

Google Cloud Certified

Professional Cloud Architect Definitive Guide



Sriram

Google Cloud Certified Professional Cloud Architect Definitive Guide

TABLE OF CONTENTS

INTRODUCTION TO GOOGLE CLOUD PLATFORM	4
CLOUD COMPUTING OVERVIEW.....	4
GOOGLE CLOUD PLATFORM OVERVIEW	15
COMPUTE SERVICES	40
COMPUTE ENGINE.....	40
APP ENGINE	56
KUBERNETES ENGINE.....	59
CLOUD FUNCTIONS.....	62
COMPUTE – INTERVIEW EXAM TIPS	63
COMPUTE QUIZ	72
STORAGE SERVICES.....	106
GCP STORAGE SERVICES	106
CLOUD STORAGE (OBJECT STORAGE).....	110
CLOUD FILESTORE (NETWORK ATTACHED STORAGE).....	119
CLOUD SQL.....	121
CLOUD SPANNER.....	126
CLOUD BIGTABLE.....	131
CLOUD DATASTORE.....	135
CLOUD FIRESTORE	138
CLOUD MEMORYSTORE.....	142
CLOUD BIGQUERY.....	145
STORAGE – INTERVIEW EXAM TIPS	153
STORAGE QUIZ	161
NETWORKING SERVICES	188
CLOUD VPC	188
CLOUD VPN.....	204
CLOUD INTERCONNECT	206
CLOUD ROUTER.....	209
CLOUD LOAD BALANCING	210
CLOUD DNS	215
CLOUD CDN	217

Google Cloud Certified Professional Cloud Architect Definitive Guide

NETWORKING – INTERVIEW EXAM TIPS	218
NETWORKING QUIZ	220
IDENTITY & SECURITY	230
CLOUD IAM	231
SECURITY – INTERVIEW EXAM TIPS.....	251
SECURITY QUIZ	256
BIGDATA SERVICES.....	269
OVERVIEW OF DATA PIPELINES	269
GCP PIPELINE COMPONENTS	270
MIGRATING HADOOP AND SPARK TO GCP	274
DATA CATALOG.....	275
DATAPREP	277
DATA STUDIO	281
DATALAB	283
CLOUD COMPOSER	285
BIGDATA – INTERVIEW EXAM TIPS.....	288
BIGDATA QUIZ	293
MACHINE LEARNING SERVICES	310
DEPLOYING MACHINE LEARNING PIPELINES.....	310
MEASURING, MONITORING & TROUBLESHOOTING MACHINE LEARNING MODELS.....	324
LEVERAGING PREBUILT MODELS AS A SERVICE	330
MACHINE LEARNING – INTERVIEW EXAM TIPS.....	340
MACHINE LEARNING QUIZ	347
MANAGEMENT TOOLS	361
MONITORING WITH STACKDRIVER	362
DEPLOYMENT MANAGER	366
CLOUD SHELL & CLOUD SDK.....	369
MANAGEMENT TOOLS – INTERVIEW EXAM TIPS.....	385
MONITORING QUIZ	386
MIGRATION SERVICES	390
MIGRATION – INTERVIEW EXAM TIPS.....	400

Google Cloud Certified Professional Cloud Architect Definitive Guide

CLOUD ARCHITECT CERTIFICATION.....401

ASSESSMENT TEST	401
CASE STUDY	410
PRACTICE EXAM	419

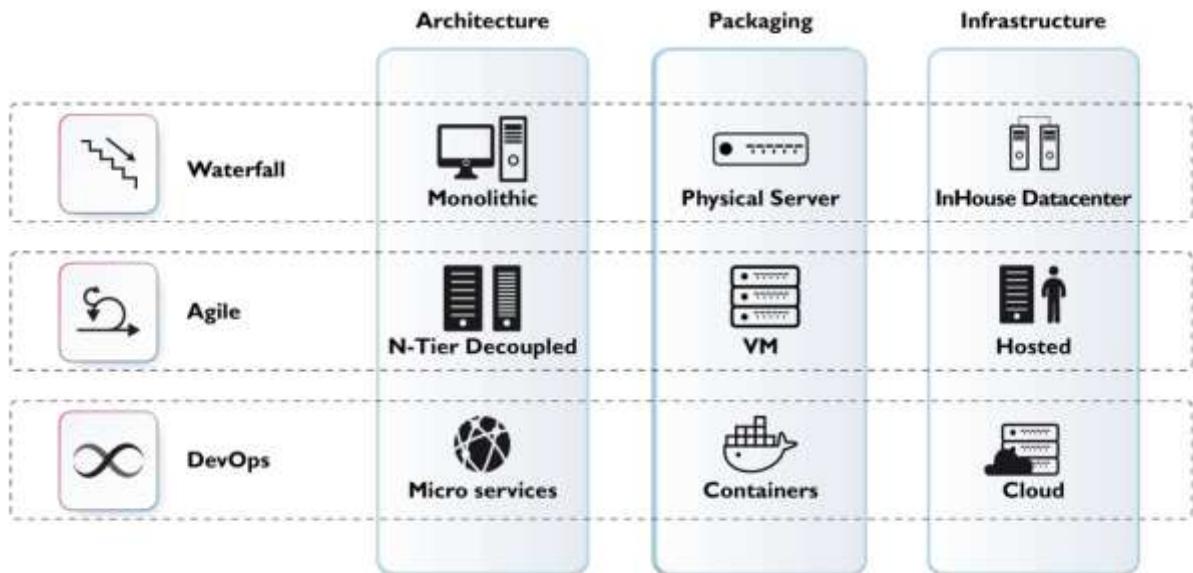
DATA ENGINEER CERTIFICATION.....501

ASSESSMENT TEST 1	501
ASSESSMENT TEST 2	510
DATA ENGINEER MODEL EXAM	533
ROI EXAM	546

Introduction to Google Cloud Platform

Cloud Computing Overview

Evolution of an IT Industry



First generation

- Adapted Waterfall SDLC model for the process
- Followed Monolithic - Tightly Coupled Architecture (2 tier | 3 tier – deployed in a single server)
- Managed Physical Server with huge capacity for less usage and wasted huge money in infra
- Hosted the application in the Inhouse Datacenter infrastructure

Second generation

- Adapted Agile SDLC model for the process which is more effective
- Followed N-tier Decoupled Architecture
- Instead of procuring Physical server for each project, started using Virtual Machines to consume less space
- Hosted the applications in third party data centers

Agile overcomes the drawback of waterfall with following measures: -

- Shorter Deployment Cycles, Faster Innovation

- Reduced Deployment Failures, Rollbacks, and Time to Recover
- Improved Communication and Collaboration
- Increased Efficiencies

Challenges in both First and Second generation

The challenges in First & Second generations are: -

- Infrastructure Readiness
- Scalability
- Quality of the Product
- Limited Release windows – Releases may slip/fail

Third generation

- Adapted Agile + DevOps for the success implementation of process and technology
- Followed Microservices Architecture without affecting the failure of any modules with zero down time
- Hosting the application through containers to free from environmental issues
- Deployed in Cloud with High Availability, Scalability, Fault Tolerance

Between Second and Third generation

Agile without DevOps leads to: -

- Release and Deployment Mismatch
- Unpredictable issues
- Blame Game
- Lack of monitoring & Feedback

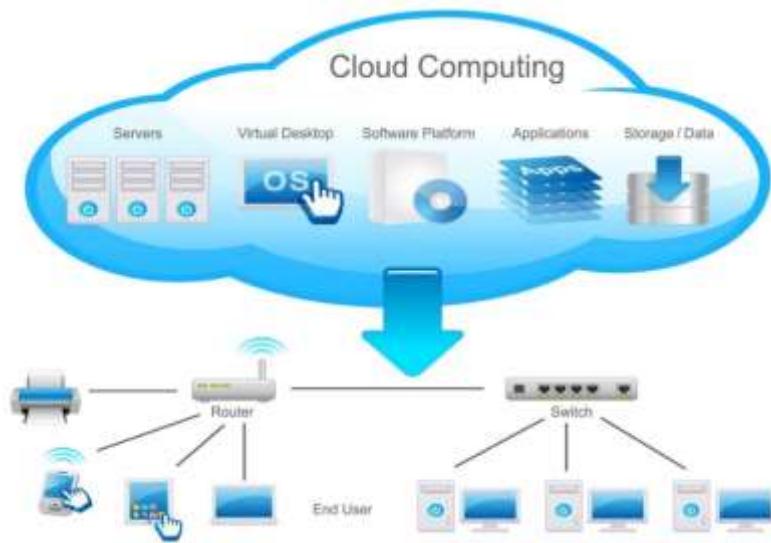
Agile with DevOps leads to: -

- Streamline Deliveries
- Team Work in Collaboration
- Continuous Monitoring & Feedback

Agile overcomes the changes in waterfall model by adapting good processes in development but not focus on deployment. So, DevOps comes to picture in third generation.

What is Cloud Computing?

Cloud is a collection of computing resources which are accessed through the Internet and the software and databases that run on those servers



Cloud computing is the on-demand delivery of IT resources and applications via the Internet with pay-as-you-go pricing. Instead of buying, owning, and maintaining data centers and servers, organizations can acquire technology such as compute power, storage, databases, and other services on an as-needed basis.

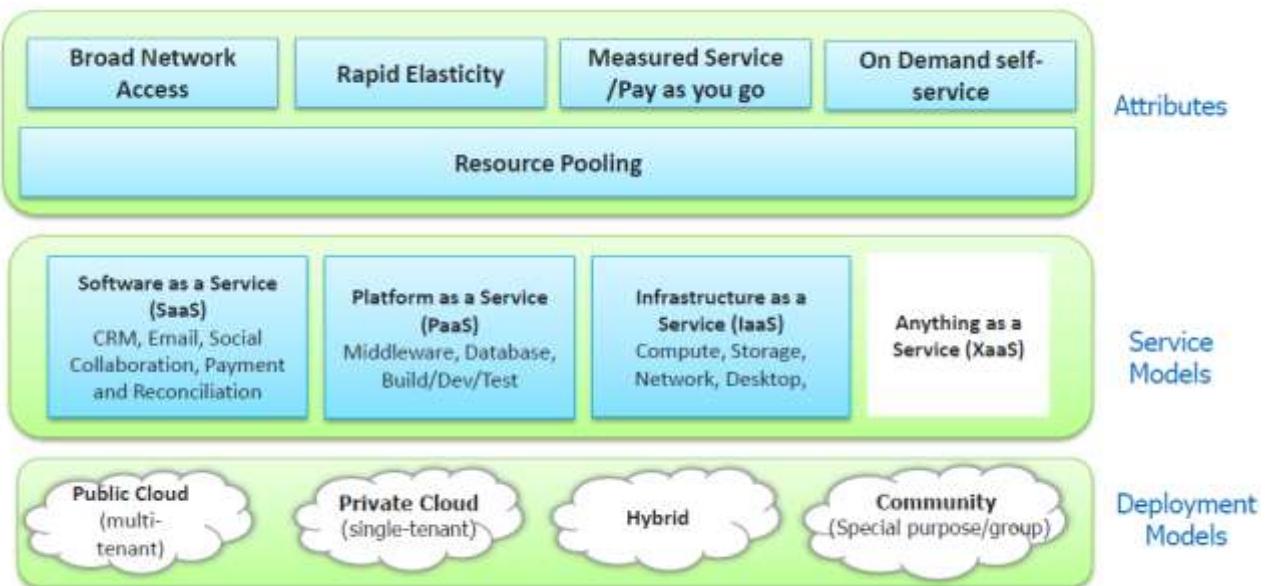
Cloud computing is built on the ability to efficiently divide physical resources into smaller but flexible virtual units. Those units can be “**rented**” by businesses on a pay-as-you-go basis and used to satisfy just about any networked application and/or workflow’s needs in an affordable, scalable, and elastic way.

Companies offering these computing services are called **Cloud providers** and typically charge for cloud computing services based on usage, similar to how you are billed for water or electricity at home. E.g.: AWS, AZURE, GOOGLE CLOUD, OCI, IBM Bluemix

The **primary reasons for moving to the cloud** are: -

- You don’t need to maintain or administer any infrastructure
- It will never run out of capacity, since it is a virtually infinite
- You can access your cloud-based applications from anywhere, you just need a device which can connect to the internet

Cloud Computing Big picture



Growth of Cloud Computing

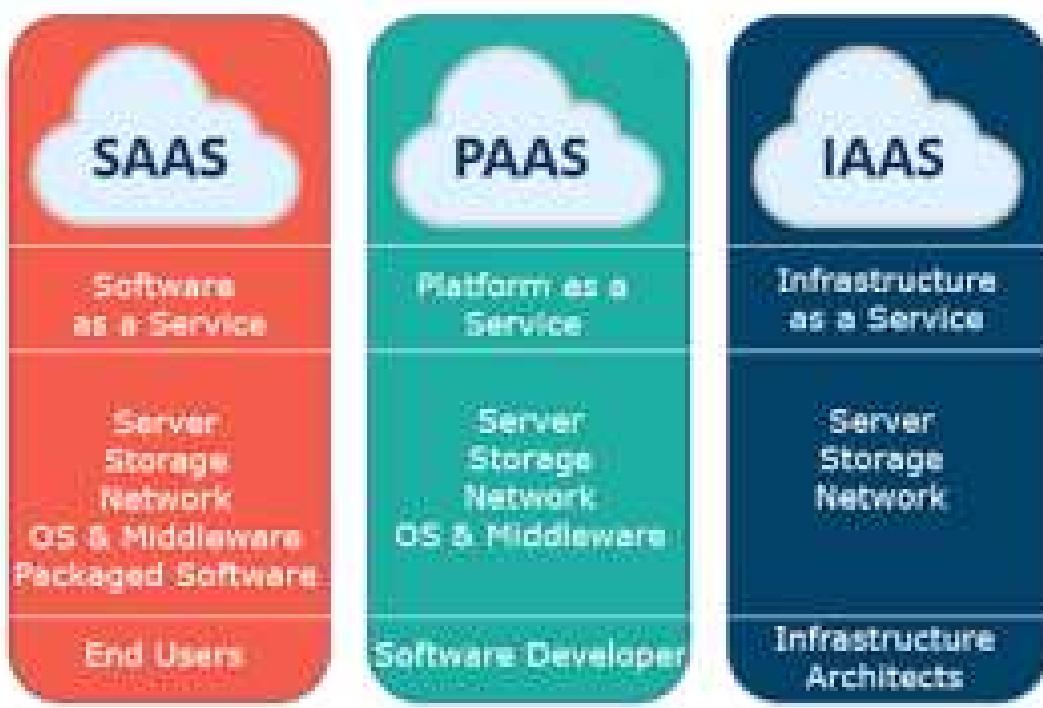


Why Cloud Computing?

The main reason for Cloud Computing is: -

- Mainly Focus on core business instead of spending time in IT Infrastructure
- Quick provisioning of services (Instead of building servers we will use services)
- Auto or Dynamic Scaling based on the demand
- Lower TCO (More Cost Effective) based on the utilization
- Better Reliability, Scalability & Sustainability
- Secure Store Management
- Low Capital Expenditure
- Frees from Internal Resources
- Utility Based
- Easy & Agile Deployment
- Device & Location Independent
- 24*7 Support

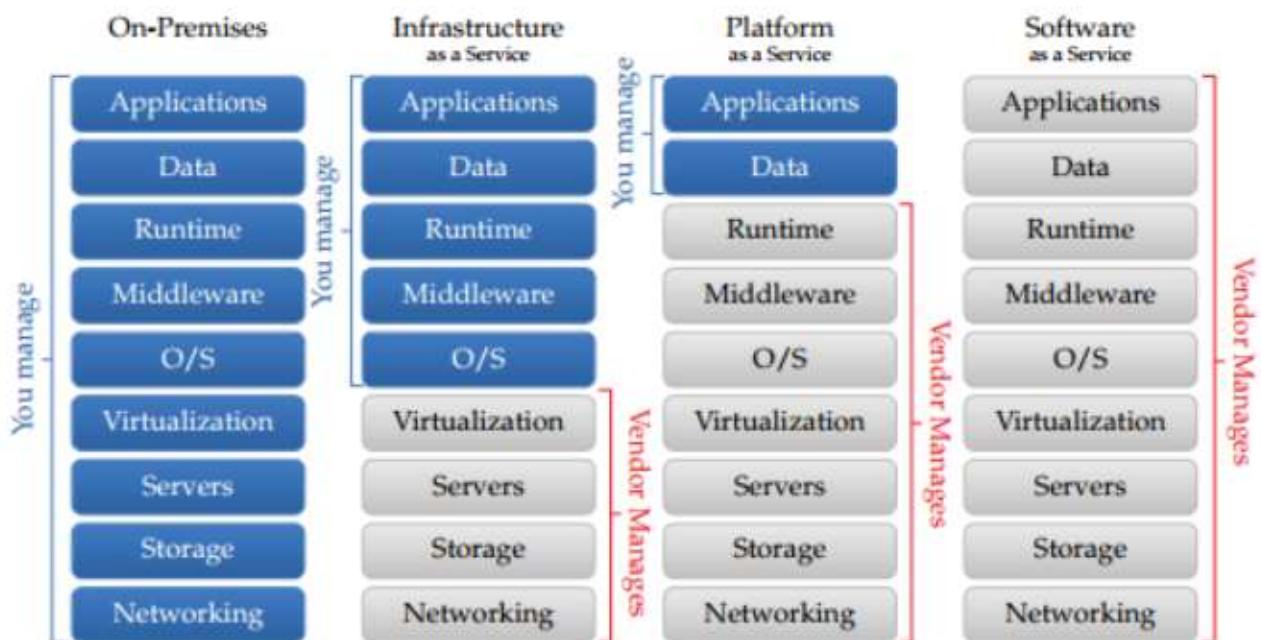
Service Models



The three service models of cloud computing are: -

- **Infrastructure as a Service (IaaS)** provides cloud infrastructure (Virtual Machine) in terms of hardware like memory, processor speed etc. Example: AWS, Azure, GCP
- **Platform as a Service (PaaS)** provides cloud application platform (Programming Environment) for the developers. Example: Elastic Beanstalk, Google AppEngine, Heroku, Salesforce.com, Apache Stratos
- **Software as a Service (SaaS)** provides cloud applications (Application Environment) which are used by the user directly without installing anything on the system. The application remains on the cloud and it can be saved and edited in there only. Example: Microsoft Office 365, CRM, Workday, NETSUITE, Athena Health Record Management, Intuit, Adobe Creative Cloud & Gmail

Comparison of On-Premise with Service Model



Deployment Model

The Four Cloud Deployment Models are: -

- Public Cloud
- Private Cloud
- Hybrid Cloud
- Community



Google Cloud Certified Professional Cloud Architect Definitive Guide

Public cloud are owned and operated by a third-party cloud service provider, which deliver their computing resources like servers and storage over the Internet. Made available to the general public or a large industry group.

Private Cloud works the same way as Public Cloud, but these services are provided to internal business units instead of to external public enterprises. Operated solely for an organization, may be managed by the organization or a third party and limit's access to enterprise and partner network.

Hybrid clouds combine public and private clouds, bound together by technology that allows data and applications to be shared between them. By allowing data and applications to move between private and public clouds, hybrid cloud gives businesses greater flexibility and more deployment options.

Companies with limited Private Cloud infrastructure may 'cloud burst' into Public Cloud for additional capacity when required. A company Cloud also have Private Cloud at their main site and use Public Cloud for their Disaster Recovery location.

Composition of two or more clouds (private, community, or public) bound together by standardized or proprietary technology that enables data and application portability

Community Cloud is similar to a traditional extranet, but with full shared data center services instead of just network connectivity between On-Premise Offices.

Characteristics of Cloud Computing

On-demand self-service	Broad network access	Resource pooling	Rapid elasticity	Measured service
No human intervention needed to get resources	Access from anywhere	Provider shares resources to customers	Get more resources quickly as needed	Pay only for what you consume

What is a Hypervisor?

A Hypervisor is a type of software used to create and run virtual machines. It integrates physical hardware resources into a platform which are distributed virtually to each user.

Hypervisor includes Oracle Virtual Box, Oracle VM for x86, VMware Fusion, VMware Workstation, and Solaris Zones.

Which data centers are deployed for cloud computing?

There are two data centers in cloud computing, one is Containerized Data centers, and another is Low-Density Data Centers

What is Hybrid cloud architecture?

It is a type of architecture where the workload is divided into two halves among which one is on public load and the other is on the local storage. It is a mix of on-premises, private cloud and third-party, and public cloud services between two platforms.

If you hold half of the workload on the public cloud whereas different half is on local storage, in such case what type of architecture can be used?

In such cases, the hybrid cloud architecture can be used.

Mention the different layers of cloud architecture?

Following are the different types of layers in cloud architecture: -

- Node controller
- Cloud controller
- Cluster controller
- Storage controller

List the Public Cloud Providers?

IaaS	SaaS	PaaS
<ul style="list-style-type: none"> ▪ Amazon Web Services ▪ MS Azure ▪ Google Compute Engine ▪ Alibaba Cloud ▪ Rackspace ▪ Digital ocean ▪ Megha ▪ TCS-Insta Compute 	<ul style="list-style-type: none"> ▪ SAP on AWS ▪ OFFICE 365 ▪ Maps ▪ Facebook 	<ul style="list-style-type: none"> ▪ Elastic Beanstalk ▪ Azure for .NET ▪ developers. Google ▪ Google app engine (GWT-RIA) ▪ developers. Facebook ▪ Cloud Foundry

List the Open-Source Cloud Provider?

Open Source in IaaS



Open Source in PaaS



What are the use cases for Cloud?



Infrastructure Transformation



SaaS (e-mail,
collaboration, etc.)



Dev and Test



Hosted
Solutions



Content Delivery
Networks



High performance
Computing



POCs



Backup



DR



VPCs/ Private
Clouds

Google Cloud Platform Overview

Today's Industry Problem

- In Today's World Multinational Company's has expanding their business in various domains like retail, logistic, travel, telecommunication, banking & finance, insurance and many more.
- Mainly there are two issues that the company is facing and wants to find out a way to resolve these are Scalability and Security
- Due to business expansion the company's software was difficult to scale and implement new capabilities such as Artificial Intelligence, Bigdata and other security aspects to handle finance regulatory compliance
- Company appointed Cloud Architect to identify the issues that the currently company's facing and informed to find a solution for their problem statement

GCP – The Ultimate Cloud Solution

- Initially, Cloud Architect planned to set up the architecture and migrate the customer business needs with GCP as pilot model to demonstrate all the service of GCP Walkthrough
- First to set up the architecture, Cloud Architect first created a GCP account for the company, then explored the essential services to meet their business needs
- GCP is the ultimate solution from Small-Mid Size-Enterprise organization with billing discount
- GCP provides various services from the below products
 - Compute
 - Storage
 - Networking
 - Database
 - Bigdata
 - Machine Learning

GCP is better than any other cloud provider

- Superb Network Performance (Fast Solutions)
- Attractive Pricing Model (Sub hour billing, Sustained use discounts, Custom machine types)
- Best PaaS Solutions
- Robust Infrastructure

What is GCP? What are its Benefits?

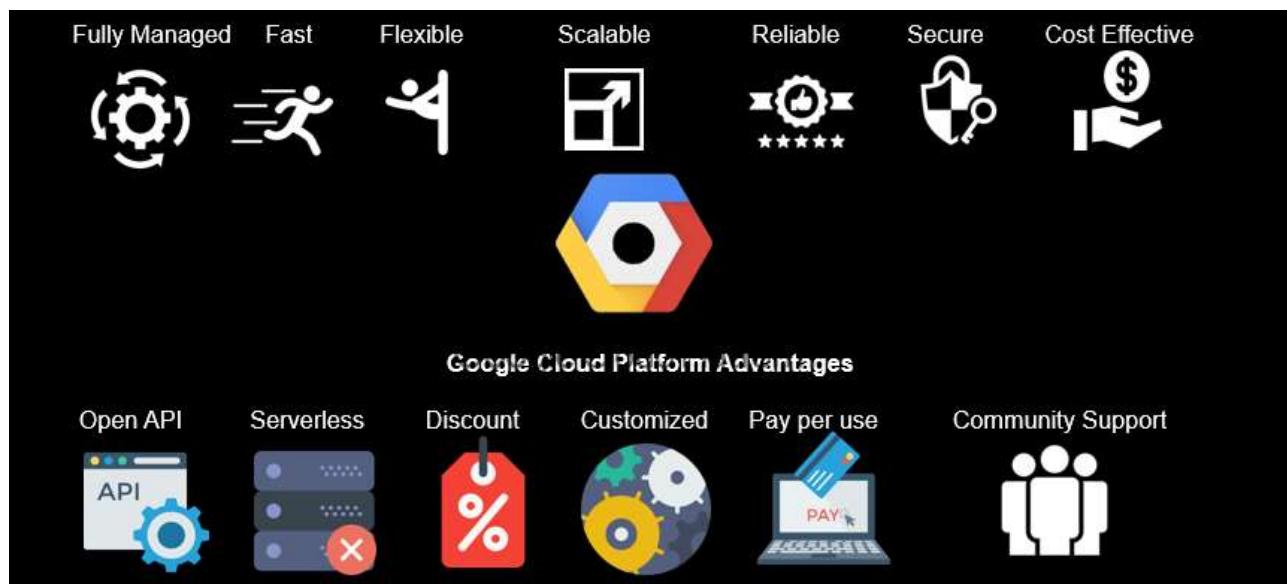
- Google Cloud Platform is a suite of public cloud computing services offered by Google, released in 2011
- Google Cloud Platform enables developers to build, test, and deploy applications on Google's highly secure, reliable and scalable infrastructure.
- GCP is a collection of Google computing resources, which are offered via services, Provides many different cloud services such as Computing, Storage, Networking, Security, Big Data and Machine Learning
- Started as PaaS and now offers IaaS as well
- Support for various programming languages : Java, Python, Ruby, Go



What are the use cases for GCP?

- Application Hosting
- Media Sharing (Image/ Video)
- Media Distribution
- Backup -Archive
- Search Engines
- Social Networking
- Scalable Applications
- SaaS / PaaS Hosting

What are the advantages of Google Cloud Platform?



The advantages of Google Cloud Platform are: -

- Provides Global Infrastructure, using private global fiber optic network
- Fully managed services & Fastest service provisioning
- Supports Hybrid and Multi-cloud environment
- Fastest and the best network among all the cloud providers
- GCP is Flexible, Scalable, Reliable Secured, highly scalable and provides high performance
- GCP is more cost-effective because it provides better pricing than other competitors with per second billing i.e. For more usage billing gets reduced
- GCP is highly secured, provides an excellent physical and digital security
- Open APIs compatibility with open-source services for Tensor Flow, Kubernetes & ForsetiSecurity
- Multi-vendor Friendly Technologies for Stack Driver, Bigdata, AI
- Support's Server based & Serverless models
- Excellent Community Support

Why GCP?

- Very Fastest growing cloud computing platform on the planet
- Third Largest public cloud computing platform on the planet
- More and More organizations are outsourcing their IT to GCP
- The GCP certifications are the most popular IT certifications right now
- Top Paid certification according to Forbes
- GCP was named as a third leader in the “IaaS Magic Quadrant” – Gartner

Figure 1. Magic Quadrant for Cloud Infrastructure and Platform Services



GCP Regions & Zones

How GCP Environment is organized?

GCP Environment is Organized into Regions and Zones



Google Cloud Platform

Announced regions



Google will continue expanding into the following new regions: -

- Melbourne, (Australia)
- Toronto, (Canada)
- Delhi, (India)
- Doha, (Qatar)

Region

A **Region** is a physical location in the world where we have multiple Zones. A grouping of GCP data centers within a specific region. Designed to be independent of other regions.

Regions are collections of zones. Zones have high-bandwidth, low-latency network connections to other zones in the same region. In order to deploy fault-tolerant applications that have high availability, Google recommends deploying applications across multiple zones and multiple regions. This helps protect against unexpected failures of components, up to and including a single zone or region. For example, the us-west1 region denotes a region on the west coast of the United States that has three zones: us-west1-a, us-west1-b, and us-west1-c.

Zones

A Zone is Subset of a Region, physically isolated & independent infrastructure, High speed connectivity, Low latency & every region has a minimum of 2 Zones

A zone is a deployment area within a region. The fully-qualified name for a zone is made up of <region>-<zone>. For example, the fully qualified name for zone a in region us-central1 is us-central1-a. Depending on how widely you want to distribute your resources, create instances across multiple zones in multiple regions for redundancy.

Edge Locations

- Edge Locations are endpoints for GCP which are used for caching content. Typically, this consists of Google Content Delivery Network (CDN). There are many more edge locations than regions. Currently there are over 100 edge locations. Based on the nearest edge location, customers can communicate and get data.
- Locations built to deliver cached data across the world. CDN utilizes this service for faster delivery to countries without AWS regions.
- Edge Location is an intermediate between the end users and servers to access the services from GCP.
- Edge Location is a small setup in different location to provide low latency connection by caching static content. Basically, it's a Content Delivery Network and used with Google CDN.

Edge Cache

Edge Cache is used to store my frequently accessed data in the server

Latency

- Latency is the measure of time required for transfer of data from client to server and again back to client
- More the latency lower is the efficiency. Low latency, higher efficiency
- GCP has its clients throughout the world, so in order to reduce latency and avoid load on servers they make use of Edge Locations

To know the latest information about GCP Regions and Zones

Visit Website: <https://cloud.google.com/compute/docs/regions-zones>

Region	Zones	Location
asia-east1	a, b, c	Changhua County, Taiwan
asia-east2	a, b, c	Hong Kong
asia-northeast1	a, b, c	Tokyo, Japan
asia-northeast2	a, b, c	Osaka, Japan
asia-northeast3	a, b, c	Seoul, South Korea
asia-south1	a, b, c	Mumbai, India
asia-southeast1	a, b, c	Jurong West, Singapore
australia-southeast1	a, b, c	Sydney, Australia
europe-north1	a, b, c	Hamina, Finland
europe-west1	b, c, d	St. Ghislain, Belgium
europe-west2	a, b, c	London, England, UK
europe-west3	a, b, c	Frankfurt, Germany
europe-west4	a, b, c	Eemshaven, Netherlands
europe-west6	a, b, c	Zürich, Switzerland
northamerica-northeast1	a, b, c	Montréal, Québec, Canada
southamerica-east1	a, b, c	Osasco (São Paulo), Brazil
us-central1	a, b, c, f	Council Bluffs, Iowa, USA
us-east1	b, c, d	Moncks Corner, South Carolina, USA
us-east4	a, b, c	Ashburn, Northern Virginia, USA
us-west1	a, b, c	The Dalles, Oregon, USA
us-west2	a, b, c	Los Angeles, California, USA
us-west3	a, b, c	Salt Lake City, Utah, USA

GCP Consumption Mechanism | Several ways of accessing GCP

			
Cloud Platform Console	Cloud Shell and Cloud SDK	Cloud Console Mobile App	REST-based API
Web user interface	Command-line interface	For iOS and Android	For custom applications

GCP is committed to environmental responsibility



GCP Products & Services

GCP Products



Compute



Storage & Database



Networking



Big Data



Developer Tools



Identity & Security



Internet of Things



Cloud AI



Management Tools



Data Transfer

Compute

It is used to process data on the cloud by making use of powerful processors which serve multiple instances at a time.

Storage

The storage as the name suggests, is used to store data in the cloud, this data can be stored anywhere but content delivery on the other hand is used to cache data nearer to the user so as to provide low latency.

Database

The database domain is used to provide reliable relational and non-relational database instances managed by GCP.

Networking

It includes services which provide a variety of networking features such as security, faster access etc.

Security and Identity

It includes services for user authentication or limiting access to a certain set of audience on your AWS resources.

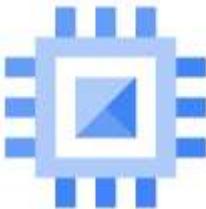
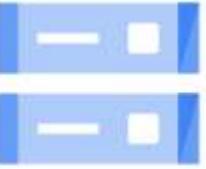
Management Tools

It includes services which can be used to manage and monitor your AWS instances.

Developer Tools

It helps the developers to build and deploy the applications

GCP Services

Compute Services	
Compute Engine: Run large scale workloads on virtual machines hosted on Google's infrastructure	
App Engine: A platform for building scalable web apps and mobile backends	
Container Engine: Run Docker containers on Google's infrastructure, powered by kubernetes	
Cloud Functions: A Serverless platform for building event-based microservices triggered by events in GCP	
Storage & Database Services	
Cloud Storage: Robust, simple and cost-effective object storage service with a global edge caching	
Cloud SQL: Data storage and management using a fully managed relational MYSQL DB	
Cloud Spanner: Fully managed relational database with unlimited scale, strong consistency & up-to 99.999% availability	

Google Cloud Certified Professional Cloud Architect Definitive Guide

Cloud Bigtable: Fast, fully managed and scalable NoSQL DB service	
Cloud Datastore: A managed, NoSQL, schema-less database for storing non-relational data	
Networking Services	
Virtual Private Cloud: Managed Networking functionality for your GCP resources	
Cloud Load Balancing: High performance, scalable load balancing on GCP	
Cloud CDN: Low Latency, Low-Cost Content Delivery Network using Google's global network	
Cloud Interconnect: Connect your infrastructure to Google's network edge with enterprise grade interconnect	
Cloud DNS: Reliable, resilient, low latency DNS serving from Google's world-wide network	

Security & Identity Services	
Cloud IAM IAM lets administrators authorize who can take action on specific resources, giving you full control and visibility to manage cloud resources centrally.	
Bigdata Services	
BigQuery: Cost-effective and fully managed data warehouse for large-scale data analytics platform	
Cloud Dataflow: Real-time data processing service for processing the batch stream data	
Cloud Dataproc: Cost-effective, easy to use and managed Spark and Hadoop Service	
Cloud Pub/Sub: Connects the services with reliable, many-to-many, asynchronous messaging hosted on Google Cloud Infrastructure	
Cloud Datalab: An easy-to-use interactive tool for large-scale data exploration, analysis, and visualization	

Machine Learning Services	
Cloud ML	
Train high-quality custom machine learning models with minimal effort and machine learning expertise. Fully managed machine learning service, Familiar notebook-based developer experience. Optimized for Google infrastructure; integrates with BigQuery and Cloud Storage	
Vision API: Derive insights from your images in the cloud or at the edge with AutoML Vision or use pre-trained Vision API models to detect emotion, understand text, and more..	
Video API: Enable powerful content discovery and engaging video experiences	
Speech-to-Text API: Speech-to-Text allows developers to convert audio to text by applying powerful neural network models in an easy-to-use API	
Text-to-Speech API: Text-to-Speech synthesizes human-like speech based on input text in a variety of voices and languages	
Natural Language API: Derive insights from unstructured text using Google machine learning	
Translation API: Dynamically translate between languages using Google machine learning	

<p>Dialogflow:</p> <p>Lifelike conversational AI with state-of-the-art virtual agents. Available in two editions: Dialogflow CX (advanced), Dialogflow ES (standard)</p> <ul style="list-style-type: none"> ○ Support rich, intuitive customer conversations, powered by Google's leading AI ○ One comprehensive development platform for chatbots and voicebots ○ Join a community of over 1.5 million developers building with Dialogflow ○ Achieve exceptional CSAT with Dialogflow, part of Contact Center AI solution 	
<p>Tensor Flow:</p> <p>An Open-source tool to build and run neural network models. Has wide platform support: CPU or GPU; mobile, server, or cloud</p>	

Management Tools	
<p>Stackdriver [Cloud Monitoring, Logging & Diagnostics]</p> <p>Provides monitoring, logging, and diagnostics for applications built on cloud infrastructure including GCP and AWS. Stackdriver provides metrics, dashboards, alerting, log management , reporting, and tracing capabilities</p>	
<p>Deployment Manager [Template-based Infrastructure Deployment]</p> <p>An infrastructure automation and management service that allows you to define templates to deploy a variety of GCP services, including Cloud Storage, Compute Engine, and Cloud SQL</p>	
<p>Cloud Shell [Browser-Based Terminal/CLI]</p> <p>Command-line access to cloud resources from within a browser, without having to install the Google Cloud SDK or other tools on your system</p>	
<p>Google Cloud Billing API [Programmatic GCP Billing Management]</p> <p>Programmatically managed billing for your GCP projects</p>	

What are the GCP Managed Services?

Installation, Upgrade, Patch, and Backup activity will be managed by the Google called Managed Services. Managed services can be serverless or non-serverless.

Serverless: Cloud Dataflow, Cloud Big Query, Cloud Pub/Sub, ⋯

Server-based: GCE, GKE, Cloud Dataproc, ⋯

Serverless Vs Server-based Service

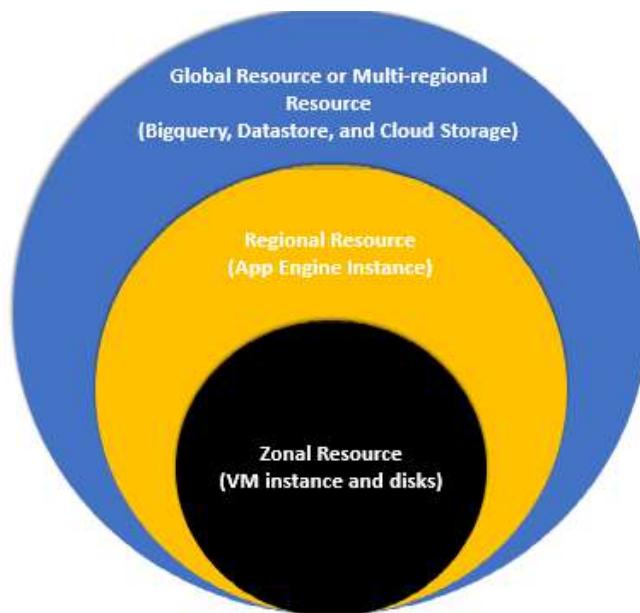
Serverless	Server-based
Serverless resources do not need to provision compute, memory, or network	Customer has to spin the virtual machine/cluster
Components for these services will be managed by Google Cloud Platform	Provision under network
The only requirement is to submit jobs	You will be paying for resources even they are idle
Google Cloud Platform will allocate resources during Job execution and deallocate when the job is complete	Customer is responsible for provisioning and de provisioning the resources

GCP Resources

What are GCP Resources?

- GCP consists of a set of physical assets, such as computers and hard disk drives, and virtual resources, such as virtual machines (VMs), that are contained in Google's data centers around the globe referred to as **GCP Resources**.
- Each data center location is in a **region**. Regions are available in Asia, Australia, Europe, North America, and South America. These locations are divided into regions and zones. You can choose where to locate your applications to meet your latency, availability and durability requirements.
- Each region is a collection of **Zones**, which are isolated from each other within the region.
- Each zone is identified by a name that combines a letter identifier with the name of the region. For example, zone a in the East Asia region is named asia-east1-a.
- Some resources can be accessed by any other resource, across regions and zones. These **global resources** include preconfigured disk images, disk snapshots, and networks.
- Some resources can be accessed only by resources that are located in the same region. These **regional resources** include static external IP addresses. Other resources can be accessed only by resources that are located in the same zone. These **zonal resources** include VM instances, their types, and disks.

GCP Resource Scope



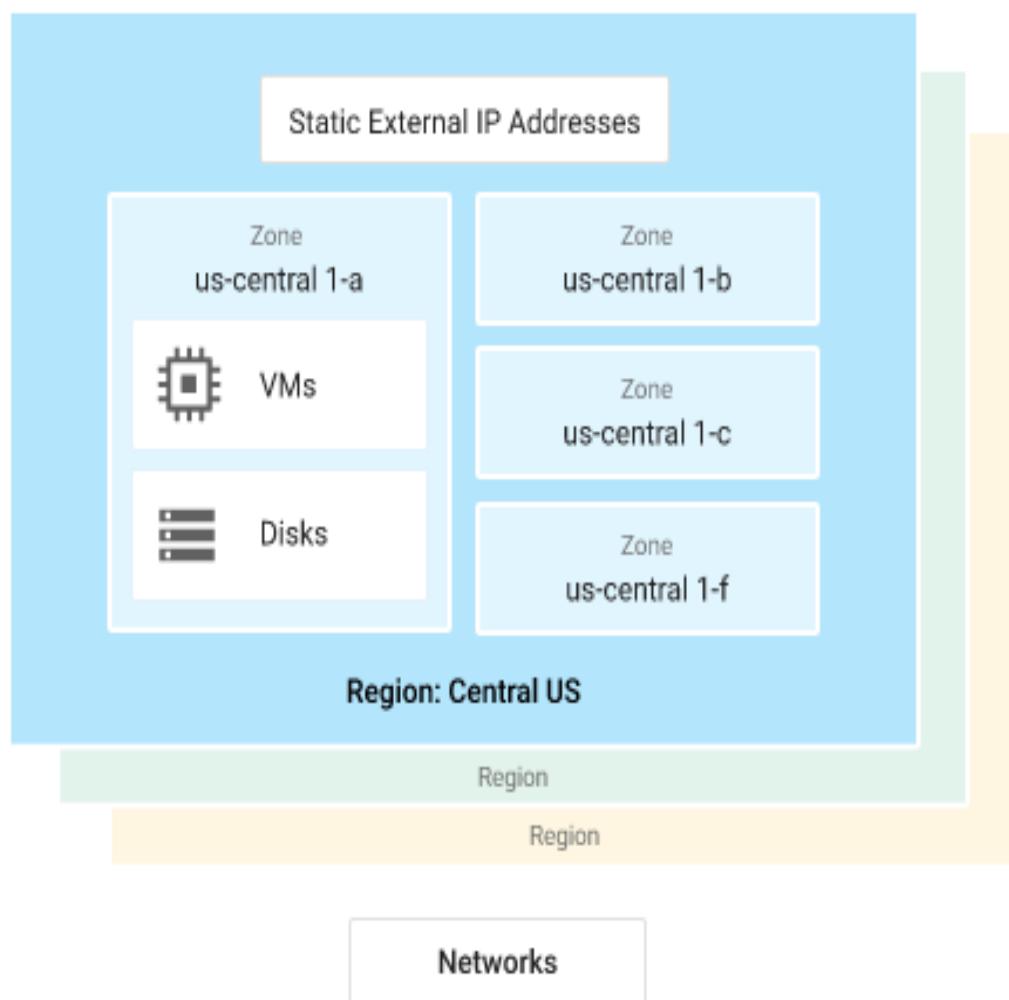
GCP Resource Hierarchy

The GCP resource hierarchy is organized as follows: -

- All resources (VMs, Storage Buckets, etc) are organized into Projects.
- These projects may be organized into folders, which can contain other folders.
- All folders and projects can be brought together under an organization node.
- Project folders and organization nodes are where policies can be defined
- Policies are inherited downstream and dictate who can access what resources
- Every resource must belong to a project and every must have a billing account associate with it

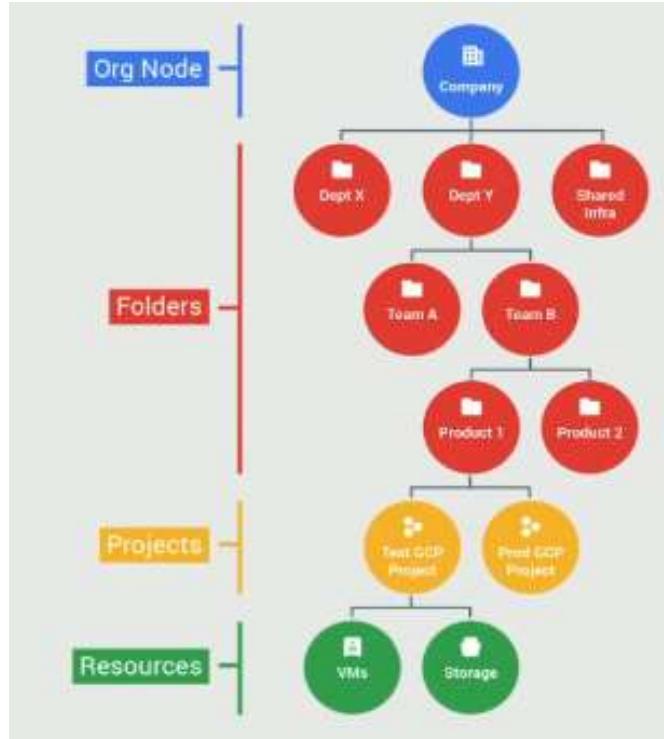
Google Cloud Platform

(Global Scope)



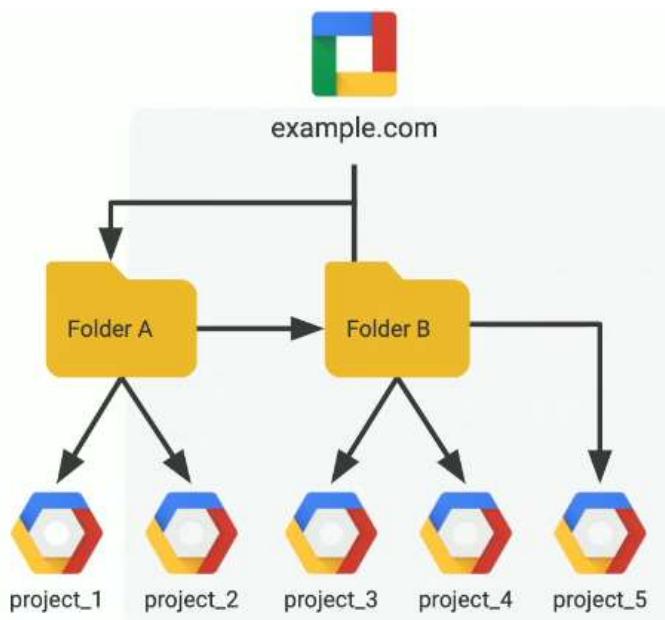
Resource hierarchy levels define the trust boundaries

- Group your resources according to your organization structure
- Levels of the hierarchy provide trust boundaries and resource isolation

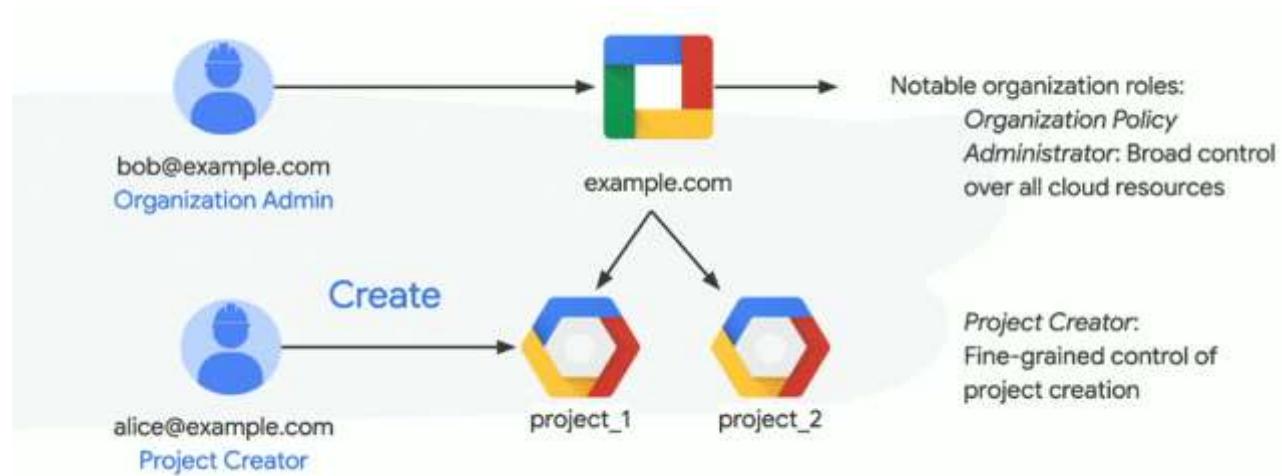


Folders offers flexible management

- Folders group projects under an organization
- Folders can contain projects, other folders, or both Use folders to assign policies



The Organization Node organize into Projects



GCP Resources with Project Association

In GCP everything is a resource. GCP services that you manage are associated with the project

In GCP Project, we can manage resources such as

- Track Resource and Quota Usage
- Enable Billing
- Managing Permissions and Credentials
- Enable Services & API
- Enable Different Devices



GCP Projects

All GCP services are associated with the Project. The Project attributes are:-

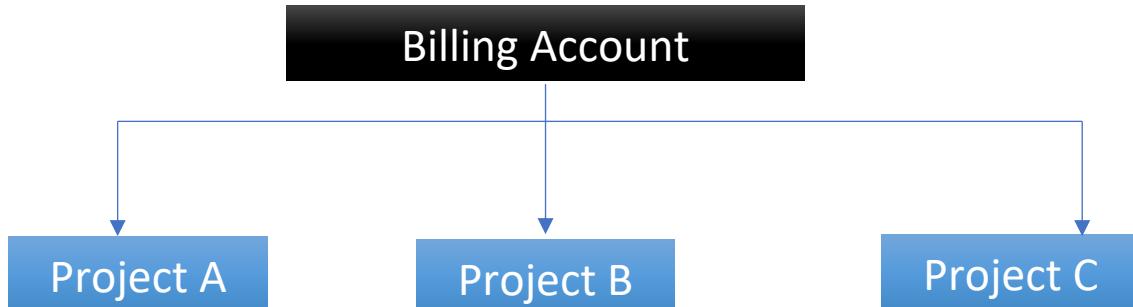
- **Project ID:** Should be Worldwide unique while creation and cannot be changed
- **Project Name:** Need not be unique, can be changed anytime
- **Project Number:** Worldwide unique, assigned by Google Cloud Platform and cannot be changed
- A project can have multiple owners, editors, viewers and billing administrators
- Provide least permission to the user based on the project requirement
- Different Roles & Permissions are: -
 - **Viewer:** Read only access
 - **Editor:** Deploy Applications, Modify Code, Configure Service
 - **Owner:** Invite Members, Remove Members, Delete Projects
 - **Billing Administrator:** Manage Billing, Add or Remove Admins

GCP Project Roles and Permissions



GCP Billing

- Resources are associated with the respective project
- Resource usage are billed at the project level
- Project level billing can be accumulated at the billing account level
- We can track all the cost of each project in one cloud billing account



In Short

All **resources** consumed associated with a **Project, Projects** are associated with **Billing Account** and **Billing happens Per-project**

GCP Pricing Model

- Per Second Billing
- Free monthly quota for the respective services like GAE, Bigquery, ...
- Usage based Charging for data storage, processing,
- Preemptible VM instances: Up to 75% discount
- Committed-use discounts: save up to 60%
- Sustained-use discounts: Up to 30%
- Custom machine types: save up to 50%
- Coldline: The archival storage with the speed of disk at the cost of tape
- All ingress traffic is free of cost. Egress traffic is charged

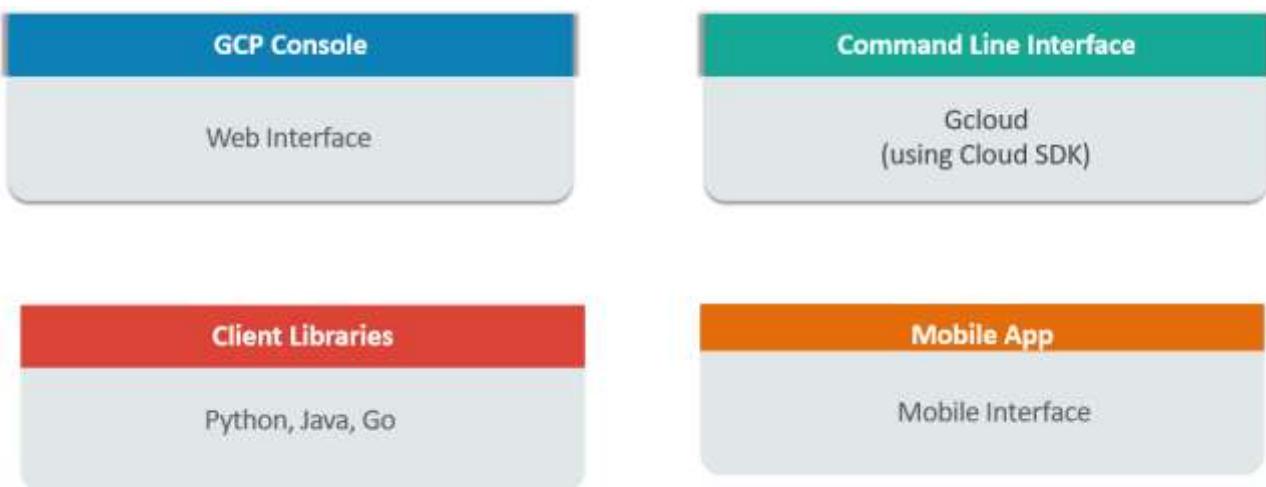
Steps to Enable Billing for Project

- Go to **API Console**
- Go to Project list and select the **Project** 
- Click on the hamburger menu icon in the top left corner and select **Billing**
- Click **Enable Billing** 
- Select the location, fill out the form, click Submit, and enable Billing
- Once the billing is enabled, the corresponding requests are billed based on the usage

Several ways of accessing the GCP resources

There are 4 ways to interact with GCP which includes: -

- GCP Console (Web Interface)
- Command Line Interface – Gcloud (using cloud SDK) , Work with gcloud , gsutil & Bq
- Client Libraries – Python, Java, Go, Rest APIs
- Mobile App



GCP Quiz

1. Does a cloud computing service let you scale your resource use up and down?

- A. Yes
- B. No

Correct Answer: A

2. To get resources from a cloud computing provider, is working with a person at the provider required?

- A. Yes
- B. No

Correct Answer: B

3. Choose fundamental characteristics of cloud computing. Mark all that are correct.

- A. Computing resources available on-demand and self-service
- B. Providers always dedicate physical resources to each customer
- C. Customers pay only for what they use or reserve
- D. Customers can scale their resource use up and down
- E. Customers are required to commit to multi-year contracts
- F. All resources are open-source based

Correct Answer: A, C, D

4. Which platforms are built on GCP?

- A. YouTube
- b. Google Search
- C. Both

Correct Answer: C

5. Google cloud platform billing is: -----

- A. Per hour basis
- B. Per minute basis
- C. Per second basis

Correct Answer: C

6. gcloud gives warning if you try to delete project?

- A. Yes
- B. No

Correct Answer: A

7. gsutil command to create a bucket is:

- A. gsutil mb
- B. gsutil create
- C. gsutil create bucket
- D. gsutil create storage

Correct Answer: A

Compute Services

Compute Engine

Highlights

- Google Compute Engine Instance is a Virtual Machines running in Google's Global Datacenter.
- Ideal for when you need compute control over your infrastructure and direct access to high performance hardware or need OS Level changes
- GCE instances are Ubuntu, Linux, and Windows OS that run on Google cloud
- Compute Engine is IaaS Option
- Configuration, Administration, Management, etc., Everything on your hand
- Selecting the appropriate configuration machine gets ready, so no need to buy machines or install OS, Dev Stack, Languages etc.,
- Use Case: Any Workload requiring a specific OS or OS configuration, currently deployed and on-premises software that you want to run in the cloud

What is Compute Engine? What are its Features?

Compute Engine is Google's infrastructure-as-a-service (IaaS) offering. It is also the building block for other services that run on top of this compute resource. The core functionality provided by Compute Engine is virtual machines. Virtual machines (VMs) in Google Cloud Platform are also known as instances.

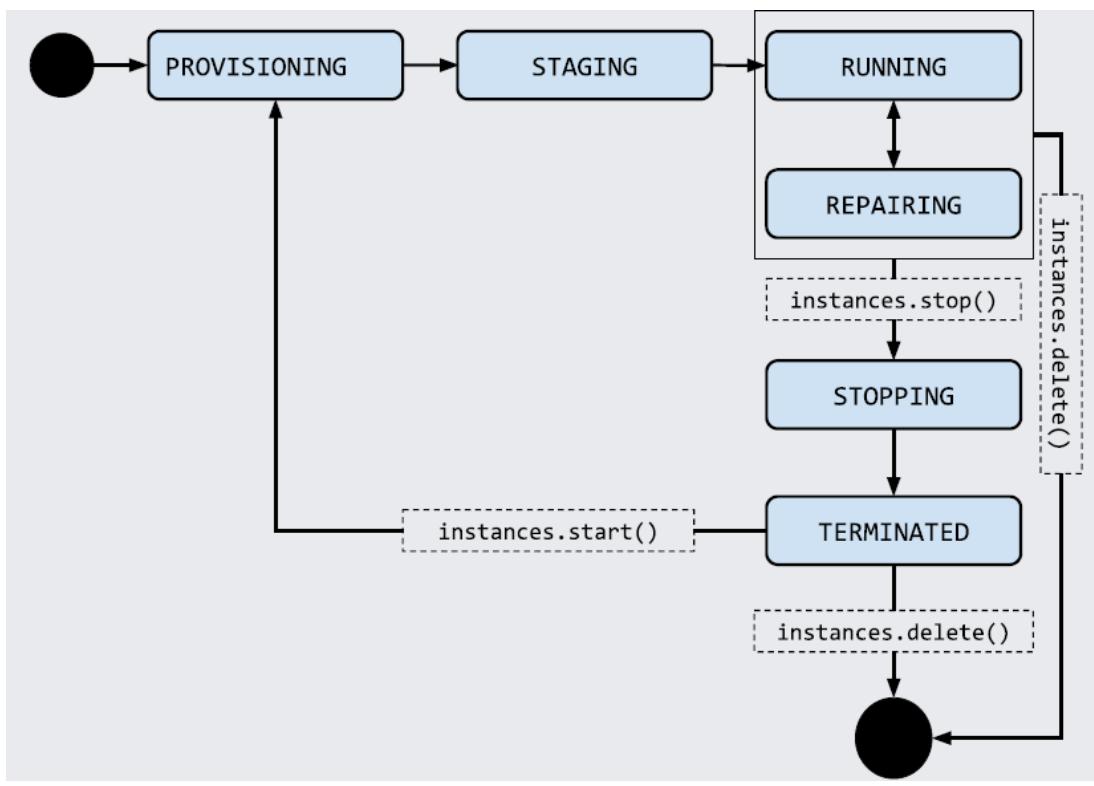
Features of Compute Engine

- Provides scalable and high-performance Virtual Machines, which Supports various operating systems such as: Windows, Ubuntu, Redhat Linux, etc
- Predefined and Custom machine with per second billing
- Preemptible VMs and Automatic Discounted prices for long-running workloads
- Global Load Balancing and CDN support & Google fibre network
- Live Migration and auto restart
- Persistent disk and balanced PDs
- OS Patch Management and sizing recommendation
- GPU and Container Support

Instances

What is an Instance?

- An instance is a virtual server in the cloud
- When creating a VM, you will need to specify several types of information such as machine type, availability status, and enhanced security controls.
- Instances are provisioned by specifying machine types, which are differentiated by the number of CPUs and the amount of memory allocated to the instance. Machine types that have balanced CPU and memory are known as standard instances.
- When creating an instance, you also specify a boot image. You can choose from a set of predefined boot images, or you can create your own based on one of the predefined boot images. When specifying a boot image, you can also specify standard persistent disk or SSD persistent disk as well as the size of the disk.



Instance Life Cycle

- An instance can transition through many instance states as part of its life cycle. When you first create an instance, Compute Engine provisions resources to start the instance. Next, the instance moves into staging, where it's prepared for the first boot, before it finally boots up and is considered running. A running instance can be stopped and restarted repeatedly during its life time

What are the different Stages in Life Cycle of an Instance?

The different stages in Life Cycle of an instance's are: -

- PROVISIONING - Resources are being allocated for the instance
- STAGING - Resources have been acquired and the instance is being prepared for
- RUNNING - The instance is booting up or running. You can connect to the instance shortly after it enters this state
- STOPPING - The instance is being stopped. This can be because a user has made a request to stop the instance or there was a failure
- REPAIRING - The instance is being repaired. This can happen because the instance encountered an internal error or the underlying machine is unavailable due to maintenance. During this time, the instance is unusable. If repair is successful, the instance returns to one of the above states.
- TERMINATED - A user stopped the instance, or the instance encountered a failure

What Instance Contains?

The Instance contains.....

- Variety of Instances
- Instance Templates
- Instance Groups
- Networks
- Content Delivery Network
- Autoscaling & Load Balancer
- Regions & Zones
- Firewall
- IP Address
- Images
- Disks

Images

What is an Image?

- An image in Compute Engine is a cloud resource that provides a reference to an immutable disk.
- Most of the public images can be used for no cost
- Some premium images may have an additional cost
- Custom images that you import to compute engine add no cost to your instance
- They incur an image storage charge when stored in your project (tar and gzipped file)
- Images are configured as a part of the instance template of a managed instance group

Image Types

- Public images for Linux and Windows Server that Google provides
- Private images that you create or import to Compute Engine
- Premium Images of other Operating Systems

Public Images

<input type="checkbox"/>	Name	Size	Created by	Family	Creation time
<input type="checkbox"/>	centos-6-v20171213	10 GB	CentOS	centos-6	Dec 14, 2017, 11:23:13 PM
<input type="checkbox"/>	centos-7-v20171213	10 GB	CentOS	centos-7	Dec 14, 2017, 11:24:16 PM
<input type="checkbox"/>	coreos-alpha-1618-0-0-v20171206	9 GB	CoreOS	coreos-alpha	Dec 7, 2017, 5:16:11 AM
<input type="checkbox"/>	coreos-beta-1590-2-0-v20171206	9 GB	CoreOS	coreos-beta	Dec 7, 2017, 5:16:36 AM
<input type="checkbox"/>	coreos-stable-1576-4-0-v20171206	9 GB	CoreOS	coreos-stable	Dec 7, 2017, 5:16:47 AM
<input type="checkbox"/>	cos-beta-63-10032-71-0	10 GB	Google	cos-beta	Dec 9, 2017, 12:01:00 AM
<input type="checkbox"/>	cos-dev-64-10176-7-0	10 GB	Google	cos-dev	Dec 8, 2017, 4:57:31 AM
<input type="checkbox"/>	cos-stable-62-9901-80-0	10 GB	Google		Dec 16, 2017, 3:20:16 AM
<input type="checkbox"/>	cos-stable-63-10032-71-0	10 GB	Google	cos-stable	Dec 8, 2017, 11:33:20 PM
<input type="checkbox"/>	debian-8-jessie-v20171213	10 GB	Debian	debian-8	Dec 14, 2017, 11:21:25 PM
<input type="checkbox"/>	debian-9-stretch-v20171213	10 GB	Debian	debian-9	Dec 14, 2017, 11:25:20 PM
<input type="checkbox"/>	rhel-6-v20171213	10 GB	RedHat	rhel-6	Dec 14, 2017, 11:28:11 PM
<input type="checkbox"/>	rhel-7-v20171213	10 GB	RedHat	rhel-7	Dec 14, 2017, 11:27:08 PM
<input type="checkbox"/>	sles-11-sp4-v20171211	10 GB	SUSE Linux Enterprise	sles-11	Dec 12, 2017, 12:57:18 AM
<input type="checkbox"/>	sles-12-sp2-sap-v20171211	10 GB	SUSE Linux Enterprise	sles-12-sp2-sap	Dec 11, 2017, 11:51:07 PM
<input type="checkbox"/>	sles-12-sp3-sap-v20171211	10 GB	SUSE Linux Enterprise	sles-12-sp3-sap	Dec 11, 2017, 11:52:46 PM
<input type="checkbox"/>	sles-12-sp3-v20171211	10 GB	SUSE Linux Enterprise	sles-12	Dec 12, 2017, 12:29:30 AM
<input type="checkbox"/>	sql2012-enterprise-windows-2012-r2-dc-v20171212	50 GB	Microsoft	sql-ent-2012-win-2012-r2	Dec 15, 2017, 3:22:23 AM
<input type="checkbox"/>	sql2012-standard-windows-2012-r2-dc-v20171212	50 GB	Microsoft	sql-std-2012-win-2012-r2	Dec 15, 2017, 2:36:59 AM
<input type="checkbox"/>	sql2012-web-windows-2012-r2-dc-v20171212	50 GB	Microsoft	sql-web-2012-win-2012-r2	Dec 15, 2017, 2:58:44 AM

Premium Images

- Additional per second charges, same charges across the world
- Red Hat Enterprise Linux, Microsoft Windows
- Charges based on the machine type used
- SQL Server images are charged per minute

Image Contents

- Boot loader
- Operating system
- File system structure
- Software
- Customizations

Images to create an instance

- A bundle of raw bytes used to create a pre-populated hard disk
- Master boot record and a bootable partition are required for the image to be bootable
- Upload the *.tar.gz to Cloud Storage and register it as an image
- Use it to create exact replicas of the original disk in any region of the platform

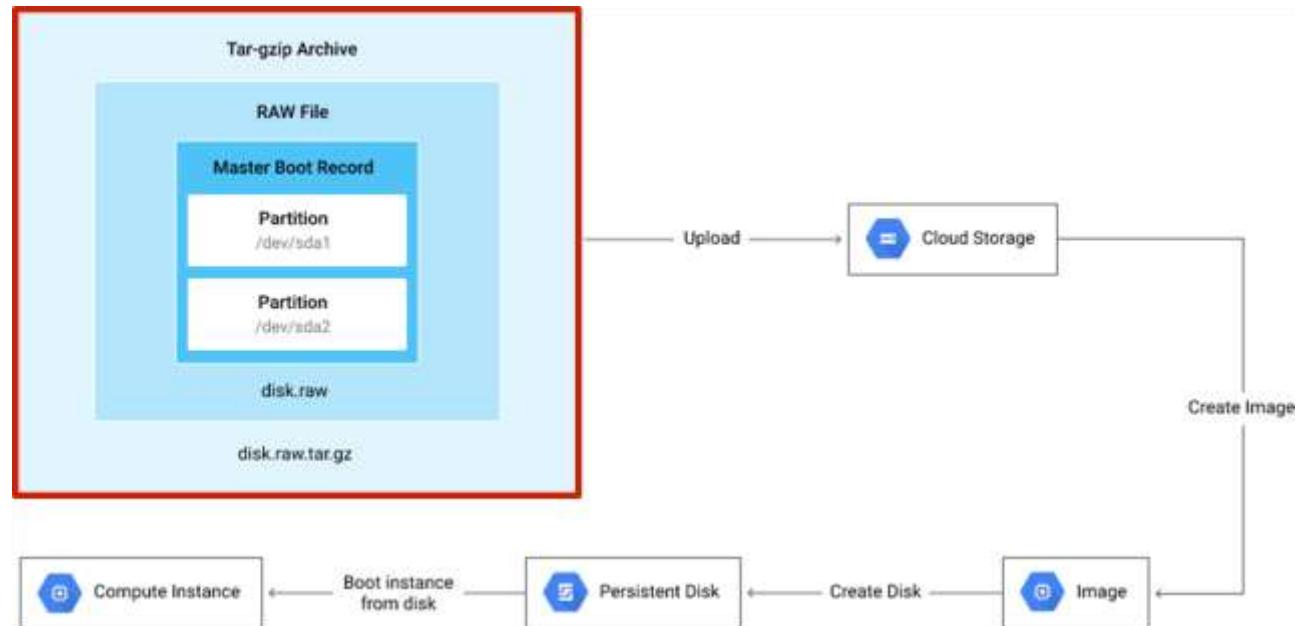
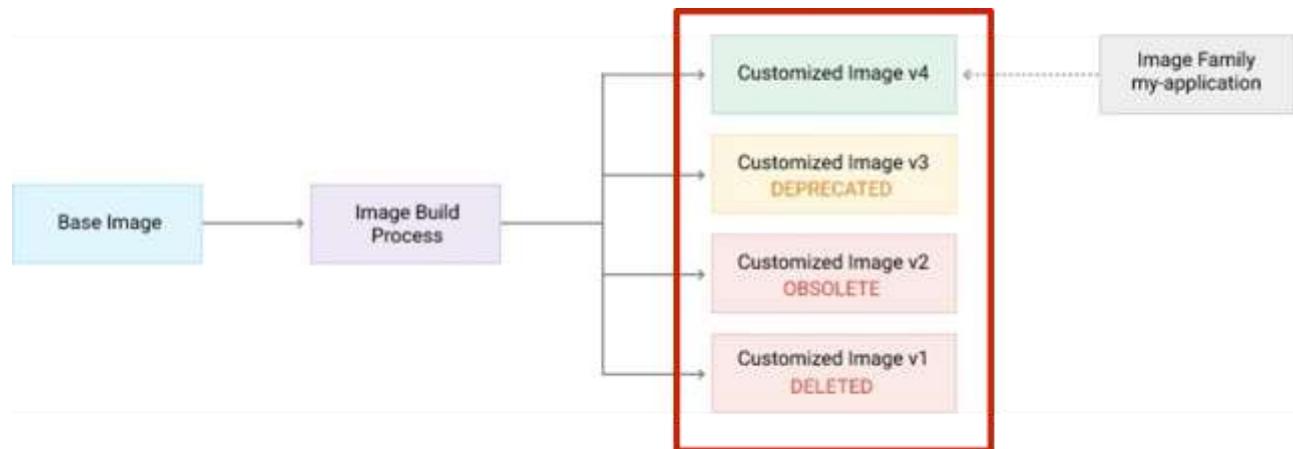
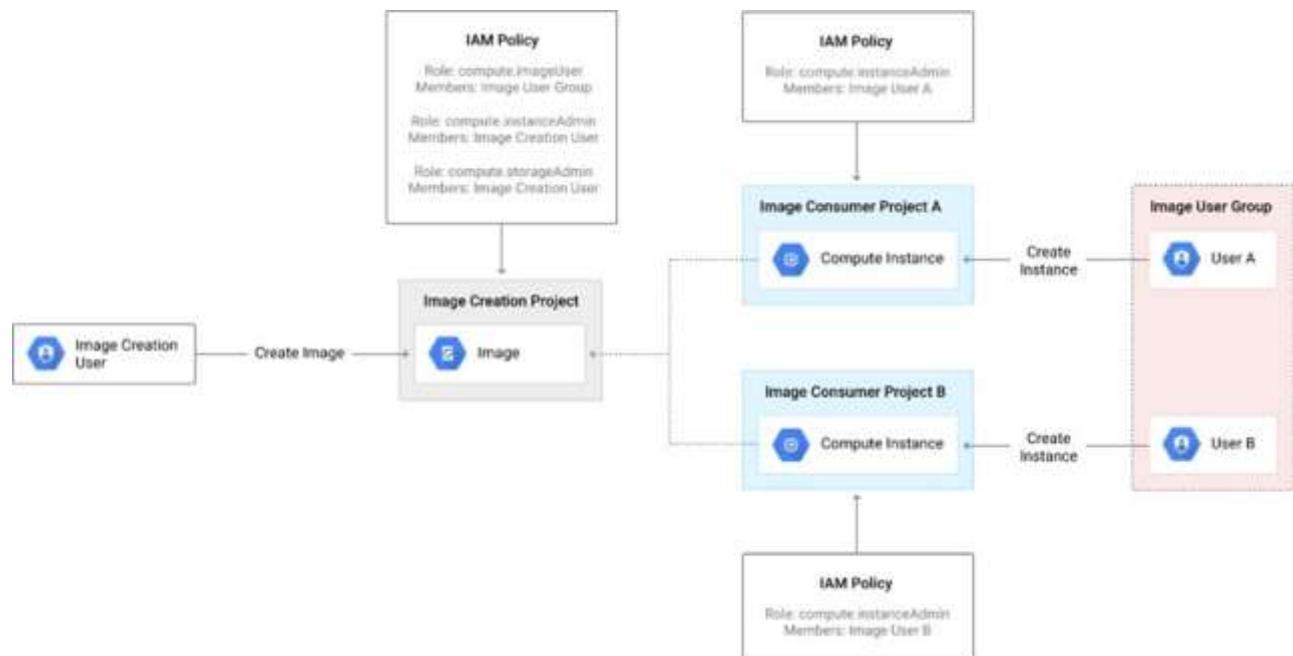


Image Life Cycle



State	Description
DEPRECATED	Images that are no longer the latest, but can still be launched by users. Users will see a warning at launch that they are no longer using the most recent image.
OBSOLETE	Images that should not be launched by users or automation. An attempt to create an instance from these images will fail. You can use this image state to archive images so their data is still available when mounted as a non-boot disk.
DELETED	Images that have already been deleted or are marked for deletion in the future. These cannot be launched, and you should delete them as soon as possible.

Sharing Images between Projects



Startup Scripts & Baking

- Startup Scripts are used to customize the instance created using a public image
- The script runs commands that deploys the application as it boots
- Script should be idempotent to avoid inconsistent or partially configured state
- Baking is a more efficient way to provision infrastructure
- Create a custom image with your configuration incorporated into the public image

Startup Scripts

- Longer for the instance to be ready & Startup scripts might fail and has to be idempotent
- Rollback has to be handled for applications and image separately
- The script will need to install dependencies during application deployment
- Each deployment might reference different versions if the latest version of the software has changed

Baking

- Much faster to go from boot to application readiness & Much more reliable for application deployments
- Version management is easier, easier to rollback to previous versions
- Fewer external dependencies during application bootstrap
- Scaling up creates instances with identical software versions

Instance Groups

Instance groups are clusters of VMs that are managed as a single unit. GCP supports two kinds of instance groups: managed and unmanaged.

Managed instance groups (MIGs) contain identically configured instances. The configuration is specified in an instance template.

Unmanaged instance groups are groups of VMs that may not be identical. They are not provisioned using an instance template. They are used only to support pre-existing cluster configurations for load balancing tasks. Unmanaged instance groups are not recommended for new configurations.

An instance template defines a machine type, boot disk image or container image, network settings, and other properties of a VM. The instance template can be used to create a single instance or a managed group of instances. Instance templates are global resources, so they can be used in any region. If you specify zonal resources in a template, however, they will have to be changed before using the template in another zone.

MIGs provide several advantages, including the following: -

- Maintaining a minimal number of instances in the MIG. If an instance fails, it is automatically replaced.
- Auto healing using application health checks. If the application is not responding as expected, the instance is restarted.
- Distribution of instances across a zone. This provides resiliency in case of zonal failures.
- Load balancing across instances in the group.
- Autoscaling to add or remove instances based on workload.
- Auto-updates, including rolling updates and canary updates.
- Rolling updates will update a minimal number of instances at a time until all instances are updated.
- Canary updates allow you to run two versions of instance templates to test the newer version before deploying it across the group.
- MIGs should be used when availability and scalability are required for VMs. This is usually the case in production systems.

Virtual Machines

When creating a VM, you will need to specify several types of information such as machine type, availability status, and enhanced security controls.

VMs can have one to eight network interfaces.

A VM also has a service account associated with it. A service account is a type of identity. Service accounts have permissions bound to them. The service account enables VMs to perform actions that would otherwise require a user to execute a task. For example, if a service account has permission to write to a Cloud Storage bucket, a process on the VM can write to that bucket using the service account.

If you need to ensure that your VMs run only on physical servers with other VMs from the same project, you can select sole tenancy when provisioning instances.

Virtual Machine Types

Broadly Machine types are categorized into Predefined and Custom

Custom Machines

- Suppose If none of the predefined machine types fit your workloads, use a custom machine type
- Save the cost of running on a machine which is more powerful than what you need
- Billed according to the number of vCPUs and the amount of memory used

Predefined

- Shared Core (For Small, Non-Resource intensive)
- General purpose | Standard
- Memory optimized | High Memory
- Compute optimized | High CPU

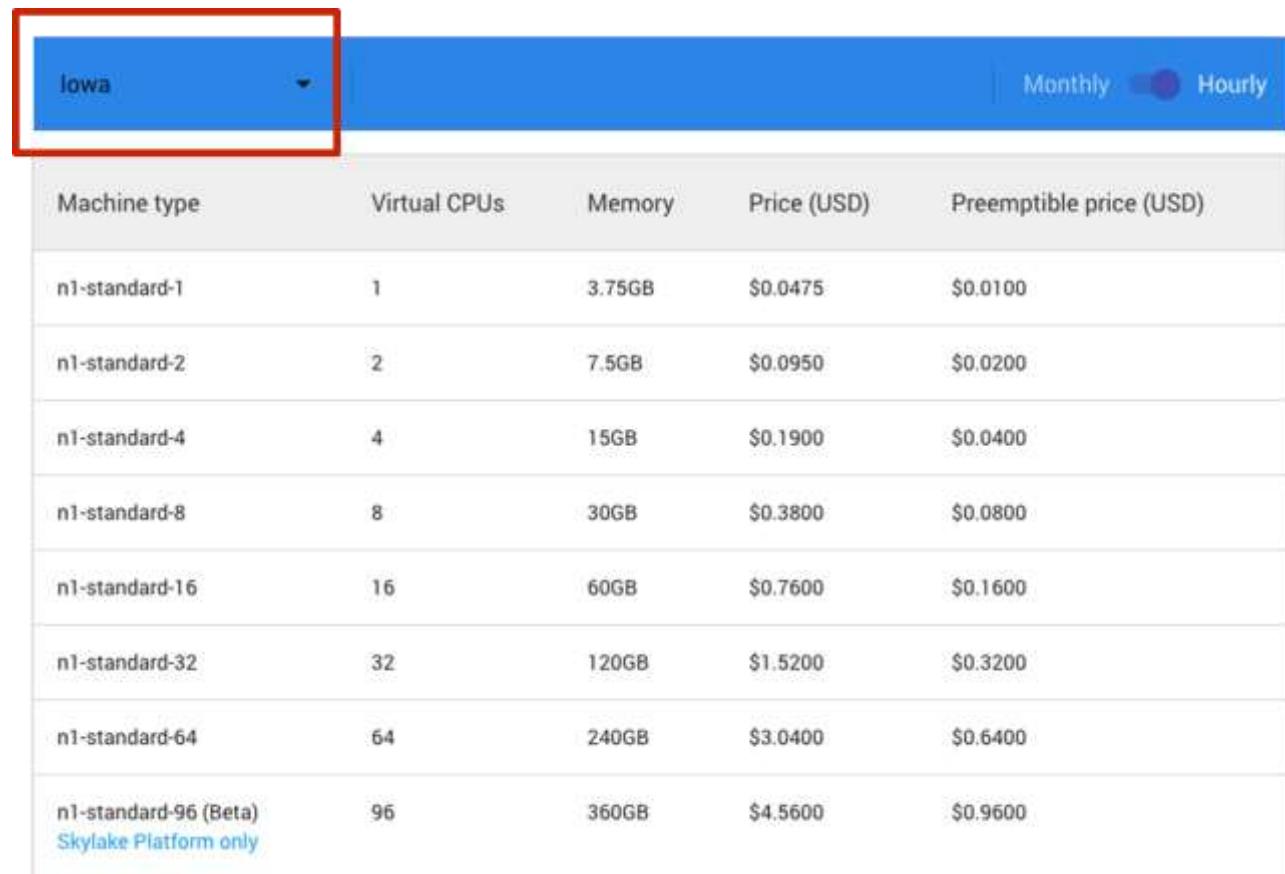
Shared Core

- Ideal for applications that do not require a lot of resources
- Small, non-resource intensive applications

Bursting

- f1-micro machine types offer bursting capabilities that allow instances to use additional physical CPU for short periods of time
- Bursting happens automatically when needed
- The instance will automatically take advantage of available CPU in bursts
- Bursts are not permanent, only possible periodically

General purpose | Standard

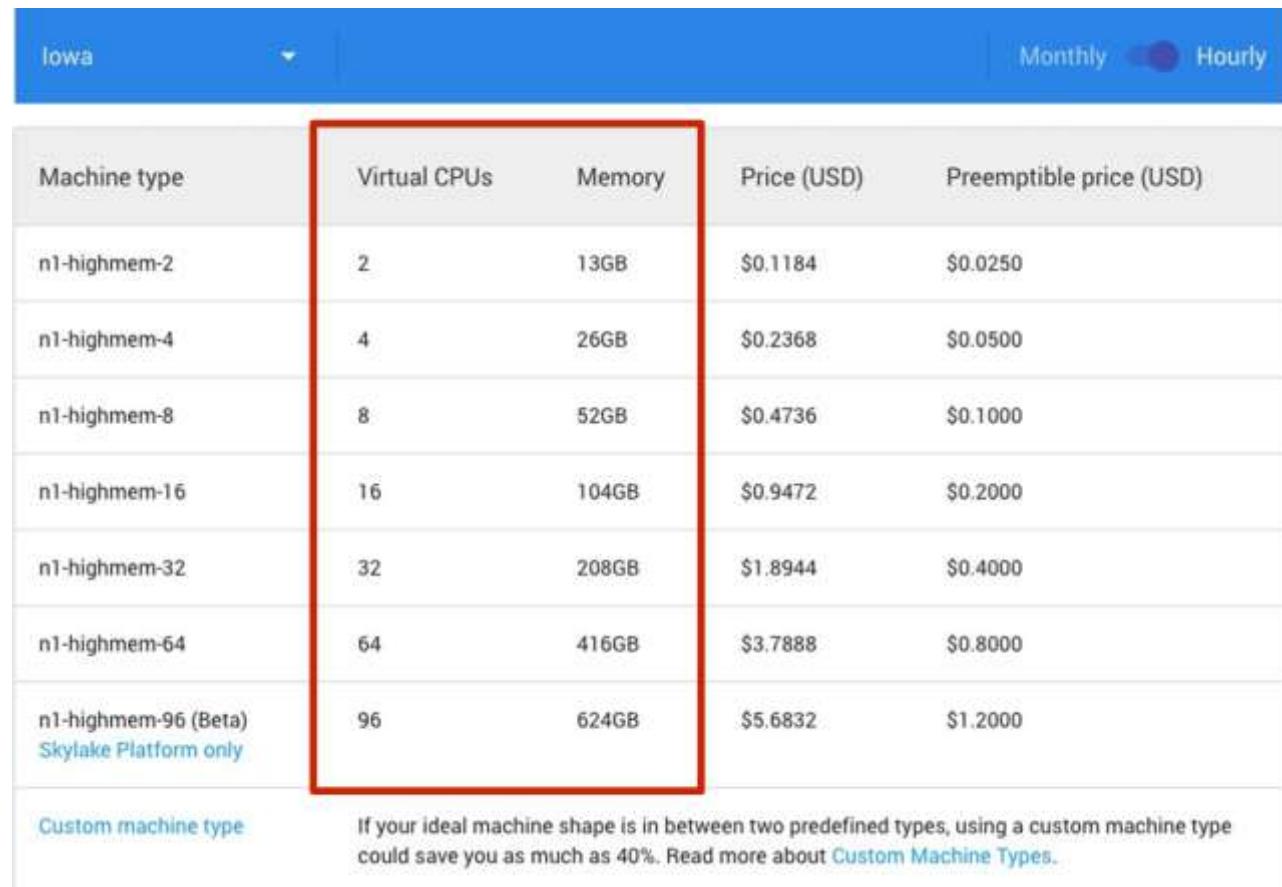


The screenshot shows a table of Google Cloud machine types for the Iowa region. The table includes columns for Machine type, Virtual CPUs, Memory, Price (USD), and Preemptible price (USD). The table lists eight machine types: n1-standard-1, n1-standard-2, n1-standard-4, n1-standard-8, n1-standard-16, n1-standard-32, n1-standard-64, and n1-standard-96 (Beta). The n1-standard-96 row has a note below it: "Skylake Platform only". The table is set against a background of a Google Cloud pricing interface with a blue header bar and a red box highlighting the region dropdown.

Machine type	Virtual CPUs	Memory	Price (USD)	Preemptible price (USD)
n1-standard-1	1	3.75GB	\$0.0475	\$0.0100
n1-standard-2	2	7.5GB	\$0.0950	\$0.0200
n1-standard-4	4	15GB	\$0.1900	\$0.0400
n1-standard-8	8	30GB	\$0.3800	\$0.0800
n1-standard-16	16	60GB	\$0.7600	\$0.1600
n1-standard-32	32	120GB	\$1.5200	\$0.3200
n1-standard-64	64	240GB	\$3.0400	\$0.6400
n1-standard-96 (Beta) Skylake Platform only	96	360GB	\$4.5600	\$0.9600

Memory optimized | High Memory

- More memory per vCPU as compared with regular machines
- Useful for tasks which require more memory as compared to processing
- 6.5 GB of RAM per core

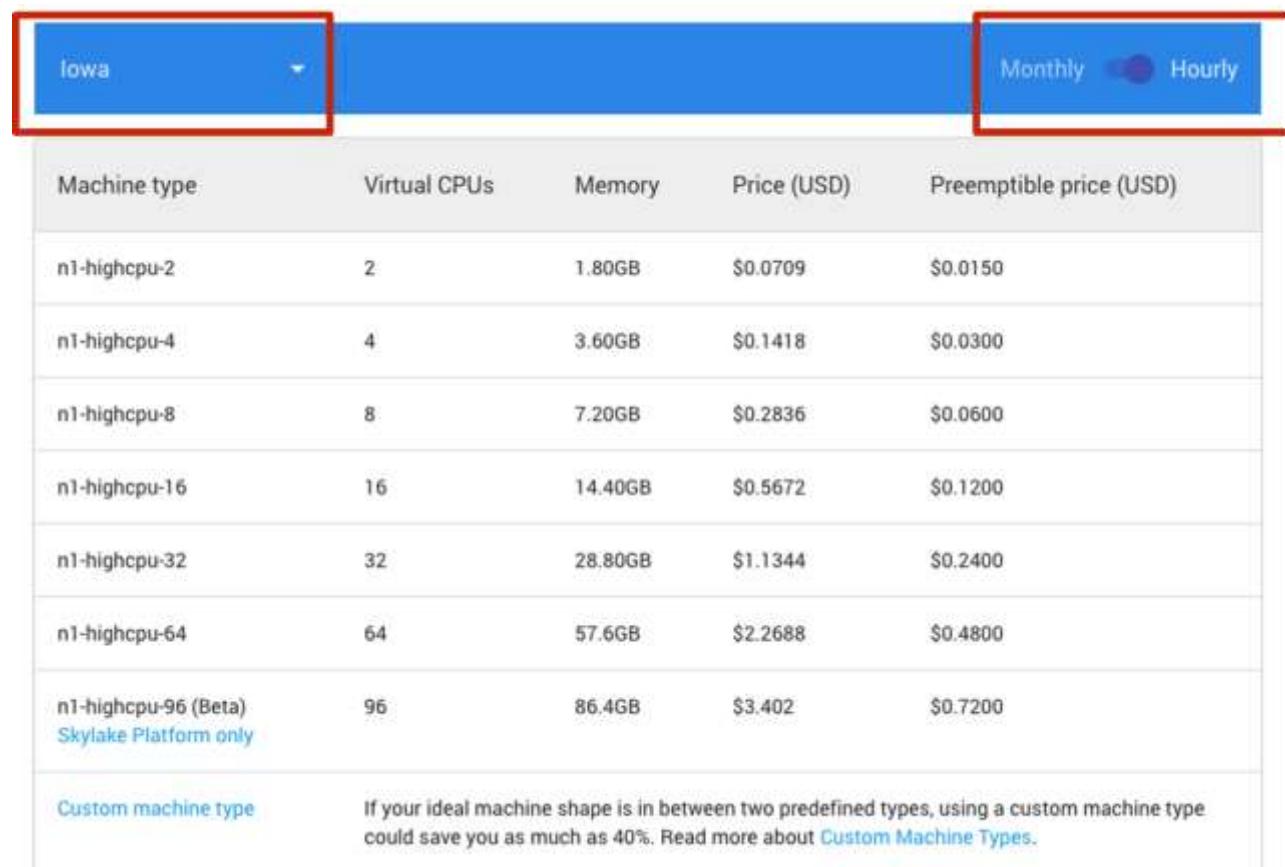


The screenshot shows a list of Google Cloud machine types categorized under "Memory optimized | High Memory". The table includes columns for Machine type, Virtual CPUs, Memory, Price (USD), and Preemptible price (USD). A red box highlights the "Virtual CPUs" and "Memory" columns. The "n1-highmem-96" row is marked as "Skylake Platform only". A note at the bottom encourages users to consider custom machine types for savings.

Machine type	Virtual CPUs	Memory	Price (USD)	Preemptible price (USD)
n1-highmem-2	2	13GB	\$0.1184	\$0.0250
n1-highmem-4	4	26GB	\$0.2368	\$0.0500
n1-highmem-8	8	52GB	\$0.4736	\$0.1000
n1-highmem-16	16	104GB	\$0.9472	\$0.2000
n1-highmem-32	32	208GB	\$1.8944	\$0.4000
n1-highmem-64	64	416GB	\$3.7888	\$0.8000
n1-highmem-96 (Beta) Skylake Platform only	96	624GB	\$5.6832	\$1.2000
Custom machine type	If your ideal machine shape is in between two predefined types, using a custom machine type could save you as much as 40%. Read more about Custom Machine Types .			

Compute optimized | High CPU

- More memory per vCPU as compared with regular machines



The screenshot shows a table of machine types for the "Iowa" region. The table includes columns for Machine type, Virtual CPUs, Memory, Price (USD), and Preemptible price (USD). The "n1-highcpu-96 (Beta)" row is marked as "Skylake Platform only". A note at the bottom explains that custom machine types can save up to 40%.

Machine type	Virtual CPUs	Memory	Price (USD)	Preemptible price (USD)
n1-highcpu-2	2	1.80GB	\$0.0709	\$0.0150
n1-highcpu-4	4	3.60GB	\$0.1418	\$0.0300
n1-highcpu-8	8	7.20GB	\$0.2836	\$0.0600
n1-highcpu-16	16	14.40GB	\$0.5672	\$0.1200
n1-highcpu-32	32	28.80GB	\$1.1344	\$0.2400
n1-highcpu-64	64	57.6GB	\$2.2688	\$0.4800
n1-highcpu-96 (Beta) Skylake Platform only	96	86.4GB	\$3.402	\$0.7200
Custom machine type	If your ideal machine shape is in between two predefined types, using a custom machine type could save you as much as 40%. Read more about Custom Machine Types .			

Right Sizing & Recommendations

- Compute Engine provides machine recommendations to help optimize resource utilization
- Automatically generated based on system metrics gathered by Stackdriver monitoring
- Uses last 8 days of data for recommendations
- Low or high CPU utilization? Use a machine type with fewer vCPUs or more vCPUs
- Low or high memory usage? Use a machine type with less or more memory

VM Discounts

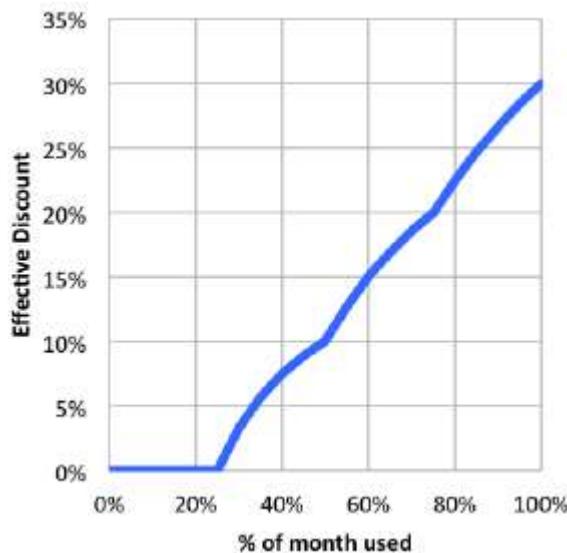
Google offers two additional kinds of discounts namely

- Sustained use
- Committed use

Sustained Use

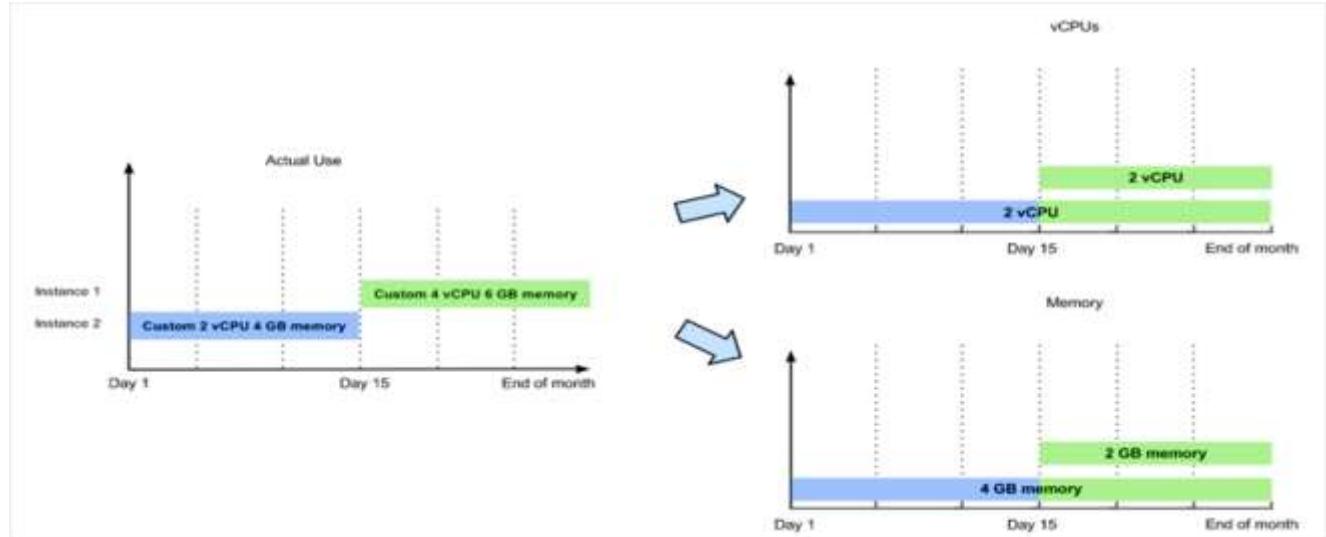
- Discounts for running a VM instance for a significant portion of the billing month
- Say you run an instance for 25% of the month, you get a discount for every incremental minute
- Applied automatically, no action to avail of these
- Applicable for instances created by Google Kubernetes Engine and Compute Engine
- Sustained use discount does NOT apply on certain machine types like E2 and A2

Usage Level (% of month)	% at which incremental is charged	Example incremental rate (USD/hour) for an n1-standard-1 instance
0%-25%	100% of base rate	\$0.0475
25%-50%	80% of base rate	\$0.0380
50%-75%	60% of base rate	\$0.0285
75%-100%	40% of base rate	\$0.0190



Sustained Discounts for Custom Machines

- Calculates sustained use discounts by combining memory and CPU usage
- Tries to combine resources to qualify for the biggest sustained usage discounts possible



Committed Use

- For workloads with predictable resource needs
- Commit for 1 year or 3 years
- Up to 70% discount based on machine type and GPUs
- Applicable for instances created by Google Kubernetes Engine and Compute Engine
- Committed use discounts does NOT apply to VMs created by App Engine flexible and Dataflow

Preemptible Virtual Machines

VMs can be standard or preemptible. Standard VMs continue to run until you shut them down or there is a failure in the VM. If a standard VM fails, it will be migrated to another physical server by default. Preemptible virtual machines run up to 24 hours before they are shut down by GCP. They can be shut down at any time before then as well. When a preemptible VM is shut down, processes on the instance will have 30 seconds to shut down gracefully.

Shielded VMs

Shielded VMs are instances with enhanced security controls, including the following: -

- Secure boot
- vTPM
- Integrity monitoring

Secure boot runs only software that is verified by digital signatures of all boot components using UEFI firmware features. If some software cannot be authenticated, the boot process fails.

Virtual Trusted Platform Module (vTPM) is a virtual module for storing keys and other secrets.

Integrity monitoring compares the boot measurements with a trusted baseline and returns true if the results match and false otherwise.

VMs can be created using the cloud console (console.cloud.google.com) or by using the GCP command-line utility, gcloud.

VMs can also be created automatically when using instance groups.

To what instances can you NOT attach a GPU?

- shared core machine types (micro & small)
- preemptible instances

When should Compute Engine be used for running containers?

When you need complete control over your container environment and your container orchestration tools.

When should Kubernetes Engine be used for running containers?

To simplify cluster management and container orchestration tasks so that you do not need to manage the underlying virtual machine instances.

Are persistent disks located in the same physical server as the computer?

No, the disks are parts of a separate storage service and physically located in separate hardware.

App Engine

What is App Engine?

App Engine is a serverless PaaS compute offering. With App Engine, users do not have to configure servers since it is a fully managed service. They provide application code that is run in the App Engine environment. There are two forms of App Engine: App Engine Standard and App Engine Flexible.

App Engine Standard

App Engine Standard is a PaaS product that allows developers to run their applications in a serverless environment. There are restrictions, however, on the languages that can be used to develop those applications. Currently, App Engine Standard provides the following language-specific runtime environments:

- Go
- Java
- PHP
- Node.js
- Python

Each instance of an application running in App Engine Standard has an instance class that is determined by the CPU speed and the amount of memory. The default instance class is F1, which has a 600 MHz CPU limit and 256 MB of memory. The largest instance class, B8, provides a 4.8 GHz CPU and 2048 MB of memory

App Engine Standard is available in two forms: first generation and second generation. Second-generation services offer more memory and more runtimes. There are no plans to deprecate first-generation services at this time.

First-generation App Engine Standard supports Python 2.7, Java 8, PHP 5.5, and Go 1.9. Network access is restricted in some cases. Specifically, Python 2.7, PHP 5.5, and Go 1.9 can access the network only through a URL Fetch API. Java 8 and Go 1.11 both have full access. Python 2.7, PHP 5.5, and Go 1.9 have no filesystem access, but Java 8 and Go 1.11 have read and write access to the /tmp directory.

The second-generation App Engine Standard improves on first-generation capabilities in that it also supports Python 3.7, PHP 7.2 and 7.3, Node.js 8 and 10, Java 11, and Go 1.11 and 1.12 (beta). All languages can use any extension or library, have full access to the network, and have read and write access to the /tmp directory.

App Engine Flexible

App Engine Flexible allows developers to customize their runtime environments by using Docker files. By default, the App Engine Flexible environment supports Java 8, Python 2.7 and Python 3.6, Node.js, Ruby, PHP, .NET core, and Go. The Java runtime does not support web-serving frameworks. Developers can customize these runtimes by providing custom Docker images or Dockerfiles. Since developers can provide custom containers, they can install custom packages and SSH into the containers. Developers can also specify how much CPU and memory is allocated to each instance.

App Engine Flexible provides health checks and automatically patches the underlying operating system. VMs are run in geographic regions specified by the GCP project containing the App Engine application. Also, VMs are restarted once a week. Operating system maintenance is performed at that time.

App Engine Standard Vs Flexible environments

Standard Environment

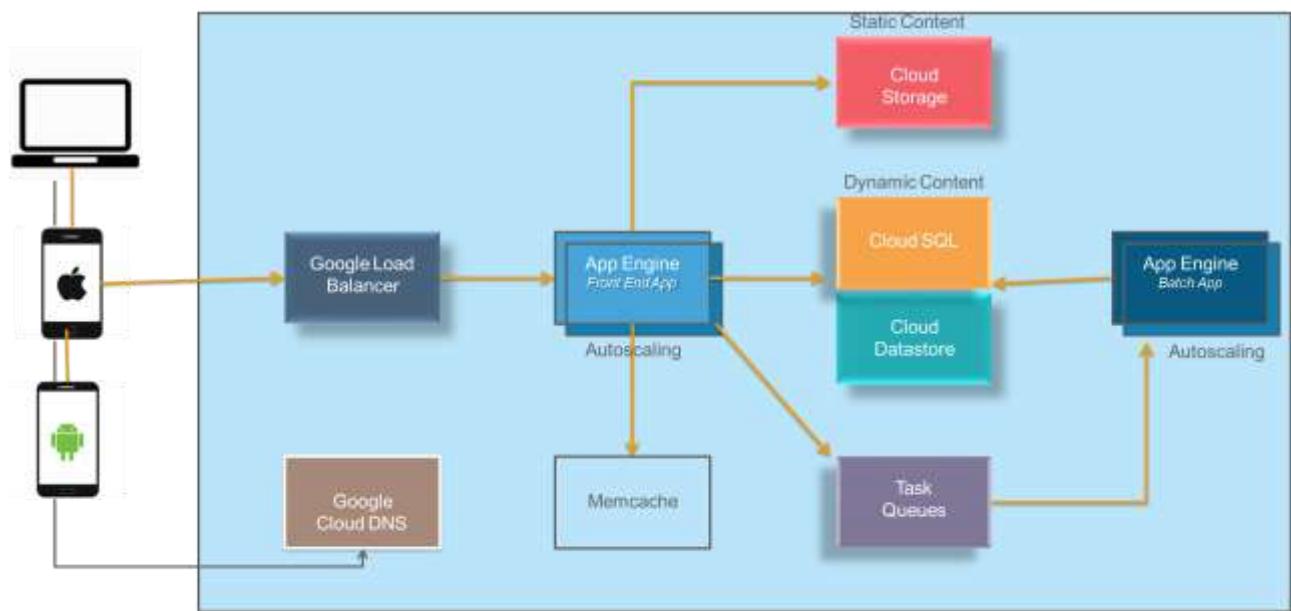
- Your application runs in a lightweight instance inside a sandbox that restricts what your application can do
- Cannot write to disk or use non-whitelisted binaries
- Limits CPU and memory options
- Best for stateless apps that respond quickly
- Scales from 0 to thousands of instances quickly

Flexible Environment

- Runs in a Docker container on Compute Engine
- Can use any language, write to disk and use any library
- Can use any compute engine machine type
- Allows for multiple processes
- Requires at least one instance, takes longer to scale

App Engine Architecture

- GCP App Engine application both development and deployment faster
- Management like memory cache, datastore, queuing up task is taken care by AppEngine environment
- Memcache: Memory cache is shared across multiple instances and it enables high speed access to information
- Task Queues: Manages the tasks on the servers. Frees long running tasks from front-end servers
- Built-in Load Balancer: Layer 3 & 7 Load Balancers are built-in



Kubernetes Engine

What is Kubernetes Engine?

Kubernetes Engine is a managed service providing *Kubernetes cluster management* and *Kubernetes container orchestration*. Kubernetes Engine allocates cluster resources, determines where to run containers, performs health checks, and manages VM lifecycles using Compute Engine instance groups.

Kubernetes Cluster Architecture

You can think of Kubernetes from the perspective of the VMs in the cluster or in terms of how applications function within the cluster.

Instances in Kubernetes

A Kubernetes cluster has two types of instances: cluster masters and nodes.

- The *cluster master* runs four core services that are part of the control plane: controller manager, API server, scheduler, and etcd.,
- The *controller manager* runs services that manage Kubernetes abstract components, such as deployments and replica sets.
- Applications interacting with the Kubernetes cluster make calls to the master using the *API server*. The API server also handles inter-cluster interactions.
- The *scheduler* is responsible for determining where to run pods, which are low-level compute abstractions that support containers.
- *etcd* is a distributed key-value store used to store state information across a cluster.
- *Nodes* are instances that execute workloads. They communicate with the cluster master through an agent called *kubelet*.

Kubernetes' Organizing Abstractions

Kubernetes introduces abstractions that facilitate the management of containers, applications, and storage services. Some of the most important are the following:

- Pods
- Services
- ReplicaSets
- Deployments
- PersistentVolumes
- StatefulSets
- Ingress

Pods are the smallest computation unit managed by Kubernetes. Pods contain one or more containers. Usually pods have just one container, but if the services provided by two containers are tightly coupled, then they may be deployed in the same container. For example, a pod may include a container running an extraction, transformation, and load process, as well as a container running ancillary services for decompressing and reformatting data. Multiple containers should be in the same pod only if they are functionally related and have similar scaling and lifecycle characteristics.

Pods are deployed to nodes by the scheduler. They are usually deployed in groups or replicas. This provides for high availability, which is especially needed with pods. Pods are ephemeral and may be terminated if they are not functioning properly. One of the advantages of Kubernetes is that it monitors the health of pods and replaces them if they are not functioning properly. Since multiple replicas of pods are run, pods can be destroyed without completely disrupting a service. Pods also support scalability. As load increases or decreases, the number of pods deployed for an application can increase or decrease.

Since pods are ephemeral, other services that depend on them need a way to discover them when needed. Kubernetes uses the service abstraction for this. A *service* is an abstraction with a stable API endpoint and stable IP address. Applications that need to use a service communicate with the API endpoints. A service keeps track of its associated pods so that it can always route calls to a functioning pod.

Google Cloud Certified Professional Cloud Architect Definitive Guide

A *ReplicaSet* is a controller that manages the number of pods running for a deployment. *Deployments* are a type of controller consisting of pods running the same version of an application. Each pod in a deployment is created using the same template, which defines how to run a pod. The definition is called a *pod specification*.

Kubernetes deployments are configured with a desired number of pods. If the actual number of pods varies from the desired state, for example if a pod is terminated for being unhealthy, then the ReplicaSet will add or remove pods until the desired state is reached.

Pods may need access to persistent storage, but since pods are ephemeral, it is a good idea to decouple pods that are responsible for computation from persistent storage, which should continue to exist even after a pod terminates. *PersistentVolumes* is Kubernetes' way of representing storage allocated or provisioned for use by a pod. Pods acquire access to persistent volumes by creating a *PersistentVolumeClaim*, which is a logical way to link a pod to persistent storage.

Pods as described so far work well for stateless applications, but when state is managed in an application, pods are not functionally interchangeable. Kubernetes uses the Stateful set abstraction, which is similar to a deployment. *StatefulSets* are used to designate pods as stateful and assign a unique identifier to them. Kubernetes uses these to track which clients are using which pods and to keep them paired.

An *Ingress* is an object that controls external access to services running in a Kubernetes cluster. An Ingress Controller must be running in a cluster for an Ingress to function.

Cloud Functions

What are Cloud Functions?

Cloud Functions is a serverless compute service well suited for event processing. The service is designed to respond to and execute code in response to events within the Google Cloud Platform. For example, if an image file is uploaded to Cloud Storage, a Cloud Function can execute a piece of code to transform the image or record metadata in a database. Similarly, if a message is written to a Cloud Pub/Sub topic, a Cloud Function may be used to operate on the message or invoke an additional process.

Events, Triggers, and Functions

Cloud Functions use three components: events, triggers, and functions. An *event* is an action that occurs in the GCP. Cloud Functions does not work with all possible events in the cloud platform; instead, it is designed to respond to five kinds of events.

- Cloud Storage
- Cloud Pub/Sub
- HTTP
- Firebase
- Stackdriver Logging

For each kind of event, there are different actions to which a Cloud Function can respond.

- *Cloud Storage* has upload, delete, and archive events.
- *Cloud Pub/Sub* recognizes message publishing events.
- *HTTP events* have five actions: GET, POST, PUT, DELETE, and OPTIONS.
- *Firebase* is a mobile application development platform that supports database triggers, remote configuration triggers, and authentication triggers.
- When a message is written to *Stackdriver Logging*, you can have it forwarded to Cloud Pub/Sub, and from there you can trigger a call to a cloud function.

A trigger in Cloud Functions is a specification of how to respond to an event. Triggers have associated functions. Currently, Cloud Functions can be written in Python 3, Go, and Node.js 8 and 10.

Compute – Interview | Exam Tips

Section -1 Compute Services

Highlights

GCP offers a number of compute services including:

- Compute Engine
- App Engine
- Kubernetes Engine
- Cloud Functions.

Compute Engine is GCP's infrastructure-as-a-service (IaaS) product. With Compute Engine, you have the greatest amount of control over your infrastructure relative to the other GCP compute services.

Kubernetes Engine is a container orchestration system, and Kubernetes Engine is a managed Kubernetes service. With Kubernetes Engine, Google maintains the cluster and assumes responsibility for installing and configuring the Kubernetes platform on the cluster. Kubernetes Engine deploys Kubernetes on managed instance groups.

App Engine is GCP's original platform-as-a-service (PaaS) offering. App Engine is designed to allow developers to focus on application development while minimizing their need to support the infrastructure that runs their applications. App Engine has two versions: App Engine Standard and App Engine Flexible.

Cloud Functions is a serverless, managed compute service for running code in response to events that occur in the cloud. Events are supported for Cloud Pub/Sub, Cloud Storage, HTTP events, Firebase, and Stackdriver Logging.

Compute Engine

Compute Engine is Google's infrastructure-as-a-service (IaaS) offering. The core functionality provided by Compute Engine is virtual machines (VMs). You have the greatest control over instances, but you also have the most management responsibility. Virtual machines (VMs) in Google Cloud Platform are also known as instances.

Instance Groups

Compute Engine supports provisioning single instances or groups of instances, known as instance groups. Instance groups are either managed or unmanaged instance groups. Managed instance groups (MIGs) consist of identically configured VMs; unmanaged instance groups allow for heterogeneous VMs, but they should be used only when migrating legacy clusters from on-premises data centers.

Benefits of MIG's

These benefits include the following:

- Auto healing based on application-specific health checks, which replace non-functioning instances
- Support for multizone groups that provide for availability in spite of zone-level failures
- Load balancing to distribute workload across all instances in the group
- Autoscaling, which adds or removes instances in the group to accommodate increases and decreases in workloads
- Automatic, incremental updates to reduce disruptions to workload processing

Virtual Machine Types

VMs can be created using the cloud console (console.cloud.google.com) or by using the GCP command-line utility, gcloud. VMs can also be created automatically when using instance groups. VMs can be standard or preemptible.

Standard VM's

Standard VMs continue to run until you shut them down or there is a failure in the VM. If a standard VM fails, it will be migrated to another physical server by default. Broadly Standard Virtual Machine types are categorized into Predefined and Custom.

Custom Machines

- Suppose If none of the predefined machine types fit your workloads, use a custom machine type
- Save the cost of running on a machine which is more powerful than what you need
- Billed according to the number of vCPUs and the amount of memory used

Predefined

- Shared Core (For Small, Non-Resource intensive)
- General purpose | Standard
- Memory optimized | High Memory
- Compute optimized | High CPU

Preemptible VM's

Preemptible virtual machines run up to 24 hours before they are shut down by GCP. They can be shut down at any time before then as well. When a preemptible VM is shut down, processes on the instance will have 30 seconds to shut down gracefully.

Shielded VMs

Shielded VM's are instances with enhanced security controls, including the following: -

- Secure boot
- vTPM
- Integrity monitoring

Secure boot runs only software that is verified by digital signatures of all boot components using UEFI firmware features. If some software cannot be authenticated, the boot process fails.

Virtual Trusted Platform Module (vTPM) is a virtual module for storing keys and other secrets.

Integrity monitoring compares the boot measurements with a trusted baseline and returns true if the results match and false otherwise.

VM's with Service Account

A VM also has a service account associated with it. A service account is a type of identity. Service accounts have permissions bound to them. The service account enables VMs to perform actions that would otherwise require a user to execute a task. For example, if a service account has permission to write to a Cloud Storage bucket, a process on the VM can write to that bucket using the service account. Managed instance groups and their features, such as autoscaling and health checks.

Provision Bigtable instances

Cloud Bigtable is a managed wide-column NoSQL database used for applications that require high-volume, low-latency writes. Bigtable has an HBase interface, so it is also a good alternative to using Hadoop HBase on a Hadoop cluster. Bigtable instances can be provisioned using the cloud console, the command-line SDK, and the REST API. When creating an instance, you provide an instance name, an instance ID, an instance type, a storage type, and cluster specifications.

Provision Cloud Dataproc

When provisioning Cloud Dataproc resources, you will specify the configuration of a cluster using the cloud console, the command-line SDK, or the REST API. When you create a cluster, you will specify a name, a region, a zone, a cluster mode, machine types, and an autoscaling policy. The cluster mode determines the number of master nodes and possible worker nodes. Master nodes and worker nodes are configured separately. For each type of node, you can specify a machine type, disk size, and disk type.

App Engine

App Engine is a platform-as-a-service (PaaS) compute offering. With App Engine, users do not have to configure servers. They provide application code that is run in the App Engine environment.

There are two forms of App Engine: App Engine Standard and App Engine Flexible. App Engine Standard uses language-specific sandboxes to execute your applications. App Engine Flexible lets you deploy containers, which you can create using Docker to customize the runtime environment. The additional services, such as the App Engine Cron Service and Task Queues.

Serverless services

Serverless services do not require conventional infrastructure provisioning but can be configured. You can configure App Engine using the app.yaml, cron.yaml, distpatch.yaml, or queue.yaml file. Cloud Functions can be configured using parameters to specify memory, region, timeout, and max instances. Cloud Dataflow parameters include job name, project ID, running, staging location, and the default and maximum number of worker nodes.

Kubernetes Engine

Kubernetes Engine is a managed service providing Kubernetes cluster management and Kubernetes container orchestration. Kubernetes Engine allocates cluster resources, determines where to run containers, performs health checks, and manages VM lifecycles using Compute Engine instance groups. It is well suited for applications built on microservices, but it also runs other containerized applications.

Containers are increasingly used to process workloads because they have less overhead than VMs and allow for finer-grained allocation of resources than VMs. A Kubernetes cluster has two types of instances: cluster masters and nodes.

Abstractions

Pods are the smallest computation unit managed by Kubernetes. Pods contain one or more containers. A ReplicaSet is a controller that manages the number of pods running for a deployment. A deployment is a higher-level concept that manages ReplicaSets and provides declarative updates. PersistentVolumes is Kubernetes' way of representing storage allocated or provisioned for use by a pod. Pods acquire access to persistent volumes by creating a PersistentVolumeClaim, which is a logical way to link a pod to persistent storage. StatefulSets are used to designate pods as stateful and assign a unique identifier to them. Kubernetes uses them to track which clients are using which pods and to keep them paired. An Ingress is an object that controls external access to services running in a Kubernetes cluster.

Cloud Functions

Cloud Functions is a serverless compute service well suited for event processing. The service is designed to respond to execute code in response to events within the Google Cloud Platform. Other issues to consider when designing infrastructure are managing state in distributed systems, data flows, and monitoring and alerting. Example a file being uploaded to Cloud Storage or a message being written to a Cloud Pub/Sub topic.

Terminology

Definitions of availability, reliability, and scalability

Availability is defined as the ability of a user to access a resource at a specific time. Availability is usually measured as the percentage of time a system is operational. Reliability is defined as the probability that a system will meet service-level objectives for some duration of time. Reliability is often measured as the mean time between failures. Scalability is the ability of a system to meet the demands of workloads as they vary over time.

When to use hybrid clouds and edge computing

The analytics hybrid cloud is used when transaction processing systems continue to run on premises and data is extracted and transferred to the cloud for analytic processing. A variation of hybrid clouds is an edge cloud, which uses local computation resources in addition to cloud platforms. This architecture pattern is used when a network may not be reliable or have sufficient bandwidth to transfer data to the cloud. It is also used when low-latency processing is required.

Messaging Services

Message brokers are services that provide three kinds of functionality: message validation, message transformation, and routing. Message validation is the process of ensuring that messages received are correctly formatted. Message transformation is the process of mapping data to structures that can be used by other services. Message brokers can receive a message and use data in the message to determine where the message should be sent. Routing is used when hub-and-spoke message brokers are used.

Distributed processing architectures

SOA is a distributed architecture that is driven by business operations and delivering business value. Typically, an SOA system serves a discrete business activity. SOAs are self-contained sets of services. Microservices are a variation on SOA architecture. Like other SOA systems, microservice architectures use multiple, independent components and common communication protocols to provide higher-level business services. Serverless functions extend the principles of microservices by removing concerns for containers and managing runtime environments.

Steps to migrate a data warehouse

At a high level, the process of migrating a data warehouse involves four stages:

- Assessing the current state of the data warehouse
- Designing the future state
- Migrating data, jobs, and access controls to the cloud
- Validating the cloud data warehouse

Purpose of Stackdriver Monitoring, Stackdriver Logging, and Stackdriver Trace

Stackdriver Metrics collect metrics on the performance of infrastructure resources and applications.

Stackdriver Logging is a service for storing and searching log data about events in infrastructure and applications. Stackdriver Trace is a distributed tracing system designed to collect data on how long it takes to process requests to services.

Section -2 Choosing Training and Serving Infrastructure

Single machines are useful for training small models

This includes when you are developing machine learning applications or exploring data using Jupyter Notebooks or related tools. Cloud Datalab, for example, runs instances in Compute Engine virtual machines.

Option of offloading some of the training load from CPUs to GPUs

GPUs have high-bandwidth memory and typically outperform CPUs on floating-point operations. GCP uses NVIDIA GPUs, and NVIDIA is the creator of CUDA, a parallel computing platform that facilitates the use of GPUs.

Distributing model training over a group of servers provides for scalability and improved availability

There are a variety of ways to use distributed infrastructure, and the best choice for you will depend on your specific requirements and development practices. One way to distribute training is to use machine learning frameworks that are designed to run in a distributed environment, such as TensorFlow.

Serving a machine learning model is the process of making the model available to make predictions for other services.

When serving models, you need to consider latency, scalability, and version management. Serving models from a centralized location, such as a data center, can introduce latency because input data and results are sent over the network. If an application needs real-time results, it is better to serve the model closer to where it is needed, such as an edge or IoT device.

Edge computing is the practice of moving compute and storage resources closer to the location at which they are needed

Edge computing devices can be relatively simple IoT devices, such as sensors with a small amount of memory and limited processing power. This type of device could be useful when the data processing load is light. Edge computing is used when low-latency data processing is needed—for example, to control machinery such as autonomous vehicles or manufacturing equipment. To enable edge computing, the system architecture has to be designed to provide compute, storage, and networking capabilities at the edge while services run in the cloud or in an on-premises data center for the centralized management of devices and centrally stored data.

Be able to list the three basic components of edge computing

Edge computing consists of edge devices, gateway devices, and the cloud platform. Edge devices provide three kinds of data: metadata about the device, state information about the device, and telemetry data. Before a device is incorporated into an IoT processing system, it must be provisioned. After a device is provisioned and it starts collecting data, the data is then processed on the device. After local processing, data is transmitted to a gateway. Gateways can manage network traffic across protocols. Data sent to the cloud is ingested by one of a few different kinds of services in GCP, including Cloud Pub/Sub, IoT Core MQTT, and Stackdriver Monitoring and Logging.

Know that an Edge TPU is a hardware device available from Google for implementing edge computing

This device is an application-specific integrated circuit (ASIC) designed for running AI services at the edge. Edge TPU is designed to work with Cloud TPU and Google Cloud services. In addition to the hardware, Edge TPU includes software and AI algorithms.

Google Cloud Certified Professional Cloud Architect Definitive Guide

Know that Cloud IoT is Google's managed service for IoT services

This platform provides services for integrating edge computing with centralized processing services. Device data is captured by the Cloud IoT Core service, which can then publish data to Cloud Pub/Sub for streaming analytics. Data can also be stored in BigQuery for analysis or used for training new machine learning models in Cloud ML. Data provided through Cloud IoT can also be used to trigger Cloud Functions and associated workflows.

Understand GPUs and TPUs

Graphic processing units are accelerators that have multiple arithmetic logic units (ALUs) that implement adders and multipliers. This architecture is well suited to workloads that benefit from massive parallelization, such as training deep learning models. GPUs and CPUs are both subject to the von Neumann bottleneck, which is the limited data rate between a processor and memory, and slow processing. TPUs are specialized accelerators based on ASICs and created by Google to improve training of deep neural networks. These accelerators are designed for the TensorFlow framework. TPUs reduces the impact of the von Neumann bottleneck by implementing matrix multiplication in the processor. Know the criteria for choosing between CPUs, GPUs, and TPUs

Compute Quiz

Section -1 Compute Services

1. You are consulting for a client that is considering moving some on-premises workloads to the Google Cloud Platform. The workloads are currently running on VMs that use a specially hardened operating system. Application administrators will need root access to the operating system as well. The client wants to minimize changes to the existing configuration. Which GCP compute service would you recommend?

- A. Compute Engine
- B. Kubernetes Engine
- C. App Engine Standard
- D. App Engine Flexible

Correct Answer: A

2. You have just joined a startup that analyses healthcare data and makes recommendations to healthcare providers to improve the quality of care while controlling costs. You have to comply with privacy regulations. A compliance consultant recommends that your startup control encryption keys used to encrypt data stored on cloud servers. You'd rather have GCP manage all encryption components to minimize your work and infrastructure management responsibilities. What would you recommend?

- A. Use default encryption enabled on Compute Engine instances.
- B. Use Google Cloud Key Management Service to store keys that you create and use them to encrypt storage used with Compute Engine instances.**
- C. Implement a trusted key store on premises, create the keys yourself, and use them to encrypt storage used with Compute Engine instances.
- D. Use an encryption algorithm that does not use keys.

Correct Answer: B

3. A colleague complains that the availability and reliability of GCP VMs is poor because their instances keep shutting down with them issuing shutdown commands. No instance has run for more than 24 hours without shutting down for some reason. What would you suggest your colleague check to understand why the instances may be shutting down?

- A. Make sure that the Stackdriver agent is installed and collecting metrics.
- B. Verify that sufficient persistent storage is attached to the instance.
- C. Make sure that the instance availability is not set to preemptible.**
- D. Ensure that an external IP address has been assigned to the instance.

Correct Answer: C

4. Your company is working on a government contract that requires all instances of VMs to have a virtual Trusted Platform Module. What Compute Engine configuration option would you enable or disable your instances?

- A. Trusted Module Setting
- B. Shielded VMs**
- C. Preemptible VMs
- D. Disable live migration

Correct Answer: B

5. You are leading a lift-and-shift migration to the cloud. Your company has several load-balanced clusters that use VMs that are not identically configured. You want to make as few changes as possible when moving workloads to the cloud. What feature of GCP would you use to implement those clusters in the cloud?

- A. Managed instance groups
- B. Unmanaged instance groups**
- C. Flexible instance groups
- D. Kubernetes clusters

Correct Answer: B

6. Your startup has a stateless web application written in Python 3.7. You are not sure what kind of load to expect on the application. You do not want to manage servers or containers if you can avoid it. What GCP service would you use?

A. Compute Engine

B. App Engine

C. Kubernetes Engine

D. Cloud Dataproc

Correct Answer: B

7. Your department provides audio transcription services for other departments in your company. Users upload audio files to a Cloud Storage bucket. Your application transcribes the audio and writes the transcript file back to the same bucket. Your process runs every day at midnight and transcribes all files in the bucket. Users are complaining that they are not notified if there is a problem with the audio file format until the next day. Your application has a program that can verify the quality of an audio file in less than two seconds. What changes would you make to the workflow to improve user satisfaction?

A. Include more documentation about what is required to transcribe an audio file successfully.

B. Use Cloud Functions to run the program to verify the quality of the audio file when the file is uploaded. If there is a problem, notify the user immediately.

C. Create a Compute Engine instance and set up a cron job that runs every hour to check the quality of files that have been uploaded into the bucket in the last hour. Send notices to all users who have uploaded files that do not pass the quality control check.

D. Use the App Engine Cron service to set up a cron job that runs every hour to check the quality of files that have been uploaded into the bucket in the last hour. Send notices to all users who have uploaded files that do not pass the quality control check.

Correct Answer: B

8. You have inherited a monolithic Ruby application that you need to keep running. There will be minimal changes, if any, to the code. The previous developer who worked with this application created a Dockerfile and image container with the application and needed libraries. You'd like to deploy this in a way that minimizes your effort to maintain it. How would you deploy this application?

- A. Create an instance in Compute Engine, install Docker, install the Stackdriver agent, and then run the Docker image.
- B. Create an instance in Compute Engine, but do not use the Docker image. Install the application, Ruby, and needed libraries. Install the Stackdriver agent. Run the application directly in the VM, not a container.
- C. Use App Engine Flexible to run the container image. App Engine will monitor as needed.**
- D. Use App Engine Standard to run the container image. App Engine will monitor as needed.

Correct Answer: C

9. You have been asked to give a presentation on Kubernetes. How would you explain the difference between the cluster master and nodes?

- A. Cluster masters manage the cluster and run core services such as the controller manager, API server, scheduler, and etcd. Nodes run workload jobs.**
- B. The cluster manager is an endpoint for API calls. All services needed to maintain a cluster are run on nodes.
- C. The cluster manager is an endpoint for API calls. All services needed to maintain a cluster are run on nodes, and workloads are run on a third kind of server, a runner.
- D. Cluster masters manage the cluster and run core services such as the controller manager, API server, scheduler, and etcd. Nodes monitor the cluster master and restart it if it fails.

Correct Answer: A

10. External services are not able to access services running in a Kubernetes cluster. You suspect a controller may be down. Which type of controller would you check?

- A. Pod
- B. Deployment
- C. Ingress Controller**
- D. Service Controller

Correct Answer: C

11. You are planning to run stateful applications in Kubernetes Engine. What should you use to support stateful applications?

- A. Pods
- B. StatefulPods
- C. StatefulSets**
- D. PersistentStorageSets

Correct Answer: C

12. Every time a database administrator logs into a Firebase database, you would like a message sent to your mobile device. Which compute service could you use that would minimize your work in deploying and running the code that sends the message?

- A. Compute Engine
- B. Kubernetes Engine
- C. Cloud Functions**
- D. Cloud Dataflow

Review Questions 89

Correct Answer: C

13. Your team has been tasked with deploying infrastructure for development, test, staging, and production environments in region us-west1. You will likely need to deploy the same set of environments in two additional regions. What service would allow you to use an Infrastructure as code (IaC) approach?

- A. Cloud Dataflow
- B. Deployment Manager**
- C. Identity and Access Manager
- D. App Engine Flexible

Correct Answer: B

14. An IoT startup collects streaming data from industrial sensors and evaluates the data for anomalies using a machine learning model. The model scales horizontally. The data collected is buffered in a server for 10 minutes. Which of the following is a true statement about the system?

- A. It is stateful.**
- B. It is stateless.
- C. It may be stateful or stateless, there is not enough information to determine.
- D. It is neither stateful nor stateless.

Correct Answer: A

15. Your team is designing a stream processing application that collects temperature and pressure measurements from industrial sensors. You estimate that for the initial release, the application will need 8 to 12 n1-highmem-32 instances. Someone on the team suggests using a Cloud Memorystore cache. What could that cache be used for?

- A. A SQL database
- B. As a memory cache to store state data outside of instances**
- C. An extraction, transformation, and load service
- D. A persistent object storage system

Correct Answer: B

16. A distributed application is not performing as well as expected during peak load periods. The application uses three microservices. The first of the microservices has the ability to send more data to the second service than the second service can process and keep up with. This causes the first microservice to wait while the second service processes data. What can be done to decouple the first service from the second service?

- A. Run the microservices on separate instances.
- B. Run the microservices in a Kubernetes cluster.
- C. Write data from the first service to a Cloud Pub/Sub topic and have the second service read the data from the topic.**
- D. Scale both services together using MIGs.

Correct Answer: C

17. A colleague has suggested that you use the Apache Beam framework for implementing a highly scalable workflow. Which Google Cloud service would you use?

- A. Cloud Dataproc
- B. Cloud Dataflow**
- C. Cloud Dataprep
- D. Cloud Memorystore

Correct Answer: B

18. Your manager wants more data on the performance of applications running in Compute Engine, specifically, data on CPU and memory utilization. What Google Cloud service would you use to collect that data?

- A. Cloud Dataprep
- B. Stackdriver**
- C. Cloud Dataproc
- D. Cloud Memorystore

Correct Answer: B

19. You are receiving alerts that CPU utilization is high on several Compute Engine instances. The instances are all running a custom C++ application. When you receive these alerts, you deploy an additional instance running the application. A load balancer automatically distributes the workload across all of the instances. What is the best option to avoid having to add servers manually when CPU utilization is high?

- A. Always run more servers than needed to avoid high CPU utilization.
- B. Deploy the instances in a MIG, and use autoscaling to add and remove instances as needed.**
- C. Run the application in App Engine Standard.
- D. Whenever you receive an alert, add two instances instead of one.

Correct Answer: B

20. A retailer has sales data streaming into a Cloud Pub/Sub topic from stores across the country. Each time a sale is made, data is sent from the point of sale to Google Cloud. The data needs to be transformed and aggregated before it is written to BigQuery. What service would you use to perform that processing and write data to BigQuery?

- A. Firebase
- B. Cloud Dataflow**
- C. Cloud Memorystore
- D. Cloud Datastore

Correct Answer: B

21. Custom machine type with 1 vCPU, 0.6 GB of total memory is a valid configuration

- A. True
- B. False**

Correct Answer: B

22. Custom machine type with 32 vCPUs, 29 GB of total memory is a valid configuration

A. True

B. False

Correct Answer: A

23. Why might a GCP customer use resources in several zones within a region?

A. For improved fault tolerance

B. For better performance

Correct Answer: A

24. Why might a GCP customer use resources in several regions around the world?

A. To bring their applications closer to users around the world, and for improved fault tolerance

B. To improve security

Correct Answer: A

25. You are choosing a technology for deploying applications, and you want to deliver them in lightweight, standalone, resource-efficient, portable packages. Which choice best meets those goals?

A. Hypervisors

B. Executable files

C. Virtual Machines

D. Containers

Correct Answer: D

26. You are classifying a number of your applications into workload types. Select the stateful applications in this list of applications. Choose all responses that are correct (2 correct responses)?

- A. Web server front end for your inventory system.
- B. A shopping application that saves user shopping cart data between sessions.**
- C. A gaming application that keeps track of user state persistently.**
- D. Image recognition application that identifies product defects from images.

Correct Answer: B, C

27. Google Compute Engine provides fine-grained control of costs. Which Compute Engine features provide this level of control?

- A. Billing budgets and alerts
- B. Per-second billing**
- C. Fully customizable virtual machines
- D. Managed instance groups
- E. Autoscaling groups

Correct Answer: B

28. You are developing a new solution and want to explore serverless application solutions. Which GCP compute services provide serverless compute resources that you can use with containers?

- A. Google Kubernetes Engine**
- B. Google App Engine
- C. Google Cloud Functions**
- D. Google Compute Engine

Correct Answer: A, C

29. You are deploying a containerized application, and you want maximum control over how containers are configured and deployed. You want to avoid the operational management overhead of managing a full container cluster environment yourself. Which Google Cloud Platform compute solution should you choose?

- A. Google Kubernetes Engine
- B. Google App Engine
- C. Google Compute Engine
- D. Google Cloud Functions

Correct Answer: A

30. You are considering deploying a solution using containers on Google cloud Platform. What Google Cloud solutions are available to you that will provide a managed compute platform with native support for containers?

- A. Compute Engine Autoscaling Groups
- B. Kubernetes Engine Clusters**
- C. Cloud Functions
- D. Container Registry

Correct Answer: B

31. You are ready to start work building an application in GCP. What GCP IAM hierarchy should you implement for this project?

- A. Create new projects for each of the component applications and create folders inside those for the resources.
- B. Create a new folder inside your organization and create projects inside that folder for the resources.
- C. Create a new organization for the project and create all projects and resources inside the new organization.
- D. Create new projects and resources inside departmental folders for the resources needed by the component applications.**

Correct Answer: D

32. You are developing a new product for a customer and need to implement control structures in GCP to help manage the GCP resources consumed by the product and the billing for the customer account. What steps should you take to manage costs for this product and customer?

- A. Configure quotas and limits for the product folders.
- B. Configure the billing account at the product folder level in the resource hierarchy.
- C. Set up budgets and alerts at the project folder level for the product.
- D. Configure the billing account for each project associated with the product.

Correct Answer:

33. You need to write some automated scripts to run periodic updates to the resources in your GCP environment. What tools can you install in your own computers to allow you to run these scripts?

- A. The Cloud Console Mobile app
- B. The Google Cloud Platform Console
- C. The Cloud Shell
- D. The Google Cloud SDK**

Correct Answer: D

34. One of the key characteristics of cloud computing is the concept of measured service. What is the primary customer benefit of the measured service aspect of cloud computing?

- A. You can get more resources as quickly as you need them.
- B. Resources can be allocated automatically.**
- C. You share resources from a large pool enabling economies of scale.**
- D. You pay only for the resources you consume.

Correct Answer: B, C

35. You can tag your Docker image for Container Registry with:

- A. docker run
- B. docker pull
- C. gcloud docker
- D. docker tag**

Correct Answer: D

36. Docker can build images automatically by reading the instructions from a ____.

- A. Container Registry
- B. requirements.txt
- C. Flask
- D. Dockerfile**

Correct Answer: D

37. What type of account is used by processes in containers in a Pod to communicate with the kube-apiserver running on the Kubernetes cluster master?

- A. A GCP Service account
- B. A Cloud IAM user account
- C. A Kubernetes Service Account**
- D. A Kubernetes normal user account

Correct Answer: C

38. You want to implement account controls that will allow you to grant junior admin users the ability to view details about production clusters and applications, and to be able to fully manage test and lab resources inside your GKE cluster environments. Which account control mechanism will provide you with the level of granular control that is required for this type of user?

- A. OAuth2
- B. Google Cloud IAM
- C. Kubernetes RBAC**
- D. Radius

Correct Answer: C

39. Which service can be used to synchronize enterprise user accounts to your GCP infrastructure, so that you can define Cloud IAM policies for them?

- A. Add your enterprise LDAP or Active Directory domain to Cloud Identity and Access Management.
- B. Deploy an LDAP or Microsoft Active Directory Server for your enterprise as a GCP compute instance.
- C. Migrate your enterprise user accounts to a Cloud Identity or G Suite organization.
- D. Use Google Cloud Directory Sync (GCDS) to synchronize the accounts between your enterprise directory and Cloud IAM.**

Correct Answer: D

40. When granting users, the permissions to access and perform actions on GCP resources using Google Cloud IAM, what type of GCP object needs to be created in order to grant roles to member accounts?

- A. A Cloud IAM user group
- B. A Cloud IAM resource binding
- C. A Cloud IAM Policy**
- D. A Cloud IAM permission

Correct Answer: C

41. Your organization is creating a new support team that will need to be able to view your Kubernetes Engine clusters and also access a number of Stackdriver features. You want to keep these users tightly controlled and need to make sure they only have the minimum level of access necessary to perform their jobs. What type of Cloud IAM role will you be using to assign these users the permissions they need?

- A. A Predefined role
- B. A Custom role**
- C. A Primitive role
- D. A Project role

Correct Answer: B

42. You have configured a new role using Kubernetes RBAC. You have supplied the action verbs you want for that role, and have added the users you require as subjects to the role. What is the next step you need to take in order to allow the users to perform those actions against the correct objects in your environment?

- A. Bind the RBAC role to the Kubernetes objects these users need to manage.**
- B. Add a GCP Service account that has the permissions necessary to manage the objects to the RBAC role.
- C. Add the RBAC role to the Cloud IAM admin policy for the clusters that contain the objects the users need to manage.
- D. Bind the RBAC role to a Kubernetes service account that has the permissions necessary to manage the objects required by the role.

Correct Answer: A

43. What security features can you control using Pod security Contexts? Choose all responses that are correct (2 correct responses).

- A. Enable App Armor, which uses security profiles to restrict individual programs actions.
- B. Limit access to GCP services and resources, for example to prevent users of the Pod from accessing Cloud Storage objects.
- C. Configure audit logging to redirect all Pod logs to an external webhook backend for persistent event auditing.
- D. Configure network access control lists, controlling which network endpoints can access, or be accessed by, the Pod.**
- E. Limit access to some Linux capabilities, for example to grant certain privileges to a process, but not all root user privileges.

Correct Answer: D & one more answer

44. You are troubleshooting an issue which happened in the last hour. You execute the command 'kubectl logs --since=3h demo-pod'. However, the events you are looking for do not appear in the output. What is the likely cause?

- A. The log file is older than 2 hours and is not included in the results.
- B. The file has been archived to Stackdriver and is no longer available locally.
- C. The log file contains more than 3 hours of data and it has been archived.
- D. The log file was greater than 100MB in size and it has been rotated.**

Correct Answer: D

45. You have a job that deploys a container that failed to run properly. How can you retrieve detailed information about the errors that happened inside the container?

- A. Execute kubectl describe in the Cloud Shell.
- B. Execute kubectl get logs in the Cloud Shell.
- C. In the GCP Console, view the GKE metrics in Stackdriver Monitoring.
- D. Execute kubectl logs in the Cloud Shell. --**

Correct Answer: D

46. You have configured both a Readiness probe and Liveness probe for a critical application component. Shortly after the application has started, the Pod is running, but the Readiness probe is failing. What effect does this have on the application's Pods and Services?

- A. Additional replica Pods are started until the Readiness Probe succeeds.
- B. The Pod will be restarted continuously until the Readiness Probe succeeds.
- C. The Service is disabled until the Readiness Probe succeeds.
- D. The Service will ignore the Pod until the Readiness Probe succeeds.**

Correct Answer: D

47. You need to monitor specific applications running on your production GKE Clusters. How should you create a logical structure for your application that allows you to selectively monitor the application's components using Stackdriver? Choose all responses that are correct (2 correct responses).

- A. Filter the Stackdriver logs using the application prefix.
- B. Filter the Stackdriver logs using the application name.
- C. Filter the Stackdriver logs using the Kubernetes labels.**
- D. Add Labels to the Pods in Kubernetes that identify the applications.**
- E. Add a prefix to all Pod names that identifies the application.

Correct Answer: C, D

48. You are using a liveness probe to monitor the state of an application. You have recently updated the application and are concerned that the boot time is now longer than before. What value must you change in the YAML manifest file to ensure that the probe is not initiated before the application is ready?

- A. spec.containers.livenessProbe.successThreshold.
- B. spec.containers.livenessProbe.initialDelaySeconds.--**
- C. spec.containers.livenessProbe.periodSeconds.
- D. spec.containers.touch.sleep

Correct Answer: B

49. You are unable to find any logs in Stackdriver for an issue with a Pod that occurred 2 months ago. You know that events were logged at the time for this issue. Why can you no longer find these logs?

- A. They have been archived by the Pod.
- B. They have been deleted by Stackdriver.**
- C. They have been exported to Cloud Storage.
- D. They have been exported to BigQuery.

Correct Answer: B

50. Kubernetes allows you to manage container clusters in multiple cloud providers.

A. True

False

Correct Answer: A

51. In Kubernetes, a node is: -----

- A. A worker machine**
- B. A machine that coordinates the scheduling and management of application containers on the cluster
- C. A tool for starting a Kubernetes cluster on a local machine

Correct Answer: A

52. What can you deploy on Kubernetes?

- A. Virtual Machines
- B. Containers**
- C. System Processes

Correct Answer: B

53. What command would you use to create a Deployment?

- A. kubectl get deployments
- B. kubectl get nodes
- C. kubectl run**

Correct Answer: C

54. How many VM instances do one need to run single App Engine application?

- A. 0**
- B. 1
- C. 2
- D. 3

Correct Answer: A

55. A startup is designing a data processing pipeline for its IoT platform. Data from sensors will stream into a pipeline running in GCP. As soon as data arrives, a validation process, written in Python, is run to verify data integrity. If the data passes the validation, it is ingested; otherwise, it is discarded. What services would you use to implement the validation check and ingestion?

- A. Cloud Storage and Cloud Pub/Sub
- B. Cloud Functions and Cloud Pub/Sub**
- C. Cloud Functions and BigQuery
- D. Cloud Storage and BigQuery

Correct Answer: B

56. Your finance department is migrating a third-party application from an on-premises physical server. The system was written in C, but only the executable binary is available. After the migration, data will be extracted from the application database, transformed, and stored in a BigQuery data warehouse. The application is no longer actively supported by the original developer, and it must run on an Ubuntu 14.04 operating system that has been configured with several required packages. Which compute platform would you use?

A. Compute Engine

B. Kubernetes Engine

C. App Engine Standard

D. Cloud Functions

Correct Answer: A

57. A team of developers has been tasked with rewriting the ETL process that populates an enterprise data warehouse. They plan to use a microservices architecture. Each microservice will run in its own Docker container. The amount of data processed during a run can vary, but the ETL process must always finish within one hour of starting. You want to minimize the amount of DevOps tasks the team needs to perform, but you do not want to sacrifice efficient utilization of compute resources. What GCP compute service would you recommend?

A. Compute Engine

B. Kubernetes Engine

C. App Engine Standard

D. Cloud Functions

Correct Answer: B

58. Your consulting company is contracted to help an enterprise customer negotiate a contract with a SaaS provider. Your client wants to ensure that they will have access to the SaaS service and it will be functioning correctly with only minimal downtime. What metric would you use when negotiating with the SaaS provider to ensure that your client's reliability requirements are met?

- A. Average CPU utilization
- B. A combination of CPU and memory utilization
- C. Mean time between failure**
- D. Mean time to recovery

Correct Answer: C

59. To ensure high availability of a mission-critical application, your team has determined that it needs to run the application in multiple regions. If the application becomes unavailable in one region, traffic from that region should be routed to another region. Since you are designing a solution for this set of requirements, what would you expect to include?

- A. Cloud Storage bucket
- B. Cloud Pub/Sub topic
- C. Global load balancer**
- D. HA VPN

Correct Answer: C

60. A startup is creating a business service for the hotel industry. The service will allow hotels to sell unoccupied rooms on short notice using the startup's platform. The startup wants to make it as easy as possible for hotels to share data with the platform, so it uses a message queue to collect data about rooms that are available for rent. Hotels send a message for each room that is available and the days that it is available. Room identifier and dates are the keys that uniquely identify a listing. If a listing exists and a message is received with the same room identifier and dates, the message is discarded. What are the minimal guarantees that you would want from the message queue?

- A. Route randomly to any instance that is building a machine learning model**
- B. Route based on the sensor identifier so identifiers in close proximity are used in the same model

- C. Route based on machine type so only data from one machine type is used for each model
- D. Route based on timestamp so metrics close in time to each other are used in the same model

Correct Answer: A

61. Sensors on manufacturing machines send performance metrics to a cloud-based service that uses the data to build models that predict when a machine will break down. Metrics are sent in messages. Messages include a sensor identifier, a timestamp, a machine type, and a set of measurements. Different machine types have different characteristics related to failures, and machine learning engineers have determined that for highest accuracy, each machine type should have its own model. Once messages are written to a message broker, how should they be routed to instances of a machine learning service?

- A. Route randomly to any instance that is building a machine learning model
- B. Route based on the sensor identifier so that identifiers in close proximity are used in the same model
- C. Route based on machine type so that only data from one machine type is used for each model**
- D. Route based on timestamp so that metrics close in time to one another are used in the same model

Correct Answer: C

62. As part of a cloud migration effort, you are tasked with compiling an inventory of existing applications that will move to the cloud. One of the attributes that you need to track for each application is a description of its architecture. An application used by the finance department is written in Java, deployed on virtual machines, has several distinct services, and uses the SOAP protocol for exchanging messages. How would you categorize this architecture?

- A. Monolithic
- B. Service-oriented architecture (SOA)**
- C. Microservice
- D. Serverless functions

Correct Answer: B

63. As part of a migration to the cloud, your department wants to restructure a distributed application that currently runs several services on a cluster of virtual machines. Each service implements several functions, and it is difficult to update one function without disrupting operations of the others. Some of the services require third-party libraries to be installed. Your company has standardized on Docker containers for deploying new services. What kind of architecture would you recommend?

- A. Monolithic
- B. Hub-and-spoke
- C. Microservices
- D. Pipeline architecture

Correct Answer: C

64. The CTO of your company is concerned about the rising costs of maintaining your company's enterprise data warehouse. Some members of your team are advocating to migrate to a cloud-based data warehouse such as BigQuery. What is the first step for migrating from the on-premises data warehouse to a cloud-based data warehouse?

- A. Assessing the current state of the data warehouse
- B. Designing the future state of the data warehouse
- C. Migrating data, jobs, and access controls to the cloud
- D. Validating the cloud data warehouse

Correct Answer: A

65. When gathering requirements for a data warehouse migration, which of the following would you include in a listing of technical requirements?

- A. Data sources, data model, and ETL scripts
- B. Data sources, data model, and business sponsor roles
- C. Data sources only
- D. Data model, data catalog, ETL scripts, and business sponsor roles

Correct Answer: A

66. In addition to concerns about the rising costs of maintaining an on-premises data warehouse, the CTO of your company has complained that new features and reporting are not being rolled out fast enough. The lack of adequate business intelligence has been blamed for a drop in sales in the last quarter. Your organization is incurring what kind of cost because of the backlog?

- A. Capital
- B. Operating
- C. Opportunity**
- D. Fiscal

Correct Answer: C

67. The data modelers who built your company's enterprise data warehouse are asking for your guidance to migrate the data warehouse to BigQuery. They understand that BigQuery is an analytical database that uses SQL as a query language. They also know that BigQuery supports joins, but reports currently run on the data warehouse are consuming significant amounts of CPU because of the number and scale of joins. What feature of BigQuery would you suggest they consider in order to reduce the number of joins required?

- A. Colossus filesystem
- B. Columnar data storage
- C. Nested and repeated fields**
- D. Federated storage

Correct Answer: C

68. While the CTO is interested in having your enterprise data warehouse migrated to the cloud as quickly as possible, the CTO is particularly risk averse because of errors in reporting in the past. Which prioritization strategy would you recommend?

- A. Exploiting current opportunities
- B. Migrating analytical workloads first
- C. Focusing on the user experience first
- D. Prioritizing low-risk use cases first**

Correct Answer: D

69. The enterprise data warehouse has been migrated to BigQuery. The CTO wants to shutdown the on-premises data warehouse but first wants to verify that the new cloud-based data warehouse is functioning correctly. What should you include in the verification process?

- A. Verify that schemas are correct and that data is loaded
- B. Verify schemas, data loads, transformations, and queries**
- C. Verify that schemas are correct, data is loaded, and the backlog of feature requests is prioritized
- D. Verify schemas, data loads, transformations, queries, and that the backlog of feature requests is prioritized

Correct Answer: B

70. A group of data scientists wants to pre-process a large dataset that will be delivered in batches. The data will be written to Cloud Storage and processed by custom applications running on Compute Engine instances. They want to process the data as quickly as possible when it arrives and are willing to pay the cost of running up to 10 instances at a time. When a batch is finished, they'd like to reduce the number of instances to 1 until the next batch arrives. The batches do not arrive on a known schedule. How would you recommend that they provision Compute Engine instances?

- A. Use a Cloud Function to monitor Stackdriver metrics, add instances when CPU utilization peaks, and remove them when demand drops.
- B. Use a script running on one dedicated instance to monitor Stackdriver metrics, add instances when CPU utilization peaks, and remove them when demand drops.
- C. Use managed instance groups with a minimum of 1 instance and a maximum of 10.**
- D. Use Cloud Dataproc with an autoscaling policy set to have a minimum of 1 instance and a maximum of 10.

Correct Answer: C

71. You are running a high-performance computing application in a managed instance group. You notice that the throughput of one instance is significantly lower than that for other instances. The poorly performing instance is terminated, and another instance is created to replace it. What feature of managed instance groups is at work here?

- A. Autoscaling
- B. Auto healing**
- C. Redundancy
- D. Eventual consistency

Correct Answer: B

72. A new engineer in your group asks for your help with creating a managed instance group. The engineer knows the configuration and the minimum and maximum number of instances in the MIG. What is the next thing the engineer should do to create the desired MIG?

- A. Create each of the initial members of the instance group using gcloud compute instance create commands
- B. Create each of the initial members of the instance group using the cloud console
- C. Create an instance template using the gcloud compute instance-templates create command**
- D. Create an instance template using the cbt create instance-template command

Correct Answer: C

73. Your team is migrating applications from running on bare-metal servers and virtual machines to running in containers. You would like to use Kubernetes Engine to run those containers. One member of the team is unfamiliar with Kubernetes and does not understand why they cannot find a command to deploy a container. How would you describe the reason why there is no deploy container command?

- A. Kubernetes uses pods as the smallest deployable unit, and pods have usually one but possibly multiple containers that are deployed as a unit.**
- B. Kubernetes uses deployments as the smallest deployable unit, and pods have usually one but possibly multiple containers that are deployed as a unit.
- C. Kubernetes uses replicas as the smallest deployable unit, and pods have usually one but possibly multiple containers that are deployed as a unit.

D. Kubernetes calls containers “pods,” and the command to deploy is kubectl deploy pod.

Correct Answer: A

74. A Kubernetes administrator wants to improve the performance of an application running in Kubernetes. They have determined that the four replicas currently running are not enough to meet demand and want to increase the total number of replicas to six. The name of the deployment is my-app-123. What command should they use?

- A. kubectl scale deployment my-app-123 --replicas 6
- B. kubectl scale deployment my-app-123 --replicas 2
- C. gcloud containers scale deployment my-app-123 --replicas 6
- D. gcloud containers scale deployment my-app-123 --replicas 2

Correct Answer: A

75. A Cloud Bigtable instance with one cluster is not performing as expected. The instance was created for analytics. Data is continuously streamed from thousands of sensors, and statistical analysis programs run continually in a batch. What would you recommend to improve performance?

- A. Use a write-optimized operating system on the nodes
- B. Use a read-optimized operating system on the nodes
- C. Add a cluster, run batch processing on one cluster, and have writes routed to the other cluster
- D. Add another node pool to the cluster in each zone that already has a node pool or that cluster

Correct Answer: C

76. A Cloud Dataproc cluster is running with a single master node. You have determined that the cluster needs to be highly available. How would you increase the number of master nodes to 3?

- A. Use the gcloud dataproc clusters update command with parameter --num-masters 3
- B. Use the gcloud dataproc clusters update command with parameter --add-masters 2
- C. Use the cbt dataproc clusters update command with parameter --add-masters 2
- D. The number of master nodes cannot be changed. A new cluster would have to be

Correct Answer: D

77. You have provisioned a Kubernetes Engine cluster and deployed an application. The application load varies during the day, and you have configured autoscaling to add replicas when CPU utilization exceeds 60 percent. How is that CPU utilization calculated?

- A. Based on all CPUs used by the deployment
- B. Based on all CPUs in the cluster
- C. Based on all CPU utilization of the most CPU-intensive pod
- D. Based on the CPU utilization of the least CPU-intensive pod

Correct Answer: A

78. A team of data scientists wants to run a Python application in a Docker container. They want to minimize operational overhead, so they decide to use App Engine. They want to run the application in a Python 3.4 environment. Which configuration file would they modify to specify that runtime?

- A. app.yaml
- B. queue.yaml
- C. dispatch.yaml
- D. cron.yaml

Correct Answer: A

79. Your team is experimenting with Cloud Functions to build a pipeline to process images uploaded to Cloud Storage. During the development stage, you want to avoid sudden spikes in Cloud Functions use because of errors in other parts of the pipeline, particularly the test code that uploads test images to Cloud Storage. How would you reduce the risk of running large numbers of Cloud Functions at one time?

- A. Use the --limit parameter when deploying the function
- B. Use the --max-instances parameter when deploying the function**
- C. Set a label with the key max-instances and the value of the maximum number of

instances

D. Set a language-specific parameter in the function to limit the number of instances

Correct Answer: B

80. Audit control requirements at your company require that all logs be kept for at least 365 days. You prefer to keep logs and log entries in Stackdriver logging, but you understand that logs with predefined retention periods of less than 1 year will require you to set up an export to another storage system, such as Cloud Storage. Which of the following logs would you need to set up exports for to meet the audit requirement?

- A. Admin activity audit logs
- B. System event audit logs
- C. Access transparency logs
- D. Data access audit logs**

Correct Answer: D

81. You would like to collect data on the memory utilization of instances running in a particular managed instance group. What Stackdriver service would you use?

- A. Stackdriver Debugger
- B. Stackdriver Logging
- C. Stackdriver Monitoring**
- D. Stackdriver Trace

Correct Answer: C

82. You would like to view information recorded by an application about events prior to the application crashing. What Stackdriver service would you use?

- A. Stackdriver Debugger
- B. Stackdriver Logging**
- C. Stackdriver Monitoring
- D. Stackdriver Trace

Correct Answer: B

83. Customers are complaining of long waits while your e-commerce site processes orders. There are many microservices in your order processing system. You would like to view information about the time each microservice takes to run. What Stackdriver service would you use?

- A. Stackdriver Debugger
- B. Stackdriver Logging
- C. Stackdriver Monitoring
- D. Stackdriver Trace**

Correct Answer: D

84. You have created a test environment for a group of business analysts to run several Cloud Dataflow pipelines. You want to limit the processing resources any pipeline can what execution parameter would you specify to limit processing resources?

- A. numWorkers
- B. maxNumWorkers**
- C. streaming
- D. maxResources

Correct Answer: B

Section -2 Choosing Training and Serving Infrastructure

1. You are in the early stages of developing a machine learning model using a framework that requires high-precision arithmetic and benefits from massive parallelization. Your data set fits within 32 GB of memory. You want to use Jupyter Notebooks to build the model iteratively and analyze results. What kind of infrastructure would you use?

- A. A TPU pod with at least 32 TPUs
- B. A single TPU
- C. A single server with a GPU**
- D. A managed instance group of 4 VMs with 64 GB of memory each

Correct Answer: C

2. You are developing a machine learning model that will predict failures in high-precision machining equipment. The equipment has hundreds of IoT sensors that send telemetry data every second. Thousands of the machines are in use in a variety of operating conditions. A year's worth of data is available for model training. You plan to use TensorFlow, a synchronous training strategy, and TPUs. Which of the following strategies would you use?

- A. MirroredStrategy
- B. CentralStorageStrategy
- C. MultiWorkerMirroredStrategy
- D. TPUStrategy**

Correct Answer: D

3. Your client has developed a machine learning model that detects anomalies in equity trading time-series data. The model runs as a service in a Google Kubernetes Engine (GKE) cluster deployed in the us-west-1 region. A number of financial institutions in New York and London are interested in licensing the technology, but they are concerned that the total time required to make a prediction is longer than they can tolerate. The distance between the serving infrastructure and New York is about 4,800 kilometres, and the distance to London is about 8,000 kilometres. This is an example of what kind of problem with serving a machine learning model?

- A. Overfitting
- B. Underfitting
- C. Latency**
- D. Scalability

Correct Answer: C

4. A study of global climate change is building a network of environmental sensors distributed across the globe. Sensors are deployed in groups of 12 sensors and a gateway. An analytics pipeline is implemented in GCP. Data will be ingested by Cloud Pub/Sub and analysed using the stream processing capabilities of Cloud Dataflow. The analysed data will be stored in BigQuery for further analysis by scientists. The bandwidth between the gateways and the GCP is limited and sometimes unreliable. The scientists have determined that they need the average temperature, pressure, and humidity measurements of each group of 12 sensors for a one-minute period. Each sensor sends data to the gateway every second. This generates 720 data points (12 sensors × 60 seconds) every minute for each of the three measurements. The scientists only need the one-minute average for temperature, pressure, and humidity. What data processing strategy would you implement?

- A. Send all 720 data points for each measurement each minute to a Cloud Pub/Sub message, generate the averages using Cloud Dataflow, and write those results to BigQuery.
- B. Average all 720 data points for each measurement each minute, send the average to a Cloud Pub/Sub message, and use Cloud Dataflow and write those results to BigQuery.**
- C. Send all 720 data points for each measurement each minute to a BigQuery streaming insert into a partitioned table.
- D. Average all 720 data points for each measurement each minute, send the average to a Cloud Pub/Sub message, and use Cloud Dataflow and write those results to BigQuery.

Correct Answer: B

5. Your DevOps team is deploying an IoT system to monitor and control environmental conditions in your building. You are using a standard IoT architecture. Which of the following components would you not use?

- A. Edge devices

B. Gateways

C. Repeater

D. Cloud platform services

Correct Answer: C

6. In the Google Cloud Platform IoT reference model, which of the following GCP services is used for ingestion?

A. Cloud Storage

B. BigQuery streaming inserts

C. Cloud Pub/Sub

D. Cloud Bigtable

Correct Answer: C

7. A startup is developing a product for autonomous vehicle manufacturers that will enable its vehicles to detect objects better in adverse weather conditions. The product uses a machine learning model built on TensorFlow. Which of the following options would you choose to serve this model?

A. On GKE using TensorFlow Training (TFJob)

B. On Compute Engine using managed instance groups

C. On Edge TPU devices in the vehicles

D. On GPUs in the vehicles

Correct Answer: C

8. In the Google Cloud Platform IoT reference model, which of the following GCP services is used for stream processing?

A. Cloud Storage

B. BigQuery streaming inserts

C. Cloud Pub/Sub

D. Cloud Dataflow

Correct Answer: D

9. You have developed a TensorFlow model using only the most basic TensorFlow operations and no custom operations. You have a large volume of data available for training, but by your estimates it could take several weeks to train the model using a 16 vCPU Compute Engine instance. Which of the following should you try instead?

- A. A 32 vCPU Compute Engine instance
- B. A TPU pod**
- C. A GKE cluster using on CPUs
- D. App Engine Second Generation

Correct Answer: B

10. You have developed a machine learning model that uses a specialized Fortran library that is optimized for highly parallel, high-precision arithmetic. You only have access to the compiled code and cannot make any changes to source code. You want to use an accelerator to reduce the training time of your model. Which of the following options would you try first?

- A. A Compute Engine instance with GPUs**
- B. A TPU pod
- C. A Compute Engine instance with CPUs only
- D. Cloud Functions

Correct Answer: A

Storage Services

GCP Storage Services

What is Storage? What are the Traditional Storage Tiers?

The Storage as the name suggests, is used to store data in the server
Traditionally we have used Volume & Vault.



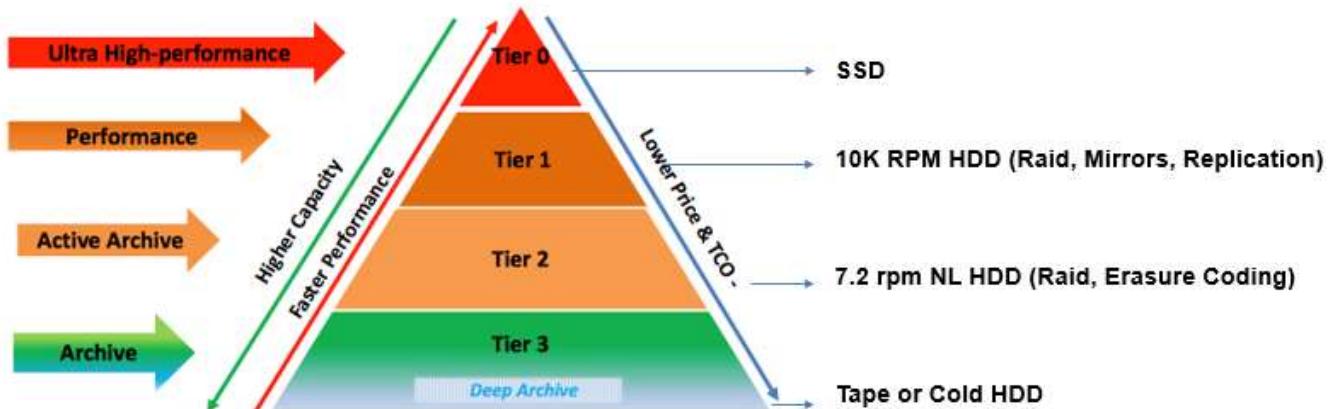
- ✓ Volume is used to store files for frequent use with limited capacity
- ✓ Vault is for archiving huge data for long term purpose



Traditional Storage Levels are represented by:

Tier 0 & Tier 1 → Volume Storage → For Short term storage

Tier 2 & Tier 3 → Vault Storage → For Long term storage



All the storage tiers are available in the cloud with low cost with high performance

Traditional Storage Vs Cloud Storage?

Impact of Cloud Storage on the following

- Low-Cost Savings
- (Reduced upto 50% of the annual budget)
- Higher Scalability
- Huge Capacity
- Easy Data Recovery
- Easy Data Accessibility
- Effective Security

Traditional Vs Cloud Data Storage



What is Cloud Storage?

The Cloud Storage is used to store data in the cloud, this data can be stored anywhere but content delivery on the other hand is used to cache data nearer to the user so as to provide low latency.

In cloud storage data should be highly secured, durable, scalable and easily managed which gives the value addition from the traditional storage

What are the different Storage options available in GCP?

The **Storage options** in GCP are: -

- Cloud Storage
- Cloud File Store

The **Database options** in GCP are: -

Relational

- Cloud SQL
- Cloud Spanner

Non-Relational

- Cloud Bigtable
- Cloud Fire Store
- Cloud Data Store

GCP offers three NoSQL databases: Bigtable, Datastore, and Cloud Firestore. All three are well suited to storing data that requires flexible schemas. Cloud Bigtable is a wide-column NoSQL database. Cloud Firestore and Cloud Datastore are document NoSQL

Memory Management

- Cloud Memorystore

Datawarehouse Management

- Cloud Bigquery

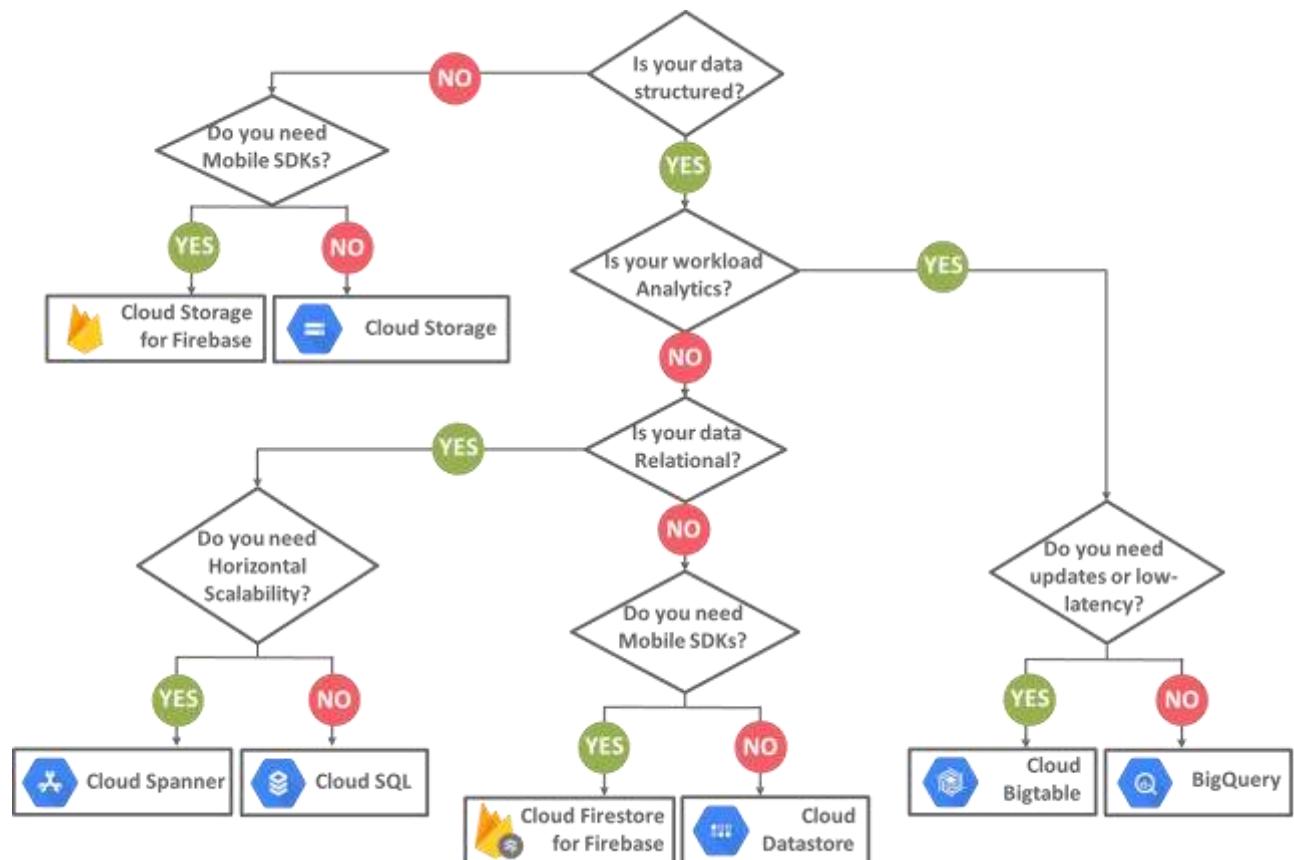
GCP Storage Use Case

	Cloud Datastore	Cloud Bigtable	Cloud Storage	Cloud SQL	Cloud Spanner	BigQuery
Type	NoSQL document	NoSQL wide column	Blobstore	Relational SQL for OLTP	Relational SQL for OLTP	Relational SQL for OLAP
Best for	Semi-structured application data, durable key-value data	"Flat" data, Heavy read/write, events, analytical data	Structured and unstructured binary or object data	Web frameworks, existing applications	Large-scale database applications (> ~2 TB)	Interactive querying, offline analytics
Use cases	Getting started, App Engine applications	AdTech, Financial and IoT data	Images, large media files, backups	User credentials, customer orders	Whenever high I/O, global consistency is needed	Data warehousing

GCP Storage Technical Details

	Cloud Datastore	Bigtable	Cloud Storage	Cloud SQL	Cloud Spanner	BigQuery
Type	NoSQL document	NoSQL wide column	Blobstore	Relational SQL for OLTP	Relational SQL for OLTP	Relational SQL for OLAP
Transactions	Yes	Single-row	No	Yes	Yes	No
Complex queries	No	No	No	Yes	Yes	Yes
Capacity	Terabytes+	Petabytes+	Petabytes+	Up to ~10 TB	Petabytes	Petabytes+
Unit size	1 MB/entity	~10 MB/cell ~100 MB/row	5 TB/object	Determined by DB engine	10,240 MiB/row	10 MB/row

GCP Storage Decision Making Chart



Cloud Storage (Object Storage)

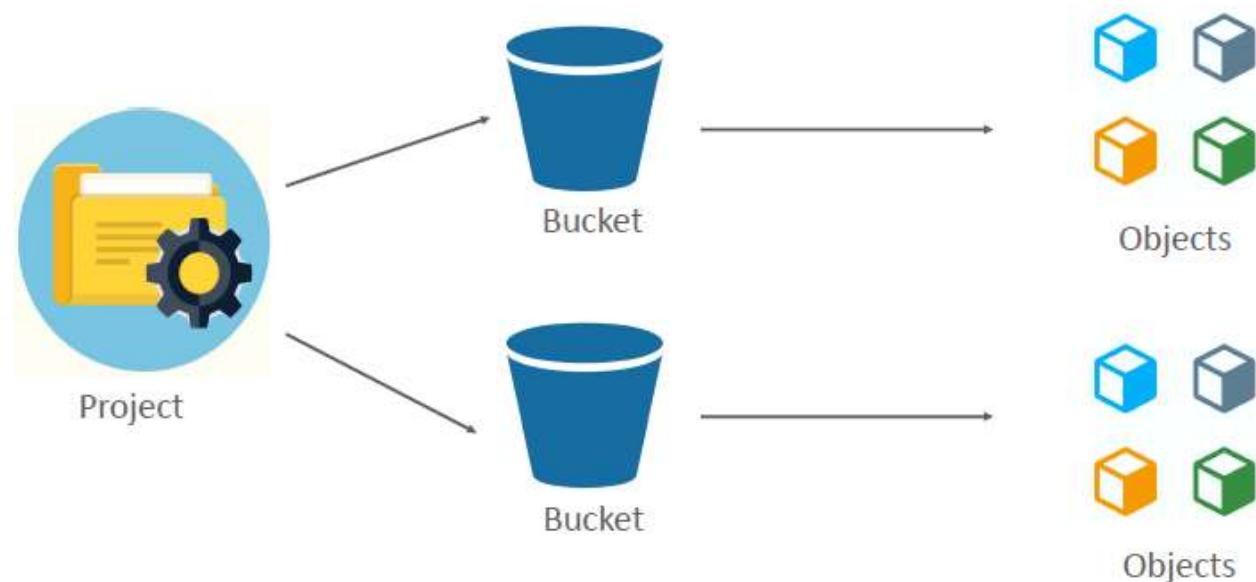
Cloud Storage Highlights

Cloud Storage is binary large object Storage: -

- High performance, internet-scale
- Simple administration
 - Does not require capacity management
- Data encryption at rest
- Data encryption in transit by default from Google to endpoint
- Online and offline import services are available
- Designed for 99.999999999% annual durability

What is Cloud Storage?

Google Cloud Storage is an object storage system. It is designed for persisting unstructured data, such as data files, images, videos, backup files, and any other data.



Google Cloud Certified Professional Cloud Architect Definitive Guide

Cloud Storage is a service for storing your objects in Google Cloud. Cloud Storage uses buckets to group objects. A bucket is a group of objects that share access controls at the bucket level. An Object is an immutable piece of data consisting of a file of any format. You store objects in containers called buckets.

Google Cloud Storage uses a global namespace for bucket names, so all bucket names must have unique names. All buckets are associated with a project (/docs/overview#projects), and you can group your project under an organization (/resource-manager/docs/cloud-platform-resource-hierarchy#organizations).

After you create a project, you can create Cloud Storage buckets (/storage/docs/creating-buckets), Upload Objects (/storage/docs/uploading-objects) to your buckets, and download objects (/storage/docs/downloading-objects) from your buckets.

You can also grant permissions to make your data accessible to members you specify, or for certain use cases such as hosting a website - accessible to everyone on the public internet (/storage/docs/access-control/making-data-public)

Bucket

- Buckets are logical containers that hold data
- All data is stored in a bucket
- Buckets have unique global name. Use globally unique identifiers (GUIDs) if creating a large number of buckets
- Bucket name should be DNS compliant so that it can be used in a DNS record as part of CNAME or A redirect
- Bucket name cannot be changed once it is created
- Labels up to 63 characters can be added for identification
- IAM provides security and permission to the bucket
- Bucket names can also be subdomain names, such as mybucket.example.com



Cloud Storage Bucket

Bucket attributes	Bucket contents
Globally unique name	Files (in a flat namespace)
Storage class	
Location (region or multi-region)	
IAM policies or Access Control Lists	Access Control Lists
Object versioning setting	
Object lifecycle management rules	

Objects

Objects are actual data saved in Cloud storage

- No limit on the number of objects in a bucket
- Two components; object data and object metadata
- Owned by original uploader and permission is based on the bucket
- Less bucket and more objects are recommended
- No Minimum size, Maximum size is 5 TB

Storage Tiers

Cloud Storage offers four tiers or types of storage.

The four types of Cloud Storage are as follows: -

- Regional
- Multiregional
- Nearline
- Coldline

Multi-regional and Regional are high-performance object storage, whereas Nearline and Coldline are backup and archival storage. All of the storage classes are accessed in analogous ways using the Cloud Storage API, and they all offer millisecond access times.

Regional

Regional storage stores multiple copies of an object in multiple zones in one region.

Multiregional

Multiregional storage mitigates the risk of a regional outage by storing replicas of objects in multiple regions. This can also improve access time and latency by distributing copies of objects to locations that are closer to the users of those objects.

Multiregional storage is also known as geo-redundant storage. Multiregional Cloud Storage buckets are created in one of the multiregions—asia, eu, or us—for data centers in Asia, the European Union, and the United States, respectively.

Nearline & Coldline

- Nearline and Coldline storage are used for storing data that is not frequently accessed.
- Data that is accessed less than once in 30 days is a good candidate for Nearline storage.
- Data that is accessed less than once a year is a good candidate for Coldline storage.
- All storage classes have the same latency to return the first byte of data.

Choosing best among Storage Classes?

	Multi-regional	Regional	Nearline	Coldline	Archive
Intended for data that is...	Most frequently accessed	Accessed frequently within a region	Accessed less than once a month	Accessed less than once a quarter	Accessed less than once a year
Availability SLA	99.95%	99.90%	99.00%	99.00%	No SLA
Access APIs	<i>Consistent APIs</i>				
Access time	<i>Millisecond access</i>				
<u>Storage price</u>	Price per GB stored per month				
<u>Retrieval price</u>	Total price per GB transferred				
Use cases	Content storage and delivery	In-region analytics, transcoding	Long-tail content, backups	Archiving, disaster recovery	Archiving, disaster recovery

Multiregional storage has a 99.95 percent availability SLA. Regional storage has a 99.90 percent availability SLA. Nearline and Coldline storage have 99.90 percent availability SLA in multiregional locations and 99.0 percent availability in regional locations.

What are the various ways to bring data into Cloud Storage?

There are several ways to bring data into Cloud Storage



Online transfer

Self-managed copies using command-line tools or drag-and-drop



Storage Transfer Service

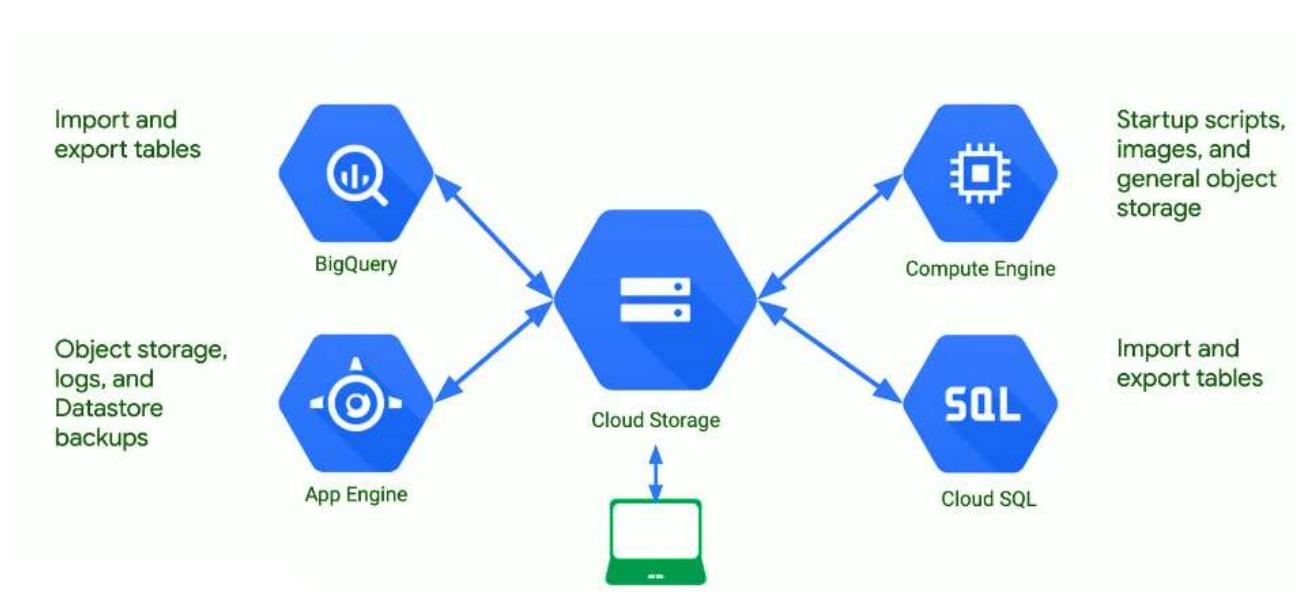
Scheduled, managed batch transfers



Transfer Appliance Beta

Rackable appliances to securely ship your data

Google Cloud Storage Integration with other Services



What are the various ways to interact with Cloud Storage?

The Various ways to interact with Cloud Storage are: -

- Cloud Shell
- GCP Console
- Command Line Tool - gsutil
- REST APIs & Client SDK

GCP Cloud Storage Use Cases

Cloud Storage is used for a few broad use cases.

- Storage of data shared among multiple instances that does not need to be on persistent attached storage. For example, log files may be stored in Cloud Storage and analysed by programs running in a Cloud Dataproc Spark cluster.
- Backup and archival storage, such as persistent disk snapshots, backups of on-premises systems, and data kept for audit and compliance requirements but not likely to be accessed.
- As a staging area for uploaded data. For example, a mobile app may allow users to upload images to a Cloud Storage bucket. When the file is created, a Cloud Function could trigger to initiate the next steps of processing.

Versioning in Cloud Storage

Cloud Storage retains a noncurrent object when there are changes to the live object

- Applies at bucket and disabled by default
- When overwriting a file, the existing file will not be deleted
- While accessing an object, it gives a live object
- Version names start with ‘#’
- Noncurrent object versions are independent of live ones
- Stopping Versioning leaves existing objects

Data Retention and Lifecycle Management

Data has something of a life as it moves through several stages, including creation, active use, infrequent access but kept online, archived, and deleted.

Cloud Storage provides **object lifecycle management policies** to make changes automatically to the way objects are stored in the object datastore. These policies contain rules for manipulating objects and are assigned to buckets. The rules apply to objects in those buckets. The rules implement lifecycle actions, including deleting an object and setting the storage class. Rules can be triggered based on the age of the object, when it was created, the number of newer versions, and the storage class of the object.

Another control for data management is **retention policies**. A retention policy uses the Bucket Lock feature of Cloud Storage buckets to enforce object retention. By setting a retention policy, you ensure that any object in the bucket or future objects in the bucket are not deleted until they reach the age specified in the retention policy. This feature is particularly useful for compliance with government or industry regulations. Once a retention policy is locked, it cannot be revoked.

What is the Data Encryption Options in GCP?

The data encryption options in GCP are: -

- Google Managed Key: Google encrypt data before it saves to disk
- Customer Supplied Encryption Key: Create and manage your own encryption key
- Customer Managed Encryption Key: Generate and manage keys
- Client-Side Encryption: Occurs before data is sent to Google cloud

Network and latency

Network latency is a consideration when designing storage systems, particularly when data is transmitted between regions with GCP or outside GCP to globally distributed devices.

Three ways of addressing network latency concerns are as follows:

- Replicating data in multiple regions and across continents
- Distributing data using Cloud CDN
- Using Google Cloud Premium Network tier

The reason to consider using these options is that the network latency without them would be too high to meet application or service requirements.

For some points of reference, note the following:

- Within Europe and Japan, expect 12ms latency.
- Within North America 30-40 ms latency is typical.
- Trans-Atlantic latency is about 70 ms.
- Trans-Pacific latency is about 100 ms.
- Latency between the Europe, Middle East, and Africa (EMEA) region and Asia Pacific is closer to 120 ms.

Data can be replicated in multiple regions under the control of a GCP service or under the control of a customer-managed service. For example, Cloud Storage multiregional storage replicates data to multiple regions. Cloud Spanner distributes data automatically among multiple regions. Cloud Firestore is designed to scale globally. Using GCP services that manage multi-region and global distribution of data is preferred to managing replication at the application levels.

Another way to reduce latency is to use GCPs Cloud CDN. This is particularly effective and efficient when distributing relatively static content globally. Cloud CDN maintains a set of globally distributed points of presence around the world. Points of presence are where the Google Cloud connects to the Internet. Static content that is frequently accessed in an area can be cached at these edge nodes.

Google Cloud Certified Professional Cloud Architect Definitive Guide

GCP offers two network service tiers. In the Standard Tier, network traffic between regions is routed over the public Internet to the destination device. With the Premium Tier, all data is routed over the Google network up to a point of presence near the destination device. The Premium Tier should be used when high-performance routing, high availability, and low latency at multi-region scales are required.

Unmanaged Databases

GCP offers a range of managed database options, there may be use cases in which you prefer to manage your own database. These are sometimes referred to as unmanaged databases, but self-managed is probably a better term.

When you manage your own databases, you will be responsible for an array of database and system administration tasks, including

- Updating and patching the operating system
- Updating and patching the database system
- Backing up and, if needed, recovering data
- Configuring network access
- Managing disk space
- Monitoring database performance and resource utilization
- Configuring for high availability and managing failovers
- Configuring and managing read replicas

The two Stackdriver components that are used with unmanaged databases are Stackdriver Monitoring and Stackdriver Logging. Instances have built-in monitoring and logging. Monitoring includes CPU, memory, and I/O metrics. Audit logs, which have information about who created an instance, is also available by default. If you would like insights into application performance, in this case into database performance, you should install Stackdriver Monitoring and Stackdriver Logging agents.

Once the Stackdriver Logging agent is installed, it can collect application logs, including database logs. Stackdriver Logging is configured with Fluentd, an open-source data collector for logs. Once the Stackdriver Monitoring agent is installed, it can collect application performance metrics. Monitoring a specific database may require a plug-in designed for the particular database, such as MySQL or PostgreSQL.

Cloud Filestore (Network Attached Storage)

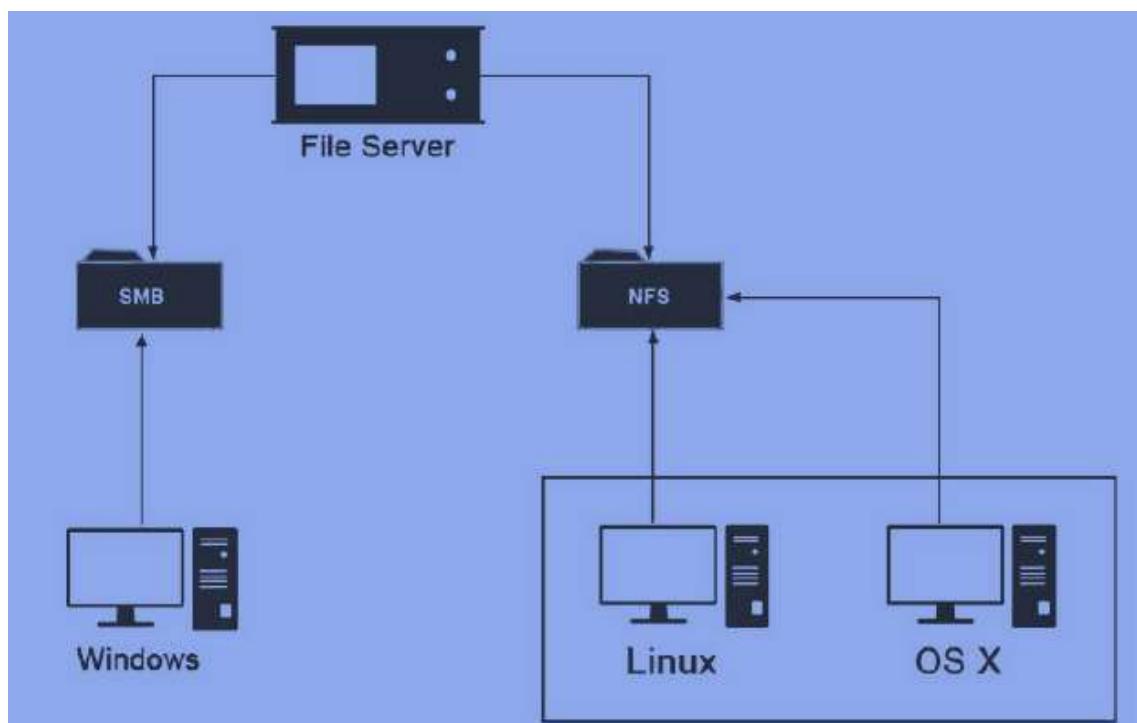
Cloud Filestore Highlights

Google Cloud Filestore provides with the following features: -

- File storage service
- Share same file system in multiple nodes
- Used as an NFS file system
- Minimum 1TB size
- Compatible with Compute Engine and GKE
- Backup
- Ability to fine tune performance and capacity
- Standard SSD support
- Access on IP address or VPC level

What is Cloud Filestore?

Cloud Filestore is a network-attached storage service that provides a filesystem that is accessible from Compute Engine and Kubernetes Engine. Cloud Filestore is designed to provide low latency and IOPS, so it can be used for databases and other performance sensitive services.



Cloud Filestore has two tiers: standard and premium.

The performance characteristics of each tier is given below: -

Feature	Standard	Premium
Maximum read throughput	100 MB/s (1 TB), 180 MB/s (10+ TB)	1.2 GB/sec
Maximum write throughput	100 MB/s (1 TB), 120 MB/s (10+ TB)	350 MB/s
Maximum IOPS	5,000	60,000
Typical availability	99.9 percent	99.9 percent

Some typical use cases for Cloud Filestore are home directories and shared directories, web server content, and migrated applications that require a filesystem.

Cloud Filestore filesystems can be mounted using operating system commands. Once mounted, file permissions can be changed as needed using standard Linux access control commands, such as chmod.

Cloud SQL

Cloud SQL Highlights

Cloud SQL is a Fully managed RDBMS: -

- Automated storage capacity management and data provisioning
- Offers MySQL and PostgreSQL databases as a service
- Integration with existing apps easily
- Automatic replication
- Managed backups
- Vertical scaling (read and write)
- Horizontal scaling (read)
- Google security
- Less maintenance cost
- High performance

What is Cloud SQL?

Cloud SQL is a fully managed service that provides MySQL and PostgreSQL databases. Cloud SQL allows users to deploy MySQL and PostgreSQL on managed virtual servers.

A managed database is one that does not require as much administration and operational support as an unmanaged database because Google will take care of core operational tasks, such as creating databases, performing backups, and updating the operating system of database instances. Google also manages scaling disks, configuring for failover, monitoring, and authorizing network connections.

Cloud SQL supports regional-level databases of up to 30 TB. If you need to store more data or need multi-regional support, consider using Cloud Spanner.

GCP manages patching database software, backups, and failovers. Key features of Cloud SQL include the following: -

- All data is encrypted at rest and in transit
- Data is replicated across multiple zones for high availability

- GCP manages failover to replicas
- Support for standard database connectors and tools is provided
- Stackdriver is integrated monitoring and logging

Relational databases, like MySQL and PostgreSQL, are used with structured data. Both require well-defined schemas, which can be specified using the data definition commands of SQL. Cloud SQL databases also support strongly consistent transactions, so there is no need to work around issues with eventual consistency. Cloud SQL is appropriate for transaction processing applications, such as e-commerce sales and inventory.

A limiting factor of Cloud SQL is that databases can scale only vertically, that is, by moving the database to a larger machine. For use cases that require horizontal scalability or support a globally accessed database, Cloud Spanner, is an appropriate choice.

What is Cloud SQL Instance?

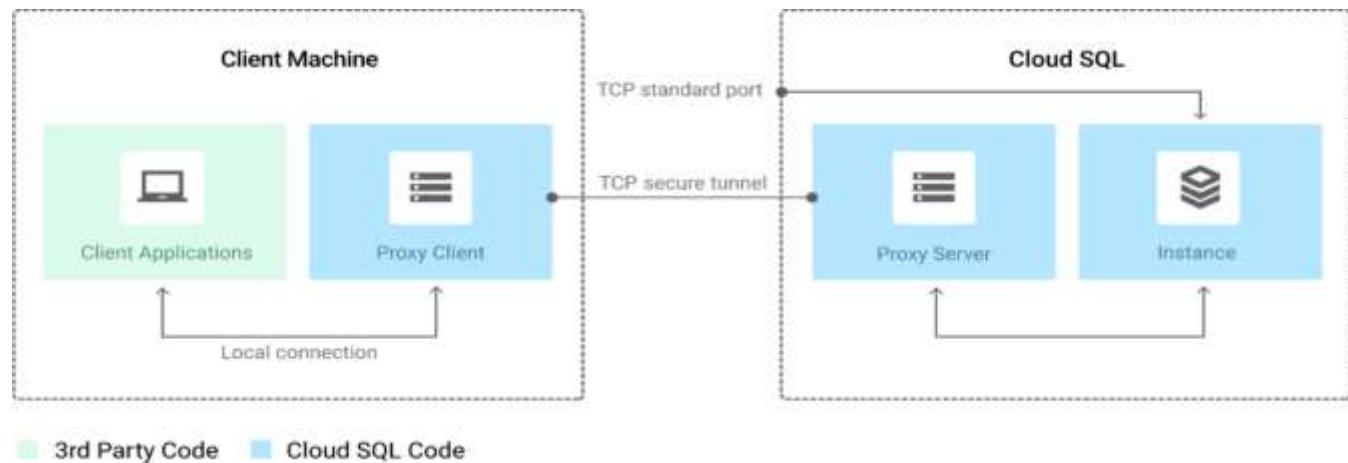
- Instances need to be created explicitly
 - Not Serverless
 - Specify region while creating instance
- First Vs Second generation instance
 - Second generation instances allow proxy support - No need to whitelist IP address or configure SSL
 - Higher availability configuration
 - Maintenance won't take down the server
- A Second-Generation instance is in a high availability configuration when it has a failover replica
- The failover replica must be in a different zone than the original instance, also call the master
- All changes made to the data on the master, including to user tables, are replicated to the failover replica using semi synchronous replication

What is Cloud Proxy?

Provides secure access to your Cloud SQL Second Generation instances without having to whitelist IP addresses or configure SSL

Secure connections: The proxy automatically encrypts traffic to and from the database; SSL certificates are used to verify client and server identities

Easier connection management: The proxy handles authentication with Google Cloud SQL, removing the need to provide static IP addresses



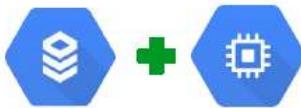
- The Cloud SQL Proxy works by having a local client, called the proxy, running in the local environment
- Your application communicates with the proxy with the standard database protocol used by your database
- The proxy uses a secure tunnel to communicate with its companion process running on the server.
- You can install the proxy anywhere in your local environment. The location of the proxy binaries does not impact where it listens for data from your application.
- When you start the proxy, need to tell it
 - What Cloud SQL instances it should establish connections to
 - Where it will listen for data coming from your application to be sent to Cloud SQL
 - Where it will find the credential's, it will use to authenticate your application to Cloud SQL

Cloud SQL can be used with other GCP Services



Cloud SQL can be used with App Engine using standard drivers.

You can configure a Cloud SQL instance to follow an App Engine application.



Compute Engine instances can be authorized to access Cloud SQL instances using an external IP address.

Cloud SQL instances can be configured with a preferred zone.



Cloud SQL can be used with external applications and clients.

Standard tools can be used to administer databases.

External read replicas can be configured.

Improving Read Performance with Read Replicas

You can improve the performance of read-heavy databases by using read replicas. A read replica is a copy of the primary instance's data that is maintained in the same region as the primary instance. Applications can read from the read replica and allow the primary instance to handle write operations.

Here are some things to keep in mind about read replicas: -

- Read replicas are not designed for high availability; Cloud SQL does not fail over to a read replica.
- Binary logging must be enabled to support read replicas.
- Maintenance on read replicas is not limited to maintenance windows—it can occur at any time and disrupt operations.
- A primary instance can have multiple read replicas, but Cloud SQL does not load-balance between them.
- You cannot perform backups on a read replica.
- Read replicas can be promoted to a standalone Cloud SQL. The operation is irreversible.
- The primary instance cannot be restored from backup if a read replica exists; it will have to be deleted or promoted.
- Read replicas must be in the same region as the primary instance.

Importing and Exporting Data

Data can be imported and exported from Cloud SQL databases. Each RDBMS has an import and an export program.

- MySQL provides mysqldump.
- PostgreSQL provides pg_dump.
- SQL Server provides bcp, which stands for bulk copy program.

In general, data is imported from and exported to Cloud Storage buckets. Files can be in either CSV or SQL dump format.

It should be noted that when you are using the command line to export a database, you must use particular flags to create a file that can later be imported. For Cloud SQL to import correctly, you will need to exclude views, triggers, stored procedures, and functions.

Those database objects will have to be re-created using scripts after a database has been imported.

Both SQL Dump and CSV formats can be compressed to save on storage costs. It is possible to import from a compressed file and export to a compressed file.

Cloud SQL is well suited for regional applications that do not need to store more than 30 TB of data in a single instance. Cloud Spanner is used for more demanding database systems.

Cloud Spanner

Cloud Spanner Highlights

Cloud Spanner is a horizontally scalable RDBMS support's:-

- A SQL RDBMS, with joins and secondary indexes
- Automatic replication & Built-in high availability
- Strong global consistency, so there is no risk of data anomalies caused by eventual consistency.
- Database sizes exceeding ~2 TB
- Managed instances with high availability
- SQL (ANSI 2011 with extensions)
- Many IOPS (Tens of thousands of reads/writes per second or more)
- Datatypes - Arrays and Structs

What is Cloud Spanner?

Cloud Spanner is Google Proprietary, more advanced than Cloud SQL. Cloud Spanner is a horizontally scalable RDBMS across the regions. i.e., bigger data, more instances, replication etc., It supports common relational features, such as schemas for structured data and SQL for querying.

Cloud Spanner provides 99.999 percent availability, which guarantees less than 5 minutes of downtime per year. Like Cloud SQL, all patching, backing up, and failover management is performed by GCP.

Cloud Spanner is a relational database, so it supports fixed schemas and is ANSI SQL 2011 compliant. Cloud Spanner provides strong consistency, so all parallel processes see the same state of the database.

This consistency is different from NoSQL databases, which are generally eventually consistent, allowing parallel processes to see different states of the database.

Cloud Spanner is highly available and does not require failover instances in the way that Cloud SQL does. It also manages automatic replication.

Google Cloud Certified Professional Cloud Architect Definitive Guide

Data is encrypted at rest and in transit. Cloud Spanner is integrated with Cloud Identity to support the use of user accounts across applications and with Cloud Identity and Access Management to control authorizations to perform operations on Cloud Spanner resources.

Cloud Spanner is used for applications that require strong consistency on a global scale. **Examples**

- Financial trading systems require a globally consistent view of markets to ensure that traders have a consistent view of the market when making trades.
- Logistics applications managing a global fleet of vehicles need accurate data on the state of vehicles.
- Global inventory tracking requires global-scale transaction to preserve the integrity of inventory data.

How Transactions works in Cloud Spanner?

- Cloud Spanner Supports serializability
- Cloud Spanner transaction support is super-strong, even stronger than traditional ACID
 - Transactions commit in an order that is reflected in their commit timestamps
 - These commit timestamps are "real time" so you can compare them to your watch
- Two transaction modes
 - Locking read-write (slow)
 - Read-only (fast)
- If making a one-off read, use something known as a “Single Read Call”
 - Fastest, no transaction checks needed!

Decision Making in Cloud Spanner

Use when

- Need high availability
- Strong consistency
- Transactional reads and writes (especially writes!)

Don't use if

- Data is not relational, or not even structured
- Want an open-source RDBMS
- Strong consistency and availability are overkill

Replication in Cloud Spanner

Cloud Spanner maintains multiple replicas of rows of data in multiple locations. Since Cloud Spanner implements globally synchronous replication, you can read from any replica to get the latest data in a row. Rows are organized into splits, which are contiguous blocks of rows that are replicated together. One of the replicas is designated the leader and is responsible for write operations. The use of replicas improves data availability as well as reducing latency because of geographic proximity of data to applications that need the data. Again, the benefit of adding nodes should be balanced with the price of those additional nodes.

The distributed nature of Cloud Spanner creates challenges for writing data. To keep all replicas synchronized, Cloud Spanner uses a voting mechanism to determine writes. Cloud Spanner uses a voting mechanism to determine the latest write-in case of a conflict value.

There are three types of replicas in Cloud Spanner:

- Read-write replicas
- Read-only replicas
- Witness replicas

Regional instances use only read-only replicas; multi-regional instances use all three types. Read-write replicas maintain full copies of data and serve read operations, and they can vote on write operations.

Read-only replicas maintain full copies of data and serve read operations, but they do not vote on write operations.

Witness replicas do not keep full copies of data but do participate in write votes. Witness replicas are helpful in achieving a quorum when voting.

Regional instances maintain three read-write replicas. In multi-regional instances, two regions are considered read-write regions and contain two replicas. One of those replicas is considered the leader replica. A witness replica is placed in a third region.

Database Design Considerations

Like NoSQL databases, Cloud Spanner can have hotspots where many read or write operations are happening on the same node instead of in parallel across multiple nodes. This can occur using sequential primary keys, such as auto-incrementing counters or timestamps. If you want to store sequential values and use them for primary keys, consider using the hash of the sequential value instead. That will evenly distribute writes across multiple nodes. This is because a hash value will produce apparently random values for each input, and even a small difference in an input can lead to significantly different hash values.

Relational databases are often normalized, and this means that joins are performed when retrieving data. For example, orders and order line items are typically stored in different tables. If the data in different tables is stored in different locations on the persistent data store, the database will spend time reading data blocks from two different areas of storage.

Cloud Spanner allows for interleaving data from different tables. This means that an order row will be stored together with order line items for that order. You can take advantage of interleaving by specifying a parent-child relationship between tables when creating the database schema.

Please note that row size should be limited to 4 GB, and that includes any interleaved rows.

Importing and Exporting Data

Data can be imported to or exported from Cloud Storage into Cloud Spanner. Exported files use the Apache Avro or CSV file formats. The export process is implemented by a Cloud Dataflow connector.

The performance of import and export operations is affected by several factors, including the following:

- Size of the database
- Number of secondary indexes
- Location of data
- Load on Cloud Spanner and number of nodes

Since Cloud Spanner can use the Avro format, it is possible to import data into Cloud Spanner from a file that was created by another application.

If you prefer to export data to CSV format, you can run a Dataflow job using the Cloud Spanner to Cloud Storage Text template.

Cloud Bigtable

Cloud Bigtable Highlights

Cloud Bigtable is managed NoSQL

- Fully managed NoSQL, wide-column database service for terabyte applications
- Integrated
 - Accessed using HBase API
 - Native compatibility with big data, Hadoop ecosystems

What is Cloud Bigtable? What are its advantages

Cloud Bigtable is designed to support petabyte-scale databases for analytic operations, such as storing data for machine learning model building, as well as operational use cases, such as streaming Internet of Things (IoT) data. It is also used for time series, marketing data, financial data, and graph data. Some of the most important features of Cloud Bigtable are:

- Cloud Bigtable is a wide-column NoSQL database used for high-volume databases that require low millisecond (ms) latency.

Bigtable features are as follows: -

- Sub 10 ms latency
- Stores petabyte-scale data
- Uses regional replication
- Queried using a Cloud Bigtable-specific command, cbt
- Supports use of Hadoop HBase interface
- Runs on a cluster of servers

Bigtable stores data in tables organized by key-value maps. Each row contains data about a single entity and is indexed by a row key. Columns are grouped into column families, which are sets of related columns. A table may contain multiple column families.

Tables in Bigtable are partitioned into blocks of contiguous rows known as tablets. Tablets are stored in the Colossus scalable filesystem. Data is not stored on nodes in the cluster. Instead, nodes store pointers to tablets stored in Colossus. Distributing read and write load across nodes yields

better performance than having hotspots where a small number of nodes are responding to most read and write requests.

Bigtable supports creating more than one cluster in a Bigtable instance. Data is automatically replicated between clusters. This is useful when the instance is performing a large number of read and write operations at the same time. With multiple clusters, one can be dedicated to responding to read requests while the other receives write requests. Bigtable guarantees eventual consistency between the replicas.

The advantages of using Bigtable are: -

- Replicated storage
- Data encryption in-flight and at rest
- Role-based ACLs
- Drives major applications such as Google Analytics and Gmail

Why Cloud Bigtable?

Customers frequently choose Bigtable if the data is: -

Big

Large quantities (>1 TB) of semi-structured or structured data

Fast

Data is high throughput or rapidly changing

NoSQL

Transactions, strong relational semantics not required

Time series

Data is time-series or has natural semantic ordering

Big data

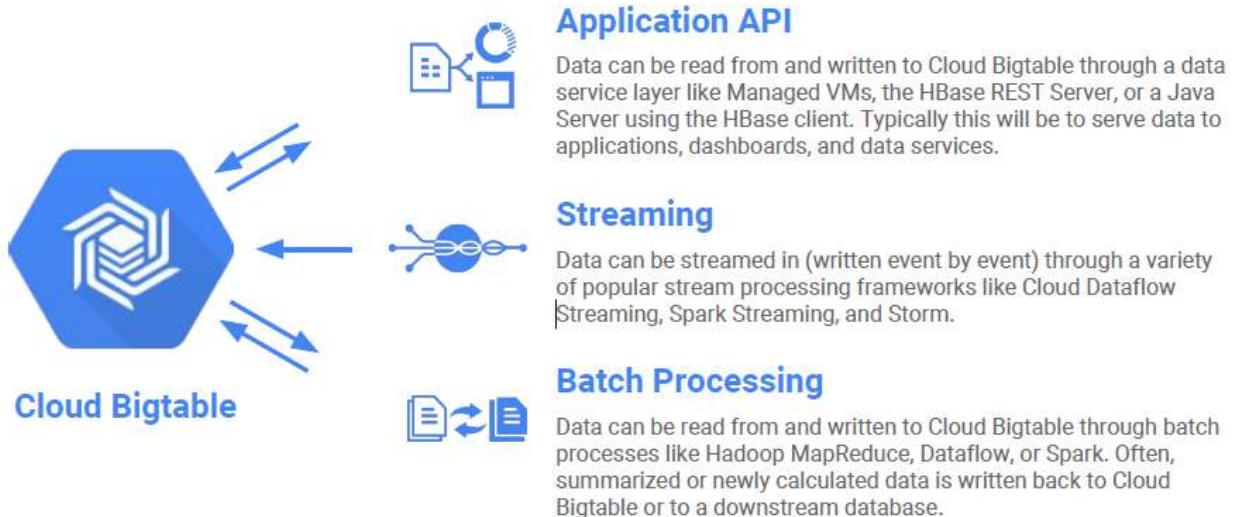
You run asynchronous batch or real-time processing on the data

Machine learning

You run machine learning algorithms on the data

Bigtable is designed to handle massive workloads at consistent low latency and high throughput, so it's a great choice for both operational and analytical applications, including IoT, user analytics, and financial data analysis

Big Table Access Patterns



Database Design Considerations

Designing tables for Bigtable is fundamentally different from designing them for relational databases. Bigtable tables are denormalized, and they can have thousands of columns.

There is no support for joins in Bigtable or for secondary indexes. Data is stored in Bigtable lexicographically by row-key, which is the one indexed column in a Bigtable table. Keeping related data in adjacent rows can help make reads more efficient.

All operations are atomic at the row level, not a transaction level. For example, if an application performs operations on two rows, it is possible for one of those operations to succeed and another one to fail. This can leave the database in an inconsistent state. To avoid this, store all the data that needs to be updated together in the same row.

Since Bigtable does not have secondary indexes, queries are executed using either row-key-based lookups or full table scans. The latter are highly inefficient and should be avoided. Instead, row-key lookups or a range of row-key lookups should be used. This requires carefully planning the row-key around the way that you will query the data. The goal when designing a row-key is to take advantage of the fact that Bigtable stores data in a sorted order.

Characteristics of a good row-key include the following:

- Using a prefix for multitenancy. This isolates data from different customers. This makes scans and reads more efficient since data blocks will have data from one customer only and customers query only their own data.
- Columns that are not frequently updated, such as a user ID or a timestamp.
- Nonsequential value in the first part of the row-key, which helps avoid hotspots.

Another way to improve performance is to use column families, which are sets of columns with related data. They are stored together and retrieved together, making reads more efficient.

Importing and Exporting

Like Cloud Spanner, Bigtable import and export operations are performed using Cloud Dataflow. Data can be exported to Cloud Storage and stored in either Avro format or SequenceFile format. Data can be imported from Avro and SequenceFile. Data can be imported from CSV files using a Dataflow process as well.

Cloud Datastore

Cloud Datastore Highlights

Cloud Datastore is a horizontally scalable NoSQL DB: -

- NoSQL designed for application backends
- Fully managed
- Uses a distributed architecture to automatically manage scaling
- Built-in redundancy
- Supports ACID transactions

The benefits of Cloud Datastore

- Schema-less access
- No need to think about underlying data structure
- Local development tools
- Includes a free daily quota
- Access from anywhere through a RESTful interface

What is Cloud Datastore?

Cloud Datastore is a managed document database, which is a kind of NoSQL database that uses a flexible JSON-like data structure called a document.

Some of the key features of Cloud Datastore are as follows: -

- ACID transactions are supported
- It is highly scalable and highly available
- It supports strong consistency for lookup by key and ancestor queries, while other queries have eventual consistency
- Data is encrypted at rest
- It provides a SQL-like query language called GQL
- Indexes are supported
- Joins are not supported
- It scales to zero. Most other database services usually require capacity to be allocated regardless of whether it is consumed.

Google Cloud Certified Professional Cloud Architect Definitive Guide

The terminology used to describe the structure of a document is different than that for relational databases. A table in a relational database corresponds to a kind in Cloud Datastore, while a row is referred to as an entity. The equivalent of a relational column is a property, and a primary key in relational databases is simply called the key in Cloud Datastore.

Cloud Datastore is fully managed. GCP manages all data management operations including distributing data to maintain performance. Also, Cloud Datastore is designed so that the response time to return query results is a function of the size of the data returned and not the size of the dataset that is queried.

The flexible data structure makes Cloud Datastore a good choice for applications like product catalog's or user profiles.

What are the features of Cloud Datastore?

The features of Cloud Datastore are: -

Atomic transactions

- Datastore can execute a set of operations where either all succeed, or none occur

High availability of reads and writes

- Datastore runs in Google data centers, which use redundancy to minimize impact from points of failure

Massive scalability with high performance

- Datastore uses a distributed architecture to automatically manage scaling. Datastore uses a mix of indexes and query constraints so your queries scale with the size of your result set, not the size of your data set

Flexible storage and querying of data

- Datastore maps naturally to object-oriented and scripting languages and is exposed to applications through multiple clients. It also provides a SQL-like query language.

Balance of strong and eventual consistency

- Datastore ensures that entity lookups and ancestor queries always receive strongly consistent data. All other queries are eventually consistent. The consistency models allow your application to deliver a great user experience while handling large amounts of data and users

Encryption at rest

- Datastore automatically encrypts all data before it is written to disk and automatically decrypts the data when read by an authorized user. For more information, see Server-Side Encryption

Fully managed with no planned downtime

- Google handles the administration of the Datastore service so you can focus on your application. Your application can still use Datastore when the service receives a planned upgrade

Cloud Firestore

Cloud Firestore Highlights

Cloud Firestore: -

- Strongly consistent
- Supports document and collection data models
- Supports real-time updates
- Provides mobile and web client libraries

What is Cloud Firestore?

Cloud Firestore is the next generation of the GCP-managed document database.

Cloud Firestore is the managed document database that is replacing Cloud Datastore. Document databases are used when the structure of data can vary from one record to another.

Cloud Firestore has features not previously available in Cloud Datastore, including

- Strongly consistent storage layer
- Real-time updates
- Mobile and web client libraries
- A new collection and document data model

Cloud Firestore operates in one of two modes: Native Mode and Cloud Datastore Mode.

The Datastore mode is backward compatible with Cloud Datastore. In Datastore mode, all transactions are strongly consistent, unlike Cloud Datastore transactions, which are eventually consistent. Other Cloud Datastore limitations on querying and writing are removed in Firebase Datastore mode.

In addition, Datastore had a limit of 25 entity groups, or ancestor/descendent relations, and a maximum of one write per second to an entity group. Those limits are removed. The new data model, real-time updates, and mobile and web client library features are available only in Native Mode.

It's best practice for customers use Cloud Firestore in Datastore mode for new server-based projects. This supports up to millions of writes per second.

Cloud Firestore in its native mode is for new web and mobile applications. This provides client libraries and support for millions of concurrent connections.

Cloud Firestore Data Model

Cloud Firestore in Datastore Mode uses a data model that consists of entities, entity groups, properties, and keys.

Entities are analogous to tables in a relational database. They describe a particular type of thing, or entity kind. Entities have identifiers, which can be assigned automatically or specified by an application. If identifiers are randomly assigned in Cloud Datastore Mode, the identifiers are randomly generated from a uniformly distributed set of identifiers. The random distribution is important to avoid hot spotting when a large number of entities are created in a short period of time.

Entities have properties, which are name-value pairs. For example, 'color':'red' is an instance of a color property with the value red. The property of values can be different types in different entities. Properties are not strongly typed. A property can also be indexed. Note that this is a difference with the other managed NoSQL databases—Bigtable does not have secondary indexes.

Properties can have one value or multiple values. Multiple values are stored as array properties. For example, a news article might have a property 'topics' and values {'technology', 'business', 'finance'}.

The value of a property can be another entity. This allows for hierarchical structures.

Here is an example of a hierarchical structure of an order:

```
{'order_id': 1,
  'Customer': "Margaret Smith",
  'Line_item': [
    {'line_id': 10,
      'description': "Blue coat",
      'quantity': 3},
    {'line_id': 20,
      'description': "Green shirt",
      'quantity': 1},
    {'line_id': 30,
      'description': "Black umbrella",
      'quantity': 1}
  ]
}
```

Indexing and Querying

Cloud Firestore uses two kinds of indexes: built-in indexes and composite indexes. Built-in indexes are created by default for each property in an entity. Composite indexes index multiple values of an entity.

Indexes are used when querying and must exist for any property referenced in filters. For example, querying for "color" = 'red' requires an index on the color property. If that index does not exist, Cloud Firestore will not return any entities, even if there are entities with a color property and value of red. Built-in indexes can satisfy simple equality and inequality queries, but more complex queries require composite indexes.

Composite indexes are used when there are multiple filter conditions in a query. For example, the filter color = 'red' and size = 'large' requires a composite index that includes both the color and size properties.

Composite indexes are defined in a configuration file called index.yaml. You can also exclude properties from having a built-in index created using the index.yaml configuration file. You may want to exclude properties that are not used in filter conditions. This can save storage space and avoid unnecessary updating of an index that is never used.

Cloud Firestore uses a SQL-like language called GQL (GraphQL). Queries consist of an entity kind, filters, and optionally a sort order specification. Here is an example query:

```
SELECT * FROM orders  
WHERE item_count > 1 AND status = 'shipping'  
ORDER BY item_count DESC
```

This query will return entities from the orders collection that have a value greater than 1 for item_count and a value of 'shipping' for status. The results will be sorted in descending order by item_count. For this query to function as expected, there needs to be an index on item_count and 'shipping' and sorted in descending order.

Importing and Exporting

Entities can be imported and exported from Cloud Firestore. Exports contain entities but not indexes. Those are rebuilt during import. Exporting requires a Cloud Storage bucket to store the exported data. You can specify a filter when exporting so that only a subset of entity kinds is exported.

Exported entities can be imported into BigQuery using the bq command, but only if an entity filter was used when creating the export. There are a number of restrictions on importing Cloud Firestore exports to BigQuery:

- The Cloud Storage URI of the export file must not have a wildcard
- Data cannot be appended to an existing table with a defined schema
- Entities in the export have a consistent schema
- Any property value greater than 64 KB is truncated to 64 KB on export

Cloud Firestore in Datastore Mode is a managed document database that is well suited for applications that require semi-structured data but that do not require low-latency writes (< 10 ms).

When low-latency writes are needed, Bigtable is a better option.

Cloud Memorystore

Cloud Memorystore Highlights

- Cloud Memorystore is a managed Redis service. Redis is an open source, in-memory data store, which is designed for sub millisecond data access.
- Cloud Memorystore supports up to 300 GB instances and 12 Gbps network throughput. Caches replicated across two zones provide 99.9 percent availability.
- As with other managed storage services, GCP manages Cloud Memorystore patching, replication, and failover. Cloud Memorystore is used for low-latency data retrieval, specifically lower latency than is available from databases that store data persistently to disk or SSD.
- Cloud Memorystore is available in two service tiers: Basic and Standard. Basic Tier provides a simple Redis cache on a single server. Standard Tier is a highly available instance with cross-zone replication and support for automatic failover.
- The caching service also has capacity tiers, called M1 through M5. M1 has 1 GB to 4 GB capacity and 3 Gbps maximum network performance. These capacities increase up to M5, which has more than 100 GB cache size and up to 12 Gbps network throughput.
- Memorystore caches are used for storing data in nonpersistent memory, particularly when low-latency access is important. Stream processing and database caching are both common use cases for Memorystore.

What is Cloud Memorystore?

Cloud Memorystore is a managed Redis service, which is commonly used for caching.

Redis instances can be created using the Cloud Console or gcloud commands.

There are only a small number of basic configuration parameters with Cloud Memorystore:

- Instance ID.
- Size specification.
- Region and zone.
- Redis version; currently the options are 3.2 and 4.0; 4.0 is recommended.
- Instance tier, which can be basic and is not highly available, or standard, which includes a failover replica in a different zone.
- Memory capacity, which ranges from 1 to 300 GB.

You also have the option of specifying Redis configuration parameters, such as maximum memory policy, and an eviction policy, such as least frequently used.

Cloud Memorystore provides support for importing and exporting from Redis; this feature is in beta as of this writing. Exporting will create a backup file of the Redis cache in a Cloud Storage bucket. During export, read and write operations can occur, but administration operations, like scaling, are not allowed. Import reads export files and overwrites the contents of a Redis cache. The instance is not available for read or write operations during the import.

Redis instances in Cloud Memorystore can be scaled to use more or less memory. When scaling a Basic Tier instance, reads and writes are blocked. When the resizing is complete, all data is flushed from the cache. Standard Tier instances can scale while continuing to support read and write operations. During a scaling operation, the replica is resized first and then synchronized with the primary. The primary then fails over to the replica. Write operations are supported when scaling Standard Tier instances, but too much write load can significantly slow the resizing operation.

When the memory used by Redis exceeds 80 percent of system memory, the instance is considered under memory pressure. To avoid memory pressure, you can scale up the instance, lower the maximum memory limit, modify the eviction policy, set time-to-live (TTL) parameters on volatile keys, or manually delete data from the instance. The TTL parameter specifies how long a key should

be kept in the cache before it becomes eligible for eviction. Frequently updated values should have short TTLs whereas keys with values that don't change very often can have longer TTLs. Some eviction policies target only keys with TTLs whereas other policies target all keys. If you find that you are frequently under memory pressure, your current eviction policy applies only to keys with TTLs, and there are keys without TTLs, then switching to an eviction policy that targets all keys may relieve some of that memory pressure.

Redis provides a number of eviction policies that determine which keys are removed from the cache when the maximum memory limit is reached. By default, Redis evicts the least recently used keys with TTLs set. Other options include evicting based on least frequently used keys or randomly selecting keys.

Although Cloud Memorystore is a managed service, you should still monitor the instance, particularly memory usage, duration periods of memory overload, cache-hit ratio, and the number of expirable keys.

Cloud Bigquery

BigQuery Highlights

BigQuery is a fully managed Petabyte Scale, Low-cost analytics Datawarehouse: -

- Provides near real-time interactive analysis of massive datasets (hundreds of TBs)
- SQL queries - with Google storage underneath
- Query using SQL syntax (SQL 2011)
- BigQuery is NoOps - No Infra to manage, so analyze data to find meaningful insights i.e., No Cluster maintenance is required
- Pay-as-you-go model
- Instead of a Dynamic Pipeline, you want to do ad-hoc SQL queries on a massive dataset BigQuery is the ultimate choice
- BigQuery is an important service for the Process and Analyze stage of the data lifecycle
- As with other databases, some common tasks are as follows:
 - Interacting with data sets
 - Importing and exporting data
 - Streaming inserts
 - Monitoring and logging
 - Managing costs
 - Optimizing tables and queries

What is BigQuery?

BigQuery is a managed data warehouse and analytics database solution. It is designed to support queries that scan and return large volumes of data, and it performs aggregations on that data. BigQuery uses SQL as a query language. Customers do not need to choose a machine instance type or storage system. BigQuery is a serverless application from the perspective of the user.

Data is stored in a columnar format, which means that values from a single column in a database are stored together rather than storing data from the same row together. This is used in BigQuery because analytics and business intelligence queries often filter and group by values in a small number of columns and do not need to reference all columns in a row.

Google Cloud Certified Professional Cloud Architect Definitive Guide

BigQuery uses the concept of a job for executing tasks such as loading and exporting data, running queries, and copying data. Batch and streaming jobs are supported for loading data.

BigQuery uses the concept of a dataset for organizing tables and views. A dataset is contained in a project. A dataset may have a regional or multiregional location. Regional datasets are stored in a single region such as us-west2 or europe-north1. Multiregional locations store data in multiple regions within either the United States or Europe.

BigQuery provides its own command-line program called bq rather than use the gcloud command line. Some of the bq commands are as follows: -

- cp for copying data
- cancel for stopping a job
- extract for exporting a table
- head for listing the first rows of a table
- insert for inserting data in newline JSON format
- load for inserting data from AVRO, CSV, ORC, Parquet, and JSON data files or from Cloud Datastore and Cloud Firestore exports
- ls for listing objects in a collection
- mk for making tables, views, and datasets
- query for creating a job to run a SQL query
- rm for deleting objects
- show for listing information about an object

BigQuery is integrated with Cloud IAM, which has several predefined roles for BigQuery: -

- **dataViewer.** This role allows a user to list projects and tables and get table data and metadata.
- **dataEditor.** This has the same permissions as dataViewer, plus permissions to create and modify tables and datasets.
- **dataOwner.** This role is similar to dataEditor, but it can also create, modify, and delete datasets.
- **metadataViewer.** This role gives permissions to list tables, projects, and datasets.
- **user.** The user role gives permissions to list projects and tables, view metadata, create datasets, and create jobs.
- **jobUser.** A jobUser can list projects and create jobs and queries.
- **admin.** An admin can perform all operations on BigQuery resources.

BigQuery is billed based on the amount of data stored and the amount of data scanned when responding to queries, or in the case of flat-rate query billing, the allocation is used based on the size of the query. For this reason, it is best to craft queries that return only the data that is needed, and filter criteria should be as specific as possible.

The BigQuery Data Transfer Service is a specialized service for loading data from other cloud services, such as Google Ads and Google Ad Managers. It also supports transferring data from Cloud Storage and AWS S3, but these are both in beta stage at the time of this writing.

GCP provides two managed relational databases and an analytics database with some relational features. **Cloud SQL** is used for transaction processing systems that do not need to scale beyond a single server. It supports MySQL and PostgreSQL. **Cloud Spanner** is a transaction processing relational database that scales horizontally, and it is used when a single server relational database is insufficient. **BigQuery** is designed for data warehousing and analytic querying of large datasets. BigQuery should not be used for transaction processing systems. If data is frequently updated after loading, then one of the other managed relational databases is a better option.

What are the features of BigQuery?

- Flexible Data Ingestion
- Global Availability
- Security and Permissions
- Cost Controls
- Highly Available
- Super-Fast Performance
- Fully Integrated
- Connect with Google Products

What are the various Ways to access Big Query?

The various Ways to access Big Query are: -

- Web UI
- REST API
- Clients
- Third Party Tools

Loading and Exporting Data

With an existing dataset, you can create tables and load data into those tables. As with querying, you can use either the UI or the command line. In both cases, you will need to specify the following:

The type of data source, which can be Google Cloud Storage, an uploaded file from local drive, G Drive, or a Cloud Bigtable table, and the data transfer service (DTS).

The data source, such as a URI to a Cloud Storage file, a Bigtable table, or a path to a local filename.

File format—one of: Avro, CSV, JSON (newline delimited), ORC, or Parquet. If you are loading from Cloud Storage, you can also load Cloud Datastore and Cloud Firestore exports

Destination table, including project, dataset, and table name.

Schema, which can be auto-detected, specified in text, or entered one column at a time, specifying the column name, type, and mode. Mode may be NULLABLE, REQUIRED, or REPEATED.

Partitioning, which can be no partitioning or partitioning by ingestion time. If a table is partitioned, you can use clustering order, which optimizes the way that column data is stored and can optimize the way that queries are executed. You can also specify that any query on a partitioned table needs to specify a partition filter in the WHERE clause. This can significantly reduce cost and improve performance by limiting the amount of data scanned when executing the query.

-

BigQuery expects data to be encoded using UTF-8. If a CSV file is not in UTF-8, BigQuery will try to convert it. The conversion is not always correct, and there can be differences in some bytes. If your CSV file does not load correctly, specify the correct encoding. If you are loading JSON files, they need to be in UTF-8 encoding.

Avro is the preferred format for loading data because blocks of data can be read in parallel, even if the file is compressed and there are no encoding issues, such as with CSV. Parquet also stores data using a column model. Uncompressed CSV and JSON files load faster than compressed files because they can be loaded in parallel, but this can lead to higher storage costs when using Cloud Storage.

The BigQuery Data Transfer Service automates loading data from Google's software as a service (SaaS) offerings, such as Google Ad Manager, Google Ads, Google Play, and YouTube Channel Reports. It can also be used to load data from Amazon S3.

Cloud Dataflow can load directly into BigQuery

Clustering, Partitioning, and Sharding Tables BigQuery provides for creating clustered tables. In clustered tables, data is automatically organized based on the contents of one or more columns. Related data is collocated in a clustered table. This kind of organization can improve the performance of some queries, specifically the queries that filter rows using the columns used to cluster data. Clustered tables can be partitioned.

It is a good practice to break large tables into smaller ones or partitions to improve query efficiency. When splitting data by date or timestamp, you can use partitions, and to split data into multiple tables by other attributes, you can try Sharding.

Partitions can be based on ingestion time or by date or timestamp in the input data. When data is partitioned by ingestion time, BigQuery creates a new partition for each day. Shards can be based on the value in a partition column, such as customer ID. The column does not have to be a date or timestamp column. It's a best practice to use partitioning by time instead of Sharding by a date. The former is more efficient with BigQuery due to the backend optimizations it creates with timestamps.

Sharding can make use of template tables, which are tables that have a schema defined in a template and that template is used to create one or more tables that have a target table name and a table suffix. The target table name is the same for all tables created with the template, but the suffix is different for each table created.

Streaming Inserts

The loading procedures just described are designed for batch loading. BigQuery also supports streaming inserts that load one row at a time. Data is generally available for analysis within a few seconds, but it may be up to 90 minutes before data is available for copy and export operations. This is not intended for transactional workloads, but rather analytical ones.

Google Cloud Certified Professional Cloud Architect Definitive Guide

Streaming inserts provide best effort de-duplication. When inserting a row, you can include an insertID that uniquely identifies a record. BigQuery uses that identifier to detect duplicates. If no insertID is provided with a row, then BigQuery does not try to de-duplicate data. If you do provide an insertID and employ de-duplication, you are limited to 100,000 rows per second and 100 MB per second. If de-duplication is not enabled, you can insert up to 1,000,000 rows per second and 1 GB per second.

The advantage of using template tables is that you do not have to create all tables in advance. For example, if you are streaming in data from a medical device and you want to have table for each device, you could use the device identifier as the suffix, and when the first data from that device arrives, a table will be created from the template.

Standard SQL makes it easy to query across template tables by allowing wildcards in a table name.

For example, if you have a set of medical device tables named

'medical_device_' + <device id>, such as 'medical_device_123', 'medical_device_124', 'medical_device_125', and so forth, you could query across those tables by using a FROM clause as follows: FROM 'med_project.med_dataset.medical_device*''

Wildcards cannot be used with views or external tables.

Monitoring and Logging in BigQuery

Stackdriver is used for monitoring and logging in BigQuery. Stackdriver Monitoring provides performance metrics, such query counts and time to run queries. Stackdriver Logging is used to track events, such as running jobs or creating tables.

Stackdriver Monitoring collects metrics on a range of operations, including

- Number of scanned bytes
- Query time
- Slots allocated
- Slots available
- Number of tables in a dataset
- Uploaded rows

Google Cloud Certified Professional Cloud Architect Definitive Guide

You can build dashboards in Stackdriver Monitoring to help track key performance indicators, such as top long-running queries and 95th percentile query time.

Stackdriver Logging tracks log entries that describe events. Events have resource types, which can be projects or datasets, and type-specific attributes, like a location for storage events. Events that are tracked include the following:

- Inserting, updating, patching, and deleting tables
- Inserting jobs
- Executing queries

Logs are useful for understanding who is performing actions in BigQuery, whereas monitoring is useful for understanding how your queries and jobs are performing.

BigQuery Cost Considerations

BigQuery costs are based on the amount of data stored, the amount of data streamed, and the workload required to execute queries. Since the prices of these various services can change, it is not important to know specific amounts, but it is helpful to understand the relative costs. That can help when choosing among different options.

BigQuery data is considered active if it was referenced in the last 90 days; otherwise, it is considered long-term data. Active Storage is currently billed at \$0.20/GB a month, and long-term data is billed at \$0.10/GB a month. The charge for long-term storage in BigQuery is currently equal to the cost of Nearline storage, so there is no cost advantage to storing long-term data in Cloud Storage unless you were to store it Coldline storage, which is currently billed at \$0.07/GB a month.

Streaming inserts are billed at \$0.01 per 200 MB, where each row is considered at least 1 KB.

On-demand queries are billed at \$5.00 per TB scanned. Monthly flat rate billing is \$10,000 per 500 slots per month. Annual flat rate billing is \$8,500 a month for 500 slots.

There is no charge for loading, copying, or exporting data, but there are charges for the storage used. There are separate charges for using BigQuery ML machine learning service (BQML), for using BigQuery's native machine learning capabilities, and for using the BigQuery Data Transfer service.

Tips for Optimizing BigQuery

One way to keep costs down is to optimize the way that you use BigQuery. Here are several ways to do this:

- Avoid using SELECT *.
- Use --dry-run to estimate the cost of a query.
- Set the maximum number of bytes billed.
- Partition by time when possible.
- Denormalize data rather than join multiple tables.

Performance of Big Query

BigQuery runs on Google's high-performance infrastructure: -

- Compute and storage are separated with a terabit network between
- You pay only for storage and process used
- Automatic discount for long-term data storage
- Easy to get data into BigQuery, we can load from Cloud Storage or Datastore or stream into BigQuery up to 100000 rows per second
- BigQuery is used in all types of organizations, from startups to Fortune 500 companies.
Smaller Organizations will come under free quotas. Bigger Organization like seamless scale and its available 99.9% SLA

Storage – Interview | Exam Tips

Section – 1 GCP Major types of Storage System

GCP provides four types of storage systems:

- Object storage
- Network attached storage
- Databases
- Caching

Cloud Storage is used for unstructured data that is accessed at the object level; there is no way to query or access subsets of data within an object. Object storage is useful for a wide array of use cases, from uploading data from client devices to storing long-term archives.

Network-attached storage is used to store data that is actively processed. Cloud Filestore provides a network filesystem, which is used to share file structured data across multiple servers.

Google Cloud offers several managed databases, including Relational and NoSQL databases.

The relational database services are Cloud SQL and Cloud Spanner, Cloud SQL is used for transaction processing systems that serve clients within a region and do not need to scale beyond a single server. Cloud Spanner provides a horizontally scalable, global, strongly consistent relational database. BigQuery is a database designed for data warehousing and analytic database applications.

The NoSQL managed databases in GCP are Bigtable, Datastore, and Firestore. Bigtable is a wide-column database designed for low-latency writes at petabyte scales. Datastore and Firestore are managed document databases that scale globally. Firestore is the next generation of document storage in GCP and has fewer restrictions than Cloud Datastore.

Cloud Memorystore is a managed Redis service. Redis is an open source, in-memory data store, which is designed for sub millisecond data access

When designing storage systems, consider data lifecycle management and network latency. GCP provides services to help implement data lifecycle management policies and offers access to the Google global network through the Premium Tier network service.

1. Object Storage

Cloud Storage

Cloud Storage is an object storage system. It is designed for persisting unstructured data, such as data files, images, videos, backup files, and any other data. It is unstructured in the sense that objects—that is, files stored in Cloud Storage—use buckets to group objects. A bucket is a group of objects that share access controls at the bucket level.

The four storage tiers are Regional, Multi-regional, Nearline, and Coldline. Multiregional storage replicates objects across multiple regions, while regional replicates data across zones within a region. Nearline is used for data that is accessed less than once in 30 days. Coldline storage is used for data that is accessed less than once a year.

Cloud Storage provides object lifecycle management policies to make changes to the way that objects are stored in the object datastore. Another control for data management is retention policies. A retention policy uses the Bucket Lock feature of Cloud Storage buckets to enforce object retention.

2. Network Attached Storage

Cloud Filestore

Cloud Filestore is a network-attached storage service that provides a filesystem that is accessible from Compute Engine and Kubernetes Engine. Cloud Filestore is designed to provide low latency and IOPs so it can be used for databases and other performance-sensitive services.

3. Database Storage (SQL | NoSQL)

SQL DB

Cloud SQL

Cloud SQL is a managed relational database that can run on a single server. Cloud SQL allows users to deploy MySQL and PostgreSQL on managed virtual servers. Database administration tasks, such as patching, backing up, and managing failover are managed by GCP.

Cloud SQL instances are created in a single zone by default, but they can be created for high availability and use instances in multiple zones. Use read replicas to improve read performance. Importing and exporting are implemented via the RDBMS-specific tool.

Cloud Spanner

Cloud Spanner is a managed database service that supports horizontal scalability across regions. Cloud Spanner is used for applications that require strong consistency on a global scale. Cloud Spanner provides 99.999 percent availability, which guarantees less than 5 minutes of downtime a year. Like Cloud SQL, all patching, backing up, and failover management is performed by GCP.

Cloud Spanner is configured as regional or multi-regional instances. Cloud Spanner is a horizontally scalable relational database that automatically replicates data. Three types of replicas are read-write replicas, read-only replicas, and witness replicas. Avoid hotspots by not using consecutive values for primary keys.

NoSQL DB

Cloud Bigtable

Cloud Bigtable is designed to support petabyte-scale databases for analytic operations. It is used for storing data for machine learning model building, as well as operational use cases, such as streaming Internet of Things (IoT) data. It is also used for time series, marketing data, financial data, and graph data.

Cloud Bigtable is a wide-column NoSQL database used for high-volume databases that require sub-10 ms latency. Cloud Bigtable is used for IoT, time-series, finance, and similar applications. For multi-regional high availability, you can create a replicated cluster in another region. All data is replicated between clusters.

Designing tables for Bigtable is fundamentally different from designing them for relational databases. Bigtable tables are denormalized, and they can have thousands of columns. There is no support for joins in Bigtable or for secondary indexes. Data is stored in Bigtable lexicographically by row-key, which is the one indexed column in a Bigtable table. Keeping related data in adjacent rows can help make reads more efficient.

Cloud Datastore

Cloud Datastore is a managed document database, which is a kind of NoSQL database that uses a flexible JSON-like data structure called a document. Cloud Datastore is fully managed. GCP manages all data management operations, including distributing data to maintain performance. Also, Cloud

Datastore is designed so that the response time to return query results is a function of the size of the data returned and not the size of the dataset that is queried. The flexible data structure makes Cloud Datastore a good choice for applications like product catalog's or user profiles. Cloud Firestore is the next generation of GCP-managed document database.

Cloud Firestore

Cloud Firestore is a document database that is replacing Cloud Datastore as the managed document database. The Cloud Firestore data model consists of entities, entity groups, properties, and keys. Entities have properties that can be atomic values, arrays, or entities. Keys can be used to lookup entities and their properties. Alternatively, entities can be retrieved using queries that specify properties and values, much like using a WHERE clause in SQL. However, to query using property values, properties need to be indexed.

4. Caching

Cloud Memorystore

Cloud Memorystore is a managed Redis service.

Redis is an open source, in-memory data store, which is designed for submillisecond data access. Cloud Memorystore supports up to 300 GB instances and 12 Gbps network throughput. Caches replicated across two zones provide 99.9 percent availability.

Redis instances can be created using the Cloud Console or gcloud commands. Redis instances in Cloud Memorystore can be scaled to use more or less memory.

When scaling a Basic Tier instance, reads and writes are blocked. When the resizing is complete, all data is flushed from the cache. Standard Tier instances can scale while continuing to support read and write operations.

When the memory used by Redis exceeds 80 percent of system memory, the instance is considered under memory pressure. To avoid memory pressure, you can scale up the instance, lower the maximum memory limit, modify the eviction policy, set time-to-live (TTL) parameters on volatile keys, or manually delete data from the instance.

BigQuery

BigQuery is a managed data warehouse and analytics database solution. BigQuery uses the concept of a dataset for organizing tables and views. A dataset is contained in a project. BigQuery provides its own command-line program called `bq` rather than use the `gcloud` command line. BigQuery is billed based on the amount of data stored and the amount of data scanned when responding to queries.

BigQuery is an analytics database that uses SQL as a query language. Datasets are the basic unit of organization for sharing data in BigQuery. A dataset can have multiple tables.

BigQuery supports two dialects of SQL: legacy and standard. Standard SQL advanced SQL features such as correlated subqueries, ARRAY and STRUCT data types, and complex join expressions. BigQuery uses the concepts of slots for allocating computing resources to execute queries.

BigQuery also supports streaming inserts, which load one row at a time. Data is generally available for analysis within a few seconds, but it may be up to 90 minutes before data is available for copy and export operations. Streaming inserts provide for best effort de-duplication. Stackdriver is used for monitoring and logging in BigQuery.

Stackdriver Monitoring provides performance metrics, such query counts and time, to run queries. Stackdriver Logging is used to track events, such as running jobs or creating tables. BigQuery costs are based on the amount of data stored, the amount of data streamed, and the workload required to execute queries.

Network Latency

Network latency is a consideration when designing storage systems, particularly when data is transmitted between regions with GCP or outside GCP to globally distributed devices. Three ways of addressing network latency concerns are replicating data in multiple regions and across continents, distributing data using Cloud CDN, and using Google Cloud Premium Network tier.

Storage Monitoring

When you manage your own databases, you will be responsible for an array of database and system administration tasks. The two Stackdriver components that are used with unmanaged databases are Stackdriver Monitoring and Stackdriver Logging. Instances have built-in monitoring and logging.

Monitoring includes CPU, memory, and I/O metrics. Audit logs, which have information about who created an instance, are also available by default. Once the Stackdriver Logging agent is installed, it can collect application logs, including database logs. Stackdriver Logging is configured with Fluentd, an open-source data collector for logs. Once the Stackdriver Monitoring agent is installed, it can collect application performance metrics.

Section – 2 Selecting Appropriate Storage Technologies

We can select how to store data in many different situations given below:

- For Temporary staging area, where it stays only seconds or less before it is read by an application and deleted.
- For Long-term archival storage for data that needs to be retained for years.
- Data engineers are increasingly called on to work with data that streams into storage constantly and in high volumes. Internet of Things (IoT) devices are an example of streaming data.
- For storing large volumes of data for batch processing, including using data to train machine learning models. Data engineers also consider the range of variety in the structure of data.
- For online transaction processing, is highly structured and varies little from one datum to the next. Other data, like product descriptions in a product Catalog, can have a varying set of attributes. Data engineers consider these and other factors when choosing a storage technology.

Data lifecycle

The four stages of the data lifecycle: ingest, storage, process and analyze, and explore and visualize.

- Ingestion is the process of bringing application data, streaming data, and batch data into the cloud.
- The storage stage focuses on persisting data to an appropriate storage system.
- Processing and analyzing is about transforming data into a form suitable for analysis.
- Exploring and visualizing focuses on testing hypotheses and drawing insights from data.

Streaming data

Streaming data is a set of data that is sent in small messages that are transmitted continuously from the data source. Streaming data may be telemetry data, which is data generated at regular intervals, and event data, which is data generated in response to a particular event. Stream ingestion services need to deal with potentially late and missing data. Streaming data is often ingested using Cloud Pub/Sub.

Batch data

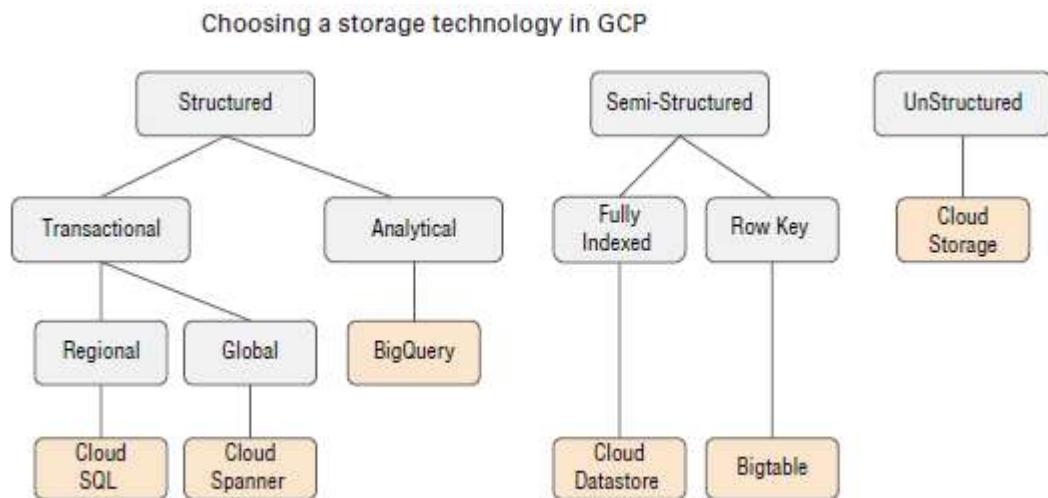
Batch data is ingested in bulk, typically infiles. Examples of batch data ingestion include uploading files of data exported from one application to be processed by another. Both batch and streaming data can be transformed and processed using Cloud Dataflow.

Technical aspects of choosing storage system

These factors include the volume and velocity of data, the type of structure of the data, access control requirements, and data access patterns.

Googles Storage Decision Tree

Google has developed a decision tree for choosing a storage system that starts distinguishing structured, semi-structured, and unstructured data.



Levels of structured data

These levels are structured, semi-structured, and unstructured. Structured data has a fixed schema, such as a relational database table. Semi-structured data has a schema that can vary; the schema is stored with data. Unstructured data does not have a structure used to determine how to store data.

Google Cloud storage services are used with the different structure types

Structured data is stored in Cloud SQL and Cloud Spanner if it is used with a transaction processing system; BigQuery is used for analytical applications of structured data. Semi-structured data is stored in Cloud Datastore if data access requires full indexing; otherwise, it can be stored in Bigtable. Unstructured data is stored in Cloud Storage.

Difference between relational and NoSQL databases

Relational databases are used for structured data whereas NoSQL databases are used for semi-structured data. The four types of NoSQL databases are key-value, document, wide-column, and graph databases.

Storage Quiz

Quiz 1 – Storage Services

1. You need to store a set of files for an extended period of time. Anytime the data in the files needs to be accessed, it will be copied to a server first, and then the data will be accessed. Files will not be accessed more than once a year. The set of files will all have the same access controls.

What storage solution would you use to store these files?

- A. Cloud Storage Coldline
- B. Cloud Storage Nearline
- C. Cloud Filestore
- D. Bigtable

Correct Answer: A

2. You are uploading files in parallel to Cloud Storage and want to optimize load performance.

What could you do to avoid creating hotspots when writing files to Cloud Storage?

- A. Use sequential names or timestamps for files.
- B. Do not use sequential names or timestamps for files.**
- C. Configure retention policies to ensure that files are not deleted prematurely.
- D. Configure lifecycle policies to ensure that files are always using the most appropriate storage class.

Correct Answer: B

3. As a consultant on a cloud migration project, you have been asked to recommend a strategy for storing files that must be highly available even in the event of a regional failure. What would you recommend?

- A. BigQuery
- B. Cloud Datastore**
- C. Multiregional Cloud Storage
- D. Regional Cloud Storage

Correct Answer: C

4. As part of a migration to Google Cloud Platform, your department will run a collaboration and document management application on Compute Engine virtual machines. The application requires a filesystem that can be mounted using operating system commands. All documents should be accessible from any instance. What storage solution would you recommend?

- A. Cloud Storage
- B. Cloud Filestore**
- C. A document database
- D. A relational database

Correct Answer: B

5. Your team currently supports seven MySQL databases for transaction processing applications. Management wants to reduce the amount of staff time spent on database administration. What GCP service would you recommend to help reduce the database administration load on your teams?

- A. Bigtable
- B. BigQuery
- C. Cloud SQL**
- D. Cloud Filestore

Correct Answer: C

6. Your company is developing a new service that will have a global customer base. The service will generate large volumes of structured data and require the support of a transaction processing database. All users, regardless of where they are on the globe, must have a consistent view of data. What storage system will meet these requirements?

- A. Cloud Spanner**
- B. Cloud SQL
- C. Cloud Storage
- D. BigQuery

Correct Answer: A

7. Your company is required to comply with several government and industry regulations, which include encrypting data at rest. What GCP storage services can be used for applications subject to these regulations?

- A. Bigtable and BigQuery only
- B. Bigtable and Cloud Storage only
- C. Any of the managed databases, but no other storage services
- D. Any GCP storage service**

Correct Answer: D

8. As part of your role as a data warehouse administrator, you occasionally need to export data from the data warehouse, which is implemented in BigQuery. What command-line tool would you use for that task?

- A. gsutil
- B. gcloud
- C. bq**
- D. cbt

Correct Answer: C

9. Another task that you perform as data warehouse administrator is granting authorizations to perform tasks with the BigQuery data warehouse. A user has requested permission to view table data but not change it. What role would you grant to this user to provide the needed permissions but nothing more?

- A. dataViewer**
- B. admin
- C. metadataViewer
- D. dataOwner

Correct Answer: A

10. A developer is creating a set of reports and is trying to minimize the amount of data each query returns while still meeting all requirements. What bq command-line option will help you understand the amount of data returned by a query without actually executing the query?

- A. --no-data
- B. --estimate-size
- C. --dry-run**
- D. --size

Correct Answer: C

11. A team of developers is choosing between using NoSQL or a relational database. What is a feature of NoSQL databases that is not available in relational databases?

- A. Fixed schemas
- B. ACID transactions
- C. Indexes
- D. Flexible schemas**

Correct Answer: D

12. A group of venture capital investors have hired you to review the technical design of a service that will be developed by a startup company seeking funding. The startup plans to collect data from sensors attached to vehicles. The data will be used to predict when a vehicle needs maintenance and before the vehicle breaks down. Thirty sensors will be on each vehicle. Each sensor will send up to 5K of data every second. The startup expects to start with hundreds of vehicles, but it plans to reach 1 million vehicles globally within 18 months. The data will be used to develop machine learning models to predict the need for maintenance. The startup is planning to use a self-managed relational database to store the time-series data. What would you recommend for a time-series database?

- A. Continue to plan to use a self-managed relational database.
- B. Use a Cloud SQL.
- C. Use Cloud Spanner.
- D. Use Bigtable.**

Correct Answer: D

13. A Bigtable instance increasingly needs to support simultaneous read and write operations. You'd like to separate the workload so that some nodes respond to read requests and others respond to write requests. How would you implement this to minimize the workload on developers and database administrators?

- A. Create two instances, and separate the workload at the application level.
- B. Create multiple clusters in the Bigtable instance, and use Bigtable replication to keep the clusters synchronized**
- C. Create multiple clusters in the Bigtable instance, and use your own replication program to keep the clusters synchronized.
- D. It is not possible to accomplish the partitioning of the workload as described.

Correct Answer: B

14. As a database architect, you've been asked to recommend a database service to support an application that will make extensive use of JSON documents. What would you recommend to minimize database administration overhead while minimizing the work required for developers to store JSON data in the database?

- A. Cloud Storage
- B. Cloud Datastore**
- C. Cloud Spanner
- D. Cloud SQL

Correct Answer: B

15. Your Cloud SQL database is close to maximizing the number of read operations that it can perform. You could vertically scale the database to use a larger instance, but you do not need additional write capacity. What else could you try to reduce the number of reads performed by the database?

- A. Switch to Cloud Spanner.
- B. Use Cloud Bigtable instead.
- C. Use Cloud Memorystore to create a database cache that stores the results of database queries.** Before a query is sent to the database, the cache is checked for the answer to the query.
- D. There is no other option—you must vertically scale.

Correct Answer: C

16. You would like to move objects stored in Cloud Storage automatically from regional storage to Nearline storage when the object is 6 months old. What feature of Cloud Storage would you use?

- A. Retention policies
- B. Lifecycle policies**
- C. Bucket locks
- D. Multiregion replication

Correct Answer: B

17. A customer has asked for help with a web application. Static data served from a data center in Chicago in the United States loads slowly for users located in Australia, South Africa, and Southeast Asia. What would you recommend to reduce latency?

- A. Distribute data using Cloud CDN.**
- B. Use Premium Network from the server in Chicago to client devices.
- C. Scale up the size of the web server.
- D. Move the server to a location closer to users.

Correct Answer: A

18. Your company wants to reduce cost on infrequently accessed data by moving it to the cloud. The data will still be accessed approximately once a month to refresh historical charts. In addition, data older than 5 years is no longer needed. How should you store and manage the data?

- A. Google Cloud Storage, Nearline bucket
- B. Google Cloud Storage, Coldline bucket
- C. Google Cloud Storage, Multi regional bucket
- D. Google Cloud Storage, Regional bucket

Correct Answer: A

19. You are developing an application that transcodes large video files. Which storage option is the best choice for your application?

- A. Cloud Storage
- B. Cloud SQL
- C. Cloud Bigtable
- D. Cloud Datastore

Correct Answer: A

20. Your company collects and stores security camera footage in Google Cloud Storage. Within the first 30 days, footage is processed regularly for threat detection, object detection, trend analysis, and suspicious behavior detection. You want to minimize the cost of storing all the data. How should you store the videos?

- A. Use Google Cloud Regional Storage for the first 30 days, and then move to Coldline Storage
- B. Use Google Cloud Nearline Storage for the first 30 days, and then move to Coldline Storage
- C. Use Google Cloud Regional Storage for the first 30 days, and then move to Nearline Storage
- D. Use Google Cloud Regional Storage for the first 30 days, and then move to Google Persistent Disk

Correct Answer: A

21. You are developing an application that transcodes large video files. Which storage option is the best choice for your application?

- A. Cloud Storage
- B. Cloud Spanner
- C. Cloud Datastore
- D. Firebase

Correct Answer: A

22. You manufacture devices with sensors and need to stream huge amounts of data from these devices to a storage option in the cloud. Which Google Cloud Platform storage option is the best choice for your application?

- A. Cloud Spanner
- B. Cloud Datastore
- C. BigQuery
- D. Cloud Bigtable**

Correct Answer: D

23. Which service should be used for 100GB of Relational DB is good for?

- A. Cloud Storage
- B. Cloud SQL**
- C. Cloud Datastore
- D. Firebase

Correct Answer: B

24. Which service should be used for 100TB of Relational DB?

- A. Cloud Storage
- B. Cloud SQL
- C. Cloud Datastore
- D. Cloud Spanner**

Correct Answer: D

25. Which service should be used for 10GB of data but Complex queries?

- A. Cloud Storage
- B. Cloud SQL**
- C. Cloud Datastore
- D. Cloud Spanner

Correct Answer: B

26. Which service should be used for Wide column of data and Complex queries?

- A. Cloud Storage
- B. Cloud SQL
- C. Cloud Datastore
- D. BigQuery**

Correct Answer: D

27. Which service should be used for Time-series data, such as CPU and memory usage over time for multiple servers?

- A. Cloud Bigtable**
- B. Cloud SQL
- C. Cloud Datastore
- D. BigQuery

Correct Answer: A

28. Which service should be used for Marketing data, such as purchase histories and customer preferences.?

- A. Cloud Bigtable**
- B. Cloud SQL
- C. Cloud Datastore
- D. BigQuery

Correct Answer: A

29. Which service should be used for financial data, such as transaction histories, stock prices, and currency exchange rates?

A. Cloud Bigtable

B. Cloud SQL

C. Cloud Datastore

D. BigQuery

Correct Answer: A

30. Which service should be used for Internet of Things data, such as usage reports from energy meters and home appliances?

A. Cloud Bigtable

B. Cloud Spanner

C. Cloud Datastore

D. BigQuery

Correct Answer: A

31. Which service should be used for Graph data, such as information about how users are connected to one another?

A. Cloud Bigtable

B. Cloud Spanner

C. Cloud Datastore

D. BigQuery

Correct Answer: A

32. Which service should be used If you need interactive querying in an online analytical processing (OLAP) system?

- A. Cloud Bigtable
- B. Cloud Spanner
- C. Cloud Datastore
- D. BigQuery**

Correct Answer: D

33. A database administrator (DBA) who is new to Google Cloud has asked for your help configuring network access to a Cloud SQL PostgreSQL database. The DBA wants to ensure that traffic is encrypted while minimizing administrative tasks, such as managing SQL certificates. What would you recommend?

- A. Use the TLS protocol
- B. Use Cloud SQL Proxy**
- C. Use a private IP address
- D. Configure the database instance to use auto-encryption

Correct Answer: B

34. You created a Cloud SQL database that uses replication to improve read performance. Occasionally, the read replica will be unavailable. You haven't noticed a pattern, but the disruptions occur once or twice a month. No DBA operations are occurring when the incidents occur. What might be the cause of this issue?

- A. The read replica is being promoted to a standalone Cloud SQL instance.
- B. Maintenance is occurring on the read replica.**
- C. A backup is being performed on the read replica.
- D. The primary Cloud SQL instance is failing over to the read replica.

Correct Answer: B

35. Your department is experimenting with using Cloud Spanner for a globally accessible database. You are starting with a pilot project using a regional instance. You would like to follow Google's recommendations for the maximum sustained CPU utilization of a regional instance. What is the maximum CPU utilization that you would target?

- A. 50%
- B. 65%**
- C. 75%
- D. 45%

Correct Answer: B

36. A Cloud Spanner database is being deployed in us-west1 and will have to store up to 20 TB of data. What is the minimum number of nodes required?

- A. 10**
- B. 20
- C. 5
- D. 40

Correct Answer: A

37. A software-as-a-service (SaaS) company specializing in automobile IoT sensors collects streaming time-series data from tens of thousands of vehicles. The vehicles are owned and operated by 40 different companies, who are the primary customers of the SaaS company. The data will be stored in Bigtable using a multitenant database; that is, all customer data will be stored in the same database. The data sent from the IoT device includes a sensor ID, which is globally unique; a timestamp; and several metrics about engine efficiency. Each customer will query their own data only. Which of the following would you use as a row-key?

- A. Customer ID, timestamp, sensor ID
- B. Customer ID, sensor ID, timestamp**
- C. Sensor ID, timestamp, customer ID
- D. Sensor ID, customer ID, timestamp

Correct Answer: B

38. A team of game developers is using Cloud Firestore to store player data, including character description, character state, and possessions. Descriptions are up to a 60-character alphanumeric string that is set when the character is created and not updated. Character state includes health score, active time, and passive time. When they are updated, they are all updated at the same time. Possessions are updated whenever the character acquires or loses a possession. Possessions may be complex objects, such as bags of items, where each item may be a simple object or another complex object. Simple objects are described with a character string. Complex objects have multiple properties. How would you model player data in Cloud Firestore?

- A. Store description and character state as strings and possessions as entities
- B. Store description, character state, and possessions as strings
- C. Store description, character state, and possessions as entities
- D. Store description as a string; character state as an entity with properties for health score, active time, and passive time; and possessions as an entity that may have embedded entities**

Correct Answer: D

39. You are querying a Cloud Firestore collection of order entities searching for all orders that were created today and have a total sales amount of greater than \$100. You have not excluded any indexes, and you have not created any additional indexes using index.yaml. What do you expect the results to be?

- A. A set of all orders created today with a total sales amount greater than \$100
- B. A set of orders created today and any total sales amount
- C. A set of with total sales amount greater than \$100 and any sales date
- D. No entities returned**

Correct Answer: D

40. You are running a Redis cache using Cloud Memorystore. One day, you receive an alert notification that the memory usage is exceeding 80 percent. You do not want to scale up the instance, but you need to reduce the amount of memory used. What could you try?

- A. Setting shorter TTLs and trying a different eviction policy.
- B. Switching from Basic Tier to Standard Tier.
- C. Exporting the cache.
- D. There is no other option—you must scale the instance.

Correct Answer: A

41. A team of machine learning engineers are creating a repository of data for training and testing machine learning models. All of the engineers work in the same city, and they all contribute datasets to the repository. The data files will be accessed frequently, usually at least once a week. The data scientists want to minimize their storage costs. They plan to use Cloud Storage; what storage class would you recommend?

- A. Regional
- B. Multi-regional
- C. Nearline
- D. Coldline

Correct Answer: A

42. Auditors have informed your company CFO that to comply with a new regulation, your company will need to ensure that financial reporting data is kept for at least three years. The CFO asks for your advice on how to comply with the regulation with the least administrative overhead. What would you recommend?

- A. Store the data on Coldline storage
- B. Store the data on multi-regional storage
- C. Define a data retention policy
- D. Define a lifecycle policy

Correct Answer: C

43. As a database administrator tasked with migrating a MongoDB instance to Google Cloud, you are concerned about your ability to configure the database optimally. You want to collect metrics at both the instance level and the database server level. What would you do in addition to creating an instance and installing and configuring MongoDB to ensure that you can monitor key instances and database metrics?

- A. Install Stackdriver Logging agent.
- B. Install Stackdriver Monitoring agent.**
- C. Install Stackdriver Debug agent.
- D. Nothing. By default, the database instance will send metrics to Stackdriver.

Correct Answer: B

44. A group of data scientists have uploaded multiple time-series datasets to BigQuery over the last year. They have noticed that their queries—which select up to six columns, apply four SQL functions, and group by the day of a timestamp—are taking longer to run and are incurring higher BigQuery costs as they add data. They do not understand why this is the case since they typically work only with the most recent set of data loaded. What would you recommend they consider in order to reduce query latency and query costs?

- A. Sort the data by time order before loading
- B. Stop using Legacy SQL and use Standard SQL dialect
- C. Partition the table and use clustering**
- D. Add more columns to the SELECT statement to use data fetched by BigQuery more efficiently

Correct Answer: C

45. You are querying a BigQuery table that has been partitioned by time. You create a query and use the --dry_run flag with the bq query command. The amount of data scanned is far more than you expected. What is a possible cause of this?

- A. You did not include _PARTITIONTIME in the WHERE clause to limit the amount of data that needs to be scanned.**
- B. You used CSV instead of AVRO file format when loading the data.
- C. Both active and long-term data are included in the query results.

D. You used JSON instead of the Parquet file format when loading the data.

Correct Answer: A

46. Your department is planning to expand the use of BigQuery. The CFO has asked you to investigate whether the company should invest in flat-rate billing for BigQuery. What tools and data would you use to help answer that question?

- A. Stackdriver Logging and audit log data
- B. Stackdriver Logging and CPU utilization metrics
- C. Stackdriver Monitoring and CPU utilization metrics
- D. Stackdriver Monitoring and slot utilization metrics**

Correct Answer: D

47. You are migrating several terabytes of historical sensor data to Google Cloud Storage. The data is organized into files with one file per sensor per day. The files are named with the date followed by the sensor ID. After loading 10 percent of the data, you realize that the data loads are not proceeding as fast as expected. What might be the cause?

- A. The file naming convention uses dates as the first part of the file name. If the files are loaded in this order, they may be creating hotspots when writing the data to Cloud Storage.**
- B. The data is in text instead of Avro or Parquet format.
- C. You are using a gcloud command-line utility instead of the REST API.
- D. The data is being written to regional instead of multi-regional storage.

Correct Answer: A

48. You are investigating long latencies in Cloud Bigtable query response times. Most queries finish in less than 20 ms, but the 99th percentile queries can take up to 400 ms. You examine a Key Visualizer heatmap and see two areas with bright colors indicating hotspots. What could be causing those hotspots?

- A. Improperly used secondary index
- B. Less than optimal partition key
- C. Improperly designed row-key**
- D. Failure to use a read replica

Correct Answer: C

49. An IoT startup has hired you to review their Cloud Bigtable design. The database stores data generated by over 100,000 sensors that send data every 60 seconds. Each row contains all the data for one sensor sent during an hour. Hours always start at the top of the hour. The row-key is the sensor ID concatenated to the hour of the day followed by the date. What change, if any, would you recommend to this design?

- A. Use one row per sensor and 60-second datasets instead of storing multiple datasets in a single row.**
- B. Start the row keyrow-key with the day and hour instead of the sensor ID.
- C. Allow hours to start at any arbitrary time to accommodate differences in sensor clocks.
- D. No change is recommended.

Correct Answer: A

50. Your company has a Cloud Bigtable database that requires strong consistency, but it also requires high availability. You have implemented Cloud Bigtable replication and specified single-cluster routing in the app profile for the database. Some users have noted that they occasionally receive query results inconsistent with what they should have received. The problem seems to correct itself within a minute. What could be the cause of this problem?

- A. Secondary indexes are being updated during the query and return incorrect results when a secondary index is not fully updated.
- B. You have not specified an app configuration file that includes single-cluster routing and use of replicas only for failover.**

- C. Tablets are being moved between nodes, which can cause inconsistent query results.
- D. The row-key is not properly designed.

Correct Answer: B

51. You have been tasked with migrating a MongoDB database to Cloud Spanner. MongoDB is a document database, similar to Cloud Firestore. You would like to maintain some of the document organization of the MongoDB design. What data type, available in Cloud Spanner, would you use to define a column that can hold a document-like structure?

- A. Array
- B. String
- C. STRUCT**
- D. JSON

Correct Answer: C

52. An application using a Cloud Spanner database has several queries that are taking longer to execute than the users would like. You review the queries and notice that they all involve joining three or more tables that are all related hierarchically. What feature of Cloud Spanner would you try in order to improve the query performance?

- A. Replicated clusters
- B. Interleaved tables**
- C. STORING clause
- D. Execution plans

Correct Answer: B

53. A Cloud Spanner database is using a natural key as the primary key for a large table. The natural key is the preferred key by users because the values are easy to relate to other data. Database administrators notice that these keys are causing hotspots on Cloud Spanner nodes and are adversely affecting performance. What would you recommend in order to improve performance?

- A. Keep the data of the natural key in the table but use a hash of the natural key as the primary key
- B. Keep the natural key and let Cloud Spanner create more splits to improve performance
- C. Use interleaved tables
- D. Use more secondary indexes

Correct Answer: A

54. You are using a UUID as the primary key in a Cloud Spanner database. You have noticed hotspotting that you did not anticipate. What could be the cause?

- A. You have too many secondary indexes.
- B. You have too few secondary indexes.
- C. You are using a type of UUID that has sequentially ordered strings at the beginning of the UUID.
- D. You need to make the maximum length of the primary key longer.

Correct Answer: C

55. You are working for a financial services firm on a Cloud Bigtable database. The database stores equity and bond trading information from approximately 950 customers. Over 10,000 equities and bonds are tracked in the database. New data is received at a rate of 5,000 data points per minute. What general design pattern would you recommend?

- A. Tall and narrow table
- B. One table for each customer
- C. One table for equities and one for bonds
- D. Option A and Option B
- E. Option A and Option C

Correct Answer: E

56. You have been brought into a large enterprise to help with a data warehousing initiative. The first project of the initiative is to build a repository for all customer-related data, including sales, finance, inventory, and logistics. It has not yet been determined how the data will be used. What Google Cloud storage system would you recommend that the enterprise use to store that data?

- A. Cloud Bigtable
- B. BigQuery
- C. Cloud Spanner
- D. Cloud Storage**

Correct Answer: D

57. Data is streaming into a BigQuery table. As the data arrives, it is added to a partition that was automatically created that day. Data that arrives the next day will be written to a different partition. The data modeler did not specify a column to use as a partition key. What kind of partition is being used?

- A. Ingestion time partitioned tables**
- B. Timestamp partitioned tables
- C. Integer range partitioned tables
- D. Clustered tables

Correct Answer: A

58. You are designing a BigQuery database with multiple tables in a single dataset. The data stored in the dataset is measurement data from sensors on vehicles in the company's fleet. Data is collected on each vehicle and downloaded at the end of each shift. After that, it is loaded into a partitioned table. You want to have efficient access to the most interesting data, which you define as a particular measurement having a value greater than 100.00.

You want to cluster on that measurement column, which is a FLOAT64. When you define the table with a timestamped partitioned table and clustering on the measurement column, you receive an error. What could that error be?

- A. You cannot use clustering on an external table.
- B. You cannot use clustering with a FLOAT64 column as the clustering key.**

- C. The table is not the FLOAT64 partition type.
- D. The clustering key must be an integer or timestamp.

Correct Answer: B

59. What data formats are supported for external tables in Cloud Storage and Google Drive?

- A. Comma-separated values only
- B. Comma-separated values and Avro
- C. Comma-separated values, Avro, and newline-delimited JSON**
- D. Comma-separated values, Avro, newline-delimited JSON, and Parquet

Correct Answer: C

Quiz 2 - Selecting Appropriate Storage Technologies

1. A developer is planning a mobile application for your company's customers to use to track information about their accounts. The developer is asking for your advice on storage technologies. In one case, the developer explains that they want to write messages each time a significant event occurs, such as the client opening, viewing, or deleting an account. This data is collected for compliance reasons, and the developer wants to minimize administrative overhead. What system would you recommend for storing this data?

- A. Cloud SQL using MySQL
- B. Cloud SQL using PostgreSQL
- C. Cloud Datastore
- D. Stackdriver Logging**

Correct Answer: D

2. You are responsible for developing an ingestion mechanism for a large number of IoT sensors. The ingestion service should accept data up to 10 minutes late. The service should also perform some transformations before writing the data to a database. Which of the managed services would be the best option for managing late arriving data and performing transformations?

- A. Cloud Dataproc
- B. Cloud Dataflow**
- C. Cloud Dataprep
- D. Cloud SQL

Correct Answer: B

3. A team of analysts has collected several CSV datasets with a total size of 50 GB. They plan to store the datasets in GCP and use Compute Engine instances to run RStudio, an interactive statistical application. Data will be loaded into RStudio using an RStudio data loading tool. Which of the following is the most appropriate GCP storage service for the datasets?

- A. Cloud Storage**
- B. Cloud Datastore

C. MongoDB

D. Bigtable

Correct Answer: A

4. A team of analysts has collected several terabytes of telemetry data in CSV datasets. They plan to store the datasets in GCP and query and analyze the data using SQL. Which of the following is the most appropriate GCP storage service for the datasets?

A. Cloud SQL

B. Cloud Spanner

C. BigQuery

D. Bigtable

Correct Answer: C

5. You have been hired to consult with a startup that is developing software for self-driving vehicles. The company's product uses machine learning to predict the trajectory of persons and vehicles. Currently, the software is being developed using 20 vehicles, all located in the same city. IoT data is sent from vehicles every 60 seconds to a MySQL database running on a Compute Engine instance using an n2-standard-8 machine type with 8 vCPUs and 16 GB of memory. The startup wants to review their architecture and make any necessary changes to support tens of thousands of self-driving vehicles, all transmitting IoT data every second. The vehicles will be located across North America and Europe. Approximately 4 KB of data is sent in each transmission. What changes to the architecture would you recommend?

A. None. The current architecture is well suited to the use case.

B. Replace Cloud SQL with Cloud Spanner.

C. Replace Cloud SQL with Bigtable.

D. Replace Cloud SQL with Cloud Datastore.

Correct Answer: C

6. As a member of a team of game developers, you have been tasked with devising a way to track players' possessions. Possessions may be purchased from a catalog, traded with other players, or awarded for game activities. Possessions are categorized as clothing, tools, books, and coins. Players may have any number of possessions of any type. Players can search for other players who have particular possession types to facilitate trading. The game designer has informed you that there will likely be new types of possessions and ways to acquire them in the future. What kind of a data store would you recommend using?

- A. Transactional database
- B. Wide-column database
- C. Document database**
- D. Analytic database

Correct Answer: C

7. The CTO of your company wants to reduce the cost of running an HBase and Hadoop cluster on premises. Only one HBase application is run on the cluster. The cluster currently supports 10 TB of data, but it is expected to double in the next six months. Which of the following managed services would you recommend to replace the on-premises cluster in order to minimize migration and ongoing operational costs?

- A. Cloud Bigtable using the HBase API**
- B. Cloud Dataflow using the HBase API
- C. Cloud Spanner
- D. Cloud Datastore

Correct Answer: A

8. A genomics research institute is developing a platform for analyzing data related to genetic diseases. The genomics data is in a specialized format known as FASTQ, which stores nucleotide sequences and quality scores in a text format. Files may be up to 400 GB and are uploaded in batches. Once the files finish uploading, an analysis pipeline runs, reads the data in the FASTQ file, and outputs data to a database. What storage system is a good option for storing the uploaded FASTQ data?

- A. Cloud Bigtable

B. Cloud Datastore

C. Cloud Storage

D. Cloud Spanner

Correct Answer: C

9. A genomics research institute is developing a platform for analyzing data related to genetic diseases. The genomics data is in a specialized format known as FASTQ, which stores nucleotide sequences and quality scores in a text format. Once the files finish uploading, an analysis pipeline runs, reads the data in the FASTQ file, and outputs data to a database. The output is in tabular structure, the data is queried using SQL, and typically queries retrieve only a small number of columns but many rows. What database would you recommend for storing the output of the workflow?

A. Cloud Bigtable

B. Cloud Datastore

C. Cloud Storage

D. BigQuery

Correct Answer: D

10. You are developing a new application and will be storing semi-structured data that will only be accessed by a single key. The total volume of data will be at least 40 TB. What GCP database service would you use?

A. BigQuery

B. Bigtable

C. Cloud Spanner

D. Cloud SQL

Correct Answer: B

11. A group of climate scientists is collecting weather data every minute from 10,000 sensors across the globe. Data often arrives near the beginning of a minute, and almost all data arrives within the first 30 seconds of a minute. The data ingestion process is losing some data because servers cannot ingest the data as fast as it is arriving. The scientists have scaled up the number of servers in their managed instance group, but that has not completely eliminated the problem. They do not wish to increase the maximum size of the managed instance group. What else can the scientists do to prevent data loss?

- A. Write data to a Cloud Dataflow stream
- B. Write data to a Cloud Pub/Sub topic**
- C. Write data to Cloud SQL table
- D. Write data to Cloud Dataprep

Correct Answer: B

12. A software developer asks your advice about storing data. The developer has hundreds of thousands of 1 KB JSON objects that need to be accessed in sub-millisecond times if possible. All objects are referenced by a key. There is no need to look up values by the contents of the JSON structure. What kind of NoSQL database would you recommend?

- A. Key-value database**
- B. Analytical database
- C. Wide-column database
- D. Graph database

Correct Answer: A

13. A software developer asks your advice about storing data. The developer has hundreds of thousands of 10 KB JSON objects that need to be searchable by most attributes in the JSON structure. What kind of NoSQL database would you recommend?

- A. Key-value database
- B. Analytical database
- C. Wide-column database
- D. Document database**

Correct Answer: D

14. A data modeler is designing a database to support ad hoc querying, including drilling down and slicing and dicing queries. What kind of data model is the data modeler likely to use?

- A. OLTP
- B. OLAP**
- C. Normalized
- D. Graph

Correct Answer: B

15. A multinational corporation is building a global inventory database. The database will support OLTP type transactions at a global scale. Which of the following would you consider as possible databases for the system?

- A. Cloud SQL and Cloud Spanner
- B. Cloud SQL and Cloud Datastore
- C. Cloud Spanner only**
- D. Cloud Datastore only

Correct Answer: C

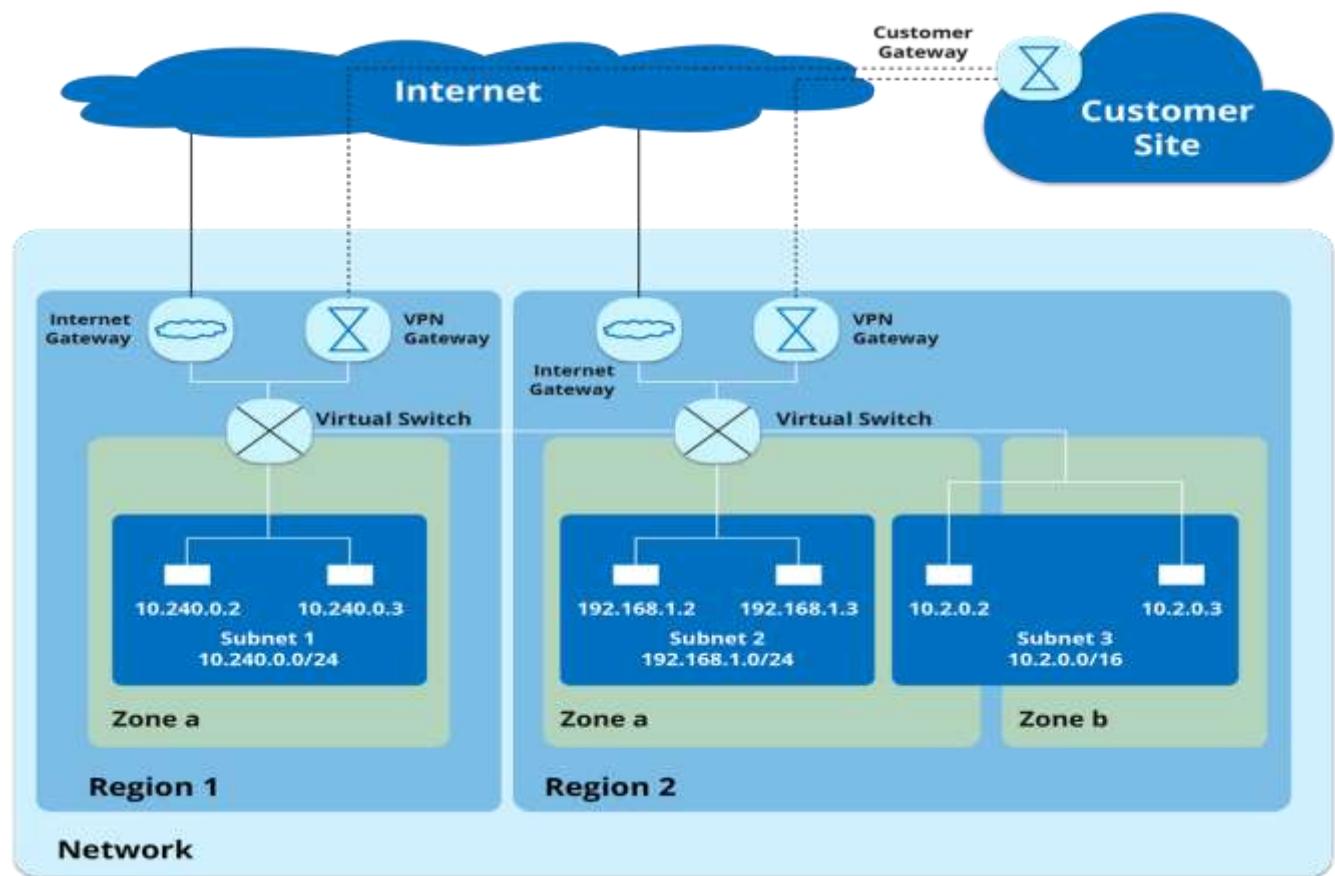
Networking Services

Cloud VPC

What is VPC?

Virtual private clouds are like a network in a data center; they are network-based organizational structures for controlling access to GCP resources. i.e., VPC is a virtual version of physical network.

- VPC's organize Compute Engine instances, App Engine Flexible instances, and GKE clusters. They are global resources, so a single VPC can span multiple regions.
- A VPC is associated with a project or an organization, and projects can have multiple VPCs. Resources within a VPC can communicate with other resources in the same VPC, subject to firewall rules. Resources can also communicate with Google APIs and services.



VPC Network Diagram shows VPN, Region, Zones, Subnet fit into end-to-end schema of things connected to customer site

Entire World's 40 % of traffic can be used by google network. One of the Fastest networks in the world is Google's Network

VPC Provides global, scalable, flexible networking for your cloud-based services

VPC Allows communication between VM instances and other resources via internal, private IP addresses

- VPC's are Global: Resources from across zones, regions
 - Different zones same | different region
 - All machines communicate using internal IP addresses
- VPC's are Scalable: Add new VMs, containers to the network
- VPC's are Multi-tenancy: VPCs can be shared across GCP projects
- VPC's are Private and secure: IAM, firewall rules

What are the Features of VPC Network?

The features of VPC Network are: -

- IAM Enabled
- Firewall Support
- VPC Sharing
- VPC Peering
- Global Private Network Communication
- Working Privately across Hybrid Environment

What is the relationship between Project and VPC Networks?

- Each project has a default auto- mode VPC network which is created with the project
- A single project has a quota of 5 networks
- A single network has a limit of 7000 instances

What are the different types of VPC Network?

The different types of VPC Network are: - Auto Mode and Custom Mode

Auto Mode

- One Subnet is created in each region automatically
- Automatically created subnets use a set of predefined IP ranges which fit within the range 10.128.0.0/9 CIDR block
- We can add subnets manually, but the IP address should range outside the mentioned CIDR block

Custom Mode

- No Subnets are automatically created, so We decide which subnets to create in which Region
- We can assign any IP address from the specified IP range

Components of VPC Network

The Components of VPC are Subnets, IP Address, Routes, Firewall, VPC Sharing & Peering

VPC Subnets

Highlights

- Subnet is a Logical partitioning of the network and Scope is Regional
- Defined by an IP address prefix range for subnetworks
- A network must have one subnet to use
- Specified in CIDR notation
- IP ranges cannot overlap between subnets
- Subnets can have resources from multiple zones | single zone
- Subnets in the GCP can contain resources only from a single region

A Subnet is a logical sub division of larger network.

A VPC can have subnets in each region in order to provide private addresses to resources in the region. Since the subnets are part of a larger network, they must have distinct IP address ranges.

When a VPC is created, it can automatically create subnets in each region, or you can specify custom subnet definitions for each region that should have a subnet. If subnets are created automatically, their IP ranges are based on the region. All automatic subnets are assigned IP addresses in the 10.nnn.0.0/20 range.

VPCs use routes to determine how to route traffic within the VPC and across subnets. Depending on the configuration of the VPC, the VPC can learn regional routes only or multiregional, global routes.

IP Address

What is IP Address?

An IP address is a string of numbers separated by periods. IP addresses are expressed as a set of four numbers — an example address might be 192.158.1.38. Each number in the set can range from 0 to 255

When you create a subnet, you will have to specify a range of IP addresses. Any resource that needs an IP address on that subnet will receive an IP address in that range. Each subnet in a VPC should have distinct, non-overlapping IP ranges.

You can specify an IP range using the **CIDR notation**. This consists of an IPv4 IP address followed by a /, followed by an integer. The integer specifies the number of bits used to identify the subnet; the remaining bits are used to determine the host address.

In CIDR Notation. For Example, 10.123.9.0/24. Contains all IP addresses in the range 10.123.9.0 to 10.123.9.255. The /24 represents the number of bits which is the network prefix

IP addresses can be assigned to GCP resources (VM, forwarding rules etc.), resources can communicate within the GCP network and outside of networks as well

Each VM instance can have 1 Primary internal IP Address + 1 or more secondary IP Address + 1 external IP Address

Types of IP Address

- Public IP
- Private IP
- External IP
- Internal IP
- Secondary IP

Public and Private IP Address

- Public IP addresses are internet routable
- Private IP addresses are internal and cannot be internet routable

Internal and External IP Address

Internal IP can be used only within the network. Every VPC and on-premise network have internal IPs. Resource with the external IP can communicate with the public internet, and IP is publicly advertised.

Internal IP Address

- Use within a VPC
- Cannot be used across VPCs unless we have special configuration (like shared VPCs or VPNs)
- Can be ephemeral or static, typically ephemeral
- VMs know their internal IP address (VM name and IP is available to the network DNS)

External IP Address

- Use to communicate across VPCs
- Traffic using external IP addresses can cause additional billing charges
- Can be ephemeral or static
- VMs are not aware of their external IP address

Secondary IP Address

- It is also referred as Alias IP
- You can assign a range of internal IP addresses as aliases to a VM's primary network interface
- It is useful if you have multiple services running on a VM and you want to assign each service with a different IP address

Routes

What is Route?

- A route is a mapping of an IP range to a destination
- Routes tell the VPC network where to send packets destined for a particular IP address
- Each VPC network comes with 2 routes. One route tells how to forward packets to every subnet in that network and other routes which is, Default internet gateway route, tells the network how to send packets out of the network

Route is made up of

- name: User-friendly name
- network: The name of the network to which this route applies
- destRange: The destination IP range that this route applies to
- instanceTags: Instance tags that this route applies to, applies to all instances if empty
- priority: Used to break ties in case of multiple matches
- nextHopInstance: Fully qualified URL. Instance must already exist
- nextHopIp: The IP address
- nextHopNetwork: URL of network
- nextHopGateway: URL of gateway
- nextHopVpnTunnel: URL of VPN tunnel

What are the Benefits of Routes?

The benefits of Routes are:-

- Many-to-one NATs
- Multiple hosts mapped to one public IP
- Transparent proxies
- Direct all external traffic to one machine

What are the different types of Routes?

- Static Routes
- Dynamic Routes
- Cloud Routes

Static Routes

Manually created routes, so topology in the network also can be added manually.

Drawbacks

- No auto-reroute suppose if the link fails
- Preferable for Small & Stable networks
- Need to manually create and delete routes whenever a network configuration change happens
- Changes are slow to converge
- Need to tear down and re-create the VPN tunnel for updating routes (this causes traffic disruption)

Dynamic Routes

- Automatically created during subnet creation
- Determines which subnets are visible to Cloud Routers
- Dynamic routes discover the changes automatically and rapidly
- Uses BGP (Border Gateway Protocol) to exchange route information between networks

There are two modes of Dynamic Routes are: -

Global dynamic routing

- Cloud router advertises all subnets in the VPC network to the on-premise router
- Reads and writes data like a messaging system
- Cloud Router advertises all subnets to the on-premise router
- Cloud Router propagates learned routes to all regions

Regional dynamic routing

- Advertises and propagates only those routes in its local region
- This is by-default mode
- Reads and writes data like a messaging system
- Cloud Router advertises and propagates routes in its local region

Cloud Routes

- Fully distributed managed routing service
- Program's dynamic/custom routes and scales with traffic Software defined
- Works over Cloud VPN or Cloud Interconnect connections to provide dynamic routing by using the BGP
- Not supported for Direct Peering or Carrier Peering connections
- Requires Cloud NAT, Cloud Interconnect, and HA VPN

Firewall

Firewall rules control network traffic by blocking or allowing traffic into (ingress) or out of (egress) a network. Two implied firewall rules are defined with VPCs: one blocks all incoming traffic, and the other allows all outgoing traffic.

You can change this behavior by defining firewall rules with higher priority. Firewall rules have a priority specified by an integer from 0 to 65535, with 0 being the highest priority and 65535 being the lowest. The two implied firewall rules have an implied priority of 65535, so you can override those by specifying a priority of less than 65535.



2 Implied Firewall Rules

- Only IPv4 addresses are supported in a firewall rule
- Firewall rules are specific to a network. They cannot be shared between networks
- Tags and service accounts cannot be used together in the same firewall rule
- A default "**allow egress**" rule. Allows all egress connections. Rule has a priority of 65535
- A default "**deny ingress**" rule. Deny all ingress connection. Rule has a priority of 65535

4 Default Firewall Rules

In addition to the two implied rules, which cannot be deleted, there are four default rules assigned to the default network in a VPC. These rules are as follows: -

- **default-allow-internal** allows ingress connections for all protocols and ports among instances in the network.
- **default-allow-ssh** allows ingress connections on TCP port 22 from any source to any instance in the network. This allows users to ssh into Linux servers.
- **Default-allow-rdp** allows ingress connections on TCP port 3389 from any source to any instance in the network. This lets users use Remote Desktop Protocol (RDP) developed by Microsoft to access Windows servers.
- **Default-allow-icmp** allows ingress ICMP traffic from any source to any instance in the network.

All of these above rules have a priority of 65534, the second-lowest priority.

Firewall rules have several attributes in addition to priority. They are as follows:

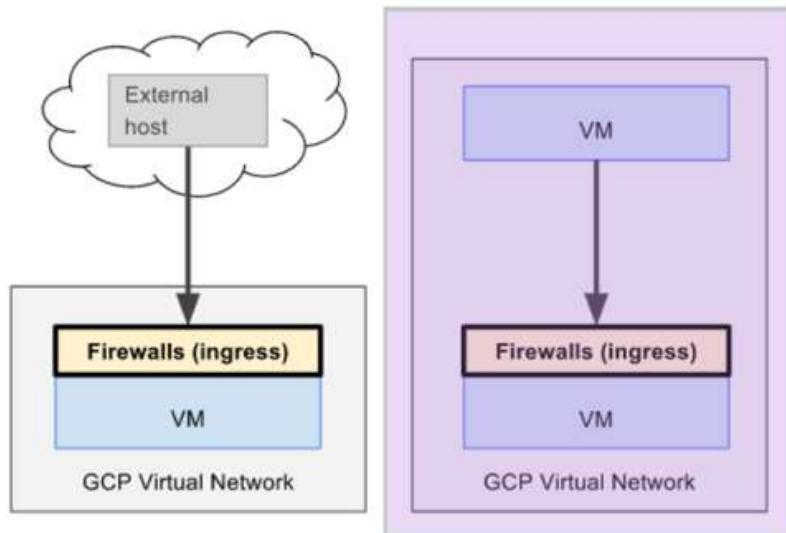
- **The direction of traffic.** This is either ingress or egress.
- **The action.** This is either allow or deny traffic.
- **The target.** This defines the instances to which the rule applies.
- **The source.** This is for ingress rules or the destination for egress rules.
- **A protocol specification.** This includes TCP, UDP, or ICMP, for example.
- **A port numbers.** A communication endpoint associated with a process.
- **An enforcement status.** This allows network administrators to disable a rule without having to delete it.

Firewall rules are global resources that are assigned to VPCs, so they apply to all VPC subnets in all regions. Since they are global resources, they can be used to control traffic between regions in a VPC.

Implied Firewall Rule - Ingress and Egress Connections

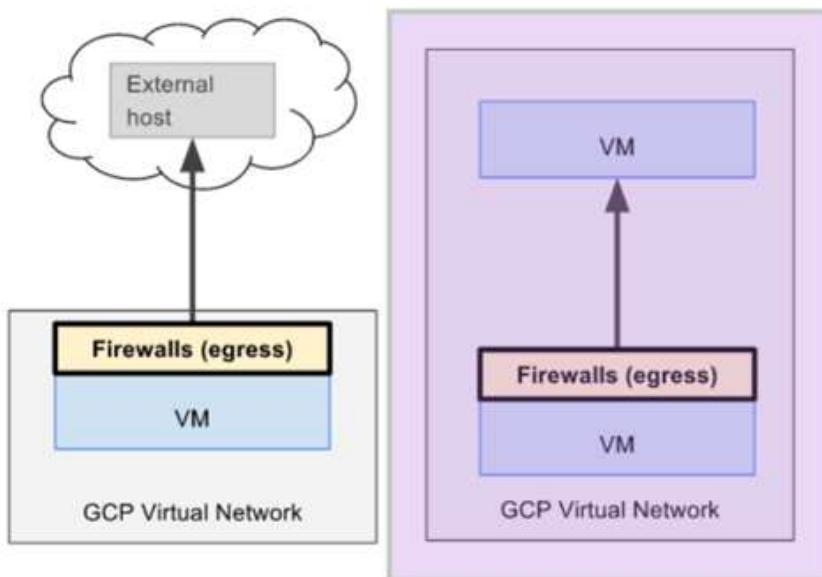
Ingress Connections

- Source CIDR ranges, Protocols, Ports
- Sources with specific tags or service accounts
- Allow: Permit matching ingress connections
- Deny: Block the matching ingress connections



Egress Connection

- Destination CIDR ranges, Protocols, Ports
- Destinations with specific tags or service accounts
- Allow: Permit matching egress connections
- Deny: Block the matching egress connections



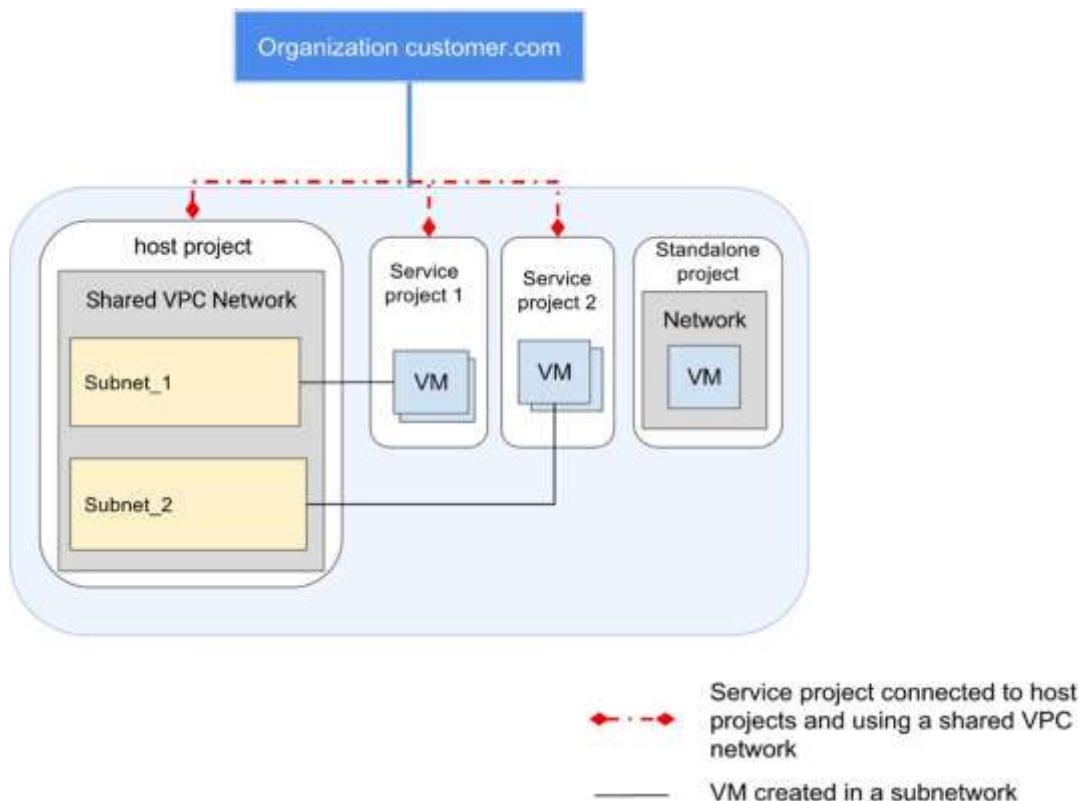
VPC Sharing

What is VPC Sharing?

- VPC can be shared across Projects, also called XPN (Cross Project Networking)
- Multiple Projects one VPC is called VPC Sharing
- Shared VPC allows communication with each other securely and efficiently using internal IPs from the network
- Subnetworks can be shared across Projects
- Helps for IAM access control
- Used typically in large organizations

How Shared VPC works?

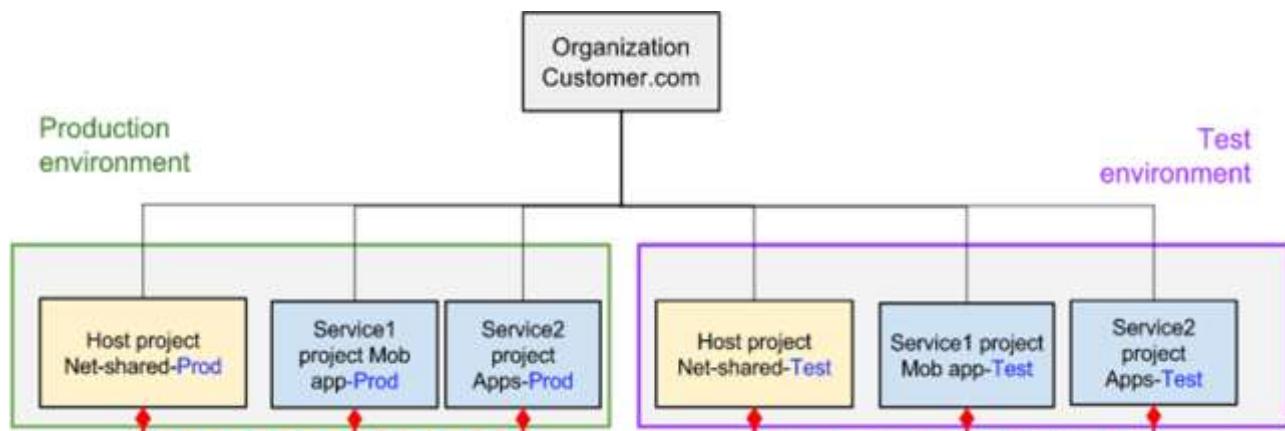
Shared VPCs are comprised of a host project and one or more service projects. The host project contains one or more Shared VPC networks. When a VPC is made a Shared VPC, all of the existing subnetworks become Shared VPC subnets. Service projects are attached to the host project at the project level.



Google Cloud Certified Professional Cloud Architect Definitive Guide

- **Host Project:** Project that hosts sharable VPC networking resources within a Cloud Organization
- **Service project:** Project that has permission to use the shared VPC networking resources from the host project
 - A service project can only be associated with a single host
 - A project cannot be a host as well as a service project at the same time
 - Instances in a project can only be assigned external IPs from the same project
 - Existing projects can use shared VPC networks - Instances on a shared VPC need to be created explicitly for the VPC
- **Standalone project:** A project that does not share networking resources with any other project
- **Shared VPC network:** A VPC network owned by the host project and shared with one or more service projects in the Cloud Organization
- **Organization:** The Cloud Organization is the top level in the Cloud Resource Hierarchy and the top-level owner of all the projects and resources created under it. A given host project and its service projects must be under the same Cloud Organization. A given host project and its service projects must be under the same Cloud Organization

Multiple Shared VPC



VPC Peering

VPC network peering enables different VPC networks to communicate using private IP address space, as defined in RFC 1918. VPC network peering is used as an alternative to using external IP addresses or using VPNs to link networks.

The following are three primary advantages of VPC network peering: -

- There is **lower latency** because the traffic stays on the Google network and is not subject to conditions on the public Internet. i.e., Lower latency as compared with public IP networking
- Services in the VPC are **inaccessible from the public Internet**, reducing the attack surface of the organization. i.e., Provides better security since services need not expose an external IP address
- There are **no egress charges** associated with traffic when using VPC network peering. i.e., Using internal IPs for traffic avoids egress bandwidth pricing on the GCP

It is important to note that peered networks manage their own resources, such as firewall rules and routes. This is different from firewall rules and routes in a VPC, which are associated with the entire VPC. Also, there is a maximum of 25 peering connections from a single VPC.

It is important to note that VPC peering can connect VPCs between organizations; VPC sharing does not operate between organizations.

What is NAT Gateway?

- NAT stands for Network Address Translation
- Enables VM instances to send outbound packets to the internet and to receive any corresponding inbound response packets without external IP addresses and private GKE clusters
- Core Features of NAT Gateway are: -
 - Highly Secured
 - High Availability
 - Superb Performance

Hybrid Cloud Networking

Hybrid-cloud networking is the practice of providing network services between an on-premise data center and a cloud. When two or more public clouds are linked together, that is called a multi-cloud network. Multi-cloud networks may also include private data centers.

Typically, architects recommend hybrid-cloud or multi-cloud environments when there are workloads that are especially well suited to run in one environment over another or when they are trying to mitigate the risk of dependency on a particular cloud service. Here are some examples: A batch processing job that uses a custom legacy application designed for a mainframe is probably best run on-premises.

Ad hoc batch processing, such as transforming a large number of image files to a new format, is a good candidate for a cloud computing environment, especially when low-cost preemptible VMs are available.

An enterprise data warehouse that is anticipated to grow well into petabyte scale is well suited to run in a cloud service such as BigQuery.

Hybrid-cloud Implementation Options

Hybrid-cloud computing is supported by three types of network links.

- Cloud VPN
- Cloud Interconnect
- Direct peering

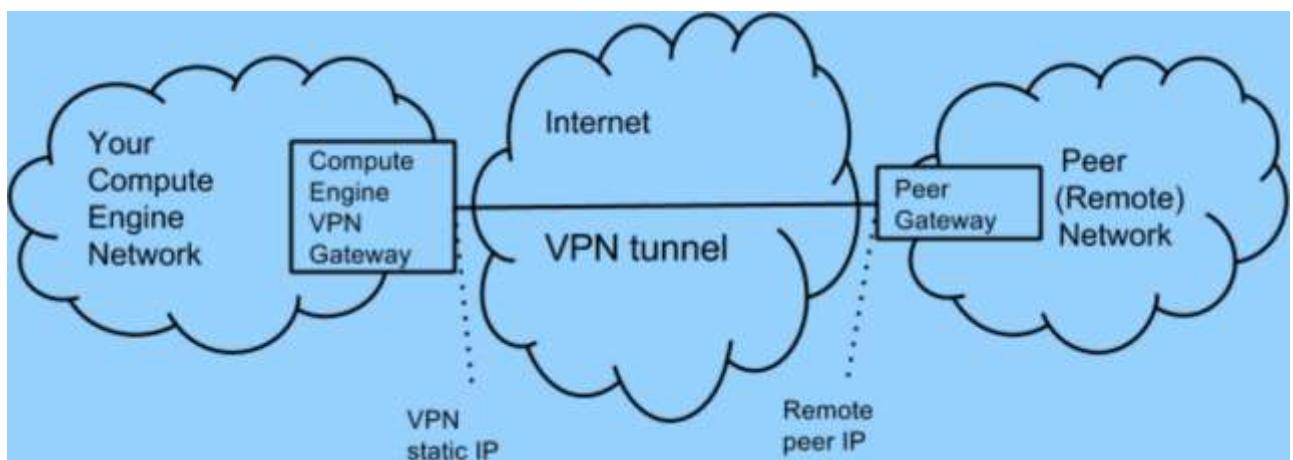
Cloud VPN

What is Cloud VPN?

- Cloud VPN is a GCP service that provides virtual private networks between GCP and on-premises networks.
- Data is transmitted over the public Internet, but the data is encrypted at the origin gateway and decrypted at the destination gateway to protect the confidentiality of data in transit. Encryption is based on the Internet Key Exchange (IKE) protocol
- Supports both static and dynamic routes for traffic between on-premise and cloud
- Supports 99.99% Service Availability
- Cloud VPN is implemented using IPsec VPNs and supports bandwidths up to 3 Gbps.

How VPN works?

- The on-premise network to be connected to the cloud and can also be another cloud VPC
- Only IPsec gateway to gateway scenarios is supported, does not work with client software on a laptop
- Must have a static external IP address
- Needs to know what destination IPs are allowed and create routes to forward packets to those IPs
- Can have multiple tunnels to a single VPN gateway, site-to-site VPN
- By this way the cloud VPC to connect to the on-premise network



What are its disadvantages of VPN?

The disadvantages of VPN are: -

- VPN traffic has to traverse the internet
- VPN will have higher latency and lower throughput as compared with dedicated interconnect and peering options

Cloud Interconnect

What is Cloud Interconnect?

The Cloud Interconnect service provides high throughput and highly available networking between GCP and on-premises networks. Cloud Interconnect is available in 10 Gbps or 100 Gbps configurations when using a direct connection between a Google Cloud access point and your data center. When using a third-party network provider, called a *Partner Interconnect*, customers have the option of configuring 50 Mbps to 10 Gbps connections.

The **advantages** of using Cloud Interconnect include the following: -

- You can transmit data on private connections. Data does not traverse the public Internet
- Private IP addresses in Google Cloud VPCs are directly addressable from on-premises devices
- You have the ability to scale up Direct Interconnects to 80 Gbps using eight 10 Gbps direct interconnects or 200 Gbps using two 100 Gbps interconnects
- You have the ability to scale up Partner Interconnects to 80 Gbps using eight 10 Gbps partner interconnects

A **disadvantage** of Cloud Interconnect is the additional cost and complexity of managing a direct or partnered connection. If low latency and high availability are not required, then using Cloud VPN will be less expensive and require less management.

An alternative to Cloud Interconnect is direct peering.

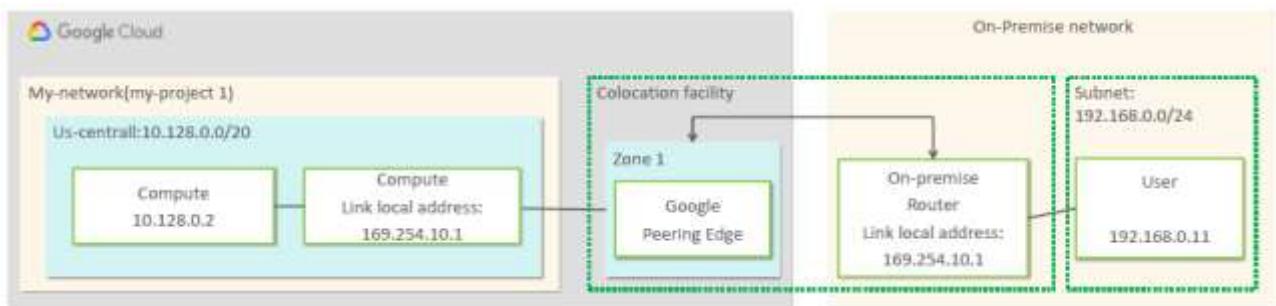
Two Types of Interconnects

There are two Types of Interconnects namely: -

- Dedicated Interconnect
- Partner Interconnect

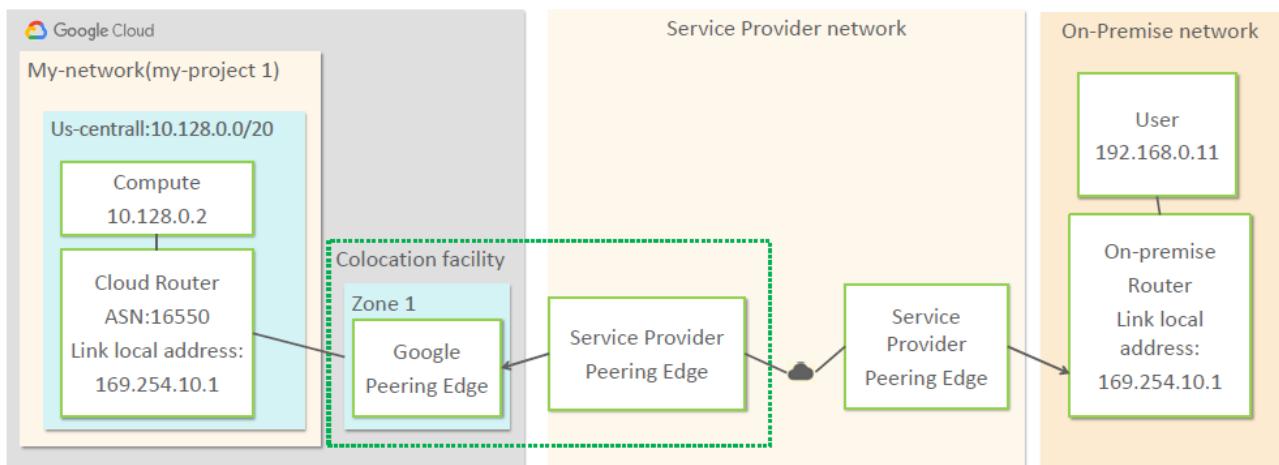
Dedicated Interconnect

- Direct connection from Google
- 10 or 100 GBPS connection
- Should be configured in on-premise
- Google provides end to end



Partner Interconnect

- From google's service provider
- Flexible speed
- BGP requires only for layer two connection
- The service provider provides SLA



Direct Peering

Network peering is a network configuration that allows for routing between networks.

Direct peering is a form of peering that allows customers to connect their networks to a Google network point of access. This kind of connection is not a GCP service—it is a lower-level network connection that is outside of GCP. It works by exchanging Border Gateway Protocol (BGP) routes, which define paths for transmitting data between networks. It does not make use of any GCP resources, like VPC firewall rules or GCP access controls.

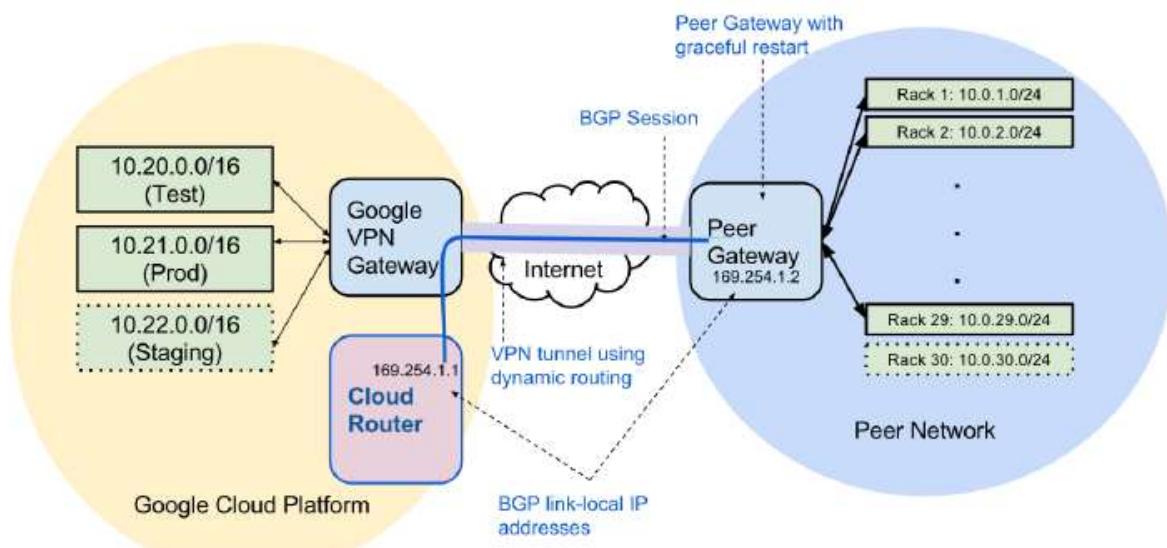
When working with hybrid computing environments, first consider workloads and where they are optimally run and how data is exchanged between networks. This can help you determine the best topology for the hybrid or multi-cloud network.

There are three options for linking networks: interconnect, VPN, and direct peering. Interconnects provide high throughput, low latency, and high availability. VPNs are a lower-cost option that does not require managing site-to-site connections, but throughput is lower. A third, not generally recommended option is direct peering. This is an option when requirements dictate that connection between networks be at the level of exchanging BGP routes.

Cloud Router

What is Cloud Router?

- Cloud Router is fully distributed and managed Google cloud service
- Dynamically exchange routes between VPC and on-premise N/W
- Uses BGP (Border Gateway Protocol)
- Peers with on premise gateway or router to exchange route information



Cloud Load Balancing

What is Load Balancing?

- Load balancing is the practice of distributing work across a set of resources
- Fully managed service, redundant and highly available.
- Route traffic to the closest VM
- Scales your application to support heavy traffic
- Detect and remove unhealthy VMs, healthy VMs automatically re-added

To determine which load balancer is an appropriate choice in a given scenario, you will have to consider three factors.

- Is the workload distributed to servers within a region or across multiple regions?
- Does the load balancer receive traffic from internal GCP resources only or from external sources as well?
- What protocols does the load balancer need to support?

Types of Load Balancer

The five types of load balancers are namely: -

Regional External

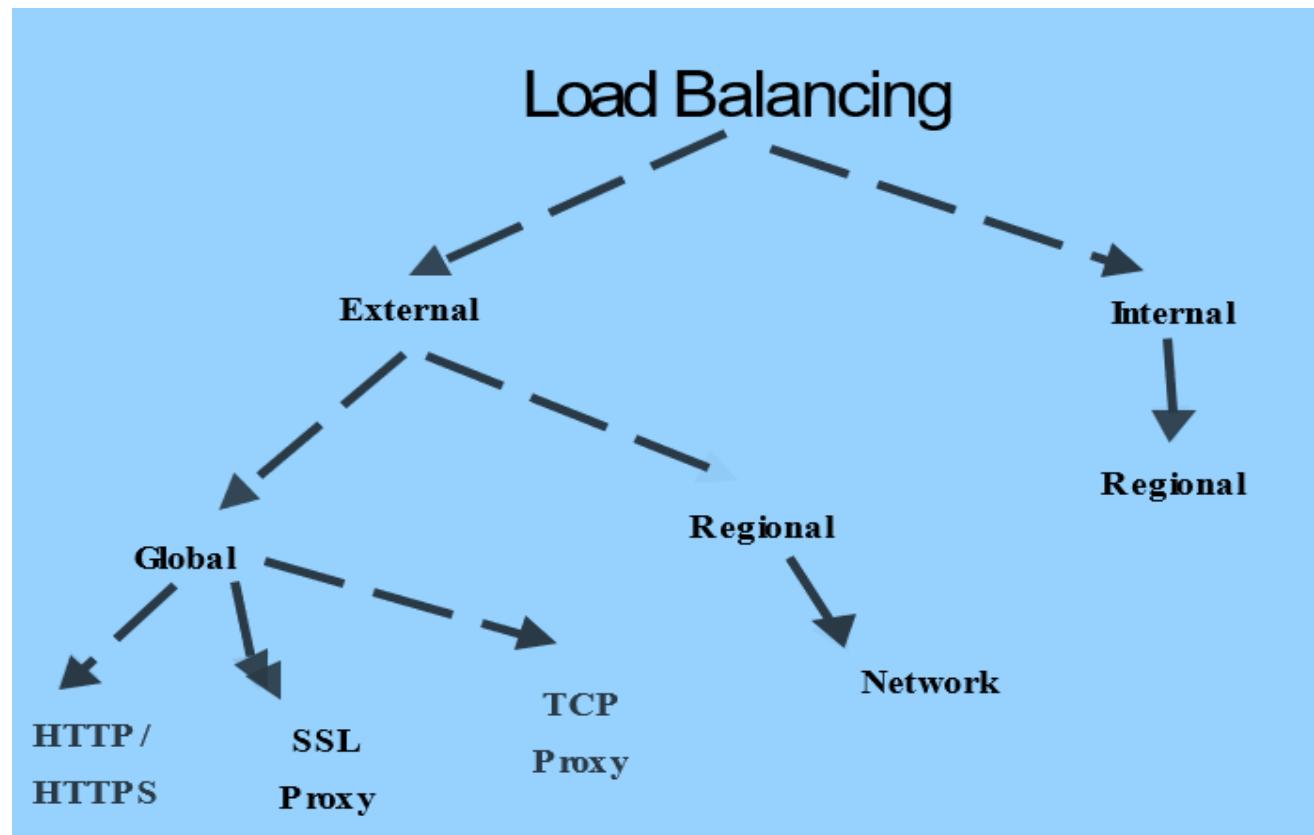
- Network TCP/UDP

Regional Internal

- Internal TCP/UDP

Global

- HTTP(S)
- SSL Proxy
- TCP Proxy



Choosing between these requires understanding if traffic will be distributed within a single region or across multiple regions, which protocols are used, and whether the traffic is internal or external to GCP

Regional External | Internal Load Balancing

The two regional load balancers are Network TCP/UDP and Internal TCP/UDP. Both work with TCP and UDP protocols as their names imply.

Network TCP/UDP

The Network TCP/UDP load balancer distributes workload based on IP protocol, address, and port. This load balancer uses forwarding rules to determine how to distribute traffic. Forwarding rules use the IP address, protocol, and ports to determine which servers, known as a target pool, should receive the traffic.

The Network TCP/UDP is a non-proxied load balancer, which means that it passes data through the load balancer without modification. This load balancer only distributes traffic to servers within the region where the load balancer is configured.

All traffic from the same connection is routed to the same instance. This can lead to imbalance if long-lived connections tend to be assigned to the same instance.

Internal TCP/UDP

The Internal TCP/UDP load balancer is the only internal load balancer. It is used to distribute traffic from GCP resources, and it allows for load balancing using private IP addresses. It is a regional load balancer.

Instances of the Internal TCP/UDP load balancer support routing either TCP or UDP packets but not both. Traffic passes through the Internal TCP/UDP load balancer and is not proxied.

The Internal TCP/UDP load balancer is a good choice when distributing workload across a set of backend services that run on a Compute Engine instance group in which all of the backend instances are assigned private IP addresses.

When traffic needs to be distributed across multiple regions, then one of the global load balancers should be used.

Global Load Balancing

The three global load balancers are the HTTP(S), SSL Proxy, and TCP Proxy Load Balancing load balancers. All global load balancers require the use of the Premium Tier of network services.

HTTP(S) Load Balancing

The HTTP(S) load balancer is used when you need to distribute HTTP and HTTPS traffic globally, or at least across two or more regions.

HTTP(S) load balancers use forwarding rules to direct traffic to a target HTTP proxy. These proxies then route the traffic to a URL map, which determines which target group to send the request to based on the URL. For example, `https://www.example.com/documents` will be routed to the backend servers that serve that kind of request, while `https://www.example.com/images` would be routed to a different target group.

The backend service then routes the requests to an instance within the target group based on capacity, health status, and zone.

In the case of HTTPS traffic, the load balancer uses SSL certificates that must be installed on each of the backend instances.

SSL Proxy Load Balancing

The SSL Proxy load balancer terminates SSL/TLS traffic at the load balancer and distributes traffic across the set of backend servers. After the SSL/TLS traffic has been decrypted, it can be transmitted to backend servers using either TCP or SSL. SSL is recommended. Also, this load balancer is recommended for non-HTTPS traffic; HTTPS traffic should use the HTTP(S) load balancer.

The SSL Proxy load balancers will distribute traffic to the closest region that has capacity. Another advantage of this load balancer is that it offloads SSL encryption/decryption for backend instances.

TCP Proxy Load Balancing

TCP Proxy Load Balancing lets you use a single IP address for all users regardless of where they are on the globe, and it will route traffic to the closest instance.

TCP Proxy load balancers should be used for non-HTTPS and non-SSL traffic.

GCP provides load balancers tailored for regional and global needs as well as specialized to protocols. When choosing a load balancer, consider the geographic distribution of backend instances, the protocol used, and whether the traffic is from internal GCP resources or potentially from external devices.

Cloud DNS

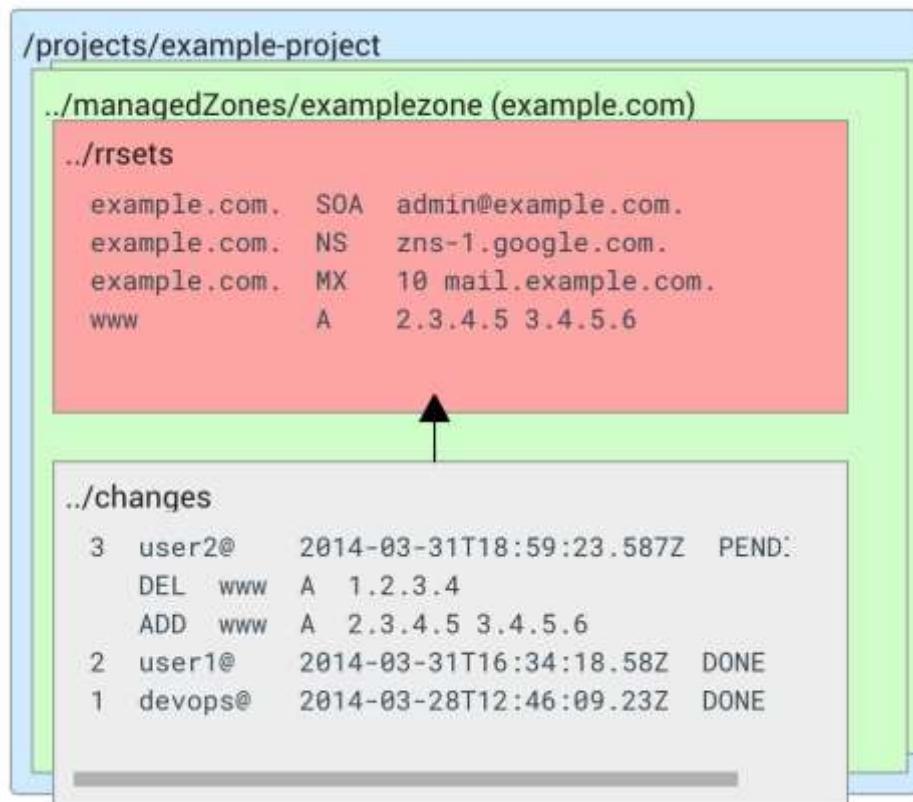
What is Cloud DNS?

- Google Cloud DNS is a high-performance, resilient, global Domain Name System (DNS) service that publishes your domain names to the global DNS in a cost-effective way.
- Translates Domain Names to IP Address
- Hierarchical distributed database that lets you store IP addresses and other data and look them up by name
- Publish zones and records in the DNS
- No burden of managing your own DNS server

What are the advantages of Cloud DNS?

- Easily Publish and Manage
- Highly Available
- Highly Scalable (Autoscaling)
- More Reliable
- More Cost Effective
- Low Latency

How Cloud DNS Works?



Managed Zone

Entity that manages DNS records for a given suffix (example.com)

Maintained by Cloud DNS

Record types

A - Address record, maps hostnames to IPv4 addresses

SOA - Start of authority - specifies authoritative information on a managed zone

MX - Mail exchange used to route requests to mail servers

NS - Name Server record, delegates a DNS zone to an authoritative server

Resource Record Changes

The changes are first made to the authoritative servers and is then picked up by the DNS resolvers when their cache expires

Cloud CDN

What is Cloud CDN?

- CDN stands for Content Delivery Network
- Cloud CDN accelerates content delivery for websites/applications using Google's global edge network to serve content closer to users
- Provides Low latency at less cost
- High Secure but no additional cost
- CDN works with Instance Groups, Zonal | Internet Network Endpoint groups
- Cloud Storage Buckets to deliver the content as backend

How CDN Works?

- For adding the content Once CDN enabled, caching happens automatically based on the user requests
- For serving the content Origin uses Http headers to indicate which response should be cached Cloud CDN uses caches in multiple locations around the globe. All the requests will be served from cache only
- When cache storage gets full, old content will be removedCaches will be usually full, and they constantly evict data generally, non-frequently accessed content is evicted. Content in HTTP(S) caches can have a configurable expiration time

Networking – Interview | Exam Tips

VPC Highlights

VPCs are virtual private clouds that define a network associated with a project. VPCs have subnets. Subnets are assigned IP ranges and all instances within a subnet are assigned IP addresses from its range. VPCs can share resources by setting up Shared VPCs. Shared VPCs have one host project and one or more service projects. VPC network peering enables different VPC networks to communicate using a private IP address space, as defined in RFC 1918. VPC network peering is used as an alternative to using external IP addresses or using VPNs to link networks.

The flow of traffic within a VPC is controlled by firewall rules. Two implied rules allow all outgoing traffic and deny most incoming traffic. Implied rules cannot be deleted, but they can be overridden by higher-priority rules. When subnets are automatically created for a VPC, a set of default rules are created to allow typical traffic patterns, such as using SSH to connect to an instance.

Hybrid-cloud networking is the practice of providing network services between an onpremise data center and a cloud. Design considerations include latency, throughput, reliability, and network topology. Hybrid cloud networks can be implemented using Cloud VPN, Cloud Interconnect, and direct peering.

Load balancing is the practice of distributing work across a set of resources. GCP provides five different load balancers: Network TCP/UDP, Internal TCP/UDP, HTTP(S), SSL Proxy, and TCP Proxy Load Balancing. Choose a load balancer based on regional or multiregional distribution of traffic, protocol, and internal or external traffic

Virtual private clouds

Virtual private clouds are like a network in a data center; they are network-based organizational structures for controlling access to GCP resources. They are global resources, so a single VPC can span multiple regions. VPCs are global resources. Subnets are regional resources.

Shared VPCs

Shared VPCs include a host VPC and one or more service VPCs. Shared VPCs are used to make resources in one project accessible to resources in other projects. Another advantage of Shared VPCs is that you can separate project and network management duties.

Firewall rules

Firewall rules control network traffic by blocking or allowing traffic into (ingress) or out of (egress) a network. Two implied rules allow all outgoing traffic and deny most incoming traffic. Implied rules cannot be deleted, but they can be overridden by higher-priority rules. When subnets are automatically created for a VPC, default rules are created to allow typical traffic patterns. These rules include default-allow-internal, default-allow-ssh, default-allow-rdp, and default-allow-icmp.

CIDR block notation

You can specify an IP range using the CIDR notation. This consists of an IPv4 IP address followed by a /, followed by an integer. The integer specifies the number of bits used to identify the subnet; the remaining bits are used to determine the host address.

Why hybrid-cloud networking is needed

When workloads are run in different environments, there will be a need for reliable networking with adequate capacity. Key considerations include latency, throughput, reliability, and network topology. Understand hybrid-cloud connectivity options and their pros and cons. Three ways to implement hybrid-cloud connectivity are Cloud VPN, Cloud Interconnect, and direct peering. Cloud VPN is a GCP service that provides virtual private networks between GCP and on-premises networks using the public Internet. The Cloud Interconnect service provides high throughput and highly available networking between GCP and an on-premises network using private network connections. Direct peering allows you to create a direct peering connection to Google Cloud edge.

Five types of load balancers and when to use them

The five types of load balancers are: Network TCP/UDP, Internal TCP/UDP, HTTP(S), SSL Proxy, and TCP Proxy. Choosing between these requires understanding if traffic will be distributed within a single region or across multiple regions, which protocols are used, and whether the traffic is internal or external to GCP.

Networking Quiz

1. Your team has deployed a VPC with default subnets in all regions. The lead network architect at your company is concerned about possible overlap in the use of private addresses. How would you explain how you are dealing with the potential problem?

- A. You inform the network architect that you are not using private addresses at all.
- B. When default subnets are created for a VPC, each region is assigned a different IP address range.
- C. You have increased the size of the subnet mask in the CIDR block specification of the set of IP addresses.
- D. You agree to assign new IP address ranges on all subnets.

Correct Answer: B

2. A data warehouse service running in GCP has all of its resources in a single project. The e-commerce application has resources in another project, including a database with transaction data that will be loaded into the data warehouse. The data warehousing team would like to read data directly from the database using extraction, transformation, and load processes that run on Compute Engine instances in the data warehouse project. Which of the following network constructs could help with this?

- A. Shared VPC
- B. Regional load balancing
- C. Direct peering
- D. Cloud VPN

Correct Answer: A

3. An intern working with your team has changed some firewall rules. Prior to the change, all Compute Engine instances on the network could connect to all other instances on the network. After the change, some nodes cannot reach other nodes. What might have been the change that causes this behavior?

- A. One or more implied rules were deleted.
- B. The default-allow-internal rule was deleted.
- C. The default-all-icmp rule was deleted.

D. The priority of a rule was set higher than 65535.

Correct Answer: B

4. The network administrator at your company has asked that you configure a firewall rule that will always take precedence over any other firewall rule. What priority would you assign?

- A. 0
- B. 1
- C. 65534
- D. 65535

Correct Answer: A

5. During a review of a GCP network configuration, a developer asks you to explain CIDR notation. Specifically, what does the 8 mean in the CIDR block 172.16.10.2/8?

- A. 8 is the number of bits used to specify a host address.
- B. 8 is the number of bits used to specify the subnet mask.
- C. 8 is the number of octets used to specify a host address.
- D. 8 is the number of octets used to specify the subnet mask.

Review Questions 125

Correct Answer: B

6. Several new firewall rules have been added to a VPC. Several users are reporting unusual problems with applications that did not occur before the firewall rule changes. You'd like to debug the firewall rules while causing the least impact on the network and doing so as quickly as possible. Which of the following options is best?

- A. Set all new firewall priorities to 0 so that they all take precedence over other rules.
- B. Set all new firewall priorities to 65535 so that all other rules take precedence over these rules.
- C. Disable one rule at a time to see whether that eliminates the problems. If needed, disable combinations of rules until the problems are eliminated.
- D. Remove all firewall rules and add them back one at a time until the problems occur and then remove the latest rule added back.

Correct Answer: C

7. An executive wants to understand what changes in the current cloud architecture are required to run compute-intensive machine learning workloads in the cloud and have the models run in production using on-premises servers. The models are updated daily. There is no network connectivity between the cloud and on-premises networks. What would you tell the executive?

- A. Implement additional firewall rules
- B. Use global load balancing
- C. Use hybrid-cloud networking
- D. Use regional load balancing

Correct Answer: C

8. To comply with regulations, you need to deploy a disaster recovery site that has the same design and configuration as your production environment. You want to implement the disaster recovery site in the cloud. Which topology would you use?

- A. Gated ingress topology
- B. Gated egress topology
- C. Handover topology
- D. Mirrored topology

Correct Answer: D

9. Network engineers have determined that the best option for linking the on-premises network to GCP resources is by using an IPsec VPN. Which GCP service would you use in the cloud?

- A. Cloud IPsec
- B. Cloud VPN
- C. Cloud Interconnect IPsec
- D. Cloud VPN IKE

Correct Answer: B

10. Network engineers have determined that a link between the on-premises network and GCP will require an 8 Gbps connection. Which option would you recommend?

- A. Cloud VPN
- B. Partner Interconnect
- C. Direct Interconnect
- D. Hybrid Interconnect

Correct Answer: B

11. Network engineers have determined that a link between the on-premises network and GCP will require a connection between 60 Gbps and 80 Gbps. Which hybrid-cloud networking services would best meet this requirement?

- A. Cloud VPN
- B. Cloud VPN and Direct Interconnect
- C. Direct Interconnect and Partner Interconnect
- D. Cloud VPN, Direct Interconnect, and Partner Interconnect

Correct Answer: C

12. The director of network engineering has determined that any links to networks outside of the company data center will be implemented at the level of BGP routing exchanges. What hybrid-cloud networking option should you use?

- A. Direct peering
- B. Indirect peering
- C. Global load balancing
- D. Cloud IKE

Correct Answer: A

13. A startup is designing a social site dedicated to discussing global political, social, and environmental issues. The site will include news and opinion pieces in text and video. The startup expects that some stories will be exceedingly popular, and others won't be, but they want to ensure that all users have a similar experience with regard to latency, so they plan to replicate content across regions. What load balancer should they use?

- A. HTTP(S)
- B. SSL Proxy
- C. Internal TCP/UDP
- D. TCP Proxy

Correct Answer: A

14. As a developer, you foresee the need to have a load balancer that can distribute load using only private RFC 1918 addresses. Which load balancer would you use?

- A. Internal TCP/UDP
- B. HTTP(S)
- C. SSL Proxy
- D. TCP Proxy

Correct Answer: A

15. After a thorough review of the options, a team of developers and network engineers have determined that the SSL Proxy load balancer is the best option for their needs. What other GCP service must they have to use the SSL Proxy load balancer?

- A. Cloud Storage
- B. Cloud VPN
- C. Premium Tier networking
- D. TCP Proxy Load Balancing

Correct Answer: C

16. True or false? In Google Cloud VPCs, subnets have regional scope.

A. True

B. False

Correct Answer: A

17. True or false: If you increase the size of a subnet in a custom VPC network, the IP addresses of virtual machines already on that subnet might be affected?

A. True

B. False

Correct Answer: B

18. Which of the following GCP network traffic types incur a charge?

A. Egress to YouTube, Maps or Drive

B. Ingress

C. Egress between zones in the same region

D. Egress to the same zone through internal IP addresses

Correct Answer: C

19. Which Network Service Tier provides a global SLA and allows for global load balancing and Cloud CDN?

A. Premium Tier

B. Standard Tier

Correct Answer: A

20. Which of the following can help break down GCP billing charges for analysis in BigQuery and visualization in Data Studio?

A. Network tags

B. Labels

Correct Answer: B

21. Sort the following steps for provisioning Shared VPC in Google Cloud Platform:

A Shared VPC Admin enables shared VPC for the host project.

An Organization Admin nominates a Shared VPC Admin.

A Shared VPC Admin delegates access to some or all subnets of a shared VPC network by granting the Network User role.

A Network User creates resources in his/her Service Project.

A. 2->1->3->4

B. 2->3->1->4

C. 4->1->3->2

D. 4->3->1->2

Correct Answer: A

22. In regards to VPC Network Peering, which of the following statements is correct?

A. Peered VPC networks do not remain administratively separate

B. Subnet IP ranges can overlap across peered VPC networks

C. Both sides of a peering association are set up in one single step

D. Transitive peering is not supported

Correct Answer: D

23. Which of the following approaches to multi-project networking, uses a centralized network administration model?

A. Shared VPC

B. Cloud VPN

C. VPC Network Peering

Correct Answer: A

24. Which of the following is not a GCP load balancing service?

A. SSL proxy load balancing

B. Hardware-defined load balancing

C. TCP proxy load balancing

D. Network load balancing

E. Internal load balancing

F. HTTP(S) load balancing

Correct Answer: B

25. Which three GCP load balancing services support IPv6 clients?

- A. HTTP(S) load balancing
- B. Network load balancing
- C. Internal load balancing
- D. TCP proxy load balancing
- E. SSL proxy load balancing

Correct Answer: A, D, E

26. Which GCP load balancing service supports GCP instance-to-instance RFC 1918 traffic?

- A. Internal load balancing
- B. TCP proxy load balancing
- C. Network load balancing
- D. HTTP(S) load balancing
- E. SSL proxy load balancing

Correct Answer: A

27. In Google Cloud Platform, a VPC network belongs to which of the following?

- A. Zone
- B. IP address range
- C. Project
- D. Region

Correct Answer: C

28. What are the three types of networks offered in Google Cloud Platform?

- A. Gigabit network, 10 gigabit network, and 100 gigabit networks
- B. Default network, auto network, and custom network

C. IPv4 unicast network, IPv4 multicast network, IPv6 network

D. Zonal, regional, and global

Correct Answer: B

29. Which GCP service translates requests for domain names into external IP addresses?

A. GCP routes

B. Compute Engine DNS

C. Alias IP Ranges

D. Cloud DNS

Correct Answer: D

30. What is the purpose of a Cloud Router and why does that matter?

A. It connects VPN on one side to Direct Peering on the other, which is faster than VPN alone.

B. A Cloud Router enables you to perform round-robin switching over multiple VPNs so that you can combine the bandwidth and get better throughput than is actually offered by the Internet.

C. It implements dynamic VPN that allows topology to be discovered and shared automatically, which reduces manual static route maintenance.

D. It is a hardware router, provided by Google but hosted in GCP.

Correct Answer: C

31. Which GCP Interconnect service requires a connection in a GCP colocation facility and provides 10 Gbps per link?

A. Cloud VPN

B. Dedicated Interconnect

C. Carrier Peering

D. Direct Peering

E. Partner Interconnect

Correct Answer: B

32. If you cannot meet Google's peering requirements, which network connection service should you choose to connect to G Suite and YouTube?

- A. Carrier Peering
- B. Direct Peering
- C. Dedicated Interconnect
- D. Partner Interconnect

Correct Answer: A

33. What are the three options for the target parameter of a firewall rule that defines the instances to which the firewall rule is intended to apply?

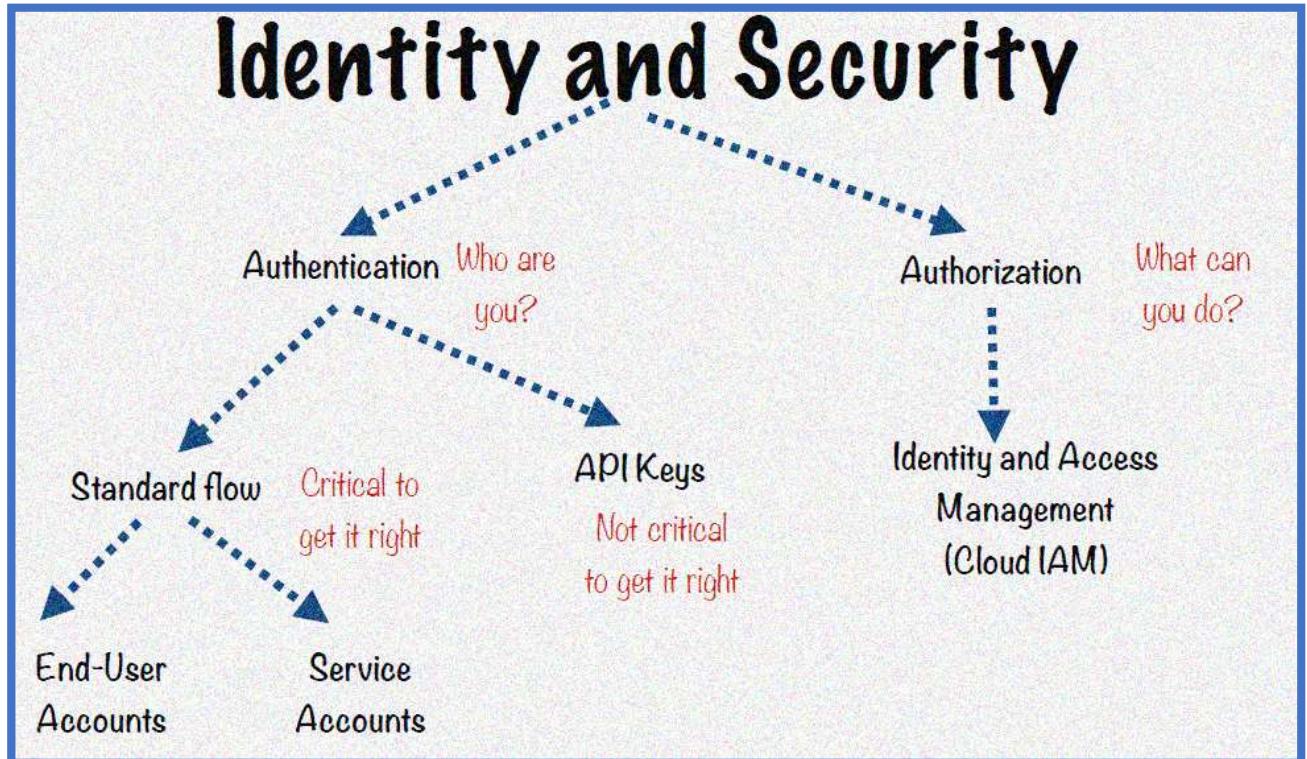
- A. Specified service accounts
- B. All instances in the network
- C. Specified IP ranges
- D. Specified target tags

Correct Answer: B, D

Identity & Security

Identity and Security

Identity & Security is divided into Authentication and Authorization



Authentication:

Authorization:

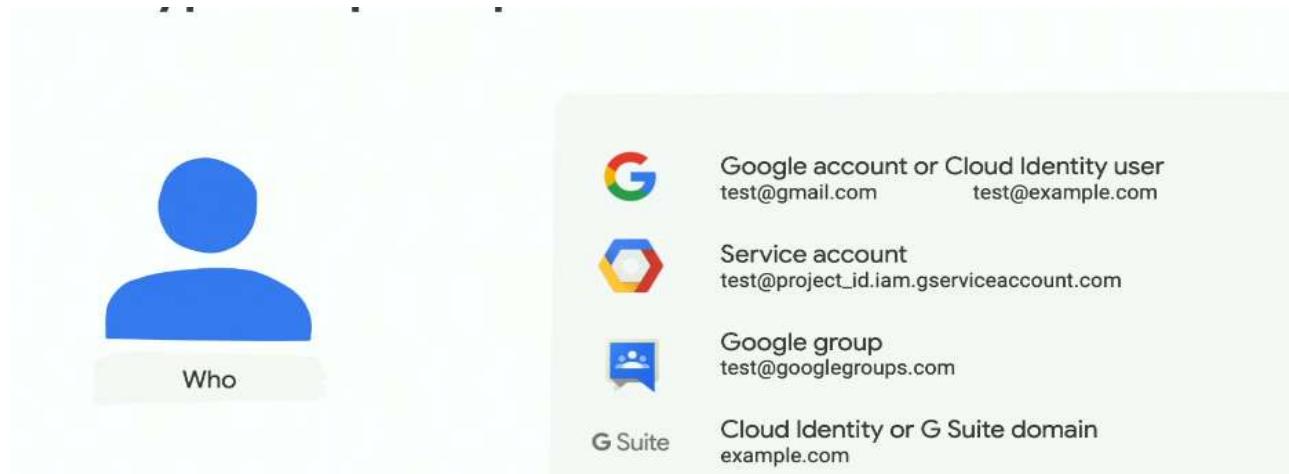
Cloud IAM

The Identity and Access Management (IAM) service is designed to allow you to specify what operations specific users can perform on particular resources.

This is also described as specifying “**Who Can do what On which resources**”

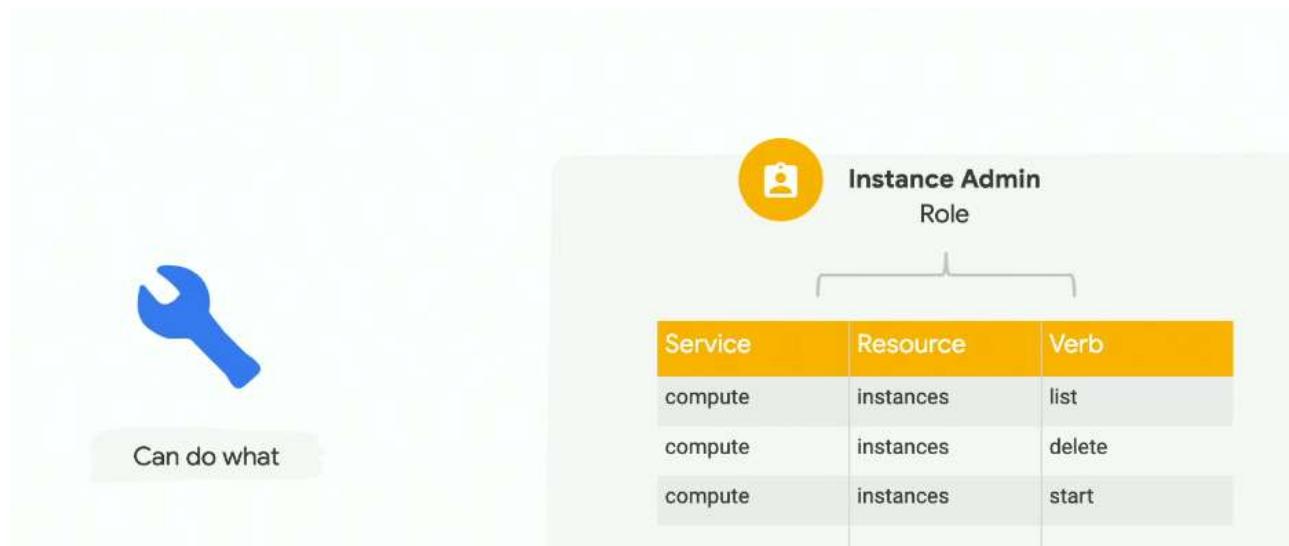
Who

IAM policies can apply to any of four types of principals



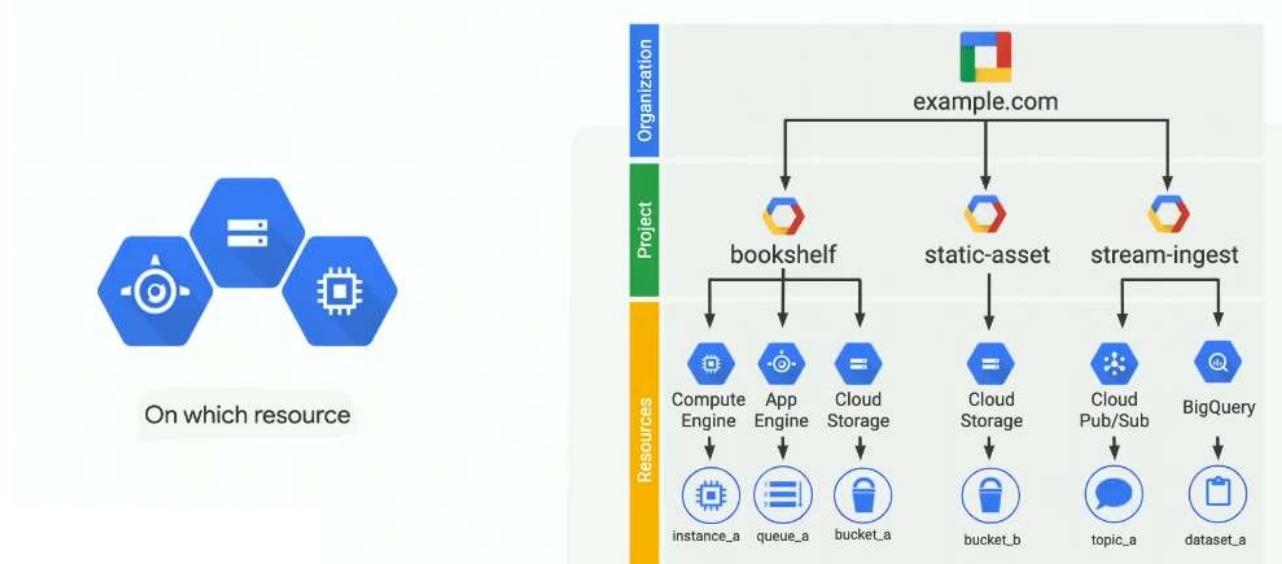
Can do what

IAM roles are collections of related permissions



On which resources

Users are organized roles on specific items



Benefits of IAM

- Easy to use
- Granular access control
- Integrated with all the GCP services
- Principle of least privilege
- Smart access control with ML Recommendation
- Chargeless Service in GCP

Elements of IAM

The primary elements of IAM are as follows: -

- Identities and groups
- Resources
- Permissions
- Roles
- Policies

Identities and groups

Identities and groups are entities that are used to grant access permissions to users.

Identities

An identity is an entity that represents a person or other agent that performs actions on a GCP resource. Identities are sometimes called members. There are several kinds of identities:

- Google account
- Service account
- Cloud Identity domain

Google accounts are used by people and represent people who interact with GCP, such as developers and administrators. These accounts are designated by an email address that is linked to a Google account. For example, `jane.doe@gmail.com` could be an identity in GCP. The domain of the email does not have to be `gmail.com`; it just has to be associated with a Google account.

Service accounts are used by applications running in GCP. These are used to give applications their own set of access controls instead of relying on using a person's account for permissions. Service accounts are also designated by email addresses. You can create multiple service accounts for an application, each with its own set of access control capabilities. When you create a service account, you specify a name of the account, such as `gcp-archexam`. IAM will then create an associated email such as `gcp-arch-exam@gcp-certs-1.iam.gserviceaccount.com`, where `gcp-certs-1` is the project ID of the project hosting the service account. Note that not all service accounts follow this pattern. When App Engine creates a service account, it uses the `appspot.gserviceaccount.com` domain.

Cloud Identity is an (IaaS) offering. Users who do not have Google accounts or G Suite accounts can use the Cloud Identity service to create an identity. It will not be linked to a Google account, but it will create an identity that can be used when assigning roles and permissions.

Groups

Related to identities are Google Groups, which are sets of Google accounts and service accounts. Groups have an associated email address. Groups are useful for assigning permissions to sets of users. When a user is added to a group, that user acquires the permissions granted to the group. Similarly, when a user is removed from the group, they no longer receive permissions from the group. Google Groups do not have login credentials, and therefore they cannot be used as an identity.

G Suite domains are another way to group identities. A G Suite domain is one that is linked to a G Suite account; that is, a G Suite account consists of users of a Google service account that bundles mail, Docs, Sheets, and so on for businesses and organizations. All users in the G Suite account are members of the associated group. Like Google Groups, G Suite domains can be used for specifying sets of users, but they are not identities.

Resources

Resources are entities that exist in the Google Cloud platform and can be accessed by users. Resources is a broad category that essentially includes anything that you can create in GCP. Resources include the following:

- Projects
- Virtual machines
- App Engine applications
- Cloud Storage buckets
- Pub/Sub topics

Google has defined a set of permissions associated with each kind of resource. Permissions vary according to the functionality of the resource.

Permissions

A permission is a grant to perform some action on a resource. Permissions vary by the type of resource with which they are associated.

Storage resources will have permissions related to creating, listing, and deleting data. For example, a user with the `bigrquery.datasets.create` permission can create tables in BigQuery. Cloud Pub/Sub has a permission called `pubsub.subscriptions.consume`, which allows users to read from a Cloud Pub/Sub topic.

Here are some examples of other permissions used by Compute Engine:

`compute.instances.get`
`compute.networks.use`
`compute.securityPolicies.list`

Here are some permissions used with Cloud Storage:

`resourcemanager.projects.get`
`resourcemanager.projects.list`
`storage.objects.create`

The permissions in IAM are fine-grained; that is, they grant permissions to do limited operations. Usually, for each Google Cloud API endpoint, there is a permission associated with it. There are API endpoints for almost every kind of action that you can take in GCP so that there is basically a one-to-one relationship between the things that you can do and the permissions to do them.

One of the reasons why it is not required to know specific permissions in detail is that GCP administrators do not have to work with them very often. Instead, they work with roles, which are collections of permissions.

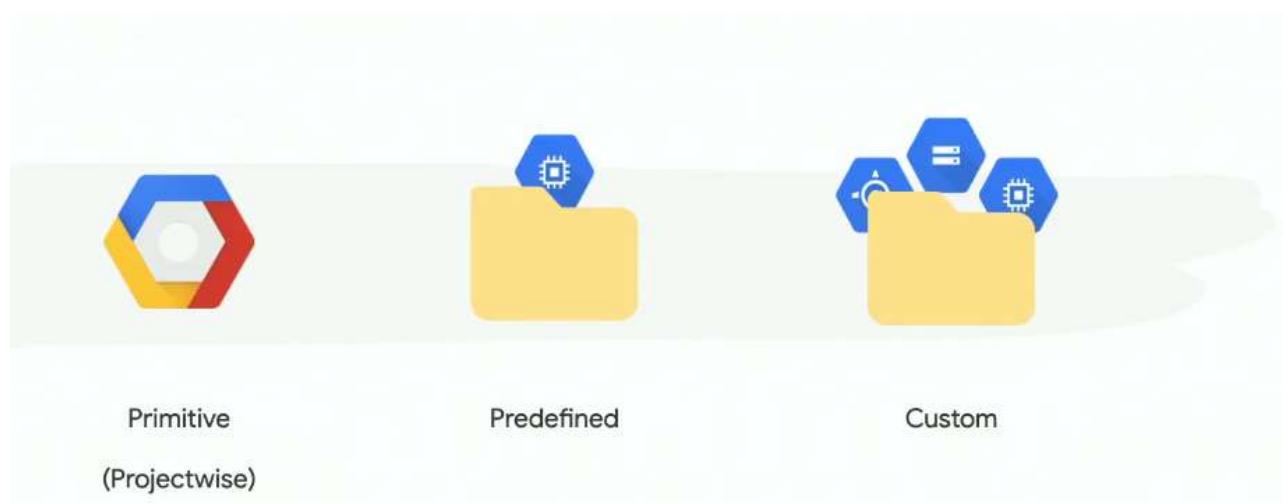
Roles

Roles are sets of permissions. One of the most important things to remember about IAM is that administrators grant roles to identities, not permissions. You cannot grant a permission directly to a user—you must grant it by assigning an identity a role.

Roles can be granted to identities. An identity can have multiple roles. Roles are granted for projects, folders, or organizations, and they apply to all resources under those. In other words, resources in those projects, folders, or organizations assume those roles when the role applies to the type of resource. For example, roles granted to a project that grants permissions to Compute Engine instances are applied to VM instances in that project.

There are three types of roles.

- Predefined
- Primitive
- Custom



Predefined roles

Predefined roles are created and managed by GCP. The roles are organized around groups of tasks commonly performed when working with IT systems, such as administering a server, querying a database, or saving a file to a storage system.

Roles have names such as the following:

- roles/bigquery.admin
- roles/bigquery.dataEditor
- roles/cloudfunction.developer
- roles/cloudsql.viewer



IAM Predefined roles offer fine grained permissions on particular services and give more granular access control

Primitive roles

Roles historically available in the Google Cloud Console. These roles are Owner, Editor, and Viewer.



The **Viewer** role grants a user read-only permission to a resource. A user with a Viewer role can see but not change the state of a resource.

The **Editor** role has all of the capabilities of the Viewer role and can also modify the state of the resource. For example, users with the Editor role on a storage system can write data to that system.

The **Owner** role includes the capabilities of the Editor role and can also manage roles and permissions for the resource to which it is assigned. For example, if the owner role is granted on a project to a user, then that user can manage roles and permissions for all resources in the project. The Owner role can also be assigned to specific resources, such as a Compute Engine instance or a Cloud Pub/Sub topic. In those cases, the permissions apply only to that specific resource.

Users with the Owner role can also set up billing.

Custom roles

- IAM Custom roles let you define a precise set of permissions.
- Create a customized role as per your need, used when Predefined roles does not meet the requirements.
- Custom role can be created with one or more permissions and then grant that custom role to users
- Custom roles are user defined and not maintained by Google
- When a custom role is created, an organization or project must be chosen to create it
- Create a custom role by combining one or more of the available Cloud IAM permissions

To assign a custom role to a user: -

- Create a custom Cloud IAM role with one or more permissions
- Grant that custom role to user

Policies

A policy is a set of statements that define a combination of users and the roles. This combination of users (or members as they are sometimes called) and a role is called a binding. Policies are specified using JSON.

From Google's IAM documentation, the role `roles/Storage_objectAdmin` is assigned to four identities, and the `roles/storage_objectViewer` is assigned to one identity—a user with the email `bob@example.com`:

```
{  
  "bindings": [  
    {  
      "role": "roles/storage.objectAdmin",  
      "members": [  
        "user:alice@example.com",  
        "serviceAccount:my-other-app@appspot.gserviceaccount.com",  
        "group:admins@example.com",  
        "domain:google.com" ]  
    },
```

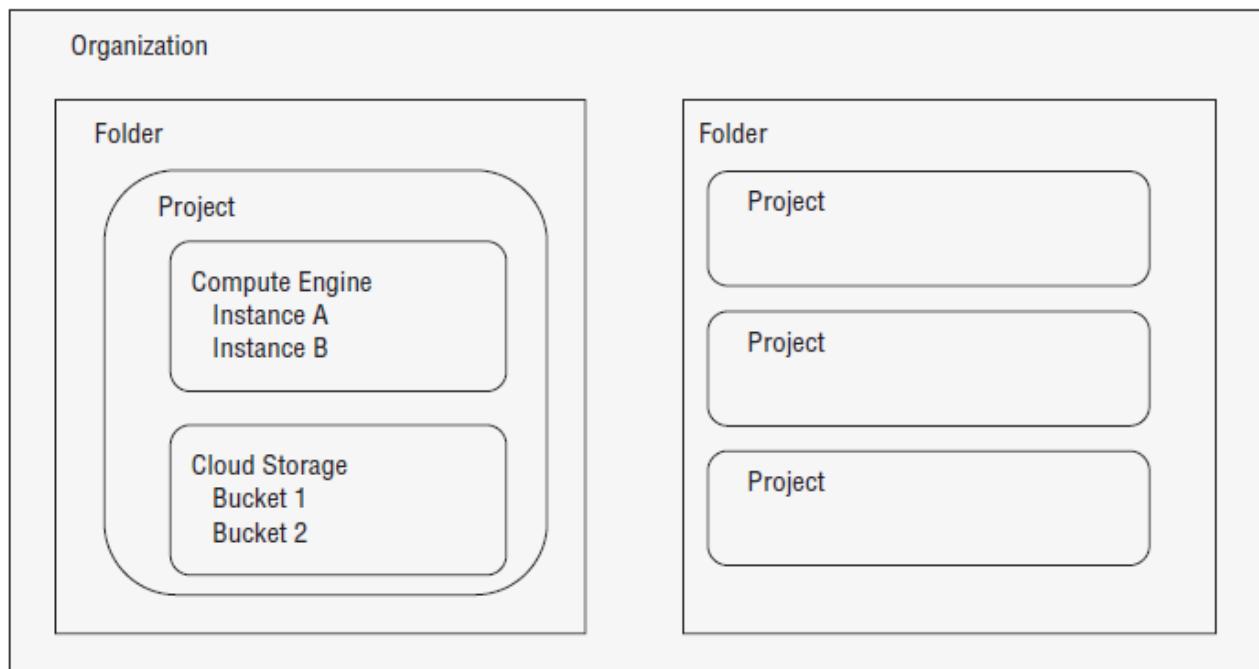
```
{  
  "role": "roles/storage.objectViewer",  
  "members": ["user:bob@example.com"]  
}  
]  
}
```

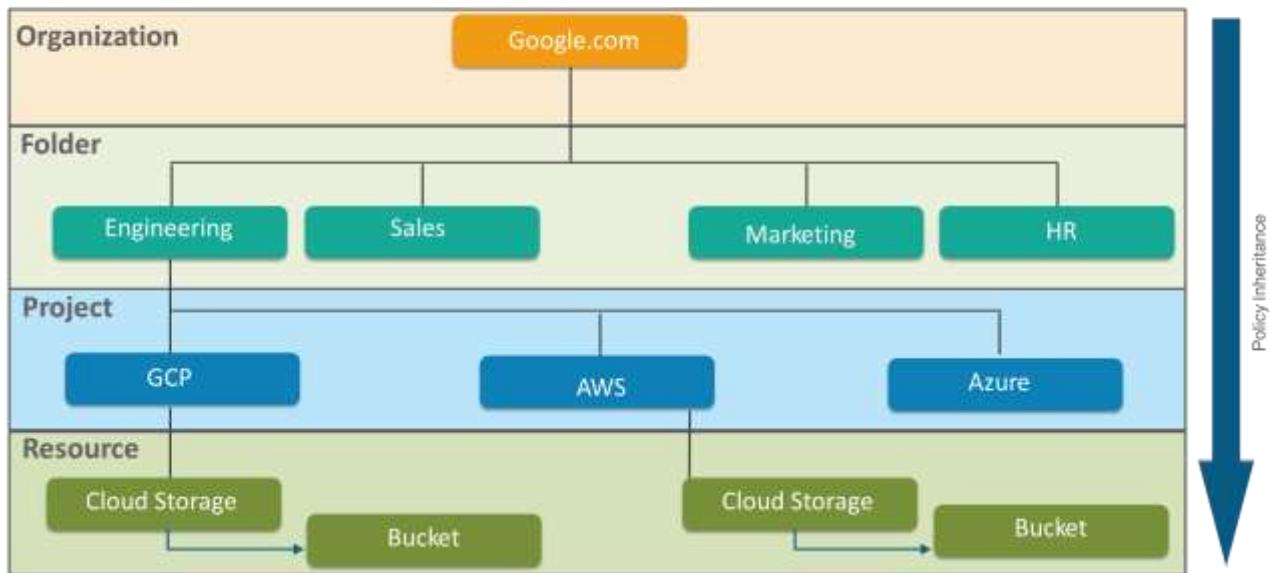
Policies can be managed using the Cloud IAM API, which provides three functions.

- `setIamPolicy` for setting policies on resources
- `getIamPolicy` for reading policies on resources
- `testIamPermissions` for testing whether an identity has a permission on a resource

Policies can be set anywhere in the resource hierarchy, such as at the organization, folder, or project level. They can also be set on individual resources. Policies set at the project level are inherited by all resources in the project, while policies set at the organization level are inherited by all resources in all folders and projects in the organization. If the resource hierarchy is changed, permissions granted by policies change accordingly.

Google Cloud Platform resource hierarchy





What are the best practices of IAM?

- Set maximum policies at organization and project level
- Always give the least amount of access to resources
- Grant role to google groups instead of individuals
- Be extra cautious while granting permission on service accounts
- Define organization policies
- Review logs regularly
- Enable multi factor authentication
- Remove unused, stale, or unnecessary IAM

What is the Authentication Mechanism in GCP?

The Authentication Mechanism in GCP are:

- End User Account
- Service Account
- API Keys

What is End User Account?

Use service accounts wherever possible, sometimes end-user authentication its unavoidable

You need to access resources on behalf of an end user of your application

For example, your application needs to access Google BigQuery datasets that belong to users of your application.

You need to authenticate as yourself (not as your application)

For example, because the Cloud Resource Manager API can create and manage projects owned by a specific user, you would need to authenticate as a user to create projects on their behalf.

What are Service Accounts? What are its types?

- A service account is a special account used by an application or resource, and not by an individual
- A service account is identified by its email address
- Service accounts do not have passwords
- They are associated with private/public RSA key pairs

The different types of Service Accounts are: -

- User Managed
- Google Managed

User Managed

A Service Account is created for every new project: -

- If Compute Engine API is enabled then it identified by:
PROJECT_NUMBER-compute@developer.gserviceaccount.com
- If App Engine application is there in the application, then its identified by:
PROJECT_ID@appspot.gserviceaccount.com

Google Managed

- Created and managed by Google
- Assigned to the new project automatically
- Gets access to the project automatically
- Represent different Google services
- Google uses these accounts to run process on your behalf
- Google managed Service account is identified: -
PROJECT_NUMBER@cloudservices.gserviceaccount.com

What are the best practices of using Service Account?

- Extra precautions should be taken when granting the serviceAccountUser Role to a user
- Always give least amount of access to service account
- Rotate user-managed service account keys
- Audit service accounts and keys

What are API Keys?

- API Keys are Simple encrypted string
- Can be used when calling certain APIs that don't need to access private user data
- Useful in clients such as browser and mobile applications that don't have a backend server
- The API key is used to track API requests associated with your project for quota and billing
- ML APIs are Natural language processing, Translation, Speech & Vision
- API Keys Creation

GCP Console => API Manager => Credentials => Create

Select “API Key”

IAM APIs - Organization Role

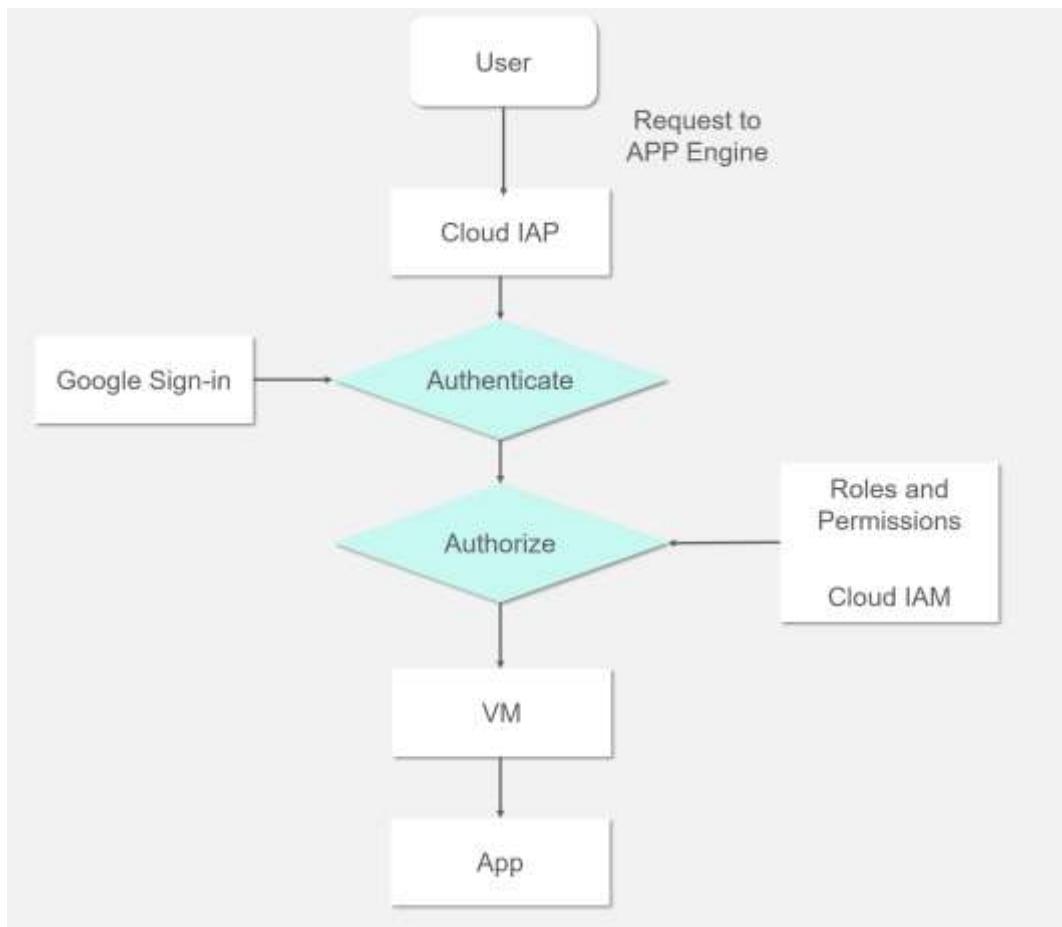
Method	REST API	Usage
create	POST /v1/{parent=organizations/*}/roles	Create new Role
delete	DELETE /v1/{name=organizations/*}/roles/*	Soft delete a role
get	GET /v1/{name=organizations/*}/roles/*	Get Role definition
list	GET /v1/{parent=organizations/*}/roles	Lists Roles defined on a resource
patch	PATCH /v1/{name=organizations/*}/roles/*	Update Role definition
undelete	POST /v1/{name=organizations/*}/roles/*:undelete	Undelete a Role - bring it back to previous state

IAM APIs - Project Role

Method	REST API	Usage
create	POST /v1/{parent=projects/*}/roles	Create new Role
delete	DELETE /v1/{name=projects/*}/roles/*	Soft delete a Role
get	GET /v1/{name=projects/*}/roles/*	Get Role definition
list	GET /v1/{parent=projects/*}/roles	Lists Roles defined on a resource
patch	PATCH /v1/{name=projects/*}/roles/*	Update Role definition
undelete	POST /v1/{name=projects/*}/roles/*:undelete	Undelete a Role - bring it back to previous state

What is Identity-Aware Proxy?

Cloud Identity-Aware Proxy: is a tool that helps control access, based on a user's identity and group membership, to applications running on Google Cloud Platform.



- IAP enables you to establish a centralized authentication layer for Apps
- IAP helps you to use Application-level Access control instead of firewalls and NW level
- IAP allows you to authenticate HTTP based apps outside of Google Cloud
- It can be used for App engine, compute engine, Kubernetes engine
- Using IAP users can access web applications from GCP securely without VPN

When to use Identity-Aware Proxy?

- Use when enforcing access control on applications and resources
- A group access can be set in such a way that the application will be accessible to employees instead of contractors

Data Security

How Security is designed in Google Cloud Technical Infrastructure?

Layer	Notable security measures (among others)
Operational security	Intrusion detection systems; techniques to reduce insider risk; employee U2F use; software development practices
Internet communication	Google Front End; designed-in Denial of Service protection
Storage services	Encryption at rest
User identity	Central identity service with support for U2F
Service deployment	Encryption of inter-service communication
Hardware infrastructure	Hardware design and provenance; secure boot stack; premises security

GCP provides multiple mechanisms for securing data in addition to IAM policies, which control access to data. Two essential services are encryption and key management.

Encryption

Encryption at rest

- Google encrypts data at rest by default. You do not have to configure any policy to enable this feature. This applies to all Google data storage services, such as Cloud Storage, Cloud SQL, and Cloud Bigtable. Encryption at rest actually occurs at multiple levels.
- Data is encrypted at multiple levels, including the application, infrastructure, and device levels
 - At the platform level, database and file data is protected using AES256 and AES128 encryption.
 - At the infrastructure level, data is grouped into data chunks in the storage system, and each chunk is encrypted using AES256 encryption.
 - At the hardware level, storage devices apply AES256 or AES128 encryption
- Data is encrypted in chunks. Each chunk has its own encryption key, which is called a data encryption key.
- Data encryption keys are themselves encrypted using a key encryption key

Encryption at transit

Encryption in transit, also called encryption in motion, is used to protect the confidentiality and integrity of data in the event that the data is intercepted in transit. GCP uses a combination of authenticating sources and encryption to protect data in transit.

Google distinguishes data in transit on the Google network and data in transit in the public Internet. Data within the boundaries of the Google network is authenticated but may not be encrypted. Data outside the physical boundaries of the Google network is encrypted.

Key Management

There are many data encryption and key encryption keys in use at any time in the Google Cloud.

Default Key Management

Google manages these keys by default for users. DEKs are stored near the data chunks that they encrypt. There is a separate DEK for each data chunk, but one KEK can be used to encrypt multiple DEKs. The KEKs are stored in a centralized key management service. The DEKs are generated by the storage service that is storing the data chunk using a common cryptographic library. The DEKs are then sent to the centralized key management service, where they are themselves encrypted using the storage system's KEK. When the storage system needs to retrieve data, it sends the DEK to the key management service,

where the calling service is authenticated, and the DEK is decrypted and sent back to the storage system.

Cloud KMS Key Management

Cloud KMS is a hosted key management service in Google Cloud. It enables customers to generate and store keys in GCP. It is used when customers want control over key management but do not need keys to reside on their own key management infrastructure.

Cloud KMS supports a variety of cryptographic keys, including AES256, RSA 2048, RSA 3072, RSA 4096, EC P256, and EC P384. It also provides functionality for automatically rotating keys and encrypting DEKs with KEKs. Cloud KMS keys can be destroyed, but there is a 24-hour delay before the key is destroyed in case someone accidentally deletes a key or in the event of a malicious act.

Cloud KMS keys can be used for application-level encryption in GCP services, including Compute Engine, BigQuery, Cloud Storage, and Cloud Dataproc.

Customer-Supplied Keys

A third alternative for key management is customer-supplied keys. Customer-supplied keys are used when an organization needs complete control over key management, including storage. In this model, keys are generated and kept on-premises and used by GCP services to encrypt the customer's data. These keys are passed with other arguments to API function calls. When the keys are sent to GCP, they are stored in memory while being used. Customer-supplied keys are not written to persistent storage.

Encryption and key management are essential components of a comprehensive security regime. Data at rest and in transit are encrypted by default. Keys are managed by default by GCP but can be managed by cloud users. They have two options: Cloud KMS, which is a hosted managed key service that generates and stores keys in the cloud on behalf of a user; the other option is customer-supplied keys, which are managed on-premises and sent to Google as part of API calls. Customer-supplied keys allow customers the greatest amount of control but also require infrastructure and management procedures that are not needed when using default encryption.

Security Evaluation

Cloud users can expend significant time and resources configuring and managing identity management services, access controls, and encryption key management. Without a formal evaluation process, however, they are in the dark about how well these measures protect their systems. Two ways to evaluate the extent of the protection provided by the combination of security measures in place are penetration testing and auditing.

Penetration Testing

Penetration testing is the process of simulating an attack on an information system in order to gain insights into potential vulnerabilities. Penetration tests are authorized by system owners. In some cases, penetration testers know something about the structure of the network, servers, and applications being tested. In other cases, testers start without detailed knowledge of the system that they are probing.

Penetration testing occurs in these five phases:

1. Reconnaissance is the phase at which penetration testers gather information about the target system and the people who operate it or have access to it. This could include phishing attacks that lure a user into disclosing their login credentials or details of software running on their network equipment and servers.
2. Scanning is the automated process of probing ports and checking for known and unpatched vulnerabilities.
3. Gaining access is the phase at which the attackers exploit the information gathered in the first two phases to access the target system.
4. In the maintaining access phase, attackers will do things to hide their presence, such as manipulating logs or preventing attacking processes from appearing in a list of processes running on a server.
5. Removing footprints, the final phase, involves eliminating indications that the attackers have been in the system. This can entail manipulating audit logs and deleting data and code used in the attack.

During a penetration test, testers will document how they gathered and exploited information, what if any vulnerabilities they exploited, and how they removed indications that they were in the system.

Auditing

Auditing is basically reviewing what has happened on your system. In the case of Google Cloud, there are a number of sources of logging information that can provide background details on what events occurred on your system and who executed those actions.

Your applications should generate logs that identify significant events, especially security-related events. For example, if a new user is granted administrator rights to an application, that should be logged. The Stackdriver Logging Agent will collect logs for widely used services, including syslog, Jenkins, Memcached, MySQL, PostgreSQL, Redis, and Zookeeper.

For a full list of logs collected, see <https://cloud.google.com/logging/docs/agent/default-logs>.

Managed services, like Compute Engine, Cloud SQL, and App Engine, log information to Stackdriver logs.

Google Cloud Certified Professional Cloud Architect Definitive Guide

Cloud Audit Logs is a GCP service that records administrative actions and data operations. Administrative actions that modify configurations or metadata of resources is always logged by Cloud Audit Logs. Data access logs record information when data is created, modified, or read. Data access logs can generate large volumes of data so that it can be configured to collect information for select GCP services.

The logs are saved for a limited period of time. Often, regulations require that audit logs be retained for longer periods of time. Plan to export audit logs from Cloud Audit Logs and save them to Cloud Storage or BigQuery. They can also be written to Cloud Pub/Sub.

Logs are exported from Stackdriver, which supports the following three export methods:

- JSON files to Cloud Storage
- Logging tables to BigQuery datasets
- JSON messages to Cloud Pub/Sub

You can use lifecycle management policies in Cloud Storage to move logs to different storage tiers, such as Nearline and Coldline storage, or delete them when they reach a specified age.

Penetration testing and logging are two recommended practices for keeping your systems secure.

Security Design Principles

As a cloud architect, you will be expected to know security design principles such as separation of duties, least privileges, and defense in depth.

Security – Interview | Exam Tips

IAM

IAM can be described as “Who Can do what On which resources”

Components of IAM

The Key components of the IAM service include identities and groups, resources, permissions, roles, and policies.

Identities

Identities can be a Google account, a service account, or a Cloud Identity account. Identities can be collected into Google Groups or G Suite groups.

Groups

Groups are useful for assigning permissions to sets of users. When a user is added to a group, that user acquires the permissions granted to the group. Similarly, when a user is removed from the group, they no longer receive permissions from the group.

Resources

Resources are entities that exist in the Google Cloud platform and can be accessed by users.

Permissions

A permission is a grant to perform some action on a resource. Permissions vary by the type of resource with which they are associated.

Roles

- Primitive roles existed prior to Cloud IAM and include Owner, Editor, and Viewer roles.
- Predefined roles are generally associated with a GCP service, such as App Engine or BigQuery, and a set of related activities, such as editing data in a database or deploying an application to App Engine. Predefined roles are preferred over primitive roles in most situations.
- With custom roles, you can assign one or more permissions to a role and then assign that role to a user, group, or service account. Custom roles are especially important when implementing the principle of least privilege, which states that users should be granted the minimal set of permissions needed for them to perform their jobs.

Policies

Policies are used to associate a set of roles and permissions with a resource.

Resource hierarchy

Organizations are at the top of the hierarchy. Organizations contain folders and projects. Folders can contain other folders as well as projects. Access controls assigned to entities in the hierarchy are inherited by entities lower in the hierarchy. Access controls assigned to an entity do not affect entities higher in the hierarchy.

Resources are entities in GCP that can be accessed by a user

Access is controlled by IAM. Resources is a broad category that essentially includes anything that you can create in GCP including projects, virtual machines, storage buckets, and Cloud Pub/Sub topics. Permissions vary by type of resource. Cloud Pub/Sub, for example, has permissions to writing messages to topics and creating subscriptions. Those permissions would not make sense for other types of resources. Some role patterns are used across entity types, such as admin and viewer.

Roles are sets of permissions

Remember that IAM permissions are granted to roles and roles are granted to identities. You cannot grant a permission directly to an identity. Google has created predefined roles that map to common organizational roles, such as administrators, viewers, and deployers. Predefined roles have all of the permissions someone in that organizational role typically needs to perform their duties. Custom roles can also be created if the predefined roles do not fit your needs.

Primitive roles should be used in limited situations

Primitive roles are the owner, editor, and viewer. These roles existed prior to IAM and grant course-grained permissions to identities. Primitive roles should be used only in cases where users need broad access, such as developers in a development environment. In general, you should favor predefined roles over primitive roles or custom roles.

Structure and function of policies

A policy consists of binding, metadata, and an audit configuration. Bindings specify how access is granted to a resource. Bindings are made up of members, roles, and conditions. The metadata of a policy includes an attribute called etag and versions. Audit configurations describe which permission types are logged and which identities are exempt from logging. Policies can be defined at different levels of the resource hierarchy, including organizations, folders, projects, and individual resources. Only one policy at a time can be assigned to an organization, folder, project, or individual resource.

Policies are used to associate a set of roles and permissions with resources

A policy is a set of statements that define a combination of users and roles. This combination of users and a role is called a binding. Policies are specified using JSON. Policies are used in addition to IAM identity-based access controls to limit access to resources.

Service Accounts

Purpose of service accounts

Service accounts are a type of identity that are used with VM instances and applications, which are able to make API calls authorized by roles assigned to the service account. A service account is identified by a unique email address. These accounts are authenticated by two sets of public/private keys. One set is managed by Google, and the other set is managed by users. Public keys are provided to API calls to authenticate the service account.

Encryption

Encryption is used to protect data in transit and at rest. Google Cloud encrypts data at rest by default. Google Cloud can manage keys, or customers can manage their own keys.

Google encrypts data at rest by default

Data is encrypted at multiple levels. At the platform level, database and file data is protected using AES256 and AES128 encryption. At the infrastructure level, data is grouped into data chunks in the storage system, and each chunk is encrypted using AES256 encryption. At the hardware level, storage devices apply AES256 or AES128 encryption.

Data at rest is encrypted with a data encryption key (DEK)

The DEK is encrypted with a KEK. Data is encrypted in chunks, and the DEK is kept near the data that it encrypts. The service writing the data has a KEK, which is used to encrypt the DEK. Google manages rotating KEKs.

Understand how Google encrypts data in transit

Google distinguishes data in transit on the Google network and data in transit in the public Internet. Data within the boundaries of the Google network is authenticated but may not be encrypted. Data outside the physical boundaries of the Google network is encrypted.

Understand data-at-rest encryption

Encryption is the process of encoding data in a way that yields a coded version of data that cannot be practically converted back to the original form without additional information. Data at rest is encrypted by default on Google Cloud Platform. Data is encrypted at multiple levels, including the application, infrastructure, and device levels. Data is encrypted in chunks. Each chunk has its own encryption key, which is called a data encryption key. Data encryption keys are themselves encrypted using a key encryption key.

Understand data-in-transit encryption

All traffic to Google Cloud services is encrypted by default. Google Cloud and the client negotiate how to encrypt data using either Transport Layer Security (TLS) or the Google-developed protocol QUICC.

Key management

Key management

Cloud KMS is a hosted key management service in the Google Cloud. It enables customers to generate and store keys in GCP. It is used when customers want control over key management. Customer-supplied keys are used when an organization needs complete control over key management, including storage.

Three types of key management

Google provides default key management in which Google generates, stores, and manages keys. With the Cloud KMS Key Management service, customers manage the generation, rotation, and destruction of keys, but the KMS service stores the keys in the cloud. Customer-supplied keys are fully managed and stored on-premises by customers.

Security

Security best practices, including separation of duties, least privilege, and defense in depth

- Separation of duties is the practice of limiting the responsibilities of a single individual in order to prevent the person from successfully acting alone in a way detrimental to the organization. Least privilege is the practice of granting only the minimal set of permissions needed to perform a duty.
- Defense in depth is the practice of using more than one security control to protect resources and data.

Usage of security controls to comply with regulations

Governments and industry organizations have developed rules and regulations to protect the privacy of individuals, ensure the integrity of business information, and make sure that a baseline level of security is practiced by organizations using information technology. Architects should understand the broad requirements of these regulations. Regulations often have common requirements around confidentiality, integrity, and availability.

Role of penetration testing and auditing

Both are forms of security evaluation. The goal of penetration testing is to find vulnerabilities in services by simulating an attack by malicious actors. You do not have to notify Google when you perform penetration testing. The purpose of auditing is to ensure that security controls are in place and functioning as expected.

Basic requirements of major regulations.

The Health Insurance Portability and Accountability Act (HIPAA) is a federal law in the United States that protects individuals' healthcare information. The Children's Online Privacy Protection Act (COPPA) is primarily focused on children under the age of 13, and it applies to websites and online services that collect information about children. The Federal Risk and Authorization Management Program (FedRAMP) is a U.S. federal government program that promotes a standard approach to assessment, authorization, and monitoring of cloud resources. The European Union's (EU) General Data Protection Regulation (GDPR) is designed to standardize privacy protections across the EU, grant controls to individuals over their private information, and specify security practices required for organizations holding private information of EU citizens.

Security Quiz

1. A company is migrating an enterprise application to Google Cloud. When running on-premises, application administrators created user accounts that were used to run background jobs. There was no actual user associated with the account, but the administrators needed an identity with which to associate permissions. What kind of identity would you recommend using when running that application in GCP?

- A. Google-associated account
- B. Cloud Identity account
- C. Service account**
- D. Batch account

Correct Answer: C

2. You are tasked with managing the roles and privileges granted to groups of developers, quality assurance testers, and site reliability engineers. Individuals frequently move between groups. Each group requires a different set of permissions. What is the best way to grant access to resources that each group needs?

- A. Create a group in Google Groups for each of the three groups: developers, quality assurance testers, and site reliability engineers. Add the identities of each user to their respective group. Assign predefined roles to each group.**
- B. Create a group in Google Groups for each of the three groups: developers, quality assurance testers, and site reliability engineers. Assign permissions to each user and then add the identities to their respective group.
- C. Assign each user a Cloud Identity, and grant permissions directly to those identities.
- D. Create a G Suite group for each of the three groups: developers, quality assurance testers, and site reliability engineers. Assign permissions to each user and then add the identities to their respective group.

Correct Answer: A

3. You are making a presentation on Google Cloud security to a team of managers in your company. Someone mentions that to comply with regulations, the organization will have to follow several security best practices, including least privilege. They would like to know how GCP supports using least privilege. What would you say?

- A. GCP provides a set of three broad roles: owner, editor, and viewer. Most users will be assigned viewer unless they need to change configurations, in which case they will receive the editor role, or if they need to perform administrative functions, in which case they will be assigned owner.
- B. GCP provides a set of fine-grained permissions and predefined roles that are assigned those permissions. The roles are based on commonly grouped responsibilities. Users will be assigned only the predefined roles needed for them to perform their duties.**
- C. GCP provides several types of identities. Users will be assigned a type of identity most suitable for their role in the organization.
- D. GCP provides a set of fine-grained permissions and custom roles that are created and managed by cloud users. Users will be assigned a custom role designed specifically for that user's responsibilities.

Correct Answer: B

4. An online application consists of a front-end service, a back-end business logic service, and a relational database. The front-end service is stateless and runs in an instance group that scales between two and five servers. The back-end business logic runs in a Kubernetes Engine cluster. The database is implemented using Cloud SQL PostgreSQL. How many trust domains should be used for this application?

- A. 1
- B. 2
- C. 3**
- D. None. These services do not need trust domains

Correct Answer: C

5. In the interest of separating duties, one member of your team will have permission to perform all actions on logs. You will also rotate the duty every 90 days. How would you grant the necessary permissions?

- A. Create a Google Group, assign roles/logging.admin to the group, add the identity of the person who is administering the logs at the start of the 90-day period, and remove the identity of the person who administered logs during the previous 90 days.
- B. Assign roles/logging.admin to the identity of the person who is administering the logs at the start of the 90-day period, and revoke the role from the identity of the person who administered logs during the previous 90 days.
- C. Create a Google Group, assign roles/logging.privateLogViewer to the group, add the identity of the person who is administering the logs at the start of the 90-day period, and remove the identity of the person who administered logs during the previous 90 days.
- D. Assign roles/logging.privateLogViewer to the identity of the person who is administering the logs at the start of the 90-day period, and revoke the role from the identity of the person who administered logs during the previous 90 days.

Correct Answer: A

6. Your company is subject to several government and industry regulations that require all personal healthcare data to be encrypted when persistently stored. What must you do to ensure that applications processing protected data encrypts it when it is stored on disk or SSD?

- A. Configure a database to use database encryption.
- B. Configure persistent disks to use disk encryption.
- C. Configure the application to use application encryption.
- D. Nothing. Data is encrypted at rest by default.**

Correct Answer: D

7. Data can be encrypted at multiple levels, such as at the platform, infrastructure, and device levels. Data may be encrypted multiple times before it is written to persistent storage. At device level, how is data encrypted in GCP?

- A. AES256 or AES128 encryption**
- B. Elliptic curve cryptography

- C. Data Encryption Standard (DES)
- D. Blowfish

Correct Answer: A

8. In GCP, each data chunk written to a storage system is encrypted with a data encryption key. The key is kept close to the data that it encrypts to ensure low latency when retrieving the key. How does GCP protect the data encryption key so that an attacker who gained access to the storage system storing the key could not use it to decrypt the data chunk?

- A. Writes the data encryption key to a hidden location on disk
- B. Encrypts the data encryption key with a key encryption key**
- C. Stores the data encryption key in a secure Cloud SQL database
- D. Applies an elliptic curve encryption algorithm for each data encryption key

Correct Answer: B

9. Data can be encrypted at different layers of the OSI network stack. Google Cloud may encrypt network data at multiple levels. What protocol is used at layer 7?

- A. IPSec
- B. TLS
- C. ALTS**
- D. ARP

Correct Answer: C

10. After reviewing security requirements with compliance specialists at your company, you determine that your company will need to manage its own encryption keys. Keys may be stored in the cloud. What GCP service would you recommend for storing keys?

- A. Cloud Datastore
- B. Cloud Firestore
- C. Cloud KMS**
- D. Bigtable

Correct Answer: C

11. The finance department of your company has notified you that logs generated by any finance application will need to be stored for five years. It is not likely to be accessed, but it has to be available if needed. If it were needed, you would have up to three days to retrieve the data. How would you recommend storing that data?

- A. Keep it in Stackdriver Logging.
- B. Export it to Cloud Storage and store it in Coldline class storage.**
- C. Export it to BigQuery and partition it by year.
- D. Export it to Cloud Pub/Sub using a different topic for each year.

Correct Answer: B

12. The legal department in your company notified software development teams that if a developer can deploy to production, then that developer cannot be allowed to perform the final code review before deploying to production. This is an example of which security best practice?

- A. Defense in depth
- B. Separation of duties**
- C. Least privilege
- D. Encryption at rest

Correct Answer: B

13. A startup has hired you to advise on security and compliance related to their new online game for children ages 10 to 14. Players will register to play the game, which includes collecting the name, age, and address of the player. Initially, the company will target customers in the United States. With which regulation would you advise them to comply?

- A. HIPAA/HITECH
- B. SOX
- C. COPPA**
- D. GDPR

Correct Answer: C

14. The company for which you work is expanding from North America to set up operations in Europe, starting with Germany and the Netherlands. The company offers online services that collect data on users. With what regulation must your company comply?

- A. HIPAA/HITECH
- B. SOX
- C. COPPA
- D. GDPR**

Correct Answer: D

15. Enterprise Self-Storage Systems is a company that recently acquired a startup software company that provides applications for small and midsize self-storage companies. The company is concerned that the business strategy of the acquiring company is not aligned with the software development plans of the software development teams of the acquired company. What IT framework would you recommend the company follow to better align business strategy with software development?

- A. ITIL**
- B. TOGAF
- C. Porters Five Forces Model
- D. Ansoff Matrix

Correct Answer: A

16. Which Cloud IAM role contains permissions to create, modify, and delete networking resources, except for firewall rules and SSL certificates?

- A. Network Viewer
- B. Network Admin**
- C. Security Admin

Correct Answer: B

17. Which type of IAM member belongs to an application or virtual machine instead of an individual end user?

- A. Google Group
- B. Cloud Identity domain
- C. Service Account**
- D. Google Account

Correct Answer: C

18. When would you choose to have an organization node? (Choose all that are correct. Choose 2 responses.)

- A. When you want to create folders.**
- B. When you want to organize resources into projects.
- C. When you want to apply organization-wide policies centrally.**
- D. There is no choice; organization nodes are mandatory.

Correct Answer: A, C

19. Order these IAM role types from broadest to finest-grained.

- A. Primitive roles, predefined roles, custom roles**
- B. Custom roles, predefined roles, primitive roles
- C. Predefined roles, custom roles, primitive roles

Correct Answer: A

20. Can IAM policies that are implemented higher in the resource hierarchy take away access that is granted by lower-level policies?

- A. Yes
- B. No**

Correct Answer: B

21. IAM access control policy can be set at which level?

- A. Organization level
- B. Project level
- C. Resource level
- D. All of the above**

Correct Answer: D

22. IAM Role can be assigned to whom?

- A. User
- B. Group
- C. Service Account
- D. All of the above**

Correct Answer: D

23. In IAM, if a policy gives you Owner permissions at the project level, your access to an individual resource in the project may be restricted to View by applying a more restrictive policy to that resource.

- A. True**
- B. False**

Correct Answer: B

24. You have been tasked with creating a pilot project in GCP to demonstrate the feasibility of migrating workloads from an on-premises Hadoop cluster to Cloud Dataproc. Three other engineers will work with you. None of the data that you will use contains sensitive information. You want to minimize the amount of time that you spend on administering the development environment. What would you use to control access to resources in the development environment?

- A. Predefined roles
- B. Custom roles
- C. Primitive roles**

D. Access control lists

Correct Answer: C

25. The auditors for your company have determined that several employees have more permissions than needed to carry out their job responsibilities. All the employees have users accounts on GCP that have been assigned predefined roles. You have concluded that the optimal way to meet the auditors' recommendations is by using custom roles. What permission is needed to create a custom role?

- A. iam.roles.create
- B. iam.custom.roles
- C. roles/iam.custom.create
- D. roles/iam.create.custom

Correct Answer: A

26. You have created a managed instance group in Compute Engine to run a high-performance computing application. The application will read source data from a Cloud Storage bucket and write results to another bucket. The application will run whenever new data is uploaded to Cloud Storage via a Cloud Function that invokes the script to start the job. You need to assign the role roles/storage.objectCreator to an identity so that the application can write the output data to Cloud Storage. To what kind of identity would you assign the roles?

- A. User.
- B. Group.
- C. Service account.**
- D. You wouldn't. The role would be assigned to the bucket.

Correct Answer: C

27. Your company has implemented an organizational hierarchy consisting of two layers of folders and tens of projects. The top layer of folders corresponds to a department, and the second layer of folders are working groups within a department. Each working group has one or more projects in the resource hierarchy. You have to ensure that all projects comply with regulations, so you have created several policies. Policy A applies to all departments. Policies B, C, D, and E are department specific. At what level of the resource hierarchy would you assign each policy?

- A. Assign policies A, B, C, D, and E to each folder
- B. Assign policy A to the organizational hierarchy and policies B, C, D, and E to each department's corresponding folder**
- C. Assign policy A to the organizational hierarchy and policies B, C, D, and E to each department's corresponding projects
- D. Assign policy A to each department's folder and policies B, C, D, and E to each project

Correct Answer: B

28. Your startup is developing a mobile app that takes an image as input and produces a list of names of objects in the image. The image file is uploaded from the mobile device to a Cloud Storage bucket. A service account is associated with the server-side application that will retrieve the image. The application will not perform any other operation on the file or the bucket. Following the principle of least privilege, what role would you assign to the service account?

- A. roles/storage.objectViewer**
- B. roles/storage.objectAdmin
- C. roles/storage.objectCreator
- D. roles/storage.objectViewer and roles/storage.objectCreator

Correct Answer: A

29. A data analyst asks for your help on a problem that users are having that involves BigQuery. The data analyst has been granted permissions to read the tables in a particular dataset. However, when the analyst runs a query, an error message is returned. What role would you think is missing from the users' assigned roles?

- A. roles/BigQuery.admin
- B. roles/BigQuery.jobUser**
- C. roles/BigQuery.metadataViewer
- D. roles/BigQuery.queryRunner

Correct Answer: B

30. Your company is subject to financial industry regulations that require all customer data to be encrypted when persistently stored. Your CTO has tasked you with assessing options for encrypting the data. What must you do to ensure that applications processing protected data encrypt it when it is stored on disk or SSD?

- A. Configure a database to use database encryption.
- B. Configure persistent disks to use disk encryption.
- C. Configure the application to use application encryption.
- D. Nothing. Data is encrypted at rest by default.**

Correct Answer: D

31. Data can be encrypted at multiple levels, such as at the platform, infrastructure, and device levels. At the device level, how is data encrypted in the Google Cloud Platform?

- A. AES256 or AES128 encryption**
- B. Elliptic curve cryptography
- C. Data Encryption Standard (DES)
- D. Blowfish

Correct Answer: A

32. In GCP, each data chunk written to a storage system is encrypted with a data encryption key. How does GCP protect the data encryption key so that an attacker who gained access to the storage system storing the key could not use it to decrypt the data chunk?

- A. GCP writes the data encryption key to a hidden location on disk.
- B. GCP encrypts the data encryption key with a key encryption key.**
- C. GCP stores the data encryption key in a secure Cloud SQL database.
- D. GCP applies an elliptic curve encryption algorithm for each data encryption key.

Correct Answer: B

33. The CTO has asked you to participate in a prototype project to provide better privacy controls. The CTO asks you to run a risk analysis job on a text file that has been inspected by the Data Loss Prevention API. What is the CTO interested in knowing?

- A. The number of times sensitive information is redacted
- B. The percentage of text that is redacted
- C. The likelihood that the data can be re-identified**
- D. What InfoType patterns were detected

Correct Answer: C

34. Your company is about to start a huge project to analyze a large number of documents to redact sensitive information. You would like to follow Google-recommended best practices. What would you do first?

- A. Identify InfoTypes to use
- B. Prioritize the order of scanning, starting with the most at-risk data**
- C. Run a risk analysis job first
- D. Extract a sample of data and apply all InfoTypes to it

Correct Answer: B

35. Your startup is creating an app to help students with math homework. The app will track assignments, how long the student takes to answer a question, the number of incorrect answers, and so on. The app will be used by students ages 9 to 14. You expect to market the app in the United States. With which of the following regulations must you comply?

- A. HIPAA
- B. GDPR
- C. COPPA**
- D. FedRAMP

Correct Answer: C

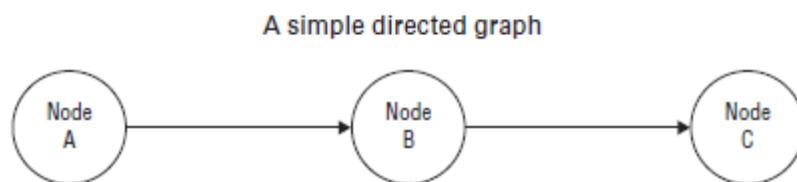
Bigdata Services

Data pipelines are sequences of operations that copy, transform, load, and analyze data. GCP services like Cloud Dataflow, Cloud Dataproc, Cloud Pub/Sub, and Cloud Composer are used to implement data pipelines.

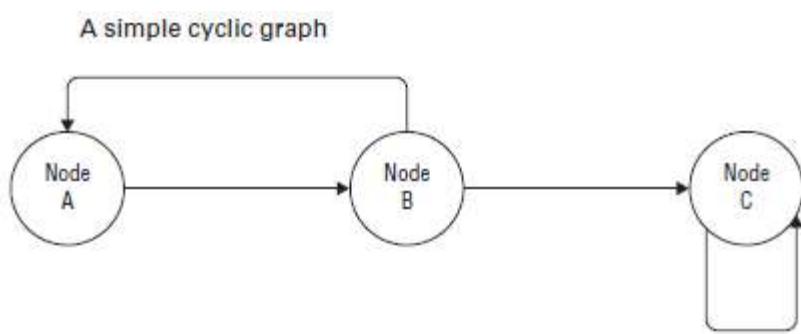
Overview of Data Pipelines

What is Data Pipeline?

A data pipeline is an abstract concept that captures the idea that data flows from one stage of processing to another. Data pipelines are modelled as directed acyclic graphs (DAGs). A graph is a set of nodes linked by edges. A directed graph has edges that flow from one node to another. Figure 3.1 shows a simple three-node graph with directed edges indicating that the flow in the graph moves from Node A to Node B and then to Node C.



Sometimes, graphs have edges that loop back to a previous node or to the node that is the origin of the edge. Below figure shows a graph with an edge that loops from Node B back to Node A and an edge from Node C to itself. Graphs with these kinds of looping back edges are known as cyclic graphs, and the loops are cycles. Cycles are not allowed in data pipelines, and for that reason the graphs that model data pipelines are directed acyclic graphs.



GCP Pipeline Components

GCP has several services that are commonly used components of pipelines, including the following:

- Cloud Pub/Sub
- Cloud Dataflow
- Cloud Dataproc
- Cloud Composer

Cloud Pub/Sub

Cloud Pub/Sub is a real-time messaging service that supports both push and pull subscription models. It is a managed service, and it requires no provisioning of servers or clusters. Cloud Pub/Sub will automatically scale and partition load as needed.

Working with Messaging Queues

Messaging queues are used in distributed systems to decouple services in a pipeline. This allows one service to produce more output than the consuming service can process without adversely affecting the consuming service. This is especially helpful when one process is subject to spikes in workload.

When working with Cloud Pub/Sub, you create a topic, which is a logical structure for organizing your messages. Once a topic is created, you create a subscription to the topic and then publish messages to the topic. Subscriptions are a logical structure for organizing the reception of messages by consuming processes.

When messaging queues receive data in a message, it is considered a publication event. Upon publication, push subscriptions deliver the message to an endpoint. Some common types of endpoints are Cloud Functions, App Engine, and Cloud Run services. Pull subscriptions are used when you want the consuming application to control when messages are retrieved from a topic. Specifically, with pull subscriptions you send a request asking for N messages, and Cloud Pub/Sub responds with the next N or fewer messages.

Cloud Dataflow

Cloud Dataflow is a managed stream and batch processing service. It is a core component for building pipelines that collect, transform, and output data.

Cloud Dataflow pipelines are written using the Apache Beam API, which is a model for combined stream and batch processing. Apache Beam incorporates Beam runners in the data pipeline; the Cloud Dataflow runner is commonly used in GCP. Apache Flink is another commonly used Beam runner.

Cloud Dataflow does not require you to configure instances or clusters—it is a no-ops service. Cloud Dataflow pipelines are run within a region. It directly integrates with Cloud Pub/Sub, BigQuery, and the Cloud ML Engine. Cloud Dataflow integrates with Bigtable and Apache Kafka.

Cloud Dataflow Concepts

Cloud Dataflow, and the Apache Beam model, are designed around several key concepts:

- Pipelines
- PCollection
- Transforms
- ParDo
- Pipeline I/O
- Aggregation
- User-defined functions
- Runner
- Triggers

Pipelines in Cloud Dataflow are, as you would expect, a series of computations applied to data that comes from a source. Each computation emits the results of computations, which become the input for the next computation in the pipeline. Pipelines represent a job that can be run repeatedly.

The *PCollection* abstraction is a dataset, which is the data used when a pipeline job is run. In the case of batch processing, the PCollection contains a fixed set of data. In the case of streaming data, the PCollection is unbounded.

Google Cloud Certified Professional Cloud Architect Definitive Guide

Transforms are operations that map input data to some output data. Transforms operate on one or more PCollections as input and can produce one or more output PCollections. The operations can be mathematical calculations, data type conversions, and data grouping steps, as well as performing read and write operations.

ParDo is a parallel processing operation that runs a user-specified function on each element in a PCollection. ParDo transforms data in parallel. ParDo receives input data from a main PCollection but may also receive additional inputs from other PCollections by using a *side input*. Side inputs can be used to perform joins. Similarly, while a ParDo produces a main output PCollection, additional collections can be output using a *side output*. Side outputs are especially useful when you want to have additional processing paths. For example, a side output could be used for data that does not pass some validation check.

Pipeline I/Os are transforms for reading data into a pipeline from a source and writing data to a sink.

Aggregation is the process of computing a result from multiple input values. Aggregation can be simple, like counting the number of messages arriving in a one-minute period or averaging the values of metrics received over the past hour.

User-defined functions (UDF) are user-specified code for performing some operation, typically using a ParDo.

Runners are software that executes pipelines as jobs.

Triggers are functions that determine when to emit an aggregated result. In batch processing jobs, results are emitted when all the data has been processed. When operating on a stream, you have to specify a window over the stream to define a bounded subset, which is done by configuring the window.

Cloud Dataproc

Cloud Dataproc is a managed Hadoop and Spark service where a preconfigured cluster can be created with one command line or console operation. Cloud Dataproc makes it easy to migrate from on-premises Hadoop clusters to GCP.

A typical Cloud Dataproc cluster is configured with commonly used components of the Hadoop ecosystem, including the following: -

Hadoop: This is an open source, big data platform that runs distributed processing jobs using the map reduce model. Hadoop writes the results of intermediate operations to disk.

Spark: This is another open source, big data platform that runs distributed applications, but in memory instead of writing the results of map reduce operations to disk.

Pig: This is a compiler that produces map reduce programs from a high-level language for expressing operations on data.

Hive: This is a data warehouse service built on Hadoop.

When working with Cloud Dataproc, you must know how to manage data storage, configure a cluster, and submit jobs.

Cloud Dataproc allows the possibility to use “ephemeral” clusters, where a large cluster can be created to run a task and then destroyed once the task is over in order to save costs.

Cloud Composer

Cloud Composer is a managed service implementing Apache Airflow, which is used for scheduling and managing workflows. As pipelines become more complex and have to be resilient when errors occur, it becomes more important to have a framework for managing workflows so that you are not reinventing code for handling errors and other exceptional cases.

Cloud Composer automates the scheduling and monitoring of workflows. Workflows are defined using Python and are directed acyclic graphs. Cloud Composer has built-in integration with

BigQuery, Cloud Dataflow, Cloud Dataproc, Cloud Datastore, Cloud Storage, Cloud Pub/Sub, and AI Platform.

Before you can run workflows with Cloud Composer, you will need to create an environment in GCP. Environments run on the Google Kubernetes Engine, so you will have to specify a number of nodes, location, machine type, disk size, and other node and network configuration parameters. You will need to create a Cloud Storage bucket as well.

Migrating Hadoop and Spark to GCP

When you are migrating Hadoop and Spark clusters to GCP, there are a few things for which you will need to plan:

- Migrating data
- Migrating jobs
- Migrating HBase to Bigtable

You may also have to shift your perspective on how you use clusters. On-premises clusters are typically large persistent clusters that run multiple jobs. They can be complicated to configure and manage. In GCP, it is a best practice to use an ephemeral cluster for each job. This approach leads to less complicated configurations and reduced costs, since you are not storing persistent data on the cluster and not running the cluster for extended periods of time.

Hadoop and Spark migrations can happen incrementally, especially since you will be using ephemeral clusters configured for specific jobs. The first step is to migrate some data to Cloud Storage. Then you can deploy ephemeral clusters to run jobs that use that data. It is best to start with low-risk jobs so that you can learn the details of working with Cloud Dataproc.

There may be cases where you will have to keep an on-premises cluster while migrating some jobs and data to GCP. In those cases, you will have to keep data synchronized between environments. Plan to implement workflows to keep data synchronized. You should have a way to determine which jobs and data move to the cloud and which stay on premises.

Google Cloud Certified Professional Cloud Architect Definitive Guide

It is a good practice to migrate HBase databases to Bigtable, which provides consistent, scalable performance. When migrating to Bigtable, you will need to export HBase tables to sequence files and copy those to Cloud Storage. Next, you will have to import the sequence files using Cloud Dataflow. When the size of data to migrate is greater than 20 TB, use the Transfer Appliance. When the size is less than 20 TB and there is at least 100 Mbps of network bandwidth available, then distcp, a Hadoop distributed copy command, is the recommended way to copy the data. In addition, it is important to know how long it will take to transfer the data and to have a mechanism for keeping the on-premises data in sync with the data in Cloud Storage.

Data Operations

Data may be operated on in several ways:

- Cataloged
 - Preprocessed
 - Visualized
 - Explored
 - Processed with workflows
-
- Data Catalog, a metadata management service supporting the discovery and management of datasets in Google Cloud.
 - Cloud Dataprep, a pre-processing tool for transforming and enriching data
 - Data Studio for visualizing data and Cloud Datalab for interactive exploration and scripting

Data Catalog

Data Catalog is a GCP metadata service for data management. It is fully managed, so there are no servers to provision or configure. Its primary function is to provide a single, consolidated view of enterprise data. Metadata is collected automatically during ingest operations to BigQuery and Cloud Pub/Sub as well through APIs and third-party tools. BigQuery metadata is collected on datasets, tables, and views. Cloud Pub/Sub topic metadata is also automatically collected.

Before you can use Data Catalog to capture metadata, you need to enable the Data Catalog API in a project that contains the resources created or accessed via the API.

Searching in Data Catalog

With the Data Catalog search capabilities, users can filter and find native metadata, which is captured from the underlying storage system that houses the subject data and user generated metadata that is collected from tags.

To be able to search metadata with Data Catalog, a user will need permissions to read metadata for the subject assets, such as a BigQuery dataset or a Pub/Sub topic. It is important to remember that Data Catalog is collecting and searching metadata, not the data in the dataset, table, topic, and so forth.

When metadata is collected from underlying storage systems, Data Catalog is a read-only service. Any changes to the metadata must be made through the underlying storage system. Data Catalog will collect metadata automatically from several resources within a project, including the following:

- Cloud Storage
- Cloud Bigtable
- Google Sheets
- BigQuery
- Cloud Pub/Sub

Metadata can also be collected manually

Tagging in Data Catalog

Tags are commonly used in GCP and other public clouds to store metadata about a resource. Tags are used for a wide range of metadata, such as assigning a department or team to all resources that they create in a project or specifying a data classification level to a Cloud Storage bucket or object. Data Catalog uses templates to help manage user-defined metadata.

Dataprep

Cloud Dataprep is a managed service designed to help reduce the time required to prepare data for analysis by providing tools to explore, cleanse, and transform data. There are no servers to provision or configure.

Cloud Dataprep is an interactive tool that relies on dragging and dropping components within a workflow rather than programming scripts. Users can read data directly from Cloud Storage and BigQuery as well as upload data from their local machines. When data is uploaded, Cloud Dataprep attempts to automatically detect the schema, the data types of values, the distribution of numeric data, and missing or anomalous values.

Figure shows the kind of summary data that users will see when working with Cloud Dataprep.



Cleansing Data

Cleansing often requires careful attention to detail about data virtually anywhere in a dataset. In some cases, only a small number of values in a column are missing or incorrectly formatted, and sometimes every value in a column needs to be corrected. Also, there are many ways that data can be incorrect, and each way requires a different procedure to correct.

The main cleansing operations in Cloud Dataprep center around altering column names, reformatting strings, and working with numeric values.

Here are some examples cleansing tasks that can be performed with Cloud Dataprep:

- Renaming columns
- Changing the datatype of a column
- Copying from one column to another
- Removing and deduplicating data
- Modifying strings
- Extracting values from strings
- Formatting dates
- Applying conditional transformations

The cleansing phase of working with a dataset is often iterative. You may find some data formats that you want to change and then begin to explore the data only to realize that additional anomalies are in the dataset that you need to correct. The interactive nature of Cloud Dataprep supports this kind of ad hoc, iterative sequence of steps.

Discovering Data

Another step-in processing data for analysis and machine learning is identifying patterns and inconsistencies in your datasets.

Cloud Dataprep supports this process by providing for the following:

- Filtering data
- Locating outliers
- Deriving aggregates, such as counts
- Calculating values across columns
- Comparing strings

In addition to performing data cleansing and discovery operations interactively, users can capture sequences of operations in a structure known as a recipe.

Enriching Data

Sometimes, datasets need to be augmented with additional columns. For example, datasets may need to be joined or appended together before using them for analysis or machine learning model building.

Cloud Dataprep supports several types of enrichment operations, including the following:

- Adding two columns
- Generating primary keys
- Adding lookup data
- Appending datasets
- Joining datasets
- Adding metadata

In the case of adding metadata, Cloud Dataprep can work with data outside the data in datasets. For example, you can reference source file path and filename, creation data, date of importing, and other metadata attributes.

Importing and Exporting Data

Cloud Dataprep can import a number of flat file formats, including

- Microsoft Excel format (XLS/XLSX)
- CSV
- JSON, including nested
- Plain text
- Tab-separated values (TSV)
- Parquet

The service can also read CSV and JSON files compressed with GZIP, BZIP, and Snappy. Avro files compressed with Snappy can also be imported.

Google Cloud Certified Professional Cloud Architect Definitive Guide

Cloud Dataprep does not change source data, but it is possible to export data after preparing it.

Data can be exported to the following:

- CSV
- JSON
- Avro
- BigQuery tables

Users can write compressed CSV and JSON files using GZIP or BZIP.

When data is imported, Cloud Dataprep creates a reference to the dataset, except when data is uploaded from a local device, in which case a copy of the dataset is created.

Structuring and Validating Data

Cloud Dataprep has functionality for more advanced transformations, including the following:

- Reshaping data
- Splitting columns
- Creating aggregations
- Pivoting data
- Manipulating arrays
- Manipulating JSON

There are also tools for validating data, including profiling source data. Profiles include information about

- Mismatched values in columns
- Missing values in columns
- Statistical breakout by quartile

Once data has been prepared using Cloud Dataprep, you can then move on to visualize the data with Data Studio.

Data Studio

Data Studio is a reporting and visualization tool. The tool is organized around reports, and it reads data from data sources and formats the data into tables and charts.

Many business use cases will require the use of Data Studio, including data warehouse reporting and monitoring with dashboards. The three basic tasks in Data Studio are connecting to data sources, visualizing data, and sharing reports.

Connecting to Data Sources

Data Studio uses the concept of a connector for working with datasets. Datasets can come in a variety of forms, including a relational database table, a Google Sheet, or a BigQuery table. Connectors provide access to all or a subset of columns in a data source. Connectors typically require you to authorize access to data.

There are three kinds of connectors:

- **Google Connectors** These are provided by Google for accessing data from other Google services, including Analytics, Ads, Sheets, and BigQuery.
- **Partner Connectors** These connectors are developed by third parties, and they provide access to non-Google data such as Facebook, GitHub, and Reddit data sources.
- **Community Connectors** These connectors are developed by anyone with a need to access a data source.

There are also three types of data sources:

- **Live Connection Data Sources** These data sources are automatically updated with changes to the underlying data source. Data is stored in the source system. Most connectors work with live data sources.
- **Extracted Data Sources** These data sources work with a static snapshot of a dataset, which can be updated on demand. These may give better performance than live connection data sources.
- **Blended Data Sources** These data sources are the result of combining data from up to five data sources.

Once you have connected to a data source in Data Studio, you can start visualizing data.

Visualizing Data

Data Studio provides components that can be deployed in a drag-and-drop manner to create reports. Data Studio reports are collections of tables and visualizations. The visualization components include the following:

- Line charts
- Bar charts
- Pie charts
- Geo maps
- Area and bubble graphs
- Paginated data tables
- Pivot tables

Users can use filters and data ranges to restrict the set of data included in report tables and charts.

Sharing Data

Developers of reports can share the report with others, who can then view or edit the report. Reports can also be made available to non-Google users with link sharing. Data Studio provides the option to schedule the running of a report and the generation of a PDF file, which can be emailed to recipients

Datalab

Cloud Datalab is an interactive tool for exploring and transforming data. Cloud Datalab runs as an instance of a container. Users of Cloud Datalab create a Compute Engine instance, run the container, and then connect from a browser to a Cloud Datalab notebook.

Cloud Datalab containers run an instance of a Jupyter Notebook.

Jupyter Notebooks

Jupyter Notebooks are documents that can contain code as well as text. Code and text are located in cells. All text and code within a cell are treated as a single unit. For example, when a cell is executed, all code in the cell is executed.

Jupyter Notebooks are widely used in data science, machine learning, and other tasks that lend themselves to interactive, iterative development. Jupyter Notebooks support a variety of programming languages, including Python and SQL.

Managing Cloud Datalab Instances

It is a relatively simple task to create and use Cloud Datalab instances. With the Cloud software development kit (SDK) already installed, including the optional Datalab component, a user can create a Cloud Datalab instance with the `datalab create` command.

For example: `datalab create --machine-type n1-highmem-2 my-datalab-instance-1`

Once the instance is created, the user can connect using the `datalab connect` command. The default port for connecting is 8081, but that can be changed by specifying the `--port` option in the `datalab create` command. The `datalab list` command provides a list of all running instances.

A Datalab instance can be deleted using the `datalab delete` command. By default, this command does not delete the persistent disk attached to the instance, assuming that the disk's configuration is also set to the default. To delete the persistent instance as well, users need to specify the `--delete-disk` option.

Adding Libraries to Cloud Datalab Instances

Data scientists and machine learning engineers often need to import libraries when working with Python.

Some of the most commonly used libraries are as follows:

- Numpy: A high-performance scientific computing package
- Scipy: An open-source package for science and engineering that uses numpy
- Pandas: An open-source package for working with tabular data
- Scikit Learn: An open-source machine learning package
- TensorFlow: An open-source machine learning package for deep learning

Many of the commonly used packages are available in Cloud Datalab, but when a user needs to add others, this is done by using either the conda install command or the pip install command.

For example, to install the data analysis package scikit-data, a user would specify the following command in a Jupyter Notebook cell: `!conda install scikit-data`

This command runs the conda installer to download the scikit-data package. Some libraries are not available through the conda installer; in that case, the pip installer can be used.

For example, the topological data analysis tool is currently not available through the conda installer, so pip should be used to install that library: `!pip install scikit-tda`

The `!` character at the beginning of the command indicates to Jupyter Notebook that the command should be run as a shell command, not a Python statement.

In some cases, exploratory data analysis is an end in itself, but in many cases, it is the first step to defining a workload that will be repeated. In those cases, you can use Cloud Composer to orchestrate those workloads.

Cloud Composer

Cloud Composer is a fully managed workflow orchestration service based on Apache Airflow. Workflows are defined as directed acyclic graphs (DAGs), which are specified in Python.

Workflows can make use of many GCP services, including the following:

- BigQuery
- Cloud Dataflow
- Cloud Dataproc
- Cloud Datastore
- Cloud Storage
- Cloud Pub/Sub
- AI Platform

Elements of workflows can run on premises and in other clouds as well as in GCP.

Airflow Environments

Apache Airflow is a distributed platform, and it requires several GCP services. When it is deployed, the GCP resources deployed are known as a Cloud Composer environment.

Environments are stand-alone deployments based on Kubernetes Engine. There can be multiple Cloud Composer environments in a single project.

Environments can be created in the GCP console or by using the command line. The SDK command to create an environment is `gcloud beta composer`.

When you create an instance, you can specify node configuration and network configuration, as well as environment variables.

Creating DAGs

Airflow DAGs are defined in Python as a set of operators and operator relationships. An operator specifies a single task in a workflow.

The most commonly used operators are as follows:

- **BashOperator** : Executes a command in the Bash shell
- **PythonOperator** : Executes a Python function
- **EmailOperator** : Sends an email message
- **SimpleHTTPOperator** : Sends HTTP requests
- **Database operators**: Includes PostgresOperator, MySQLOperator, SQLiteOperator, and JdbcOperator
- **Sensor**: Waits for a certain event, such as a specific time or the creation of a file or other resource

The order of operators is specified using the `>>` symbol. For example, assuming that you have created a `write_files_python` PythonOperator and a `delete_temp_files_bash` BashOperator , you can have `write_files_python` executed first followed by `delete_temp_files_bash` as follows:

```
write_files_python >> delete_temp_files_bash
```

Airflow Logs

The Airflow environment creates two types of logs: Airflow logs and streaming logs. Airflow logs are associated with a single DAG task. These files are stored in the Cloud Storage logs folder of the Cloud Composer environment. Logs are retained after an environment is shut down. You will need to delete logs manually. Streaming logs are a superset of Airflow logs. These logs are stored in Stackdriver and can be viewed using the Logs viewer. You can also use log-based metrics for monitoring and alerting.

Airflow generates several logs, including the following:

- **Airflow-database-init-job:** For database initialization
- **Airflow-scheduler:** For logs generated by the scheduler
- **Airflow-webserver:** For logs generated by the web interface
- **Airflow-worker:** For logs generated as DAGs are executed
- **Airflow-monitoring:** For logs generated by Airflow monitoring
- **Airflow:** For otherwise uncategorized logs

Cloud Composer are that it is a workflow orchestration service that runs within Kubernetes Engine and executes tasks specified in a Python script composed of operators that execute tasks. Tasks can be executed on a schedule, manually, or in response to an external event.

Bigdata – Interview | Exam Tips

Model of data pipelines

A data pipeline is an abstract concept that captures the idea that data flows from one stage of processing to another. Data pipelines are modeled as directed acyclic graphs (DAGs). A graph is a set of nodes linked by edges. A directed graph has edges that flow from one node to another.

Four stages in a data pipeline

Ingestion is the process of bringing data into the GCP environment. Transformation is the process of mapping data from the structure used in the source system to the structure used in the storage and analysis stages of the data pipeline. Cloud Storage can be used as both the staging area for storing data immediately after ingestion and also as a long-term store for transformed data. BigQuery and Cloud Storage treat data as external tables and query them. Cloud Dataproc can use Cloud Storage as HDFS-compatible storage. Analysis can take on several forms, from simple SQL querying and report generation to machine learning model training and data science analysis.

Structure and function of data pipelines will vary according to the use case to which they are applied

Three common types of pipelines are data warehousing pipelines, stream processing pipelines, and machine learning pipelines.

Common patterns in data warehousing pipelines

Extract, transformation, and load (ETL) pipelines begin with extracting data from one or more data sources. When multiple data sources are used, the extraction processes need to be coordinated. This is because extractions are often time based, so it is important that extracts from different sources cover the same time period. Extract, load, and transformation (ELT) processes are slightly different from ETL processes. In an ELT process, data is loaded into a database before transforming the data. Extraction and load procedures do not transform data. This kind of process is appropriate when data does not require changes from the source format. In a change data capture approach, each change is a source system that is captured and recorded in a data store. This is helpful in cases where it is important to know all changes over time and not just the state of the database at the time of data extraction.

Unique processing characteristics of stream processing

This includes the difference between event time and processing time, sliding and tumbling windows, late arriving data and watermarks, and missing data. Event time is the time that something occurred at the place where the data is generated. Processing time is the time that data arrives at the endpoint where data is ingested. Sliding windows are used when you want to show how an aggregate, such as the average of the last three values, change over time, and you want to update that stream of averages each time a new value arrives in the stream. Tumbling windows are used when you want to aggregate data over a fixed period of time—for example, for the last one minute.

Components of a typical machine learning pipeline

This includes data ingestion, data preprocessing, feature engineering, model training and evaluation, and deployment. Data ingestion uses the same tools and services as data warehousing and streaming data pipelines. Cloud Storage is used for batch storage of datasets, whereas Cloud Pub/Sub can be used for the ingestion of streaming data. Feature engineering is a machine learning practice in which new attributes are introduced into a dataset. The new attributes are derived from one or more existing attributes.

Cloud Pub/Sub is a managed message queue service

Cloud Pub/Sub is a real-time messaging service that supports both push and pull subscription models. It is a managed service, and it requires no provisioning of servers or clusters. Cloud Pub/Sub will automatically scale as needed. Messaging queues are used in distributed systems to decouple services in a pipeline. This allows one service to produce more output than the consuming service can process without adversely affecting the consuming service. This is especially helpful when one process is subject to spikes.

Cloud Dataflow is a managed stream and batch processing service

Cloud Dataflow is a core component for running pipelines that collect, transform, and output data. In the past, developers would typically create a stream processing pipeline (hot path) and a separate batch processing pipeline (cold path). Cloud Dataflow is based on Apache Beam, which is a model for combined stream and batch processing. Understand these key Cloud Dataflow concepts:

- Pipelines
- PCollection

- Transforms
- ParDo
- Pipeline I/O
- Aggregation
- User-defined functions
- Runner
- Triggers

Cloud Dataproc is a managed Hadoop and Spark service

Cloud Dataproc makes it easy to create and destroy ephemeral clusters. Cloud Dataproc makes it easy to migrate from on-premises Hadoop clusters to GCP. A typical Cloud Dataproc cluster is configured with commonly used components of the Hadoop ecosystem, including Hadoop, Spark, Pig, and Hive. Cloud Dataproc clusters consist of two types of nodes: master nodes and worker nodes. The master node is responsible for distributing and managing workload distribution.

Cloud Composer is a managed service implementing Apache Airflow

Cloud Composer is used for scheduling and managing workflows. As pipelines become more complex and have to be resilient when errors occur, it becomes more important to have a framework for managing workflows so that you are not reinventing code for handling errors and other exceptional cases. Cloud Composer automates the scheduling and monitoring of workflows. Before you can run workflows with Cloud Composer, you will need to create an environment in GCP.

Consideration for migrating from on-premises Hadoop and Spark to GCP

Factors include migrating data, migrating jobs, and migrating HBase to Bigtable. Hadoop and Spark migrations can happen incrementally, especially since you will be using ephemeral clusters configured for specific jobs. There may be cases where you will have to keep an on-premises cluster while migrating some jobs and data to GCP. In those cases, you will have to keep data synchronized between environments. It is a good practice to migrate HBase databases to Bigtable, which provides consistent, scalable performance.

Data Catalog is a metadata service for data management

Data Catalog is fully managed, so there are no servers to provision or configure. Its primary function is to provide a single, consolidated view of enterprise data. Metadata is collected automatically during ingest operations to BigQuery and Cloud Pub/Sub, as well through APIs and third-party tools.

Data Catalog will collect metadata automatically from several GCP sources

These sources include Cloud Storage, Cloud Bigtable, Google Sheets, BigQuery, and Cloud Pub/Sub. In addition to native metadata, Data Catalog can collect custom metadata through the use of tags.

Cloud Dataprep is an interactive tool for preparing data for analysis and machine learning

Cloud Dataprep is used to cleanse, enrich, import, export, discover, structure, and validate data. The main cleansing operations in Cloud Dataprep center around altering column names, reformatting strings, and working with numeric values. Cloud Dataprep supports this process by providing for filtering data, locating outliers, deriving aggregates, calculating values across columns, and comparing strings.

Data Studio as a reporting and visualization tool

The Data Studio tool is organized around reports, and it reads data from data sources and formats the data into tables and charts. Data Studio uses the concept of a connector for working with datasets. Datasets can come in a variety of forms, including a relational database table, a Google Sheet, or a BigQuery table. Connectors provide access to all or to a subset of columns in a data source. Data Studio provides components that can be deployed in a drag-and-drop manner to create reports. Reports are collections of tables and visualization.

Cloud Datalab is an interactive tool for exploring and transforming data

Cloud Datalab runs as an instance of a container. Users of Cloud Datalab create a Compute Engine instance, run the container, and then connect from a browser to a Cloud Datalab notebook, which is a Jupyter Notebook. Many of the commonly used packages are available in Cloud Datalab, but when users need to add others, they can do so by using either the conda install command or the pip install command.

Cloud Composer is a fully managed workflow orchestration service based on Apache Airflow

Workflows are defined as directed acyclic graphs, which are specified in Python. Elements of workflows can run on premises and in other clouds as well as in GCP. Airflow DAGs are defined in Python as a set of operators and operator relationships. An operator specifies a single task in a workflow. Common operators include Bash Operator and Python Operator.

Bigdata Quiz

1. A large enterprise using GCP has recently acquired a startup that has an IoT platform. The acquiring company wants to migrate the IoT platform from an on-premises data center to GCP and wants to use Google Cloud managed services whenever possible. What GCP service would you recommend for ingesting IoT data?

- A. Cloud Storage
- B. Cloud SQL
- C. Cloud Pub/Sub**
- D. BigQuery streaming inserts

Correct Answer: C

2. You are designing a data pipeline to populate a sales data mart. The sponsor of the project has had quality control problems in the past and has defined a set of rules for filtering out bad data before it gets into the data mart. At what stage of the data pipeline would you implement those rules?

- A. Ingestion
- B. Transformation**
- C. Storage
- D. Analysis

Correct Answer: B

3. A team of data warehouse developers is migrating a set of legacy Python scripts that have been used to transform data as part of an ETL process. They would like to use a service that allows them to use Python and requires minimal administration and operations support.

Which GCP service would you recommend?

- A. Cloud Dataproc
- B. Cloud Dataflow**
- C. Cloud Spanner
- D. Cloud Dataprep

Correct Answer: B

4. You are using Cloud Pub/Sub to buffer records from an application that generates a stream of data based on user interactions with a website. The messages are read by another service that transforms the data and sends it to a machine learning model that will use it for training. A developer has just released some new code, and you notice that messages are sent repeatedly at 10-minute intervals. What might be the cause of this problem?

- A. The new code release changed the subscription ID.
- B. The new code release changed the topic ID.
- C. The new code disabled acknowledgments from the consumer.**
- D. The new code changed the subscription from pull to push.

Correct Answer: C

5. It is considered a good practice to make your processing logic idempotent when consuming messages from a Cloud Pub/Sub topic. Why is that?

- A. Messages may be delivered multiple times.**
- B. Messages may be received out of order.
- C. Messages may be delivered out of order.
- D. A consumer service may need to wait extended periods of time between the delivery of messages.

Correct Answer: A

6. A group of IoT sensors is sending streaming data to a Cloud Pub/Sub topic. A Cloud Dataflow service pulls messages from the topic and reorders the messages sorted by event time. A message is expected from each sensor every minute. If a message is not received from a sensor, the stream processing application should use the average of the values in the last four messages. What kind of window would you use to implement the missing data logic?

- A. Sliding window**
- B. Tumbling window
- C. Extrapolation window
- D. Crossover window

Correct Answer: A

7. Your department is migrating some stream processing to GCP and keeping some on-premises. You are tasked with designing a way to share data from on-premises pipelines that use Kafka with GPC data pipelines that use Cloud Pub/Sub. How would you do that?

- A. Use CloudPubSubConnector and Kafka Connect
- B. Stream data to a Cloud Storage bucket and read from there
- C. Write a service to read from Kafka and write to Cloud Pub/Sub
- D. Use Cloud Pub/Sub Import Service

Correct Answer: A

8. A team of developers wants to create standardized patterns for processing IoT data.

Several teams will use these patterns. The developers would like to support collaboration and facilitate the use of patterns for building streaming data pipelines. What component should they use?

- A. Cloud Dataflow Python Scripts
- B. Cloud Dataproc PySpark jobs
- C. Cloud Dataflow templates**
- D. Cloud Dataproc templates

Correct Answer: C

9. You need to run several map-reduce jobs on Hadoop along with one Pig job and four PySpark jobs. When you ran the jobs on premises, you used the department's Hadoop cluster. Now you are running the jobs in GCP. What configuration for running these jobs would you recommend?

- A. Create a single cluster and deploy Pig and Spark in the cluster.
- B. Create one persistent cluster for the Hadoop jobs, one for the Pig job and one for the PySpark jobs.
- C. Create one cluster for each job, and keep the cluster running continuously so that you do not need to start a new cluster for each job.
- D. Create one cluster for each job and shut down the cluster when the job completes.**

Correct Answer: D

10. You are working with a group of genetics researchers analyzing data generated by gene sequencers. The data is stored in Cloud Storage. The analysis requires running a series of six programs, each of which will output data that is used by the next process in the pipeline. The final result set is loaded into BigQuery. What tool would you recommend for orchestrating this workflow?

- A. Cloud Composer
- B. Cloud Dataflow
- C. Apache Flink
- D. Cloud Dataproc

Correct Answer: A

11. An on-premises data warehouse is currently deployed using HBase on Hadoop. You want to migrate the database to GCP. You could continue to run HBase within a Cloud Dataproc cluster, but what other option would help ensure consistent performance and support the HBase API?

- A. Store the data in Cloud Storage
- B. Store the data in Cloud Bigtable**
- C. Store the data in Cloud Datastore
- D. Store the data in Cloud Dataflow

Correct Answer: B

12. The business owners of a data warehouse have determined that the current design of the data warehouse is not meeting their needs. In addition to having data about the state of systems at certain points in time, they need to know about all the times that data changed between those points in time. What kind of data warehousing pipeline should be used to meet this new requirement?

- A. ETL
- B. ELT
- C. Extraction and load
- D. Change data capture**

Correct Answer: D

13. What types of jobs can be run using Dataproc?

- A. Spark
- B. Hive
- C. MapReduce
- D. All of the above**

Correct Answer: D

14. Can a cluster be resized using Dataproc?

- A. Yes**
- B. No

Correct Answer: A

15. Can data be shared across pipeline?

- A. Yes
- B. No**

Correct Answer: B

16. What are the common big data challenges that you will be building solutions for in this course? (Check all that apply)

- A. Migrating existing on-premise workloads to the cloud--**
- B. Analyzing large datasets at scale--**
- C. Building containerized applications for web development
- D. Building streaming data pipelines--**
- E. Applying machine learning to your datasets--**

Correct Answer: A, B, D, E

17. You have a large enterprise that will likely have many teams using their own Google Cloud Platform projects and resources. What should you be sure to have to help manage and administer these resources? (Check all that apply)

- A. A defined Organization--**
- B. Folders for teams and/or products--**
- C. A defined access control policy with Cloud IAM--**
- D. A Kubernetes or Hadoop cluster for each project**

Correct Answer: A, B, C

18. Which of the following is NOT one of the advantages of Google Cloud security

- A. Google Cloud will automatically manage and curate your content and access policies to be safe for the public--**
- B. Google Cloud will secure the physical hardware that is running your applications and infrastructure--**
- C. Google Cloud has tools like Cloud IAM that help you administer and set company-wide security policies--**
- D. Google Cloud will manage audit logging of access and use of resources in your account**

Correct Answer: B, C

19. If you don't have a large dataset of your own but still want to practice writing queries and building pipelines on Google Cloud Platform, what should you do?

- A. Practice with the datasets in the Google Cloud Public Datasets program--**
- B. Find other public datasets online and upload them into BigQuery--**
- C. Work to create your own dataset and then upload it into BigQuery for analysis--**

Correct Answer: A, B, C

20. As you saw in the demo, Compute Engine nodes on GCP are:

- A. One of ~50 choices in terms of CPU and memory
- B. Expensive to create and teardown
- C. Pre-installed with all the software packages you might ever need.
- D. Allocated on demand, and you pay for the time that they are up. —**

Correct Answer: D

21. Quick quiz. You need a table to hold the dataset

- A. True
- B. False--**

Correct Answer: B

22. Quick quiz. Check which you can use to access BigQuery

- A. checkThird-party tools--**
- B. checkMake calls to BigQuery REST API--**
- C. checkWeb UI--**
- D. checkCommand line tool--

Correct Answer: A, B, C

23. You should feed your machine learning model your _____ and not your _____. It will learn those for itself!

- A. data, rules
- B. if/then statements, data
- C. rules, data--**

Correct Answer: C

24. True or False: Cloud SQL is a big data analytics warehouse

A. True--

B. False

Correct Answer: A

Correct - Cloud SQL is a transaction RDBMS or relational database management system. It is designed for many more WRITES than READS Whereas BigQuery is a big data analytics warehouse which is optimized for reporting READS.

25. True or False: If you are migrating your Hadoop workload to the cloud, you must first rewrite all your Spark jobs to be compliant with the cloud.

A. True

B. False—

Correct Answer: B

Correct - you can run your same Spark job code running on the same Hadoop software but running on cloud hardware with Cloud Dataproc.

26. You are thinking about migrating your Hadoop workloads to the cloud and you have a few workloads that are fault-tolerant (they can handle interruptions of individual VMs gracefully). What are some architecture considerations you should explore in the cloud? Choose all that apply

Use PVMs or Preemptible Virtual Machines

- A. Migrate your storage from on-cluster HDFS to off-cluster Google Cloud Storage (GCS)--**
- B. Consider having multiple Cloud Dataproc instances for each priority workload and then turning them down when not in use--**

Correct Answer: A, B

27. Google Cloud Storage is a good option for storing data that: (Select the 2 correct options below).

- A. May be required to be read at some later time (i.e. load a CSV file into BigQuery)--**
- B. Is ingested in real-time from sensors and other devices and supports SQL-based queries
- C. Will be accessed frequently and updated constantly with new transactions from a front-end and needs to be stored in a relational database
- D. May be imported from a bucket into a Hadoop cluster for analysis--**

Correct Answer: A, D

28. Relational databases are a good choice when you need:

- A. Transactional updates on relatively small datasets**
- B. Fast queries on terabytes of data
- C. Streaming, high-throughput writes
- D. Aggregations on unstructured data

Correct Answer: A

29. Cloud SQL and Cloud Dataproc offer familiar tools (MySQL and Hadoop/Pig/Hive/Spark).

What is the value-add provided by Google Cloud Platform?

(Select the 2 correct options below)

- A. It's the same API, but Google implements it better
- B. Google-proprietary extensions and bug fixes to MySQL, Hadoop, and so on
- C. Fully-managed versions of the software offer no-ops—**

Yes. No-ops is the main value-add here.

- D. Running it on Google infrastructure offers reliability and cost savings—**

Yes. You pay only for the resources you use. Cloud SQL can be shut down when it's not being used.

Hadoop clusters can be of preemptible nodes, and so on.

Correct Answer: C, D

30. Which of the below are the core services that make up BigQuery? (choose the correct 2)

- A. Query service--
- B. Storage service--
- C. Data Optimization service
- D. Machine Learning service

Correct Answer: A, B

31. You want to know how many rows are in the BigQuery Public Dataset on San Francisco Bike Shares. What could you do?

Run the below query:

```
SELECT  
SUM(*) AS total_trips  
FROM  
'bigquery-public-data.san_francisco_bikeshare.bikeshare_trips'
```

#In the BigQuery Web UI, find the table and click the details tab and view the rows.--

Run the below query:--

```
SELECT  
COUNT(*) AS total_trips  
FROM  
'bigquery-public-data.san_francisco_bikeshare.bikeshare_trips'--
```

True or False: You can query a Google Spreadsheet directly from BigQuery without loading it in first.

- A. True-- this is a federated query
- B. False

Correct Answer: A

32. You have a taxi service data schema that has three columns:

- ride_id
- ride_timestamp
- ride_status

You want to use BigQuery for reporting but you don't want to split your table into multiple sub-tables. What native features of BigQuery data types should you explore? (check all that apply)

Consider adding lat / long geographic data points as new columns and using GIS Functions to quickly plot the distances your fleet has travelled.--

Consider making ride_timestamp an ARRAY of timestamp values so each ride_id row in your table could still be unique and easy to report off of.--

Consider renaming the ride_id column to 'label' so you can use it in a BigQuery ML model to predict the ride_id of the next ride.

Complete the following

In ML, a row of data is called a(n) _____ and a column of data is called a(n) _____.

We mark one or more columns as _____ which we know for historical data and are trying to predict for future data.

- a. labels
- b. instance or observation
- c. feature
- d. instance or observation
- e. labels
- f. Feature
- g. instance or observation--**
- h. feature--**
- i. labels--**

Correct Answer: g, h, i

33. If you have an image classification task for identifying whether a car is present in a photo or not, which solution should you try first?

- A. Try the Cloud Vision API first--
- B. Try AutoML Vision first
- C. Try a custom model in BQML first
- D. Try a custom model in TensorFlow first

Correct Answer: A

34. Which of the following are the jobs of a data engineer?

- A. Get the data to where it can be useful--
- B. Get the data into a usable condition--
- C. Add new value to the data--
- D. Manage the data--
- E. Productionize data processes--

Correct Answer: A, B, C, D, E

35. Which statements are true?

- A. Cloud SQL is optimized for high-throughput writes--
- B. BigQuery is optimized for high-read data--
- C. BigQuery is a row-based storage
- D. Cloud SQL is optimized for high-read data

Correct Answer: A, B

36. Which statement best describes a data lake?

- A. The place where you capture every aspect of your business operations. Data is stored in its natural, raw format. --**
- B. Data storage intended for analytics.
- C. Storage optimized for high-throughput writes.
- D. Storage for current/historical data intended for reporting. --**

Correct Answer: A, D

37. Which of the following statements on Cloud Storage are true?

- A. Cloud Storage simulates a file system--**
- B. Cloud Storage allows you to set retention policies on all objects in a bucket--**
- C. Cloud Storage implements both Cloud IAM policy and Access Control Lists--**
- D. Data in Cloud Storage is not encrypted

Correct Answer: A, B, C

38. Which of the following statements on BigQuery are true?

- A. Data is run length-encoded and dictionary-encoded--**
- B. Data on BigQuery is physically stored in a redundant way separate from the compute cluster--**
- C. A BigQuery slot is a combination of CPU, memory, and networking resources--**
- D. The number of slots allotted to a query is independent of query complexity

Correct Answer: A, B, C

39. True or False: ARRAYS can be part of regular fields or STRUCTS in BigQuery?

- A. True--**
- B. False

Correct Answer: A

40. Analysts and data scientists at your company ask for your help with data preparation. They currently spend significant amounts of time searching for data and trying to understand the exact definition of the data. What GCP service would you recommend that they use?

A. Cloud Composer

B. Data Catalog

C. Cloud Dataprep

D. Data Studio

Correct Answer: B

41. Machine learning engineers have been working for several weeks on building a recommendation system for your company's e-commerce platform. The model has passed testing and validation, and it is ready to be deployed. The model will need to be updated every day with the latest data. The engineers want to automate the model building process that includes running several Bash scripts, querying databases, and running some custom Python code. What GCP service would you recommend that they use?

A. Cloud Composer

B. Data Catalog

C. Cloud Dataprep

D. Data Studio

Correct Answer: A

42. A business intelligence analyst has just acquired several new datasets. They are unfamiliar with the data and are especially interested in understanding the distribution of data in each column as well as the extent of missing or misconfigured data. What GCP service would you recommend they use?

A. Cloud Composer

B. Cloud Catalog

C. Cloud Dataprep

D. Data Studio

Correct Answer: C

43. Line-of-business managers have asked your team for additional reports from data in a data warehouse. They want to have a single report that can act as a dashboard that shows key metrics using tabular data as well as charts. What GCP service would you recommend?

- A. Cloud Composer
- B. Data Catalog
- C. Cloud Dataprep
- D. Data Studio**

Correct Answer: D

44. You are using Cloud Dataprep to prepare datasets for machine learning. Another team will be using the data that you prepare, and they have asked you to export your data from Cloud Dataprep. The other team is concerned about file size and asks you to compress the files using GZIP. What formats can you use in the export file?

- A. CSV only
- B. CSV and JSON only**
- C. CSV and AVRO only
- D. JSON and AVRO only

Correct Answer: B

45. The finance department in your company is using Data Studio for data warehouse reporting. Their existing reports have all the information they need, but the time required to update charts and tables is longer than expected. What kind of data source would you try to improve the query performance?

- A. Live data source
- B. Extracted data source**
- C. Compound data source
- D. Blended data source

Correct Answer: B

46. A DevOps team in your company uses Data Studio to display application performance data. Their top priority is timely data. What kind of connection would you recommend they use to have data updated in reports automatically?

- A. Live data source
- B. Extracted data source
- C. Compound or blended data source
- D. Extracted or live data source

Correct Answer: A

47. A machine learning engineer is using Data Studio to build models in Python. The engineer has decided to use a statistics library that is not installed by default. How would you suggest that they install the missing library?

- A. Using conda install or pip install from a Cloud shell
- B. Using conda install or pip install from within a Jupyter Notebook**
- C. Use the Linux package manager from within a Cloud shell
- D. Download the source from GitHub and compile locally

Correct Answer:B

48. A DevOps engineer is working with you to build a workflow to load data from an on-premises database to Cloud Storage and then run several data pre-processing and analysis programs. After those are run, the output is loaded into a BigQuery table, an email is sent to managers indicating that new data is available in BigQuery, and temporary files are deleted. What GCP service would you use to implement this workflow?

- A. Cloud Dataprep
- B. Cloud Dataproc
- C. Cloud Composer**
- D. Data Studio

Correct Answer: C

49. You have just received a large dataset. You have comprehensive documentation on the dataset and are ready to start analyzing. You will do some visualization and data filtering, but you also want to be able to run custom Python functions. You want to work interactively with the data. What GCP service would you use?

- A. Cloud Dataproc
- B. Cloud Datalab**
- C. Cloud Composer
- D. Data Studio

Correct Answer: B

Machine Learning Services

Deploying Machine Learning Pipelines

ML pipelines include several stages, beginning with data ingestion and preparation, then data segregation, followed by model training and evaluation

Structure of ML Pipelines

Machine learning projects begin with a problem definition. This could involve how to improve sales revenue by making product recommendations to customers, how to evaluate medical images to detect tumors, or how to identify fraudulent financial transactions.

These are three distinct types of problems, but they can all be solved using machine learning techniques deployed via ML pipelines.

The stages of a machine learning pipeline are as follows:

- Data ingestion
- Data preparation
- Data segregation
- Model training
- Model evaluation
- Model deployment
- Model monitoring

ML pipelines are more cyclic than linear. This is a difference with dataflow pipelines, like those used to ingest, transform, and store data, which are predominantly linear.

1. Data Ingestion

Data ingestion for machine learning can be either batch or streaming.

1.1 Batch Data Ingestion

Batch data ingestion should use a dedicated process for ingesting each distinct data source.

1.2 Streaming Data Ingestion

Cloud Pub/Sub is designed for scalable messaging, including streaming data ingestion. There are several advantages to using Cloud Pub/Sub for streaming data ingestion. Cloud Pub/Sub is a fully managed, serverless service that is available globally.

2. Data Preparation

Data preparation is the process of transforming data from its raw form into a structure and format that is amenable to analysis by machine learning algorithms. There are three steps to data preparation:

- Data exploration
- Data transformation
- Feature engineering

2.1 Data Exploration

Data exploration is the first step to working with a new data source or a data source that has had significant changes. The goal of this stage is to understand the distribution of data and the overall quality of data.

2.2 Data Transformation

Data transformation is the process of mapping data from its raw form into data structures and formats that allow for machine learning. Transformations can include the following:

- Replacing missing values with a default value
- Replacing missing values with an inferred value based on other attributes in a record
- Replacing missing values with an inferred value based on attributes in other records
- Changing the format of numeric values, such as truncating or rounding real numbers to integers
- Removing or correcting attribute values that violate business logic, such as an invalid product

identifier

- Deduplicating records
- Joining records from different data sets
- Aggregating data, such as summing values of metrics into hour or minute totals

Cloud Dataprep can be used for interactive data transformation, and it is especially useful when you’re working with new datasets. For large volumes of data or cases in which the set of needed transformations is well defined, Cloud Dataflow is a good option for implementing transformations. Cloud Dataflow supports both batch and stream processing.

2.3 Feature Engineering

Feature engineering is the process of adding or modifying the representation of features to make implicit patterns more explicit.

Another common feature engineering technique is one-hot encoding. This process maps a list of categorical values to a series of binary numbers that have a single value set to 1 and all other values set to 0.

3. Data Segregation

Data segregation is the process splitting a data set into three segments: training, validation, and test data.

3.1 Training Data

Machine learning algorithms create models, which are functions that map from some input to a predicted output. The mappings are generated based on data rather than by manually programming them. The data used to generate these mappings is known as the training data.

3.2 Validation Data

Although machine learning algorithms can learn some values, known as parameters, from training data, there are some values that are specified by a machine learning engineer.

These values are known as hyperparameters. Hyperparameters are values that configure a model, and they can include the number of layers in a neural network or the maximum depth of trees in a random forest model. Hyperparameters vary by machine learning algorithm.

The validation data set is used to tune hyperparameters. This is often done by experimenting with different values for hyperparameters. For example, in neural networks, you can specify the number of layers and number of nodes in each layer of neural networks, as well as a learning rate.

The number of layers and number of nodes in each layer of a neural network determine both how well a model will function and how long it will take to train. There is no calculation available to determine the optimal number of layers and nodes—you have to experiment with different combinations of layers and nodes per layer to find the best configuration. The learning rate determines how much neural network weights are adjusted when training. A large learning rate will make larger changes to weights, so the model may learn faster. However, you also risk missing an optimal set of weights because large changes in weights can overshoot the optimal weights. Smaller learning rates will learn more slowly but are more likely to find the optimal set of weights.

Each time you try another combination of hyperparameter values, you will need to train the model and evaluate the results. You should not use the same data to assess your hyperparameter choices as you use to train the model. The data that is used for evaluating hyperparameter choices is known as validation data.

3.3 Test Data

The third data segment is test data. This is data that is not used for either training or hyperparameter tuning. By using data to which the model has never been exposed and that has never been used to tune hyperparameters, you can get an estimate of how well the model will perform with other previously unseen data.

The main criteria for deciding how to split data are as follows:

- Ensuring that the test and validation datasets are large enough to produce statistically meaningful results
- Ensuring that the test and validation datasets are representative of the data as a whole
- Ensuring that the training dataset is large enough for the model to learn from in order to make

After data is segmented, the next step in ML pipelines is model training.

4. Model training

Model training is the process of using training data to create a model that can be used to make predictions.

4.1 Feature selection

Feature selection is another part of model training. This is the process of evaluating how a particular attribute or feature contributes to the predictiveness of a model. The goal is to have features of a dataset that allow a model to learn to make accurate predictions.

Features are selected to do the following:

- Reduce the time and amount of data needed to train a model
- Make models easier to understand
- Reduce the number of features that have to be considered; this is known as the curse of dimensionality
- Reduce the risk of overfitting the model to the training data, which in turn would reduce the model accuracy on data not encountered during training

There are a number of different algorithms for selecting features. The simplest approach is to train a model with each subset of features and see which subset has the best performance. Although it is simple and easy to implement, this naive approach is not scalable.

Another approach is to start with a subset of features, measure its performance, and then make an incremental modification in the subset of features. If the performance of the modified subset is better, the modification is retained, and the process is repeated.

Otherwise, the modification is discarded, and another one is tried. This technique is known as greedy hill climbing.

4.2 Underfitting, Overfitting, and Regularization

Two problems that can occur during model training are underfitting and overfitting. Underfitting creates a model that is not able to predict values of training data correctly or new data that was not used during training. If the model performs poorly across multiple algorithms, and when evaluating the model using the same data that was used to train it, then that is underfitting, and it is likely caused by insufficient training data.

The problem of underfitting may be corrected by increasing the amount of training data, using a different machine learning algorithm, or modifying hyperparameters. In the case of changing hyperparameters, this could include using more nodes or layers in a neural network, which is a collection of interconnected artificial neurons used to make a calculation, or increasing the number and depth of trees in a random forest—in other words, a collection decision trees, which are an ordered conjunction of features that determine a label.

Overfitting occurs when a model fits the training data too well. This happens when there is noise in the data and the model fits the noise as well as the correct data points. Noise is a term borrowed from signal processing to describe data points that are in the dataset but that are not generated by the underlying processes being modelled. A problem in the network, for instance, could corrupt a measurement sent from a sensor, in which case that data point would be noise.

One way to compensate for the impact of noise in the data and reduce the risk of overfitting is by introducing a penalty for data points that make the model more complicated. This process is called regularization. Two kinds of regularization are L1 regularization, which is also known as Lasso Regularization, for Least Absolute Shrinkage and Selection Operator, and L2, or Ridge Regression.

Both L1 and L2 regularization add a penalty to the function that computes errors in a prediction, which is called the cost function. Machine learning algorithms use these cost functions to improve the quality of prediction. In L1 regularization, the penalty is based on magnitude or absolute value of coefficients in a model. In L2 regularization, the penalty is based on the square of the magnitude of the coefficient.

L1 regularization can shrink a feature's coefficient to zero, effectively eliminating the feature from consideration.

5. Model evaluation

There are a variety of ways to understand and evaluate the quality of a machine learning model, including

- Individual evaluation metrics
- K-fold cross validation
- Confusion matrices
- Bias and variance

5.1 Individual Evaluation Metrics

Machine learning models can be evaluated based on a number of metrics, including the following:

Accuracy This is the measure of how often a machine learning algorithm makes a correct prediction, specifically $\text{Accuracy} = (\text{True Positive} + \text{True Positive}) / \text{Total Predicted}$

Precision This is the proportion of positive cases that were correctly identified. This measure is also known as the positive predictive value. The formula is as follows: $\text{Precision} = \text{True Positive} / (\text{True Positive} + \text{False Positive})$

Recall This is the number of actual positive cases that were correctly identified as opposed to negative cases identified as positive. The formula for recall is as follows: $\text{Recall} = \text{True Positive} / (\text{True Positive} + \text{False Negative})$

F1 Score This is a measure that combines precision and recall. The formula for calculating the F measure is as follows: $2 * ((\text{precision} * \text{recall}) / (\text{precision} + \text{recall}))$

There are trade-offs to optimizing precision or recall. The F1 score combines both precision and recall, so it is often used to evaluate models.

5.2 K-Fold Cross Validation

K-fold cross validation is a technique for evaluating model performance by splitting a data set into k segments, where k is an integer. For example, if a dataset were split into five equal-sized subsets, that would be a fivefold cross-validation dataset.

K-fold cross-validation datasets are used as follows. The dataset is shuffled or randomly sorted and split into k groups. One group is held out for evaluation purposes, and the remaining groups are used to train the model. When the model is trained, it is evaluated using the dataset that was held out. The evaluation results are stored, and the process is repeated, holding out another group and training with the remaining groups until all groups have been used for evaluation.

The value of k should be chosen so that each segment is representative of the dataset at large. In practice, setting k to 10 seems to work well in many cases. When k is set equal to the number of records in a dataset, that is called leave-one-out-cross-validation

5.3 Confusion Matrices

Confusion matrices are used with classification models to show the relative performance of a model. In the case of a binary classifier, a confusion matrix would be 2×2 , with one column and one row for each value. One advantage of using a confusion matrix is that it helps to see quickly the accuracy of a model.

5.4 Bias and Variance

The errors made by predictive models can be understood in terms of bias and variance. Data scientists typically run many cycles of training models with different sets and sizes of training datasets. This allows them to understand bias and variance better. Bias is the difference between the average prediction of a model and the correct prediction of a model. Models with high bias tend to have oversimplified representations of the process that generates the training data; this is underfitting the model. When data scientists run training models with different datasets, bias is measured by the difference of the average predicted value from a target value.

Variance is the variability in model predictions. It helps to understand how model prediction differs across multiple training datasets. It does not measure how accurate a prediction is from a target value, but rather the variability of a model to predict. Models with high variance tend to overfit training data so that the model works well when making predictions on the training data, but it does not generalize to data that the model has not seen before.

Ideally, models should have both low bias and low variance, and this is achieved by lowering the mean squared error (MSE) of the model by working through multiple training datasets.

6. Model deployment

Machine learning models are programs not unlike other programs. When they are deployed, they should be managed using the same best practices used for manually coded applications, including version control, testing, and continuous integration and deployment.

7. Model monitoring

Machine learning models should be monitored like other applications. Performance monitoring should be used to understand how well the model is using resources. Stackdriver Monitoring can be used to collect metrics on resource utilization, and Stackdriver Logging can be used to capture information on events that occur while the model is running.

In addition to performance monitoring, models should be monitored for accuracy. This can be done using the model to predict the value of new instances on which the model was not trained. For example, if a model was deployed one week ago and new data has become available since then, and the actual values of the predicted feature are known, then that new data can be used to assess the accuracy of the model.

Keep in mind, however, that even if the model does not change, the accuracy of the model can change if there are changes in the process or the entity being modelled. For example, consider that a model is developed to predict how much a customer will spend on a particular type of product. The model is trained with data collected at some point in time. Now imagine that the company changes its pricing strategy. The data used to train the model does not reflect how customers respond to the new pricing strategy and therefore will make predictions based on the prior strategy. In this case, the model should be updated by training it on the data collected since the new strategy was adopted.

ML pipelines are abstractions that can be implemented in several ways in GCP.

GCP Options for Deploying Machine Learning Pipeline

Data engineers have multiple options for deploying machine learning workflows in GCP, from running custom programs in Compute Engine to using a fully managed machine learning service. In this section, we will look at four options that take advantage of the machine learning capabilities of several different GCP services.

They are as follows:

- Cloud AutoML
- BigQuery ML
- Kubeflow
- Spark Machine Learning

Each option is particularly well suited to a specific workload

Cloud AutoML

Cloud AutoML is a machine learning service designed for developers who want to incorporate machine learning into their applications without having to learn many of the details of ML. The service uses a GUI to train and evaluate models, which reduces the level of effort required to get started building models.

There are several AutoML products, including the following:

- AutoML Vision
- AutoML Video Intelligence (in beta as of this writing)
- AutoML Natural Language
- AutoML Translation
- AutoML Tables (in beta as of this writing)

AutoML Vision has several capabilities:

AutoML Vision Classification -This service enables users to train their own machine learning models to classify images.

Google Cloud Certified Professional Cloud Architect Definitive Guide

AutoML Vision Edge - Image Classification This capability enables users to build custom image classification models that can be deployed to edge devices. This is useful if an application requires local classification and real-time responses to the results of the classification.

AutoML Vision Object Detection - This feature is used to train models to detect objects in images and provide information about those objects, such as their location in the image.

AutoML Vision Edge - Object Detection This feature enables object detection capabilities at the edge of a distributed system that includes cloud and remote or edge processing.

AutoML Video Intelligence Classification can be used to train machine learning models to classify segments of video using a custom set of labels. AutoML Video Intelligence Object Tracking supports the training of machine learning models that can detect and track multiple objects through video segments.

AutoML Natural Language enables developers to deploy machine learning applications that can analyze documents and classify them, identify entities in the text, and determine sentiment or attitudes from text.

AutoML Translation provides developers with the ability to create custom translation models. This is particularly useful if you are developing a translation application for a domain with its own nomenclature, such as a field of science or engineering.

AutoML Tables builds machine learning models based on structured data. AutoML provides tools to clean and analyze datasets, including the ability to detect missing data and determine the distribution of data for each feature. It also performs common feature engineering tasks such as normalizing numeric values, creating buckets for continuous value features, and extracting date and time features from timestamps. AutoML builds models using multiple machine learning algorithms, including linear regression, deep neural network, gradient-boosted decision trees, AdaNet, and ensembles of models generated by a variety of algorithms. This approach allows AutoML Tables to determine the best algorithm for each use case.

Finally, AutoML provides comparable capabilities as BigQueryML.

BigQuery ML

BigQuery ML enables users of the analytical database to build machine learning models using SQL and data in BigQuery datasets. Making machine learning methods available through SQL functions, combined with not having to move data, is a key advantage to using BigQuery ML over other options.

BigQuery ML can be accessed through the following:

- BigQuery web user interface
- bq command-line tool
- BigQuery REST API
- External tools, including Jupyter Notebooks
- BigQuery ML supports several types of machine learning algorithms, including
 - Linear regression for forecasting
 - Binary logistic regression for classification with two labels
 - Multiple logistic regression for classification with more than two labels
 - K-means clustering for segmenting datasets
 - TensorFlow model importing to allow BigQuery users access to custom TensorFlow models

Sometimes, a use case could make use of either AutoML Tables or BigQueryML. AutoML Tables may be a better option when you want to optimize your model, without a lot of experimentation, with different algorithms and feature engineering. If you have many features, AutoML maybe a better option since it automates common feature engineering tasks. AutoML typically takes longer to return a model since it tests a variety of models, so if you need to minimize model generation time, then consider BigQueryML.

Kubeflow

Kubeflow is an open-source project for developing, orchestrating, and deploying scalable and portable machine learning workloads. Kubeflow is designed for the Kubernetes platform.

Kubeflow originally began life as a tool to help run TensorFlow jobs on Kubernetes, but it expanded to a multi-cloud framework for running ML pipelines. Kubeflow can be used to run machine learning workloads in multiple clouds or in a hybrid cloud environment.

Kubeflow includes the following components:

- Support for training TensorFlow models
- TensorFlow Serving, which is used to deploy trained models and make them available to other services
- A JupyterHub installation, which is a platform for spawning and managing multiple instances of a single-user Jupyter Notebook server
- Kubeflow Pipelines, which are used to define machine learning workflows

Kubeflow Pipelines are descriptions of machine learning workflows, including all the components needed to execute a workflow. Pipelines are packaged as Docker images.

Kubeflow is a good choice for running machine learning workloads if you are already working with Kubernetes Engine and have some experience with building machine learning models. Kubeflow supports the scalable use of machine learning models but does not provide some of the features of AutoML or the simplicity of BigQuery ML.

Spark Machine Learning

Cloud Dataproc is a managed Spark and Hadoop service. Included with Spark is a machine learning library called MLib. If you are already using Cloud Dataproc or a self-managed Spark cluster, then using Spark MLib may be a good choice for running machine learning workloads.

Spark MLib contains several tools, including

- Machine learning algorithms for classification, regression, clustering, and collaborative filtering, which are used with recommendation engines
- Support for feature engineering, data transformation, and dimensionality reduction
- ML pipelines for executing multistep workloads
- Other utilities, such as math libraries and other data management tools

Google Cloud Certified Professional Cloud Architect Definitive Guide

Spark MLlib's API supports the chaining together of multiple steps in a machine learning pipeline.

- Pipelines are constructed from several types of components, including
- Data Frames Tabular structures for storing data in memory
- Transformers Applies functions to data frames, including applying a machine learning model to generate predictions
- Estimators Algorithms that are used to apply machine learning algorithms to create models
- Parameters Used by transformers and estimators

Spark MLlib has a wide variety of machine learning algorithms. If you need an algorithm not available in other GCP machine learning services, consider using Spark MLlib.

The available algorithms include the following:

- Support vector machines for classification
- Linear regression for forecasting
- Decision trees, random forests, and gradient-boosted trees for classification
- Naive Bayes for classification
- K-means clustering and streaming k-means for segmenting
- Latent Dirichlet allocation for segmenting
- Singular value decomposition and principal component analysis for dimensionality reduction
- Frequent Pattern (FP)-growth and association rules for frequent pattern mining

Summary

Data engineers have a variety of options for running their machine learning workloads in GCP. Cloud AutoML is designed for developers who want to build on existing machine learning models and tools that automate some machine learning tasks, like feature engineering. BigQuery ML allows SQL users to build models within BigQuery and avoid having to export data and develop models using Python or Java. Kubeflow supports deploying scalable ML pipelines in Kubernetes. Spark MLlib is a comprehensive set of machine learning tools that can be used when deploying Cloud Dataproc clusters.

Cloud AutoML and Cloud BigQuery are readily used by developers and analysts with limited or no experience building machine learning models. Both Kubeflow and Spark MLlib require some knowledge of machine learning techniques and practices.

Measuring, Monitoring & Troubleshooting Machine Learning Models

Three Types of Machine Learning Algorithms

Machine learning algorithms have traditionally been categorized into supervised, unsupervised, and reinforcement learning algorithms

Supervised Learning

Supervised learning algorithms are used to make predictions. When an ML model is used to predict a discrete value, it is known as a classification model. When a model is used to predict a continuous value, it is known as a regression model.

Classification

Supervised algorithms learn from examples. If you wanted to build an ML model to predict whether or not an animal is a mammal, you could use data such as that shown in the table. Supervised learning algorithms use one of the columns of data to represent the value to be predicted. For example, you would use the Mammal column for the predicted value. In ML terminology, this column is called a label

The simplest form of a classifier is called a binary classifier. A classifier that predicts if an animal is a mammal or not a mammal is an example of a binary classifier. More realistic examples of binary classifiers are those used to predict if a credit card transaction is fraudulent or legitimate, and a machine vision ML model that predicts if a medical image contains or does not contain evidence of a malignant tumor.

A multiclass classification model assigns more than two values. In the animal data example, a model that predicts a habitat from the other features is an example of a multiclass classifier. Another example is a classifier that categorizes news stories into one of several categories, such as politics, business, health, and technology.

Google Cloud Certified Professional Cloud Architect Definitive Guide

Some commonly used supervised algorithms are support vector machines (SVMs), decision trees, and logistic regression.

Support vector machines (SVMs) represent instances in a multidimensional space. A simple example is a set of points in a three-dimensional grid. SVMs find a boundary between classes that maximize the distance between those classes. Figure shows an example of a plane separating two groups of instances. If a new point were added to the space, it could be categorized based on which side of the plane it is located.

In general, when features are numerical, you can imagine that each instance is a point in an n-dimensional space, where n is the number of numerical features. Thinking of instances in terms of points located in space makes some useful concepts possible for thinking about machine learning algorithms. A decision boundary is a line, plane, or hyperplane that separates instances into categories.

A decision tree is a type of supervised learning algorithm that builds a model as a series of decisions about features, such as the amount of a transaction or the frequency of a particular word in a news story. For example, Figure 11.2 shows how a decision tree can be used to predict the type of an animal.

Logistic regression is a statistical model based on the logistic function, which produces an S-shaped curve known as a sigmoid and is used for binary classification.

Regression

Supervised learning algorithms can also be used to predict a numeric value. These are called regression algorithms. Consider the data which shows the average number of months that it takes someone to find a job. Regression algorithms could be used with data such as this to create a model for predicting the number of months it will take a person to find a new job.

Regression algorithms map one set of variables to other continuous variables. They are used, for example, to predict the lifetime value of a customer, to estimate the time before a piece of equipment fails given a set of operating parameters, or to predict the value of a stock at some point in the future.

Simple linear regression models use one feature to predict a value, such as using a person's age to predict their height.

Multiple linear regression models use two or more features to predict a value, such as age and gender, to predict a person's height.

Simple nonlinear regression models use a single feature but fit a non-straight line to data. For example, a nonlinear regression model may fit a quadratic curve (one with one bend in the curve). Multiple nonlinear regression models use two or more features and fit a curved line to the data.

Unsupervised Learning

Unsupervised learning algorithms find patterns in data without using predefined labels. Two commonly used forms of unsupervised learning are clustering and anomaly detection.

Unsupervised algorithms learn by starting with unlabeled data and then identifying salient features, such as groups or clusters, and anomalies in a data stream. Unlike with supervised algorithms, there is no one feature or label that is being predicted.

Clustering, or cluster analysis, is the process of grouping instances together based on common features.

K-means clustering is a technique used for partitioning a dataset into k partitions and assigning each instance to one of the partitions. It works by minimizing the variance between the instances in a cluster.

Another clustering algorithm is the K-nearest neighbors' algorithm, which takes as input the k closest instances and assigns a class to an instance based on the most common class of the nearest neighbors.

Anomaly Detection

Anomaly detection is the process of identifying unexpected patterns in data. Anomalies come in a variety of forms. Point anomalies are outliers, such as a credit card transaction with an unusually large amount. Contextual anomalies are unusual patterns given other features of an instance. For example, purchasing a large number of umbrellas in a desert area is contextually anomalous. Collective anomalies are sets of instances that together create an anomaly, such as a large number of credit card transactions from different geographic areas in a short period of time.

Anomaly detection algorithms use density-based techniques, such as k-nearest neighbors, cluster analysis, and outlier detection approaches.

Reinforcement Learning

Reinforcement learning is an approach to learning that uses agents interacting with an environment and adapting behavior based on rewards from the environment. This form of learning does not depend on labels.

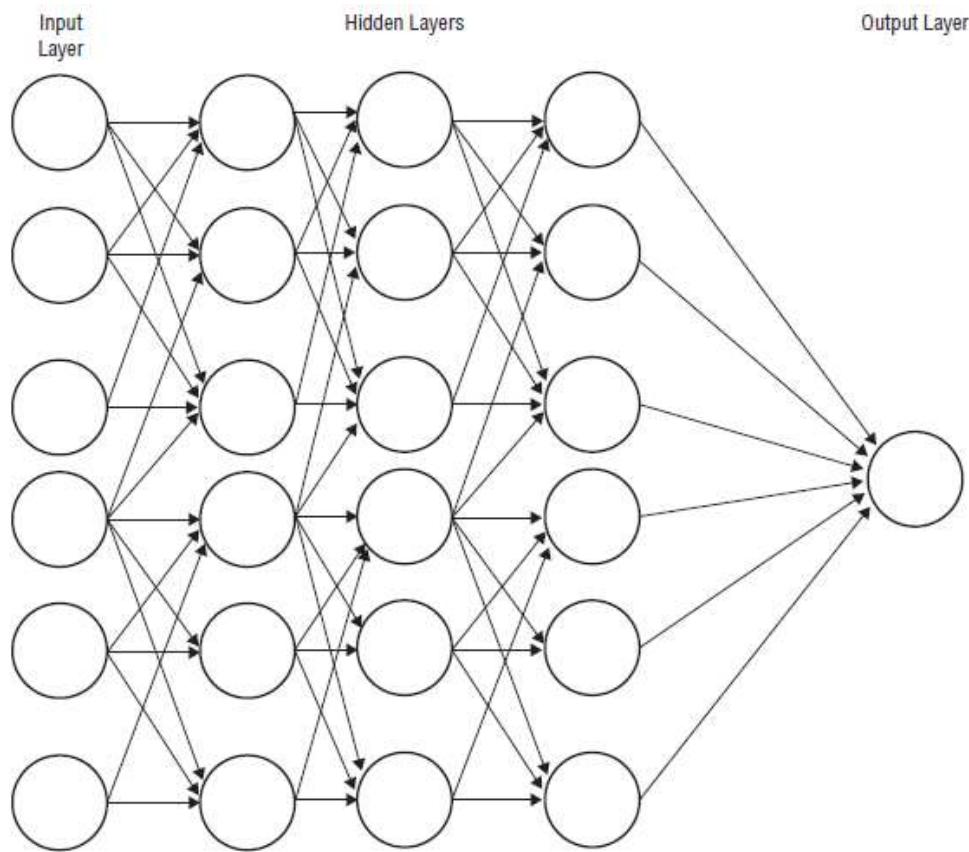
Reinforcement learning is modeled as an environment, a set of agents, a set of actions, and a set of probabilities of transitioning from one state to another after a particular action is taken. A reward is given after the transition from one state to another following an action.

Reinforcement learning is useful when a problem requires a balance between short-term and long-term trade-offs. It is especially useful in robotics and game playing.

Deep Learning

Deep learning uses the concept of an artificial neuron as a building block. These neurons or nodes have one or more inputs and one output. The inputs and outputs are numeric values. The output of one neuron is used as the input of one or more other neurons. A simple, single neuron model is known as a perceptron and is trained using the perceptron algorithm. By adding layers of neurons on top of each other, you can build more complex, deep networks, such as the one shown in Figure below

An example deep learning network



The first layer is the input layer. This is where feature values are input to the network. The middle layers are known as hidden layers. The final layer is the output layer. The output of that node is the output of the model.

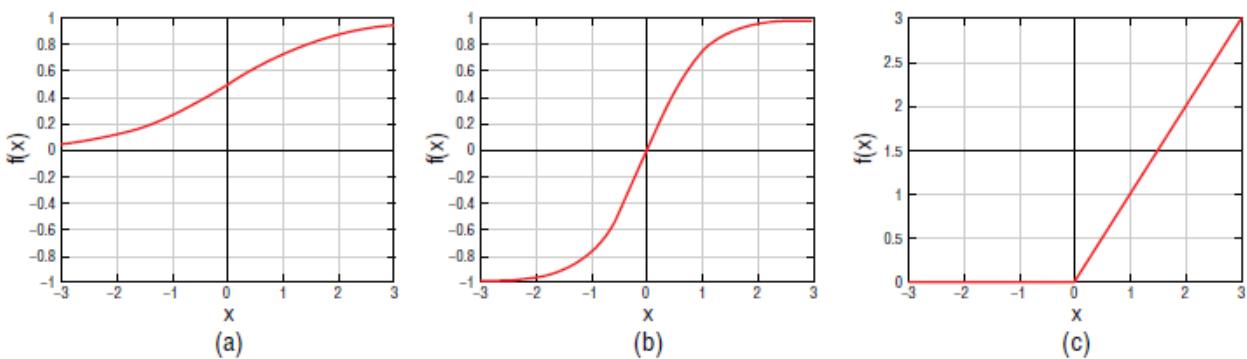
When deep learning models are trained, parameters in each node are adjusted so that as values or signals are sent through the network, the correct value is produced by the output node. One of the most important features of neural networks is that they can learn virtually any mapping function, from inputs to outputs.

There are a few neural network and deep learning-specific terms with which you should be familiar. Each neuron in a neural network has a set of inputs and produces an output.

The function that maps from inputs to outputs is known as an activation function. A commonly used activation function is the rectified linear unit (ReLU) function. When a weighted sum of inputs is

equal to or less than zero, the ReLU outputs a 0. When the weighted sum of inputs is greater than zero, the ReLU outputs the weighted sum.

Other activation functions include the hyperbolic tangent (Tanh) function and the sigmoid function. Both of these are nonlinear functions. If a neural network used linear activation functions, it would be possible to combine all of the linear activation functions into a single linear function, but that is not possible with nonlinear functions. Since neural networks use nonlinear activation functions, they are able to create networks that model nonlinear relationships, and that is one of the reasons why they are so successfully used on problems that do not lend themselves to solutions with linear models. Figure shows graphs of the three activation functions described.



Examples of a (a) sigmoid function, (b) hyperbolic tangent function, and (c) rectilinear unit function

Since deep learning networks can learn virtually any mapping function, it is possible for networks essentially to memorize the training data and the correct responses. This is an extreme example of overfitting. Even if a network does not memorize the training data and labels, they can still overfit the training data. Deep learning models can make use of dropout, which is a form of regularization. During training, a random number of nodes are ignored when calculating weighted sums. This simulates removing a node from a network and, in effect, reduces the network's ability to learn and tends to produce models less likely to overfit the training data.

During training, the error of the prediction is propagated through the layers, and the size of the error propagated decreases with each layer. In deep networks, this can lead the vanishing gradient problem, which is that early layers in the network take much longer to train. Machine learning researchers have investigated many ways to tackle the vanishing gradient problem, and the most common solution is to use ReLU as an activation function.

Leveraging Prebuilt Models as a Service

Google Cloud Platform provides several services that use pretrained machine learning models to help developers build and deploy intelligent services more quickly. The services are broadly grouped into the following categories:

- Sight
- Conversation
- Language
- Structured data

These services are available through APIs or Cloud AutoML services. Cloud AutoML uses the APIs to provide easier-to-use services such as AutoML Vision.

1. Sight

GCP has two APIs for working with sight-related intelligence: Vision AI and Video AI.

There is some similarity between the services. For example, both services provide functions to identify objects and filter content. In addition, the Video Intelligence AI has video-specific features, such as tracking objects across frames and transcribing audio.

Vision AI

The Vision AI API is designed to analyze images and identify text using OCR, to enable the search of images, and to filter explicit images. Images are sent to Vision AI by specifying a URI path to an image or by sending the image data as Base64-encoded text.

There are three options for calling the Vision AI API: Google-supported client libraries, REST, and gRPC. Google recommends using the client libraries, which are available for C#, Go, Java, Node.js, Python, PHP, and Ruby.

For each image sent to the API, the following operations can be performed:

- Detecting text in images
- Detecting handwriting in images
- Detecting text in PDF, TIFF, and other types of files
- Detecting faces

- Detecting hints for suggested vertices for a cropped region of an image
- Detecting image properties
- Detecting landmarks
- Detecting logos
- Detecting multiple objects
- Detecting explicit content (Safe Search)
- Detecting web entities and pages

Video AI

Video AI provides models that can extract metadata; identify key persons, places, and things; and annotate video content. This service has pretrained models that automatically recognize objects in videos.

Specifically, this API can be used to perform the following:

- Identifying objects, locations, activities, animal species, products, and so on
- Detecting shot changes
- Detecting explicit content
- Tracking objects
- Detecting text
- Transcribing videos

Videos are sent to the Video AI API by specifying a URI path to a video or by encoding the image data as a Base64 text and passing it into the content field of the request when using the REST API. The gRPC client can accept binary data directly. Google recommends embedding Base64-encoded files into a variable in code and passing that variable into API function calls.

The Video AI API also includes a service for transcribing audio tracks in a video. That service supports the following:

- Ability to include up to 30 alternative translations for a word, listed in descending order of confidence
- Profanity filtering
- Transcription hints, which are unusual phrases used in the audio that may be otherwise difficult to transcribe

- Audio track detection
- Support for identifying multiple speakers in a video
- Automatically adding punctuation

This API supports several video formats, including MOV, MPEG4, MP4, AVI, and formats that can be decoded from FFmpeg, an open-source suite of libraries supporting cross-platform use of video, audio, and multimedia files.

2. Conversation

Three APIs support conversation services:

- Dialogflow
- Cloud Text-to-Speech API
- Cloud Speech-to-Text API

Dialogflow is used for creating conversational interfaces, whereas the other services can be used for interactive or batch processing that transforms speech to text and text to speech.

Dialogflow

Dialogflow is used for chatbots, interactive voice response (IVR), and other dialogue-based interactions with human speech. This service is based on natural language-understanding technology that is used to identify entities in a conversation and extract numbers, dates, and time, as well as custom entities that can be trained using examples. Dialogflow also provides prebuilt agents that can be used as templates.

A Dialogflow Agent is a virtual agent that participates in a conversation. Conversations can be audio or text based. Agents can be configured with the following:

- Agent settings for language options and behavior control
- Intents for categorizing a speaker, or end user's, intentions
- Entities to identify and extract data from interactions
- Knowledge to parse documents, such as FAQs
- Integrations for applications that process end-user interactions
- Fulfillment to connect the service to integrations

Google Cloud Certified Professional Cloud Architect Definitive Guide

A key component of Dialogflow is intents. Intents categorize a speaker’s intention for a single statement in a conversation. The speaker’s statement is assigned an intent classification.

For example, a question about the weather may be mapped to a forecast intent, which would then be configured to process that statement to extract information, such as the time and location of the desired weather forecast. Intents are structures that include the following:

- Training phrases
- Actions
- Parameters to support extraction
- Responses to provide the speaker with answers

Dialogflow also supports contexts to model natural language contexts. Contexts help match intent. Contexts are used to analyze anaphora, which are words that refer to other entities. For example, in the phrase “they went to the store,” they refer’s to some group of people presumably mentioned earlier in the conversation. Context is used to resolve that reference to the entities mentioned earlier.

Dialogflow is accessible through REST, gRPC, and client libraries. Client libraries are available for C#, Go, Java, Node.js, PHP, Python, and Ruby.

Cloud Text-to-Speech API

GCP’s Cloud Text-to-Speech API maps natural language texts to human-like speech. The API works with more than 30 languages and has more than 180 humanlike voices. The service is based on speech synthesis technology called WaveNet, which is a deep generative model developed by DeepMind.

The API works with plain text or Speech Synthesis Markup Language (SSML) and audio files, including MP3 and WAV files. To generate speech, you call a synthesize function of the API. That function returns a Base64-encoded string that has to be decoded into an audio file. Linux users can use the Base64 command-line utility to perform that conversion.

One of the parameters needed for synthesis is a voice specification. The voices vary by language, gender, and, in some languages, dialects, such as French as spoken in France versus French as spoken

Google Cloud Certified Professional Cloud Architect Definitive Guide

in Canada. Supported languages include Arabic, Czech, Danish, Dutch, English, French, Greek, Hindi, Hungarian, Indonesian, Italian, Korean, Mandarin Chinese, Norwegian, Polish, Portuguese, Swedish, Turkish, and Vietnamese. These are available in standard voice or WaveNet voice, which is higher quality. WaveNet synthesis costs more than standard.

Another synthesis parameter is the device specification. Cloud Text-to-Speech can optimize the generated speech for particular devices, such as wearable devices, headphones, small Bluetooth speakers, and large home entertainment devices.

Cloud Speech-to-Text API

The Cloud Speech-to-Text API is used to convert audio to text. The service is based on deep learning technology and supports 120 languages and variants. The service can be used for transcribing audio files as well as for supporting voice-activated interfaces. Cloud Speech-to-Text automatically detects the language being spoken. This feature is in beta, but it is already available for a large number of languages. Generated text can be returned as a stream of text or in batches as a text file.

The service has several pretrained models that are optimized for particular kinds of speech, including the following:

- Voice commands
- Phone calls
- Video
- Default voice for other domains

Other features of the service include noise handling; automatic punctuation; speaker diarization, which identifies the speaker of each utterance; and the ability to handle streaming audio.

Google has several recommendations for best results. Audio should be captured at a sampling rate of 16,000 Hz or higher. Use a lossless codec, such as FLAC or LINEAR16, for recording and transmitting audio. When recording multiple individuals, use a separate channel for each individual. If the speakers are using a specialized vocabulary, use word and phrase hints to improve accuracy. Finally, for short queries or commands, use the StreamingRecognize function with single_utterance set to true.

Google Cloud Certified Professional Cloud Architect Definitive Guide

The Cloud Speech-to-Text API can be called using client libraries. You can also use the gcloud ml speech command from the command line.

3. Language

GCP has two APIs to support language processing: a translation and an analysis service.

Translation

Google's translation technology is available for use through the Cloud Translation API. The basic version of this service, Translation API Basic, enables the translation of texts between more than 100 languages. An advanced API, Translation API Advanced, is also available that supports customization for domain-specific and context-specific terms and phrases.

Translation API Basic can translate text and HTML content between languages. It can automatically detect languages, so users do not need to specify the source language. The basic API supports REST but not gRPC. Translation API Basic has several client APIs, including ones for C#, Go, Java, Python, Node.js, PHP, and Ruby.

Translation API Advanced has most of the features of Translation API Basic, plus support for glossaries, batch translations, custom models, and a gRPC API.

Natural Language

The Natural Language API uses machine learning-derived models to analyze texts. With this API, developers can extract information about people, places, events, addresses, and numbers, as well as other types of entities. The service can be used to find and label fields within semi-structured documents, such as emails. It also supports sentiment analysis.

The Natural Language API has a set of more than 700 general categories, such as sports and entertainment, for document classification. It can also be combined with other machine learning services, like Speech-to-Text API, to extract information from audio content.

For more advanced users, the service performs syntactic analysis that provides parts of speech labels and creates parse trees for each sentence. Users of the API can specify domain-specific keywords and phrases for entity extraction and custom labels for content classification. It is also

Google Cloud Certified Professional Cloud Architect Definitive Guide

possible to use spatial structure understanding to improve the quality of entity extraction. For example, you may be able to take advantage of the layout of text to improve custom entity extraction.

The API can support working with up to 5,000 classification labels and 1 million documents. Documents may be up to 10 MB in size.

The Natural Language API includes functions to perform a variety of text analysis operations, including the following:

- Identifying entities
- Analyzing sentiment associated with each entity
- Analyzing sentiment of the overall text
- Generating syntactic analysis, including parts-of-speech tags and syntax trees
- Classifying documents into categories

Here are some example high-level content categories:

- Arts and Entertainment
- Autos and Vehicles
- Business and Industrial
- Computer and Electronics
- Food and Drink
- Games
- Health
- People and Society
- Law and Government
- News

Each of these high-level categories have finer-grained categories as well;

4. Structured Data

GCP has three machine learning services for structured data: AutoML Tables, and the Recommendations AI API and Cloud Inference API.

Recommendations AI API

The Recommendations AI API is a service for suggesting products to customers based on their behavior on the user's website and the product catalog of that website. The service builds a recommendation model specific to the site. Recommendations AI is currently in beta.

The product catalog contains information on products that are sold to customers, such as names of products, prices, and availability. End-user behavior is captured in logged events, such as information about what customers search for, which products they view, and which products they have purchased.

The two primary functions of the Recommendations AI API are ingesting data and making predictions. Data is ingested using either the `catalogItems.create` function for individual items or the `catalogItems.import` method for bulk loading. Google recommends providing as much detail as possible and updating the product catalog information as needed to keep the model up to date with the actual product catalog.

Customer activity records also need to be ingested. User events that are useful for the recommendation service include clicking a product, adding an item to a shopping cart, and purchasing a product. Customer events can be recorded using a JavaScript pixel placed on the website to record actions, using the Google Tag Manager to tag and record events, or sending events directly to the Recommendations AI API using the `userEvents.write` method.

In addition to loading data, users of the Recommendations AI API will need to specify some configuration parameters to build a model, including recommendation types.

Google Cloud Certified Professional Cloud Architect Definitive Guide

Recommendation types are one of the following:

- Others you may like: These are additional items that the customer is most likely to purchase.
- Frequently bought together: These are items often purchased during the same session.
- Recommended for you: This predicts the next product with which the customer is likely to engage.
- Recently viewed: This is simply a set of catalog IDs of products with which the customer has recently interacted.

You will also need to specify an optimization objective. There are three such objectives:

- Click-through rate (CTR): This is the default optimization, and it maximizes the likelihood that the user engages the recommendation.
- Revenue per order: This is the default objective for the frequently bought together recommendation type, and it cannot be used with other types.
- Conversion rate: This rate maximizes the likelihood that the user purchases the recommended product.

The Recommendations AI API tracks metrics to help you evaluate the performance of recommendations made. The metrics include the following:

- Total revenue from all recorded purchase events: This is the sum of all revenue from all purchases.
- Recommender-engaged revenue: This is the revenue from purchase events that include at least one recommended item.
- Recommendation revenue: This is the revenue from recommended items.
- Average order value (AOV): This is the average of orders from all purchase events.
- Recommender-engaged AOV: This is the average value of orders that include at least one recommended item.
- Click-through rate: This is the number of product views from a recommendation.
- Conversion rate: This is the number of times that an item was added to a cart divided by the total number of recommendations.
- Revenue from recommendations: This is the total revenue from all recommendations.

The Recommendations AI API is tailored for interacting with customers on an e-commerce site. The Cloud Inference API is a machine learning service designed to help analyze time-series data.

Cloud Inference API

Many activities and operations of interest to business can be captured in time-series data. This can include tracking the number of customers visiting a website or physical store in some time period, collecting sensor measurements from manufacturing machinery, and collecting telemetry data from fleet vehicles. The Cloud Inference API provides real-time analysis of time-series data. The Cloud Inference API is currently in alpha.

The Cloud Inference API provides for processing time-series datasets, including ingesting from JSON formats, removing data, and listing active datasets. It also supports inference queries over datasets, including correlation queries, variation in frequency over time, and probability of events given evidence of those events in the dataset.

Machine Learning – Interview | Exam Tips

Section 1 Deploying Machine Learning Pipelines

Stages of ML pipelines

Data ingestion, data preparation, data segregation, model training, model evaluation, model deployment, and model monitoring are the stages of ML pipelines. Although the stages are listed in a linear manner, ML pipelines are more cyclic than linear, especially relating to training and evaluation.

Batch and streaming ingestion

Batch data ingestion should use a dedicated process for ingesting each distinct data source. Batch ingestion often occurs on a relatively fixed schedule, much like many data warehouse ETL processes. It is important to be able to track which batch data comes from, so include a batch identifier with each record that is ingested. Cloud Pub/Sub is designed for scalable messaging, including ingesting streaming data. Cloud Pub/Sub is a good option for ingesting streaming data that will be stored in a database, such as Bigtable or Cloud Firebase, or immediately consumed by machine learning processes running in Cloud Dataflow, Cloud Dataproc, Kubernetes Engine, or Compute Engine. When using BigQuery, you have the option of using streaming inserts.

Three kinds of data preparation

The three kinds of data preparation are [data exploration](#), [data transformation](#), and [feature engineering](#). Data exploration is the first step in working with a new data source or a data source that has had significant changes. The goal of this stage is to understand the distribution of data and the overall quality of data. Data transformation is the process of mapping data from its raw form into data structures and formats that allow for machine learning. Transformations can include replacing missing values with a default value, changing the format of numeric values, and deduplicating records. Feature engineering is the process of adding or modifying the representation of features to make implicit patterns more explicit. For example, if a ratio of two numeric features is important to classifying an instance, then calculating that ratio and including it as a feature may improve the model quality. Feature engineering includes the understanding of key attributes (features) that are meaningful for machine learning objectives at hand. This includes dimensional reduction.

Data Segregation

The data segregation is the process splitting a dataset into three segments: training, validation, and test data. Training data is used to build machine learning models. Validation data is used during hyperparameter tuning. Test data is used to evaluate model performance. The main criteria for deciding how to split data are to ensure that the test and validation datasets are large enough to produce statistically meaningful results, that test and validation datasets are representative of the data as a whole, and that the training dataset is large enough for the model to learn to make accurate predictions with reasonable precision and recall.

Training a model

Know that feature selection is the process of evaluating how a particular attribute or feature contributes to the predictiveness of a model. The goal is to have features of a dataset that allow a model to learn to make accurate predictions. Know that underfitting creates a model that is not able to predict values of training data correctly or new data that was not used during training.

Underfitting, overfitting, and regularization

The problem of underfitting may be corrected by increasing the amount of training data, using a different machine learning algorithm, or modifying hyperparameters. Understand that overfitting occurs when a model fits the training data too well. One way to compensate for the impact of noise in the data and reduce the risk of overfitting is by introducing a penalty for data points, which makes the model more complicated. This process is called regularization. Two kinds of regularization are L1 regularization, which is also known as Lasso Regularization, for Least Absolute Shrinkage and Selection Operator, and L2 or Ridge Regression.

Various ways to evaluate a model

Methods for evaluation a model include individual evaluation metrics, such as accuracy, precision, recall, and the F measure; k-fold cross-validation; confusion matrices; and bias and variance. K-fold cross-validation is a technique for evaluating model performance by splitting a data set into k segments, where k is an integer. Confusion matrices are used with classification models to show the relative performance of a model. In the case of a binary classifier, a confusion matrix would be 2×2 , with one column and one row for each value.

Understand bias and variance

Bias is the difference between the average prediction of a model and the correct prediction of a model. Models with high bias tend to have oversimplified models; this is underfitting the model. Variance is the variability in model predictions. Models with high variance tend to overfit training data so that the model works well when making predictions on the training data but does not generalize to data that the model has not seen before.

Options for deploying machine learning workloads on GCP

These options include Cloud AutoML, BigQuery ML, Kubeflow, and Spark MLib. Cloud AutoML is a machine learning service designed for developers who want to incorporate machine learning in their applications without having to learn many of the details of ML. BigQuery ML enables users of the analytical database to build machine learning models using SQL and data in BigQuery datasets. Kubeflow is an open-source project for developing, orchestrating, and deploying scalable and portable machine learning workloads. Kubeflow is designed for the Kubernetes platform. Cloud Dataproc is a managed Spark and Hadoop service. Included with Spark is a machine learning library called MLib, and it is a good option for machine learning workloads if you are already using Spark or need one of the more specialized algorithms included in Spark MLib.

Section 2 Measuring, Monitoring, and Troubleshooting Machine Learning Models

Three types of machine learning algorithms

supervised, unsupervised, and reinforcement learning. Supervised algorithms learn from labeled examples. Unsupervised learning starts with unlabeled data and identifies salient features, such as groups or clusters, and anomalies in a data stream. Reinforcement learning is a third type of machine learning algorithm that is distinct from supervised and unsupervised learning. It trains a model by interacting with its environment and receiving feedback on the decisions that it makes.

Supervised learning is used for classification and regression

Classification models assign discrete values to instances. The simplest form is a binary classifier that assigns one of two values, such as fraudulent/not fraudulent, or has malignant tumor/does not have malignant tumor. Multiclass classification models assign more than two values. Regression models map continuous variables to other continuous variables.

How unsupervised learning differs from supervised learning

Unsupervised learning algorithms find patterns in data without using predefined labels. Three types of unsupervised learning are clustering, anomaly detection, and collaborative filtering. Clustering, or cluster analysis, is the process of grouping instances together based on common features. Anomaly detection is the process of identifying unexpected patterns in data.

How reinforcement learning differs from supervised and unsupervised techniques

Reinforcement learning is an approach to learning that uses agents interacting with an environment and adapting behavior based on rewards from the environment. This form of learning does not depend on labels. Reinforcement learning is modeled as an environment, a set of agents, a set of actions, and a set of probabilities of transitioning from one state to another after a particular action is taken. A reward is given after the transition from one state to another following an action.

Structure of neural networks, particularly deep learning networks

Neural networks are systems roughly modeled after neurons in animal brains and consist of sets of connected artificial neurons or nodes. The network is composed of artificial neurons that are linked together into a network. The links between artificial neurons are called connections. A single neuron is limited in what it can learn. A multilayer network, however, is able to learn more functions. A multilayer neural network consists of a set of input nodes, hidden nodes, and an output layer.

Machine learning terminology

This includes general machine learning terminology, such as baseline and batches; feature terminology, such as feature engineering and bucketing; training terminology, such as gradient descent and backpropagation; and neural network and deep learning terms, such as activation function and dropout. Finally, know model evaluation terminology such as precision and recall.

Common sources of errors, including data-quality errors, unbalanced training sets & bias

Poor-quality data leads to poor models. Some common data-quality problems are missing data, invalid values, inconsistent use of codes and categories, and data that is not representative of the population at large. Unbalanced datasets are ones that have significantly more instances of some categories than of others. There are several forms of bias, including automation bias, reporting bias, and group attribution.

Section 3 Leveraging Prebuilt Models as a Service

Functionality of the Vision AI API

The Vision AI API is designed to analyze images and identify text, enable the search of images, and filter explicit images. Images are sent to the Vision AI API by specifying a URI path to an image or by sending the image data as Base64-encoded text. There are three options for calling the Vision AI API: Google-supported client libraries, REST, and gRPC.

Functionality of the Video Intelligence API

The Video Intelligence API provides models that can extract metadata; identify key persons, places, and things; and annotate video content. This service has pretrained models that automatically recognize objects in videos. Specifically, this API can be used to identify objects, locations, activities, animal species, products, and so on, and detect shot changes, detect explicit content, track objects, detect text, and transcribe videos.

Functionality of Dialogflow

Dialogflow is used for chatbots, interactive voice response (IVR), and other dialogue-based interactions with human speech. The service is based on natural language-understanding technology that is used to identify entities in a conversation and extract numbers, dates, and time, as well as custom entities that can be trained using examples. Dialogflow also provides prebuilt agents that can be used as templates.

Functionality of the Cloud Text-to-Speech API

GCP's Cloud Text-to-Speech API maps natural language texts to human-like speech. The API works with more than 30 languages and has more than 180 humanlike voices. The API works with plaintext or Speech Synthesis Markup Language (SSML) and audio files, including MP3 and WAV files. To generate speech, you call a synthesize function of the API.

Functionality of the Cloud Speech-to-Text API

The Cloud Speech-to-Text API is used to convert audio to text. This service is based on deep learning technology and supports 120 languages and variants. The service can be used for transcribing audios as well as for supporting voice-activated interfaces. Cloud Speech-to-Text automatically detects the language being spoken. Generated text can be returned as a stream of text or in batches as a text file.

Functionality of the Cloud Translation API

Google's translation technology is available for use through the Cloud Translation API. The basic version of this service, Translation API Basic, enables the translation of texts between more than 100 languages. There is also an advanced API, Translation API Advanced, which supports customization for domain-specific and context-specific terms and phrases.

Functionality of the Natural Language API

The Natural Language API uses machine learning-derived models to analyze texts. With this API, developers can extract information about people, places, events, addresses, and numbers, as well as other types of entities. The service can be used to find and label fields within semi-structured documents, such as emails. It also supports sentiment analysis. The Natural Language API has a set of more than 700 general categories, such as sports and entertainment, for document classification. For more advanced users, the service performs syntactic analysis that provides parts of speech labels and creates parse trees for each sentence. Users of the API can specify domain-specific keywords and phrases for entity extraction and custom labels for content classification.

Functionality of the Recommendations AI API

The Recommendations AI API is a service for suggesting products to customers based on their behavior on the user's website and the product catalog of that website. The service builds a recommendation model specific to the site. The product catalog contains information on products that are sold to customers, such as names of products, prices, and availability. End-user behavior is captured in logged events, such as information about what customers search for, which products they view, and which products they have purchased. There are two primary functions the Recommendations AI API: ingesting data and making predictions.

Functionality of the Cloud Inference API

The Cloud Inference API provides real-time analysis of time-series data. The Cloud Inference API provides for processing time-series datasets, including ingesting from JSON formats, removing data, and listing active datasets. It also supports inference queries over datasets, including correlation queries, variation in frequency over time, and probability of events given evidence of those events in the dataset.

Machine Learning Quiz

Section 1 - Deploying Machine Learning Pipelines

1. You have been tasked with helping to establish ML pipelines for your department. The models will be trained using data from several sources, including several enterprise transaction processing systems and third-party data provider datasets. Data will arrive in batches. Although you know the structure of the data now, you expect that it will change, and you will not know about the changes until the data arrives. You want to ensure that your ingestion process can store the data in whatever structure it arrives in. After the data is ingested, you will transform it as needed. What storage system would you use for batch ingestion?

- A. Cloud Storage
- B. Cloud Spanner
- C. Cloud Dataprep
- D. Cloud Pub/Sub

Correct Answer: A

2. A startup company is building an anomaly detection service for manufacturing equipment. IoT sensors on manufacturing equipment transmit data on machine performance every 30 seconds. The service will initially support up to 5,000 sensors, but it will eventually grow to millions of sensors. The data will be stored in Cloud Bigtable after it is pre-processed and transformed by a Cloud Dataflow workflow. What service should be used to ingest the IoT data?

- A. Cloud Storage
- B. Cloud Bigtable
- C. BigQuery Streaming Insert
- D. Cloud Pub/Sub**

Correct Answer: D

3. A machine learning engineer has just started a new project. The engineer will be building a recommendation engine for an e-commerce site. Data from several services will be used, including data about products, customers, and inventory. The data is currently available in a data lake and stored in its raw, unprocessed form. What is the first thing you would recommend the machine learning engineer do to start work on the project?

- A. Ingest the data into Cloud Storage
- B. Explore the data with Cloud Dataprep**
- C. Transform the data with Cloud Dataflow
- D. Transform the data with BigQuery

Correct Answer: B

4. A machine learning engineer is in the process of building a model for classifying fraudulent transactions. They are using a neural network and need to decide how many nodes and layers to use in the model. They are experimenting with several different combinations of number of nodes and number of layers. What data should they use to evaluate the quality of models being developed with each combination of settings?

- A. Training data
- B. Validation data**
- C. Test data
- D. Hyperparameter data

Correct Answer: B

5. A developer with limited knowledge of machine learning is attempting to build a machine learning model. The developer is using data collected from a data lake with minimal data preparation. After models are built, they are evaluated. Model performance is poor. The developer has asked for your help to reduce the time needed to train the model and increase the quality of model predictions. What would you do first with the developer?

- A. Explore the data with the goal of feature engineering**
- B. Create visualizations of accuracy, precision, recall, and F measures
- C. Use tenfold cross-validation
- D. Tune hyperparameters

Correct Answer: A

6. A developer has built a machine learning model to predict the category of new stories. The possible values are politics, economics, business, health, science, and local news. The developer has tried several algorithms, but the model accuracy is poor even when evaluating the model on using the training data. This is an example of what kind of potential problem with a machine learning model?

- A. Overfitting
- B. Underfitting**
- C. Too much training data
- D. Using tenfold cross-validation for evaluation

Correct Answer: B

7. A developer has built a machine learning model to predict the category of new stories. The possible values are politics, economics, business, health, science, and local news. The developer has tried several algorithms, but the model accuracy is quite high when evaluating the model using the training data but quite low when evaluating using test data. What would you recommend to correct this problem?

- A. Use confusion matrices for evaluation
- B. Use L1 or L2 regularization when evaluating
- C. Use L1 or L2 regularization when training**
- D. Tune the hyperparameters more

Correct Answer: C

8. Your e-commerce company deployed a product recommendation system six months ago. The system uses a machine learning model trained using historical sales data from the previous year. The model performed well initially. When customers were shown product recommendations, the average sale value increased by 14 percent. In the past month, the model has generated an average increase of only 2 percent. The model has not changed since it was deployed six months ago. What could be the cause of the decrease in effectiveness, and what would you recommend to correct it?

- A. The model is overfitting—use regularization.

- B. The data used to train the model is no longer representative of current sales data, and the model should be retrained with more recent data.
- C. The model should be monitored to collect performance metrics to identify the root cause of the decreasing effectiveness of the model.
- D. The model is underfitting—train with more data.

Correct Answer: B

9. A startup company is developing software to analyze images of traffic in order to understand congestion patterns better and how to avoid them. The software will analyze images that are taken every minute from hundreds of locations in a city. The software will need to identify cars, trucks, cyclists, pedestrians, and buildings. The data on object identities will be used by analysis algorithms to detect daily patterns, which will then be used by traffic engineers to devise new traffic flow patterns. What GCP service would you use for this?

- A. AutoML Vision Object Detection
- B. AutoML Vision Edge - Object Detection
- C. AutoML Video Intelligence Classification
- D. Auto ML Video Intelligence Object Tracking

Correct Answer: A

10. An analyst would like to build a machine learning model to classify rows of data in a dataset. There are two categories into which the rows can be grouped: Type A and Type B. The dataset has over 1 million rows, and each row has 32 attributes or features. The analyst does not know which features are important. A labelled training set is available with a sufficient number of rows to train a model. The analyst would like the most accurate model possible with the least amount of effort on the analyst's part. What would you recommend?

- A. Kubeflow
- B. Spark MLlib
- C. AutoML Tables
- D. AutoML Natural Language

Correct Answer: C

11. The chief financial officer of your company would like to build a program to predict which customers will likely be late paying their bills. The company has an enterprise data warehouse in BigQuery containing all the data related to customers, billing, and payments. The company does not have anyone with machine learning experience, but it does have analysts and data scientists experienced in SQL, Python, and Java. The analysts and data scientists will generate and test a large number of models, so they prefer fast model building. What service would you recommend using to build the model?

- A. Kubeflow
- B. Spark MLlib
- C. BigQuery ML**
- D. AutoML Tables

Correct Answer: C

12. A team of researchers is analyzing buying patterns of customers of a national grocery store chain. They are especially interested in sets of products that customers frequently buy together. The researchers plan to use association rules for this frequent pattern mining. What machine learning option in GCP would you recommend?

- A. Cloud Dataflow
- B. Spark MLlib**
- C. BigQuery ML
- D. AutoML Tables

Correct Answer: B

Section 2 - Measuring, Monitoring, and Troubleshooting

1. You are building a machine learning model to predict the sales price of houses. You have 7 years of historical data, including 18 features of houses and their sales price. What type of machine learning algorithm would you use?

- A. Classifier
- B. Regression**
- C. Decision trees
- D. Reinforcement learning

Correct Answer: B

2. You have been asked to build a machine learning model that will predict if a news article is a story about technology or another topic. Which of the following would you use?

- A. Logistic regression**
- B. K-means clustering
- C. Simple linear regression
- D. Multiple linear regression

Correct Answer: A

3. A startup is collecting IoT data from sensors placed on manufacturing equipment. The sensors send data every five seconds. The data includes a machine identifier, a timestamp, and several numeric values. The startup is developing a model to identify unusual readings. What type of unsupervised learning technique would they use?

- A. Clustering
- B. K-means
- C. Anomaly detection**
- D. Reinforcement learning

Correct Answer: C

4. You want to study deep learning and decide to start with the basics. You build a binary classifier using an artificial neuron. What algorithm would you use to train it?

- A. Perceptron
- B. SVM
- C. Decision tree
- D. Linear regression

Correct Answer: A

5. A group of machine learning engineers has been assigned the task of building a machine learning model to predict the price of gold on the open market. Many features could be used, and the engineers believe that the optimal model will be complex. They want to understand the minimum predictive value of a model that they can build from the data that they have.

What would they build?

- A. Multiclass classifier
- B. K clusters
- C. Baseline model**
- D. Binary classifier

Correct Answer: C

6. You are preparing a dataset to build a classifier. The data includes several continuous values, each in the range 0.00 to 100.00. You'd like to have a discrete feature derive each continuous value. What type of feature engineering would you use?

- A. Bucketing**
- B. Dimension reduction
- C. Principal component analysis
- D. Gradient descent

Correct Answer: A

7. You have been tasked with developing a classification model. You have reviewed the data that you will use for training and testing and realize that there are a number of outliers that you think might lead to overfitting. What technique would you use to reduce the impact of those outliers on the model?

- A. Gradient descent
- B. Large number of epochs
- C. L2 regularization**
- D. Backpropagation

Correct Answer: C

8. You have built a deep learning neural network that has 8 layers, and each layer has 100 fully connected nodes. The model fits the training data quite well with an F1 score of 98 out of 100. The model performs poorly when the test data is used, resulting in an F1 score of 62 out of 100. What technique would you use to try to improve performance of this model?

- A. User more epochs
- B. Dropout**
- C. Add more layers
- D. ReLU

Correct Answer: B

9. Your team is building a classifier to identify counterfeit products on an e-commerce site. Most of the products on the site are legitimate, and only about 3 percent of the products are counterfeit. You are concerned that, as is, the dataset will lead to a model that always predicts that products are legitimate. Which of the following techniques could you use to prevent this?

- A. Undersampling**
- B. Dropout
- C. L1 regularization
- D. AUC

Correct Answer: A

10. You are reviewing a dataset and find that the data is relatively high quality. There are no missing values and only a few outliers. You build a model based on the dataset that has high accuracy, precision, and recall when applied to the test data. When you use the model in production, however, it renders poor results. What might have caused this condition?

- A. Applying L1 regularization
- B. Applying dropout
- C. Reporting bias**
- D. Automation bias

Correct Answer: C

Section 3 - Leveraging Prebuilt Models as a Service

1. You are building a machine learning model to analyze unusual events in traffic through urban areas. Your model needs to distinguish cars, bikes, pedestrians, and buildings. It is especially important that the model be able to identify and track moving vehicles. Video will be streamed to your service from cameras mounted on traffic lights. What GCP service would you use for the object analysis and tracking?

- A. Cloud Video Intelligence API
- B. Cloud Vision API
- C. Cloud Inference API
- D. Cloud Dataflow

Correct Answer: A

2. A startup is building an educational support platform for students from ages 5-18. The platform will allow teachers to post assignments and conduct assessments. Students will be able to upload content, including text and images. The founder of the startup wants to make sure that explicit images are not uploaded. What GCP service would you use?

- A. Cloud Video Intelligence API
- B. Cloud Vision API**
- C. Cloud Inference API
- D. Cloud Dataprep

Correct Answer: B

3. You are using the Cloud Vision API to detect landmarks in images. You are using the batch processing with asynchronous requests. The source images for each batch is in a separate Cloud Storage bucket. There are between 1 and 5,000 images in each bucket. Each batch request processes one bucket. All buckets have the same access controls. Sometimes, the operations succeed and sometimes they fail. What could be the cause of the errors?

- A. Cloud Video Intelligence API
- B. Some buckets have more than 2,000 images.**
- C. There is an issue with IAM settings.

D. Images have to be uploaded directly from a device, not a Cloud Storage bucket.

Correct Answer: B

4. Your team is building a chatbot to support customer support. Domain experts from the customer support team have identified several kinds of questions that the system should support, including questions about returning products, getting technical help, and asking for product recommendations. You will use Dialogflow to implement the chatbot. What component of Dialogflow will you configure to support the three question types?

- A. Entities
- B. Fulfillments
- C. Integrations
- D. Intents**

Correct Answer: D

5. A developer asks for your help tuning a text-to-speech service that is used with a health and wellness app. The app is designed to run on watches and other personal devices. The sound quality is not as good as the developer would like. What would you suggest trying to improve the quality of sound?

- A. Change the device specification to optimize for a wearable device
- B. Change from standard to WaveNet-quality voice
- C. Encode the text in Base64**
- D. Options A and B
- E. Options A, B, and C

Correct Answer: C

6. A developer asks for your help tuning a speech-to-text service that is used to transcribe text recorded on a mobile device. The quality of the transcription is not as good as expected. The app uses LINEAR16 encoding and a sampling rate of 12,000 Hz. What would you suggest to try to improve the quality?

- A. Use WaveNet option
- B. Increase the sampling rate to at least 16,000 Hz**

C. Use Speech Synthesis Markup Language to configure conversion parameters

D. Options A and B

Correct Answer: B

7. You have developed a mobile app that helps travellers quickly find sites of interest. The app uses the GCP Translation service. The initial release of the app used the REST API, but adoption has grown so much that you need higher performance from the API and plan to use gRPC instead. What changes do you need to make to the way that you use the Translation service?

A. Use the WaveNet option

B. Use Translation API Basic

C. Use Translation API Advanced

D. Option A or B

Correct Answer: C

8. You are experimenting with the GCP Translation API. You have created a Jupyter Notebook and plan to use Python 3 to build a proof-of-concept system. What are the first two operations that you would execute in your notebook to start using the Translation API?

A. Import Translation libraries and create a translation client

B. Create a translation client and encode text in UTF-8

C. Create a translation client, and set a variable to TRANSLATE to pass in as a parameter to the API function call

D. Import Translation libraries, and set a variable to TRANSLATE to pass in as a parameter to the API function call

Correct Answer: A

9. You have been hired by a law firm to help analyze a large volume of documents related to a legal case. There are approximately 10,000 documents ranging from 1 to 15 pages in length. They are all written in English. The lawyers hiring you want to understand who is mentioned in each document so that they can understand how those individuals worked together. What functionality of the Natural Language API would you use?

A. Identifying entities

- B. Analyzing sentiment associated with each entity
- C. Analyzing sentiment of the overall text
- D. Generating syntactic analysis

Correct Answer: A

10. As a founder of an e-commerce startup, you are particularly interested in engaging with your customers. You decide to use the GCP Recommendations AI API using the “others you may like” recommendation type. You want to maximize the likelihood that users will engage with your recommendations. What optimization objective would you choose?

- A. Click-through rate (CTR)**
- B. Revenue per order
- C. Conversation rate
- D. Total revenue

Correct Answer: A

11. Your e-commerce startup has been growing rapidly since its launch six months ago. You are starting to notice that the rate of revenue growth is slowing down. Your board of directors is asking you to develop a strategy to increase revenue. You decide to personalize each customer’s experience. One of the ways in which you plan to implement your strategy is by showing customers products that they are likely to interact with next. What recommendation type would you use?

- A. Others you may like
- B. Frequently bought together
- C. Recommended for you**
- D. Recently viewed

Correct Answer: C

12. You work for an enterprise with a large fleet of vehicles. The vehicles are equipped with several sensors that transmit data about fuel utilization, speed, and other equipment operating characteristics. The chief of operations has asked you to investigate the feasibility of building a predictive maintenance application that can help identify breakdowns before they occur. You decide to prototype an anomaly detection model as a first step. You want to build this as quickly as possible, so you decide to use a machine learning service. Which GCP service would you use?

- A. Cloud Inference API
- B. AutoML Tables
- C. AutoML Vision
- D. Cloud Anomaly Detection API

Correct Answer: A

Management Tools

What are the GCP Management Tools?

- Monitoring with Stackdriver
- Deployment Manager
- Cloud Shell & Cloud SDK

Monitoring with Stackdriver

What is Monitoring?

Monitoring is the practice of collecting measurements of key aspects of infrastructure and applications. Examples include average CPU utilization over the past minute, the number of bytes written to a network interface, and the maximum memory utilization over the past hour. These measurements, which are known as *metrics*, are made repeatedly over time and constitute a time series of measurements.

Metrics

Metrics have a particular pattern that includes some kind of property of an entity, a time range, and a numeric value. GCP has defined metrics for a wide range of entities, including the following:

- GCP services, such as BigQuery, Cloud Storage, and Compute Engine.
- Operating system and application metrics which are collected by Stackdriver agents that run on VMs.
- Anthos, which includes metrics include Kubernetes and Istio metrics
- AWS metrics that measure performance of Amazon Web Service resources, such as EC2 instances.
- External metrics including metrics defined in Prometheus, a popular open-source monitoring tool.

In addition to the metric name, value, and time range, metrics can have labels associated with them. This is useful when querying or filtering resources that you are interested in monitoring.

Time Series

A time series is a set of metrics recorded with a time stamp. Often, metrics are collected at a specific interval, such as every second or every minute. A time series is associated with a monitored entity. Stackdriver Monitoring provides an API for working with time-series metrics.

The API supports the following: -

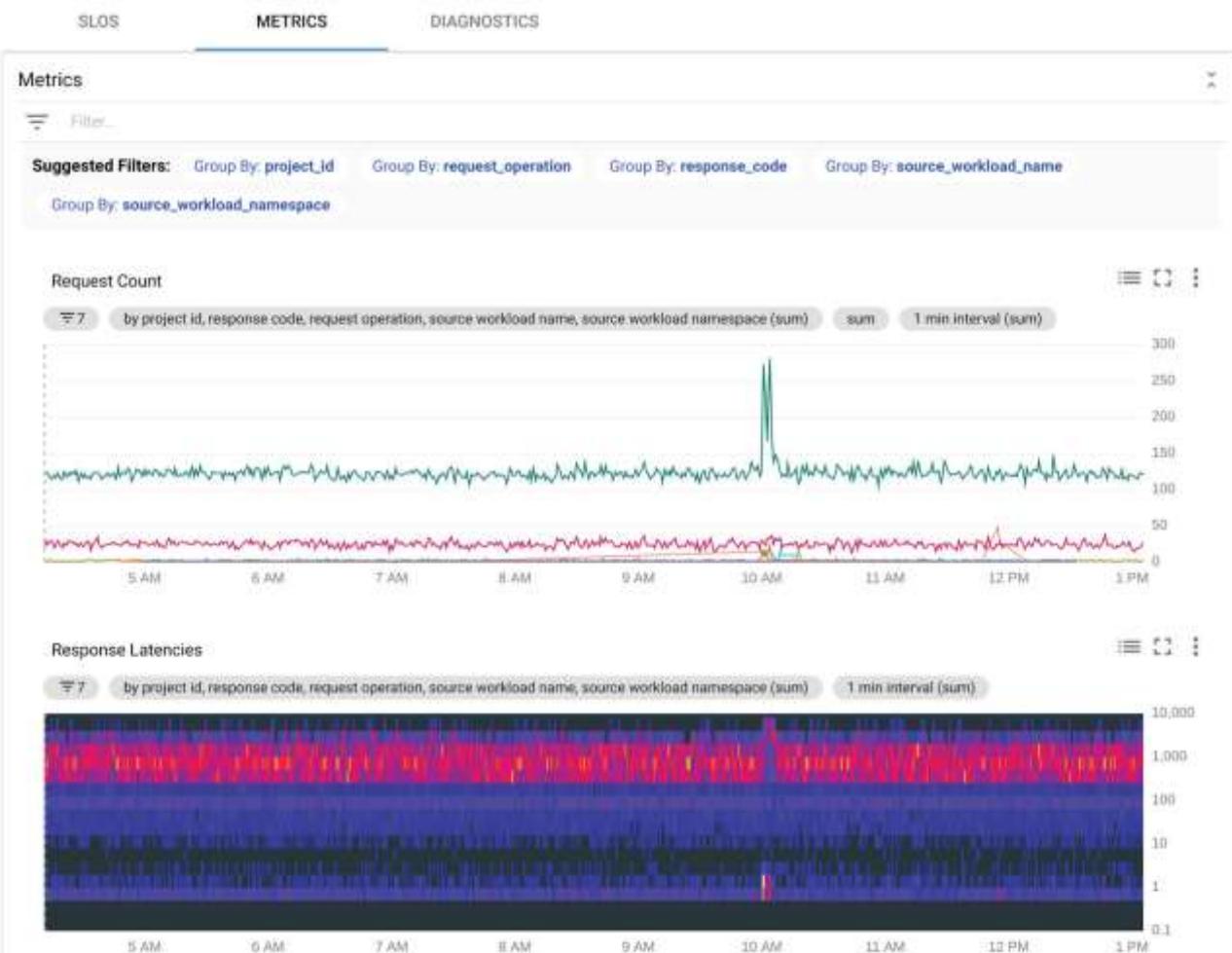
- Retrieving time series in a project, based on metric name, resource properties, and other attributes

- Grouping resources based on properties
- Listing group members
- Listing metric descriptors
- Listing monitored entities descriptors

Common ways of working with metrics are using dashboards and alerting.

Dashboards

Dashboards are visual displays of time series. For example, Below Figure shows a simple dashboard with a history of request count and latency of a service.



Dashboards are customized by users to show data that helps monitor and meet service-level objectives or diagnose problems with a particular service. Dashboards are especially useful for determining correlated failures. For example, one dashboard may indicate that you are not meeting your service-level objective of response time for purchase transactions.

You might then switch to a dashboard that shows performance metrics for your application server, database, and cache. From there, you could determine which of the main three components of the systems are having performance issues. If, for example, the cache hit rate is unusually low, that can lead to greater IO load on the database, which can increase the latency of purchase transactions. Crafting informative dashboards is often an iterative process. You may know key performance indicators that you should monitor when you first create a service, but over time you may monitor additional metrics. For example, if you had an incident because of a lag in a messaging service, you may want to monitor several metrics of the messaging service to catch potential problems before they occur.

Of course, watching a dashboard for potential problems is not the most efficient use of DevOps engineers' time. Reliable systems also depend on alerting and automation to notify engineers when a problem arises and needs their attention.

Stackdriver

Stackdriver is GCP's tool for monitoring, logging, and diagnostics. Stackdriver gives you access to many different kinds of signals from your infrastructure platforms, virtual machines, containers, middleware, and application tier: logs, metrics, traces. It gives your insight into your application's health, performance, and availability, so if issues occur you can fix them faster.

Stackdriver offers capabilities in six areas

Monitoring



Platform, system, and application metrics

Uptime/health checks

Dashboards and alerts

Error Reporting



Error notifications

Error dashboard

Logging



Platform, system, and application logs

Log search, view, filter, and export

Log-based metrics

Debugger



Debug applications

Trace



Latency reporting and sampling

Per-URL latency and statistics

Profiler



Continuous profiling of CPU and memory consumption

Google Cloud Certified Professional Cloud Architect Definitive Guide

The core components of Google Stackdriver: monitoring, logging, trace, error reporting, and debugging.

Stackdriver Monitoring checks the endpoints of web applications and other internet-accessible services running on your cloud environment. You can configure uptime checks associated with URLs, groups, or resources, such as instances and load balancers. You can set up alerts on interesting criteria, like when health check results or uptimes fall into levels that need action. You can use Monitoring with a lot of popular notification tools. And you can create dashboards to help you visualize the state of your application.

Stackdriver Logging lets you view logs from your applications, and filter and search on them. Logging also lets you define metrics based on log contents that are incorporated into dashboards and alerts. You can also export logs to BigQuery, Cloud Storage, and Cloud Pub/Sub.

Stackdriver Error Reporting tracks and groups the errors in your cloud applications, and it notifies you when new errors are detected.

With **Stackdriver Trace**, you can sample the latency of App Engine applications and report per-URL statistics.

Stackdriver Debugger works best when your application's source code is available, such as in Cloud Source Repositories, although it can be in other repositories too.

Stackdriver Profiler is a statistical, low-overhead profiler that continuously gathers CPU usage and memory-allocation information from your production applications. It attributes that information to the application's source code, helping you identify the parts of the application consuming the most resources, and otherwise illuminating the performance characteristics of the code.

Deployment Manager

Highlights

- Infrastructure management service
- Create a yaml template describing your environment and use Deployment Manager to
- create resources
- Provides repeatable deployments

What is Deployment Manager?

Deployment Manager is an infrastructure management service that automates the creation and management of your Google Cloud Platform resources for you.

Deployment Manager service that allows you to specify infrastructure as code. It is a good practice to define infrastructure as code, since it allows teams to reproduce environments rapidly. It also lends itself to code reviews, version control, and other software engineering practices. Deployment Manager uses declarative templates that describe what should be deployed.

Setting up your environment in GCP can entail many steps: setting up compute, network, and storage resources and keeping track of their configurations. You can do it all by hand if you want to, taking an imperative approach. But it is more efficient to use a template. That means a specification of what the environment should look like, declarative rather than imperative.

GCP provides Deployment Manager to let do just that. It's an infrastructure management service that automates the creation and management of your Google Cloud Platform resources for you.

To use Deployment Manager, you create a template file, using either the YAML markup language or Python, that describes what you want the components of your environment to look like. Then you give the template to Deployment Manager, which figures out and does the actions needed to create the environment your template describes. If you need to change your environment, edit your template, and then tell Deployment Manager to update the environment to match the change.

Google Cloud Certified Professional Cloud Architect Definitive Guide

You can store and version-control your Deployment Manager templates in Cloud Source Repositories.

Sample Deployment

A sample of a template specifying an f1-micro instance that would be created in the us-west1 region using a project with the project ID of gcp-arch-exam-project, a boot disk with the Centos 7 operating system installed, and an external IP address on the network interface:

resources:

- type: compute.v1.instance
name: gcp-arch-exam-vm1

properties:

```
zone: us-west1-a  
machineType: https://www.googleapis.com/compute/v1/projects/gcp-arch-examproject/  
zones/us-central1-f/machineTypes/f1-micro  
disks:  
- deviceName: boot  
type: PERSISTENT  
boot: true  
autoDelete: true  
initializeParams:  
sourceImage: https://www.googleapis.com/compute/v1/projects/gce-uefiimages/  
global/images/family/centos-7  
networkInterfaces:  
- network: https://www.googleapis.com/compute/v1/projects/gcp-arch-examproject  
/global/networks/default  
accessConfigs:  
- name: External NAT  
type: ONE_TO_ONE_NAT
```

Google Cloud Certified Professional Cloud Architect Definitive Guide

Sets of resource templates can be grouped together into deployments. When a deployment is run or executed, all of the specified resources are created.

Access Control is done using IAM and uses the following roles:

`roles/deploymentmanager.editor`

`roles/deploymentmanager.typeEditor`

`roles/deploymentmanager.viewer`

`roles/deploymentmanager.typeViewer`

Cloud Shell & Cloud SDK

GCP Consumption Mechanism

Cloud Platform Console	Cloud Shell and Cloud SDK	Cloud Console Mobile App	REST-based API
Web user interface	Command-line interface	For iOS and Android	For custom applications

1. GCP Console

- Centralized control for all projects
- Manage, create project resources
- Dashboard for monitoring resources
- Developer tools

Launching GCP Console

Launch GCP Console using: <https://console.cloud.google.com>

The screenshot shows the Google Cloud Platform console dashboard for the project "GCP Champion".

- Project Details:** Shows the project name "GCP Champion", project ID "gcp-champion", and project number "699887619371". It also includes a link to "ADD PEOPLE TO THIS PROJECT".
- Resource Utilization:** Shows the utilization of Compute Engine resources over time. A message indicates "No data is available for the selected time frame." Below this, there is a chart with a single data point at 0.0% utilization.
- Billing:** Shows estimated charges of INR ₹0.00 for the billing period 1–8 Apr 2021. It includes a "Billing Info" button.
- Google Cloud Platform status:** Shows "All services normal".
- Quick access:** Lists recent items such as "Getting started – GCP Champion", "APIs & Services – APIs & Services", "VM instances – Compute Engine", and "Quotas – APIs & Services".
- Search bar:** Allows searching for products and resources.
- Navigation:** Includes links for DASHBOARD, ACTIVITY, and RECOMMENDATIONS, along with a "Project Name" dropdown and a "CUSTOM" link.

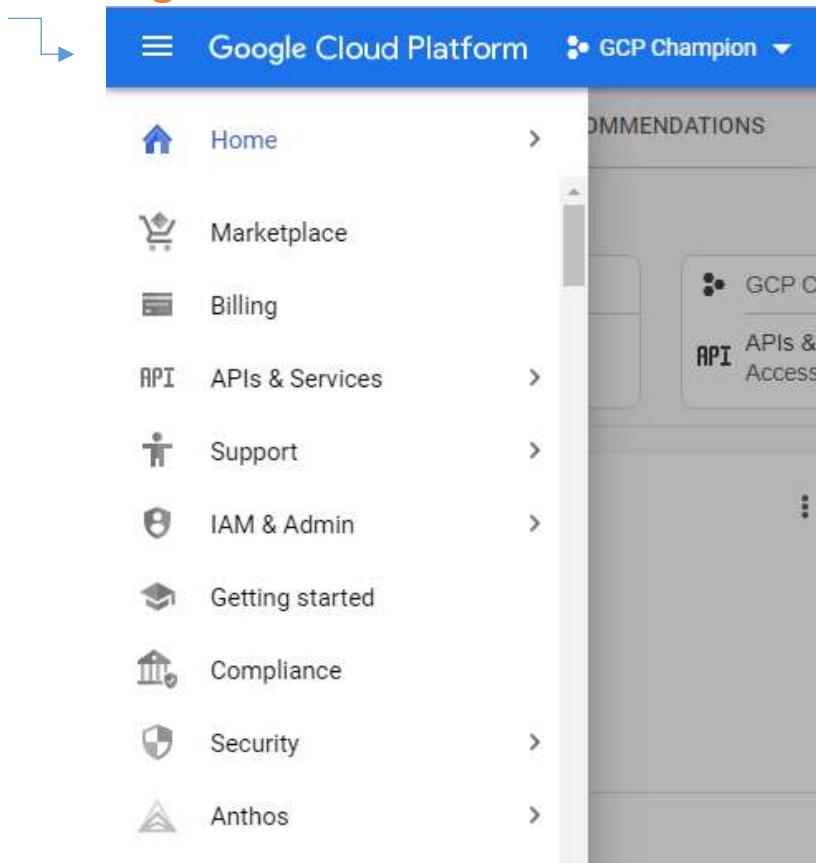
GCP Dashboard Provides High-level view of Google Cloud resources

- Project Info
- Provides resource utilization details
- Billing details of project
- API-wise usage details
- Links to documentations
- Monitoring
- Error Reporting
- News

Navigation | Console Menu

- Will list all the service offerings by the Google Cloud Platform
- It is categorized into Compute, Storage, Network, Big Data, AI, Operations, and Tools
- The top list displays top management and support options. You can also pin services and products you often use

Hamburger



Viewing Console Activity

Get the log of all activities

DASHBOARD	ACTIVITY	RECOMMENDATIONS
02/04/2021		
12:44 google.api.serviceusage.v1.ServiceUsage.EnableS... gcpchampion20@gmail.com has executed google.api.serviceusage.v1.Service...		
12:44	Create agent in project	gcpchampion20@gmail.com created agent in global
12:26	google.longrunning.Operations.GetOperation	gcpchampion20@gmail.com has executed google.longrunning.Operations.Get...
12:26	Completed:google.api.serviceusage.v1.ServiceUsa...	gcpchampion20@gmail.com has executed google.api.serviceusage.v1.Service...
12:26	google.longrunning.Operations.GetOperation	gcpchampion20@gmail.com has executed google.longrunning.Operations.Get...
12:26	google.api.serviceusage.v1.ServiceUsage.EnableS...	gcpchampion20@gmail.com has executed google.api.serviceusage.v1.Service...

What are the key factors of GCP Console?

The key factors of GCP Console are: -

- Resource Management
- Powerful Data Management
- Connect to VM using SSH browser
- Billing
- Activity Updates
- Cloud Shell
- Marketplace
- Perform Diagnostics

2. Cloud Shell & Cloud SDK

Command Line Interface Tools

- CLI Tools can be accessed via Cloud Shell or Cloud SDK
- Various CLI Tools are: -
 - gcloud
 - gsutil
 - bq

Cloud Shell

Google Cloud Shell provides you with gcloud command-line access to computing resources hosted on the Google Cloud Platform. Cloud Shell is a Debian-based virtual machine with a persistent 5GB home directory, which makes it easy for you to manage your GCP projects and resources.

Google Cloud Shell

Free, pre-installed with the tools you need for the Google Cloud Platform. [Learn More](#)

```
Starting update of app: test-project, version: 1
10:35 PM Cloning 1 static file.
10:35 PM Cloning 5 application files.
10:35 PM Compilation starting.
10:35 PM Compilation completed.
10:35 PM Starting deployment.
10:35 PM Checking if deployment succeeded.
10:35 PM Deployment successful.
10:35 PM Checking if updated app version is serving.
10:35 PM Completed update of app: test-project, version: 1
vardhanm92@cloudshell:~/appengine-example$
```

Real Linux environment

- Linux Debian-based OS
- 5GB persisted home directory
- Add, edit and save files

Configured for Google Cloud

- Google Cloud SDK
- Google App Engine SDK
- Docker
- Git
- Text editors
- Build tools
- View more ↗

Popular language support

- Python
- Java
- Go
- Node.js

[CANCEL](#) [START CLOUD SHELL](#)

Features

- Cloud shell provides VM instance of Cloud SDK Linux already installed
- Helpful for automation of tasks
- Set of CLI tools: Gcloud, Gsutil, BigQuery
- Write and run scripts using Cloud shell to control resource
- Cloud shell: no need to install anything on local machine
- Can be accessed from browser

Cloud SDK

The Cloud SDK gcloud and other utilities you need come pre-installed in Cloud Shell, which allows you to get up and running quickly.

- Cloud SDK can be installed via apt-get, yum, or using an installer
- We can run the tools gcloud, gsutil & bq interactively or using automated scripts
- Cloud SDK supports Windows, Linux, macOS, after installed, the SDK supports updates and additional features
- Cloud SDK commands to manage resources, Cloud Storage & Bigquery

The key factors of Cloud SDK are: -

- Command-line interface for Google Cloud Platform products and services
- Available as Docker image
- Available as VM instance via Cloud shell
- Can be installed on any OS: Windows, Linux, Mac
- Also, can be invoked using Cloud Shell in console
- Set of CLI tools: gcloud, gsutil, BigQuery, kubectl Tool

Components of Cloud SDK

By default, following components are installed when Cloud SDK is installed:

- gcloud: Tool for interacting with GCP
- gsutil: Tool for performing tasks related to Google Cloud Storage
- Bq: Tool for working with Data in BigQuery
- Package: Containing dependencies for SDK
- Core: Libraries used internally by the SDK tools

Installing Cloud SDK on Windows

Download Cloud SDK installer, Launch and complete installation

<https://dl.google.com/dl/cloudsdk/channels/rapid/GoogleCloudSDKInstaller.exe>

Alternatively download Cloud SDK installer from Codelabs

<https://codelabs.developers.google.com/codelabs/cloud-windows-powershell/index.html?index=..%2F..index#0>



Initializing & Updating Cloud SDK on Windows

- Initialize Cloud SDK using command called *gcloud init*
 - In your browser, log in to Google user account
 - Click "Allow" - To grant permission to access GCP resources
 - Select the Project on the command prompt – If there is only 1 project, no need to select
- Updating SDK to latest Cloud SDK version using command called *gcloud components update*

Installing Cloud SDK on Linux

Pre-requisite

Check Python version on the system i.e., It should be more than 2.7.9

```
//Check Python version
```

```
>> python2 –version
```

Download Cloud SDK for Linux

For 64bit Linux

https://dl.google.com/dl/cloudsdk/channels/rapid/downloads/google-cloud-sdk-190.0.1-linux-x86_64.tar.gz

For 32-bit Linux

<https://dl.google.com/dl/cloudsdk/channels/rapid/downloads/google-cloud-sdk-190.0.1-linux-x86.tar.gz>

Add Cloud SDK in your path on your Linux

```
//Run install script to add Cloud SDK in path
```

```
./google-cloud-sdk/install.sh
```

Initialize the Cloud SDK on Linux

```
./google-cloud-sdk/bin/gcloud init
```

CLI Tools

gcloud

What is gcloud?

- Gcloud is a CLI tool for GCP
- Used for automation
- Can be invoked from command line and from within scripts
- Part of Google Cloud SDK
- We can Download from the below link
https://cloud.google.com/sdk/docs/#install_the_latest_cloud_tools_version_cloudsdk_current_version

Usage of gcloud

- gcloud is used for automation
- gcloud can be invoked from command line and from within scripts
- gcloud is a part of Google Cloud SDK, so we can manage, create, delete, activate, and change gcloud configurations
- gcloud helps to create, delete GCP Compute Engine Services, Kubernetes Engine, and Dataproc Clusters
- Gcloud helps to manage jobs, submit jobs, and view logs

Managing gcloud

gcloud can be used for Creating and Managing: -

- Google Compute Engine Virtual Machine instances and other resources
- Google Cloud SQL instances
- Google Container Engine clusters Google Cloud Dataproc clusters and jobs
- Google Cloud DNS managed zones and record sets
- Google Cloud Deployment manager deployments
- Deploy App engine application

gcloud Commands

gcloud Initializing

- After installation gcloud init helps to perform a setup the below tasks: -
- Authorizing the cloud SDK tools to use your user account credentials to access your GCP resources.
- Setting up a configuration covering the active account, current project, and optionally, the default compute engine region and zone.
- Initializing gcloud using command called

gcloud init

gcloud Configuration

- gcloud Configuration is set of properties that help establish the gcloud command line behavior and cloud SDK tools
- When we install cloud SDK tools, gcloud has a default configuration
- We can also create our custom configuration. For example, we can have one configuration for development, testing, and production
- To view the current configuration commands
gcloud config list
- To list all the configuration in the project
gcloud config configurations list
- To view all set values for the configuration
gcloud config configurations describe <configuration_name>

gcloud Command – Help

- The gcloud commands to get help

gcloud help

To Create & Activate Configuration

- To create a new configuration

gcloud config configurations create dev

- To activate configuration

gcloud config configurations activate dev

To Set Region & Zone

The gcloud commands to set the Region & Zone: -

- To set the region for the current configuration

gcloud config set compute/region <region>

- To set the zone for the current configuration

gcloud config set compute/zone <zone name>

To Set Google Account & Set Project

The gcloud commands to set the Google Account & Project

- To set Google Account email address or a Google Service account email address

gcloud config set core/account <user@example.com>

- To Set the project for the current configuration

gcloud config set project <projectname>

Components

- To check the list of installed components

gcloud components list

- To Install A component

gcloud components install COMPONENT_ID

- To Update components

gcloud components update

- To Remove a component

gcloud components remove COMPONENT_ID

gcloud Command – Compute

- List of instances

gcloud compute instances list

- List of instances in zone asia-northeast1-a

gcloud compute instances list --filter="zone:asia-northeast1-a"

- List compute resources

gcloud compute instances list --

format='table[box,title=Instances](name:sort=1,zone:title=zone,status)'

- List Instance template

gcloud compute instance-templates list

- Provision a VM in London Region

gcloud compute instances create learn-gcp-gce-shell --zone=europe-west2-a

gsutil

What is gsutil?

- During Cloud SDK installation, gsutil gets installed by default
- gsutil is a command utility to interact with the cloud storage bucket in GCP up to Petabyte size
- We can easily manage, create, delete a bucket and upload or download files from a bucket
- We can manage user access and controls objects and bucket
- Syntax: To know the Structure of Cloud Storage Object

gs://<bucketname>/<object name>

Where *gs://* -> prefix

gsutil Commands – General

- Built in help
 - gsutil help*
- Built in help for command or topic
 - gsutil help cp*
- gsutil top level command line options
 - gsutil help options*
- Version of gsutil
 - gsutil version -1*

gsutil Commands – File Management

- Upload a file using the cp command
 - gsutil cp <local_file> gs://<bucketname>/*
- Download a file using the cp command
 - gsutil cp gs://<bucketname>/<remote_file>*
- Transfer a file between buckets
 - gsutil cp gs://<bucket_A>/<remote_file> gs://<bucket_B>/*

- List the objects inside the bucket

```
gsutil ls gs://<bucketname>/
```

- Check the space used by all the objects

```
gsutil du h gs://<bucketname>/
```

- Make the content of a target folder identical to the content of a source folder

```
gsutil m rsync r d ./ myfolder gs://<bucketname>
```

gsutil Commands – Bucket

- To Create a bucket

```
gsutil mb gs://<bucketname> | gsutil mb gs://learning-gcp-in nutshell
```

- To List all buckets

```
gsutil ls
```

- To List the objects inside the bucket

```
gsutil ls gs://<bucketname>/
```

- To Delete a bucket

```
gsutil rb gs://<bucket_name>
```

gsutil Commands – Access Control List

- To lists the permissions on a given object

```
gsutil acl get gs://<bucketname>/
```

- To sets the permissions on a given object

```
gsutil acl set acl.txt gs://bucket/<filename>
```

- Changes and modifies the current permissions on a given object

```
gsutil acl ch -u <username>:R gs://<bucketname>/
```

Bigquery (bq)

- A BigQuery is a collection of tables, views and Models
- To create a Bigquery dataset
 - Open a cloud shell / google cloud sdk
 - To create a dataset -> bq mk <dataset name>
 - To view the dataset -> bq show <dataset name>
- Example To Create and View the dataset
 - bq mk digitalasset
 - bq show digitalasset

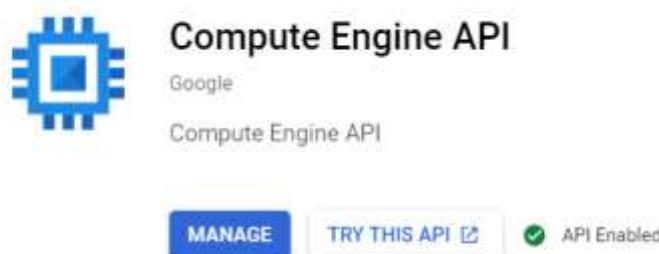
Cloud SDK API Explorer

API Explorer

- A Browser based interactive tool to explore Google APIs
- We can check Available APIs and provides ease of usage & execute request and see the response in real time
- API Explorer is a mechanism to interact with Google Cloud Services programmatically
- The API library offers quick access, documentation, and configuration options for 200+ Google APIs.

Compute Engine API

- Find library for service from the menu
- Click on "Try this API", search, and select the action to perform
- Fill the properties; this generates a JSON



Try this API

Request parameters

project	curl --request POST \
string	'https://compute.googleapis.com/compute/v1/projects/[PROJECT]/zones/[ZONE]/instances?key=[YOUR_API_KEY]
zone	--header 'Authorization: Bearer [YOUR_ACCESS_TOKEN]' \
string	--header 'Accept: application/json' \
requestId	--data '{"machineType": "", "name": ""}' \
string	--compressed

Show standard parameters ▾

Request body

```
{
  "machineType": "",
  "name": ""
}
```

For suggestions, press control+space or click one of the blue "add" circles.

Credentials ⓘ

Google OAuth 2.0

List API using Web Console | Command Line

In Web Console, to list the APIs and services that are available in a project: -

- Go to the Cloud Console API Library page
- Click Select to choose the GCP project
- To list the APIs using Command Line: -

```
gcloud config set <project>
gcloud services list - --available
gcloud services enable/disable <service name>
```

Example: gcloud services disable pubsub.googleapis.com

Client Libraries

- Open-source libraries
- Available for most used programming languages e.g. Python, .NET, Node.js, Java etc.
- Provide interface for calling Google Cloud APIs

The screenshot shows the Google Cloud Platform API Library. On the left, there's a sidebar with a 'Filter by' dropdown set to 'Maps' and a 'CATEGORY' sidebar with various service links. The main area is divided into sections: 'Maps' (with four cards: Google Maps Android API, Google Maps SDK for iOS, Google Maps JavaScript API, and Google Places API for Android) and 'Machine learning' (with four cards: Dialogflow API, Google Cloud Vision API, Google Cloud Natural Language API, and Google Cloud Speech API). There are 'VIEW ALL' buttons for each section.

To install the client libraries

- Go to => <https://cloud.google.com/apis/docs/cloud-client-libraries>
- Select the language & Install the Client Libraries

Google Cloud Client Library	Installation & Reference
Go	<ul style="list-style-type: none">▪ GitHub Repo▪ Library Reference
Java	<ul style="list-style-type: none">▪ GitHub Repo▪ Library Reference
Node.js	<ul style="list-style-type: none">▪ GitHub Repo▪ Library Reference
Python	<ul style="list-style-type: none">▪ GitHub Repo▪ Library Reference
Ruby	<ul style="list-style-type: none">▪ GitHub Repo▪ Library Reference
PHP	<ul style="list-style-type: none">▪ GitHub Repo▪ Library Reference
C#	<ul style="list-style-type: none">▪ GitHub Repo▪ Library Reference

Management Tools – Interview | Exam Tips

GCP

Monitoring Quiz

1. As an SRE, you are assigned to support several applications. In the past, these applications have had significant reliability problems. You would like to understand the performance characteristics of the applications, so you create a set of dashboards. What kind of data would you display on those dashboards?

- A. Metrics and time-series data measuring key performance attributes, such as CPU utilization
- B. Detailed log data from syslog
- C. Error messages output from each application
- D. Results from the latest acceptance tests

Correct Answer: A

2. After determining the optimal combination of CPU and memory resources for nodes in a Kubernetes cluster, you want to be notified whenever CPU utilization exceeds 85 percent for 5 minutes or when memory utilization exceeds 90 percent for 1 minute. What would you have to specify to receive such notifications?

- A. An alerting condition
- B. An alerting policy
- C. A logging message specification
- D. An acceptance test

Correct Answer: B

3. A compliance review team is seeking information about how your team handles high-risk administration operations, such as granting operating system users root privileges. Where could you find data that shows your team tracks changes to user privileges?

- A. In metric time-series data
- B. In alerting conditions
- C. In audit logs
- D. In ad hoc notes kept by system administrators

Correct Answer: C

4. Release management practices contribute to improving reliability by which one of the following?

- A. Advocating for object-oriented programming practices
- B. Enforcing waterfall methodologies
- C. Improving the speed and reducing the cost of deploying code
- D. Reducing the use of stateful services

Correct Answer: C

5. A team of software engineers is using release management practices. They want developers to check code into the central team code repository several times during the day. The team also wants to make sure that the code that is checked in is functioning as expected before building the entire application. What kind of tests should the team run before attempting to build the application?

- A. Unit tests
- B. Stress tests
- C. Acceptance tests
- D. Compliance tests

Correct Answer: A

6. Developers have just deployed a code change to production. They are not routing any traffic to the new deployment yet, but they are about to send a small amount of traffic to servers running the new version of code. What kind of deployment are they using?

- A. Blue/Green deployment
- B. Before/After deployment
- C. Canary deployment
- D. Stress deployment

Correct Answer: C

7. You have been hired to consult with an enterprise software development that is starting to adopt agile and DevOps practices. The developers would like advice on tools that they can use to help them collaborate on software development in the Google Cloud. What version control software might you recommend?

- A. Jenkins and Cloud Source Repositories
- B. Syslog and Cloud Build
- C. GitHub and Cloud Build
- D. GitHub and Cloud Source Repositories

Correct Answer: D

8. A startup offers a software-as-a-service solution for enterprise customers. Many of the components of the service are stateful, and the system has not been designed to allow incremental rollout of new code. The entire environment has to be running the same version of the deployed code. What deployment strategy should they use?

- A. Rolling deployment
- B. Canary deployment
- C. Stress deployment
- D. Blue/Green deployment

Correct Answer: D

9. A service is experiencing unexpectedly high volumes of traffic. Some components of the system are able to keep up with the workload, but others are unable to process the volume of requests. These services are returning a large number of internal server errors. Developers need to release a patch as soon as possible that provides some relief for an overloaded relational database service. Both memory and CPU utilization are near 100 percent. Horizontally scaling the relational database is not an option, and vertically scaling the database would require too much downtime. What strategy would be the fastest to implement?

- A. Shed load
- B. Increase connection pool size in the database
- C. Partition the workload
- D. Store data in a Pub/Sub topic

Correct Answer: A

10. A service has detected that a downstream process is returning a large number of errors. The service automatically slows down the number of messages it sends to the downstream process. This is an example of what kind of strategy?

- A. Load shedding
- B. Upstream throttling
- C. Rebalancing
- D. Partitioning

Correct Answer: B

11. Google Stackdriver supports following cloud platforms:

- A. AWS
- B. GCP
- C. Azure
- D. All of the above

Correct Answer: D

Migration Services

What is the reason for migrating to Cloud?

The primary reason for migrating to Cloud are:

- On-Demand Resourcing
- Auto Scaling
- More Cost Benefit
- Disaster Recovery
- Managed Services with 24*7 support

What are the decision-making steps to be considered before migration?

The decision-making steps before migration are:

HW Requirement

Compare cost of local Hardware vs What do you need on the cloud?

SW Requirement/Licensing

What kind of operating system/SW to be used? and Do we need licenses?

Dependencies

Apart from VMs, what other services are required by your application, what is the cost of that?

Other factors are

- Stability
- Reliability
- Scalability
- Application Performance Metrics
- Resource Cost
- Consider Data Privacy Aspect
- Application Scalability
- Do we need to do Refactoring?
- How DevOps will work
- Available Migration Tools

What are the different migration approaches?

There are two types of migration approaches

- Self-Service
- Partner Assisted

In Self-Service

- You need to do all the assessment, planning and migration on your own
- Free
- Recommended to follow Best Practices

In Partner Assisted

- Work with Google partners for your migration
- This is paid service

Google recommends to go with partner-assisted migration

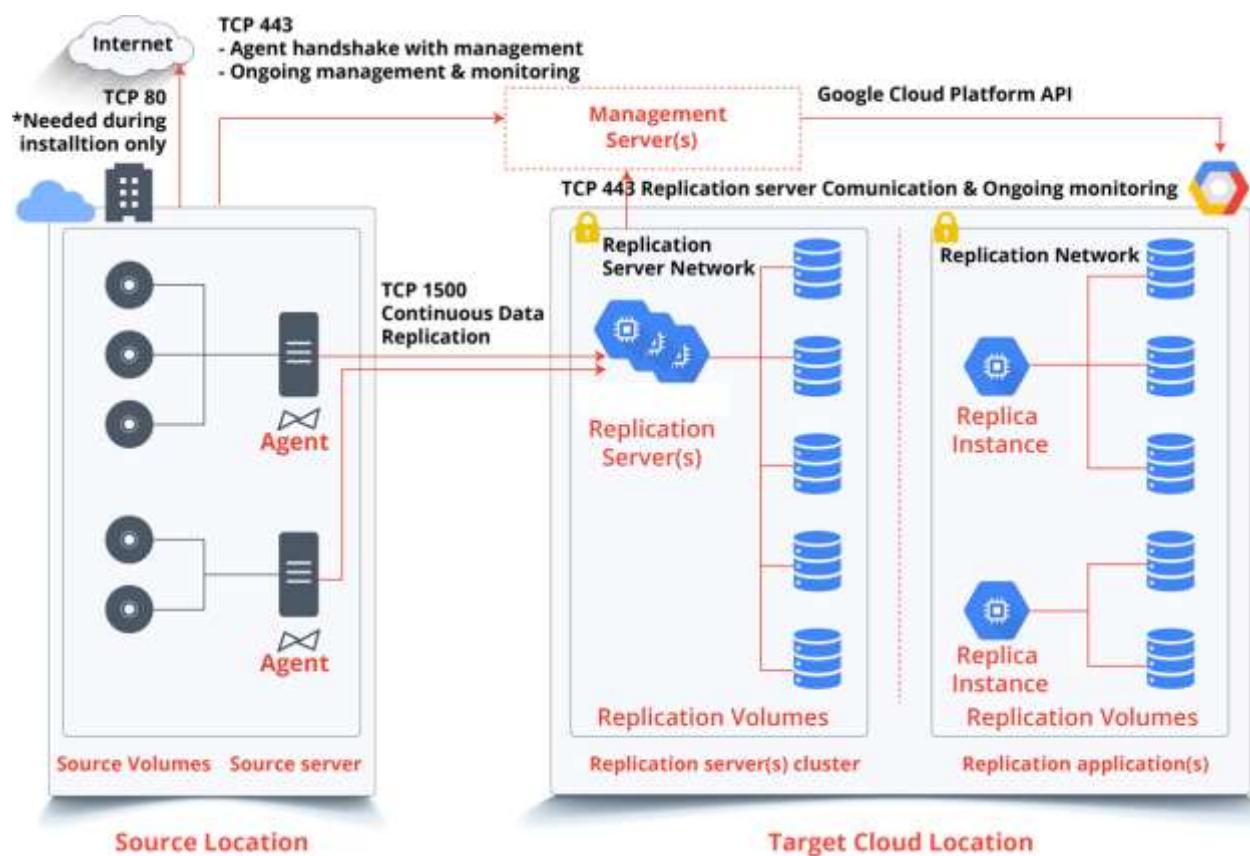
Activities	Self Service Migration	Partner Assisted Migration
Assessment	Done by Internal Team	Done by Partner. Application clusters are automatically extracted.
Planning	Done by Internal Team	Partners determine the appropriate machine types.
N/W Configuration	Internal Team configure network rules to match the internal systems	Partner solutions enable networking to match the existing N/W topology.
Replication	Use GCP's free migration service.	Partners replicate the virtual machines into the cloud

What are the best practices for migration?

The best practices for migration are:

- Provide Detailed Cost Estimation for each service
- Assign the items to migration based on the categories
- Designing the migration environment
- Establish the governance with GCP Account
- Creating a target network (Decide how they connected), think various options to connect various resources such as VPN, Dedicated Interconnect or Direct Peering
- Planning for operations - Using Cloud deployment manager to integrate DevOps tool & using tools for Monitoring and logging
- Kick off Migration i.e., Start migrating your first application, if needed the process has to be refined for each migration

What are the Steps involved to Migrate VM/Server to Compute Engine?



Google Cloud Certified Professional Cloud Architect Definitive Guide

- To use Migration service, you need to link project using service key which will be required on Tool
- Disruption of service does NOT happen during migration
- Charges will be made only for the cloud resources. Source machine should be running on Windows & Linux

The steps involved in migrating VM to Compute Engine:

- Creating VM
- Configure your Tool account
- Select replication options
- Install the migration agent
- Configure the target machine
- verify initial sync completion
- Test the Creation of Target VM Instances
- Test the Availability of the Target VMs
- Cut Over to GCP

Explain Migration Architecture?

These are the possible scenarios while migrating to cloud

Complete Redesigning

Involves complete redesigning of an application for cloud

Lift-and-shift migration

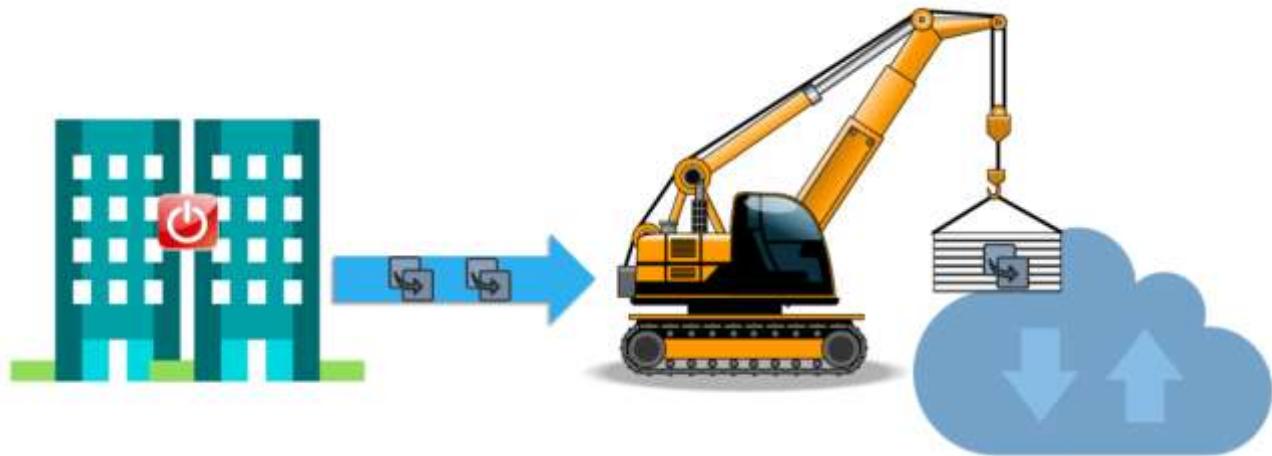
Easiest approach, need minimal changes in application

Hybrid Migration

Partial movement of application to cloud

Lift-and-shift migration

- Simplest migration
- Shut down the existing application
- Copy the data and virtual machines to cloud



Also known as “rehosting” or the “forklift approach,” lift and shift is a term used to describe the strategy of removing workloads and tasks from one storage location and placing them in another, usually cloud-based, location.

The advantage of a lift and shift cloud migration is that it allows organizations to move their applications quickly and easily without having to re-architect them.

While the process is sometimes characterized as “cutting and pasting,” it requires a great deal more forethought, documentation, and planning to ensure that data sets will be matched with handling systems in the new environment and that the applications will have all the resources they need to operate effectively.

Moving the Data

- Move the necessary data of application to cloud storage such as data stored in databases, static assets etc.
- Move data which would be required by the application when virtual machines are launched

Moving the VMs

- You may visit the Link provided that contains multiple options to move VMs: [Link](#)
- It is a third-party partner service

Testing the application

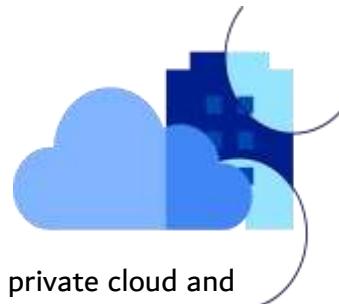
- Run the application in test mode
- Verify performance metrics
- Verify that app is properly deployed, monitoring and logging framework is working fine

Moving to Production

- Inform users regarding maintenance window
- Take down application from its current location
- Switch the DNS entries
- Bring up the application in the cloud

Hybrid Migration

- Application is moved to the cloud partially
- Most probably the front end and the application logic are moved to the cloud
- Back end and associated services does not move to cloud



Hybrid cloud is a cloud computing environment that uses a mix of on-premises, private cloud and third-party, public cloud services with orchestration between the two platforms.

By allowing workloads to move between private and public clouds as computing needs and costs change, hybrid cloud gives businesses greater flexibility and more data deployment options.

Determining the network connection

- Create a connection between your existing resources and Cloud Platform.
- Type of connection depends upon your need. Example: VPN or dedicated and high-speed line.

Migrating the VMs

- You may visit the [Link](#) provided that contains multiple options to move VMs: [Link](#)
- It is a third-party partner service

Moving to production

- Inform users regarding maintenance window
- Take down application from its current location.
- Switch the DNS entries
- Bring up the application in the cloud

Advantages of Hybrid Model

- Enterprise Security
- Enterprise Networking
- Step towards full migration
- Easy Control
- Policy based Orchestration
- Service Management using Istio
- API Management
- Greater Visibility

What are the different ways to connect from On-Premises to GCP?

Different applications and workloads require different network connectivity solutions.

The two major options are interconnect and peering.

Interconnect

Cloud Interconnect provides two options for extending your on-premises network to your Google Cloud Platform (GCP)

Virtual Private Cloud (VPC) networks. You can create a dedicated connection (Dedicated Interconnect) or use a service provider (Partner Interconnect) to connect to VPC networks. When choosing an interconnect type, consider your connection requirements, such as the connection location and capacity.

If you have high bandwidth needs, Dedicated Interconnect can be a cost-effective solution.

If you require a lower-bandwidth solution, Dedicated Interconnect and Partner Interconnect provide a variety of capacity options starting at 50 Mbps.

If you can't physically meet Google's network in a colocation facility to reach your Virtual Private Cloud networks, you can use Partner Interconnect to connect to a variety of service providers that connect directly to Google.

There are two types of Interconnects namely Dedicated Interconnect, IPSec VPN.

Dedicated Interconnect

- Allows you to directly connect your on-premises network to your GCP VPC
- Useful for high amount of data
- Dedicated Interconnect offers enterprise-grade connections to GCP
- Useful to connect to your VPC
- To extend corporate data center's IP space into the Google cloud (In hybrid environments)
- Can be configured to offer a 99.9% or a 99.99% uptime SLA

IPsec VPN

- Useful to connect to your VPC over the public internet
- Uses Google Cloud VPN
 - Traffic between the on-prem and GCP network is encrypted by one VPN gateway and decrypted by another VPN gateway.
- Useful for low volume data connections.

Peering

Peering is a process by which two Internet networks connect and exchange traffic. It allows them to directly hand off traffic between each other's customers, without having to pay a third party to carry that traffic across the Internet for them.

Peering is distinct from transit, the more usual way of connecting to the Internet, in which an end user or network operator pays another, usually larger, network operator to carry all their traffic for them.

There are two types of Peering namely Direct Peering & Carrier Peering

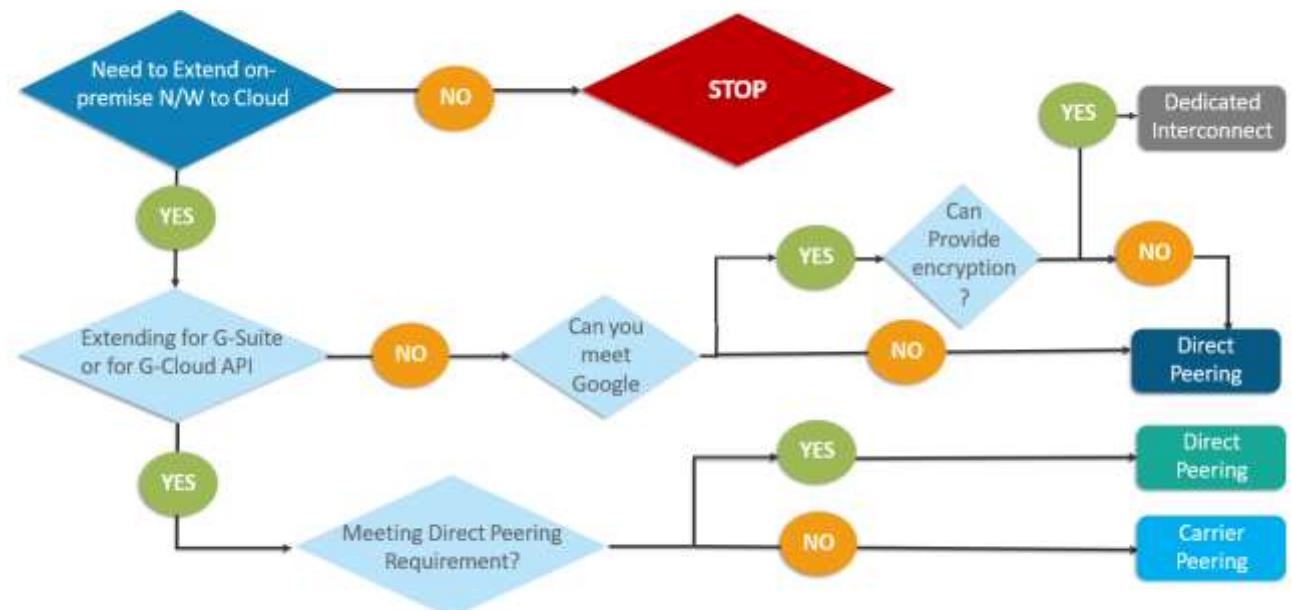
Direct Peering

- For Direct peering connection, one needs to satisfy the peering requirement
- Cheaper than IPsec VPN option

Carrier Peering

- Used if you cannot satisfy Google's peering requirements

Connecting from On-Premises to GCP Decisions Chart



Comparison of Interconnect Options

	Dedicated Interconnect	IPSec VPN Tunnel
Access Type	Internal IP addresses in RFC 1918 address space	Internal IP addresses in RFC 1918 address space
Capacity	10 Gbps for each link	1.5-3 Gbps for each tunnel
Cost	Reduced egress costs Fee for each link and VLAN	Egress is billed same as general network pricing Fee for each tunnel

Comparison of Peering Options

	Direct	Carrier
Access Type	Public IP addresses	Public IP addresses
Capacity	10 Gbps for each link	Varies based on partner offering
Cost	Settlement free peering Reduced cost for egress	Cost based on partner offering Reduced cost for egress

Migration – Interview | Exam Tips

GCP

Cloud Architect Certification

Assessment Test

1. Building for Builders LLC manufactures equipment used in residential and commercial building. Each of its 500,000 pieces of equipment in use around the globe has IoT devices collecting data about the state of equipment. The IoT data is streamed from each device every 10 seconds. On average, 10 KB of data is sent in each message. The data will be used for predictive maintenance and product development. The company would like to use a managed database in Google Cloud. What would you recommend?

- A. Apache Cassandra
- B. Cloud Bigtable**
- C. BigQuery
- D. CloudSQL

Correct Answer: B

2. You have developed a web application that is becoming widely used. The frontend runs in Google App Engine and scales automatically. The backend runs on Compute Engine in a managed instance group. You have set the maximum number of instances in the backend managed instance group to five. You do not want to increase the maximum size of the managed instance group or change the VM instance type, but there are times the frontend sends more data than the backend can keep up with and data is lost. What can you do to prevent the loss of data?

- A. Use an unmanaged instance group
- B. Store ingested data in Cloud Storage
- C. Have the frontend write data to a Cloud Pub/Sub topic, and have the backend read from that topic**
- D. Store ingested data in BigQuery

Correct Answer: C

3. You are setting up a cloud project and want to assign members of your team different permissions. What GCP service would you use to do that?

- A. Cloud Identity
- B. Identity and Access Management (IAM)**
- C. Cloud Authorizations
- D. LDAP

Correct Answer: B

4. You would like to run a custom container in a managed Google Cloud Service. What are your two options?

- A. App Engine Standard and Kubernetes Engine
- B. App Engine Flexible and Kubernetes Engine**
- C. Compute Engine and Kubernetes Engine
- D. Cloud Functions and App Engine Flexible

Correct Answer: B

5. PhotosForYouToday prints photographs and ships them to customers. The frontend application uploads photos to Cloud Storage. Currently, the backend runs a cron job that checks Cloud Storage buckets every 10 minutes for new photos. The product manager would like to process the photos as soon as they are uploaded. What would you use to cause processing to start when a photo file is saved to Cloud Storage?

- A. A Cloud Function**
- B. An App Engine Flexible application
- C. A Kubernetes pod
- D. A cron job that checks the bucket more frequently

Correct Answer: A

Google Cloud Certified Professional Cloud Architect Definitive Guide

6. The chief financial officer of your company believes that you are spending too much money to run an on-premises data warehouse and wants to migrate to a managed cloud solution. What GCP service would you recommend for implementing a new data warehouse in GCP?

A. Compute Engine

B. BigQuery

C. Cloud Dataproc

D. Cloud Bigtable

Correct Answer: B

7. A government regulation requires you to keep certain financial data for seven years. You are not likely to ever retrieve the data, and you are only keeping it to be in compliance. There are approximately 500 GB of financial data for each year that you are required to save. What is the most cost-effective way to store this data?

A. Cloud Storage multiregional storage

B. Cloud Storage Nearline storage

C. Cloud Storage Coldline storage

D. Cloud Storage persistent disk storage

Correct Answer: C

8. Global Games Enterprises Inc. is expanding from North America to Europe. Some of the games offered by the company collect personal information. With what additional regulation will the company need to comply when it expands into the European market?

A. HIPAA

B. PCI-DS

C. GDPR

D. SOX

Correct Answer: C

9. Your team is developing a Tier 1 application for your company. The application will depend on a PostgreSQL database. Team members do not have much experience with PostgreSQL and want to implement the database in a way that minimizes their administrative responsibilities for the database. What managed service would you recommend?

- A. Cloud SQL
- B. Cloud Dataproc
- C. Cloud Bigtable
- D. Cloud PostgreSQL

Correct Answer: A

10. What is a service-level indicator?

- A. A metric collected to indicate how well a service-level objective is being met**
- B. A type of log
- C. A type of notification sent to a sysadmin when an alert is triggered
- D. A visualization displayed when a VM instance is down

Correct Answer: A

11. Developers at MakeYouFashionable have adopted agile development methodologies. Which tool might they use to support CI/CD?

- A. Google Docs
- B. Jenkins**
- C. Apache Cassandra
- D. Clojure

Correct Answer: B

12. You have a backlog of audio files that need to be processed using a custom application. The files are stored in Cloud Storage. If the files were processed continuously on three n1-standard-4 instances, the job could complete in two days. You have 30 days to deliver the processed files, after which they will be sent to a client and deleted from your systems. You would like to minimize the cost of processing. What might you do to help keep costs down?

- A. Store the files in coldline storage
- B. Store the processed files in multiregional storage
- C. Store the processed files in Cloud CDN
- D. Use preemptible VMs**

Correct Answer: D

13. You have joined a startup selling supplies to visual artists. One element of the company's strategy is to foster a social network of artists and art buyers. The company will provide e-commerce services for artists and earn revenue by charging a fee for each transaction. You have been asked to collect more detailed business requirements. What might you expect as an additional business requirement?

- A. The ability to ingest streaming data
- B. A recommendation system to match buyers to artists**
- C. Compliance with SOX regulations
- D. Natural language processing of large volumes of text

Correct Answer: B

14. You work for a manufacturer of specialty die cast parts for the aerospace industry. The company has built a reputation as the leader in high-quality, specialty die cast parts, but recently the number of parts returned for poor quality is increasing. Detailed data about the manufacturing process is collected throughout every stage of manufacturing. To date, the data has been collected and stored but not analyzed. There is a total of 20 TB of data. The company has a team of analysts familiar with spreadsheets and SQL. What service might you recommend for conducting preliminary analysis of the data?

- A. Compute Engine
- B. Kubernetes Engine
- C. BigQuery**
- D. Cloud Functions

Correct Answer: C

15. A client of yours wants to run an application in a highly secure environment. They want to use instances that will only run boot components verified by digital signatures. What would you recommend they use in Google Cloud?

- A. Preemptible VMs
- B. Managed instance groups**
- C. Cloud Functions
- D. Shielded VMs

Correct Answer: B

16. You have installed the Google Cloud SDK. You would now like to work on transferring files to Cloud Storage. What command-line utility would you use?

- A. bq
- B. gsutil**
- C. cbt
- D. gcloud

Correct Answer: B

17. Kubernetes pods sometimes need access to persistent storage. Pods are ephemeral—they may shut down for reasons not in control of the application running in the pod. What mechanism does Kubernetes use to decouple pods from persistent storage?

- A. PersistentVolumes**
- B. Deployments
- C. ReplicaSets
- D. Ingress

Correct Answer: A

18. An application that you support has been missing service-level objectives, especially around database query response times. You have reviewed monitoring data and determined that a large number of databases read operations is putting unexpected load on the system. The database uses MySQL, and it is running in Compute Engine. You have tuned SQL queries, and

the performance is still not meeting objectives. Of the following options, which would you try next?

- A. Migrate to a NoSQL database.
- B. Move the database to Cloud SQL.
- C. Use Cloud Memorystore to cache data read from the database to reduce the number of reads on the database.**
- D. Move some of the data out of the database to Cloud Storage.

Correct Answer: C

19. You are running a complicated stream processing operation using Apache Beam. You want to start using a managed service. What GCP service would you use?

- A. Cloud Dataprep
- B. Cloud Dataproc
- C. Cloud Dataflow**
- D. Cloud Identity

Correct Answer: C

20. Your team has had a number of incidents in which Tier 1 and Tier 2 services were down for more than 1 hour. After conducting a few retrospective analyses of the incidents, you have determined that you could identify the causes of incidents faster if you had a centralized log repository. What GCP service could you use for this?

- A. Stackdriver Logging**
- B. Cloud Logging
- C. Cloud SQL
- D. Cloud Bigtable

Correct Answer: A

21. A Global 2000 company has hired you as a consultant to help architect a new logistics system. The system will track the location of parts as they are shipped between company

facilities in Europe, Africa, South America, and Australia. Anytime a user queries the database, they must receive accurate and up-to-date information; specifically, the database must support strong consistency. Users from any facility may query the database using SQL. What GCP service would you recommend?

- A. Cloud SQL
- B. BigQuery
- C. Cloud Spanner**
- D. Cloud Dataflow

Correct Answer: C

22. A database architect for a game developer has determined that a NoSQL document database is the best option for storing players' possessions. What GCP service would you recommend?

- A. Cloud Datastore**
- B. Cloud Storage
- C. Cloud Dataproc
- D. Cloud Bigtable

Correct Answer: A

23. A major news agency is seeing increasing readership across the globe. The CTO is concerned that long page-load times will decrease readership. What might the news agency try to reduce the page-load time of readers around the globe?

- A. Regional Cloud Storage
- B. Cloud CDN**
- C. Fewer firewall rules
- D. Virtual private network

Correct Answer: B

24. What networking mechanism allows different VPC networks to communicate using private IP address space, as defined in RFC 1918?

- A. ReplicaSets
- B. Custom subnets
- C. VPC network peering**
- D. Firewall rules

Correct Answer: C

25. You have been tasked with setting up disaster recovery infrastructure in the cloud that will be used if the on-premises data center is not available. What network topology would you use for a disaster recovery environment?

- A. Meshed topology
- B. Mirrored topology**
- C. Gated egress topology
- D. Gated ingress topology

Correct Answer: B

Case Study

The Google Cloud Professional Cloud Architect Certification exam uses three case studies as the basis for some questions on the exam. Each case study includes a company overview, solution concept, technical requirements, business requirements, and an executive statement.

The three case studies are available online here:

Dress4Win

<https://cloud.google.com/certification/guides/cloud-architect/casestudy-dress4win-rev2>

Mountkirk Games

<https://cloud.google.com/certification/guides/cloud-architect/casestudy-mountkirkgames-rev2>

TerramEarth

<https://cloud.google.com/certification/guides/cloud-architect/casestudy-terramearth-rev2>

Case Study 1: Dress4Win

Company Overview

Dress4Win is a web-based company that helps their users organize and manage their personal wardrobe using a web app and mobile application. The company also cultivates an active social network that connects their users with designers and retailers. They monetize their services through advertising, e-commerce, referrals, and a freemium app model. The application has grown from a few servers in the founder's garage to several hundred servers and appliances in a collocated data center. However, the capacity of their infrastructure is now insufficient for the application's rapid growth. Because of this growth and the company's desire to innovate faster, Dress4Win is committing to a full migration to a public cloud.

Solution Concept

For the first phase of their migration to the cloud, Dress4Win is moving their development and test environments. They are also building a disaster recovery site because their current infrastructure is at a single location. They are not sure which components of their architecture can be migrated as is and which components need to be changed before migrating them.

Existing Technical Environment

The Dress4Win application is served out of a single data center location. All servers run Ubuntu LTS v16.04.

Databases:

MySQL: 1 server for user data, inventory, and static data:

- MySQL 5.7
- 8 core CPUs
- 128 GB of RAM
- 2x 5 TB HDD (RAID 1)

Redis: 3 server cluster for metadata, social graph, and caching. Each server consists of:

- Redis 3.2
- 4 core CPUs
- 32GB of RAM

Compute:

40 Web application servers providing micro-services-based APIs and static content.

- Tomcat - Java
- Nginx
- 4 core CPUs
- 32 GB of RAM

20 Apache Hadoop/Spark servers:

- Data analysis
- Real-time trending calculations
- Eight core CPUs
- 128 GB of RAM
- 4x 5 TB HDD (RAID 1)

3 RabbitMQ servers for messaging, social notifications, and events:

- Eight core CPUs
- 32GB of RAM

Miscellaneous servers:

- Jenkins, monitoring, bastion hosts, and security scanners
- Eight core CPUs
- 32GB of RAM

Storage appliances:

- iSCSI for VM hosts
- Fiber channel SAN - MySQL databases
 - 1 PB total storage; 400 TB available
- NAS - image storage, logs, backups
 - 100 TB total storage; 35 TB available

Business Requirements

- Build a reliable and reproducible environment with scaled parity of production.
- Improve security by defining and adhering to a set of security and Identity and Access Management (IAM) best practices for the cloud.
- Improve business agility and speed of innovation through rapid provisioning of new resources.
- Analyze and optimize architecture for performance in the cloud.

Technical Requirements

- Easily create non-production environments in the cloud.
- Implement an automation framework for provisioning resources in cloud.
- Implement a continuous deployment process for deploying applications to the on-premises data center or cloud.
- Support failover of the production environment to the cloud during an emergency.
- Encrypt data on the wire and at rest.
- Support multiple private connections between the production data center and cloud environment.

Executive Statement

Our investors are concerned about our ability to scale and contain costs with our current infrastructure. They are also concerned that a competitor could use a public cloud platform to offset their up-front investment and free them to focus on developing better features. Our traffic patterns are highest in the mornings and weekend evenings; during other times, 80 percent of our capacity is sitting idle.

Our capital expenditure is now exceeding our quarterly projections. Migrating to the cloud will likely cause an initial increase in spending, but we expect to transition completely before our next hardware refresh cycle. Our total cost of ownership (TCO) analysis over the next five years for a public cloud strategy achieves a cost reduction of 30–50 percent over our current model.

Case Study 2: Mountkirk Games

Company Overview

Mountkirk Games makes online, session-based, multiplayer games for mobile platforms. They build all of their games using some server-side integration. Historically, they have used cloud providers to lease physical servers. Due to the unexpected popularity of some of their games, they have had problems scaling their global audience, application servers, MySQL databases, and analytics tools. Their current model is to write game statistics to files and send them through an ETL tool that loads them into a centralized MySQL database for reporting.

Solution Concept

Mountkirk Games is building a new game, which they expect to be very popular. They plan to deploy the game's backend on the Google Compute Engine so that they can capture streaming metrics, run intensive analytics, and take advantage of its autoscaling server environment and integrate with a managed NoSQL database.

Business Requirements

- Increase to a global footprint
- Improve uptime (downtime is loss of players)
- Increase efficiency of the cloud resources they use
- Reduce latency to all customers

Technical Requirements

Requirements for Game Backend Platform

1. Dynamically scale up or down based on game activity.
2. Connect to a transactional database service to manage user profiles and game state.
3. Store game activity in a time series database service for future analysis.
4. As the system scales, ensure that data is not lost due to processing backlogs.
5. Run hardened Linux distro.

Requirements for Game Analytics Platform

1. Dynamically scale up or down based on game activity.
2. Process incoming data on the fly directly from the game servers.
3. Process data that arrives late because of slow mobile networks.
4. Allow queries to access at least 10 TB of historical data.
5. Process files that are regularly uploaded by users' mobile devices.

Executive Statement

Our last successful game did not scale well with our previous cloud provider, resulting in lower user adoption and affecting the game's reputation. Our investors want more key performance indicators (KPIs) to evaluate the speed and stability of the game, as well as other metrics that provide deeper insight into usage patterns so that we can adapt the game to target users. Additionally, our current technology stack cannot provide the scale we need, so we want to replace MySQL and move to an environment that provides autoscaling, low-latency load balancing, and frees us up from managing physical servers.

Case Study 3: TerramEarth

Company Overview

TerramEarth manufactures heavy equipment for the mining and agricultural industries. About 80 percent of their business is from mining and 20 percent is from agriculture. They currently have over 500 dealers and service centers in 100 countries. Their mission is to build products that make their customers more productive.

Solution Concept

There are 20 million TerramEarth vehicles in operation that collect 120 fields of data per second. Data is stored locally on the vehicle, and it can be accessed for analysis when a vehicle is serviced. The data is downloaded via a maintenance port. This same port can be used to adjust operational parameters, allowing the vehicles to be upgraded in the field with new computing modules.

Approximately 200,000 vehicles are connected to a cellular network, allowing TerramEarth to collect data directly. At a rate of 120 fields of data per second, with 22 hours of operation per day, TerramEarth collects a total of about 9 TB of data per day from these connected vehicles.

Existing Technical Environment

TerramEarth's existing architecture is composed of Linux and Windows-based systems that reside in a single U.S. west coast-based data center. These systems gzip CSV files from the field, upload via FTP, and place the data in their data warehouse. Because this process takes time, aggregated reports are based on data that is three weeks old. With this data, TerramEarth has been able to stock replacement parts preemptively and reduce unplanned downtime of their vehicles by 60 percent. However, because the data is stale, some customers are without their vehicles for up to four weeks while they wait for replacement parts.

Business Requirements

- Decrease unplanned vehicle downtime to less than one week
- Support the dealer network with more data on how their customers use their equipment to position new products and services better.
- Have the ability to partner with different companies—especially with seed and fertilizer suppliers in the fast-growing agricultural business—to create compelling joint offerings for their customers.

Technical Requirements

- Expand beyond a single data center to decrease latency to the American Midwest and east coast
- Create a backup strategy
- Increase security of data transfer from equipment to the data center
- Improve data in the data warehouse
- Use customer and equipment data to anticipate customer needs

Application 1: Data ingest

A custom Python application reads uploaded data files from a single server and writes to the data warehouse.

Compute

Windows Server 2008 R2

- 16 CPUs
- 128 GB of RAM
- 10 TB local HDD storage

Application 2: Reporting

An off-the-shelf application that business analysts use to run a daily report to see what equipment needs repair. Only 2 analysts of a team of 10 (5 west coast, 5 east coast) can connect to the reporting application at a time.

Compute

Off-the-shelf application. License tied to number of physical CPUs.

Windows Server 2008 R2

- 16 CPUs
- 32 GB of RAM
- 500 GB HDD

Data warehouse

- A single PostgreSQL server
- RedHat Linux
- 64 CPUs
- 128 GB of RAM
- 4x 6TB HDD in RAID 0

Executive Statement

Our competitive advantage has always been in our manufacturing process, with our ability to build better vehicles for lower cost than our competitors. However, new products with different approaches are constantly being developed, and I'm concerned that we lack the skills to undergo the next wave of transformations in our industry. My goals are to build our skills while addressing immediate market needs through incremental innovations.

Practice Exam

Your web application has several VM instances running within a VPC. You want to restrict communications between instances to only the paths and ports you authorize, but you don't want to rely on static IP addresses or subnets because the app can autoscale. How should you restrict communications?

- A. Use separate VPCs to restrict traffic
- B. Use firewall rules based on network tags attached to the compute instances**
- C. Use Cloud DNS and only allow connections from authorized hostnames
- D. Use service accounts and configure the web application particular service accounts to have access

Correct Answer: B

You are using Cloud SQL as the database backend for a large CRM deployment. You want to scale as usage increases and ensure that you don't run out of storage, maintain 75% CPU usage cores, and keep replication lag below 60 seconds. What are the correct steps to meet your requirements?

- A. 1. Enable automatic storage increase for the instance.**
2. Create a Stackdriver alert when CPU usage exceeds 75%, and change the instance type to reduce CPU usage.
3. Create a Stackdriver alert for replication lag, and shard the database to reduce replication time.

- B. 1. Enable automatic storage increase for the instance.**
2. Change the instance type to a 32-core machine type to keep CPU usage below 75%.
3. Create a Stackdriver alert for replication lag, and shard the database to reduce replication time.

- C. 1. Create a Stackdriver alert when storage exceeds 75%, and increase the available storage on the instance to create more space.**
2. Deploy memcached to reduce CPU load.
3. Change the instance type to a 32-core machine type to reduce replication lag.

Google Cloud Certified Professional Cloud Architect Definitive Guide

- D. 1. Create a Stackdriver alert when storage exceeds 75%, and increase the available storage on the instance to create more space.
2. Deploy memcached to reduce CPU load.
3. Create a Stackdriver alert for replication lag, and change the instance type to a 32-core machine type to reduce replication lag.

Correct Answer: A

You are tasked with building an online analytical processing (OLAP) marketing analytics and reporting tool. This requires a relational database that can operate on hundreds of terabytes of data. What is the Google-recommended tool for such applications?

- A. Cloud Spanner, because it is globally distributed
- B. Cloud SQL, because it is a fully managed relational database
- C. Cloud Firestore, because it offers real-time synchronization across devices
- D. BigQuery, because it is designed for large-scale processing of tabular data**

Correct Answer: D

You have deployed an application to Kubernetes Engine, and are using the Cloud SQL proxy container to make the Cloud SQL database available to the services running on Kubernetes. You are notified that the application is reporting database connection issues. Your company policies require a post-mortem. What should you do?

- A. Use gcloud sql instances restart.
- B. Validate that the Service Account used by the Cloud SQL proxy container still has the Cloud Build Editor role.
- C. In the GCP Console, navigate to Stackdriver Logging. Consult logs for Kubernetes Engine and Cloud SQL.**
- D. In the GCP Console, navigate to Cloud SQL. Restore the latest backup. Use kubectl to restart all pods.

Correct Answer: C

Your company is running a stateless application on a Compute Engine instance. The application is used heavily during regular business hours and lightly outside of business hours. Users are reporting that the application is slow during peak hours. You need to optimize the application's performance. What should you do?

- A. Create a snapshot of the existing disk. Create an instance template from the snapshot. Create an autoscaled managed instance group from the instance template.
- B. Create a snapshot of the existing disk. Create a custom image from the snapshot. Create an autoscaled managed instance group from the custom image.
- C. Create a custom image from the existing disk. Create an instance template from the custom image. Create an autoscaled managed instance group from the instance template.**
- D. Create an instance template from the existing disk. Create a custom image from the instance template. Create an autoscaled managed instance group from the custom image.

Correct Answer: C

You are running a cluster on Kubernetes Engine to serve a web application. Users are reporting that a specific part of the application is not responding anymore.

You notice that all pods of your deployment keep restarting after 2 seconds.

The application writes logs to standard output. You want to inspect the logs to find the cause of the issue. Which approach can you take?

- A. Review the Stackdriver logs for each Compute Engine instance that is serving as a node in the cluster.
- B. Review the Stackdriver logs for the specific Kubernetes Engine container that is serving the unresponsive part of the application.**
- C. Connect to the cluster using gcloud credentials and connect to a container in one of the pods to read the logs.
- D. Review the Serial Port logs for each Compute Engine instance that is serving as a node in the cluster.

Correct Answer: B

Google Cloud Certified Professional Cloud Architect Definitive Guide

You are using a single Cloud SQL instance to serve your application from a specific zone. You want to introduce high availability. What should you do?

- A. Create a read replica instance in a different region
- B. Create a failover replica instance in a different region**
- C. Create a read replica instance in the same region, but in a different zone
- D. Create a failover replica instance in the same region, but in a different zone

Correct Answer: B

Your company wants to start using Google Cloud resources but wants to retain their on-premises Active Directory domain controller for identity management. What should you do?

- A. Use the Admin Directory API to authenticate against the Active Directory domain controller.
- B. Use Google Cloud Directory Sync to synchronize Active Directory usernames with cloud identities and configure SAML SSO.**
- C. Use Cloud Identity-Aware Proxy configured to use the on-premises Active Directory domain controller as an identity provider.
- D. Use Compute Engine to create an Active Directory (AD) domain controller that is a replica of the on-premises AD domain controller using Google Cloud Directory Sync.

Correct Answer: B

Your customer wants to capture multiple GBs of aggregate real-time key performance indicators (KPIs) from their game servers running on Google Cloud Platform and monitor the KPIs with low latency. How should they capture the KPIs?

- A. Store time-series data from the game servers in Google Bigtable, and view it using Google Data Studio.
- B. Output custom metrics to Stackdriver from the game servers, and create a Dashboard in Stackdriver Monitoring Console to view them.**
- C. Schedule BigQuery load jobs to ingest analytics files uploaded to Cloud Storage every ten minutes, and visualize the results in Google Data Studio.
- D. Insert the KPIs into Cloud Datastore entities, and run ad hoc analysis and visualizations of them in Cloud Datalab.

Correct Answer: B

You have a Python web application with many dependencies that requires 0.1 CPU cores and 128 MB of memory to operate in production. You want to monitor and maximize machine utilization. You also want to reliably deploy new versions of the application. Which set of steps should you take?

A. Perform the following:

1. Create a managed instance group with f1-micro type machines.
2. Use a startup script to clone the repository, check out the production branch, install the dependencies, and start the Python app.
3. Restart the instances to automatically deploy new production releases.

B. Perform the following:

1. Create a managed instance group with n1-standard-1 type machines.
2. Build a Compute Engine image from the production branch that contains all of the dependencies and automatically starts the Python app.
3. Rebuild the Compute Engine image, and update the instance template to deploy new production releases.

C. Perform the following:

1. Create a Kubernetes Engine cluster with n1-standard-1 type machines.
2. Build a Docker image from the production branch with all of the dependencies, and tag it with the version number.
3. Create a Kubernetes Deployment with the imagePullPolicy set to "IfNotPresent" in the staging namespace, and then promote it to the production namespace after testing.

D. Perform the following:

1. Create a Kubernetes Engine cluster with n1-standard-4 type machines.
2. Build a Docker image from the master branch with all of the dependencies, and tag it with "latest".
3. Create a Kubernetes Deployment in the default namespace with the imagePullPolicy set to "Always". Restart the pods to automatically deploy new production releases.

Correct Answer: C

Google Cloud Certified Professional Cloud Architect Definitive Guide

You need to upload files from your on-premises environment to Cloud Storage. You want the files to be encrypted on Cloud Storage using customer-supplied encryption keys. What should you do?

- A. Supply the encryption key in a .boto configuration file. Use gsutil to upload the files.
- B. Supply the encryption key using gcloud config. Use gsutil to upload the files to that bucket.
- C. Use gsutil to upload the files, and use the flag --encryption-key to supply the encryption key.
- D. Use gsutil to create a bucket, and use the flag --encryption-key to supply the encryption key. Use gsutil to upload the files to that bucket.

Correct Answer: A

You are creating an App Engine application that uses Cloud Datastore as its persistence layer. You need to retrieve several root entities for which you have the identifiers. You want to minimize the overhead in operations performed by Cloud Datastore. What should you do?

- A. Create the Key object for each Entity and run a batch get operation
- B. Create the Key object for each Entity and run multiple get operations, one operation for each entity
- C. Use the identifiers to create a query filter and run a batch query operation
- D. Use the identifiers to create a query filter and run multiple query operations, one operation for each entity

Correct Answer: A

You are designing an application for use only during business hours. For the minimum viable product release, you'd like to use a managed product that automatically "scales to zero" so you don't incur costs when there is no activity. Which primary compute resource should you choose?

- A. Cloud Functions
- B. Compute Engine
- C. Kubernetes Engine**
- D. AppEngine flexible environment

Correct Answer: C

Google Cloud Certified Professional Cloud Architect Definitive Guide

You need to evaluate your team readiness for a new GCP project. You must perform the evaluation and create a skills gap plan incorporates the business goal of cost optimization. Your team has deployed two GCP projects successfully to date. What should you do?

- A. Allocate budget for team training. Set a deadline for the new GCP project.
- B. Allocate budget for team training. Create a roadmap for your team to achieve Google Cloud certification based on job role.**
- C. Allocate budget to hire skilled external consultants. Set a deadline for the new GCP project.
- D. Allocate budget to hire skilled external consultants. Create a roadmap for your team to achieve Google Cloud certification based on job role.

Correct Answer: B

The development team has provided you with a Kubernetes Deployment file. You have no infrastructure yet and need to deploy the application. What should you do?

- A. Use gcloud to create a Kubernetes cluster. Use Deployment Manager to create the deployment.
- B. Use gcloud to create a Kubernetes cluster. Use kubectl to create the deployment.**
- C. Use kubectl to create a Kubernetes cluster. Use Deployment Manager to create the deployment.
- D. Use kubectl to create a Kubernetes cluster. Use kubectl to create the deployment.

Correct Answer: B

You need to set up Microsoft SQL Server on GCP. Management requires that there's no downtime in case of a data center outage in any of the zones within a GCP region. What should you do?

- A. Configure a Cloud SQL instance with high availability enabled.
- B. Configure a Cloud Spanner instance with a regional instance configuration.
- C. Set up SQL Server on Compute Engine, using Always on Availability Groups using Windows Failover Clustering. Place nodes in different subnets.
- D. Set up SQL Server Always on Availability Groups using Windows Failover Clustering. Place nodes in different zones.**

Correct Answer: D

Your web application must comply with the requirements of the European Union's General Data Protection Regulation (GDPR). You are responsible for the technical architecture of your web application. What should you do?

- A. Ensure that your web application only uses native features and services of Google Cloud Platform, because Google already has various certifications and provides "pass-on" compliance when you use native features.
- B. Enable the relevant GDPR compliance setting within the GCP Console for each of the services in use within your application.**
- C. Ensure that Cloud Security Scanner is part of your test planning strategy in order to pick up any compliance gaps.
- D. Define a design for the security of data in your web application that meets GDPR requirements.

Correct Answer: B

Your company is using BigQuery as its enterprise data warehouse. Data is distributed over several Google Cloud projects. All queries on BigQuery need to be billed on a single project. You want to make sure that no query costs are incurred on the projects that contain the data. Users should be able to query the datasets, but not edit them. How should you configure users' access roles?

- A. Add all users to a group. Grant the group the role of BigQuery user on the billing project and BigQuery dataViewer on the projects that contain the data.
- B. Add all users to a group. Grant the group the roles of BigQuery dataViewer on the billing project and BigQuery user on the projects that contain the data.
- C. Add all users to a group. Grant the group the roles of BigQuery jobUser on the billing project and BigQuery dataViewer on the projects that contain the data.**
- D. Add all users to a group. Grant the group the roles of BigQuery dataViewer on the billing project and BigQuery jobUser on the projects that contain the data.

Correct Answer: C

Google Cloud Certified Professional Cloud Architect Definitive Guide

You have developed an application using Cloud ML Engine that recognizes famous paintings from uploaded images. You want to test the application and allow specific people to upload images for the next 24 hours. Not all users have a Google Account. How should you have users upload images?

- A. Have users upload the images to Cloud Storage. Protect the bucket with a password that expires after 24 hours.
- B. Have users upload the images to Cloud Storage using a signed URL that expires after 24 hours.**
- C. Create an App Engine web application where users can upload images. Configure App Engine to disable the application after 24 hours. Authenticate users via Cloud Identity.
- D. Create an App Engine web application where users can upload images for the next 24 hours. Authenticate users via Cloud Identity.

Correct Answer: B

You have an application deployed on Kubernetes Engine using a Deployment named echo-deployment. The deployment is exposed using a Service called echoservice. You need to perform an update to the application with minimal downtime to the application. What should you do?

- A. Use kubectl set image deployment/echo-deployment <new-image>
- B. Use the rolling update functionality of the Instance Group behind the Kubernetes cluster
- C. Update the deployment yaml file with the new container image. Use kubectl delete deployment/echo-deployment and kubectl create -f <yaml-file>
- D. Update the service yaml file which the new container image. Use kubectl delete service/echo-service and kubectl create -f <yaml-file>**

Correct Answer: D

You are working in a highly secured environment where public Internet access from the Compute Engine VMs is not allowed. You do not yet have a VPN connection to access an on-premises file server. You need to install specific software on a Compute Engine instance. How should you install the software?

Google Cloud Certified Professional Cloud Architect Definitive Guide

- A. Upload the required installation files to Cloud Storage. Configure the VM on a subnet with a Private Google Access subnet. Assign only an internal IP address to the VM. Download the installation files to the VM using gsutil.
- B. Upload the required installation files to Cloud Storage and use firewall rules to block all traffic except the IP address range for Cloud Storage. Download the files to the VM using gsutil.
- C. Upload the required installation files to Cloud Source Repositories. Configure the VM on a subnet with a Private Google Access subnet. Assign only an internal IP address to the VM. Download the installation files to the VM using gcloud.
- D. Upload the required installation files to Cloud Source Repositories and use firewall rules to block all traffic except the IP address range for Cloud Source Repositories. Download the files to the VM using gsutil.

Correct Answer: A

You are working in a highly secured environment where public Internet access from the Compute Engine VMs is not allowed. You do not yet have a VPN connection to access an on-premises file server. You need to install specific software on a Compute Engine instance. How should you install the software?

- A. Upload the required installation files to Cloud Storage. Configure the VM on a subnet with a Private Google Access subnet. Assign only an internal IP address to the VM. Download the installation files to the VM using gsutil.
- B. Upload the required installation files to Cloud Storage and use firewall rules to block all traffic except the IP address range for Cloud Storage. Download the files to the VM using gsutil.
- C. Upload the required installation files to Cloud Source Repositories. Configure the VM on a subnet with a Private Google Access subnet. Assign only an internal IP address to the VM. Download the installation files to the VM using gcloud.
- D. Upload the required installation files to Cloud Source Repositories and use firewall rules to block all traffic except the IP address range for Cloud Source Repositories. Download the files to the VM using gsutil.

Correct Answer: A

Google Cloud Certified Professional Cloud Architect Definitive Guide

Your company is moving 75 TB of data into Google Cloud. You want to use Cloud Storage and follow Google-recommended practices. What should you do?

- A. Move your data onto a Transfer Appliance. Use a Transfer Appliance Rehydrator to decrypt the data into Cloud Storage.
- B. Move your data onto a Transfer Appliance. Use Cloud Dataprep to decrypt the data into Cloud Storage.
- C. Install gsutil on each server that contains data. Use resumable transfers to upload the data into Cloud Storage.
- D. Install gsutil on each server containing data. Use streaming transfers to upload the data into Cloud Storage.**

Correct Answer: D

You are deploying an application on App Engine that needs to integrate with an on-premises database. For security purposes, your on-premises database must not be accessible through the public Internet. What should you do?

- A. Deploy your application on App Engine standard environment and use App Engine firewall rules to limit access to the open on-premises database.
- B. Deploy your application on App Engine standard environment and use Cloud VPN to limit access to the on-premises database.
- C. Deploy your application on App Engine flexible environment and use App Engine firewall rules to limit access to the on-premises database.
- D. Deploy your application on App Engine flexible environment and use Cloud VPN to limit access to the on-premises database.**

Correct Answer: D

You have an application that will run on Compute Engine. You need to design an architecture that takes into account a disaster recovery plan that requires your application to fail over to another region in case of a regional outage. What should you do?

- A. Deploy the application on two Compute Engine instances in the same project but in a different region. Use the first instance to serve traffic, and use the HTTP load balancing service to fail over to the standby instance in case of a disaster.

Google Cloud Certified Professional Cloud Architect Definitive Guide

- B. Deploy the application on a Compute Engine instance. Use the instance to serve traffic, and use the HTTP load balancing service to fail over to an instance on your premises in case of a disaster.
- C. Deploy the application on two Compute Engine instance groups, each in the same project but in a different region. Use the first instance group to serve traffic, and use the HTTP load balancing service to fail over to the standby instance group in case of a disaster.**
- D. Deploy the application on two Compute Engine instance groups, each in separate project and a different region. Use the first instance group to server traffic, and use the HTTP load balancing service to fail over to the standby instance in case of a disaster.

Correct Answer: C

Google Cloud Platform resources are managed hierarchically using organization, folders, and projects. When Cloud Identity and Access Management (IAM) policies exist at these different levels, what is the effective policy at a particular node of the hierarchy?

- A. The effective policy is determined only by the policy set at the node
- B. The effective policy is the policy set at the node and restricted by the policies of its ancestors
- C. The effective policy is the union of the policy set at the node and policies inherited from its ancestors**
- D. The effective policy is the intersection of the policy set at the node and policies inherited from its ancestors

Correct Answer: C

You are migrating you're on-premises solution to Google Cloud in several phases. You will use Cloud VPN to maintain a connection between your on-premises systems and Google Cloud until the migration is completed. You want to make sure all you're on-premises systems remain reachable during this period. How should you organize your networking in Google Cloud?

- A. Use the same IP range on Google Cloud as you use on-premises
- B. Use the same IP range on Google Cloud as you use on-premises for your primary IP range and use a secondary range that does not overlap with the range you use on-premises
- C. Use an IP range on Google Cloud that does not overlap with the range you use on-premises**
- D. Use an IP range on Google Cloud that does not overlap with the range you use on-premises for your primary IP range and use a secondary range with the same IP range as you use on-premises

Correct Answer: C

You have found an error in your App Engine application caused by missing Cloud Datastore indexes. You have created a YAML file with the required indexes and want to deploy these new indexes to Cloud Datastore. What should you do?

- A. Point gcloud datastore create-indexes to your configuration file
- B. Upload the configuration file to the App Engine's default Cloud Storage bucket, and have App Engine detect the new indexes
- C. In the GCP Console, use Datastore Admin to delete the current indexes and upload the new configuration file
- D. Create an HTTP request to the built-in python module to send the index configuration file to your application

Correct Answer: A

Your company has multiple on-premises systems that serve as sources for reporting. The data has not been maintained well and has become degraded over time. You want to use Google-recommended practices to detect anomalies in your company data. What should you do?

- A. Upload your files into Cloud Storage. Use Cloud Datalab to explore and clean your data.
- B. Upload your files into Cloud Storage. Use Cloud Dataprep to explore and clean your data.**
- C. Connect Cloud Datalab to your on-premises systems. Use Cloud Datalab to explore and clean your data.
- D. Connect Cloud Dataprep to your on-premises systems. Use Cloud Dataprep to explore and clean your data.

Correct Answer: B

Your company is migrating its on-premises data center into the cloud. As part of the migration, you want to integrate Kubernetes Engine for workload orchestration. Parts of your architecture must also be PCI DSS-compliant. Which of the following is most accurate?

- A. App Engine is the only compute platform on GCP that is certified for PCI DSS hosting.
- B. Kubernetes Engine cannot be used under PCI DSS because it is considered shared hosting.

C. Kubernetes Engine and GCP provide the tools you need to build a PCI DSS-compliant environment.

D. All Google Cloud services are usable because Google Cloud Platform is certified PCI-compliant.

Correct Answer: C

You have an outage in your Compute Engine managed instance group: all instance keep restarting after 5 seconds. You have a health check configured, but autoscaling is disabled. Your colleague, who is a Linux expert, offered to look into the issue. You need to make sure that he can access the VMs. What should you do?

- A. Grant your colleague the IAM role of project Viewer
- B. Perform a rolling restart on the instance group
- C. Disable the health check for the instance group. Add his SSH key to the project-wide SSH keys**
- D. Disable autoscaling for the instance group. Add his SSH key to the project-wide SSH Keys

Correct Answer: C

You are building a continuous deployment pipeline for a project stored in a Git source repository and want to ensure that code changes can be verified deploying to production. What should you do?

- A. Use Spinnaker to deploy builds to production using the red/black deployment strategy so that changes can easily be rolled back.
- B. Use Spinnaker to deploy builds to production and run tests on production deployments.
- C. Use Jenkins to build the staging branches and the master branch. Build and deploy changes to production for 10% of users before doing a complete rollout.**
- D. Use Jenkins to monitor tags in the repository. Deploy staging tags to a staging environment for testing. After testing, tag the repository for production and deploy that to the production environment.

Correct Answer: C

Google Cloud Certified Professional Cloud Architect Definitive Guide

You are analyzing and defining business processes to support your startup's trial usage of GCP, and you don't yet know what consumer demand for your product will be. Your manager requires you to minimize GCP service costs and adhere to Google best practices. What should you do?

- A. Utilize free tier and sustained use discounts. Provision a staff position for service cost management.
- B. Utilize free tier and sustained use discounts. Provide training to the team about service cost management.**
- C. Utilize free tier and committed use discounts. Provision a staff position for service cost management.
- D. Utilize free tier and committed use discounts. Provide training to the team about service cost management.

Correct Answer: B

As part of implementing their disaster recovery plan, your company is trying to replicate their production MySQL database from their private data center to their GCP project using a Google Cloud VPN connection. They are experiencing latency issues and a small amount of packet loss that is disrupting the replication. What should they do?

- A. Configure their replication to use UDP.
- B. Configure a Google Cloud Dedicated Interconnect.**
- C. Restore their database daily using Google Cloud SQL.
- D. Add additional VPN connections and load balance them.
- E. Send the replicated transaction to Google Cloud Pub/Sub.

Correct Answer: B

Your customer support tool logs all email and chat conversations to Cloud Bigtable for retention and analysis. What is the recommended approach for sanitizing this data of personally identifiable information or payment card information before initial storage?

- A. Hash all data using SHA256
- B. Encrypt all data using elliptic curve cryptography
- C. De-identify the data with the Cloud Data Loss Prevention API**

Google Cloud Certified Professional Cloud Architect Definitive Guide

D. Use regular expressions to find and redact phone numbers, email addresses, and credit card numbers

Correct Answer: C

You are using Cloud Shell and need to install a custom utility for use in a few weeks. Where can you store the file so it is in the default execution path and persists across sessions?

- A. ~/bin
- B. Cloud Storage
- C. /google/scripts
- D. /usr/local/bin

Correct Answer: A

You want to create a private connection between your instances on Compute Engine and your on-premises data center. You require a connection of at least 20 Gbps. You want to follow Google-recommended practices. How should you set up the connection?

- A. Create a VPC and connect it to your on-premises data center using Dedicated Interconnect.**
- B. Create a VPC and connect it to your on-premises data center using a single Cloud VPN.
- C. Create a Cloud Content Delivery Network (Cloud CDN) and connect it to your on-premises data center using Dedicated Interconnect.
- D. Create a Cloud Content Delivery Network (Cloud CDN) and connect it to your on-premises datacenter using a single Cloud VPN.

Correct Answer: A

You are designing a mobile chat application. You want to ensure people cannot spoof chat messages, by providing a message were sent by a specific user. What should you do?

- A. Tag messages client side with the originating user identifier and the destination user.
- B. Encrypt the message client side using block-based encryption with a shared key.
- C. Use public key infrastructure (PKI) to encrypt the message client side using the originating user's private key.**
- D. Use a trusted certificate authority to enable SSL connectivity between the client application and the server.

Correct Answer: C

Your organization wants to control IAM policies for different departments independently, but centrally. Which approach should you take?

- A. Multiple Organizations with multiple Folders
- B. Multiple Organizations, one for each department
- C. A single Organization with Folders for each department**
- D. A single Organization with multiple projects, each with a central owner

Correct Answer: C

You created a pipeline that can deploy your source code changes to your infrastructure in instance groups for self-healing. One of the changes negatively affects your key performance indicator. You are not sure how to fix it, and investigation could take up to a week. What should you do?

- A. Log in to a server, and iterate on the fix locally
- B. Revert the source code change, and rerun the deployment pipeline**
- C. Log into the servers with the bad code change, and swap in the previous code
- D. Change the instance group template to the previous one, and delete all instances

Correct Answer: B

You deploy your custom Java application to Google App Engine. It fails to deploy and gives you the following stack trace. What should you do?

```
java.lang.SecurityException: SHA1 digest error for
com/Altostrat/CloakedServlet.class
    at com.google.appengine.runtime.Request.process
-d36f818a24b8cf1d (Request.java)
    at
sun.security.util.ManifestEntryVerifier.verify
(ManifestEntryVerifier.java:210)
    at java.util.jar.JarVerifier.processEntry
(JarVerifier.java:218)
    at java.util.jar.JarVerifier.update
(JarVerifier.java:205)
    at
java.util.jar.JarVerifiersVerifierStream.read
(JarVerifier.java:428)
    at sun.misc.Resource.getBytes
(Resource.java:124)
    at java.net.URLClassLoader.defineClass
(URLClassLoader.java:273)
    at sun.reflect.GeneratedMethodAccessor5.invoke
(Unknown Source)
    at
sun.reflect.DelegatingMethodAccessorImpl.invoke
(DelegatingMethodAccessorImpl.java:43)
    at java.lang.reflect.Method.invoke
(Method.java:616)
    at java.lang.ClassLoader.loadClass
(ClassLoader.java:266)
```

- A. Upload missing JAR files and redeploy your application.
- B. Digitally sign all of your JAR files and redeploy your application
- C. Recompile the Cloaked Servlet class using and MD5 hash instead of SHA1

Correct Answer: A

Google Cloud Certified Professional Cloud Architect Definitive Guide

Your organization has a 3-tier web application deployed in the same network on Google Cloud Platform. Each tier (web, API, and database) scales independently of the others. Network traffic should flow through the web to the API tier and then on to the database tier. Traffic should not flow between the web and the database tier. How should you configure the network?

- A. Add each tier to a different subnetwork
- B. Set up software-based firewalls on individual VMs
- C. Add tags to each tier and set up routes to allow the desired traffic flow
- D. Add tags to each tier and set up firewall rules to allow the desired traffic flow**

Correct Answer: D

Your development team has installed a new Linux kernel module on the batch servers in Google Compute Engine (GCE) virtual machines (VMs) to speed up the nightly batch process. Two days after the installation, 50% of the batch servers failed the nightly batch run. You want to collect details on the failure to pass back to the development team. Which three actions should you take? Choose 3 answers.

- A. Use Stackdriver Logging to search for the module log entries
- B. Read the debug GCE Activity log using the API or Cloud Console
- C. Use gcloud or Cloud Console to connect to the serial console and observe the logs**
- D. Identify whether a live migration event of the failed server occurred, using in the activity log**
- E. Adjust the Google Stackdriver timeline to match the failure time, and observe the batch server metrics**
- F. Export a debug VM into an image, and run the image on a local server where kernel log messages will be displayed on the native screen

Correct Answer: C, D, E

Your company wants to try out the cloud with low risk. They want to archive approximately 100 TB of their log data to the cloud and test the analytics features available to them there, while also retaining that data as a long-term disaster recovery backup. Which two steps should you take? Choose 2 answers.

- A. Load logs into Google BigQuery**
- B. Load logs into Google Cloud SQL

- C. Import logs into Google Stackdriver
- D. Insert logs into Google Cloud Bigtable
- E. Upload log files into Google Cloud Storage**

Correct Answer: A, E

A development manager is building a new application. He asks you to review his requirements and identify what cloud technologies he can use to meet them. The application must:

- 1. Be based on open-source technology for cloud portability**
- 2. Dynamically scale compute capacity based on demand**
- 3. Support continuous software delivery**
- 4. Run multiple segregated copies of the same application stack**
- 5. Deploy application bundles using dynamic templates**
- 6. Route network traffic to specific services based on URL**

Which combination of technologies will meet all of his requirements?

- A. Google Kubernetes Engine, Jenkins, and Helm
- B. Google Kubernetes Engine and Cloud Load Balancing
- C. Google Kubernetes Engine and Cloud Deployment Manager
- D. Google Kubernetes Engine, Jenkins, and Cloud Load Balancing**

Correct Answer: D

You have created several pre-emptible Linux virtual machine instances using Google Compute Engine. You want to properly shut down your application before the virtual machines are pre-empted. What should you do?

- A. Create a shutdown script named k99.shutdown in the /etc/rc.6.d/ directory
- B. Create a shutdown script registered as a xinetd service in Linux and configure a Stackdriver endpoint check to call the service
- C. Create a shutdown script and use it as the value for a new metadata entry with the key shutdown-script in the Cloud Platform Console when you create the new virtual machine instance**
- D. Create a shutdown script, registered as a xinetd service in Linux, and use the gcloud compute instances add-metadata command to specify the service URL

as the value for a new metadata entry with the key shutdownscript-url

Correct Answer: C

You are designing a large distributed application with 30 microservices. Each of your distributed microservices needs to connect to a database back-end. You want to store the credentials securely. Where should you store the credentials?

- A. In the source code
- B. In an environment variable
- C. In a secret management system**
- D. In a config file that has restricted access through ACLs

Correct Answer: C

A lead engineer wrote a custom tool that deploys virtual machines in the legacy data center. He wants to migrate the custom tool to the new cloud environment.

You want to advocate for the adoption of Google Cloud Deployment Manager.

What are two business risks of migrating to Cloud Deployment Manager? Choose 2 answers.

- A. Cloud Deployment Manager uses Python
- B. Cloud Deployment Manager APIs could be deprecated in the future
- C. Cloud Deployment Manager is unfamiliar to the company's engineers**
- D. Cloud Deployment Manager requires a Google APIs service account to run
- E. Cloud Deployment Manager can be used to permanently delete cloud resources
- F. Cloud Deployment Manager only supports automation of Google Cloud resources**

Correct Answer: C, F

Google Cloud Certified Professional Cloud Architect Definitive Guide

Your organization requires that metrics from all applications be retained for 5 years for future analysis in possible legal proceedings. Which approach should you use?

- A. Grant the security team access to the logs in each Project
- B. Configure Stackdriver Monitoring for all Projects, and export to BigQuery
- C. Configure Stackdriver Monitoring for all Projects with the default retention policies
- D. Configure Stackdriver Monitoring for all Projects, and export to Google Cloud Storage**

Correct Answer: D

Your company has decided to build a backup replica of their on-premises user authentication PostgreSQL database on Google Cloud Platform. The database is 4 TB, and large updates are frequent. Replication requires private address space communication. Which networking approach should you use?

- A. Google Cloud Dedicated Interconnect
- B. Google Cloud VPN connected to the data center network**
- C. A NAT and TLS translation gateway installed on-premises
- D. A Google Compute Engine instance with a VPN server installed connected to the data center network

Correct Answer: B

Auditors visit your teams every 12 months and ask to review all the Google Cloud Identity and Access Management (Cloud IAM) policy changes in the previous 12 months. You want to streamline and expedite the analysis and audit process. What should you do?

- A. Create custom Google Stackdriver alerts and send them to the auditor
- B. Enable Logging export to Google BigQuery and use ACLs and views to scope the data shared with the auditor**
- C. Use cloud functions to transfer log entries to Google Cloud SQL and use ACLs and views to limit an auditor's view
- D. Enable Google Cloud Storage (GCS) log export to audit logs into a GCS bucket and delegate access to the bucket

Correct Answer: B

During a high traffic portion of the day, one of your relational databases crashes, but the replica is never promoted to a master. You want to avoid this in the future. What should you do?

- A. Use a different database
- B. Choose larger instances for your database
- C. Create snapshots of your database more regularly
- D. Implement routinely scheduled failovers of your databases**

Correct Answer: D

A small number of API requests to your microservices-based application take a very long time. You know that each request to the API can traverse many services. You want to know which service takes the longest in those cases. What should you do?

- A. Set timeouts on your application so that you can fail requests faster
- B. Send custom metrics for each of your requests to Stackdriver Monitoring
- C. Use Stackdriver Monitoring to look for insights that show when your API latencies are high
- D. Instrument your application with Stackdriver Trace in order to break down the request latencies at each microservice**

Correct Answer: D

One of the developers on your team deployed their application in Google Container Engine with the Dockerfile below. They report that their application deployments are taking too long. You want to optimize this Dockerfile for faster deployment times without adversely affecting the app's functionality. Which two actions should you take? Choose 2 answers.

- A. Remove Python after running pip
- B. Remove dependencies from requirements.txt
- C. Use a slimmed-down base image like Alpine Linux**
- D. Use larger machine types for your Google Container Engine node pools**
- E. Copy the source after the package dependencies (Python and pip) are installed

Correct Answer: C, D

Your solution is producing performance bugs in production that you did not see in staging and test environments. You want to adjust your test and deployment procedures to avoid this problem in the future. What should you do?

- A. Deploy fewer changes to production
- B. Deploy smaller changes to production
- C. Increase the load on your test and staging environments**
- D. Deploy changes to a small subset of users before rolling out to production

Correct Answer: C

You want to optimize the performance of an accurate, real-time, weather-charting application. The data comes from 50,000 sensors sending 10 readings a second, in the format of a timestamp and sensor reading. Where should you store the data?

- A. Google BigQuery
- B. Google Cloud SQL
- C. Google Cloud Bigtable**
- D. Google Cloud Storage

Correct Answer: C

Your company's user-feedback portal comprises a standard LAMP stack replicated across two zones. It is deployed in the us-central1 region and uses autoscaled managed instance groups on all layers, except the database. Currently, only a small group of select customers have access to the portal. The portal meets a 99,99% availability SLA under these conditions. However next quarter, your company will be making the portal available to all users, including unauthenticated users. You need to develop a resiliency testing strategy to ensure the system maintains the SLA once they introduce additional user load. What should you do?

- A. Capture existing user's input, and replay captured user load until autoscale is triggered on all layers. At the same time, terminate all resources in one of the zones
- B. Create synthetic random user input, replay synthetic load until autoscale logic is triggered on at least one layer, and introduce "chaos" to the system by terminating random resources on both zones

C. Expose the new system to a larger group of users, and increase group size each day until autoscale logic is triggered on all layers. At the same time, terminate random resources on both zones

D. Capture existing user's input, and replay captured user load until resource utilization crosses 80%. Also, derive estimated number of users based on existing user's usage of the app, and deploy enough resources to handle 200% of expected load

Correct Answer: C

Your company is forecasting a sharp increase in the number and size of Apache Spark and Hadoop jobs being run on your local datacenter. You want to utilize the cloud to help you scale this upcoming demand with the least amount of operations work and code change. Which product should you use?

- A. Google Cloud Dataflow
- B. Google Cloud Dataproc**
- C. Google Compute Engine
- D. Google Kubernetes Engine

Correct Answer: B

The database administration team has asked you to help them improve the performance of their new database server running on Google Compute Engine. The database is for importing and normalizing their performance statistics and is built with MySQL running on Debian Linux. They have an n1-standard-8 virtual machine with 80 GB of SSD persistent disk. What should they change to get better performance from this system?

- A. Increase the virtual machine's memory to 64 GB
- B. Create a new virtual machine running PostgreSQL
- C. Dynamically resize the SSD persistent disk to 500 GB**
- D. Migrate their performance metrics warehouse to BigQuery
- E. Modify all of their batch jobs to use bulk inserts into the database

Correct Answer: C

Google Cloud Certified Professional Cloud Architect Definitive Guide

Your application needs to process credit card transactions. You want the smallest scope of Payment Card Industry (PCI) compliance without compromising the ability to analyze transactional data and trends relating to which payment methods are used. How should you design your architecture?

- A. Create a tokenizer service and store only tokenized data
- B. Create separate projects that only process credit card data
- C. Create separate subnetworks and isolate the components that process credit card data
- D. Streamline the audit discovery phase by labeling all of the virtual machines (VMs) that process PCI data
- E. Enable Logging export to Google BigQuery and use ACLs and views to scope the data shared with the auditor**

Correct Answer: E

You have been asked to select the storage system for the click-data of your company's large portfolio of websites. This data is streamed in from a custom website analytics package at a typical rate of 6,000 clicks per minute. With bursts of up to 8,500 clicks per second. It must have been stored for future analysis by your data science and user experience teams. Which storage infrastructure should you choose?

- A. Google Cloud SQL
- B. Google Cloud Bigtable**
- C. Google Cloud Storage
- D. Google Cloud Datastore

Correct Answer: B

You are creating a solution to remove backup files older than 90 days from your backup Cloud Storage bucket. You want to optimize ongoing Cloud Storage spend. What should you do?

- A. Write a lifecycle management rule in XML and push it to the bucket with gsutil
- B. Write a lifecycle management rule in JSON and push it to the bucket with gsutil**
- C. Schedule a cron script using gsutil ls -lr gs://backups/** to find and remove items older than 90 days

Google Cloud Certified Professional Cloud Architect Definitive Guide

D. Schedule a cron script using gsutil ls -l gs://backups/** to find and remove items older than 90 days and schedule it with cron

Correct Answer: B

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

- A. Google Cloud Dataproc
- B. Google Cloud Dataflow**
- C. Google Container Engine with Bigtable
- D. Google Compute Engine with Google BigQuery

Correct Answer: B

Your customer is receiving reports that their recently updated Google App Engine application is taking approximately 30 seconds to load for some of their users. This behavior was not reported before the update. What strategy should you take?

- A. Work with your ISP to diagnose the problem
- B. Open a support ticket to ask for network capture and flow data to diagnose the problem, then roll back your application
- C. Roll back to an earlier known good release initially, then use Stackdriver Trace and Logging to diagnose the problem in a development/test/staging environment**
- D. Roll back to an earlier known good release, then push the release again at a quieter period to investigate. Then use Stackdriver Trace and Logging to diagnose the problem

Correct Answer: C

Google Cloud Certified Professional Cloud Architect Definitive Guide

A production database virtual machine on Google Compute Engine has an ext4-formatted persistent disk for data files. The database is about to run out of storage space. How can you remediate the problem with the least amount of downtime?

- A. In the Cloud Platform Console, increase the size of the persistent disk and use the resize2fs command in Linux.**
- B. Shut down the virtual machine, use the Cloud Platform Console to increase the persistent disk size, then restart the virtual machine
- C. In the Cloud Platform Console, increase the size of the persistent disk and verify the new space is ready to use with the fdisk command in Linux
- D. In the Cloud Platform Console, create a new persistent disk attached to the virtual machine, format and mount it, and configure the database service to move the files to the new disk
- E. In the Cloud Platform Console, create a snapshot of the persistent disk restore the snapshot to a new larger disk, unmount the old disk, mount the new disk and restart the database service

Correct Answer: A

You write a Python script to connect to Google BigQuery from a Google Compute Engine virtual machine. The script is printing errors that it cannot connect to BigQuery. What should you do to fix the script?

- A. Install the latest BigQuery API client library for Python
- B. Run your script on a new virtual machine with the BigQuery access scope enabled**
- C. Create a new service account with BigQuery access and execute your script with that user
- D. Install the bq component for gcloud with the command gcloud components install bq.

Correct Answer: B

Your customer is moving an existing corporate application to Google Cloud Platform from an on-premises data center. The business owners require minimal user disruption. There are strict security team requirements for storing passwords. What authentication strategy should they use?

- A. Use G Suite Password Sync to replicate passwords into Google
- B. Federate authentication via SAML 2.0 to the existing Identity Provider
- C. Provision users in Google using the Google Cloud Directory Sync tool**

D. Ask users to set their Google password to match their corporate password

Correct Answer: C

You need to reduce the number of unplanned rollbacks of erroneous production deployments in your company's web hosting platform. Improvement to the QA/Test processes accomplished an 80% reduction. Which additional two approaches can you take to further reduce the rollbacks? Choose 2 answers.

- A. Introduce a green-blue deployment model
- B. Replace the QA environment with canary releases
- C. Fragment the monolithic platform into microservices**
- D. Reduce the platform's dependency on relational database systems
- E. Replace the platform's relational database systems with a NoSQL database

Correct Answer: A, C

To reduce costs, the Director of Engineering has required all developers to move their development infrastructure resources from on-premises virtual machines (VMs) to Google Cloud Platform. These resources go through multiple start/stop events during the day and require state to persist. You have been asked to design the process of running a development environment in Google Cloud while providing cost visibility to the finance department.

Which two steps should you take? Choose 2 answers.

- A. Use the --no-auto-delete flag on all persistent disks and stop the VM**
- B. Use the --auto-delete flag on all persistent disks and terminate the VM
- C. Apply VM CPU utilization label and include it in the BigQuery billing export
- D. Use Google BigQuery billing export and labels to associate cost to groups**
- E. Store all state into local SSD, snapshot the persistent disks, and terminate the VM
- F. Store all state in Google Cloud Storage, snapshot the persistent disks, and terminate the VM

Correct Answer: A, D

Your company wants to track whether someone is present in a meeting room reserved for a scheduled meeting. There are 1000 meeting rooms across 5 offices on 3 continents. Each room is equipped with a motion sensor that reports its status every second. The data from the motion detector includes only a sensor ID and several different discrete items of information. Analysts will use this data, together with information about account owners and office locations. Which database type should you use?

- A. Flat file
- B. NoSQL
- C. Relational**
- D. Blobstore

Correct Answer: C

You set up an autoscaling instance group to serve web traffic for an upcoming launch. After configuring the instance group as a backend service to an HTTP(S) load balancer, you notice that virtual machine (VM) instances are being terminated and re-launched every minute. The instances do not have a public IP address. You have verified the appropriate web response is coming from each instance using the curl command. You want to ensure the backend is configured correctly. What should you do?

- A. Ensure that a firewall rule exists to allow source traffic on HTTP/HTTPS to reach the load balancer.
- B. Assign a public IP to each instance and configure a firewall rule to allow the load balancer to reach the instance public IP.
- C. Ensure that a firewall rule exists to allow load balancer health checks to reach the instances in the instance group.**
- D. Create a tag on each instance with the name of the load balancer. Configure a firewall rule with the name of the load balancer as the source and the instance tag as the destination.

Correct Answer: C

Google Cloud Certified Professional Cloud Architect Definitive Guide

An application development team believes their current logging tool will not meet their needs for their new cloud-based product. They want a better tool to capture errors and help them analyze their historical log data. You want to help them find a solution that meets their needs. What should you do?

- A. Direct them to download and install the Google StackDriver logging agent
- B. Send them a list of online resources about logging best practices
- C. Help them define their requirements and assess viable logging tools
- D. Help them upgrade their current tool to take advantage of any new features

Correct Answer: A

Your company plans to migrate a multi-petabyte data set to the cloud. The data set must be available 24hrs a day. Your business analysts have experience only with using a SQL interface. How should you store the data to optimize it for ease of analysis?

- A. Load data into Google BigQuery
- B. Insert data into Google Cloud SQL
- C. Put flat files into Google Cloud Storage
- D. Stream data into Google Cloud Datastore

Correct Answer: A

The operations manager asks you for a list of recommended practices that she should consider when migrating a J2EE application to the cloud. Which three practices should you recommend? Choose 3 answers.

- A. Port the application code to run on Google App Engine
- B. Integrate Cloud Dataflow into the application to capture real-time metrics
- C. Instrument the application with a monitoring tool like Stackdriver Debugger
- D. Select an automation framework to reliably provision the cloud infrastructure
- E. Deploy a continuous integration tool with automated testing in a staging environment
- F. Migrate from MySQL to a managed NoSQL database like Google Cloud Datastore or Bigtable

Correct Answer: A, C, E

A news feed web service has the following code running on Google App Engine. During peak load, users report that they can see news articles they already viewed. What is the most likely cause of this problem?

```
import news
from flask import Flask, redirect, request
from flask.ext.api import status
from google.appengine.api import users

app = Flask(__name__)
sessions = {}

@app.route("/")
def homepage():
    user = users.get_current_user()
    if not user:
        return "Invalid login",
status.HTTP_401_UNAUTHORIZED

    if user not in sessions:
        sessions[user] = {"viewed": []}

    news_articles = news.get_new_news (user, sessions [user]
["viewed"])
    sessions [user] ["viewed"] += [n["id"] for n
in news_articles]

    return news.render(news_articles)

if __name__ == "__main__":
    app.run()
```

- A. The session variable is local to just a single instance
- B. The session variable is being overwritten in Cloud Datastore**
- C. The URL of the API needs to be modified to prevent caching
- D. The HTTP Expires header needs to be set to -1 stop caching

Correct Answer: B

One of your primary business objectives is being able to trust the data stored in your application. You want to log all changes to the application data. How can you design your logging system to verify authenticity of your logs?

- A. Write the log concurrently in the cloud and on premises
- B. Use a SQL database and limit who can modify the log table
- C. Digitally sign each timestamp and log entry and store the signature**
- D. Create a JSON dump of each log entry and store it in Google Cloud Storage

Correct Answer: C

Your company has decided to make a major revision of their API in order to create better experiences for their developers. They need to keep the old version of the API available and deployable, while allowing new customers and testers to try out the new API. They want to keep the same SSL and DNS records in place to serve both APIs. What should they do?

- A. Configure a new load balancer for the new version of the API
- B. Reconfigure old clients to use a new endpoint for the new API
- C. Have the old API forward traffic to the new API based on the path
- D. Use separate backend pools for each API path behind the load balancer**

Correct Answer: D

Your marketing department wants to send out a promotional email campaign. The development team wants to minimize direct operation management. They project a wide range of possible customer responses, from 100 to 500,000 clickthrough per day. The link leads to a simple website that explains the promotion and collects user information and preferences. Which infrastructure should you recommend? Choose 2 answers.

- A. Use Google App Engine to serve the website and Google Cloud Datastore to store user data.**
- B. Use a Google Container Engine cluster to serve the website and store data to persistent disk.
- C. Use a managed instance group to serve the website and Google Cloud Bigtable to store user data.**
- D. Use a single Compute Engine virtual machine (VM) to host a web server, backend by Google Cloud SQL.

Correct Answer: A, C

Your company just finished a rapid lift and shift to Google Compute Engine for your compute needs. You have another 9 months to design and deploy a more cloud-native solution. Specifically, you want a system that is no-ops and autoscaling. Which two compute products should you choose? Choose 2 answers.

- A. Compute Engine with containers
- B. Google Kubernetes Engine with containers**
- C. Google App Engine Standard Environment**
- D. Compute Engine with custom instance types
- E. Compute Engine with managed instance groups

Correct Answer: B, C

Your company places a high value on being responsive and meeting customer needs quickly. Their primary business objectives are release speed and agility.

You want to reduce the chance of security errors being accidentally introduced.

Which two actions can you take? Choose 2 answers.

- A. Ensure every code check-in is peer reviewed by a security SME
- B. Use source code security analysers as part of the CI/CD pipeline**
- C. Ensure you have stubs to unit test all interfaces between components
- D. Enable code signing and a trusted binary repository integrated with your CI/CD pipeline
- E. Run a vulnerability security scanner as part of your continuous-integration /continuous-delivery (CI/CD) pipeline**

Correct Answer: B, E

You want to enable your running Google Kubernetes Engine cluster to scale as demand for your application changes. What should you do?

- A. Add additional nodes to your Kubernetes Engine cluster using the following command:gcloud container clusters resize CLUSTER_Name – -size 10
- B. Add a tag to the instances in the cluster with the following command:gcloud compute instances add-tags INSTANCE – -tags enableautoscaling max-nodes=10
- C. Update the existing Kubernetes Engine cluster with the following command:gcloud alpha container clusters update mycluster – -enableautoscaling- -min-nodes=1 – -max-nodes=10**
- D. Create a new Kubernetes Engine cluster with the following command: gcloud alpha container clusters create mycluster – -enableautoscaling- -min-nodes=1 – -max-nodes=10 and redeploy your application

Correct Answer: C

Your company runs several databases on a single MySQL instance. They need to take backups of a specific database at regular intervals. The backup activity needs to complete as quickly as possible and cannot be allowed to impact disk performance. How should you configure the storage?

- A. Configure a cron job to use the gcloud tool to take regular backups using persistent disk snapshots.
- B. Mount a Local SSD volume as the backup location. After the backup is complete, use gsutil to move the backup to Google Cloud Storage.**
- C. Use gcsfuse to mount a Google Cloud Storage bucket as a volume directly on the instance and write backups to the mounted location using mysqldump.
- D. Mount additional persistent disk volumes onto each virtual machine (VM) instance in a RAID10 array and use LVM to create snapshots to send to Cloud Storage

Correct Answer: B

Google Cloud Certified Professional Cloud Architect Definitive Guide

You are helping the QA team to roll out a new load-testing tool to test the scalability of your primary cloud services that run on Google Compute Engine with Cloud Bigtable.

Which three requirements should they include? Choose 3 answers.

- A. Ensure that the load tests validate the performance of Cloud Bigtable
- B. Create a separate Google Cloud project to use for the load-testing environment**
- C. Schedule the load-testing tool to regularly run against the production environment
- D. Ensure all third-party systems your services use is capable of handling high load
- E. Instrument the production services to record every transaction for replay by the load-testing tool**
- F. Instrument the load-testing tool and the target services with detailed logging and metrics collection**

Correct Answer: B, E, F

Your customer is moving their corporate applications to Google Cloud Platform. The security team wants detailed visibility of all projects in the organization. You provision the Google Cloud Resource Manager and set up yourself as the org admin. What Google Cloud Identity and Access Management (Cloud IAM) roles should you give to the security team?

- A. Org viewer, project owner
- B. Org viewer, project viewer**
- C. Org admin, project browser
- D. Project owner, network admin

Correct Answer: B

You want to make a copy of a production Linux virtual machine in the US-Central region. You want to manage and replace the copy easily if there are changes on the production virtual machine. You will deploy the copy as a new instance in a different project in the US-East region. What steps must you take?

- A. Use the Linux dd and netcat commands to copy and stream the root disk contents to a new virtual machine instance in the US-East region.
- B. Create a snapshot of the root disk and select the snapshot as the root disk when you create a new virtual machine instance in the US-East region.

Google Cloud Certified Professional Cloud Architect Definitive Guide

C. Create an image file from the root disk with Linux dd command, create a new virtual machine instance in the US-East region

D. Create a snapshot of the root disk, create an image file in Google Cloud Storage from the snapshot, and create a new virtual machine instance in the US-East region using the image file the root disk.

Correct Answer: D

A recent audit revealed that a new network was created in your GCP project. In this network, a GCE instance has an SSH port open to the world. You want to discover this network's origin.

What should you do?

A. Search for Create VM entry in the Stackdriver alerting console

B. Navigate to the Activity page in the home section. Set category to Data Access and search for Create VM entry

C. In the Logging section of the console, specify GCE Network as the logging section. Search for the Create Insert entry

D. Connect to the GCE instance using project SSH keys. Identify previous logins in system logs, and match these with the project owners list

Correct Answer: C

The application reliability team at your company this added a debug feature to their backend service to send all server events to Google Cloud Storage for eventual analysis. The event records are at least 50 KB and at most 15 MB and are expected to peak at 3,000 events per second. You want to minimize data loss. Which process should you implement?

A. – Append metadata to file body

- Compress individual files
- Name files with serverName – Timestamp
- Create a new bucket if bucket is older than 1 hour and save individual files to the new bucket. Otherwise, save files to existing bucket.

B. – Batch every 10,000 events with a single manifest file for metadata

- Compress event files and manifest file into a single archive file

- Name files using serverName – EventSequence
- Create a new bucket if bucket is older than 1 day and save the single archive file to the new bucket.
Otherwise, save the single archive file to existing bucket.

C. – Compress individual files

- Name files with serverName – EventSequence
- Save files to one bucket
- Set custom metadata headers for each object after saving

D. – Append metadata to file body

- Compress individual files
- Name files with a random prefix pattern
- Save files to one bucket

Correct Answer: D

For this question, refer to the Dress4Win case study. You want to ensure that your on-premises architecture meets business requirements before you migrate your solution. What change in the on-premises architecture should you make?

- A. Replace RabbitMQ with Google Pub/Sub.
- B. Downgrade MySQL to v5.7, which is supported by Cloud SQL for MySQL.
- C. Resize compute resources to match predefined Compute Engine machine types.**
- D. Containerize the micro services and host them in Google Kubernetes Engine.

Correct Answer: C

Your company's test suite is a custom C++ application that runs tests throughout each day on Linux virtual machines. The full test suite takes several hours to complete, running on a limited number of on-premises servers reserved for testing. Your company wants to move the testing infrastructure to the cloud, to reduce the amount of time it takes to fully test a change to the system, while changing the tests as little as possible. Which cloud infrastructure should you recommend?

- A. Google Compute Engine unmanaged instance groups and Network Load Balancer

B. Google Compute Engine managed instance groups with auto-scaling

- C. Google Cloud Dataproc to run Apache Hadoop jobs to process each test
- D. Google App Engine with Google StackDriver for logging

Correct Answer: B

A lead software engineer tells you that his new application design uses websockets and HTTP sessions that are not distributed across the web servers. You want to help him ensure his application will run properly on Google Cloud Platform. What should you do?

- A. Help the engineer to convert his websocket code to use HTTP streaming
- B. Review the encryption requirements for websocket connections with the security team
- C. Meet with the cloud operations team and the engineer to discuss load balancer options**
- D. Help the engineer re-design the application to use a distributed user session service that does not rely on websockets and HTTP sessions.

Correct Answer: C

For this question, refer to the Dress4Win case study. You are responsible for the security of data stored in Cloud Storage for your company, Dress4Win. You have already created a set of Google Groups and assigned the appropriate users to those groups. You should use Google best practices and implement the simplest design to meet the requirements. Considering Dress4Win's business and technical requirements, what should you do?

- A. Assign custom IAM roles to the Google Groups you created in order to enforce security requirements. Encrypt data with a customer-supplied encryption key when storing files in Cloud Storage.
- B. Assign custom IAM roles to the Google Groups you created in order to enforce security requirements. Enable default storage encryption before storing files in Cloud Storage.
- C. Assign predefined IAM roles to the Google Groups you created in order to enforce security requirements. Utilize Google's default encryption at rest when storing files in Cloud Storage.**
- D. Assign predefined IAM roles to the Google Groups you created in order to enforce security requirements. Ensure that the default Cloud KMS key is set before storing files in Cloud Storage.

Correct Answer: C

Google Cloud Certified Professional Cloud Architect Definitive Guide

For this question, refer to the Dress4Win case study. Dress4Win is expected to grow to 10 times its size in 1 year with a corresponding growth in data and traffic that mirrors the existing patterns of usage. The CIO has set the target of migrating production infrastructure to the cloud within the next 6 months. How will you configure the solution to scale for this growth without making major application changes and still maximize the ROI?

- A. Migrate the web application layer to App Engine, and MySQL to Cloud Datastore, and NAS to Cloud Storage. Deploy RabbitMQ, and deploy Hadoop servers using Deployment Manager.
- B. Migrate RabbitMQ to Cloud Pub/Sub, Hadoop to BigQuery, and NAS to Compute Engine with Persistent Disk storage. Deploy Tomcat, and deploy Nginx using Deployment Manager.
- C. Implement managed instance groups for Tomcat and Nginx. Migrate MySQL to Cloud SQL, RabbitMQ to Cloud Pub/Sub, Hadoop to Cloud Dataproc, and NAS to Compute Engine with Persistent Disk storage.
- D. Implement managed instance groups for the Tomcat and Nginx. Migrate MySQL to Cloud SQL, RabbitMQ to Cloud Pub/Sub, Hadoop to Cloud Dataproc, and NAS to Cloud Storage.**

Correct Answer: D

For this question, refer to the Dress4Win case study. Considering the given business requirements, how would you automate the deployment of web and transactional data layers?

- A. Deploy Nginx and Tomcat using Cloud Deployment Manager to Compute Engine. Deploy a Cloud SQL server to replace MySQL. Deploy Jenkins using Cloud Deployment Manager.**
- B. Deploy Nginx and Tomcat using Cloud Launcher. Deploy a MySQL server using Cloud Launcher. Deploy Jenkins to Compute Engine using Cloud Deployment Manager scripts.
- C. Migrate Nginx and Tomcat to App Engine. Deploy a Cloud Datastore server to replace the MySQL server in a high-availability configuration. Deploy Jenkins to Compute Engine using Cloud Launcher.
- D. Migrate Nginx and Tomcat to App Engine. Deploy a MySQL server using Cloud Launcher. Deploy Jenkins to Compute Engine using Cloud Launcher.

Correct Answer: A

Google Cloud Certified Professional Cloud Architect Definitive Guide

For this question, refer to the Dress4Win case study. Which of the compute services should be migrated as -is and would still be an optimized architecture for performance in the cloud?

- A. Web applications deployed using App Engine standard environment
- B. RabbitMQ deployed using an unmanaged instance group
- C. Hadoop/Spark deployed using Cloud Dataproc Regional in High Availability mode**
- D. Jenkins, monitoring, bastion hosts, security scanners services deployed on custom machine types

Correct Answer: C

For this question, refer to the Dress4Win case study. To be legally compliant during an audit, Dress4Win must be able to give insights in all administrative actions that modify the configuration or metadata of resources on Google Cloud. What should you do?

- A. Use Stackdriver Trace to create a trace list analysis.
- B. Use Stackdriver Monitoring to create a dashboard on the project's activity.
- C. Enable Cloud Identity-Aware Proxy in all projects, and add the group of Administrators as a member.
- D. Use the Activity page in the GCP Console and Stackdriver Logging to provide the required insight.**

Correct Answer: D

Dress4Win has end-to-end tests covering 100% of their endpoints. They want to ensure that the move to the cloud does not introduce any new bugs. Which additional testing methods should the developers employ to prevent an outage?

- A. They should enable Google Stackdriver Debugger on the application code to show errors in the code.
- B. They should add additional unit tests and production scale load tests on their cloud staging environment.
- C. They should run the end-to-end tests in the cloud staging environment to determine if the code is working as intended.**
- D. They should add canary tests so developers can measure how much of an impact the new release causes to latency.

Correct Answer: C

You want to ensure Dress4Win's sales and tax records remain available for infrequent viewing by auditors for at least 10 years. Cost optimization is your top priority. Which cloud services should you choose?

- A. Google Cloud Storage Coldline to store the data, and gsutil to access the data.
- B. Google Cloud Storage Nearline to store the data, and gsutil to access the data.
- C. Google Bigtable with US or EU as location to store the data, and gcloud to access the data.
- D. BigQuery to store the data, and a web server cluster in a managed instance group to access the data. Google Cloud SQL mirrored across two distinct regions to store the data, and a Redis cluster in a managed instance group to access the data.

Correct Answer: A

The current Dress4win system architecture has high latency to some customers because it is located in one data center. As of a future evaluation and optimizing for performance in the cloud, Dressss4win wants to distribute its system architecture to multiple locations when Google cloud platform. Which approach should they use?

- A. Use regional managed instance groups and a global load balancer to increase performance because the regional managed instance group can grow instances in each region separately based on traffic.
- B. Use a global load balancer with a set of virtual machines that forward the requests to a closer group of virtual machines managed by your operations team.
- C. Use regional managed instance groups and a global load balancer to increase reliability by providing automatic failover between zones in different regions.
- D. Use a global load balancer with a set of virtual machines that forward the requests to a closer group of virtual machines as part of a separate managed instance groups.**

Correct Answer: D

Google Cloud Certified Professional Cloud Architect Definitive Guide

Dress4Win would like to become familiar with deploying applications to the cloud by successfully deploying some applications quickly, as is. They have asked for your recommendation. What should you advise?

- A. Identify self-contained applications with external dependencies as a first move to the cloud.
- B. Identify enterprise applications with internal dependencies and recommend these as a first move to the cloud.
- C. Suggest moving their in-house databases to the cloud and continue serving requests to on-premise applications.**
- D. Recommend moving their message queuing servers to the cloud and continue handling requests to on-premise applications.

Correct Answer: C

Dress4Win has asked you for advice on how to migrate their on-premises MySQL deployment to the cloud. They want to minimize downtime and performance impact to their on-premises solution during the migration. Which approach should you recommend?

- A. Create a dump of the on-premises MySQL master server, and then shut it down, upload it to the cloud environment, and load into a new MySQL cluster.
- B. Setup a MySQL replica server/slave in the cloud environment, and configure it for asynchronous replication from the MySQL master server on-premises until cutover.**
- C. Create a new MySQL cluster in the cloud, configure applications to begin writing to both on-premises and cloud MySQL masters, and destroy the original cluster at cutover.
- D. Create a dump of the MySQL replica server into the cloud environment, load it into: Google Cloud Datastore, and configure applications to read/write to Cloud Datastore at cutover.

Correct Answer: B

Dress4Win has configured a new uptime check with Google Stackdriver for several of their legacy services. The Stackdriver dashboard is not reporting the services as healthy. What should they do?

- A. Install the Stackdriver agent on all of the legacy web servers.
- B. In the Cloud Platform Console download the list of the uptime servers' IP addresses and create an inbound firewall rule**

Google Cloud Certified Professional Cloud Architect Definitive Guide

- C. Configure their load balancer to pass through the User-Agent HTTP header when the value matches GoogleStackdriverMonitoring-UptimeChecks (<https://cloud.google.com/monitoring>)
- D. Configure their legacy web servers to allow requests that contain user-Agent HTTP header when the value matches GoogleStackdriverMonitoring- UptimeChecks (<https://cloud.google.com/monitoring>)

Correct Answer: B

As part of their new application experience, Dress4Wm allows customers to upload images of themselves. The customer has exclusive control over who may view these images. Customers should be able to upload images with minimal latency and also be shown their images quickly on the main application page when they log in. Which configuration should Dress4Win use?

- A. Store image files in a Google Cloud Storage bucket. Use Google Cloud Datastore to maintain metadata that maps each customer's ID and their image files.**
- B. Store image files in a Google Cloud Storage bucket. Add custom metadata to the uploaded images in Cloud Storage that contains the customer's unique ID.
- C. Use a distributed file system to store customers' images. As storage needs increase, add more persistent disks and/or nodes. Assign each customer a unique ID, which sets each file's owner attribute, ensuring privacy of images.
- D. Use a distributed file system to store customers' images. As storage needs increase, add more persistent disks and/or nodes. Use a Google Cloud SQL database to maintain metadata that maps each customer's ID to their image files.

Correct Answer: A

At Dress4Win, an operations engineer wants to create a low-cost solution to remotely archive copies of database backup files. The database files are compressed tar files stored in their current data center. How should he proceed?

- A. Create a cron script using gsutil to copy the files to a Coldline Storage bucket.**
- B. Create a cron script using gsutil to copy the files to a Regional Storage bucket.
- C. Create a Cloud Storage Transfer Service Job to copy the files to a Coldline Storage bucket.
- D. Create a Cloud Storage Transfer Service job to copy the files to a Regional Storage bucket.

Correct Answer: A

Dress4Win has asked you to recommend machine types they should deploy their application servers to. How should you proceed?

- A. Perform a mapping of the on-premises physical hardware cores and RAM to the nearest machine types in the cloud.
- B. Recommend that Dress4Win deploy application servers to machine types that offer the highest RAM to CPU ratio available.
- C. Recommend that Dress4Win deploy into production with the smallest instances available, monitor them over time, and scale the machine type up until the desired performance is reached.
- D. Identify the number of virtual cores and RAM associated with the application server virtual machines align them to a custom machine type in the cloud, monitor performance, and scale the machine types up until the desired performance is reached.**

Correct Answer: D

As part of Dress4Win's plans to migrate to the cloud, they want to be able to set up a managed logging and monitoring system so they can handle spikes in their traffic load. They want to ensure that:

- * The infrastructure can be notified when it needs to scale up and down to handle the ebb and flow of usage throughout the day* Their administrators are notified automatically when their application reports errors.**
- * They can filter their aggregated logs down in order to debug one piece of the application across many hosts**

Which Google StackDriver features should they use?

- A. Logging, Alerts, Insights, Debug
- B. Monitoring, Trace, Debug, Logging
- C. Monitoring, Logging, Alerts, Error Reporting
- D. Monitoring, Logging, Debug, Error Report**

Correct Answer: D

Google Cloud Certified Professional Cloud Architect Definitive Guide

For this question, refer to the TerramEarth case study. You are asked to design a new architecture for the ingestion of the data of the 200,000 vehicles that are connected to a cellular network. You want to follow Google-recommended practices. Considering the technical requirements, which components should you use for the ingestion of the data?

- A. Google Kubernetes Engine with an SSL Ingress
- B. Cloud IoT Core with public/private key pairs**
- C. Compute Engine with project-wide SSH keys
- D. Compute Engine with specific SSH keys

Correct Answer: B

The Dress4Win security team has disabled external SSH access into production virtual machines (VMs) on Google Cloud Platform (GCP).

The operations team needs to remotely manage the VMs, build and push Docker containers, and manage Google Cloud Storage objects. What can they do?

- A. Grant the operations engineer access to use Google Cloud Shell.
- B. Configure a VPN connection to GCP to allow SSH access to the cloud VMs.**
- C. Develop a new access request process that grants temporary SSH access to cloud VMs when an operations engineer needs to perform a task.
- D. Have the development team build an API service that allows the operations team to execute specific remote procedure calls to accomplish their tasks.

Correct Answer: B

For this question, refer to the TerramEarth case study. Considering the technical requirements, how should you reduce the unplanned vehicle downtime in GCP?

- A. Use BigQuery as the data warehouse. Connect all vehicles to the network and stream data into BigQuery using Cloud Pub/Sub and Cloud Dataflow. Use Google Data Studio for analysis and reporting.**
- B. Use BigQuery as the data warehouse. Connect all vehicles to the network and upload gzip files to a Multi-Regional Cloud Storage bucket using gcloud. Use Google Data Studio for analysis and reporting.

Google Cloud Certified Professional Cloud Architect Definitive Guide

C. Use Cloud Dataproc Hive as the data warehouse. Upload gzip files to a MultiRegional Cloud Storage bucket. Upload this data into BigQuery using gcloud. Use Google data Studio for analysis and reporting.

D. Use Cloud Dataproc Hive as the data warehouse. Directly stream data into partitioned Hive tables. Use Pig scripts to analyze data.

Correct Answer: A

For this question, refer to the TerramEarth case study. You need to implement a reliable, scalable GCP solution for the data warehouse for your company, TerramEarth.

Considering the TerramEarth business and technical requirements, what should you do?

A. Replace the existing data warehouse with BigQuery. Use table partitioning.

B. Replace the existing data warehouse with a Compute Engine instance with 96 CPUs.

C. Replace the existing data warehouse with BigQuery. Use federated data sources.

D. Replace the existing data warehouse with a Compute Engine instance with 96 CPUs. Add an additional Compute Engine pre-emptible instance with 32 CPUs.

Correct Answer: A

For this question, refer to the TerramEarth case study. A new architecture that writes all incoming data to BigQuery has been introduced. You notice that the data is dirty, and want to ensure data quality on an automated daily basis while managing cost. What should you do?

A. Set up a streaming Cloud Dataflow job, receiving data by the ingestion process. Clean the data in a Cloud Dataflow pipeline.

B. Create a Cloud Function that reads data from BigQuery and cleans it. Trigger it. Trigger the Cloud Function from a Compute Engine instance.

C. Create a SQL statement on the data in BigQuery, and save it as a view. Run the view daily, and save the result to a new table.

D. Use Cloud Dataprep and configure the BigQuery tables as the source. Schedule a daily job to clean the data.

Correct Answer: D

Google Cloud Certified Professional Cloud Architect Definitive Guide

For this question, refer to the TerramEarth case study. TerramEarth has decided to store data files in Cloud Storage. You need to configure Cloud Storage lifecycle rule to store 1 year of data and minimize file storage cost. Which two actions should you take?

- A. Create a Cloud Storage lifecycle rule with Age: "30", Storage Class: "Standard", and Action: "Set to Coldline", and create a second GCS life-cycle rule with Age: "365", Storage Class: "Coldline", and Action: "Delete".**
- B. Create a Cloud Storage lifecycle rule with Age: "30", Storage Class: "Coldline", and Action: "Set to Nearline", and create a second GCS life-cycle rule with Age: "91", Storage Class: "Coldline", and Action: "Set to Nearline".
- C. Create a Cloud Storage lifecycle rule with Age: "90", Storage Class: "Standard", and Action: "Set to Nearline", and create a second GCS life-cycle rule with Age: "91", Storage Class: "Nearline", and Action: "Set to Coldline".
- D. Create a Cloud Storage lifecycle rule with Age: "30", Storage Class: "Standard", and Action: "Set to Coldline", and create a second GCS life-cycle rule with Age: "365", Storage Class: "Nearline", and Action: "Delete".

Correct Answer: A

For this question, refer to the TerramEarth case study. You need to implement a reliable, scalable GCP solution for the data warehouse for your company, TerramEarth. Considering the TerramEarth business and technical requirements, what should you do?

- A. Replace the existing data warehouse with BigQuery. Use table partitioning.**
- B. Replace the existing data warehouse with a Compute Engine instance with 96 CPUs.
- C. Replace the existing data warehouse with BigQuery. Use federated data sources.
- D. Replace the existing data warehouse with a Compute Engine instance with 96 CPUs. Add an additional Compute Engine pre-emptible instance with 32 CPUs.

Correct Answer: A

Google Cloud Certified Professional Cloud Architect Definitive Guide

For this question, refer to the TerramEarth case study. To be compliant with European GDPR regulation, TerramEarth is required to delete data generated from its European customers after a period of 36 months when it contains personal data. In the new architecture, this data will be stored in both Cloud Storage and BigQuery. What should you do?

- A. Create a BigQuery table for the European data, and set the table retention period to 36 months. For Cloud Storage, use gsutil to enable lifecycle management using a DELETE action with an Age condition of 36 months.
- B. Create a BigQuery table for the European data, and set the table retention period to 36 months. For Cloud Storage, use gsutil to create a SetStorageClass to NONE action when with an Age condition of 36 months.
- C. Create a BigQuery time-partitioned table for the European data, and set the partition expiration period to 36 months. For Cloud Storage, use gsutil to enable lifecycle management using a DELETE action with an Age condition of 36months.**
- D. Create a BigQuery time-partitioned table for the European data, and set the partition period to 36 months. For Cloud Storage, use gsutil to create a SetStorageClass to NONE action with an Age condition of 36 months.

Correct Answer: C

TerramEarth's 20 million vehicles are scattered around the world. Based on the vehicle's location, its telemetry data is stored in a Google Cloud Storage (GCS) regional bucket (US, Europe, or Asia). The CTO has asked you to run a report on the raw telemetry data to determine why vehicles are breaking down after 100 K miles. You want to run this job on all the data.

What is the most cost-effective way to run this job?

- A. Move all the data into 1 zone, then launch a Cloud Dataproc cluster to run the job**
- B. Move all the data into 1 region, then launch a Google Cloud Dataproc cluster to run the job
- C. Launch a cluster in each region to pre-process and compress the raw data, then move the data into a multi-region bucket and use a Dataproc cluster to finish the job
- D. Launch a cluster in each region to pre-process and compress the raw data, then move the data into a region bucket and use a Cloud Dataproc cluster to finish the job

Correct Answer: A

Google Cloud Certified Professional Cloud Architect Definitive Guide

TerramEarth has equipped all connected trucks with servers and sensors to collect telemetry data. Next year they want to use the data to train machine learning models. They want to store this data in the cloud while reducing costs. What should they do?

- A. Have the vehicle's computer compress the data in hourly snapshots, and store it in a Google Cloud Storage (GCS) Nearline bucket
- B. Push the telemetry data in real-time to a streaming dataflow job that compresses the data, and store it in Google BigQuery
- C. Push the telemetry data in real-time to a streaming dataflow job that compresses the data, and store it in Cloud Bigtable
- D. Have the vehicle's computer compress the data in hourly snapshots, and store it in a GCS Coldline bucket**

Correct Answer: D

Your agricultural division is experimenting with fully autonomous vehicles. You want your architecture to promote strong security during vehicle operation. Which two architectures should you consider? (Choose two.)

- A. Treat every micro service call between modules on the vehicle as untrusted.**
- B. Require IPv6 for connectivity to ensure a secure address space.
- C. Use a trusted platform module (TPM) and verify firmware and binaries on boot.**
- D. Use a functional programming language to isolate code execution cycles.
- E. Use multiple connectivity subsystems for redundancy.
- F. Enclose the vehicle's drive electronics in a Faraday cage to isolate chips.

Correct Answer: A, C

Operational parameters such as oil pressure are adjustable on each of TerramEarth's vehicles to increase their efficiency, depending on their environmental conditions. Your primary goal is to increase the operating efficiency of all 20 million cellular and unconnected vehicles in the field. How can you accomplish this goal?

- A. Have your engineers inspected the data for patterns, and then create an algorithm with rules that make operational adjustments automatically

- B. Capture all operating data, train machine learning models that identify ideal operations, and run locally to make operational adjustments automatically**
- C. Implement a Google Cloud Dataflow streaming job with a sliding window, and use Google Cloud Messaging (GCM) to make operational adjustments automatically
- D. Capture all operating data, train machine learning models that identify ideal operations, and host in Google Cloud Machine Learning (ML) Platform to make operational adjustments automatically

Correct Answer: B

To speed up data retrieval, more vehicles will be upgraded to cellular connections and be able to transmit data to the ETL process. The current FTP process is error-prone and restarts the data transfer from the start of the file when connections fail, which happens often. You want to improve the reliability of the solution and minimize data transfer time on the cellular connections. What should you do?

- A. Use one Google Container Engine cluster of FTP servers. Save the data to a Multi-Regional bucket. Run the ETL process using data in the bucket
- B. Use multiple Google Container Engine clusters running FTP servers located in different regions. Save the data to multi-regional buckets in US, EU, and Asia. Run the ETL process using the data in the bucket
- C. Directly transfer the files to different Google Cloud Multi-Regional Storage bucket locations in US, EU, and Asia using Google APIs over HTTP(S). Run the ETL process using the data in the bucket**
- D. Directly transfer the files to a different Google Cloud Regional Storage bucket location in US, EU, and Asia using Google APIs over HTTP(S). Run the ETL process to retrieve the data from each regional bucket

Correct Answer: C

Google Cloud Certified Professional Cloud Architect Definitive Guide

TerramEarth plans to connect all 20 million vehicles in the field to the cloud. This increases the volume to 20 million 600 byte records a second for 40 TB an hour. How should you design the data ingestion?

- A. Vehicles write data directly to GCS
- B. Vehicles write data directly to Google Cloud Pub/Sub**
- C. Vehicles stream data directly to Google BigQuery
- D. Vehicles continue to write data using the existing system (FTP)

Correct Answer: B

You analysed TerramEarth's business requirement to reduce downtime, and found that they can achieve a majority of time saving by reducing customer's wait time for parts. You decided to focus on reduction of the 3 weeks aggregate reporting time.

Which modifications to the company's processes should you recommend?

- A. Migrate from CSV to binary format, migrate from FTP to SFTP transport, and develop machine learning analysis of metrics
- B. Migrate from FTP to streaming transport, migrate from CSV to binary format, and develop machine learning analysis of metrics
- C. Increase fleet cellular connectivity to 80%, migrate from FTP to streaming transport, and develop machine learning analysis of metrics**
- D. Migrate from FTP to SFTP transport, develop machine learning analysis of metrics, and increase dealer local inventory by a fixed factor

Correct Answer: C

Which of TerramEarth's legacy enterprise processes will experience significant change as a result of increased Google Cloud Platform adoption?

- A. Opex/capex allocation, LAN changes, capacity planning
- B. Capacity planning, TCO calculations, opex/capex allocation**
- C. Capacity planning, utilization measurement, data center expansion
- D. Data Center expansion, TCO calculations, utilization measurement

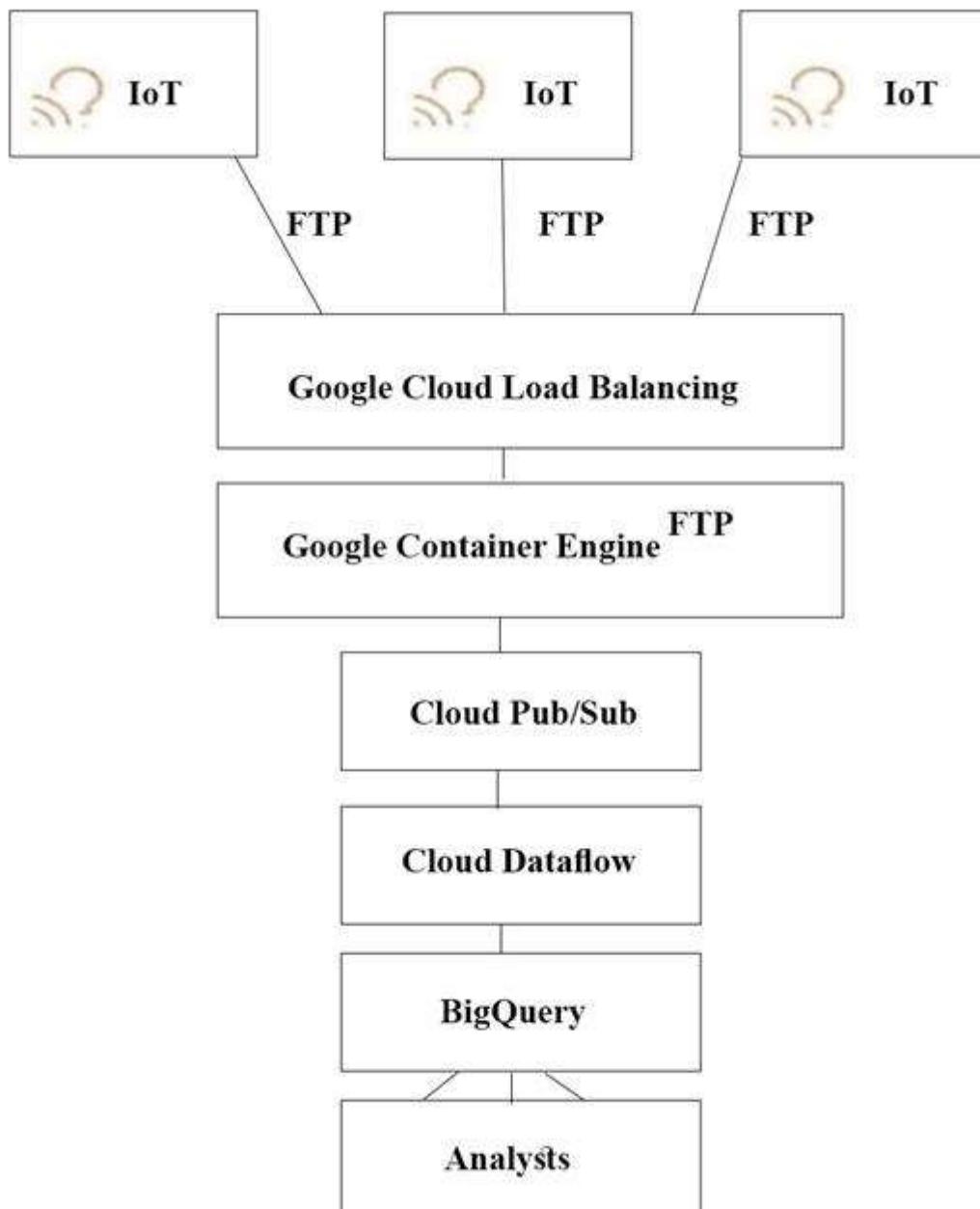
Correct Answer: B

TerramEarth's CTO wants to use the raw data from connected vehicles to help identify approximately when a vehicle in the field will have a catastrophic failure.

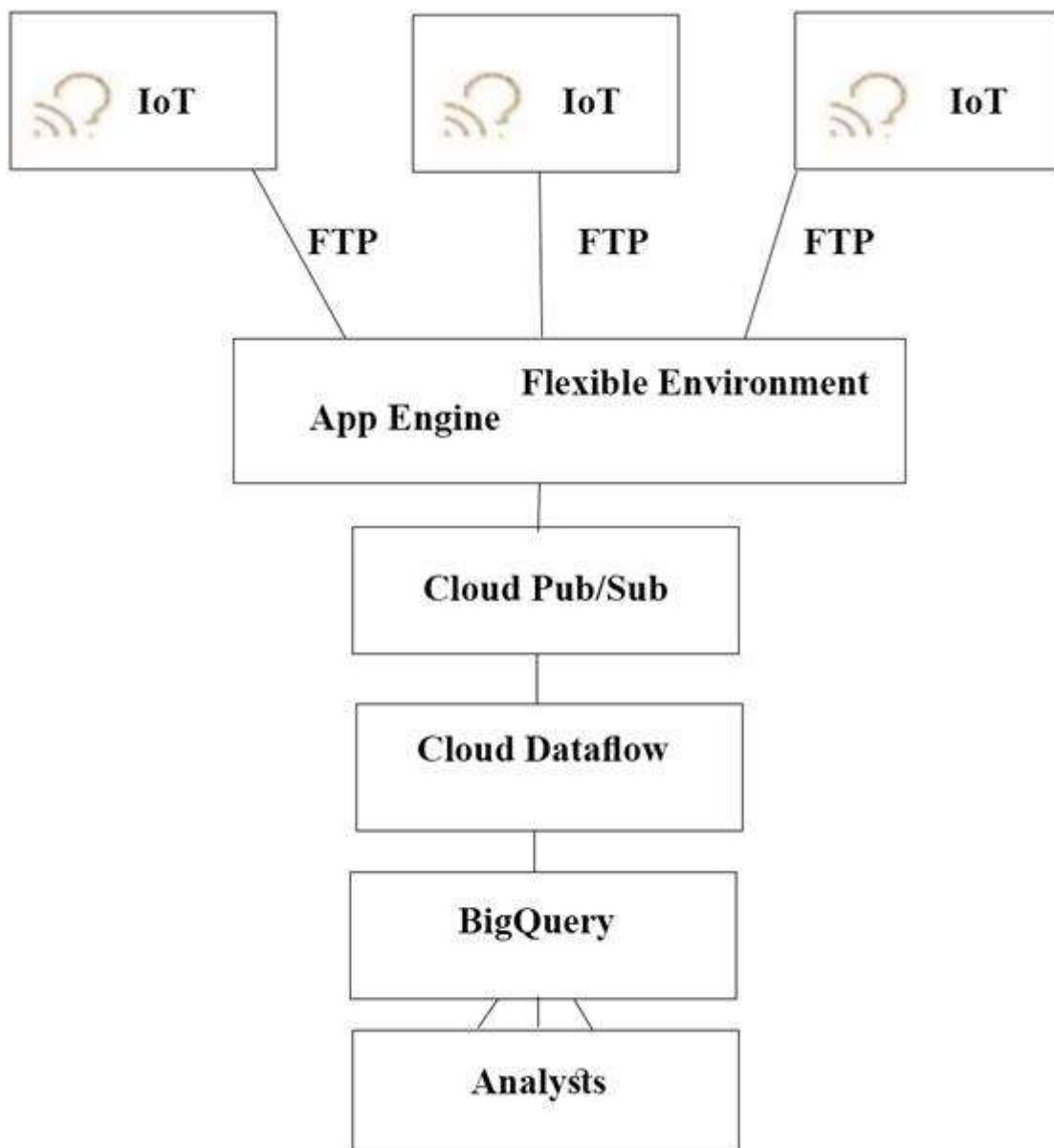
You want to allow analysts to centrally query the vehicle data.

Which architecture should you recommend?

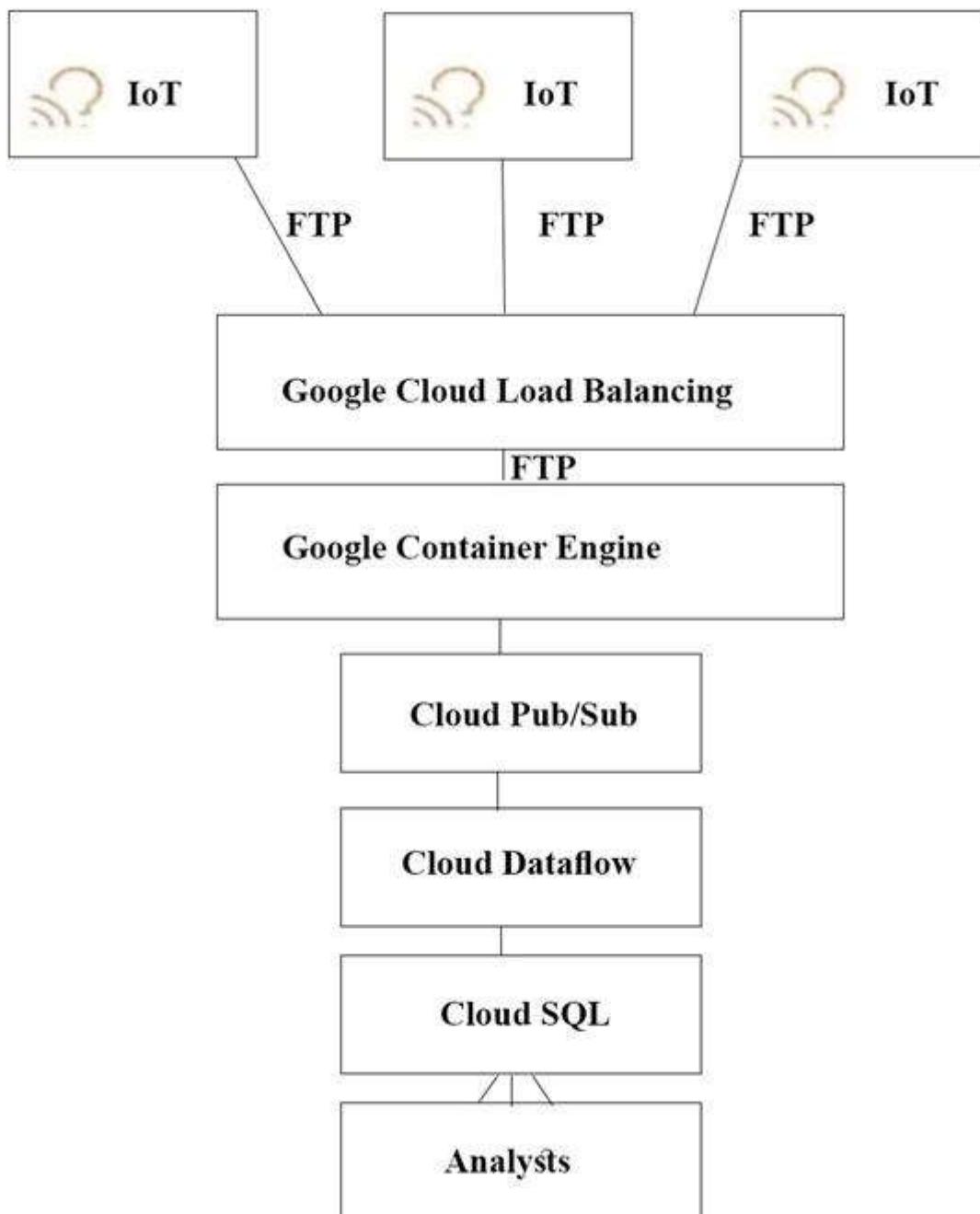
A.



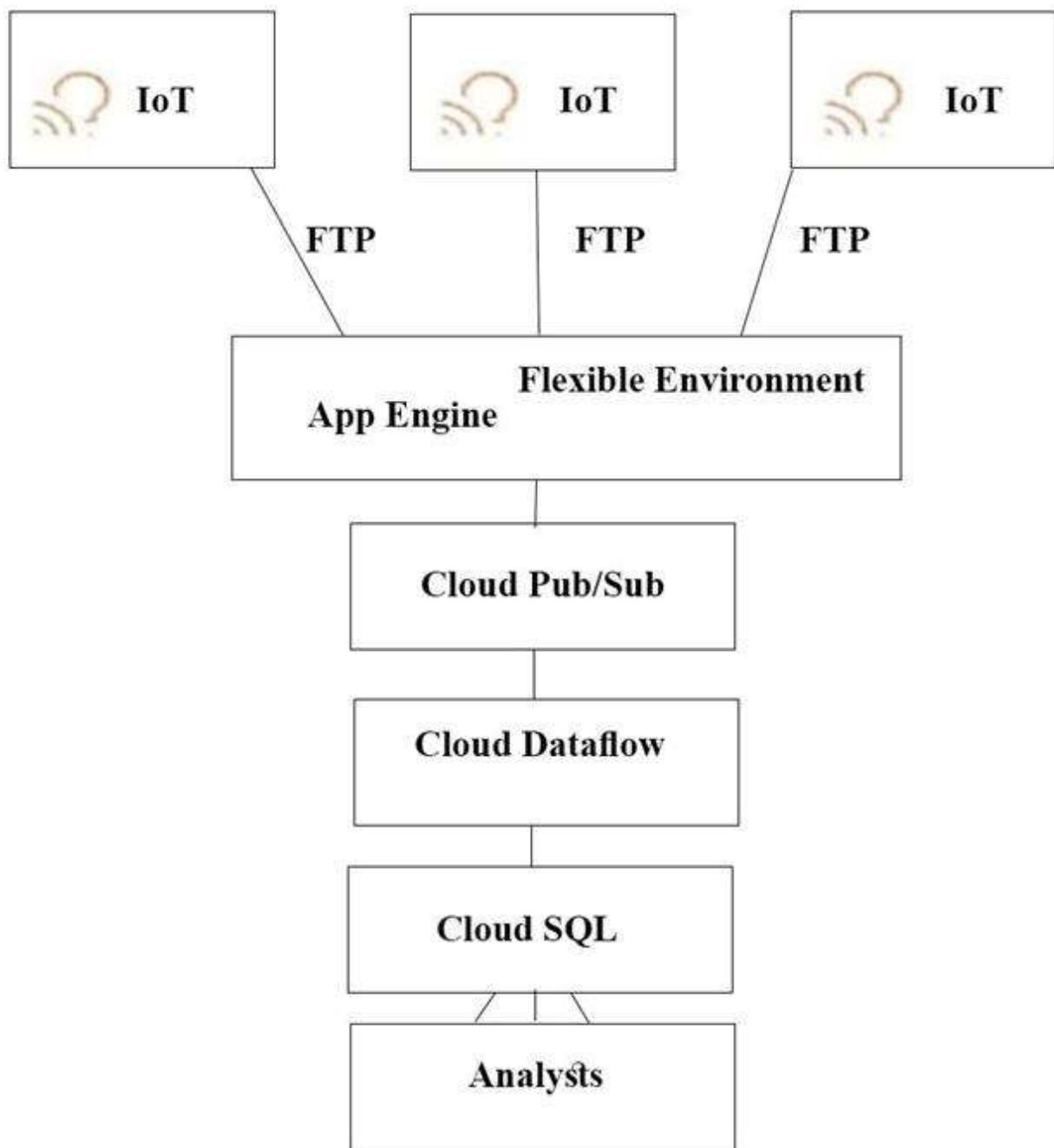
B.



C.



D.



Correct Answer: A

Google Cloud Certified Professional Cloud Architect Definitive Guide

The TerramEarth development team wants to create an API to meet the company's business requirements. You want the development team to focus their development effort on business value versus creating a custom framework. Which method should they use?

- A. Use Google App Engine with Google Cloud Endpoints. Focus on an API for dealers and partners
- B. Use Google App Engine with a JAX-RS Jersey Java-based framework. Focus on an API for the public
- C. Use Google App Engine with the Swagger (Open API Specification) framework. Focus on an API for the public
- D. Use Google Container Engine with a Django Python container. Focus on an API for the public
- E. Use Google Container Engine with a Tomcat container with the Swagger (Open API Specification) framework. Focus on an API for dealers and partners

Correct Answer: A

Your development team has created a structured API to retrieve vehicle data. They want to allow third parties to develop tools for dealerships that use this vehicle event data. You want to support delegated authorization against this data. What should you do?

- A. Build or leverage an OAuth-compatible access control system
- B. Build SAML 2.0 SSO compatibility into your authentication system
- C. Restrict data access based on the source IP address of the partner systems
- D. Create secondary credentials for each dealer that can be given to the trusted third party

Correct Answer: A

For this question, refer to the Mountkirk Games case study. You are in charge of the new Game Backend Platform architecture. The game communicates with the backend over a REST API. You want to follow Google-recommended practices. How should you design the backend?

- A. Create an instance template for the backend. For every region, deploy it on a multi-zone managed instance group. Use an L4 load balancer.
- B. Create an instance template for the backend. For every region, deploy it on a single-zone managed instance group. Use an L4 load balancer.

- C. Create an instance template for the backend. For every region, deploy it on a multi-zone managed instance group. Use an L7 load balancer.
- D. Create an instance template for the backend. For every region, deploy it on a single-zone managed instance group. Use an L7 load balancer.

Correct Answer: C

For this question, refer to the Mountkirk Games case study. Which managed storage option meets Mountkirk's technical requirement for storing game activity in a time series database service?

- A. Cloud Bigtable
- B. Cloud Spanner
- C. BigQuery
- D. Cloud Datastore

Correct Answer: A

For this question, refer to the Mountkirk Games case study. You need to analyze and define the technical architecture for the compute workloads for your company, Mountkirk Games. Considering the Mountkirk games business and technical requirements, what should you do?

- A. Create network load balancers. Use preemptible Compute Engine instances.
- B. Create network load balancers. Use non-preemptible Compute Engine instances.
- C. Create a global load balancer with managed instance groups and autoscaling policies. Use preemptible Compute Engine instances.
- D. Create a global load balancer with managed instance groups and autoscaling policies. Use non-preemptible Compute Engine instance.**

Correct Answer: D

For this question, refer to the Mountkirk Games case study. You need to analyze and define the technical architecture for the database workloads for your company, Mountkirk Games. Considering the business and technical requirements, what should you do?

- A. Use Cloud SQL for time series data, and use Cloud Bigtable for historical data queries.

Google Cloud Certified Professional Cloud Architect Definitive Guide

- B. Use Cloud SQL to replace MySQL, and use Cloud Spanner for historical data queries.
- C. Use Cloud Bigtable to replace MySQL, and use BigQuery for historical data queries.
- D. Use Cloud Bigtable for time series data, use Cloud Spanner for transactional data, and use BigQuery for historical data queries.**

Correct Answer: D

For this question, refer to the Mountkirk Games case study. Mountkirk Games wants you to design a way to test the analytics platform's resilience to changes in mobile network latency.

What should you do?

- A. Deploy failure injection software to the game analytics platform that can inject additional latency to mobile client analytics traffic.
- B. Build a test client that can be run from a mobile phone emulator on a Compute Engine virtual machine, and run multiple copies in Google Cloud Platform regions all over the world to generate realistic traffic.**
- C. Add the ability to introduce a random amount of delay before beginning to process analytics files uploaded from mobile devices.
- D. Create an opt-in beta of the game that runs on players' mobile devices and collects response times from analytics endpoints running in Google Cloud Platform regions all over the world.

Correct Answer: B

For this question, refer to the Mountkirk Games case study. Mountkirk Games wants to design their solution for the future in order to take advantage of cloud and technology improvements as they become available. Which two steps should they take? (Choose two.)

- A. Store as much analytics and game activity data as financially feasible today so it can be used to train machine learning models to predict user behavior in the future.**
- B. Begin packaging their game backend artifacts in container images and running them on Kubernetes Engine to improve the availability to scale up or down based on game activity.**
- C. Set up a CI/CD pipeline using Jenkins and Spinnaker to automate canary deployments and improve development velocity.
- D. Adopt a schema versioning tool to reduce downtime when adding new game features that require storing additional player data in the database.

Google Cloud Certified Professional Cloud Architect Definitive Guide

E. Implement a weekly rolling maintenance process for the Linux virtual machines so they can apply critical kernel patches and package updates and reduce the risk of 0-day vulnerabilities.

Correct Answer: A, B

For this question, refer to the Mountkirk Games case study. You need to analyze and define the technical architecture for the compute workloads for your company, Mountkirk Games. Considering the Mountkirk Games business and technical requirements, what should you do?

- A. Create network load balancers. Use preemptible Compute Engine instances.
- B. Create network load balancers. Use non-preemptible Compute Engine instances.
- C. Create a global load balancer with managed instance groups and autoscaling policies. Use preemptible Compute Engine instances.
- D. Create a global load balancer with managed instance groups and autoscaling policies. Use non-preemptible Compute Engine instances.**

Correct Answer: D

Mountkirk Games needs to create a repeatable and configurable mechanism for deploying isolated application environments. Developers and testers can access each other's environments and resources, but they cannot access staging or production resources. The staging environment needs access to some services from production.

What should you do to isolate development environments from staging and production?

- A. Create a project for development and test and another for staging and production**
- B. Create a network for development and test and another for staging and production
- C. Create one subnetwork for development and another for staging and production
- D. Create one project for development, a second for staging and a third for production

Correct Answer: A

Mountkirk Games wants to set up a real-time analytics platform for their new game. The new platform must meet their technical requirements.

Which combination of Google technologies will meet all of their requirements?

- A. Kubernetes Engine, Cloud Pub/Sub, and Cloud SQL

B. Cloud Dataflow, Cloud Storage, Cloud Pub/Sub, and BigQuery

- C. Cloud SQL, Cloud Storage, Cloud Pub/Sub, and Cloud Dataflow
- D. Cloud Dataproc, Cloud Pub/Sub, Cloud SQL, and Cloud Dataflow
- E. Cloud Pub/Sub, Compute Engine, Cloud Storage, and Cloud Dataproc

Correct Answer: B

For this question, refer to the Mountkirk Games case study. Mountkirk Games wants to migrate from their current analytics and statistics reporting model to one that meets their technical requirements on Google Cloud Platform.

Which two steps should be part of their migration plan? (Choose two.)

- A. Evaluate the impact of migrating their current batch ETL code to Cloud Dataflow.**
- B. Write a schema migration plan to de-normalize data for better performance in BigQuery.**
- C. Draw an architecture diagram that shows how to move from a single MySQL database to a MySQL cluster.
- D. Load 10 TB of analytics data from a previous game into a Cloud SQL instance, and run test queries against the full dataset to confirm that they complete successfully.
- E. Integrate Cloud Armor to defend against possible SQL injection attacks in analytics files uploaded to Cloud Storage.

Correct Answer: A, B

Mountkirk Games wants you to design their new testing strategy. How should the test coverage differ from their existing backends on the other platforms?

- A. Tests should scale well beyond the prior approaches**
- B. Unit tests are no longer required, only end-to-end tests
- C. Tests should be applied after the release is in the production environment
- D. Tests should include directly testing the Google Cloud Platform (GCP) infrastructure

Correct Answer: A

Mountkirk Games has deployed their new backend on Google Cloud Platform (GCP). You want to create a thorough testing process for new versions of the backend before they are released to

Google Cloud Certified Professional Cloud Architect Definitive Guide

the public. You want the testing environment to scale in an economical way. How should you design the process?

- A. Create a scalable environment in GCP for simulating production load
- B. Use the existing infrastructure to test the GCP-based backend at scale
- C. Build stress tests into each component of your application using resources internal to GCP to simulate load
- D. Create a set of static environments in GCP to test different levels of load – for example, high, medium, and low

Correct Answer: A

Mountkirk Games wants to set up a continuous delivery pipeline. Their architecture includes many small services that they want to be able to update and roll back quickly. Mountkirk Games has the following requirements:

Services are deployed redundantly across multiple regions in the US and Europe

Only frontend services are exposed on the public internet

They can provide a single frontend IP for their fleet of services

Deployment artifacts are immutable

Which set of products should they use?

- A. Google Cloud Storage, Google Cloud Dataflow, Google Compute Engine
- B. Google Cloud Storage, Google App Engine, Google Network Load Balancer
- C. Google Kubernetes Registry, Google Container Engine, Google HTTP(S) Load Balancer**
- D. Google Cloud Functions, Google Cloud Pub/Sub, Google Cloud Deployment Manager

Correct Answer: C

Mountkirk Games' gaming servers are not automatically scaling properly. Last month, they rolled out a new feature, which suddenly became very popular. A record number of users are trying to use the service, but many of them are getting 503 errors and very slow response times.

What should they investigate first?

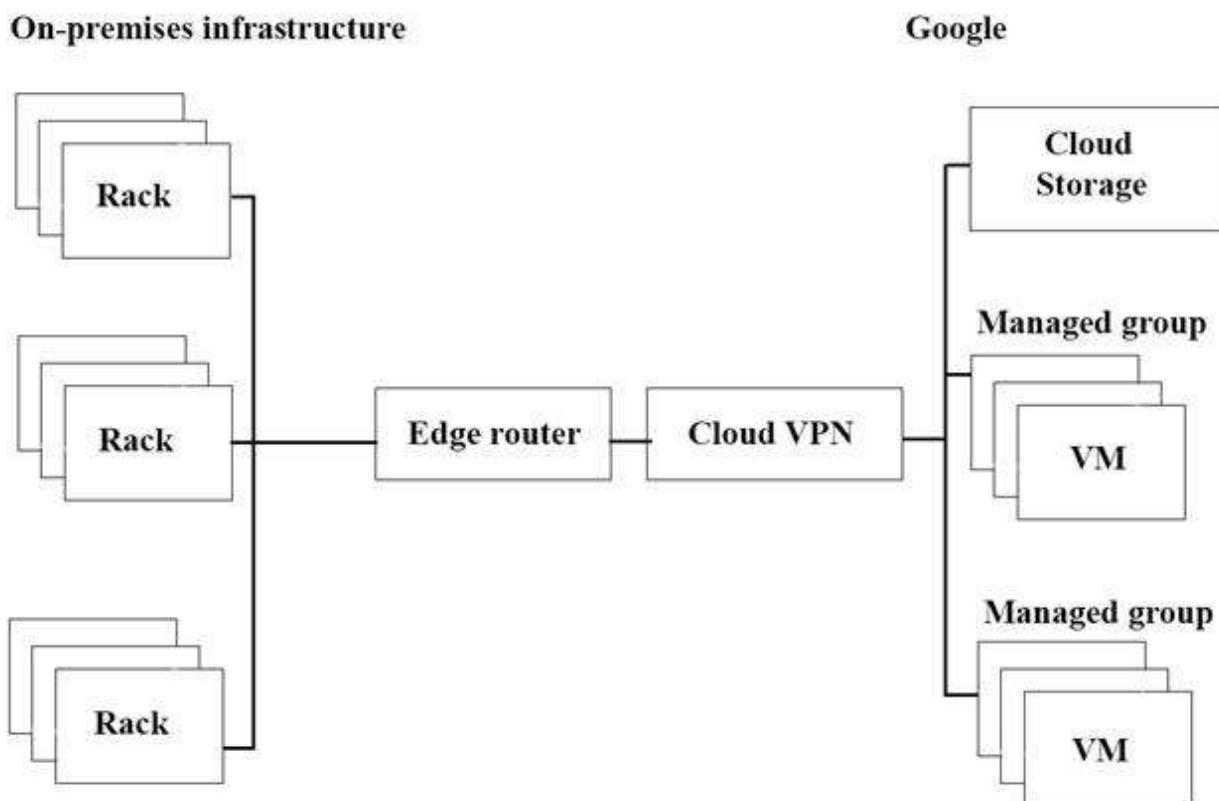
- A. Verify that the database is online
- B. Verify that the project quota hasn't been exceeded**
- C. Verify that the new feature code did not introduce any performance bugs

D. Verify that the load-testing team is not running their tool against production

Correct Answer: B

The migration of JencoMart's application to Google Cloud Platform (GCP) is progressing too slowly. The infrastructure is shown in the diagram. You want to maximize throughput.

What are three potential bottlenecks? Choose 3 answers.



- A. A single VPN tunnel, which limits throughput
- B. A tier of Google Cloud Storage that is not suited for this task
- C. A copy command that is not suited to operate over long distances
- D. Fewer virtual machines (VMs) in GCP than on-premises machines
- E. A separate storage layer outside the VMs, which is not suited for this task
- F. Complicated internet connectivity between the on-premises infrastructure and GCP

Correct Answer: A, C, E

Google Cloud Certified Professional Cloud Architect Definitive Guide

JencoMart wants to move their User Profiles database to Google Cloud Platform.

Which Google Database should they use?

- A. Cloud Spanner
- B. Google BigQuery
- C. Google Cloud SQL
- D. Google Cloud Datastore**

Correct Answer: D

The JencoMart security team requires that all Google Cloud Platform infrastructure is deployed using a least privilege model with separation of duties for administration between production and development resources. What Google domain and project structure should you recommend?

- A. Create two G Suite accounts to manage users: one for development/test/staging and one for production. Each account should contain one project for every application
- B. Create two G Suite accounts to manage users: one with a single project for all development applications and one with a single project for all production applications
- C. Create a single G Suite account to manage users with each stage of each application in its own project
- D. Create a single G Suite account to manage users with one project for the development/test/staging environment and one project for the production environment**

Correct Answer: D

A few days after JencoMart migrates the user credentials database to Google Cloud Platform and shuts down the old server, the new database server stops responding to SSH connections. It is still serving database requests to the application servers correctly.

What three steps should you take to diagnose the problem? Choose 3 answers.

- A. Delete the virtual machine (VM) and disks and create a new one
- B. Delete the instance, attach the disk to a new VM, and investigate
- C. Take a snapshot of the disk and connect to a new machine to investigate**
- D. Check inbound firewall rules for the network the machine is connected to**
- E. Connect the machine to another network with very simple firewall rules and investigate

F. Print the Serial Console output for the instance for troubleshooting, activate the interactive console, and investigate

Correct Answer: C, D, F

JencoMart has decided to migrate user profile storage to Google Cloud Datastore and the application servers to Google Compute Engine (GCE). During the migration, the existing infrastructure will need access to Datastore to upload the data. What service account key-management strategy should you recommend?

- A. Provision service account keys for the on-premises infrastructure and for the GCE virtual machines (VMs)
- B. Authenticate the on-premises infrastructure with a user account and provision service account keys for the VMs
- C. Provision service account keys for the on-premises infrastructure and use Google Cloud Platform (GCP) managed keys for the VMs**
- D. Deploy a custom authentication service on GCE/Google Kubernetes Engine (GKE) for the on-premises infrastructure and use GCP managed keys for the VMs

Correct Answer: C

JencoMart has built a version of their application on Google Cloud Platform that serves traffic to Asia. You want to measure success against their business and technical goals.

Which metrics should you track?

- A. Error rates for requests from Asia
- B. Latency difference between US and Asia
- C. Total visits, error rates, and latency from Asia
- D. Total visits and average latency for users from Asia**
- E. The number of character sets present in the database

Correct Answer: D

Practice Exam 1 (20)

Mountkirk Games wants to set up a continuous delivery pipeline. Their architecture includes many small services that they want to update and roll back quickly. Mountkirk Games has the following requirements:

- Services are deployed redundantly across multiple regions in the US and Europe.
- Only frontend services are exposed on the public internet.
- They can reserve a single frontend IP for their fleet of services.
- Deployment artifacts are immutable.

Which set of products should they use?

- A. Cloud Storage, Cloud Dataflow, Compute Engine
- B. Cloud Storage, App Engine, Cloud Load Balancing
- C. Container Registry, Google Kubernetes Engine, Cloud Load Balancing**
- D. Cloud Functions, Cloud Pub/Sub, Cloud Deployment Manager

Correct Answer: C

Because you do not know every possible future use for the data TerramEarth collects, you have decided to build a system that captures and stores all raw data in case you need it later. How can you most cost-effectively accomplish this goal?

- A. Have the vehicles in the field stream the data directly into BigQuery.
- B. Have the vehicles in the field pass the data to Cloud Pub/Sub and dump it into a Cloud Dataproc cluster that stores data in Apache Hadoop Distributed File System (HDFS) on persistent disks.
- C. Have the vehicles in the field continue to dump data via FTP, adjust the existing Linux machines, and use a collector to upload them into Cloud Dataproc HDFS for storage.
- D. Have the vehicles in the field continue to dump data via FTP, and adjust the existing Linux machines to immediately upload it to Cloud Storage with gsutil.**

Correct Answer: D

Google Cloud Certified Professional Cloud Architect Definitive Guide

Today, TerramEarth maintenance workers receive interactive performance graphs for the last 24 hours (86,400 events) by plugging their maintenance tablets into the vehicle. The support group wants support technicians to view this data remotely to help troubleshoot problems. You want to minimize the latency of graph loads. How should you provide this functionality?

- A. Execute queries against data stored in a Cloud SQL.
- B. Execute queries against data indexed by vehicle_id, timestamp in Cloud Bigtable.
- C. Execute queries against data stored on daily partitioned BigQuery tables.**
- D. Execute queries against BigQuery with data stored in Cloud Storage via BigQuery federation.

Correct Answer: B

Your agricultural division is experimenting with fully autonomous vehicles. You want your architecture to promote strong security during vehicle operation. Which two architecture characteristics should you consider?

- A. Use multiple connectivity subsystems for redundancy.
- B. Require IPv6 for connectivity to ensure a secure address space.**
- C. Enclose the vehicle's drive electronics in a Faraday cage to isolate chips.
- D. Use a functional programming language to isolate code execution cycles.**
- E. Treat every microservice call between modules on the vehicle as untrusted.
- F. Use a Trusted Platform Module (TPM) and verify firmware and binaries on boot.**

Correct Answer: B, D, F

Which of TerramEarth's legacy enterprise processes will experience significant change as a result of increased Google Cloud Platform adoption?

- A. OpEx/CapEx allocation, LAN change management, capacity planning
- B. Capacity planning, TCO calculations, OpEx/CapEx allocation**
- C. Capacity planning, utilization measurement, data center expansion
- D. Data center expansion, TCO calculations, utilization measurement

Correct Answer: B

You analysed TerramEarth's business requirement to reduce downtime and found that they can achieve a majority of time saving by reducing customers' wait time for parts. You decided to focus on reduction of the 3 weeks' aggregate reporting time. Which modifications to the company's processes should you recommend?

- A. Migrate from CSV to binary format, migrate from FTP to SFTP transport, and develop machine learning analysis of metrics.
- B. Migrate from FTP to streaming transport, migrate from CSV to binary format, and develop machine learning analysis of metrics.
- C. Increase fleet cellular connectivity to 80%, migrate from FTP to streaming transport, and develop machine learning analysis of metrics.**
- D. Migrate from FTP to SFTP transport, develop machine learning analysis of metrics, and increase dealer local inventory by a fixed factor.

Correct Answer: C

Your company wants to deploy several microservices to help their system handle elastic loads. Each microservice uses a different version of software libraries. You want to enable their developers to keep their development environment in sync with the various production services. Which technology should you choose?

- A. RPM/DEB
- B. Containers**
- C. Chef/Puppet
- D. Virtual machines

Correct Answer: B

Your company wants to track whether someone is present in a meeting room reserved for a scheduled meeting. There are 1000 meeting rooms across 5 offices on 3 continents. Each room is equipped with a motion sensor that reports its status every second. You want to support the data upload and collection needs of this sensor network. The receiving infrastructure needs to account for the possibility that the devices may have inconsistent connectivity. Which solution should you design?

Google Cloud Certified Professional Cloud Architect Definitive Guide

- A. Have each device create a persistent connection to a Compute Engine instance and write messages to a custom application.
- B. Have devices poll for connectivity to Cloud SQL and insert the latest messages on a regular interval to a device specific table.
- C. Have devices poll for connectivity to Cloud Pub/Sub and publish the latest messages on a regular interval to a shared topic for all devices.**
- D. Have devices create a persistent connection to an App Engine application fronted by Cloud Endpoints, which ingest messages and write them to Cloud Datastore.

Correct Answer: C

Your company wants to try out the cloud with low risk. They want to archive approximately 100 TB of their log data to the cloud and test the analytics features available to them there, while also retaining that data as a long-term disaster recovery backup. Which two steps should they take?

- A. Load logs into BigQuery.**
- B. Load logs into Cloud SQL.
- C. Import logs into Stackdriver.
- D. Insert logs into Cloud Bigtable.
- E. Upload log files into Cloud Storage.**

Correct Answer: A, E

You set up an autoscaling instance group to serve web traffic for an upcoming launch. After configuring the instance group as a backend service to an HTTP(S) load balancer, you notice that virtual machine (VM) instances are being terminated and re-launched every minute. The instances do not have a public IP address. You have verified that the appropriate web response is coming from each instance using the curl command. You want to ensure that the backend is configured correctly. What should you do?

- A. Ensure that a firewall rule exists to allow source traffic on HTTP/HTTPS to reach the load balancer.
- B. Assign a public IP to each instance, and configure a firewall rule to allow the load balancer to reach the instance public IP.

C. Ensure that a firewall rule exists to allow load balancer health checks to reach the instances in the instance group.

D. Create a tag on each instance with the name of the load balancer. Configure a firewall rule with the name of the load balancer as the source and the instance tag as the destination.

Correct Answer: C

Your organization has a 3-tier web application deployed in the same network on Google Cloud Platform. Each tier (web, API, and database) scales independently of the others. Network traffic should flow through the web to the API tier, and then on to the database tier. Traffic should not flow between the web and the database tier. How should you configure the network?

- A. Add each tier to a different subnetwork.
- B. Set up software-based firewalls on individual VMs.
- C. Add tags to each tier and set up routes to allow the desired traffic flow.
- D. Add tags to each tier and set up firewall rules to allow the desired traffic flow.**

Correct Answer: D

Your organization has 5 TB of private data on premises. You need to migrate the data to Cloud Storage. You want to maximize the data transfer speed. How should you migrate the data?

- A. Use gsutil.**
- B. Use gcloud.
- C. Use GCS REST API.
- D. Use Storage Transfer Service.

Correct Answer: A

You are designing a mobile chat application. You want to ensure that people cannot spoof chat messages by proving that a message was sent by a specific user. What should you do?

- A. Encrypt the message client-side using block-based encryption with a shared key.
- B. Tag messages client-side with the originating user identifier and the destination user.
- C. Use a trusted certificate authority to enable SSL connectivity between the client application and the server.

D. Use public key infrastructure (PKI) to encrypt the message client-side using the originating user's private key.

Correct Answer: D

You are designing a large distributed application with 30 microservices. Each of your distributed microservices needs to connect to a database backend. You want to store the credentials securely. Where should you store the credentials?

- A. In the source code
- B. In an environment variable
- C. In a key management system**
- D. In a config file that has restricted access through ACLs

Correct Answer: C

Mountkirk Games wants to set up a real-time analytics platform for their new game. The new platform must meet their technical requirements. Which combination of Google technologies will meet all of their requirements?

- A. Kubernetes Engine, Cloud Pub/Sub, and Cloud SQL
- B. Cloud Dataflow, Cloud Storage, Cloud Pub/Sub, and BigQuery**
- C. Cloud SQL, Cloud Storage, Cloud Pub/Sub, and Cloud Dataflow
- D. Cloud Pub/Sub, Compute Engine, Cloud Storage, and Cloud Dataproc

Correct Answer: B

Mountkirk Games has deployed their new backend on Google Cloud Platform (GCP). You want to create a thorough testing process for new versions of the backend before they are released to the public. You want the testing environment to scale in an economical way. How should you design the process?

- A. Create a scalable environment in GCP for simulating production load.**
- B. Use the existing infrastructure to test the GCP-based backend at scale.
- C. Build stress tests into each component of your application and use resources from the already deployed production backend to simulate load.

Google Cloud Certified Professional Cloud Architect Definitive Guide

D. Create a set of static environments in GCP to test different levels of load—for example, high, medium, and low.

Correct Answer: A

Your customer is moving their corporate applications to Google Cloud Platform. The security team wants detailed visibility of all resources in the organization. You use Resource Manager to set yourself up as the org admin. What Cloud Identity and Access Management (Cloud IAM) roles should you give to the security team?

- A. Org viewer, Project owner
- B. Org viewer, Project viewer**
- C. Org admin, Project browser
- D. Project owner, Network admin

Correct Answer: B

To reduce costs, the Director of Engineering has required all developers to move their development infrastructure resources from on-premises virtual machines (VMs) to Google Cloud Platform. These resources go through multiple start/stop events during the day and require state to persist. You have been asked to design the process of running a development environment in Google Cloud while providing cost visibility to the finance department. Which two steps should you take?

- A. Use persistent disks to store the state. Start and stop the VM as needed.
- B. Use the --auto-delete flag on all persistent disks before stopping the VM.
- C. Apply VM CPU utilization label and include it in the BigQuery billing export.**
- D. Use BigQuery billing export and labels to relate cost to groups.**
- E. Store all state in local SSD, snapshot the persistent disks, and terminate the VM.
- F. Store all state in Cloud Storage, snapshot the persistent disks, and terminate the VM.

Correct Answer: C, D

Your company has decided to make a major revision of their API in order to create better experiences for their developers. They need to keep the old version of the API available and

deployable, while allowing new customers and testers to try out the new API. They want to keep the same SSL and DNS records in place to serve both APIs. What should they do?

- A. Configure a new load balancer for the new version of the API.
- B. Reconfigure old clients to use a new endpoint for the new API.
- C. Have the old API forward traffic to the new API based on the path.
- D. Use separate backend services for each API path behind the load balancer.**

Correct Answer: D

The database administration team has asked you to help them improve the performance of their new database server running on Compute Engine. The database is used for importing and normalizing the company's performance statistics. It is built with MySQL running on Debian Linux. They have an n1-standard-8 virtual machine with 80 GB of SSD zonal persistent disk. What should they change to get better performance from this system in a cost-effective manner?

- A. Increase the virtual machine's memory to 64 GB.
- B. Create a new virtual machine running PostgreSQL.
- C. Dynamically resize the SSD persistent disk to 500 GB.**
- D. Migrate their performance metrics warehouse to BigQuery.

Correct Answer: C

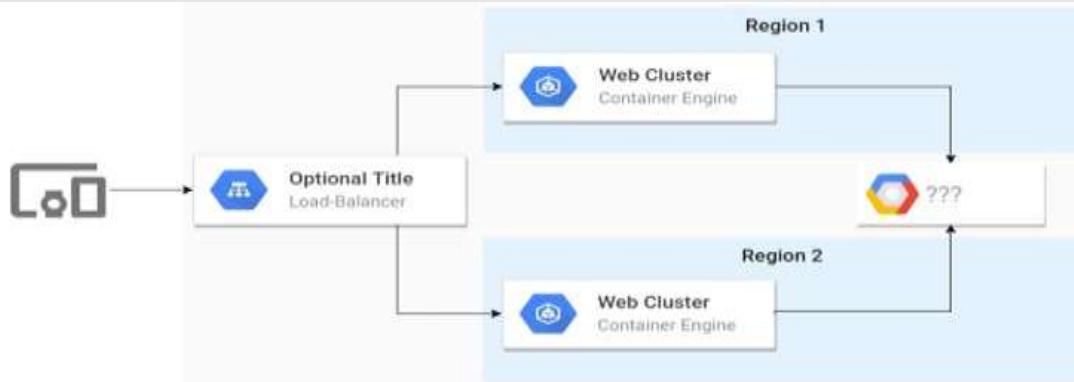
Practice Exam 2 (20)

Your company wants to reduce cost on infrequently accessed data by moving it to the cloud. The data will still be accessed approximately once a month to refresh historical charts. In addition, data older than 5 years is no longer needed. How should you store and manage the data?

- A. In Google Cloud Storage and stored in a Multi-Regional bucket. Set an Object Lifecycle Management policy to delete data older than 5 years.
- B. In Google Cloud Storage and stored in a multi-Regional bucket. Set an Object Lifecycle Management policy to change the storage class to Coldline for data older than 5 years.
- C. In Google Cloud Storage and stored in a Nearline bucket. Set an Object Lifecycle Management policy to delete data older than 5 years.**
- D. In Google Cloud Storage and stored in a Nearline bucket. Set an Object Lifecycle Management policy to change the storage class to Coldline for data older than 5 years.

Correct Answer: C

Your company's architecture is shown in the diagram. You want to keep data in sync across Region 1 and Region 2. Which product should you use?



- A. Google Cloud SQL
- B. Google Cloud Bigtable
- C. Google Cloud Storage**
- D. Google Cloud Datastore

Correct Answer: C

Google Cloud Certified Professional Cloud Architect Definitive Guide

Your company is building a large-scale web application. Each team is responsible for its own service component of the application and wants to manage its own individual projects. You want each service to communicate with the others over RFC1918 address space. What should you do?

- A. Deploy each service into a single project within the same VPC.
- B. Configure Shared VPC, and add each project as a service of the Shared VPC project.**
- C. Configure each service to communicate with the others over HTTPS protocol.
- D. Configure a global load balancer for each project, and communicate between each service using the global load balancer IP addresses.

Correct Answer: B

Your company collects and stores security camera footage in Google Cloud Storage. Within the first 30 days, footage is processed regularly for threat detection, object detection, trend analysis, and suspicious Behaviour detection. You want to minimize the cost of storing all the data. How should you store the videos?

- A. Use Google Cloud Regional Storage for the first 30 days, and then move to Coldline Storage.**
- B. Use Google Cloud Nearline Storage for the first 30 days, and then move to Coldline Storage.
- C. Use Google Cloud Regional Storage for the first 30 days, and then move to Nearline Storage.
- D. Use Google Cloud Regional Storage for the first 30 days, and then move to Google Persistent Disk.

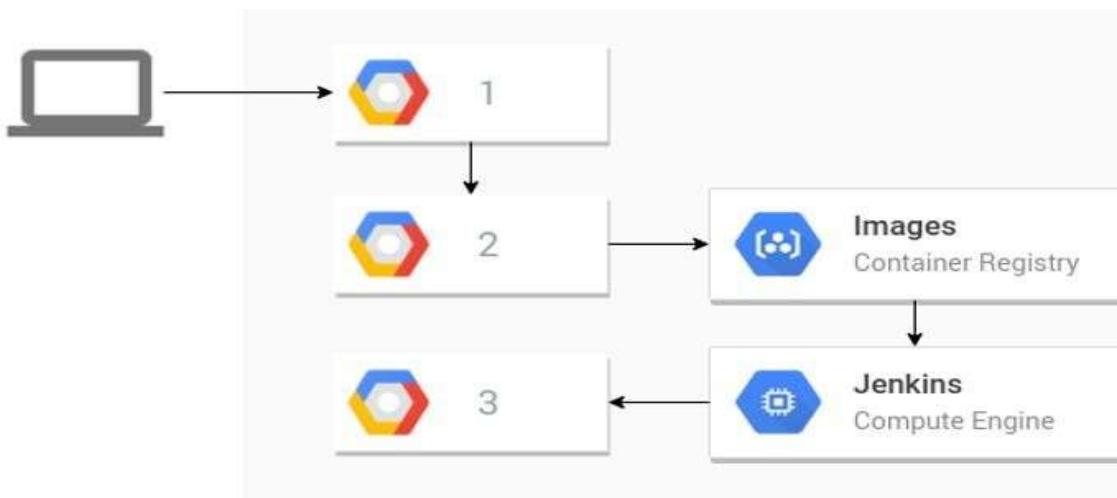
Correct Answer: A

You are trying to deploy the Google Cloud Deployment Manager manifest shown below. The team wants the instance group to scale to 3 instances, but only 1 instance is starting. You want the deployment to succeed. What should you do?

- A. Choose a different zone.
- B. Use a custom source Image.
- C. Remove the network interface.
- D. Remove the second disk from the instance group template**

Correct Answer: D

Your CI/CD pipeline process is shown in the diagram. Which GCP services should you use in boxes 1, 2, and 3?



- A. Google Cloud Storage, Google Cloud Pub/Sub, Google Compute Engine
- B. Google Cloud Storage, Google Cloud Shell, Google Container Engine
- C. Google Cloud Source Repositories, Google Cloud Storage, Google Container Engine
- D. Google Cloud Source Repositories, Google Cloud Container Builder, Google Container Engine**

Correct Answer: D

You are running an application in Google App Engine that is serving production trace. You want to deploy a risky but necessary change to the application. It could take down your service if not properly coded. During development of the application, you realized that it can only be properly tested by live user trace. How should you test the feature?

- A. Deploy the new application version temporarily, and then roll it back.
- B. Create a second project with the new app in isolation, and onboard users.
- C. Set up a second Google App Engine service, and then update a subset of clients to hit the new service.
- D. Deploy a new version of the application, and use trace splitting to send a small percentage of trace to it.**

Correct Answer: D

Google Cloud Certified Professional Cloud Architect Definitive Guide

One of the Microservices in your application has an intermittent performance problem. You have not observed the problem when it occurs but when it does, it triggers a particular burst of log lines. You want to debug a machine while the problem is occurring. What should you do?

- A. Log into one of the machines running the microservice and wait for the log storm
- B. In the Stackdriver Error Reporting dashboard, look for a pattern in the times the problem occurs
- C. Configure your microservice to send traces to Stackdriver Trace so you can find what is taking so long
- D. Set up a log metric in Stackdriver Logging, and then set up an alert to notify you when the number of log lines increases past a threshold**

Correct Answer: D

For future phases, Dress4Win is looking at options to deploy data analytics to the Google Cloud. Which option meets their business and technical requirements?

- A. Run current jobs from the current technical environment on Google Cloud Dataproc.**
- B. Review all current data jobs. Identify the most critical jobs and create Google BigQuery tables to store and query data.
- C. Review all current data jobs. Identify the most critical jobs and develop Google Cloud Dataflow pipelines to process data.
- D. Deploy a Hadoop/Spark cluster to Google Compute Engine virtual machines. Move current jobs from the current technical environment and run them on the Hadoop/Spark cluster.

Correct Answer: A

The Dress4Win developers are evaluating using Google Cloud Platform. They have identified some applications that can easily move to Google App Engine Flexible Environment. The developers will deploy their code using the Google Cloud SDK tools. Which two of their stated technical requirements does this solution meet?

- A. Encrypt data on the wire and at rest.
- B. Use managed services whenever possible.**
- C. Identify production services that can migrate to the cloud to save capacity.
- D. Support failover of the production environment to the cloud during an emergency.
- E. Evaluate and choose an automation framework for provisioning resources in the cloud.**

Google Cloud Certified Professional Cloud Architect Definitive Guide

F. Support multiple VPN connections between the production data center and the cloud environment

Correct Answer: B, E

The architecture diagram below shows an event-based processing pipeline that Dress4win is building to label and compress user uploaded images. Which GCP products should they use in boxes 1, 2 and 3?

- A. Google App Engine, Google Cloud Datastore, Google Cloud Dataflow
- B. Google App Engine, Google Cloud Dataflow, Google Cloud Functions
- C. Google Cloud Storage, Google Cloud Pub/Sub, Google Cloud Dataflow**
- D. Google Cloud Dataflow, Google Cloud Pub/Sub, Google Cloud Functions

Correct Answer: C

Dress4Win wants to do penetration security scanning on the test and development environments deployed to the cloud. The scanning should be performed from an end user perspective as much as possible. How should they conduct the penetration testing?

- A. Notify Google to begin conducting regular penetration security scanning on behalf of Dress4Win.
- B. Deploy the security scanners into the cloud environments and conduct penetration testing within each environment.
- C. Use the on-premises scanners to conduct penetration testing on the cloud environments routing trace over the VPN.
- D. Use the on-premises scanners to conduct penetration testing on the cloud environments routing trace over the public internet**

Correct Answer: D

Your company's architecture is shown in the diagram. You want to automatically and simultaneously deploy new code to each Google Container Engine cluster. Which method should you use?



- A. Use an automation tool, such as Jenkins.
- B. Change the clusters to activate federated mode.
- C. Use Parallel SSH with Google Cloud Shell and kubectl.
- D. Use Google Cloud Container Builder to publish the new images.

Correct Answer: A

Your company plans to migrate a multi-petabyte data set to the cloud. The data set must be available 24hrs a day. Your business analysts have experience only with using a SQL interface. How should you store the data to optimize it for ease of analysis?

- A. Load data into Google BigQuery.
- B. Insert data into Google Cloud SQL.
- C. Put files into Google Cloud Storage.
- D. Stream data into Google Cloud Datastore.

Correct Answer: A

You want to make a copy of a production Linux virtual machine in the US Central region. You want to manage and replace the copy easily if there are changes on the production virtual machine. You will deploy the copy as a new instance in a different project in the US-East region. What steps must you take?

- A. Use the Linux dd and netcat commands to copy and stream the root disk contents to a new virtual machine instance in the US-East region.
- B. Create a snapshot of the root disk and select the snapshot as the root disk when you create a new virtual machine instance in the US-East region.
- C. Create an image file from the root disk with Linux dd command, create a new disk from the image file, and use it to create a new virtual machine instance in the US-East region.
- D. Create a snapshot of the root disk, create an image file in Google Cloud Storage from the snapshot, and create a new virtual machine instance in the US-East region using the image file for the root disk.**

Correct Answer: D

Your customer is moving their storage product to Google Cloud Storage (GCS). The data contains personally identifiable information (PII) and sensitive customer information. What security strategy should you use for GCS?

- A. Use signed URLs to generate time bound access to objects.
- B. Grant IAM read-only access to users, and use default ACLs on the bucket.
- C. Grant no Google Cloud Identity and Access Management (Cloud IAM) roles to users, and use granular ACLs on the bucket.**
- D. Create randomized bucket and object names. Enable public access, but only provide specific file URLs to people who do not have Google accounts and need access.

Correct Answer: C

Google Cloud Certified Professional Cloud Architect Definitive Guide

Your customer is moving their corporate applications to Google Cloud Platform. The security team wants detailed visibility of all projects in the organization. You provision the Google Cloud Resource Manager and set up yourself as the org admin. Which Google Cloud Identity and Access Management (Cloud IAM) roles should you give to the security team?

- A. Org viewer and project owner
- B. Org viewer and project viewer**
- C. Org admin and project browser
- D. Project owner and network admin

Correct Answer: B

A recent software update to an e-commerce website running on Google Cloud has caused the website to crash for several hours. The CTO decides that all critical changes must now have a backout/roll-back plan. The website is deployed on hundreds of virtual machines (VMs), and critical changes are frequent. Which two actions should you take to implement the back-out/roll-back plan? (Choose two)

- A. Create a Nearline copy for the website static data les stored in Google Cloud Storage.
- B. Enable object versioning on the website's static data les stored in Google Cloud Storage.**
- C. Use managed instance groups with the “update-instances” command when starting a rolling update.**
- Enable Google Cloud Deployment Manager (CDM) on the project, and define each change with a new CDM template.
- D. Create a snapshot of each VM prior to an update, and recover the VM from the snapshot in case of a new version failure.

Correct Answer: B, C

You need to reduce the impact of unplanned rollbacks of erroneous production deployments in your company's web hosting platform. Improvement to the QA processes accomplished an 80% reduction. Which additional two approaches can you take to further reduce the impact of rollbacks? (Choose two)

- A. Introduce a green-blue deployment model.**
- B. Fragment the monolithic platform into Microservices.**

- C. Remove the QA environment. Start executing canary releases.
- D. Remove the platform's dependency on relational database systems.
- E. Replace the platform's relational database systems with a NoSQL database

Correct Answer: A, B

A lead software engineer tells you that his new application design uses websockets and HTTP sessions that are not distributed across the web servers. You want to help him ensure his application will run properly on Google Cloud Platform. What should you do?

- A. Help the engineer to convert his websocket code to use HTTP streaming.
- B. Review the encryption requirements for websocket connections with the security team.
- C. Meet with the cloud operations team and the engineer to discuss load balancer options.**
- D. Help the engineer redesigns the application to use a distributed user session service that does not rely on websockets and HTTP sessions

Correct Answer: C

Data Engineer Certification

Assessment Test 1

1. You are migrating your machine learning operations to GCP and want to take advantage of managed services. You have been managing a Spark cluster because you use the MLlib library extensively. Which GCP managed service would you use?

- A. Cloud Dataprep
- B. Cloud DataProc**
- C. Cloud Dataflow
- D. Cloud Pub/Sub

Correct Answer: B

2. Your team is designing a database to store product catalog information. They have determined that you need to use a database that supports flexible schemas and transactions. What service would you expect to use?

- A. Cloud SQL
- B. Cloud BigQuery
- C. Cloud Firestore**
- D. Cloud Storage

Correct Answer: C

3. Your company has been losing market share because competitors are attracting customers with a more personalized experience on their e-commerce platforms, including providing recommendations for products that might be of interest to them. The CEO has stated that your company will provide equivalent services within 90 days. What GCP service would you use to help meet this objective?

- A. Cloud Bigtable
- B. Cloud Storage
- C. AI Platform**
- D. Cloud Datastore

Correct Answer: C

4. The finance department at your company has been archiving data on premises. They no longer want to maintain a costly dedicated storage system. They would like to store up to 300 TB of data for 10 years. The data will likely not be accessed at all. They also want to minimize cost. What storage service would you recommend?

- A. Cloud Storage multi-regional storage
- B. Cloud Storage Nearline storage
- C. Cloud Storage Coldline storage**
- D. Cloud Bigtable

Correct Answer: C

5. You will be developing machine learning models using sensitive data. Your company has several policies regarding protecting sensitive data, including requiring enhanced security on virtual machines (VMs) processing sensitive data. Which GCP service would you look to for meeting those requirements?

- A. Identity and access management (IAM)
- B. Cloud Key Management Service
- C. Cloud Identity
- D. Shielded VMs**

Correct Answer: D

6. You have developed a machine learning algorithm for identifying objects in images. Your company has a mobile app that allows users to upload images and get back a list of identified objects. You need to implement the mechanism to detect when a new image is uploaded to Cloud Storage and invoke the model to perform the analysis. Which GCP service would you use for that?

- A. Cloud Functions**
- B. Cloud Storage Nearline
- C. Cloud Dataflow
- D. Cloud Dataproc

Correct Answer: A

7. An IoT system streams data to a Cloud Pub/Sub topic for ingestion, and the data is processed in a Cloud Dataflow pipeline before being written to Cloud Bigtable. Latency is increasing as more data is added, even though nodes are not at maximum utilization. What would you look for first as a possible cause of this problem?

- A. Too many nodes in the cluster
- B. A poorly designed row key**
- C. Too many column families
- D. Too many indexes being updated during write operations

Correct Answer: B

8. A health and wellness startup in Canada has been more successful than expected. Investors are pushing the founders to expand into new regions outside of North America. The CEO and CTO are discussing the possibility of expanding into Europe. The app offered by the startup collects personal information, storing some locally on the user's device and some in the cloud. What regulation will the startup need to plan for before expanding into the European market?

- A. HIPAA
- B. PCI-DSS
- C. GDPR**
- D. SOX

Correct Answer: C

9. Your company has been collecting vehicle performance data for the past year and now has 500 TB of data. Analysts at the company want to analyze the data to understand performance differences better across classes of vehicles. The analysts are advanced SQL users, but not all have programming experience. They want to minimize administrative overhead by using a managed service, if possible. What service might you recommend for conducting preliminary analysis of the data?

- A. Compute Engine
- B. Kubernetes Engine

C. BigQuery

D. Cloud Functions

Correct Answer: C

10. An airline is moving its luggage-tracking applications to Google Cloud. There are many requirements, including support for SQL and strong consistency. The database will be accessed by users in the United States, Europe, and Asia. The database will store approximately 50 TB in the first year and grow at approximately 10 percent a year after that. What managed database service would you recommend?

A. Cloud SQL

B. BigQuery

C. Cloud Spanner

D. Cloud Dataflow

Correct Answer: C

11. You are using Cloud Firestore to store data about online game players' state while in a game. The state information includes health score, a set of possessions, and a list of team members collaborating with the player. You have noticed that the size of the raw data in the database is approximately 2 TB, but the amount of space used by Cloud Firestore is almost 5 TB. What could be causing the need for so much more space?

A. The data model has been denormalized.

B. There are multiple indexes.

C. Nodes in the database cluster are misconfigured.

D. There are too many column families in use.

Correct Answer: B

12. You have a BigQuery table with data about customer purchases, including the date of purchase, the type of product purchases, the product name, and several other descriptive attributes. There is approximately three years of data. You tend to query data by month and

then by customer. You would like to minimize the amount of data scanned. How would you organize the table?

- A. Partition by purchase date and cluster by customer**
- B. Partition by purchase date and cluster by product
- C. Partition by customer and cluster by product
- D. Partition by customer and cluster by purchase date

Correct Answer: A

13. You are currently using Java to implement an ELT pipeline in Hadoop. You'd like to replace your Java programs with a managed service in GCP. Which would you use?

- A. Data Studio
- B. Cloud Dataflow**
- C. Cloud Bigtable
- D. BigQuery

Correct Answer: B

14. A group of attorneys has hired you to help them categorize over a million documents in an intellectual property case. The attorneys need to isolate documents that are relevant to a patent that the plaintiffs argue has been infringed. The attorneys have 50,000 labeled examples of documents, and when the model is evaluated on training data, it performs quite well. However, when evaluated on test data, it performs quite poorly. What would you try to improve the performance?

- A. Perform feature engineering
- B. Perform validation testing
- C. Add more data
- D. Regularization**

Correct Answer: D

15. Your company is migrating from an on-premises pipeline that uses Apache Kafka for ingesting data and MongoDB for storage. What two managed services would recommend as replacements for these?

- A. Cloud Dataflow and Cloud Bigtable
- B. Cloud Dataprep and Cloud Pub/Sub
- C. Cloud Pub/Sub and Cloud Firestore**
- D. Cloud Pub/Sub and BigQuery

Correct Answer: C

16. A group of data scientists is using Hadoop to store and analyze IoT data. They have decided to use GCP because they are spending too much time managing the Hadoop cluster. They are particularly interested in using services that would allow them to port their models and machine learning workflows to other clouds. What service would you use as a replacement for their existing platform?

- A. BigQuery
- B. Cloud Storage
- C. Cloud Dataproc**
- D. Cloud Spanner

Correct Answer: C

17. You are analyzing several datasets and will likely use them to build regression models. You will receive additional datasets, so you'd like to have a workflow to transform the raw data into a form suitable for analysis. You'd also like to work with the data in an interactive manner using Python. What services would you use in GCP?

- A. Cloud Dataflow and Data Studio
- B. Cloud Dataflow and Cloud Datalab**
- C. Cloud Dataprep and Data Studio
- D. Cloud Datalab and Data Studio

Correct Answer: B

Google Cloud Certified Professional Cloud Architect Definitive Guide

18. You have a large number of files that you would like to store for several years. The files will be accessed frequently by users around the world. You decide to store the data multi-regional Cloud Storage. You want users to be able to view files and their metadata in a Cloud Storage bucket. What role would you assign to those users? (Assume you practicing the principle of least privilege.)

- A. roles/storage.objectCreator
- B. roles/storage.objectViewer**
- C. roles/storage.admin
- D. roles/storage.bucketList

Correct Answer: B

19. You have built a deep learning neural network to perform multiclass classification. You find that the model is overfitting. Which of the following would not be used to reduce overfitting?

- A. Dropout
- B. L2 Regularization
- C. L1 Regularization
- D. Logistic regression**

Correct Answer: D

20. Your company would like to start experimenting with machine learning, but no one in the company is experienced with ML. Analysts in the marketing department have identified some data in their relational database that they think may be useful for training a model. What would you recommend that they try first to build proof-of-concept models?

- A. AutoML Tables**
- B. Kubeflow
- C. Cloud Firestore
- D. Spark MLLib

Correct Answer: A

21. You have several large deep learning networks that you have built using TensorFlow. The models use only standard TensorFlow components. You have been running the models on an n1-highcpu-64 VM, but the models are taking longer to train than you would like. What would you try first to accelerate the model training?

- A. GPUs
- B. TPUs**
- C. Shielded VMs
- D. Preemptible VMs

Correct Answer: B

22. Your company wants to build a data lake to store data in its raw form for extended periods of time. The data lake should provide access controls, virtually unlimited storage, and the lowest cost possible. Which GCP service would you suggest?

- A. Cloud Bigtable
- B. BigQuery
- C. Cloud Storage**
- D. Cloud Spanner

Correct Answer: C

23. Auditors have determined that your company's processes for storing, processing, and transmitting sensitive data are insufficient. They believe that additional measures must be taken to prevent sensitive information, such as personally identifiable government-issued numbers, are not disclosed. They suggest masking or removing sensitive data before it is transmitted outside the company. What GCP service would you recommend?

- A. Data loss prevention API**
- B. In-transit encryption
- C. Storing sensitive information in Cloud Key Management
- D. Cloud Dataflow

Correct Answer: A

24. You are using Cloud Functions to start the processing of images as they are uploaded into Cloud Storage. In the past, there have been spikes in the number of images uploaded, and many instances of the Cloud Function were created at those times. What can you do to prevent too many instances from starting?

- A. Use the --max-limit parameter when deploying the function.
- B. Use the --max-instances parameter when deploying the function.**
- C. Configure the --max-instance parameter in the resource hierarchy.
- D. Nothing. There is no option to limit the number of instances.

Correct Answer: B

25. You have several analysis programs running in production. Sometimes they are failing, but there is no apparent pattern to the failures. You'd like to use a GCP service to record custom information from the programs so that you can better understand what is happening. Which service would you use?

- A. Stackdriver Debugger
- B. Stackdriver Logging**
- C. Stackdriver Monitoring
- D. Stackdriver Trace

Correct Answer: B

26. The CTO of your company is concerned about the rising costs of maintaining your company's enterprise data warehouse. The current data warehouse runs in a PostgreSQL instance. You would like to migrate to GCP and use a managed service that reduces operational overhead and one that will scale to meet future needs of up to 3 PB. What service would you recommend?

- A. Cloud SQL using PostgreSQL
- B. BigQuery**
- C. Cloud Bigtable
- D. Cloud Spanner

Correct Answer: B

Assessment Test 2

1. You have Cloud Functions written in Node.js that pull messages from Cloud Pub/Sub and send the data to BigQuery. You observe that the message processing rate on the Pub/Sub topic is orders of magnitude higher than anticipated, but there is no error logged in Stackdriver Log Viewer. What are the two most likely causes of this problem? Choose 2 answers.

- A. Publisher throughput quota is too small.
- B. Total outstanding messages exceed the 10-MB maximum.
- C. Error handling in the subscriber code is not handling run-time errors properly.**
- D. The subscriber code cannot keep up with the messages.
- E. The subscriber code does not acknowledge the messages that it pulls.**

Correct Answer: C, E

2. You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?

- A. Add a SideInput that returns a Boolean if the element is corrupt.
- B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.**
- C. Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.
- D. Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.

Correct Answer: B

3. You work for a shipping company that uses handheld scanners to read shipping labels. Your company has strict data privacy standards that require scanners to only transmit recipients' personally identifiable information (PII) to analytics systems, which violates user privacy rules. You want to quickly build a scalable solution using cloud-native managed services to prevent exposure of PII to the analytics systems. What should you do?

- A. Create an authorized view in BigQuery to restrict access to tables with sensitive data.

- B. Install a third-party data validation tool on Compute Engine virtual machines to check the incoming data for sensitive information.
- C. Use Stackdriver logging to analyze the data passed through the total pipeline to identify transactions that may contain sensitive information.
- D. Build a Cloud Function that reads the topics and makes a call to the Cloud Data Loss Prevention API. Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.**

Correct Answer: D

4. You have developed three data processing jobs. One executes a Cloud Dataflow pipeline that transforms data uploaded to Cloud Storage and writes results to BigQuery. The second ingests data from on-premises servers and uploads it to Cloud Storage. The third is a Cloud Dataflow pipeline that gets information from third-party data providers and uploads the information to Cloud Storage. You need to be able to schedule and monitor the execution of these three workflows and manually execute them when needed. What should you do?

- A. Create a Direct Acyclic Graph in Cloud Composer to schedule and monitor the jobs.**
- B. Use Stackdriver Monitoring and set up an alert with a Webhook notification to trigger the jobs.
- C. Develop an App Engine application to schedule and request the status of the jobs using GCP API calls.
- D. Set up cron jobs in a Compute Engine instance to schedule and monitor the pipelines using GCP API calls.

Correct Answer: A

5. You are managing a Cloud Dataproc cluster. You need to make a job run faster while minimizing costs, without losing work in progress on your clusters. What should you do?

- A. Increase the cluster size with more non-preemptible workers.
- B. Increase the cluster size with preemptible worker nodes, and configure them to forcefully decommission.
- C. Increase the cluster size with preemptible worker nodes, and use Cloud Stackdriver to trigger a script to preserve work.

D. Increase the cluster size with preemptible worker nodes, and configure them to use graceful decommissioning.

Correct Answer: D

6. You want to automate execution of a multi-step data pipeline running on Google Cloud. The pipeline includes Cloud Dataproc and Cloud Dataflow jobs that have multiple dependencies on each other. You want to use managed services where possible, and the pipeline will run every day. Which tool should you use?

- A. cron
- B. Cloud Composer**
- C. Cloud Scheduler
- D. Workflow Templates on Cloud Dataproc

Correct Answer: B

7. You used Cloud Dataprep to create a recipe on a sample of data in a BigQuery table. You want to reuse this recipe on a daily upload of data with the same schema, after the load job with variable execution time completes. What should you do?

- A. Create a cron schedule in Cloud Dataprep.
- B. Create an App Engine cron job to schedule the execution of the Cloud Dataprep job.
- C. Export the recipe as a Cloud Dataprep template, and create a job in Cloud Scheduler.**
- D. Export the Cloud Dataprep job as a Cloud Dataflow template, and incorporate it into a Cloud Composer job.

Correct Answer: C

8. You have a data stored in BigQuery. The data in the BigQuery dataset must be highly available. You need to define a storage, backup, and recovery strategy of this data that minimizes cost. How should you configure the BigQuery table?

- A. Set the BigQuery dataset to be regional. In the event of an emergency, use a point-in-time snapshot to recover the data.

B. Set the BigQuery dataset to be regional. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup. In the event of an emergency, use the backup copy of the table.

C. Set the BigQuery dataset to be multi-regional. In the event of an emergency, use a point-in-time snapshot to recover the data.

D. Set the BigQuery dataset to be multi-regional. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup. In the event of an emergency, use the backup copy of the table.

Correct Answer: C

9. You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate. What should you do?

- A. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.**
- B. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.
- C. Use the BigQuery streaming the stream changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.
- D. Use the BigQuery bulk loader to batch load inventory changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.

Correct Answer: A

10. You need to copy millions of sensitive patient records from a relational database to BigQuery. The total size of the database is 10 TB. You need to design a solution that is secure and time-efficient. What should you do?

- A. Export the records from the database as an Avro file. Upload the file to GCS using gsutil, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.**
- B. Export the records from the database as an Avro file. Copy the file onto a Transfer Appliance and send it to Google, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.
- C. Export the records from the database into a CSV file. Create a public URL for the CSV file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the CSV file into BigQuery using the BigQuery web UI in the GCP Console.
- D. Export the records from the database as an Avro file. Create a public URL for the Avro file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

Correct Answer: A

11. You have a requirement to insert minute-resolution data from 50,000 sensors into a BigQuery table. You expect significant growth in data volume and need the data to be available within 1 minute of ingestion for real-time analysis of aggregated trends. What should you do?

- A. Use bq load to load a batch of sensor data every 60 seconds.
- B. Use a Cloud Dataflow pipeline to stream data into the BigQuery table.**
- C. Use the INSERT statement to insert a batch of data every 60 seconds.
- D. Use the MERGE statement to apply updates in batch every 60 seconds.

Correct Answer: B

12. You have a query that filters a BigQuery table using a WHERE clause on timestamp and ID columns. By using bq query --dry_run you learn that the query triggers a full scan of the table, even though the filter on timestamp and ID select a tiny fraction of the overall data. You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries. What should you do?

- A. Create a separate table for each ID.

B. Use the LIMIT keyword to reduce the number of rows returned.

C. Recreate the table with a partitioning column and clustering column.

D. Use the bq query --maximum_bytes_billed flag to restrict the number of bytes billed.

Correct Answer: C

13. Your company needs to upload their historic data to Cloud Storage. The security rules don't allow access from external IPs to their on-premises resources. After an initial upload, they will add new data from existing on-premises applications every day. What should they do?

A. Execute gsutil rsync from the on-premises servers.

B. Use Cloud Dataflow and write the data to Cloud Storage.

C. Write a job template in Cloud Dataproc to perform the data transfer.

D. Install an FTP server on a Compute Engine VM to receive the files and move them to Cloud Storage.

Correct Answer: D

14. You store historic data in Cloud Storage. You need to perform analytics on the historic data. You want to use a solution to detect invalid data entries and perform data transformations that will not require programming or knowledge of SQL. What should you do?

A. Use Cloud Dataflow with Beam to detect errors and perform transformations.

B. Use Cloud Dataprep with recipes to detect errors and perform transformations.

C. Use Cloud Dataproc with a Hadoop job to detect errors and perform transformations.

D. Use federated tables in BigQuery with queries to detect errors and perform transformations.

Correct Answer: B

15. You are designing an Apache Beam pipeline to enrich data from Cloud Pub/Sub with static reference data from BigQuery. The reference data is small enough to fit in memory on a single worker. The pipeline should write enriched results to BigQuery for analysis. Which job type and transforms should this pipeline use?

A. Batch job, PubSubIO, side-inputs

B. Streaming job, PubSubIO, JdbcIO, side-outputs

C. Streaming job, PubSubIO, BigQueryIO, side-inputs

D. Streaming job, PubSubIO, BigQueryIO, side-outputs

Correct Answer: C

16. You have a data pipeline that writes data to Cloud Bigtable using well-designed row keys. You want to monitor your pipeline to determine when to increase the size of your Cloud Bigtable cluster. Which two actions can you take to accomplish this? Choose 2 answers.

- A. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.
- B. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.
- C. Monitor the latency of write operations. Increase the size of the Cloud Bigtable cluster when there is a sustained increase in write latency.**
- D. Monitor storage utilization. Increase the size of the Cloud Bigtable cluster when utilization increases above 70% of max capacity.
- E. Monitor latency of read operations. Increase the size of the Cloud Bigtable cluster if read operations take longer than 100 ms.

Correct Answer: A, C

17. You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps.

You have the following requirements:

You will batch-load the posts once per day and run them through the Cloud Natural Language API.

You will extract topics and sentiment from the posts.

You must store the raw posts for archiving and reprocessing.

You will create dashboards to be shared with people both inside and outside your organization.

You need to store both the data extracted from the API to perform analysis as well as the raw social media posts for historical archiving. What should you do?

- A. Store the social media posts and the data extracted from the API in BigQuery.
- B. Store the social media posts and the data extracted from the API in Cloud SQL.

C. Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.

D. Feed to social media posts into the API directly from the source, and write the extracted data from the API into BigQuery.

Correct Answer: C

18. You need to migrate a 2TB relational database to Google Cloud Platform. You do not have the resources to significantly refactor the application that uses this database and cost to operate is of primary concern. Which service do you select for storing and serving your data?

- A. Cloud Spanner
- B. Cloud Bigtable
- C. Cloud Firestore
- D. Cloud SQL**

Correct Answer: D

19. You’re using Bigtable for a real-time application, and you have a heavy load that is a mix of read and writes. You’ve recently identified an additional use case and need to perform hourly an analytical job to calculate certain statistics across the whole database. You need to ensure both the reliability of your production application as well as the analytical workload. What should you do?

- A. Export Bigtable dump to GCS and run your analytical job on top of the exported files.
- B. Add a second cluster to an existing instance with a multi-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.**
- C. Add a second cluster to an existing instance with a single-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.
- D. Increase the size of your existing cluster twice and execute your analytics workload on your new resized cluster.

Correct Answer: B

20. You work for a bank. You have a labelled dataset that contains information on already granted loan application and whether these applications have been defaulted. You have been asked to train a model to predict default rates for credit applicants. What should you do?

- A. Increase the size of the dataset by collecting additional data.
- B. Train a linear regression to predict a credit default risk score.**
- C. Remove the bias from the data and collect applications that have been declined loans.
- D. Match loan applicants with their social profiles to enable feature engineering.

Correct Answer: B

21. You are deploying MariaDB SQL databases on GCE VM Instances and need to configure monitoring and alerting. You want to collect metrics including network connections, disk IO and replication status from MariaDB with minimal development effort and use StackDriver for dashboards and alerts. What you do?

- A. Install the OpenCensus Agent and create a custom metric collection application with a StackDriver exporter.
- B. Place the MariaDB instances in an Instance Group with a Health Check.
- C. Install the StackDriver Logging Agent and configure fluentd in_tail plugin to read MariaDB logs.**
- D. Install the StackDriver Agent and configure the MySQL plugin.

Correct Answer: C

22. You’re training a model to predict housing prices based on an available dataset with real estate properties. Your plan is to train a fully connected neural net, and you’ve discovered that the dataset contains latitude and longitude of the property. Real estate professionals have told you that the location of the property is highly influential on price, so you’d like to engineer a feature that incorporates this physical dependency. What should you do?

- A. Provide latitude and longitude as input vectors to your neural net.
- B. Create a numeric column from a feature cross of latitude and longitude.
- C. Create a feature cross of latitude and longitude, bucketize at the minute level and use L1 regularization during optimization.**

D. Create a feature cross of latitude and longitude, bucketize it at the minute level and use L2 regularization during optimization

Correct Answer: C

23. Your team is responsible for developing and maintaining ETLs in your company. One of your Dataflow jobs is failing because of some errors in the input data, and you need to improve reliability of the pipeline (incl. being able to reprocess all failing data). What should you do?

- A. Add a filtering step to skip these types of errors in the future, extract erroneous rows from logs.
- B. Add a try... catch block to your DoFn that transforms the data, extract erroneous rows from logs.
- C. Add a try... catch block to your DoFn that transforms the data, write erroneous rows to PubSub directly from the DoFn.**
- D. Add a try... catch block to your DoFn that transforms the data, use a side Output to create a PCollection that can be stored to PubSub later.

Correct Answer: C

24. You have migrated a Hadoop job from an on-prem cluster to dataproc and GCS. Your Spark job is a complicated analytical workload that consists of many shuffling operations and initial data are parquet files (on average 200-400 MB size each). You see some degradation in performance after the migration to Dataproc, so you'd like to optimize for it. You need to keep in mind that your organization is very cost-sensitive, so you'd like to continue using Dataproc on preemptible (with 2 non-preemptible workers only) for this workload. What should you do?

- A. Increase the size of your parquet files to ensure them to be 1 GB minimum.
- B. Switch to TFRecords formats (appr. 200MB per file) instead of parquet files.
- C. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.
- D. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.**

Correct Answer: D

25. You have an Apache Kafka Cluster on-prem with topics containing web application logs. You need to replicate the data to Google Cloud for analysis in BigQuery and Cloud Storage. The preferred replication method is mirroring to avoid deployment of Kafka Connect plugins. What should you do?

- A. Deploy a Kafka cluster on GCE VM Instances. Configure your on-prem cluster to mirror your topics to the cluster running in GCE. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.
- B. Deploy a Kafka cluster on GCE VM Instances with the PubSub Kafka connector configured as a Sink connector. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.
- C. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Source connector. Use a Dataflow job to read from PubSub and write to GCS.
- D. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Sink connector. Use a Dataflow job to read from PubSub and write to GCS.**

Correct Answer: D

26. You are a head of BI at a large enterprise company with multiple business units that each have different priorities and budgets. You use on-demand pricing for BigQuery with a quota of 2K concurrent on-demand slots per project. Users at your organization sometimes don't get slots to execute their query and you need to correct this. You'd like to avoid introducing new projects to your account. What should you do?

- A. Convert your batch BQ queries into interactive BQ queries.
- B. Create an additional project to overcome the 2K on-demand per-project quota.
- C. Switch to flat-rate pricing and establish a hierarchical priority model for your projects.**
- D. Increase the amount of concurrent slots per project at the Quotas page at the Cloud Console.

Correct Answer: C

27. You are a retailer that wants to integrate your online capabilities with different in-home assistants, such as Google Home. You need to interpret customer voice commands and issue an order to the backend systems. Which solutions should you choose?

- A. Cloud Speech-to-Text API
- B. Cloud Natural Language API

C. Dialogflow Enterprise Edition

d. Cloud AutoML Natural Language

Correct Answers: C

28. Your company has a hybrid cloud initiative. You have a complex data pipeline that moves data between cloud provider services and leverages services from each of the cloud providers. Which cloud-native service should you use to orchestrate the entire pipeline?

A. Cloud Dataflow

B. Cloud Composer

C. Cloud Dataprep

D. Cloud Dataproc

Correct Answers: B

29. You use a dataset in BigQuery for analysis. You want to provide third-party companies with access to the same dataset. You need to keep the cost of data sharing low and ensure that the data is current. Which solution should you choose?

A. Create an unauthorized view on the BigQuery table to control data access, and provide third-party companies with access to that view

B. Use Cloud Scheduler to export the data on a regular basis to cloud storage, and provide third-party companies with access to the new dataset

C. Create a separate dataset in BigQuery that contains the relevant data to share, and provide third-party companies with access to the new dataset

d. Create a Cloud Dataflow job that reads the data in frequent time intervals, and writes it to the relevant Bigquery dataset or Cloud Storage bucket for third-party companies to use

Correct Answers: A

30. You need to set access to BigQuery for different departments within your company. Your solution should comply with the following requirements: Each department should have access only to their data. Each department will have one or more leads who need to be able to create and update tables and provide them to their team. Each department has data analysts who need to be able to query but not to modify data. How should you set access to the data in BigQuery?

Google Cloud Certified Professional Cloud Architect Definitive Guide

- A. Create a dataset for each department. Assign the department leads the role of OWNER, and assign the data analysts the role of WRITER on their dataset
- B. Create a dataset for each department. Assign the department leads the role of WRITER, and assign the data analysts the role of READER on their dataset
- C. Create a table for each department, Assign the department leads the role of Editor, and assign the data analysts the role of Viewer on the project the table is in
- D. Create a table for each department, Assign the department leads the role of Editor, and assign the data analysts the role of viewer on the project table is in**

Correct Answers: D

31. You are operating a Cloud Dataflow streaming pipeline. The pipeline aggregates events from a Cloud Pub/Sub subscription source, within a window, and sinks the resulting aggregation to a Cloud Storage bucket. the source has consistent throughput. You want to monitor an alert on Behaviour of the pipeline with Cloud Stackdriver to ensure that it is processing data. Which Stackdriver alerts should you create?

- A. An alert based on a decrease of subscription/num-undelivered_messages for the source and a rate of change increase/storage/used_bytes for the destination
- B. An alert based on a increase of subscription/num-undelivered_messages for the source and a rate of change decrease/storage/used_bytes for the destination**
- C. An alert based on a decrease of instance/storage/used_bytes for the source and a rate of change increase of subscription/num-undelivered_messages for the destination
- D. An alert based on a increase of instance/storage/used_bytes for the source and a rate of change decrease of subscription/num-undelivered_messages for the destination

Correct Answer: B

32. You currently have a single on-premises Kafka cluster in a data center in the us-east region that is responsible for ingesting messages from IoT devices globally. Because large parts of the globe have poor internet connectivity, message sometimes batch at the edge, come in all at once, and cause a spike in load on your Kafka Cluster. This is becoming difficult to manage and

prohibitively expensive. What is the Google recommended could native architecture for the scenario?

- A. Edge TPUs as sensor devices for storing and transmitting the messages
- B. Cloud Dataflow connected to the Kafka cluster to scale the processing of incoming messages
- C. An IoT gateway connected to Cloud Pub/Sub, with Cloud Dataflow to read and process the messages from Cloud Pub/Sub**
- D. A Kafka cluster virtualized on Computing Engine in US-East with Cloud Load Balancing to connect to the devices around the world.

Correct Answer: C

33. You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-Time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to achieve these snapshots for a long time. Which two methods can accomplish this? Choose 2 answers.

- A. Use managed export and store the data in a Cloud Storage bucket using Nearline or Coldline class**
- B. Use managed export, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export
- C. Use managed export, and then import the data into BigQuery table created just for that export, and delete temporary export files
- D. Write an application that uses Cloud Datastore client libraries to read all the entities. Treat each entity as a BigQuery streaming insert. Assign an export timestamp for each export and attach it as an extra column for each row. Make sure that the BigQuery table is partitioned using the export timestamp column
- E. Write an application that uses Cloud Datastore client libraries to read all the entities. Format the exported data into a JSON file. Apply compression before storing the data in Cloud Source Repositories**

Correct Answer: A, E

34. You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis. Every hour, thousands of transactions are updated with a new status. The size of the initial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt?

- A. Denormalize the data as much as possible
- B. Preserve the structure of the data as much as possible
- C. Use BigQuery UPDATE to further reduce the size of the dataset**
- D. Develop a data pipeline where status updates are appended to BigQuery instead of updated
- E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro file. Use BigQuery's support for external data sources to query**

Correct Answers: C and E

35. You are designing a cloud-native historical data processing system to meet the following conditions: The data being analysed in CSV, Avro and PDF formats and will be accessed by multiple analysis tools including Cloud Dataproc, BigQuery and Compute Engine. A streaming data pipeline stores new data daily. Performance is not a factor in the solution. The solution design should maximize availability. How should you design data storage for this solution?

- A. Create a Cloud Dataproc cluster with high availability. Store the data in HDFS, and perform analysis as needed.
- B. Store the data in BigQuery. Access the data using BigQuery connector or Cloud Dataproc and Compute Engine
- C. Store the data in a regional Cloud Storage Bucket. Access the bucket directly using Cloud Dataproc, BigQuery and Compute Engine
- D. Store the data in a multi-regional Cloud Storage bucket. Access the data directly using Cloud Dataproc, BigQuery and Compute Engine**

Correct Answers: D

Google Cloud Certified Professional Cloud Architect Definitive Guide

36. You have a petabyte of analytics data and need to design a storage and processing platform for it. You must be able to perform data warehouse-style analytics on the data in Google Cloud and expose the dataset as files for batch analysis tools in other cloud providers. What should you do?

- A. Store and process the entire dataset in BigQuery
- B. Store and process the entire dataset in Cloud Bigtable**
- C. Store the full dataset in BigQuery, and store a compressed copy of the data in Cloud Storage Bucket
- D. Store the warm data as files in Cloud Storage, and store the active data in BigQuery. Keep this ratio as 80% warm and 20% active

Correct Answers: B

37. Your team wants to implement an app to help warehouse workers recognize incoming components based on a photo of the component. You want to implement the first working version of this app (AS Proof-Of-Concept) within a few working days. What should you do?

- A. Use Cloud Vision AutoML with the existing dataset
- B. Use Cloud Vision AutoML but reduce your dataset twice**
- C. Use Cloud Vision API by providing custom labels as recognition hints
- D. Train your own image recognition model leveraging transfer learning techniques

Correct Answers: B

38. You are working on a niche product in the image recognition domain. Your team has developed a model that is dominated by custom c++ Tensor Flow ops your team has implemented. These ops are used inside your main training loop and are performing bulky matrix multiplications. It currently takes up to several days to train a model. You want to decrease this time significantly and keep the cost low by using an accelerator on Google Cloud. What should you do?

- A. Use Cloud TPUs without any additional adjustment to your code
- B. Use Cloud TPUs after implementing GPU Kernel support for your custom ops**
- C. Use Cloud GPUs after implementing GPU Kernel support for your customs ops
- D. Stay on CPUs, and increase the size of the cluster you're training your model on

Correct Answers: B

39. You work on a regression problem in a natural language processing domain, and you have 100M labelled examples in your dataset. You have randomly shuffled your data and split your dataset into train and test samples (in a 90/10 ratio). After you trained the neural network and evaluated your model on a test set, you discover that the root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set. How should you improve the performance of your model?

- A. Increase the share of the test sample in the train-test-split
- B. Try to collect more data and increase the size of your dataset
- C. Try out regularization techniques (e.g., dropout or batch normalization) to avoid overfitting
- D. Increase the complexity of your model by e.g., introducing an additional layer or increasing the size of vocabularies or n grams used**

Correct Answers: D

40. The marketing team at your organization provides regular updates of a segment of your customer dataset. The marketing team has given you a CSV with 1 million records that must be updated in BigQuery. When you use the UPDATE statement in BigQuery, you receive a quota Exceeded error. What should you do?

- A. Reduce the number of records updated each day to stay within the BigQuery UPDATE DML statement limit
- B. Increase the BigQuery UPDATE DML statement limit in the quota management section of the Google Cloud Platform Console
- C. Split the source CSV file into smaller CSV file in Cloud Storage to reduce the number of BigQuery UPDATE DML statement per BigQuery job**
- D. Import the new records from the CSV file into a new BigQuery table. Create a BigQuery job that merges the new records with the existing records and writes the results to a new BigQuery table

Correct Answers: C

41. As your organization expands its usage of GCP, many teams have started to create their own projects. Projects are further multiplied to accommodate different stages of deployments and

Google Cloud Certified Professional Cloud Architect Definitive Guide

target audiences. Each project requires unique access control configurations. The central IT team needs to have access to all projects. Furthermore, data from Cloud Storage buckets and BigQuery datasets must be shared for the use in other projects. Which two steps should you take? Choose 2 answers.

- A. Use Cloud Deployment Manager to automate access provision**
- B. Introduce resource hierarchy to leverage access control policy inheritance
- C. Create distinct groups for various teams, and specify groups in Cloud IAM policies**
- D. Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets
- E. For each Cloud Storage bucket or BigQuery dataset, decide which projects need access. Find all the active members who have access to these projects and create a Cloud IAM policy to grant access to all these users

Correct Answers: A and C

42. Your United States based company has created an application for assessing and responding to user actions. The primary table's data volume grows by 250,000 records per seconds. Many third parties use your application's APIs to build the functionality in to their own frontend applications. Your application's APIs should comply with the following requirements: Single global endpoint ANSI SQL support consistent access to the most up-to-date data. What should you do?

- A. Implement BigQuery with no region selected for storage or processing
- B. Implement Cloud Spanner with the leader in North America and read-only replicas in ASIA and EUROPE**
- C. Implement Cloud SQL for PostgreSQL with the master in North America and read replicas in ASIA and EUROPE
- D. Implement Cloud Bigtable with the primary cluster in North America and secondary clusters in ASIA and EUROPE

Correct Answers: B

43. A data scientist has created a BigQuery ML model and asks you to create an ML pipeline to serve predictions. You have a REST API application with the requirement to serve predictions for an individual user ID with latency under 100 milliseconds. You use the following query to

Google Cloud Certified Professional Cloud Architect Definitive Guide

generate predictions: SELECT predicted_label, user_id from ML.PREDICT(MODEL 'dataset.model', table user_features), How should you create the ML pipeline?

- A. Add a WHERE clause to the query, and grant the BigQuery Data Viewer role to the application service account
- B. Create an Authorized view with the provided query. Share the dataset that contains the view with the application service account
- C. Create a Cloud Dataflow pipeline using BigQueryIO to read results from the query. Grant the Dataflow Worker role to the application service account
- D. Create a Cloud Dataflow pipeline using BigQueryIO to read predictions for all users from the query. Write the results to Cloud BigtableIO. Grant the Bigtable Reader role to the application service account so that the application can read predictions for individual users from Cloud Bigtable**

Correct Answers: D

44. You operate a database that stores stock trades and an application that retrieves averages stock price for a given company over an adjustable window of time. The data is stored in Cloud Bigtable where the datetime of the stock trade is the beginning of the row key. Your application has thousands of concurrent users and you notice that performance is starting to degrade as more stocks are added. What should you do to improve the performance of your application?

- A. Change the row key syntax in your Cloud Bigtable table to begin with the stock symbol**
- B. Change the row key syntax in your Cloud Bigtable table to begin with a random number per second
- C. Change the data pipeline to use BigQuery for storing stock trades, and update your application
- D. Use Cloud Dataflow to write summary of each day's stock trades to an AVRO file on Cloud Storage. Update your application to read from Cloud Storage and Cloud Bigtable to compute the responses

Correct Answers: A

45. You use BigQuery as your centralized analytics platform. New data is loaded every day, and an ETL pipeline modifies the original data and prepares it for the final users. This ETL pipeline is regularly modified and can generate errors, but sometime the errors are detected only after 2 weeks. You need to provide a method to recover from these errors and your backups should

be optimized for storage costs. How should you organize your data in BigQuery and store your backups?

- A. Organize your data in a single table, export and compress and store the BigQuery data in Cloud Storage
- B. Organize your data in separate tables for each month, and export, compress, and store the data in Cloud Storage
- C. Organize your data in separate tables for each month, and duplicate your data on a separate dataset in BigQuery
- D. Organize your data in separate tables for each month, and use snapshot decorators to restore the table to a time prior to the corruption**

Correct Answers: D

46. You have several Spark jobs that run on a Cloud Dataproc cluster on a schedule. Some of the jobs run in sequence, and some of the jobs run concurrently. You need to automate the process. What should you do?

- A. Create a Cloud Dataproc Workflow Template
- B. Create an initialization action to execute the jobs
- C. Create a Directed Acyclic Graph in Cloud Composer**
- D. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster

Correct Answers: C

47. You have historical data covering the last three years in BigQuery and a data pipeline that delivers new data to BigQuery daily. You have noticed that when the Data Science team runs a query filtered on a date column and limited to 30-90 days of data, the query scans the entire table. You also noticed that your bill is increasing more quickly than you expected. You want to resolve the issue as cost-effectively as possible while maintaining the ability to conduct SQL queries. What should you do?

- A. Re-create the tables using DDL. Partition the tables by a column containing a TIMESTAMP or DATE Type**

Google Cloud Certified Professional Cloud Architect Definitive Guide

- B. Recommend that the Data Science team export the table to a CSV file on Cloud Storage and use Cloud Datalab to explore the data by reading the files directly
- C. Modify your pipeline to maintain the last 30-90 days of data in one table and the longer history in a different table to minimize full table scans over the entire history
- D. Write an Apache Beam pipeline that creates a BigQuery table per day. Recommend that the Data Science team use wildcards on the table name suffixes to select the data they need

Correct Answers: A

48. You are designing a data processing pipeline. The pipeline must be able to scale automatically as load increases. Messages must be processed at least once and must be ordered within a window of 1 hour. How should you design the solution?

- A. Use Apache Kafka for message ingestion and use Cloud Dataproc for streaming analysis
- B. Use Apache Kafka for message ingestion and use Cloud Dataflow for streaming analysis
- C. Use Cloud Pub/Sub for message ingestion and Cloud Dataproc for streaming analysis
- D. Use Cloud Pub/Sub for message ingestion and Cloud Dataflow for streaming analysis**

Correct Answers: D

49. A shipping company has live package tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the life cycle of a package. The table was originally created with ingest-data partitioning. Over time the query performance in BigQuery, what should you do?

- A. Implement clustering in BigQuery on the ingest date column
- B. Implement clustering in BigQuery on the package-tracking ID column
- C. Tier older data onto Cloud Storage files, and leverage extended tables
- D. Re-create the table using data partitioning on the package delivery date**

Correct Answers: D

50. You operate a logistic company, and you want to improve event delivery reliability for vehicle-based sensors. You operate small data centers around the world to capture these

Google Cloud Certified Professional Cloud Architect Definitive Guide

events, but leased lines that provide connectivity from your event processing infrastructure are unreliable, with unpredictable latency. You want to address this issue in the most cost-effective way. What should you do?

- A. Deploy small kafka clusters in your data centers to buffer events
- B. Have the data acquisition devices publish data to Cloud Pub/Sub
- C. Establish a Cloud Interconnect between all remote data centers and Google
- D. Write a Cloud Dataflow pipeline that aggregates all data in Session windows**

Correct Answers: D

51. Your company is migrating their 30-node Apache Hadoop cluster to the cloud. They want to re-use Hadoop jobs they have already created and minimize the management of the cluster as much as possible. They also want to be able to persist data beyond the life of the cluster. What should you do?

- A. Create a Google Cloud Dataflow job to process the data.
- B. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.
- C. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.
- D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.**
- E. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.

Correct Answer: D

52. You are developing an application on Google Cloud that will automatically generate subject labels for users' blog posts. You are under competitive pressure to add this feature quickly, and you have no additional developer resources. No one on your team has experience with machine learning. What should you do?

- A. Call the Cloud Natural Language API from your application. Process the generated Entity Analysis as labels.**
- B. Call the Cloud Natural Language API from your application. Process the generated Sentiment Analysis as labels.
- C. Build and train a text classification model using TensorFlow. Deploy the model using Cloud Machine Learning Engine. Call the model from your application and process the results as labels.

D. Build and train a text classification model using TensorFlow. Deploy the model using a Kubernetes Engine cluster. Call the model from your application and process the results as labels.

Correct Answer: A

53. After migrating ETL jobs to run on BigQuery, you need to verify that the output of the migrated jobs is the same as the output of the original. You have loaded a table containing the output of the original job and want to compare the contents with output from the migrated job to show that they are identical. The tables do not contain a primary key column that would enable you to join them together for comparison. What should you do?

- A. Select random samples from the tables using the RAND() function and compare the samples.
- B. Select random samples from the tables using the HASH() function and compare the samples.**
- C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sorting. Compare the hashes of each table.
- D. Create stratified random samples using the OVER() function and compare equivalent samples from each table.

Correct Answer: B

Data Engineer Model Exam

1. You are building storage for files for a data pipeline on Google Cloud. You want to support JSON files. The schema of these files will occasionally change. Your analyst teams will use running aggregate ANSI SQL queries on this data. What should you do?

- A. Use BigQuery for storage. Provide format files for data load. Update the format files as needed.
- B. Use BigQuery for storage. Select "Automatically detect" in the Schema section.**
- C. Use Cloud Storage for storage. Link data as temporary tables in BigQuery and turn on the "Automatically detect" option in the Schema section of BigQuery.
- D. Use Cloud Storage for storage. Link data as permanent tables in BigQuery and turn on the "Automatically detect" option in the Schema section of BigQuery.

Correct Answer: B

2. You use a Hadoop cluster both for serving analytics and for processing and transforming data. The data is currently stored on HDFS in Parquet format. The data processing jobs run for 6 hours each night. Analytics users can access the system 24 hours a day. Phase 1 is to quickly migrate the entire Hadoop environment without a major re-architecture. Phase 2 will include migrating to BigQuery for analytics and to Cloud Dataflow for data processing. You want to make the future migration to BigQuery and Cloud Dataflow easier by following Google-recommended practices and managed services. What should you do?

- A. Lift and shift Hadoop/HDFS to Cloud Dataproc.
- B. Lift and shift Hadoop/HDFS to Compute Engine.
- C. Create a single Cloud Dataproc cluster to support both analytics and data processing, and point it at a Cloud Storage bucket that contains the Parquet files that were previously stored on HDFS.
- D. Create separate Cloud Dataproc clusters to support analytics and data processing, and point both at the same Cloud Storage bucket that contains the Parquet files that were previously stored on HDFS.**

Correct Answer: D

3. You are building a new real-time data warehouse for your company and will use Google BigQuery streaming inserts. There is no guarantee that data will only be sent in once but you do have a unique ID for each row of data and an event timestamp. You want to ensure that duplicates are not included while interactively querying data. Which query type should you use?

- A. Include ORDER BY DESC on timestamp column and LIMIT to 1.
- B. Use GROUP BY on the unique ID column and timestamp column and SUM on the values.
- C. Use the LAG window function with PARTITION by unique ID along with WHERE LAG IS NOT NULL.
- D. Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1.**

Correct Answer: D

4. You are designing a streaming pipeline for ingesting player interaction data for a mobile game. You want the pipeline to handle out-of-order data delayed up to 15 minutes on a per-player basis and exponential growth in global users. What should you do?

- A. Design a Cloud Dataflow streaming pipeline with session windowing and a minimum gap duration of 15 minutes. Use "individual player" as the key. Use Cloud Pub/Sub as a message bus for ingestion.**
- B. Design a Cloud Dataflow streaming pipeline with session windowing and a minimum gap duration of 15 minutes. Use "individual player" as the key. Use Apache Kafka as a message bus for ingestion.
- C. Design a Cloud Dataflow streaming pipeline with a single global window of 15 minutes. Use Cloud Pub/Sub as a message bus for ingestion.
- D. Design a Cloud Dataflow streaming pipeline with a single global window of 15 minutes. Use Apache Kafka as a message bus for ingestion.

Correct Answer: A

5. Your company is loading comma-separated values (CSV) files into Google BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file. What is the most likely cause of this problem?

- A. The CSV data loaded in BigQuery is not flagged as CSV.
- B. The CSV data had invalid rows that were skipped on import.

C. The CSV data loaded in BigQuery is not using BigQuery's default encoding.

D. The CSV data has not gone through an ETL phase before loading into BigQuery.

Correct Answer: C

6. Your company is migrating their 30-node Apache Hadoop cluster to the cloud. They want to re-use Hadoop jobs they have already created and minimize the management of the cluster as much as possible. They also want to be able to persist data beyond the life of the cluster. What should you do?

- A. Create a Google Cloud Dataflow job to process the data.
- B. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.
- C. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.
- D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.**
- E. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.

Correct Answer: D

7. You work for an economic consulting firm that helps companies identify economic trends as they happen. As part of your analysis, you use Google BigQuery to correlate customer data with the average prices of the 100 most common goods sold, including bread, gasoline, milk, and others. The average prices of these goods are updated every 30 minutes. You want to make sure this data stays up to date so you can combine it with other data in BigQuery as cheaply as possible. What should you do?

- A. Load the data every 30 minutes into a new partitioned table in BigQuery.
- B. Store and update the data in a regional Google Cloud Storage bucket and create a federated data source in BigQuery.**
- C. Store the data in Google Cloud Datastore. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Cloud Datastore.
- D. Store the data in a file in a regional Google Cloud Storage bucket. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

Correct Answer: B

8. You have 250,000 devices which produce a JSON device status event every 10 seconds. You want to capture this event data for outlier time series analysis. What should you do?

- A. Ship the data into BigQuery. Develop a custom application that uses the BigQuery API to query the dataset and displays device outlier data based on your business requirements.
- B. Ship the data into BigQuery. Use the BigQuery console to query the dataset and display device outlier data based on your business requirements.
- C. Ship the data into Cloud Bigtable. Use the Cloud Bigtable cbt tool to display device outlier data based on your business requirements.**
- D. Ship the data into Cloud Bigtable. Install and use the HBase shell for Cloud Bigtable to query the table for device outlier data based on your business requirements

Correct Answer: C

9. You are selecting a messaging service for log messages that must include final result message ordering as part of building a data pipeline on Google Cloud. You want to stream input for 5 days and be able to query the current status. You will be storing the data in a searchable repository. How should you set up the input messages?

- A. Use Cloud Pub/Sub for input. Attach a timestamp to every message in the publisher.**
- B. Use Cloud Pub/Sub for input. Attach a unique identifier to every message in the publisher.
- C. Use Apache Kafka on Compute Engine for input. Attach a timestamp to every message in the publisher.
- D. Use Apache Kafka on Compute Engine for input. Attach a unique identifier to every message in the publisher.

Correct Answer: A

10. You want to publish system metrics to Google Cloud from a large number of on-prem hypervisors and VMs for analysis and creation of dashboards. You have an existing custom monitoring agent deployed to all the hypervisors and your on-prem metrics system is unable to handle the load. You want to design a system that can collect and store metrics at scale. You don't want to manage your own time series database. Metrics from all agents should be written to the same table but agents must not have permission to modify or read data written by other agents. What should you do?

A. Modify the monitoring agent to publish protobuf messages to Cloud PubSub. Use a Dataproc cluster or Dataflow job to consume messages from Pubsub and write to BigTable.

- B. Modify the monitoring agent to write protobuf messages directly to BigTable.
- C. Modify the monitoring agent to write protobuf messages to HBase deployed on GCE VM Instances
- D. Modify the monitoring agent to write protobuf messages to Cloud Pubsub. Use a Dataproc cluster or Dataflow job to consume messages from Pubsub and write to Cassandra deployed on GCE VM Instances.

Correct Answer: A

11. You are designing storage for CSV files and using an I/O-intensive custom Apache Spark transform as part of deploying a data pipeline on Google Cloud. You intend to use ANSI SQL to run queries for your analysts. How should you transform the input data?

- A. Use BigQuery for storage. Use Cloud Dataflow to run the transformations.
- B. Use BigQuery for storage. Use Cloud Dataproc to run the transformations.**
- C. Use Cloud Storage for storage. Use Cloud Dataflow to run the transformations.
- D. Use Cloud Storage for storage. Use Cloud Dataproc to run the transformations.

Correct Answer: B

12. You are designing a relational data repository on Google Cloud to grow as needed. The data will be transnationally consistent and added from any location in the world. You want to monitor and adjust node count for input traffic, which can spike unpredictably. What should you do?

- A. Use Cloud Spanner for storage. Monitor storage usage and increase node count if more than 70% utilized.
- B. Use Cloud Spanner for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.**
- C. Use Cloud Bigtable for storage. Monitor data stored and increase node count if more than 70% utilized.
- D. Use Cloud Bigtable for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.

Correct Answer: B

13. You have a Spark application that writes data to Cloud Storage in Parquet format. You scheduled the application to run daily using DataProcSparkOperator and Apache Airflow DAG by Cloud Composer. You want to add tasks to the DAG to make the data available to BigQuery users. You want to maximize query speed and configure partitioning and clustering on the table. What should you do?

- A. Use "BashOperator" to call "bq insert".
- B. Use "BashOperator" to call "bq cp" with the "--append" flag.
- C. Use "GoogleCloudStorageToBigQueryOperator" with "schema_object" pointing to a schema JSON in Cloud Storage and "source_format" set to "PARQUET".**
- D. Use "BigQueryCreateExternalTableOperator" with "schema_object" pointing to a schema JSON in Cloud Storage and "source_format" set to "PARQUET".

Correct Answer: C

14. You have a website that tracks page visits for each user and then creates a Cloud Pub/Sub message with the session ID and URL of the page. You want to create a Cloud Dataflow pipeline that sums the total number of pages visited by each user and writes the result to BigQuery. User sessions timeout after 30 minutes. Which type of Cloud Dataflow window should you choose?

- A. A single global window
- B. Fixed-time windows with a duration of 30 minutes
- C. Session-based windows with a gap duration of 30 minutes**
- D. Sliding-time windows with a duration of 30 minutes and a new window every 5 minute

Correct Answer: C

15. You are designing a basket abandonment system for an ecommerce company. The system will send a message to a user based on these rules: a). No interaction by the user on the site for 1 hour b). Has added more than \$30 worth of products to the basket c). Has not completed a transaction. You use Google Cloud Dataflow to process the data and decide if a message should be sent. How should you design the pipeline?

- A. Use a fixed-time window with a duration of 60 minutes.

- B. Use a sliding time window with a duration of 60 minutes.
- C. Use a session window with a gap time duration of 60 minutes.**
- D. Use a global window with a time-based trigger with a delay of 60 minutes.

Correct Answer: C

16. You need to stream time-series data in Avro format, and then write this to both BigQuery and Cloud Bigtable simultaneously using Cloud Dataflow. You want to achieve minimal end-to-end latency. Your business requirements state this needs to be completed as quickly as possible.

What should you do?

- A. Create a pipeline and use ParDo transform.
- B. Create a pipeline that groups the data into a PCollection and uses the Combine transform.
- C. Create a pipeline that groups data using a PCollection and then uses Cloud Bigtable and BigQueryIO transforms.**
- D. Create a pipeline that groups data using a PCollection, and then use Avro I/O transform to write to Cloud Storage. After the data is written, load the data from Cloud Storage into BigQuery and Cloud Bigtable.

Correct Answer: C

17. Your company's on-premises Apache Hadoop servers are approaching end-of-life, and IT has decided to migrate the cluster to Google Cloud Dataproc. A like-for-like migration of the cluster would require 50 TB of Google Persistent Disk per node. The CIO is concerned about the cost of using that much block storage. You want to minimize the storage cost of the migration.

What should you do?

- A. Put the data into Google Cloud Storage.**
- B. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.
- C. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.
- D. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.

Correct Answer: A

18. You are designing storage for two relational tables that are part of a 10-TB database on Google Cloud. You want to support transactions that scale horizontally. You also want to optimize data for range queries on non-key columns. What should you do?

- A. Use Cloud SQL for storage. Add secondary indexes to support query patterns.
- B. Use Cloud SQL for storage. Use Cloud Dataflow to transform data to support query patterns.
- C. Use Cloud Spanner for storage. Add secondary indexes to support query patterns.**
- D. Use Cloud Spanner for storage. Use Cloud Dataflow to transform data to support query patterns.

Correct Answer: C

19. Your company is streaming real-time sensor data from their factory floor into Bigtable and they have noticed extremely poor performance. How should the row key be redesigned to improve Bigtable performance on queries that populate real-time dashboards?

- A. Use a row key of the form <timestamp>.
- B. Use a row key of the form <sensorid>.
- C. Use a row key of the form <timestamp>#<sensorid>.
- D. Use a row key of the form <sensorid>#<timestamp>.**

Correct Answer: D

20. You are developing an application on Google Cloud that will automatically generate subject labels for users' blog posts. You are under competitive pressure to add this feature quickly, and you have no additional developer resources. No one on your team has experience with machine learning. What should you do?

- A. Call the Cloud Natural Language API from your application. Process the generated Entity Analysis as labels.**
- B. Call the Cloud Natural Language API from your application. Process the generated Sentiment Analysis as labels.
- C. Build and train a text classification model using TensorFlow. Deploy the model using Cloud Machine Learning Engine. Call the model from your application and process the results as labels.
- D. Build and train a text classification model using TensorFlow. Deploy the model using a Kubernetes Engine cluster. Call the model from your application and process the results as labels.

Correct Answer: A

21. Your company is using WILDCARD tables to query data across multiple tables with similar names. The SQL statement is currently failing with the error shown below. Which table name will make the SQL statement work correctly?

```
# Syntax error: Expected end of statement but got "--" at [4:11]
SELECT age
FROM
    bigquery-public-data.noaa_gsod.gsod
WHERE
    age != 99
    AND _TABLE_SUFFIX = '1929'
ORDER BY
    age DESC
```

- A. `bigquery-public-data.noaa_gsod.gsod`
- B. bigquery-public-data.noaa_gsod.gsod*
- C. 'bigquery-public-data.noaa_gsod.gsod*'`
- D. `bigquery-public-data.noaa_gsod.gsod*`

Correct Answer: D

22. You are working on an ML-based application that will transcribe conversations between manufacturing workers. These conversations are in English and between 30-40 sec long. Conversation recordings come from old enterprise radio sets that have a low sampling rate of 8000 Hz, but you have a large dataset of these recorded conversations with their transcriptions. You want to follow Google-recommended practices. How should you proceed with building your application?

- A. Use Cloud Speech-to-Text API, and send requests in a synchronous mode.
- B. Use Cloud Speech-to-Text API, and send requests in an asynchronous mode.
- C. Use Cloud Speech-to-Text API, but resample your captured recordings to a rate of 16000 Hz.
- D. Train your own speech recognition model because you have an uncommon use case and you have a labelled dataset.

Correct Answer: A

Google Cloud Certified Professional Cloud Architect Definitive Guide

23. You are developing an application on Google Cloud that will label famous landmarks in users' photos. You are under competitive pressure to develop a predictive model quickly. You need to keep service costs low. What should you do?

- A. Build an application that calls the Cloud Vision API. Inspect the generated MID values to supply the image labels.
- B. Build an application that calls the Cloud Vision API. Pass client image locations as base64-encoded strings.**
- C. Build and train a classification model with TensorFlow. Deploy the model using Cloud Machine Learning Engine. Pass client image locations as base64-encoded strings.
- D. Build and train a classification model with TensorFlow. Deploy the model using Cloud Machine Learning Engine. Inspect the generated MID values to supply the image labels.

Correct Answer: B

24. You are building a data pipeline on Google Cloud. You need to select services that will host a deep neural network machine-learning model also hosted on Google Cloud. You also need to monitor and run jobs that could occasionally fail. What should you do?

- A. Use Cloud Machine Learning to host your model. Monitor the status of the Operation object for 'error' results.
- B. Use Cloud Machine Learning to host your model. Monitor the status of the Jobs object for 'failed' job states.**
- C. Use a Kubernetes Engine cluster to host your model. Monitor the status of the Jobs object for 'failed' job states.
- D. Use a Kubernetes Engine cluster to host your model. Monitor the status of Operation object for 'error' results.

Correct Answer: B

25. You work on a regression problem in a natural language processing domain, and you have 100M labelled examples in your dataset. You have randomly shuffled your data and split your dataset into training and test samples (in a 90/10 ratio). After you have trained the neural network and evaluated your model on a test set, you discover that the root-mean-squared error

(RMSE) of your model is twice as high on the train set as on the test set. How should you improve the performance of your model?

- A. Increase the share of the test sample in the train-test split.
- B. Try to collect more data and increase the size of your dataset.
- C. Try out regularization techniques (e.g., dropout or batch normalization) to avoid overfitting.
- D. Increase the complexity of your model by, e.g., introducing an additional layer or increasing the size of vocabularies or n-grams used to avoid underfitting.**

Correct Answer: D

26. You are using Cloud Pub/Sub to stream inventory updates from many point-of-sale (POS) terminals into BigQuery. Each update event has the following information: product identifier "prodSku", change increment "quantityDelta", POS identification "termId", and "messageId" which is created for each push attempt from the terminal. During a network outage, you discovered that duplicated messages were sent, causing the inventory system to over-count the changes. You determine that the terminal application has design problems and may send the same event more than once during push retries. You want to ensure that the inventory update is accurate. What should you do?

- A. Inspect the "publishTime" of each message. Make sure that messages whose "publishTime" values match rows in the BigQuery table are discarded.
- B. Inspect the "messageId" of each message. Make sure that any messages whose "messageId" values match corresponding rows in the BigQuery table are discarded.
- C. Instead of specifying a change increment for "quantityDelta", always use the derived inventory value after the increment has been applied. Name the new attribute "adjustedQuantity".
- D. Add another attribute orderId to the message payload to mark the unique check-out order across all terminals. Make sure that messages whose "orderId" and "prodSku" values match corresponding rows in the BigQuery table are discarded.**

Correct Answer: D

27. You designed a database for patient records as a pilot project to cover a few hundred patients in three clinics. Your design used a single database table to represent all patients and their visits, and you used self-joins to generate reports. The server resource utilization was at

50%. Since then, the scope of the project has expanded. The database table must now store 100 times more patient records. You can no longer run the reports, because they either take too long or they encounter errors with insufficient compute resources. How should you adjust the database design?

- A. Add capacity (memory and disk space) to the database server by the order of 200.
- B. Shard the tables into smaller ones based on date ranges, and only generate reports with pre-specified date ranges.
- C. Normalize the master patient-record table into the patients table and the visits table, and create other necessary tables to avoid self-join.**
- D. Partition the table into smaller tables, with one for each clinic. Run queries against the smaller table pairs, and use unions for consolidated reports.

Correct Answer: C

28. Your startup has never implemented a formal security policy. Currently, everyone in the company has access to the datasets stored in Google BigQuery. Teams have the freedom to use the service as they see fit, and they have not documented their use cases. You have been asked to secure the data warehouse. You need to discover what everyone is doing. What should you do first?

- A. Use Google Stackdriver Audit Logs to review data access.**
- B. Get the identity and access management (IAM) policy of each table.
- C. Use Stackdriver Monitoring to see the usage of BigQuery query slots.
- D. Use the Google Cloud Billing API to see what account the warehouse is being billed to.

Correct Answer: A

29. You created a job which runs daily to import highly sensitive data from an on-premises location to Cloud Storage. You also set up a streaming data insert into Cloud Storage via a Kafka node that is running on a Compute Engine instance. You need to encrypt the data at rest and supply your own encryption key. Your key should not be stored in the Google Cloud. What should you do?

- A. Create a dedicated service account, and use encryption at rest to reference your data stored in Cloud Storage and Compute Engine data as part of your API service calls.

Google Cloud Certified Professional Cloud Architect Definitive Guide

- B. Upload your own encryption key to Cloud Key Management Service, and use it to encrypt your data in Cloud Storage. Use your uploaded encryption key and reference it as part of your API service calls to encrypt your data in the Kafka node hosted on Compute Engine.
- C. Upload your own encryption key to Cloud Key Management Service, and use it to encrypt your data in your Kafka node hosted on Compute Engine.
- D. Supply your own encryption key, and reference it as part of your API service calls to encrypt your data in Cloud Storage and your Kafka node hosted on Compute Engine.**

Correct Answer: D

30. You are working on a project with two compliance requirements. The first requirement states that your developers should be able to see the Google Cloud Platform billing charges for only their own projects. The second requirement states that your finance team members can set budgets and view the current charges for all projects in the organization. The finance team should not be able to view the project contents. You want to set permissions. What should you do?

- A. Add the finance team members to the default IAM Owner role. Add the developers to a custom role that allows them to see their own spend only.
- B. Add the finance team members to the Billing Administrator role for each of the billing accounts that they need to manage. Add the developers to the Viewer role for the Project.**
- C. Add the developers and finance managers to the Viewer role for the Project.
- D. Add the finance team to the Viewer role for the Project. Add the developers to the Security Reviewer role for each of the billing accounts.

Correct Answer: B

ROI Exam

1. Storage of JSON files with occasionally changing schema for ANSI SQL queries

- A. Store in BigQuery. Provide format files for data load and update them as needed.
- B. Store in BigQuery. Select "Automatically detect" in the Schema section.**
- C. Store in Cloud Storage. Link data as temporary tables in BigQuery and turn on the "Automatically detect" option in the Schema section of BigQuery.
- D. Store in Cloud Storage. Link data as permanent tables in BigQuery and turn on the "Automatically detect" option in the Schema section of BigQuery.

Correct Answer: B

2. Low-cost one-way one-time migration of two 100-TB file servers to GCP; data will only be accessed from Germany

- A. Use Transfer Appliance. Transfer to a Cloud Storage Regional storage bucket.**
- B. Use Transfer Appliance. Transfer to a Cloud Storage Multi-Regional bucket.
- C. Use Storage Transfer Service. Transfer to a Cloud Storage Regional bucket.
- D. Use Storage Transfer Service. Transfer to a Cloud Storage Multi-Regional bucket.

Correct Answer: A

3. Cost-effective backup to GCP of multi-TB databases from another cloud including monthly DR drills

- A. Use Transfer Appliance. Transfer to Cloud Storage Nearline bucket.
- B. Use Transfer Appliance. Transfer to Cloud Storage Coldline bucket.
- C. Use Storage Transfer Service. Transfer to Cloud Storage Nearline bucket.**
- D. Use Storage Transfer Service. Transfer to Cloud Storage Coldline bucket.

Correct Answer: C

4. 250,000 devices produce a JSON device status every 10 seconds. How do you capture event data for outlier time series analysis?

- A. Capture data in BigQuery. Develop a BigQuery API custom application to query the dataset and display device outlier data.
- B. Capture data in BigQuery. Use the BigQuery console to query the dataset and display device outlier data.
- C. Capture data in Cloud Bigtable. Use the Cloud Bigtable cbt tool to display device outlier data.**
- D. Capture data in Cloud Bigtable. Install and use the HBase shell for Cloud Bigtable to query the table for device outlier data.

Correct Answer: C

5. Event data in CSV format to be queried for individual values over time windows. Which storage and schema to minimize query costs?

- A. Use Cloud Bigtable. Design tall and narrow tables, and use a new row for each single event version.**
- B. Use Cloud Bigtable. Design short and wide tables, and use a new column for each single event version.
- C. Use Cloud Storage. Join the raw file data with a BigQuery log table.
- D. Use Cloud Storage. Write a Cloud Dataprep job to split the data into partitioned tables.

Correct Answer: A

6. Customer wants to maintain investment in existing Apache Spark code data pipeline.

- A. BigQuery
- B. Cloud Dataflow
- C. Cloud Dataproc**
- D. Cloud Dataprep

Correct Answer: C

7. Host a deep neural network machine learning model on GCP. Run and monitor jobs that could occasionally fail.

- A. Use Cloud Machine Learning to host your model. Monitor the status of the Operation object for 'error' results.
- B. Use Cloud Machine Learning to host your model. Monitor the status of the Jobs object for 'failed' job states.**
- C. Use a Kubernetes Engine cluster to host your model. Monitor the status of the Jobs object for 'failed' job states.
- D. Use a Kubernetes Engine cluster to host your model. Monitor the status of the Operation object for 'error' results.

Correct Answer: B

8. Cost effective way to run non critical Apache Spark jobs on Cloud Dataproc?

- A. Set up a cluster in high availability mode with high memory machine types. Add 10 additional local SSDs.
- B. Set up a cluster in high availability mode with default machine types. Add 10 additional preemptible worker nodes.
- C. Set up a cluster in standard mode with high memory machine types. Add 10 additional preemptible worker nodes.**
- D. Set up a cluster in standard mode with the default machine types. Add 10 additional local SSDs.

Correct Answer: C

9. Promote a Cloud Bigtable solution with a lot of data from development to production and optimize for performance.

- A. Change your Cloud Bigtable instance type from Development to Production, and set the number of nodes to at least 3. Verify that the storage type is HDD.
- B. Change your Cloud Bigtable instance type from Development to Production, and set the number of nodes to at least 3. Verify that the storage type is SSD.**
- C. Export the data from your current Cloud Bigtable instance to Cloud Storage. Create a new Cloud Bigtable Production instance type with at least 3 nodes. Select the HDD storage type. Import the data into the new instance from Cloud Storage.

Google Cloud Certified Professional Cloud Architect Definitive Guide

D. Export the data from your current Cloud Bigtable instance to Cloud Storage. Create a new Cloud Bigtable Production instance type with at least 3 nodes. Select the SSD storage type. Import the data into the new instance from Cloud Storage.

Correct Answer: B

10. As part of your backup plan, you want to be able to restore snapshots of Compute Engine instances using the fewest steps.

- A. Export the snapshots to Cloud Storage. Create disks from the exported snapshot files. Create images from the new disks.
- B. Export the snapshots to Cloud Storage. Create images from the exported snapshot files.**
- C. Use the snapshots to create replacement disks. Use the disks to create instances as needed.
- D. Use the snapshots to create replacement instances as needed.

Correct Answer: D

11. You want to minimize costs to run Data Studio reports on BigQuery queries by using prefetch caching.

- A. Set up the report to use the Owner's credentials to access the underlying data in BigQuery, and direct the users to view the report only once per business day (24-hour period).
- B. Set up the report to use the Owner's credentials to access the underlying data in BigQuery, and verify that the 'Enable cache' checkbox is selected for the report.**
- C. Set up the report to use the Viewer's credentials to access the underlying data in BigQuery, and also set it up to be a 'view only' report.
- D. Set up the report to use the Viewer's credentials to access the underlying data in BigQuery, and verify that the 'Enable cache' checkbox is not selected for the report.

Correct Answer: B

12. A Data Analyst is concerned that a BigQuery query could be too expensive.

- A. Use the LIMIT clause to limit the number of values in the results.
- B. Use the SELECT clause to limit the amount of data in the query. Partition data by date so the query can be more focused.**

Google Cloud Certified Professional Cloud Architect Definitive Guide

- C. Set the Maximum Bytes Billed, which will limit the number of bytes processed but still run the query if the number of bytes requested goes over the limit.
- D. Use GROUP BY at the end so the results will be grouped into fewer output values.

Correct Answer: B

13. BigQuery data is stored in external CSV files in Cloud Storage; as the data has increased, the query performance has dropped.

- A. Import the data into BigQuery for better performance.**
- B. Request more slots for greater capacity to improve performance.
- C. Divide the data into partitions based on date.
- D. Time to move to Cloud Bigtable; it is faster in all cases.

Correct Answer: A

14. Source data is streamed in bursts and must be transformed before use.

- A. Use Cloud Bigtable for fast input and cbt for ETL.
- B. Ingest data to Cloud Storage. Use Cloud Dataproc for ETL.
- C. Use Cloud Pub/Sub to buffer the data, and then use BigQuery for ETL.
- D. Use Cloud Pub/Sub to buffer the data, and then use Cloud Dataflow for ETL.**

Correct Answer: D

15. Calculate a running average on streaming data that can arrive late and out of order.

- A. Use Cloud Pub/Sub and Cloud Dataflow with Sliding Time Windows.**
- B. Use Cloud Pub/Sub and Google Data Studio.
- C. Cloud Pub/Sub can guarantee timely arrival and order.
- D. Use Cloud Dataflow's built-in timestamps for ordering and filtering.

Correct Answer: A

