Designing and Building

Secure Data Environment

Services for handling sensitive data in a research context

Google Cloud

HEAnet
Ireland's National Education & Research Network

Hariprasad Radhakrishnan (Hari)
Customer Engineer
Healthcare and Research
Google Cloud

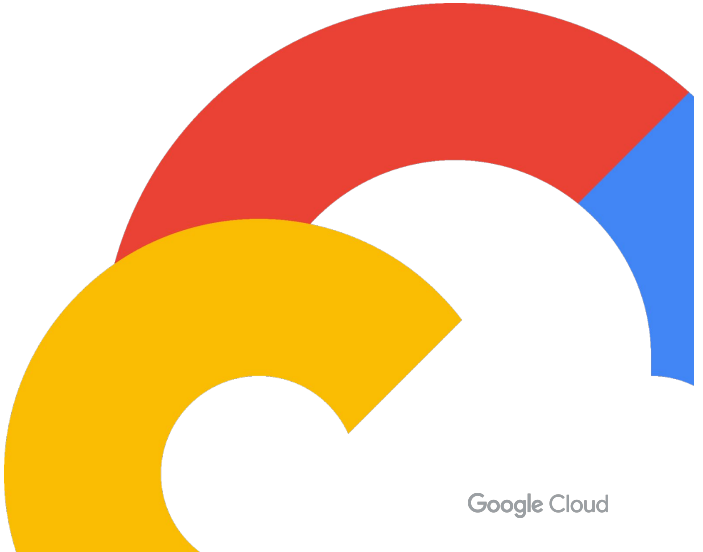
5 years @ Google

Work with Research, Higher Ed and Healthcare organizations across EMEA. Passionate about all things Life Sciences, Healthcare, genomics and fundamental research.

Google Cloud

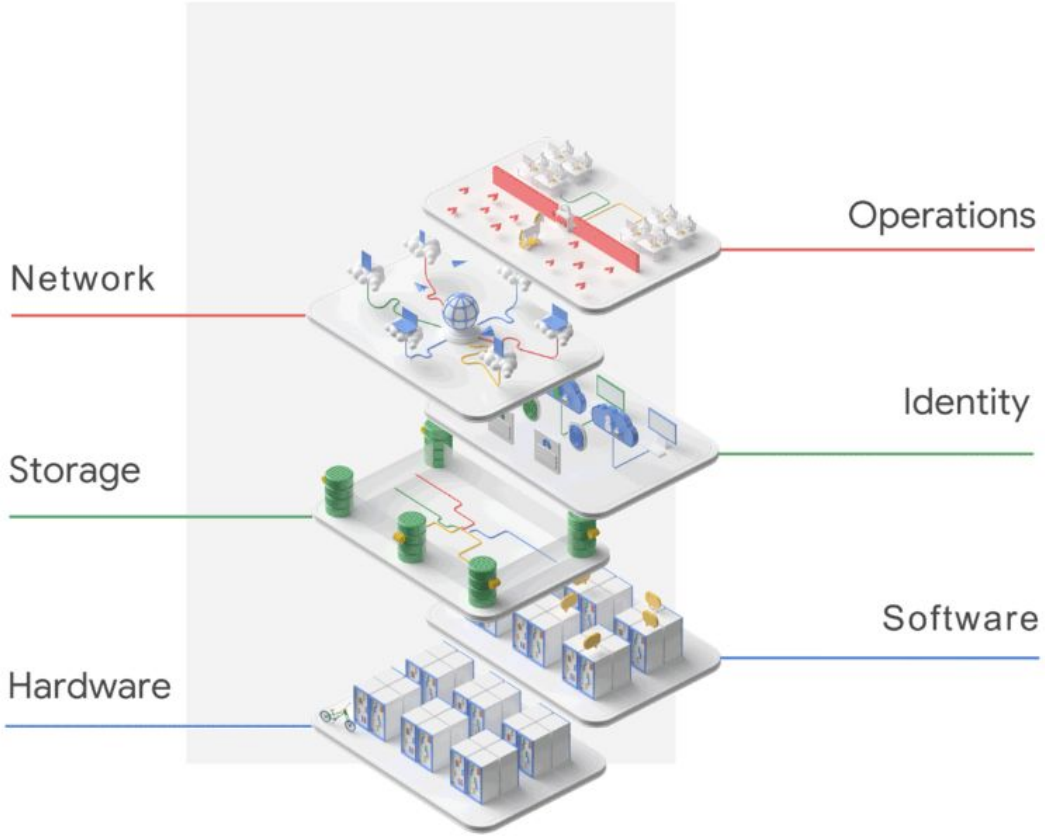
Google Cloud

A Global High-performance infrastructure for **cloud** computing, data analytics & machine learning.



Security that's built in

Defense in depth by default and at scale



Google Cloud

Contents

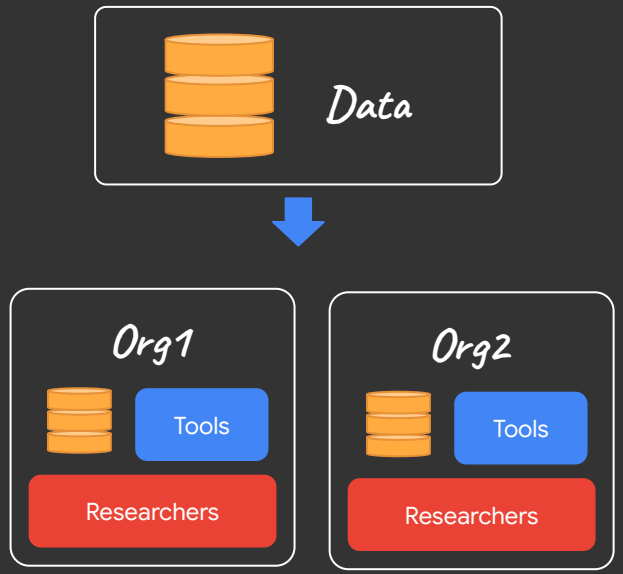
01 Introductions
02 Five safes, use cases
03 Secure Data environment
04 GCP Building blocks
05 Blueprints & Deployable Solutions
06 Q&A

Data Protection Challenges

Key concerns when it comes to sensitive data

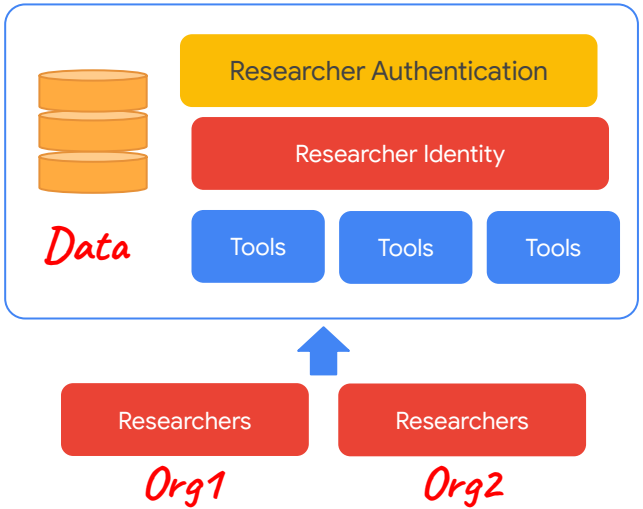
- How do I protect sensitive data or my IP?
- How do I stay compliant with data protection regulations?
- How do I collaborate with other companies processing their sensitive data?
- How do I protect my clients' and users' data?

When Data Leaves your system



Varying security, auditability, least control

When Users come to your system for Data



Consistent experience, shared tools , security, auditability and control

Secure Data environment - Things to consider

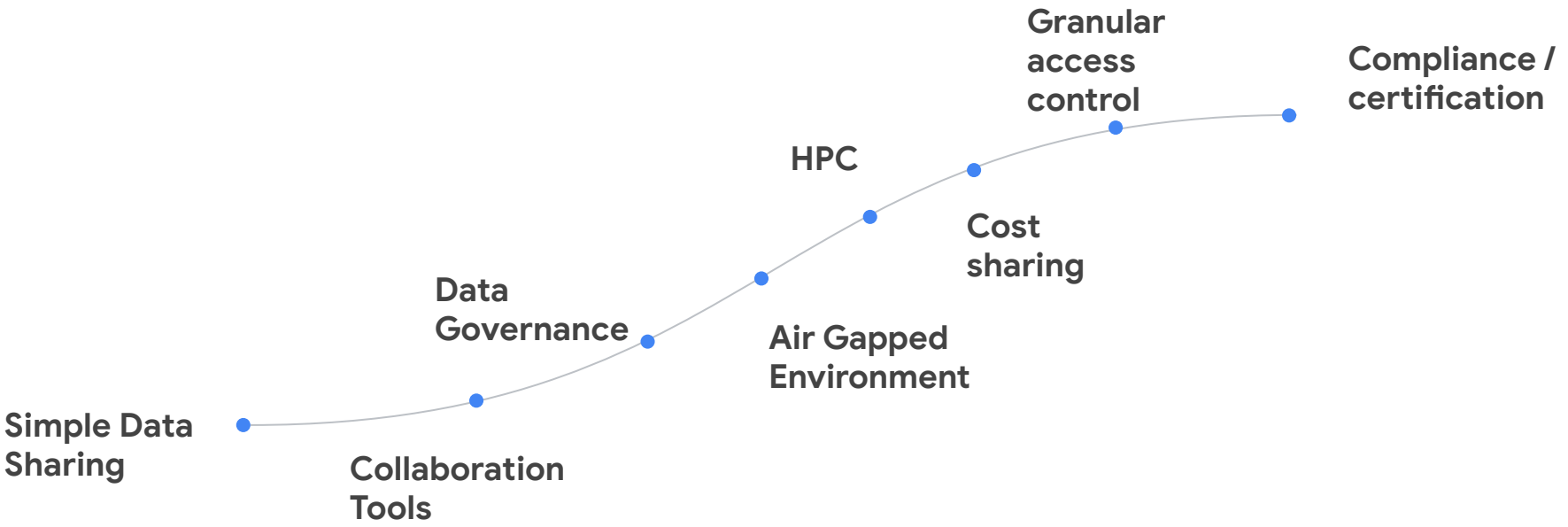
Safe

- Data
- People
- Projects
- Outputs
- Setting
- Computing*

Identify sensitive or PII Data	Is there a real need to share sensitive information	Can we get things done by using anonymised data instead - re-identification risks
Who is the custodian of the data and what Data Governance needs to go in place	Who needs to access the data and for how long, who should not have access	Is there a need to have data exfiltration controls
What Data is being extracted out of the system	Ethical Principles	Complying with Law, transparency and auditability
Security		

Ref: https://www.hdruk.ac.uk/access-to-health-data/trusted-research-environments/
https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/

Need vs Acceptable Complexity vs Cost



Use cases we see for SDEs / TREs

- Government National Health Initiatives requiring SDE / TRE service on Google Cloud.
- Bio-Bank initiatives requiring secure access to genetic data to researchers
- Pharma looking for SDE / TRE services for their external collaboration with Academia.
- Research organisations hosting sensitive data for cross academia, industry, research collaboration projects.

Example: Cancer Gateway data on GCP

National Cancer Institute

ISB Cancer Genomics Cloud

Industry: Healthcare & Life Sciences
Country: United States

https://github.com/isb-cgc/

https://isb-cgc.appspot.com/

All Ref to: ISB-CGC - Cancer Gateway in the Cloud

ISB-CGC: Gateway to cancer research in Google cloud

Data

- Upload your data to cloud
- Access 15+ data types from 25 projects
- Integrated image viewers

Tools

- Choice of workflow technology
- Web-based tools
- Google BigQuery
- Fully customizable Google VMs

Analysis

- Interactive notebooks from Jupyter Lab, R-Studio, or Google Colab
- Pre-designed bio-stats notebooks
- Plot-based visual analysis

Results

- Export charts from notebooks for publication

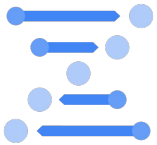
More information at: **https://isb-cgc.appspot.com/**

Google

Existing SDE / TRE on GCP

National Health Organisation
Patient Genomics and Health Data

Industry: Healthcare & Life Sciences
Country: Nordics



(GWAS)

National Health Data - SDE / TRE for 500K Genetic and phenotype data

Air gapped environment for National population health data comprising of both Genetic data and patient health information

Ability to run large scale Genomic pipelines

Access to bioinformatic tools through a Vdi solution

Data governance and approval workflows for extracting summary data

Access to pharmaceutical industry partners

Exfiltration controls using perimeter security

Existing SDE / TRE on GCP

UK Consortium of Research & Universities with Pharma partners

Industry: Healthcare & Life Sciences
Country: United Kingdom

Project SDE / TRE hosting 50,000 patient cohort data to investigate genetic contribution to diseases.

Jointly analyse genetic data and electronic health record data inside a certified SDE / TRE on Google Cloud

Workspaces for different groups of scientists working on sensitive data

Ability to share data between workspaces by researchers when needed

Access to bioinformatic tools through a virtual desktop solution

ABility for researchers to run High performance computing workloads within their workspace

IGV Browser support for analysing genetic variants

Exfiltration controls using perimeter security

Existing SDE / TRE on GCP

Big Pharma - Europe

Industry: Healthcare & Life Sciences
Country: EMEA

NVIDIA MONAI

Open Health Imaging Foundation

jupyter

3DSlicer

slim

SDE / TRE for biomedical Images and FHIR

SDE / TRE to support ingestion of biomedical and radiology images along with phenotype information in FHIR format.

Ability to natively store DICOM and FHIR standard data.

Data harmonization workflows with PII data scans and redaction.

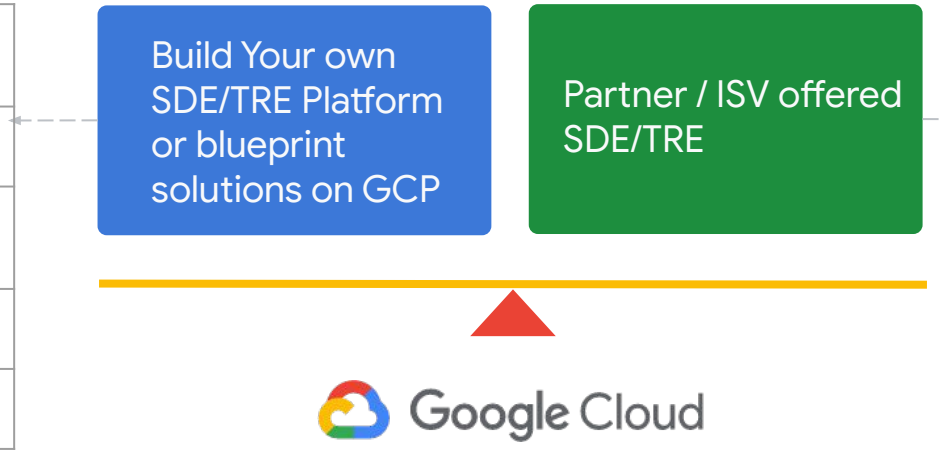
Support for DICOM viewer and annotation tools for Researchers using the SDE / TRE

Support for Workspaces for different pools of researchers and clinical teams

Ability to collaborate and securely share data in SDE / TRE with external collaborators

Choosing your approach

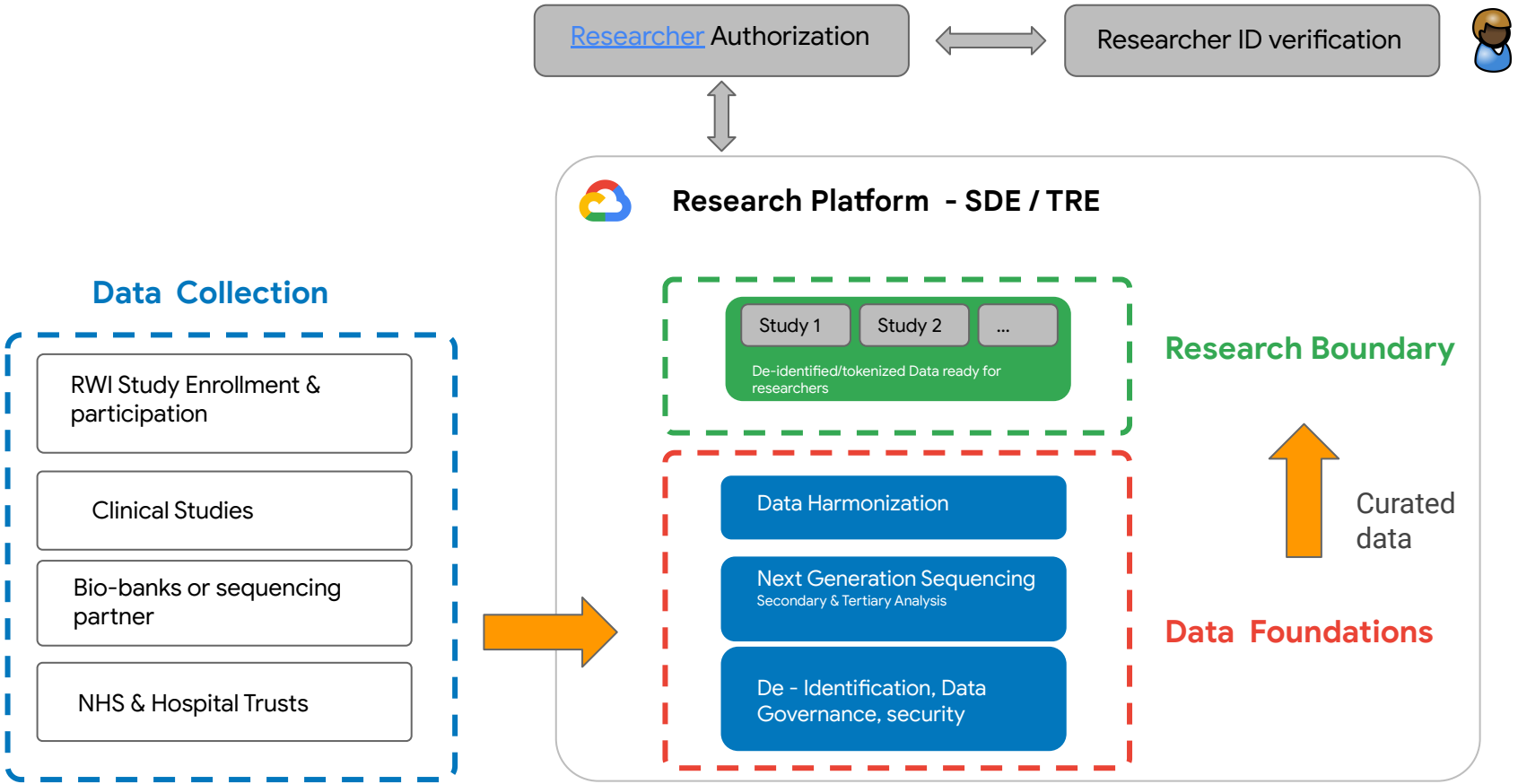
Technical Expertise	High
Resourcing	High
Customize to Needs	High
Time to Market	Slow
Licensing costs	None



Technical Expertise	Low
Resourcing	Low
Customize to Needs	Low
Time to market	Quick
Licensing costs	High

Building blocks of Secure Data Environment (SDE / TRE)

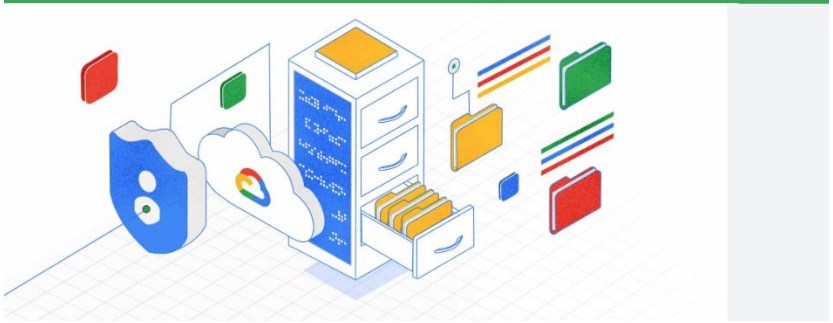
Secure Data Environment (SDE / TRE)



1

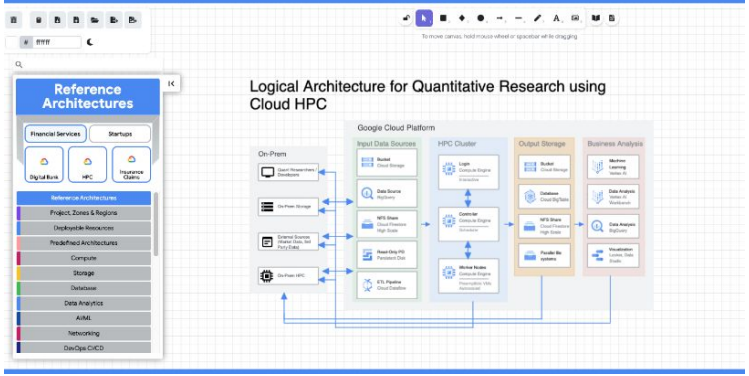
GCP Building Blocks

Services to handle and protect sensitive data



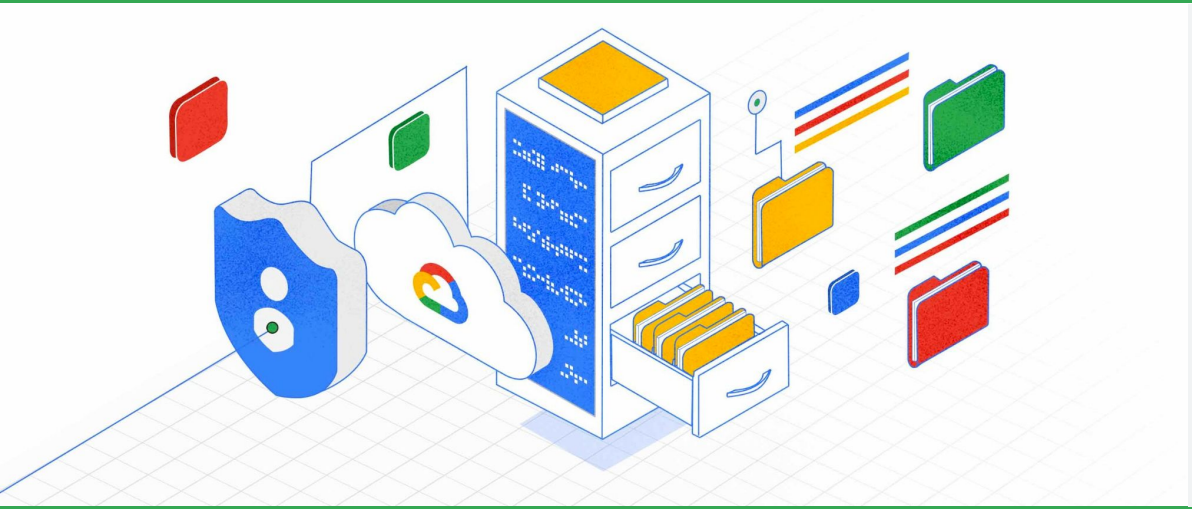
2

Blue prints



GCP Building Blocks

Services to handle and protect sensitive data



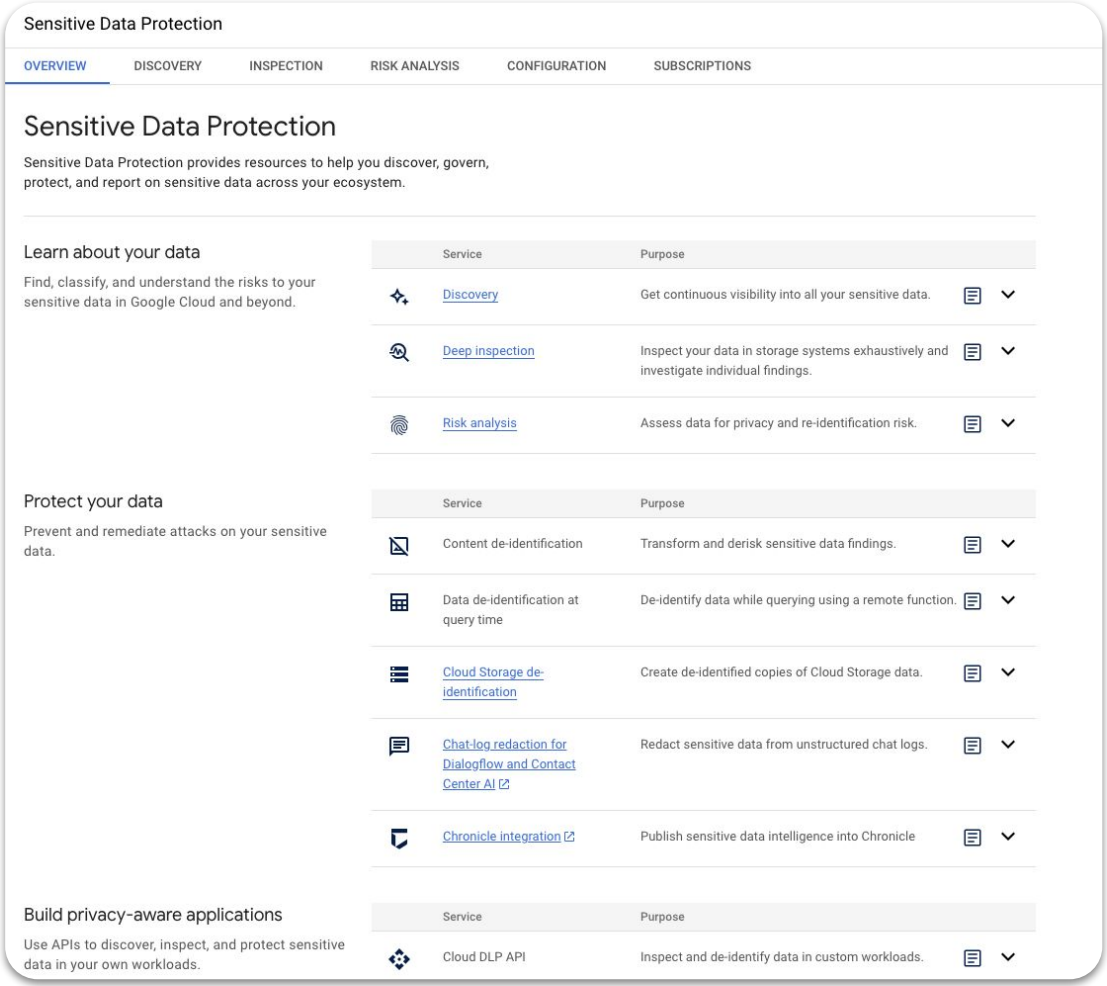
Sensitive Data Protection

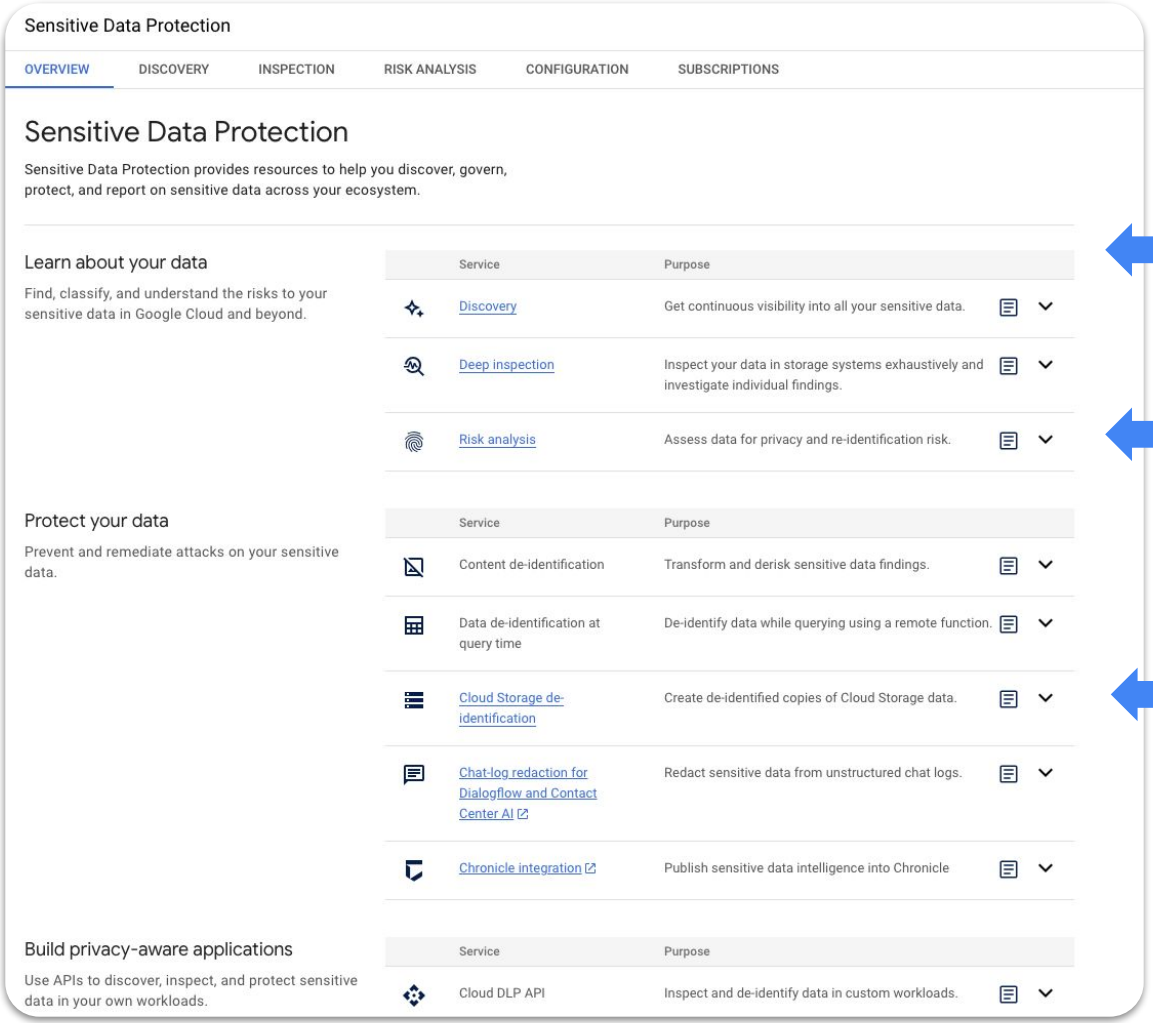
Discover-Govern-Protect-Report

Sensitive Data Protection

Sensitive Data Protection provides resources to help you discover, govern, protect, and report on sensitive data across your ecosystem.

Low Operational Overhead





Discover your sensitive data

Sensitive Data Discovery, a service that continuously profiles your sensitive data so that you can understand and protect it.

Risk Analysis

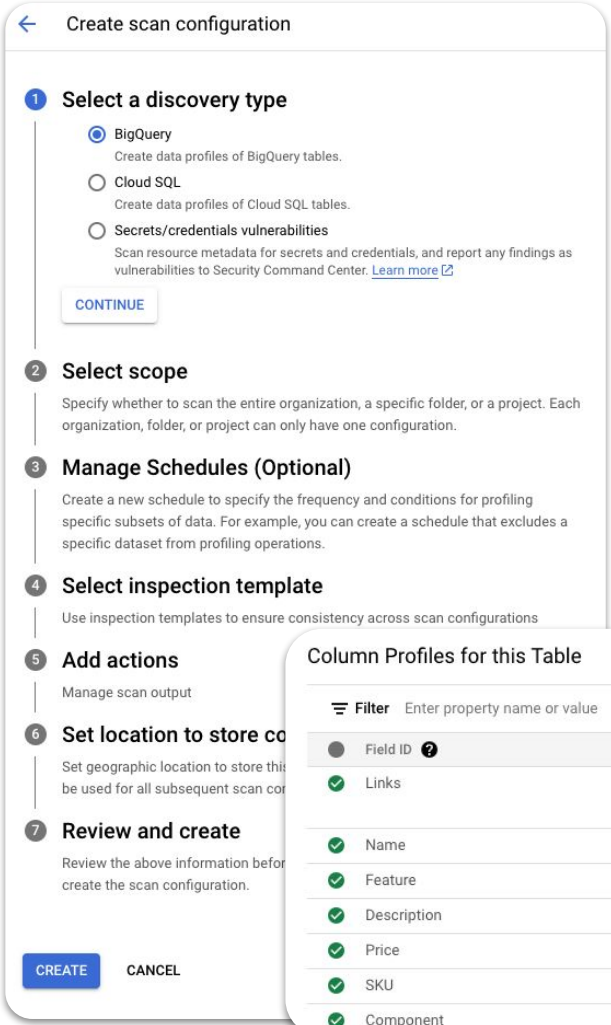
Assess data for privacy and re-identification risk.

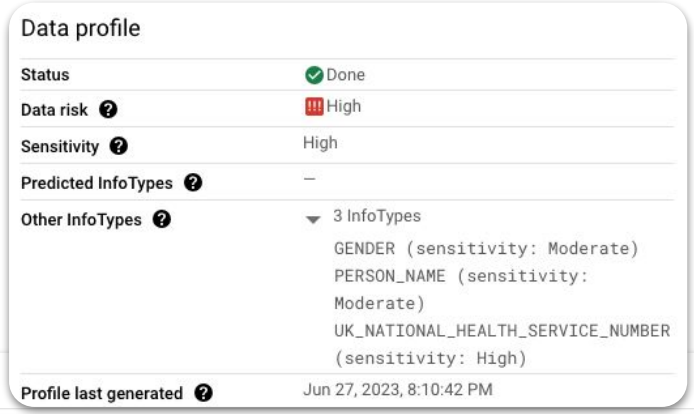
Risk analyses can help you see how the size, shape, and distribution of data can increase re-identification risk.

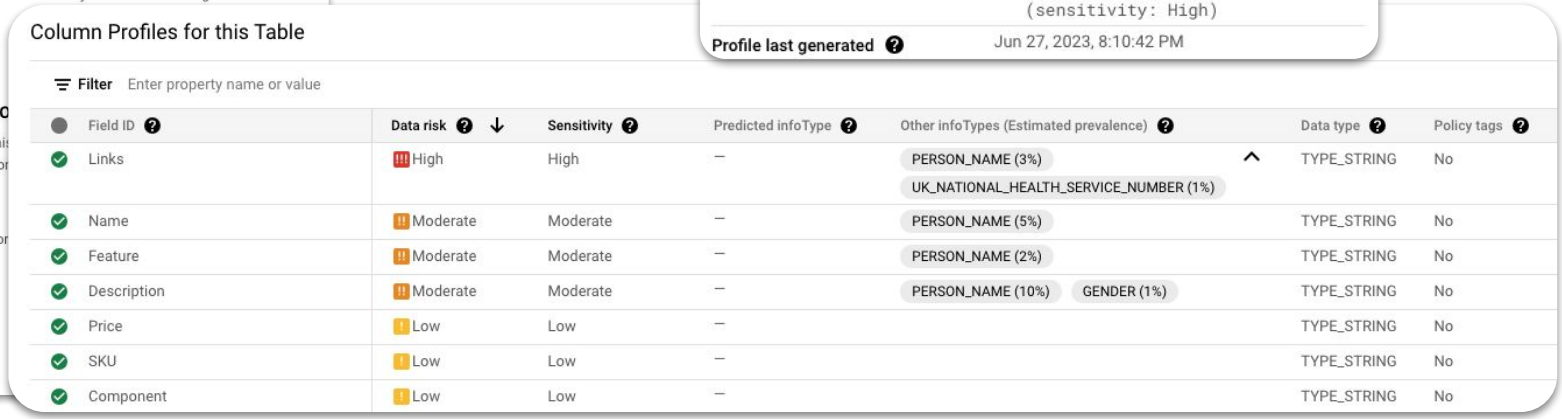
Protect data by de-identifying it

De-identification is the process of removing identifying information from data. Sensitive Data Protection can de-identify sensitive data in text content, including text stored in container structures such as tables.

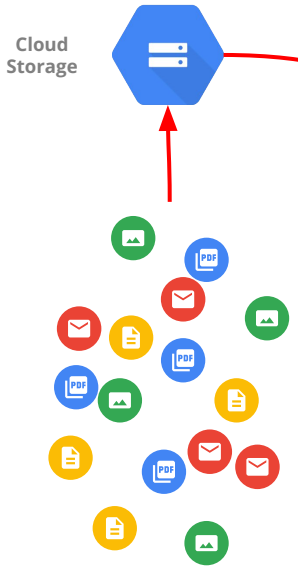
Running Discovery scans for sensitive data







Scan documents



pii-scanner

projects/sandbox-hari/dlpJobs/i-pii-scanner

Done

Findings	Bytes scanned
14882	52.23 MB

Job results

Filter table

Name	Display name	Description	Total	% Total
PERSON_NAME	Person name	A full person name, which can include first names, middle names o...	5013	33.68%
LOCATION	Street addresse...	A physical address or location. Note: Not recommended for use du...	4896	32.9%
DATE	Date or time	A date. This includes most date formats, as well as the names of c...	1495	10.05%
TIME	Human readabl...	A timestamp of a specific time of day. Note: Not recommended for ...	1158	7.78%
AUSTRALIA_TAX_FILE_NUMBER	Australia tax fil...	An Australian tax file number (TFN) is a number issued by the Austr...	639	4.29%
US_HEALTHCARE_NPI	USA healthcare...	The US National Provider Identifier (NPI) is a unique 10-digit identifi...	558	3.75%
PHONE_NUMBER	Phone number	A telephone number.	484	3.25%
ETHNIC_GROUP	Ethnic Group	A person's ethnic group.	343	2.3%
AGE	Age of an indivi...	An age measured in months or years.	215	1.44%
DATE_OF_BIRTH	Date of birth	A date of birth. Note: Not recommended for use during latency sen...	66	0.44%

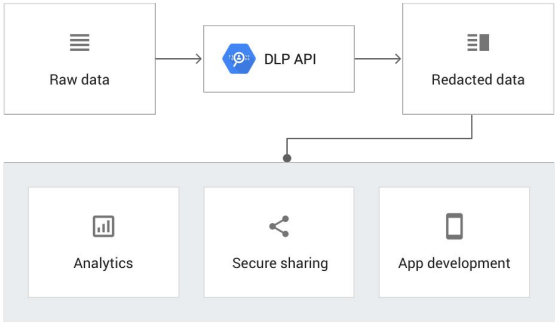
Google Cloud

Data Loss Prevention

Identifying, redacting and cataloging sensitive PII

Data Loss Prevention (DLP) API

Cloud DLP provides access to a powerful sensitive data inspection, classification, and de-identification platform.



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

De-ID: Masking and Tokenization

Input → “This is my phone number: (858)867-5309”

Partial Masking

Output → “This is my phone number: (858)XXX-XXXX”

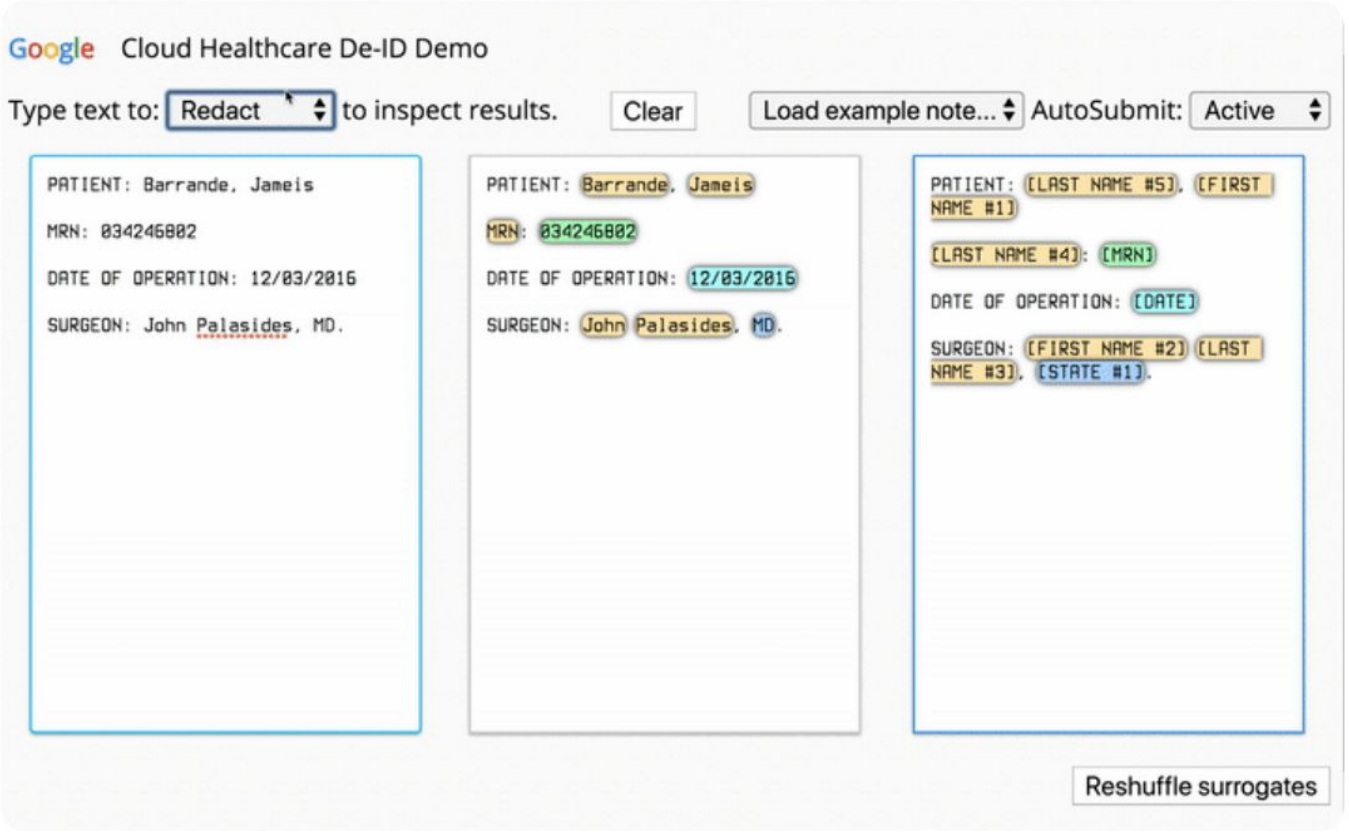
Hashing or Tokenizing

Output → “This is my phone number: ga+32mx32s2as8cw38AEfknsFthc”

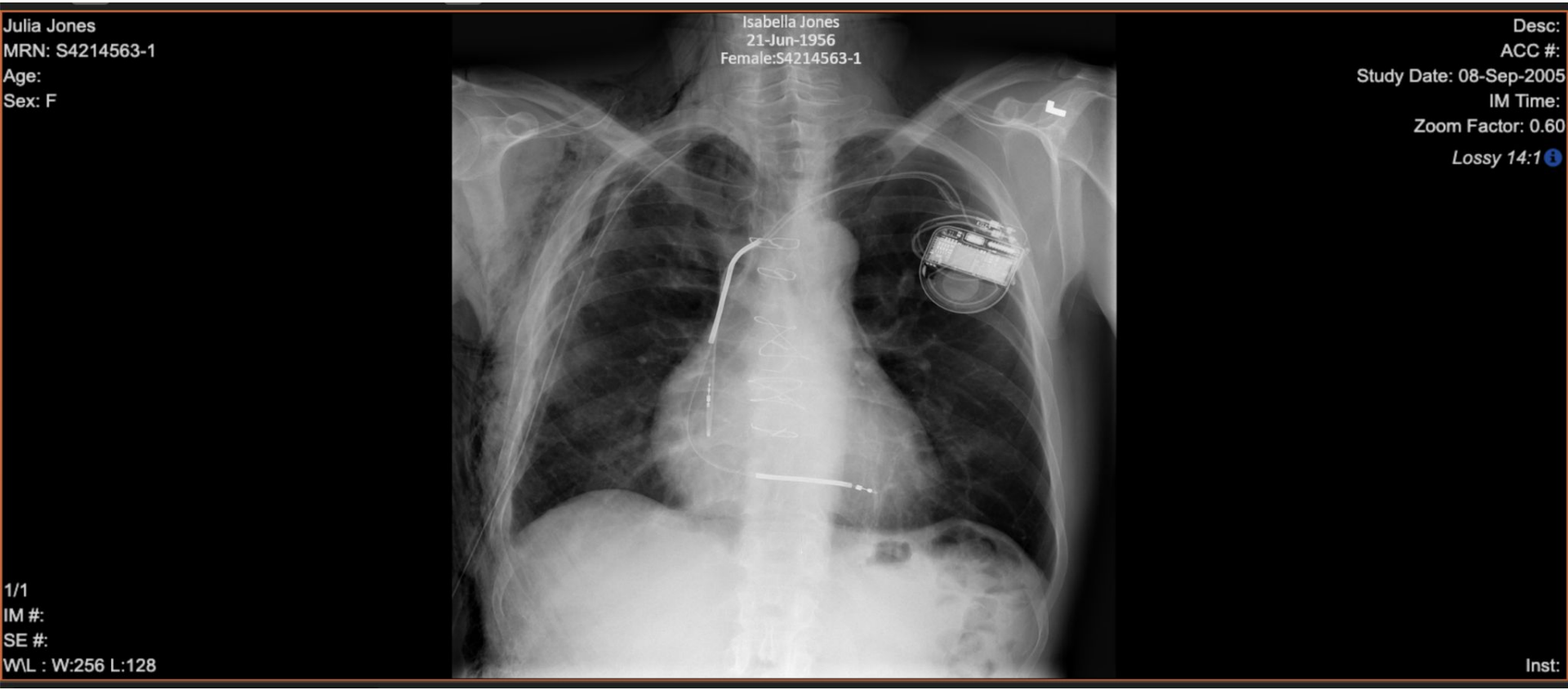
Format Preserving Encryption or Tokenization

Output → “This is my phone number: (431)582-6528”

De-ID Healthcare data

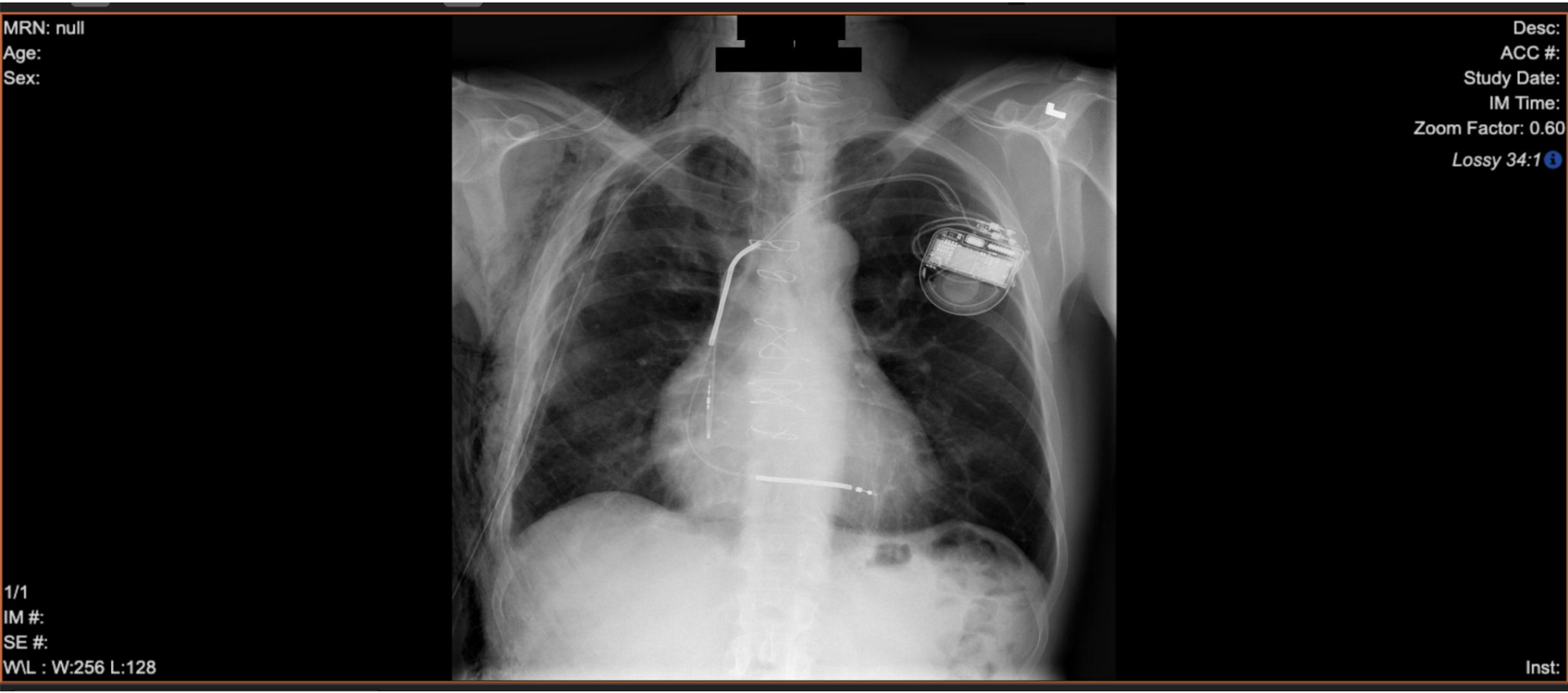


De Identification for images



The first image shows an x-ray image with sample PII and PHI data appearing in both metadata and burned-in text.

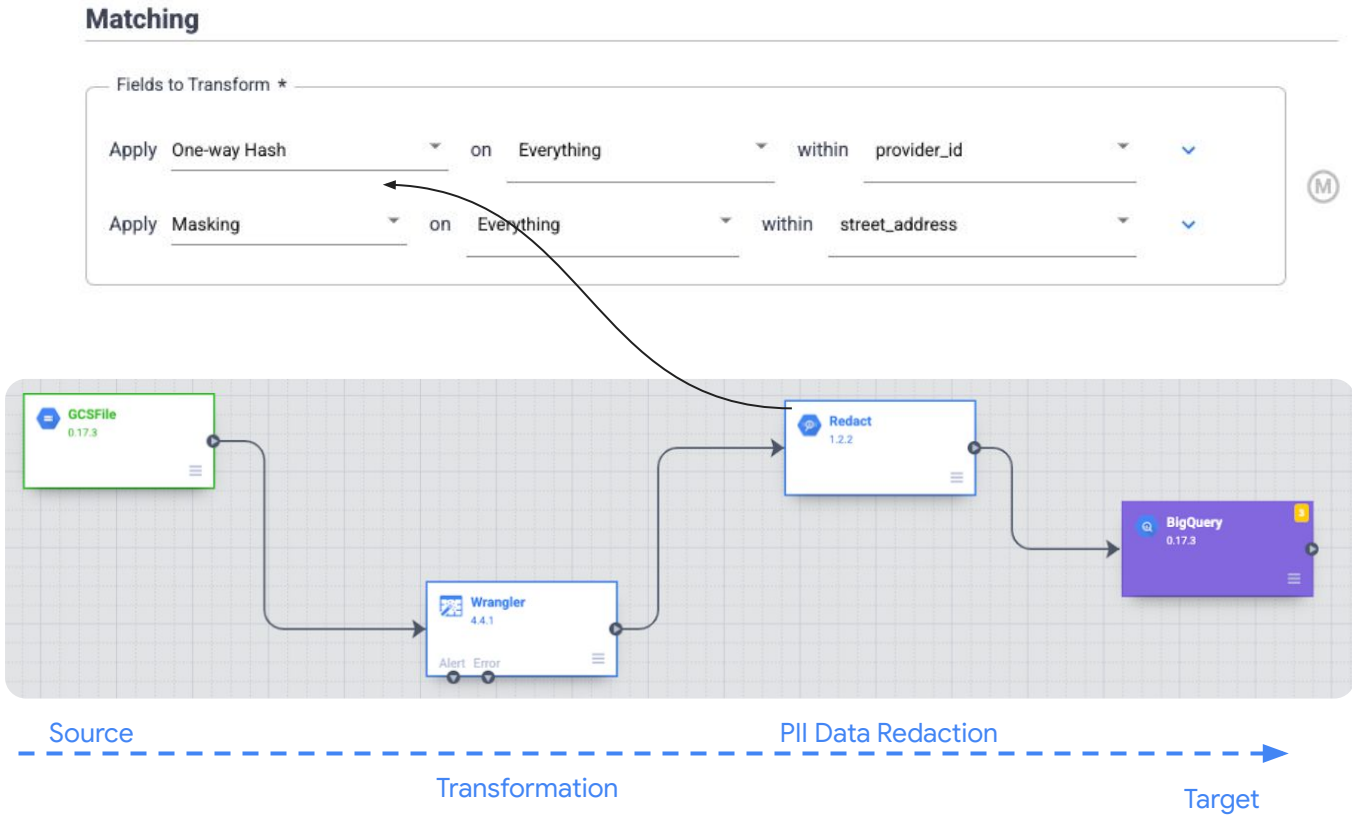
De Identification for images



The second image shows the same x-ray image with all sample PII and PHI metadata removed or obscured.

Data Fusion - De Identification - ETL Pipelines

Cloud Data Fusion is a fully managed, code-free data integration service that helps efficiently build and manage ETL/ELT data pipelines.



Data Fusion - Source to Target

provider_id	agency_name	street_address	city	state	zip_code	total_episodes_r	distinct_users_n	total_hha_charg
337290	AMERICARE CE	5923 STRICKLA	BROOKLYN	NY	11234	3148	2310	14112445
58217	BRADBOURNE	16029 ARROW	HAWINDALE	CA	91706	620	297	1786721
58419	INSIGHT HEALT	500 S KRAEMEI	BREA	CA	92821	19	15	127300
108167	D & N HOME HE	1140 W 50 ST S	HIALEAH	FL	33012	208	63	1399981
109023	PSN HEALTH C	17670 NW 78 A\	MIAMI GARDEN	FL	33015	59	18	460554
109029	RAINBOW HOM	15165 NW 77 A\	MIAMI LAKES	FL	33014	127	54	554558
109181	MARTINS FLOR	6501 NW 36ST,	VIRGINIA GARD	FL	33166	297	77	1586487
109383	LIFETIME HOMI	8785 SW 165TH	MIAMI	FL	33193	218	134	1002418
109437	SUPREME HOM	8910 MIRAMAR	MIRAMAR	FL	33025	100		
109589	ADVANCE HOM	9835 SW 72ND	MIAMI	FL	33173	165		
109694	POTENTIAL HO	9560 SW 107 A\	MIAMI	FL	33176	76		
178089	MERCY HOME	2102 EAST 21S	WICHITA	KS	67214	70		
248048	BARNABAS HE	223 WASHINGT	BRAINERD	MN	56401	47		
368049	OMNICARE HO	24800 CHAGRI	BEACHWOOD	OH	44122	19		

Hashing or Tokenizing

Masking

GENE_DI... GENE_DI... MEDICARE

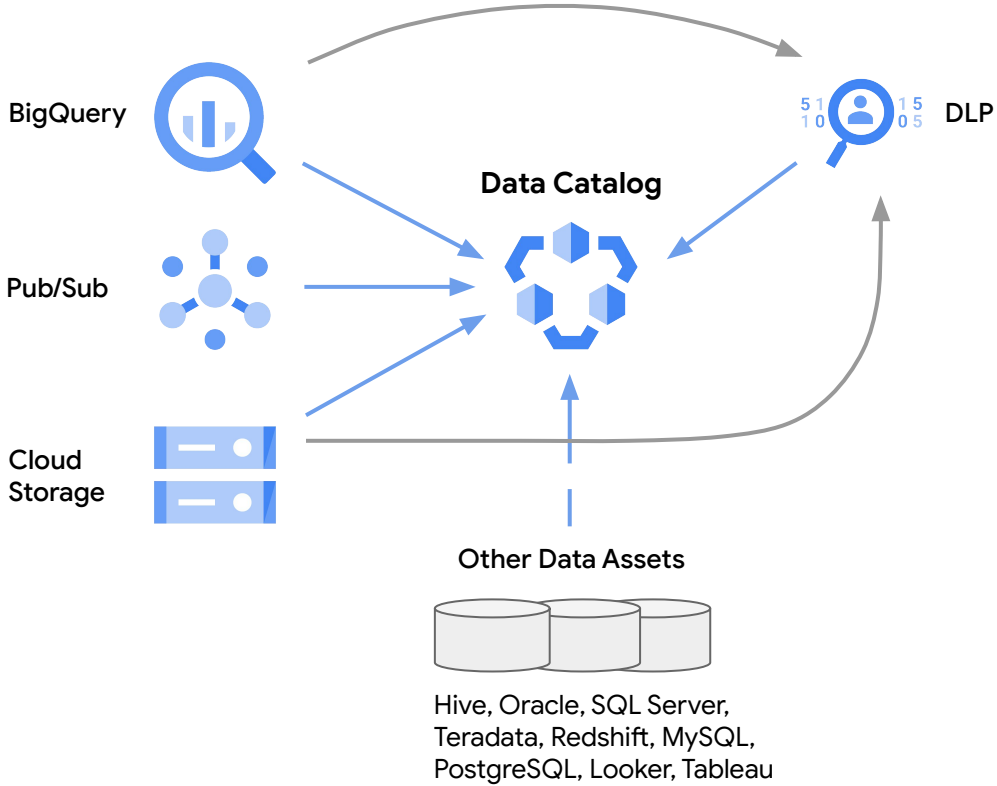
medicare

SCHEMA DETAILS PREVIEW

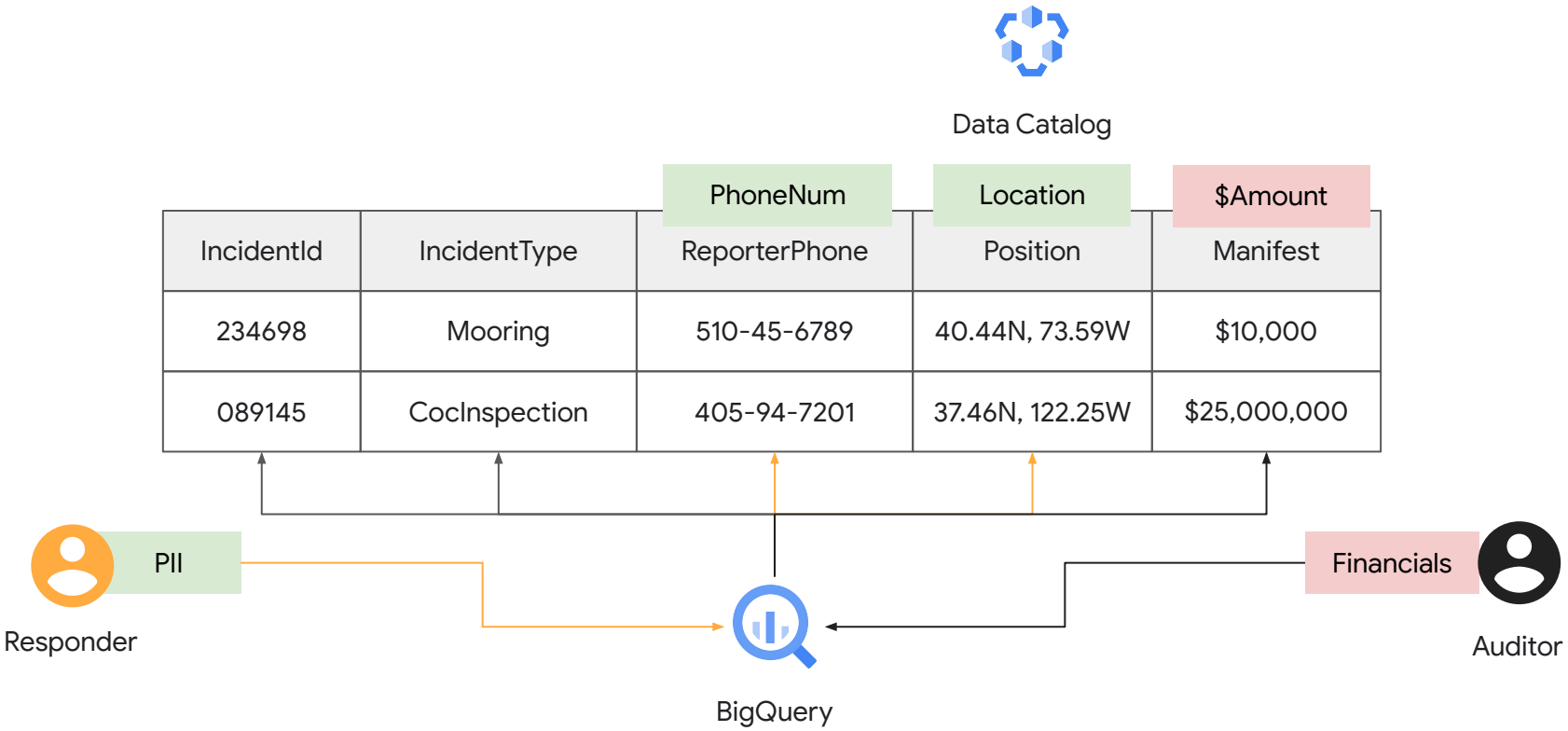
Row	provider_id	agency_name	street_address	city
1	iFTGFGYM4NrkSFKatwpCddis9rUXeyLtRIDhZK/9njE=	CUIDADO CASERO EL GIGANTE	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	ADJUNTAS
2	aZ2srYlPPl3Bcp2lp9NHqrQ21Zflk/6oMY5dj01/fZU=	DD HOME CARE SAN FRANCISCO DE ASIS, INC	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	AGUADILLA
3	ZUwJS/VgYVl0Q0yB5laG7Z5VmPY62PnRrgd06oDH0cQ=	ST LUKES HCP	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	AGUADILLA
4	f021vitE1ui/b80DTor8uvrA0Rs5B82HgSrYPho9FCk=	ARECIBO MEDICAL HCP INC	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	ARECIBO
5	Fj3p1ixSbGG35cd6l+1zZRzjJaWt/ryKQoLsmlTMXQc=	ST LUKES HCP	xxxxxxxxxxxxxxxxxxxxxxxxxxxx	ARECIBO
6	d+fxOui4jvVZbHqUz6ayjtxtU3VO4ABoySCH4cn4rHc=	ST LUKES HCP	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	LARES
7	7hmcaxKjHyEl29FMegMQBYfKm5HYdZF2BHj8rplruBo=	CORPORACION DE LAS VEGAS, INC	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	MANATI
8	aZhdmA/iS7blRm075ggykYRGfHGFRedwGWH8TVVBuVE=	LUZ DE ESPERANZA HOME CARE INC	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	ALT DE LUCH
9	2y00N0aSJELBT5W1scWaEe8L3w54rKkkcO/8hbvsCb8=	ST LUKES HCP	xxxxxxxxxxxxxx	MAYAGUEZ
10	mUKleYlBPFfCgLX13ilK9NSNGjFlre1T8KEuXQZwrvQ=	FIRST HOME CARE CENTER INC.	xxxxxxxxxxxxxxxxxxxxx	MAYAGUEZ
11	3HZF1kAWRz/L5YTf5X7xkLRv/d7d1QzWM5hMP7tptfc=	CUIDADO CASERO MAYAGUEZ INC	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	MAYAGUEZ
12	q4NJEWdzJjSY2FpCll1Bg3j5kWJDJOl2LJ5wOHpvv1k=	VISITING HEALTH SERVICES PROGRAM HOSPITAL DAMAS	xxxxxxxxxxxxxxxxx	PONCE
13	Jv/YYe4uNlO8AunCVGDXpgAcc1Qf0nLyDl5T248FFUM=	ATENCION MEDICA EN EL HOGAR INC	xxxxxxxxxxxxxxxxxxxxxxx	CAGUAS

Auto Tag sensitive data - Dataplex - Data Catalog

1. Auto-syncs ‘technical metadata’ from GCP data assets in near real-time
2. **Auto-tags PII data** through DLP integration
3. Supports non-GCP data assets through open-source connectors



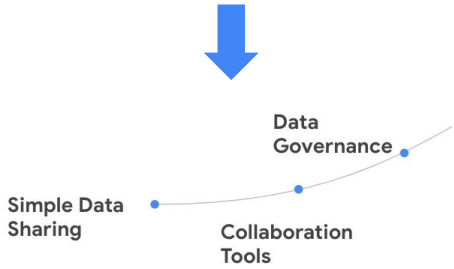
BigQuery - Using Metadata tags to restrict access



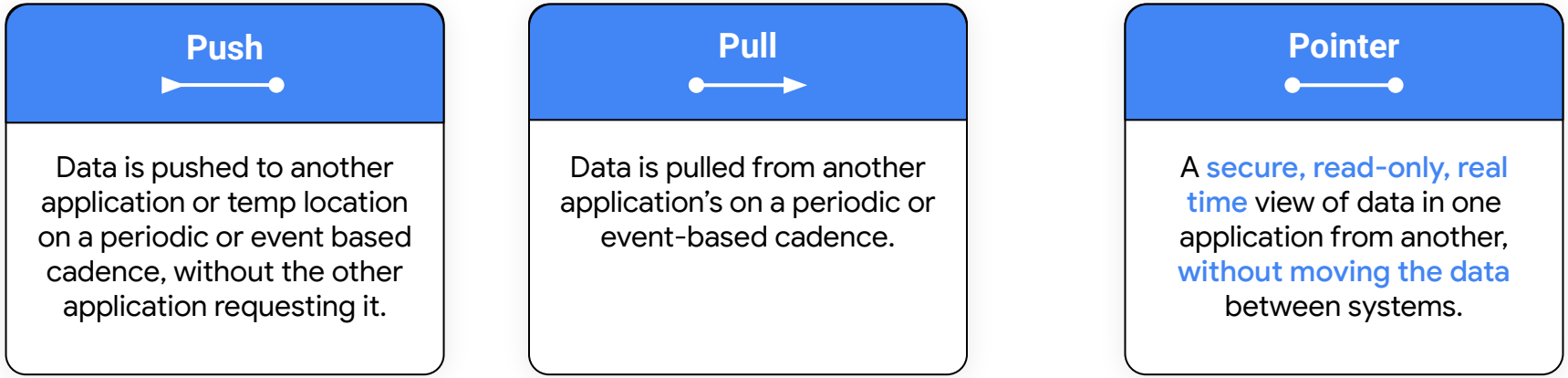
Analytics Hub

Data Clean Rooms

Privacy-centric data sharing and analysis



The Evolution of Data Sharing



Traditional Model

Data is copied, adding network and storage costs

Asynchronous feeds: Incremental data is complex

Set-up and governance is shared and can be fragile

No visibility of data usage once shared

Next Gen Model

Data remains in place, reducing costs

Updates are available in realtime

User friendly Publish/Subscribe model

Usage Metrics automatically available

Analytics Hub

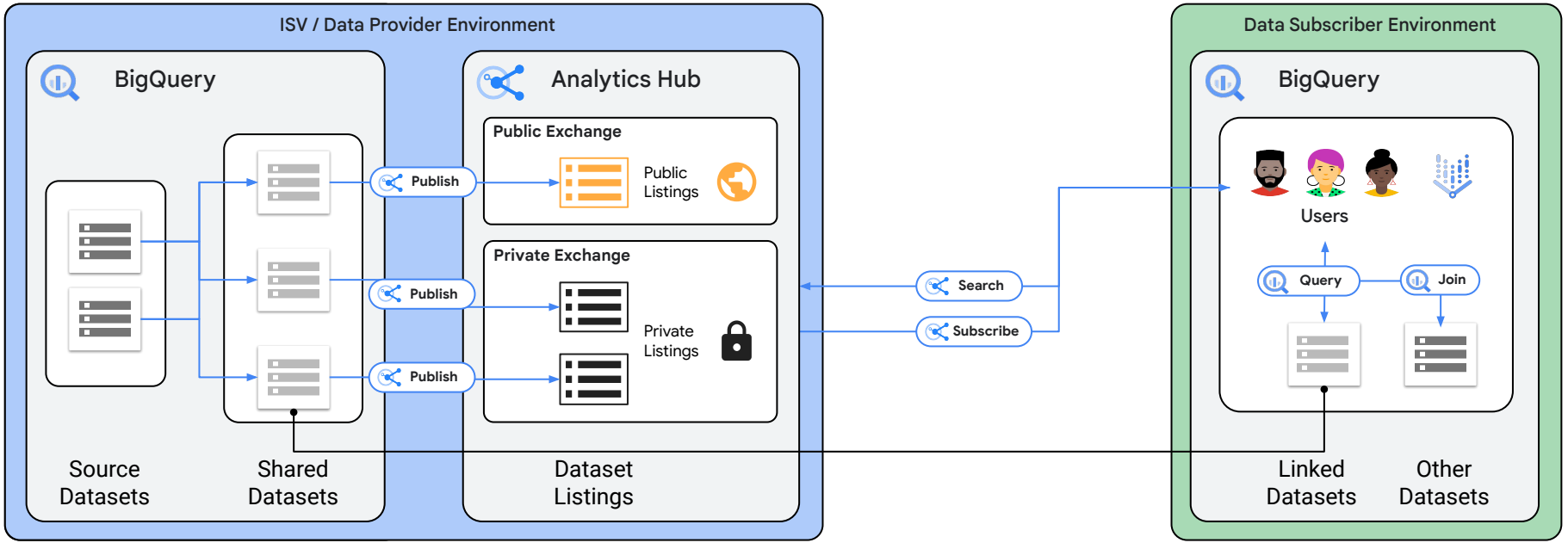
Simple, secure, zero-copy data and analytics assets sharing

via BigQuery

Analytics Hub Ecosystem

250+ data providers | 2,900+ listings | Over **350Pb** Shared p/w

Data Sharing Architecture



Shared Datasets are collections of tables and views defined by the **Data Provider** and are the unit of cross- project / cross-org sharing.

Exchanges are collections of **Shared Dataset Listings**. Exchange visibility can be Private, Public, Internal, or Restricted. Permissions are set by the Exchange Admin.

Data Subscribers can search across **Exchanges** and Subscribe to **Dataset Listings** that they are authorized to.

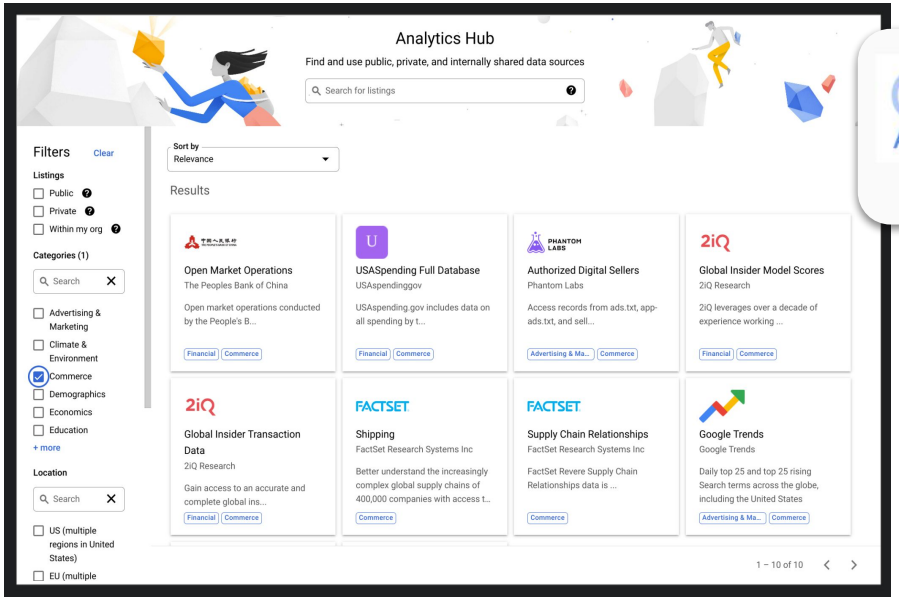
Data Subscribers get an opaque, read-only **Linked Dataset** inside their project that they can query and combine with their own datasets.

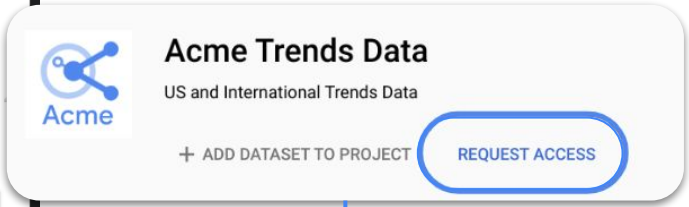
Providers pay for storage

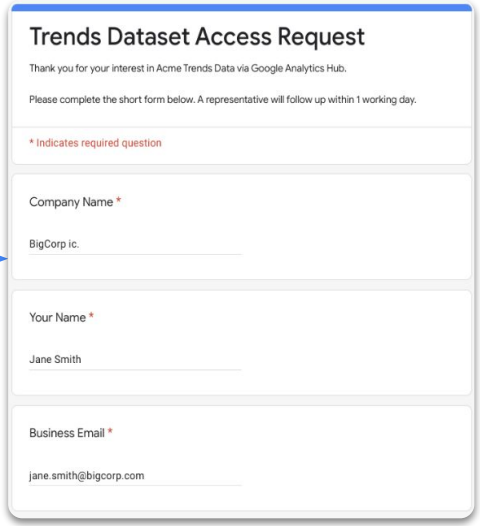
No cost for Exchanges & Listings

Subscribers pay for queries

End User (Researcher) Searches and requests access to subscribe to the data.

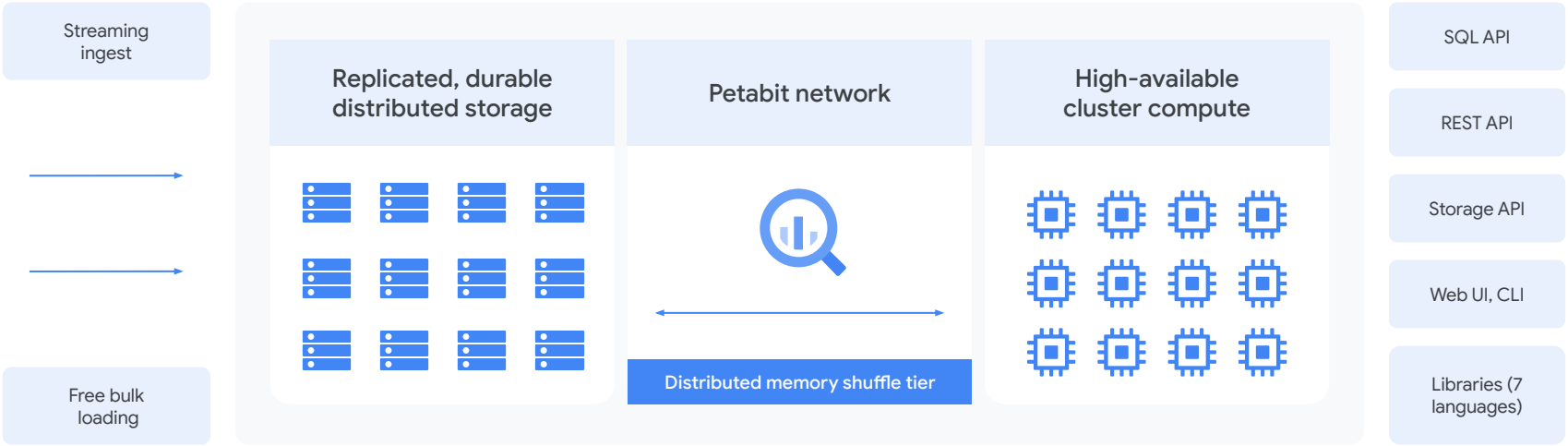






Building on BigQuery

Decoupled storage and compute for maximum flexibility



BigQuery sees more than **6000+ organizations** sharing over **275+ PB's** of data per week.

A data clean room is a secure collaboration environment which allows two or more participants to leverage data assets for specific, mutually agreed upon uses, while guaranteeing the enforcement of strict data access limitations, e.g., not revealing or exposing the personal data of their customers to other parties.

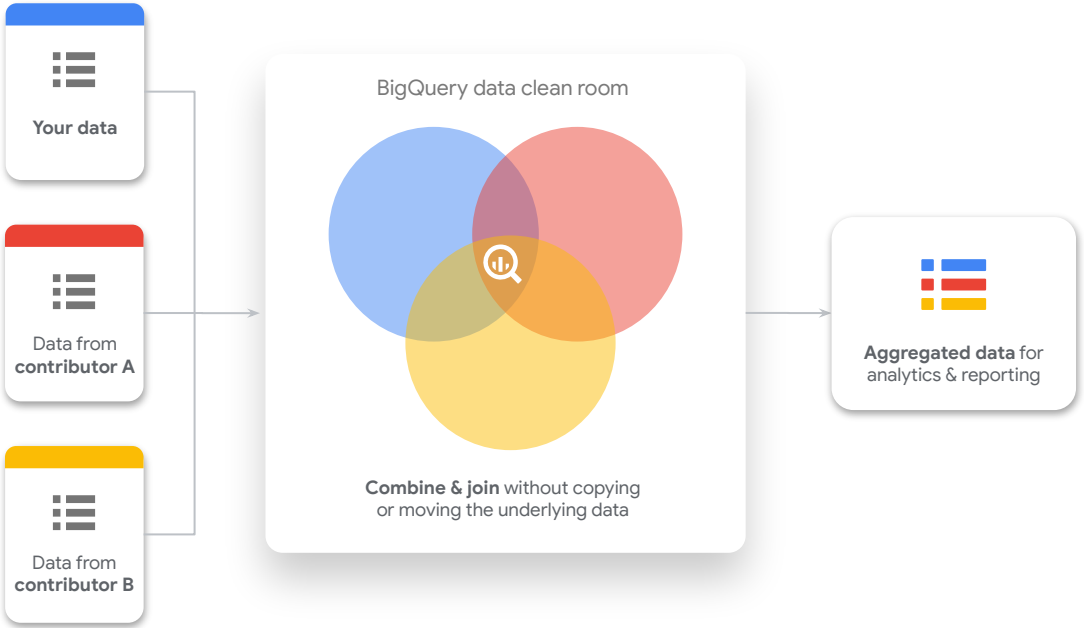
Jeffrey Bustos
VP. Measurement, Addressability & Data Center at IAB

Source: iab, April 2023

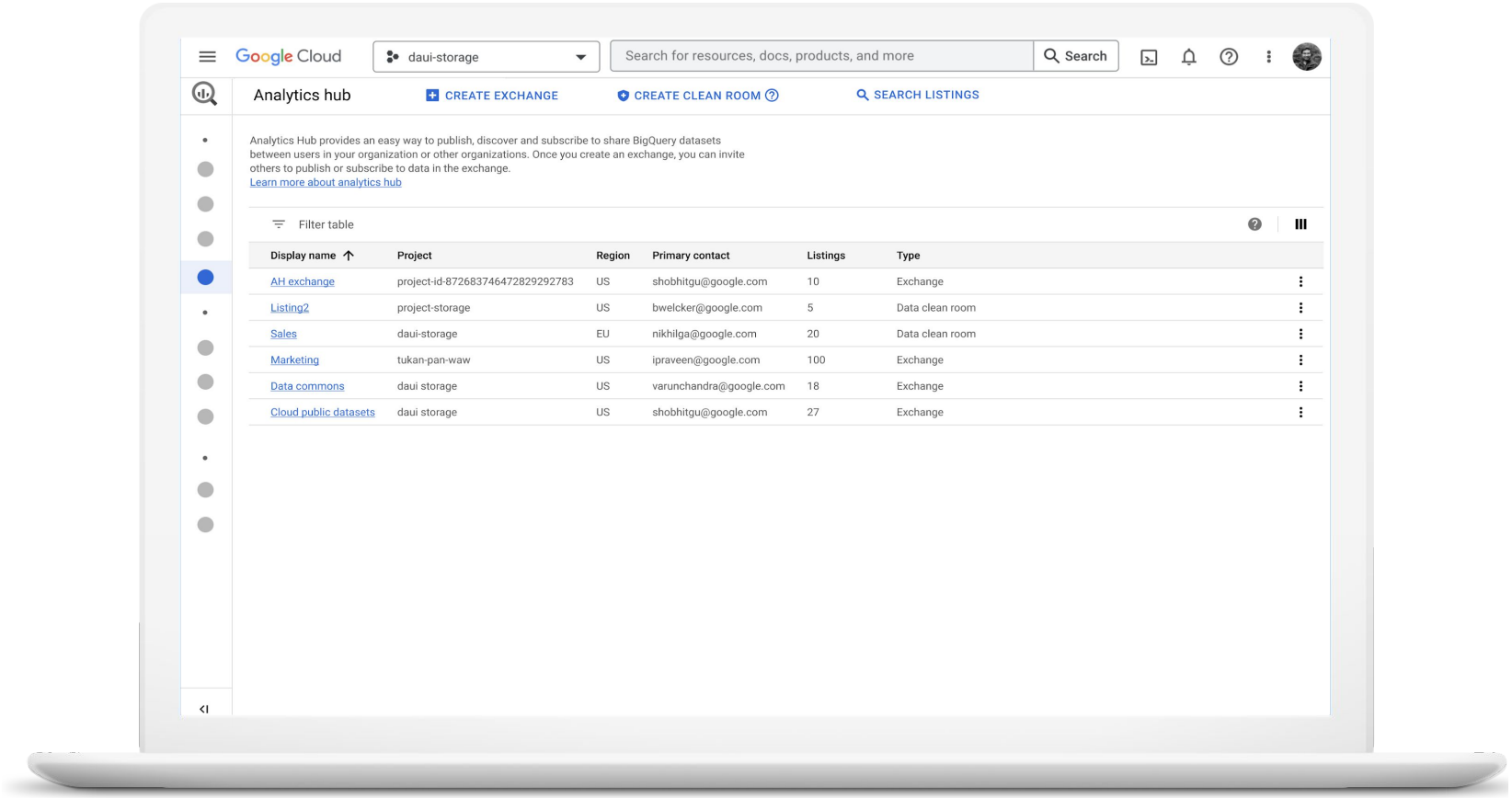
BigQuery - Data clean rooms

Combine, analyze, and securely share sensitive data in a privacy-centric way.

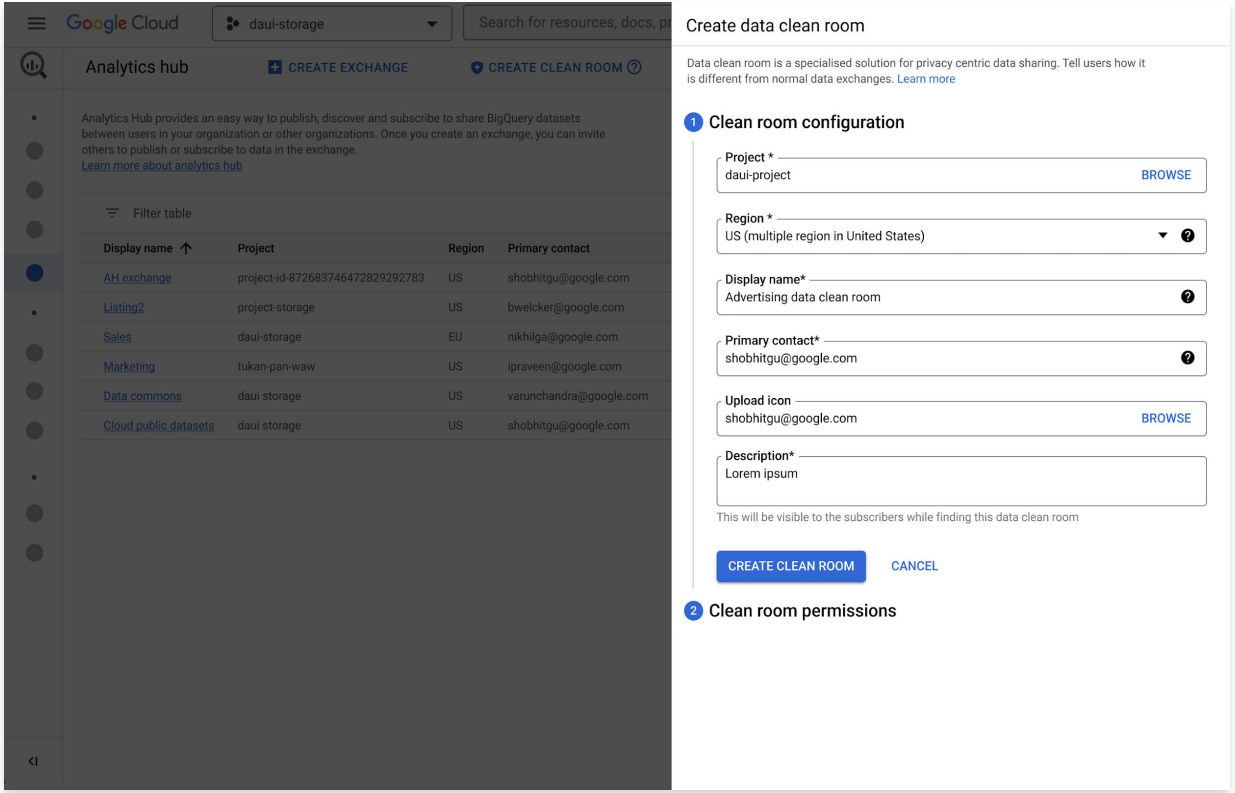
- Create a low-trust environment for you and your partners to collaborate without copying or moving the underlying data.
- Perform privacy-enhancing transformations in BigQuery SQL interfaces.
- Monitor usage to detect privacy threats on shared data.



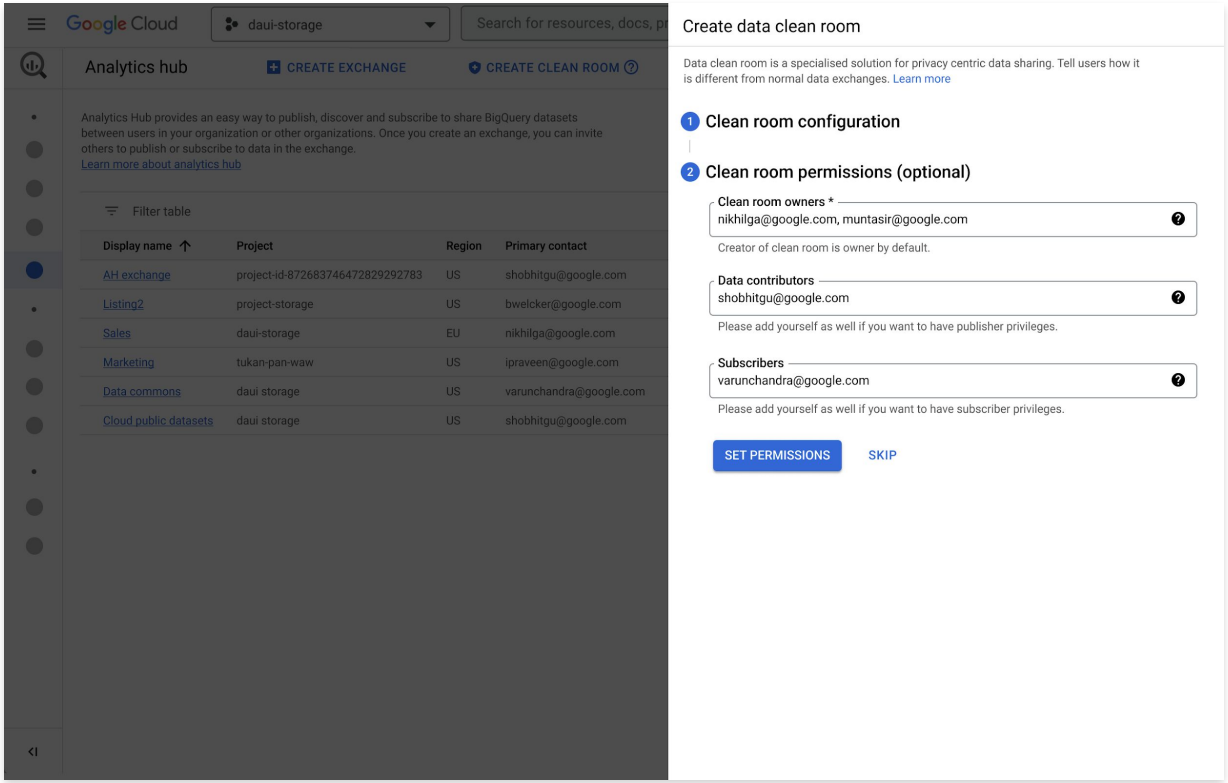
Data Clean Room - Walkthrough



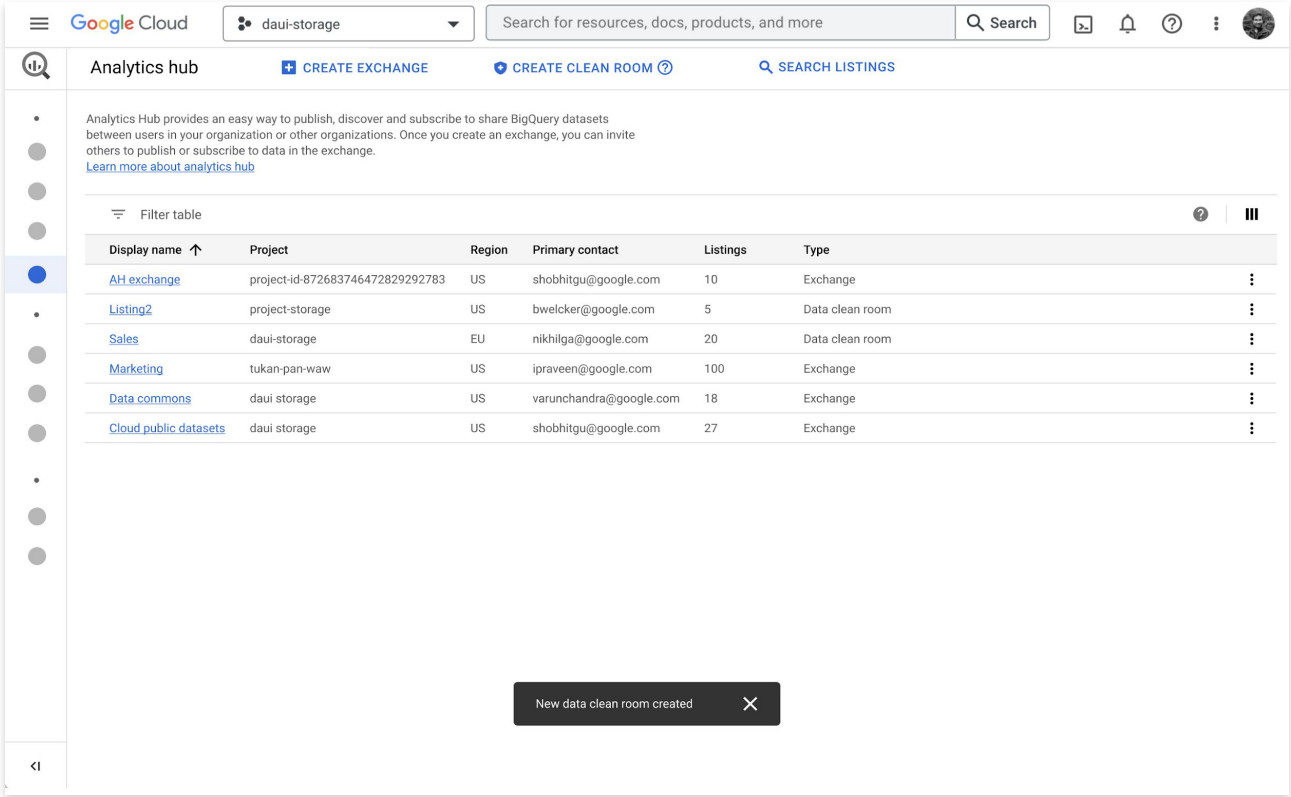
Create data clean room



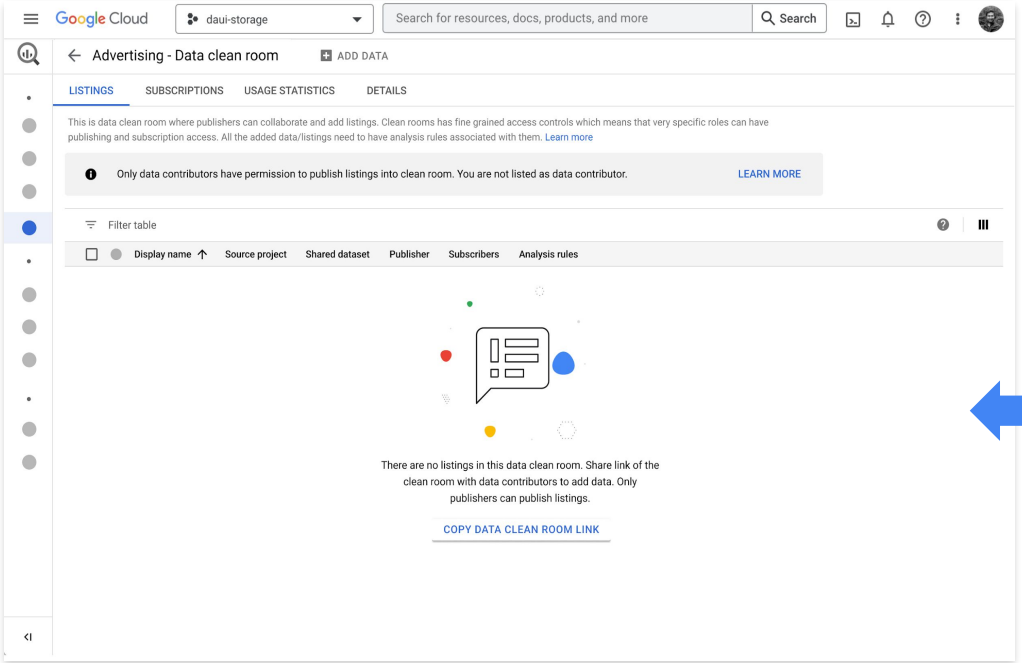
Set clean room permissions

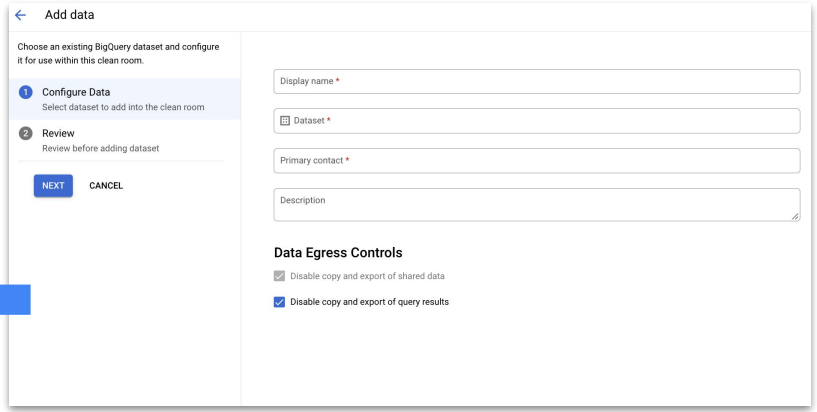


Clean room created

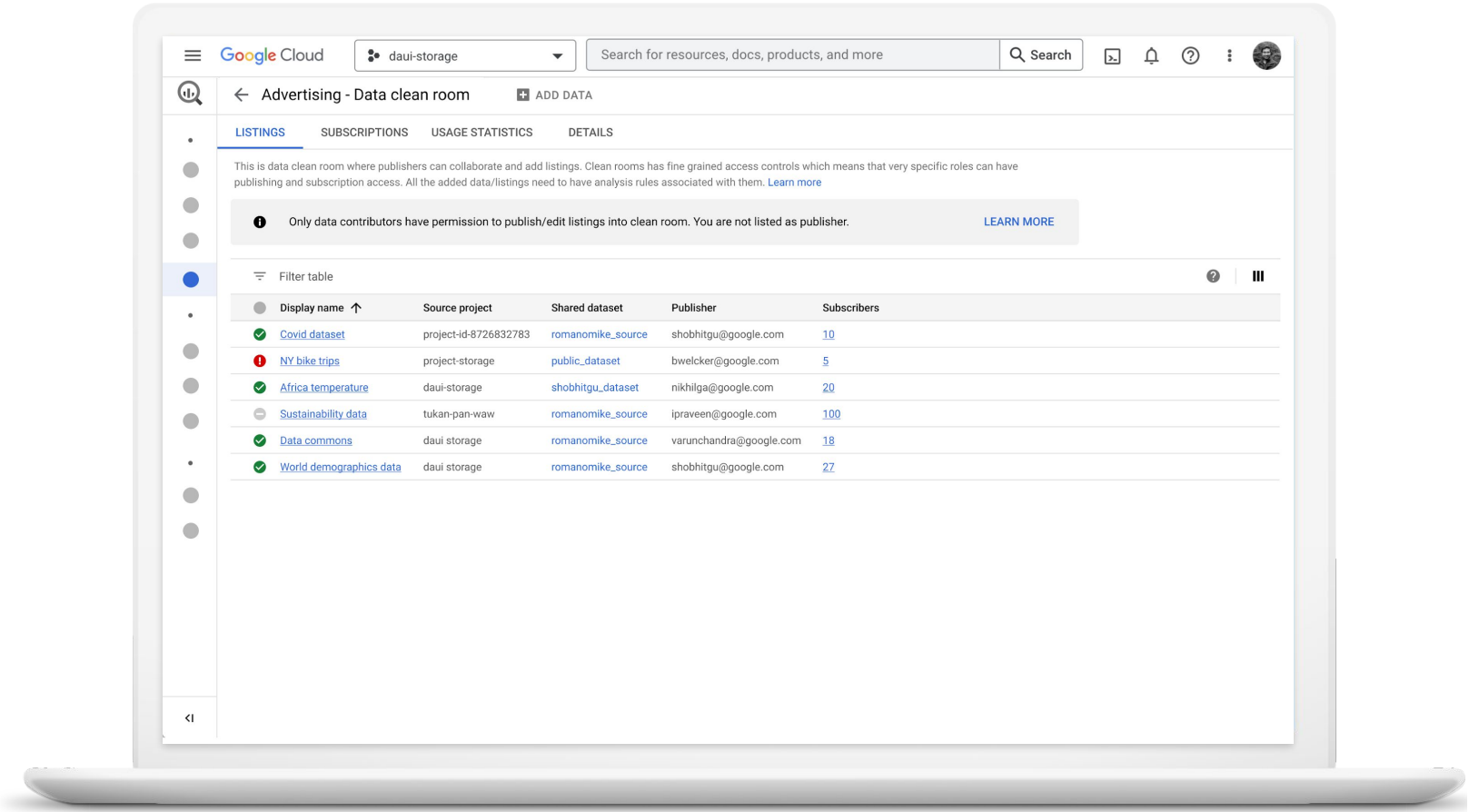


First time view - Add your data

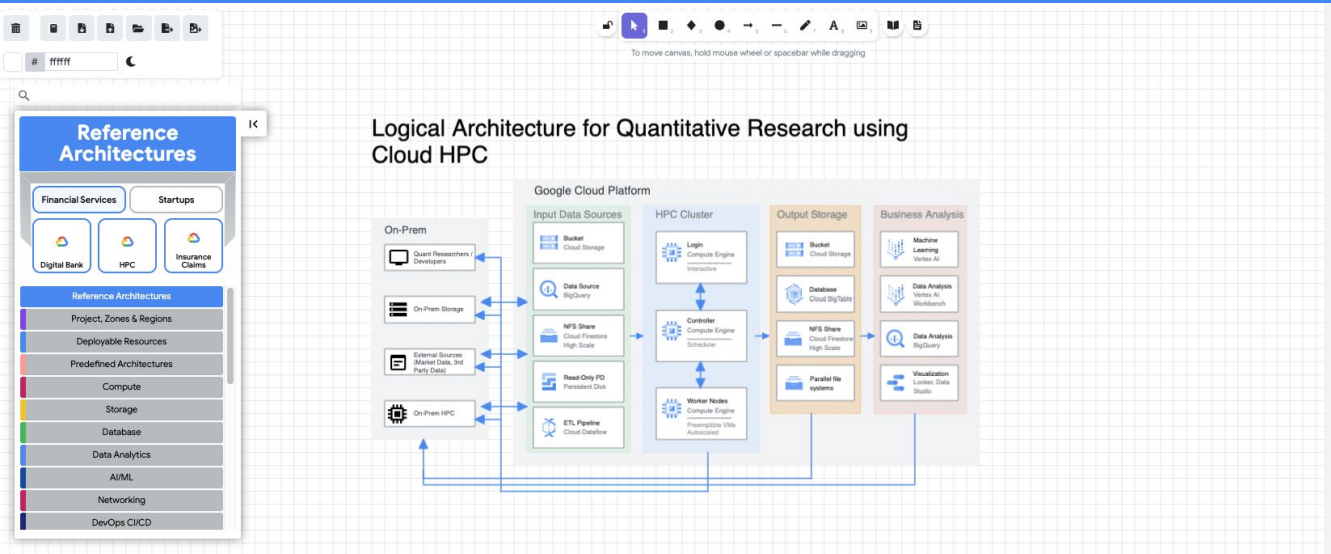




Published listings in a clean room



Blue prints


To move canvas, hold mouse wheel or spacebar while dragging
Reference Architectures
Financial Services
Startups
Digital Bank
HPC
Insurance Claims
Reference Architectures
Project, Zones & Regions
Deployable Resources
Predefined Architectures
Compute
Storage
Database
Data Analytics
AI/ML
Networking
DevOps CI/CD
Logical Architecture for Quantitative Research using Cloud HPC
Google Cloud Platform
On-Prem
Input Data Sources
HPC Cluster
Output Storage
Business Analysis

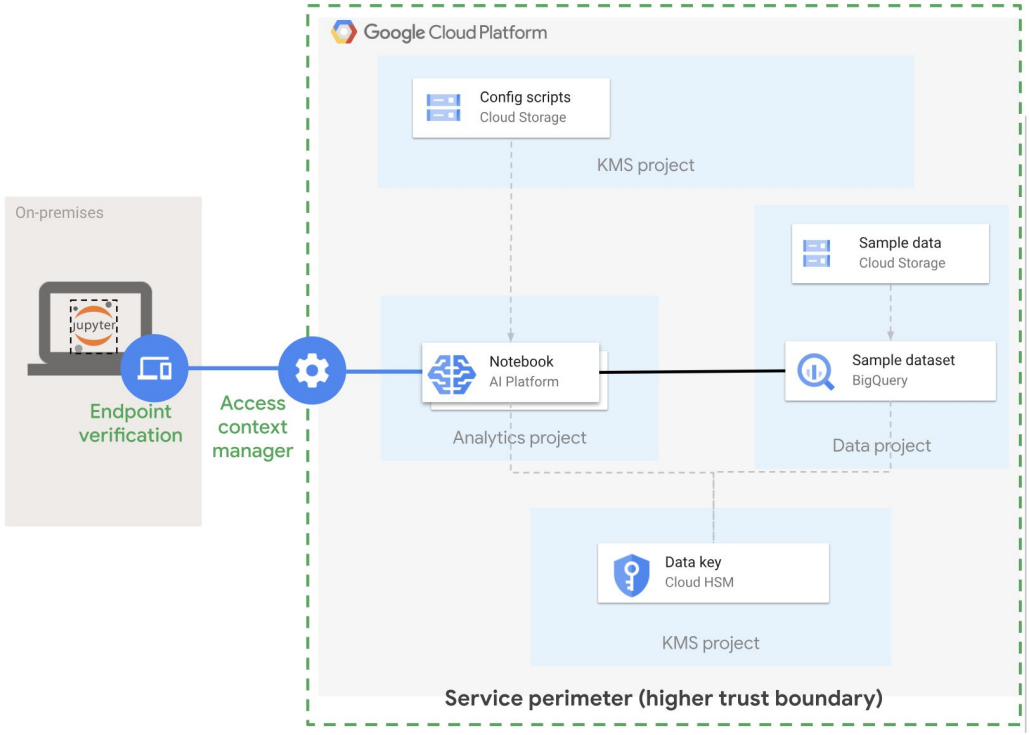
Secure Notebook and Perimeter controls

Blueprint

Protecting confidential data in Vertex AI Workbench user-managed notebooks

The architecture also creates security controls that help you to do the following:

- Mitigate the risk of data exfiltration to a device that is used by data scientists in your enterprise.
- Protect your notebooks instances from external network traffic.
- Limit access to the VM that hosts the notebook instances.



[GitHub repository]

Secure Data Warehouse

Blueprint

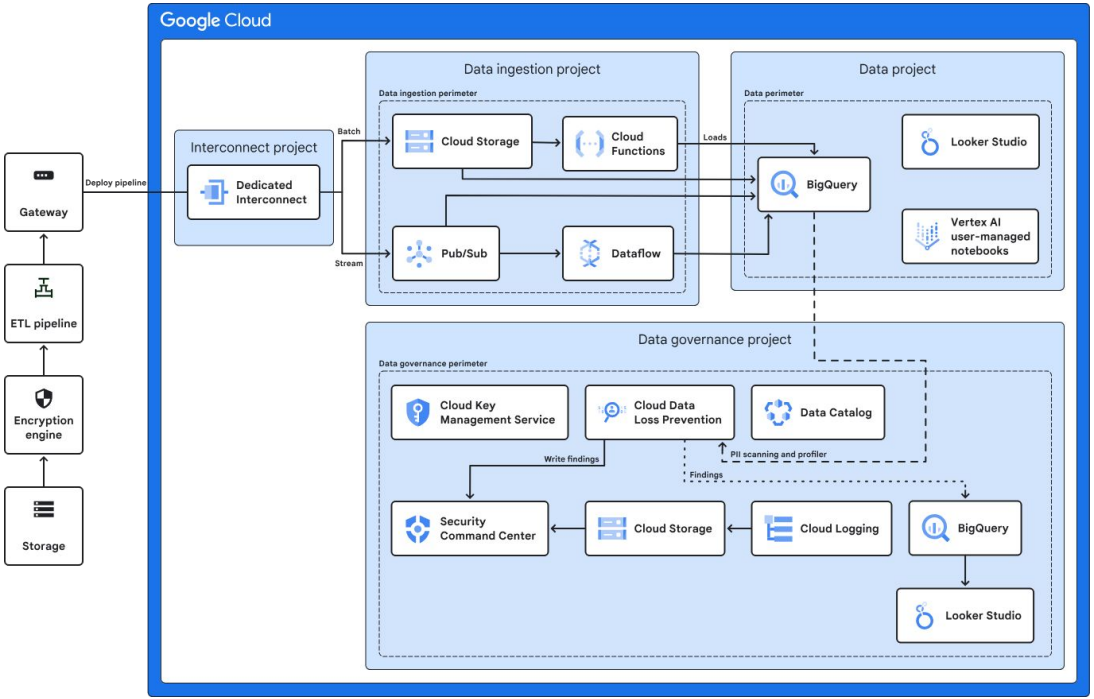
Secure Data Warehouse

If your Project includes confidential data, you must take measures to preserve the security, confidentiality, integrity, and availability of the business data while it is stored, while it is in transit, or while it is being analyzed.

The blueprint helps

- Configure controls that help secure access to confidential data.
- Configure controls that help secure the data pipeline.
- Configure an appropriate separation of duties for different personas.
- Set up templates to find and de-identify confidential data.
- Set up appropriate security controls and logging to help protect confidential data.
- Use data classification and policy tags to restrict access to specific columns in the data warehouse.

Secure Data Warehouse Blueprint



LEARN
Tutorial

Deploy a secured data warehouse to store confidential data

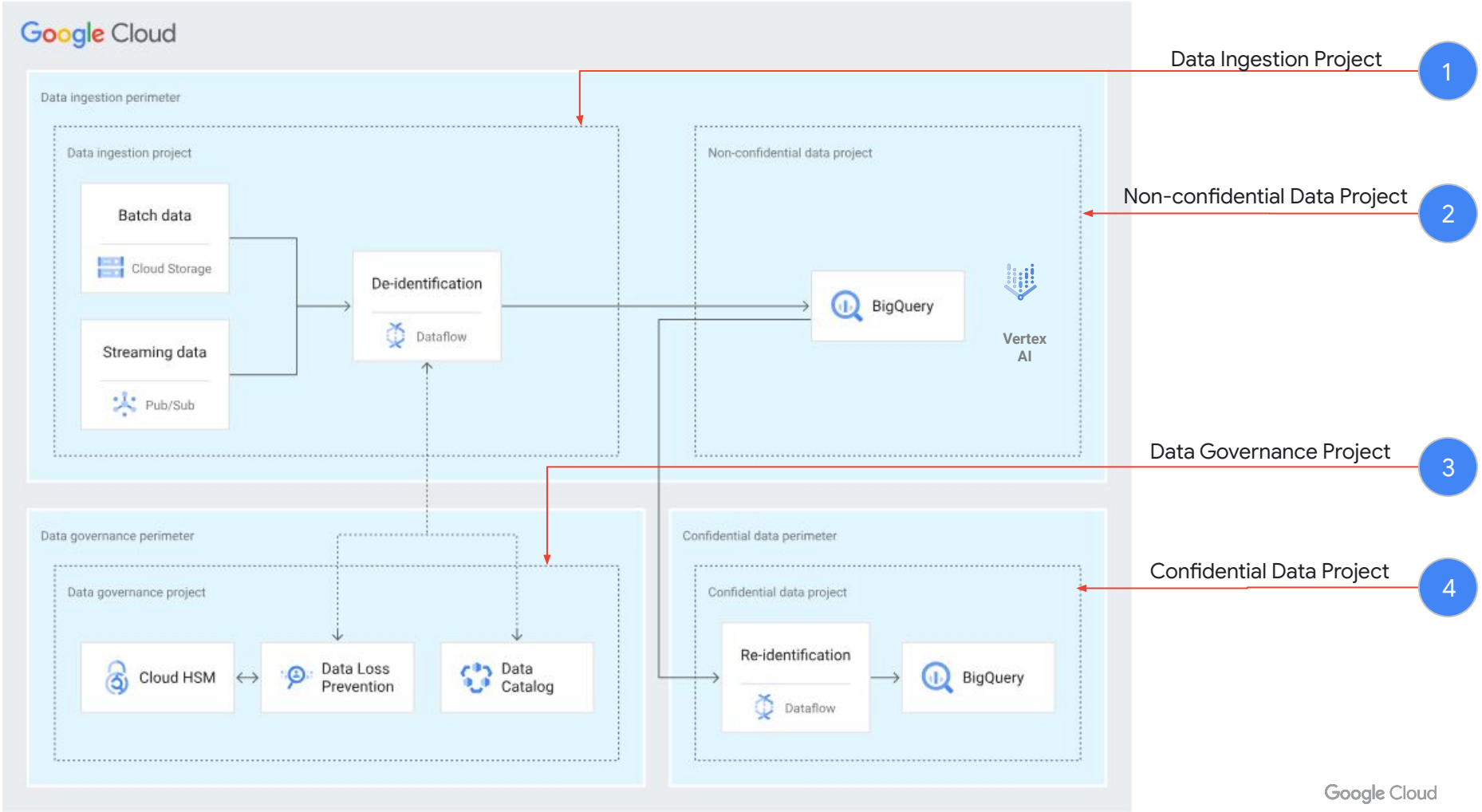
Learn how to use Terraform to deploy a demo of a secured BigQuery data warehouse that can store confidential data. This tutorial describes how to do the following:

1. Set up your environment (projects, service accounts, groups, and so on).
2. Deploy the Terraform code required to create a BigQuery data warehouse.
3. View the security controls in the deployed environment.
4. Clean up to avoid billing charges.

Before you start

1. Verify that your user identity has the `iam.serviceAccountUser` and `iam.serviceAccountTokenCreator` roles for your organization's development folder, as described in Organization structure. If you do not have a folder that you use for demos

START

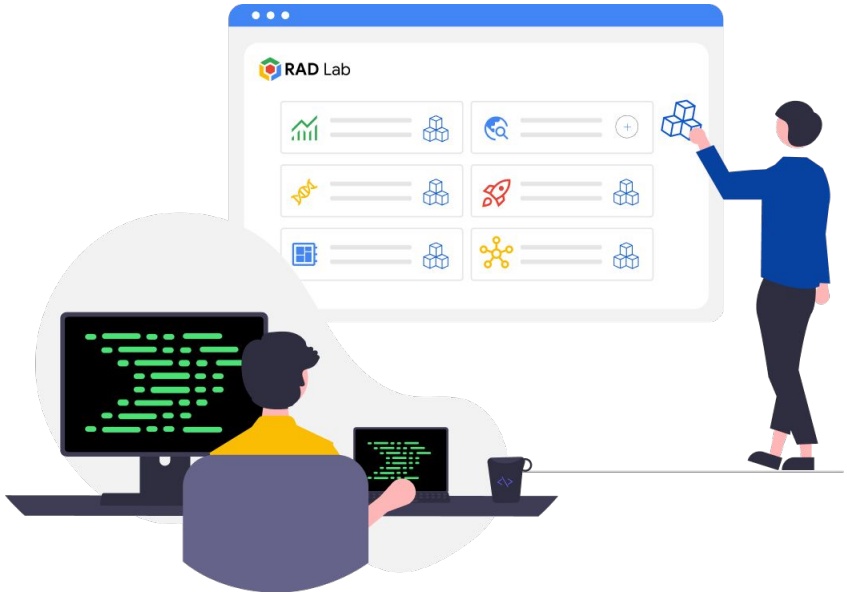

Google Cloud
Data ingestion perimeter
Data ingestion project
Batch data
Cloud Storage
Streaming data
Pub/Sub
De-identification
Dataflow
Non-confidential data project
BigQuery
Vertex AI
Data governance perimeter
Data governance project
Cloud HSM
Data Loss Prevention
Data Catalog
Confidential data perimeter
Confidential data project
Re-identification
Dataflow
BigQuery
Data Ingestion Project
1
Non-confidential Data Project
2
Data Governance Project
3
Confidential Data Project
4
Google Cloud

RAD Lab

Deploying Secure Data Warehouse Solution

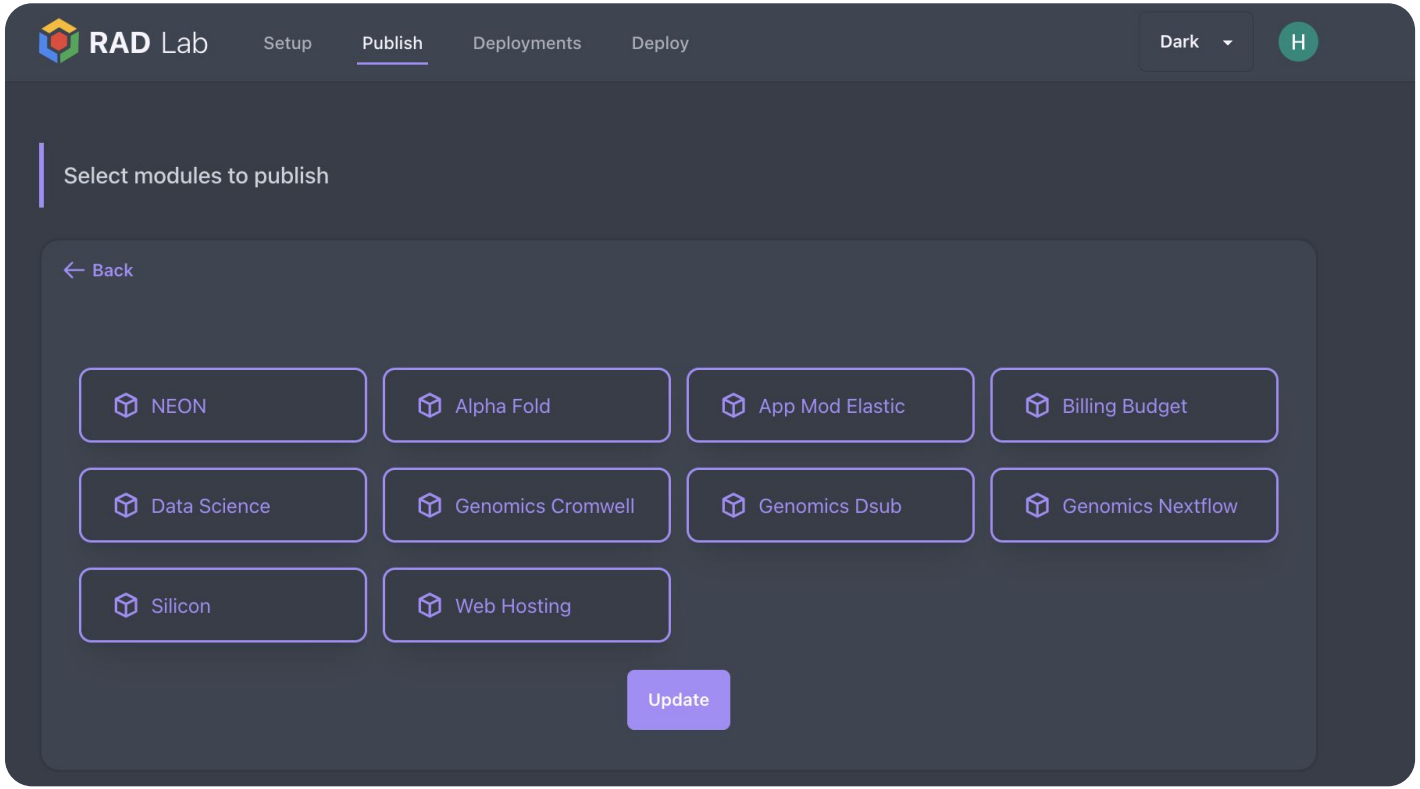
RAD Lab

Cloud admins configure **RAD Lab UI** for their organization to **enable self-service cloud deploys**

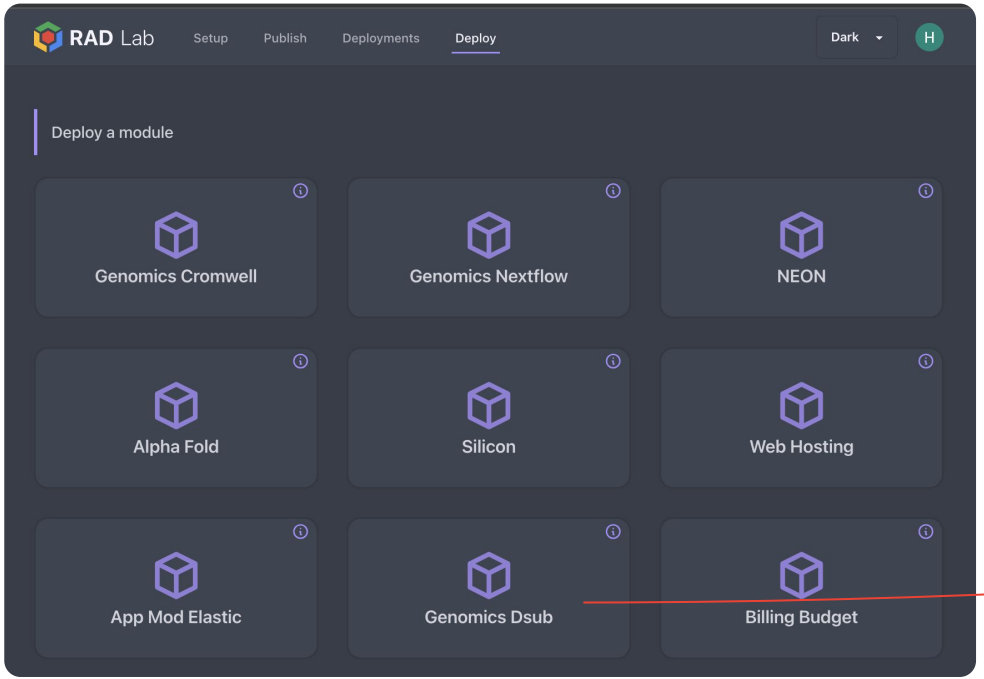


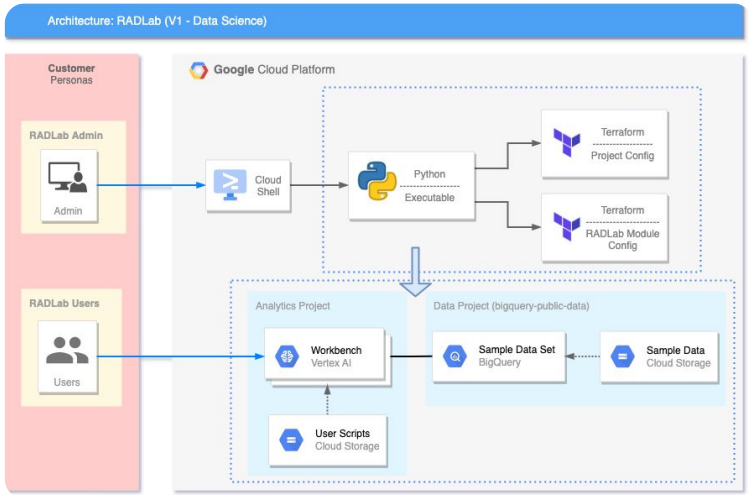
Researchers, analysts, and other professionals select a **customized cloud environment** for their work and **deploy with a few clicks**

RAD Lab UI


RAD Lab
Setup
Publish
Deployments
Deploy
Dark
H
Select modules to publish
Back
NEON
Alpha Fold
App Mod Elastic
Billing Budget
Data Science
Genomics Cromwell
Genomics Dsub
Genomics Nextflow
Silicon
Web Hosting
Update

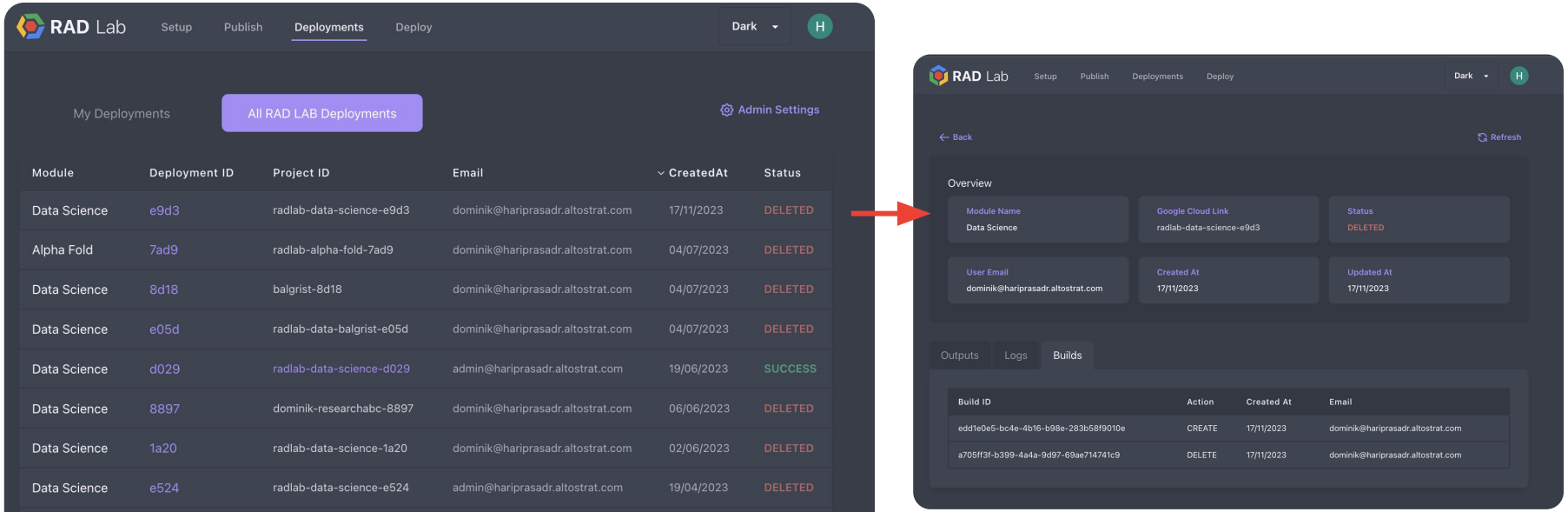
RAD Lab UI - self service model for researchers





Click to deploy

Admin - Monitor and Delete deployments



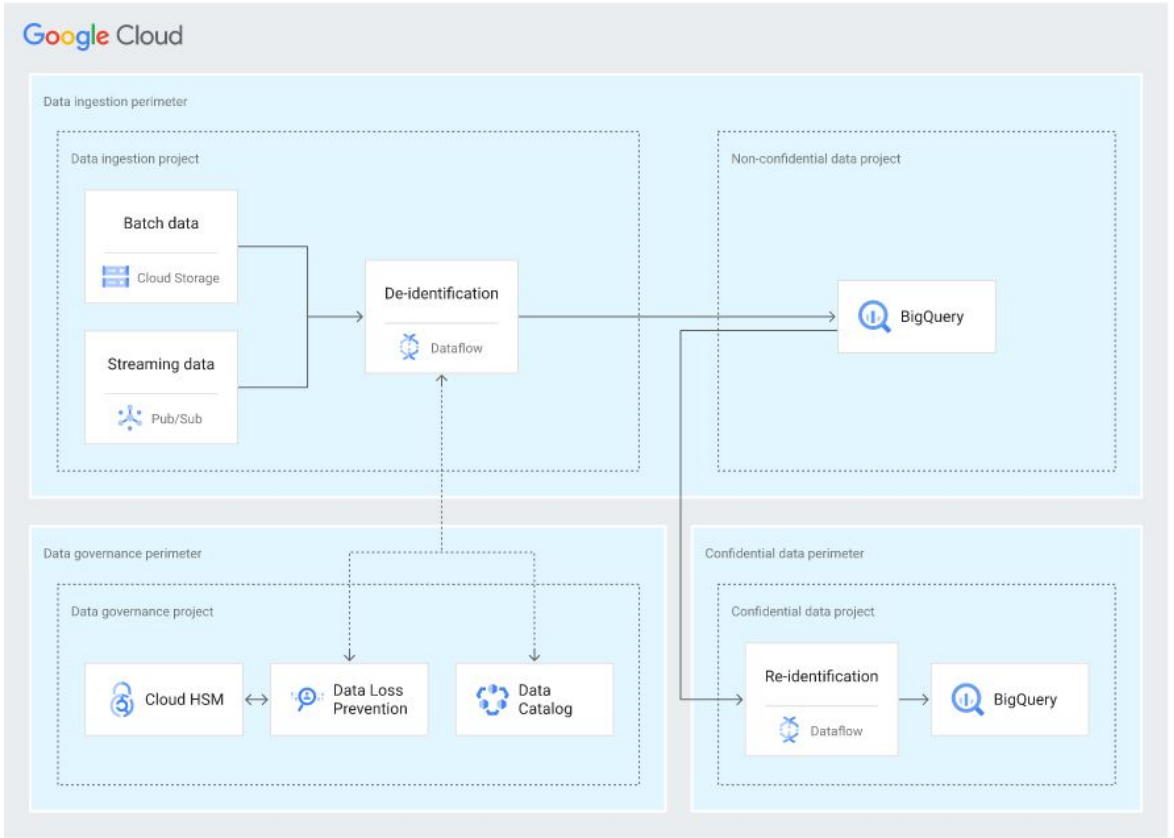
Secure data warehouse

RAD Lab

Quick deploy using RAD Lab as a self serve module

OPEN IN GOOGLE CLOUD SHELL

https://github.com/GoogleCloudPlatform/rad-lab



Thank you.

Google Cloud