**HSBC Pune C - Notes - 6 & 7 July 2020**

Start time: 9:00am Pune Time

Please use your personal devices for the links below

**Link to the Slide Deck:**

[**https://tinyurl.com/ybk3qhmt**](https://tinyurl.com/ybk3qhmt)

**Link to this document (Classroom Notes):**

[**https://tinyurl.com/yalng5z3**](https://tinyurl.com/yalng5z3)

Link to Google’s Cloud Architect Certification Site: <https://cloud.google.com/certification/cloud-architect>

We will resume at 4:50pm

In the meantime - if you can please complete the course survey

<https://bit.ly/2YfuFyU>

(Link is also in the Chat)

This document is where Grant will type his notes.

Link to Google’s Cloud Architect Certification Site: <https://cloud.google.com/certification/cloud-architect>

I want to build a timesheet collection program - what should I use???

Current have it written as a JAVA app - how should run it on Google

Pattern: Heavy usage on particular days (but not every day, end of 2-weeks pay / end of month)

Computing: 6 choices on Google

First Decision: Do you want to provide CODE or a CONTAINER

Code-base (python, java, php, go, c#, etc)

Virtual Machine

App Engine Standard - we give google code, they run it on their environment (python, java, php, go)

Cloud Functions - we give them a function - not an entire program - node, python, go

Container-based (microservices, docker, etc)

Kubernetes - complex container (dozens or more containers to run together)

App Engine Flex - simple container deployments (think one docker container running isolated)

Cloud Run - run a stateless docker container for 100ms to seconds at a time

If I choose compute: Do I need a REALLY BIG VM or can it be distributed across many small VMs

Vertical Scalability - Make a VM bigger

Horizontal Scalability - Can it run on more than one VM

Compute Engine

Virtual MAchines

CPU+RAM + Disk

N1-standard-1 (1x CPU, 3.75GB) vs n1-standard-4 (1xCPU, 15GB RAM)

1 CPU/standard ratio of RAM vs 4 CPU/standard ratio of RAM

N1-highmem-2 - 2 CPUs, higher ratio of RAM to CPU (2x CPU, 13GB RAM)

N1-highcpu-2 - 2 CPUs, lower ratio of RAM to CPU (2x CPU, 1.8GB RAM)

Disks on VMs

Persistent Disks (spinning hard drives) - cheaper than persistent SSD ($0.04/GB)

Larger disks are faster than smaller disks

SSDs are faster than Persistent Disks

Persistent SSD (SSD-based disks) - more expensive than persistent disk ($0.17/GB)

PAY FOR WHAT YOU ALLOCATE (so if you allocate a 500GB disk - even if you don’t store data in it - you are paying for it)

Key concepts on VMs

To increase disk IOPS - increase disk size or move to Persistent SSD

To increase network bandwidth - INCREASE CPU

Assume a HOST has 96 CPU Cores

What if I build a 1CPU VM - and consume all the networking? They won’t let me do that. I’ll be throttled to 2giga-bits per second of bandwidth. Each CPU gets you more bandwidth

YOU DO NOT add more networks to get more network capacity. ONE NETWORK CARD PER NETWORK on GCP

So if you have ONE VPC Network defined - each VM can only have ONE Network Card

If you have TWO VPC Networks defined (maybe a mainnet and dmznet) - you might have a VM that has TWO networks - one on each network

What to expect on the Exam

50 Multiple Choice Questions

Select the correct answer from the list of 4 (select 1 of 4)

Multiple-Select - select TWO items from the list below that solve the problem

On Multiple Select - they will TELL YOU how many to choose

If they say select 2, you must choose 2 - or the answer will be graded incorrect

No PARTIAL CREDIT - you must get BOTH (so if one was right, one was wrong - question is wrong)

Suggested Experience

3+ years of industry experience - TCP/IP (knowledge of IP, TCP/UDP, basic firewaalls, what routers, etc), Security Least Privilege, Separation of Duties, etc, Basic Application design (monolithic, microservices, etc), how to build docker container (Dockerfile)

YOU DO NOT have to know specific languages (Python, Java, etc) - general programming basics

1+ years designing and managing solutions using GCP - this is what you have been doing for the last 2 months (coursera, qwiklabs, design projects and this workshop) - so you are getting it accelerated

Common Patterns on Google

AppEngine + Firestore + Cloud Storage (GCS) (all three are GCP-native, autoscale, no pre-allocation)

PostgreSQL, MySQL or Microsoft SQL 2017 - CloudSQL

Batch Processing into a Data Warehouse (we will see this is case studies)

GCS -> Dataflow -> BigQuery

Streaming Processing into a Data Warehouse

Pub/Sub -> Dataflow -> BigQuery

Hadoop on GCP - Hadoop/HDFS

1. Lift and Shift to **Google Dataproc** (Run the cluster and the storage (HDFS)
   1. Stateful cluster - that must run 24x7 because the HDFS has files in it
2. Move the HDFS into GCS, and update references hdfs://folder/files to gs://bucket/folder/files
   1. Now we have a STATELESS cluster (compute only) - GCS has our files in it - so we can create/destroy the cluster as needed

Complex Container Orchestration - Kubernetes

Full Control of the Operating System - VMs, Compute Engine, GCE

Node.js on Google - Cloud Functions

Host a STATIC CONTENT WEBSITE - DNS -> Bucket (http) or DNS->Load Balancer->Bucket (https)

LocalSSD - High performance cache - but non permanent

Backups -> Store Backups in Buckets

Major Migration Pathways

Lift and Shift (Run what you ran on Premise up in GCP)

VM on Premise -> VM on Google

Kubernetes on Premise -> GKE - Google Kuberetes Engine on Google

(I doubt you would see a question that the correct answer is build your OWN kubernetes cluster on GCP - GKE would be the best answer)

MySQL on Premise -> VM on Google OR CloudSQL running MySQL

PostgreSQL on Premise -> VM on Google OR CloudSQL running PostgreSQL

If face with the choice - CloudSQL running PostgreSQL is the more “googlie-answer”

(it’s google’s exam - choose the more googlie-answer, unless they convince you otherwise)

Replatform/Re-engineer

VM on Premise (running Java or Python) -> App Engine (going to take a little work), but app is becoming more cloud native/Google handles more of the management

VM on Premise (with node.js) -> Cloud Function

Moving from MySQL to Spanner or Firestore/DataStore

(we can’t just lift and shift - there WILL be code changes)

Questions are NOT going to ask trivial - but you should know scale

No pricing - but you should know that per GB

GCS is cheaper than

Persistent Disks which are cheaper than

Persistent SSD

GCS bills based on consumption (how much data stored for how long)

Persistent Disks and Persistent SSD - bill on what you ALLOCATE and for how long you had it allocated

Example:

GCS Multi-regional, standard storage is 2.6c/per GB/per month (and I only pay for what I store)

Persistent Disk in us-central1 is 4c/per GB/per month (and I pay for what I allocate)

Persistent SSD in us-central1 is 17c/per GB/per month (and I pay for what I allocate)

Pay for CPU and RAM when the VM is running - you pay for the disk from the time to create it until you delete it (so if a VM is running - CPU+RAM+Disk, if a VM is stopped - you just pay for the disk)

Persistent Disks live in the Disk Infrastructure

There is ONE type of disk that lives on the HOST - Local SSD

Local SSDs CANNOT be used for the operating system

Local SSDs are erased at shutdown/NON-persistent. SWAP/Temporary/Ephermeral Disk

Add at BOOT time, removed at Shutdown time - pay only while the VM is running

EXTREMELY FAST - but they don’t survive a VM shutdown/poweroff/crash/etc

CANNOT ADD/REMOVE LOCAL SSD while the VM is running. (because they had to put my VM on specific hardware that had localSSDs available) 1 to 8 Local SSDs per VM (new configuration also allows 16 or 24)

Only come 375GB at a time - you can’t have half a local SSD, 100GB - must be a full unit of 375GB

CloudSQL has a VM behind the scenes (you defined CPU+RAM) and you are paying for CloudSQL from creation until deletion - no way to STOP cloud SQL.

Cloud SQL has a maximum supported disk of 30TB

Question will steer you to Cloud SQL or Cloud Spanner often based on SIZE

CloudSQL is GB to TB (up to 30TB)

Cloud Spanner is usually TB to PB (no limits - but expensive)

**MAKE THEM CONVINCE YOU OF THESE PRODUCTS**

Cloud Spanner (minimum of $600/month, TB to PB scale)

BigTable (minimum of $1500/month - 3-node production cluster, TB to PB scale)

Local SSD (temporary/swap - do NOT persist)

**Additional Google Products we should know on a HIGH LEVEL**

**(one or two sentences - what they do)**

**DLP (Data Loss Prevention) API -** put data (text) or images thru DLP - and identify and/or redact PII information.

**Shielded VMs** - extra security on VMs - vTMP (virtual Trusted Platform Module), UEFI boot (instead of old BIOS boot), Attestations about Boot (boot 1, shutdown1, boot2, shutdown2 in logging), and low level Boot/rootkit protection

**Storage Transfer Appliance** - 100TB or 480TB sizes - google ships you the device

USING the device’s web interface - you set your encryption

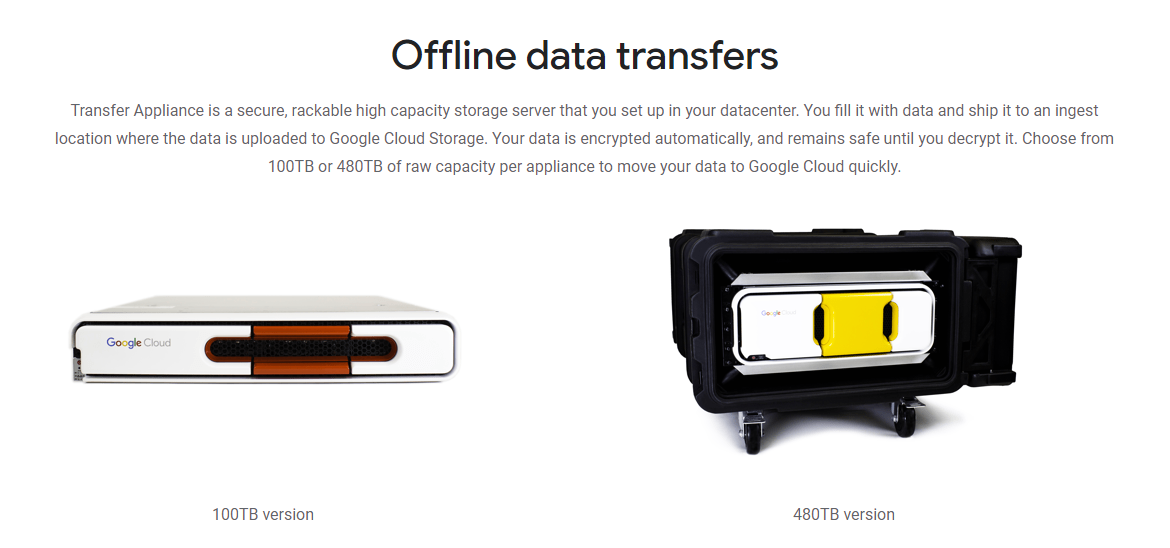
Copy your data onto the device

Ship the device back to Google

Google stages your ENCRYPTED data

YOU - **REHYDRATE** YOUR DATA -- using the **REHYDRATOR** Appliance from the Marketplace (VM)

Good for 40TB/60TB or more - or data would take longer than 1 week to upload



**Cloud Armor** - Firewall for Load Balancer. Distributed Denial of Service (DDoS) protection, BlockAllow IP address ranges/County/Region IP ranges (block NZ)

**CDN** - stores Static content -

Load cache upon first pull of an cache item

Serves that content from the cache

Caches are closer to the users - **reduce latency to users**

Caches serving content are cheaper than service the content directly (**lower egress costs**)

We do NOT specifically load the Google CDN

Content is CACHES on the fly thru the Load Balancer

**IAP** - Identity Aware Proxy - send you to a LOGIN page before you can access a web application running on Load Balancer or IAP

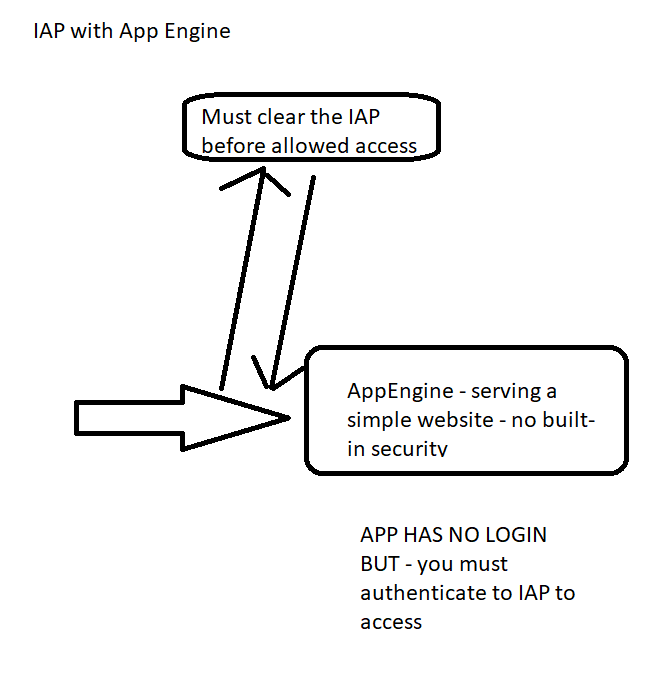
So I’m forced to a Google Login - where I have to provide a credential

Two-factor (One-time Code, Confirm in Chrome)

Since I’m entering my credential - it could be redirect back to my SAML2.0 infrastructure for Single-Sign-On (SSO)

Grant’s @roitraining.com - went to Azure AD

HSBC’s Single-Sign-On - goes to a SAML2.0 interface and verifies password against Active Directory, and returns a token that Google identifies as proof of credentials



**Access Transparency (not likely on Cloud Architect exam - it’s on the Security Exam)**

Normally - all access is logged with Stackdriver logging - via the APIs being called

BUT WHAT IF GOOGLE NEEDS TO DO SOMETHING ON THE BACK-end as part of troubleshooting - that isn’t done thru a traditional API

“Expand visibility and control over your cloud provider with near real-time logs and approval controls”

**Cloud DataPrep** - program-less (we write recipes) to transform our data - in conjunction with Trifacta

Executes the recipe using **Dataflow** as the underlying engine

Load Dataset -> Write a recipe (using Trifacta syntax) - aggregate, remove/correct data, etc

**Cloud Data Fusion** - writing workflows and executes on a HADOOP (DataProc) (not a third-party - it uses an opensource workflow platform and executes on a DataProc cluster)

**IoT Core** - Internet of Things - Sensors sending data into GCP for Analysis

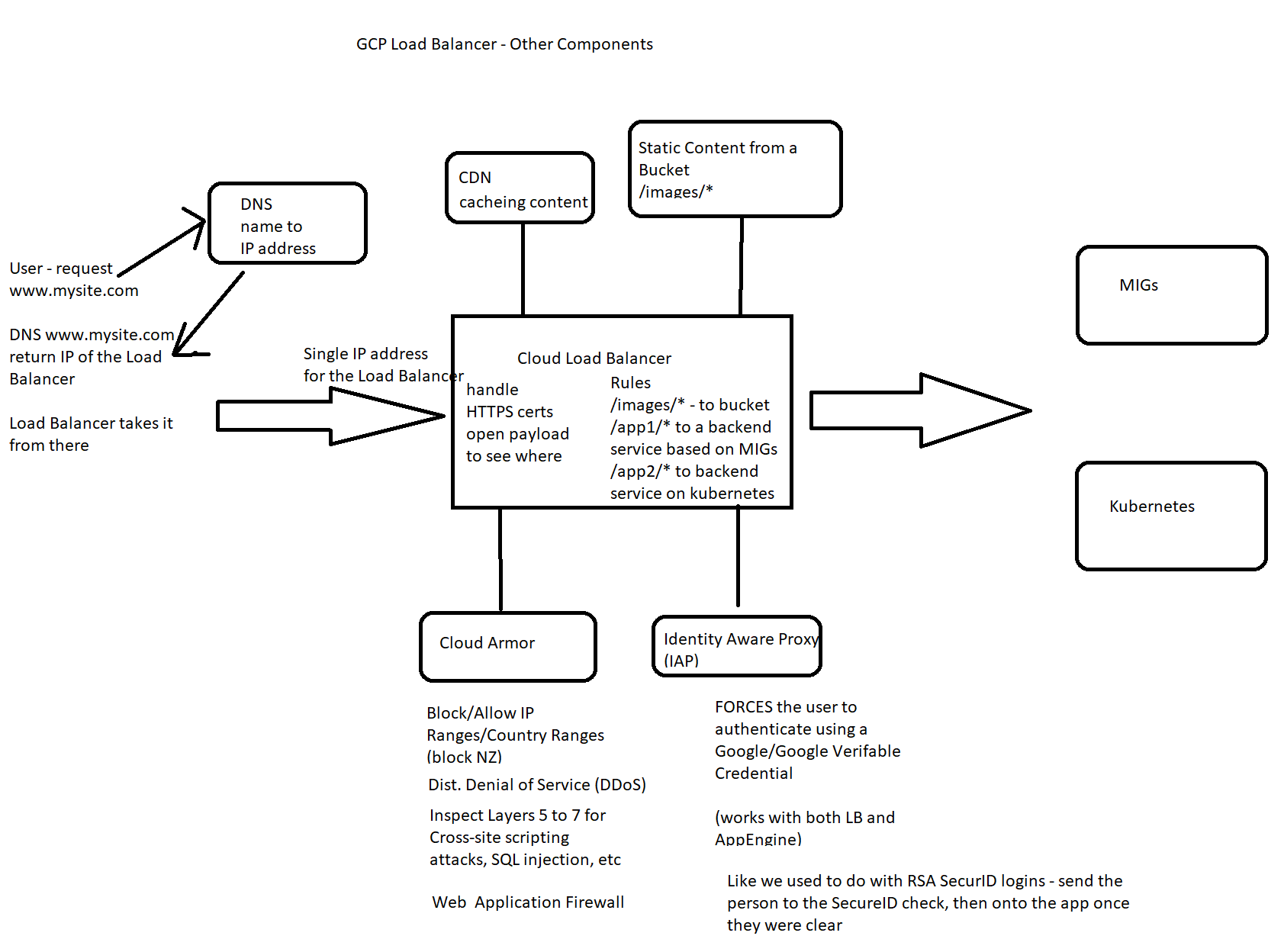
IoT Core is a Managed Service for Interfacing the devices to Pub/Sub and then into BigTable or BigQuery, Processing, etc

If I want to see EVERY sensor reading for a specific sensor - data in BigTable makes sense

(30,000 rows per seconds read or write w/3 Nodes)

If I want to AVERAGE ALL sensor data - I need it in BigQueryu

SOMETIMES - I put it in BOTH BigQuery (for Analystics) and BigTable (individual/range of row retrieval)



DRESS FOR WIN:

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.

Day 1 - what can you give them on Google

Show them a website “we’re sorry - our website is down”

Static Content in a Bucket. Change DNS record to point to GCP (http)

Static Content in a Bucket, Google Load Balancer (for the HTTPS), change DNS record to point to GCP Load Balancer IP

They are not sure which components of their architecture they can migrate as is and which components they need to change before migrating them.

What can they Lift and Shift

What do they need to Replatform

All servers run Ubuntu LTS v16.04 -- so they are using VMs or Raw Hardware on-premise

Lift and Shift - VM on GCP (GCE)

* 40 web application servers providing micro-services based APIs and static content
  + Tomcat - Java
  + Nginx
  + Four core CPUs
  + 32 GB of RAM

Suggestions for Lift and Shift

Tomcat/Java - VM running LTS 16.04 + Tomcat

Suggestions for Re-platforming

Java - AppEngine Standard (tomcat just hosts Java apps)

Containerize (docker) Tomcat/Java (dockerhub has Tomcat) -> AppEngine Flex or Kubernetes

Nginx - What is Nginx? Web Server, Load Balancer, API Gateway, Proxy, Cache, etc (opensource)

Suggestions for Lift and Shift?

VM on GCE - running current Stack (Nginx) (one big VM?)

Introduce: MIGs, Autoscalers and Load Balancer

Why serve static content from CPUs - buckets don’t consume CPUs - begin replatform

Re-platform:

Web Server static content? Global Load Balancer + Buckets + CDN

Load Balancing - Global Load Balancer

Caching - Bucket + CDN + Global Load Balancer

Proxy - Global Load Balancer

SSL/TLS Offload (HTTPS->HTTP) - Global Load Balancer (Managed Certificates)

Cloud EndPoints is API Gateway based on Nginx

HSBC-specific: Mulesoft = API-Gateway-Enterprise-Edition (APIGEE)

Cloud EndPoints is a simpler product, not as powerful - but still an API Gateway based on NGINX

* 20 Apache Hadoop/Spark servers:
  + Data analysis
  + Real-time trending calculations
  + Eight core CPUs
  + 128 GB of RAM
  + 4x 5 TB HDD (RAID 1)

Pattern: Lift and Shift Hadoop to Data Proc (it will have Hadoop/Spark jobs + HDFS storage)

THEN - move the HDFS to GCS - replace hdfs://folder/files references with gs://bucket/folder/files

Now cluster can be created/destroyed as needed, to the scale needed

Three RabbitMQ servers for messaging, social notifications, and events:

Messaging - IBM MQ, Kafka, MS MQ, etc

Replatform to Pub/Sub (code changes - so it’s not a lift and shift) - they are NOT going to ask you about lift and shift (or if they did - run them in a VM)

Pub/Sub is Google’s messaging infrastructure

* Miscellaneous servers:
  + Jenkins - Continuous Integration (CI)
    - L&S: VM or Kubernetes Cluster running Jenkins
      * Pre-build in the marketplace
      * Jenkins is Good for OnPremise AND GCP
    - Continuous Deployment (CD) can happen on Jenkins
      * In the world of Containers - Spinnaker is for CD

iSCSI for VM hosts, Fibre channel SAN = ON GOOGLE - this is persistent Disk

SAN = Storage Area Network - Block-level - blocks of data stored onto a disk infrastructure = Persistent Disks attached to VMs

NAS = Network Attached Storage - File-level - store/retrieve files: GCP = Cloud Storage/Buckets

LEARN TWO CI/CD Pipelines on GCP

Jenkins + Spinnaker (Open-Source, run on VMs or Containers)

Source Repositories->Cloud Build (CI/CD)->Container Registry(GCR) (Google-Native Managed Services)(all services in GCP Project)

Cloud Build can build from

Dockerfile (build containers - and stores into GCR)

YAML file with instructions on what to do (build java, container, run tests, push to QA, other operations)

Cloud Build could build a container for Cloud Run (Cloud Functions can read code from Source Repository, Cloud Build needs to pull the code from GCR)

* Monitoring - Operations (formerly stackdriver)
  + Logging/Monitoring agents have to be installed - collect more information
  + By default - google does NOT know what the insides of your VMs are doing
    - They can tell disk IO, network IO, CPU usage
    - They don’t know know what programs
* bastion hosts
  + Jump Stations, etc
  + Gcloud commands to talk to GCP, ssh into VMs, etc
* security scanners
  + If I plan on running a nessus scan against my GCP resources - do I have to tell google?
  + NO - I do NOT need to tell google - <https://support.google.com/cloud/answer/6262505>

Security Command Center (SCC) - an ORGANIZATIONAL LEVEL scanning

Looks at projects and their resources

Inventory, Vulnerabilities, alerts, etc

Recommendations

Web Security scanner (WSS) - scans an application for vulnerabilities (XSS, SQL Injection, known vulnerabilities, etc

Forseti Security (Inventory System/Policy Checker)

Business Requirements

Build a reliable and reproducible environment with scaled parity of production

CloudSQL to run a Dev copy of the DB, Build VMs, scale up/down for simulations, etc

* Improve security by defining and adhering to a set of security and identity and access management (IAM) best practices for cloud
  + Users -> Groups -> GCP Roles (GCP Roles give permissions)
* Improve business agility and speed of innovation through rapid provisioning of new resources
  + Infrastructure Automation (building VMs, Building networks/firewall rules, etc) = Terraform or Deployment Manager
  + CI/CD Pipelines for CODE/Containers (NOT INFRASTRUCTURE)- Jenkins/Spinnaker and Source Repos->Cloud Build->GCR
* Analyze and optimize architecture for performance in the cloud
  + This very generic - no specific answer - but a question could be asked - leading towards replatforming (recommendations for replatforming)

### Technical requirements

* Easily create non-production environments in the cloud

VMs, CloudSQL, simulating onpremise network with a VPC

* Implement an **automation** framework for provisioning resources in cloud
  + Terraform or Deployment Manager
* Implement a continuous deployment process for deploying applications to the on-premises data center or cloud
  + Jenkins/Spinnaker (onpremise and cloud) over Sources Repos->Cloud Build->GCR (less on-premise)
  + Kubernetes runs OnPremise, OnCloud and Other Cloud
* Support failover of the production environment to cloud during an emergency
  + Static Website: DNS->Bucket (http), DNS->Load Balancer(https)->Bucket
  + Dynamic Content: DNS->Load Balancer->Do you have the APP migrated to Google yet? (this stage is just dev and test) - it may be more “Our website is down”, this application is un-available, etc
* Encrypt data on the wire and at rest
  + Everything on wire is encrypted by default
  + Everything at rest is encrypted by default (google managed Encryption)
    - Customer Managed Encryption Keys (CMEK) - google has keys in KMS (think a vault) - you control key rotations, etc
    - Customer Supplied Encryption Keys (CSEK) - customer has keys
* Support multiple private connections between the production data center and cloud environment.

What command lines do I need to be an ~~expert~~ have basic knowledge on for GCP Exam

Gcloud - command line management of Google

Gsutil - everything BUCKETS

Bq - everything BIGQUERY

Which of the following commands will create a bucket is:

gcloud mb gs://bucketname

gcloud gs://bucketname mk

gcloud mb gs://bucketname --multiregional

gsutil mb gs://bucketname

I need to **create** a kubernetes cluster on gcp

kubectl

**Gcloud - I need the cluster before I can kubectl**

I need to deploy an application on a kubernetes cluster

**kubectl**

Gcloud

I need to add a new node pool to a kubernetes cluster (node pool is a group of VMs)

Kubectl

gcloud

I need to CREATE a bigtable cluster

gcloud

cbt (which does exist - cloud big table)

I need to build a new table on a Bigtable cluster

Gcloud

cbt

Where did Hadoop/Big Data Come From:

Late 90/early 2000’s - Google needs a place to store large amounts of data (copies of websites) and way to process it in parallel

90’s/2000’s - you share concepts - Scientists and Engineers like to write papers

GFS (distributed filesystem) -> Yahoo reads the paper -> Yahoo build HDFS (Hadoop Distributed FS)

MapReduce (programming model) -> Yahoo paper -> Yahoo builds Hadoop (their implementation of map-reduce)

BigTable (distributed database) > Yahoo reads paper -> Yahoo build HBase (Hadoop Database)

MapReduce -> BigQuery (writing MapReduce as SQL commands) and Dataflow/Beam - JAva and python library for doing MapReduce (parallel) operations

Why would google OFFER HADOOP on GCP - when they have their own tools ?

Lots of customers run Hadoop - and they need to bring the workload (but more importantly - the DATA) to GCP - so that it can be processed on Google Native Tools in the future

Offer DataProc - A Managed Hadoop infrastructure so customer can bring their Hadoop workloads

Hadoop / HDFS / HBase ->

Hadoop/HDFS -> lift and shift to DataProc

Once in DataProc - you can begin moving the HDFS (storage) into GCS

Once it’s in GCS - you can continue to run your Hadoop/Spark/Pig/Hive but against the data stored in GCS

NEW DEVELOPMENT:

Import the GCS data into BigQuery

Process the GCS data with Dataflow / Apache Beam (Parallelization for Python and Java)

ORG->Folder->Project Hierarchy

Permissions are INHERITED

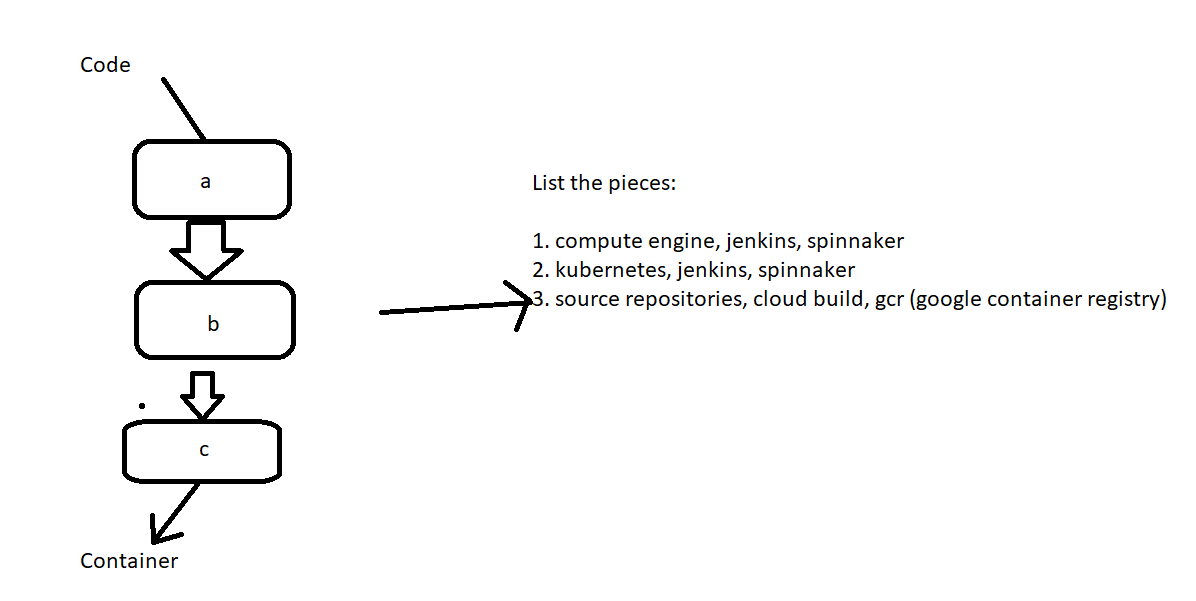
Roles grant Permissions

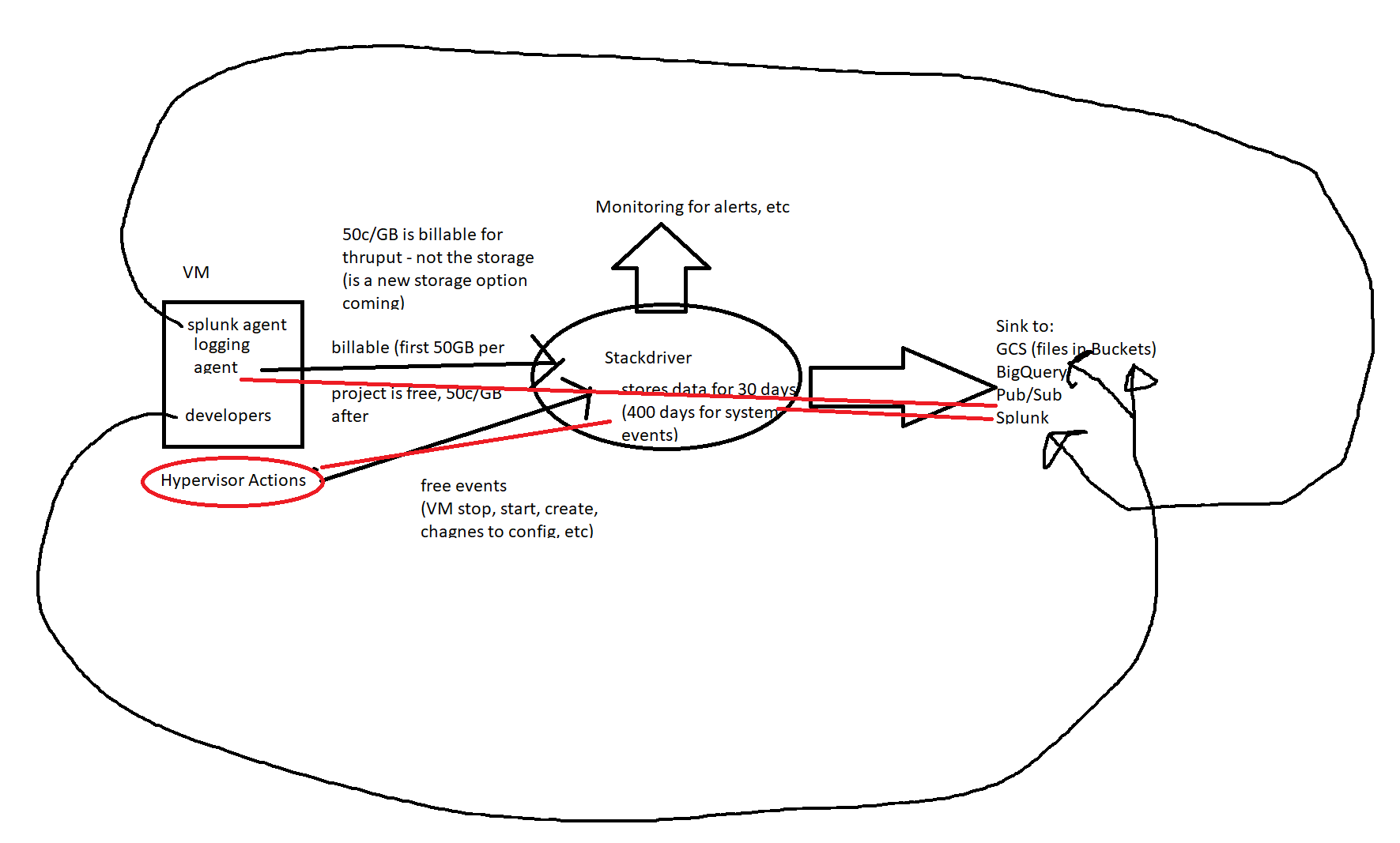
(you don’t directly give permissions - you grant users/groups roles)

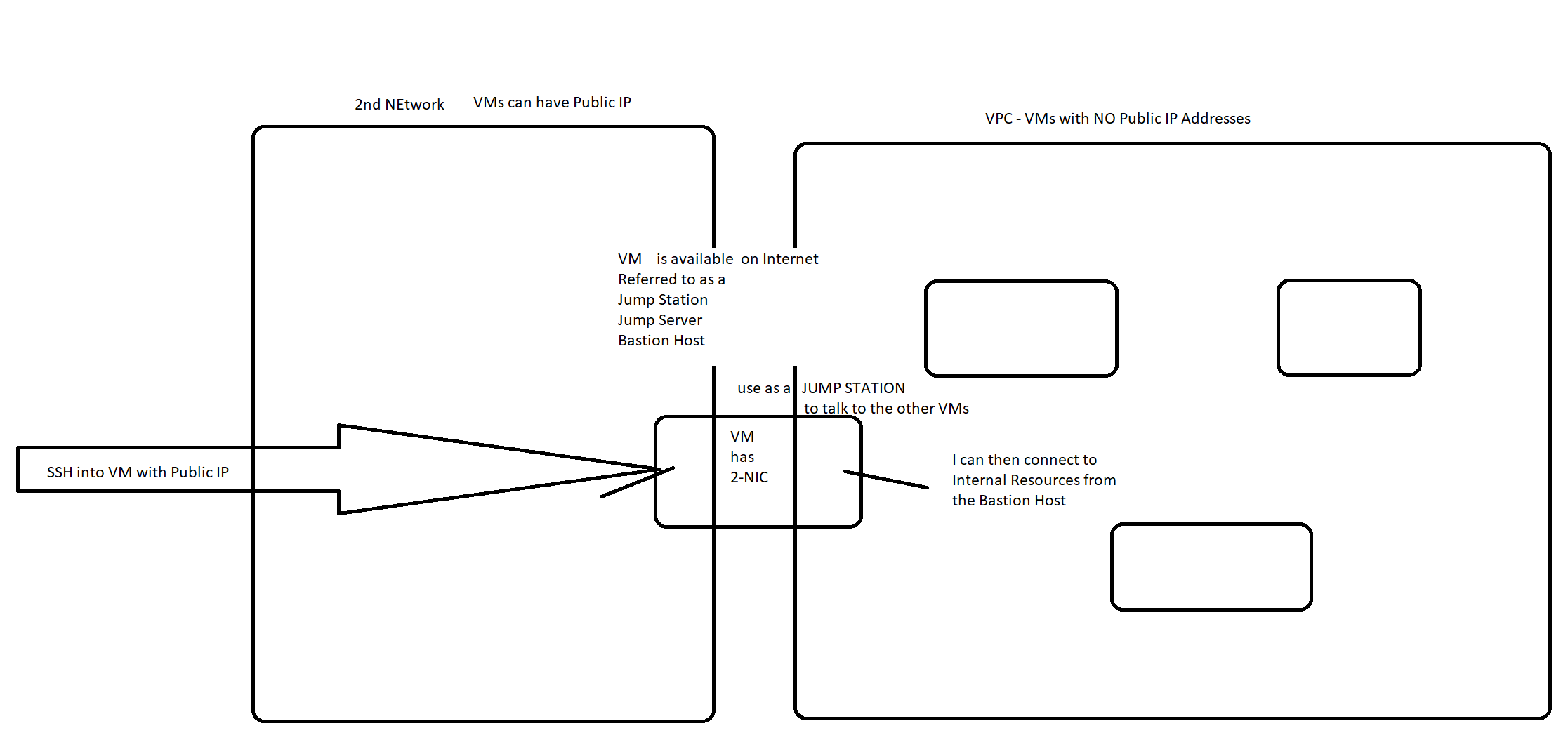
Primitives - Owner, Editor, Viewer

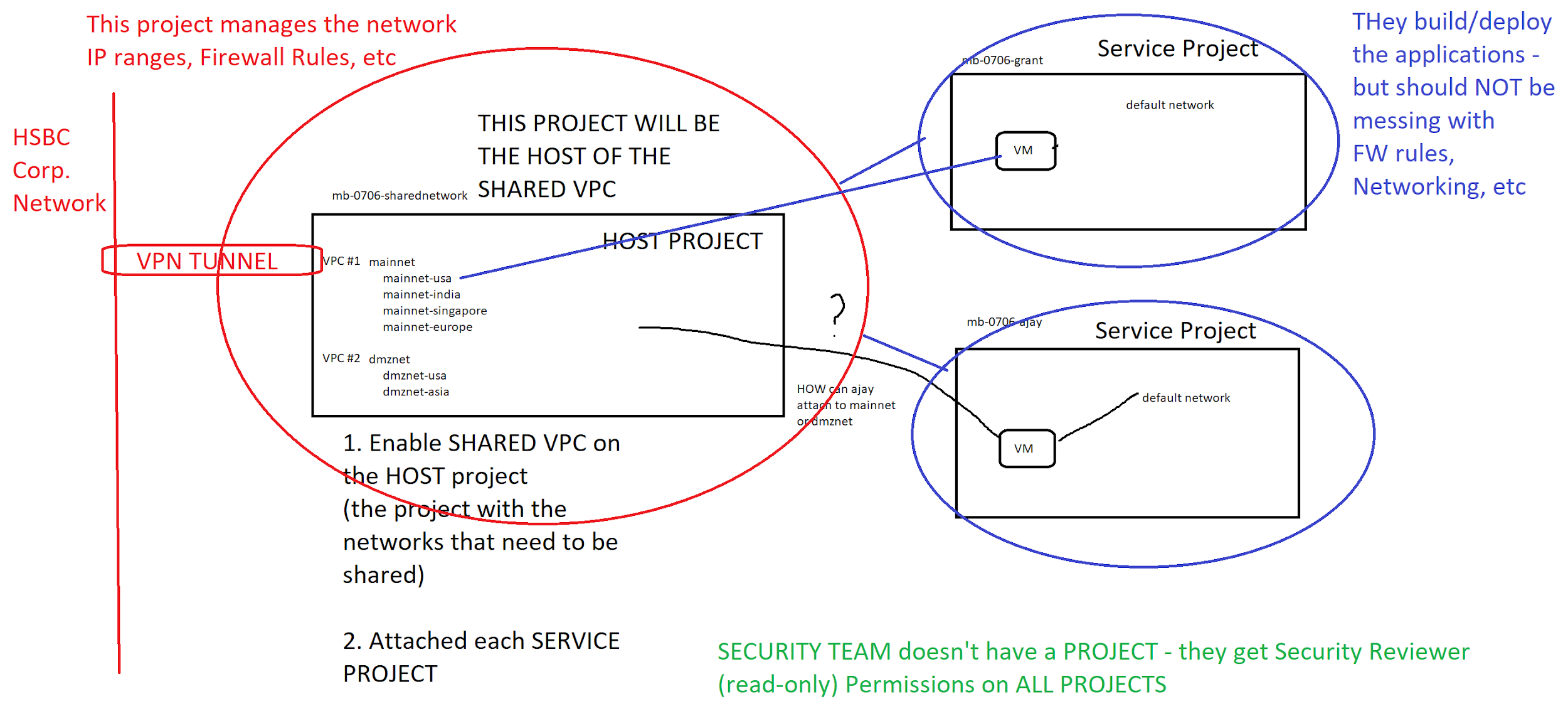
Pre-defined Roles - over 500 pre-defined roles in Google

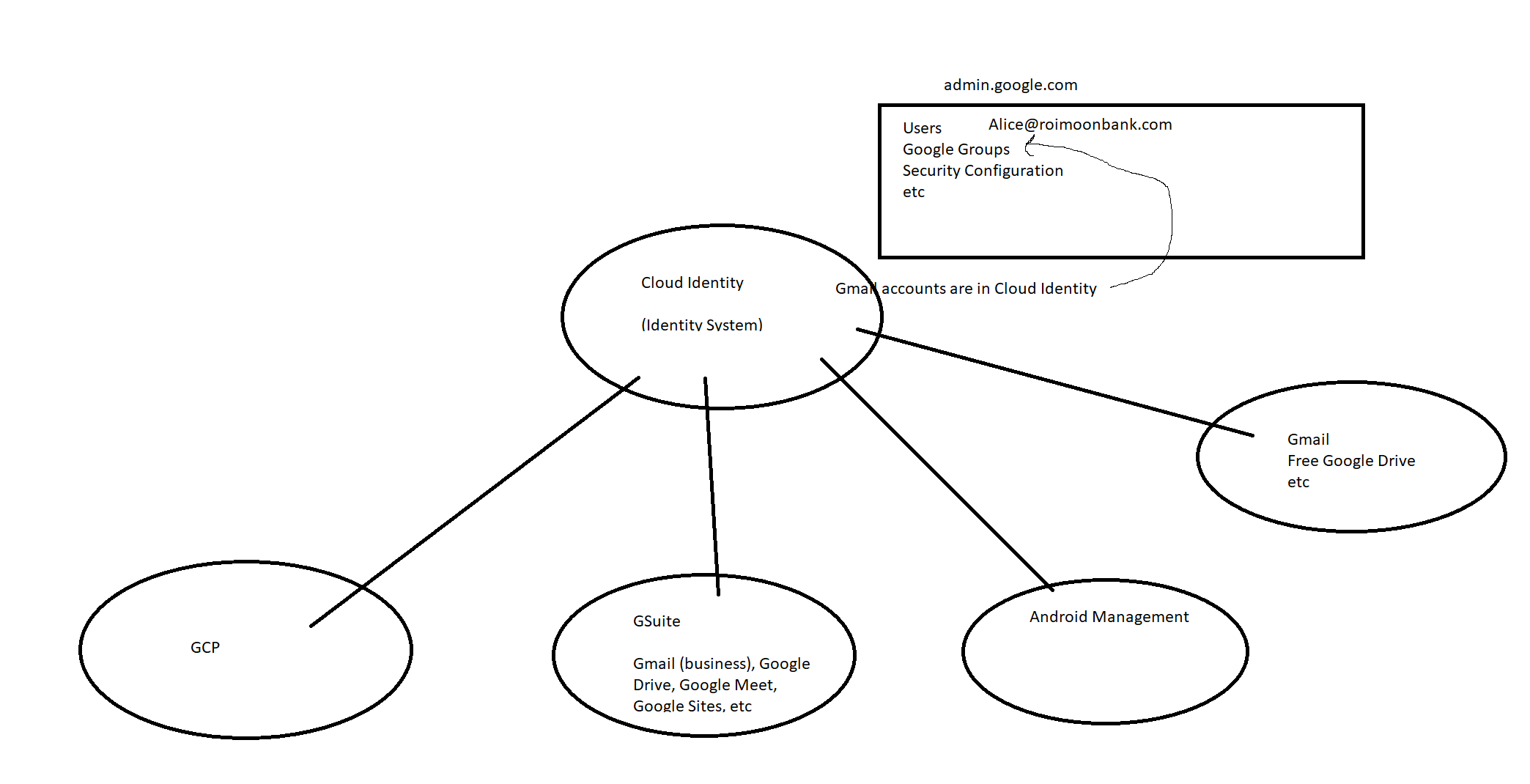
Custom Roles - you create the Role and assign it permissions

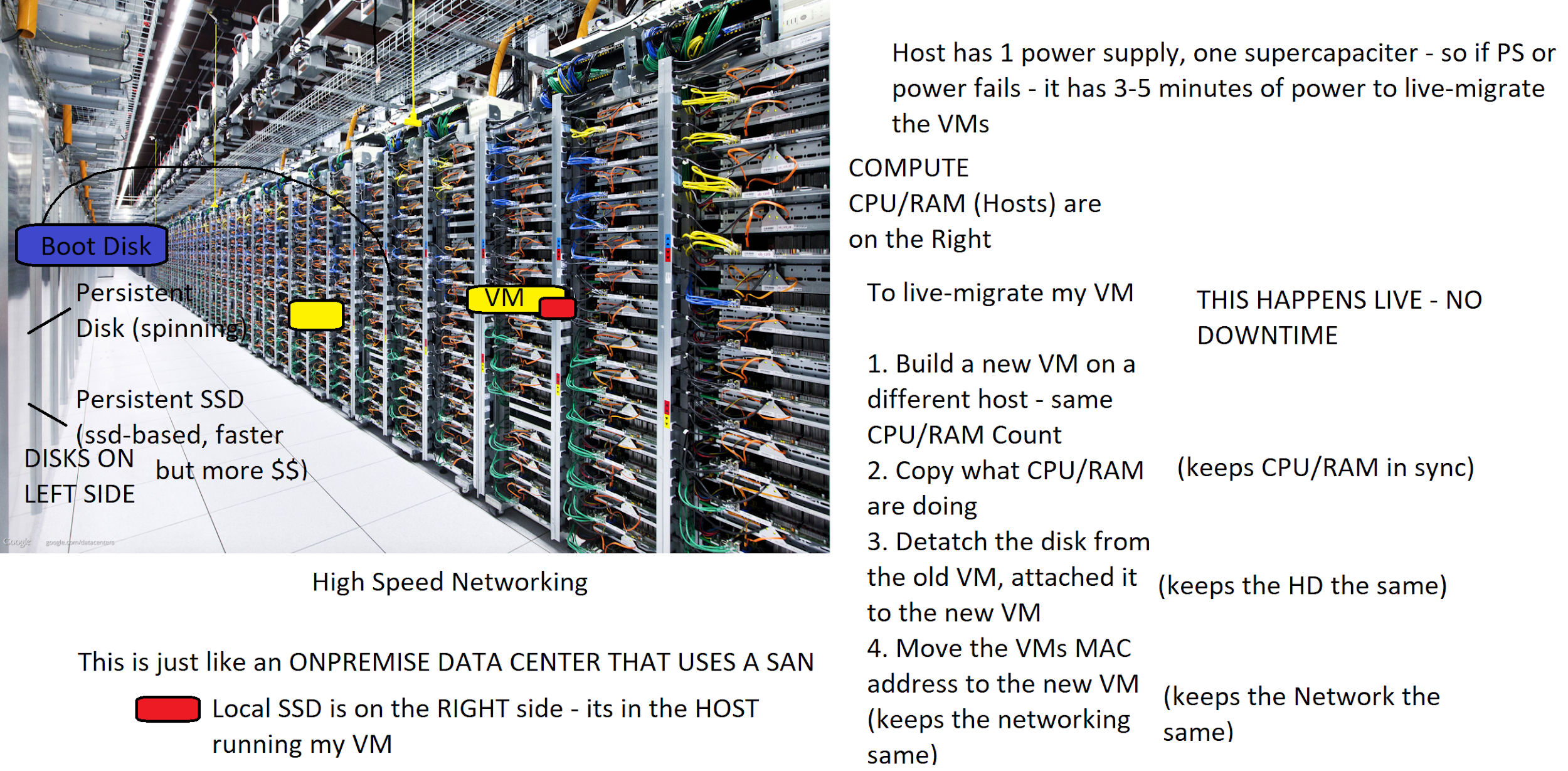












How do I managed things on GCP

Cloud Web Console

Google Cloud SDK (Software Development Kit)

If you install the SDK - you get gcloud, gsutil, bq and other tools can be added

Google Cloud Shell (which is a small VM (technically a container) that has the SDK already installed)

AT HSBC - which tools will you use for managed GCP?

Web Console - NOT on production environments

Cloud Shell - NOPE - too risky - the tool has security concerns by HSBC (grant agrees)

So - as you use Google at HSBC (Hoogle)

Installed the Google SDK onto your laptop (so you have gcloud, gsutil, etc)

Maybe some access to the Web Console

**Automation Deployment/Infrastructure as code**

**(not Code-deployment - RESOURCE deployment)**

Automation of Project->Network->Subnets->Firewall Rules->Managed Instance Groups, Templates, Load Balancers, VMs, Kubernetes Clusters, etc -

Cloud Agnostic (any cloud) - Terraform

Google-specific - Deployment Manager

Here is the deployment manager dilemma

1. You are NOT going to use it at HSBC
2. It will take about 3-4 hours to fully understand Deployment Manager
3. There is likely to be 1 to 2 questions on Deployment manager

So - HOW SHOULD YOU PREPARE FOR THOSE 2 QUESTIONS?

Spend 4 hour learning a product you will never user

Spend those 4 hours on other topics - and try your best on the Deployment Manager Questions

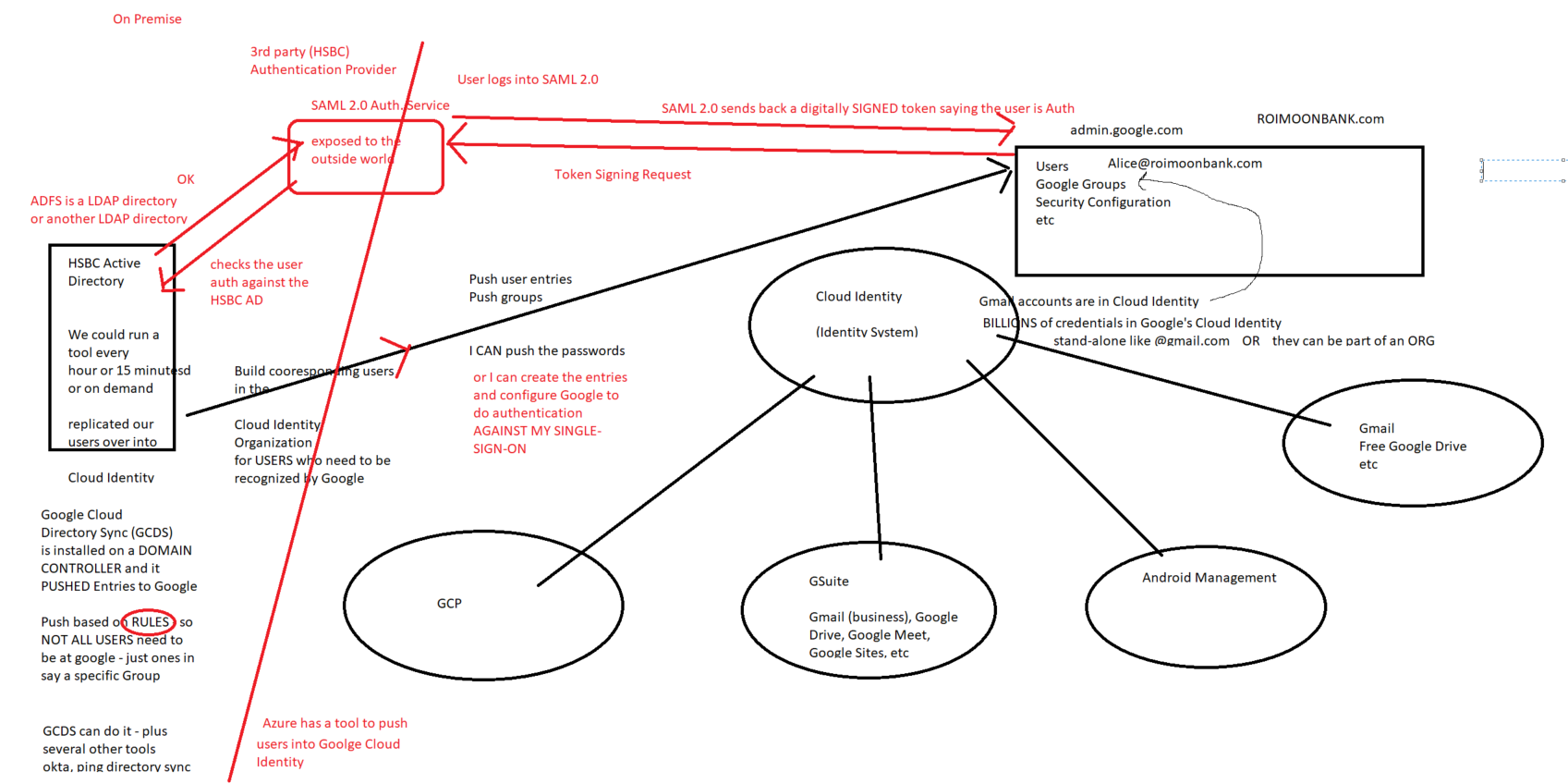
Questions MIGHT Show you the YAML file used for deployment manager - and say something is wrong:

Expect you troubleshoot what the deployment manager issue is

Building Repeatable Deployments: VMs, Networking, Firewall - Deployment Manager/Terraform

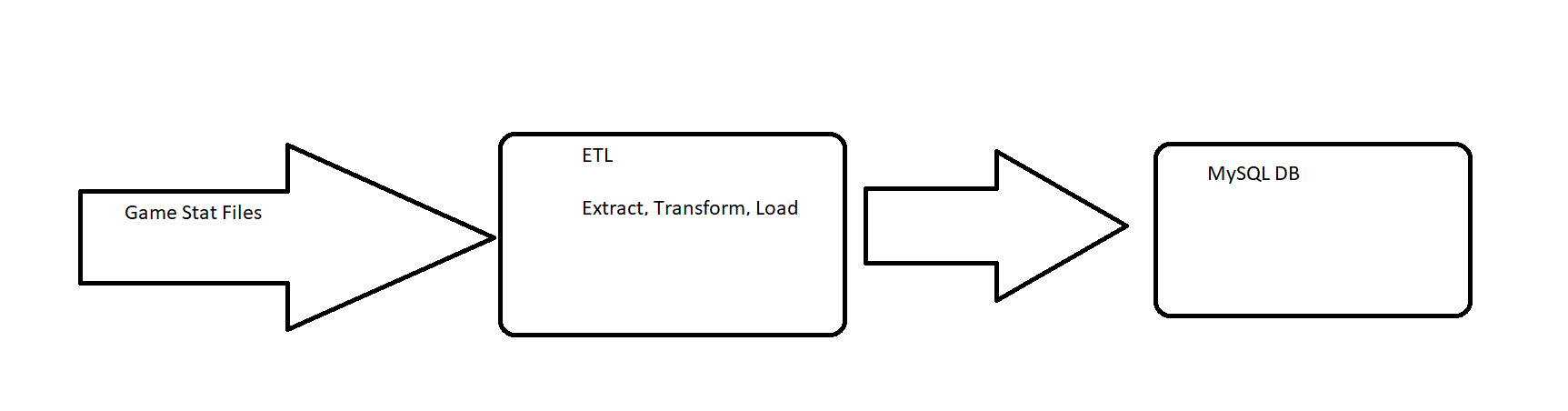
Repeatable deployment of a Cloud Function?

Python, node, etc might be the code-base - but Deployment Manager or Terraform are going to be used to DEPLOY the function



**MountKirk Games Breakdown**

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.



Mountkirk Games is building a **new game**, which they expect to be very popular. They plan to deploy the game’s backend on **Compute Engine** so they can capture streaming metrics, run **intensive analytics**, and take advantage of its **autoscaling server environment** and integrate with a **managed NoSQL database**.

Criteria: VMs (Compute engine)

Capture Streaming Metrics

Intensive Analytics

Take advantage of autoscaling server environment

Managed NoSQL Database

* Increase to a global footprint

Run in multiple regions, use load balancing, CDN

* Improve uptime—downtime is loss of players

**Highly Available + Disaster Recovery**

How do you get High Availability in GCP: multiple zones

HA SQL - ONE REGION, TWO ZONES

Google can lose a Zone - and your app should keep going

Multiple Multi-zone Managed Instance Groups + Load Balancer sending load to different MIGs

How do you get Disaster Recovery in GCP: multiple regions

If google has a REGION destroyed (Asteroid, Volcano, earthquake, flooding, natural disaster, terror attack, riots, aliens- Tokyo: Godzilla

* Increase efficiency of the cloud resources we use

Use Managed Services or Services products where possible

* Reduce latency to all customers

PUT THE FRONT-END SERVERS closer to the users (multi-regional Load Balancer, targeting multi-zone Managed Instance Groups and CDN (cache static content close to users)

### Technical requirements

* Requirements for game backend platform
  1. Dynamically scale up or down based on game activity

COMPUTE ENGINE (already told us this) - MIGs / Autoscaler

* 1. Connect to a transactional database service to manage user profiles and game state

It does **NOT** say “Relational” or “Highly Relational”, does NOT say “ANSI SQL”

They said “ integrate with a **managed NoSQL database.**”

Firestore/Datastore is OK

* 1. Store game activity in a timeseries database service for future analysis

BigTable - if I need single-row and range of row retrieval

If they want to know all moves a specific player made BT

BigQuery - if I need aggregation / data analys

If they want to know the average number of moves to complete a level, etc across all players - BigQuery

* 1. As the system scales, ensure that data is not lost due to processing backlogs

Use Pub/Sub to buffer the data that needs to be written - in case hte back-end is overloaded - messages in Pub/Sub can be processed and then removed from Pub/Sub upon process completion

* 1. Run hardened Linux distro

Compute Engine (they want full control over the OS)

Shielded VMs

What is a VM? Hard Disk + Configuration File

Where are secrets/keys/etc - on the Hard Disk or in the Configuration

Where are secrets/keys held on a corporate laptop/deskops - held in a CHIP on the Motherboard called a TPM - Trusted Platform Module v2

Requirements for **game analytics platform (Data Warehouse, Analytics is BIgQuery in GCP world)**

Dynamically scale up or down based on game activity (analytics - not LB/MIG)

Means use Cloud Dataflow (autoscaling cluster)

Process incoming data on the fly directly from the game servers

Streaming Pipeline: pub/sub- > Dataflow->Bigquery

Process data that arrives late because of slow mobile networks

This is a capability of Dataflow / Apache Beam

Allow queries to access at least 10 TB of historical data

Free-bee answer (the 10TB used to be a hint not to use CloudSQL prior to summer of 2019 maximum CloudSQL was 10TB. It’s now 30TB - but still - it’s a hint don’t use cloud SQL)

THIS IS AN ANALYTIC DATABASE - NOT A TRANSACTION

No CloudSQL, No Spanner (these are OLTP) Transaction

BigQuery (OLAP) Analytic

Process files that are regularly uploaded by users’ mobile devices

Batch Pipeline: gcs->Dataflow->BigQuery

**Databases on GCP**

Transaction, Highly Relational, ANSI SQL Compatible Databases

CloudSQL or Spanner or run a

CloudSQL - MySQL, PostgreSQL or Microsoft SQL 2017

Google for GB to 30TB databases

Off-shelf-shelf - google does NOT change the MySQL, PostgreSQL or MS SQL code

They just manage the server, OS, and DB software

You build the Databases/Tables/etc

Needs a 1-hour week Maintenance window

To change configuration needs to shutdown and startup (underlying VM changes)

1 Region, 1 Zone or 1 Region, 2 Zone (high-availability - if zone fails, google with automatically fail over to 2nd node highavailiability)

Additional READ-replica can be created, but to fail over to these - you have to manually do the failover

Spanner - Google’s in-house built ANSI SQL Database

Good for TB to PB - $600+ a month to run

Understands SQL (INSERT, UPDATE, SELECT, DELETE, Drop table, truncate, create table,etc)

Scales horizontally - adding/removing spanner nodes will increase/decrease performance - you have to do it (NOT autoscaling by google)

No maintenance window, no downtime, 99.99 (regionla) / 99.999 (multi-regional)

No downtime to add/remove nodes - you are just adding/removing computing power

NoSQL on Google: Key-Value, Document DB, Wide-Column DB

Key-Value: Memorystore (Managed Redis) or Memcache (built-into AppEngine)

Think of the Windows Registry for Key-Value example

Does NOT use SQL queries - store Keys w/Value in them - read back

Document DB: Complex Key-value - a row is a block of JSON data - which a bunch of key:values stored within - including repetitive fields, etc

Read/Write the row

` Does NOT use SQL queries

Each ROW can be different layout than other rows (JSON key:value pairs) - “each row has it’s own SCHEMA

Example: MongoDB is very popular Document DB.

Firestore/Datastore is a Document DB on GCP - pay per million reads/write

No provisioning of scale or disk needed - scales to infinity and beyond!

GB/TB and even PB

Why would DocumentDB (Firestore)be cheaper than CloudSQL:

Not necessarily.

Cloud SQL - you have to define the server - so it doesn’t automatically scale

Cloud SQL n1-standard-1-db 1 CPU/3.75GB of RAM - that’ sis your computing power

Firestore - you pay per million reads/writes

0.60 per million reads

1.80 per million writes

Plus the amount of disk space your DB takes up

Wide Column - Lots of columns - but ONLY ONE Key native in the database (the row ordering)

Very high read/write throughput

No indexes (updates don’t need to update several key/indexes)

Built for single-row and range-of-row retrieval

What is Dataflow and how does it relate to Apache Beam

Google needed a way for their developers to write MapReduce Jobs

(remember - google published Google File System Paper, Google MapReduce Paper and BigTable Paper)

GFS-> yahoo created Hadoop Distributed File System (HDFS based on the paper

MapReduce -> yahoo create Hadoop based on the paper

BigTable -> yahoo created Hadoop Database (HBase) based on the paper

How do you PROGRAM MapReduce on Google though?

They wrote an object class for Java called “BEAM”

Beam handles parallelization (parallel loops called pipelines)

Also implemented for Python

Create a pipeline (based on files in a bucket, rows in a query, messages in a queue)

Task1 on object

Task2 on object

…

Aggregation

End pipeline

Available both Java and Python

Google handed BEAM over to Apache Foundation - so it would be more widely adopted

WRite locally - these parallelizable loops

Submit a Python or Java code to google using BEAM - google will create a parallel cluster

Encryption on GCP

VM - A VM’s Hard Disk can be encrypted one of three ways (per disk configuration)

Google Managed Encryption Keys ~~(GMEK)~~

Google Manages ALL ASPECTS of the encryption

No visibility - you trust the documentation

No extra charges

Customer Managed Encryption (CMEK)

Google has a Key MAnagement Service (KMS)

You Create/Manage the Key - but GOOGLE STORES/USES the keys on your behalf

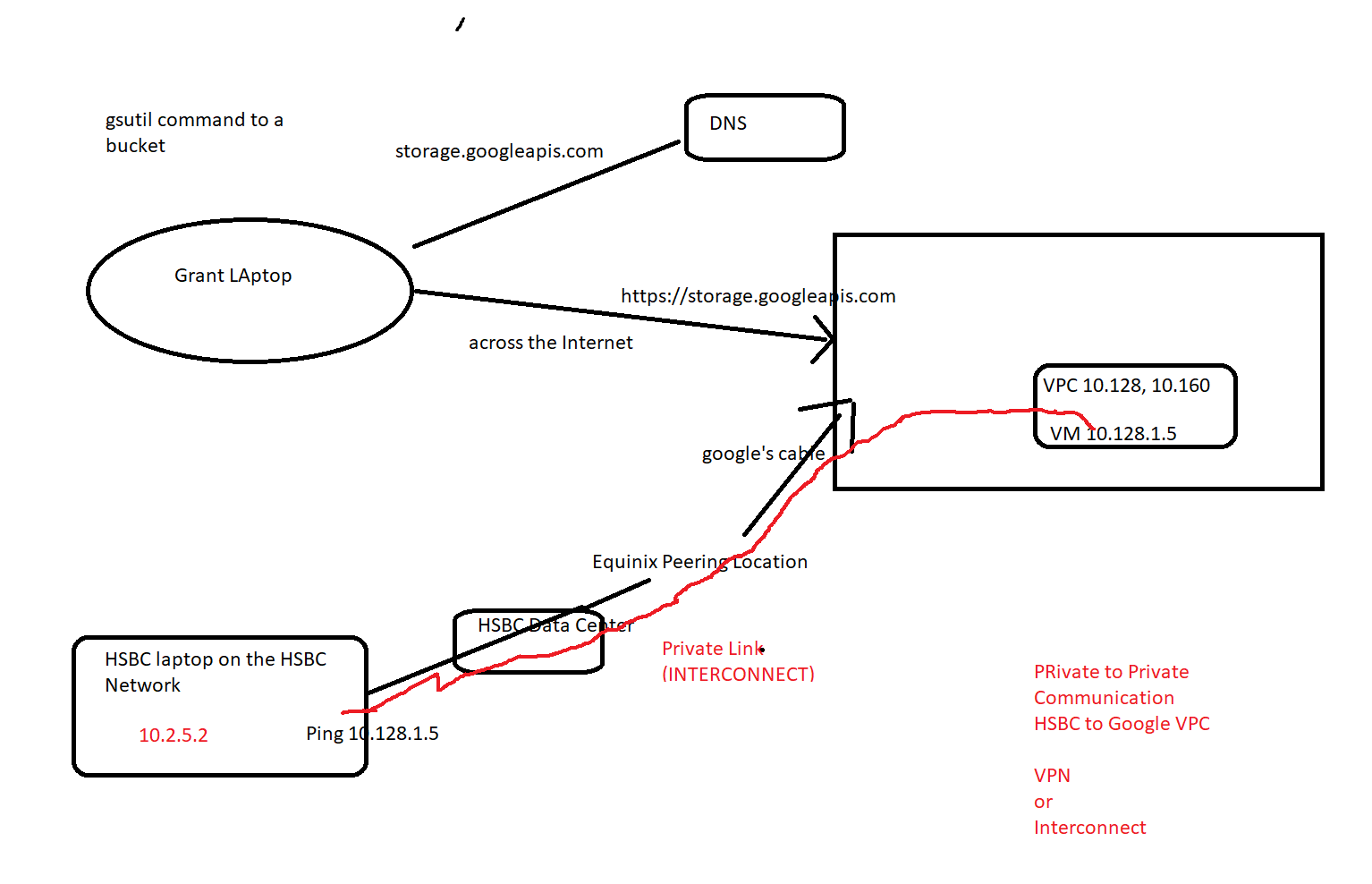
You control rotation, key destruction, and you get auditability (makes the auditors happy)

Pay per key + per 10,000 key uses

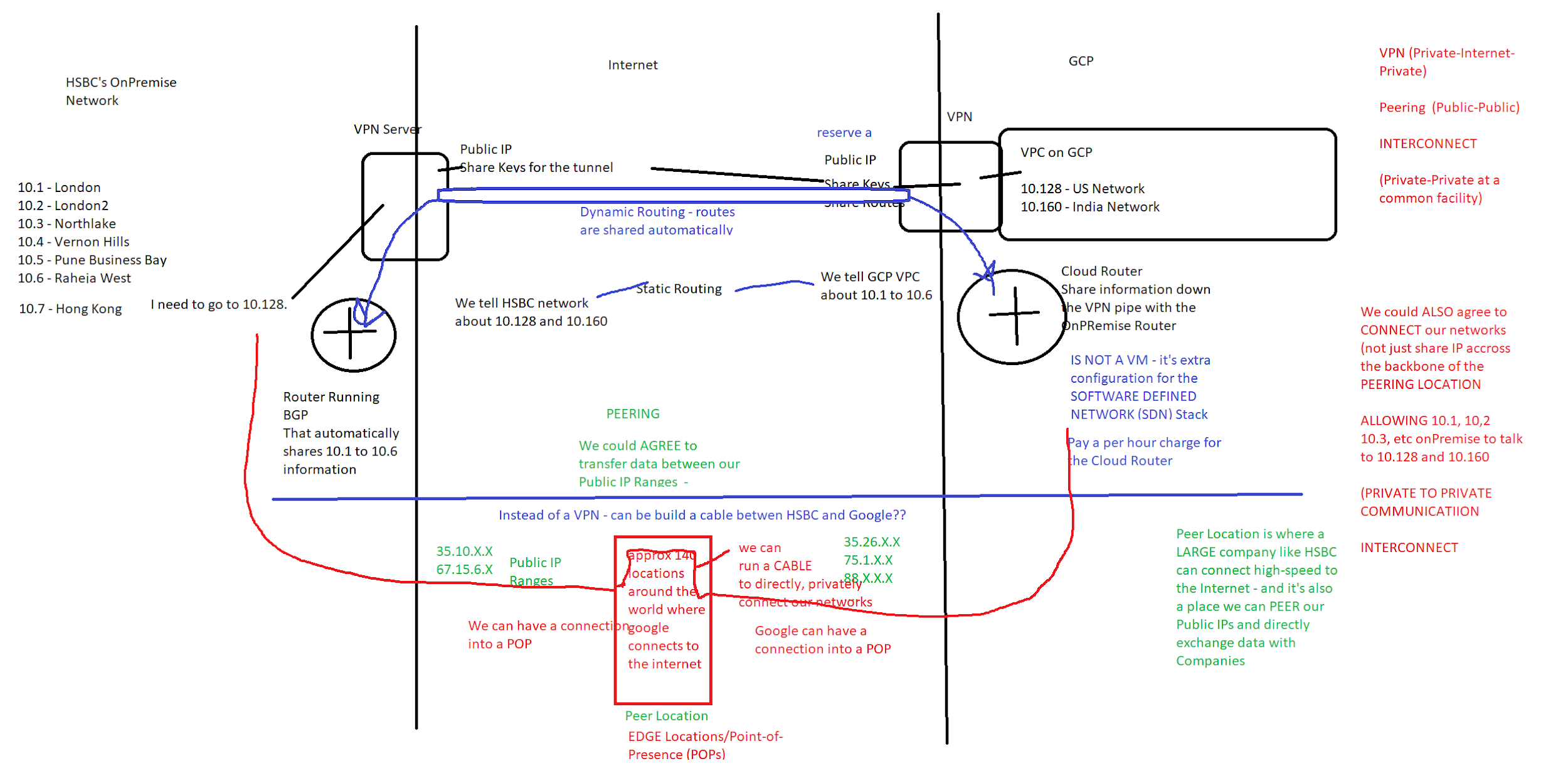
Customer Supplied Encryption Keys (CSEK)

YOU SUPPLY the encryption key at time of use

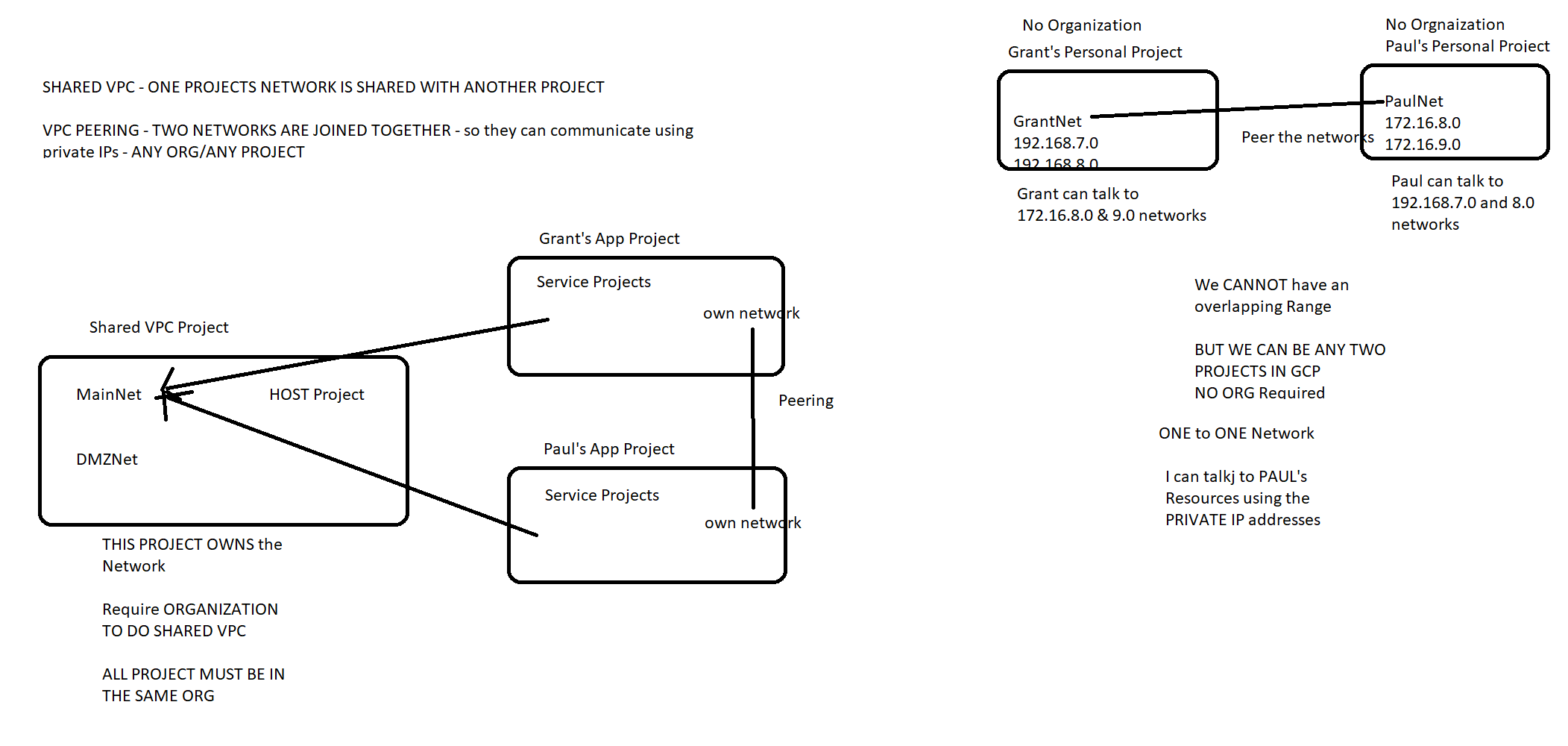
To start a VM - you provide the key



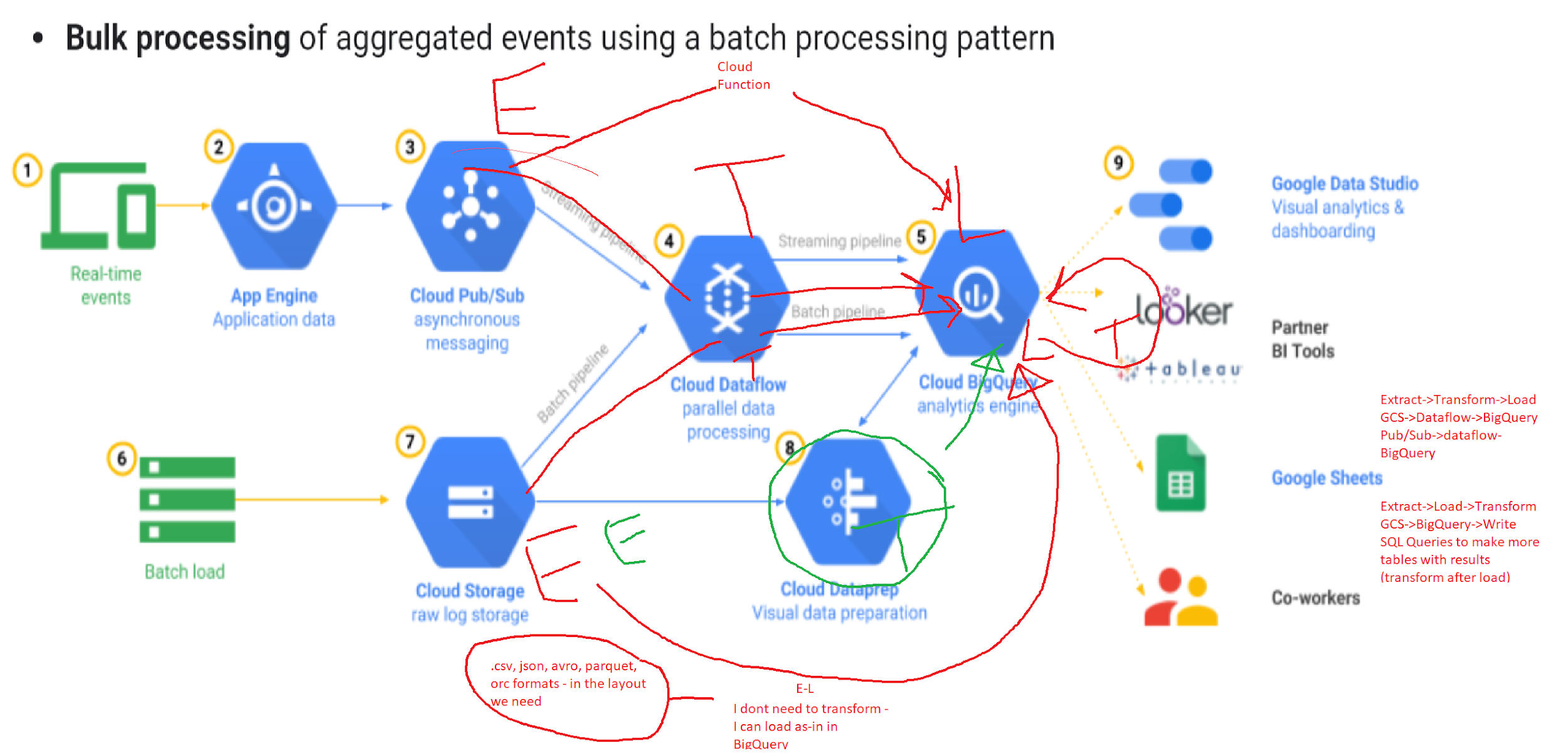
VPN and Interconnect



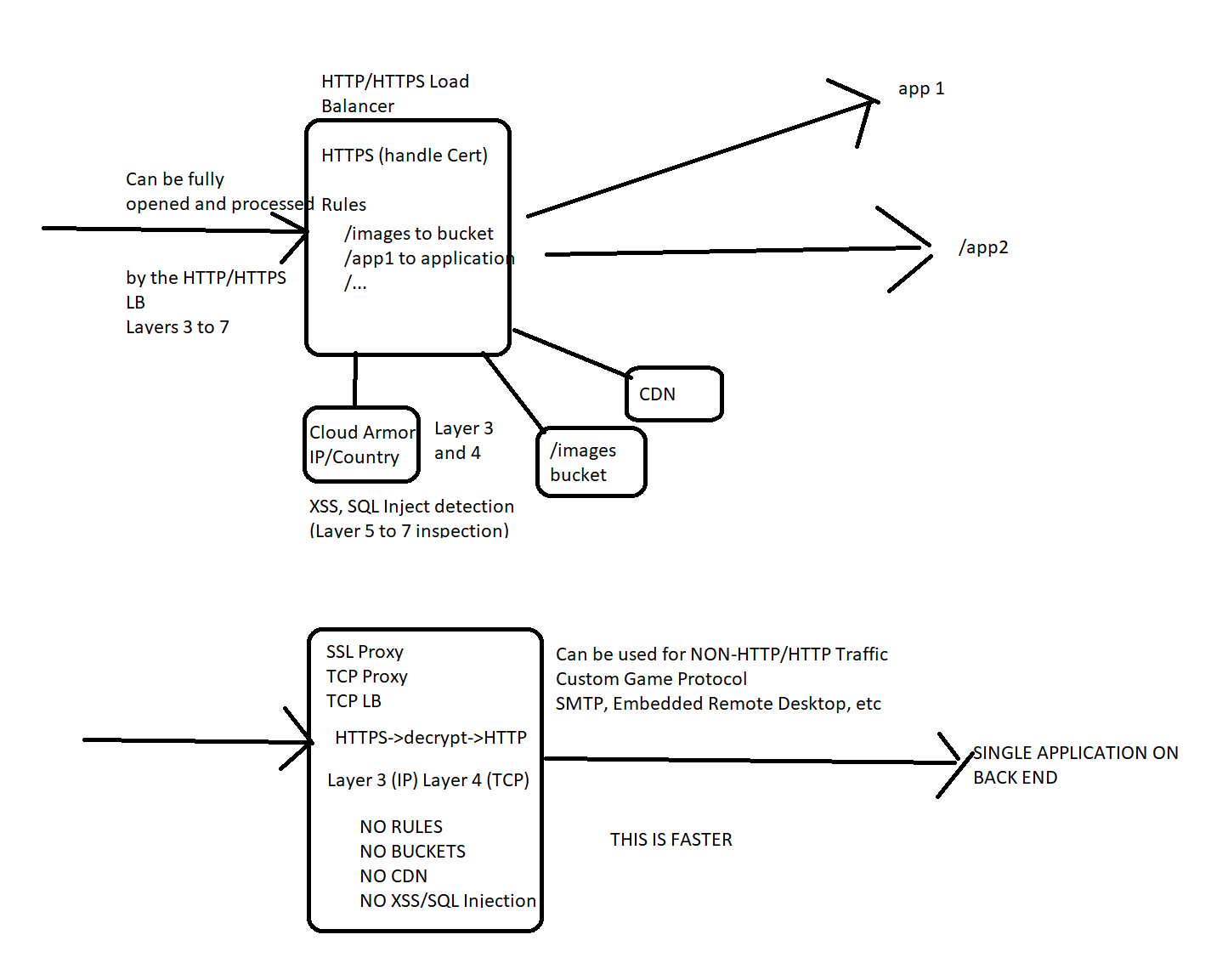
Shared VPC vs VPC Peering



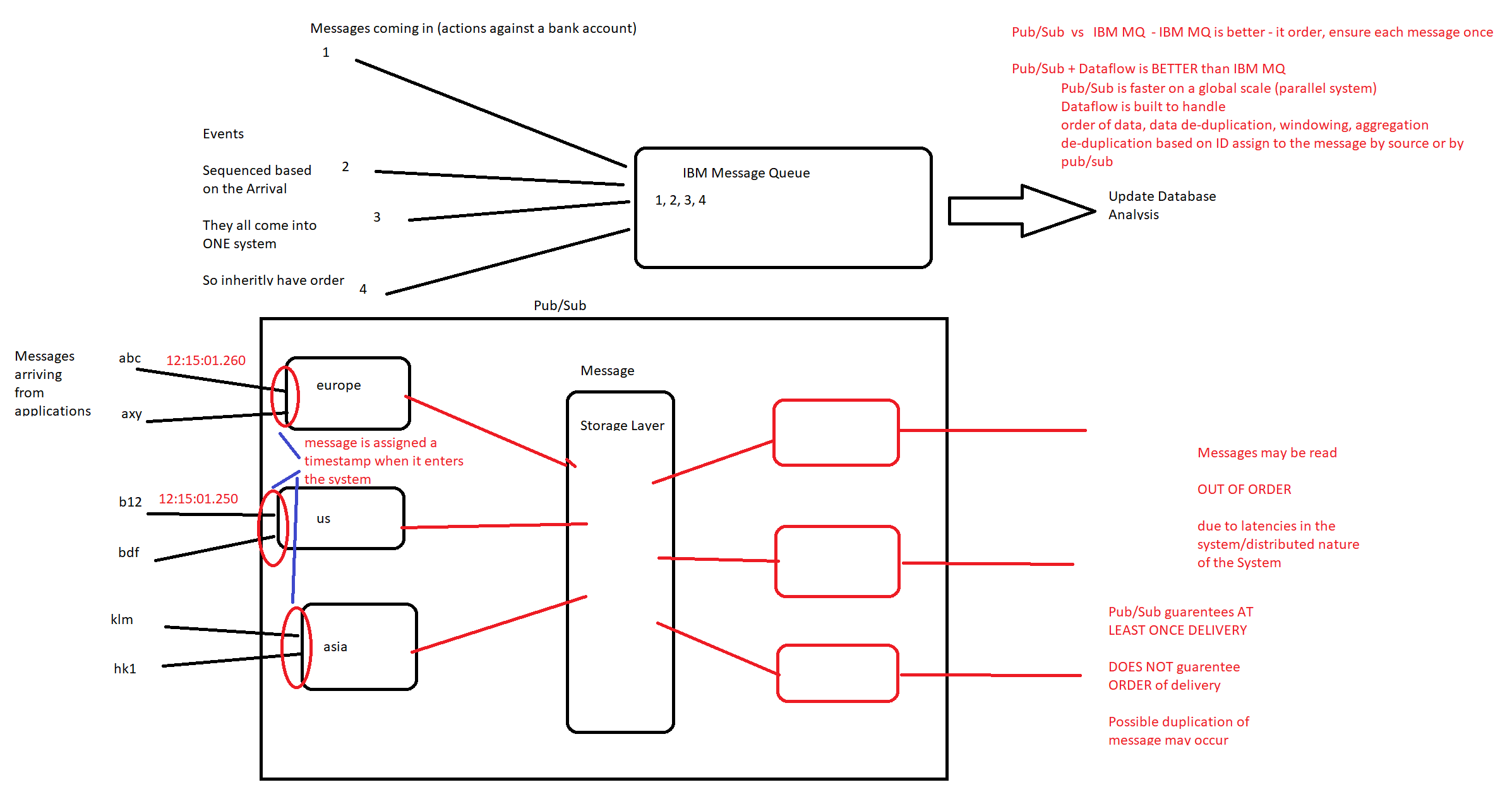
Mobile Game Analytics

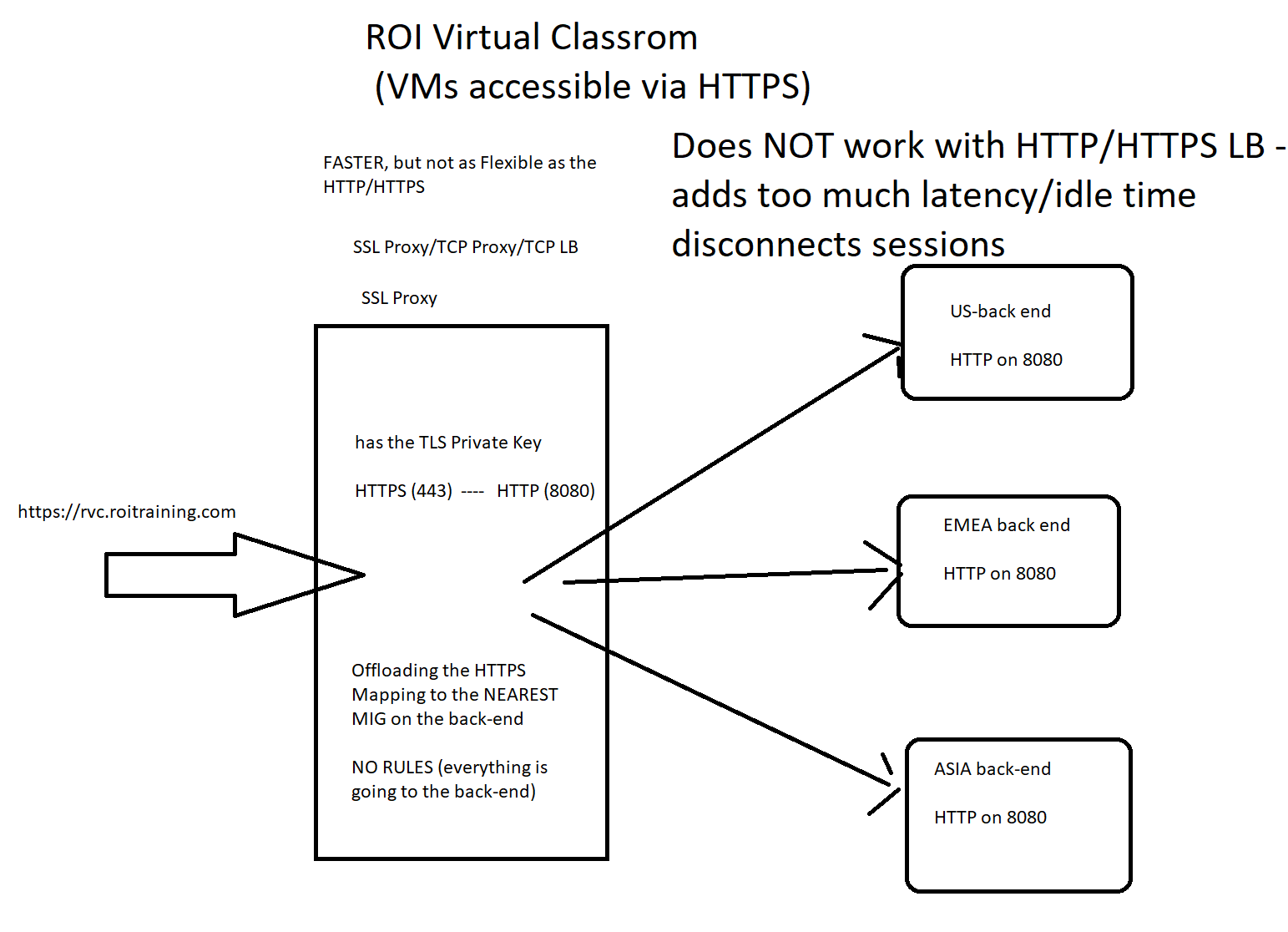


TCP Load Balancer vs HTTPS Load Balancer

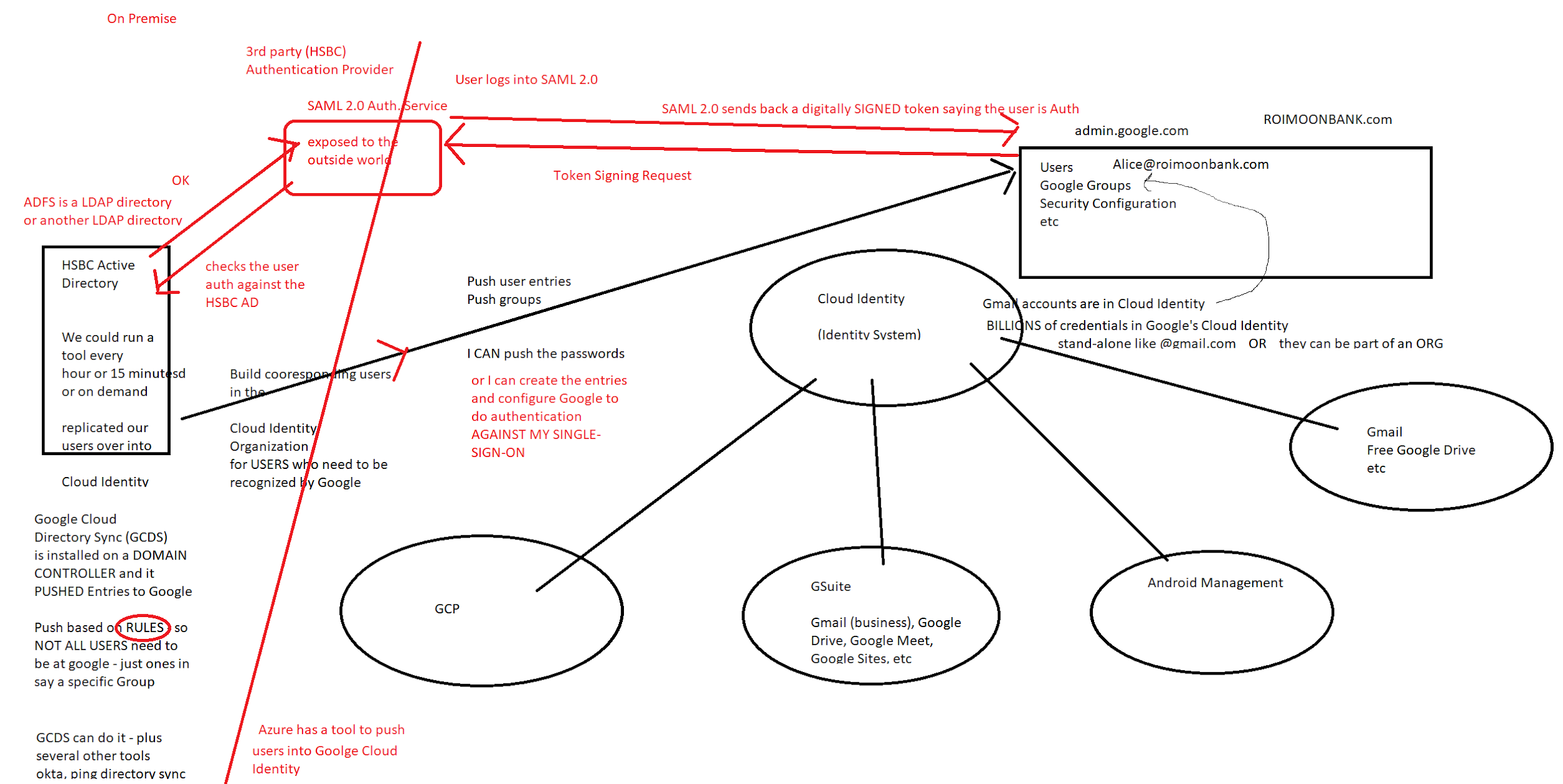


Pub/Sub+Dataflow vs IBM MQ





Authentication Visualization on Google - Cloud Identity



TerramEarth Case Study

