CHAPTER 2

Overview of Cloud Computing and Google Cloud

In this chapter, we'll cover

- Cloud computing basics
- How Google Cloud differs from other cloud platforms
- The various Google Cloud products and services of relevance to the cloud architect
- The various ways you can access and manage your cloud
- The business and technical context of cloud architecture

The development of the Internet has been one of the most transformational events in the history of human civilization, especially with regard to social evolution. In more than two million years, there has never been a point at which we humans have been able to communicate globally and in real time to share knowledge and data that can be digested by virtually anyone who wants to consume it. The Internet has eliminated nearly all communication barriers and has "open-sourced" the availability of knowledge.

We are no longer forced to rely solely on potentially biased sources of information to gain knowledge. Before the Internet, education, media, and everyday speech were very much predisposed to bias because of the extremely limited access to and availability of sources of information that could provide accurate knowledge and help build perspective. Before the Internet, if you were a curious "free thinker" with access to books, funding, and time, you could dive into as many opportunities as possible to become an expert. But that was not so easy for the average person. Although biases still exist on the Internet, nowadays you can do a quick Google search and find hundreds of perspectives that can help you formulate a more rational understanding of virtually any topic.

The Internet has also provided a blank canvas of infinite possibilities, enabling humans to conduct transactions, connect and develop relationships, and solve problems, all without being physically present. And the engineers of the world continue to build new possibilities. Although the Internet has given humans the ability to analyze data at scale, as the scale has continued to grow, we've encountered many performance bottlenecks along the way that hindered our ability to progress at the pace we wanted. Today, cloud computing provides massive-scale computing resources to nearly everyone, enabling petabytes of data to be analyzed and surfaced back to the end user in microseconds.

Cloud computing enables innovation and on-demand growth that is unconstrained by the resource bottlenecks of traditional data centers that powered the world since the inception of the Internet. But the cloud has also presented many new conflicts for the world. With massive-scale computing, we've been able to develop algorithms that provide us with new ethical challenges. We've trusted corporate entities with enormous troves of our personal data, which they correlate with petabyte-sized databases to build digital profiles of our individual identities and activities. We now have machine-learning algorithms that can mimic our voices, simply by listening to us speak a single sentence. Deepfake algorithms can record and render our faces into artificial scenes that look entirely real. Hackers can analyze data at light speeds to identify vulnerable users, and then use these massive computing resources to attack one user or millions of users in an instant. How can we maintain the security, privacy, neutrality, and transparency of data stored on the Internet considering all of the constantly emerging technologies that represent only the tip of the iceberg of possibility?

Although these philosophical meanderings are not part of the Professional Cloud Architect exam, as certified Cloud Architects, we should understand the philosophical impacts of this technology, especially with regard to how we can build ethically and protect ourselves and the world against unethical activities. It's important that we're purposeful in the way we develop solutions and that we understand how our work can ensure a safer future for our children. We must build solutions that are representative of goodness and morality across all walks of life.

As I'm writing this book, we are undergoing a significant historical event in American society with regard to the acknowledgment of the inequalities faced by African Americans and people of color. The serious consequences of the situation are widespread; we're also experiencing inequalities created by technology with regard to the enormous amounts of data collected about individuals. Consider facial recognition technologies, for example. Amazon recently announced a yearlong moratorium on police use of its facial recognition tool (Amazon Rekognition) because the tool has been unable to properly identify African Americans. The Rekognition API has wrongfully tagged innocent African Americans as criminals because of a lack of proper police training on the algorithm and because of many other built-in blind spots inadvertently included by developers when they built the tool. In addition, IBM announced it would no longer offer, develop, or research facial recognition technology for many of the same reasons.

Despite these issues, it's immature to assume that because these big-tech companies are not investing in facial recognition software, it doesn't exist anywhere outside that space. In fact, the open source community has already provided many free and open solutions that do the same thing. No matter how much we may think that the progression of technology is in the hands of big tech, it turns out that it's not necessarily so; it's actually in the hands of the many technologists who are advancing this profession at work or in their off hours. Technology will move at the pace it's going to move at, unhindered by world governments. Our job is to ensure that we minimize the side effects and promote goodness.

Facial recognition tools are also used for good ends, such as to identify and rescue human-trafficking victims, to secure access to our homes, and to simplify our lives. But we need to be constantly aware of all of the implications of using this technology to ensure that we have developed rationality ourselves and to ensure that our diverse set of voices is heard beyond the technical solutions we develop.

It's important that you think about the philosophy of your work and take pride in shaping the future of the world—and that you use that power for good. Share your ideas with your peers, ask for help and guidance, and ask others to assess your blind spots. In the last 30 years, we've seen technology change the fabric of society at an exponential speed. Think about how you want the next 30 years to look for yourself and for society and how that reflects in the work you do to give back to the world.

Overview of Cloud Computing

Cloud computing is the on-demand availability of computing services over the Internet that offers faster innovation, flexible resources, and economies of scale. This includes providing servers, storage, databases, networking, software, analytics, and business intelligence to businesses and consumers alike, without requiring these users to maintain any physical infrastructure. In the traditional world, where we used to think of three-tier architectures, cloud computing has eliminated all traditional computing paradigms and brought forward the notion of service-based architectures in the cloud, where the myriad computing activities performed are broken down into the most miniscule services and are decoupled from the former monolithic computing model. With microservices architecture, deployment gets simplified as the functionality is separated into smaller units. There is more freedom to use various languages and frameworks in development, there are fewer dependencies between teams, and it's easier to design for reliability. Google has been doing microservices-based architecture prior to the advent of the cloud, notably with Borg, the company's internal corporate cluster manager that led to the creation of Docker and Kubernetes.

In the cloud, customers and cloud service providers operate using fundamentally different responsibility models. In traditional computing environments, a business typically would own the entirety of a data center environment and all of the labor associated with it, or it would be contracting with managed data center companies and purchasing equipment to be stored in the contractor's environment, paying them to perform physical management and some level of logical management. Figure 2-1 shows the Google Cloud shared services model.

In this shared responsibility model, the cloud offers a few new concepts, namely, *Infrastructure as a Service* (IaaS), *Platform as a Service* (PaaS), *Software as a Service* (SaaS), and as a bonus, serverless, or *Function as a Service* (FaaS), a term used these days to describe fully serverless environments.

IaaS provides the most flexible cloud computing model and enables you to retain the most control over your infrastructure. Google Compute Engine is an example. With an IaaS solution, you can deploy virtual machine environments onto servers, giving you full ownership of the infrastructure end-to-end, without having to manage the physical

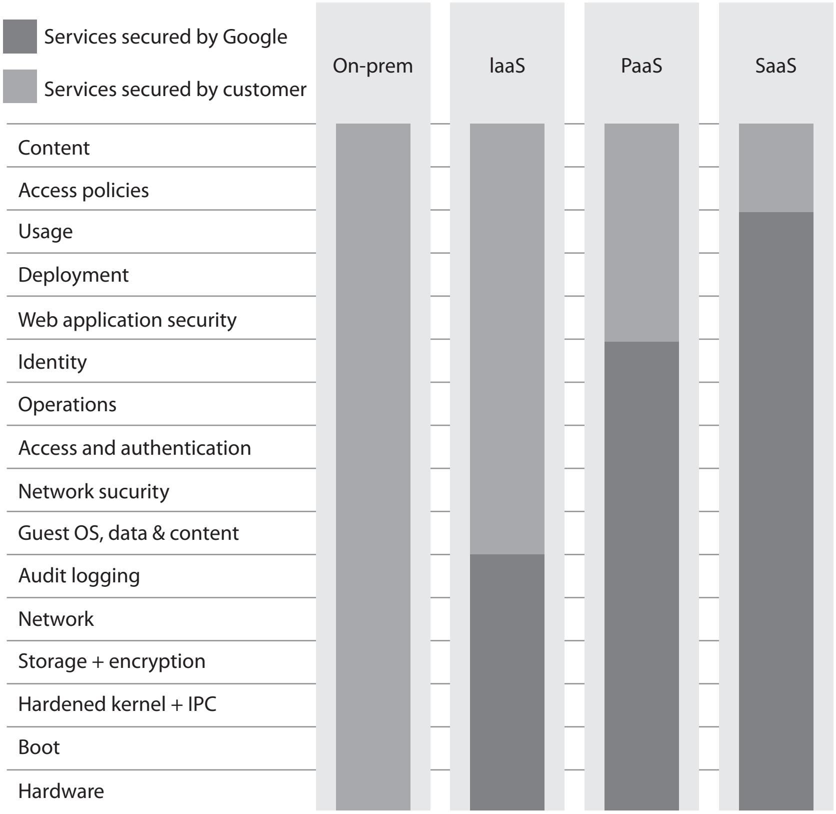

Figure 2-1 The Google Cloud shared services model

servers yourself. IaaS does have limitations as well; for example, you need more overhead to manage your resources, since you retain a lot of control. That inherently means you have a lot more responsibility for the security of your environments.

The PaaS model offers a simple, cost-effective solution to developing and deploying applications on a scalable and highly available platform. PaaS typically offers development teams a lot more speed for application deployment, and because it typically ramps up and down based on usage, a PaaS solution can be more cost-effective. Google App Engine is an example; however, while developer expertise has evolved, App Engine is no longer the default model, and the current trend for consuming cloud computing by application developers is to use Google Kubernetes Engine (GKE), which falls somewhere in the middle between PaaS and IaaS. Some of the limitations of PaaS concern data security, since the cloud service provider controls the underlying infrastructure; vendor lock-in (although this is not an issue with GKE); operational limitations; and a lack of full developer flexibility.

The SaaS model is the most familiar to everyone, in which an application is delivered over the Internet through a web browser, without the need to download or install anything on the client side. SaaS solutions are advantageous for software that is designed to perform a general set of tasks and to disallow a developer end user to customize or modify the application. Examples of SaaS are Salesforce, Intuit QuickBooks, and Google Workspace. There are several limitations to SaaS: these solutions are designed to solve only certain use cases and are not designed as solutions that enable developers the freedom to build. Google Drive, for example, is a SaaS solution designed purely for file storage on the Internet. End users don't have much control here beyond the service catalog of options Google Drive provides. Major implications of SaaS are vendor lock-in, lack of interoperability, lack of control and customization, and concerns about data security.

With a move to the cloud, businesses are typically concerned about the following:

- Reducing capital expenditures and turning them into operational expenditures
- Scaling resource expenses to actual end-user demand rather than initial projected demand, significantly reducing program risk
- Eliminating or transitioning much of the overhead of IT resources and letting companies focus on building great products and experiences for their customers
- Accelerating the pace of innovation to achieve a competitive edge or unlocking new business opportunities

If you're not dabbling in the cloud by 2020, that's pretty much a negative. Companies that have full, on-premises environments are looked down upon as not being "technically savvy" (minus companies such as Facebook, which has its own "cloud"). Most talented engineers want to focus on solving complex business problems and don't want to worry about the mundane and often painful tasks required to build and operate infrastructure. As such, the best talent these days doesn't want to work at traditional enterprises.

Lastly, you'll need to know the ins and outs of the following key cloud concepts for the exam:

- A *public cloud* offers cloud services to the public; think of the big providers such as Google Cloud, Amazon Web Services (AWS), Microsoft Azure, and Alibaba Cloud.
- A *private cloud* is developed in-house, specific to a business. Think of the Facebook Cloud, which Facebook had built for its own use. Or think of an on-premises cloud, where all servers, storage, and networks are dedicated to the company and hosted in a dedicated data center.
- A *hybrid cloud* uses a mixture of a public cloud and a private cloud to create a more diverse environment. Customers with highly sensitive data, such as financial services or healthcare, like to have their own environments under their full control, where they can store their most highly sensitive workloads and data. Or maybe the customer hasn't gotten around to migrating fully to a public cloud because of a lack of commitment, features, time, and so on.

- A *multi-cloud* is similar to a hybrid cloud, in that multiple clouds are at play, but typically in a multi-cloud environment, a business will be using multiple *public* clouds. Many top companies today use this model, such as Snapchat, which uses both Google Cloud and AWS for its workloads.
- A *community cloud* (which is not covered in the exam) is a collaborative effort in which infrastructure is shared between several organizations from a specific community with common concerns. It may be managed internally or by a third party, and can be hosted internally or externally.

Google Cloud vs. Other Clouds

Although several other cloud service providers focus on the whole spectrum of offerings, from IaaS to highly specialized PaaS or SaaS offerings, the key players in this space globally are Amazon Web Services (AWS), Google Cloud, Alibaba Cloud, and Microsoft Azure.

In 2018, Alphabet (Google's holding company) disclosed the revenues of Google Cloud at $1 billion per quarter. In its Q1 2020 earnings call, Google Cloud reported that it generated $2.78 billion in revenue—that's 52 percent higher than one year prior and almost triple its growth since 2018. Although AWS is dominating the market at $10 billion per quarter and Microsoft Azure is inherently present across enterprises because of Microsoft's global dominance in computing, Google Cloud is continuing to grow at a rapid pace as an alternative cloud service provider with several advantages. A major trend today within many companies is to employ a multi-cloud strategy. Given the tenant isolation design advantage of designing public cloud solutions from traditional on-premises architectures, different workloads can and should be enabled to run on the most suitable public cloud platform to achieve the desired business outcome. Google Cloud Platform is gaining traction into every major enterprise globally, signing big names like Equifax, Home Depot, McKesson, Disney, Snap, Salesforce, PayPal, and HSBC. What is it about Google Cloud that is becoming so appealing to these massive enterprises?

Although Amazon built the AWS platform from scratch, Google had already designed a massive global computing platform and networking backbone that served all of its employees and users worldwide. In a way, Google used its existing internal infrastructure to develop a new layer of abstraction to externalize it to customers worldwide—the Google Cloud Platform (GCP). Four key core competencies and a set of principles based on system design provide the framework for GCP's design.

Security First

Security and data protection is at the core of Google and its products. As a customer of the Google Cloud, you own your data and control how it is used. GCP also has strong internal controls and auditing features that protect customers against insider access to their data. It offers continuous security monitoring and several security features as part of its shared responsibility model, providing its customers confidence that their businesses are safe from malicious activities. Lastly, one of the major benefits of using GCP is that it is built from Google's already existing private global backbone, so Google is able to encrypt data at rest and in transit by default. Google controls the majority of the service delivery, and thus the user experience, all from within its own infrastructure.

Open Cloud

It's become increasingly evident throughout the evolution of the cloud that customers don't want to lock themselves into one cloud provider; instead, they often use the strengths of various cloud providers for different aspects of their business and for business continuity. Google offers an open cloud that enables customers to leverage multiple different clouds and follow a common development and operations approach to deliver their applications. Customers can innovate, build, and scale rapidly while minimizing the constraints of a single technology. Google strongly emphasizes this approach because of its pioneering efforts in building and cultivating the open source community, which is a key element of its corporate philosophy.

NOTE Check out Google's corporate philosophy to read the "Ten things we know to be true," which have held true since the company's beginning: https://www.google.com/about/philosophy.html.

Analytics and Artificial Intelligence

Analytics and artificial intelligence (AI) remain two of Google's strong points as a company that is heavily data driven. GCP offers fully managed, serverless analytics products and services that eliminate the constraints of scale, performance, and cost. GCP empowers customers to leverage real-time insights, enabling them to improve their decision-making and accelerate innovation—all without having to manage any infrastructure. Google has historically been on the forefront of researching and improving AI, offering innovations such as MapReduce, Dremel, Apache Beam, and TensorFlow, which both customers and Google can use to power its products with more AI capabilities. Simply stated, Google provides superior analytics and AI products as part of the great selling points for customers migrating to GCP.

Global Data Centers and Network

As mentioned, Google Cloud was built on the same infrastructure that Google uses to serve more than 100,000 employees and billions of consumers worldwide. This massive private network consists of more than 24 regions, 73 zones, and 144 network edge locations and is available in more than 200 countries and territories. This is arguably the largest and most advanced software-defined network, delivering the highest level of performance and availability in a secure and sustainable way. This global backbone has been tested and vetted with billions of users worldwide, using all of Google's products and internal technology. Building a cloud on top of this backbone only makes sense, especially when reliability is increasingly one of the more important performance indicators for successful businesses. Talk to someone on the presales side at Google Cloud, and they will talk your ears off about the Google network.

Principles of System Design

Google Cloud follows a very strict framework that enables them to build robust, secure, and scalable systems. Four principles provide guidance on designing systems for internal users and customers, as outlined next.

Operational Excellence

Operational excellence is the principle of building a foundation that successfully enables reliability across your infrastructure by efficiently running, managing, and monitoring systems that deliver business value. Three key strategies drive this principle:

- *Automating build, test, and deployment by using continuous integration and continuous deployment pipelines.* This enables customers to programmatically do rapid deployment and iterations based on a continuous feedback loop.
- *Monitoring business objective metrics by defining, measuring, and alerting on key metrics.* Data needs to be measured and output to your business leaders to give insight into where they have the competitive edge and where they can further optimize or reassess.
- *Conducting disaster recovery testing proactively and periodically.* Disaster can strike a company in so many different ways, often causing financial and reputational business harm. This is overlooked too many times by customers until disaster strikes and ends up costing them exponentially more than it would cost had they been prepared.

Security, Privacy, and Compliance

It is critical for any customer doing business in the cloud to ensure their intellectual property is protected and their customers are safe from malicious activity. This is a core principle of Google's system design. Four key strategies drive this principle:

- *Implementing least privileges with identity and authorization controls.* Centralizing your identity management system and designing your access management structure in a way that allows users to do only what they're intended to do, while ensuring *nonrepudiation* (a user cannot deny their activity) and audit logs that are available to be consumed by automated and manual detection mechanisms.
- *Building a layered security approach.* Also known as defense-in-depth, this involves the implementation of a variety of security controls at each level of the infrastructure and applications designed on top of the infrastructure. The idea is to assume that any security control can be breached, and when it is breached, several other layers of defense are available to protect intellectual property.
- *Automating deployment of sensitive tasks.* Humans continue to be the weakest link in performance and security of administrative tasks. By automating the deployment of these tasks, you can eliminate the dependency on humans.

- *Implementing security monitoring.* Part of a strong security model is to prevent, detect, and respond to malicious activity. By implementing automated tools to monitor your infrastructure, you can gather data to continue protecting your weak points and prevent malicious activity from occurring in your environment and harming your business.

Reliability

Google sees reliability as the most important feature of any application. Without reliability, users begin to churn (stop using the product). Google suggests 15 strategies to achieve reliability; here are three key ones:

- *Reliability is defined by the user*. Many data points can encompass all sorts of important factors in your workload, but truly measuring using key performance indicators (KPIs) requires an understanding of user actions, and the metrics define the success of those actions for reliability.
- *Use sufficient reliability*. There's no need to overinvest in reliability if you're meeting user satisfaction. Figure out what sort of availability keeps your users happy and retained, and ensure that you continually assess reliability as your infrastructure grows.
- *Create redundancy*. Always assume that if you depend on a single point to provide a function, that point can and will fail someday. When building your infrastructure and applications, always try to leverage resource redundancy across resources that can fail independently.

Performance and Cost Optimization

Managing the performance of your applications and the associated costs is a balancing act, as highly performant environments often end up costing more to maintain. Understanding where you've met your minimum performance requirements and where you need to optimize cost is an important principle for system design. These three strategies are relevant here:

- *Evaluate performance requirements*. Determine the minimum performance you need from your applications.
- *Use scalable design patterns*. Leverage automatically scaling products and services where applicable to minimize cost to what is necessary.
- *Identify and implement cost-saving approaches*. Understand the priority of each of your services with respect to its application to your business objectives. Use these priorities to optimize for service availability and cost.

A 10,000-Foot Overview of GCP

We're going to dive into a bit of an overview of GCP to help you understand the overall elements of the cloud, including several of the products and services you'll need to know about for the exam. It's a lot of content, so don't worry about memorizing everything right now; we'll be diving into these concepts and services to a greater depth throughout the book. It'll be good to get some initial exposure, so that the next time you read about these ideas in the book, you'll be able to memorize their salient points.

You can always refer back to this discussion if you need to do some quick memorization exercises when preparing for your exam. You'll notice that I've mentioned AWS comparisons where I could—if you are familiar with AWS, this may be beneficial for you. If you are not familiar with AWS, you're getting free multi-cloud knowledge. You can leave a tip at the door! It's important that a Google Cloud Architect understand multi-cloud deployments, so there's no reason to avoid discussing other major clouds in this book.

NOTE I need to shout out to Ryan Kroonenberg from A Cloud Guru for introducing me to this 10,000-foot overview concept; it's brilliant!

Lastly, there are hundreds and hundreds of concepts, products, and services in GCP—we'll cover those that you're most likely going to be tested on in the exam. As a Google Cloud Architect, you should go beyond the scope of this discussion in your job and familiarize yourself with the entire GCP portfolio.

EXAM TIP When you're done with this book, come back here and review this entire section to ensure that you know the function of each and every product outlined here. Remember that the exam will ask you for keywords that may differentiate similar solutions. For example, think about the various database types and get to know the differences between them as well as when to use each technology.

Compute Solutions

GCP includes various computing- and application-level offerings.

Google Compute Engine

Google Compute Engine (GCE) is an IaaS solution that enables users to launch virtual machines (VMs) on demand. With GCE, users manage the entire underlying infrastructure associated with the VM instances, including the machine types. VMs can be launched on predefined or custom machine sizes. GCE supports live migration, OS patch management, preemptible VMs (PVMs), and more. It is similar to Amazon Elastic Compute Cloud (EC2).

Preemptible Virtual Machine

Preemptible virtual machines (PVMs) are low-cost, short-term instances that are intended to run batch jobs and fault-tolerant workloads on Compute Engine. They offer significant cost savings, typically up to 80 percent, while still offering the same performance and capabilities of regular VMs. It is similar to Amazon EC2 Spot Instances.

Google App Engine

Google App Engine (GAE) is a PaaS solution that offers a fully managed, serverless application platform for building and deploying applications, without users having to manage the underlying infrastructure. With no server management and no configuration deployments, developers can focus on building applications. GAE supports popular development languages such as Go, Ruby, PHP, Java, Node.js, Python, C#, and .NET Framework, and you can bring your own language runtimes and frameworks. It is similar to AWS Elastic Beanstalk.

Google Kubernetes Engine

Google Kubernetes Engine (GKE) is a PaaS solution that offers a secure managed Kubernetes (K8s) service. GKE offers enterprise-ready containerized solutions with prebuilt deployment templates, enabling customers to ensure portability, with simplified licensing and consolidated billing. GKE is the direction that most modern enterprises and cloud-natives are heading, and although you may not encounter much about it on the exam, it's very important for the modern Google Cloud Architect to learn. It is similar to Amazon Elastic Kubernetes Service (EKS).

Cloud Run

Cloud Run is a PaaS solution that offers a fully managed compute platform for deploying and scaling containerized applications. Cloud Run eliminates infrastructure management and is able to scale up and down on demand, charging only for the exact resources used. It supports any language, library, or binary and is built upon the open standard Knative. It is similar to AWS Fargate.

Cloud Functions

Cloud Functions is a FaaS offering and is an event-driven, serverless computing platform. With Cloud Functions, you can run your code locally or in the cloud without having to provision any servers. It scales up or down on demand, so it is cost-effective, and you pay only for what you use. Developers can write code, and Google Cloud does the rest. It is similar to AWS Lambda.

EXAM TIP Take a look at the actual Google Cloud products website and skim through it, check out some blog posts of people who've recently passed the exam, and gather as much data as you can. The exam is notorious for asking questions that are related to the certification subject matter but are not specific to GCP.

Storage Solutions

Various storage offerings are available on GCP.

Google Cloud Storage

Google Cloud Storage (GCS) is a globally unified, scalable, and highly durable object storage offering. It offers object life cycle management to move your data automatically to lower-cost storage classes based on criteria you define to optimize your cost. GCS is often used for content delivery, data lakes, and backup. It offers varying service level agreement (SLA) availability levels depending on the storage class, ranging from 99.0 to 99.95 percent. It is similar to Amazon Simple Storage Service (S3).

Cloud Filestore

Cloud Filestore provides high-performance, managed file storage for applications that require a file system. Like the Network File System (NFS) protocol, Filestore offers the ability to stand up a network-attached storage on your GCE or GKE instances. Filestore is highly consistent, fast, fully managed, and scalable using Elastifile to grow or shrink your clusters. Filestore offers a 99.9 percent SLA availability level. It is similar to Amazon Elastic File System (EFS).

Persistent Disk

Persistent Disk (PD) provides high-performance, durable block storage for solid-state drive (SSD) and hard disk drive (HDD) devices, which can be attached to GCE or GKE instances. Storage volumes can be resized and backed up and support simultaneous reads. It is similar to Amazon Elastic Block Store (EBS).

Local SSD

Local solid-state drives (SSDs) are high-performance, ephemeral block storage disks that are physically attached to the servers that host your VM instances. They offer superior performance, high input/output operations per second (IOPS), and ultra–low latency compared to other block storage options. They are typically used for temporary storage use cases such as caching or scratch processing space—think of workloads such as high-performance computing (HPC), media rendering, and data analytics. It is similar to Amazon EC2 SSD-based instance store volumes.

Database Solutions

Various database offerings are provided on GCP.

Cloud Bigtable

Cloud Bigtable is a fully managed and scalable NoSQL database for large analytical and operational workloads. It's able to handle millions of requests per second at a consistent sub-10ms latency. Bigtable is ideal for things like personalization engines, advertising technology (ad-tech), digital media, and Internet of Things (IoT), and it connects easily to other database services such as BigQuery and the Apache ecosystem. Bigtable offers a 99.99 percent SLA availability level. It is similar to Amazon DynamoDB.

Cloud SQL

Cloud SQL is a fully managed relational database for MySQL, PostgreSQL, and SQL Server, offering a simple integration from just about any application such as GCE, GKE, or GAE. You can use BigQuery to directly query your Cloud SQL databases. CloudSQL offers a 99.95 percent SLA availability level. It is similar to the Amazon Relational Database Service (RDS).

Cloud Spanner

Cloud Spanner is a fully managed, scalable, relational database for regionally and globally distributed application data. It offers the benefits of a relational database structure while scaling horizontally like a nonrelational database, allowing for strong consistency across rows, regions, and contents with a 99.999 percent SLA availability level. Cloud Spanner solved a major issue with traditional databases by eliminating the trade-off between scale and consistency with its horizontally scaling, low latency, and highly consistent characteristics. Cloud Spanner is similar to Amazon Aurora, but Aurora's biggest benefit is performance over RDS and MySQL/PostgreSQL compatibility. Cloud Spanner promises a high-performance, globally distributed RDBMS, which is not MySQL/PostgreSQL compatible.

Cloud Firestore

Cloud Firestore is a fully managed, fast, serverless, cloud-native NoSQL document database that is designed for mobile, web, and IoT applications at global scale. Firestore is the next generation of Datastore, which was the original highly scalable NoSQL database for mobile and web-based applications. Firestore offers a 99.999 percent SLA availability level. It is similar to Amazon DynamoDB. The key differentiator between Firestore and Bigtable is that Firestore is designed for mobile applications and Bigtable is designed for analytical workloads.

Memorystore

Memorystore is a scalable, secure, and highly available in-memory service for Redis and Memcached. It enables you to build application caches that provide sub-millisecond data access, and it's entirely compatible with open source Redis and Memcached. Memorystore provides a 99.9 percent SLA availability level. It is similar to Amazon ElastiCache.

Data Analytics

Various data analytics offerings are available on GCP.

BigQuery

BigQuery is a highly scalable, cost-effective serverless solution for data warehousing in the cloud. It enables you to analyze petabyte-scale data with zero operational overhead. BigQuery is one of Google Cloud's top products and is based on the Dremel query engine that Google developed. It has a 99.9 percent SLA availability level. There are no direct comparisons with AWS products, because BigQuery is an industry leader and is in a class of its own.

Dataproc

Dataproc is a fully managed data and analytics processing solution based on open source tools. You can build fully managed Apache Spark, Apache Hadoop, Presto, and other open source clusters. A very cost-effective solution, Dataproc is pay as you go and offers per-second pricing. It is similar to Amazon Elastic MapReduce and AWS Batch.

Dataflow

Dataflow is a serverless, cost-effective, unified stream and batch data processing service that is fully managed and supports the Apache Beam SDK and runs on a system of workers and jobs. If you see a question about Apache Beam on the exam, look for an answer that refers to Dataflow. It is similar to AWS Batch and Amazon Kinesis.

Pub/Sub

Pub/Sub is a global messaging and event ingestion solution that provides you a simple and reliable staging location for your event-based data before it gets processed, stored, and analyzed. Pub/Sub offers at-least-once delivery, exactly once processing, no provisioning, and is global by default. Pub/Sub offers a 99.95 percent SLA availability level. It is similar to Amazon Simple Queue Service (SQS), Amazon Simple Notification Service (SNS), and Amazon Kinesis.

Cloud Composer

Cloud Composer is a fully managed workflow orchestration service built on Apache Airflow that simplifies orchestration and empowers you to author, schedule, and monitor pipelines across clouds and on-premises environments. It is similar to AWS Data Pipeline and AWS Glue.

Networking Solutions

Various networking offerings are available on GCP.

Global Resources

Global resources can be accessed in any zone within the same project. These resources include such things as images, snapshots, Virtual Private Cloud (VPC) networks, firewalls, and their associated routes.

Region

Regions are independent geographic areas that contain multiple zones (or data centers). Regional resources offer redundancy by being deployed across multiple zones within a region. Some services, such as Datastore, BigQuery, Bigtable, and Cloud Storage, are distributed within and across regions—known as multiregional deployments.

Zone

Zones are deployment areas for resources within a region. One zone is typically a data center within a region and should be considered as a single failure domain. In fault-tolerant application deployments, the best practice is to deploy applications across multiple zones within a region, and ideally to deploy across multiple regions. If a zone becomes unavailable, all of the zonal resources will be unavailable until the services are restored.

Virtual Private Cloud

A virtual private cloud (VPC) is a virtual network that provides connectivity for resources within a project. Projects can contain multiple VPC networks, and by default new projects start with a default auto-mode VPC network that also includes one subnet in each region. Custom-mode VPC networks start with no subnet. VPC networks are global resources and are not associated with any particular region or zone.

Subnet

Subnets, or subnetworks, are logical partitions within a VPC network with one primary IP range and zero or more secondary IP ranges. Subnets are regional resources, and each subnet defines a range of IP addresses. You can create more than one subnet per region. When an auto-mode VPC network is created, one subnet from each region is automatically created within it using predefined IP ranges. When a custom-mode VPC network is created, no subnets are automatically created, giving you complete control over the subnets and IP ranges. Custom-mode VPC networks are better suited for enterprises and production environments.

Shared VPC

A Shared VPC network enables an organization to connect resources from multiple projects to the same VPC. This enables project resources to communicate securely using internal IP addressing from that network. In the Shared VPC model, you designate one project as a host project and attach one or more services projects to it. A shared VPC is referred to as "XPN" in the console and CLI.

Cloud DNS

Cloud DNS offers a reliable, resilient, low-latency authoritative Domain Name System (DNS) service that guarantees 100-percent availability. It provides automatic scaling, enabling users to create and update millions of DNS records. Cloud DNS is a simple and very cost-effective solution to individuals who host their own DNS servers or leverage other third-party DNS providers. It is similar to Amazon Route 53.

VPC Flow Logs

VPC Flow Logs are used for network monitoring, forensics, security analysis, and cost optimization. These logs provide a sample of network flows sent and received by VM instances or GKE nodes within a network. VPC Flow Logs can be very expensive to use, so it is not recommended to leave them on indefinitely.

Firewall

When deploying a VPC, you can use firewall rules to allow or deny connections to and from your application instances based on the rules you deploy. Each firewall rule can apply to ingress or egress connections, but not both. Rules are enforced at the instance level, but the configuration is associated with the VPC network—so you cannot share firewall rules among VPC networks, including peered networks. VPC firewall rules are stateful. Once a session has been established, firewall rules allow bidirectional traffic. It is similar to security groups in AWS.

Cloud Content Delivery Network

Cloud Content Delivery Network (CDN) is a fast, reliable web and video content delivery network with global scale and reach. It provides edge caches, known as points of presence (PoPs), that are peered with nearly every major Internet service provider (ISP) worldwide, and it uses the Anycast architecture to provide a single global IP address for global distribution. CDN leverages Google's proprietary fiber-optic backbone to carry network traffic globally.

Cloud Load Balancing

Google Cloud Load Balancer (GCLB) offers a fully distributed, high-performance, scalable load balancing service across GCP, with a variety of load balancer options. With GCLB, you get a single Anycast IP that fronts all your backend instances across the globe, including multiregion failover. In addition, software-defined load balancing services enable you to apply load balancing to your HTTP(S), TCP/SSL, and UDP traffic. You can also terminate your SSL traffic with an SSL proxy and HTTPS load balancing. Internal load balancing enables you to build highly available internal services for your internal instances without requiring any load balancers to be exposed to the Internet.

Cloud NAT

Cloud NAT is GCP's managed network address translation service that enables users to provision application instances without public IPs and to access the Internet for updates, patching, configuration management, and more. It does not allow outside resources to access any of the private instances behind the NAT gateway. Cloud NAT works with both GCE and GKE and offers regional high availability. It is similar to a NAT Gateway on AWS.

Cloud VPN

Cloud VPN enables users to connect their on-premises environment or other public cloud networks to their VPC networks securely over an encrypted IPSec virtual private network (VPN) tunnel for data bandwidth needs up to 3.0 Gbps. This is useful for low-volume data connections. It offers an incredible 99.99 percent availability. One of the newer features supported by Cloud VPN is the support of multiple tunnels; you can use this functionality to augment data bandwidth beyond 3.0 Gbps. It is similar to AWS Client VPN.

Cloud Interconnect

Cloud Interconnect offers an enterprise-grade connection to your VPC networks via either a Dedicated Interconnect or a Partner Interconnect. Using a Dedicated Interconnect, you can deploy a connection directly to a Google edge network, choosing between a 10-Gbps or 100-Gbps pipe. Using a Partner Interconnect, you can deploy a connection to Google through a supported third-party service provider, choosing between a 50-Mbps or 10-Gbps pipe. SLAs vary with regard to the type of connection you select. It is similar to AWS Direct Connect.

Peering

With peering, you can establish a direct connection between your network and Google while cutting egress fees, if you meet the requirements to connect directly with Direct Peering or through a partner with Carrier Peering. The recommended methods for accessing Google Cloud are through a Dedicated Interconnect or Partner Interconnect.

VPC Network Peering

VPC Network Peering enables internal IP address connectivity across two VPC networks, including VPC networks that do not belong to the same project or the same organization. VPC Network Peering enables two VPC networks to communicate internally on Google's software-defined network, and it does not traverse the public Internet. This is advantageous to using external IP addresses or VPNs to connect because it improves network latency and provides network security, as traffic does not get exposed to the Internet. It also minimizes costs; there are no egress costs because traffic communicates using internal IPs.

Private Google Access Options

There are four main access options for privately accessing Google Cloud: Private Google Access, Private Google Access for on-premises hosts, Private Services Access, and Serverless VPC Access. Each access option enables virtual machine instances that have internal IP addresses to access certain APIs and services. This is helpful for scenarios in which you don't want to assign an external IP address for your VM instances to connect to APIs or services outside of your internal network.

Operations Solutions

Various operations offerings are available on GCP.

Cloud Logging

Cloud Logging, previously known as Stackdriver Logging, is a real-time log management and analysis tool that enables you to store, search, analyze, monitor, and alert on log data and events. It allows for ingestion of any custom log data from any source and is a fully managed service. Integration into Cloud Monitoring enables you to define alerts based on certain metrics you select. It is similar to Amazon CloudWatch logs.

Cloud Monitoring

Cloud Monitoring is a full-stack, fully managed monitoring solution that gives you visibility into the performance, uptime, and overall health of your applications. It integrates with AWS out of the box, and it enables you to define custom metrics for key alerts your business is looking to monitor for. It is similar to Amazon CloudWatch monitoring.

Cloud Trace

Cloud Trace is a distributed tracing service that you can use to collect latency from your applications and track how requests propagate through your application. It can provide in-depth latency reports to surface performance issues, and it works across VMs, containers, or GAE projects. It is similar to AWS X-Ray.

Developer Tools

Various developer tools are available to us on GCP.

Cloud SDK

The Cloud SDK is a set of command-line tools and libraries that enable you to interact with Google Cloud products and services directly from the command line. The SDK supports popular languages such as Java, Python, NodeJS, Ruby, Go, .NET Framework, and PHP. The **gcloud** command-line tool is used for interacting with your cloud environment, along with other product-specific command-line tools such as **gsutil** for Cloud Storage, **bq** for BigQuery, and **kubectl** for GKE.

Cloud Source Repositories

Cloud Source Repositories is a private Git repository service that you can use to design, develop, and securely manage your code. It enables you to extend your Git workflow by connecting to other tools such as publish/subscribe (pub/sub) messaging, Cloud Monitoring, Cloud Logging, and more. You can mirror code from GitHub or BitBucket to get powerful code search, browsing, and diagnostic capabilities. You can also use regular expressions to refine your search across the directories.

Container Registry

Container Registry is a private Docker repository that enables you to store, manage, and secure your Docker container images. You can also perform vulnerability analysis and manage access control to the container images. With Container Registry, you can integrate your continuous integration/continuous delivery (CI/CD) pipelines to design fully automated Docker pipelines. It is similar to JFrog Artifactory or Amazon Elastic Container Registry (ECR).

Hybrid Cloud and Multi-Cloud Solutions

Several hybrid cloud and multi-cloud offerings are available on GCP.

Anthos

Anthos is a fairly new offering from Google Cloud. It is Google Cloud's solution to the increasing need for hybrid and multi-cloud PaaS requirements and for preventing vendor lock-in. With Anthos, you can run, manage, and govern applications in a hybrid or multi-cloud environment. Anthos GKE enables you to run enterprise-grade container orchestration and management in cloud and on-premises environments.

With Anthos Config Management, you can govern configuration policies across your environments. Anthos Service Mesh (powered by Istio) is a service mesh architecture that eliminates a lot of networking and traffic routing concerns by leveraging mutual Transport Layer Security (mTLS) to secure your service-to-service or end user–to–service communications, so that your developers can focus on building applications. It also lets you easily make role-based access controls and fine-grained access controls. Anthos Security is a tool that enables you to define and enforce security controls across your environments.

Migration Solutions

Various migration offerings are available on GCP.

Storage Transfer Service

Using Storage Transfer Service, you can complete large-scale online data transfers to your Cloud Storage buckets. Use Google's high-bandwidth network pipes to leverage ultra-high-speed connections to transfer petabyte-scale data—if you have a strong network yourself. For massive scale data transfers, it is advised that you use a transfer appliance. It is similar to AWS DataSync.

Transfer Appliance

Transfer Appliance is a physical device that Google provides in increments of either 100TB or 480TB models that enable you to accelerate the speed at which you transfer data to Google Cloud. It is similar to AWS Snowball.

Security and Identity Solutions

Various security offerings are available on GCP.

Cloud Asset Inventory

Cloud Asset Inventory is a metadata inventory service that enables you to view, monitor, and analyze all of your GCP resources and policies. You can export your entire inventory, analyze changes, build real-time notification when assets are changed, and sift through your resources and identity and access management (IAM) policies. It gives you a deep and detailed view of all the resource metadata and is similar to AWS Config.

Security Command Center

Security Command Center is a security management and data risk platform that provides a straightforward view of your cloud security vulnerabilities, threats, and compliance issues. Security Command Center includes an underlying suite of tools that provide all the logic for its capabilities, some of which come only with the paid premium version. Security Health Analytics does scanning for security misconfigurations and compliance violations. Event Threat Detection does log-based threat detection using Google's threat intelligence engine. Web Security Scanner identifies common web-based vulnerabilities on public-facing endpoints. All detections are surfaced as "Findings" into the Security Command Center dashboard, and all the findings can be exported into a customer's security information and event management (SIEM) platform.

Cloud Audit Logs

Cloud Audit Logs give you visibility into all user activity in your Google Cloud. It provides a full view of all administrative activities, access to data, and a hardened, always-on trail that cannot be disabled. Audit trails are immutable and reside in highly protected storage. You can leverage these logs for incident management and to track user activity for your security operations teams. This is similar to AWS CloudTrail.

VPC Service Controls

VPC Service Controls enable you to define a security perimeter for constraining your managed GCP services such as Cloud Storage, BigQuery, and Bigtable to your VPC network, so that you can ensure that malicious users cannot exfiltrate data in the event of a misconfigured access control or configuration.

Access Transparency

Access Transparency logs are near-real-time logs that show you when a Google administrator accesses your data. Though Cloud Audit Logs provide visibility into the actions of the privileged users in your environment, sometimes Google administrators may need to access your environment (for example, to respond to an outage, or when you opened up a support ticket that required data access). These events are logged as Access Transparency logs.

Cloud Data Loss Prevention

Cloud Data Loss Prevention (DLP) is a fully managed service that minimizes the risk of data exfiltration by enabling you to discover, classify, and protect your sensitive data. With Cloud DLP, you can use de-identification methods with streaming and stored data, and you can also continuously scan for environments where data does not meet your classification requirements.

Cloud Key Management Service

With Cloud Key Management Service (KMS), you can manage your cryptographic keys on Google Cloud. KMS offers the ability to generate and manage the key encryption keys (KEKs) that protect sensitive data by using customer-managed encryption keys (CMEKs). KMS also supports customer-supplied encryption keys, although that service has not seen much development and may be replaced by External Key Manager (EKM), a service that will enable you to store your own supplied encryption keys at a third-party colocation. KMS has integration with Cloud HSM, enabling you the ability to create a key protected by a Federal Information Processing Standards (FIPS) 140-2 Level 3 device.

Cloud HSM

Cloud HSM is a managed, cloud-hosted hardware security module (HSM) that enables you to protect your cryptographic keys in a FIPS 140-2 Level 3–certified HSM. This is critical for financial services customers who need to meet compliance requirements, for example. HSM easily integrates with Cloud KMS, and you pay for what you use.

Interacting with the GCP

You can interact with the GCP in several ways: via the Google Cloud Console, via the command-line interface (CLI), and via client libraries. As a cloud architect, you should have hands-on experience interacting with the platform through all three mechanisms.

Google Cloud Console

You can use a web-based GUI to manage your projects and resources. Within the Cloud Console, you can also access the Cloud Shell, which enables you to manage Google Cloud projects and perform more complex development tasks. You can SSH into your instances directly though the browser. With both a native iOS and Android application, you can perform some functionality on the go.

Command-Line Interface

For users who prefer to work in a terminal environment, the Google Cloud SDK provides the **gcloud** command-line tool that gives you access to a more familiar interface for engineers. You can use **gcloud** for both your development workflow and your GCP resources.

Other important tools to know for the exam (and that we'll cover later in the book) are the **bq**, **gsutil**, and **kubectl** command-line tools. You may see a few questions on the exam regarding syntax of these tools, so it's a great idea to review the reference links included in the "Additional References" section later in this chapter. Here is a quick reference guide:

- **gcloud** is the primary command-line tool to create and manage resources:

  ```
  gcloud GROUP | COMMAND
  ```

- **bq** is the BigQuery command-line tool. You probably won't see questions about this on the exam, but remember that **bq** = BigQuery.

  ```
  bq --global_flag argument bq_command --command-specific_flag argument
  ```

- **cbt** is the Cloud Bigtable command-line tool. You probably won't see questions about this on the exam, but remember that **cbt** = Cloud Bigtable.

  ```
  cbt [-<option> <option-argument>] <command> <required-argument>
  [optional-argument]
  ```

- **gsutil** is the Google Cloud Storage command-line tool:

  ```
  gsutil [command] [OPTIONS] [BUCKET_NAME]/[OBJECT_NAME]
  ```

- **kubectl** is the Kubernetes command-line tool:

  ```
  kubectl [command] [TYPE] [NAME] [flags]
  ```

EXAM TIP You may see a question about how to use the command line to make a change to a GCS bucket. If you have four answers, two of which start with **bq** and two of which start with **gs**, you can immediately eliminate two of those answers from the question, as **bq** would not be the correct syntax to use here.

Exercise 2-1: CLI Example

In this exercise we're going to use the **gcloud** command-line tool to perform pub/sub messaging operations. We'll create a topic, subscribe to the topic, publish a message, and receive the message.

NOTE Before you begin, take a look at the "gcloud pubsub" section in the gcloud reference guide at https://cloud.google.com/sdk/gcloud/reference/pubsub.

Syntax:

```
gcloud pubsub GROUP [GCLOUD_WIDE_FLAG …]
```

1. Initialize the Cloud SDK:
   ```
   gcloud init
   ```
2. Create a topic:
   ```
   gcloud pubsub topics create my-topic
   ```
3. Subscribe to the topic:
   ```
   gcloud pubsub subscriptions create --topic my-topic my-sub
   ```
4. Publish a message to the topic:
   ```
   gcloud pubsub topics publish my-topic --message "hello"
   ```
5. Receive the message:
   ```
   gcloud pubsub subscriptions pull --auto-ack my-sub
   ```

Client Libraries

With client libraries you can call Google Cloud APIs by exposing application APIs and administrative APIs. You can also use the Google API client library to access APIs for products such as Google Maps, Google Drive, and YouTube. Application APIs provide access to services. They're optimized to support languages such as Node.js and Python. Use administrative APIs to manage your resources.

Business and Technical Context for the Google Cloud Architect

A vitally important skill for a cloud architect, beyond being a senior engineer, is the ability to understand business objectives and translate those into requirements and design cues. Remember that a cloud architect is the most trusted advisor to an organization's business and technical leaders. You should be an excellent collaborator on both sides of the organization and should be able to convert the organization's needs into tangible solutions. It's not the technology that pays your salary, after all; the business does that. And without a business, there is no technology.

EXAM TIP The biggest element of the Google Cloud Professional Cloud Architect exam is your ability to understand and look for business goals and technical requirements. You'll be presented with scenario-based questions as well as questions based on case studies. As you're reading through the questions, carefully look for keywords that describe elements of the solution. Once you start to parse through a question and gather all of the keywords, make mental bullet-point notes (as I don't believe you're given paper and pencil for this exam) of all these words and phrases, because they'll help you sift through the answers presented.

Assessing Business Requirements

Business requirements are typically broader objectives that are necessary for a business to achieve an operational goal. These may include requirements about reducing expenditures, improving the organization's security, improving data reliability, reducing downtime, minimizing disruptions to services or the impacts of an incident, and so on. Think about how you can capture all of these important business elements when you're assessing a problem or in the room with a stakeholder. A more seasoned cloud architect will know how to probe for important requirements in discovery from stakeholders, rather than waiting on stakeholders to define them. Oftentimes, the stakeholders don't have their requirements entirely fleshed out.

Reducing Expenditures

One of the biggest value points in moving to the cloud is the ability to convert your capital expenditures into operational expenditures. In traditional computing, the business's capital expenditures (CapEx) relate to the money spent up-front on buying servers and equipment, which you'd hope to use for years and years to come. Depreciation, maintenance, and the rapid pace of technology often render hardware obsolete, and this makes traditional capital expenditures a tough challenge in the on-premises environment. With the shift to the cloud, businesses are able to convert their capital expenditures into operational expenditures (OpEx) by paying for services as they're consumed and only for what they consume. Although that can make the cost of your technology cheaper, it doesn't mean companies should stop optimizing their OpEx. It's very important that a business experiences financial gain by investing in the cloud, which is known as the return on investment (ROI). Typically, as companies are looking to move to the cloud, the chief financial officer (CFO) is looking to understand what will be the total cost of ownership (TCO), or the total cost of all the direct and indirect technologies.

EXAM TIP You'll see several questions on the exam that mention the words "most cost-effective" or something similar. This is an example of a key phrase that will help you identify a potential solution.

There are many ways to save money in the cloud. Some engineering teams may be accustomed to setting up an entire VM infrastructure just to perform some services that could easily be done through a managed service. Imagine, for example, that you have to trigger a batch job based on an event that gets published in pub/sub, where you may have considered setting up a VM to trigger this job. You have to pay for all the costs of your VM infrastructure, and you have to pay for all the labor-hours of your resources who are maintaining this VM environment. You have a lot more security overhead to manage, and your VMs don't automatically know to spin up and spin down when a job has been triggered. In such a scenario, you could substitute an entire VM-based architecture for a managed service (Cloud Functions) or even for a preemptible VM. Preemptible VMs may be a great option to save money in this case, because they're designed for short-lived batch jobs or fault-tolerant workloads and are up to 80 percent cheaper than normal VMs.

EXAM TIP When you see questions on your exam that mention "cost-efficiency," think about managed services, serverless services, and things like preemptible VMs.

SLIs, SLOs, SLAs… So What?!

It's very important that you understand the level of service that you offer your users. It's almost impossible to manage a service well if you don't know what is important for that service and how to measure its behavior. To that end, it's important that you understand service level indicators (SLIs), service level objectives (SLOs), and service level agreements (SLAs). You may not see a question on the exam that asks you to define SLIs and SLOs, but you may be presented with a data point or a requirement indicating that the business is looking to maintain or achieve a higher SLA with a new architecture. SLA seems to be a broad term that the industry uses to represent a variety of meanings, so let's break down these three terms here.

A *service level indicator* is a quantitative measure of a chosen characteristic of the level of service that is provided from a product or service. If the characteristic is availability, the SLI could be a percentage of time, often expressed in the number of "nines" (for example, 99.99 percent is "four nines"). Remember that SLA does not mean service level availability—it means service level agreement, even though the availability is oftentimes expressed in the agreement. The actual number itself, or the range of numbers, is the *service level objective*. If we're expecting 99.999 percent availability for a system, "five nines" is our objective and the availability is our measure. What's left if we don't meet these requirements is described in our *service level agreement*. The SLA is a contract that describes the expectations and consequences of meeting or missing an SLO. For example, if Google Cloud Storage doesn't meet its availability targets of 99.95 percent (for the storage class—there are other SLOs for other storage classes), customers are eligible to receive financial credits of a percentage of the monthly bill for the service.

In short, the SLI is the indicator, the measure itself—availability, error rate, and so on. The SLO is the objective, the numerical value that describes the expectation of the measure—99.99 percent availability, 2 to 5 percent error rate, and so on. And the SLA is the agreement, the contracted expectations of using the product or service and what happens if the objectives aren't met. Remember how important reliability is to a successful product and business, and think about how you can use these acronyms effectively in your business meetings.

Assessing Technical Requirements

Technical requirements are based on the functional and nonfunctional requirements of a system that were defined in Chapter 1. As a refresher, functional requirements are the "what"—what is the system supposed to do? For example, the system needs to extract data from this API and load it into a storage bucket, or the system needs to process orders in this format. Nonfunctional requirements are the "how"—how should my system perform? How do we deal with the constraints? For example, the system needs to process at least X amount of data objects per second. It's not so important that you understand the difference between functional and nonfunctional requirements on the exam because you're just parsing for technical requirements, but this is helpful in your day-to-day work as a cloud architect.

Exam Strategies for Case Studies

On the exam, you'll be tested on technical requirements after being presented with scenario-based questions or questions that are based on case studies. Let's look at the technical requirements from the Mountkirk Games case study, as presented in the "Case Studies" section of Chapter 1.

If you refer back to the solution concept on the full case study, you know that the plan is to deploy a new game on Compute Engine and also to leverage a managed NoSQL database. Dynamically scaling the game up or down provides both cost savings and scalability. Scaling up enables the game to support new users as needed without engineering bottlenecks. Scaling down enables the system to save on expensive and unnecessary resource costs. Although many other services dynamically scale up and down, including managed services that cost less than GCE, the case study stated in the solutions concept that the company has decided to build this on Compute Engine. For the sake of this exam, you won't have to sway the exam writers in another direction for the answers. There are much better solutions in the real world than using GCE here, but we'll save that one for another time. In this case, using managed instance groups (MIGs) on GCE would be useful, because MIGs provide autoscaling that automatically adds or deletes instances based on load.

For the second requirement, Mountkirk Games is looking to connect to a transactional database to manage user profiles and game state. The company also wants to integrate with a managed NoSQL database. Firestore (or Cloud Datastore if the exam uses the old terminology) would be a great solution here, because it's a highly scalable NoSQL database built for global applications.

For the third requirement, Mountkirk Games wants to store game activity in a time-series database for future analysis. Immediately Bigtable comes to mind, because Bigtable is a time-series database, but ultra-low latency is not important here because the company wants to store the game activity "for future analysis." BigQuery is another time-series database, and Mountkirk Games wants to query 10TB of historical data on its analytics platform. Tricky, right? You can't get an answer just by reading one line. You know that they want to store the data for future analysis in one requirement, and they want to query 10TB of historical data in another requirement. For that reason, BigQuery is the right answer here.

If you read one keyword and think you can come to a conclusion, hold your horses—and keep reading for additional data points. You might have a feasible solution, and then come across another requirement that completely changes that answer, and that initial solution you thought of may be one of the answers they trick you with on the exam! So be careful as you work through your questions.

(continued)

Strategy for Scenario-Based Questions

Consider the following test scenario:

> CatSnap, a popular cat videos application, wants to build a solution that enables their extended workforce—contractors and temporary staff—to access an environment in which they can upload and download marketing materials for the marketing team.

How do you turn this into a solution? A lot more information is needed here. Luckily, on the exam, you'll get all of the information you need (though in real life, you'll have to probe a little deeper).

Here's an example of a scenario-based question:

> CatSnap, a popular cat videos application, needs to store 50TB of data in an environment where they can share it with extended staff that does not have CatSnap credentials, so that these staff members can upload and download marketing materials that they will be editing. The data needs to have non-repudiation of who accessed it for auditing and monitoring, and data that is older than six months needs to be moved to an archive, where it'll be accessed at most once a year. What is the most secure, cost-effective, and fastest way to do this?
>
> **A.** Provision a private GCS bucket, apply object life cycle policies to move it to coldline after six months, onboard the extended workforce with a CatSnap identity account, and enable bucket logging for the security team to review.
>
> **B.** Provision a private GCS bucket, apply object life cycle policies to move it to archive after six months, onboard the extended workforce with a CatSnap identity account, and enable bucket logging for the security team to review.
>
> **C.** Provision a private GCS bucket, apply object life cycle policies to move it to coldline after six months, enable data owners to create signed URLs that will be provided to the extended workforce as needed, and enable bucket logging for the security team to review.
>
> **D.** Provision a private GCS bucket, apply object life cycle policies to move it to archive after six months, enable data owners to create signed URLs that will be provided to the extended workforce as needed, and enable bucket logging for the security team to review.

So here's what you'd want to parse from this question:

- 50TB of object storage
- Shared user environment
- Untrusted users without credentials

- Upload and download permissions
- Nonrepudiation of each audit log entry
- After six months, move to a new storage class
- Archive is accessed once a year
- Most secure
- Cost-effective
- Fastest

You may have an answer already, but if you look at the four potential answers provided, you can identify another pattern and gather another data point:

- Provisioning a private GCS bucket is a given across all answers.
- Applying object life cycle policies is next, but what's the difference between coldline and archive storage classes? Well, if you knew that the data is accessed once a year and they're looking for the most cost-effective solution, it sounds like archive is the answer here. Coldline would still work, though, because you can access it once a year or more as well, but the key words here are "most… cost-effective."
- Ah, here's an interesting one—do we onboard and provision users with CatSnap identities, or do we use signed URLs? It says the fastest way, so granting signed URLs is the fastest way here. But wait, there's also a requirement of nonrepudiation of all user accesses, so can I have nonrepudiation if my users are using signed URLs? That requirement is an example of a distractor: "fastest" doesn't matter here, because the fastest solution does not satisfy all requirements.
- Bucket logging is enabled across all four answers.

As you start to dissect each exam question, you'll need to have this mind-set: What are patterns I can identify? Where can I find more requirements or keywords in my questions and in the answers provided? How can I eliminate multiple questions at once? While all four of the answers are technically correct, at the end of the day, if you parse through this question properly, the answer should be B, because you cannot use signed URLs as a means to prove nonrepudiation of all the users who could be accessing your data.

Chapter Review

This long chapter provided an overview of a wide variety of complex concepts, including the basics of cloud computing. When you're done reading this book, come back to this chapter and review it to ensure that you know this information. The rest of the book will be more focused on one cloud concept at a time.

This chapter discussed a philosophical overview of your profession and why it's so important for you to think outside of the box of your job. You're on the path to becoming (or already are) someone who designs systems that can fundamentally change society. Beyond your job, think about the moral and ethical duty you have to ensure that the work you do serves a positive purpose to society and to ensure that your voice is always heard as a rational voice in a room with stakeholders who may be a bit one-sided.

We also covered a quick overview of cloud computing, much of which you already know, and discussed some of the key differences between Google Cloud and the other public clouds. Remember that security, open cloud, analytics and AI, and Google's massive global private network are some of the key differentiators between Google, AWS, and Azure. Think about how you structure your conversations with your stakeholders when it comes to deciding where to develop your workloads if you work in a multi-cloud environment. For the exam, you most likely won't see any questions around this topic.

We took a 10,000-foot overview of the Google Cloud Platform products and services that are most likely to be the focus of your exam, as well as the ways you can manage your cloud. It's important that you understand all the Google technologies identified in this chapter, though this discussion is not inclusive of all that could be on your exam (in other words, additional things may show up on your exam).

Lastly, we discussed the mind-set you'll need when attending meetings or taking the exam and parsing for business and technical requirements. Work on some practice questions and try to understand how certain words may trigger an entirely different meaning for your solution. Unfortunately, the exam can be more difficult for non-English speakers or non-Japanese speakers, because the exam is offered in only those two languages as of this writing. But don't fret, because you're doing great, and you're going to know a lot about Google Cloud Platform by the end of this book.

Additional References

If you'd like more information about the topics discussed in this chapter, check out these sources:

- **Google Cloud Architecture Framework** https://cloud.google.com/architecture/framework
- **Google Cloud Adoption Framework** https://cloud.google.com/adoption-framework
- **Kubernetes CLI** https://kubernetes.io/docs/reference/kubectl/overview/#syntax
- **Gsutil Tool** https://cloud.google.com/storage/docs/gsutil
- **BQ CLI** https://cloud.google.com/bigquery/docs/bq-command-line-tool
- **Gcloud CLI** https://cloud.google.com/sdk/gcloud/reference
- **Jayendra Patil's PCA Blog Post** https://jayendrapatil.com/google-cloud-professional-cloud-architect-certification-learning-path/

Questions

1. You are looking to buy a new computer for personal use. You want a powerful enough computer to surf websites and run (mostly gaming) applications. But you realize that you are short on cash. What computer should you buy?

   A. MacBook Pro

   B. Chromebook

   C. Windows 10 desktop

   D. Windows 10 laptop

2. Your company has made plans to roll out OpenShift, a Kubernetes platform solution offered by IBM Red Hat, across all its on-premises and public cloud environments. Given that you are the lead architect responsible for your company's GCP deployments, what type of shared responsibility model will this deployment entail for you?

   A. On-premises

   B. IaaS

   C. PaaS

   D. SaaS

3. VPC networks are:

   A. Global

   B. Regional

   C. Zonal

   D. Local

4. Subnets are:

   A. Global

   B. Regional

   C. Zonal

   D. Local

5. You need to attach high-performance storage with very high IOPS and low latency to your VM instance. Which technology should you use?

   A. Google Cloud Storage

   B. Local SSD

   C. Cloud FileStore

   D. Persistent SSD disk

6. Google's operational excellence principle demands the building of a foundation to enable reliability successfully across your infrastructure by efficiently running, managing, and monitoring systems that deliver business value. Which of the following is not a key strategy that drives this principle?

   A. End-to-end automation

   B. Monitoring business objectives

   C. Performance and cost optimization

   D. Disaster recovery

7. Which of the following service level measures are considered a legally enforceable contract between the service provider and the service consumer?

   A. SLA

   B. SLE

   C. SLO

   D. SLI

8. Your development team is building a new business-critical application using virtual machines to be deployed in a dedicated production project. As part of this effort, the team is looking to implement a dedicated application testing environment within a development project. The tests generally take less than an hour to complete. They need to keep the testing machine costs low, but consistent with their production environment. Which type of virtual machine optimization strategy would you use for their testing project environment?

   A. Use sole-tenant nodes.

   B. Automate the VM life cycle.

   C. Use preemptible VMs.

   D. Use purchase commitments.

9. Which Google Cloud Platform database offering is best suited for integration with client-side mobile and web applications, gaming leaderboards, and user presence at global scale?

   A. BigQuery

   B. Cloud Memorystore

   C. Cloud Bigtable

   D. Cloud Firestore

10. You decide to use GCP to host a simple website using Drupal Content Management from Google Cloud Platform Marketplace to run a Google Compute Engine VM instance. You have global ambitions but a limited budget. What feature would you enable to provide a better experience to your global audience?

A. Cloud Interconnect

B. Cloud DNS

C. Cloud CDN

D. Cloud Load Balancer

Answers

1. C. This question is about parsing requirements. First, the requirement of surfing websites can be accommodated by all the computers listed. Second, running locally installed applications, including games, eliminates Chromebook from the picture and even the MacBook. Third, you have a requirement for the most performance with the least cost. Lastly, while both the Windows desktop and laptop seem to be strong answers, the prerequisite knowledge of the gaming PC space will help you answer this question effectively, making the Windows desktop the best answer, because these computers typically offer the most bang for the buck. You'll see questions on the exam that are very tricky, just like this one.

2. B. The key to remember here is that for a service provided (GCP in this case) to take responsibility for its PaaS, it must offer the service as a managed service. GCP offers its own Kubernetes platform called GKE. But OpenShift is not a Google-offered PaaS solution. As such, Google will not take responsibility for the backend operations and design of your OpenShift environments. You will need to manage all the VMs that OpenShift will provision as part of its GCP deployment. So this is an IaaS deployment from a shared responsibility model perspective.

3. A. VPC networks are global in GCP. Most questions on the exam are multilayered. So you will generally not get a straight question like this. Knowing this element will help you answer multilayered questions more effectively.

4. B. Subnets are regional within GCP. That means that you can create instances within your subnet across multiple zones. Keep that in mind when building highly available solutions. Many other cloud service providers support only zonal subnets.

5. B. From a storage perspective, you can attach each of the provided answers to your VMs. Google Cloud Storage can be mounted to a VM. But this is not the highest performer. The highest performance possible is achieved when the storage is included with the CPU and not attached via the network. The only answer that satisfies that criteria is Local SSD. Its read and write IOPS are almost ten times higher than persistent network attached storage options.

6. **C.** Operational excellence does not include performance and cost optimization; that is a different system design principle.
7. **A.** A service level agreement (SLA) is an enforceable legal contract between a service provider and a service customer. A service level indicator (SLI) is a measure of the service level provided by a service provider to a customer. SLIs form the basis of service level objectives (SLOs), which in turn form the basis of SLAs. In this way, SLAs define the level of service expected by a customer from a supplier, laying out the metrics by which that service is measured and the remedies or penalties, if any, should the agreed-upon service levels not be achieved.
8. **B.** This is typically the type of question you might see on the exam. You're provided a set of requirements that can often be met by a number of approaches. Between B and C, there are many cost savings. In fact, I would try to use automated preemptible VMs to get the most bang for my dollars. Preemptible machines give you up to 80 percent off and would satisfy the ability to use the same type of machines as the production systems (as do purchase commitments). Furthermore, with a one-hour timetable to run your tests, preemptible VMs are ideal. Even if you get preempted, you can try again later. If you automate your VM life cycle, you immediately get about a 96 percent savings. That means you run the machine for only 1 hour of 24 hours a day. The ideal answer in real life could be to use both preemptible instances and VM life cycle management as part of an automated CICD build and test process. But that is not how the question presented things. So the biggest contributor to cost savings would be automation of the VM life cycle.
9. **D.** You can argue that certain database technologies are cross-purpose and could be deployed in other situations. You would be right. But, again, the key to passing the test is to provide the best answer based on the key information presented. Firestore would be the best database to integrate with mobile and web apps.
10. **C.** The only answer that helps with the content distribution is Cloud CDN (Content Delivery Network). Although DNS may be in use, it won't make an impact on the global audience, which this question is asking for. Furthermore, since this is a simple VM, adding a load balancer to the architecture does not buy you anything. To leverage a load balancer, you would want multiple VMs within multiple regions, which is clearly not the case in this question.