Google Cloud

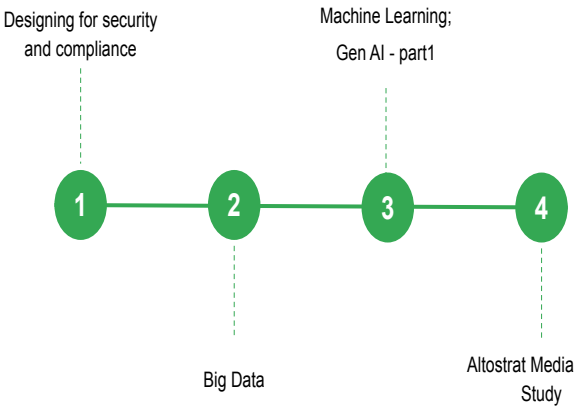
Professional Cloud Architect

Preparing for Professional Cloud Architect Journey for AWS Professionals

Plan:

- ~10 mins overall slide. Cover ETL / ELT / EL; Bigquery storage vs query; advanced services (Looker), ETL coordination / management (Fusion, Composer) and Data Catalog
- ~20 rest of Big Data
- ~8 min Machine Learning
- ~5 mins Service Mesh
- ~15 mins case study with questions

Session 6 topics

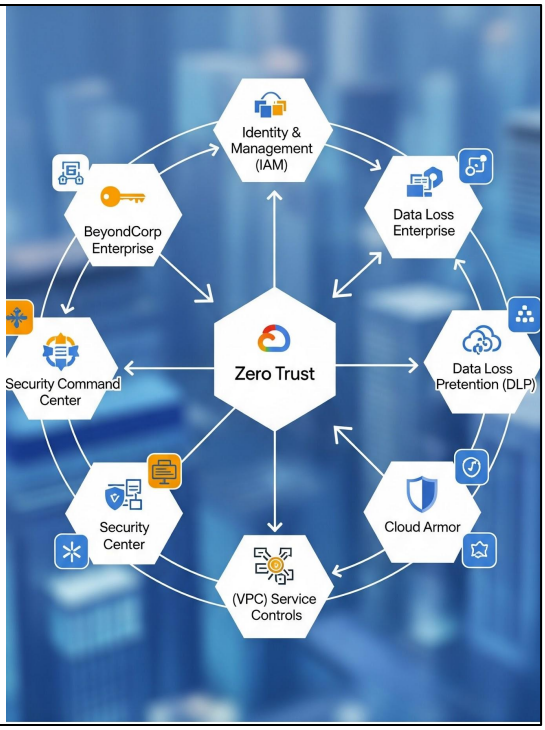


Designing for security and compliance

Google's approach to Zero Trust

.... and other security measures

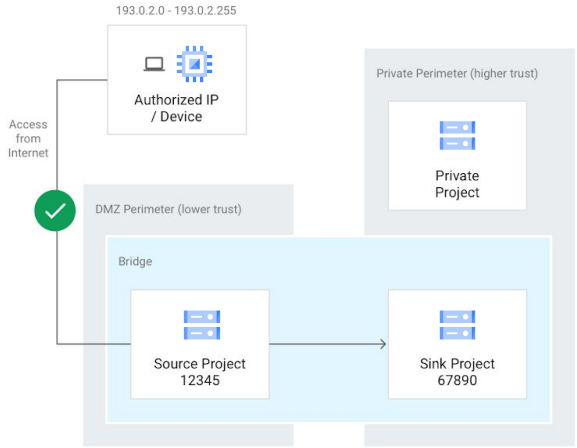
- Intro to Zero Trust
  - Zero Trust at Google
- Security by design
- Shared responsibility and shared fate
- Shifting security left
- Secure and responsible AI
- How to use AI for security



- go through the links and emphasize the most important areas
- AI-related security shall be covered next week

VPC Service Controls (no AWS equivalent)

It can also allow communication between two perimeters using a service perimeter bridge



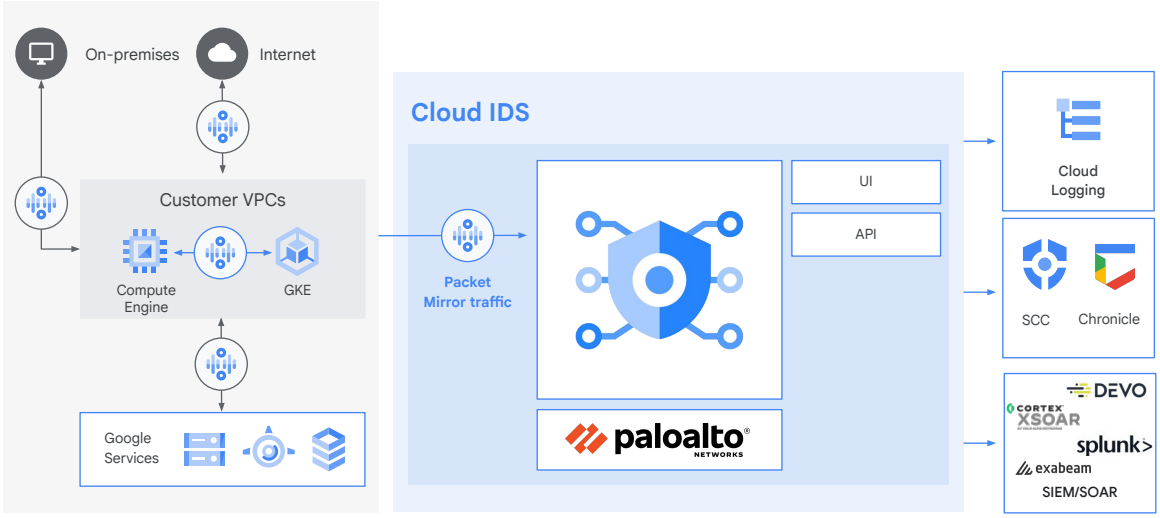
Can VPC Service Controls be used in a hybrid cloud environment? Yes, they can!

Perimeter bridges can be used to enable communication between projects in different service perimeters. Keep in mind that a project can belong to more than one perimeter *bridge* but can only be included in one service perimeter.

Protecting sensitive data with Sensitive Data Protection (Macie, Comprehend)

- Find Personally Identifiable Information (PII), such as name, email, address, and credit card number
- Allow custom data types to be defined
- Works with data in Cloud Storage (Macie scans S3 only), BigQuery, Datastore, and other locations, including data outside Google Cloud (Comprehend is text only, up to 100 KB per API call)
- Works with structured and semistructured data (.pdf, .docx, .csv, .pptx, etc.) as well as unstructured data such as images (.bmp, .jpeg, .png, etc.)
- Optionally redact (natively with DLP; via integration with Comprehend and Lambda in AWS) sensitive information (replacement character, data type [e.g. name], encrypted version of the data, hash, etc.)

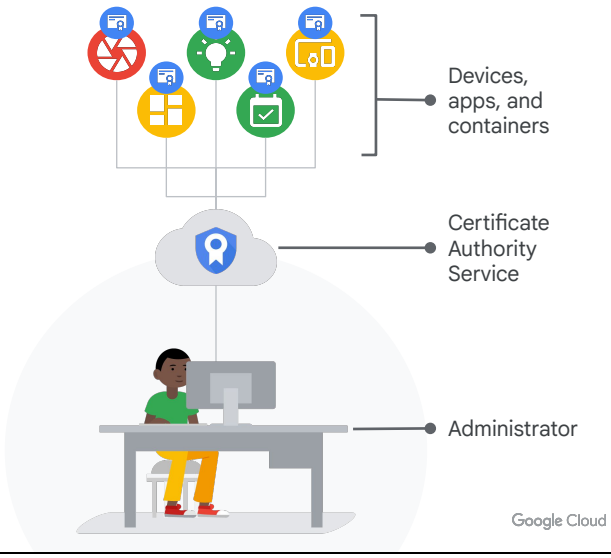
Cloud IDS



Certificate Authority Service (Certificate Manager Private Certificate Authority)

Simplify and automate the deployment and management of private certificate authorities (CAs) while staying in control of your private keys.

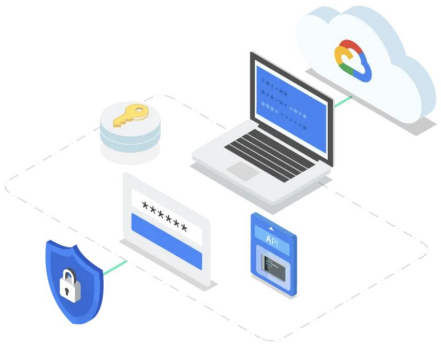
- Simpler deployment and management
- Tailored for you
- Enterprise-ready



Secret management (Secrets Manager)

Secret Manager

- Secure storage system for API keys, passwords, certificates, and other sensitive data
- Single source of truth to manage, access, and audit (via CloudWatch or CloudTrail) secrets across Google Cloud (AWS)
- Every interaction generates an audit log
- Simple lifecycle management with versioning and the ability to pin requests to the latest version of a secret.
- IAM used to control access



Secret Manager

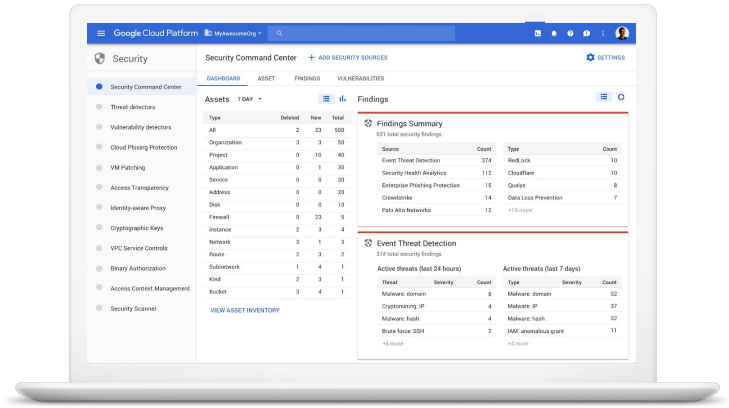
Security Command Center (Security Hub / GuardDuty / Audit Manager / Systems Manager Explorer / Inspector)

Gain centralized visibility and control over your Google Cloud (AWS) data and resources

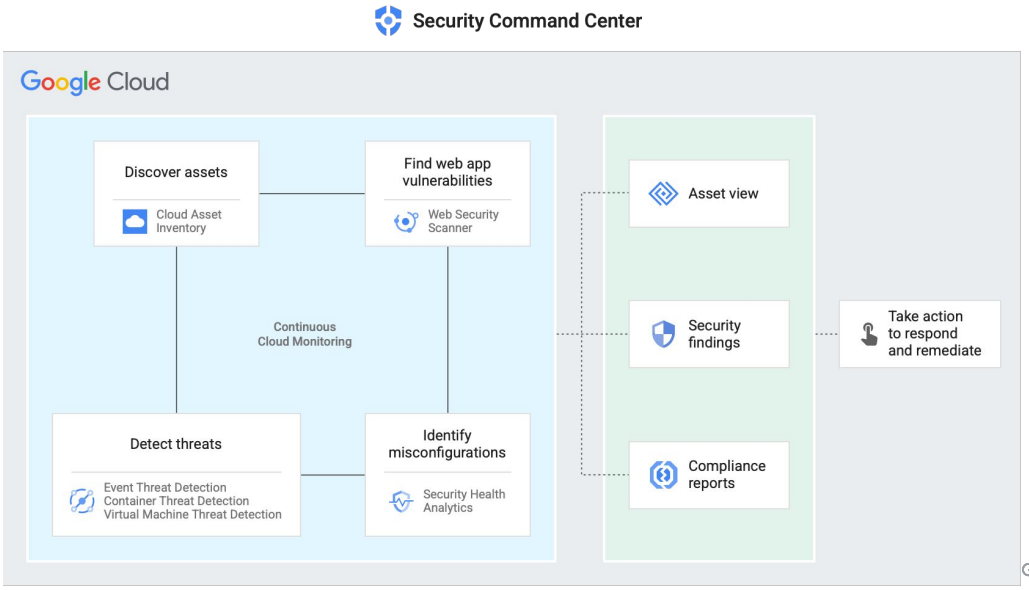
Find and fix risky misconfigurations

Report on and maintain compliance (NIST, PCI, etc.)

Detect threats targeting your Google Cloud (AWS) assets (platform and container levels)



Security Command Center



Docs:
https://cloud.google.com/security-command-center/docs/concepts-security-command-center-overview

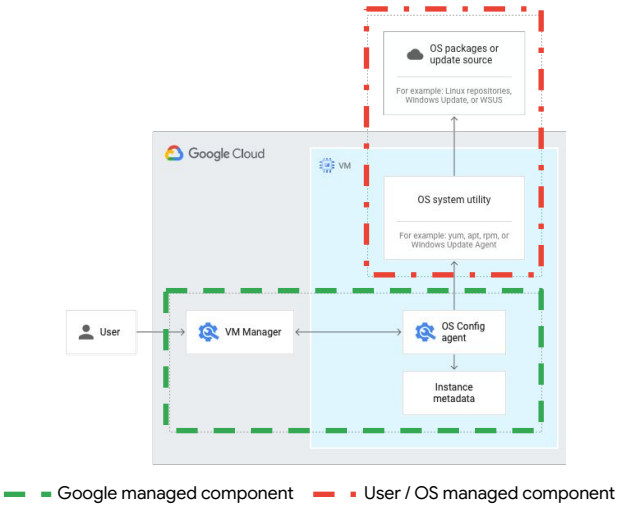
Compliance Resource Center (Compliance Programs web page)

- Google Cloud conforms to global and industry-specific compliance standards
  - GDPR in the EU
  - FedRAMP and SOX in the US
  - HIPPA for the healthcare industry in US
  - PCI for banking
  - Many others...

Data security
Data residency
Data privacy

- **As a Professional Cloud Architect, you should be aware of compliance standards that are required for applications you are building**
  - See: https://cloud.google.com/security/compliance/

VM Manager (Systems Manager)



Components:

- OS Patch Management
  - On-demand and scheduled patching
  - Patch compliance reporting
- OS Inventory Management
  - View installed operating systems and versions
  - View installed apps and versions
  - View missing packages and updates vs. compliance baseline
  - View vulnerability reports
- OS Configuration Management
  - Deploy, configure, and maintain software configurations via OS Policies

VM Manager: https://cloud.google.com/compute/docs/vm-manager

Firewall Insights helps you better understand and safely optimize your firewall rules

Firewall Insights, a component product of Network Intelligence Center, produces metrics and insights that let you make better decisions about your firewall rules. It provides data about how your firewall rules are being used, exposes misconfigurations, and identifies rules that could be made more strict.

Firewall Insights uses Cloud Monitoring metrics and Recommender insights.

Cloud Monitoring collects measurements to help you understand how your applications and system services are performing. A collection of these measurements is generically called a metric. The applications and system services being monitored are called monitored resources. Measurements might include the latency of requests to a service, the amount of disk space available on a machine, the number of tables in your SQL database, the number of widgets sold, and so forth. Resources might include virtual machines, database instances, disks, and so forth.

Recommender is a service that provides recommendations and insights for using resources on Google Cloud. These recommendations and insights are per-product or per-service, and are generated based on heuristic methods, machine learning, and current resource usage. You can use insights independently from recommendations. Each insight has a specific insight type. Insight types are specific to a single Google Cloud product and resource type. A single product can have multiple insight types, where each provides a different type of insight for a different resource.

Using Cloud Monitoring for metrics:

https://cloud.google.com/monitoring/api/v3/metrics

Using Recommender for insights:
https://cloud.google.com/recommender/docs/insights/using-insights

Compliance in GCP - 1/2

- **ISO 27001**
  - Requirements for an information security management system (ISMS), specifies a set of best practices
  - ONLY GUIDANCE, lays out allow Google to ensure a comprehensive and continually improving model for security management.
- **SOC 2**
  - The purpose of this report is to evaluate an organization's information systems relevant to security, availability, processing integrity, confidentiality, and privacy.
  - Relevant are different services: VPC Service Controls, DLP, Cloud Security Command Center, Cloud Armor etc
- **PCI DSS**
  - Appropriate practices that merchants and service providers should follow to protect cardholder data.
  - Relevant are MANY GCP services: networking, logging, encryption etc
- **FIPS 140-2**
  - A security standard that sets forth requirements for cryptographic modules, including hardware, software, and/or firmware, for U.S. federal agencies.
  - Google Cloud Platform uses a FIPS 140-2 validated encryption module called BoringCrypto (certificate 3318) in our production environment. This means that both data in transit to the customer and between data centers, and data at rest are encrypted using FIPS 140-2 validated encryption.

Google Cloud

[In short, GCP fulfills a ton of different regulations and undergoes regular external audits which confirm that the platform is compliant with PCI-DSS, DGPR, HIPAA and others. Now what do you think: does it mean that your systems deployed to GCP are also compliant with those, and why? What do you think?
A: …
Also, please be aware that the exam will focus on SECURITY, so even if a question mentions one of those compliance regulations by name, there will be more information, like it will not only say you need to comply with GDPR, but it will also translate it to technical term, for example asking you "how would you design this or that making sure your data stays in a specific GCP region?"

Compliance in GCP - 2/2

- **HIPAA**
  - Healthcare-related.
  - Complying with HIPAA is a shared responsibility between the customer and Google.
  - Google Cloud Platform supports HIPAA compliance (within the scope of a Business Associate Agreement) but ultimately customers are responsible for evaluating their own HIPAA compliance.
- **FedRAMP**
  - Government-wide program that provides a standardized approach to security assessment, authorization, and continuous monitoring for cloud products and services.
  - Risk impact levels (Low, Moderate, or High)
  - Google is one of the first hyperscale commercial cloud providers to achieve FedRAMP High on a commercial public cloud offering, and is one of the largest providers of FedRAMP services available on the market today.
  - NO SEPARATE 'GOVERNMENT' REGIONS EXIST IN GCP.
- **GDPR**
  - PII data protection in Europe.
  - Our customers own their data and we believe they should have the strongest levels of control over data stored in the cloud. Our public cloud provides customers with world-class levels of visibility and control over their data through our services.
  - Storing data in Europe, optionally manage encryption keys and store them outside of GCP, External Key Manager etc.

Google Cloud

*[****** Transcript on previous slide ******]*

How do you ensure compliance?

By implementing “security-relevant” options!

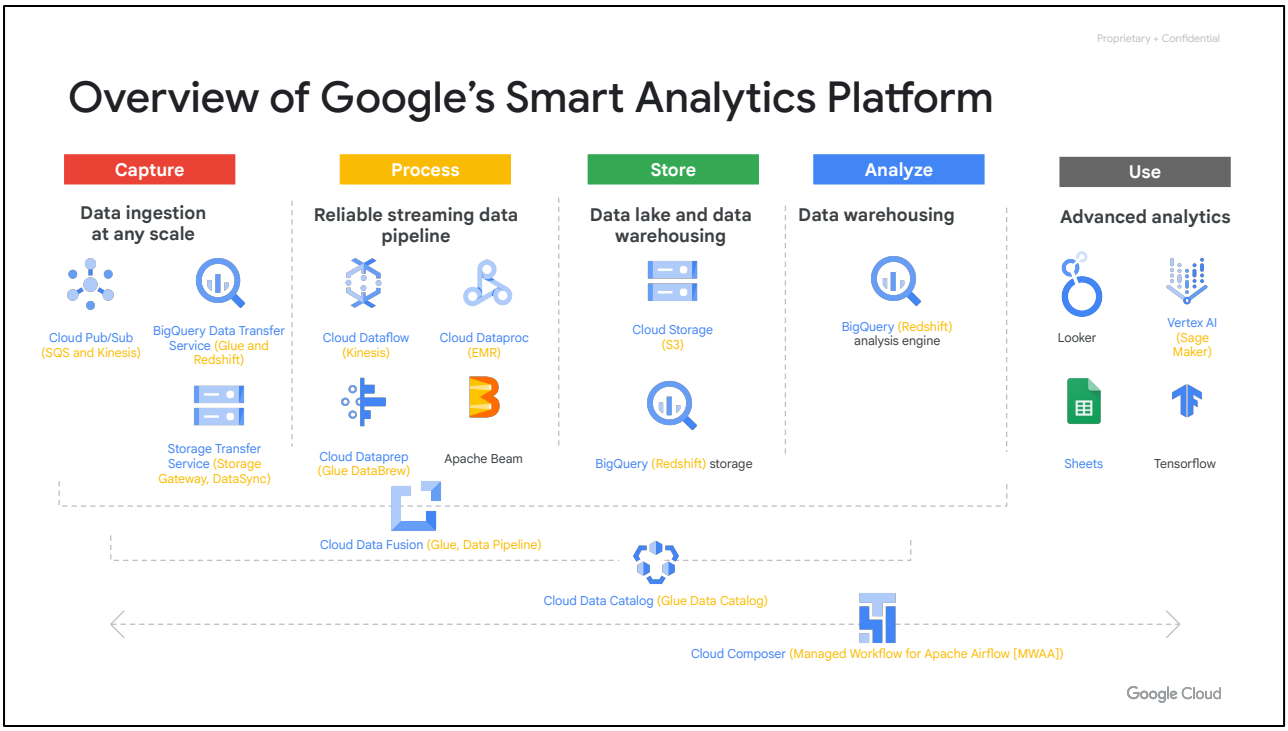
Google Security Overview	Shielded VMs	Identity and Access Management
Access Transparency	Confidential Computing	IAM Conditions
GCP Compliance offerings	Shared VPC	Identity-Aware Proxy
Binary Authorization	VPC Service Controls	Resource Manager
Data Loss Prevention	Cloud Armor	Private Service Connect
Key Management Service	DNSSEC	Private Google Access
Organization Policy Service	Cloud VPN	Serverless VPC Access
Anthos Service Mesh	VPC Flow Logs	Web Security Scanner
Cloud Asset Inventory	Firewall Insights	Cloud Audit Logs
OS Login	Packet Mirroring	Centralized Telemetry

and more…

Google Cloud

[As already mentioned, you should definitely focus on security, which is embedded in each of GCP services you choose. There is no single, global security service, but instead you will increase or decrease security posture of your solutions when planning each part, like when deciding between standard or shared VPCs, using Cloud Armor, choosing how users will ssh into VMs and so on.
And if you’d like to better understand how Google thinks about security on high level, have a look at the red link in the top left corner.]

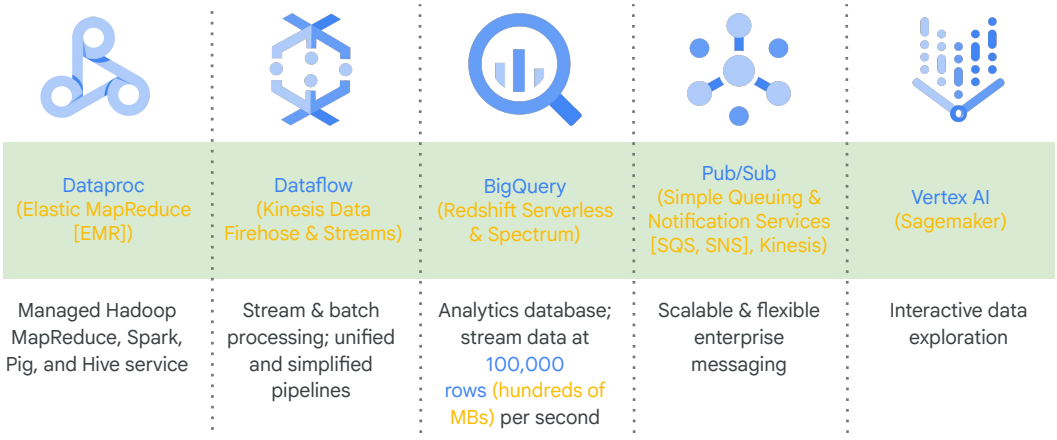
BigData



BigQuery is part of Google Cloud's comprehensive data analytics platform that covers the analytics value chain from Ingest >> process >> store >> advanced analytics and collaboration. BigQuery is deeply integrated with the GCP's analytical and data processing offering, allowing customers to build an enterprise ready cloud native data warehouse.

1. Cloud Pub/sub - Scaled messaging platform
2. DTS - Ads data for marketing cloud
3. Beam - Stream and batch processing with single programing model with Dataflow
4. Dataproc - Managed Hadoop and Spark platform
5. Dataprep - Analyst can now do data prep using visual tool
6. Data Fusion - Fully managed, code-free data integration service to manage ETL/ELT pipelines and also track lineage of that data.
7. BigQuery cloud-native, highly scalable data warehouse
8. GCS as your data lake for structured and unstructured data
9. Cloud ML Engine & Tensorflow for machine learning on top of data on BQ and GCS
10. Data studio and Sheet for your analysis

Google Cloud's big data services are fully managed and scalable



Source: Demo Template

Google Cloud Big Data solutions are designed to help you transform your business and user experiences with meaningful data insights. It is an integrated, serverless platform. "Serverless" means you don't have to provision compute instances to run your jobs. The services are fully managed, and you pay only for the resources you consume. The platform is "integrated" so Google Cloud data services work together to help you create custom solutions.

Dataproc

Elastic MapReduce

Dataproc (Elastic MapReduce) is managed Hadoop

- Fast, easy, managed way to run Hadoop and Spark/Hive/Pig on Google Cloud (AWS).
- Create VM clusters in 90 seconds (several minutes) or less on average.
- Scale clusters up and down even when jobs are running.
- Available in a serverless version as well

Source: Demo Template
Apache Hadoop is an open-source framework for big data. It is based on the MapReduce programming model, which Google invented and published. The MapReduce model, at its simplest, means that one function -- traditionally called the “map” function -- runs in parallel across a massive dataset to produce intermediate results; and another function -- traditionally called the “reduce” function -- builds a final result set based on all those intermediate results. The term “Hadoop” is often used informally to encompass Apache Hadoop itself and related projects, such as Apache Spark, Apache Pig, and Apache Hive.

Dataproc is a fast, easy, managed way to run Hadoop, Spark, Hive, and Pig on Google Cloud. All you have to do is to request a Hadoop cluster. It will be built for you in 90 seconds or less, on top of Compute Engine virtual machines whose number and type you can control. If you need more or less processing power while your cluster’s running, you can scale it up or down. You can use the default configuration for the Hadoop software in your cluster, or you can customize it. And you can monitor your cluster using Stackdriver.

Why use Dataproc?

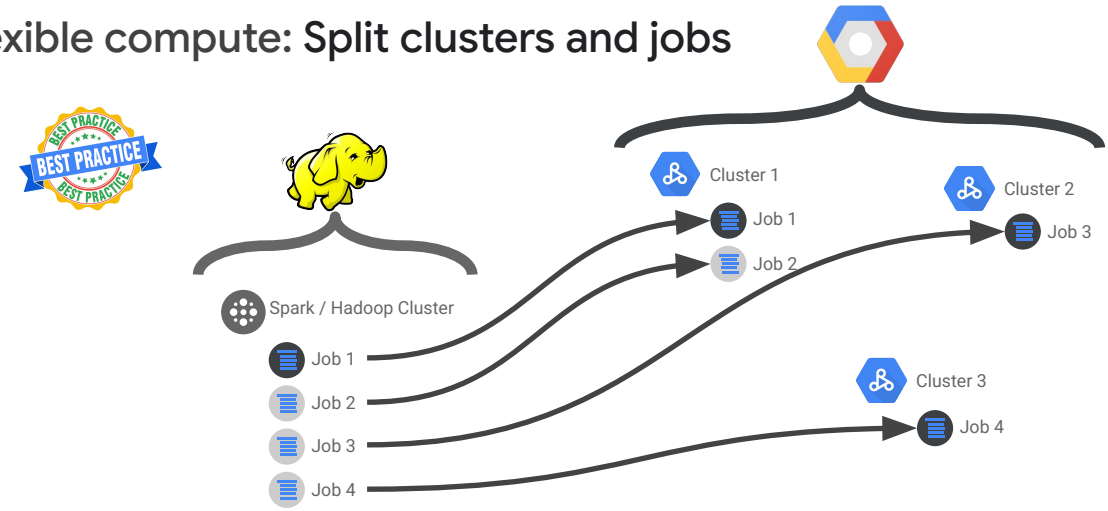
- Easily migrate on-premises Hadoop jobs to the cloud.
- Quickly analyze data (like log data) stored in Cloud Storage (S3); create a cluster quickly, and then delete it immediately or use the serverless option.
- Use Spark/Spark SQL to quickly perform data mining and analysis.
- Use Spark Machine Learning Libraries (MLlib) to run classification algorithms.

Source: Demo Template
Running on-premises Hadoop jobs requires a hardware investment. On the other hand, running these jobs in Dataproc allows you to pay only for hardware resources during the life of the ephemeral customer you create. You can further save money using preemptible instances for batch processing.

You can also save money by telling Dataproc to use preemptible Compute Engine instances for your batch processing. You have to make sure that your jobs can be restarted cleanly if they're terminated and you get a significant break in the cost of the instances. At the time this video was made, preemptible instances were around 80% cheaper. Be aware that the cost of the Compute Engine instances isn't the only component of the cost of a Dataproc cluster, but it's a significant one.

Once your data is in a cluster, you can use Spark and Spark SQL to do data mining, and you can use MLlib, which is Apache Spark's Machine Learning Libraries, to discover patterns through machine learning.

Flexible compute: Split clusters and jobs



Exam Tips:

- *When thinking about **Dataproc**, you should really think about per-job, ephemeral, auto-scaling clusters with auto-shutdown after the task is completed.*
- *Using Spot/Preemptible VMs for **secondary** Dataproc workers is a common pattern.*
- *Switching from HDFS to GCS is also a best practice in most cases.*

Google Cloud

[Bear in mind that Google thinks about Dataproc in a specific way: instead of having a long-term, static, expensive cluster, you would define your Hadoop jobs and schedule them on Dataproc clusters, which are created per job and then either stopped to cost-optimize the resources, or deleted altogether.
Second thing is that In order to delete a Dataproc cluster and not lose the results the job calculated, you would need to offload those results somewhere outside. So you will learn about two best practices, where Google would like you to use GCS buckets instead of HDFS and BigTable instead of HBase - and those rules are also a kind of best practice.
And finally, remember about the isolation -> if you have dev/staging/prod Hadoop jobs in on-premises, you might want to run them on different Dataproc clusters in GCP]

Diagnostic Question Discussion

You need to migrate Hadoop jobs for your company's Data Science team without modifying the underlying infrastructure. You want to minimize costs and infrastructure management effort.

What should you do?

A. Create a Dataproc cluster using standard and spot worker instances.

B. Create a Dataproc cluster using spot worker instances only.

C. Manually deploy a Hadoop cluster on Compute Engine using standard instances.

D. Manually deploy a Hadoop cluster on Compute Engine using spot instances.

https://cloud.google.com/dataproc/docs/concepts/compute/secondary-vms

Google Cloud

Eliminate C & D
Eliminate B -> Spot cannot be used ONLY

- Using preemptible VMs does not always save costs since preemptions can cause longer job execution with resulting higher job costs

Diagnostic Question Discussion

You need to migrate Hadoop jobs for your company's Data Science team without modifying the underlying infrastructure. You want to minimize costs and infrastructure management effort.

What should you do?

A. **Create a Dataproc cluster using standard and spot worker instances.**

B. Create a Dataproc cluster using spot worker instances only.

C. Manually deploy a Hadoop cluster on Compute Engine using standard instances.

D. Manually deploy a Hadoop cluster on Compute Engine using spot instances.

https://cloud.google.com/dataproc/docs/concepts/compute/secondary-vms

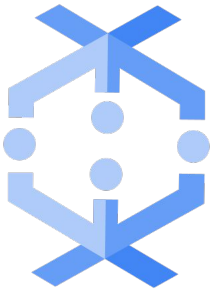
Eliminate C & D
Eliminate B -> Spot cannot be used ONLY

Dataflow

Kinesis Data Firehose (batch) and Streams (streaming), Glue

Dataflow (Kinesis Data Firehose (batch) and Streams (streaming), Glue): managed data pipelines

- Processes data using Compute Engine instances.
  - Clusters are sized for you.
  - Automated scaling, no instance provisioning required (serverless).
- Write code once and get batch (Firehose) and streaming (by adding Kinesis Data Streams).
  - Transform-based programming model.



Source: Demo Template

Dataproc is great when you have a dataset of known size, or when you want to manage your cluster size yourself. But what if your data shows up in realtime? Or it's of unpredictable size or rate? That's where Dataflow is a particularly good choice. It's both a unified programming model and a managed service, and it lets you develop and execute a big range of data processing patterns: extract-transform-and-load, batch computation, and continuous computation. You use Dataflow to build data pipelines, and the same pipelines work for both batch and streaming data.

Dataflow is a unified programming model and a managed service for developing and executing a wide range of data processing patterns including ETL, batch computation, and continuous computation. Dataflow frees you from operational tasks like resource management and performance optimization.

Dataflow features:

Resource Management: Dataflow fully automates management of required processing resources. No more spinning up instances by hand.

On Demand: All resources are provided on demand, enabling you to scale to meet your business needs. No need to buy reserved compute instances.

Intelligent Work Scheduling: Automated and optimized work partitioning which can dynamically rebalance lagging work. No more chasing down “hot keys” or pre-processing your input data.

Auto Scaling: Horizontal auto scaling of worker resources to meet optimum throughput requirements results in better overall price-to-performance.

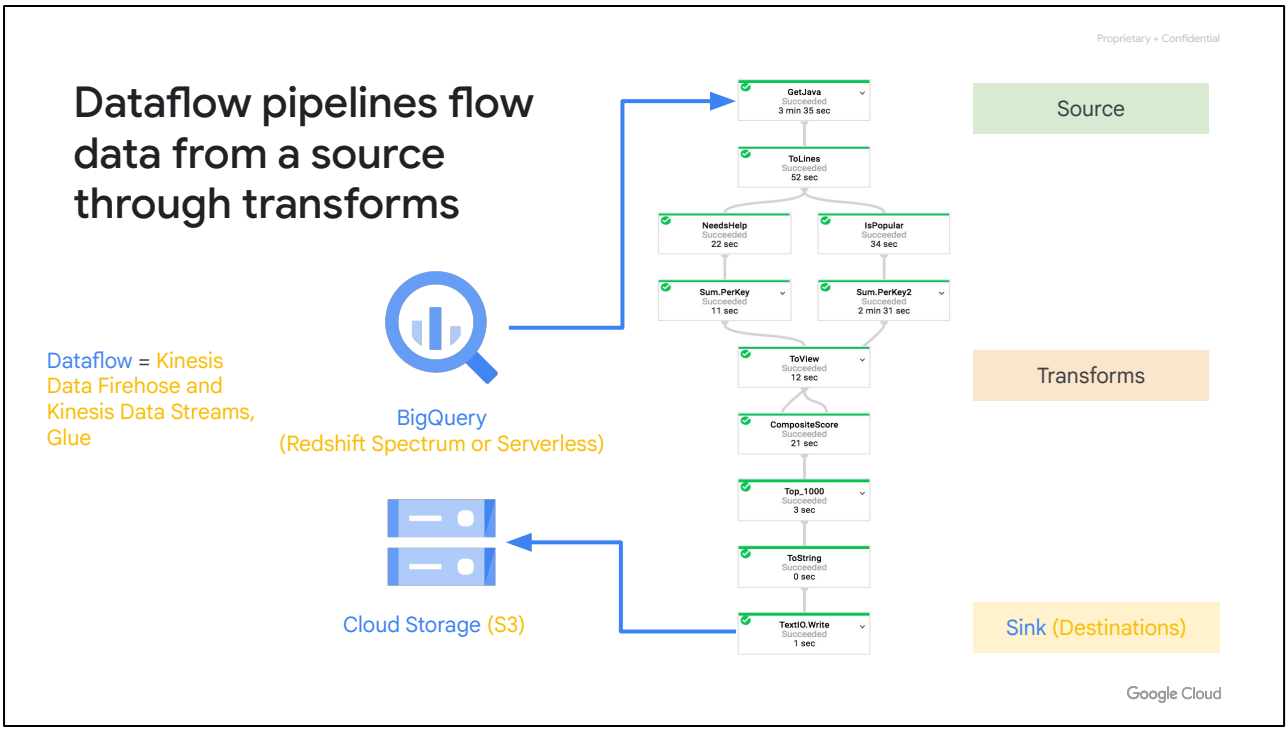
Unified Programming Model: The Dataflow API enables you to express MapReduce like operations, powerful data windowing, and fine grained correctness control regardless of data source.

Open Source: Developers wishing to extend the Dataflow programming model can fork and or submit pull requests on the Java-based Dataflow SDK. Dataflow pipelines can also run on alternate runtimes like Spark and Flink.

Monitoring: Integrated into the Cloud Console, Dataflow provides statistics such as pipeline throughput and lag, as well as consolidated worker log inspection—all in near-real time.

Integrated: Integrates with Cloud Storage, Pub/Sub, Datastore, Cloud Bigtable, and BigQuery for seamless data processing. And can be extended to interact with others sources and sinks like Apache Kafka and HDFS.

Reliable & Consistent Processing: Dataflow provides built-in support for fault-tolerant execution that is consistent and correct regardless of data size, cluster size, processing pattern or pipeline complexity.

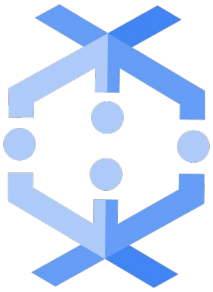


Source: Demo Template
This example Dataflow pipeline reads data from a BigQuery table (the “source”), processes it in various ways (the “transforms”), and writes its output to Cloud Storage (the “sink”). Some of those transforms you see here are map operations, and some are reduce operations. You can build really expressive pipelines.

Each step in the pipeline is elastically scaled. There is no need to launch and manage a cluster. Instead, the service provides all resources on demand. It has automated and optimized work partitioning built in, which can dynamically rebalance lagging work. That reduces the need to worry about “hot keys” -- that is, situations where disproportionately large chunks of your input get mapped to the same custer.

Why use Dataflow?

- *ETL* (extract/transform/load) pipelines to move, filter, enrich, shape data.
- *Data analysis*: batch (via Firehose) computation or continuous computation using streaming (with Streams and Firehose).
- *Orchestration*: create pipelines that coordinate services, including external services.
- Integrates with Google Cloud (AWS) services like Cloud Storage (S3), Pub/Sub, BigQuery (Redshift), and Cloud Bigtable (DynamoDB).
  - Open source Java, Python, .NET. node.js, and Ruby SDKs.

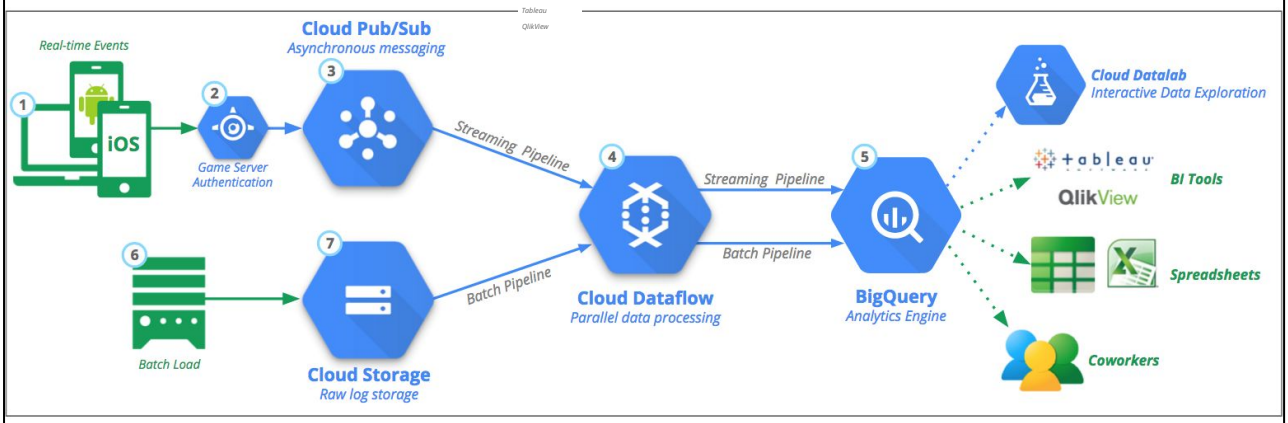


Source: Demo Template
People use Dataflow in a variety of use cases. For one, it serves well as a general-purpose ETL tool.

And its use case as a data analysis engine comes in handy in things like these: fraud detection in financial services; IoT analytics in manufacturing, healthcare, and logistics; and clickstream, Point-of-Sale, and segmentation analysis in retail.

And, because those pipelines we saw can orchestrate multiple services, even external services, it can be used in real time applications such as personalizing gaming user experiences.

Example architecture for data analytics



[And finally, let's just have a look at a very typical data analytics pipeline in GCP. So whenever you get a question about some kind of data ingestion from source (regardless of what that source is) in order to analyze this data in GCP, this is usually a preferred architecture. Dataflow is in the center, it get either batch data from GCS or other service, plus streaming data flowing from some devices via Pub/Sub, and then it cleans up and transforms data so that it's in the proper formats. After data is processed, it goes to BigQuery, where you can directly execute SQL queries, or perform analysis using other tools]

Diagnostic Question Discussion

Your company has successfully migrated to the cloud and wants to analyze their data stream to optimize operations. They do not have any existing code for this analysis, so they are exploring all their options. These options include a mix of batch and stream processing, as they are running some hourly jobs and live- processing some data as it comes in.

Which technology should they use for this?

A. Google Cloud Dataproc

B. Google Cloud Dataflow

C. GKE with Bigtable

D. Google Compute Engine with BigQuery

B

Diagnostic Question Discussion

Your company has successfully migrated to the cloud and wants to analyze their data stream to optimize operations. They do not have any existing code for this analysis, so they are exploring all their options. These options include a mix of batch and stream processing, as they are running some hourly jobs and live- processing some data as it comes in.

Which technology should they use for this?

A. Google Cloud Dataproc

B. Google Cloud Dataflow

C. GKE with Bigtable

D. Google Compute Engine with BigQuery

B

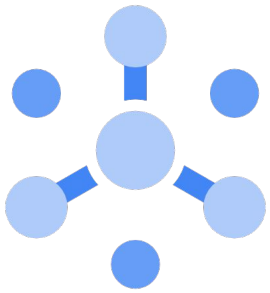
Pub/Sub

Simple Query Service (SQS) / Simple Notification Service (SNS)

https://cloud.google.com/pubsub/docs/subscriber#subscription_type_comparison

Pub/Sub (Simple Query Service (SQS) / Simple Notification Service (SNS)): scalable, reliable messaging

- Supports many-to-many (many-to one [SQS] and one-to-many [SNS]) asynchronous messaging.
  - Application components make push/pull subscriptions to topics (push to SQS queues, pull from SQS queues, SNS pushes to subscribers).
- Includes support for offline consumers.
- Based on proven Google (AWS) technologies.
- Integrates with Dataflow for data processing pipelines.



Source: Demo Template
Pub/Sub is a fully managed real-time messaging service that allows you to send and receive messages between independent applications. You can leverage Pub/Sub's flexibility to decouple systems and components hosted on Google Cloud or elsewhere on the internet. By building on the same technology Google uses, Pub/Sub is designed to provide "at least once" delivery at low latency with on-demand scalability to 1 million messages per second (and beyond).

Pub/Sub features:

Highly Scalable
Any customer can send up to 10,000 messages per second, by default—and millions per second and beyond, upon request.

Push and Pull Delivery
Subscribers have flexible delivery options, whether they are accessible from the internet or behind a firewall.

Encryption
Encryption of all message data on the wire and at rest provides data security and protection.

Replicated Storage
Designed to provide "at least once" message delivery by storing every message on

multiple servers in multiple zones.

Message Queue
Build a highly scalable queue of messages using a single topic and subscription to support a one-to-one communication pattern.

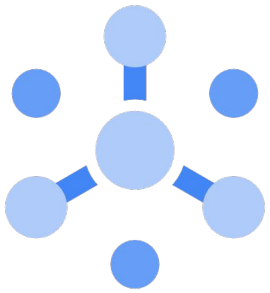
End-to-End Acknowledgement
Building reliable applications is easier with explicit application-level acknowledgements.

Fan-out
Publish messages to a topic once, and multiple subscribers receive copies to support one-to-many or many-to-many communication patterns.

REST API
Simple, stateless interface using JSON messages with API libraries in many programming languages.

Why use Pub/Sub?

- Pub/Sub (Kinesis Data Stream and Firehose) is building block for data ingestion in Dataflow, Internet of Things (IoT), Marketing Analytics, etc.
- Pub/Sub is foundation for Dataflow streaming.
- Pub/Sub (SNS) provides push notifications for cloud-based applications.
- Pub/Sub (SQS) connects applications across Google Cloud (AWS) (push/pull between components (e.g. GCE [EC2] and App Engine [Elastic Beanstalk]).



Source: Demo Template
Pub/Sub builds on the same technology Google uses internally. It's an important building block for applications where data arrives at high and unpredictable rates, like Internet of Things systems. If you're analyzing streaming data, Dataflow is a natural pairing with Pub/Sub.

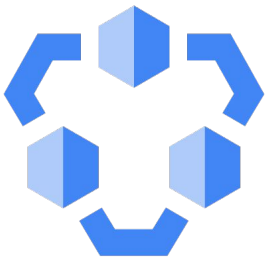
Pub/Sub also works well with applications built on Google Cloud's compute platforms. You can configure your subscribers to receive messages on a "push" or a "pull" basis. In other words, subscribers can get notified when new messages arrive for them, or they can check for new messages at intervals.

Dataplex

Glue Data Catalog

Data Management

Cloud Data Catalog (now part of Dataplex) (Glue Data Catalog): a fully managed and highly scalable data discovery and metadata management service



Organizations faced with a wealth of data spread across disjointed systems need an **effective solution for data discovery**

Offers **unified data discovery** of all data assets, spread across multiple projects and systems

Empowers users to **annotate business metadata** in a collaborative manner

Provides the **foundation** for data **governance**, data **lineage** and data **access control**

AWS v/s Google Cloud: Big data ingestion and storage

Resource	AWS	GCP
Real time ingestion	Kinesis Data Stream, Managed Streaming for Apache Kafka (MSK)	Pub/Sub
Data Lake / Raw Data	S3, Lake Formation	Cloud Storage
Managed ETL	Kinesis Data Streams, Firehose, Glue	Dataflow, BigQuery
Data Catalog	Glue Data Catalog	Cloud Data Catalog
Data Warehouse	Redshift	BigQuery

AWS v/s Google Cloud: Big data analytics, visualization, and protection

Resource	AWS	GCP
Managed Spark / Hadoop	Elastic MapReduce (EMR)	Dataproc
Serverless Analytics	Athena, Redshift Serverless	BigQuery
Data Visualization	Quicksight	Data Studio, Looker
Sensitive data protection	Macie	Data Loss Prevention
Data exfiltration protection	N/A	VPC Service Controls
Messaging	Simple Queue Service (SQS)	Pub/Sub
Notifications	Simple Notification Service (SNS)	Pub/Sub

Data Lake building blocks in GCP

A data lake is a centralized repository designed to store, process, and secure large amounts of structured, semistructured, and unstructured data.
It can store data in its original format and process any variety of it.

Google Cloud offers a suite of autoscaling services that can be used to build a data lake.

Let's explore the options.

First is data ingestion.

Dataflow and Cloud Data Fusion let you ingest data to a data lake.

Dataflow provides unified stream and batch data processing that is serverless, fast, and cost-effective.

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

Next is storage. Cloud Storage provides globally unified, scalable, and highly durable object storage for developers and enterprises.

Lastly, Dataproc and BigQuery support data and analytics processing.

Dataproc makes open source data and analytics processing fast, easy, and more secure in the cloud.

BigQuery is a serverless, highly scalable, and cost-effective cloud data warehouse with built-in machine learning (ML) and business intelligence (BI) that works across clouds.

Data lake services: GCP vs AWS

Topic	AWS	Google Cloud
Data ingestion	AWS Lake Formation and AWS Glue	Dataflow and Cloud Data Fusion
Storage	Amazon S3	Cloud Storage
Data processing	Amazon EMR, Amazon Redshift, and Amazon Athena	Dataproc, BigQuery, and Dataflow

Additional resource: Comparison of data ingestion, storage and data processing in AWS and Google Cloud

AWS and Google Cloud both offer services that can be used to build enterprise data lakes.

Let's explore a high-level view of how the services map across platforms.

For data ingestion:

- AWS uses AWS Lake Formation and AWS Glue, while
- Google Cloud uses Dataflow and Cloud Data Fusion

For storage:

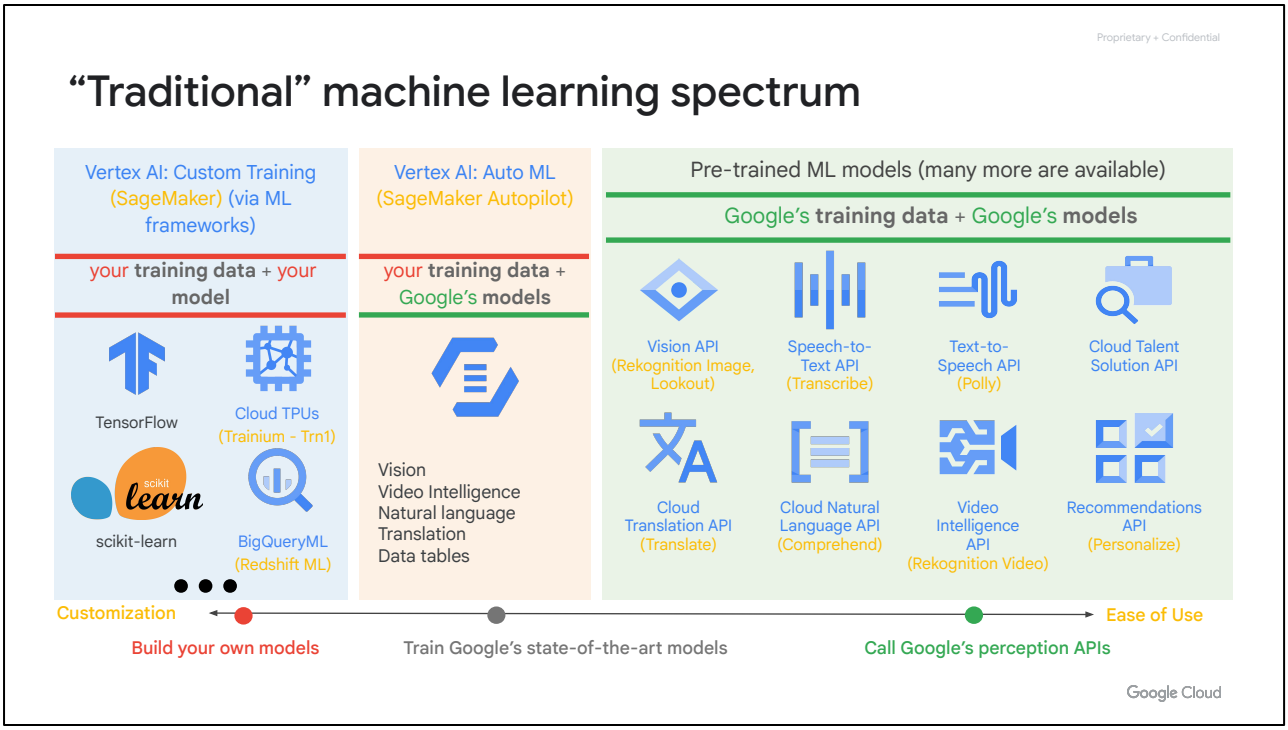
- AWS uses Amazon S3, and
- Google Cloud relies on Cloud Storage

And for data processing:

- AWS uses Amazon EMR, Amazon Redshift or Amazon Athena, while
- Google Cloud relies on Dataproc, BigQuery and Dataflow

Refer to the additional resource “Comparison of data ingestion, storage and data processing in AWS and Google Cloud” to get more insight into the various options.

Machine Learning

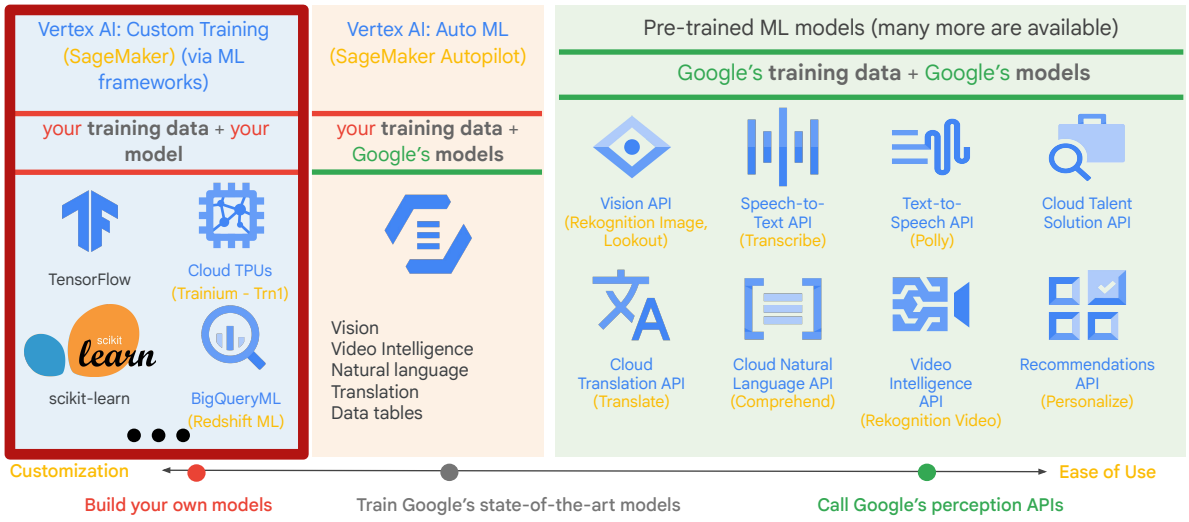


Source: Demo Template
Different options exist when it comes to leveraging machine learning. Advanced users, who want more control over the building and training of ML models, will use tools that offer the levels of flexibility they are looking for. This would involve developing custom models through an ML library like TensorFlow, that's supported on Cloud ML Engine, which is now a part of AI Platform. This option works for data scientists with the skills and the need to create a TensorFlow model.

But increasingly, you don't have to do that. Google makes the power of ML available to you even if you have a limited knowledge of ML. You can use AutoML to build on Google's ML capabilities to create your own custom ML models that are tailored to specific business needs, and then integrate those models into applications and web sites.

Alternatively, Google has a range of pre-trained ML models that are ready for immediate use within applications in ways that the respective APIs are designed to support. Such pretrained models are excellent ways to replace user input with ML.

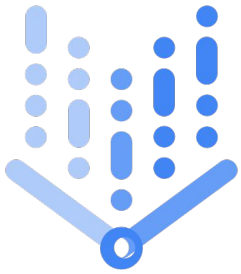
For those who need maximum control and customization



Proprietary + Confidential

Vertex AI (SageMaker):a service to get your data science projects up and running in minutes

- Managed JupyterLab experience.
- Secure development and controlled user access.
- Advanced networking.
- Support for data science frameworks and optimized for machine learning.
- Git support.
- Bring your own container / model.



Google Cloud

Source: Demo Template
AI Platform Notebooks is a managed service that offers an integrated and secure JupyterLab environment for data scientists and machine learning developers to experiment, develop, and deploy models into production. Users can create instances running JupyterLab that come pre-installed with the latest data science and machine learning frameworks in a single click.

AI Platform Notebooks features:

Managed JupyterLab experience:
AI Platform Notebooks is built on the industry standard JupyterLab. So you can use it with the RPython and R data science community and customize your environment by installing JupyterLab plugins.

Secure development:
AI Platform Notebooks supports popular enterprise security architectures through VPC-SC, shared VPC, and private IP controls. You can also encrypt your data on disk with CMEK.

Controlled user access:
You can choose between two predefined user access modes: restrict AI Platform Notebooks to a single-user or use a service account. You can also customize access based on your enterprise security architecture based on Cloud Identity and Access Management.

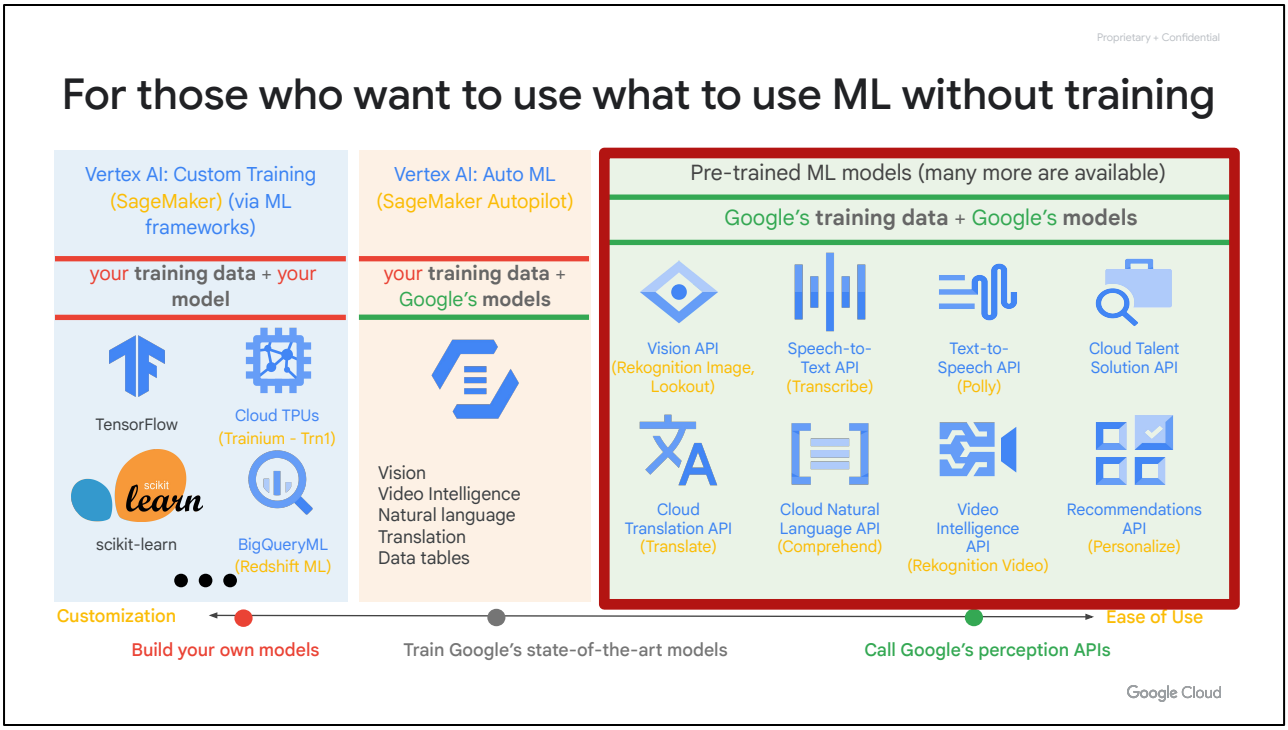
Advanced networking:
You can select any virtual private cloud for their AI Platform Notebook instances, provided that they have access either through Google Private Access or the internet to Cloud Storage. You can also turn off public IP address and access your instance via proxy.

Support for data science frameworks:
Google provides a pre-configured environment that supports the most popular data science libraries, including R, pandas, NumPy, SciPy, scikit-learn, and Matplotlib, and ML frameworks like TensorFlow, Keras, fast.ai, RAPIDS, XGBoost, and PyTorch.

Optimized for machine learning:
AI Platform Notebooks' optimized versions of TensorFlow and PyTorch enable you to get the most out of Google Cloud hardware and seamlessly add and remove GPUs from your instance.

Git support:
It’s easy to pull and push notebooks from your Git repository, making it also easy to share your notebooks with colleagues.

Bring your own container:
You can run a AI Platform Notebook instance on a container of your choice. This provides you the flexibility to install specific libraries mandated by your organization or preconfigure the environment running JupyterLab to your preference.



Source: Demo Template
Different options exist when it comes to leveraging machine learning. Advanced users, who want more control over the building and training of ML models, will use tools that offer the levels of flexibility they are looking for. This would involve developing custom models through an ML library like TensorFlow, that's supported on Cloud ML Engine, which is now a part of AI Platform. This option works for data scientists with the skills and the need to create a TensorFlow model.

But increasingly, you don't have to do that. Google makes the power of ML available to you even if you have a limited knowledge of ML. You can use AutoML to build on Google's ML capabilities to create your own custom ML models that are tailored to specific business needs, and then integrate those models into applications and web sites.

Alternatively, Google has a range of pre-trained ML models that are ready for immediate use within applications in ways that the respective APIs are designed to support. Such pretrained models are excellent ways to replace user input with ML.

Use the Vision API (Rekognition Image) to understand image content

Detect and label | Extract text | Identify entities

Source: Demo Template
Let's start with the Vision API. There are three major components that all roll up into this REST API, and behind-the-scenes each of these are powered by many ML models and years of research.

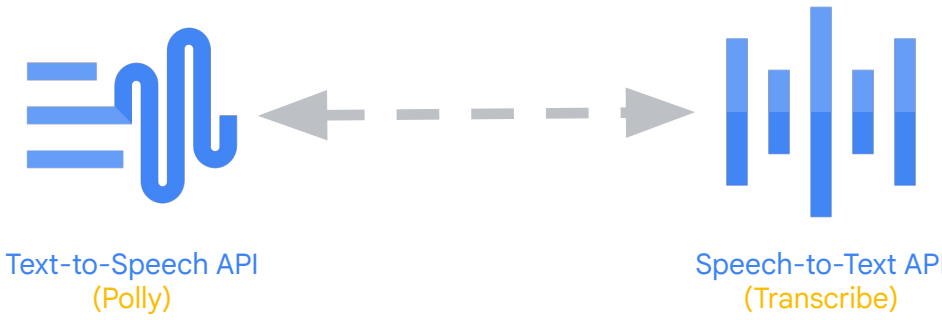
The first is detecting what an image is and classifying it. The Vision API picks out the dominant entity, for example a car or a cat, within an image from a broad set of object categories. This allows you to easily detect broad sets of objects in your images. Facial detection can detect when a face appears in photos, along with associated facial features such as eye, nose and mouth placement, and likelihood of over 8 attributes like joy and sorrow. Facial recognition however, isn't supported and Google doesn't store facial detection information on any Google server. You can use the API to easily build metadata on your image catalog, enabling new scenarios like image based searches or recommendations.

Next, are images with text, like scanned documents or signs. The Vision API uses optical character recognition, or OCR, to extract the text of a wide range of languages into a selectable, searchable format.

Lastly is a bit of intuition from the web and uses the power of Google Image Search. Does the image contain entities we know, like the Eiffel tower or a famous person? Landmark detection allows you to identify popular natural and manmade structures, along with the associated latitude and longitude of the landmark, and logo detection allows you to identify product logos within an image.

You can build metadata on your image catalog, extract text, moderate offensive content, or enable new marketing scenarios through image sentiment analysis. You can also analyze images uploaded in the request or integrate with an image storage on Cloud Storage.

Convert speech to text and vice versa



Text-to-Speech API
(Polly)

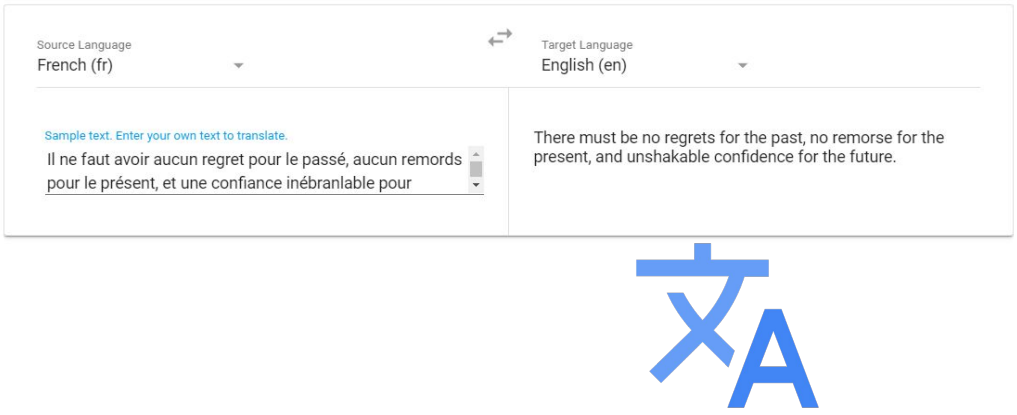
Speech-to-Text API
(Transcribe)

Source: Demo Template
There are two APIs that apply to speech.

The Text-to-Speech API converts text into human-like speech in more than 180 voices across more than 30 languages and variants. It applies research in speech synthesis and Google's powerful neural networks to deliver high-fidelity audio. With this API, you can create lifelike interactions with users that transform customer service, device interaction, and other applications.

The Speech-to-Text API enables you to convert real-time streaming or prerecorded audio to text. The API recognizes 120 languages and variants to support a global user base. You can enable voice command-and-control, transcribe audio from call centers, and so on.

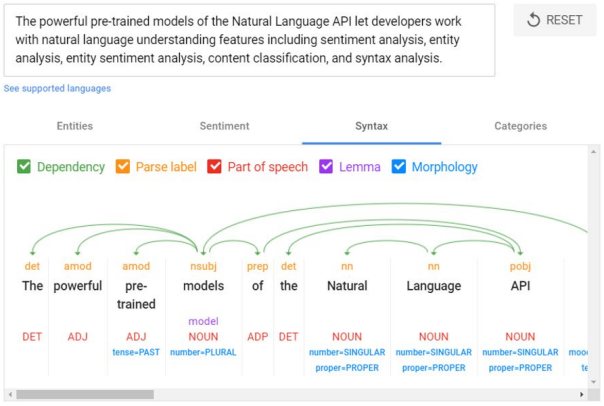
Dynamically translate between languages using the Cloud Translation API (Translate)



Source: Demo Template
The Cloud Translation API provides a simple programmatic interface for translating an arbitrary string into any supported language. The Cloud Translation API is highly responsive, so websites and applications can integrate with the API for fast, dynamic translation of source text from the source language to a target language, for example from French to English. Language detection is also available in cases where the source language is unknown.

Let's look at a short video that shows how Bloomberg, a global leader in business and financial data, news and insight, applied the Cloud Translation API to reach all of their customers regardless of language.

Derive insights from unstructured text with the Cloud Natural Language API (Comprehend)

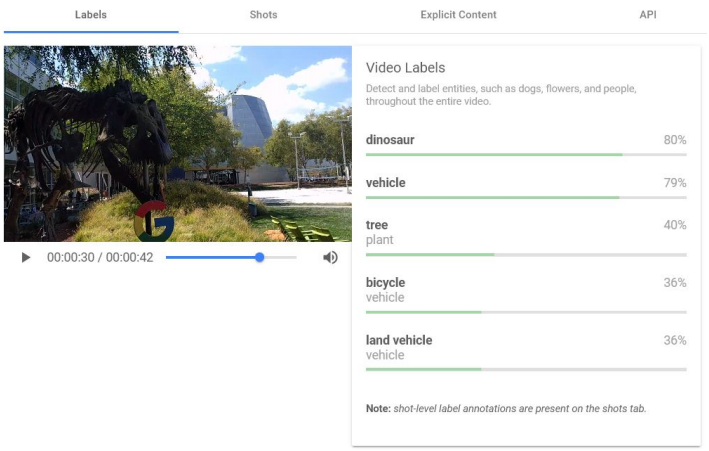


Source: Demo Template
The Cloud Natural Language API offers a variety of natural language understanding technologies. It can do syntax analysis, breaking down sentences into tokens, identify the nouns, verbs, adjectives, and other parts of speech, and figuring out the relationships among the words.

It can also do entity recognition, in other words, it can parse text and flag mentions of people, organizations, locations, events, products and media.

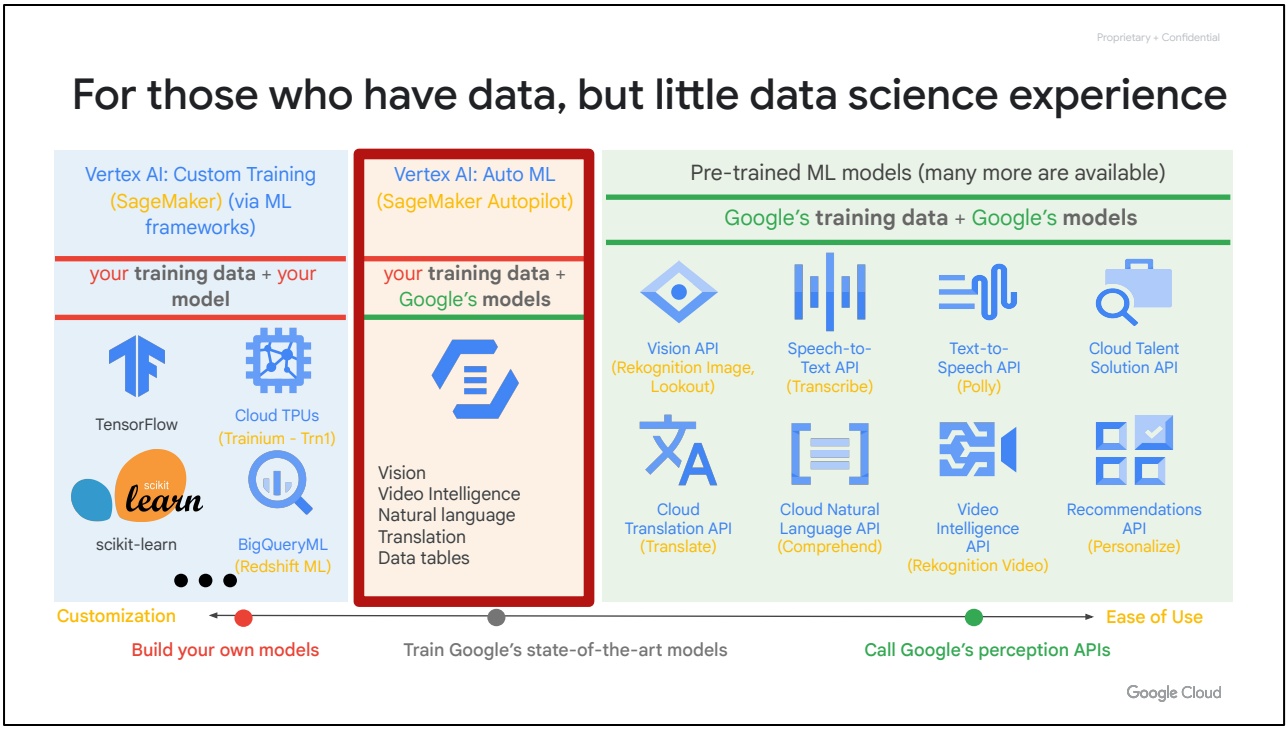
Sentiment analysis allows you to understand customer opinions to find actionable product and UX insights.

Make your media more discoverable with the Video Intelligence API (Rekognition Video)



Source: Demo Template
The Video Intelligence API allows users to use Google video analysis technology as part of their applications. The REST API enables users to annotate videos stored in Cloud Storage with video and 1 frame-per-second contextual information. It helps you identify key entities -- that is, nouns -- within your video, and when they occur. You can use it to make video content searchable and discoverable.

The API supports the annotation of common video formats, including .MOV, .MPEG4, .MP4, and .AVI.

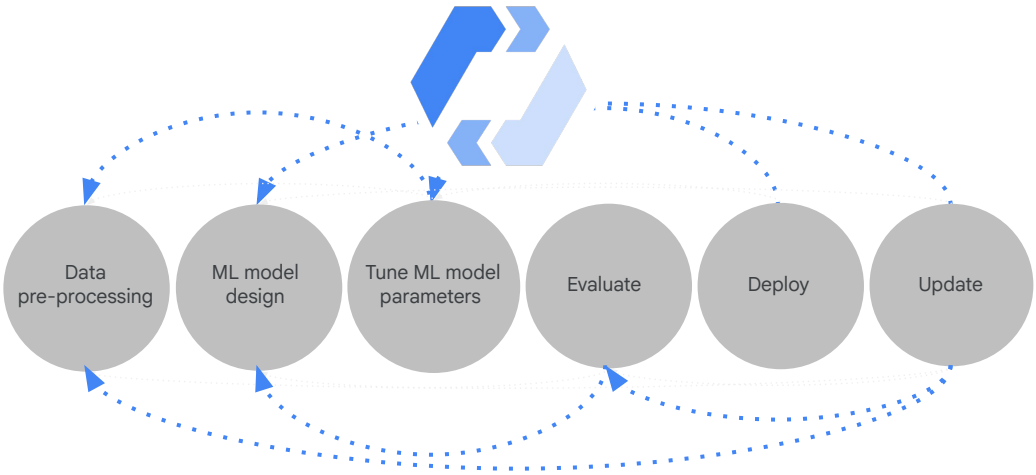


Source: Demo Template
Different options exist when it comes to leveraging machine learning. Advanced users, who want more control over the building and training of ML models, will use tools that offer the levels of flexibility they are looking for. This would involve developing custom models through an ML library like TensorFlow, that's supported on Cloud ML Engine, which is now a part of AI Platform. This option works for data scientists with the skills and the need to create a TensorFlow model.

But increasingly, you don't have to do that. Google makes the power of ML available to you even if you have a limited knowledge of ML. You can use AutoML to build on Google's ML capabilities to create your own custom ML models that are tailored to specific business needs, and then integrate those models into applications and web sites.

Alternatively, Google has a range of pre-trained ML models that are ready for immediate use within applications in ways that the respective APIs are designed to support. Such pretrained models are excellent ways to replace user input with ML.

AutoML simplifies the process



Source: Demo Template

The ability of AutoML to efficiently solve an ML problem is largely due to how it simplifies these complex steps that are associated with custom ML model building.

Use AutoML for what you can see

AutoML Vision

Derive insights from images in the cloud or at the edge.

AutoML Video Intelligence

Enable powerful content discovery and engaging video experiences.

Source: Demo Template
There are two AutoML products that apply to what you can see.

With AutoML Vision, you simply upload images and train custom image models through an easy-to-use graphical interface You can optimize your model for accuracy, latency, and size. AutoML Vision Edge allows you to export your custom trained models to an application in the cloud, or to an array of devices at the edge. You can train models to classify images through labels you choose. Alternatively Google's data labeling service allows you to use their team to help annotate your images, videos, and text. Later, we'll complete a lab where we'll use AutoML Vision to train a custom model to recognize different types of clouds.

AutoML Video Intelligence makes it easy to train custom models to classify and track objects within videos. It's ideal for projects that require custom entity labels to categorize content which aren't covered by the pre-trained Video Intelligence API.

Use AutoML for what you can hear

AutoML Natural Language

Reveal the structure and meaning of text through machine learning.

AutoML Translation

Dynamically translate between languages.

Source: Demo Template
There are also two AutoML products that apply to what you can hear.

With AutoML Natural Language, you can train custom ML models to classify, extract, and detect sentiment. This allows you to identify entities within documents and label them based on your own domain-specific keywords or phrases. The same applies to being able to understand the overall opinion, feeling, or attitude expressed in a block of text that's tuned to domain-specific sentiment scores.

AutoML Translation allows you to upload translated language pairs and it will train a custom model which translation queries return results specific to your domain, and that you can scale and adapt to meet your needs.

Use AutoML to turn structured data into predictive insights

AutoML Tables

Automatically build and deploy state-of-the-art machine learning models on structured data.

Source: Demo Template
AutoML Tables reduces the time it takes to go from raw data to top-quality, production-ready machine learning models from months to just a few days.

There are many different use cases for AutoML Tables. For example, if you're in retail, you can better predict customer demand so you can preemptively fill gaps and maximize your revenue by optimizing product distribution, promotions, and pricing. In insurance, you could foresee and optimize a policyholder portfolio's risk and return by zeroing in on the potential for large claims and likelihood of fraud. In marketing, you can better understand your customer. For example, What's your average customer's lifetime value? You can make the most of marketing spend by using AutoML Tables to estimate predicted purchasing value, volume, frequency, lead conversion probability, and churn likelihood.

BigQuery ML (Redshift ML) makes AI easy

Train and deploy ML models in SQL

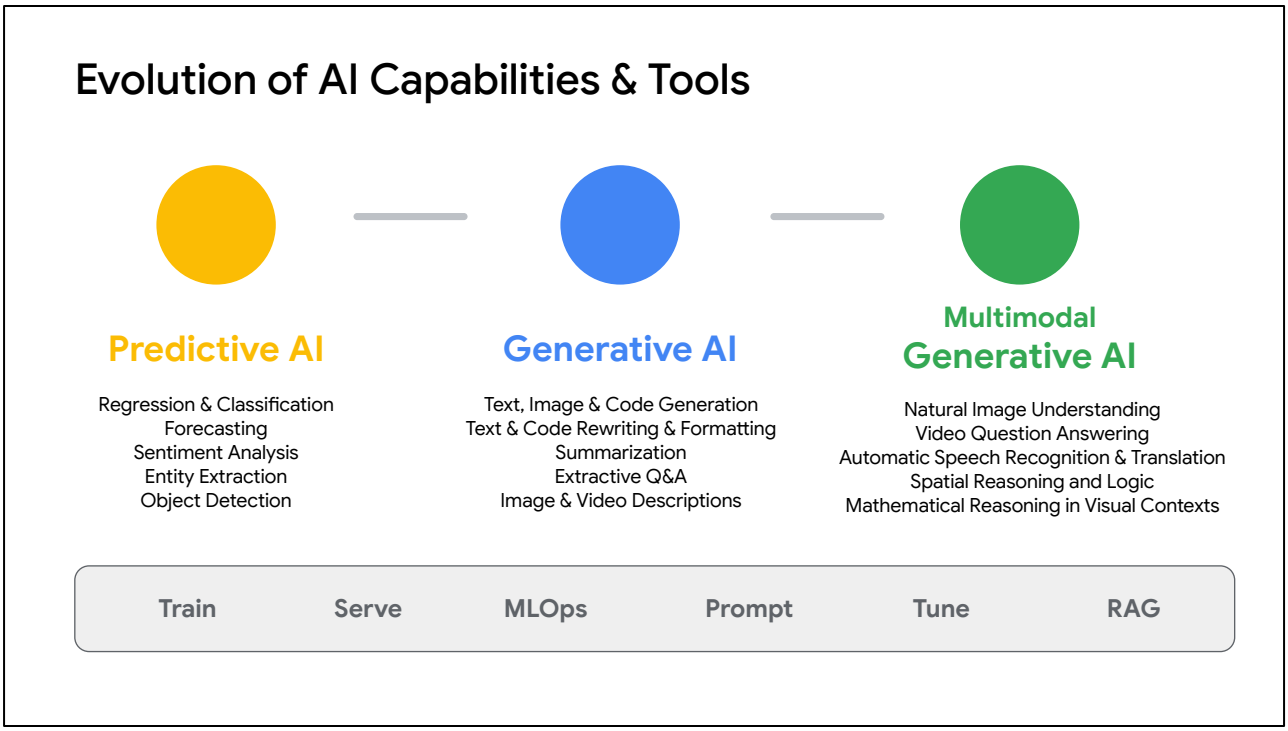
Execute ML workflows without moving data from BigQuery

Automate common ML tasks

Built-in infrastructure management, security & compliance

Without needing to move your data out of BigQuery, with BigQuery ML, you can train and deploy machine learning models directly using SQL. That means you've got data storage, data analytics, **and** machine learning all within BigQuery.

Gen AI - Part 1



[Historically, AI excelled at tasks like identifying patterns, finding anomalies, and making predictions on specific topics. For example, it could act as a specialized expert to detect credit card fraud, but it would not be able to perform other tasks.
However, we're now witnessing a profound shift in AI's capabilities. The next generation of AI, powered by generative models, is like highly advanced learner who have absorbed an entire library and can perform many tasks (writer, coder, summarizer, translator and so on). And most importantly - it can now create entirely new content.

Now, it's also about how we interact with the model and what type of output the model can generate. While initial Gen AI models were mostly focusing on chatbot-like functionality, it's now more and more common to have those "Multimodal" models, where you can both write text to a model, upload an image, give it a link to a youtube video and so on. And the model itself can also answer in different forms, instead of just a text. For example, it can give you a new picture based on your what you explained using your voice.]

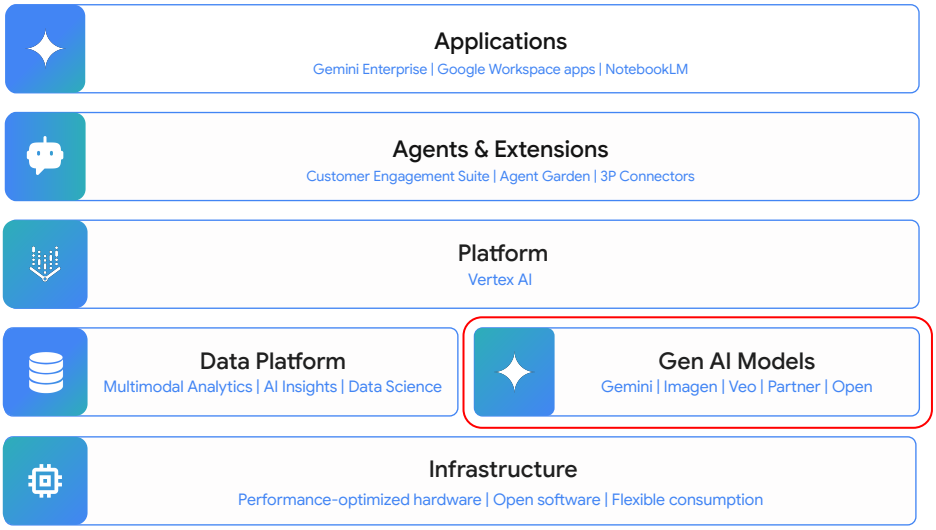
– introduce the "multimodality"

---

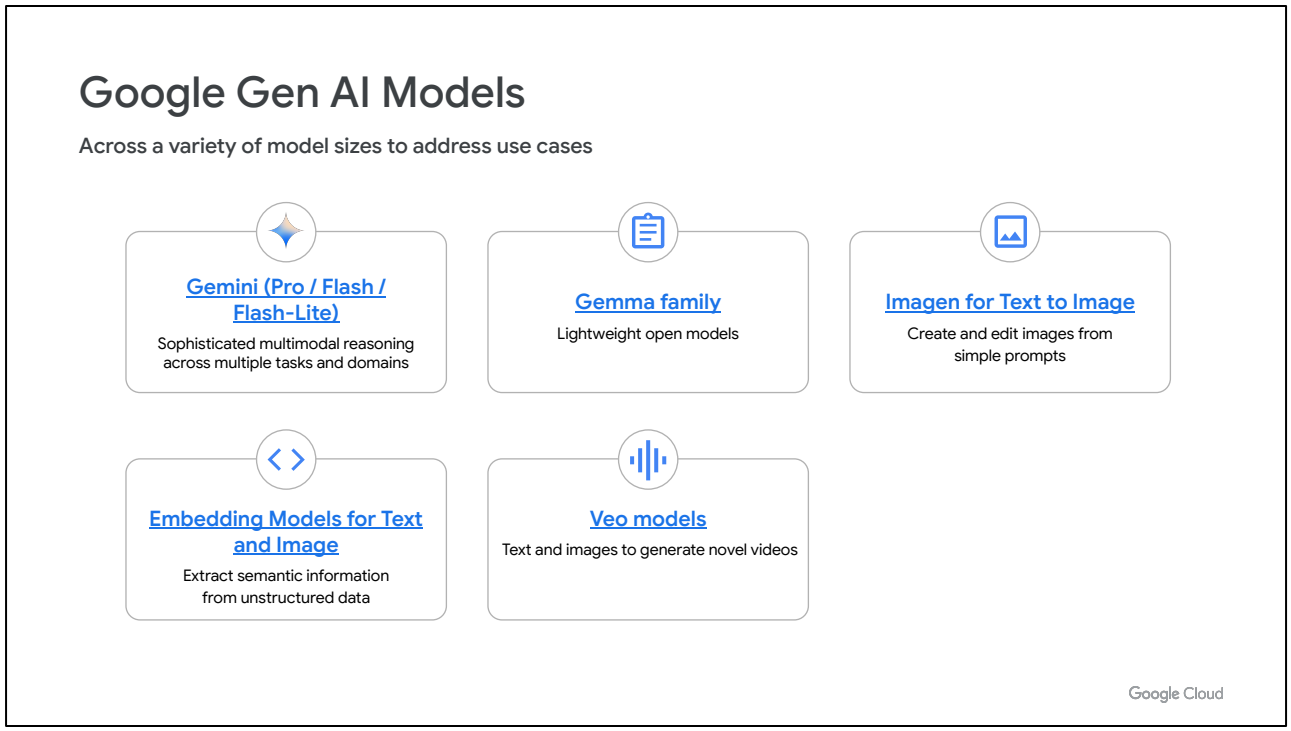
predictive = predicting a very well defined output usually a number or label (cat/dog, $100 forecast, pixels on an image that surround an object, fraud/not fraud)

generative = generating a chunk of output that isn't easy to measure with predictive ai metrics like accuracy/prediction/recall (json structure, code, image, copy, summary, etc.)

Google's Unified AI technology stack



Google defined so-called "AI technology stack" with layers, each responsible for a different aspect. *The first layer is infrastructure and I don't think we need to go deeper into this one. The bottom line i: in order to make Gen AI work on massive scale, a lot of high-performance computing, storage and bandwidth is needed, so each hyperscaler invests heavily in their data centers, GPUs, TPUs and so on.*
Data Platform layer is mostly about Big Data, differentiation between structured and unstructured data, labeled and unlabeled data, and overall data quality. Let's also assume that you know enough about those, and going much deeper would be rather a task for Data Engineers.
The first layer I'd like to explore a bit more is about Models.



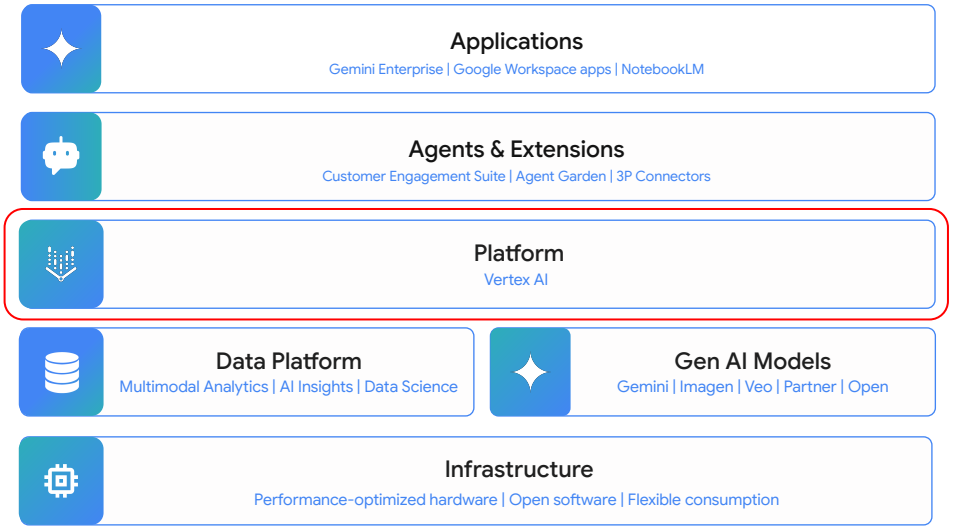
Here, you'd probably want to know a bit about Google's model portfolio.
https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models

Examples of Google's models:
• **Gemini**: (AS A MODEL, NOT AN APP YET! -> app, including Gems etc will be covered next week) A **multimodal** model that can understand and operate across diverse data formats like text, images, audio, and video. It supports advanced conversational AI, content creation, and nuanced question answering.
• **Gemma**: A family of lightweight, open models built upon the research and technology behind Gemini, offering user-friendly and customizable solutions for local deployments and specialized AI applications.
• **Imagen**: A powerful text-to-image diffusion model that excels at generating high-quality images from textual descriptions.
** NOW: Nano banana? But NOT on the exam!
• **Embedding Models:**: Converts text data into vector representations for semantic search, classification, clustering, and similar tasks.
• **Veo**: A model capable of generating video content based on text descriptions or still images.

All of those are **foundation** models, but some of those are **LLMs** (like Gemini or Gemma), and others (Veo or Image) are **diffusion**.

– emphasize it’s important to differentiate between Gemini Models and Gemini App!
– you can host Partner / open-source models in GCP as well, but you won’t be asked for details of those on the exam.

Google's Unified AI technology stack



Let's now shift to the Vertex AI Platform, which allows you to build and scale models;

What is Vertex AI?

Vertex AI

cloud.google.com/vertex-ai

- Managed, End-to-End **AI & ML Platform on Google Cloud**
- **Model Garden** - Generative & Predictive AI Models from Google, Partners and Open Source
- **Vertex AI Studio** - Experiment with Models
- **Custom Models** - Training/Prediction Pipelines
- **Vector Search** for Embeddings
- **Colab Enterprise** for Jupyter Notebooks
- **Enterprise-Grade** Security/Reliability

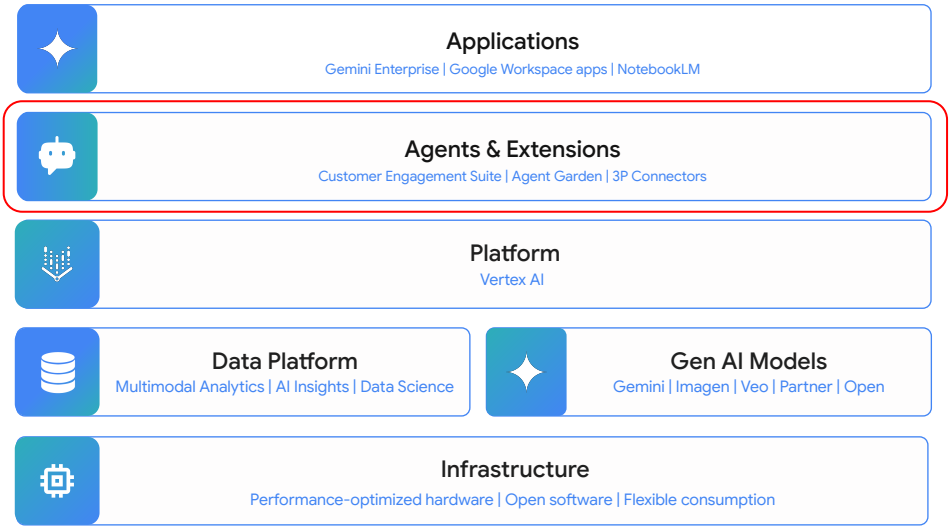
Google Cloud

Vertex AI is Google Cloud's managed, end-to-end AI platform. It's not just one tool, but a whole suite of tools gathered under one umbrella.

****** DEMO VERTEX AI -> open GCP console and go through most important functionalities

It includes functionalities the Model Garden, where you can access different models; It has Vertex AI Studio for experimenting and prompt design, tools for custom model training, Colab Enterprise for notebooks and so on. All of this is built with enterprise-grade security, scalability, and compliance.

Google's Unified AI technology stack



Last layer I'd like to cover today is "Agents & Extensions". This one is becoming more and more important as so-called "Agentic AI" seems to be gaining momentum as we speak.

Observe, Act, Achieve

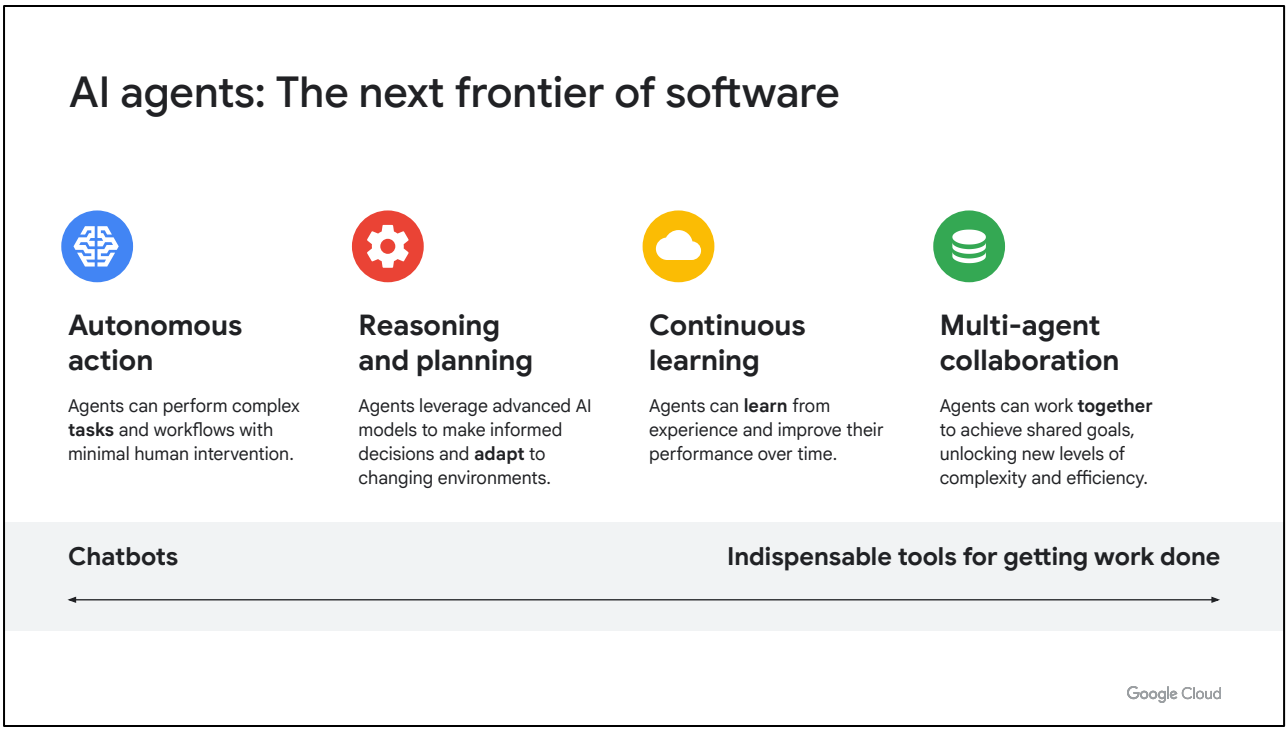
An AI Agent is an **application** that tries to achieve a **goal** by **observing** the world and **acting** upon it using the **tools** it has at its disposal.

Google Cloud

First, let's define an "agent". The simplest definition is that's an application that tries to achieve a goal.
It does this by observing the world (or a data source), reasoning about what to do, and then acting upon it using the **tools** it has. The key concept here is that it moves beyond just answering a question to actively *doing* something.

Agents are applications that use generative AI models to ***think*** **and** ***act*** **towards** ***goals***

Key message: Agents reason on how to best achieve a goal based on inputs and **tools** at its disposal, leveraging models, tools, and orchestration.



Agents differ from simple chatbots because they have four key properties:

1) **Autonomous action:** They can perform complex workflows with minimal human intervention.

2) **Reasoning and planning:** They use advanced models to make decisions.

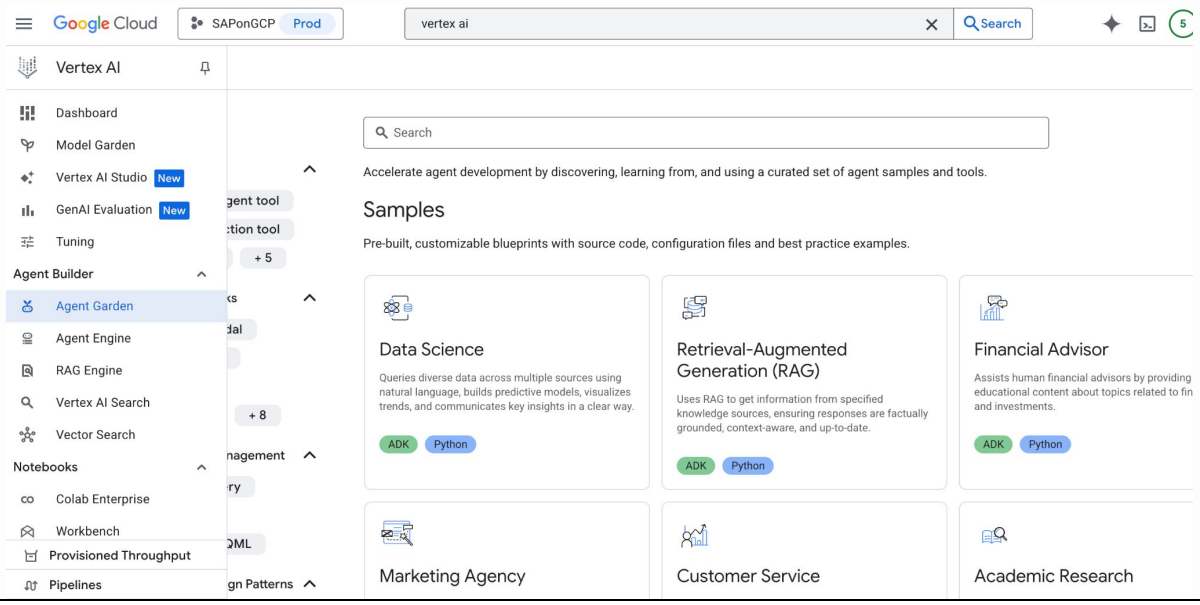
3) **Continuous learning:** They can improve their performance over time.

4) **Multi-agent collaboration:** They can even work together to achieve shared goals.

Ok, but where do you take the agents from? There are a couple of options actually:

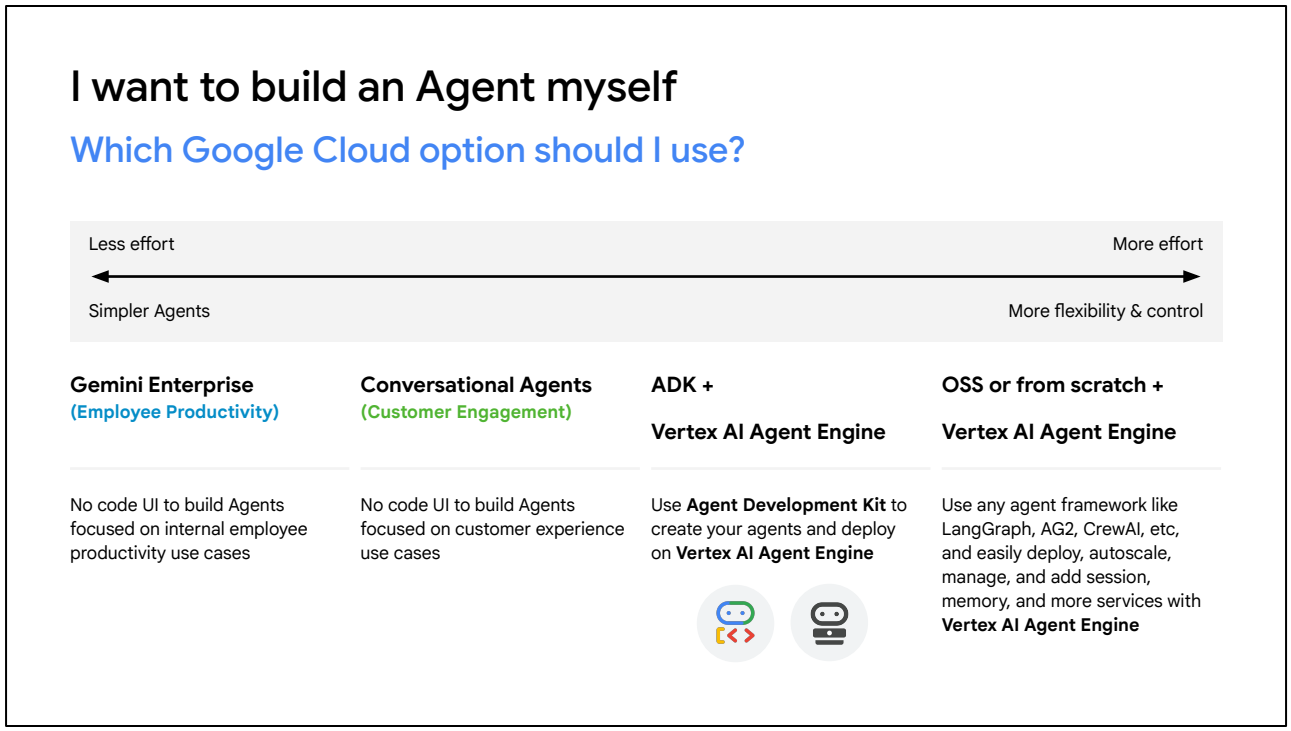
- You can build your own agents, using “no-code” or a bit more advanced approach
- Or you can use agents that Google or 3rd parties created.

Use Google Cloud's ready-packaged agents



Show Vertex AI -> Agent Garden

- Compare to "model garden"



If you'd like to create your own agent, there are a couple of paths. The ones you see on the right are a bit more complex and require some coding, but it's also possible to use UI-based methods:

– show https://cloud.google.com/gemini-enterprise/agents?hl=en -> "Make your own agents" and zoom into "News Summarizer" agent

1. (pre-preview) Agent Designer for no-code custom agents for Gemini Enterprise: https://cloud.google.com/gemini/enterprise/docs/agent-designer
2. Build custom agents using the Agent Development Kit (ADK) in Vertex AI, then deploy and govern them in Gemini Enterprise: https://github.com/google/adk-samples/tree/main/python/agents
3. Upload, access, deploy, and govern agents you've built on external platforms.

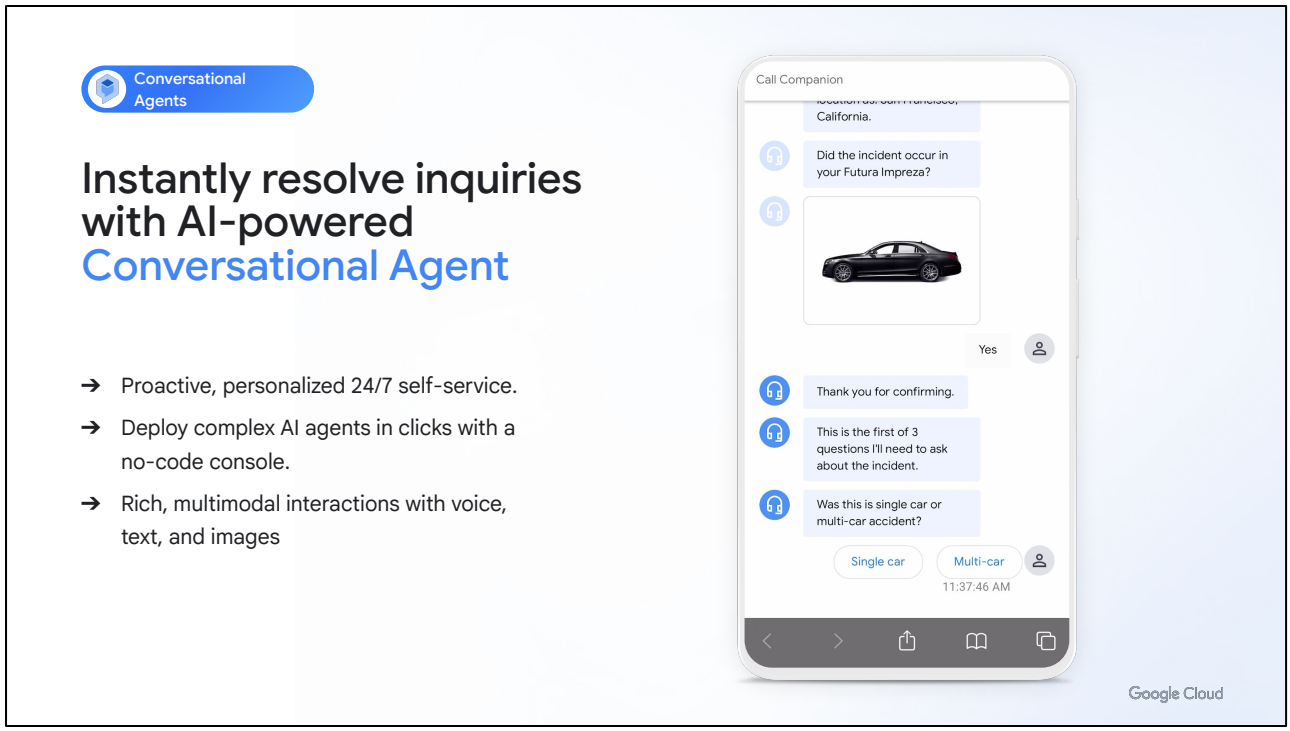
Building AI agents can be tailored to different needs:

- **For Business Users (Simpler Agents, Less Effort):**
  - **Agentspace:** A no-code interface perfect for building agents focused on internal employee productivity.
  - **Conversational Agents:** Another no-code UI, but geared towards enhancing customer engagement.
- **For Developers (More Flexibility & Control, More Effort):**
  - **ADK + Vertex AI Agent Engine:** Use the Agent Development Kit to craft highly customized agents and deploy them on Vertex AI Agent Engine.
  - **OSS or from scratch + Vertex AI Agent Engine:** For ultimate control, you can build from open-source tools or from scratch, then leverage Vertex AI Agent Engine for deployment and management.

Customer Engagement Suite with Google AI

Google Cloud

Now, let's explore the Customer Engagement Suite, which is a kind of an umbrella on top of some agent-based functionalities

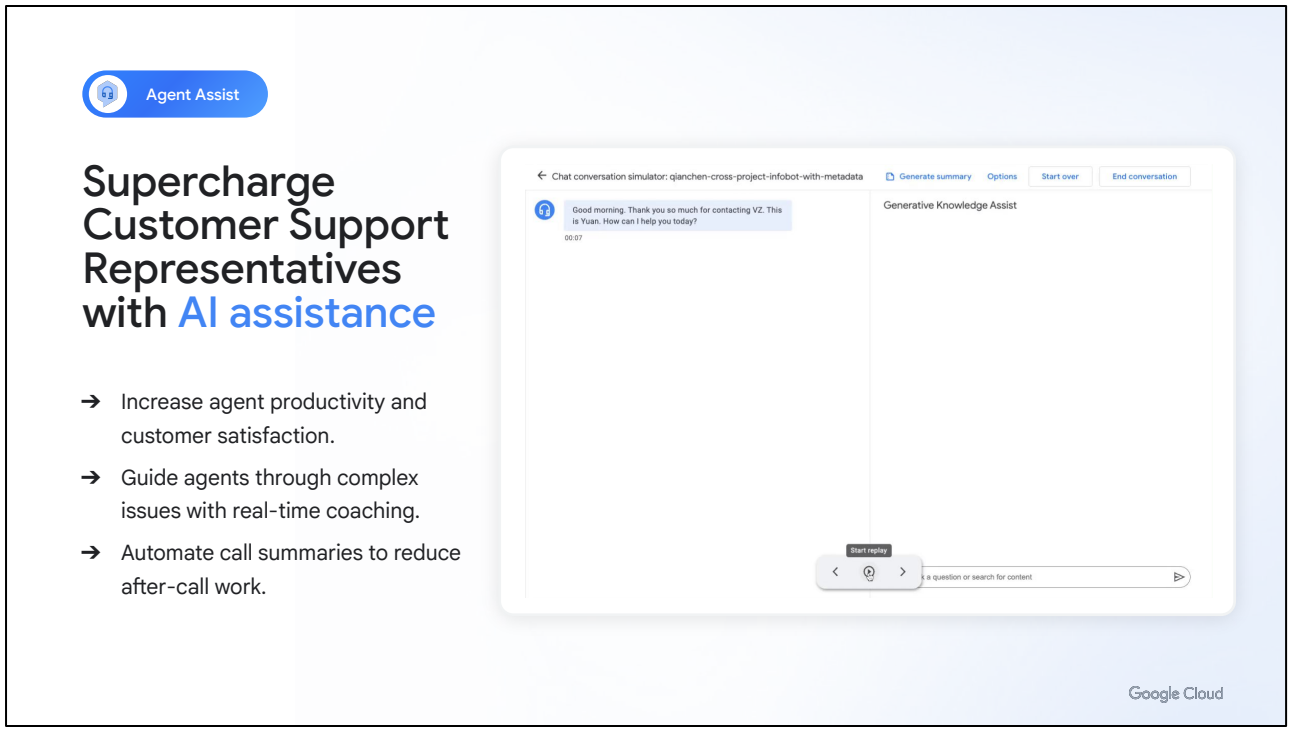


First, Conversational Agents.

They allows you to use out-of-the box agents or build them yourself with a simple, no-code interface. And then use those for customer service needs. Since Gemini is models are being used here, those agents can use human-like voices, they try to understand emotions, and they even support streaming video, so that the agent can interpret and respond to what they see in real-time when shared from customer devices. What's most important, they can be well grounded in your company resources so that they operate only within a scope you let them.

– go through this example: https://cloud.google.com/dialogflow/cx/docs/concept/playbook/prebuilt/department-of-motor-vehicles

https://cloud.google.com/blog/products/ai-machine-learning/next-generation-customer-engagement-suite-ai-agents



Of course, not every issue can be resolved with automation. For your human experts, Agent Assist might be considered. It works by providing real-time, step-by-step guidance, suggesting the next best action or response. It can also provide live translation, and after the call, it uses Gemini to automatically summarize the entire conversation. It can also provide training recommendations to customer care representatives for improving their proficiency in the long run.

Conversational Insights and Quality AI

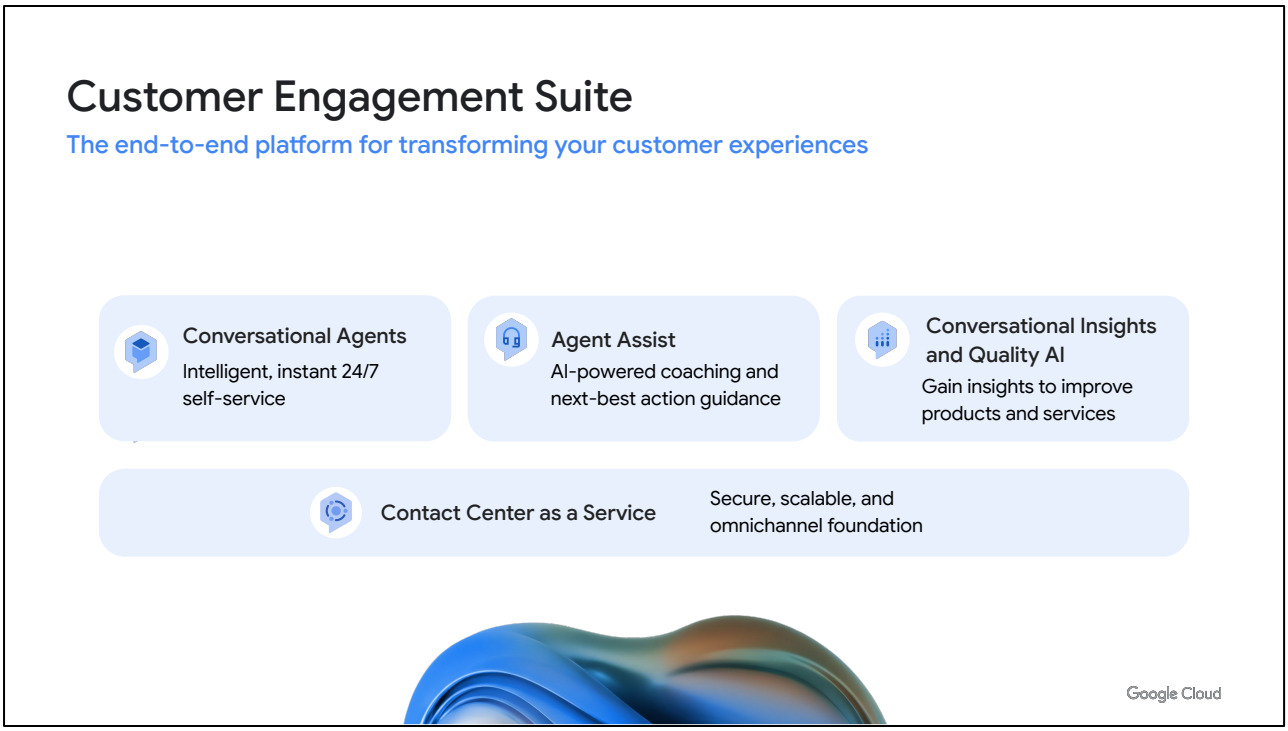
Turn customer conversations into business-wide intelligence

- Automatically review every conversation for quality.
- Identify top customer issues and agent coaching needs.
- Use customer feedback to improve products and services.

Google Cloud

This brings us to the most strategic component of the suite: Conversational Insights. This is how you transform your contact center from a reactive cost center into a proactive business intelligence engine.

It allows managers detect and visualize patterns in the contact center data, based on as much data as you'd like to import from interactions with your customers. It helps to understand customer sentiment at scale, and pinpoint exactly where you can improve your operations and even your products or services.



So how do all of those play together under the Customer Engagement Suite umbrella?

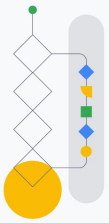
– PLAY https://www.youtube.com/watch?v=Z0GwPJncNqg from 0:40 till the end (5 mins)

https://cloud.google.com/contact-center/ccai-platform/docs

The Generative AI Landscape:
Workflow Agents

Workflow agents are designed to streamline your work and make sure things get done efficiently and correctly by automating tasks or going through complex processes.

- **You provide input:** You define a task or trigger a process like submitting a form, uploading a file, initiating a scheduled event, or even ordering a product online.
- **The agent understands:** The agent is the software that automates those steps. It interprets the task's requirements and defines the series of steps needed to complete the task.
- **The agent calls a tool:** Based on the workflow's definition, the agent executes a series of actions. This could involve data transformation, file transfer, sending notifications, integrating with external systems, or initiating other automated processes using APIs.
- **The agent generates a result/output:** It compiles the outcome of the executed actions, which might be a report, a data file, a confirmation message, or an updated status within a system.
- **The agent delivers the result/output:** The agent delivers the output to the designated recipient(s) or system(s), such as via email, a dashboard, a database update, or a file storage location.



- **Ecommerce order fulfillment:** An agent automatically processes orders, updates inventory, sends shipping notifications, and handles returns.
- **Customer onboarding:** An agent guides new customers through account setup, provides tutorials, and answers frequently asked questions.
- **Automated research:** An agent can conduct in-depth research on a given topic by autonomously browsing the web, summarizing relevant content, and generating comprehensive reports. (Try this out with Gemini Deep Research.)
- **Security Log Parsing:** An agent that inspects incoming security logs for abnormalities and can flag them, open a ticket, begin triage, and assign to humans for review when necessary.

Google Cloud

Once we're done with Conversational Agents, let's also make sure we can differentiate those from so-called "workflow agents".

Those are built to streamline complex processes by automating multi-step tasks. Unlike conversational agents, they are triggered by a defined task or process. The agent interprets the task, executes a series of actions (potentially integrating with external systems or APIs), and generates a structured output or result.

– go through this example on high level:
https://cloud.google.com/blog/products/ai-machine-learning/build-kyc-agentic-workflows-with-googles-adk

Altostrat Media

case study

https://services.google.com/fh/files/misc/v6.1_pca_altostrat_media_case_study_english.pdf

Proposed Technical Solution

- Altostrat currently utilizes GKE for content management and delivery, and Cloud Run for serverless tasks like video transcoding and metadata extraction
    - Implement Istio / Anthos / ASM / Cloud Service Mesh / Fleets to provide a centralized management platform for both Google Cloud and the on-premises legacy systems
    - See more about container orchestration
    - Continue using Cloud Run functions for event-driven tasks that require serverless execution, such as video transcoding and personalized content recommendations, which scales with minimal latency
- Use Cloud Storage Lifecycle Management to optimize costs for growing media library (documents, audio, video).
    - Use Storage Transfer Service to migrate large-scale on-premises archival data (over 1 TB) to Cloud Storage
- Real-time Processing and Data Flow:
    - Integrate Pub/Sub with Dataflow to create streaming pipelines for real-time parallel data processing, preparing data for analysis in BigQuery
- Continue leveraging BigQuery in combination with BI tools (like Looker or Tableau) for interactive data exploration and decision-making on content strategy
- AI-related:
    - **Content Enrichment and Metadata Extraction:** Use pre-trained AI services such as Video Intelligence API and Natural Language API to automatically extract rich metadata from media assets, enabling content discovery, dynamic pricing, and targeted marketing
    - **Harmful Content Detection:** use features such as Model Armor, use Vertex AI content filters and / or develop custom AI-powered detection.
    - **Generative AI for User Experience and Content Virality:** Build AI chatbots leveraging LLMs and Conversational AI (e.g., Dialogflow or specific Vertex AI features). Also, implement Generative AI to automatically generate concise summaries of diverse media content
    - **Model Management and Auditing:** Utilize Vertex AI functionalities such as: Model evaluation, Explainable AI, Vertex AI Workbench or Colab Notebooks, Model Monitoring, Vertex AI Model Registry etc.
- Plus others: Cloud Build for CI/CD modernization, Hybrid Connectivity options, use Google Cloud Observability platform etc

[Altostrat Media case study] Diagnostic Question #1

Altostrat Media stores a vast and growing library of video content in Cloud Storage. The majority of their archived documentaries (which comprise 60% of their total volume) are accessed less than four times per year, and many are retained for long-term regulatory purposes. Altostrat needs to immediately optimize cloud storage costs for these growing media volumes while ensuring long-term retention requirements are met and maintaining high availability for serving the content globally

Which storage solution best balances cost optimization, global availability, and long-term regulatory compliance?

A. Use a Multi-Regional Cloud Storage bucket and apply an Object Lifecycle Management policy to transition objects accessed less than once a year to Archive Storage.

B. Use a Regional Coldline Storage bucket, and rely on Object Versioning for long-term retention assurance.

C. Store all content in a Multi-Regional Standard Storage bucket to ensure high availability, and manually move documentary assets to Archive Storage only after 12 months using a scheduled data job.

D. Use a Dual-Region Nearline Storage bucket, ensuring objects are moved to Coldline Storage after 90 days of inactivity using Object Lifecycle Management

Google Cloud

A

Multi-regional for HA and serving content globally, plus object lifecycle management to optimize costs.

[Altostrat Media case study] Diagnostic Question #1

Altostrat Media stores a vast and growing library of video content in Cloud Storage. The majority of their archived documentaries (which comprise 60% of their total volume) are accessed less than four times per year, and many are retained for long-term regulatory purposes. Altostrat needs to immediately optimize cloud storage costs for these growing media volumes while ensuring long-term retention requirements are met and maintaining high availability for serving the content globally

Which storage solution best balances cost optimization, global availability, and long-term regulatory compliance?

A. Use a Multi-Regional Cloud Storage bucket and apply an Object Lifecycle Management policy to transition objects accessed less than once a year to Archive Storage.

B. Use a Regional Coldline Storage bucket, and rely on Object Versioning for long-term retention assurance.

C. Store all content in a Multi-Regional Standard Storage bucket to ensure high availability, and manually move documentary assets to Archive Storage only after 12 months using a scheduled data job.

D. Use a Dual-Region Nearline Storage bucket, ensuring objects are moved to Coldline Storage after 90 days of inactivity using Object Lifecycle Management

Google Cloud

A

Multi-regional for HA and serving content globally, plus object lifecycle management to optimize costs.

[Altostrat Media case study] Diagnostic Question #2

Altostrat Media aimed for enhanced reach and personalization. From an architectural perspective, how would you design a solution on GCP to dynamically segment audiences and deliver personalized ad content at scale, considering both batch and real-time data processing needs?

A. Utilize Dataflow for ETL from disparate sources into Cloud SQL, then integrate with Marketing Platform.

B. Ingest real-time events via Pub/Sub to Dataflow for stream processing and feature engineering, storing results in BigQuery for ML model training with Vertex AI, and serving predictions via custom APIs on Cloud Run.

C. Store all data in Cloud Storage buckets, use Dataproc for occasional batch processing, and manually update ad campaigns.

D. Implement App Engine to host custom audience segmentation logic and connect directly to ad networks.

Google Cloud

Answer: B.

Explanation:

- **Option B** describes a robust, scalable, and real-time architecture for personalization.
    - **Pub/Sub** is the ideal choice for ingesting high-volume, real-time events (clicks, impressions, user behavior).
    - **Dataflow** (managed Apache Beam) is perfect for complex stream processing, ETL, and feature engineering on that real-time data. It can also handle batch.
    - **BigQuery** serves as the central analytical data store for processed data and historical context.
    - **Vertex AI** provides a comprehensive MLOps platform for training sophisticated models (e.g., recommendation engines, segmentation models).
    - **Cloud Run** is excellent for serving low-latency, stateless prediction APIs from these ML models, scaling dynamically to handle ad-serving traffic.
- **Option A** is too limited; Cloud SQL won't handle the scale and analytical nature well.
- **Option C** is primarily batch-oriented and lacks real-time capabilities and

- robust ML integration for dynamic personalization.
- **Option D** uses App Engine, which could host logic, but it doesn't describe a full data-driven ML pipeline for scale and real-time processing.

[Altostrat Media case study] Diagnostic Question #2

Altostrat Media aimed for enhanced reach and personalization. From an architectural perspective, how would you design a solution on GCP to dynamically segment audiences and deliver personalized ad content at scale, considering both batch and real-time data processing needs?

A. Utilize Dataflow for ETL from disparate sources into Cloud SQL, then integrate with Marketing Platform.

B. Ingest real-time events via Pub/Sub to Dataflow for stream processing and feature engineering, storing results in BigQuery for ML model training with Vertex AI, and serving predictions via custom APIs on Cloud Run.

C. Store all data in Cloud Storage buckets, use Dataproc for occasional batch processing, and manually update ad campaigns.

D. Implement App Engine to host custom audience segmentation logic and connect directly to ad networks.

Google Cloud

Answer: B.

Explanation:

- **Option B** describes a robust, scalable, and real-time architecture for personalization.
    - **Pub/Sub** is the ideal choice for ingesting high-volume, real-time events (clicks, impressions, user behavior).
    - **Dataflow** (managed Apache Beam) is perfect for complex stream processing, ETL, and feature engineering on that real-time data. It can also handle batch.
    - **BigQuery** serves as the central analytical data store for processed data and historical context.
    - **Vertex AI** provides a comprehensive MLOps platform for training sophisticated models (e.g., recommendation engines, segmentation models).
    - **Cloud Run** is excellent for serving low-latency, stateless prediction APIs from these ML models, scaling dynamically to handle ad-serving traffic.
- **Option A** is too limited; Cloud SQL won't handle the scale and analytical nature well.
- **Option C** is primarily batch-oriented and lacks real-time capabilities and

- robust ML integration for dynamic personalization.
- **Option D** uses App Engine, which could host logic, but it doesn't describe a full data-driven ML pipeline for scale and real-time processing.

[Altostrat Media case study] Diagnostic Question #3

A key business requirement is to enable natural language interaction with the platform and provide 24/7 personalized user support via advanced chatbots. This chatbot needs to utilize Natural Language Understanding (NLU) to answer complex queries about content and suggest personalized recommendations.

Which Google Cloud service combination should be used to build and deploy this critical user engagement component?

A. Use a custom Python script deployed on a managed instance group (MIG) for NLU, and integrate it with BigQuery ML for recommendations.

B. Use a monolithic Node.js application deployed on GKE to manage all user interactions and recommendation logic

C. Utilize the Natural Language API directly within Cloud Functions to analyze user text input, avoiding any need for stateful conversational platforms

D. Leverage Conversational AI (such as Dialogflow or Vertex AI's capabilities) to handle user interaction and intent recognition

Google Cloud

D

The requirement is to create advanced chatbots that provide 24/7 personalized user support and utilize natural language understanding (NLU) for complex interactions.
• Conversational AI: The overall AI/ML strategy is to leverage LLMs and Conversational AI for personalized experiences and content virality. Dedicated conversational AI services (like Dialogflow or Vertex AI's specific tools) are purpose-built to manage complex, stateful, NLU-driven conversations, which is necessary for advanced 24/7 personalized assistance.

[Altostrat Media case study] Diagnostic Question #3

A key business requirement is to enable natural language interaction with the platform and provide 24/7 personalized user support via advanced chatbots. This chatbot needs to utilize Natural Language Understanding (NLU) to answer complex queries about content and suggest personalized recommendations.

Which Google Cloud service combination should be used to build and deploy this critical user engagement component?

A. Use a custom Python script deployed on a managed instance group (MIG) for NLU, and integrate it with BigQuery ML for recommendations.

B. Use a monolithic Node.js application deployed on GKE to manage all user interactions and recommendation logic

C. Utilize the Natural Language API directly within Cloud Functions to analyze user text input, avoiding any need for stateful conversational platforms

D. Leverage Conversational AI (such as Dialogflow or Vertex AI's capabilities) to handle user interaction and intent recognition

Google Cloud

D

The requirement is to create advanced chatbots that provide 24/7 personalized user support and utilize natural language understanding (NLU) for complex interactions.
• Conversational AI: The overall AI/ML strategy is to leverage LLMs and Conversational AI for personalized experiences and content virality. Dedicated conversational AI services (like Dialogflow or Vertex AI's specific tools) are purpose-built to manage complex, stateful, NLU-driven conversations, which is necessary for advanced 24/7 personalized assistance.

[Altostrat Media case study] Diagnostic Question #4

Altostrat currently relies on Cloud Monitoring and Prometheus, but alerts for critical system issues are primarily delivered via email. To accelerate and enhance the reliability of operational workflows, the architecture team needs to implement a low-latency, highly reliable alerting mechanism for major issues (e.g., GKE cluster failure or high-priority transcoding errors)

What action should the architect take to improve the immediacy and reliability of critical alerts?

A. Set up a dedicated Compute Engine instance to continuously parse incoming email alerts and manually execute runbooks.

B. Increase the retention period of Prometheus metrics to allow for better post-mortem analysis

C. Configure Cloud Monitoring to integrate with an external service like PagerDuty or Slack via Notification Channels

D. Use Cloud Deployment Manager to standardize the deployment of existing Cloud Monitoring dashboards

Google Cloud

C

While Altostrat uses Cloud Monitoring, relying on email alerts is not ideal for low-latency, highly reliable critical incident response.
• Notification Channels: Integrating Cloud Monitoring alerts with dedicated notification channels (like PagerDuty or Slack) allows for immediate, actionable, and reliable escalation policies, significantly accelerating the response time and enhancing the reliability of operational workflows compared to email.

[Altostrat Media case study] Diagnostic Question #4

Altostrat currently relies on Cloud Monitoring and Prometheus, but alerts for critical system issues are primarily delivered via email. To accelerate and enhance the reliability of operational workflows, the architecture team needs to implement a low-latency, highly reliable alerting mechanism for major issues (e.g., GKE cluster failure or high-priority transcoding errors)

What action should the architect take to improve the immediacy and reliability of critical alerts?

A. Set up a dedicated Compute Engine instance to continuously parse incoming email alerts and manually execute runbooks.

B. Increase the retention period of Prometheus metrics to allow for better post-mortem analysis

C. Configure Cloud Monitoring to integrate with an external service like PagerDuty or Slack via Notification Channels

D. Use Cloud Deployment Manager to standardize the deployment of existing Cloud Monitoring dashboards

Google Cloud

C

While Altostrat uses Cloud Monitoring, relying on email alerts is not ideal for low-latency, highly reliable critical incident response.
• Notification Channels: Integrating Cloud Monitoring alerts with dedicated notification channels (like PagerDuty or Slack) allows for immediate, actionable, and reliable escalation policies, significantly accelerating the response time and enhancing the reliability of operational workflows compared to email.

[Altostrat Media case study] Diagnostic Question #5

Altostrat deploys new applications with stateful information (like user management data) on GKE and uses Cloud SQL for the managed relational database backend. To ensure low latency and improved network security, the Cloud SQL instance is configured with a Private IP.

Which is the most secure and recommended way to ensure the GKE pods can connect reliably to the Cloud SQL instance?

A. Enable a Public IP address on the Cloud SQL instance and use SSL certificates for connection

B. Set up a dedicated VM as a jump host within the VPC network to proxy all database connections

C. Use the Cloud SQL Auth Proxy running as a sidecar container within the GKE pods to manage secure, IAM-based connectivity over the Private Service Access network

D. Configure the Cloud SQL instance with an Authorized Network (CIDR block) that encompasses the entire GKE node IP range

Google Cloud

C

Cloud SQL Auth Proxy: The proxy is the recommended option for connecting to Cloud SQL, even when behind a Private IP, because it handles strong encryption and uses IAM for secure authentication. Using a Private IP address alone provides lower network latency and improved network security (traffic is not exposed to the public internet), but the proxy ensures secure, managed authentication on top of that private connection.

[Altostrat Media case study] Diagnostic Question #5

Altostrat deploys new applications with stateful information (like user management data) on GKE and uses Cloud SQL for the managed relational database backend. To ensure low latency and improved network security, the Cloud SQL instance is configured with a Private IP.

Which is the most secure and recommended way to ensure the GKE pods can connect reliably to the Cloud SQL instance?

A. Enable a Public IP address on the Cloud SQL instance and use SSL certificates for connection

B. Set up a dedicated VM as a jump host within the VPC network to proxy all database connections

C. Use the Cloud SQL Auth Proxy running as a sidecar container within the GKE pods to manage secure, IAM-based connectivity over the Private Service Access network

D. Configure the Cloud SQL instance with an Authorized Network (CIDR block) that encompasses the entire GKE node IP range

Google Cloud

C

Cloud SQL Auth Proxy: The proxy is the recommended option for connecting to Cloud SQL, even when behind a Private IP, because it handles strong encryption and uses IAM for secure authentication. Using a Private IP address alone provides lower network latency and improved network security (traffic is not exposed to the public internet), but the proxy ensures secure, managed authentication on top of that private connection.

Make sure to...

Enjoy the journey as much as the destination!

Now that you know about the overall setup of this course and how to use the workbook, let's get started by exploring section 1 of the exam guide.