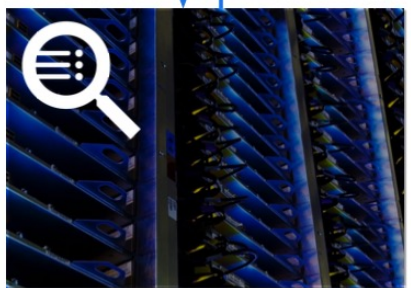


Data transformation and streaming pipelines

Dataflow

Dataproc

Cloud Data Fusion

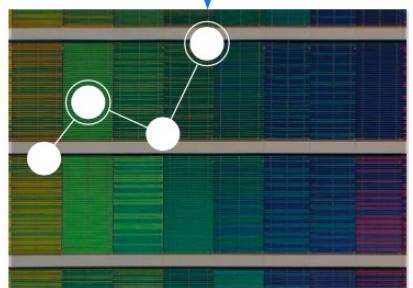


Data warehousing and data lake

BigQuery

Cloud Storage

Cloud BigTable



Advanced analytics

Vertex AI

Google Data Studio

Notebooks

Looker

BI tools from Technology Partners

Google Sheets

Orchestration

Cloud Composer

Proprietary + Confidential

ETL to Solve Data Quality Issues

Proprietary + Confidential

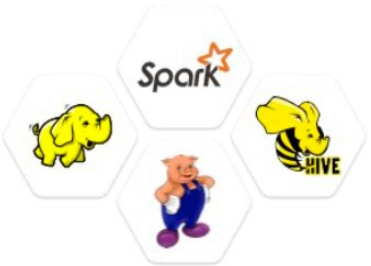
Cases when you look beyond Dataflow and BigQuery

Issue	Solution
Latency, throughput	Dataflow to Bigtable
Reusing Spark pipelines	Dataproc
Need for visual pipeline building	Cloud Data Fusion

Proprietary + Confidential

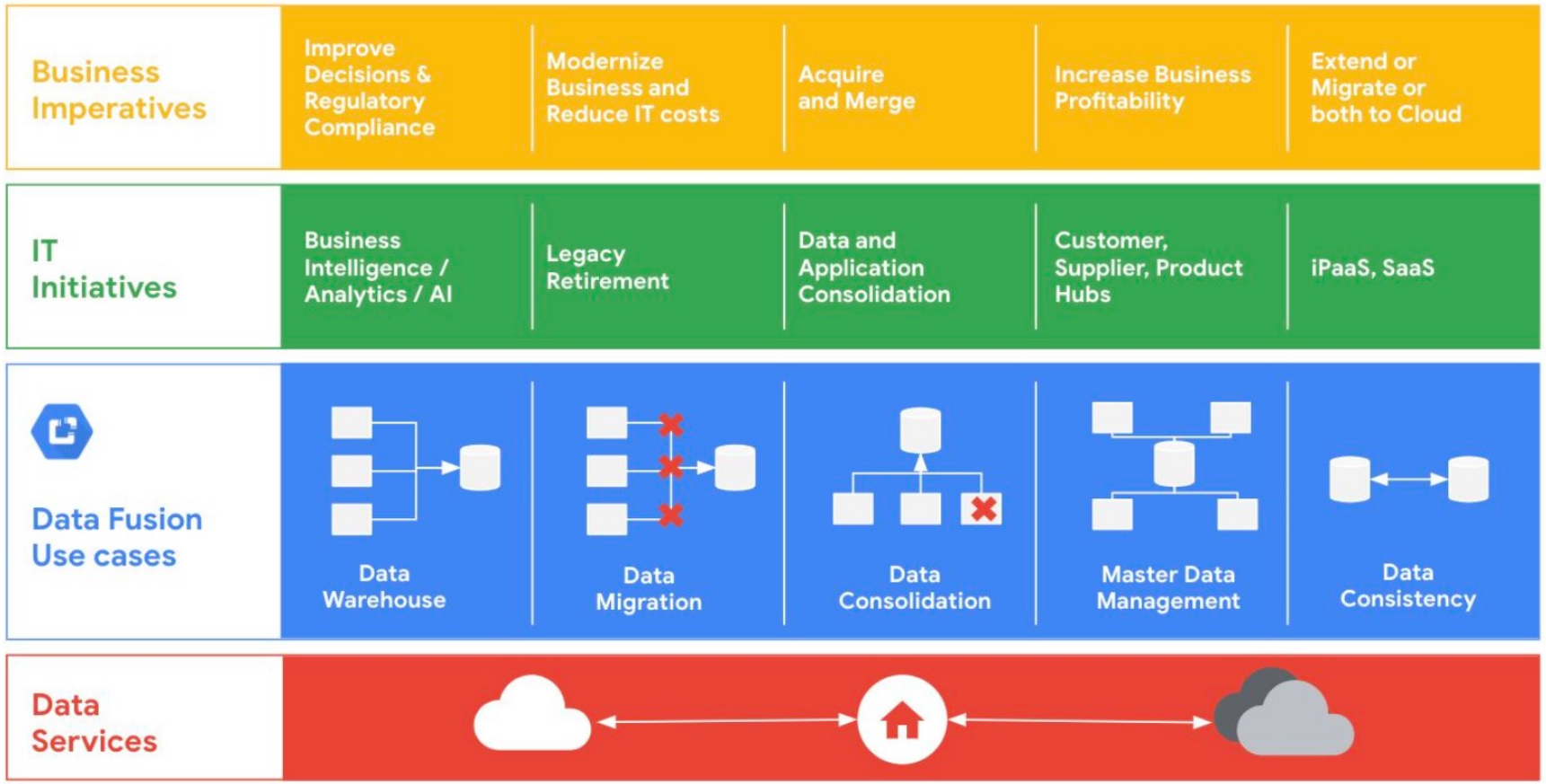
Dataproc is a managed service for batch processing, querying, streaming, and ML

Google Cloud



Proprietary + Confidential

Build and manage data pipelines quickly with Cloud Data Fusion



Proprietary + Confidential

Tracking lineage in ETL pipelines can be important

Discovery: Find the data you need

Where it came from

The processes it has been through

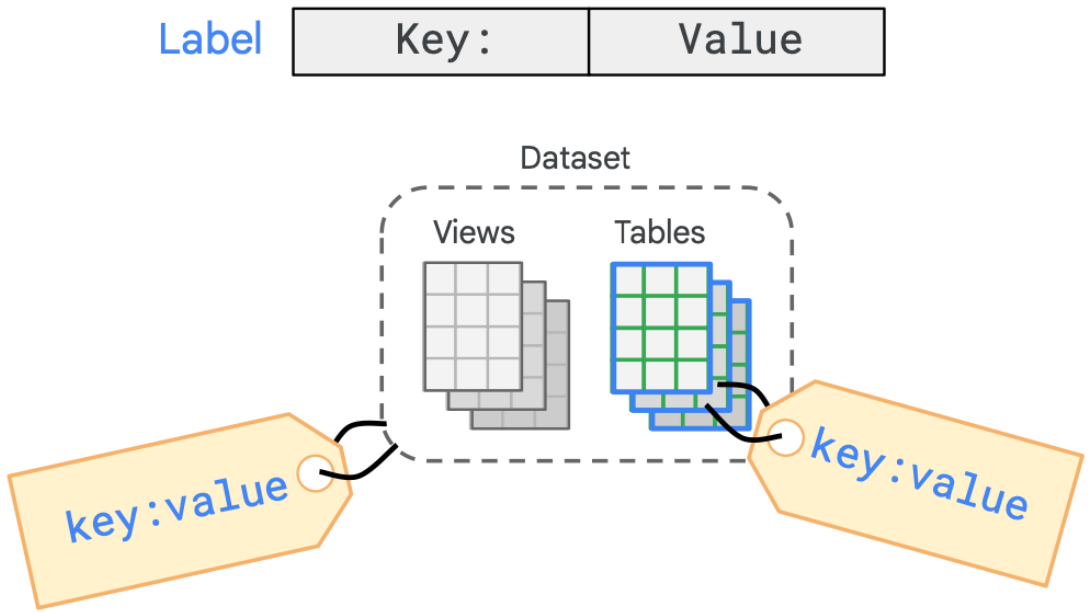
Its present location and condition

Lineage: Metadata about the data

- What format is it in?
- What qualities does it have?
- Is it fit for the intended use?
- Can you transform or process it to make it fit for the intended use?

Proprietary + Confidential

Labels on datasets, tables, and views can help track lineage



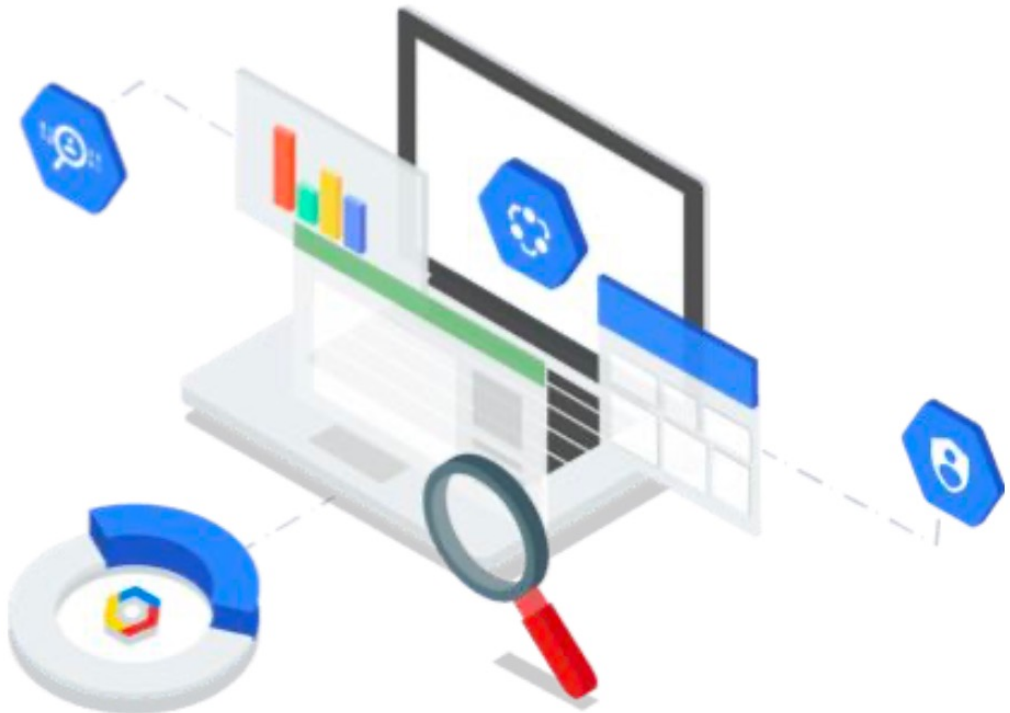
Example

A series of similar tables:

Salesdata: Europe
Salesdata: March
Salesdata: Repeat customers

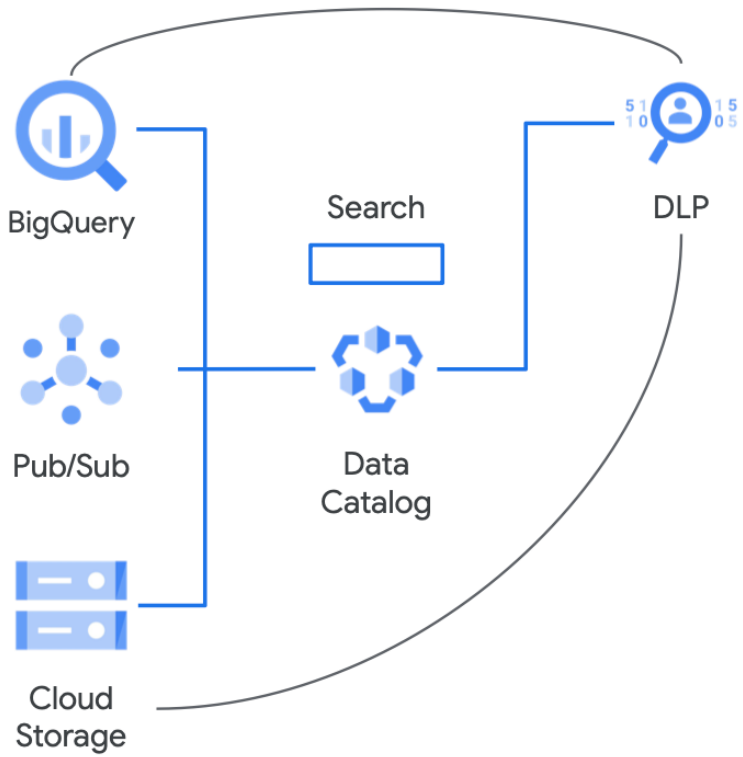
Proprietary + Confidential

View your datasets and labels in Data Catalog



Proprietary + Confidential

Data Catalog for managed data discovery



Data Catalog

- **Simplify data discovery at any scale:**
  Fully managed metadata management service with no infrastructure to set up or manage.
- **Unified view of all datasets:**
  Central and secure data catalog across Google Cloud with metadata capture and tagging.
- **Data governance foundation:**
  Security compliance with access level controls along with Cloud Data Loss Prevention integration for handling sensitive data.