# Big Data

# Agenda

- Big data
- Big data process
- Big data management
- Summary

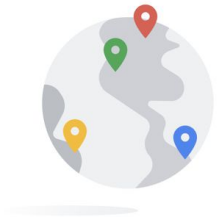
# The six Vs of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: *volume*, *variety and velocity*. Over time, other Vs have been added to descriptions of big data:

| VOLUME | VARIETY | VELOCITY | VERACITY | VALUE | VARIABILITY |
|---|---|---|---|---|---|
| The amount of data from myriad sources. | The types of data: structured, semi-structured, unstructured. | The speed at which big data is generated. | The degree to which big data can be trusted. | The business value of the data collected. | The ways in which the big data can be used and formatted. |

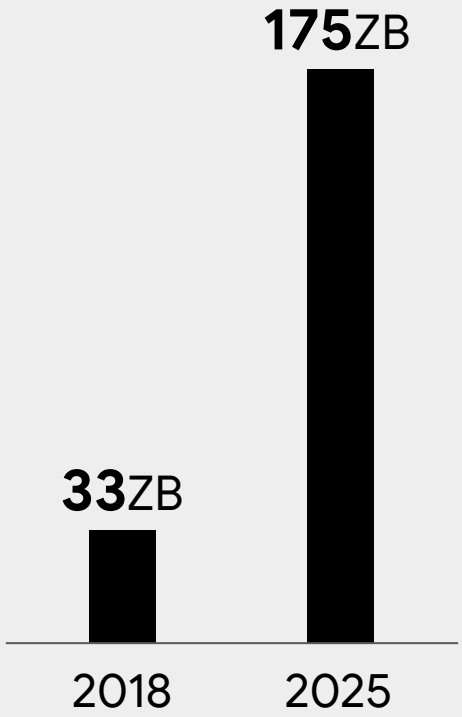




# The world is generating more data than ever

By 2025, the world datasphere will be **175 zettabytes**.



# Data analytics remains untapped

69%

69% of companies report that they have not created a data-driven organization

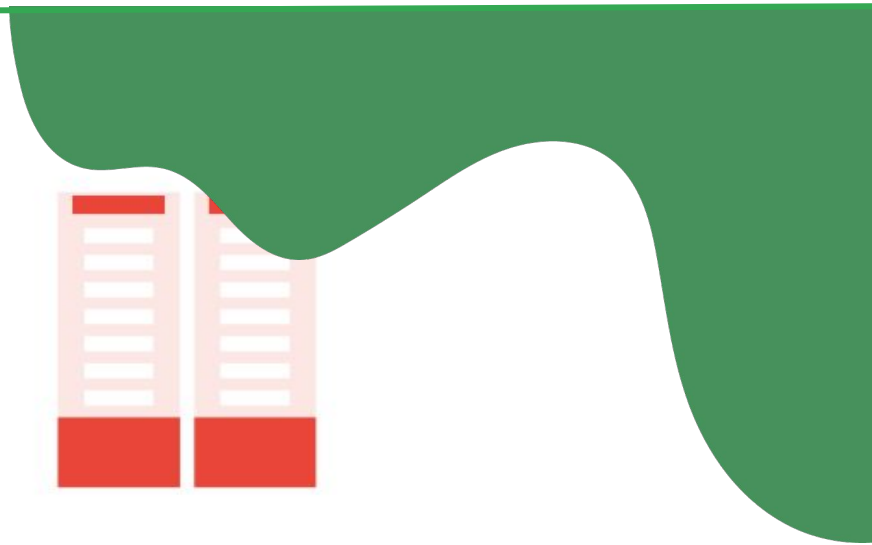
71%

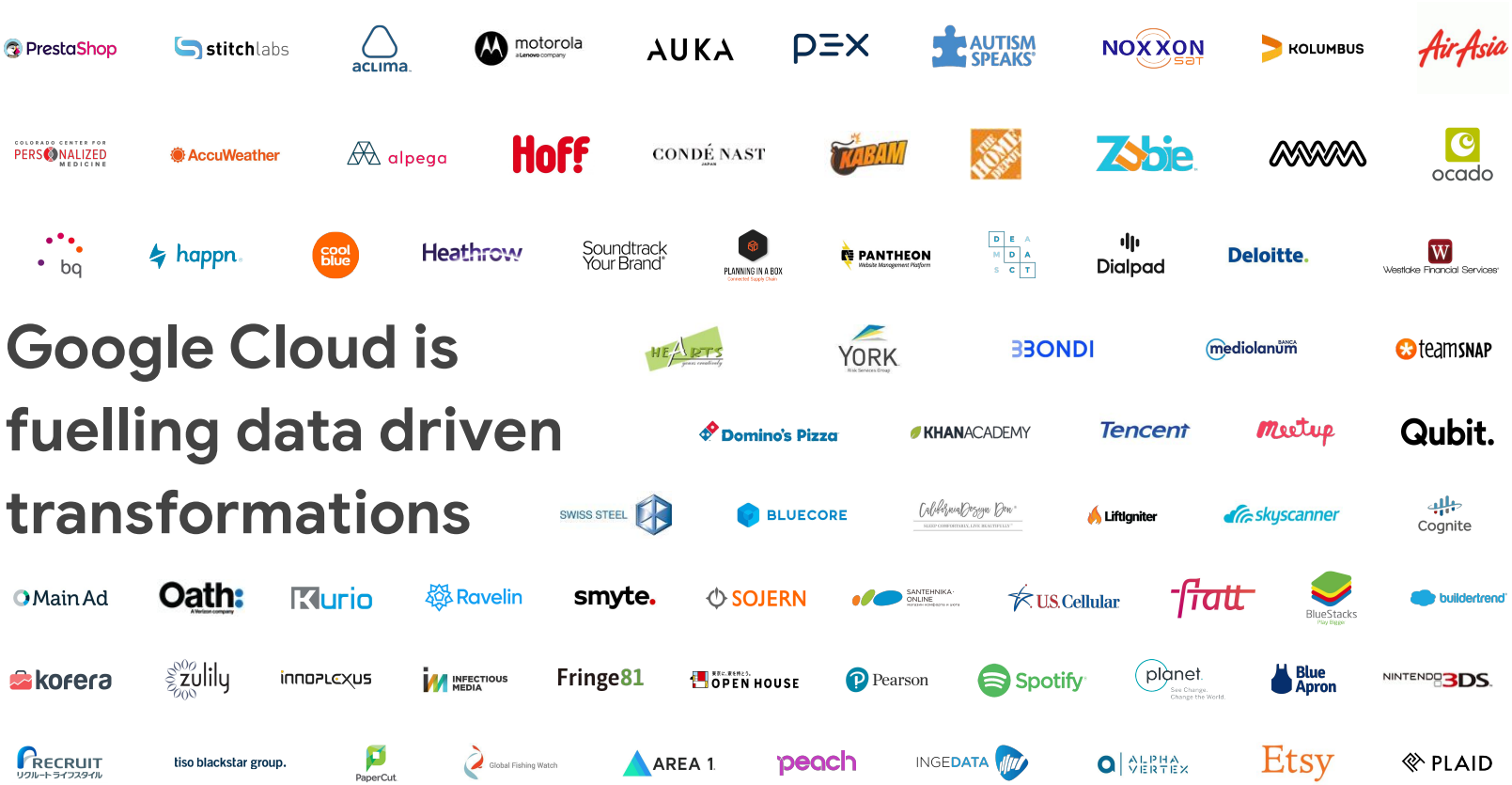
And 71% report that they have yet to forge a data culture

*HBR, New Vantage Partners Feb 2019

# And the time is ripe...

...as traditional data warehouses melt under data growth






# Google Cloud is fuelling data driven transformations

# Delivers a double-digit uplift in revenue per session

## With help from Google Cloud

- ★ Provides personalized recommendations to customers, improving engagement and experience
- ★ Supports Hanes Australasia's growth as a leading ecommerce retailer in Australia and New Zealand
- ★ Enables the business to obtain real commercial value from data

HANES Brands Inc

"The product is extremely easy to use—Google Cloud has provided the expertise, functionality, and performance, so we do not need to be data scientists to make the most of it."

—Peter Luu, Online Analytics Manager, Hanes Australasia

# Manages up to 5TB of data per day

## With help from Google Cloud

- ★ Supports 1 million motorcycle drivers with rapid access to riders and optimized routes
- ★ Enables demand forecasting and pricing adjustments
- ★ Positions business for international expansion


“Google Maps Platform now resides at the core of our engine to help us figure out optimized routes and estimated times of arrival for our drivers.”

—Ajey Gore, Group Chief Technology Officer, GO-JEK

# Saves 1,000 man-hours by switching to automated tagging

## With help from Google Cloud

- ★ Empowers staff to work more productively and easily locate the photos they need
- ★ Enables a wider pool of contributors by automating tagging and labeling tasks
- ★ Grows engagement among Singaporeans thanks to a faster photo rollout

Sport SINGAPORE

"Before Google Cloud, we used a static hard disk system, so people had to physically come to the office to access files. Because of the speed at which content is created and updated, we want a better way to get fellow Singaporeans to work together with us in content creation."

—Darren Ho, Team Lead for Digital Operations, Sport Singapore

# Big Data Process

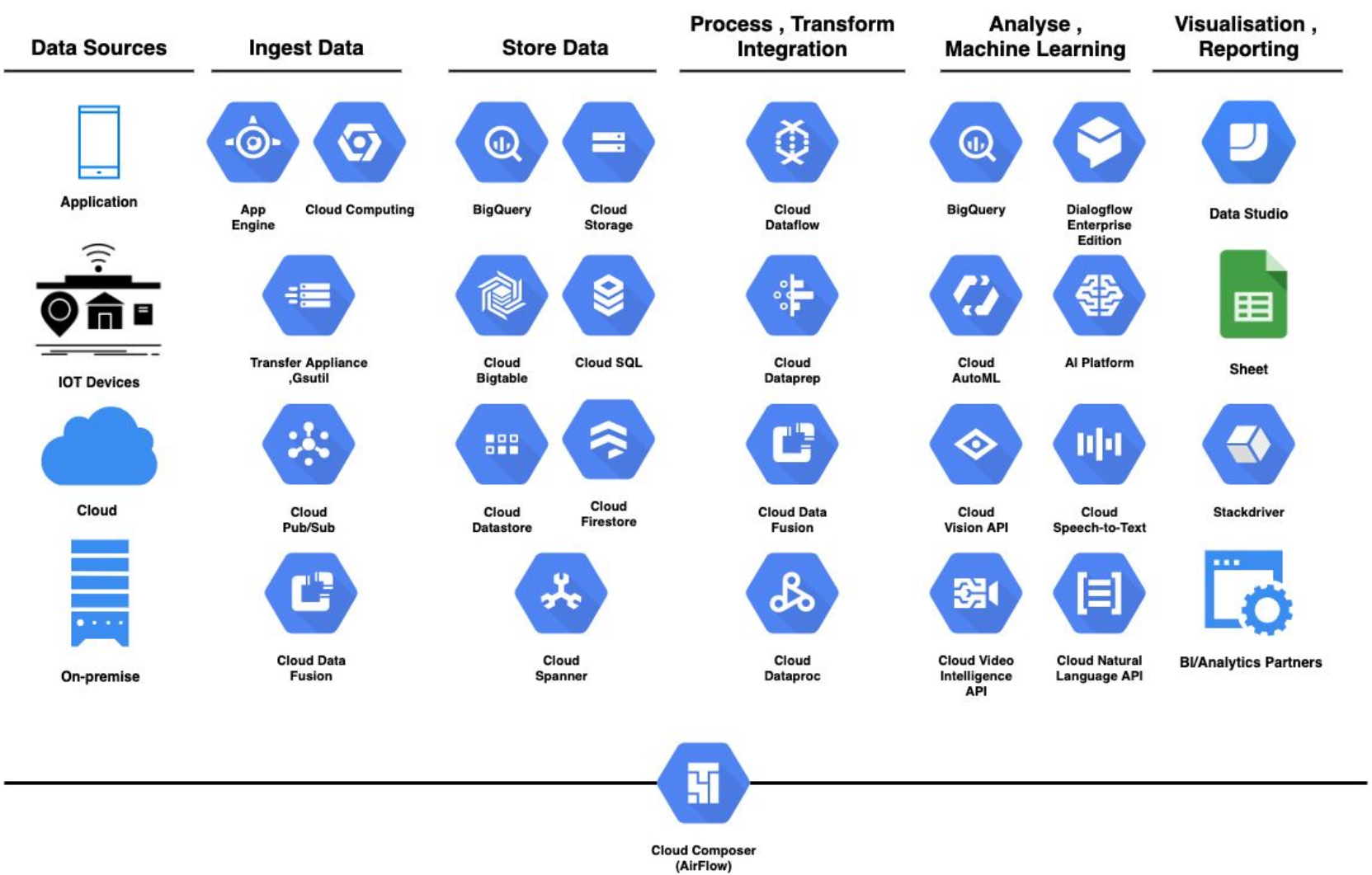
# Simplify Data Analytics Process


- Data ingestion
- IoT

- Data lake
- Database

- Data pipeline
- Data transformation
- Data warehouse
- Data mart

- Data analytics
- Visualization
- Machine Learning


Data Sources
Ingest Data
Store Data
Process , Transform Integration
Analyse , Machine Learning
Visualisation , Reporting
Application
App Engine
Cloud Computing
BigQuery
Cloud Storage
Cloud Dataflow
BigQuery
Dialogflow Enterprise Edition
Data Studio
IOT Devices
Transfer Appliance ,Gsutil
Cloud Bigtable
Cloud SQL
Cloud Dataprep
Cloud AutoML
AI Platform
Sheet
Cloud
Cloud Pub/Sub
Cloud Datastore
Cloud Firestore
Cloud Data Fusion
Cloud Vision API
Cloud Speech-to-Text
Stackdriver
On-premise
Cloud Data Fusion
Cloud Spanner
Cloud Dataproc
Cloud Video Intelligence API
Cloud Natural Language API
BI/Analytics Partners
Cloud Composer (AirFlow)

# EL / ELT / ETL

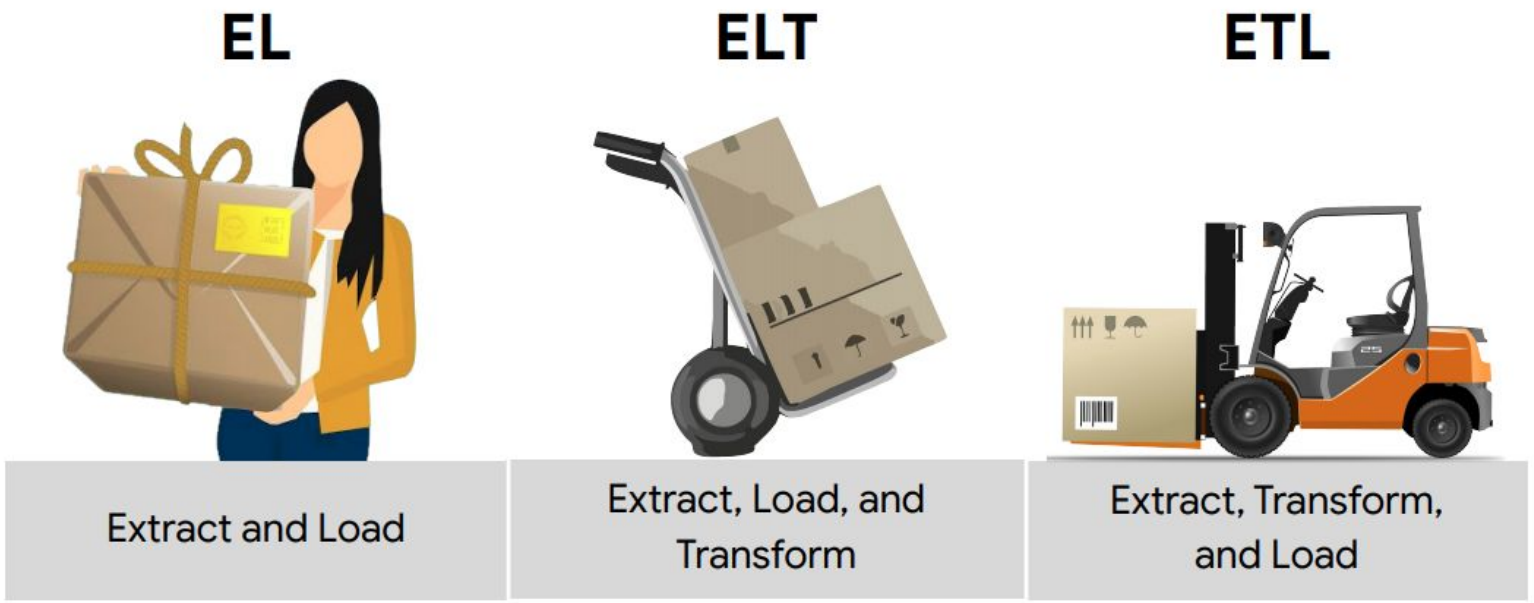
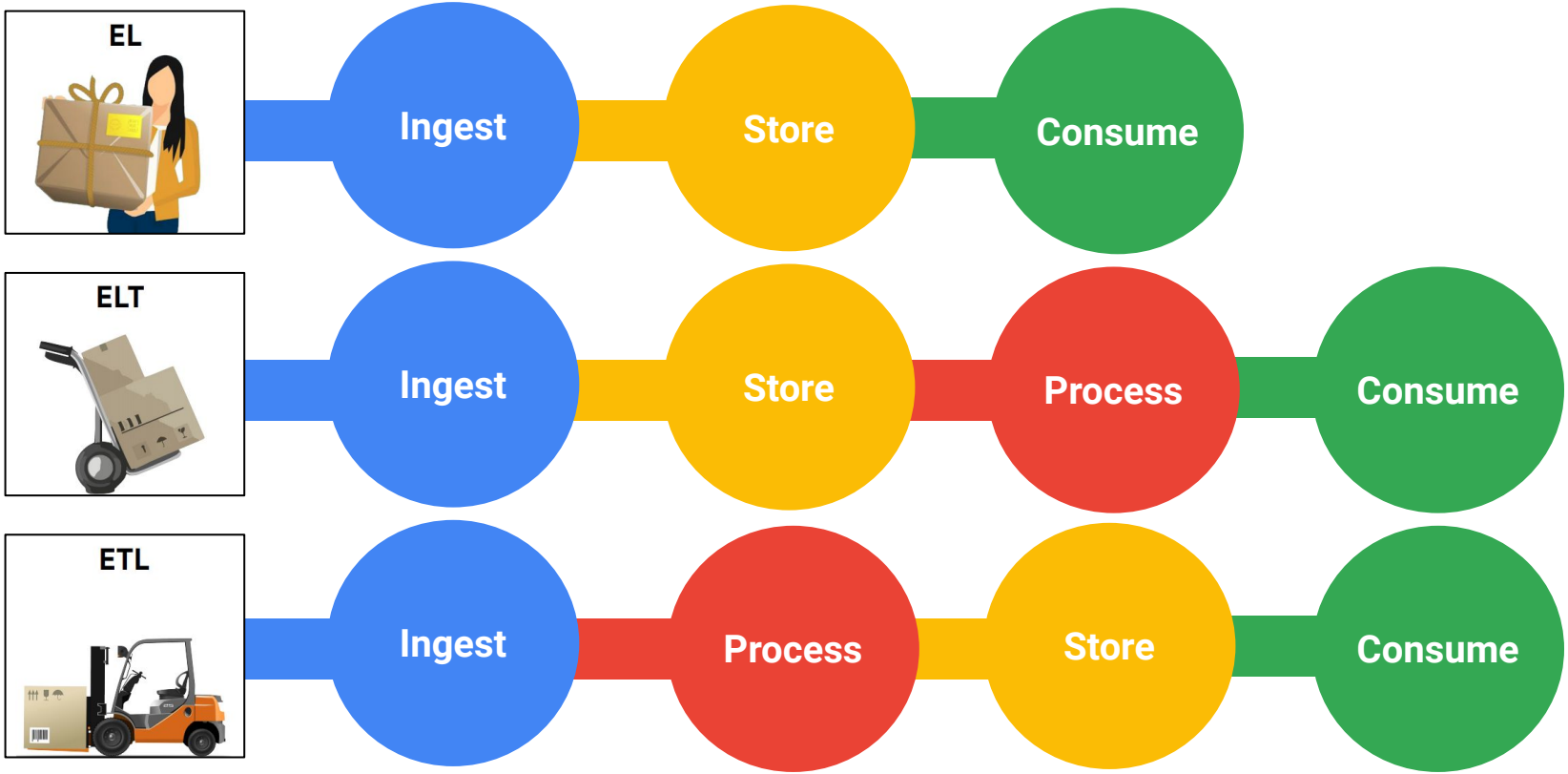
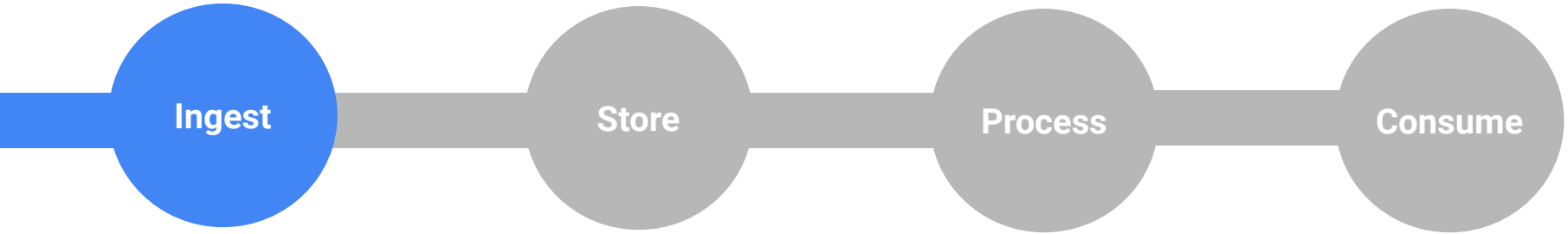



**Image Ref** : Data Engineering course, Coursera

# EL / ELT / ETL


EL
Ingest
Store
Consume
ELT
Ingest
Store
Process
Consume
ETL
Ingest
Process
Store
Consume

# Simplify Data Analytics Process




- Data ingestion
- IoT

- Data lake
- Database

- Data pipeline
- Data transformation
- Data warehouse
- Data mart

- Data analytics
- Visualization
- Machine Learning

# Data ingestion

## Batch

การประมวลผลแบบ Batch คือ การประมวลผลของกลุ่มข้อมูลที่ถูกเก็บไว้ในช่วงระยะเวลาหนึ่ง เช่น การประมวลผลในรอบ สัปดาห์ หรือรอบหนึ่งเดือน เหมาะสำหรับข้อมูลที่มีขนาดใหญ่

## Streaming

การประมวลผลแบบ Stream คือ การประมวลผลแบบเรียลไทม์ เช่น การเก็บค่าของอุปกรณ์ IOT, GPS เป็นต้น

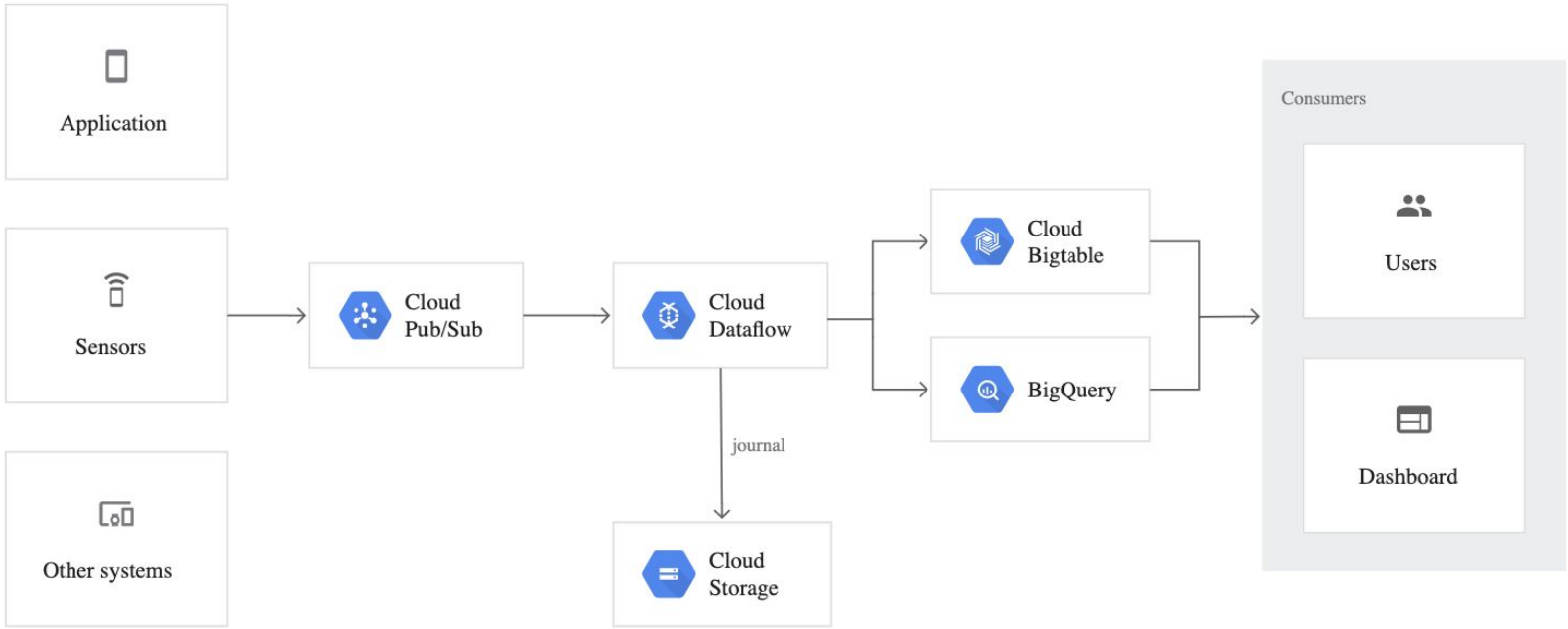
Cloud Pub/Sub

Data Transfer Service

Cloud IoT Core

Storage Transfer Service

# Real-time analytics Iot Device



# Simplify Data Analytics Process


- Data ingestion
- IoT

- Data lake
- Database

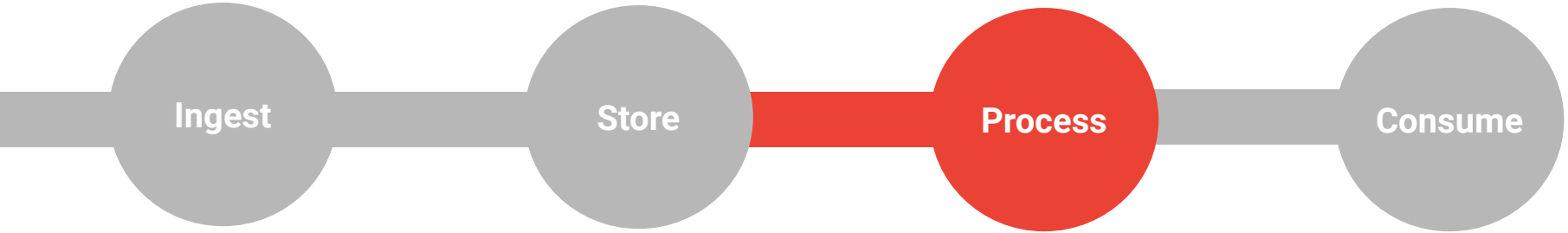
- Data pipeline
- Data transformation
- Data warehouse
- Data mart

- Data analytics
- Visualization
- Machine Learning

# Storage


In Memory
Relational
NoSQL
Warehouse
Object
Block
File
Cloud Memorystore
Cloud SQL
Cloud Spanner
Cloud Datastore, Firestore
Cloud Bigtable
BigQuery
Cloud Storage
Persistent Disk
Cloud Filestore
Good for: In memory datastore
Good for: Relational database service
Good for: Scalable relational database
Good for: Serverless NoSQL document
Good for: NoSQL key-value and wide-column
Good for: Enterprise DW
Good for: Unstructured data, objects or blobs
Good for: Local VM file storage
Good for: Lift/shift apps requiring file

# Simplify Data Analytics Process




- Data ingestion
- IoT

- Data lake
- Database

- Data pipeline
- Data transformation
- Data warehouse
- Data mart

- Data analytics
- Visualization
- Machine Learning

# Data Pipeline


Data pipeline is a set of data process connected in series which ingest raw data from disparate sources and move the data to a destination for storage and analysis.

# Data transformations

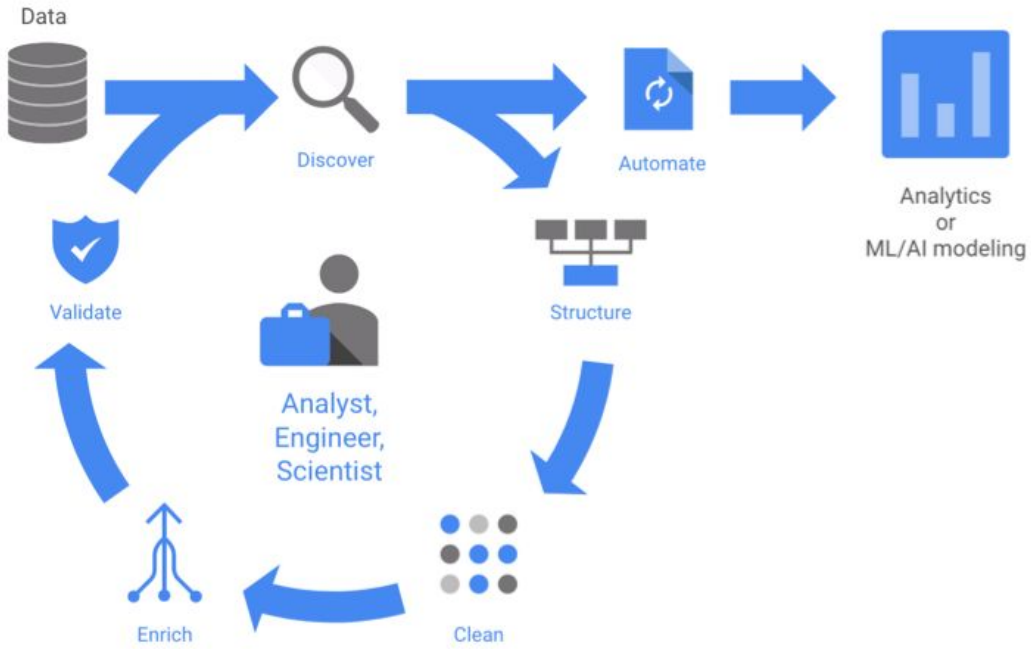
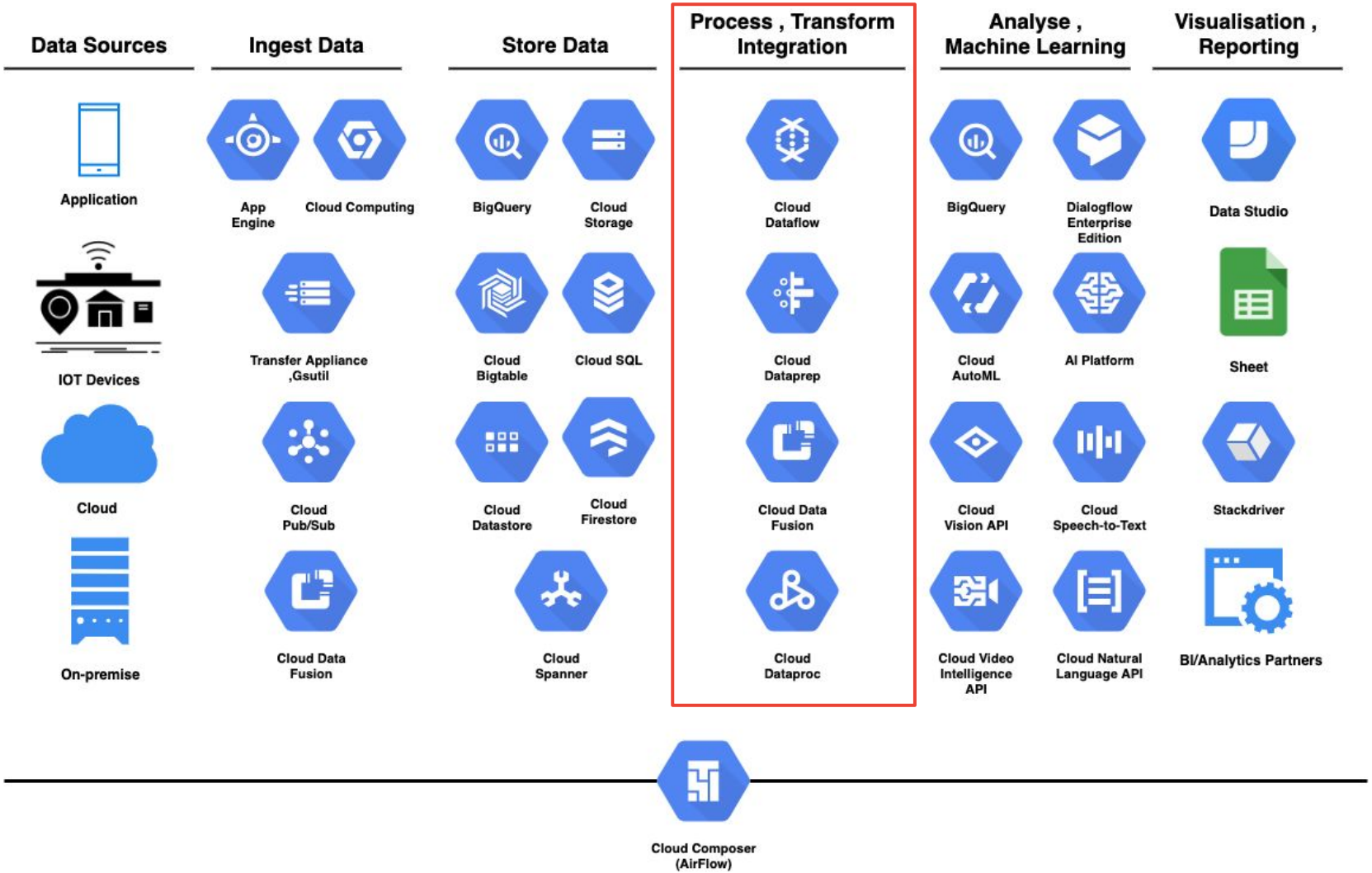
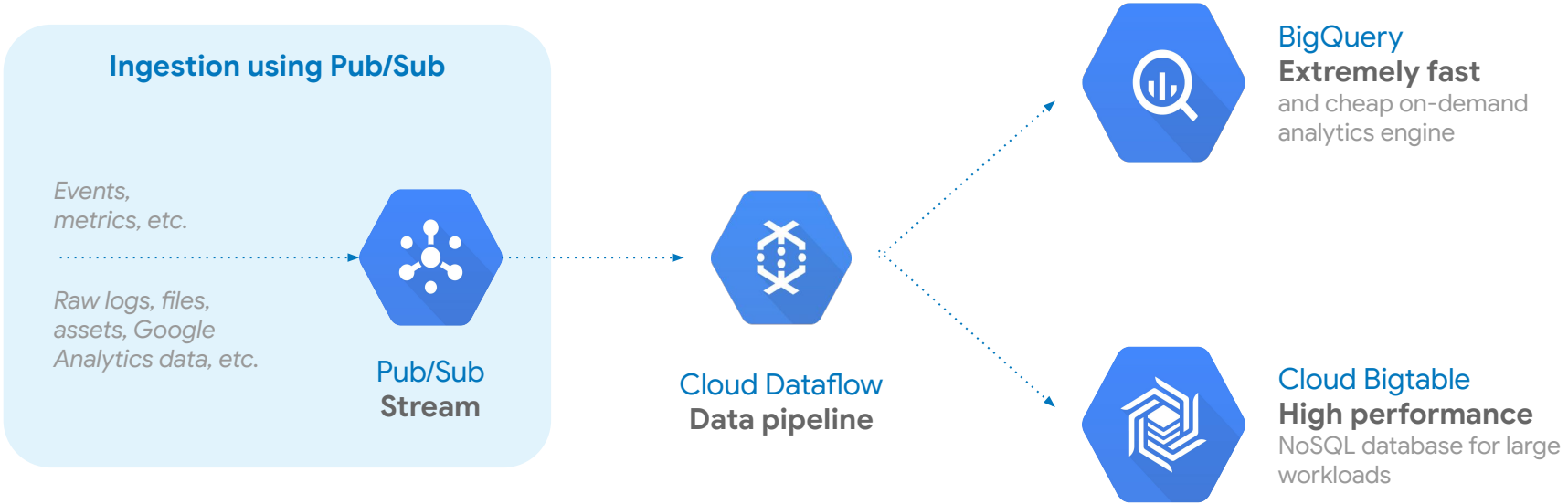


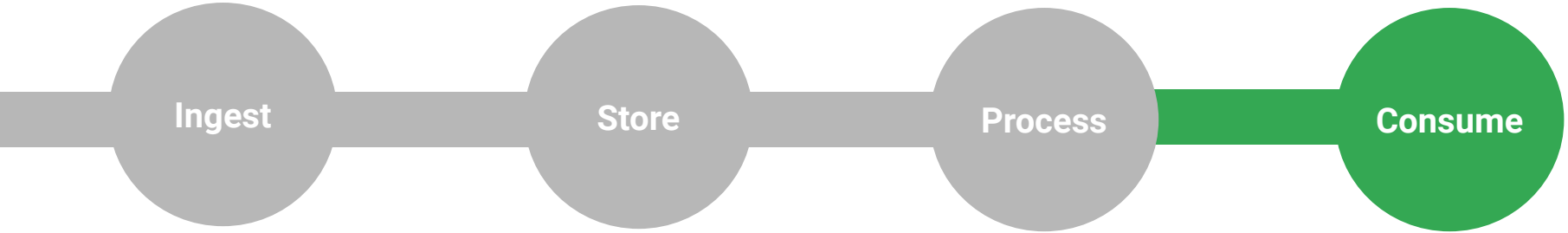

**Image Ref** : https://www.trifacta.com/blog/general-availability-google-cloud-dataprep/


Data Sources
Ingest Data
Store Data
Process , Transform Integration
Analyse , Machine Learning
Visualisation , Reporting
Application
App Engine
Cloud Computing
BigQuery
Cloud Storage
Cloud Dataflow
BigQuery
Dialogflow Enterprise Edition
Data Studio
IOT Devices
Transfer Appliance ,Gsutil
Cloud Bigtable
Cloud SQL
Cloud Dataprep
Cloud AutoML
AI Platform
Sheet
Cloud
Cloud Pub/Sub
Cloud Datastore
Cloud Firestore
Cloud Data Fusion
Cloud Vision API
Cloud Speech-to-Text
Stackdriver
On-premise
Cloud Data Fusion
Cloud Spanner
Cloud Dataproc
Cloud Video Intelligence API
Cloud Natural Language API
BI/Analytics Partners
Cloud Composer (AirFlow)

# Stream data pipeline



# Simplify Data Analytics Process




- Data ingestion
- IoT

- Data lake
- Database

- Data pipeline
- Data transformation
- Data warehouse
- Data mart

- Data analytics
- Visualization
- Machine Learning

Ingest Store Process Consume

# Google Cloud AI Platform

**Perception Services**


Vision


Video Intelligence


Speech


Natural Language

Translation

**Platform, libraries, tools**


Cloud ML Engine


BigQuery


DataLab


**AI infrastructure**


Cloud Storage

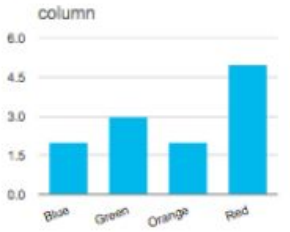

Networking

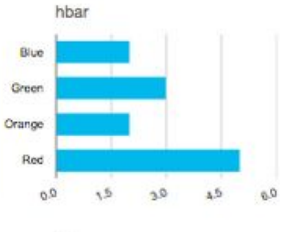

Compute Engine

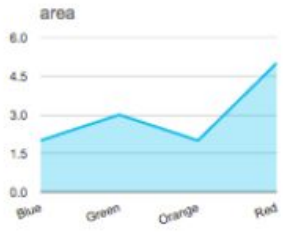

Cloud GPUs

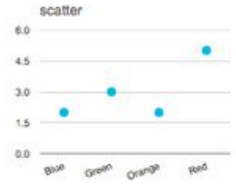

Cloud TPUs

# Data visualization



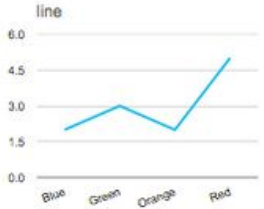



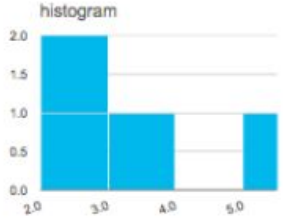



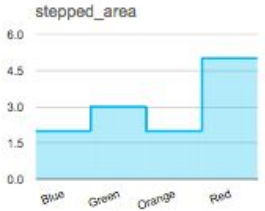




table

| Choose a color | Submissions |
|---|---|
| Blue | 2 |
| Green | 3 |
| Orange | 2 |
| Red | 5 |


# Building an IoT Analytics Pipeline on Google Cloud

1 hour 10 minutes  5 Credits  Rate Lab

**GSP088**

# Big Data Management

# Simplify Data Analytics Process


- Data ingestion
- IoT

- Data lake
- Database

- Data pipeline
- Data transformation
- Data warehouse
- Data mart

- Data analytics
- Visualization
- Machine Learning

# Big Data Management

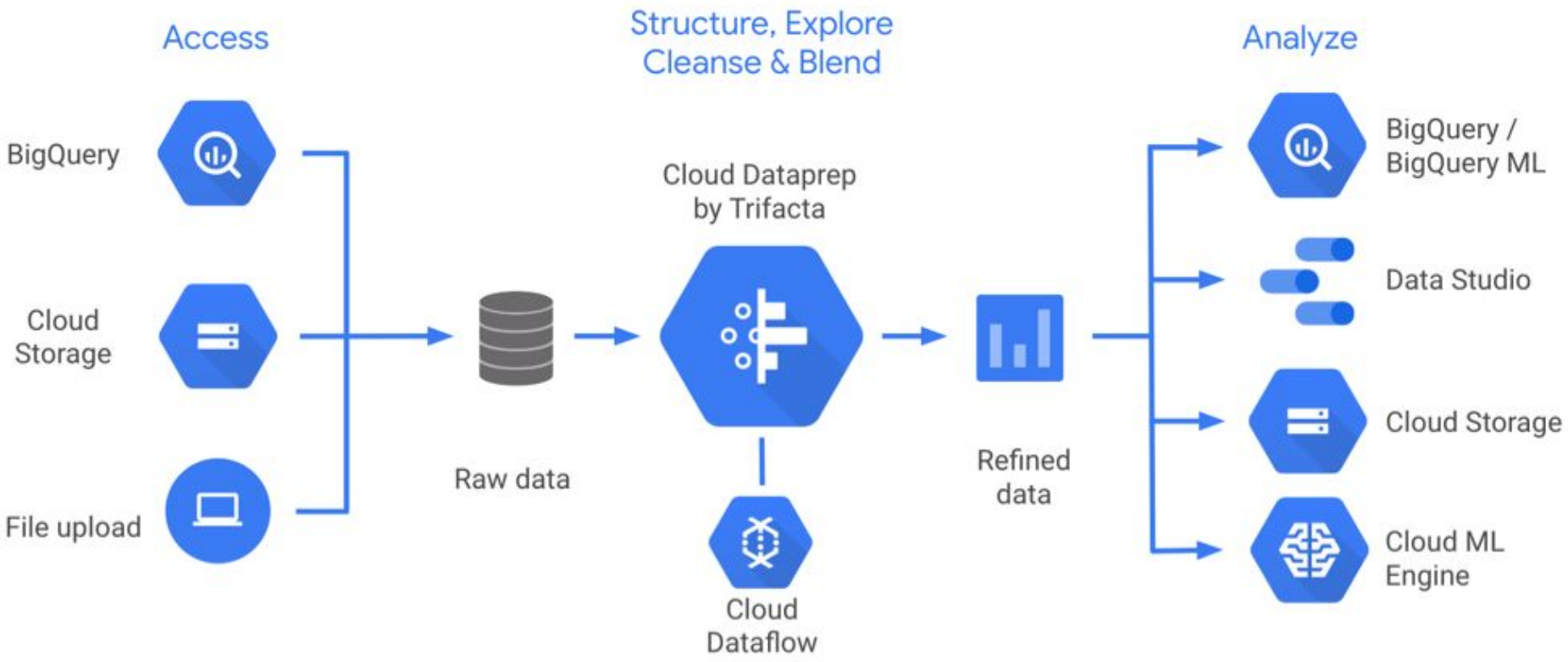
the creation and implementation of architectures, policies, and procedures that manage the full data lifecycle needs of an organization.

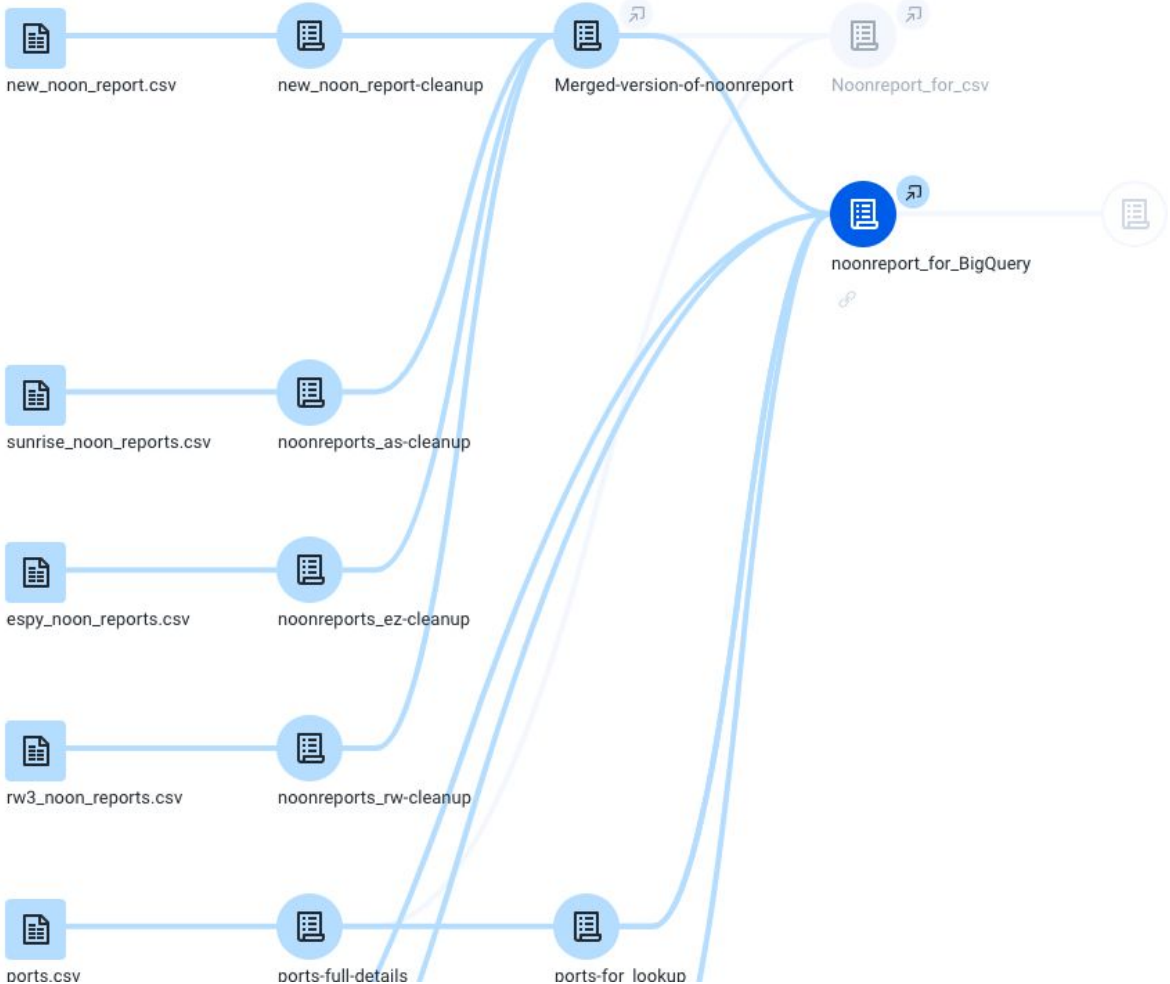
- **Data preparation**
- **Data pipelines**
- **Data extract, transform, load (ETL)**
- **Data catalogs**
- **Data warehouses**
- **Data governance**
- **Data architecture**
- **Data security**

# The Benefits of Good Big Data Management

- Audit processes are more easily, reduce their risk of data loss, increase customer trust, improve operations productivity, optimize IT agility and costs.
- Better data discovery which makes Data-driven decision making more efficient.
- Easy to find and understand the information that in need.
- Provides the structure for information to be easily shared with others
- Allows information to be stored for future reference and easy retrieval.

# Big Data Management: Data preparation






## Details

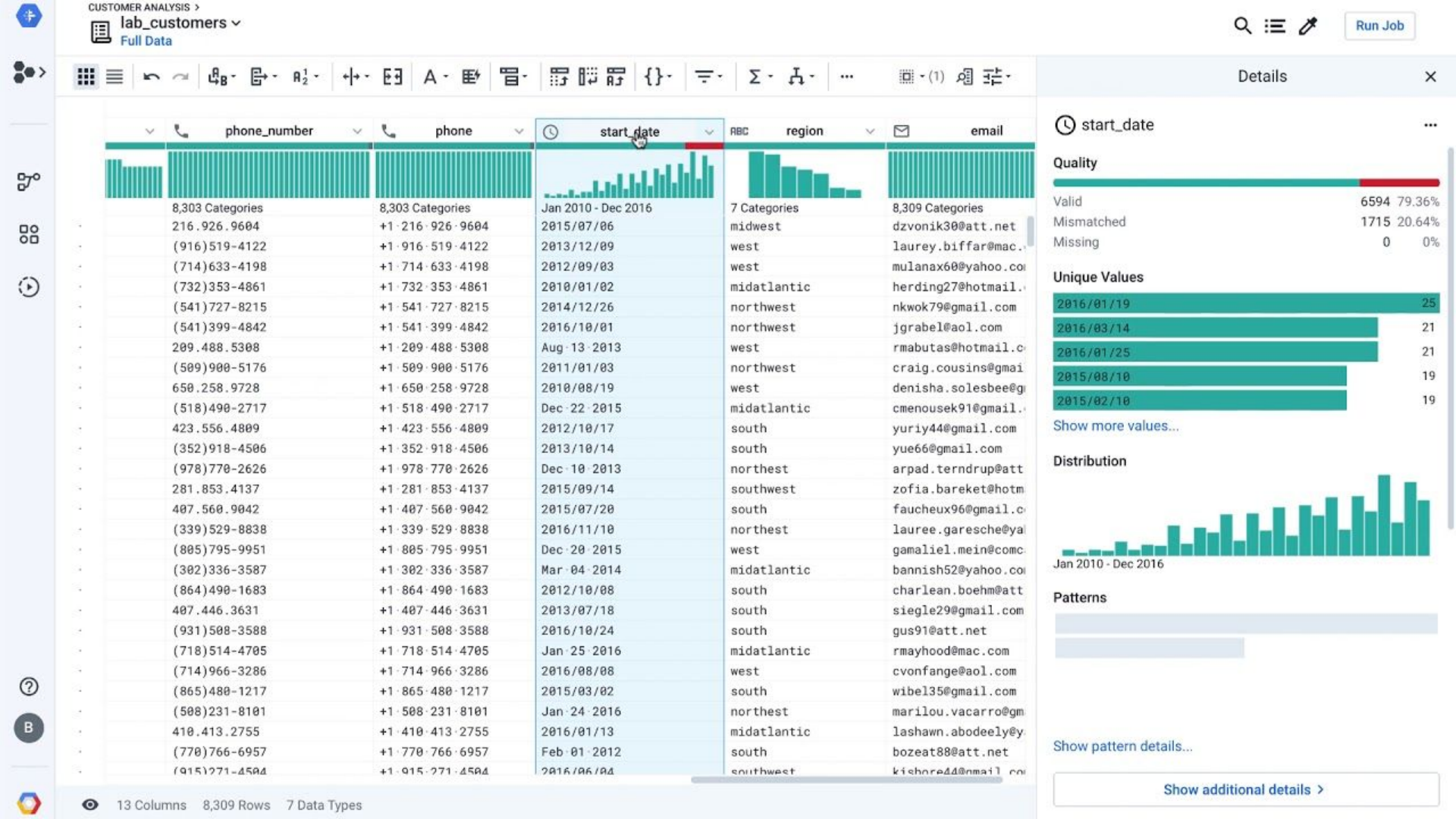
noonreport_for_BigQuery
This is the noonreport for BigQuery

Edit Recipe | Add new Recipe

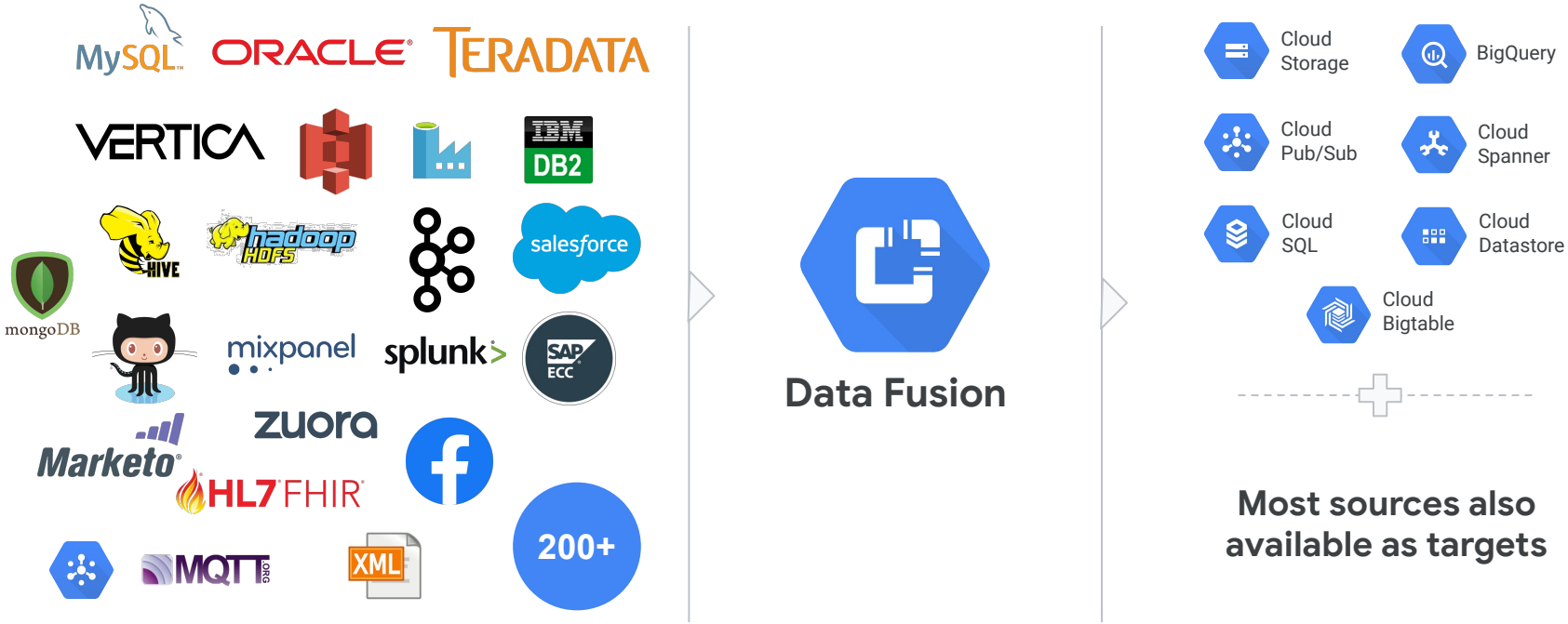
Recipe | Data

Steps Preview

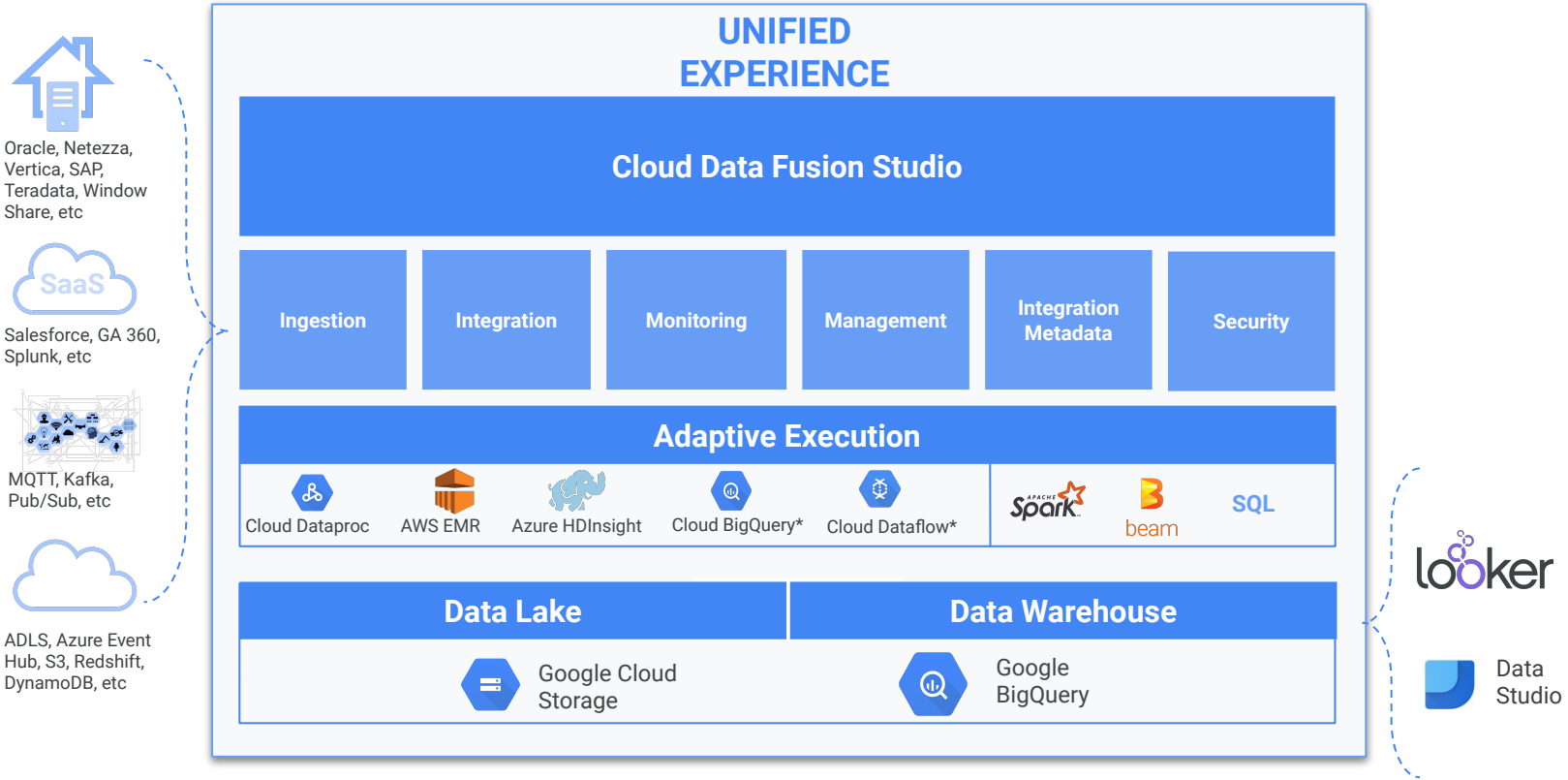
1. **Replace matches** of `4` from ship with 'MYRW'
2. **Replace matches** of `2` from ship with 'MYEZ'
3. **Replace matches** of `3` from ship with 'MYAS'
4. **Lookup** countryid against countryid in old-noonreport-countries as country
5. **Delete** countryid
6. **Rename** country to 'countryid'
7. **Change** local_time **type** to Datetime, yy-mm-dd hh:mm:ss, yyyymmddHHMMSS
8. **Change** local_time **type** to Datetime, yy-mm-dd hh:mm:ss, yyyy*mm*dd*HH:MM:SS
9. **Change** ETD **type** to Datetime, yy-mm-dd hh:mm:ss, yyyy*mm*dd*HH:MM:SS
10. **Replace matches** of `.` starting after `{start}{digit}+.{digit}+` ending before `{digit}{end}` from position_lat with ''

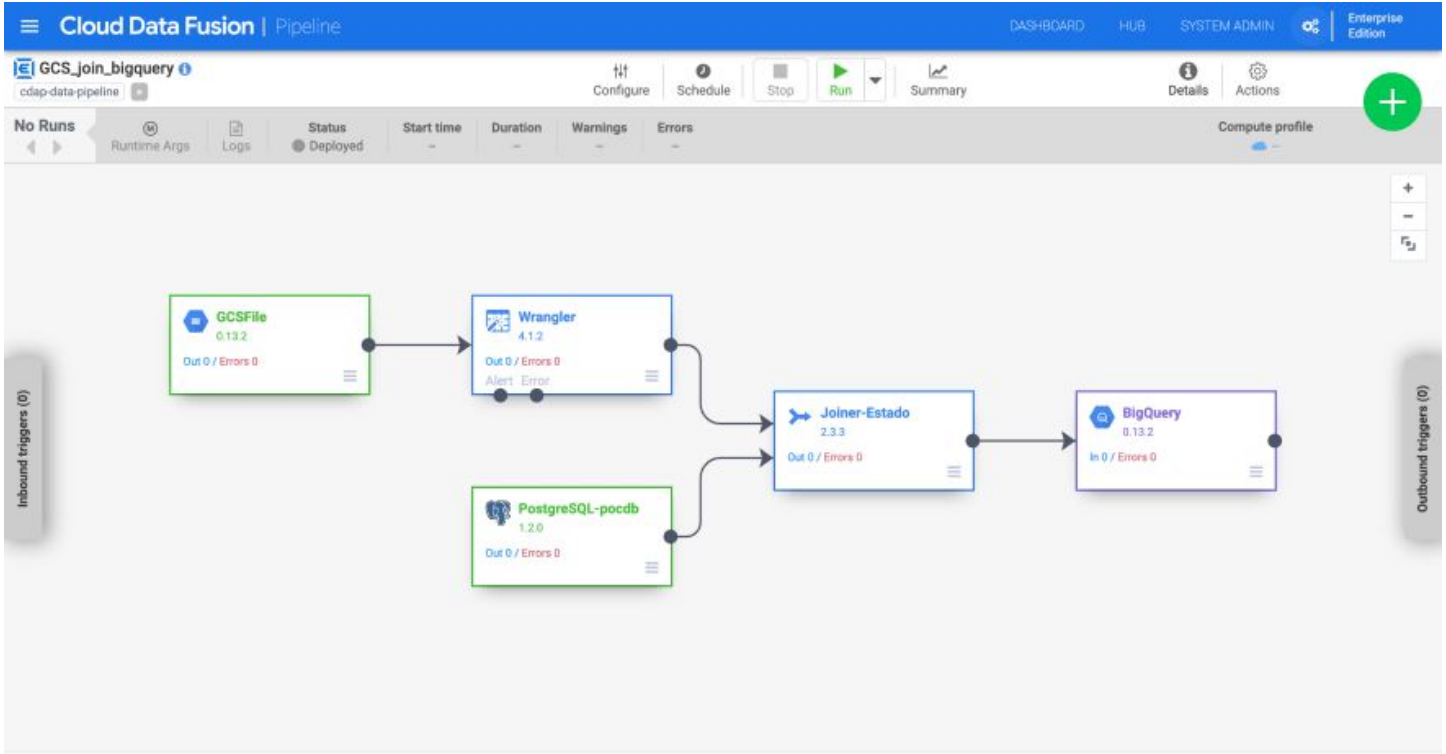

CUSTOMER ANALYSIS >
lab_customers
Full Data
Run Job
phone_number
phone
start_date
region
email
8,303 Categories
8,303 Categories
Jan 2010 - Dec 2016
7 Categories
8,309 Categories
216.926.9604 +1 216 926 9604 2015/07/06 midwest dzvonik30@att.net
(916)519-4122 +1 916 519 4122 2013/12/09 west laurey.biffar@mac.
(714)633-4198 +1 714 633 4198 2012/09/03 west mulanax60@yahoo.co
(732)353-4861 +1 732 353 4861 2010/01/02 midatlantic herding27@hotmail.
(541)727-8215 +1 541 727 8215 2014/12/26 northwest nkwok79@gmail.com
(541)399-4842 +1 541 399 4842 2016/10/01 northwest jgrabel@aol.com
209.488.5308 +1 209 488 5308 Aug 13 2013 west rmabutas@hotmail.c
(509)900-5176 +1 509 900 5176 2011/01/03 northwest craig.cousins@gmai
650.258.9728 +1 650 258 9728 2010/08/19 west denisha.solesbee@g
(518)490-2717 +1 518 490 2717 Dec 22 2015 midatlantic cmenousek91@gmail.
423.556.4809 +1 423 556 4809 2012/10/17 south yuriy44@gmail.com
(352)918-4506 +1 352 918 4506 2013/10/14 south yue66@gmail.com
(978)770-2626 +1 978 770 2626 Dec 10 2013 northest arpad.terndrup@att
281.853.4137 +1 281 853 4137 2015/09/14 southwest zofia.bareket@hotm
407.560.9042 +1 407 560 9042 2015/07/20 south faucheux96@gmail.c
(339)529-8838 +1 339 529 8838 2016/11/10 northest lauree.garesche@ya
(805)795-9951 +1 805 795 9951 Dec 20 2015 west gamaliel.mein@comc
(302)336-3587 +1 302 336 3587 Mar 04 2014 midatlantic bannish52@yahoo.co
(864)490-1683 +1 864 490 1683 2012/10/08 south charlean.boehm@att
407.446.3631 +1 407 446 3631 2013/07/18 south siegle29@gmail.com
(931)508-3588 +1 931 508 3588 2016/10/24 south gus91@att.net
(718)514-4705 +1 718 514 4705 Jan 25 2016 midatlantic rmayhood@mac.com
(714)966-3286 +1 714 966 3286 2016/08/08 west cvonfange@aol.com
(865)480-1217 +1 865 480 1217 2015/03/02 south wibel35@gmail.com
(508)231-8101 +1 508 231 8101 Jan 24 2016 northest marilou.vacarro@gm
410.413.2755 +1 410 413 2755 2016/01/13 midatlantic lashawn.abodeely@y
(770)766-6957 +1 770 766 6957 Feb 01 2012 south bozeat88@att.net
13 Columns
8,309 Rows
7 Data Types
Details
start_date
Quality
Valid 6594 79.36%
Mismatched 1715 20.64%
Missing 0 0%
Unique Values
2016/01/19 25
2016/03/14 21
2016/01/25 21
2015/08/10 19
2015/02/10 19
Show more values...
Distribution
Jan 2010 - Dec 2016
Patterns
Show pattern details...
Show additional details >

# Big Data Management: Data pipeline



# Big Data Management: Data pipeline



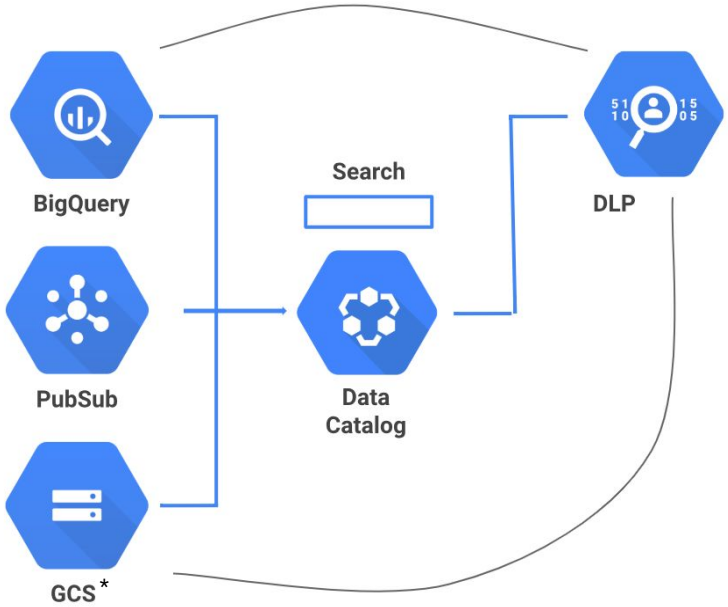

Cloud Data Fusion | Pipeline
DASHBOARD
HUB
SYSTEM ADMIN
Enterprise Edition
GCS_join_bigquery
cdap-data-pipeline
Configure
Schedule
Stop
Run
Summary
Details
Actions
No Runs
Runtime Args
Logs
Status
Deployed
Start time
Duration
Warnings
Errors
Compute profile
Inbound triggers (0)
Outbound triggers (0)
GCSFile
0.13.2
Out 0 / Errors 0
Wrangler
4.1.2
Out 0 / Errors 0
Alert Error
Joiner-Estado
2.3.3
Out 0 / Errors 0
BigQuery
0.13.2
In 0 / Errors 0
PostgreSQL-pocdb
1.2.0
Out 0 / Errors 0

# Big Data Management: Data catalog

Data Catalog is a fully managed and highly scalable data discovery and metadata management service

| | |
|---|---|
| Fully managed & scalable | Easy to get started, there's no infrastructure to set up or manage |
| Simplified data discovery | Simple and easy-to-use search interface, powered by Google search technology that supports Gmail and Drive |
| Built-in governance | Cloud DLP and Cloud IAM integrations provide a foundation for governance |

# Big Data Management: Data catalog



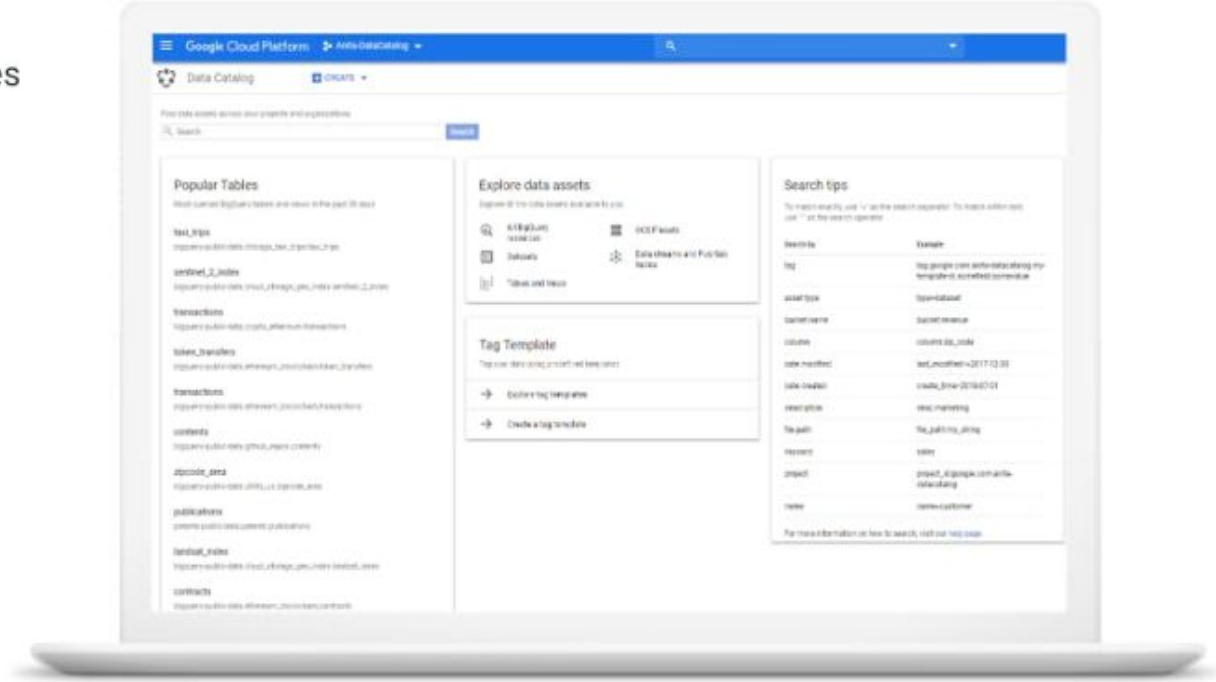

## Data Catalog feature highlights

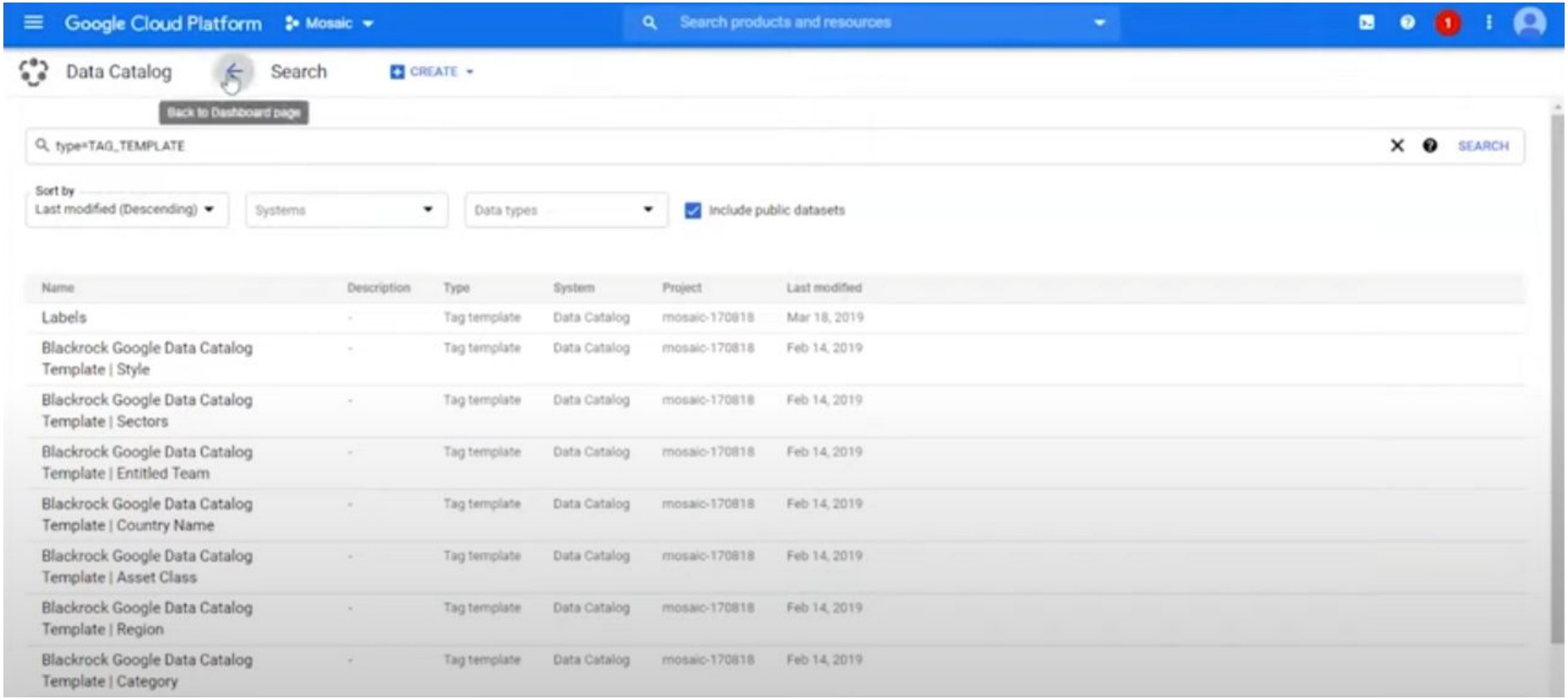
1. Provides a simple search interface for data discovery
2. Supports UI and API for all metadata operations
3. Supports business metadata through schematized tags
4. Auto-ingests technical metadata from GCP data assets
5. Enforces ACL controls on metadata
6. Auto-tags PII data through DLP integration

**Simple keyword search interface** enables both, business and technical users

**Facet search** enables power users

**projectid**:my_proj_id
**system**=bigquery
**type**=table.view
**column**:keyword
**description**:keyword
**tag**:has_pii=true
**createtime**>2020-01
**Updatetime**:2020-02-02




Google Cloud Platform
Mosaic
Search products and resources
Data Catalog
Search
CREATE
Back to Dashboard page
type=TAG_TEMPLATE
SEARCH
Sort by
Last modified (Descending)
Systems
Data types
Include public datasets
Name
Description
Type
System
Project
Last modified
Labels
Tag template
Data Catalog
mosaic-170818
Mar 18, 2019
Blackrock Google Data Catalog Template | Style
Feb 14, 2019
Blackrock Google Data Catalog Template | Sectors
Blackrock Google Data Catalog Template | Entitled Team
Blackrock Google Data Catalog Template | Country Name
Blackrock Google Data Catalog Template | Asset Class
Blackrock Google Data Catalog Template | Region
Blackrock Google Data Catalog Template | Category

# Big Data Management: Data warehouse

Google Cloud Platform's **enterprise data warehouse** for analytics

Gigabyte- to **petabyte-scale** storage and SQL queries

**Encrypted,** durable,
And highly available

Fully managed and **serverless** for maximum agility and scale

Unique

**Real-time** insights from streaming data

Unique

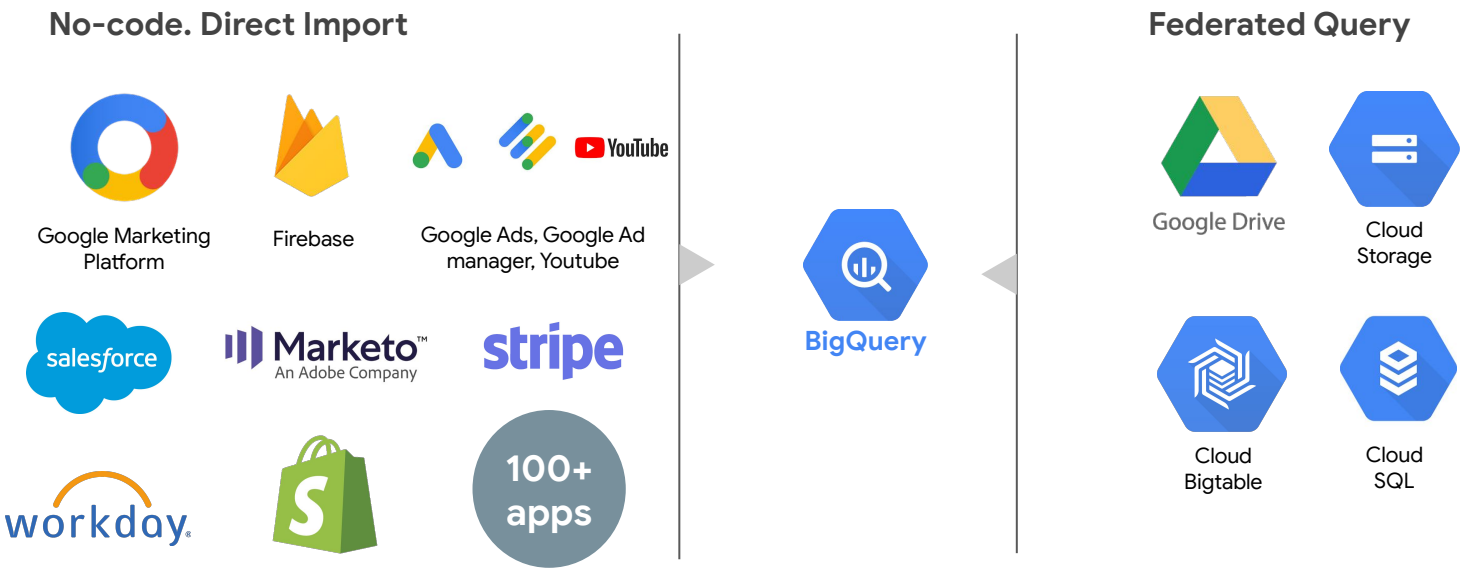
Built-in **ML** for out-of-the-box predictive insights

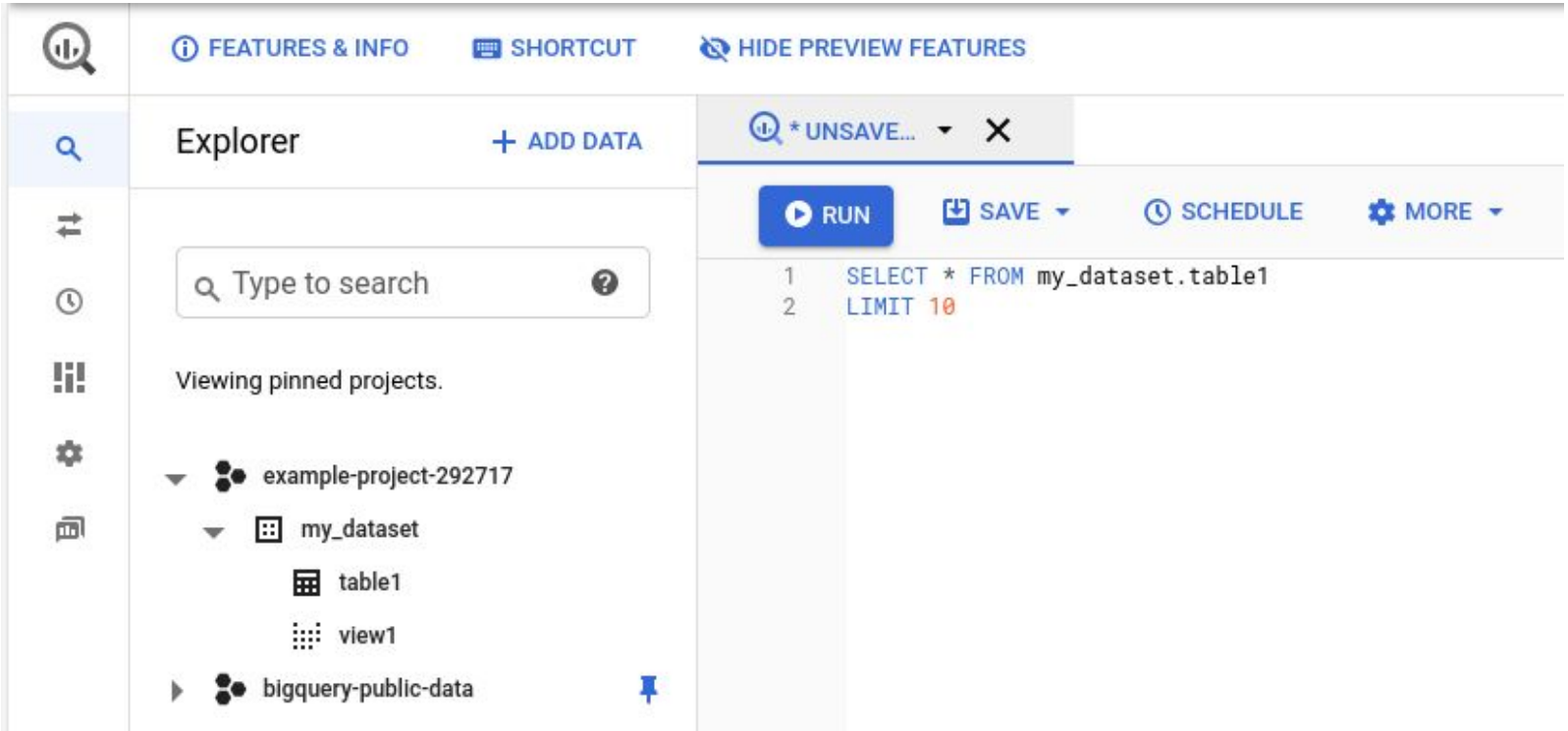
Unique

High-speed, in-memory **BI Engine** for faster reporting and analysis

Unique

# Big Data Management: Data warehouse




FEATURES & INFO
SHORTCUT
HIDE PREVIEW FEATURES
Explorer
ADD DATA
* UNSAVE...
RUN
SAVE
SCHEDULE
MORE
Type to search
Viewing pinned projects.
example-project-292717
my_dataset
table1
view1
bigquery-public-data
1 SELECT * FROM my_dataset.table1
2 LIMIT 10

## Query editor

```
SELECT
  mutation.case_barcode,
  mutation.Variant_Type
FROM
  `isb-cgc-bq.TCGA_versioned.somatic_mutation_hg19_DCC_2017_02` AS mutation
WHERE
  mutation.Hugo_Symbol = 'CDKN2A'
  AND project_short_name = 'TCGA-BLCA'
GROUP BY
  mutation.case_barcode,
  mutation.Variant_Type
ORDER BY
  mutation.case_barcode
```

Run | Save query | Save view | Schedule query | More

## Query results

SAVE RESULTS | EXPLORE DATA

Query complete (0.7 sec elapsed, 192.3 MB processed)

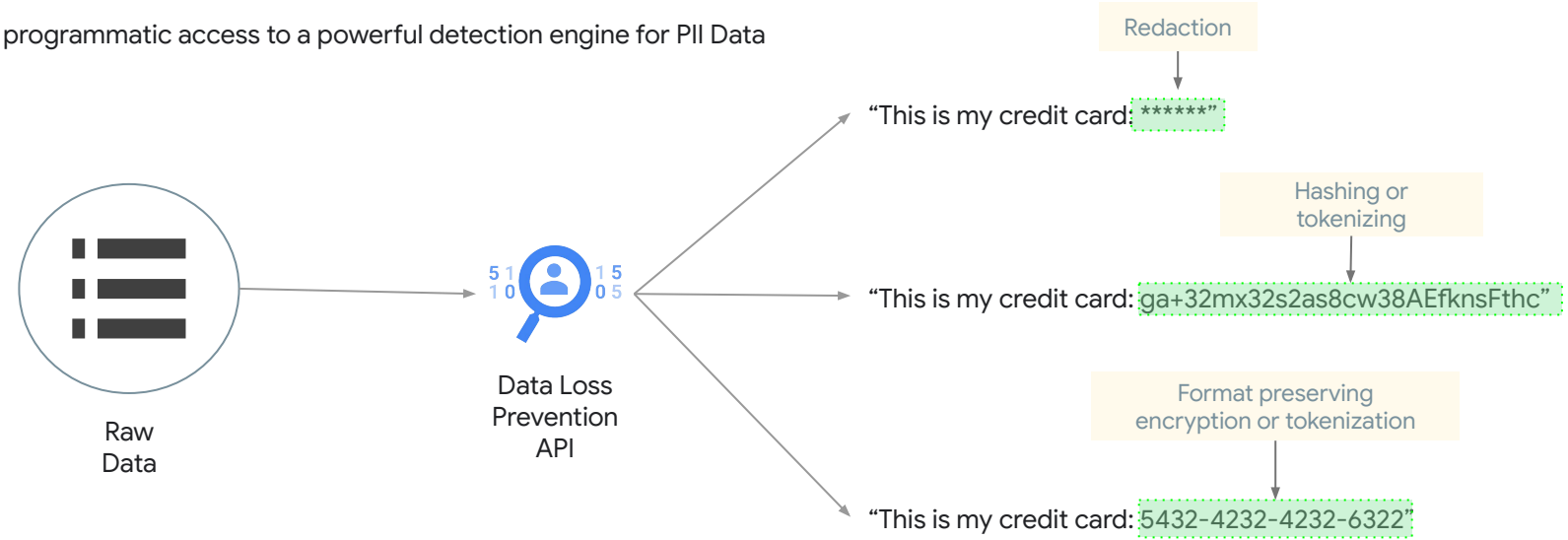
Job information | Results | JSON | Execution details

| Row | case_barcode | Variant_Type |
|---|---|---|
| 1 | TCGA-4Z-AA7R | SNP |
| 2 | TCGA-DK-A6B6 | SNP |
| 3 | TCGA-DK-AA6Q | DEL |
| 4 | TCGA-DK-AA6S | SNP |
| 5 | TCGA-E7-A677 | SNP |
| 6 | TCGA-E7-A7XN | DEL |
| 7 | TCGA-FD-A3N5 | INS |

# Big Data Management: Data Security

## Cloud DLP (Data Loss Prevention)

Provides programmatic access to a powerful detection engine for PII Data



# Cloud DLP: Predefined content detectors

## Canada

Quebec Health Insurance Number (QHIN)
Ontario Health Insurance Plan (OHIP)
British Columbia Personal Health Number (PHN)
Social Insurance Number (SIN)

## United States

Social Security Number
Driver's License number
Drug Enforcement Administration (DEA) Number
ABA Routing Number
National Provider Identifier (NPI)
CUSIP
FDA Approved Prescription Drugs

## United Kingdom

Driver's License Number
National Health Service (NHS) Number
National Insurance Number (NINO)

## Spain

NIF Number
NIE Number

## Netherlands

National Identification Number (BSN)

## France

National ID Card (CNI)
Social Security Number (NIR)

## Brazil

CPF Number

## India

Personal Permanent Account Number (PAN)

## Australia

Medicare Account Number
Tax File Number (TFN)

## Global

Credit Card Number
Bank Account Number (IBAN)
Bank Account Number (SWIFT)
ICD 9-CM Lexicon Global
ICD 10-CM Lexicon

# Cloud DLP: Data masking (de-identification)

| ID | Job Title | Phone | Comments |
|---|---|---|---|
| 359740 | Senior Engineer | 307-964-0673 | Please email them at jane@imadethisup.com |
| 981587 | VP, Engineer | 713-910-6787 | none |
| 394091 | Lawyer | 692-398-4146 | Updated phone to: 692-398-4146 |
| 986941 | Senior Ops Manager | 294-967-5508 | none |
| 490456 | Junior Ops Manager | 791-954-3281 | Tried to verify account with their SSN 222-44-5555 |

### Easy to use transformations

Redaction, masking, pseudonymization, tokenization, format-preserving encryption, date-shifting and more.

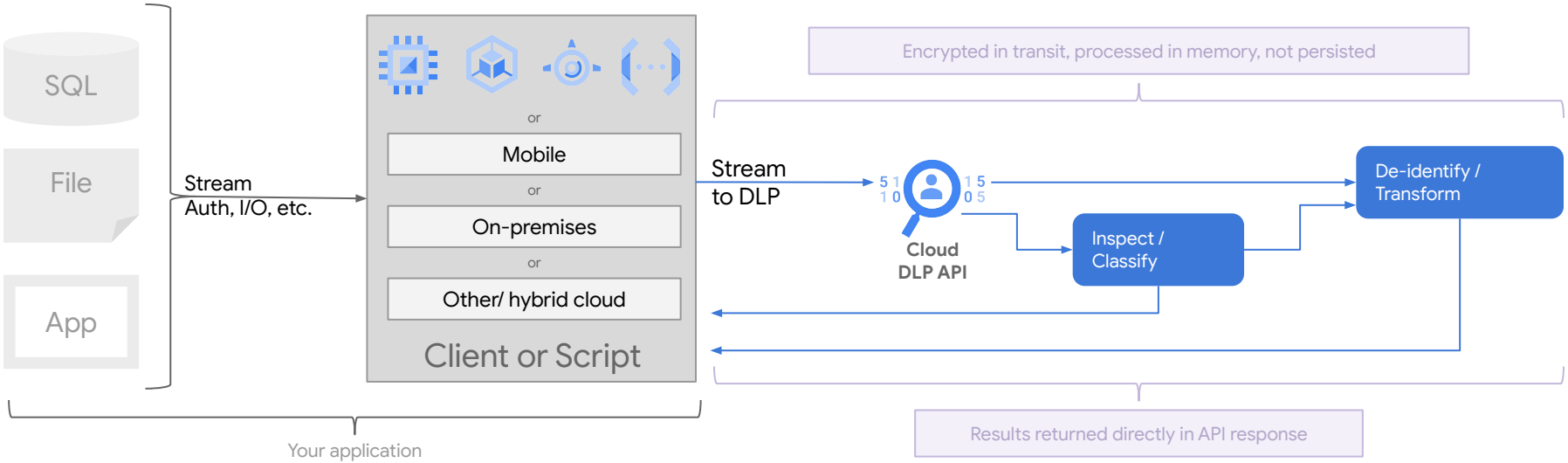
### Handle structured and unstructured data

Apply transforms to an entire column or based on classified data in a “blob of text” or both.

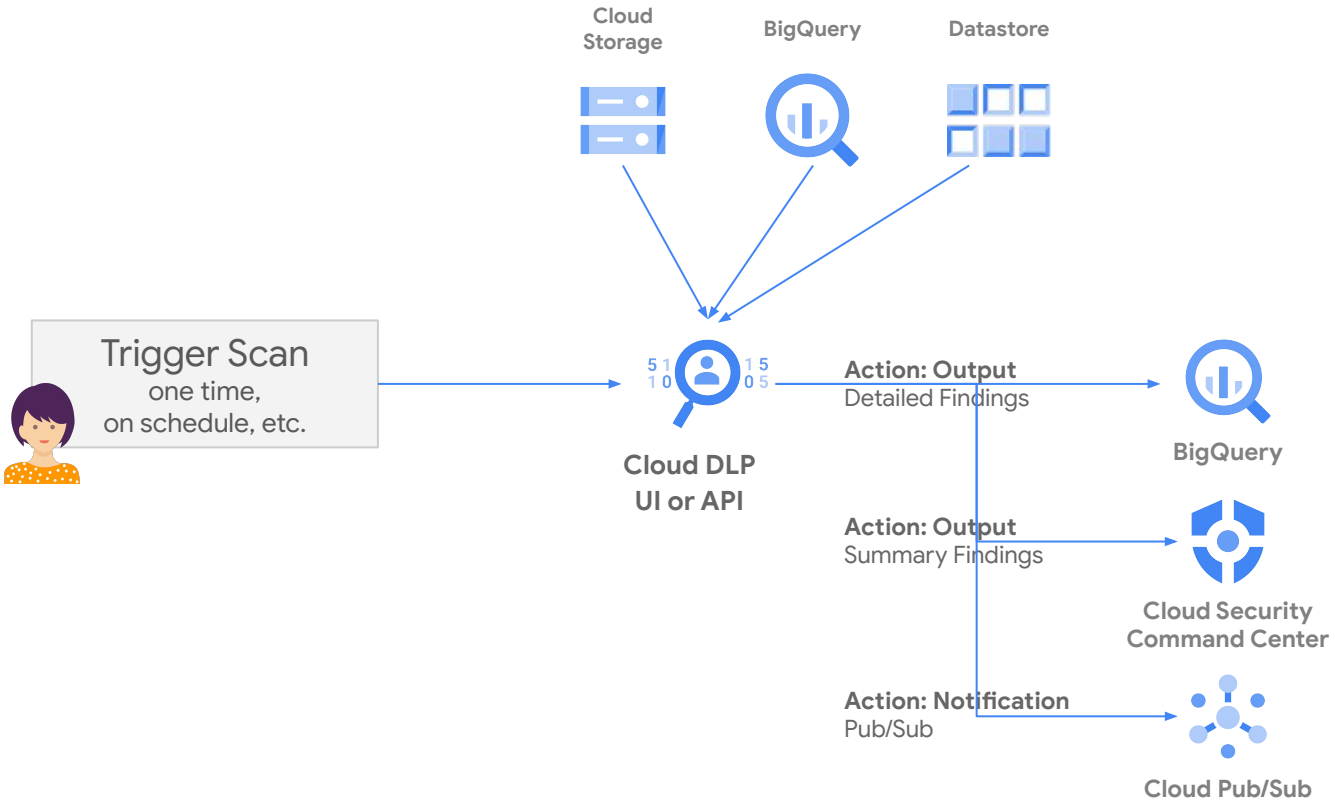
### Mask images

Generate redacted images based on findings or remove all text.

# Implementing Cloud DLP (content methods)



# Implementing Cloud DLP (storage methods)

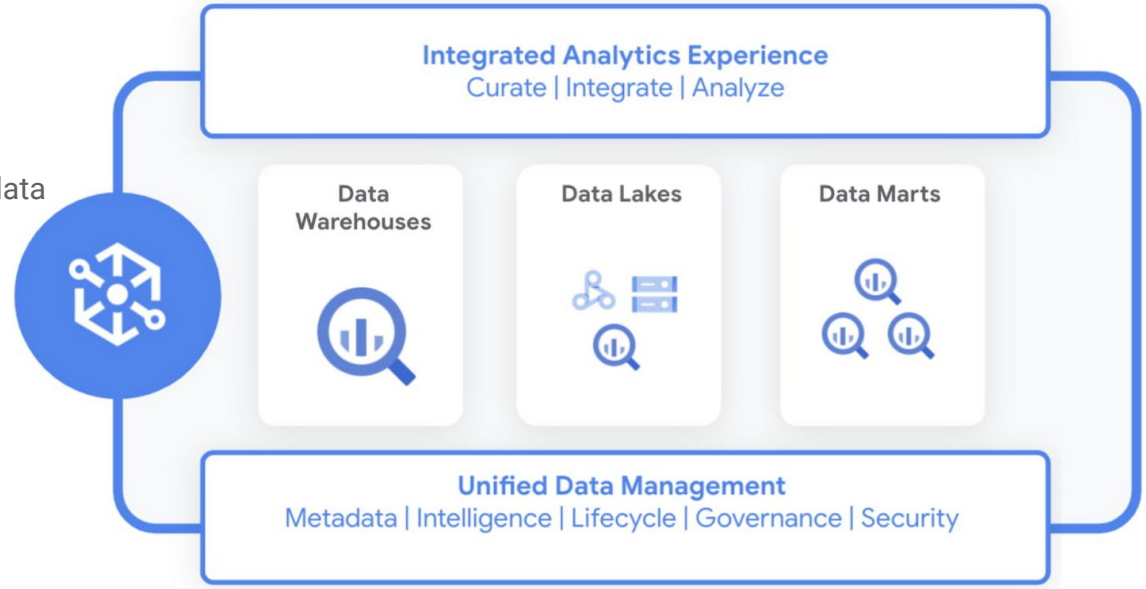


# New services for Unified Data Cloud Strategy

# Dataplex PREVIEW

Centrally manage, monitor and govern your data across data lakes, data warehouses and data marts, and make this data securely accessible to a variety of analytics and data science tools from a single view

- Integrated analytics experience
- Data life cycle management
- Data intelligence
- Centralized security and governance
- Built for distributed data



# Cloud Dataplex: Key features

Integrated analytics experience
Seamless integration with Google-native and open source tools allowing an integrated analytics experience to curate, secure, integrate, and analyze data at scale. Additionally, access to fully managed analytics environments with one-click access to notebooks and SQL scripts with the ability to run Apache Spark, SparkSQL, and push down queries to BigQuery.

Data life cycle management
Logically organize your data that spans multiple storage services in data lakes and data zones to map to your business domains. Provide a task-driven single pane of control to ingest, organize, curate, secure, and archive your data with one-click templates and integrations with Dataflow and Data Fusion.

Data intelligence
Automate data discovery, data classification, schema detection, metadata harvesting and registration for any type of data, global data quality checks, and more with built-in data intelligence powered by Google's leading machine learning and AI.

Centralized security and governance
Central policy management, monitoring, and auditing for data authorization and classification. Distributed data ownership with global monitoring and governance across data and related artifacts like machine learning models.

Built for distributed data
Unify your data without movement or duplication. Leave your data where it is to minimize cost and maximize performance.

# Datastream PREVIEW

Serverless and easy-to-use change data capture and replication service

Move and synchronize data between heterogeneous databases, storage and applications reliably to support real-time analytics, database replication and event-driven architectures

- Streaming data from Oracle and MySQL
- Normalized data types across sources
- Schema drift resolution
- Secure by design

# Cloud Datastream: Key features

Streaming data from Oracle and MySQL
Datastream reads and delivers every change—insert, update, and delete—from your Oracle and MySQL databases to load data into BigQuery, Cloud SQL, Cloud Storage, and Cloud Spanner. Agentless and Google-native, it reliably streams every event as it happens.

Normalized data types across sources
Datastream normalizes every event's data type from the source database's type into a unified Datastream type. This lossless data type normalization across sources means easier downstream processing in a source-agnostic way, regardless of where the data originated.

Schema drift resolution
As source schemas change, Datastream allows for fast and seamless schema drift resolution. Datastream rotates files, creating a new file in the destination bucket, on every schema change. Original source data types are just an API-call away with the up-to-date, versioned Schema Registry.

Secure by design
Datastream supports multiple secure, private connectivity methods to protect data in transit. In addition, data is encrypted in transit and at rest so you can rest easy knowing your data is protected as it streams.

# Datastream


(*) See documentation for full list of currently supported sources

Analytics over BigQuery

DB Migration with DMS

Event-driven with Cloud Functions

Stream Processing with Dataflow

AI/ML with Google Brain

Hybrid Cloud

# Analytics Hub PREVIEW

Analytics Hub efficiently and securely exchanges data analytics assets across organizations to address challenges of data reliability and cost. Create and access a curated library of internal and external assets, including unique datasets like Google Trends, backed by the power of BigQuery.

- Built on a decade of data sharing in BigQuery
- Curation and self-service through exchanges
- A sharing model for scalability, security, and flexibility

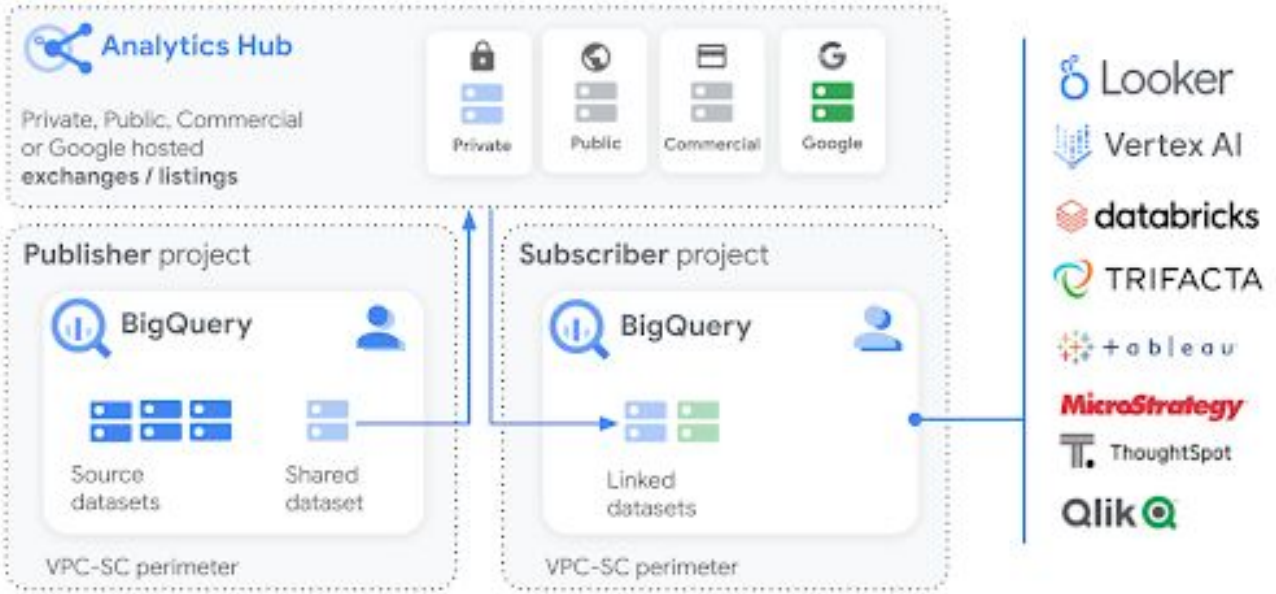
# Analytics Hub: Key features

Built on a decade of data sharing in BigQuery
Since 2010, BigQuery has supported always-live, in-place data sharing within an organization's security perimeter (intra-organizational sharing) as well as data sharing across boundaries to external organizations, e.g., in your vendor or partner ecosystem. Looking at usage over a one week period in April of this year (2021), more than 3,000 organizations shared over 200 petabytes of data in BigQuery, not accounting for intra-organizational sharing. Analytics Hub makes the administration of sharing assets across any boundary even easier and more scalable, while retaining access to key capabilities of BigQuery like its built-in ML, real-time, and geospatial analytics.

Curation and self-service through exchanges
Exchanges are collections of data and analytics assets designed for sharing. Administrators can easily curate an exchange by managing the dataset listings within the exchange. Rich metadata can help subscribers find the data they're looking for, and even leverage analytics assets associated with that data. Exchanges within Analytics Hub are private by default, but granular roles and permissions can be set easily for you to deliver data at scale to exactly the right audiences.

A sharing model for scalability, security, and flexibility
Shared datasets are collections of tables and views in BigQuery defined by a data publisher and make up the unit of cross-project / cross-organizational sharing. Data subscribers get an opaque, read-only, linked dataset inside their project and VPC perimeter that they can combine with their own datasets and connect to solutions from Google Cloud or our partners. For example, a retailer might create a single exchange to share demand forecasts to the 1,000's of vendors in their supply chain–having joined historical sales data with weather, web clickstream, and Google Trends data in their own BigQuery project, then sharing real-time outputs via Analytics Hub. The publisher can add metadata, track subscribers, and see aggregated usage metrics.

# Analytics Hub



# Redacting Sensitive Data with the DLP API

1 hour 1 Credit

## Overview

In this lab, you set up the Cloud Data Loss Prevention API (DLP API) and and use the API to inspect a string of data for sensitive information. The DLP API helps you better understand and manage sensitive data.

It provides fast, scalable classification and redaction for sensitive data elements like credit card numbers, names, social security numbers, US and selected international identifier numbers, phone numbers and Google Cloud credentials.

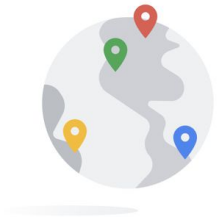
# Data Catalog: Qwik Start

30 minutes 1 Credit Rate Lab

GSP729

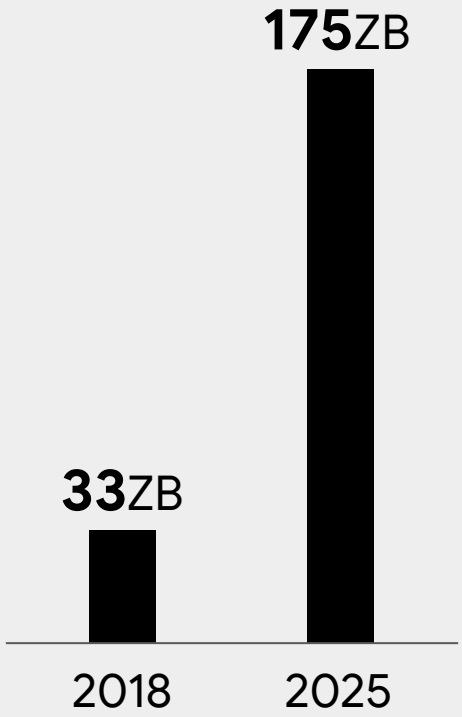
Google Cloud Self-Paced Labs

# Summary



# The world is generating more data than ever

By 2025, the world datasphere will be **175 zettabytes**.



# Simplify Data Analytics Process


- Data ingestion
- IoT

- Data lake
- Database

- Data pipeline
- Data transformation
- Data warehouse
- Data mart

- Data analytics
- Visualization
- Machine Learning

# Big Data Management

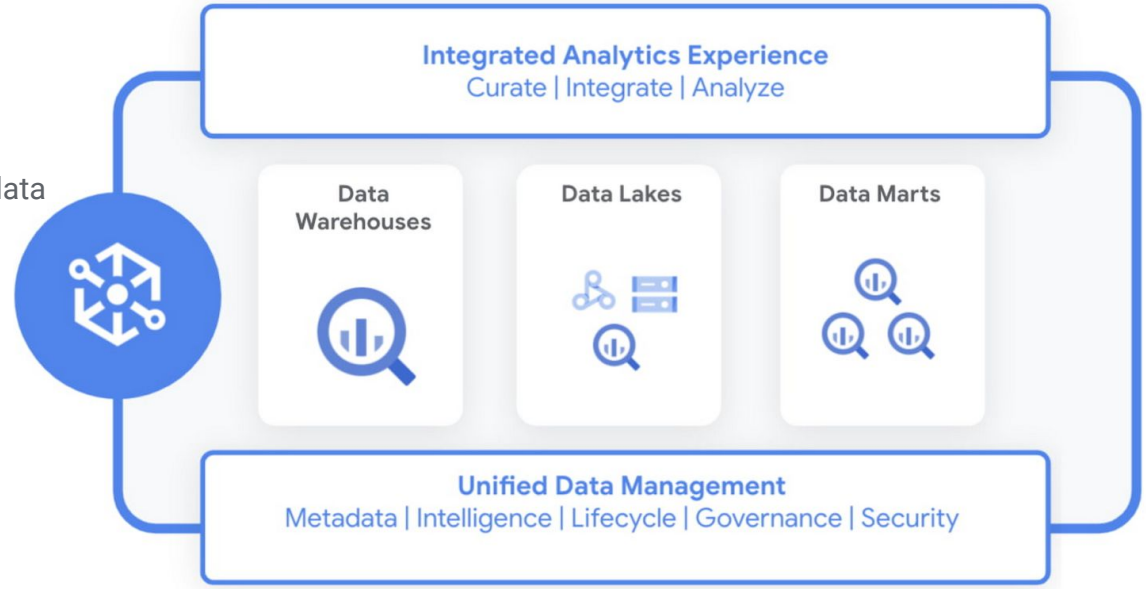
the creation and implementation of architectures, policies, and procedures that manage the full data lifecycle needs of an organization.

- **Data preparation**
- **Data pipelines**
- **Data extract, transform, load (ETL)**
- **Data catalogs**
- **Data warehouses**
- **Data governance**
- **Data architecture**
- **Data security**

# Dataplex PREVIEW

Centrally manage, monitor and govern your data across data lakes, data warehouses and data marts, and make this data securely accessible to a variety of analytics and data science tools from a single view

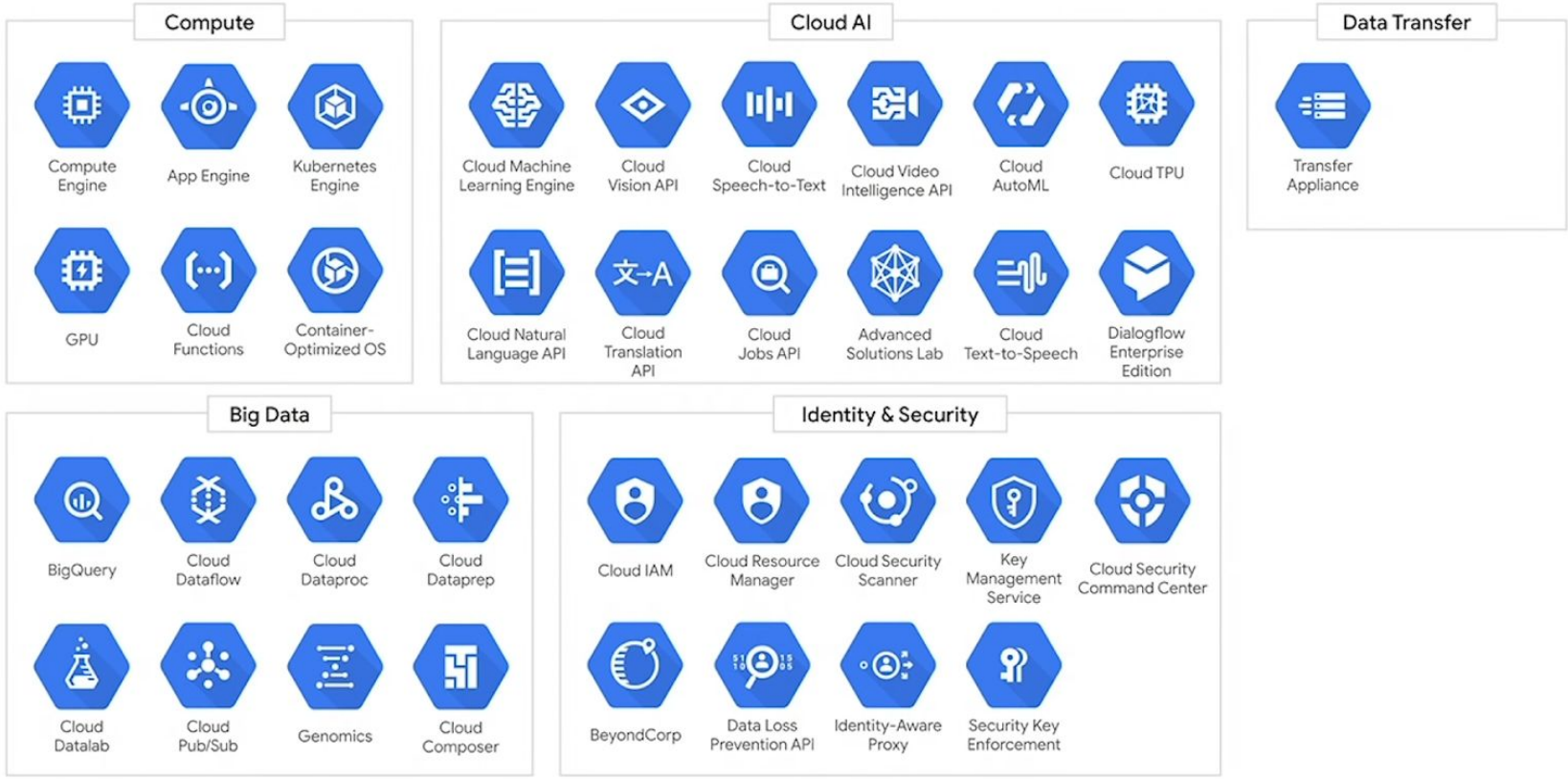
- Integrated analytics experience
- Data life cycle management
- Data intelligence
- Centralized security and governance
- Built for distributed data



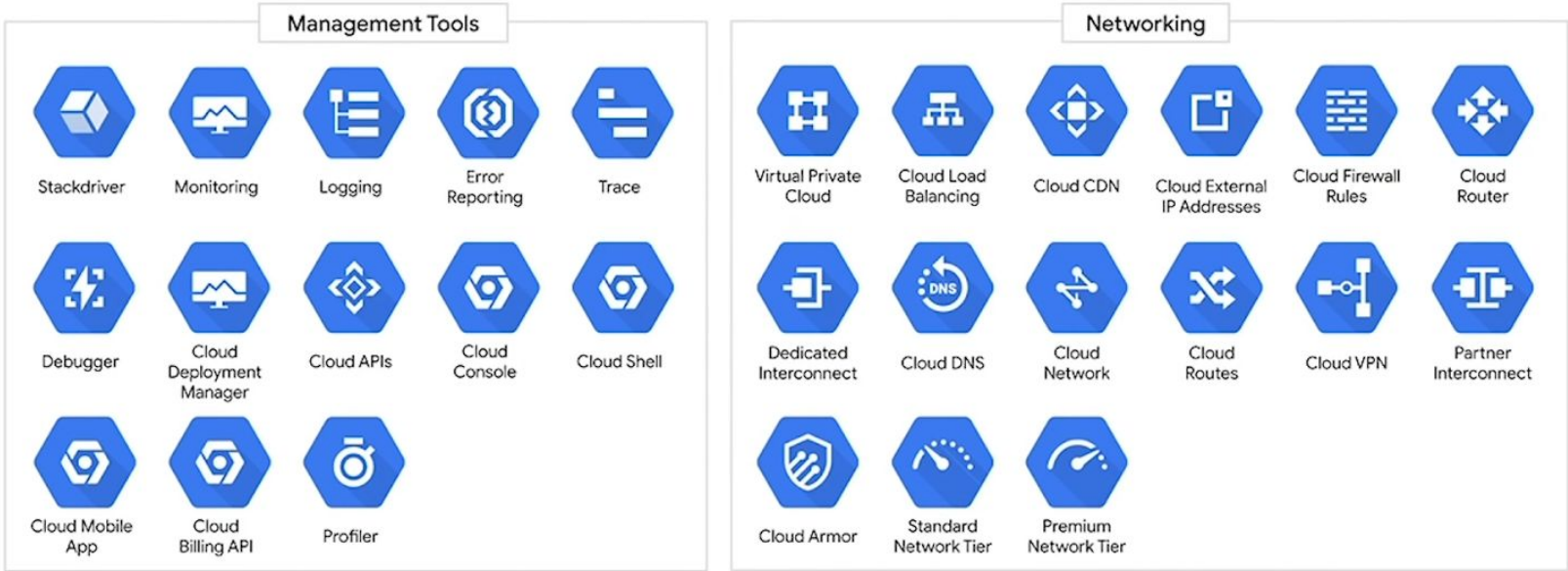

Dataplex in
a minute
Cloud Bytes

There are a ton
Of Google Cloud Products

# Google Cloud



## Google Cloud



# Google Cloud