Google Cloud Professional Cloud Architect Master Cheat Sheet

Google's infrastructure

Backbone

Google has built a global, meshed backbone network to interconnect their data centers and to deliver traffic to their Edge Points of Presence (POPs)

PoPs

70+ edge PoPs in 33 countries interconnected via the Backbone network

Edge Caching

This is a caching platform that sits on top of their infrastructure network. Edge locations can be found in virtually every country.

Regions

- Specific geographical locations where you can run your resources
- Collection of zones
- Regional resources are available to resources in any zone in the region

Example: *us-central1*

Multi Regions

- A multi-regional location is a **general** geographical area, such as the United States.
    - multi-region locations contain **multiple** *regional* locations.
- A regional location is a **specific** geographical area, such as South Carolina.
    - All regional locations are seperated from other regional locations by at least 100 miles.

Example: Multi-region = *us-central*

Zones

- Isolated locations within a region
- Zonal resources are only available in that zone

Example: *us-central1-a*

GCP Networking Fundamentals

VPC Networks

Pricing

Resources

- Sub-hour billing (billed in minutes)

Network traffic

- Ingress is free
- Egress is charged
  - Egress to some GCP services sometimes free

Sustained-use discounts

- Sustained use discounts are applied automatically; there is no action required on your part to enable these discounts.
- Discounts increase
- You can get up to a 30% net discount for instances that run the entire month.
- Machines that run more for than 25% of the month (incremental discount for each percent past 25%)
- No upfront costs and not tied to machine type

Example: https://cloudplatform.googleblog.com/2014/04/introducing-sustained-use-discounts.html

Note: Understand, watch a video about it, read Joe's notes.

Committed-use discounts

- Similar to AWS Reserved Instances (RIs) TBA

Security

- All data is encrypted at rest
- Network encryption ** All control information is encrypted ** All WAN traffic to be encrypted automatically ** Moving towards encrypting all local traffic within data centres

Google advises you to always "distrust the network". They have created a security model called Beyond Corp that shifts access controls from the network permiteter to individuals devices and users, allows employees to work securely from any location without needing a VPN

Main Product Offerings

Organization & Projects

Google Resource Hierarchy

- Organization
- Projects
- Resources

It's important to remember that a resource can only have one parent project

Projects

Control access to resources

Components:

- Project Name (Friendly Name)
- Project ID (App ID)
  - Must be globally unique
  - Cannot be changed once set
- Project Number

  - Used in places to identify resources that belong to specific projects

Interacting with Google Cloud Platform

Cloud Shell

- Pre-configured Google SDK Linux Instance
- Automatic auth based upon GCP Console login
- Accessible via any web browser
- All client libs for web apps pre-installed
- 5GB persistent storage

Note

- Direct/interactive use only
  - If in violation, session can be terminated without notice

CLI

gcloud

Allows you to manage Google Cloud Platform resources and developer workflow

Format: `gcloud [GROUP] [GROUP] [COMMAND] --arguments` Example: `gcloud compute instances create instance-1 --zone us-central1-a`
Note

- `gcloud alpha...`
  - Feature is typically not ready for Production
- `gcloud beta...`
  - Feature on the other hand is normally a completed feature, that is being tested to be production ready.

gsutil

API

Cloud IAM

Provides granular access to resources, prevents unwanted access to other resources and adopts the security principle of least privilege.

Core Components:

- Members (Who)
    - Person (Google Account)
    - Google Group
    - Service Account
        - Special type of account belonging to your application and can be identified by `<project_number/id>@developer.gserviceaccount.com`
- Permissions & Roles (What)
    - Role
        - Collection of *permissions* to use or manage GCP resources
        - Assigned to users
    - Permissions
        - Give access to a given *resource*
        - Identified by `<service>.<resource>.<verb>`
            - E.g. `pubsub.subscriptions.consume`
- Resources
- Policies
    - Bind Members to Roles at a hierarchy level
        - Such as Organisation, Folder, Project or Resource
    - Collection of *Roles* that define who has what type of access
    - Are hierarchally defined, with parent overruling child policy

Service Accounts

- **Global**
- Similar to AWS Role
- This is a special type of Google account that represents an application, not an end user
- Can be "assumed" by applications or individual users when authorised

Service Account Keys

GCP-managed keys

- Keys used by GCP services such as App Engince and Compute Engine

- Key cannot be downloads
- Rotated automatically on a weekly basis

User-managed keys

- Keys are created, downloadable, and managed by users
- Expire 10 years from creation

Primitive vs Predefined Roles

Primitive

Historical roles before Cloud IAM was implemented, they are applied at the *Project* level and the scope is very broad.

Types:

- Viewer
  - Read only
- Editor
  - Read + Write
- Owner
  - Read + Write
  - Manage access to Project and resources
  - Setup project billing

Predefined

Much more granular access, they are applied at the *Resource* level

Cloud Identity

- **Global**
- Identity as a Service (IDaaS) to provision and manage users and groups
- Supports MFA and enforcement, including security keys
- Identities can be used to SSO with other apps via OIDC, SAML, OAuth2
- Can sync from AD and LDAP directories via Google Cloud Directory Sync
- Free Google Accounts for non Google Suite users, tied to a verified domain

Cloud Resource Manager

- **Global**
- Hierarchically manage resources by project, folder, and organization
    - Organisation is root node in hierarchy
- Provides a Recycle bin which allows you to undelete projects
- You can define custom IAM roles at the organisation level
- Can apply IAM policies at organisation, folder or project levels

Cloud Audit Logging

- **Global**
- Who did what, where, and when?
- Maintains two audit logs for each project and organisation:
    - Admin activity
        - 400 day retention which is free
    - Data Access
        - 7 day retention which is free but if you require 30 days of retention then you have to pay
- This is a Stackdriver service

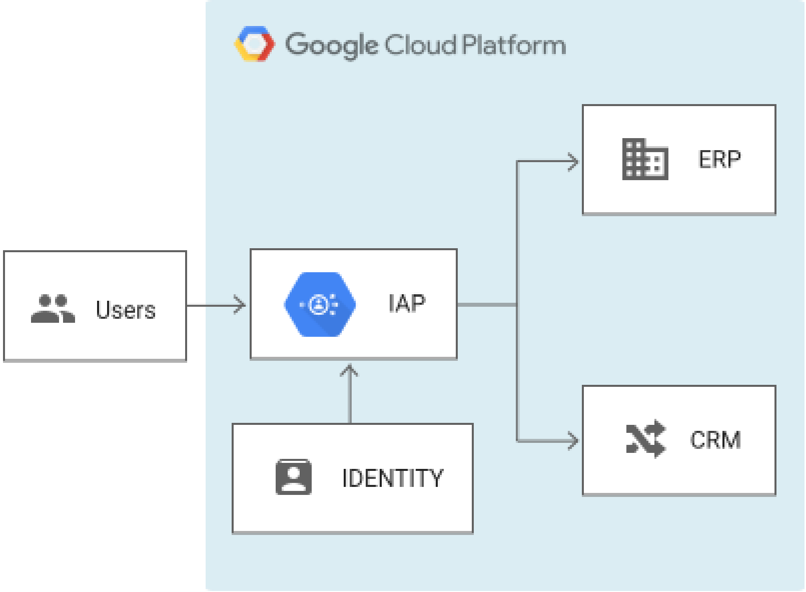
Cloud KMS

- **Regional** and **Global**
- Low-latency service to manage and use AES256 encryption keys, to protect secrets
- Rotate keys automatically or on demand
    - Keeps old active key version, to allow decypting
- Key deletion has 24 hour delay to prevent accidental or malicious data loss
- You pay for:
    - the active key versions stored over time
    - key use operations such as encyption and decryption

Cloud IAP (Identity Aware Proxy)

- Guards apps running on GCP through identity verification instead of VPN access
- Based on CLB and IAM and woll only pass authorised requests

- Pay for load balancing / protocol forwarding rules and traffic



Security Scanner

- **Global**
- Free but limited GAE application vulnerability scanner with "very low false positive rates"
- Crawler based
- Can detect:
  - XSS (Cross-site-scripting)
  - Flash injection
  - Mixed content (Is there HTTP content within HTTPS)
  - Outdated/insecure libraries

Global, Regional, and Zonal Resources

Resource	Global	Regional	Zonal
Images	x		
Snapshots	x		
Instance Templates	x		

Resource	Global	Regional	Zonal
VPC Network	x		
Firewalls	x		
Routes	x		
Addresses		x	
Subnets		x	
Regional Managed Instance Groups		x	
Instances			x
Disks			x
Machine Types			x
Zonal Managed Instance Groups			x

Compute options

GCE

- **Zonal**
- IAAS
- VMs referred to as Instances
- Offers complete control and most flexibility at the cost of the following adminstrative burdens;
    - CPU/GPU
    - Memory
    - Disk Space
    - OS
    - Firewall Controls

    - Network Connection/management (VPN/Load Balancing)

Instance Groups

You can create and manage groups of VM instances so that you don't have to individually control each instance in your project. Compute Engine offers two different types of instance groups:

- managed
    - zonal managed instance groups
    - regional managed instance groups
- unmanaged

Managed

Uses instance templates to create a group of identical resources. Making changes to instances will make the changes to the whole instance group, benefits of homongenous grouping of VM instances are:

- Automatic scaling
- Work with Load balancing to distribute traffic to all of the instances in the group
- If an instance in a group stops, crashes, or is deleted then the group automatically recreates the instance

Zonal

A Zonal managed instance group will contain instances from the same zone.

Note: Choose zonal if you want to avoid the slightly higher latency incurred by cross-zone communication or if you need fine-grained control of the sizes of your groups in each zone.

Regional

A Regional group will contain instances from multiple zones across the region.

Note: This is general recommended group over Zonal as it protects against zonal failures and unforeseen scenarios where an entire group of instances in a single zone malfunctions.

Unmanaged

Unmanaged groups are groups of dissimilar instances that you can arbitrarily add and remove from the group. Unmanaged instance groups do not offer autoscaline, rolling update support, or the use of instance templates.

Note: Use if you need to apply load balancing to your pre-existing configurations or to groups of dissimilar instances.

GKE

- **Regional**
- Managed Applications not machines
- Powered by Kubernetes
- Deploy containerized applications
    - De-couples app components from OS
    - Run app in multiple envs, regardless of OS
    - Kubernetes DNS on by default
        - No need for Consul unless wanted
- No IAM integration
    - To connect with other GCP services you have top manage these secrets more manually.
- Production clusters require >=3 nodes
- Integrates with Persistent Storage on underlying GCE components

App Engine

- **Regional**
- Managed Service
    - No adminstration is needed for underlying infrastructure
    - Deployment, maintenance, and scalability handled
- Developers can focus on writing the code, while Google handles the rest
- Build scalable web apps and mobile backends

Standard

Supports:

- Python
- Java
- Go
- PHP

Flexible

Supports:

- Java 8
- Servlet 3.1
- Jetty 9
- Python 2.7 & 3.5
- Node.js
- Ruby
- PHP
- .NET core
- Go
- AND any other custom runtime if using a custom Docker image

Cloud Functions

- **Regional**
- FaaS (Functions as a Service, I.e. Serverless)
- Runs Node.js code in response to an event
  - Triggers can include:
    - GCS Objects
    - Pub/Sub Messages
    - HTTP Request
- Pay for CPU & RAM assigned to function per 100ms (mins. 100ms)
- Massive scalability (horizontally)

Firebase

TBA

Storage, DB, & Transfer

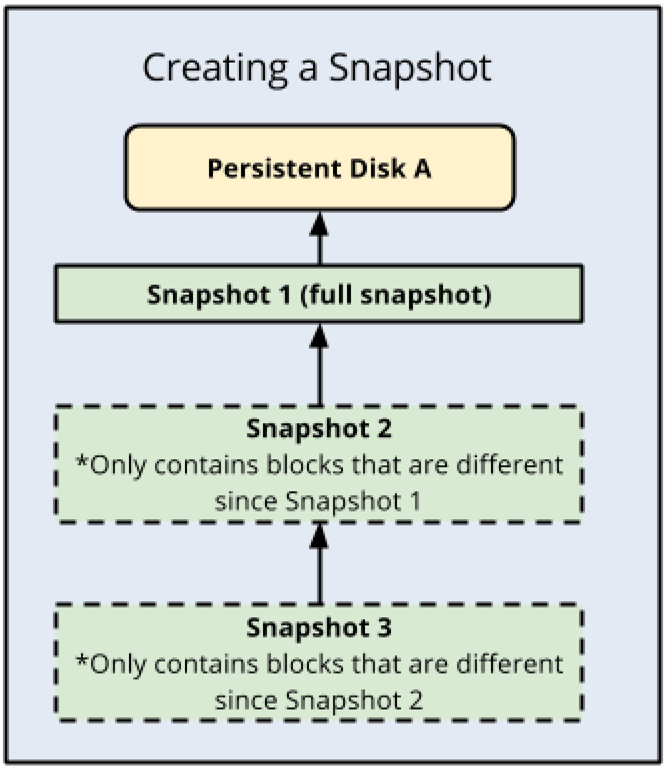
Local SSD

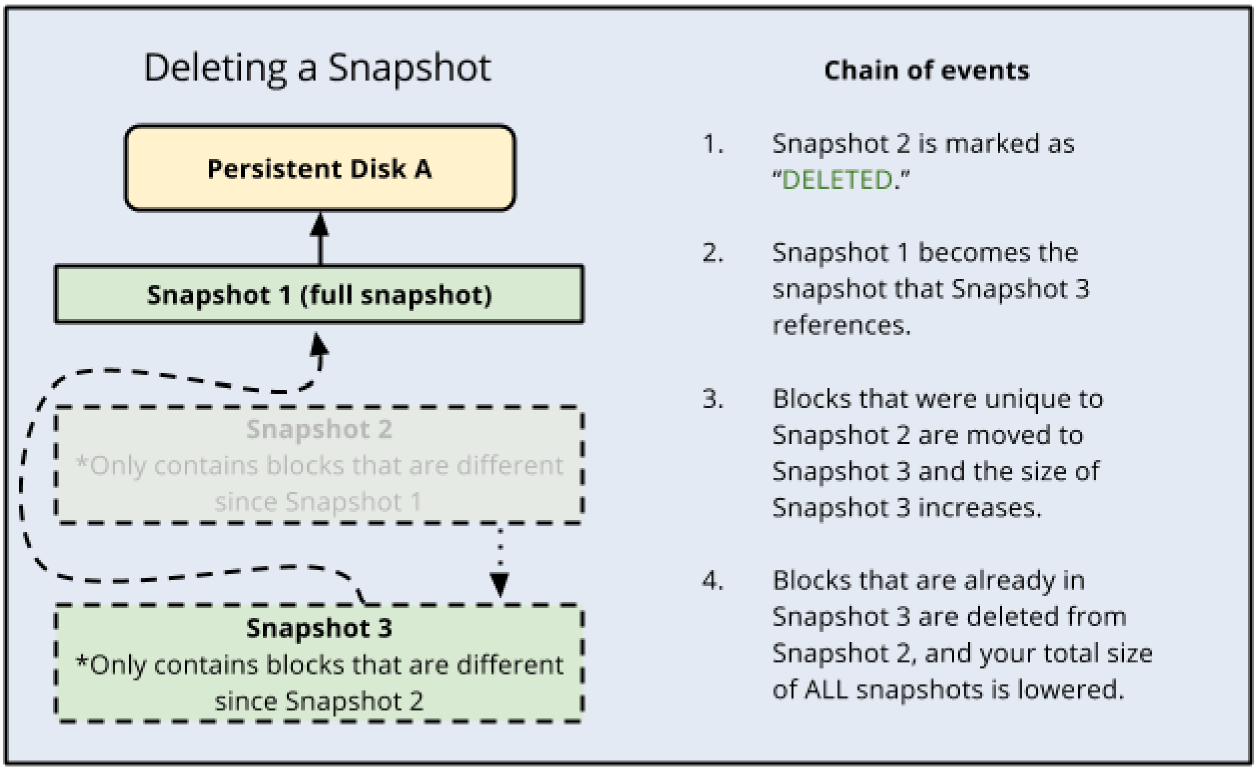
- **Zonal**
- Data is encypted at rest
- 375GB SSD attached to each server

  - Similar to the ephemeral disk on AWS
- Data is lost whenever the instance shuts down
- Data survives live migrations
- Pay for GB-month provisioned

Persistent Disk

- **Zonal**
- Persistent disks
- Performance scales with wolume size
  - Performance is way below that of a Local SSD but is still fast.
- Can resize while in use but will need file system update within VM
  - Max file size: 10TB
- Pricing = Incremental storage difference * ($ * time)


Creating a Snapshot
Persistent Disk A
Snapshot 1 (full snapshot)
Snapshot 2
*Only contains blocks that are different since Snapshot 1
Snapshot 3
*Only contains blocks that are different since Snapshot 2


Deleting a Snapshot
Persistent Disk A
Snapshot 1 (full snapshot)
Snapshot 2
*Only contains blocks that are different since Snapshot 1
Snapshot 3
*Only contains blocks that are different since Snapshot 2

Chain of events

1. Snapshot 2 is marked as "DELETED."
2. Snapshot 1 becomes the snapshot that Snapshot 3 references.
3. Blocks that were unique to Snapshot 2 are moved to Snapshot 3 and the size of Snapshot 3 increases.
4. Blocks that are already in Snapshot 3 are deleted from Snapshot 2, and your total size of ALL snapshots is lowered.

Snapshots

- **Global**
- Can Snapshot with Persistent Disk and make machine images out of them
- Similar to EBS

Cloud SQL

- **Regional**
- Fully Managed
- Databases:
  - MySQL
  - PostgreSQL
- Supports automatice replication, backup, failover
- Scaling is manual (both vertically and horizontally)

Cloud Spanner

- **Regional / Multi-Regional / Global**
- Horizontally scalable
- Strongly consistent
  - Provides external consistency which is far more than stongly
- Relational database structure with non-relational horizontal scale
- Supports SQL to query data
- ACID transactions
- Scales from 1 to thousands of nodes
- Requires 3 nodes for a Production environment
- Use only for large systems
  - Not for small apps/systems
- Pay for provisioned node time (region/multi-region) and used storage time
  - Production systems are very costly

BigQuery

- **Multi-Regional**
- Serverless column-store data warehouse
- Supports SQL to query data
- Pay for:
  - GBs actually considered (scanned) during queries
    - Attemtps to reuse cached results, which are free
  - Data stored (GB-months)
    - Relatively inexpensive
    - This gets cheaper when tables are not modified for 90 days
  - Streaming inserts paid per GB

Cloud Datastore

- **Regional / Multi-Regional**
- Fully Managed
- NoSQL DB
- Similar to DynamoDB
- Capabilities:
  - ACID transactions
  - SQL-like queries
  - Indexes
  - RESTful interface
- Pay for GB-months of storage used
- Pay for IO operations (r,w,deletes) performed

Notes:

- Cloud Datastore was born as the structured data store for AppEngine
- Scales from 0 to terabytes worth of data as your application grows

Bigtable

- **Regional**
- Fully Managed
- Low latency
- High throughput
- NoSQL DB
- Used for **large** operational and analytical applications
- Supports:
  - HBase API
- Integrates with:
  - Hadoop
  - Dataflow
  - Dataproc
- Automatic storage scaling
- Manual processing nodes scaling
- Pay per processing node hours
- Pay for GB-hours used for storage
  - Cheap HDD or fast SSD

When should I use it?

- Storing 1TB and more of structured data
- When there is a very high high volume of writes
- When read and write latency reqwuirements are that of a single digit millisecond range with strong consistency
- When a clear, straightforward migration from HBase to a managed cloud service is required

Cloud Storage

- **Regional Multi-Regional**
- Fully Managed
- Strongly Consistent (for overwirte PUTs and DELETEs)
- Durability = 11 9'S
- Can provide site hosting funtionality
- Lifecycle features

Data Transfer Appliance

- Rackable, high-capacity storage server
- Physically transfer (ship) data from your data centre to GCS
- Similar to AWS Snowball
- Ingest only
- 100 or 480TB versions

Storage Transfer Service

- **Global**
- If data is not in your own data centre then you can use this
- Destination is always GCS bucket
- Source can be:
  - S3
  - HTTP/HTTPS endpoint
  - GCS Bucket
- Pay for it's actions, such as data transfer

Google Domains

- **Global**
- Google's registrar for domain names
- Built-in DNS or custom nameservers
- Supports DNSSEC

Cloud DNS

- **Global**
- DNS service
- 100% uptime guarantee
- Low latency globally
- Supports DNSSEC
- Pay for:
  - Hosted zone, fixed fee
  - DNS lookups (i.e. usage)

Static IP

- **Regional Global**
- Two types:
  - Regional Static IP
    - GCE Instances
    - Network Load Balancers
  - Global Static IP (Anycast IP)
    - Global Load Balancers
      - HTTP(S)
      - SSL Proxy
      - TCP Proxy *Note:*
- Pay for IPs that are not in use

Cloud Load Balancing (CLB)

- **Regional Global**

- Built into their Software Defined Networking (SDN) system that can naturally handle spikes without any prewarming

- Two types of availability:

  - Regional Network Load Balancer

        - Supports:
            - Session Affinity
                - Setting up uses forwarding rules based on IP, protocol (TCP/UDP), and (optionally) port
            - Round Robin
            - Health Checks
    - Global Load Balancer
        - Supports:
            - Multi-region failover for HTTP(S), SSL Proxy, and TCP Proxy

- LB Types
    - HTTPS(S) Load Balancing
        - HTTP LB
        - HTTPS LB
        - Internet facing or single and multi-region
    - TCP Load Balancing
        - TCP LB
        - SSL Proxy
        - TCP Proxy
        - Internet facing or single and multi-region
    - UDP Load Balancing
        - UDP LB
        - Internet facing or single region

Cloud CDN

- **Global**
- Low-latency content delivery
- Based on HTTP(S) CLB & integrated with GCE & GCS
- Supports GCP only
    - HTTP(S) LBs
        - Backend can be a GCS Bucket
    - Does not support custom origins
- Supports the following protocols of HTTP/2 and HTTPS

- Pay for
  - POP to client network egress
  - HTTP(S) request volume
  - Per cache invalidation request (not per resource)
- VPC
- **Global**
- Global IPv4 unicast SDN for GCP resources
- **Subnets are Regional**
- Can:
  - Be shared across multiple Projects
  - Be peered with other VPCs
  - Enable private (internal IP) access to some GCP services (e.g. BQ, GCS)
- Pay for:
  - Certain services (e.g. VPN)
  - Network egress

Cloud Interconnect

- **Regional Multi-Regional**

Use Case

Say you have an application running within GCP on a GCE instance but you need to let the application access data from a business system on-premise then you would choose to Cloud Interconnect

- Connecting external networks to Google's network

Direct access to RFC1918 IPs in your VPC - with SLA (Private Connections)

- Dedicated Interconnect
- Cloud VPN

Access to Google public IPs only - without SLA Peering

- Direct Peering
- Carrier Peering

Cloud VPN

- **Regional**
- IPsec VPN
- To connect to VPC via public internet for low-volume data connections
- Persistent, static connections between gateways
  - Not for a Dynamic client
- VPN Gateways must have static IP
- **Encrypted link to VPC, into one subnet**
- Supports both Static and Dynamic routing
  - Dynamic is preferred to stop the need to re-establish the connection
- 99.9% availability SLA
- Pay per tunnel-hour
- Normal traffic charges apply

Dedicated Interconnect

- **Regional Multi-Regional**
- Direct physical link between VPC and on-prem for high-volume data connections
- VLAN attachment is private connection to VPC in one region; no public GCP APIs
- Link are private but not encrypted
  - You need to layer your own encryption in order to achieve encrypted traffic
- Redundant connections are advised to provide high availabilty achieving 99.99% SLA.
  - Without redundant conneciotns the SLA is 99.9%
- Pay fee per 10Gbps link, plus small fee per VLAN attachment

Cloud Router

- **Regional**

- Dynamic routing using BGP for hybrid networks linking GCP VPCs to external networks
- Works with Cloud VPN and Dedicated Interconnect
- Automatically learns subnets in VPC and announces them to on-prem network
- Without Cloud Router you must manage static routes for VPN
- Free to setup
- Pay for VPC egress

CDN Interconnect

- **Regional Multi-Regional**
- Direct, low-latency connectivity to certain CDN providers, with cheaper egress
- For external CDNs, not GCP's CDN service
- Supports:
  - Akami
  - Cloudflare
  - Fastly
- Contact CDN provider to set up for GCP project and which regions
- Free to enable, then pay less for the egress you configured

Big Data & IoT

Big Data Lifecycle

IoT Core

- **Global**
- Fully Managed
- A service to connect, manage, and ingest data from devices globally
- Devices connect securely using IoT industry-standard MQTT or HTTPS protocols
- CA signed certs can be used to verify device ownership on first connect
- Pay per MB of data exchanged with devices

Device Manager

- Handles device identity, authentication, config, and control

Protocol Bridge

- Publishes incoming telemtry to Cloud PubSub for processing

Cloud Pub/Sub (Publish/Subscribe)

- **Global**
- Infinitely scalable
- At least once messaging for ingestion, decoupling etc...
- Can be thought of as the "glue" that links everything together
- Pay for data volume; min 1KB per publish/push/pull request, not charged per message
- Can even end up being the replacement for things such as AWS Kinesis or Apache Kafka

Components

- Topic and Subscribers
  - A publisher sends a message to that topic which will then get sent to all the subscribers

Messages

- Can be up to 10MB
- Undelivered messages are strored for 7 days
  - There is no DLQ (Dead Letter Queue)

Modes

Push

- Delivers to HTTPS endpoints
- Will delete messages when it receives an HTTP success code
- Uses a "slow-start" algorithm which ramps up on success and backs off & retries, on failures

Pull

- Delivers messages to requesting clients and waits for ACK to delete or until the timer expires
- Lets clients set rate of consumption, and suppors batching and long-polling (Similar to AWS SQS)

Cloud Dataprep

- **Global**
- Visually explore, clean, and prep data for analysis without running servers
- Ad-hoc ETL, for BA's and not IT professionals
- Managed version of Trifacta Wrangler
- Source data can be from the services and types below, formatted in CSV, JSON, or relational:
  - GCS
  - BigQuery
  - File Upload
- Automatically detects schemas, datatypes, possible joins, and various anomalies
- Pay for underlying Dataflow job, plus management overhead charge on top of the services accessed

Cloud Dataproc

- **Zonal**
- Batch MapReduce processing via configurable, managed Spark & Hadoop clusters
- Scales, by removing or adding nodes, even whhile jobs are running
- Integrates with:
  - GCS
  - BigQuery
  - Bigtable
  - Some Stackdriver services
- Pay for:
  - underlying GCE servers used in the cluster
  - a Cloud Dataproc management fee per vCPU-hour in the cluster
- **You should use this service to move *existing* Spark/Hadoop setups to GCP**
  - You should use Cloud Dataflow for new data processing pipelines

Cloud Dataflow

- **Zonal**
- Fully Managed Apache Beam
- Smartly-autoscaled and dynamically redistributes lagging work, mid-job, to optimise run time
- Batch or Stream MapReduce-like processing
- Integrates with:
    - Cloud Pub/Sub
    - Datastore,
    - BigQuery
    - Bigtable
    - Cloud ML
    - Stackdriver
- Pay for underlying worker GCE via consildated charges
    - Pay per second for vCPUs, RAM GBs, and Persistent Disks
    - Dataflow *Shuffle* charged for time per GB used

Cloud Datalab

- **Regional**
- Interactive tool for data exploration, analysis, visualization and machine learning
- Uses Jupyter Notebook

Cloud Data Studio

- **Global**
- Big Data Visualisation tool for dashboards and reporting
- Similiar to AWS Quicksight and Tableau

Operations and Management

Stackdriver

- **Global**

- Family of services for monitoring, logging, and diagnosing apps on GCP and AWS

Stackdriver Monitoring

- **Global**
- Provides visibility into perf, uptime, and overall health of cloud apps
    - Based on collectd
- Includes built in custom metrics, dashboards, global uptime monitoring and alerts
- Can follow a trail, such as Linking from an alert, then to the dashboards, to logs, and then to the traces
- Premium Monitorigin can support AWS
- Sends alerts via email, and GCP Mobile App
    - Premium can send to SMS, Slack, SNS, HipChat, webhook, etc…
- Pay per time series per month for custom logs-based metrics allotment overages

Stackdriver Logging

- **Global**
- Similar to Splunk and Cloudwatch Logs
- Store, search, analyse, and alert on log data and events
    - Based on Fluentd
- Send any logs through API alongwith built in support for some GCP services and AWS with an agent
- Create real-time metrics from log data, then alert or chart them on dashboards
- Send real-time log data to BigQuery for advanced analytics and SQL-like querying
- When logs are about to expire you can export to GCS
- Pay per project per month; pay for premium to get more per hour

Stackdriver Reporting

- **Global**
- Counts, analyses, aggregates, and tracks crashes within a centralised interface
- Alert when a new application error cannot be grouped with existing ones
- Link directly from notifications to error details

- Exception stack trace parser knows:
  - Java
  - Python
  - JavaScript
  - Ruby
  - C#
  - PHP
  - Go

Stackdriver Trace

- **Global**
- Tracks and displays call tree and timings across distributed systems to debug performance
- Automaticall captures traces from App Engine
- Zipkin collector allows Zipkin tracers to submit data to Trace
- Generate reports on demand and get daily auto reports per traced app

Stackdriver Debugger

- **Global**
- Grabs program state (callstate, vars, expressions) in live deploys
- Source view supports:
  - Cloud Source Repository
  - Github
  - Bitbucket
  - Local and Upload
- Share debuggin session with others all you need to do is send the URL
- Free to use

Cloud Deployment Manager

- **Global**
- Similar to Terraform and Cloudformation
- Create and Manage resources via declarative tempaltes
- Templates written in:
  - YAML
  - Python

  - Jinja2
- Supports input and output parameters
- Create and update of deployments both support preview

Cloud Billing API

- **Global**
- Programmatically manage billing for GCP projects and get GCP pricing

Development & APIs

Cloud Endpoints

- **Global**
- Handles auth, monitoring, logging, and API keys for APIs backed by GCP
- Based on NGINX and runs on a container (running on instances), called an ESP (Extensible Service Proxy) which is super fast and hook into the Cloud Load Balancer
- Uses JWT
- Integrates with:
  - Firebase
  - Auth0
  - Google Auth
- Pay per call to your API

Research

- xinetd
- helm
  - Kubernetes, Jenkins, Helm (Reference)
- Apache Beam (Relates to Data Flow)
- Organsiational setup suing Projects for isolation etc...
- CAPEX vs OPEX
- `gcloud alpha` VS `gcloud beta` VS `gcloud`
- How do you share a VPC network from one project to another within an organisation.
  - You have to configure a Shared VPC

        - What is a Shared VPC?
            - It's this

Training Exercises

- Kubernetes and Helm within GCP