Google Cloud

Digital Transformation with Google Cloud

This study guide covers the materials from the first course of the six-course series for the Cloud Digital Leader certification.

Certification exam focus

The Cloud Digital Leader certification exam is designed to assess knowledge of cloud literacy, common business use cases, and how cloud solutions and the capabilities of Google Cloud products support an enterprise.

What is the cloud?

The cloud refers to the on-demand delivery of computing power, database storage, apps, and other IT resources via the internet with pay-as-you-go pricing.

What is digital transformation?

When an organization uses modern digital technologies such as cloud platforms to create or modify business processes, culture, and customer experiences to meet changing business and market dynamics, it's considered digital transformation.

The common drivers of digital transformation include:

- Staying ahead of competitors.
- Avoiding integration tax.
- New security challenges.
- The need for a connected architecture and flexible solutions.

Cloud deployment models and strategies

Models:

- **On-premises:** When an organization hosts and manages its own apps and data on servers located in its own data center or facility.
- **Private cloud:** Cloud computing resources used exclusively by one organization. Resources can be physically located on the organization's site (on-premises) or managed by a third-party provider.
- **Public cloud:** Cloud services offered by a third-party provider (like Google Cloud) over the public internet. These resources are shared by multiple organizations or users.

Strategies:

- **Hybrid cloud:** Infrastructure environment that combines a public cloud and a private cloud (or on-premises infrastructure), allowing data and apps to be shared between them.
- **Multicloud:** The use of cloud services from more than one different public cloud provider.

Benefits of cloud technology

Scalability

Agility

Cost-effectiveness

High availability

Strategic value

Flexibility

Speed

Global reach

Data-driven insights

Security

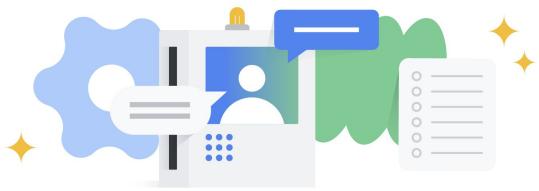
Google Cloud

Digital Transformation with Google Cloud

Google Cloud's strengths

- World-leading AI
- Openness and interoperability
- AI Hypercomputer
- AI-optimized data platform
- Security
- Global network

Agentic AI: Refers to AI systems that can independently plan, use tools, and interact with other systems to complete complex, multistep workflows.



Cloud computing service models

Infrastructure as a Service (IaaS)

Similar to leasing a car, IaaS provides on-demand access to core resources like compute and storage without the burden of owning hardware. It offers high scalability and a pay-as-you-go model.

Example: Compute Engine, Cloud Storage

Platform as a Service (PaaS)

Operating like a taxi service, PaaS provides a complete environment for developers to build and run applications while the provider manages patches and infrastructure.

Example: Gemini Enterprise Agent Platform

Software as a Service (SaaS)

Comparable to riding a bus, SaaS delivers complete, ready-to-use applications directly through a web browser. It eliminates the need for local installation or maintenance, providing a cost-effective, subscription-based solution.

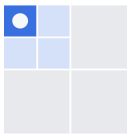
Example: Google Workspace

Networking terms and concepts

- **IP address:** Unique numerical label that acts as a digital mailing address for a device.
- **Domain Name System (DNS):** Translates human-readable web addresses into numerical IP addresses.
- **Latency:** Time it takes for a data packet to travel from its source to its destination.
- **Bandwidth:** The maximum amount of data that can be transmitted over a network connection in a given amount of time.

Google Cloud's network

Regions: Independent geographic areas containing multiple zones.



Zones: Isolated locations within regions to protect against single-failure events.

Edge locations: Globally distributed points of presence that cache content near users.

Open standard versus open source

- **Open standard:** A publicly available specification for a technology (for example, TCP/IP and SQL). By providing a common language or framework, open standards prevent vendor lock-in and ensure that diverse systems can communicate seamlessly.
- **Open source:** Software for which the source code is made publicly available and can be inspected, modified, and enhanced by anyone (for example, Linux and PyTorch). This model encourages rapid innovation, as a global community of developers can contribute to the software.

Google Cloud

Exploring Data Transformation with Google Cloud

This study guide covers the materials from the second course of the six-course series for the Cloud Digital Leader certification.

Certification exam focus

The Cloud Digital Leader certification exam is designed to assess knowledge of cloud literacy, common business use cases, and how cloud solutions and the capabilities of Google Cloud products support an enterprise.

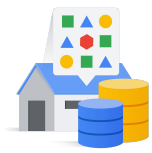
The value of data

Data represents raw information—both structured and unstructured—that an organization collects from every interaction and process. The importance of data lies in its ability to empower smarter, real-time decision-making and personalized customer experiences.

By leveraging an intelligent data cloud to analyze everything from financial reports to social media sentiment, organizations can perceive emerging trends, predict future needs, and maintain a competitive edge in a rapidly changing market.

Data storage

Database: An organized collection of data stored in tables and accessed electronically from a computer system.



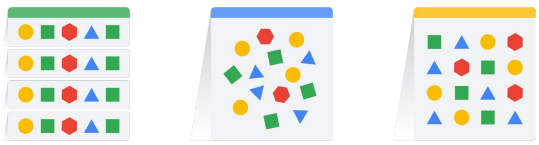
Data warehouse: An enterprise system used for the analysis and reporting of structured and semistructured data from multiple sources.

Data lake: A repository designed to ingest, store, explore, process, and analyze any type or volume of raw data regardless of the source.

It is important to maintain openness and interoperability with data management platforms. This allows data to move freely between different tools, ensuring it doesn't get trapped in isolated systems or tied to one specific provider.

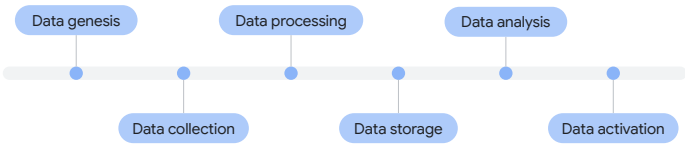
Types of data

- **First-party:** Proprietary customer datasets collected directly by a business through its own audience transactions and digital interactions.
- **Second-party:** Data owned by another organization, such as a partner, that is used to augment a company's own internal datasets.
- **Third-party:** External datasets collected by organizations with no direct customer relationship; often purchased to provide broader context and insights.



- **Structured:** Highly organized and well-defined information typically stored in tables with clear relationships between rows and columns.
- **Unstructured:** Information that lacks a predefined data model or organization, including formats like text and video files, which historically required machine learning to analyze.
- **Semistructured:** Data that falls between categories, lacking a formal database structure but containing tags such as in JSON files, making it easier to analyze than unstructured data.

Data supply chain



The steps of the data supply chain include data genesis, data collection, data processing, data storage, data analysis, and data activation.

Google Cloud

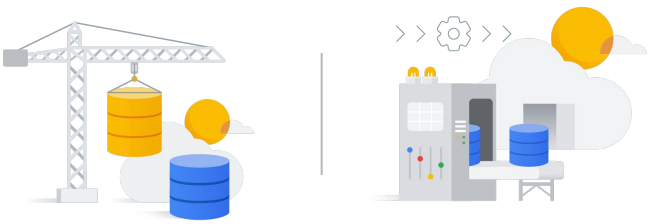
Exploring Data Transformation with Google Cloud

Key data management terms

- **Relational:** Organizes structured data into rows and columns within tables. It uses a predefined schema, defines relationships via keys, and is managed using SQL.
- **Non-relational:** Database approach designed to handle large volumes of rapidly changing data that doesn't fit well into the rigid, tabular structure of a relational database.
- **Object storage:** Data storage architecture that manages data as objects in a flat address space rather than in a file hierarchy. This is ideal for unstructured data at massive scale.
- **SQL:** The standard, declarative language used to manage, query, and manipulate relational databases.
- **NoSQL:** A term for database systems that do not use SQL as their primary query language. NoSQL offers flexible schema models, often prioritizing availability and scalability over strict consistency.

Database migration and modernization

Organizations can modernize databases via a simple lift-and-shift for minimal upheaval or choose a fully managed migration using tools like Google Cloud's Database Migration Service and Datastream to offload maintenance.



Google Cloud data management

Cloud Storage	Object storage service for unstructured data, like images, videos, logs, and backups.
Cloud Spanner	Relational database service that offers both traditional relational structure and horizontal scalability.
Cloud SQL	Fully managed, relational database service for MySQL, PostgreSQL, and SQL Server databases on Google Cloud.
AlloyDB	Fully managed, PostgreSQL-compatible database service for demanding enterprise workloads.
Bigtable	Fully managed, NoSQL database service for large analytical and operational workloads.
BigQuery	Serverless, highly scalable, and cost-effective cloud data warehouse.
Firestore	Flexible, scalable, NoSQL document database for mobile, web, and server development.

Cloud Storage classes

- **Standard:** Highest storage cost but no retrieval fees. Best for data accessed frequently.
- **Nearline:** Lower storage cost with a small retrieval fee. Ideal for data accessed about once a month.
- **Coldline:** Even lower storage cost but higher retrieval fees. Best for data accessed about once a quarter.
- **Archive:** Lowest storage cost with the highest retrieval fees. Best for data accessed less than once a year.
- **Autoclass:** Automatically transitions objects between classes based on access patterns to optimize costs.

Google Cloud

Exploring Data Transformation with Google Cloud

Looker

Looker is a web-based Google Cloud business intelligence platform that democratizes data by allowing anyone to explore data and answer their own questions without technical expertise, empowering organizations to uncover and act on insights immediately.

Pairing Looker with SQL databases, like Google Cloud's BigQuery, allows organizations to create real-time reports and dashboards and integrate data into workflows.

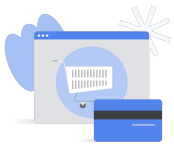
Data governance

Data governance is a framework of processes, roles, and standards that ensures information is used effectively to achieve organizational goals. It establishes clear rules for who can access and manage data, as well as how and when they can do so.

Streaming analytics

Streaming analytics enables businesses to process and analyze data records continuously as they are generated rather than waiting for traditional delayed batches. This real-time approach allows organizations to react instantly to time-sensitive events, ensuring decisions are made using the freshest data possible.

Use cases include:



Ecommerce: User clickstreams can be analyzed to optimize the shopping experience with real-time pricing, promotions, and inventory management.

Financial services: Account activity can be analyzed to detect abnormal behaviors in the data stream and generate a security alert.

News media: User click records can be streamed from various news source platforms, and data can then be enriched with demographic information to better serve articles that are relevant to the targeted audience.

Google Cloud's data processing tools

Pub/Sub

A fully managed, asynchronous messaging service used to publish and subscribe to real-time data streams. It ensures reliable event delivery and seamless data distribution for streaming analytics.

Dataflow

A fully managed service for executing data processing pipelines. It automatically scales resources, optimizes processing, and is built on the open-source Apache Beam model.

Managed Service for Apache Spark

A fully managed Apache Spark and Apache Hadoop service on Google Cloud. It allows users to run large-scale data processing jobs with reduced overhead for cluster management.

Google Cloud

Innovating with Google Cloud Artificial Intelligence

This study guide covers the materials from the third course of the six-course series for the Cloud Digital Leader certification.

Certification exam focus

The Cloud Digital Leader certification exam is designed to assess knowledge of cloud literacy, common business use cases, and how cloud solutions and the capabilities of Google Cloud products support an enterprise.

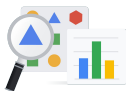
What is artificial intelligence (AI)?

AI is a broad field of computer science dedicated to building systems that can simulate human intelligence to perform tasks. Tasks include learning, reasoning, problem-solving, and decision-making.

Fundamental AI concepts

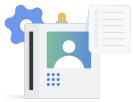
Machine learning (ML): A subset of AI with the ability to automatically learn and improve from experience without being explicitly programmed. ML uses algorithms to analyze data, identify patterns, and make predictions or decisions.

Generative AI (gen AI): A type of AI that can produce new content, including text, images, audio, and synthetic data. Common uses include chatbots, content generation, and document synthesis.



Data analytics: The science of examining raw data to draw conclusions. It involves various techniques and processes to clean, transform, and interpret data to uncover insights and support decision-making.

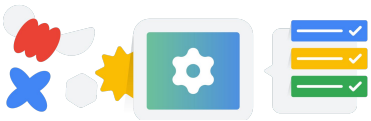
Business intelligence: A set of strategies, processes, apps, data, and technologies used to collect, analyze, present, and monitor business information. The primary goal is to improve business decisions by providing historical, current, and predictive views of business operations.



Agentic AI: A type of AI that is fundamentally reshaping the way people work. Individuals can use AI systems to interact with each other to complete complex, multistep workflows. Some use cases include workforce productivity, customer support, product innovation, operations, and research.

Responsible and explainable AI

Responsible AI refers to a governance framework covering fairness and accountability of AI systems.



Explainable AI is a specific subset of this framework that is focused on the internal logic and interpretability of model outputs. Responsible and explainable AI both aim to build trust and mitigate risk by ensuring technology is used ethically and safely.

The importance of high-quality data

The accuracy of ML predictions relies on large volumes of data. If data has errors, is not aligned to the problem, or is biased in some way, it is considered low quality. Data is evaluated against six dimensions to ensure that it is high quality:

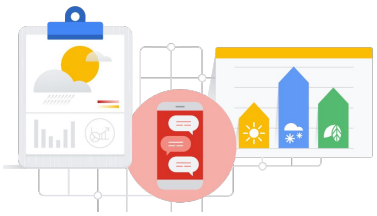
- Completeness
- Uniqueness
- Timeliness
- Validity
- Accuracy
- Consistency

Google Cloud

Innovating with Google Cloud Artificial Intelligence

Business problems that AI can help solve

- Replacing or simplifying rule-based systems.
- Deriving business insights from large datasets (both structured and unstructured).
- Scaling business decisions.

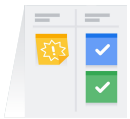


Key benefits of Google Cloud AI offerings

- Best infrastructure for AI
- AI-ready data cloud
- Sophisticated 1P models
- All-in-one AI developer platforms
- Prebuilt AI agents and apps

Consider these factors when selecting a Google Cloud AI solution for your business:

Implementation speed. How quickly do you need to get your model to production?



Potential for business differentiation. How unique is your model?

Technical expertise requirements. Does your staff need upskilling and training to implement an AI strategy?

Choice and flexibility. Does the solution require customization, or will it use managed services?

Development effort. How complex is the problem, and how much data is available?

Gemini Enterprise Agent Platform: A comprehensive platform to build, scale, govern, and optimize agents.

Google Cloud pretrained APIs

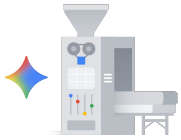
Gemini Enterprise Agent Platform API: Builds, orchestrates, and governs autonomous enterprise AI agents. Use case: Workflow automation.

Vision API: Extracts text, objects, and insights from visual imagery. Use case document OCR.

Cloud translation API: Dynamically translates text across thousands of language pairs. Use case: Website localization.

Speech-to-text API: Converts spoken audio into accurate written text. Use case: Call transcription.

Other Google Cloud AI solutions



Gemini: Google's advanced multimodal reasoning and generation model. Use case: Software development.

Agent Studio on Gemini Enterprise Agent Platform: Development workspace used to build autonomous, high-value business agents by grounding foundation models in proprietary data, visually designing operational logic, and integrating APIs to automate complex workflows.

AI Hypercomputer: Maximizes AI workload performance and efficiency by integrating optimized hardware like GPUs and TPUs with open software standards and flexible, cost-controlled consumption models. Use case: Supercomputing workloads.

BigQuery ML: Builds and executes ML models using standard SQL. Use case: Churn prediction.

Google Cloud

Modernize Infrastructure and Applications with Google Cloud

This study guide covers the materials from the fourth course of the six-course series for the Cloud Digital Leader certification.

Certification exam focus

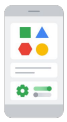
The Cloud Digital Leader certification exam is designed to assess knowledge of cloud literacy, common business use cases, and how cloud solutions and the capabilities of Google Cloud products support an enterprise.

Cloud migration terms

- **Compute:** The ability of a machine to process information, like handling tasks such as storage, retrieval, and analysis.
- **Discovery and assessment:** The initial phase of a migration project where an organization inventories its existing on-premises environment.
- **Workload:** A specific app, service, or capability that can run in the cloud or on-premises.

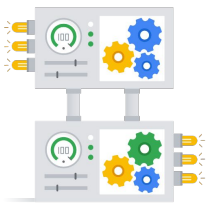
- **Retired workload:** When a workload is removed from a platform.
- **Retained workload:** When a workload is intentionally kept on a platform.
- **Rehosted workload (lift and shift):** When a workload is migrated to the cloud without changing anything in the its code or architecture.
- **Replatformed workload (move and improve):** When a workload is migrated to the cloud after making changes to its code or architecture.
- **Reimagine:** The process of rethinking how an organization uses technology to achieve its business goals.

App: A computer program or software that helps users do something.



Cloud infrastructure terms

- **Virtual machines (VMs):** A software emulation of a physical computer that runs its own OS using shared hardware resources.



- **Spot VMs:** Discounted, excess compute resources that Google Cloud can reclaim at any time.
- **Managed services:** Cloud services where the provider handles maintenance and operations, letting users focus entirely on the app.
- **Containerization and containers:** A lightweight virtualization method that packages code and its dependencies to ensure consistent performance across different environments.

Compute Engine: Google Cloud's Infrastructure as a Service (IaaS) product that lets users create and run VMs on Google infrastructure.

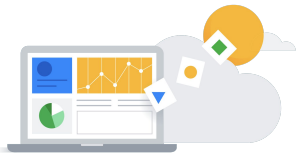
- No upfront investments.
- Each VM functions as a complete OS.

Google Cloud

Modernize Infrastructure and Applications with Google Cloud

Cloud architecture terms

Microservices: When an app is structured as a collection of small, autonomous services that communicate over a network, rather than one large, monolithic block of code.



Serverless computing: An architectural pattern where the developer focuses on the logic while the execution environment is abstracted away.

Load balancing: Distributing incoming network traffic across a group of backend servers to ensure no single server becomes overwhelmed.

Kubernetes: An open-source system for automating the deployment, scaling, and management of containerized apps.

Google Kubernetes Engine (GKE):

A Google-hosted managed Kubernetes service in the cloud. The GKE environment consists of multiple Compute Engine instances, grouped to form a cluster.

Benefits of modern cloud app development

- Microservices architecture
- Deployment with managed services
- Pay-as-you-go pricing
- Scalability
- Highly available and resilient
- Monitoring and management tools

Google Cloud serverless computing products

Cloud Run

A fully managed environment for running containerized apps that can answer multiple events at the same time.

Cloud Run functions

A platform for hosting simple, single-purpose functions that are attached to events one at a time that are emitted from your cloud infrastructure and services.

App Engine

A service to build and deploy web apps.

Google Cloud products supported on multicloud and hybrid cloud environments

AlloyDB Omni | **BigQuery Omni** | **GKE Enterprise** | **Cloud SQL** | **Looker**

Google Cloud

Modernize Infrastructure and Applications with Google Cloud

Application programming interface (API):
A set of protocols and definitions that enables distinct software apps to interface and exchange data. It serves as a standardized translation layer, ensuring predictable, secure communication between systems.

Apigee API Management

Google Cloud's API management service to operate APIs with enhanced scale, security, and automation.

- It helps organizations secure their APIs.
- It tracks and analyzes API usage.
- It helps with developing and deploying APIs.
- It offers API versioning, API documentation, and even API throttling.

Google Cloud

Trust and Security with Google Cloud

This study guide covers the materials from the fifth course of the six-course series for the Cloud Digital Leader certification.

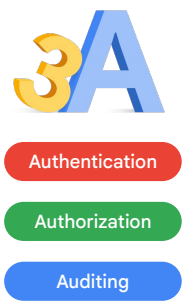
Certification exam focus

The Cloud Digital Leader certification exam is designed to assess knowledge of cloud literacy, common business use cases, and how cloud solutions and the capabilities of Google Cloud products support an enterprise.

Cyber attacks and threats

- **Social engineering:** Using deceptive emails (phishing) to trick people into downloading malware, sharing passwords, or exposing data.
- **Malware, viruses, and ransomware:** Malicious software that aims to disrupt operations, inflict damage, or gain unauthorized access to computer systems.
- **DDoS attacks:** Flooding a service with illegitimate traffic from compromised systems, denying access to genuine users.
- **Large language model (LLM) attacks:** Targeting generative AI (gen AI) to manipulate output, extract sensitive data, bypass safety guardrails, or abuse computing resources.

The three As of cloud identity management



- **Authentication:** Presenting unique credentials, such as passwords, physical tokens, or biometric data like fingerprints or voice recognition.
- **Authorization:** Determining what a user or system is allowed to do within a system.
- **Auditing/accounting:** Monitoring and tracking user activities within a system.

Concepts related to securing sensitive data

- **Privileged access:** Grants specific users access to a broader set of resources than ordinary users.
- **Least privilege:** Advocates granting users only the access they need to perform their job responsibilities.
- **Zero-trust architecture:** Assumes that no user or device can be trusted by default.

Concepts related to mitigating cyber threats

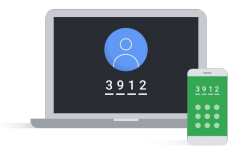
- **Security by default:** Emphasizes integrating security measures into systems and apps from the initial stages of development.
- **Security posture:** The overall security status of a cloud environment, measured against defined baselines.
- **Cyber resilience:** An organization's ability to withstand and recover quickly from cyber attacks.

Concepts related to securing cloud resources

- **Firewall:** A network device that regulates traffic based on predefined security rules.
- **Data Loss Prevention:** Strategies and tools to automatically identify, monitor, and protect sensitive information.

Two-step verification (2SV)

A security feature that requires users to provide two different forms of identification to log in.



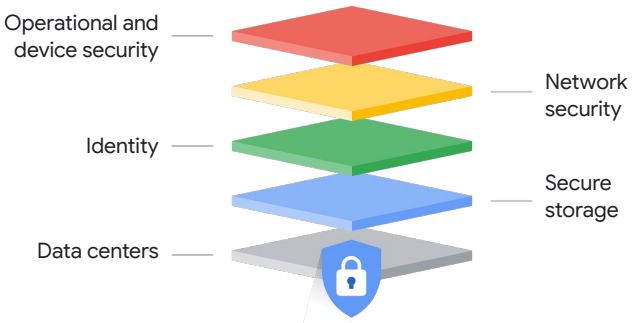
Google Cloud

Trust and Security with Google Cloud

Cloud security components

- **Control:** The policies and technical safeguards used to mitigate security risks.
- **Compliance:** The adherence to laws, regulations, and industry standards.
- **Confidentiality:** Keeping important information safe and secret.
- **Integrity:** Keeping data accurate and trustworthy.
- **Availability:** Ensuring cloud systems and services are accessible and ready for use by the right people when needed.

Google's multilayered security approach



Encryption and decryption

- **Encryption:** The process of converting data into an unreadable format by using an algorithm and a secret key.

  Data encryption in different states: In use, in transit, at rest.

- **Decryption:** The use of an encryption key to restore encrypted data back to its original form.

Google Threat Intelligence

A Google solution that provides curated data on malicious actors, indicators of compromise, and known attack patterns. It's powered by five unique data sources:

Google Threat Intelligence

1. Frontline intelligence
2. Crowdsourced threat intelligence
3. Open-source threat intelligence (OSINT)
4. Human-curated threat intelligence
5. Google insights

Security Command Center

A centralized security management and risk platform designed to protect cloud workloads. It provides:

- **Posture management:** It discovers cloud assets, scans for misconfigurations, and continuously monitors compliance against major industry standards.
- **Vulnerability detection:** It identifies weaknesses in virtual machines, containers, and web apps before they can be exploited.
- **Threat detection:** It monitors system logs and runtime environments in real time to detect active threats like malware, unauthorized access, and resource abuse.

Security operations (SecOps)

The continuous collaboration between security and IT teams to monitor, detect, and mitigate cyber threats in real time.

Google Cloud

Trust and Security with Google Cloud

Google security offerings

- **Cloud VPN**
- **Cloud Interconnect**
- **Cloud Armor**
- **Cloud Logging**
- **Identity and Access Management (IAM)**
- **Confidential Computing**
- **Certificate Manager**
- **Identity-Aware Proxy**

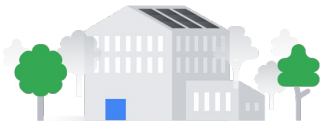
Google's AI security offerings

- **Gemini in Google SecOps:** An advanced gen AI-powered feature integrated into the Google SecOps platform that acts as a virtual security analyst, dramatically speeding up threat detection, investigation, and response workflows.
- **AI Protection**: Integrated with Security Command Center, it's a set of capabilities designed to safeguard AI workloads and data across clouds and models, irrespective of the platforms you choose to use.
- **Model Armor:** A fully managed Google Cloud security service designed to enhance the security and safety of LLM apps, particularly gen AI models.

Google provides valuable insights and accountability through transparency reports and third-party audits, which shed light on government and corporate actions that affect privacy, security, and access to information.

Google-built global network and data centers

- **Physical security:** Access is managed through multiple physical layers, including custom-designed electronic access cards, biometric scanners, laser barriers, and 24/7 human security monitoring.
- **Networking isolation**: The global network is a purpose-built private fiber network that handles all customer traffic. This dramatically reduces exposure to the public internet and enables fast, secure communication across the globe.



Google Cloud trust principles

1. You own your data, not Google.
2. Google does not sell customer data to third parties.
3. Google Cloud does not use customer data for advertising.
4. All customer data is encrypted by default.
5. We guard against insider access to your data.
6. We never give any government entity "backdoor" access.
7. Our privacy practices are audited against international standards.

Concepts about storing and keeping data secure

- **Data sovereignty:** The legal concept that data is subject to the laws and regulations of the country where it resides.
- **Data residency:** The physical location where data is stored or processed.

Google Cloud

Scaling with Google Cloud Operations

This study guide covers the materials from the last course of the six-course series for the Cloud Digital Leader certification.

Certification exam focus

The Cloud Digital Leader certification exam is designed to assess knowledge of cloud literacy, common business use cases, and how cloud solutions and the capabilities of Google Cloud products support an enterprise.

Expenditures and total cost of ownership (TCO)

Moving from on-premises to the cloud shifts costs form capital expenditures (CapEx) to operational expenditures (OpEx).

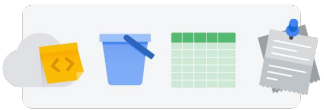
CapEx: Large, upfront investments in fixed assets, which may require lead time to procure, provision, and configure.

OpEx: Pay-as-you-go expenses for services used. There are no upfront costs, but spending is variable; scaling up or down based on real-time usage.

Shifting from CapEx to an OpEx cloud model frees teams from the burden of buying and maintaining physical hardware, allowing for agility and speed.

Google Cloud resource hierarchy

Google Cloud's resource hierarchy organizes assets into four key levels—Resources, Projects, Folders, and the Organization node—which collectively act as a structural blueprint for your cloud footprint.

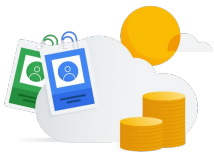


Benefits of this resource hierarchy include:

- **Granular access control:** Precise permission mapping at every level.
- **Inheritance and propagation rules:** Reduces the need to manually configure individual resources.
- **Security and compliance:** Least-privilege enforcement minimizes risk exposure.
- **Strong visibility and auditing capabilities:** Centralized tracking simplifies compliance reviews.

Cloud financial governance best practices

Effective cloud financial governance requires a collaborative, cross-functional partnership between people, processes, and technology.



People: Identify who manages cloud costs.

Processes: Establish a routine so usage is monitored.

Technology: Use Google Cloud's tools to help gain visibility, automate cost control, and receive recommendations for optimizing resources.

Controlling cloud consumption

Google Cloud provides several tools to help you control consumption:

Resource quota policies: Prevent overspending by limiting resource consumption.

Budget threshold rules: Early warning alerts for cost overruns.

Cloud Billing reports: Reactive tracking and data analysis.

Dynamic Workload Scheduler: Cost-efficient scheduling for nonurgent workloads.

Spot VMs: Deeply discounted, preemptible virtual machines.

Committed use discounts (CUDs): Cheaper rates for predictable long-term usage.

Google Cloud

Scaling with Google Cloud Operations

Key cloud operations terms

- **Operational excellence:** The execution of a business strategy more consistently and reliably than the competition, often through automation and continuous improvement.
- **Reliability:** The probability that a system will perform its intended function without failure for a specified period of time.
- **High availability:** A system design that ensures a prearranged level of operational performance (uptime) is met during a specified period.

Modernizing with Google Cloud Observability

Cloud Monitoring

Track system performance and alert on resource health.

Cloud Logging

Centralize log data for fast incident troubleshooting.

Cloud Trace

Analyze application latency to pinpoint bottlenecks quickly.

Cloud Profiler

Optimize code execution to reduce CPU consumption.

Error Reporting

Aggregate and track application crashes for developers.

Developer operations (DevOps) and site reliability engineering (SRE)

DevOps is a collaborative approach that unites development and operations teams to accelerate and improve software delivery. SRE serves as a specific practice within this framework to guarantee the reliability and efficiency of cloud-deployed systems.

Three main concepts of SRE

Service level indicators (SLIs): Quantitative measurements of how a service is performing. Examples include latency, error rate, and system uptime.

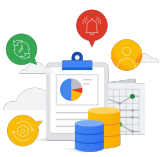
Service level objectives (SLOs): The specific reliability target set for a service level indicator. It defines the goal, such as aiming for the system to be available for 99.9% of the month.

Service level agreements (SLAs): A binding contract between a service provider and a customer. It guarantees specific SLOs and outlines the penalties if the provider fails to meet them.

Measuring performance and reliability

System performance and reliability are measured using the four golden signals: Latency tracks the time it takes to service a request, while traffic measures the overall demand. Saturation monitors the utilization of resources like CPU, while errors measure the rate of failed requests, so system operates successfully.

Designing resilient infrastructure



- **Redundancy:** Eliminate single points of failure to prevent downtime.
- **Replication:** Ensure real-time data copy availability across multiple regions.
- **Scalable infrastructure:** Adapt to shifting demand by dynamically provisioning resources.
- **Backups:** Secure historical data for recovery during a disaster.