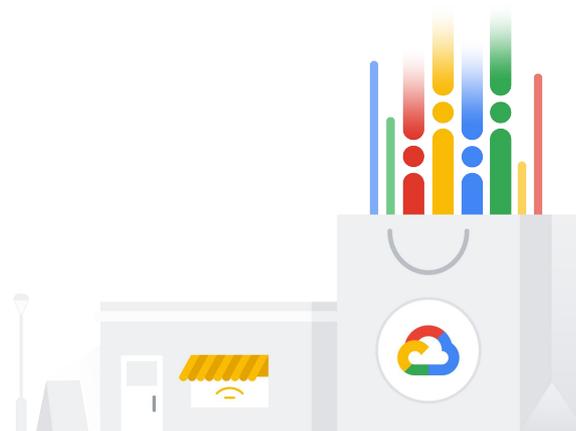# Agenda

- Alphabet / Google Overview
- GCP Healthcare & Life Sciences Overview
- STRIDES Initiative Overview & Benefits to NCI Researchers
- Questions & Answers

Google Cloud

Google Cloud

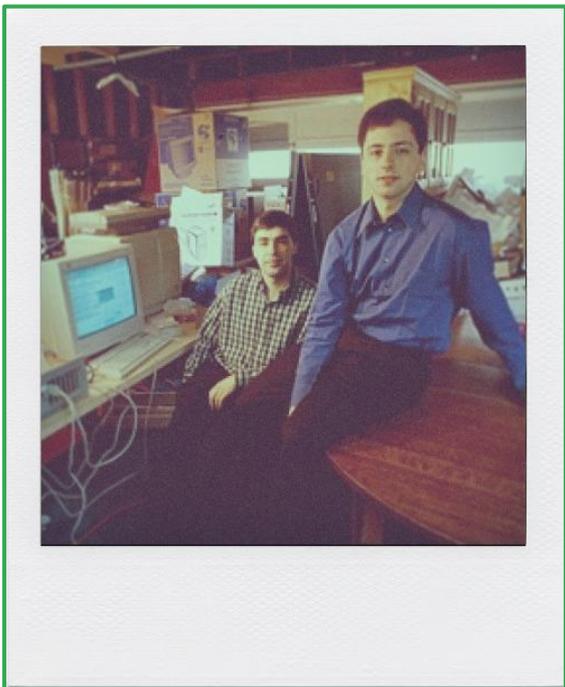# Introduction to Alphabet / Google

Alphabet / Google Overview Agenda

- 🔴 **Corporate Structure**

- 🟡 **Cloud Services & Partners**

- 🟢 **Cloud Capabilities**

- 🔵 **Products and Solutions**

Google Cloud

Google's Structure

Google in 1998   Alphabet in 2022

>160,000
Alphabet employees in
Q2 2022

$69B
quarterly revenue in
Q2 2022

150
offices in 60 countries

23%
Growth YoY

Google Cloud

Alphabet

Google Ventures
Venture & capital funding

Google X
Innovation Lab & Research

Google

Calico
Longevity Research

DeepMind
Artificial Intelligence & Machine Learning

Online Advertising Platform
Google Ads

GMP
Unified Ads & Analytics Platform

Verily
Improving Quality of Life

Nest
Connected Home Devices

Google Pay
Digital Wallet Platform

Google Cloud
Cloud Services & Workspace

SideWalk Labs
Solving Big Urban Problems

Waymo
Self Driving Vehicles

Maps
Mapping, Location Services & Logistics

YouTube
Internet Video Service
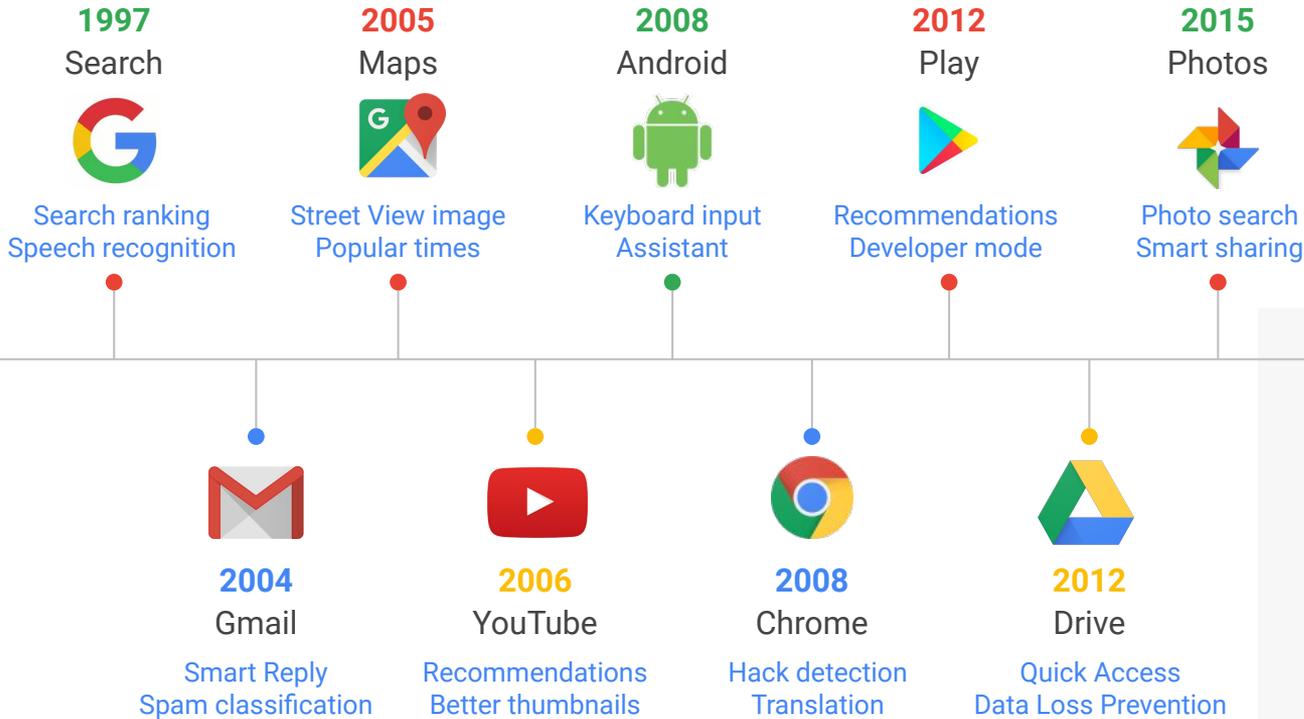
Jigsaw
Online Global Security Solutions

Google Fiber
High Speed Internet Services

Android
Mobile Operating System

Firebase
Mobile & Web No-Ops Platform

Google Cloud

Our products are **how** we innovate

1997
Search

**Search ranking
Speech recognition**

2005
Maps

**Street View image
Popular times**

2008
Android

**Keyboard input
Assistant**

2012
Play

**Recommendations
Developer mode**

2015
Photos

**Photo search
Smart sharing**

2004
Gmail

**Smart Reply
Spam classification**

2006
YouTube

**Recommendations
Better thumbnails**

2008
Chrome

**Hack detection
Translation**

2012
Drive

**Quick Access
Data Loss Prevention**

Google Cloud

9
Billion User
Applications

100%
Machine Learning
Powered

Google is a world leader in applying AI/ML to real-world situations

Search
Search ranking
Speech recognition

Translate
Text, graphic and
speech translations

Photos
Photos search

Gmail
Smart reply
Spam classification

Self Driving Car
1.5MM miles driven

Data Center Power Usage
Reduced cooling energy 40%

AlphaGo
First AI to beat a world Go champion (2016)

YouTube
Video recommendations
Better thumbnails

Google Cloud

Google Cloud Services & Partners

More than 60 Google Cloud Platform services

Compute

- Compute Engine
- App Engine
- Container Engine
- Container Registry
- Cloud Functions

Storage and Databases

- Cloud Storage
- Cloud Bigtable
- Cloud Datastore
- Cloud SQL
- Cloud Spanner
- Persistent Disk

Networking

- Virtual Private Cloud
- Cloud Load Balancing
- Cloud CDN
- Cloud Router
- External IP
- Cloud Interconnect
- Cloud DNS
- Networking
- VPN

Big Data

- BigQuery
- Cloud Dataflow
- Cloud Dataproc
- Cloud Datalab
- Cloud Pub/Sub
- Genomics

Machine Learning

- Cloud Machine Learning
- Cloud Vision API
- Cloud Speech API
- Cloud Natural Language API
- Cloud Translation API
- Cloud Jobs API

Identity & Security

- Cloud IAM
- Cloud Resource Manager
- Cloud Security Scanner
- Key Management Service
- Firewall
- BeyondCorp
- Data Loss Prevention
- Identity-Aware Proxy
- Security Key Enforcement

Google Cloud

More than 60 Google Cloud Platform services

Management Tools

Stackdriver

Monitoring

Logging

Error Reporting

Trace

Debugger

Cloud Deployment Manager

Cloud Endpoints

Cloud Console

Cloud Shell

Cloud Mobile App

Cloud Billing API

Cloud APIs

Developer Tools

Cloud SDK

Cloud Deployment Manager

Cloud Source Repositories

Cloud Tools for Android Studio

Cloud Tools for IntelliJ

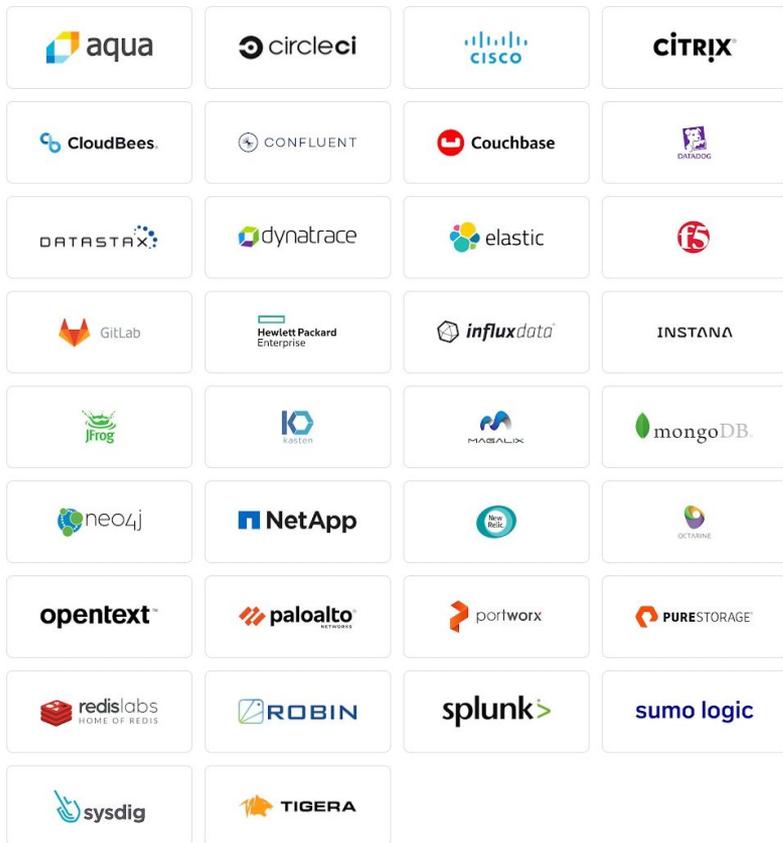**Cloud Tools for PowerShell**

Cloud Tools for Visual Studio
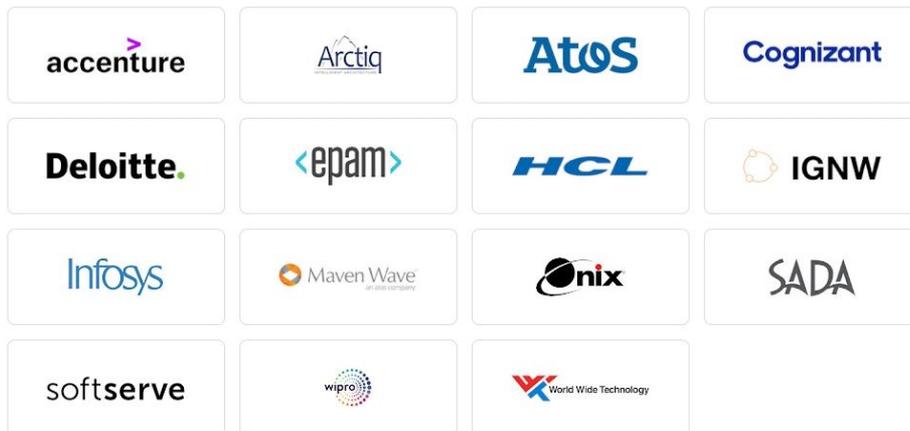
Google Plug-in for Eclipse

Cloud Test Lab

Google Cloud

Partners

Technology partners

aqua • circleci • CISCO • CITRIX

CloudBees • CONFLUENT • Couchbase • DATADOG

DATASTAX • dynatrace • elastic • f5

GitLab • Hewlett Packard Enterprise • influxdata • INSTANA

JFrog • kasten • MAGALIX • mongoDB

neo4j • NetApp • New Relic • OCTARINE

opentext • paloalto NETWORKS • portworx • PURESTORAGE

redislabs HOME OF REDIS • ROBIN • splunk> • sumo logic

sysdig • TIGERA

Platform partners

Atos • CISCO • DELLEMC • Hewlett Packard Enterprise

intel • Lenovo • NetApp

Service partners

accenture • Arctiq • Atos • Cognizant

Deloitte. • <epam> • HCL • IGNW

Infosys • Maven Wave • nix • SADA

softserve • wipro • World Wide Technology
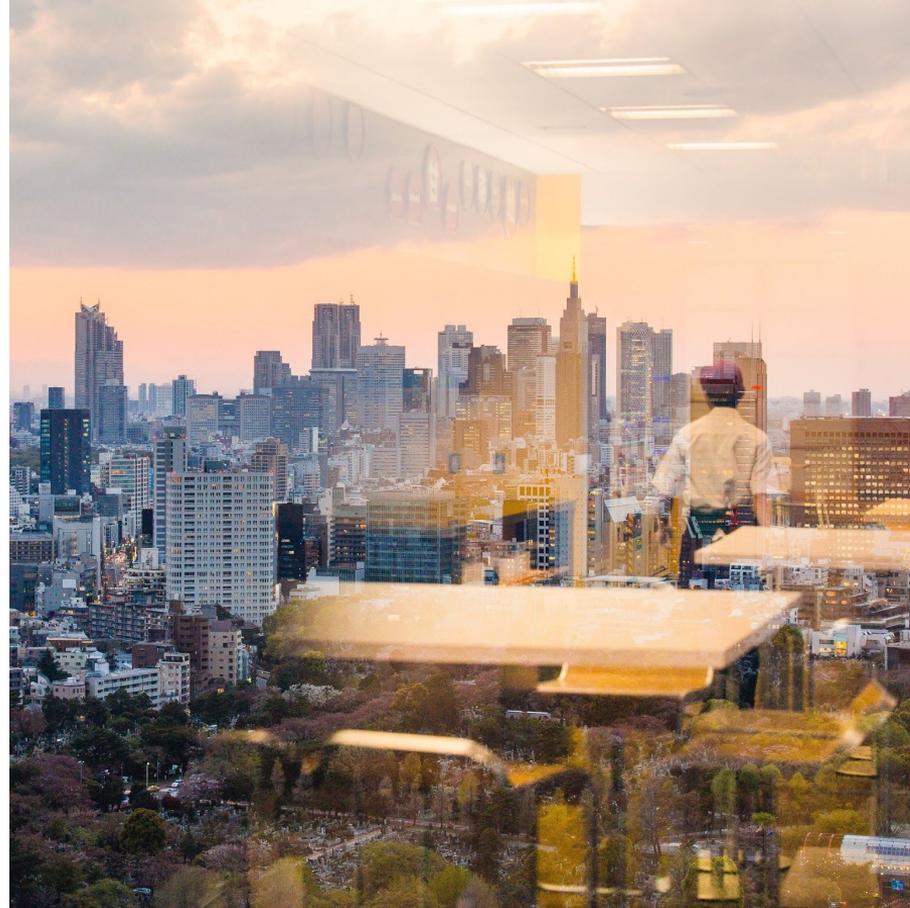
Google Cloud

Google Cloud Capabilities

Google Cloud

What makes Google Cloud different

Best-in-class Security ❯ Protect systems, data, and users

Hybrid & Multi-Cloud ❯ Enables choice

Fully Managed No Ops ❯ Ease of use with serverless

Embedded AI & ML ❯ Intelligence in everything

Best of Google ❯ Bringing culture of innovation to customers and partners

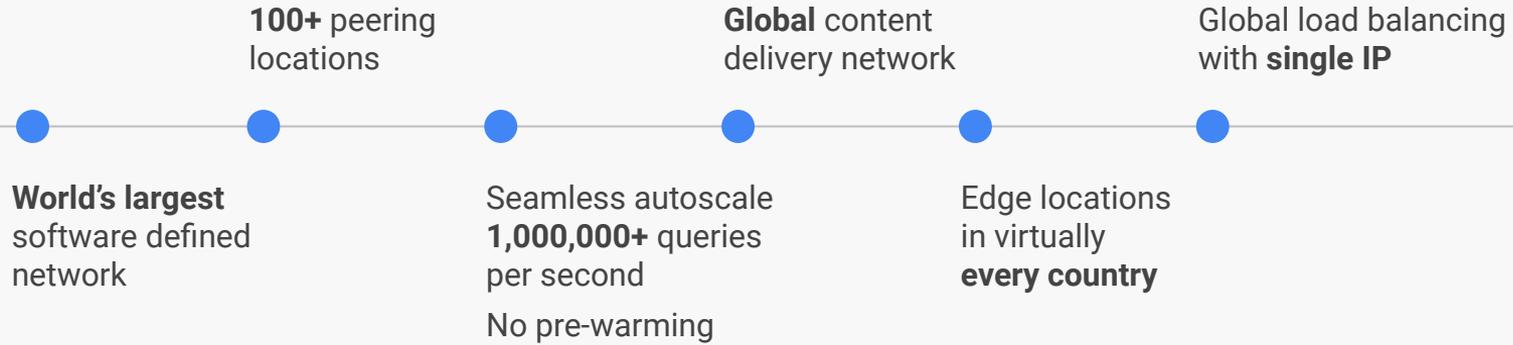# Helping Customers Meet Their Environmental Goals

Google is committed to environmental responsibility

100% carbon neutral since 2007

One of the world's largest corporate purchasers of renewable energy

First data centers to achieve ISO 14001 certification

Global network

100+ peering locations

Global content delivery network

Global load balancing with **single IP**

World's largest software defined network

Seamless autoscale **1,000,000+** queries per second

No pre-warming

Edge locations in virtually **every country**

Google Cloud

Continued global expansion

35

Regions

Oregon
Salt Lake City
Los Angeles
Las Vegas
Dallas
Iowa
Columbus
Toronto
Montréal
N. Virginia
S. Carolina

São Paulo
Santiago

Finland
Netherland
London
Belgium
Paris
Madrid
Berlin
Warsaw
Frankfurt
Zurich
Turin
Milan

Tel Aviv
Dammam
Doha
Delhi
Mumbai

Seoul
Tokyo
Osaka
Taiwan
Hong Kong

Singapore
Jakarta

Sydney
Melbourne

Current region with 3 zones
Future region with 3 zones

106
Zones available

Google Cloud

18

Google Cloud Platform

Edge point of presence · ○ Dedicated Interconnect · — Network
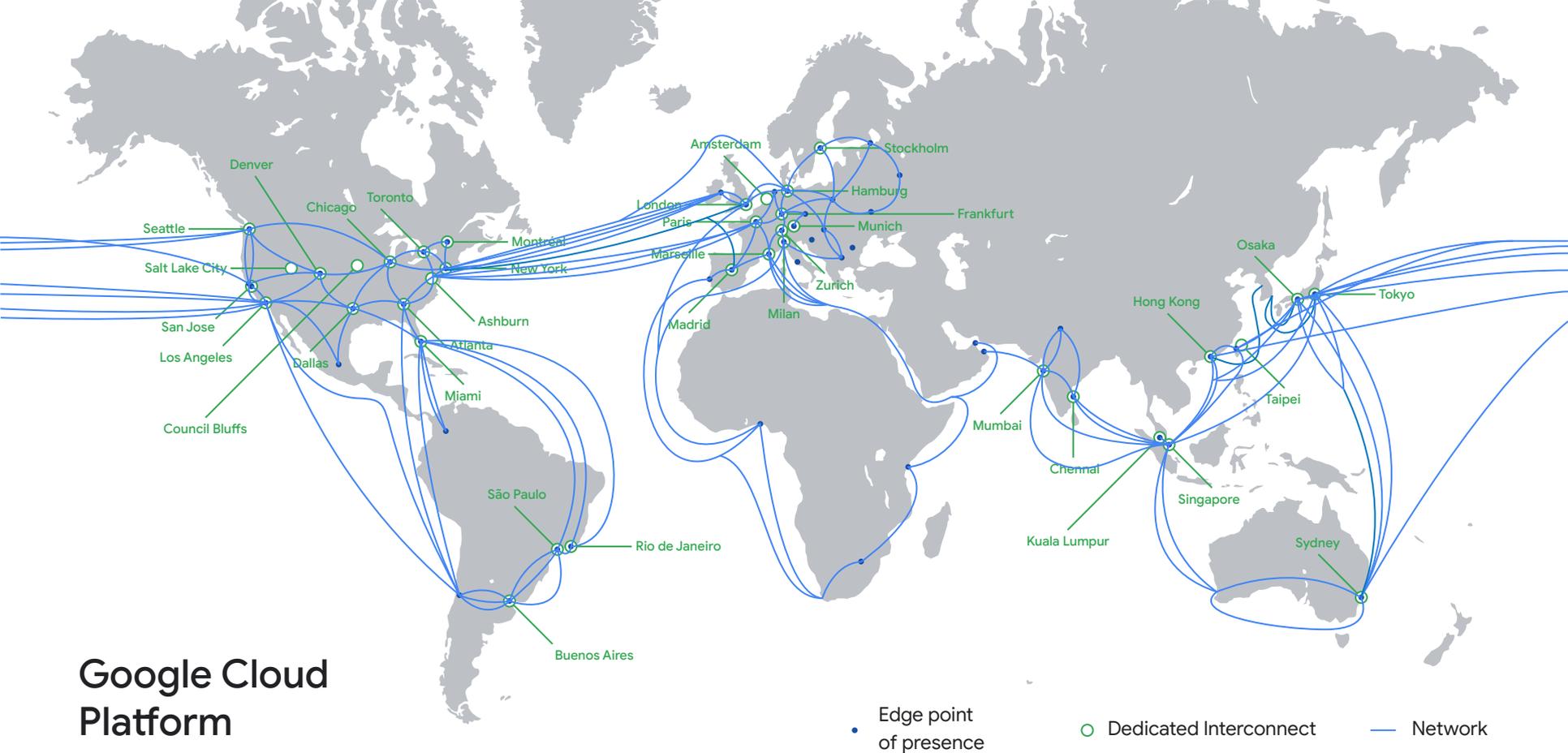
173 PoPs, 81 Dedicated Interconnects and 14 Subsea Cables

Google Cloud   19

The Network Matters

Typical Cloud Provider

Cloud Provider

User

Google Cloud

Google Cloud

Google Pop

Google Pop

ISP

User
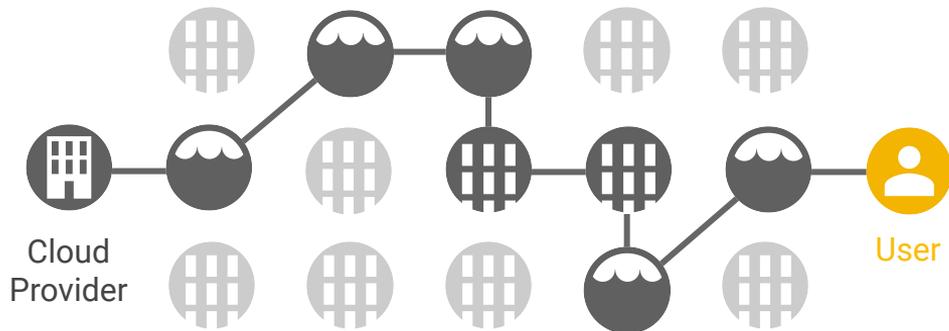
Google Cloud
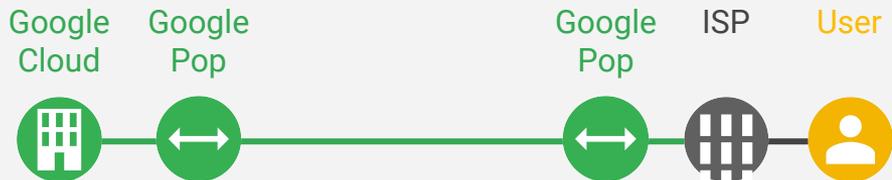
Our approach to security in two words

Trust Nothing

Google Cloud

Defense in depth, at scale, by default

Internet
Communication

Operational and
Device Security

Identity

Storage
Services

Hardware
Infrastructure

Service
Deployment

Google Cloud
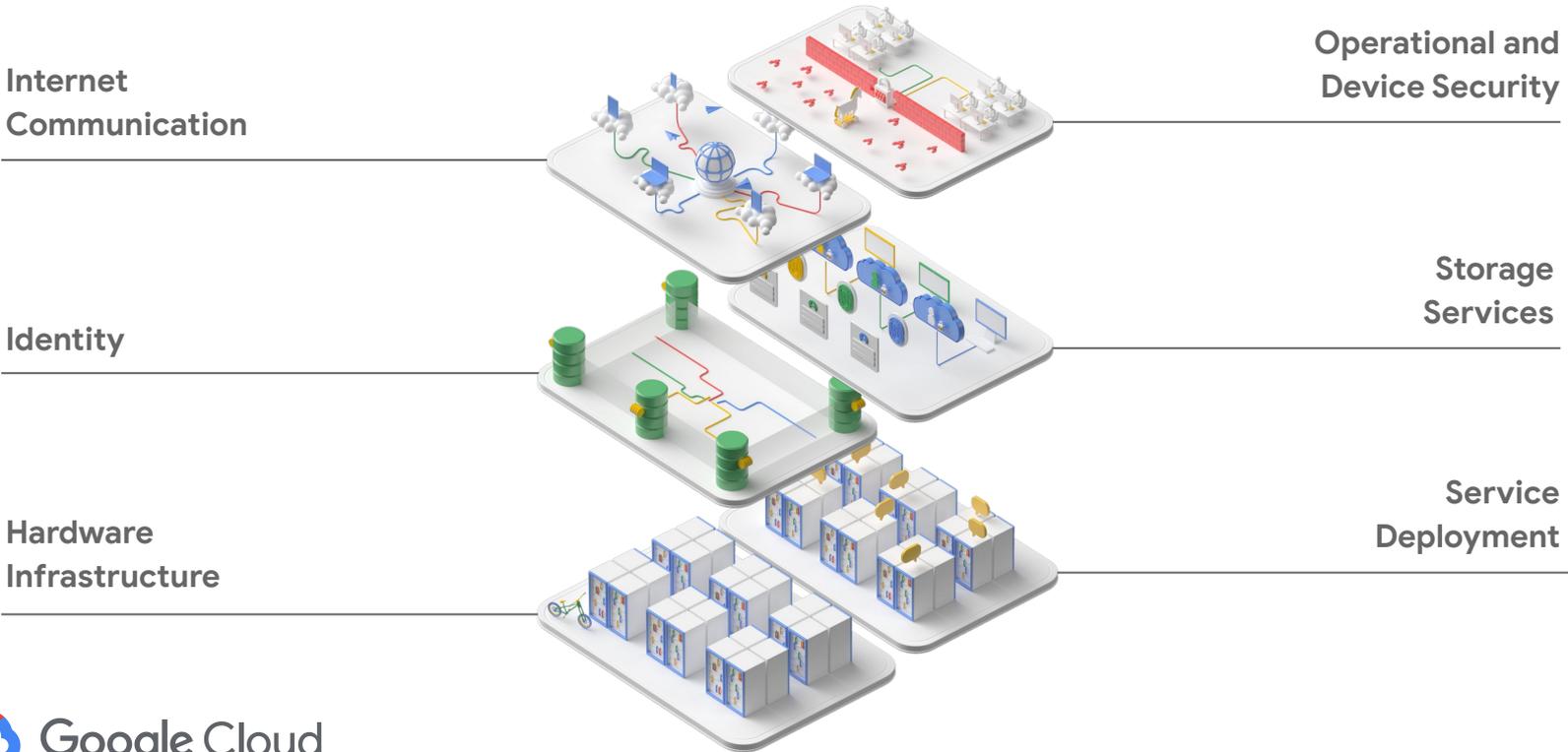
End-to-end provenance & attestation

Titan

Google's purpose-built chip to establish a hardware root of trust for both machines and peripherals in cloud infrastructure.
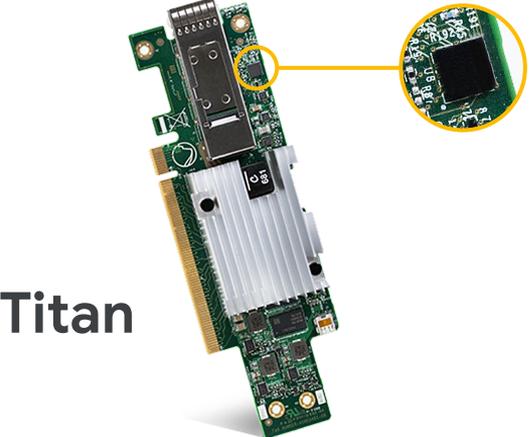
Purpose-built **chips**

Purpose-built **servers**

Purpose-built **storage**

Purpose-built **network**

Purpose-built **data centers**

Reduced "vendor in the middle" risk

Google Cloud

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 222-44-555

Data Loss Prevention

Sophisticated Machine Learning to find and redact sensitive information

Google Cloud

Encrypt at rest
by default

All connections use TLS

Data is chunked and each chunk is encrypted with its own data encryption key

Data encryption keys (DEKs) are wrapped using a key encryption key (KEK)

Encrypted chunks and wrapped encryption keys are distributed across Google's storage infrastructure

Google Cloud

Compliance offerings

Global

ISO/IEC 27001
ISO/IEC 27017
ISO/IEC 27018
ISO/IEC 27701
SOC 1
SOC 2
SOC 3
PCI DSS
CSA STAR
MPAA
Independent Security Evaluators Audit
GxP

Americas

USA

HIPAA
HITRUST
FedRAMP
FIPS 140-2
COPPA
FERPA
NIST 800-53
NIST 800-171
NIST 800-34
Sarbanes- Oxley
SEC Rule 17a-4(f)
CFTC Rule 1.31(c)-(d)
FINRA Rule 4511(c)
HECVAT
DISA IL2
CCPA
CJIS

Canada

PIPEDA
Personal Health Information Protection Act

Argentina

Personal Data Protection Law

Europe, Middle East & Africa

Europe

GDPR
EU Model Contract Clauses
TISAX
EBA Guidelines

Germany

BSI C5

Switzerland

FINMA

France

HDS

Spain

Esquema Nacional de Seguridad

South Africa

POPI

UK

NCSC Cloud Security Principles
NHS IG Toolkit

Asia Pacific

Australia

Australian Privacy Principles
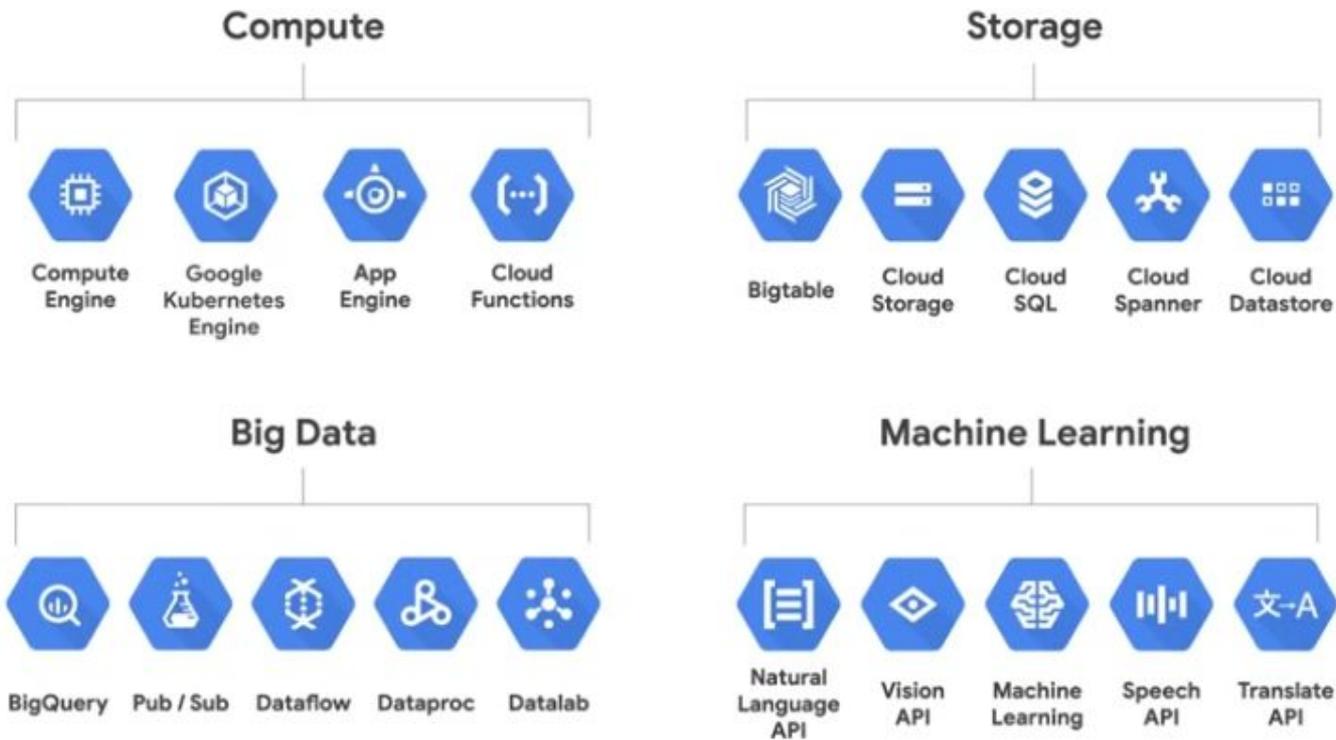Australian Prudential Regulatory Authority Standards
IRAP

Japan

FISC
My Number Act
NISC
CSV Guidelines
3G3M

Singapore

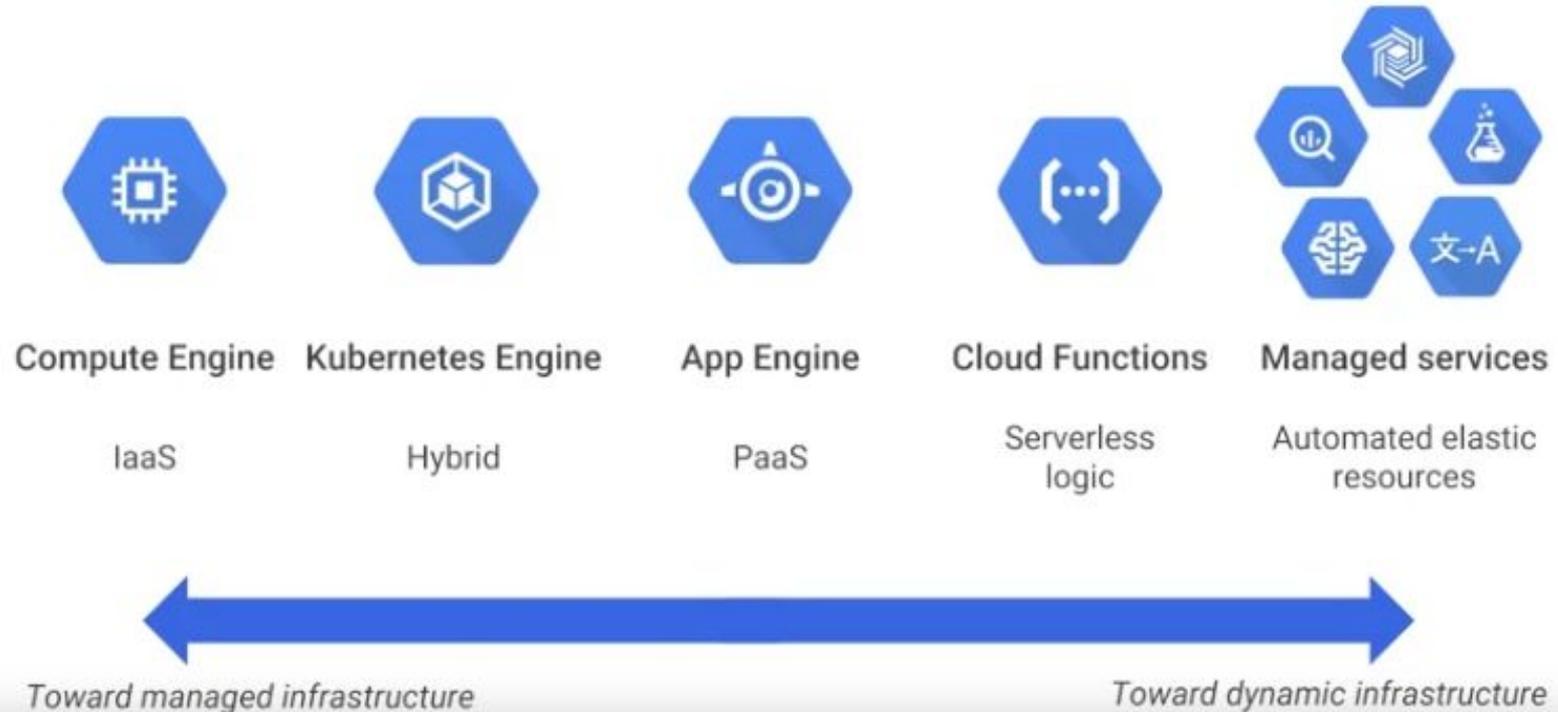MTCS Tier 3
OSPAR
MAS Guidelines
ABS Guide

Google Cloud

Google Cloud Products & Solutions

Google Cloud

GCP Core Products and Solutions

Infrastructure Solutions

Google Cloud

Computing Architectures

GCP computing architectures meet you where you are

Compute Engine   Kubernetes Engine   App Engine   Cloud Functions   Managed services

IaaS   Hybrid   PaaS   Serverless logic   Automated elastic resources

Toward managed infrastructure   *Toward dynamic infrastructure*
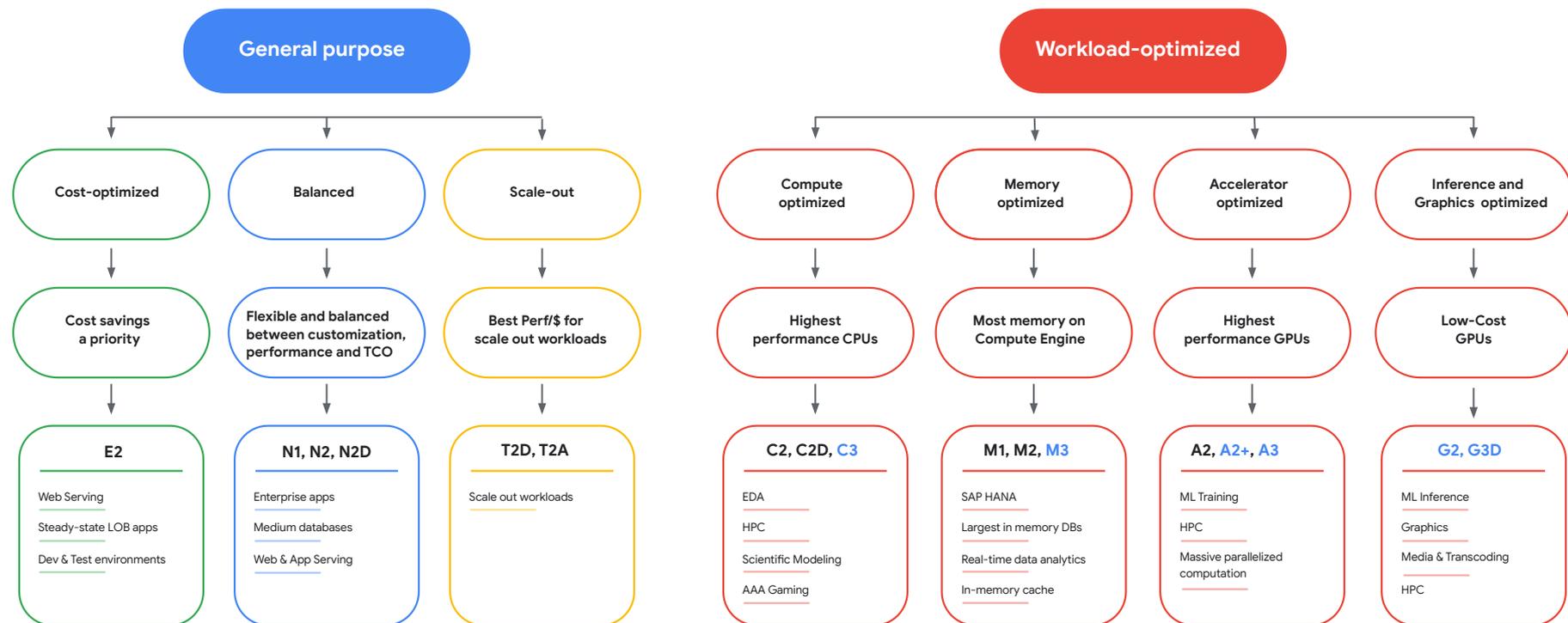
Compute Engine

- Live Migration = Less downtime

- Custom Machine Types (incl. GPU)

- Industry-leading I/O performance

- Only pay for what you use, per minute

- Super-fast startup: 1,000 VMs in < 5 min

- Resize disks with no downtime
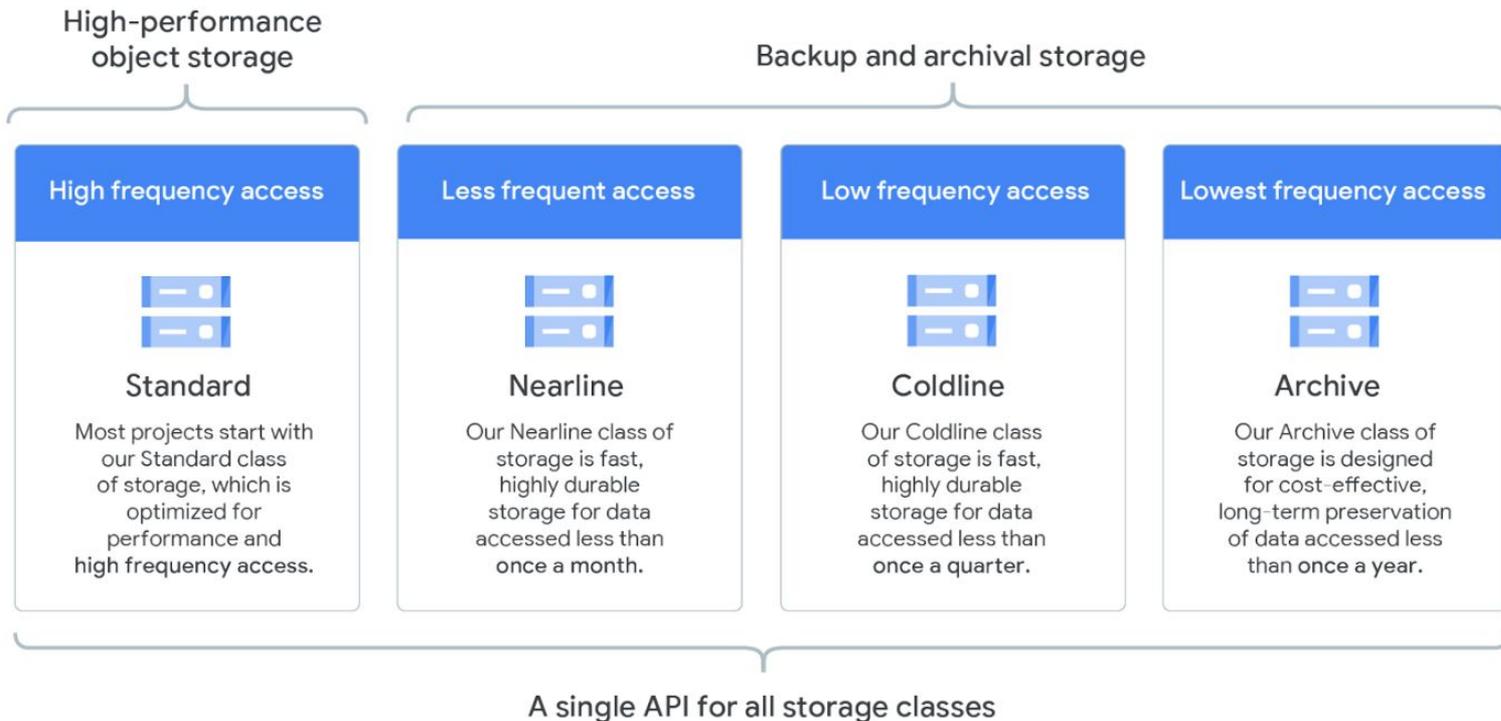
- Zonal & Regional Autoscaling
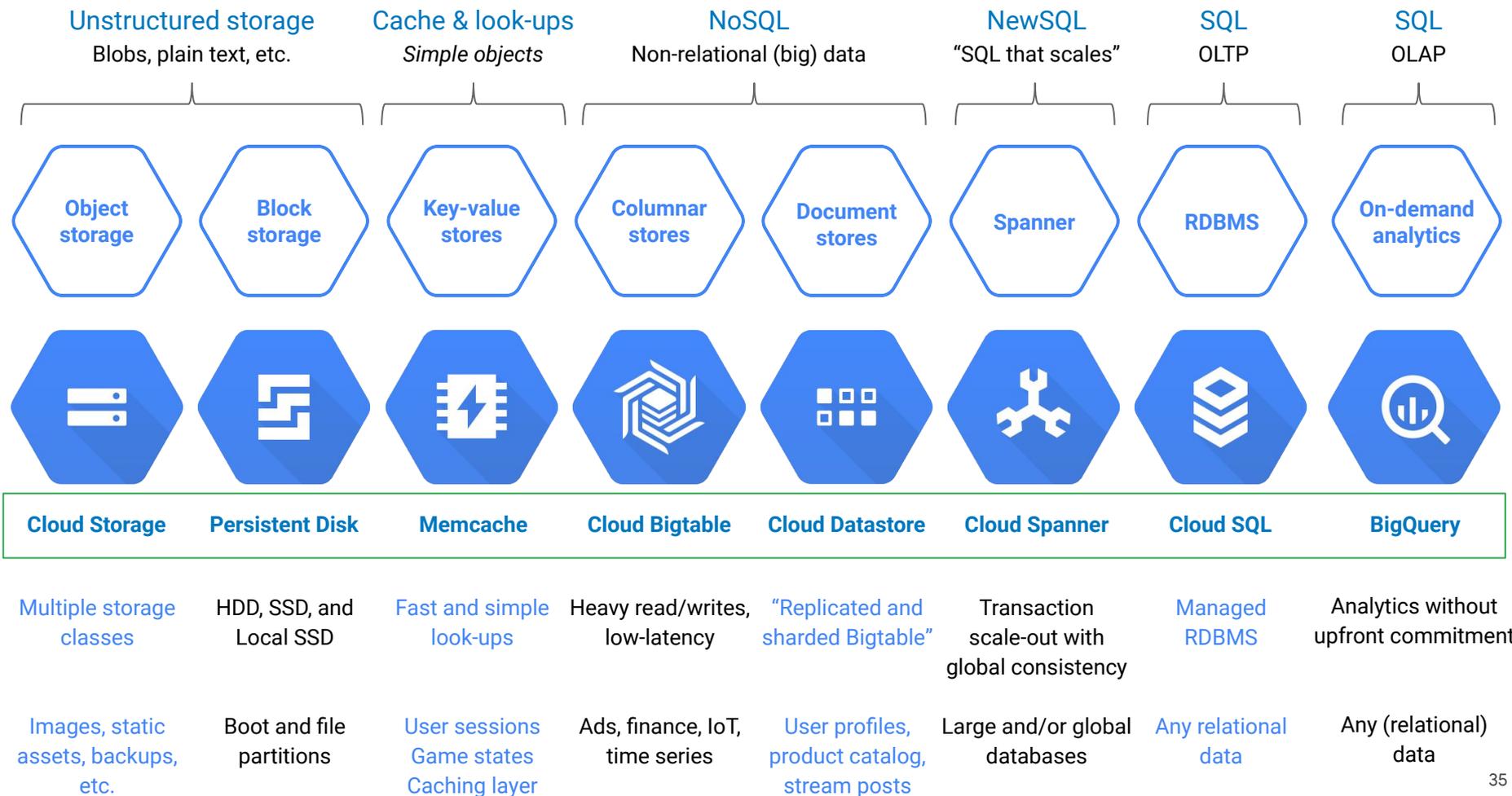
VM Families For All Workloads
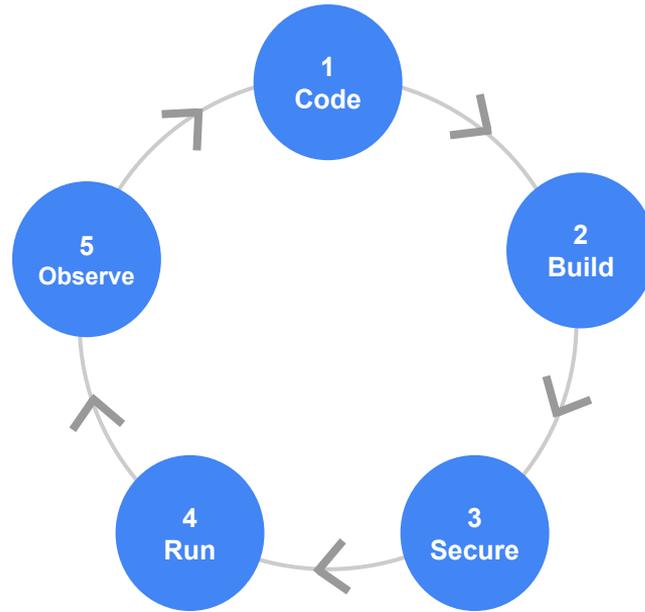
The right capabilities for all your workloads

General purpose

Cost-optimized	Balanced	Scale-out
Cost savings a priority	Flexible and balanced between customization, performance and TCO	Best Perf/$ for scale out workloads

E2
- Web Serving
- Steady-state LOB apps
- Dev & Test environments

N1, N2, N2D
- Enterprise apps
- Medium databases
- Web & App Serving

T2D, T2A
- Scale out workloads

Workload-optimized

Compute optimized	Memory optimized	Accelerator optimized	Inference and Graphics optimized
Highest performance CPUs	Most memory on Compute Engine	Highest performance GPUs	Low-Cost GPUs

C2, C2D, C3
- EDA
- HPC
- Scientific Modeling
- AAA Gaming

M1, M2, M3
- SAP HANA
- Largest in memory DBs
- Real-time data analytics
- In-memory cache

A2, A2+, A3
- ML Training
- HPC
- Massive parallelized computation

G2, G3D
- ML Inference
- Graphics
- Media & Transcoding
- HPC

Storage and Database Solutions

Google Cloud

GCP Storage Considerations - Lifecycle Management

High-performance object storage

Backup and archival storage

High frequency access	Less frequent access	Low frequency access	Lowest frequency access
Standard	**Nearline**	**Coldline**	**Archive**
Most projects start with our Standard class of storage, which is optimized for performance and high frequency access.	Our Nearline class of storage is fast, highly durable storage for data accessed less than once a month.	Our Coldline class of storage is fast, highly durable storage for data accessed less than once a quarter.	Our Archive class of storage is designed for cost-effective, long-term preservation of data accessed less than once a year.

A single API for all storage classes

Google Cloud

Unstructured storage		Cache & look-ups	NoSQL		NewSQL	SQL	SQL
Blobs, plain text, etc.		Simple objects	Non-relational (big) data		"SQL that scales"	OLTP	OLAP
Object storage	Block storage	Key-value stores	Columnar stores	Document stores	Spanner	RDBMS	On-demand analytics
Cloud Storage	**Persistent Disk**	**Memcache**	**Cloud Bigtable**	**Cloud Datastore**	**Cloud Spanner**	**Cloud SQL**	**BigQuery**
Multiple storage classes	HDD, SSD, and Local SSD	Fast and simple look-ups	Heavy read/writes, low-latency	"Replicated and sharded Bigtable"	Transaction scale-out with global consistency	Managed RDBMS	Analytics without upfront commitment
Images, static assets, backups, etc.	Boot and file partitions	User sessions Game states Caching layer	Ads, finance, IoT, time series	User profiles, product catalog, stream posts	Large and/or global databases	Any relational data	Any (relational) data

35

Application Development Solutions

Google Cloud

Application Development Workflow

Containers
Flexibility

Serverless
Velocity

Code → Build → Run

Introducing Google Cloud's Anthos

Anthos lets you build and manage modern **hybrid and multi-cloud** applications without lock-in

Build once, to run anywhere, across your existing on-premise infrastructure and all major public cloud providers

Google Cloud

All Major Clouds

Anthos

On-Premise Data Center
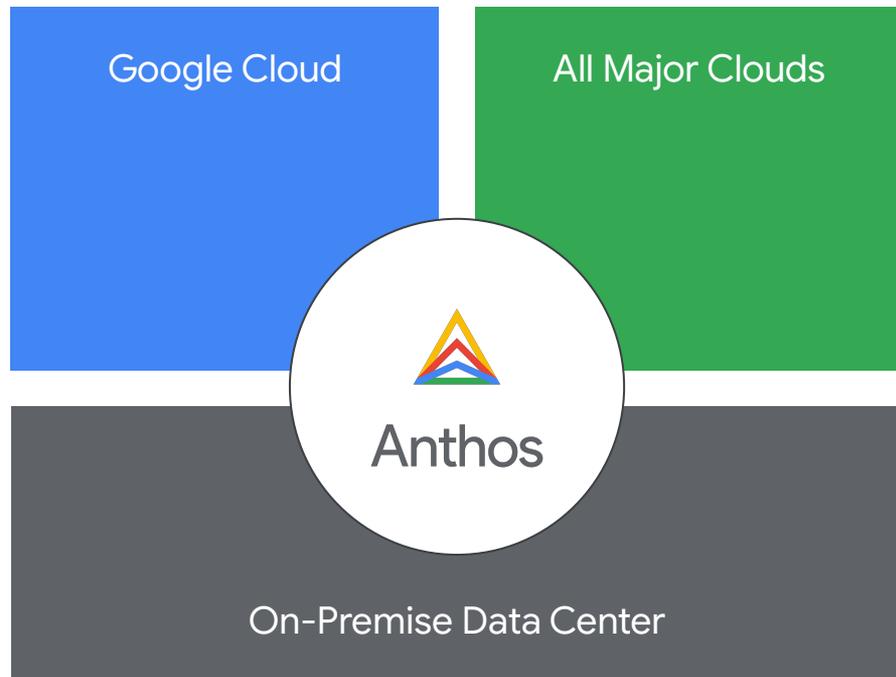
Google Cloud

Data, Analytics & Machine Learning

Google Cloud

Data Lifecycle in Google Cloud
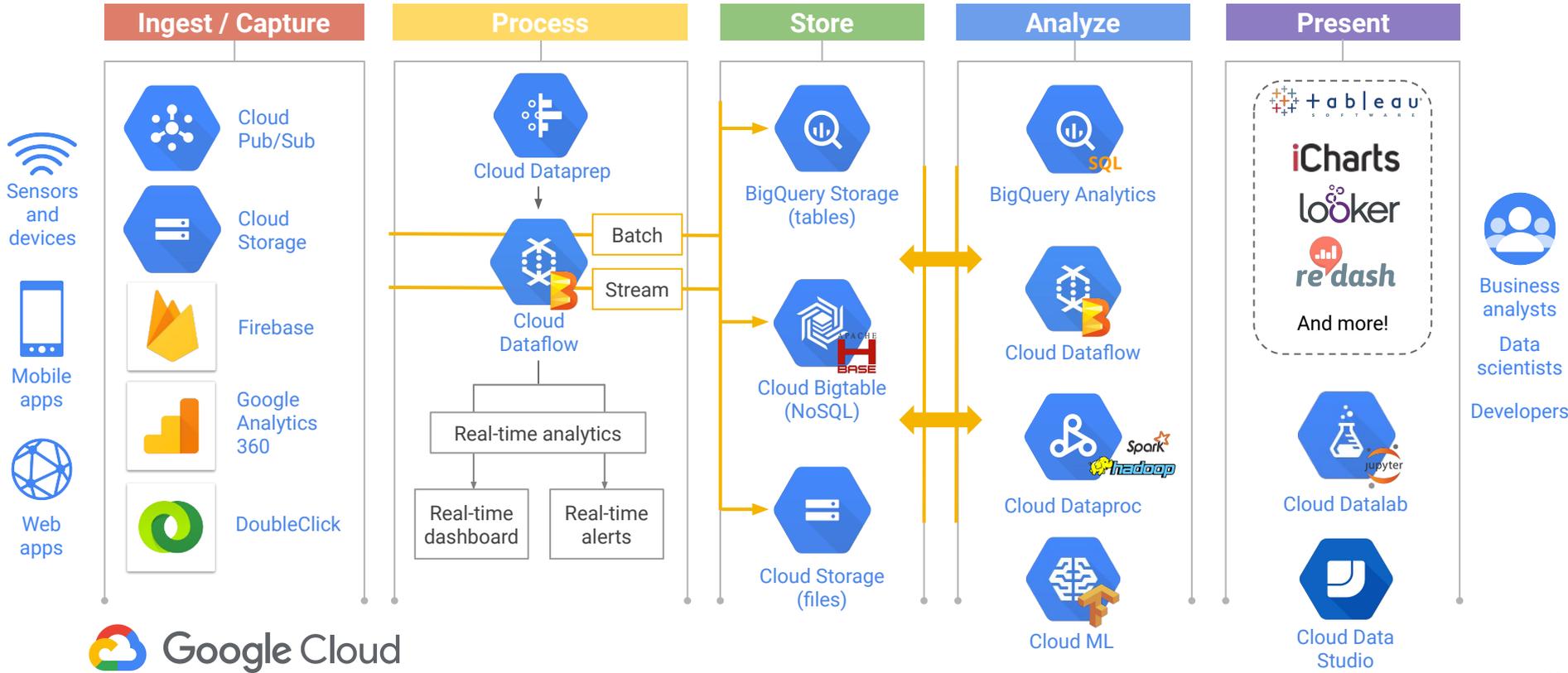
Data and Analytics Partner Ecosystem

Machine Learning with Google

Use Our Models

Take advantage of Google's domain expertise

No tools or AI expertise required

Extend or customize with AutoML

OR

Train Your Own

Build on your own specialized domain expertise

Use Google tools for building and training models

Google Cloud

Powered by Open source

AI Platform

AI Building Blocks

Vision   Video Intelligence   Speech   Natural Language   Translation

Platform, Libraries, Tools

Cloud ML Engine   BigQuery   DataLab   TensorFlow   Spark   beam   R   TORCH   K   Spark MLlib   learn

AI Infrastructure

Cloud Storage   Networking   Compute Engine   Cloud GPUs   Cloud TPUs

Google Cloud

Google Cloud

Introduction to Google Healthcare & Life Sciences (HCLS)
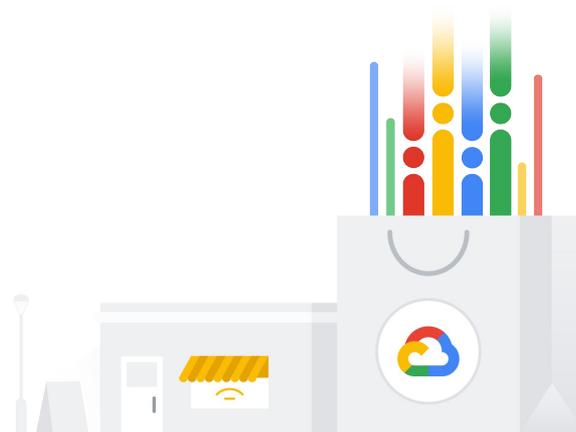
At Google Cloud, our mission is why we innovate

Google

Organize the **world's information** and make it universally **accessible** and **useful.**

Data

APIs

Intelligence

Google Cloud
Health & Life Sciences

Organize the **world's health & life sciences information** and make it universally **accessible, secure,** and **useful.**

Data

Open standards and APIs

Privacy and Compliance

AI and Analytics

Our values

Respect the user

Respect the opportunity

Respect each other

Google Cloud

We do **search** but also lot of **research**!

Alphabet

Google

Google Ventures Venture & Capital Funding	**Loon** Flight Connection	**Google X** Innovation Lab & Research
Waymo Self Driving Vehicles	**Fiber** High Speed Internet Services	**Calico** Longevity Research
SideWalk Labs Solving Big Urban Problems	**Wing** Drone-based Delivery of Freight	**Verily** Improving Quality of Life
Jigsaw Global Online Security Solutions	**CapitalG** Private Equity	**DeepMind** AI & Machine Learning

Search Advertising SEM	**Google Cloud** Healthcare and Life Sciences
YouTube Internet Video Service	**Google Health** AI and Consumer Health Products
Maps Mapping, Location Services & Logistics	**Google Marketing Platform** Data Analytics Suite of Tools
Devices and Services Pixel, Nest Chromecast	**Android** Mobile Operating System

Google

Agenda

1. Cloud Life Sciences API
2. Variant Transforms
3. Cloud Healthcare API
4. De-identification
5. Medical Imaging Suite

Google Cloud

Life Sciences API
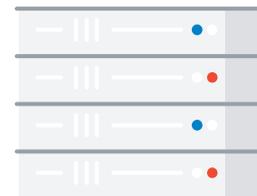
01

Google Cloud

Genomics is critical for the race to drug discovery but harnessing the exponential growth in genomics data will require significant resources

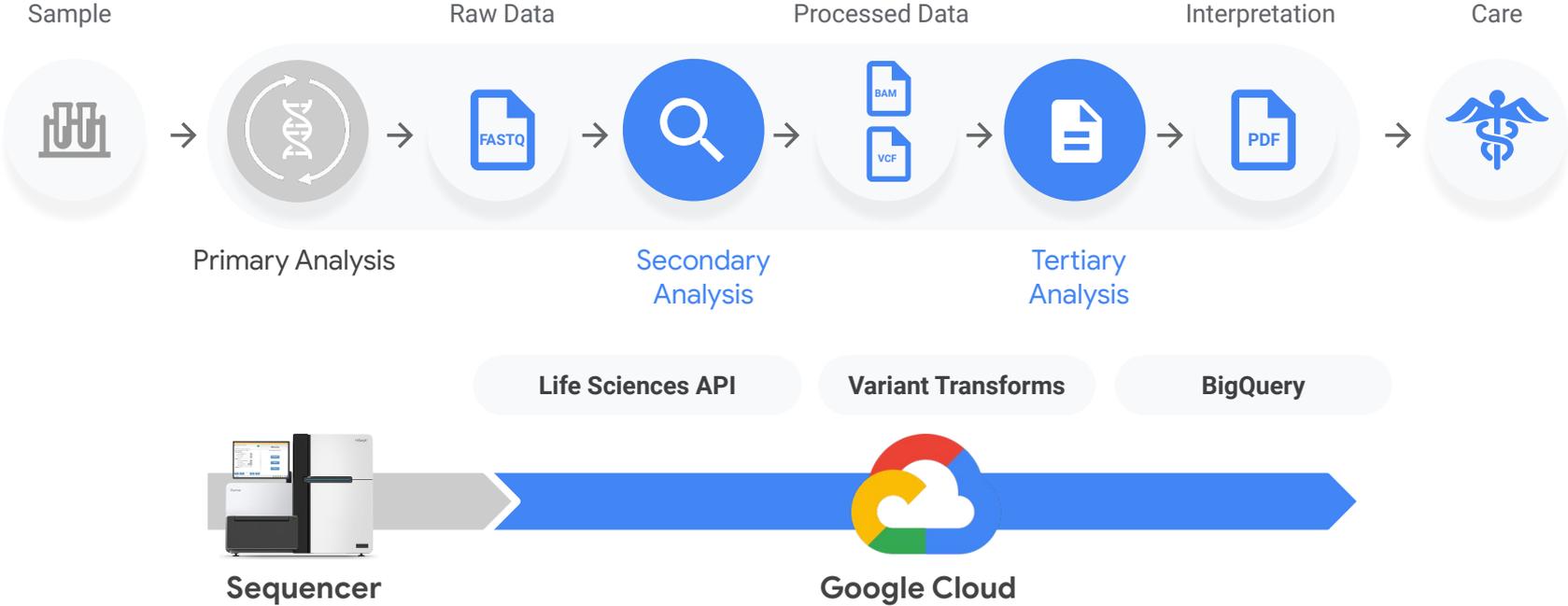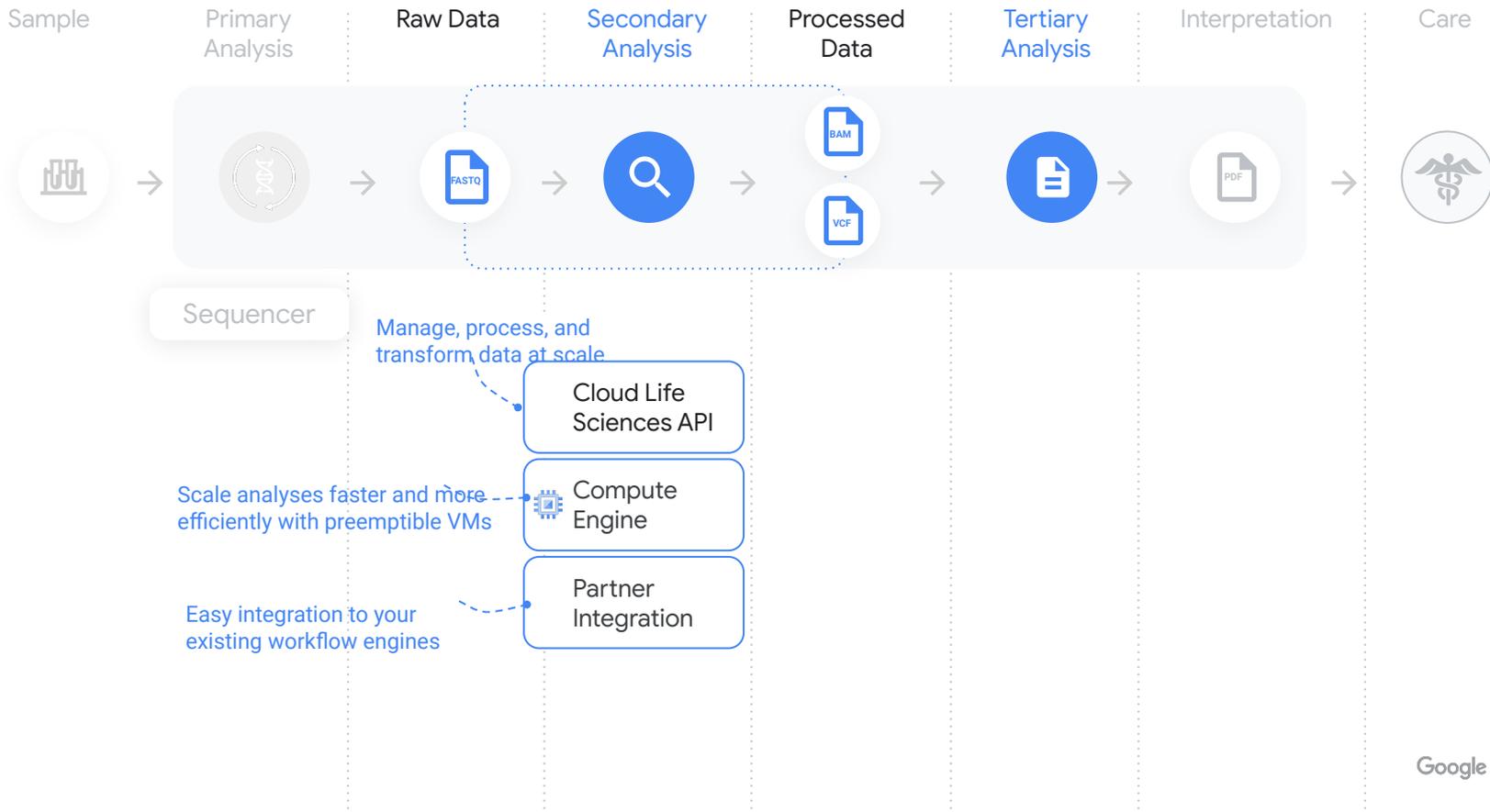Drug targets based on genomic analysis are **2x** as likely to be approved.[1]

By **2025,** more than **100 million** genomes will have been sequenced - accumulating over **20 billion gigabytes** of raw data.[2]

Sources: [1] PLOS Genetics, King et al, December 2019, [2] Labiotech, January 2020

Google Cloud

Process genomic data at scale with Google Cloud

How it works

Sample → Primary Analysis → Raw Data → Secondary Analysis → Processed Data → Tertiary Analysis → Interpretation → Care

Sequencer

BAM

VCF

FASTQ

PDF

Manage, process, and transform data at scale

Cloud Life Sciences API

Scale analyses faster and more efficiently with preemptible VMs

Compute Engine

Easy integration to your existing workflow engines

Partner Integration

Google Cloud

Genomics Secondary Analysis with Pipelines API

Run bioinformatics pipelines at scale and with low cost, using industry standard tools and frameworks (e.g. GATK) as well as optimized novel ML-based tools (e.g. DeepVariant).

Optimize for turnaround time with parallel execution and for cost with PVMs, GPUs, TPUs and regional storage.

Ecosystem of platform partners and workflow engines to help manage your workloads.

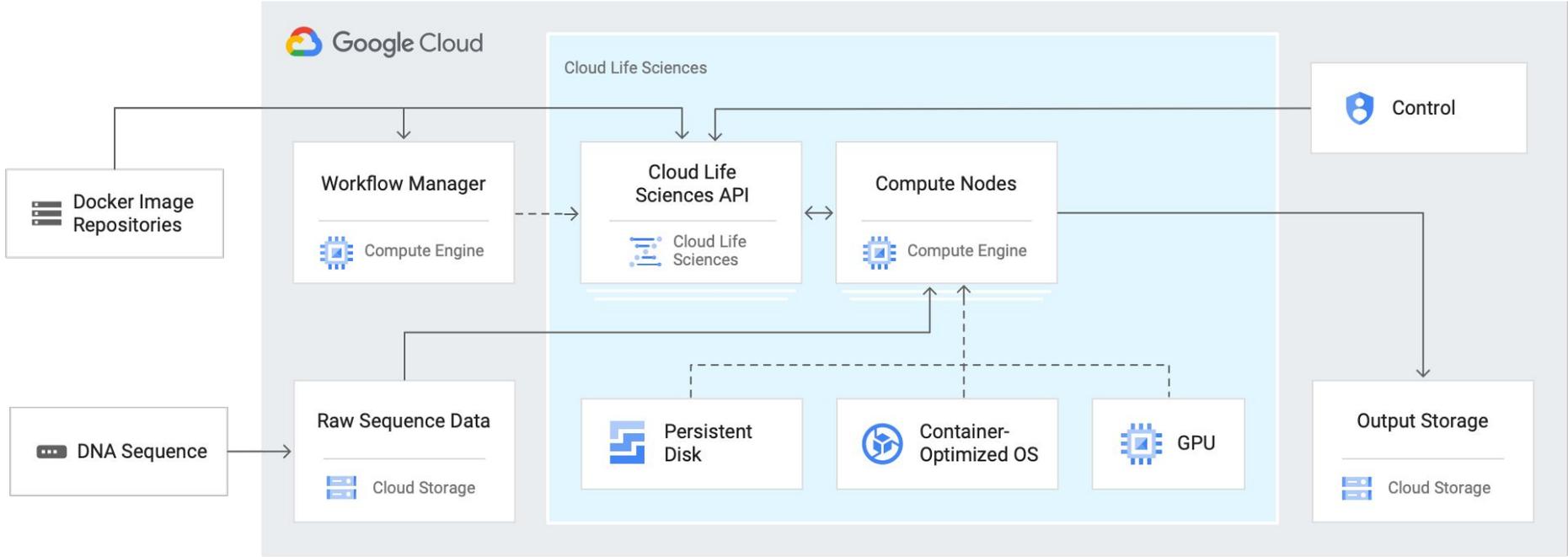Terra

nextflow

SevenBridges

BC Platforms

Sentieon

ROSALIND

Snakemake

dsub

NVIDIA
CLARA

DNASTACK

Galaxy

Google Cloud

Google Cloud

Cloud Life Sciences

Control

Docker Image Repositories

Workflow Manager

Compute Engine

Cloud Life Sciences API

Cloud Life Sciences

Compute Nodes

Compute Engine

DNA Sequence

Raw Sequence Data

Cloud Storage

Persistent Disk

Container-Optimized OS

GPU

Output Storage

Cloud Storage

Data path   Control path

Google Cloud

Variant Transforms

Google Cloud

02

How it works

Sample · Primary Analysis · Raw Data · Secondary Analysis · Processed Data · Tertiary Analysis · Interpretation · Care

Sequencer

BAM
VCF

BigQuery

Run analytics including cross-modality or clinical analytics

Variant Transforms

Standardize schema to easily push data into BigQuery

Variant Annotations libraries

Enrich meaning of your variant data

Google Cloud

Using Variant Transforms to load a VCF to BigQuery

How it works

Sample → Primary Analysis → Raw Data → Secondary Analysis → Processed Data → Tertiary Analysis → Interpretation → Care

Sequencer

BigQuery

Run analytics including cross-modality or clinical analytics

Variant Transforms

Standardize schema to easily push data into BigQuery

Use familiar tools within managed notebook environments

Vertex AI Notebooks

Variant Annotations libraries

Enrich meaning of your variant data

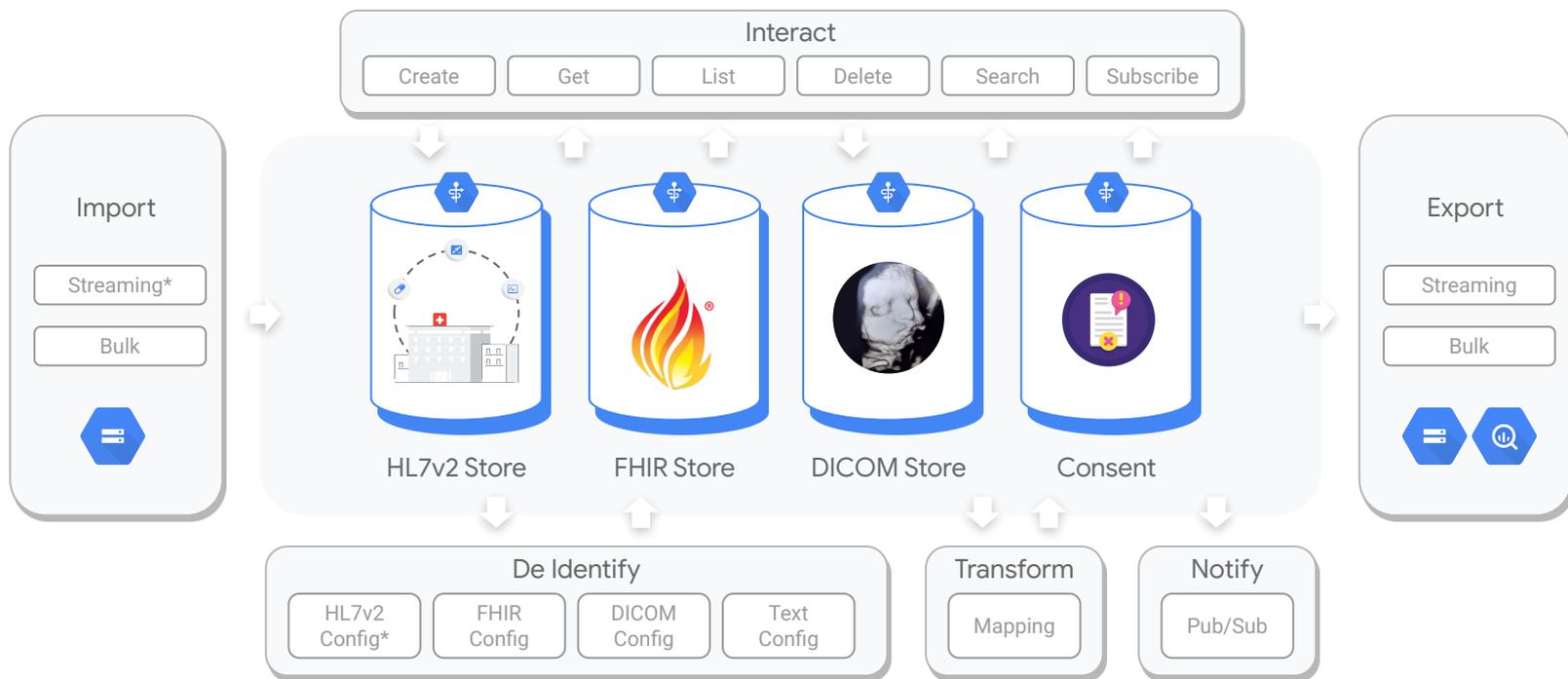Google Cloud

Healthcare API

Google Cloud

Traditional database stores persist data to disk and offer a API, like Structured Query Language

SQL

Google Cloud
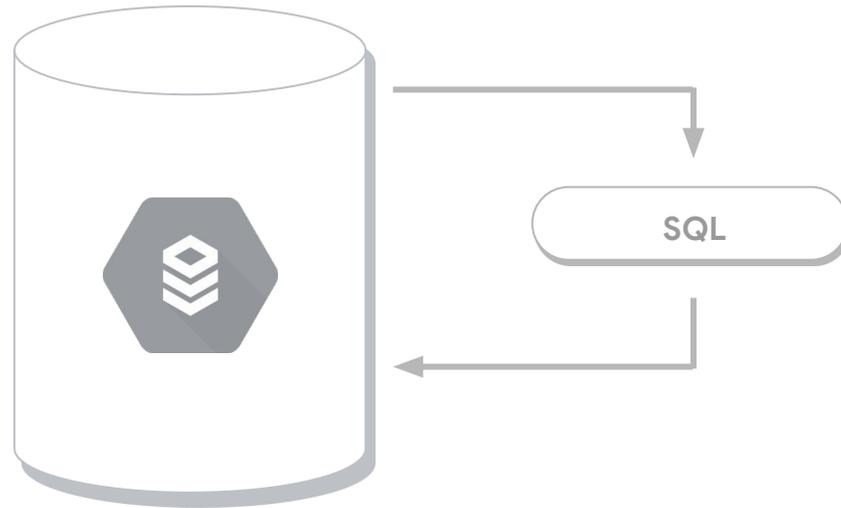
With Cloud Healthcare API, you create data stores that implement healthcare-native APIs

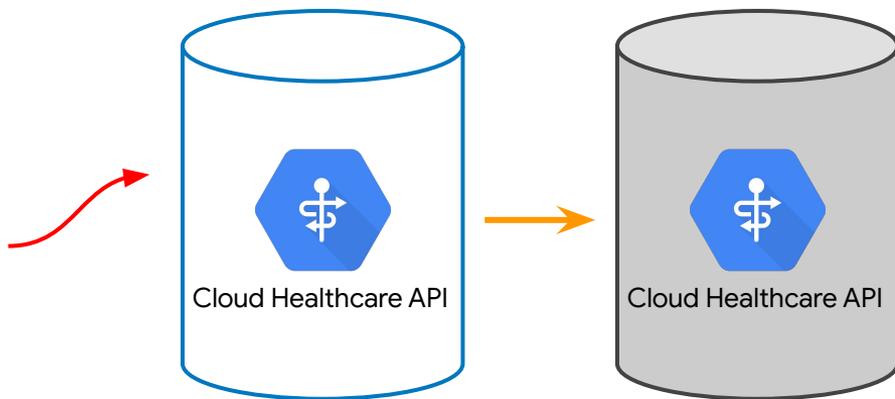HL7v2, FHIR, DICOM

Google Cloud
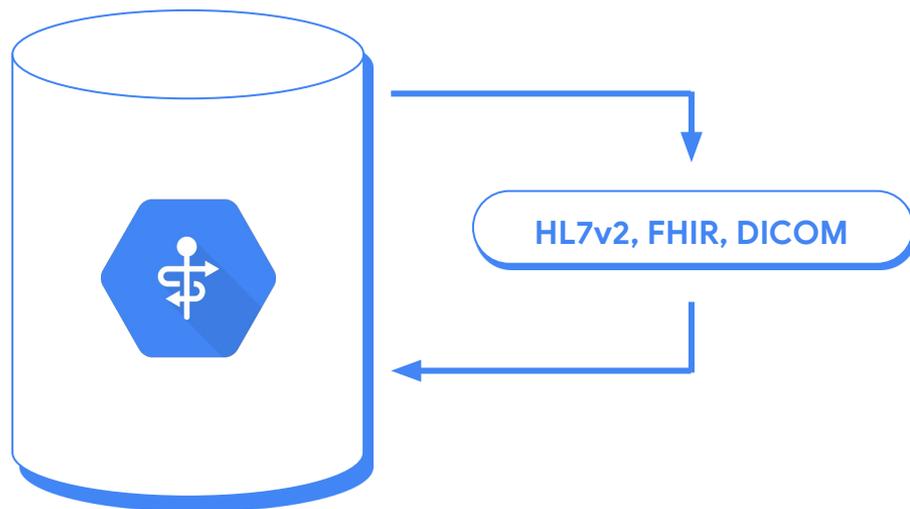
Healthcare API REST Endpoints

De-identification

Google Cloud

Example : De-identify an entire healthcare dataset

Store your EHR
Data in
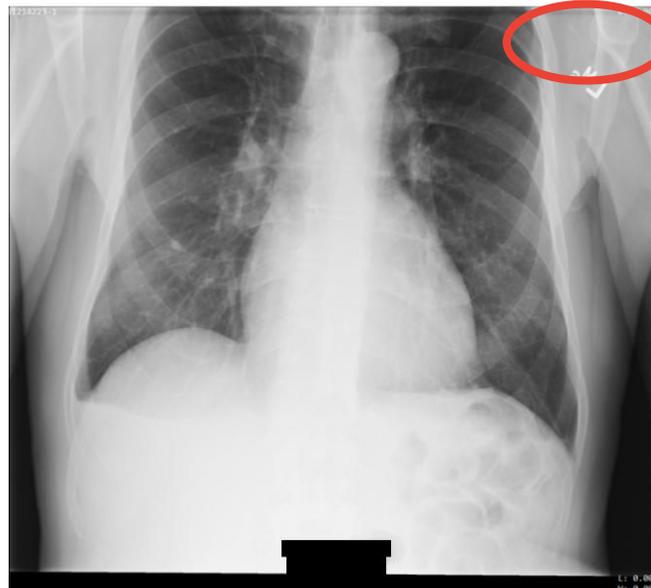Healthcare API
(FHIR, DICOM, HL7v2)

Cloud Healthcare API

Cloud Healthcare API

Create de-identified
dataset of you EHR
data

De-identification for images

Example : De-identification of a DICOM Image

Google Cloud

Google Cloud Medical Imaging Suite: Reference Architecture

Healthcare API

2 Imaging Lab

Flywheel on GCP

MONAI †
NVIDIA

Open Health
Imaging Foundation

Flywheel

4 Imaging AI Pipelines

Vertex AI

Kubernetes

NVIDIA
CLARA

3 Imaging Datasets & Dashboards

BigQuery

Looker

1 Imaging Store

NFS → Dataflow → **DICOM Data Store** ← Functions ← Storage ← ONTAP® on GCP

NetApp

Virtual Archival Appliance

Edge Devices

Networked Devices (PACS, DICOM)

Data Center

NetApp

5 Imaging Deployment

Browser

Anthos

Remote Visualization

Model Development

EDGE DEVICES

Q&A

Google Cloud

An Overview of the NIH STRIDES Initiative

NCI's Containers and Workflows Interest Group Webinar Series

Philip Meacham, PhD [C]   October 14, 2022

Philip.Meacham@nih.gov
Cloud Instructional Development Specialist, STRIDES
Center for Information Technology

NIH — National Institutes of Health

NIH STRIDES
Accelerating biomedical research

Disclosures

- Dr. Philip Meacham is a contractor with Deloitte Consulting LLP, supporting the NIH STRIDES Initiative

Common Barriers to Accessing the Cloud

- When institutions and researchers want to use the cloud, a host of complexities arise:

  - Setting up acquisition vehicles and access to cloud service providers

  - Budgeting and paying for usage / optimizing costs / preventing overspends

  - Learning new tools and new ways of working

  - Growing, securing, and maintaining prototype capabilities as more robust infrastructure, systems, and services

NIH STRIDES Initiative

Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES)

Serving **both the NIH Intramural and extramural research communities**, the STRIDES Initiative accelerates biomedical research in the cloud by simplifying access, reducing costs, lowering technological barriers, and improving processes.

Core motivations for STRIDES include:

- Democratization of computational research and data science:
  - Leveling the playing field for those traditionally underrepresented in biomedical research
- Cost savings and efficiencies for the research community at large:
  - More usage begets more savings and greater overall discounts for all
- Strong partnerships with cloud providers:
  - Resulting in collaborative R&D engagements and more direct focus and support on research

Partnerships with aws Google Cloud Microsoft Azure

76

NIH STRIDES Initiative

STRIDES Is...	STRIDES Is Not...
An NIH program, and part of NIH's data science portfolio	A destination (i.e., there is no "STRIDES cloud")
A mechanism for NIH and NIH-funded researchers to access and use cloud compute, storage, and related services	A service for researchers to store or analyze research data *(though it can help provide access to cloud capabilities for storing and analyzing data)*
One method for using the cloud to support biomedical research	The only method for using the cloud to support biomedical research
Encouraged by NIH	Required by NIH

Partnerships with   aws   Google Cloud   Microsoft Azure

STRIDES by the Numbers*

25
NIH ICs
PARTICIPATING

109
EXTRAMURAL INSTITUTIONS
ENROLLED

995+
PROGRAM/PROJECTS
ONBOARDED

200+
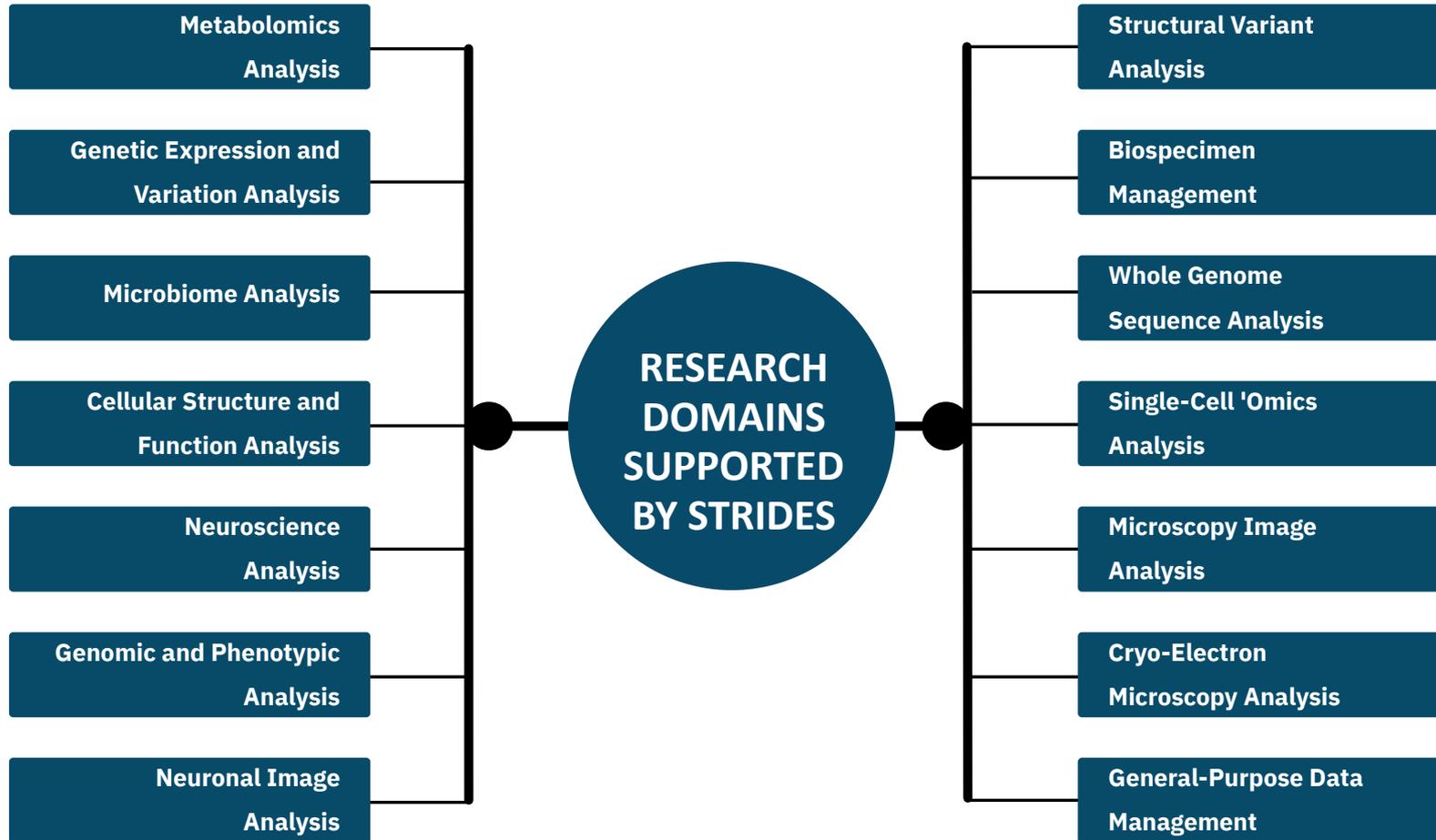PETABYTES
OF DATA STORED

4,700+
TRAININGS COMPLETED

274M+
COMPUTE HOURS

$41M+
COST SAVINGS TO ICs

Impact as of August 31, 2022

Metabolomics Analysis

Genetic Expression and Variation Analysis

Microbiome Analysis

Cellular Structure and Function Analysis

Neuroscience Analysis

Genomic and Phenotypic Analysis

Neuronal Image Analysis

RESEARCH DOMAINS SUPPORTED BY STRIDES

Structural Variant Analysis

Biospecimen Management

Whole Genome Sequence Analysis

Single-Cell 'Omics Analysis

Microscopy Image Analysis

Cryo-Electron Microscopy Analysis

General-Purpose Data Management

79

NIH> NATIONAL CANCER INSTITUTE

NCI Projects with STRIDES

- Currently 150+ NCI projects with STRIDES accounts
  - 25 extramural and 8 intramural projects are using GCP

- NCI has received funding for several important projects through Cloud Services Support via the STRIDES Initiative. Funding supported projects across NCI including within the Cancer Research Data Commons (CRDC)

Awarded Projects

- HALO: Enterprise-wide 2-D Imaging and Digital Pathology Cloud Platform
- A Sustainable Medical Imaging Challenge Cloud Infrastructure (MedICCI)
- DCEG Analytic Tools Suite
- COMETS Analytics
- TP53 Website Migration to NCI CRDC from WHO

NIH> National Institutes of Health

NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

Success Story

The NCI Cancer Research Data Commons (CRDC) is a **cloud-based data science infrastructure** that provides secure access to a large, comprehensive, and expanding collection of cancer research data. Users can explore and use analytical and visualization tools for data analysis in the cloud.

- Data Commons Framework and Data Repositories
  - Genomic Data Commons in both GCP and AWS
  - Proteomic Data Commons in AWS
  - Imaging Data Commons in GCP
- Cloud Resources
  - ISB-CGC hosts several specialty databases (e.g., TP53 database) on GCP

The CRDC contains over **7PB worth of data** and has benefitted greatly from the STRIDES discounts, cloud credits, and free space allocated to the CRDC in both **Google Cloud and AWS**, allowing the CRDC to focus these savings on additional repository and tool building within the CRDC**.**

Cost Savings

- **Total across CSPs: ~$2.6M**

National Institutes of Health

CRDC Radiogenomics: Machine Learning Research in the Cloud

Goal: Use deep learning and radiomics to predict mutation status of gliomas from pre-operative MRI scans.

> The days when a researcher could download data to the computer under their desk are rapidly fading. The NCI Imaging Data Commons, with its connections to the other data types (genomics, proteomics, clinical) in the Cancer Research Data Commons, provides an **efficient means to solve important multimodal AI problems using cloud-scale resources** that will advance biomedical science and the care of patients.
>
> –**Bradley Erickson**, MD, PhD, Professor of Radiology and Medical Director of AI at Mayo Clinic

IDC
- Imaging Data Commons (IDC)
- Cohort exploration
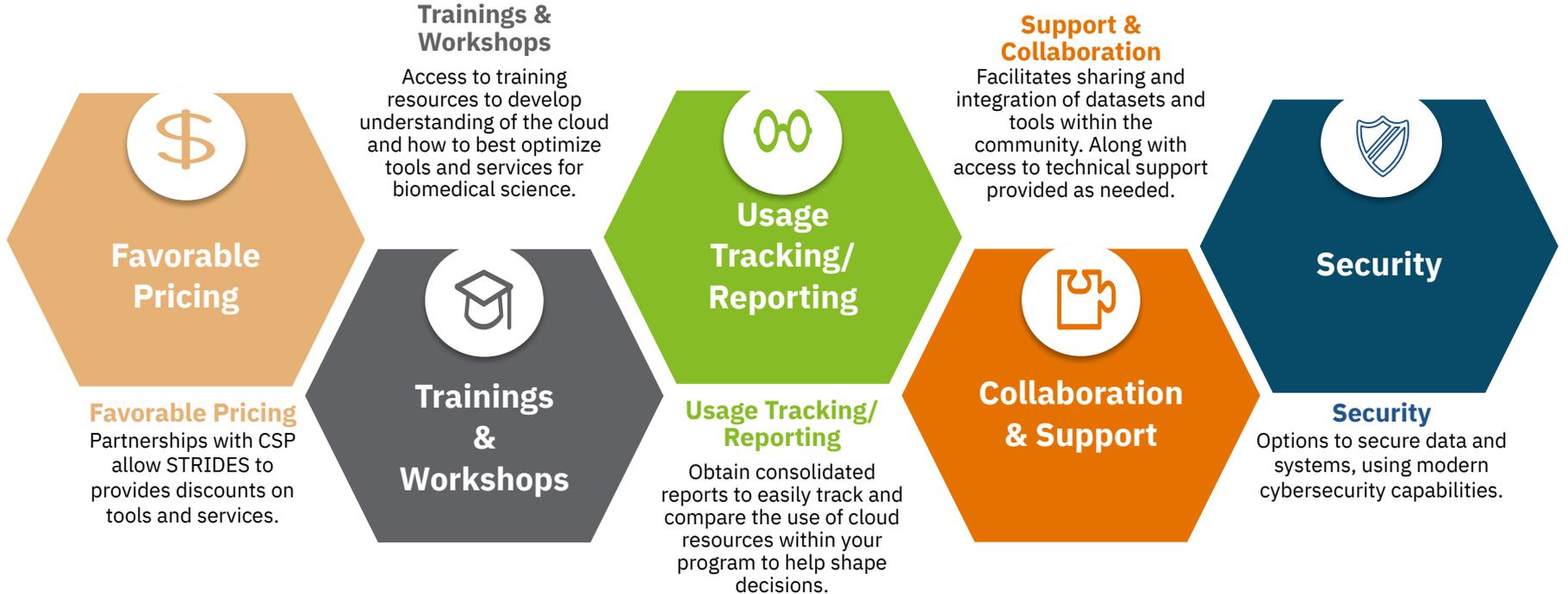- Imaging data preparation and QA

GDC
- Genomics Data Commons (GDC)
- Obtain mutation status
- Obtain demographics

GCP
- Google Cloud Platform (GCP)
- Match imaging & genomic data
- ML model development & evaluation

Slide courtesy of Conte and Erickson (Mayo Clinic)

Benefits of the NIH STRIDES Initiative

Benefits of STRIDES

Trainings & Workshops
Access to training resources to develop understanding of the cloud and how to best optimize tools and services for biomedical science.

Support & Collaboration
Facilitates sharing and integration of datasets and tools within the community. Along with access to technical support provided as needed.

Favorable Pricing

Usage Tracking/ Reporting

Security

Trainings & Workshops

Collaboration & Support

Favorable Pricing
Partnerships with CSP allow STRIDES to provides discounts on tools and services.

Usage Tracking/ Reporting
Obtain consolidated reports to easily track and compare the use of cloud resources within your program to help shape decisions.

Security
Options to secure data and systems, using modern cybersecurity capabilities.

STRIDES Support Services

Cloud Service Provider (CSP) Technical Account Management (TAM)
Designated technical touchpoint for CSP utilization and account support

CSP Subject Matter Expertise/Architecture Support
Consultative architectural guidance delivered in the context of your applications and use cases

Professional Services
Complex solutioning led by CSP experts in a particular cloud technology such as high-performance computing or machine learning, or areas of study such as genomics or RNA sequencing

Google Cloud

Enterprise Support
- Available to intramural partners
- 24/7 Technical Support / Incident Response
- Basic architecture and security recommendations
- Migration best practices
- Office Hours – Standing meeting with Google engineers and guest speakers, open NIH GCP community

Professional Services
- Available to intramural/extramural partners
- No discounts
- https://www.carahsoft.com/google/NIH-STRIDES-initiative/

85

STRIDES Training

- Course offerings range from fundamentals, to research support to technical topics

- Custom courses with content and examples specific to biomedical research, meant to address researcher needs and challenges

Instructor-Led Training   On-Demand Training   Office Hours with Technical Experts & SMEs   Topical Webinars

Contact the STRIDES Training Team at: STRIDESTraining@nih.gov

Visit the STRIDES Training website at: cloud.nih.gov/training

Upcoming GCP Courses

- 10/20: GCP Fundamentals – Big Data & ML
- 11/2: Introduction to Biomedical Data Science in Google Cloud (Custom)
- 11/16: Data Driven Transformation with Google Cloud
- 12/2: Getting Started with Terraform for Google Cloud
- 12/12: Introduction to Biomedical Data Science in Google Cloud (Custom)
- 12/16: Governance and Cost Optimization for Google Cloud Projects

View the **STRIDES Training calendar** for all upcoming trainings with all CSPs: cloud.nih.gov/training/calendar

Barriers to Cloud Still Exist

- Not knowing, *a priori,* which cloud platform to use or how different services compare to one another (within or across cloud platforms)

- Having to coordinate internal funding at NIH (MOUs and DCCs) and/or contractual agreements with resellers for extramural institutions

- Not understanding how much to budget

- Needing to understand and identify all the different roles/responsibilities needed to support a researcher's work on the cloud before the work begins

- Not understanding fully what the transition to cloud will entail, and being hesitant to "jump in with both feet"

NIH Cloud Lab

NIH Cloud Lab is a no-cost, 90-day pilot program that **enables NIH researchers to try commercial cloud services** in a NIH-approved environment. Trainings and guardrails are provided to protect against financial and security risks.
Let us know you're interested at: cloud.nih.gov/resources/cloudlab

Exploring the Cloud Consoles with Full Access
Researchers can gain an understanding of the look and feel of cloud environments before they jump into a full STRIDES account for research. Examples of actions include:

- Deploy a full range of resources
- CPU or GPU VMs
- Managed Jupyter notebooks
- Advanced AI/ML capabilities
- Bioinformatic workflow managers
- Access to compute clusters

Supplementing Cloud Training with Biomedical Tutorials
Researchers can use the sandbox to strengthen their understanding of cloud training or follow along with training content in a separate environment. Examples of included tutorials (with more being added) are:

- Variant Calling
- GWAS
- Medical Imaging
- RNA seq
- Single Cell RNA seq
- Proteomics
- Using HPC environments in the cloud

Experimenting with Simple Cloud Solutions
Researchers interested in solutions for specific scientific tasks can use the sandbox to build proof of concept or other simple solutions to understand LOE and other details for production.

Benchmarking Costs
Testing out different tools and configurations (instance types, sizes, etc.) to optimize research analyses

89

NIH Cloud Lab Tutorials

NIH Cloud Lab for GCP

There are a lot of resources available to learn about GCP, which can be overwhelming. NIH Cloud Lab's goal is to make cloud very easy and accessible for you, so that you can spend less time on administrative tasks and focus on your research.

Use this repository to learn about how to use GCP by exploring the linked resources and walking through the tutorials. If you are a beginner, we suggest you begin with this Jumpstart section. If you already have foundational knowledge of GCP and cloud, feel free to skip ahead to the tutorials section for in-depth examples of how to run specific workflows such as genomic variant calling and medical image analysis.

Overview of Page Contents

- Getting Started
- Overview
- Command Line Tools
- Ingest and Store Data
- Virtual Machines
- Disk Images
- Jupyter Notebooks
- Creating Conda Environments
- Managing Containers
- Serverless Functionality
- Clusters
- Billing and Benchmarking
- Cost Optimization
- Managing Your Code
- Getting Support
- Additional Training
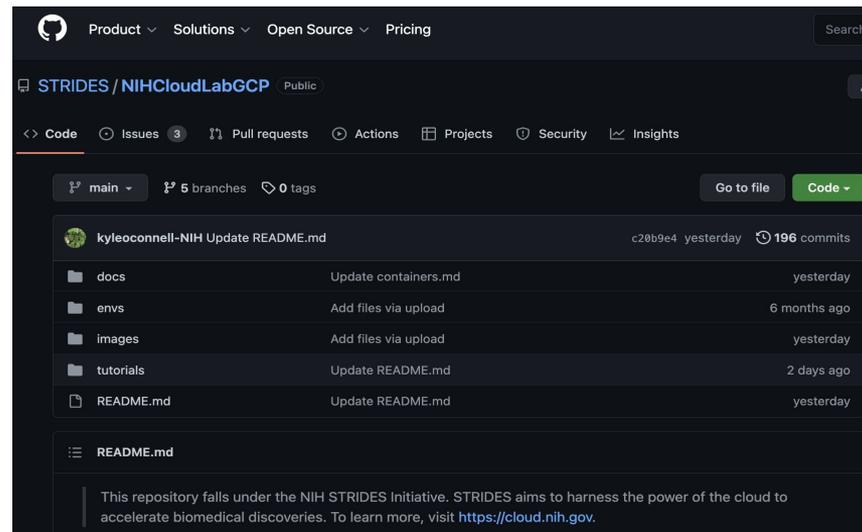
A collection of bioinformatic and other scientific and data science tutorials are available on GitHub at the links below. These will be continuously updated with additional tutorials.

AWS: github.com/STRIDES/NIHCloudLabAWS

GCP: github.com/STRIDES/NIHCloudLabGCP

Getting Started

- NCI projects should first contact their internal IT organization CBIIT

  (Center for Biomedical Informatics and IT)

  - Reviews any request for a cloud account

  - Validates use case and funding lines

  - Directs to STRIDES team for account provisioning

- Contact information
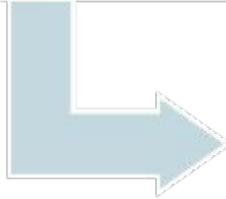
  Contact Sue Pan at: pansu@nih.gov

  NCI CBIIT: datascience.cancer.gov

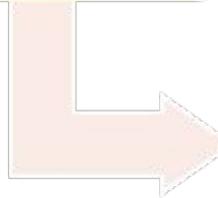# Reaching the Cloud via NIH STRIDES

Select CSP

- Select the Cloud Service Provider(s) on whose platform you will work
  - AWS (Amazon Web Services)
  - Google Cloud Platform (Google)
  - Azure (Microsoft)

Establish Funding Pathway

- Determine the invoicing pathway that best suits your needs
  - NIH Invoicing
  - Awardee Invoicing

Determine Cloud Management

- Determine the cloud management approach that best suits your needs
  - NIH Enterprise Cloud
  - External Cloud

To begin the enrollment process, please contact the STRIDES team at STRIDES@nih.gov or visit cloud.nih.gov/enrollment

Engage with the CIT Cloud Services (aka STRIDES) Team

Are you looking for additional information?

Visit our website at https://cloud.nih.gov/ to learn more about the STRIDES Initiative

Book a consultation with a STRIDES team member about onboarding, costs, security, or technology at: https://outlook.office365.com/owa/calendar/STRIDES@bookings.nih.gov/bookings/

Create a service ticket using ServiceNow: https://myitsm.nih.gov/sp
(Select 'NIH Scientific Computing Services' > 'Cloud Computing Request') – *NOTE: available for internal NIH staff only*

Do you have specific questions?

Contact the STRIDES team at: STRIDES@nih.gov

Questions?

Please visit the NIH STRIDES website at cloud.nih.gov

Or contact the NIH STRIDES team at STRIDES@nih.gov

NIH National Institutes of Health

NIH
STRIDES
INITIATIVE

Thank You!

Google Cloud