Reliability, Policy, and Security

Ensuring reliability

Data and service reliability

Available

Durable

Alternative Failover Backup Disaster Recovery (DR)

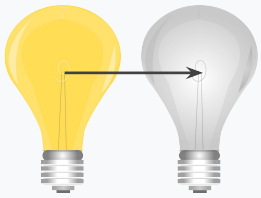
3

TIP: Available and durable are real-world values, usually not 100%.

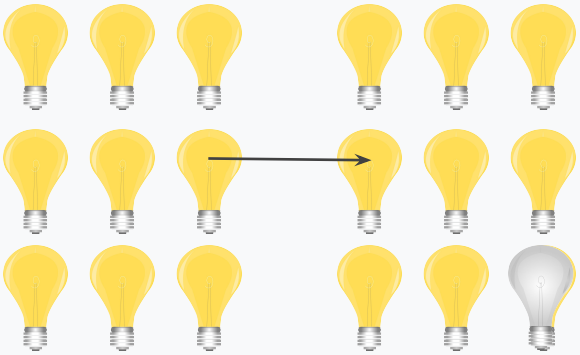
Reliable - Produces consistent outputs, operates as expected. A measure of how long the item performs its intended function.
Available - Accessible on demand. A measure of the % of time the item is in an operable state.
Durable - Data does not disappear, information is not lost. More accurately, a measure of the rate at which data is lost. Example: a diamond is durable.

https://pixabay.com/en/precious-diamond-jewelry-expensive-1199183/
https://pixabay.com/en/brick-wall-wall-brickwork-masonry-3170274/

Distributing for scale may improve reliability



1/1
100% out



1/9th
11% out

Distributing work for scale often reduces the impact of a single loss and increases reliability.

TIP

4

May improve reliability … If the solution is designed to be fault tolerant.

https://pixabay.com/en/light-bulb-electric-electric-bulb-146595/

Performing quality control

Monitor BigQuery with Stackdriver

TIP

You can monitor infrastructure and data services with Stackdriver.

- Available for all BigQuery customers
- Fully interactive GUI. Customers can create custom dashboards displaying up to 13 BigQuery metrics, including:
  - Slots Utilization
  - Queries in Flight
  - Uploaded Bytes (not shown)
  - Stored Bytes (not shown)

6

TIP: Integrated monitoring across services can simplify monitoring.
It is possible to surface application values as custom metrics in Stackdriver.

These charts show Slot Utilization, Slots available and queries in flight for a 1 hr period.

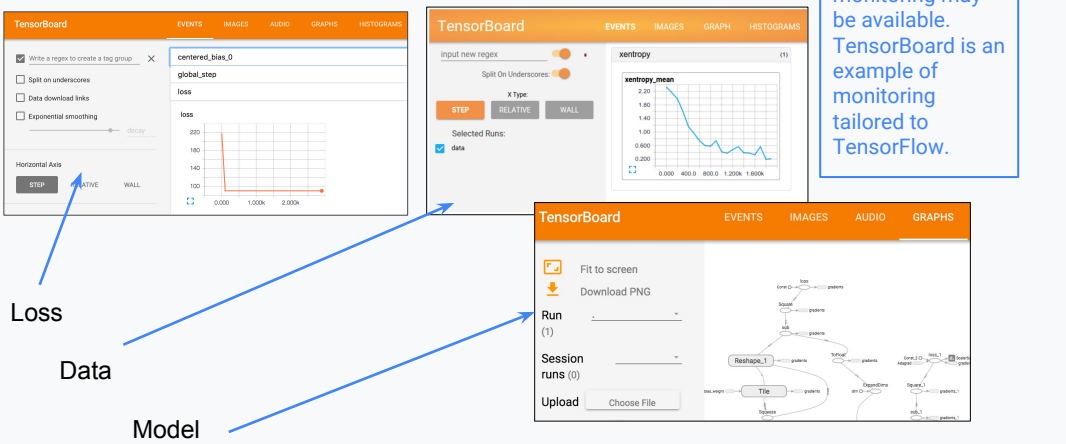
The Stackdriver charting tools offer

- Graphical User Interface to create custom dashboards for multiple GCP Products
- virtually real time data on many parameters (the lag on slot utilization for example is less than 5 minutes)
- Interactive graphical controls (zooming, creating new charts, selecting display modes, etc)

Known Issues:

- There is a known issue when Stackdriver reports slots available for customers that have subreservations. Please direct any questions to me.

Use TensorBoard to monitor training



TIP

Service-specific monitoring may be available. TensorBoard is an example of monitoring tailored to TensorFlow.

TensorBoard is a collection of visualization tools designed specifically to help you visualize TensorFlow.

- TensorFlow graph
- Plot quantitative metrics
- Pass and graph additional data

Events at top left shows “loss”.

Graphs at bottom right shows the linear model graph as built by TensorFlow.

Point TensorBoard at model output directory.

https://www.tensorflow.org/programmers_guide/summaries_and_tensorboard

Estimator comes with a method that handles distributed training and evaluation

```
estimator = tf.estimator.LinearRegressor(
                        model_dir=output_dir,
                        feature_columns=feature_cols)


...

tf.estimator.train_and_evaluate(estimator,
                                train_spec,
                                eval_spec)
```

TIP

In TensorFlow, data is often divided into training and evaluation sets, which defines a path for measuring effectiveness and for improvement.

Pass in:

1. Estimator
2. Train Spec
3. Eval Spec

- Distribute the graph
- Share variables
- Evaluate occasionally
- Handle machine failures
- Create checkpoint files
- Recover from failures
- Save summaries for TensorBoard

Some job skills are part of each technology

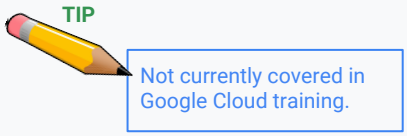
Assessing, **troubleshooting**, and **improving** data representations and improving data processing infrastructure

Troubleshooting and improving data quality and processing performance are distributed through all the technologies.

TIP: Troubleshooting and improving data quality and processing performance is distributed through all the technologies.

Some job skills are not technical

Advocating policies and
publishing data and reports



SCRIPT: "The training *does* cover the mechanics of generating and reports."

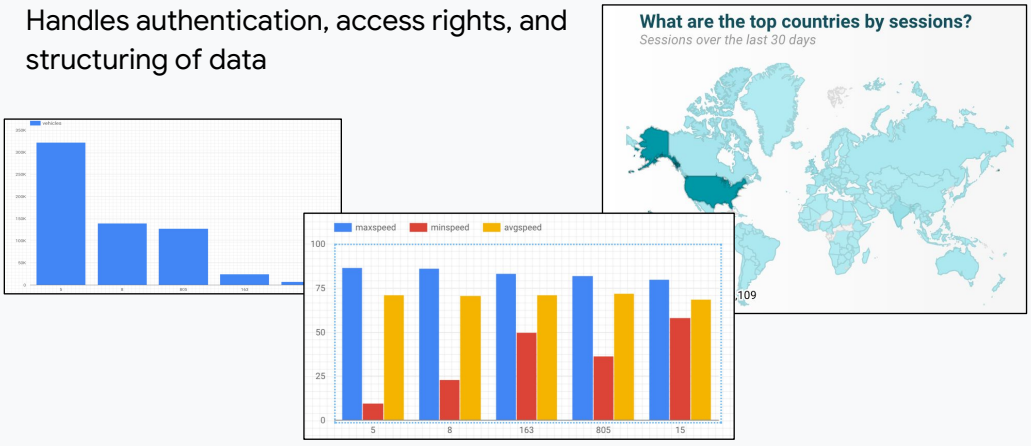
TIP: This subject is in the exam outline. It is a general Job skill rather than a technical skill and not specifically covered in Google technical training.

Visualizing data and advocating policy

Google Data Studio lets you build dashboards and reports

Easy to read, share, and fully customizable

Handles authentication, access rights, and structuring of data



Use Data Studio to visualize YouTube titles and aggregated view counts summarized over 30 days and segmented by Country Code in the fewest steps.

A. Set up a YouTube data source for your channel data for Data Studio. Set Views as the metric, and set Video Title as a report dimension. Set Country Code as a filter.
B. Set up a YouTube data source for your channel data for Data Studio. Set Views as the metric, and set Video Title and Country Code as report dimensions.
C. Export your YouTube views to Cloud Storage. Set up a Cloud Storage data source for Data Studio. Set Views as the metric, and set Video Title as a report dimension. Set Country Code as a filter.
D. Export your YouTube views to Cloud Storage. Set up a Cloud Storage data source for Data Studio. Set Views as the metric, and set Video Title and Country Code as report dimensions.

DEPE

Use Data Studio to visualize YouTube titles and aggregated view counts summarized over 30 days and segmented by Country Code in the fewest steps.

A. Set up a YouTube data source for your channel data for Data Studio. Set Views as the metric, and set Video Title as a report dimension. Set Country Code as a filter.
B. Set up a YouTube data source for your channel data for Data Studio. Set Views as the metric, and set Video Title and Country Code as report dimensions. ✓
C. Export your YouTube views to Cloud Storage. Set up a Cloud Storage data source for Data Studio. Set Views as the metric, and set Video Title as a report dimension. Set Country Code as a filter.
D. Export your YouTube views to Cloud Storage. Set up a Cloud Storage data source for Data Studio. Set Views as the metric, and set Video Title and Country Code as report dimensions.

DEPE

Solution

B is correct because there is no need to export; you can use the existing YouTube data source. Country Code is a dimension because it's a string and should be displayed as such, that is, showing all countries, instead of filtering.

A is not correct because you cannot produce a summarized report that meets your business requirements using the options listed.

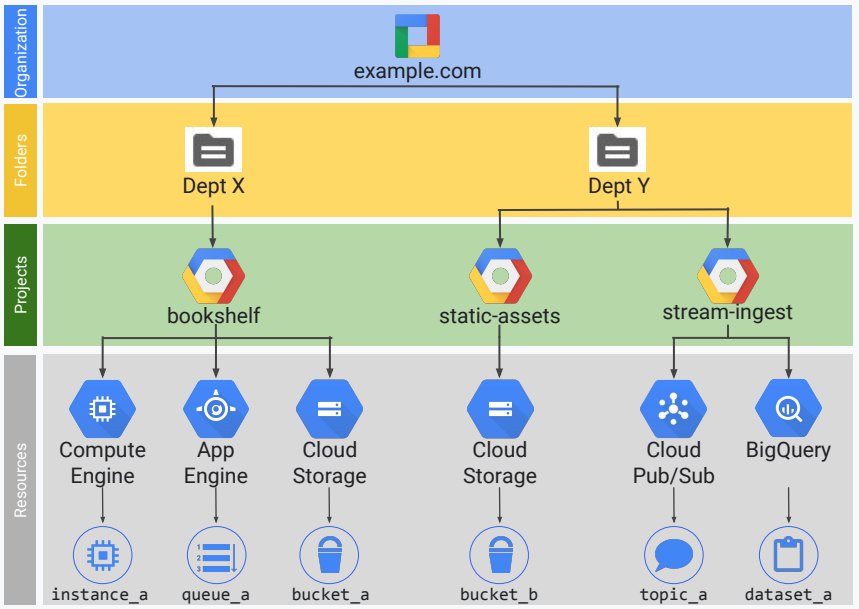
C and D are not correct because you do not need to export data from YouTube to Cloud Storage; you can simply use the existing YouTube data source.

Article: "About dimensions and metrics" in Data Studio dashboard Help.
Article: "Manage segments" in Data Studio dashboard Help.

DEPE
Obviously a connector. Country code as filter would simply drop out, not segment. Dimensions describe and group data, so it would have the effect of segmenting the report, however, Data Studio includes a feature called segments which is set separately for using Google Analytics Segments.

Designing for security and compliance

Cloud IAM Resource Hierarchy



A policy is set on a resource, and each policy contains a set of:
Roles
Role members
Resources inherit policies from parent:
Resource policies are a union of parent and resource.
If parent policy is less restrictive, it overrides a more restrictive resource policy.

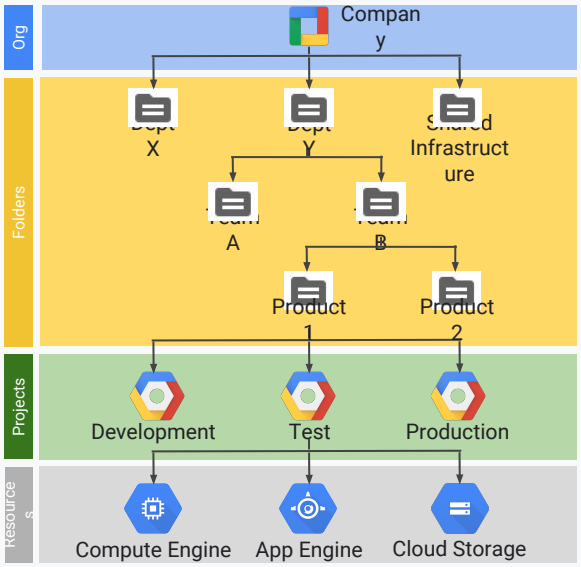
https://cloud.google.com/iam/
https://cloud.google.com/iam/docs/
https://cloud.google.com/iam/docs/concepts
https://cloud.google.com/iam/docs/understanding-roles
https://cloud.google.com/iam/docs/service-accounts

Folders

Additional grouping mechanism and isolation boundaries between projects:

- Different legal entities
- Departments
- Teams

Folders allow delegation of administration rights.



Folder map well to organization structure. It is a way to isolate organizations or users or products while still having them share billing and corporate resources.

https://cloud.google.com/resource-manager/docs/managing-multiple-orgs
https://cloud.google.com/resource-manager/docs/access-control-org

https://cloud.google.com/resource-manager/docs/creating-managing-projects
https://cloud.google.com/resource-manager/docs/access-control-proj

https://cloud.google.com/resource-manager/docs/creating-managing-folders
https://cloud.google.com/resource-manager/docs/access-control-folders

Identity and access

Separate responsibilities.

Always have a backup or alternative in case the responsible person is unreachable.

Have a separate maintenance path when the normal paths aren't working (e.g., bastion host).

Use groups to allocate permissions, then separately manage group membership.

Customize roles for greater granularity of permissions.

Give each group only the permissions they need to perform that job or task.

Place critical functions on service machines to create accountability trail (login log, activity monitoring).

Backup/spare logs and records; have a review, analysis, and monitoring strategy (ex: monthly reports).

TIP: Commit a security checklist to memory. Sometimes just running down a list will rapidly identify a solution.

https://pixabay.com/en/data-security-keyboard-computer-1590455/

Encryption of VM disks and Cloud Storage buckets

Default Encryption	Customer-Managed Encryption Keys (CMEK)	Customer-Supplied Encryption Keys (CSEK)	Client-Side Encryption
Data is automatically encrypted before being written to disk. Each encryption key is itself encrypted with a set of master keys.	Google-generated **data** encryption key (DEK) is still used. Allows you to create, use, and revoke the **key** encryption key (KEK). Uses Cloud Key Management Service (Cloud KMS).	Keep keys on premises, and use them to encrypt your cloud services. Google can't recover them. Disk encryption on VMs Cloud Storage encryption. Keys are never stored on disk unencrypted. You provide your key at each operation, and Google purges it from its servers when each operation completes.	Data is encrypted before it is sent to the cloud. Your keys; your tools Google doesn't know whether your data is encrypted before it's uploaded. No way to recover keys. If you lose your keys, remember to delete the objects!

20

Encryption options
https://cloud.google.com/security/encryption-at-rest/

Customer Managed Encryption Keys (CMEK) *using Cloud KMS*
When you use Cloud Dataproc, cluster and job data is stored on Persistent Disks (PDs) associated with the Compute Engine VMs in your cluster and in a Cloud Storage bucket. This PD and bucket data is encrypted using a Google-generated data encryption key (DEK) and key encryption key (KEK). The CMEK feature allows you to create, use, and revoke the key encryption key (KEK). Google still controls the data encryption key (DEK).

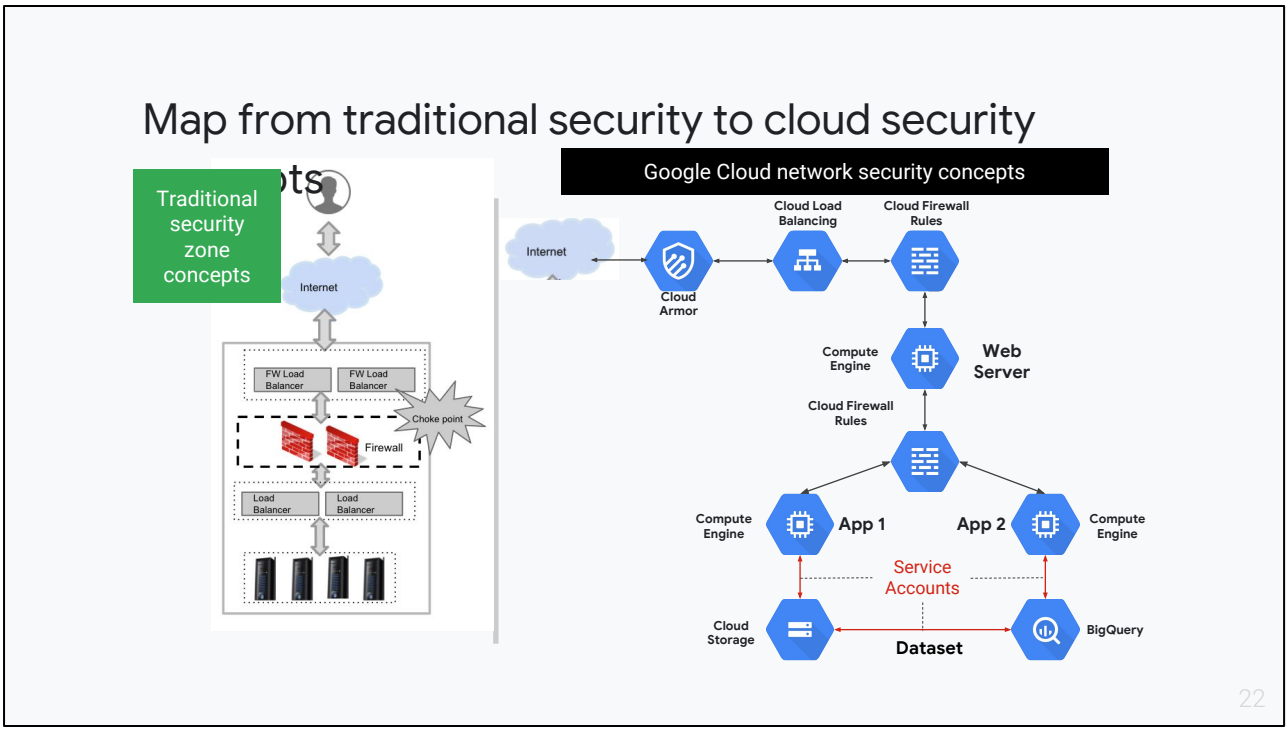
Default Encryption
Encryption at rest uses the Key Management System (KMS) to generate KEKs and DEKs.
https://cloud.google.com/security/encryption-at-rest/default-encryption/#key_management

https://pixabay.com/en/key-old-skeleton-lock-metal-door-30417/

Key Management Service (KMS): https://cloud.google.com/kms/
AES256 keys
Generate keys
Usage includes off-cloud
Key rotation
When a key is destroyed, there is a 24-hour delay
API support
Envelope DEK/KEK

Client-Side Encryption

https://cloud.google.com/storage/docs/encryption/client-side-keys



Key concepts: Cloud Armor, Cloud Load Balancing, Cloud Firewall Rules, Service Accounts, separation into front-end and back-end, isolation of resources using separate service accounts between services.

TIP: Because of pervasive availability of firewall rules, you don't have to install a router in the network at a particular location to get firewall protection. That means you can layer the firewalls as shown in this example.

TIP: Because if pervasive support for Service accounts, you can "lock down" connections between components.

Security

1. Security is an "umbrella" term for many specific services. Access. Authorization. Accountability. Privacy. Authentication. Encryption.
2. Layering security measures provides a greater deterrent due to the synergy of multiple methods working together.
3. In some cases, practical security involves raising the effort required to bypass the security above the value of the item being protected. Sometimes called "castle logic". The walls of the castle (and depth of the moat) should be higher (and deeper) than the value of the treasure in the castle.

When faced with a security question on an exam (or in practice), determine which of the specific technologies/services is being discussed (authentication, encryption) for example. Then determine exactly what the goals are for sufficient security. Is it deterrence? Is it meeting a standard for compliance? Is the goal to eliminate a particular risk or vulnerability? This will help you define the scope of a solution, whether on an exam or in application.

Groups Analyst1 and Analyst2 should not have access to each other's BigQuery data.

A. Place the data in separate tables, and assign appropriate group access.
B. Analyst1 and Analyst2 must be in separate projects, along with the data.
C. Place the data in separate datasets, and assign appropriate group access.
D. Place the data in separate tables, but encrypt each table with a different group key.

Groups Analyst1 and Analyst2 should not have access to each other's BigQuery data.

A. Place the data in separate tables, and assign appropriate group access.
B. Analyst1 and Analyst2 must be in separate projects, along with the data.
C. Place the data in separate datasets, and assign appropriate group access. ✓
D. Place the data in separate tables, but encrypt each table with a different group key.

Solution

C is correct. BigQuery data access is controlled at the dataset level.

A is not correct because BigQuery does not provide IAM access control to the individual table. B is not correct because the Analyst groups can be in the same project. D is incorrect because encryption does not determine access.

https://cloud.google.com/bigquery/docs/access-control

Provide Analyst3 secure access to BigQuery query results, but not the underlying tables or datasets.

A. Export the query results to a public Cloud Storage bucket.
B. Create a BigQuery Authorized View and assign a project-level user role to Analyst3.
C. Assign the bigquery.resultsonly.viewer role to Analyst3.
D. Create a BigQuery Authorized View and assign an organization-level role to Analyst3.

Provide Analyst3 secure access to BigQuery query results, but not the underlying tables or datasets.

A. Export the query results to a public Cloud Storage bucket.
B. Create a BigQuery Authorized View and assign a project-level user role to Analyst3. ✓
C. Assign the bigquery.resultsonly.viewer role to Analyst3.
D. Create a BigQuery Authorized View and assign an organization-level role to Analyst3.

Solution

B is correct. You need to copy/store the query results in a separate dataset and provide authorization to view and/or use that dataset.

A is not secure. C: The readonly.viewer role does not exist AND secure access cannot be applied to a query. D: An organizational role is too broad and violates the principle of "least privilege."

https://cloud.google.com/bigquery/docs/share-access-views

Ensuring Solution Quality

Exam Guide Review

Security

Designing for security and compliance.	Identity and access management Data security Ensuring privacy Legal compliance

Tip: IAM -- Understand permissions and custom roles. Under what conditions are custom roles preferred over standard predefined roles?

Tip: Data security -- DLP -- Cloud DLP allows you to minimize what you collect, store, expose, or copy. Classify or automatically redact sensitive data from text streams before you write to disk, generate logs or perform analysis.

Be familiar with all these:

- Cloud IAM
- Encryption, Key Management
- Data Loss Prevention API
- HIPPA, COPPA, FedRAMP, GDPR

Efficiency

Ensuring scalability and efficiency.	Building and running test suites Pipeline monitoring Assessing, troubleshooting, and improving data representations and data processing infrastructure Resizing and autoscaling resources

Tip: A lot of administration over resources is presented in the console. But a lot of runtime information such as logs, performance, and so forth is presented and reported in Stackdriver. Stackdriver provides information for troubleshooting both functional and performance issues.

- Stackdriver

Reliability

Ensuring reliability and fidelity.	Performing data preparation and quality control Verification and monitoring Planning, executing, and stress testing data recovery Choosing between ACID, idempotent, eventual consistency requirements

Tip: Establishing standard data quality at ingress using Cloud Dataprep or by running an ETL pipeline can prevent many problems later in processing that would be difficult to troubleshoot.

Tip: Keep in mind the business purpose of the data processing. How resilient does the application need to be? For example, financial transactions usually cannot be dropped and must not be duplicated. But a statistical analysis might be equally valid if a small amount of data is lost. These assumptions influence the approach to rerunning failed jobs.

Study these:

- Cloud Dataprep
- Fault-tolerance
- Rerunning failed jobs
- Performing retrospective re-analysis

Portability

Ensuring flexibility and portability.	Mapping to current and future business requirements Designing for data and application portability

Tip: Where is the official authoritative data (sometimes called the source of truth) and where are the replicas? How frequently does data need to be shared or updated? Can smaller parts of the data be synchronized to reduce costs?

Tip: Where is the data stored? Where is the data going to be processed? Can data storage and data processing be in locations near one another?

Tip: When will the data need to be exported? How difficult and expensive will it be to make this happen? For example, you might want to store data in a different location or in a different type of storage to meet business requirements for portability.

- Multi-cloud data residency requirements


Google Cloud Training and Google Cloud