GCP

Google Cloud

Professional Data Engineer

Google Certified

Professional Data Engineer

Professional Data Engineer

Google Cloud Certified
PROFESSIONAL
DATA ENGINEER

- Pay attention for 5 minutes, before we dive in.
- Challenging certification, and course is long so have patience.
- Advance professional certification, Expect basics Associate level GCP knowledge
- Learn by Doing
- So with every exam objective, There is hand-on Lab – 80+

GCP certifications

Google Cloud Certified
PROFESSIONAL
DATA ENGINEER
Google Cloud Certified
PROFESSIONAL
CLOUD ARCHITECT
Google Cloud Certified
PROFESSIONAL
CLOUD DEVELOPER
Google Cloud Certified
PROFESSIONAL
CLOUD DEVOPS ENGINEER
Google Cloud Certified
PROFESSIONAL
CLOUD SECURITY ENGINEER
Google Cloud Certified
PROFESSIONAL
COLLABORATION ENGINEER
Google Cloud Certified
PROFESSIONAL
CLOUD NETWORK ENGINEER

Google Cloud Certified
ASSOCIATE
CLOUD ENGINEER

https://cloud.google.com/certification/guides/data-engineer

Cloud Cost for this course

- $0 – for GCP account
- GCP Free trial
- $300 for next 3 months https://cloud.google.com/free
- Length: Two hours
- Registration fee: $200 (plus tax where applicable)
- Languages: English, Japanese
- Exam format: Multiple choice and multiple select,

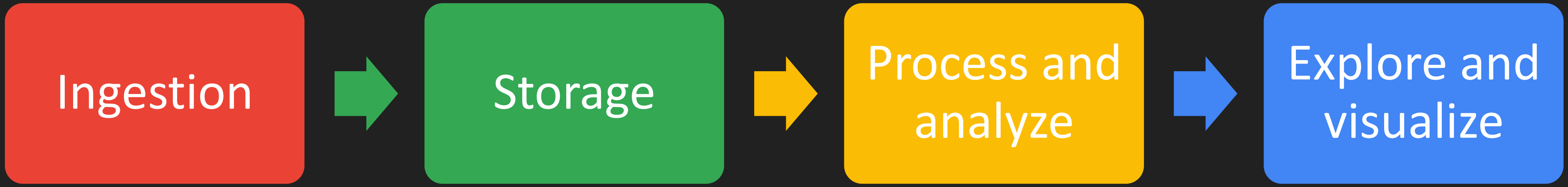
Google Cloud Certified PROFESSIONAL DATA ENGINEER

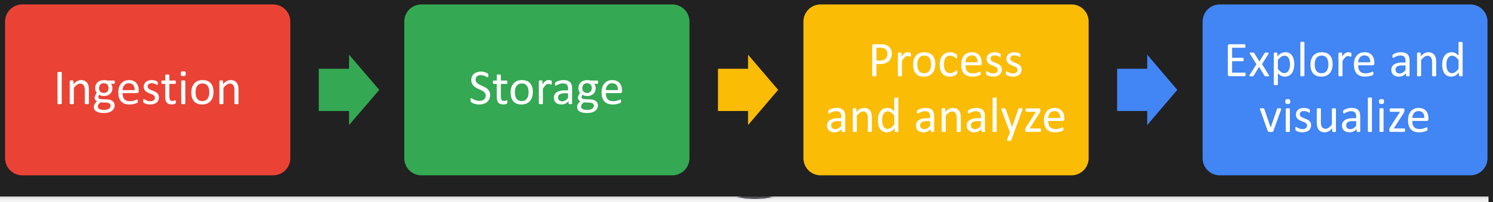
Data Engineering Concepts

Data Pipelines Basics

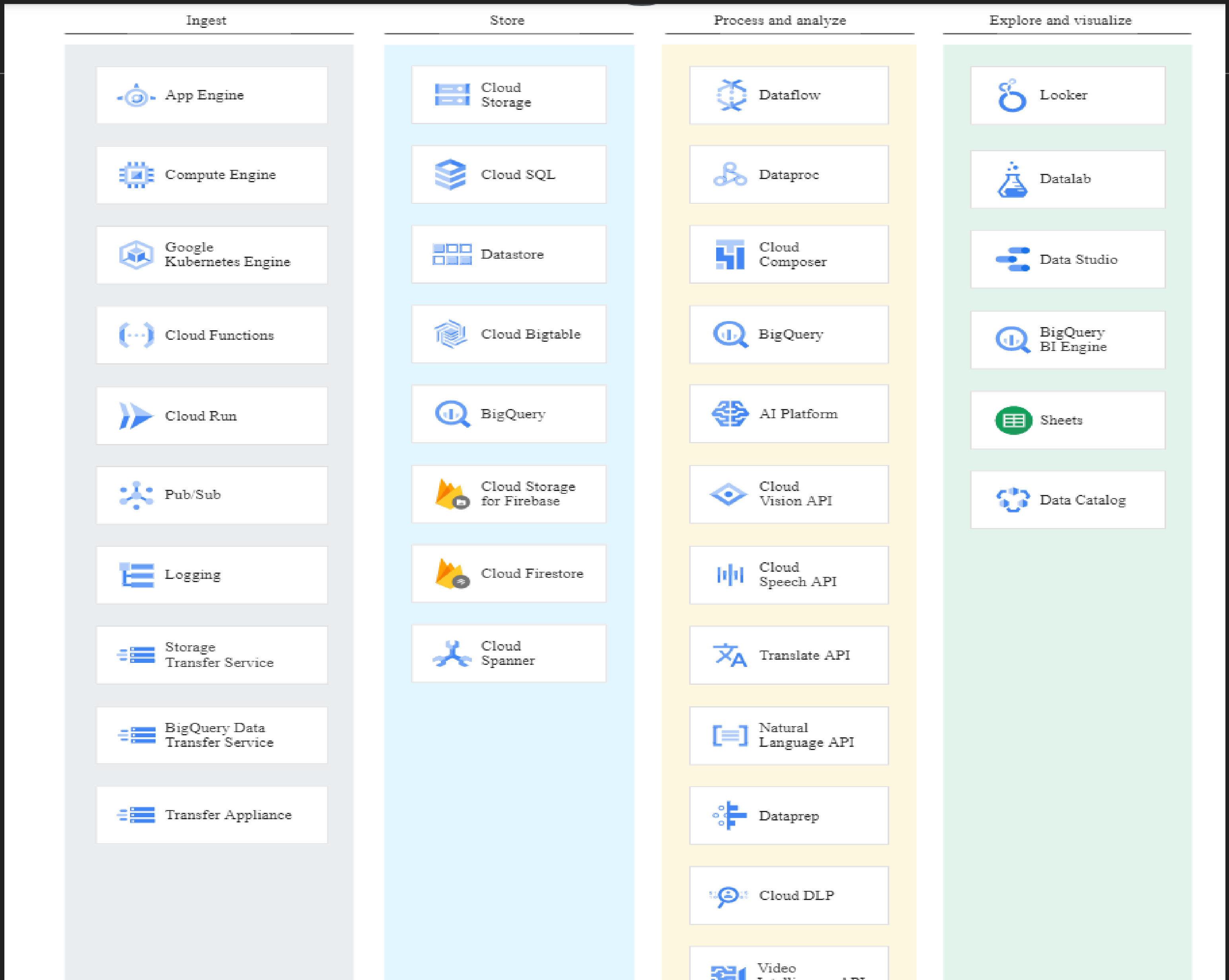
Data engineering Overview

- Data Pipelines
- How Data Flows





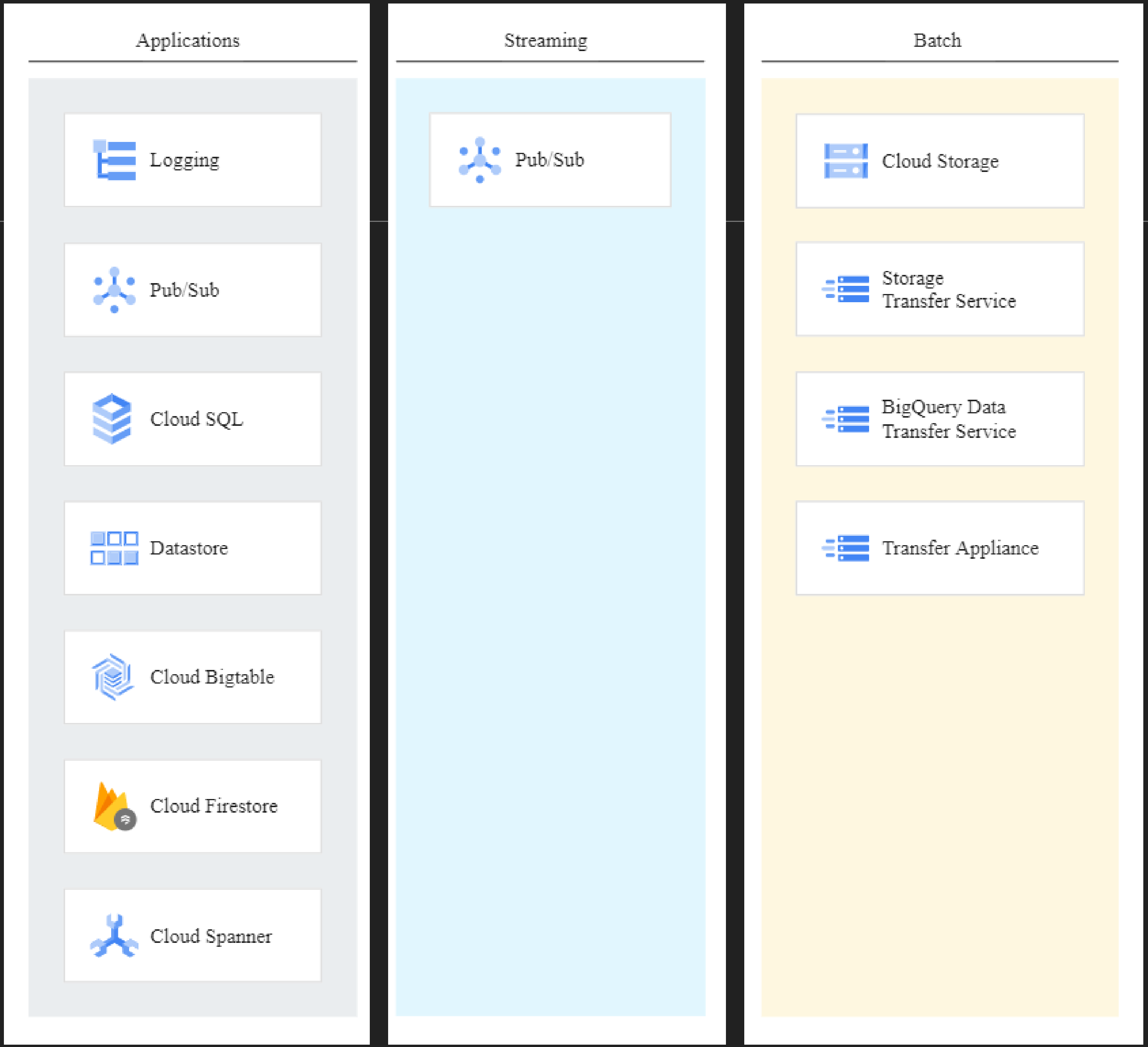
Ingestion → Storage → Process and analyze → Explore and visualize



Ingest	Store	Process and analyze	Explore and visualize
App Engine	Cloud Storage	Dataflow	Looker
Compute Engine	Cloud SQL	Dataproc	Datalab
Google Kubernetes Engine	Datastore	Cloud Composer	Data Studio
Cloud Functions	Cloud Bigtable	BigQuery	BigQuery BI Engine
Cloud Run	BigQuery	AI Platform	Sheets
Pub/Sub	Cloud Storage for Firebase	Cloud Vision API	Data Catalog
Logging	Cloud Firestore	Cloud Speech API	
Storage Transfer Service	Cloud Spanner	Translate API	
BigQuery Data Transfer Service		Natural Language API	
Transfer Appliance		Dataprep	
		Cloud DLP	
		Video Intelligence API	

Ingest

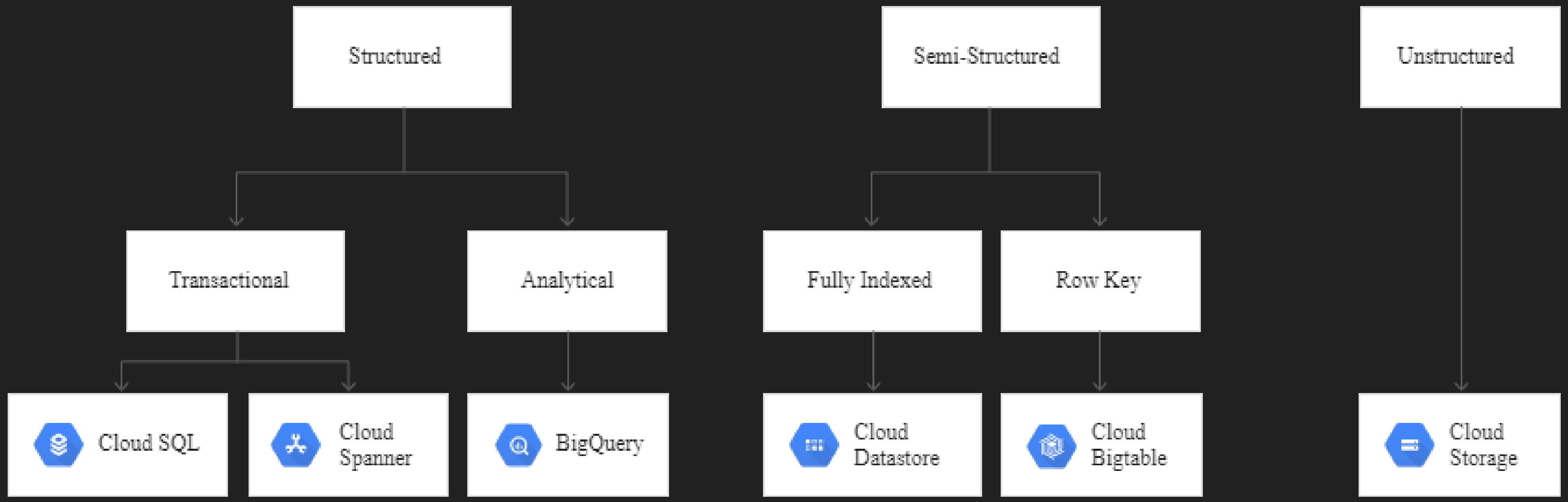
- ➢ Gather Data from multiple sources
- ➢ Data gather from App
  - ➢ Event Log, Click stream Data, e-commerce Transaction
- ➢ Streaming Ingest
  - ➢ PubSub
- ➢ Batch Ingest
  - ➢ Different Transfer services
  - ➢ GCS - gsutil



Applications	Streaming	Batch
Logging	Pub/Sub	Cloud Storage
Pub/Sub		Storage Transfer Service
Cloud SQL		BigQuery Data Transfer Service
Datastore		Transfer Appliance
Cloud Bigtable		
Cloud Firestore		
Cloud Spanner		

Store

- Cost efficient & durable data storage



Structured

Transactional

Cloud SQL

Cloud Spanner

Analytical

BigQuery

Semi-Structured

Fully Indexed

Cloud Datastore

Row Key

Cloud Bigtable

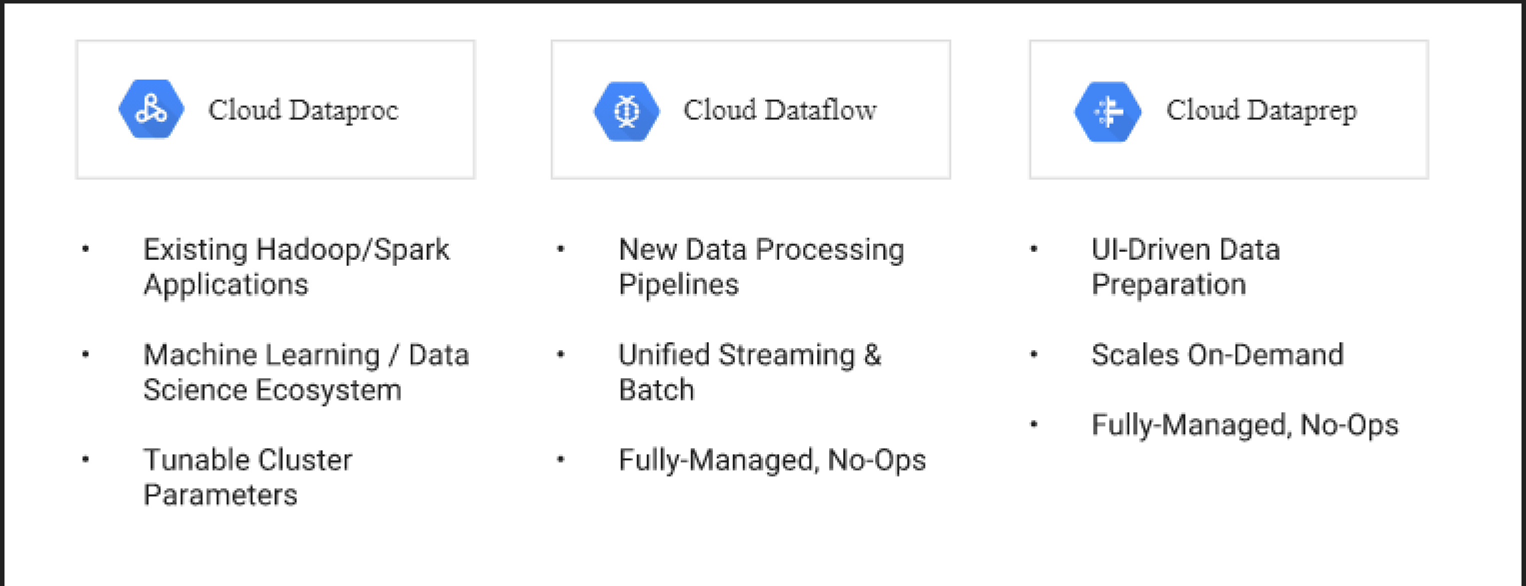
Unstructured

Cloud Storage

https://cloud.google.com/architecture/data-lifecycle-cloud-platform#store

Process & Analyze

- What kind of outcome you want
- What analysis you want to perform
- Convert Data into meaning
- Analyze Data with BigQuery
- Apply ML with
  - BigQuery ML
  - Spark ML with DataProc
  - Vertex API
    - Build ML Model with Auto ML/Custom Model



Cloud Dataproc

- Existing Hadoop/Spark Applications
- Machine Learning / Data Science Ecosystem
- Tunable Cluster Parameters

Cloud Dataflow

- New Data Processing Pipelines
- Unified Streaming & Batch
- Fully-Managed, No-Ops

Cloud Dataprep

- UI-Driven Data Preparation
- Scales On-Demand
- Fully-Managed, No-Ops

https://cloud.google.com/architecture/data-lifecycle-cloud-platform

Explore and visualize

- Google Data Studio – Easy to use BI Engine
  - Dashboard & Visualization
- Datalab
  - Interactive Jupyter Notebook
  - Support for all Data Science Library
- ML Prebuilt API
  - Vision API
  - Speech API

Types of Data - Structure

Structured

Semi-Structured

Unstructured

Structured Data

- Tabular Data
- Represented by Rows & Columns
- SQL can be used to interact with data
- Fixed Schema
- Each row has same number of columns
- Relational Database are structured
- MySQL, Oracle SQL, PostgreSQL , MSSQL
- In GCP, Cloud SQL, Cloud Spanner

Book_id	Book_name	Author_id
100	C	1
101	Java	1
102	Python	2

Author_id	Author_name
1	John
2	Alice

Semi-Structured Data

- Each Record has variable number of Properties
- No Fixed Schema
- Flexible structure
- NoSQL kind of Data
- Store data as key-value pair
- JSON – Java Script object Notation are base way to represent semi structure data
- MongoDB, Cassandra, Redis, Neo4j
- In GCP, BigTable, DataStore, memoryStore

```
Doc #1{
        “studentID” : 100,
        “name” : “john”,
         “score” : 78,
        “country” : “US”
},
Doc #2{
        “studentID” : 101,
        “name” : “Alice”,
        “rank” : 7,
},
```

UnStructured Data

- No Pre define Structure in Data
- Image
  - video data,
  - natural Language are example of unstructured data
- Google Cloud Storage, File store inside GCP to store Unstructure data

Batch Data vs Streaming Data

- Batch Data Processing
  - Defined Start & End of data – data size is known
  - Processing High volume of data after certain periodic interval
  - Long time to process data
  - Payment processing
- Streaming Data
  - Unbounded, No End defined
  - Data is processed as it arrives
  - Size is unknown
  - No much heavy processing – take millisecond - seconds to process data
  - Stock data processing

GCP Fundamental

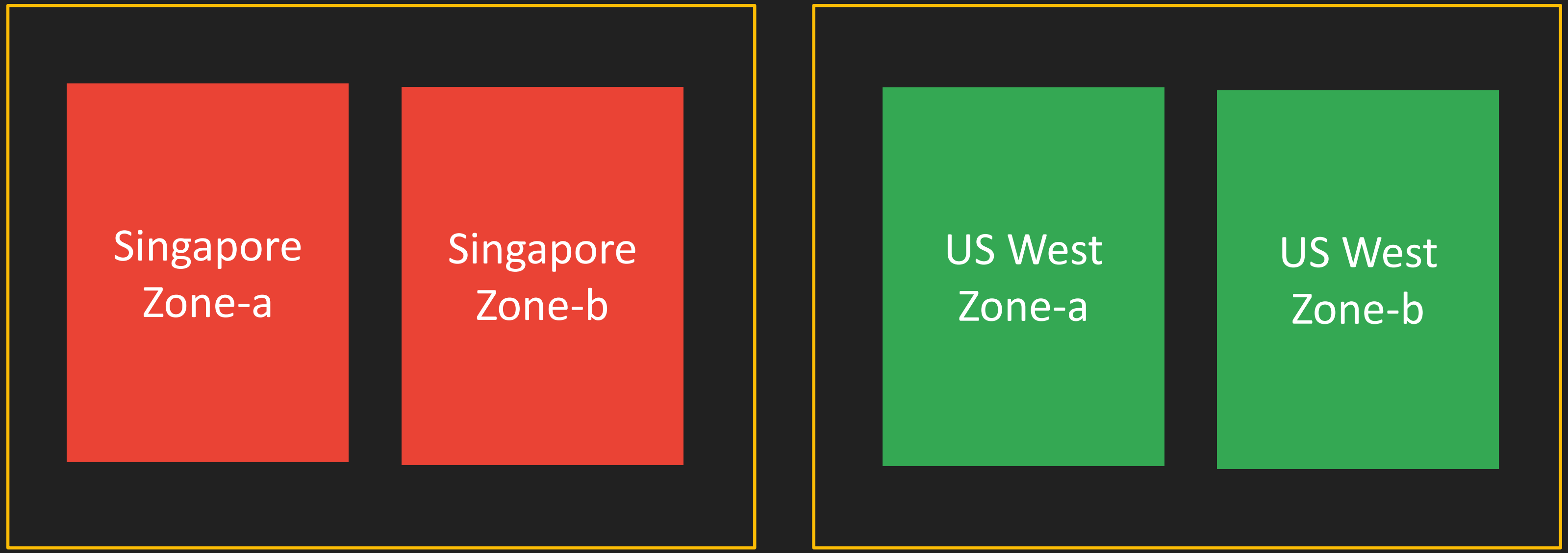
[redacted]

[redacted]

GCP Regions & Zones

Why Zones & Regions

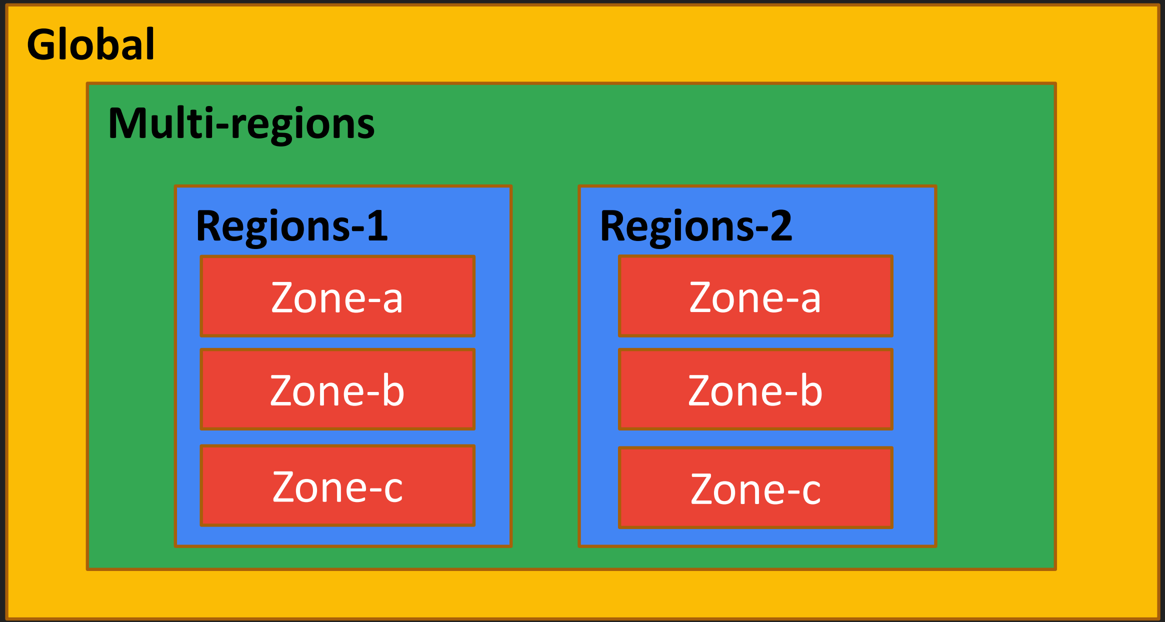
- Low latency
- Follow Government rules
- High availability
- Disaster recovery



GCP (Zones & Region)

Fascinating Number: Google Is Now 40% Of The Internet (forbes.com)

- Zones – Independent data Center
- Region – Geographical area
- Multi-region : Collection of Geographical
- Global - Anywhere



Global Locations - Regions & Zones | Google Cloud

Create GCP Free Tier Account

GCP Services

Storage & Database

- Cloud Storage, File Store
- SQL, Spanner, Big Query, Big table

ML/AI

- Vertex AI
- Prebuilt API
- Custom Model
- Auto ML

Data processing

- Data Proc, Data Flow
- Data Prep
- Data Catalog
- Big Query, PubSub
- Composer, Data Fusion

Secondary Services

- Compute Engine
- Kubernetes
- App Engine – Cloud Run

GCP Basic Services

Basic Infrastructure services in GCP

IAM

Compute Engine

Kubernetes

App Engine

Cloud Run

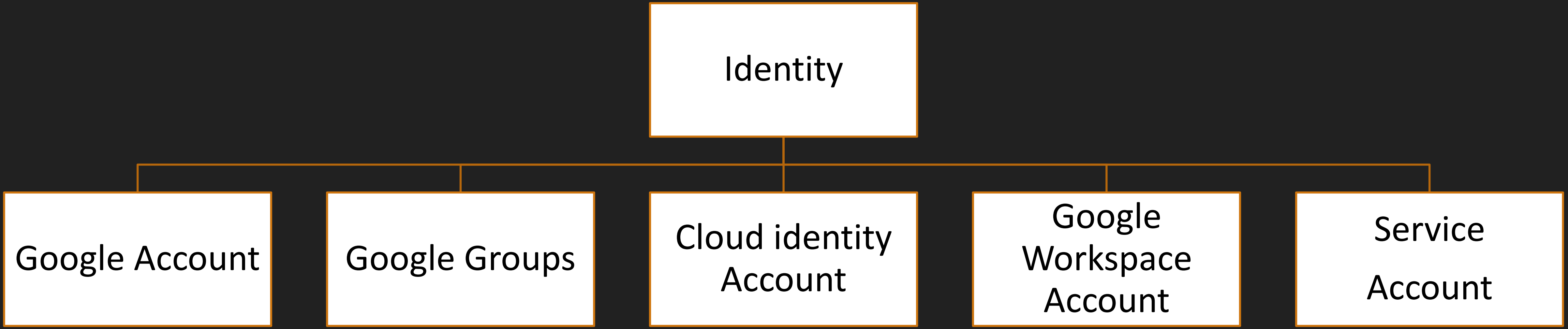
Cloud Function

VPC

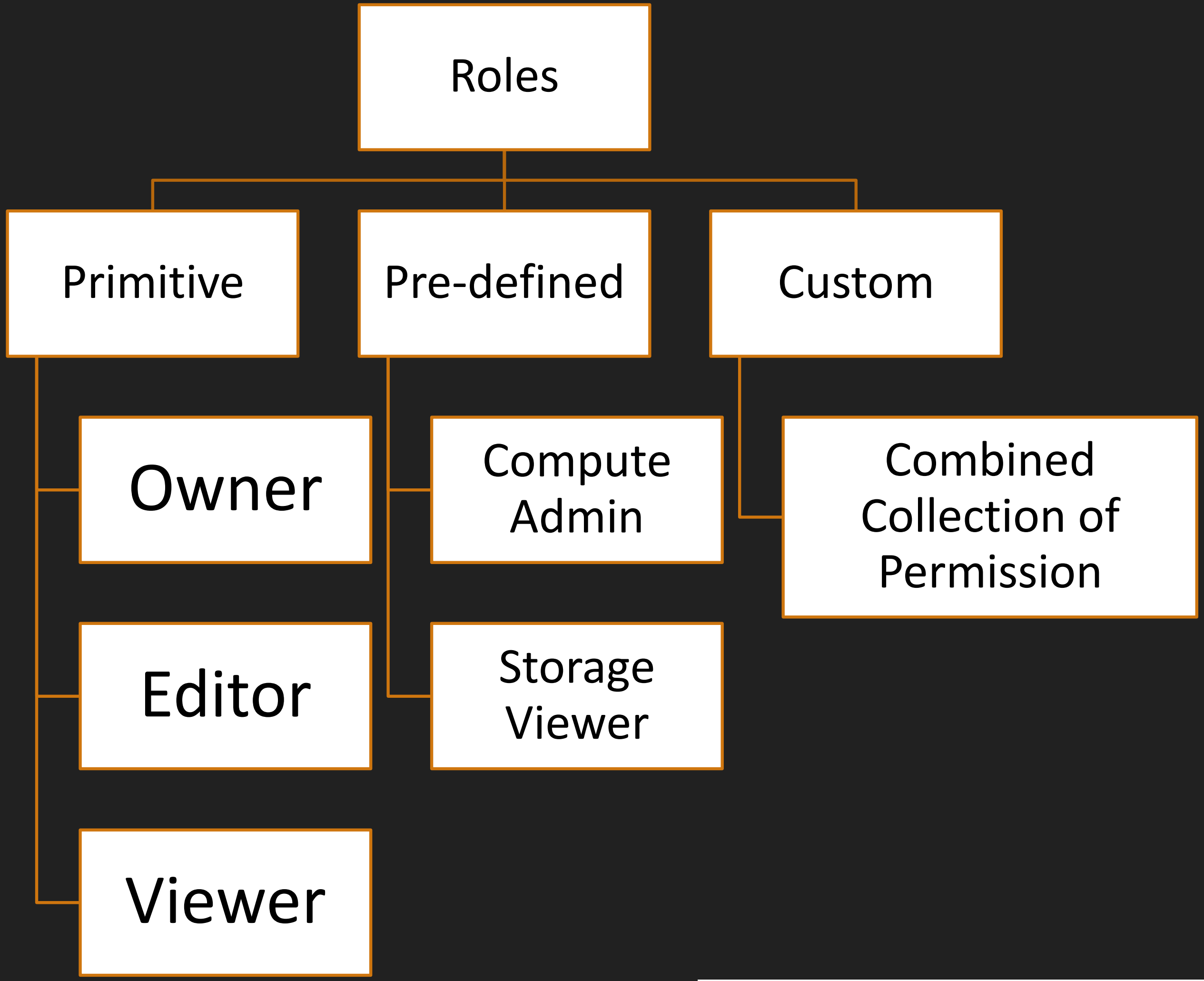
IAM

- Identity & access management
- Who can do What on Which resources
- Who - Identity
- What - Action : Create, Update, Delete
- Which – Resources, Compute Engine, App Engine, Cloud Storage
- Roles : Collections of Permissions
- Built-in Roles
- Custom Role
- Service Account

IAM – Identity



IAM – Roles & Permission



- Roles are collection of permissions
- One can assign Role to identity, but Can not assign permission directly.

Assign Roles to identity

Service Account

- For non human – like for Apps, services
- Service Account is identity for Compute engine
- Service account keys can be used for authentication
- Max 10 keys per Service Account
- Max 100 Service Account per project
- Let’s Explore Service Account

Provision Virtual machine

- Basic Building block of any Cloud
  - Compute Engine
- IAAS – Full Control, more flexibility, more responsibility
- Important parameter :
  - Zone
  - Service Account
  - Machine family – CPU, RAM
  - Boot Disk
  - Storage
  - Virtual Private Cloud

App Engine Deployment

- PAAS Solution
- No Server management
- Deploy HTTP based application
- Focus on Code
- Standard & Flexible mode
- Support many runtime engine
  - Python, java, Go, Node JS
- Let’s Deploy Hello World NodeJS App

GKE – kubernetes Engine

- Let’s say
  - you want to create 100’s of container to scale your app
  - need some automate approach which fully manage all container lifecycle
- Kubernetes is the solution for it
- Open source
- Orchestration system for containerized application.
- Open Source – Developed by Google , Launched in 2014 – Kubernetes
- In 2015, Google Launched cloud version - GKE
- Cloud Agnostics
- Written in Go language

- Let’s Deploy image to GKE
  1. Create Cluster
  2. Deploy Workload (Container image)
  3. Expose Outside World

Google Kubernetes Engine

- Orchestration system for containerized application.
- Open Source – Developed by Google
- Launched in 2014 – Kubernetes
- In 2015, Google Launched cloud version - GKE
- Cloud Agnostics
- Written in Go language
- Let’s deploy App on Kubernetes

Deploy Containerized app on Google Kubernetes Engine - GKE (CAAS)

Google Cloud Run

- Next Generation App Engine
- Deploy Containerized Workload
- No Server management
- Auto-scaling as Traffic grows
- Let’s See in action

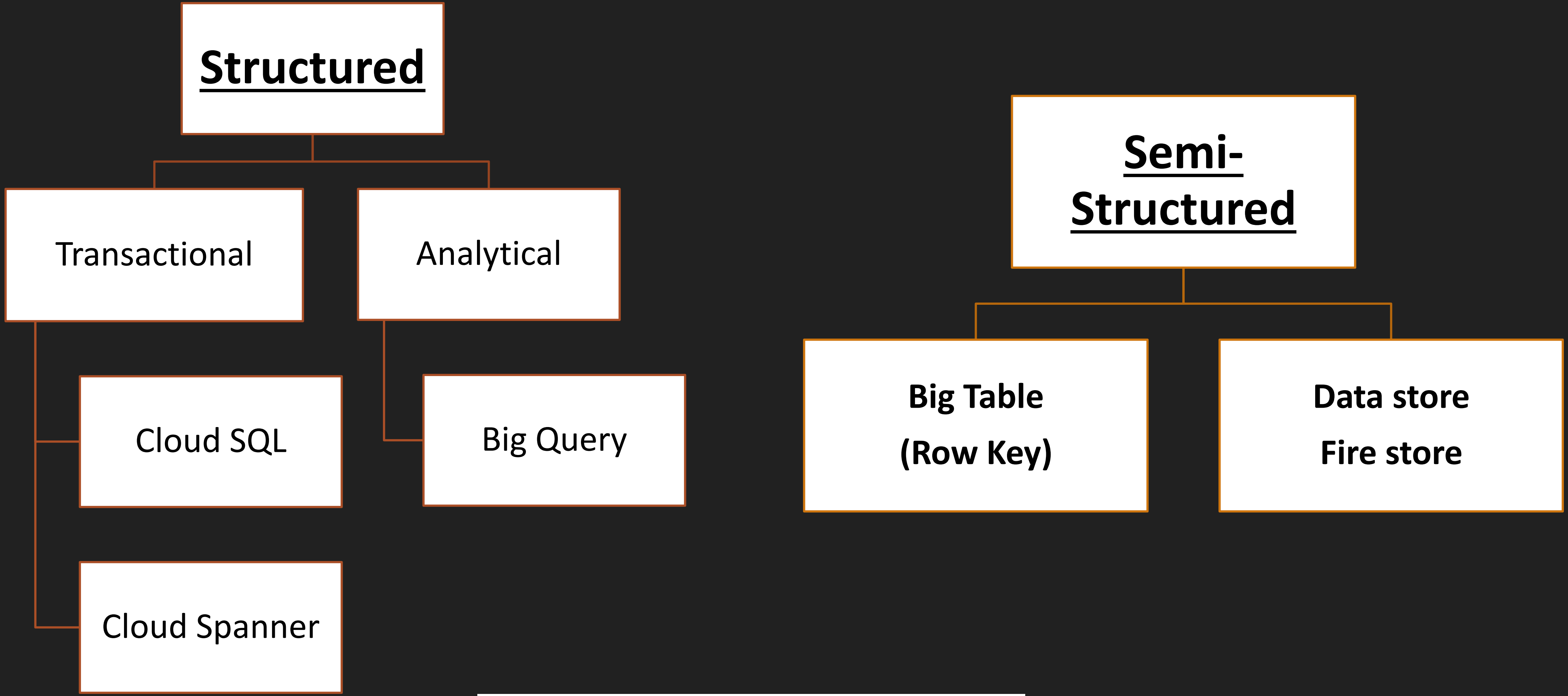
Google Cloud Function

- Server less
- Fully managed
- Build small micro service
- Auto scaling as traffic increase
- Event based trigger
  - Http
  - File upload etc.
  - Message pushed to pub/sub
- Let’s See in action

Different storage product (GCP)

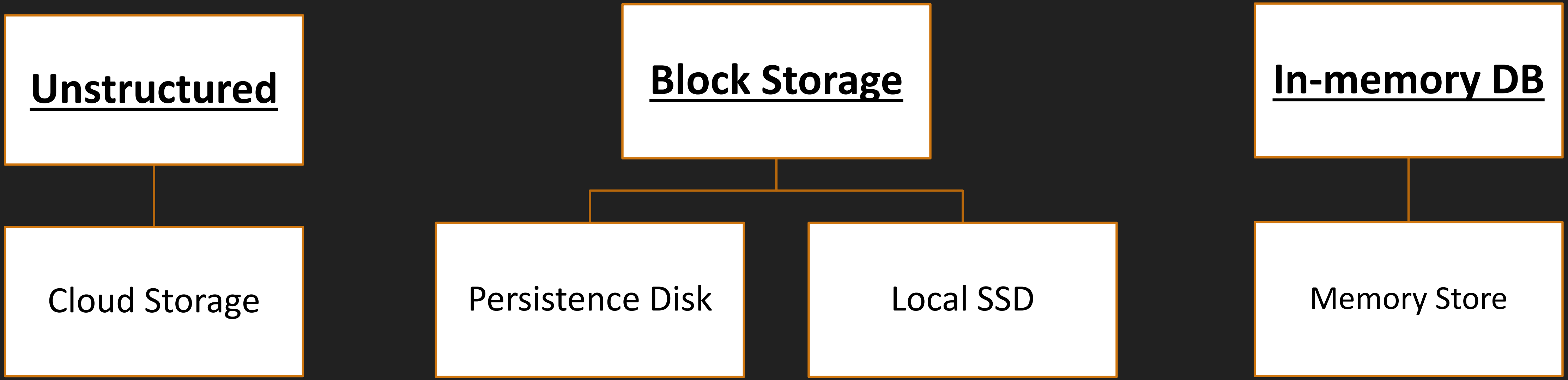
[redacted]

Different storage product



- **Structured**
  - Transactional
    - Cloud SQL
    - Cloud Spanner
  - Analytical
    - Big Query
- **Semi-Structured**
  - **Big Table (Row Key)**
  - **Data store Fire store**

Different storage product



Unstructured

Cloud Storage

Block Storage

Persistence Disk

Local SSD

In-memory DB

Memory Store

Google Cloud Storage

Google Cloud Storage

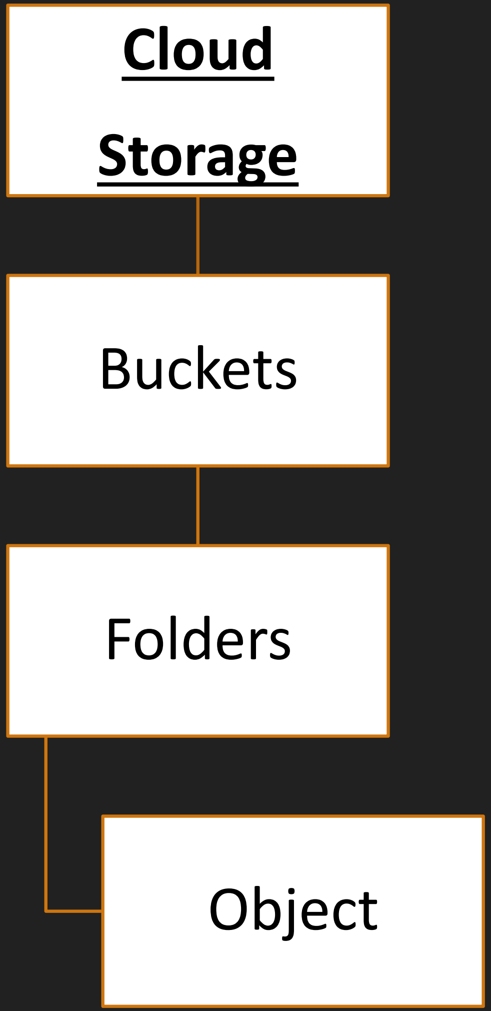
- Object storage solution in GCP
- Unstructured Data storage
  - Image
  - Video
  - Binary File, etc…
- Cloud storage can be used for long term archival storage
- Can be access object over http, Rest API
- No capacity planning required
  - Scale to Exabyte
- Unlimited data can be stored
- By Default Data is encrypted at rest
- In transit also by default encryption.

Google Cloud Storage

- No minimum Object Size
- Max object size is 5 TB
- High Durability – 99.999999999% annual
- Object can be Globally access
- Single API to access across multiple storage class
- Data is geo - redundant
  - Due to Multiregional
  - Dual-Region storage

Object Organization

- Global unique name for bucket
- Example access URL :
  - https://storage.cloud.google.com/[bucket]/[objectname]
- Bucket name appear in URL
- So, be careful while naming bucket
- Does not store anything like file system
  - Folder are virtual
- Bucket level lock with data retention policy
- Object are immutable
- Object can be versioned



Storage Location

Region

- Lowest latency within a single region
- Replicated data across multiple zone in single region

Dual-region

- High availability and low latency across 2 regions (Paired region)
- Auto-failover

Multi-region

- Highest availability across continent area – US, EU, Asia
- Auto-failover

Storage Class

- How frequently access data
- How much amount of data

Standard

- Good for Hot data
- High frequency access
- Storage Costliest
- Access cost is very low
- Low latency
- SLA :
  - 99.95% Multi/Dual
  - 99.9% Regional

Near line

- Low Frequency access Once in a 30 days
- Storage is Cheaper than standard
- Access cost will increase
- Back up
- SLA : 99.9% Multi/Dual
  - 99.0% for Regional

Cold line

- Very low frequency to access
- Once in 90 days
- Storage is Cheaper than Near line
- SLA :
  - 99.9% Multi/Dual
  - 99.0% for Regional

Archive

- Offline data
- Backup
- Data access is once in year
- Storage Cheapest
- Access cost very high
- No SLA

[Hands-on]
Google Cloud Storage

[redacted]

Object Lifecycle management

- Based on condition what action needs to perform on object.
- Condition
  - Object age
  - Object file type
  - after some specific date
- Action
  - Transition to different storage class for high performance
  - Like – Standard to Nearline
  - Coldline to Delete

Secure Data with Encryption

- Encryption
  - Google managed Encryption keys
    - No Configuration
    - Fully managed
  - Customer managed Encryption keys
    - Create keyring in Cloud KMS
    - key will be managed by customer. Like Key rotation
  - Customer supplied Encryption keys
    - We will generate Key with : openssl rand =base64 32
    - gsutil – encrypt with CSEK

Object Versioning

- Help to prevent accidental deletion of object
- Enable/Disable versioning at bucket level
- Get access to older version with (object key + version number)
- If you don’t need earlier version, delete it & reduce storage cost
- If you don’t specify version number, always retrieve latest version
- Let’s see in action

Controlling access

- Who can do what on GCS at what level
- Permissions
- Apply at Bucket level
  - Uniform level access
    - No Object level permission
    - Apply uniform at all object inside bucket
  - Fine grained permission
    - Access Control List – ACL For Each object Separately

- Apply Project level
  - IAM
  - Different Predefined Role
    - Storage Admin
    - Storage Object Admin
    - Storage Object Creator
    - Storage Object Viewer
  - Create Custom Role
- Assign Bucket level Role
  - Select bucket & assign role
  - To user
  - To other GCP services or product

Bucket Retention Policy

- Minimum duration for which bucket will be protected from
  - Deletion
  - modification
- Let’s see How to Configure it.

Signed URL

- Temporary access
- you can give access to user who doesn’t have Google Account.
- URL expired after time period defined.
- Max period for which URL is valid is 7 days.
- gsutil signurl -d 10m -u gs://<bucket>/<object>

GCS – Pricing

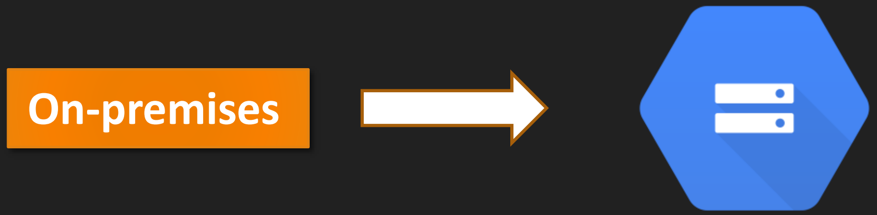
- Storage Pricing
- Data access Pricing
- Go to Cloud Console & create Bucket, observe pricing

Data Transfer Services

[redacted]

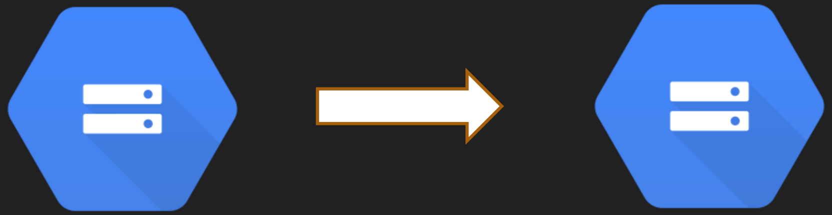
Data Transfer Service

- From On-premises to Google Cloud Storage (GCS)

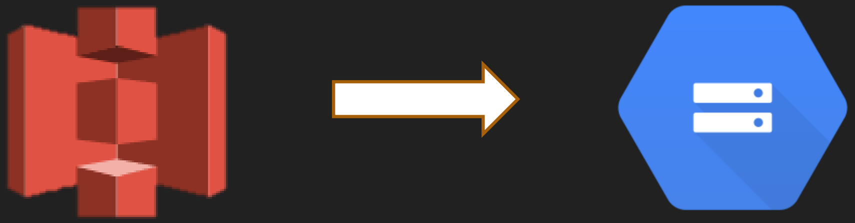


On-premises

- From One bucket to another bucket inside same GCP



- From Other public cloud Amazon S3, Azure Container to GCS



On-premises to (GCS)

- gsutil – command line utility
  - Online mode of transfer
  - install locally Google Cloud SDK
  - gsutil -m cp large_number_of_small_files (-m for parallel upload)
  - Should we go for it or not?
    - Good Network
    - Follow chart in next slide
- Transfer Service for on-premises data
  - This will quickly and securely move your data from private data centers into Google Cloud Storage
  - Two step process
    - installing an agent
    - create a transfer job

Transfer Appliance

- Transfer Appliance
  - Physical device which securely transfer large amounts of data to Google Cloud Platform
  - When data that exceeds 20 TB or would take more than a week to upload.

Online vs Offline transfer

Close

Far

Data Size

100 PB	124 days	3 years	34 years	340 years	3,404 years	34,048 years
10 PB	12 days	124 days	3 years	34 years	340 years	3,404 years
1 PB	30 hours	12 days	124 days	3 years	34 years	340 years
100 TB	3 hours	30 hours	12 days	124 days	3 years	34 years
10 TB	18 minutes	3 hours	30 hours	12 days	124 days	3 years
1 TB	2 minutes	18 minutes	3 hours	30 hours	12 days	124 days
100 GB	11 seconds	2 minutes	18 minutes	3 hours	30 hours	12 days
10 GB	1 second	11 seconds	2 minutes	18 minutes	3 hours	30 hours
1 GB	0.1 seconds	1 second	11 seconds	2 minutes	18 minutes	3 hours
	100 Gbps	10 Gbps	1 Gbps	100 Mbps	10 Mbps	1 Mbps

Network Bandwidth

Transfer Service|cloud data

- This will quickly and securely transfer data into Google Cloud Storage
- From various sources
  - Amazon S3
  - Azure Blob Storage
  - Move data between Cloud Storage buckets
- Create Transfer Job
- Onetime run or recurring

Google Block Storage

[redacted]

Google Block Storage

- Block storage – hard Disk storage
  - Direct attached Storage
  - Network attached Storage

Direct attached – Local SSD

- Local SSD
- Physically attached to VM
- Very High Performance – 10x to 100x of Persistence Disk
- Costlier than Persistence Disk
- You can not re attach to other VM
- Once VM destroy, Local SSD will be deleted
- Lower Availability
- Temporary/Ephemeral Storage
- No Snapshot
- Let’s see in action.

Local SSD with Compute Engine

Network attached storage

- Network attached hard disk
- Persistent Disks
- Zonal, Regional
- Not attached directly to any VM
- Can be re-attached with other VM
- Very Flexible – resize easily
- Permanent storage
- Snapshot supported
- Cheaper than Local SSD

Persistence Disk with Compute Engine

[redacted]

FileStore

Storage
Which storage to use when

- Cloud Storage
  - Unstructured data storage
  - Video stream, Image
  - Staging environment
  - Compliance
  - Backup
  - Data lake
- Persistent Disk
  - Attach Disk with VM & Containers
  - Share read-only disk with multiple VM
  - Database storage
- Local Disk
  - Temporary high performance attach Disk
- File Store
  - Performance predictable
  - Lift-shift millions of Files

Structured data Solution in GCP

[redacted]

Structured data

Cloud SQL

Cloud Spanner

[redacted]

Few Concept

- Relational data
  - OLTP
  - OLAP
- RTO vs RPO
- Vertical vs Horizontal Scaling
- Availability & Durability

OLTP & OLAP

[redacted]

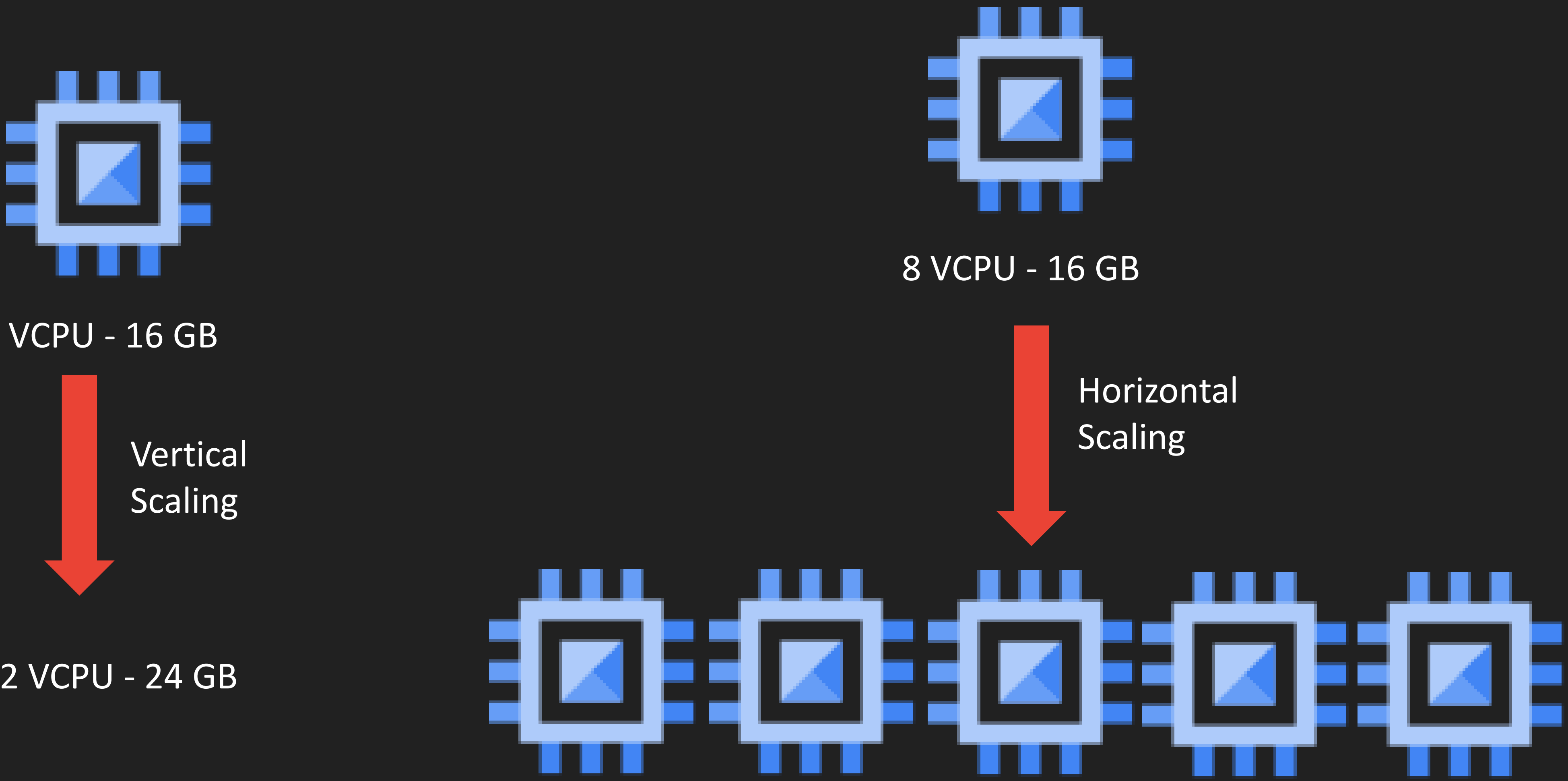
OLTP

- OLTP – Online Transaction Processing
- Simple Query
- Large number of small transaction
- Traditional RDBMS
- Database modification
- Popular Database - MySQL, PostgreSQL, Oracle, MSSQL
- ERP, CRM, Banking application
- GCP - Cloud SQL, Cloud spanner

OLAP

- OLAP – Online Analytical Processing
- Data warehousing
- Data is collected from multiple sources
- Complex Query
- Data analysis
- Google Cloud Big Query – Petabyte Data warehouse
- Reporting Application, Web click analysis, BI Dashboard app

Vertical – Horizontal Scaling



8 VCPU - 16 GB

Vertical Scaling

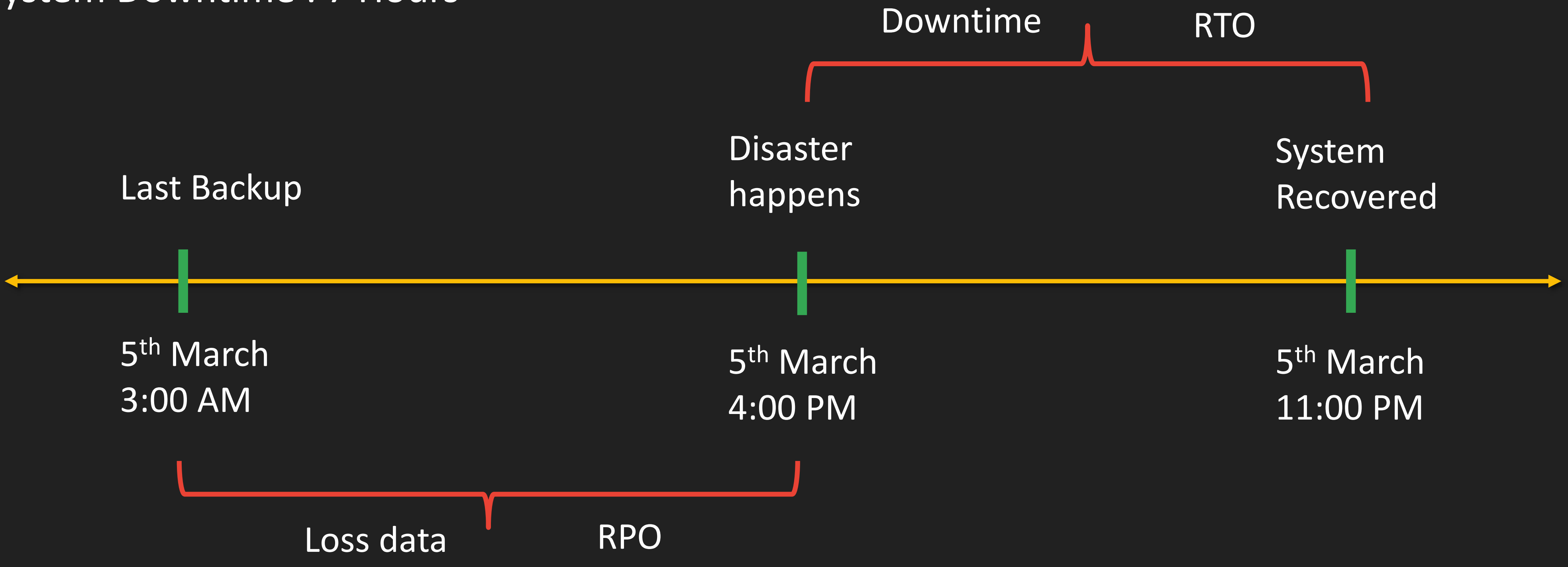
12 VCPU - 24 GB

8 VCPU - 16 GB

Horizontal Scaling

RTO & RPO

- Data loss : 13 hours
- System Downtime : 7 Hours



RTO & RPO

- RTO – Recovery Time objective
  - Maximum time for which system can be down
- RPO - Recovery Point objective
  - Maximum time for which organization can tolerate Dataloss

RTO & RPO

- RTO – Recovery Time objective
  - Maximum time for which system can be down
- RPO - Recovery Point objective
  - Maximum time for which organization can tolerate Dataloss

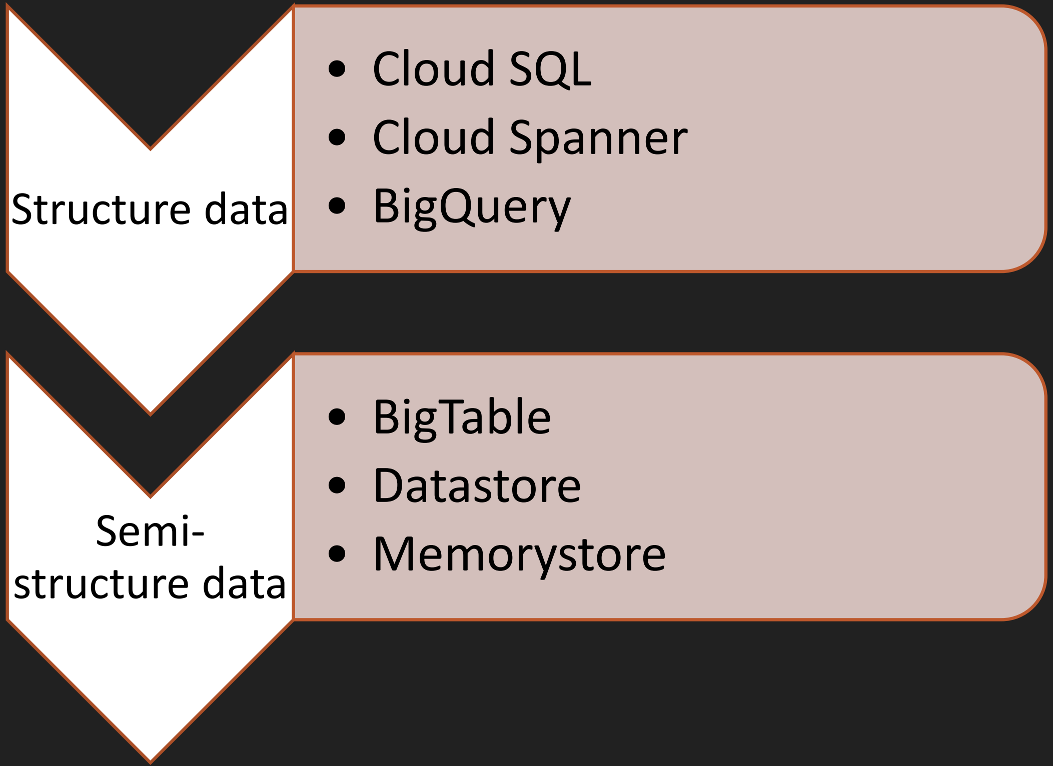
Durability

- If you loose data means
  - business is down
  - No business afford to loose data
- How healthy & resilient your data is
- Object Storage provider measure durability in terms of number of 9’s
- Example : 99.9999999999999% - 11 9’s
- That means that even with one billion objects, you would likely go a hundred years without losing a single one!
- https://cloud.google.com/blog/products/storage-data-transfer/understanding-cloud-storage-11-9s-durability-target

Availability

- If region goes down where your data stored
  - Replicate data across many region
- How much amount of time data is up/available to access.
- Data replicated across multiple regions, means higher Availability
- SLA – service level agreement
- SLA – 99.99% : four 9’s
- https://uptime.is
- https://cloud.google.com/terms/sla

GCP Database products



Structure data

- Cloud SQL
- Cloud Spanner
- BigQuery

Semi-structure data

- BigTable
- Datastore
- Memorystore

Google Cloud SQL

[redacted]

Google Cloud SQL

- Fully managed Relational database services for MySQL, PostgreSQL & SQL Server
- Lift & shift above database
- Regional Database with 99.95% SLA
- Storage up to 30 TB
- Scale up to 96 core & 416 GB Memory
- No Horizontal Scaling
- Data is encrypted with Google managed key or CMEK
- Cloud SQL can be accessed from anywhere like – App Engine, Compute Engine…
- Used for storing Transactional database
- Ecommerce, CRM kind application backend.

Google Cloud SQL

- No maintenance & auto update
- Back-up Database
  - On-demand Backup
  - Schedule backup
- Database migration service (DMS)
  - migrate data from different SQL system to Cloud SQL
- Point-in Time Recovery
- Scale with Read replicas – To transfer workload to other instance
- Export data
  - gcloud utility or Cloud Console
  - In SQL/CSV format

[Hands-on]
Create Google Cloud SQL

[Hands-on]
Connect Cloud SQL & IP Whitelisting

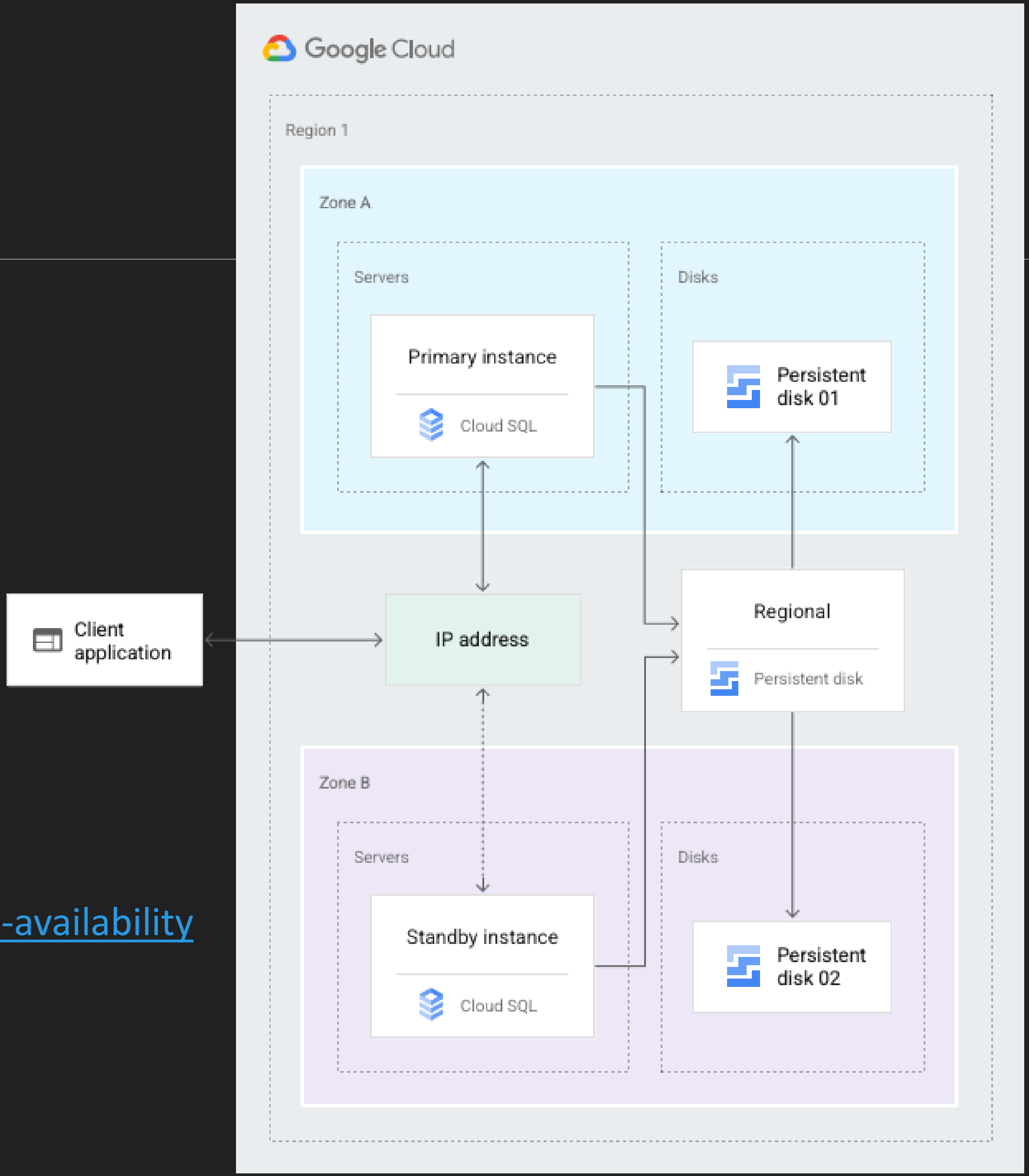
[Hands-on]
Data migration to Cloud SQL Demo

[redacted]

[Hands-on]
Cloud SQL failover demo

Google Cloud SQL Failover

https://cloud.google.com/sql/docs/mysql/high-availability



Cloud SQL Explore

Cloud SQL Export

[redacted]

Google Cloud Spanner

[redacted]

Google Cloud Spanner

- Distributed & scalable solution for RDBMS in GCP
- Fully managed, Mission critical application
- Horizontal Scalability
- use when Data volume > 2 TB
- Costlier than Cloud SQL
- Cloud SQL has just Read replicas,
  - where as in cloud spanner horizontal read/write across region
- Highly scalable, Petabyte scale
- Data is strongly typed.
  - Must define schema database
  - Datatype for each column of each table must be defined.
- 99.999% availability

- Cloud native solution – specific to GCP
  - Lift & Shift not possible, Not recommended
- Spanner = Cloud SQL + Horizontal Scalable
- Scale to petabyte
- Regional/ Multi-region level instance can be created
- Data export
  - can not export with gcloud
  - Cloud Console or Cloud Dataflow Job

Spanner vs RDBMS-SQL

	Spanner	Cloud SQL
Availability	High	During failover little downtime
Scalable	Horizontal	vertical
Price	Costly	Cheaper than spanner
SQL/Schema Support	yes	yes
Replication	High	Only Read Replica

[Hands-on] Cloud Spanner

[redacted]

Cloud Spanner Demo

- Create Spanner Instance
- Create database edu_db
- Create 2 Table
  - Author
    - AuthorID
    - AuthorName
  - Book
    - BookId
    - Bookname
    - AuthorId

After Job Done make sure to delete Spanner Instance

Which database to use when

- Cloud SQL
  - Lift & Shift SQL based system
  - CRM, Ecommerce App
  - Max Data size is 30 TB

- Cloud Spanner
  - Horizontal scalability
  - Low latency
  - High scalability in terms of storage + compute
  - if Data Storage requirement is beyond TB

Semi-Structured data Solution in GCP

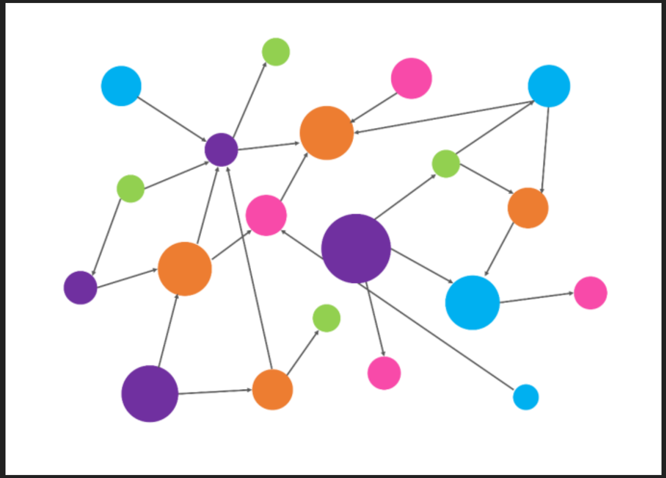
[redacted]

NoSQL Introduction

- Not a SQL
- Flexible Schema
- Variable number of property
- Data model will be like
  - Document
  - key-value pair
  - Graph based

```
{
        “studentID” : 100,
        “name” : “john”,
         “score” : 78,
        “country” : “US”
},
{
        “studentID” : 101,
        “name” : “Alice”,
        “rank” : 7,
},
```

key	value
100	std1
101	std2



Semi-structured data

BigTable

Datastore

Firestore

Google Cloud Datastore

Cloud Datastore

- Highly scalable NoSQL database
- Serverless
- Document kind data storage – MongoDB
- App Engine + Datastore
- SQL Like Queries – GQL
- Support ACID Transaction
- Multiple indexes
- Data replication across different region
- Use case
    - Session Info
    - Product catlog
- Export data from gcloud utility only

Datastore	RDBMS
Kind	Table
Entity	Row
Property	Column
Key	Primary Key

[Hands-on] Datastore

Google Cloud Firestore

Cloud Firestore

- Firestore is the next generation of Datastore
- Highly scalable NoSQL database
- Collection & Document Model
- Two mode
  - native Mode
  - datastore mode
- Real-time updates
- Mobile and Web client libraries
- Let’s see in Action

Firestore	RDBMS
Collection group	Table
Document	Row
Field	Column
Document ID	Primary Key

[Hands-on] Firestore

Datastore & Firestore Pricing

Google Cloud Memorystore

[redacted]

Cloud Memorystore

- Fully managed Inmemory database
- sub-millisecond data access
- Two engine supported
  - Redis
  - Memcached
- Only Internal IP
- Highly available with 99.9% SLA
- Import/Export data from Cloud Storage to memory store
- Let’s create memory store Instance

[Hands-on] Memorystore

[redacted]

Google Cloud BigTable

[redacted]

Cloud BigTable

- Fully managed
- Wide column NOSQL database
- Not serverless
- Scale horizontally with Multiple Node
- Scale to huge Volume of data
- Data stored at column wise
- Column are grouped into column family
- Milli second latency
- Handles millions of request per second
- How to access
  - cbt – command line (part of cloud sdk)
  - Hbase API

- No Multi column index
  - Only Row key based indexing
- Design Row Key is very important
- Design Row key by keeping in your mind
  - which is your frequent query in application
  - No Hot spotting
  - Don't use monotonically increasing key
- Seamless integration with
  - Warehouse – BigQuery
  - Machine Learning Product
- Used for
  - Financial data
  - Time series Data

Cloud BigTable

Row Key	Personal_data_cf		Professional_data_cf		
	name	age	salary	designation	company
1					
2					
3					

Professional_data_cf:salary

[Hands-on]
Google Cloud BigTable

BigTable Pricing

GCP Data Processing Solution

- BigQuery
  - Analytical Workload – Storage + Processing
- DataFlow – Apache Beam
- DataProc – Spark, Hadoop
- Data Fusion
- Cloud Composer
- Data Prep
- Cloud PubSub

Google Cloud BigQuery

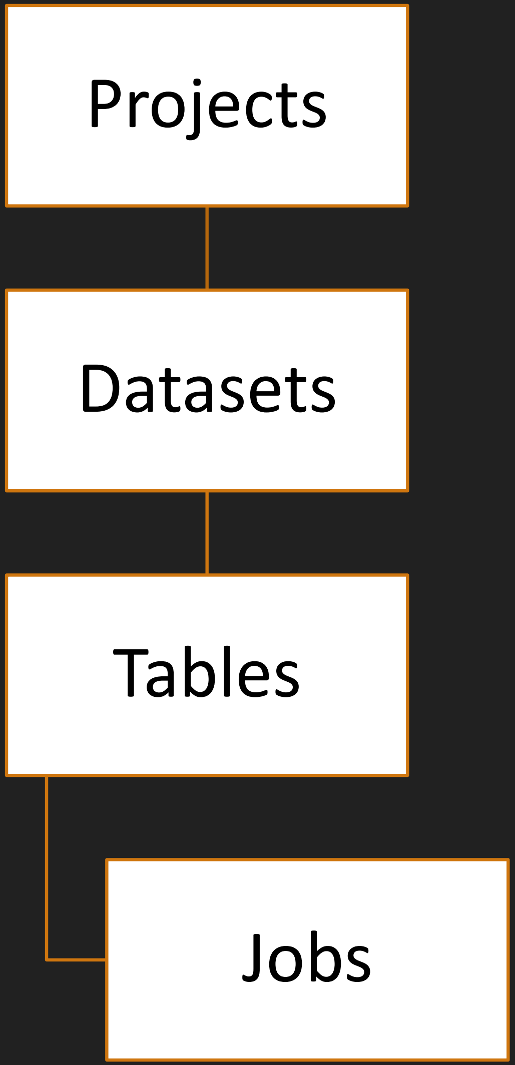
Cloud BigQuery

- Data warehouse solution in GCP
- Like Relational database – SQL schema
- Serverless
- Built using BigTable + GCP Infrastructure
- BigQuery is Columnar storage
- This is for Analytical database
  - not for Transactional purpose
- Exabyte scale
- Query using
  - Standard SQL
  - legacy SQL
- Big Query can query from external data source.
  - Cloud storage, SQL, Big Table

- Biquery can load data from various sources.
  - CSV, JSON, Avro, SQL and many more
- Query is very expensive
- $5 approx. for 1 TB of data scanned
- Before query execution do dry run.
- Alternative to OpenSource Apache Hive
- How to access BigQuery
  - Cloud Console
  - bq – command line tool
  - Client library - written in C#, Go, Java, Node.js, PHP, Python, and Ruby

BigQuery Data organization

- Projects are top level container in GCP
- Dataset hold multiple tables
- Each table must belong to dataset
- Assign Role at the organization, project, and dataset level
- Tables – contain data
  - It has Schema
- Types of Tables
  - Native tables, External tables, Views
- Jobs
  - manage asynchronous tasks
  - Types of Job
    - Load, Query, Extract, Copy



Projects

Datasets

Tables

Jobs

When BigQuery should be used

- When workload is analytical
- When Data doesn’t change in database, as bigquery use built –in cache
- For complex query
- When query takes more execution time.
- off-load some workload from primary transaction DB
- When you large volume of data
- No Join is preferred.
  - When you data is denormalized

[Hands-on]
Cloud BigQuery Explore + Public dataset

[Hands-on]
Cloud BigQuery +
Local data

[redacted]

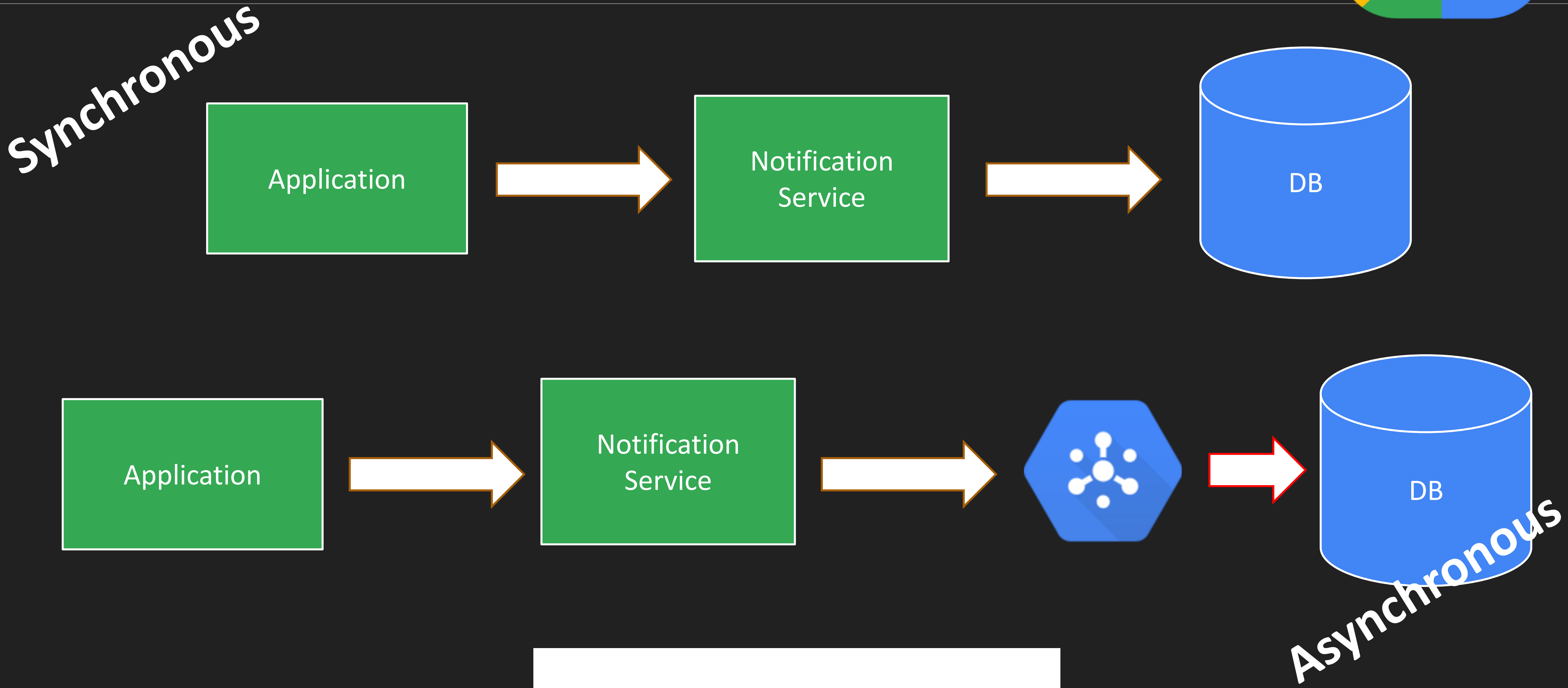
Cloud BigQuery Pricing

- On-demand
  - Pay for what you use
- Flat rat Pricing
  - Allocate compute & storage capacity

Google Cloud PubSub

[redacted]

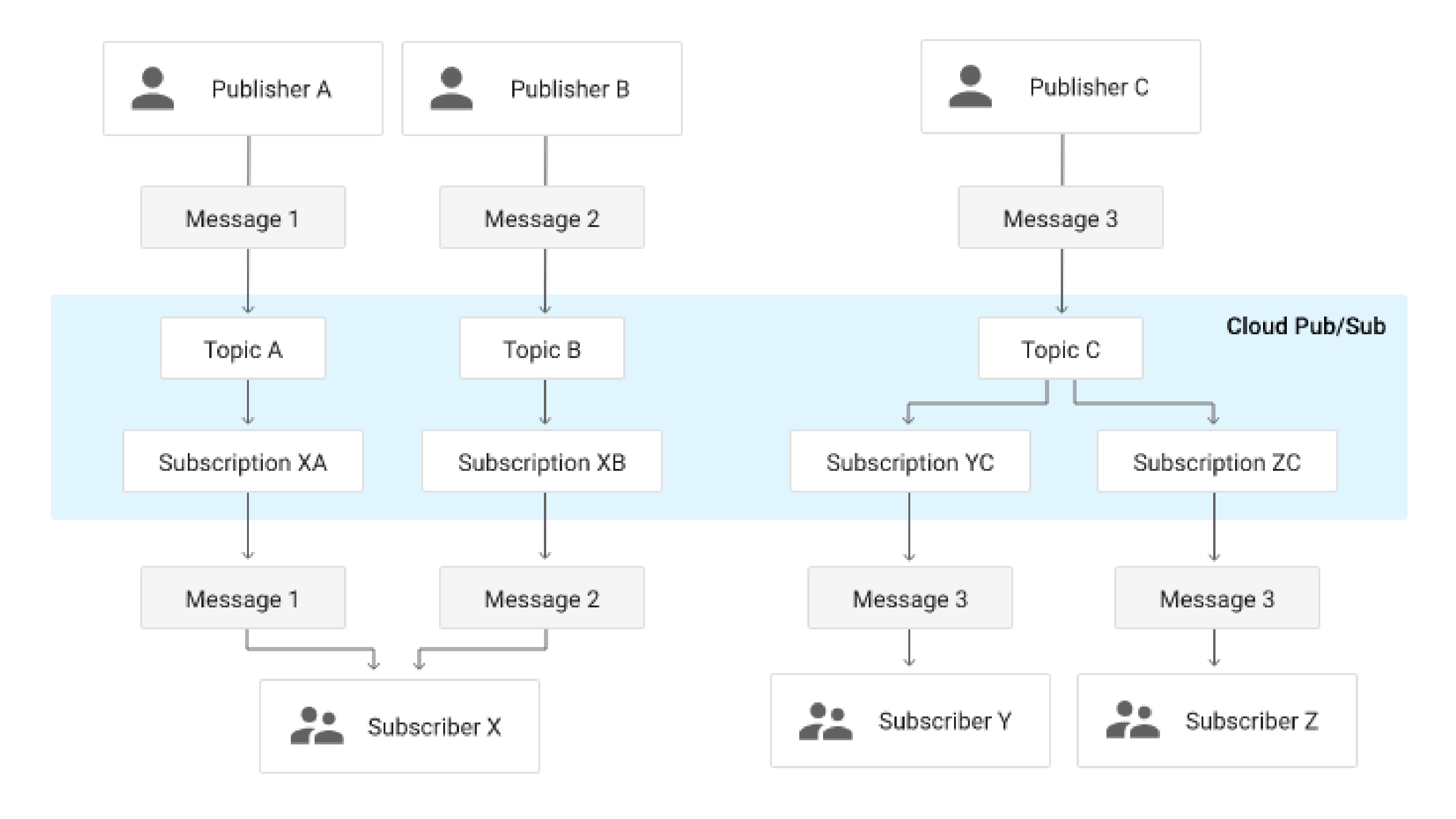
PubSub


Synchronous
Application
Notification Service
DB
Application
Notification Service
DB
Asynchronous

How PubSub works

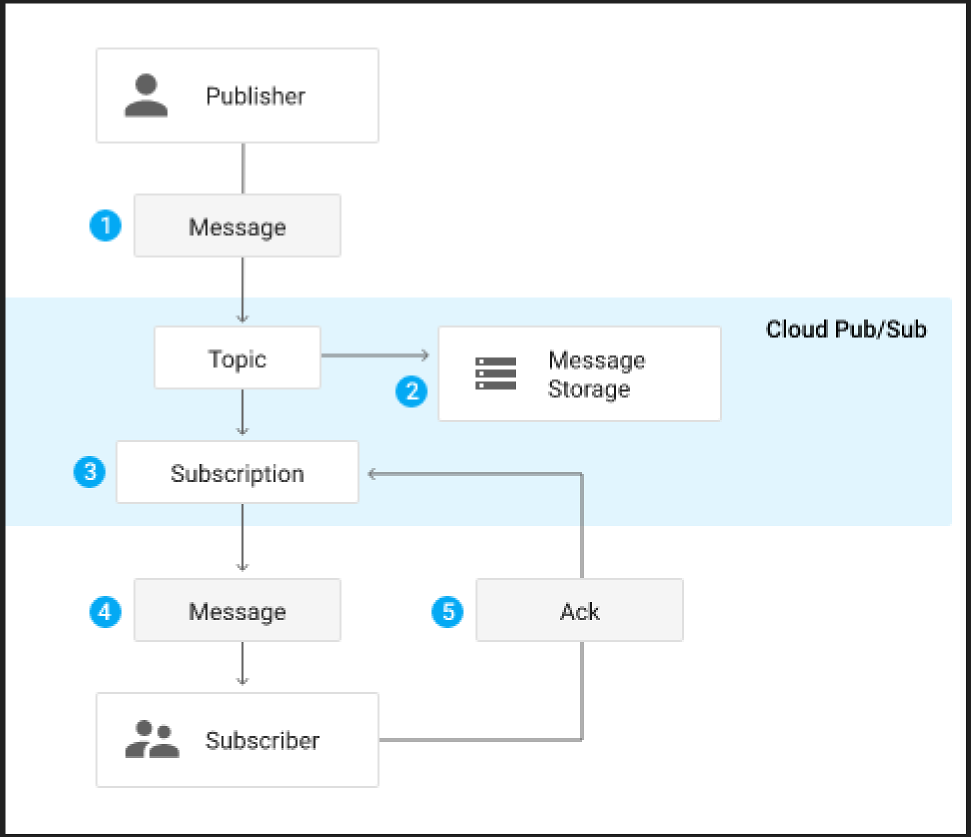
- Fully-managed asynchronous messaging service
- Scale to billions of message per day
- Publisher – App send message to Topic
- Push & Pull way to access messages
  - Pull – Subscriber pull message
  - Push – Message will be sent to subscriber via webhook
- One topic – Multiple Subscriber
- One subscriber – Multiple Topic

https://cloud.google.com/pubsub/docs/overview



Cloud PubSub

- Fully-managed Pubsub system inside Google Cloud
- Serverless
- Auto-scaling and auto-provisioning with support from zero to hundreds of GB/second
- Topic – Storage reference
- Publisher send message to topic at pubsub.googleapis.com
- Push – Pull way to access message
- Once subscriber receive message ack is sent.
- Cloud Pubsub act as staging environment for many GCP services



Advantage PubSub

- Durability of data will increase
- Highly Scalable, Scalable
- Decoupling between both system (Publisher & Subscriber)
  - Application don’t synchronously communicate with Notification service
  - Application (Publisher) is not dependent on Notification service (Subscriber)

[Hands-on] Cloud PubSub

Cloud DataFlow

- Managed service for variety of data processing
- An advanced unified programming model to implement batch and streaming data processing jobs that run on various execution engine/ runner
- Cloud version of Apache Beam = (Batch + Stream)
- Serverless, Fully managed
- Horizontal autoscaling of worker
- Jobs created with
  - Pre-define template
  - Notebook instance
    - Write Data Pipeline job in Java, Python, SQL
  - From Cloud Shell/Local Machine

How DataFlow Works

- Write Job in Java, Python Go
- Unified API for both batch + stream Processing
  - No Need to separately handle Batch & streaming data
- Execution
  - Direct Runner
    - Scaling issue
  - Apache Flink
  - Apache Spark
  - Cloud DataFlow

Apache Beam

- Pipeline
  - A pipeline is a graph of transformations that a user constructs that defines the data processing they want to do.
- IO-Transform
  - https://beam.apache.org/documentation/io/built-in/
- Pcollection
  - Fundamental data type in Beam
- Ptransform
  - The operations executed within a pipeline
  - https://beam.apache.org/documentation/programming-guide/#transforms
- Runner - Execution engine

[Hands-on] Cloud DataFlow

PRE-DEFINE TEMPLATE

[redacted]

[Hands-on] Cloud DataFlow

NOTEBOOK INSTANCE

[Hands-on] Cloud DataFlow

EXECUTE JOB FROM SHELL WITH DATAFLOW

Cloud DataProc

- Managed Hadoop & Spark Services inside GCP
- Lift/Shift Existing Hadoop/Spark based Job
- Cluster type
  - Standard (1 master, N workers)
  - Single Node (1 master, 0 workers)
  - High Availability (3 masters, N workers)
- Worker node regular VM or Preemptible VM (Cost reduction)
- Job Supported :
  - Hadoop, SparkR, Spark, SparkSQL, Hive, Pig, PySpark
- Demo
  - Spark, PySpark, Notebook Instance

[Hands-on]
Create Cloud DataProc Cluster

[Hands-on] Cloud DataProc

SPARK JOB

[redacted]

[Hands-on] Cloud DataProc

SUBMIT PYSPARK JOB

[Hands-on] Cloud DataProc

NOTEBOOK INSTANCE

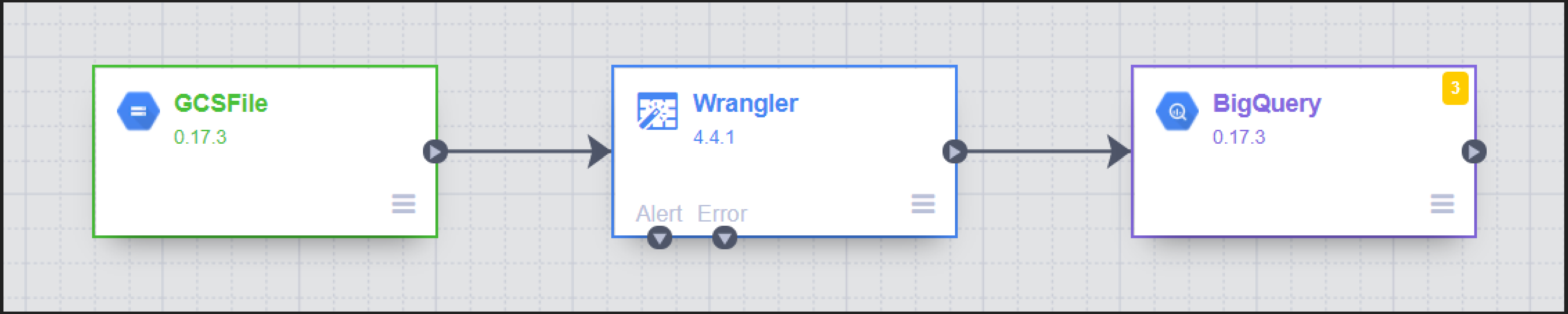
Cloud Fusion

[redacted]

Cloud Data Fusion

- Fully-managed, cloud native solution to quickly building data pipelines
- Code free, Drag-n-drop tool
- 150+ preconfigured connectors & transformations
- Built with Open-source CDAP
- 3 Edition are available
  - Developer
  - Basic
  - Enterprise
- Pricing :
  - https://cloud.google.com/data-fusion/pricing#cloud-data-fusion-pricing
- Let’s see in Action – create Cloud Fusion Instance

Cloud Data Fusion Demo



Cloud Composer

Cloud Composer

- Fully Managed Apache Airflow which in GCP
- Airflow is a workflow & orchestration engine
- With Airflow, one can programmatically schedule and monitor workflows
- Workflows are defined as directed acyclic graphs (DAGs)
- DAGs are written in Python 3.x
- Built-in integration for Other GCP services
  - Google BigQuery,
  - Cloud Dataflow & Dataproc,
  - Cloud Datastore
  - Cloud Storage,
  - Cloud Pub/Sub, and Cloud ML Engine

[Hands-on]
Create Cloud Composer Instance

Write first DAG in Cloud Composer

Data Loss Prevention API

[redacted]

[redacted]

Data Loss Prevention API

- Fully managed service designed to help you discover, classify, and protect your most sensitive data.
- PII data
  - Person’s name, Credit Card Number, SSN
- Apply API on Cloud Storage, Big Query Data
- DLP work upon Free form Text, Structured & Unstructured data (image)
- What to do with this Data
  - Identify sensitive data
  - De-identify data
    - Masking and Encryption
  - re-identify (In case want to recover original data)

De-Identification of Data

- Redacation – remove sensitive data
- Replacement – replace with some tokens (Like Info_type)
- Masking – Replace one/more character with some other char
- Encryption – Encrypt Sensitive Data

TEMPLETES, INFOTYPES &

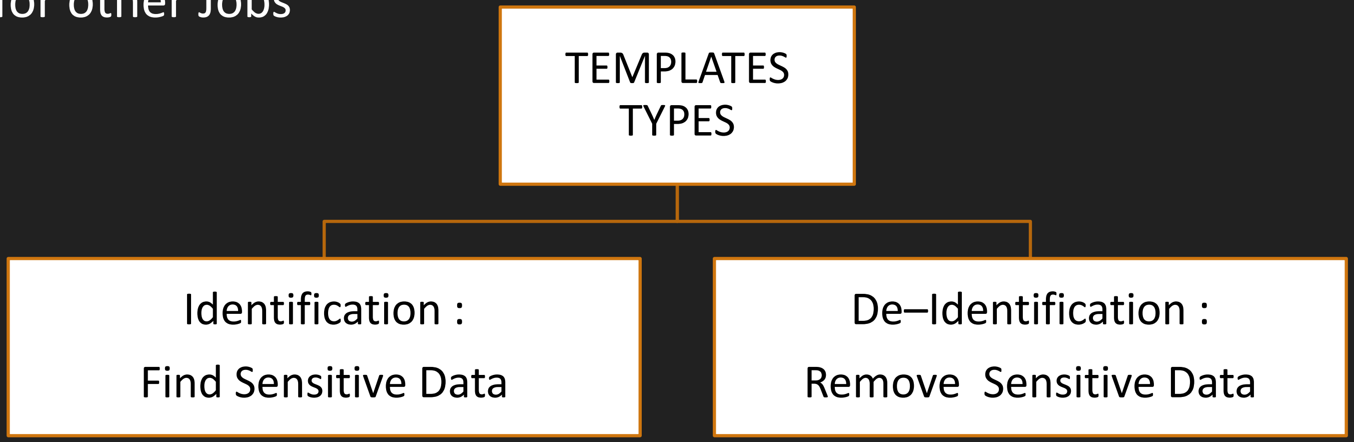
MATCH LIKELIHOOD

[redacted]

[redacted]

TEMPLATES

- Configuration which define for
  - Inspection of Jobs
  - De-identification of Jobs
- Once Template defined , can be reused for other Jobs



TEMPLATES TYPES

Identification :

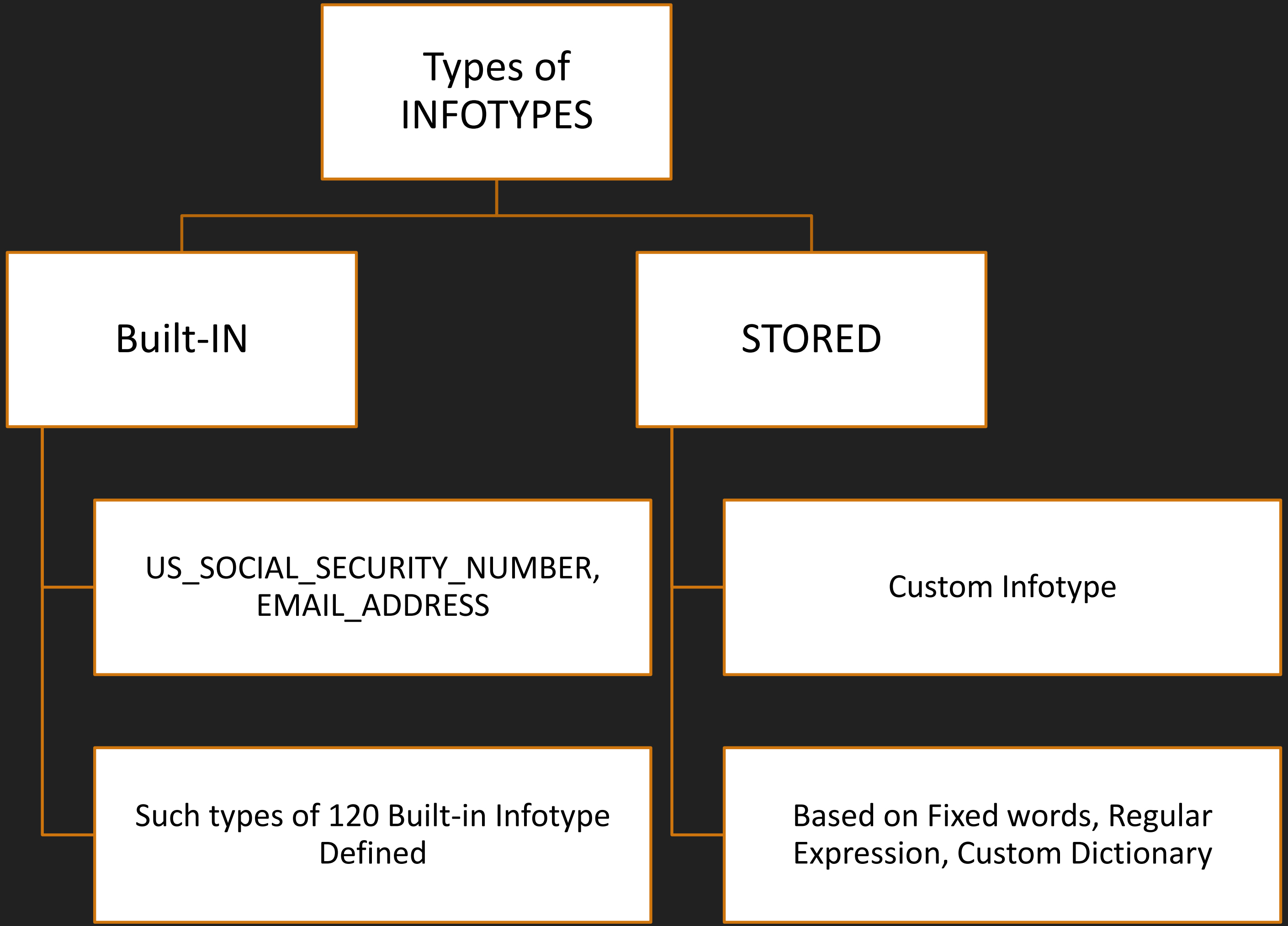
Find Sensitive Data

De–Identification :

Remove Sensitive Data

INFOTYPES

- What to Scan For
  - Like Credit Card
  - SSN
  - Age



Types of INFOTYPES

Built-IN

- US_SOCIAL_SECURITY_NUMBER, EMAIL_ADDRESS
- Such types of 120 Built-in Infotype Defined

STORED

- Custom Infotype
- Based on Fixed words, Regular Expression, Custom Dictionary

MATCH LIKELIHOOD

LIKELIHOOD_UNSPECIFIED	Default value; same as POSSIBLE.
VERY_UNLIKELY	It is very unlikely that the data matches the given InfoType.
UNLIKELY	It is unlikely that the data matches the given InfoType.
POSSIBLE	It is possible that the data matches the given InfoType.
LIKELY	It is likely that the data matches the given InfoType.
VERY_LIKELY	It is very likely that the data matches the given InfoType.

DLP API Demo

https://cloud.google.com/dlp/demo/#!/

[redacted]

[redacted]

Create INFO_TYPE

(Hands-on)

[redacted]

[redacted]

Create TEMPLETES

(Hands-on)

[redacted]

[redacted]

Create Job for Inspection (Hands-on)

[redacted]

[redacted]

Create Template for De-identification

[redacted]

[redacted]

Apply Some more rules to template

[redacted]

[redacted]

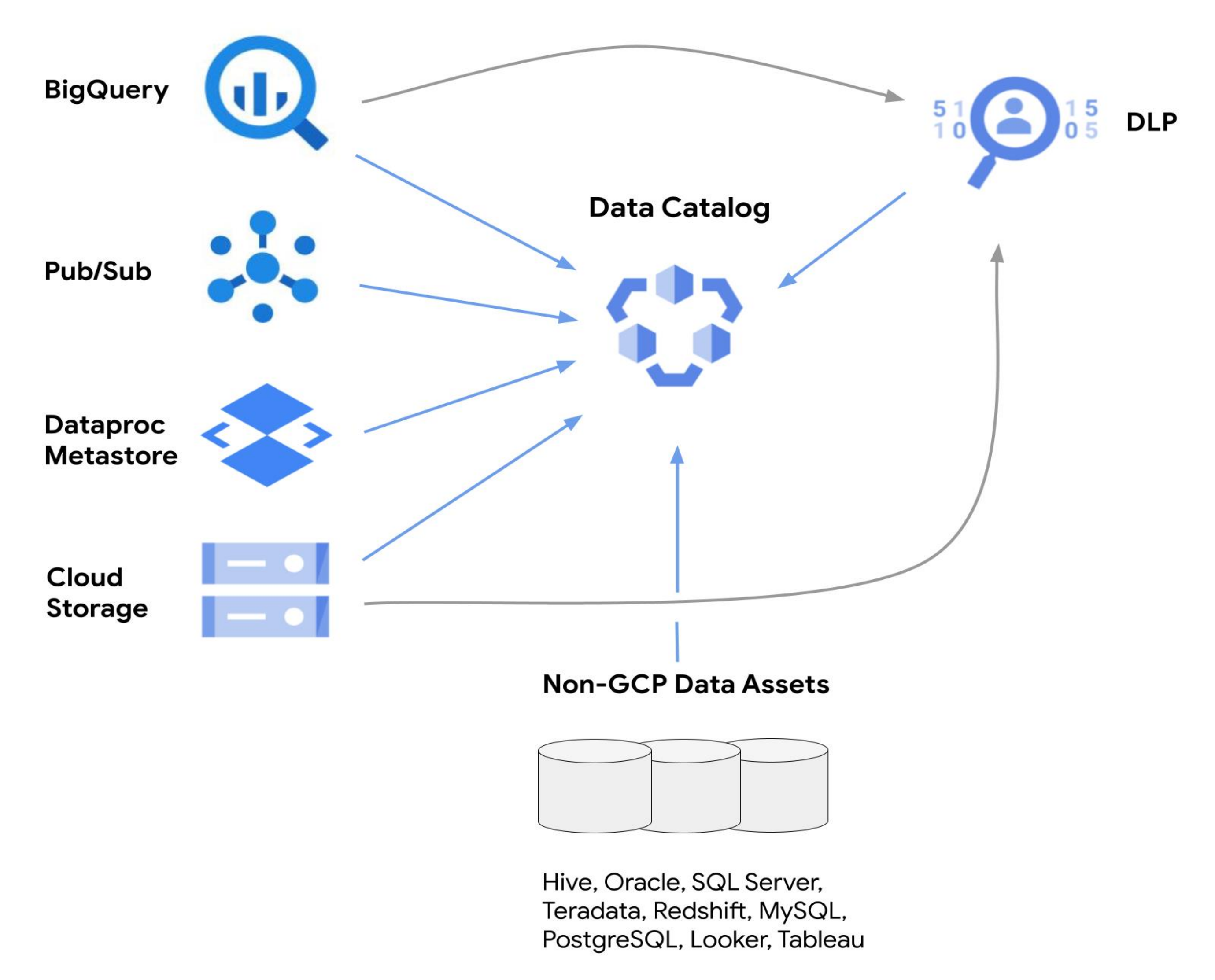
Data Catalog

[redacted]

Data Catalog

- Most organizations today are dealing with a large and growing number of data assets.
- Data stakeholders (consumers, producers, and administrators) face a number of challenges:
  - Searching for insightful data
  - Understanding data
  - Making data useful
- Data Catalog
  - A fully managed and highly scalable data discovery and metadata management service.
  - Single place to discover all data, asset across all project
- Using Data catalog
  - Search data assets
  - tag data

How Data Catalog works



Metadata

- Technical Metadata
  - For BigQuery, Pubsub these metadata resides inside individual products
  - Technical meta data being registered by catalog automatically
- Business Metadata
  - Attach Tag to existing data asset
  - Define some Tag template and attach metadata

[Hands-on] Data Catalog

ML/AI Module Introduction

Machine Learning

[redacted]

Machine Learning – GCP

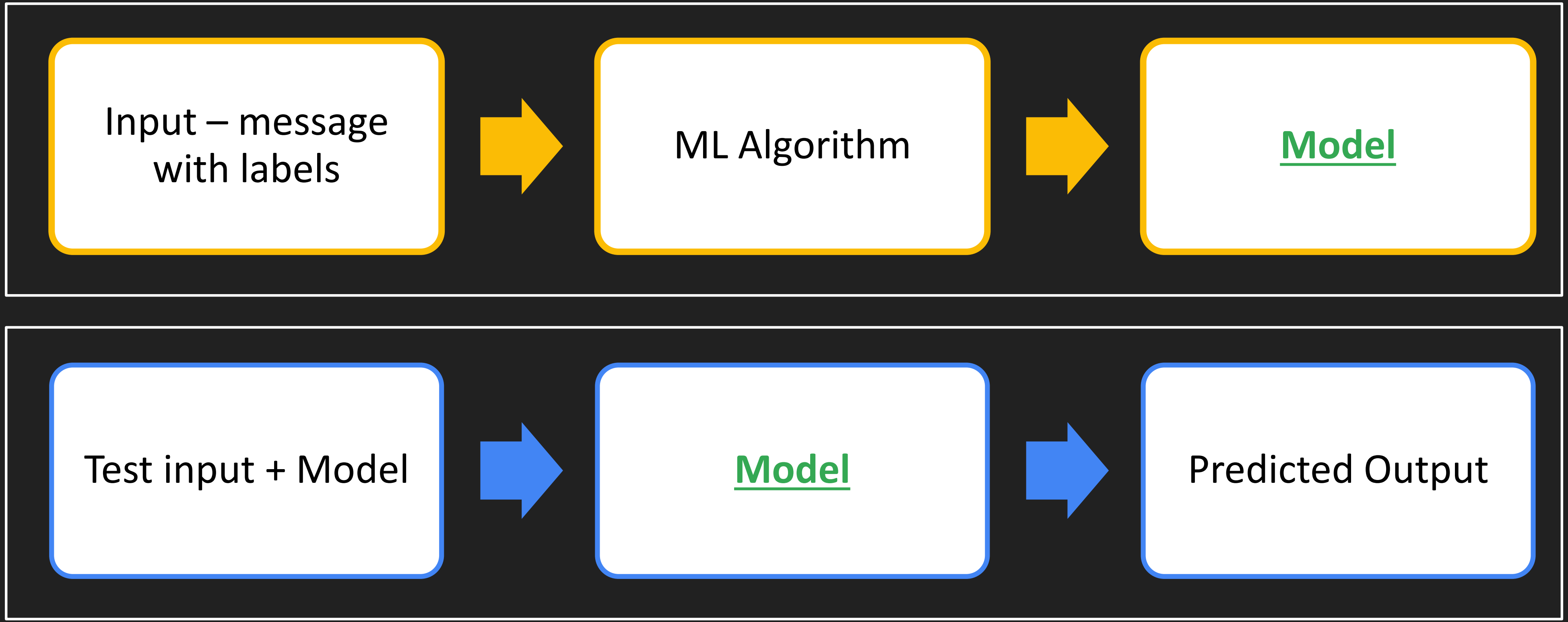
- Concept behind Machine Learning
- Types of ML System
- Pre trained Model
- Custom Model
- TPU – tensor processing unit

Machine Learning

- Design Spam email classification system
- How to design?
- What rules you will code inside system
  - If message coming from some specified list of senders, spam it
  - If message contain word like lottery, promotion, spam it
- But how many such rule you will define inside system.
- It is very difficult & cumbersome task to design such way.
- If spammer start sending spam which is not part of rule book.
- So, need some intelligent approach,
- Machine Learning is the solution behind it.

Machine Learning

- Rather than define such rule,
- In machine learning, system learn from data
- Training + Testing kind of system



Input – message with labels → ML Algorithm → Model

Training Process

Test input + Model → Model → Predicted Output

Testing Process

Types of ML System

- ML Types
- Supervised learning
  - Label has been given
  - Regression
  - Classification
- Unsupervised learning
  - No labels
  - Find Structure within data

[redacted]

Regression

- Output prediction is continuous in nature
- Example
  - House Price prediction
- Regression ML Algorithm :
  - Linear Regression
  - SVR
  - Decision Tree Regressor

Area	No of Bedroom	Price
5434	5	3536
2342	5	3564
243	1	4564
987	4	7675

Classification

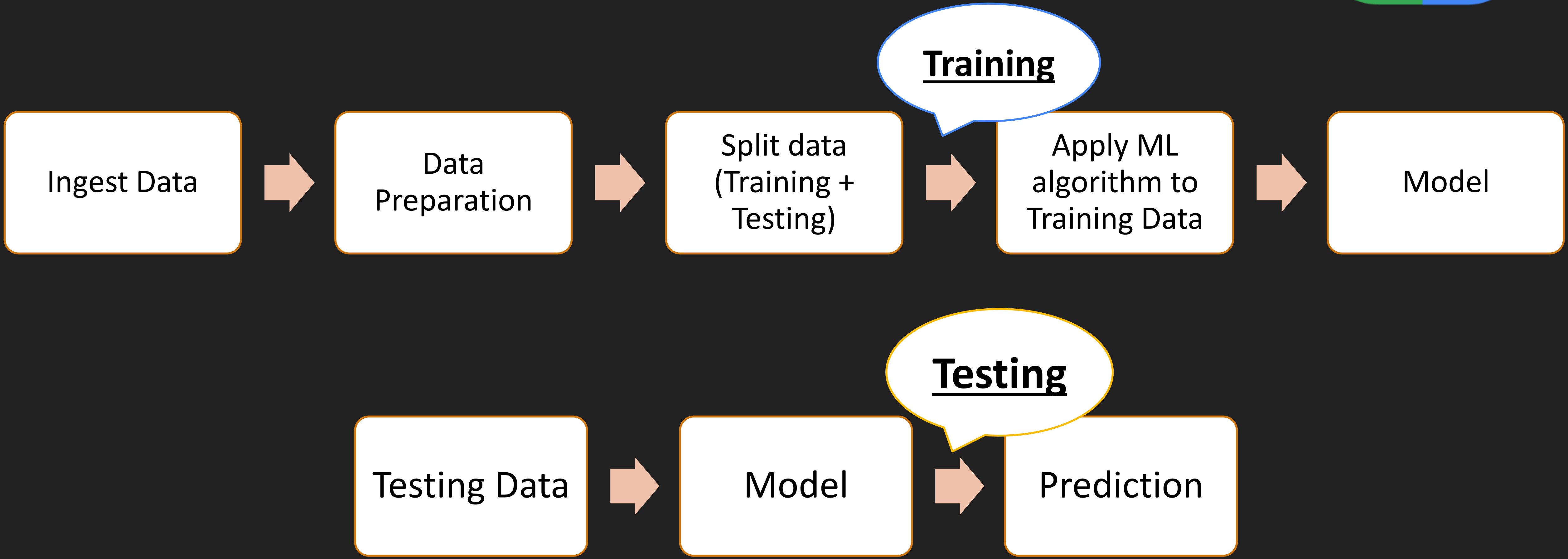
- Output prediction is discrete in nature
- Example
  - Sentiment analysis of review : +ve/-ve
    - This product is very much helpful. +ve
  - Is it Orange?
    - Yes/No

- Classification Algorithm :
  - Logistic Regression
  - SVM
  - KNN
  - Decision Tree Classification

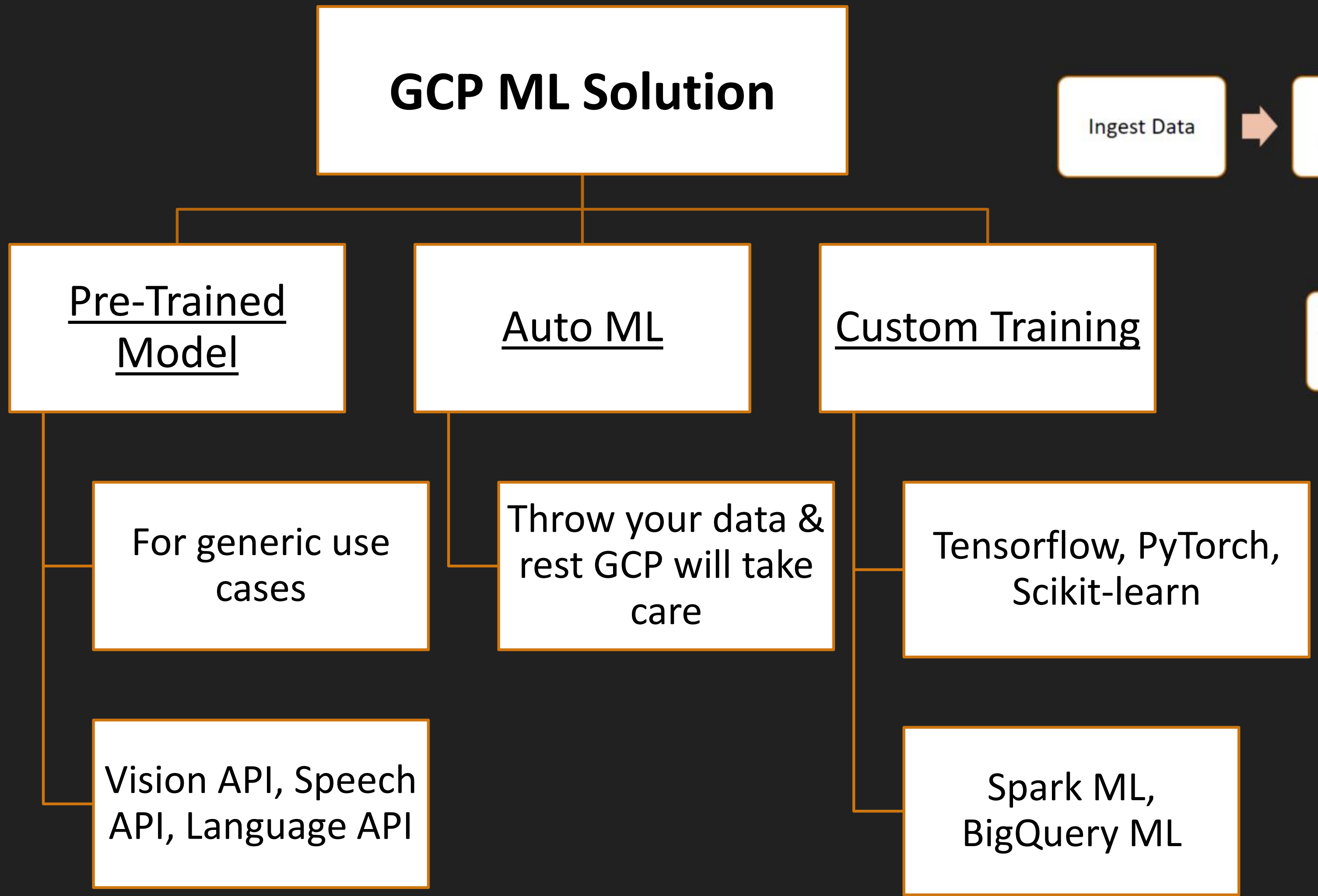
Unsupervised Learning

- No label Given
- Find Structure within data
- Clustering is type of Unsupervised Learning
- Some clustering Algorithm :
  - K-Means
  - hierarchical

Machine Learning workflow



Machine Learning + GCP



GCP ML Solution

Pre-Trained Model

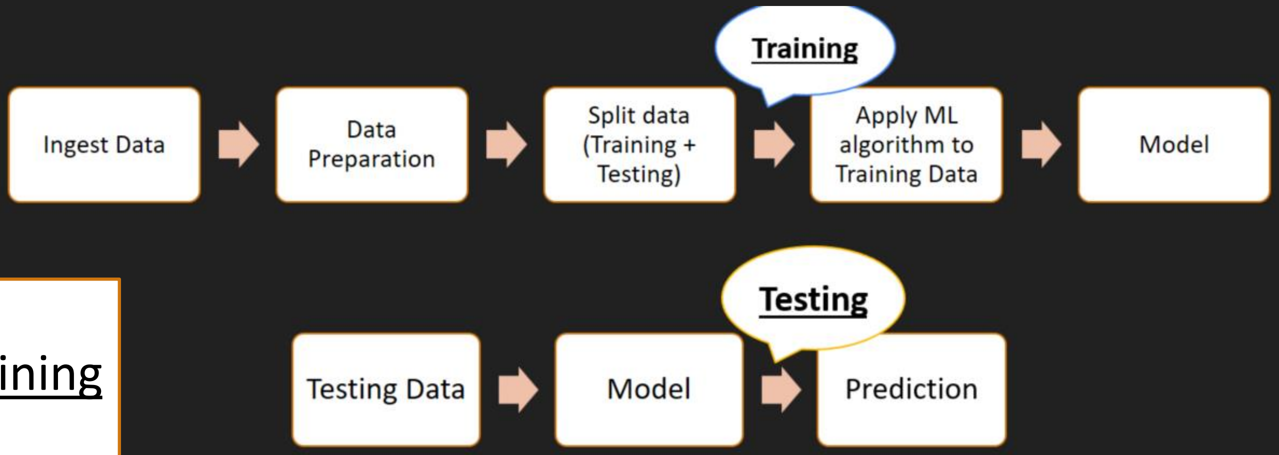
- For generic use cases
- Vision API, Speech API, Language API

Auto ML

- Throw your data & rest GCP will take care

Custom Training

- Tensorflow, PyTorch, Scikit-learn
- Spark ML, BigQuery ML



Training

Ingest Data → Data Preparation → Split data (Training + Testing) → Apply ML algorithm to Training Data → Model

Testing

Testing Data → Model → Prediction

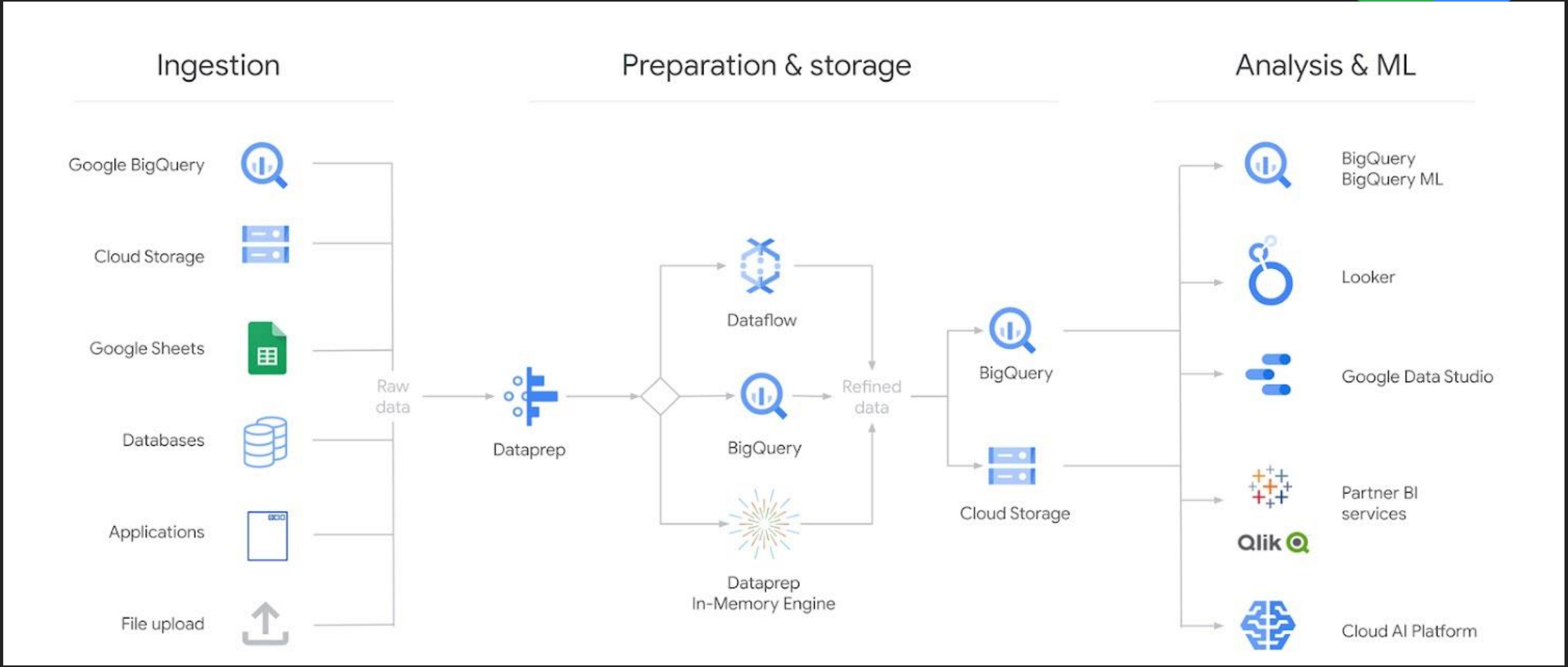
Data Preparation with DataPrep

[redacted]

DataPrep

- Intelligent Data Preparation tool
- Visually explore, clean, and prepare data for analysis and machine learning
- Build by Trifacta – Third Party tool, not cloud native one
- Play with this tool without any code, with just click
- Dataprep is serverless and works at any scale
- No infrastructure to deploy or manage
- Automatically detect schema, anomalies
- Do all time consuming task easily
- Concern – Need to share data with Trifacta

DataPrep ETL Pipeline



Let’s see DataPrep in Action

Pre-trained Model

Pre-Trained Model

- Google has huge amount of data
- Google has already trained ML/AI algorithm to build model
- For generic use case like
  - Object recognition/detection – Vision API
  - OCR
  - Speech to Text
  - Language Translation
  - NLP API – to get insight from natural language
- You can take advantage of pre-built model.
- No Training required from customer
- Use already built Rest API for above use cases

Vision API

Natural Language API

Speech to Text API

Text to Speech API

Vision API

- Derive insights from your images
  - Detect printed and handwritten text
  - Detect objects
  - Identify popular places and product logos
  - Moderate content
  - Celebrity recognition
- How to use
  - Web UI – Just for Testing
  - https://cloud.google.com/vision#section-2
- gcloud – CLI
- Python SDK

Natural Language API

- Derive insights from unstructured text using Google machine learning
  - Identify entities within documents
  - Sentiment analysis
  - Content classification
- How to use
  - gcloud – CLI
  - Python SDK

Speech ⇔ Text API

- Speech to Text API
  - Accurately convert speech into text using an API powered by Google’s AI technologies.
  - 125 languages support
  - Streaming speech recognition
  - Content filtering
  - Automatic punctuation
  - https://cloud.google.com/speech-to-text#section-2
- Text to Speech API
  - Convert text into natural-sounding speech using an API powered by Google’s AI technologies.
  - https://cloud.google.com/text-to-speech#section-2

ML API Pricing

[redacted]

Auto Machine Learning

Auto Machine Learning

- AutoML – Auto Machine Learning
- Your use case is not generic
- You have some custom requirement
- Vision API – recognize shoes
- Auto ML – Different types of shoes detection
  - Adidas, Nike shoes
- Throw your data & GCP will create best model for you
- State-of art Transfer learning technology
- Throw your data & Google AI will create model
- 2 use cases Demo :
  - Flower species recognition
  - Text Classification

Text Classification

- Dataset creation

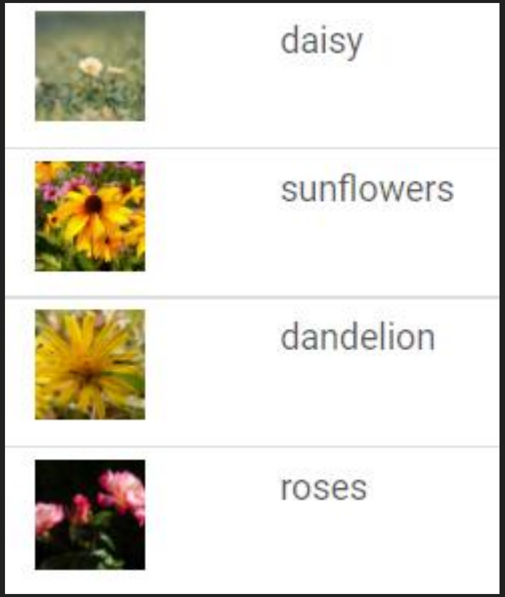
Data	Class
My eldest son who is 27 just got word he has a new job after finishing his bachelors degree. This made me very happy!	achievement
I visited my best friend at her school on St. Patrick's day.	bonding
My mom cooked some delicious rice for me with curd.	affection
Today I make Eye contact with my crush. She Also look into my Eyes For a Seconds or Two. I can still Memorize his Beautiful Eyes.	affection

achievement
Affection
bonding
enjoy_the_moment
exercise
leisure
nature

- Train Model
- Analyze Model
- Deploy for Prediction

Flower Classification

- Dataset creation



daisy

sunflowers

dandelion

roses

daisy
dandelion
roses
sunflowers
tulips

- Train Model
  - Define Node hours
- Analyze Model
- Deploy for Prediction

TPU

- TPU – Tensor Processing Unit
- Machine Learning Training is one of the most time consuming process
- It may take hours to days to sometime week
- Training time depend upon Ml Algorithm + Amount of dataset
- Google introduce Tensorflow framework to do Machine Learning which powers their own ML Product
- Tensor are basic building block of this framework.
- So, To do training faster Google created ASIC based in-house dedicated computing for Tensor Processing
- Speed up training by 20x to 30x
- Work with VM, GKE, AI Platform
- Quickly experiment with number of ML Models creation

Train your own model

Custom Model

- You have your own dataset
- You want to train your own model
- You have team of data scientist
- They want to write own algorithm based on
  - Scikit-learn, XGBoost
  - Tensorflow
  - PyTorch Framework
- Notebook instance
- Build Logistic Regression model for flower species recognition
- Save model in pickle file
- Deploy model endpoint

BigQuery ML

- Create ML Model in SQL
- No Python, No Java.
- No need to export data to other environment
- Model support
  - Linear Regression
  - Multiclass Logistic Regression
  - K-Means
  - XGBoost
  - Tensorflow – Import
- use case demo
  - Flower species recognition
  - From BigQuery public dataset

Main BigQuery Function

- Create MODEL
  - Model Type – Linear Reg, Logistic
  - Label Column
  - Learning Rate etc…
- Evaluate Model
  - ML.Evaluate
  - Provide Model & Test Data
  - Determine how good model performance on Test data
- Prediction
  - ML.Prediction
  - Apply Live data to Model to get prediction

[Hands-on] BigQuery ML

Cloud Data Studio

Data Studio

- BI tool from Google
- Connect your data from spreadsheets, Analytics, Google Ads, Google BigQuery and many more connectors
- Drag & drop, no code
- Create reports & Dashboards
- Free
- Let’s see in Action

THANK YOU