Google Cloud Platform: Healthcare Solutions Playbook

Cloud training, tools, and reference architectures for researchers - Prepared for the National Institutes of Health (NIH) STRIDES.

National Institutes of Health

Author(s): Adrienne Wilkins, Trenton Leslie
Title: Technical Account Manager (TAM), Intern

Contents

Introduction 2

Learning 2

Data Management Solutions 4

Reference Architectures 5

- Genomics and Secondary Analysis 5
- Patient Monitoring 9
- Variant Analysis 11
- Healthcare API for Machine Learning and Analytics 14
- Radiological Image Extraction 18

Appendix **21**

Introduction

The National Institutes of Health (NIH) established The Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) initiative to provide biomedical researchers with access to advanced, cost-effective, cloud-based computational infrastructure, tools, and services. Through STRIDES, researchers can take advantage of emerging data management methodologies, technological expertise, computational platforms, and tools to support cutting-edge experimentation and innovation. NIH has partnered with Google Cloud to support the STRIDES initiative through cloud services. In support of STRIDES, we've developed sets of playbooks to help enable researchers to build healthcare and life sciences solutions on Google Cloud Platform (GCP).

The goal of this playbook it to aid researchers in developing healthcare systems and workflows on GCP. This playbook will provide GCP architecture and solutions examples for implementing ***genomics and secondary analysis***, ***patient monitoring***, ***variant analysis***, ***radiological image extraction***, and ***usage of the healthcare API***. Additionally, this playbook will outline training and digital resources to help upskill and enable researchers to build on Google Cloud, while highlighting the appropriate products and services to use when architecting on GCP.

Learning

Generally, cloud adopters fall under one of three categories:

- **Cloud Novice**- Little to no understanding of the cloud
- **Cloud Ready**- Familiar with the cloud, some experience
- **Cloud Native** - Lots of cloud experience, expert-level knowledge

Understanding this broad spectrum of experience levels, we've highlighted key training resources to help upskill researchers on Google Cloud. Additionally, Google offers onsite, instructor-led training to enable large groups of participants across your organization.

Cloud Novice Training	Cloud Ready Training	Cloud Native Training
Video: Welcome to GCP	Documentation: All GCP Products & Services	Documentation: All GCP Products & Services
Video: Intro to GCP for Students	Documentation: GCP for Data Center Professionals	Documentation: GCP Solutions
Video: GCP 101	Documentation: GCP for AWS Professionals	Documentation: Cloud Healthcare API
Video: GCP Essentials	Documentation: GCP for Azure Professionals	Video: Healthcare in the Cloud
Documentation: GCP Conceptual Overview	Documentation: GCP for OpenStack Users	Video: Predictions with Healthcare API
Documentation: About GCP Services	Video: The Future of Health	Video: Transform Healthcare with ML
Documentation: GCP Development & Admin Tools	Video: Uncovering Clinical Insights	Video: Agility in Healthcare
Documentation: All GCP Products & Services	Video: Medical Imaging 2.0	Video: Bringing Clinical Data to GCP
Virtual Course: GCP Fundamentals - Core Infrastructure	Virtual Course: GCP Fundamentals - Core Infrastructure	Video: Mining Clinical Notes with GCP
Virtual Lab: GCP Essentials	Virtual Course: Essential Cloud Infrastructure - Foundation	Video: Health Taxonomy - NLP for Health Research
	Virtual Course: Essential Cloud Infrastructure - Core Services	Video: Genomic Analysis on GCP
	Virtual Course: Elastic Cloud Infrastructure - Scaling and Automation	Virtual Course: Architecting with Google Cloud Platform
	Virtual Course: Elastic Cloud Infrastructure - Containers and Services	Virtual Course: Big Data and Machine Learning Fundamentals
	Virtual Lab: Cloud Architecture	Virtual Course: Leveraging Unstructured Data with Cloud Dataproc on Google Cloud Platform
	Virtual Lab: Cloud Engineering	Virtual Course: Serverless Data Analysis with Google BigQuery and Cloud Dataflow
	VIrtual Lab: Cloud Development	Virtual Course: Serverless Machine Learning with Tensorflow on Google Cloud Platform

		Virtual Course: Building Resilient Streaming Systems on Google Cloud Platform Virtual Lab: Data Engineering Virtual Lab: Data Science on GCP Virtual Lab: Data Science on GCP - Machine Learning Virtual Lab: Google Cloud Solutions II - Data and Machine Learning

Data Management Solutions

Google offers a number of healthcare and life sciences solutions to aid researchers in realizing the potential of data through holistic insights.

Google's Healthcare API can be used to accelerate the ingestion, storage, and integration of key data types, such as FHIR, HL7, and DICOM. The Healthcare API can also be used with the Apigee Healthcare APIx to help organizations easily develop healthcare and patient applications that are securely hosted, easy to use, and capable of analyzing patient data.

Tools such as BigQuery, Cloud Machine Learning Engine and Tensorflow enable researchers to apply analytics and artificial intelligence to data. With these tools, users can establish mechanisms to auto-detect patterns, predict clinical outcomes, and to quickly analyze large amounts of data.

Additionally, both G Suite and GCP are FedRAMP compliant and HITRUST CSF certified; prescriptive security controls are implemented throughout Google Cloud for processing, storing, and transmitting sensitive data. Both G Suite and GCP also support HIPAA compliance across dozens of products in all regions, zones, network paths, and points of presence for Google Cloud.

Learn more about Google Cloud's healthcare solutions and innovative capabilities.

Reference Architectures

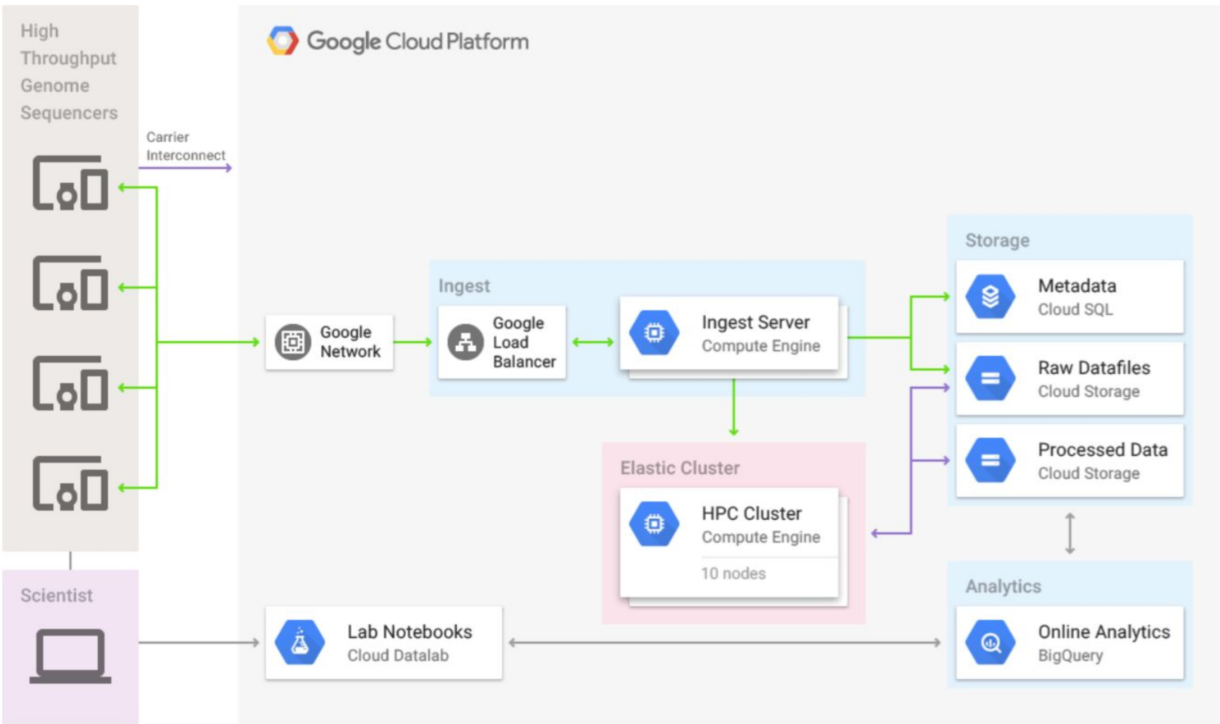
(link)

Google Cloud supports a wide variety of healthcare systems and workflows, from genomics and variant analysis to patient monitoring and machine learning on health records. Here's a few common healthcare use cases, and how to implement those solutions in Google Cloud.

Genomics and Secondary Analysis

(link)

Send sequence data to Google Cloud for processing, analytic analysis, and share results to interactive lab notebooks.



Solution Summary:

Use the cloud to process data from your existing, on-premise sequencers. Establish a high-bandwidth, Interconnect from on-premise to your Google Cloud VPC network. As data is sent from sequencers over Interconnect, it's load balanced to ingest server(s) that subsequently stores the raw data in Google Cloud Storage (GCS) and metadata about the raw data in Cloud SQL. After the raw data is stored in GCS, a job is triggered and submitted to an HPC cluster in the cloud. The HPC cluster processes the raw sequence reads into contigs or aligned BAM files. The processed data is then made accessible to scientists through interactive notebooks, and BigQuery is used to further analyze the sequence data.

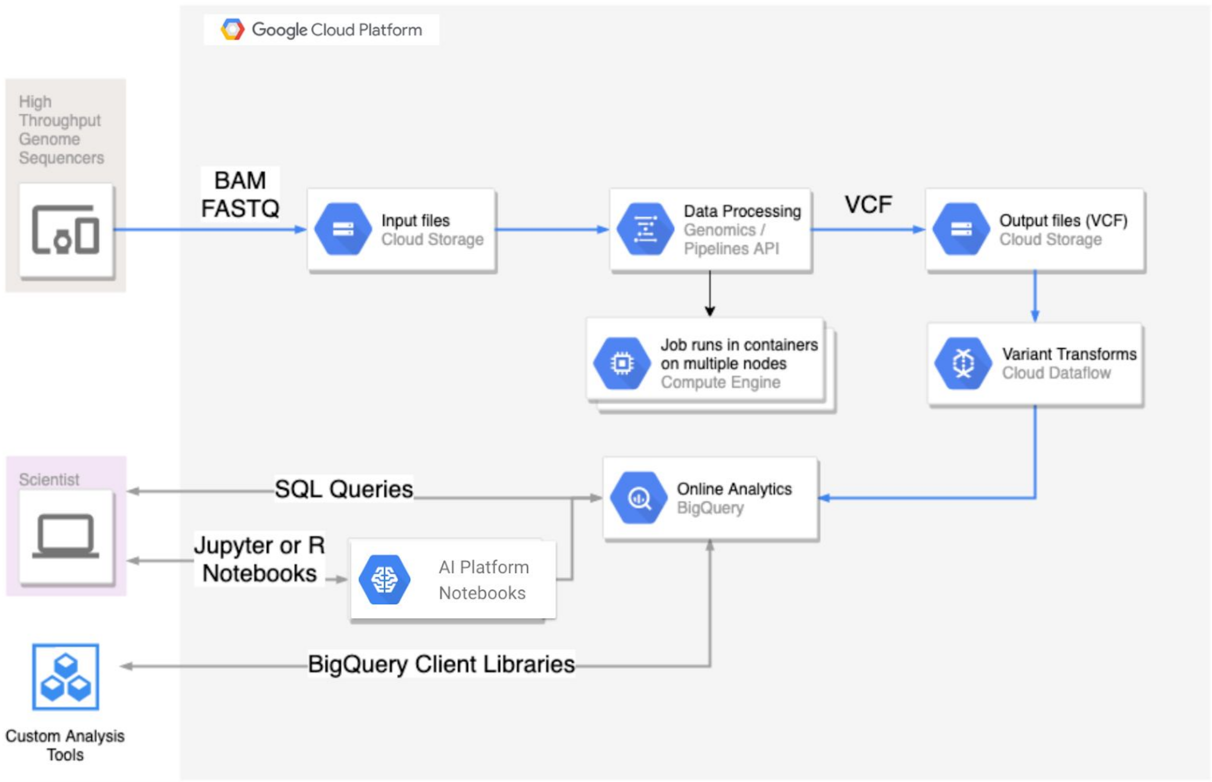
Suggested GCP Products and Services:

- Cloud Interconnect - for high-capacity connectivity from on-premise to your GCP network VPC through a Dedicated connection or using a Partner/Carrier connection
- Cloud VPC Network - Virtual Private Cloud provides global, scalable, and flexible networking for your GCP resources and services without compromising integrity on public-facing connections
- Google Load Balancer - To distribute load-based resources while meeting high availability requirements
- Google Compute Engine (GCE) - VM instances for processing data
- Cloud Genomics: Pipelines API - Google's API for running workflows in containers on GCE instances.
- Google Datalab - Google's interactive lab notebook built on Jupyter
- AI Platform Notebooks - A **managed service** that offers an integrated JupyterLab environment; create instances running JupyterLab that come pre-installed with the latest data science and machine learning frameworks.
- Cloud SQL - Fully **managed** relational database to support MySQL or PostgreSQL data storage.
- Cloud Storage - Unified object storage for storing raw data files and processed data. Supports regional, multi-regional, archive, and infrequently accessed data.
- BigQuery - Analytics data warehouse for large-scale datasets.

High-level Setup Steps:

1. Create a Google Cloud Project and use Cloud IAM to manage who has access to the resources within the GCP project
2. In the GCP project, create a VPC network to logically isolate your project resources
3. Provision a Dedicated Interconnect or a Partner Interconnect to extend your on-premise environment to GCP over a high-bandwidth connection
4. In the VPC network, use GCE VMs to create a high performance compute (HPC) cluster for data processing, and spin up ingest servers to receive the data
5. Create a MySQL or PostgreSQL instance that'll store metadata from the raw sequence data that's sent to the ingest servers
6. Create a Cloud Storage bucket that will store the raw sequence data from the ingest servers
7. Configure the load balancer service to distribute incoming sequence data to the ingest servers

8. Use Cloud Functions to create a GCS trigger that kicks off a job on the HPC cluster to process the raw sequence data reads from GCS. The HPC job should process the raw sequence reads into contigs or aligned BAM files
9. Load the GCS data into BigQuery or use BigQuery to query the data directly from GCS, to run further analysis
10. Create Cloud Datalab instances to share the processed data via notebooks, or use AI Platform (recommended) to create notebooks with native Jupyter support.



High-level Setup Steps (using Pipelines API):

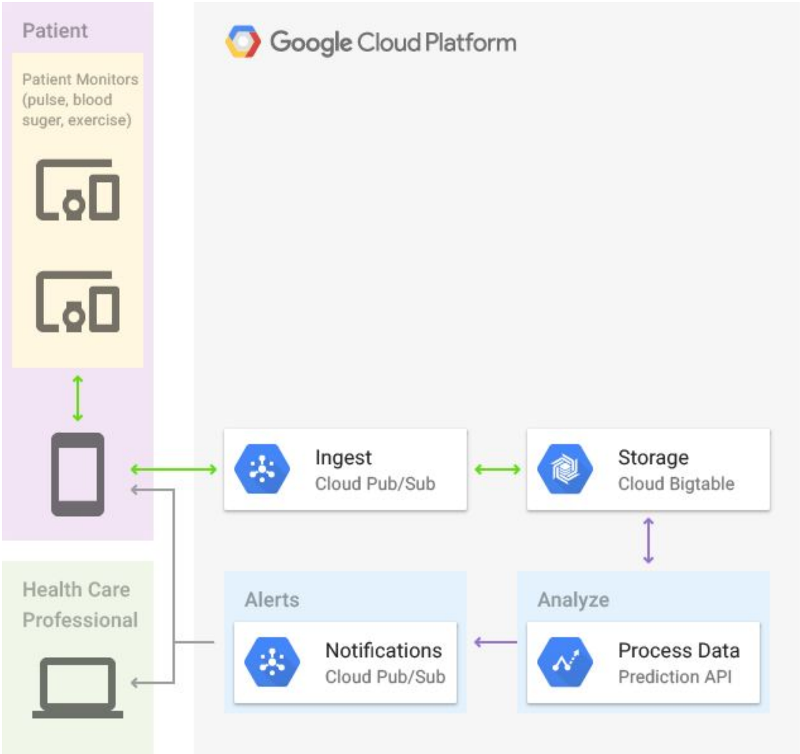
1. Create a Google Cloud Project and use Cloud IAM to manage who has access to the resources within the GCP project
2. Provision a Dedicated Interconnect or a Partner Interconnect to extend your on-premise environment to GCP over a high-bandwidth connection
3. Create a Cloud Storage bucket that will store the raw sequence data from the ingest servers
4. Configure the load balancer service to distribute incoming sequence data to the ingest servers
5. Package the analytic workflow into containers and define the workflow in a WDL or CWL config file or use an existing workflow such as GATK (tutorial).

6. Use a workflow engine (such as Cromwell or dsub) to schedule and run the analytics workflow using Pipelines API.
7. Use Variant Transforms to load the variant files (VCF) to BigQuery for further analysis.
8. Use BigQuery to analyze the variants.
9. Create Cloud Datalab instances to share the processed data via notebooks, or use AI Platform (recommended) to create notebooks with native Jupyter support.

Patient Monitoring

(link)

Securely send patient data to Google Cloud for processing, analytic analysis, and collaboration on results.



Solution Summary:

Information is ingested from patient monitoring devices (eg. blood sugar readings, heart rate, etc.) into connected cloud services for storage and analysis. As data is stored, advanced analytics using the prediction API or Tensorflow, are updated to determine patient risk. Notifications are triggered based on predetermined criteria, and can be sent to the patient, additional health care professionals, and researchers for further collaboration.

Suggested GCP Products and Services:

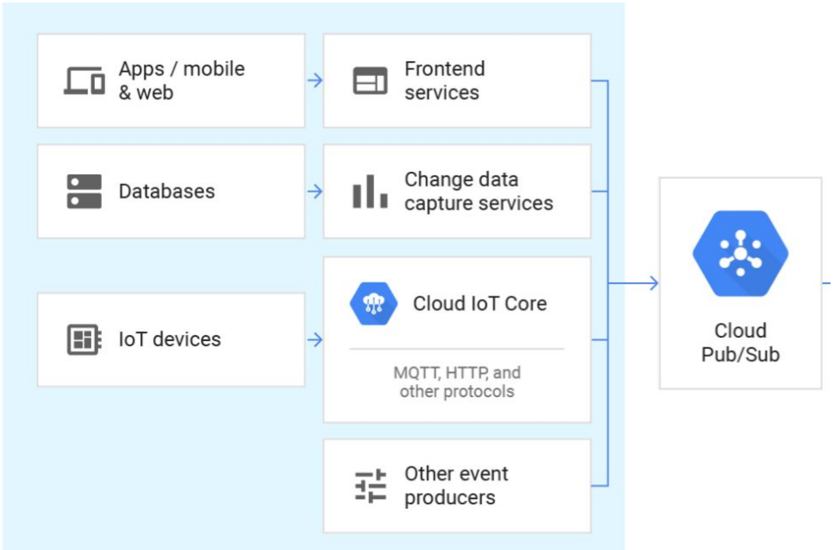
- Cloud VPC Network - Virtual Private Cloud provides global, scalable, and flexible networking for your GCP resources and services without compromising integrity on public-facing connections
- Cloud IoT Core - A fully managed service that provides a complete solution for collecting, processing, analyzing, and visualizing IoT (Internet of Things) data in real time

- Cloud Pub/Sub - a simple and reliable staging location for pushing (pub) and pulling (sub) event data on its journey towards processing, storage, and analysis
- Cloud Bigtable - A large-scale, fully managed NoSQL database service for operational workloads
- Cloud Machine Learning Engine (now part of the Google AI Platform) - The successor to Google Prediction API, MLE is a managed service that allows developers and data scientists to build and run production-level machine learning models with the purpose of optimizing data analysis and data flow.

High-level Setup Steps:

1. Create a Google Cloud Project and use Cloud IAM to manage who has access to the resources within the GCP project.
2. In the GCP project, create a VPC network to logically isolate your project resources.
3. Build out a Cloud Pub/Sub system to ingest events & messages data from patient monitoring devices.
4. (Optional) Use Cloud IoT Core to create registries of IoT devices, making it easy to securely connect, manage, and ingest data from millions of globally dispersed devices. Device data captured by Cloud IoT Core gets published to Cloud Pub/Sub for downstream analytics.
5. Create a Google Cloud Storage bucket that will store event and message notifications exported from Pub/Sub.
6. In AI Platform, create an ML training job against your patient device data stored in GCS, then create predictions from the trained models you developed.
7. Cloud Pub/Sub integrates with various Cloud APIs, including machine learning, big data, and storage APIs. Use this integration to send notifications to patients, other healthcare professionals, researchers, and end devices.

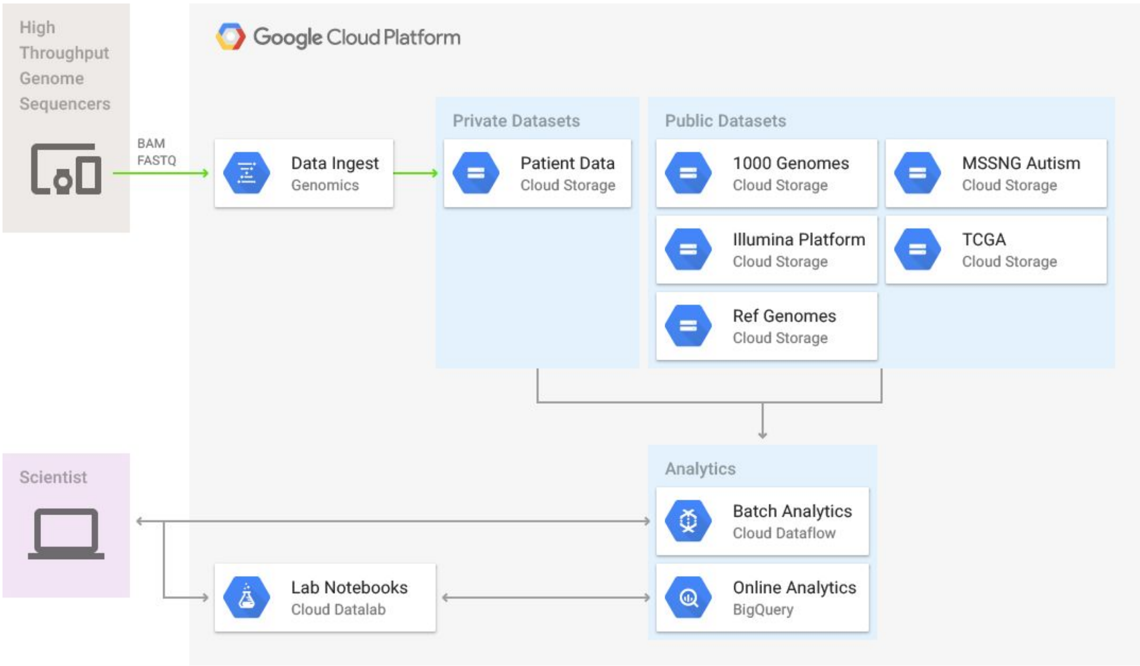
Optional Cloud IoT setup:



Variant Analysis

(link)

Efficiently process and analyze large genomic datasets using Google Genomics. Run experiments in parallel to get results sooner, sharing results and insights with collaborators and broader communities.



Solution Summary:

Cloud Genomics offers petabyte scale genomic data processing and analysis to meet the needs of bioinformatic researchers. Genomics API leverages Google's unique big data technology to store and manage large volumes of sequence data in FASTQ or BAM format. Using the API and command line tools, users can upload files into private and shared repositories.

In addition to private datasets, users can access a large corpus of public datasets available natively in GCS. Both batch analytics using Cloud Dataflow, or interactive analytics using BigQuery and DataLab are supported.

Suggested GCP Products and Services:

- Cloud VPC Network - Virtual Private Cloud provides global, scalable, and flexible networking for your GCP resources and services without compromising integrity on public-facing connections
- Google Genomics - Ask bigger questions by efficiently processing petabytes of genomic data
- BigQuery - Analytics data warehouse for large-scale datasets
- Cloud Storage - Unified object storage for developers and enterprises
- Cloud Datalab - An easy-to-use interactive tool for data exploration, analysis, visualization, and machine learning.
- AI Platform Notebooks - A **managed service** that offers an integrated JupyterLab environment; create instances running JupyterLab that come pre-installed with the latest data science and machine learning frameworks.
- Cloud Dataflow - Simplified stream and batch data processing, with equal reliability and expressiveness

High-level Setup Steps:

1. Create a Google Cloud Project and use Cloud IAM to manage who has access to the resources within the GCP project.
2. In the GCP project, create a VPC network to logically isolate your project resources.
3. Create Google Cloud Storage buckets to store public and private datasets in the cloud; select storage classes based on availability requirements and bucket locations that are close to users. Use IAM to control bucket access policies. Implement additional encryption for sensitive data, and consider a requester pays model for qualified datasets.
    a. Review strategies for transferring big datasets to GCP, and Google's various data transfer options including Online Transfer, GCS Transfer, Transfer Appliance, and BigQuery Transfer Service.
    b. Use Cloud Genomics Public Datasets, which have already been loaded into Cloud Genomics BigQuery and GCS buckets
4. Use Cloud Genomics to create an ingest pipeline of sequence data & metadata from BAM and FASTQ files. Store and load genomic variants in Google Cloud Storage buckets.
    a. Use Cloud Genomics Public Datasets, which have already been loaded into Cloud Genomics BigQuery and GCS buckets
5. Create dataset(s) in BigQuery, which will store the pipeline data for further analysis. Create BigQuery jobs to load, export, query, or copy data from GCS on your behalf.
6. Create an Apache Beam pipeline and use Cloud Dataflow to run the batch analytics pipeline, which will analyze data stored in GCS using BigQuery. Use Dataflow to read data from and write data to BigQuery.
7. Use Cloud Datalab or AI Platform (recommended) to create notebooks for visualizing BigQuery data. AI Platform is recommended for native Jupyter notebook integration.

8. Use notebooks and Dataflow to share analytics data, visualizations, and insights from BigQuery with additional teams and researchers.

Healthcare API for Machine Learning and Analytics

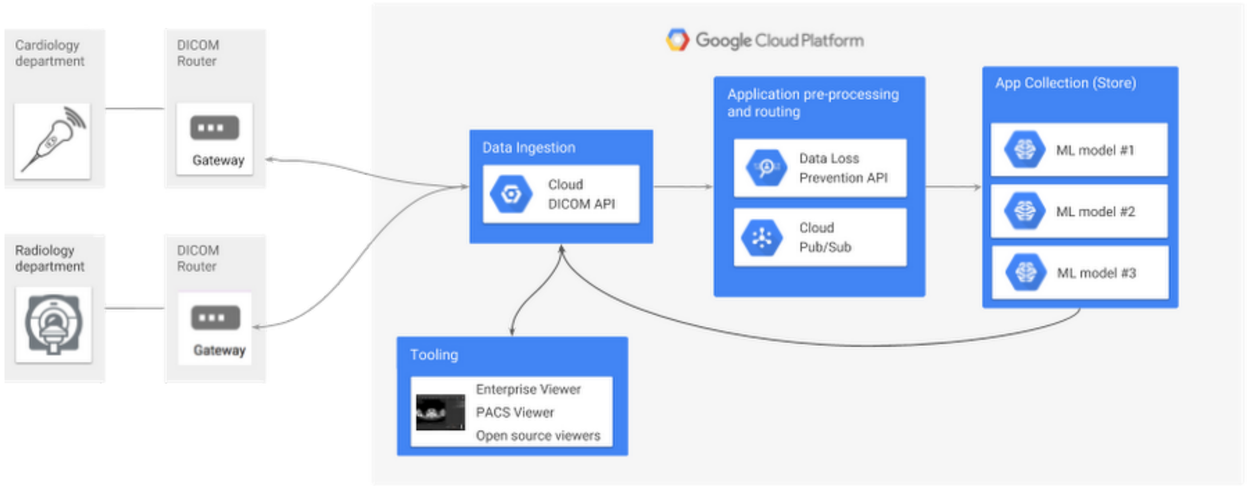
Solution Summary

Use the Healthcare API to easily integrate machine learning's powerful analytical insights on top of new and existing medical data. With a focus on healthcare applications, the Google Healthcare API gives developers an efficient tool set to transform propositions into solutions at work. Google's machine learning has precise recognition that can be trained with high accuracy to serve an array of predictive functions.

Suggested GCP Products and Services:

- Cloud VPC Network - Virtual Private Cloud provides global, scalable, and flexible networking for your GCP resources and services without compromising integrity on public-facing connections.
- Cloud Healthcare API - Standards-based APIs powering actionable healthcare insights for security and compliance-focused environments.
- Cloud Pub/Sub - a simple and reliable staging location for pushing (pub) and pulling (sub) event data on its journey towards processing, storage, and analysis.
- Cloud Storage - Unified object storage for storing raw data files and processed data. Supports regional, multi-regional, archive, and infrequently accessed data.
- Cloud Dataflow - Simplified stream and batch data processing, with equal reliability and expressiveness.
- Cloud Dataproc - A fast, easy-to-use, fully managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler, more cost-efficient way.
- BigQuery - Analytics data warehouse for large-scale datasets.
- Cloud Datalab - An easy-to-use interactive tool for data exploration, analysis, visualization, and machine learning.
- Data Loss Prevention API - 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 GCP credentials.
- Cloud AutoML - Suite of machine learning products that enables developers with limited machine learning expertise to train high-quality models specific to their business needs.

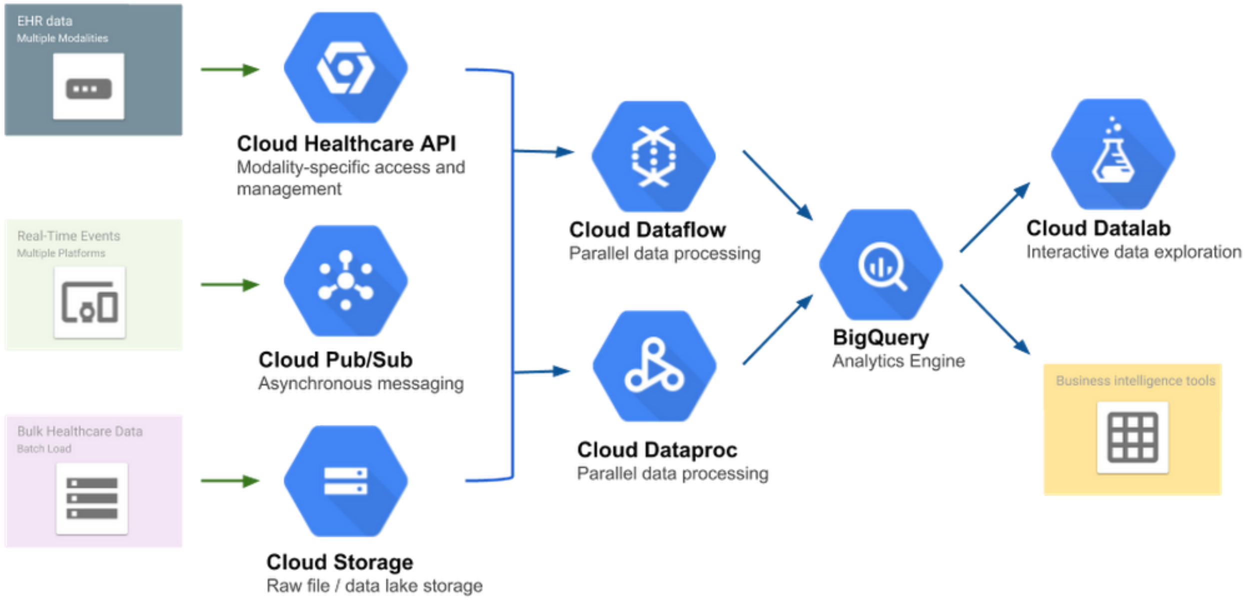
Machine Learning Reference (link)



High-level Setup Steps:

1. Create a Google Cloud Project and use Cloud IAM to manage who has access to the resources within the GCP project.
2. In the GCP project, create a VPC network to logically isolate your project resources.
3. Create service account(s) that will enable you to authenticate to the Healthcare API.
4. To use the Cloud Healthcare API for ingesting, storing, analyzing, and integrating healthcare data with cloud-based applications, create dataset(s) to support the healthcare data models you want to analyze.
   a. Overview of Healthcare API datasets and data stores.
   b. Enable dataset de-identification to remove PII or obscure sensitive information
5. Create DICOM, HL7, and/or FHIR stores that will house ingested medical imaging, clinical event messaging, and clinical resource data.
6. Cloud Pub/Sub integrates with various Cloud APIs, including machine learning, big data, and storage APIs. Use this integration to ingest notifications and messages from Healthcare API and send them to CloudAutoML for things like natural language processing, ML modeling, translation, video intelligence, and vision classification.

Analytics Reference (link)



High-level Setup Steps:

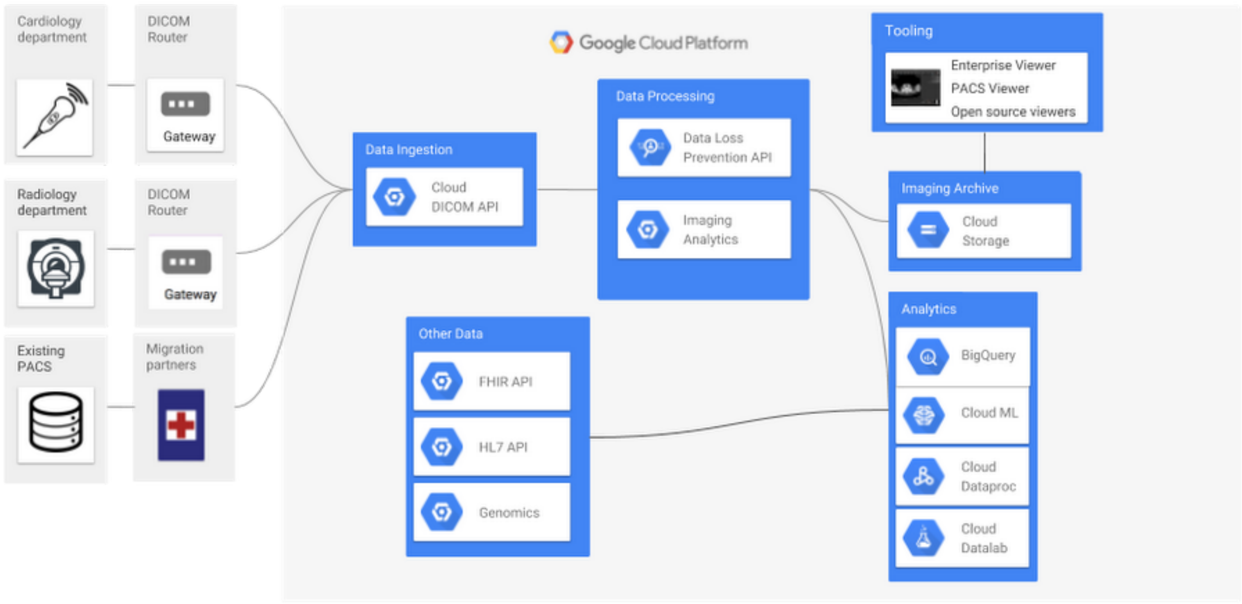
1. Create a Google Cloud Project and use Cloud IAM to manage who has access to the resources within the GCP project.
2. In the GCP project, create a VPC network to logically isolate your project resources.
3. Create service account(s) that will enable you to authenticate to the Healthcare API.
4. To use the Cloud Healthcare API for ingesting, storing, analyzing, and integrating healthcare data with cloud-based applications, create dataset(s) to support the healthcare data models you want to analyze.
    a. Overview of Healthcare API datasets and data stores.
    b. Enable dataset de-identification to remove PII or obscure sensitive information
5. Create DICOM, HL7, and/or FHIR stores that will house ingested medical imaging, clinical event messaging, and clinical resource data.
6. Create Google Cloud Storage buckets to store raw electronic health records and medical files. Select storage classes based on availability requirements and bucket locations that are close to users. Use IAM to control bucket access policies. Implement additional encryption for sensitive data, and consider a requester pays model for qualified datasets.
    a. Review strategies for transferring big datasets to GCP, and Google's various data transfer options including Online Transfer, GCS Transfer, Transfer Appliance, and BigQuery Transfer Service.
    b. Create a Google Cloud Storage bucket that will store event and message notifications exported from Pub/Sub.

7. Cloud Pub/Sub integrates with various Cloud APIs, including machine learning, big data, and storage APIs. Use Pub/Sub to great an ingestion pipeline of messaging streams for Dataproc or Dataflow
8. Create dataset(s) in BigQuery, which will store the pipeline data for further analysis. Create BigQuery jobs to load, export, query, or copy data from GCS on your behalf.
9. Create an Apache Beam pipeline and use Cloud Dataflow to run the batch analytics pipeline, which will analyze data stored in GCS using BigQuery. Use Dataflow to read data from and write data to BigQuery.
10. Create a Dataproc cluster and get the BigQuery Connector for Dataproc, to submit a Dataproc job that further processes the ingested medical data.
11. Use Cloud Datalab or AI Platform (recommended) to create notebooks for visualizing BigQuery data. AI Platform is recommended for native Jupyter notebook integration.
12. Use notebooks and Dataflow to share analytics data, visualizations, and insights from BigQuery with additional teams and researchers.

Radiological Image Extraction

(link)

Use Google Cloud to gain insights from devices and equipment. Ingest common medical data formats for interoperable processing and analytics.



Solution Summary:

Google Cloud is home to a collection of products and services that can be customized to fit unique customer needs. In the case of Radiological Image Extraction, healthcare professionals and their associates can connect to the Google Cloud Platform to upload, analyze and share findings in near-real time.

With trainable machine learning, GCP becomes a powerful tool in analyzing large datasets of images to produce meaningful results. On a further step, one can use Google's Healthcare API's to establish unique systems to easily parse discrete information from analytic results and compare this current data against historical data.

Suggested GCP Products and Services:

- Cloud VPC Network - Virtual Private Cloud provides global, scalable, and flexible networking for your GCP resources and services without compromising integrity on public-facing connections.

- Cloud Healthcare API (supports/replaced DICOM, FHIR, HL7 APIs) - Supports existing healthcare data standards, formats, and protocols to bridge the gap between traditional on-premise care applications & systems, enabling them in the cloud.
  - DICOM - Digital Imaging and Communications in Medicine (DICOM) is an international standard used for medical images such as X-rays, MRIs, ultrasounds, and other medical imaging modalities. The Cloud Healthcare API provides operations for reading DICOM instances, studies, and series that are consistent with the DICOMweb standard, and supports the DICOM DIMSE C-STORE operation using an open-source adapter.
  - FHIR - Fast Healthcare Interoperability Resources (FHIR) is a healthcare standard for representing and exchanging electronic medical information.
  - HL7v2 - Health Level Seven International Version 2 (HL7v2) is a clinical messaging format that provides data about events that occur inside an organization.
- Data Loss Prevention API - 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 GCP credentials.
- AI Platform - Collection of machine learning and data analytics tools for developers, data scientists, and data engineers.
  - Cloud Machine Learning Engine - The successor to Google Prediction API, MLE is a managed service that allows developers and data scientists to build and run production-level machine learning models with the purpose of optimizing data analysis and data flow.
- BigQuery - Analytics data warehouse for large-scale datasets.
- BigQuery ML (use instead of Cloud ML) - Enables users to create and execute machine learning models in BigQuery using standard SQL queries.
- Cloud Storage - Unified object storage for developers and enterprises.
- Cloud Dataproc - A fast, easy-to-use, fully managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler, more cost-efficient way.
- Cloud Datalab - An easy-to-use interactive tool for data exploration, analysis, visualization, and machine learning.

High-level Setup Steps:

1. Create a Google Cloud Project and use Cloud IAM to manage who has access to the resources within the GCP project.
2. In the GCP project, create a VPC network to logically isolate your project resources.
3. Create service account(s) that will enable you to authenticate to the Healthcare API.
4. To use the Cloud Healthcare API for ingesting, storing, analyzing, and integrating healthcare data with cloud-based applications, create dataset(s) to support the healthcare data models you want to analyze.
   a. Overview of Healthcare API datasets and data stores.

   b. Enable dataset de-identification to remove PII or obscure sensitive information
5. Create DICOM, HL7, and/or FHIR stores that will house ingested medical imaging, clinical event messaging, and clinical resource data.
6. Create Google Cloud Storage buckets to store imaging and clinical data archives in the cloud. Select storage classes based on availability requirements and bucket locations that are close to users. Use IAM to control bucket access policies. Implement additional encryption for sensitive data, and consider a requester pays model for qualified datasets.
7. Create dataset(s) in BigQuery, which will store the analyzed data.
8. (Optional) Create a Dataproc cluster and get the BigQuery Connector for Dataproc, to submit a Dataproc job that further processes the ingested medical data.
9. Using the Healthcare API, ingest and manage medical imaging, clinical events and clinical messaging data. Configure DICOM routers and/or use GCP's data transfer options to migrate medical data to the cloud.
10. Export DICOM, HL7, and/or FHIR data to GCS for backup or archive storage. Additionally, send DICOM, HL7, and/or FHIR data to BigQuery for analysis.
11. (Optional) Use BigQuery ML to create and train ML models based on the medical data collected and analyzed in BigQuery
12. Use Cloud Datalab or AI Platform (recommended) to create notebooks for visualizing BigQuery data; AI Platform is recommended for native Jupyter notebook integration.

Appendix

Healthcare Solutions Reference Architectures (link)

Reference Architecture for Creating a HIPAA-Aligned Project in GCP (link)

Importing FHIR clinical data into the cloud using the Cloud Healthcare API (link)

Getting to know the Google Cloud Healthcare API - Part 1 (link)

Getting to know the Google Cloud Healthcare API - Part 2 (link)

Getting to know the Google Cloud Healthcare API - Part 3 (link)