ROI TRAINING
MAXIMIZE YOUR TRAINING INVESTMENT

2018
Partner of the Year
Google Cloud

Welcome!

Google Cloud Certification Workshop—Data Engineer

ROI TRAINING
MAXIMIZE YOUR TRAINING INVESTMENT

The following course materials are copyright protected materials. They may not be reproduced or distributed and may only be used by students attending the *Google Cloud Certification Workshop—Data Engineer* course.

ROITRAINING
MAXIMIZE YOUR TRAINING INVESTMENT

Google Cloud
Premier Partner

- Google Cloud Global Training Partner of the Year – 2018 & 2017
- 50,000+ people trained in 30+ countries across 6 continents
- Largest team of Google Cloud Platform Authorized Instructors
- Helped more IT professionals achieve Google Cloud Platform certification than any other partner

google.roitraining.com

Course Versions

- This course is offered in several versions and durations
  - The delivery will vary depending on the duration
- For example:
  - Hands-on exercises are normally provided for experience and as extra practice to be done outside of class hours
  - Some versions of the class will perform the exercises during class hours

ROI TRAINING
MAXIMIZE YOUR TRAINING INVESTMENT

Google Cloud Certification Workshop—Data Engineer

Introduction

Course Objectives

In this course, we will learn how to:

- Prepare for the GCP Data Engineer certification exam
- Choose the appropriate GCP data storage solution
- Architect batch and streaming data processing pipelines on GCP
- Leverage GCP tools for data manipulation, analysis, and visualization
- Build machine learning models with GCP tools

Prerequisites for the Exam

- The exam tests one's understanding of how to design, build, and maintain secure, reliable data processing systems on Google Cloud Platform
- Experience with Google Cloud Platform at the level of course Data Engineering on Google Cloud Platform
- Real-world experience using Google Cloud Platform

The Course Workbook

- The course materials are written as a workbook
- Make a copy by:
  - Going to the URL: http://tinyurl.com/yxg9d3cs
  - In Google slides, choose **File | Make a Copy**
- This allows you to edit the course materials and do the activities provided

ROI TRAINING
MAXIMIZE YOUR TRAINING INVESTMENT

Google Cloud Certification Workshop—Data Engineer

Chapter 1:
Data Engineer Certification Overview

Chapter Objectives

In this chapter, we will:

- Review the certification exam format and guidelines

Chapter Concepts

Exam Overview

Exam Prep

What Is a Data Engineer's Job?

- As a class, let's come up with a list of things a data engineer would need to do or understand

What You Are Tested On

- You are ***not*** tested on trivia (*what is the maximum number of this or that*)
- You are ***not*** required to code or configure a service
- You are tested on your ability to:
  - Build storage solutions on Google Cloud Platform
  - Architect storage and processing systems using GCP services
  - Transform data for efficient analysis and machine learning
  - Analyze and visualize data using GCP tools
- Can you choose which services to use for various cases optimizing:
  - Cost, performance, security, fault-tolerance, etc.

Exam Format

- Multiple-choice questions
- You have 2 hours to take the exam
  - Time should not be a problem if you are prepared
- Grading is Pass/Fail

Taking the Exam

- Must register to take the exam
- Exam must be taken in-person at a Kryterion testing center location
- Bring 2 forms of ID
- Exam fee is $200

Chapter Concepts

Exam Overview

Exam Prep

Do Now: Understanding the Exam

- Take a few minutes to read the following pages:
  - https://cloud.google.com/certification/data-engineer
  - https://cloud.google.com/certification/faqs
  - https://cloud.google.com/certification/guides/data-engineer-2/

Google Cloud Certification Workshop—Data Engineer

Chapter 2: Google Storage Fundamentals

Chapter Objectives

In this chapter, we will learn how to:

- Choose the right storage solution based cost, availability, durability, and consistency
- Differentiate GCP's storage products by use case
- Architect data processing solutions using GCP storage services

Chapter Concepts

Storage Overview

GCP Storage Options

Architecting Data Processing Solutions

Exam Prep

Do Now: Types of Data

- Below, list different types of data you need to store

Do Now: Data Storage Considerations

- In addition to the type of data, list other considerations you should consider when choosing an appropriate storage option?

Choosing Data Storage Solutions

- The right storage solution depends on many factors
- What type of data are you storing?
  - Images, videos, relational, text, code, etc.
- How much data do you have?
  - How you store MBs of data is different than how you would store PBs
- What are the security requirements?
  - Public data distributed online
  - Sensitive personal information
- How will you process the data?
- Is it multi-user data?
- Etc.

Storage Cost

- Cost varies widely depending on the storage solution
  - Knowing how storage services are priced helps determine which to use
  - Need to know what storage requirements your application needs
- Characteristics of a storage service have an impact on price
  - Spanner is 99.999% available, but do you really need that?
  - BigTable supports huge amounts of data, but do you have much data?

Availability and Durability

- Availability is the percentage of time the data can be accessed
  - Achieved by deploying services to multiple zones and/or regions
- Durability defines the likelihood of losing data because of a hardware failure
  - Achieved by writing data to multiple physical disks
  - The more disks, the higher the durability

Do Now: Cost vs. Availability vs. Durability

- Fill in the table below:

Scenario	Use Cases
High availability is the most important consideration	
Cost and durability are more important than availability	
Cost is the only consideration	

Consistency

- ***Transactional consistency*** – When a transaction completes, all operations must be successful or all are rolled back
  - The data must conform to all rules specified by the database
  - The state of data is known by all nodes in a distributed system
- ***Eventual consistency*** – After data is updated, the system guarantees that all copies of the data in a distributed system will "eventually" be the same
  - It is possible that requests to different nodes will return different results after an update
- ***Strong consistency*** – All nodes in a distributed system that have received the same update will be in the same state once the transaction completes. All nodes would thus return the same result to a query.

Data Warehousing

- Combines data from multiple sources
  - Relational databases
  - Logs
  - Web data
  - Etc.
- Allows data from multiple sources to be combined and analyzed
- Historical archive of data

Chapter Concepts

Storage Overview

GCP Storage Options

Architecting Data Processing Solutions

Exam Prep

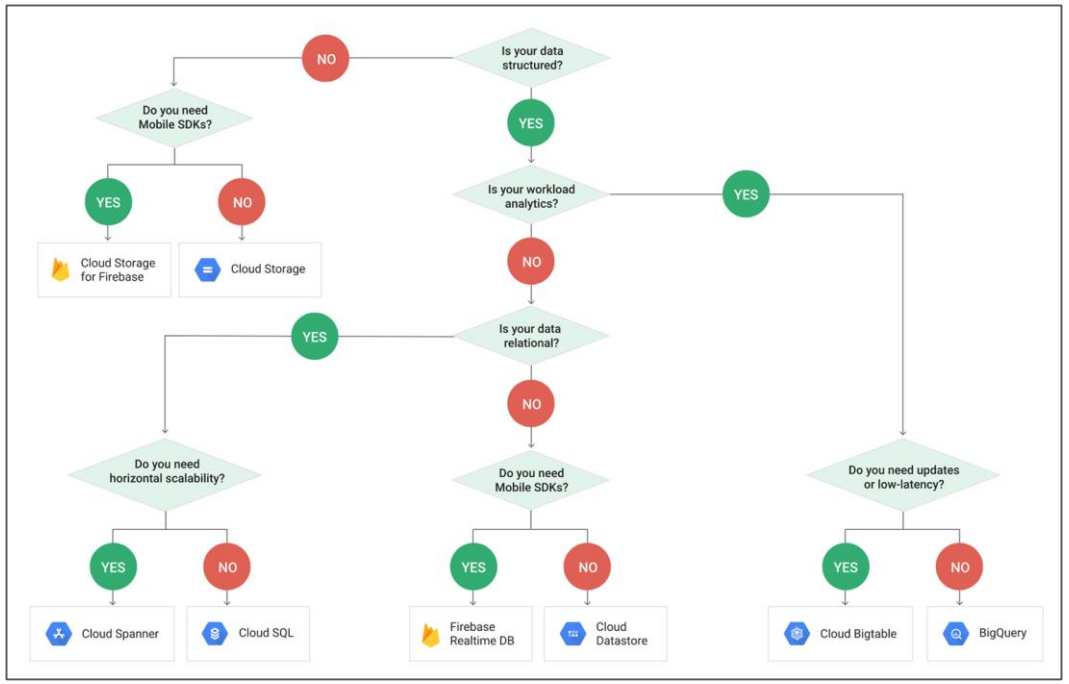
Google's Storage Service Choices

Choosing a Storage Service

- Choose your storage service based on your use case
- Balance, cost, availability, durability, consistency, scalability
- Know the characteristics of the different services

Do Now: Choosing the Right Storage Service

- Visit the Choosing a Storage Option link
  - Examine the flowchart on choosing a storage option
  - Read the table that describes the various storage options and their use cases



Chapter Concepts

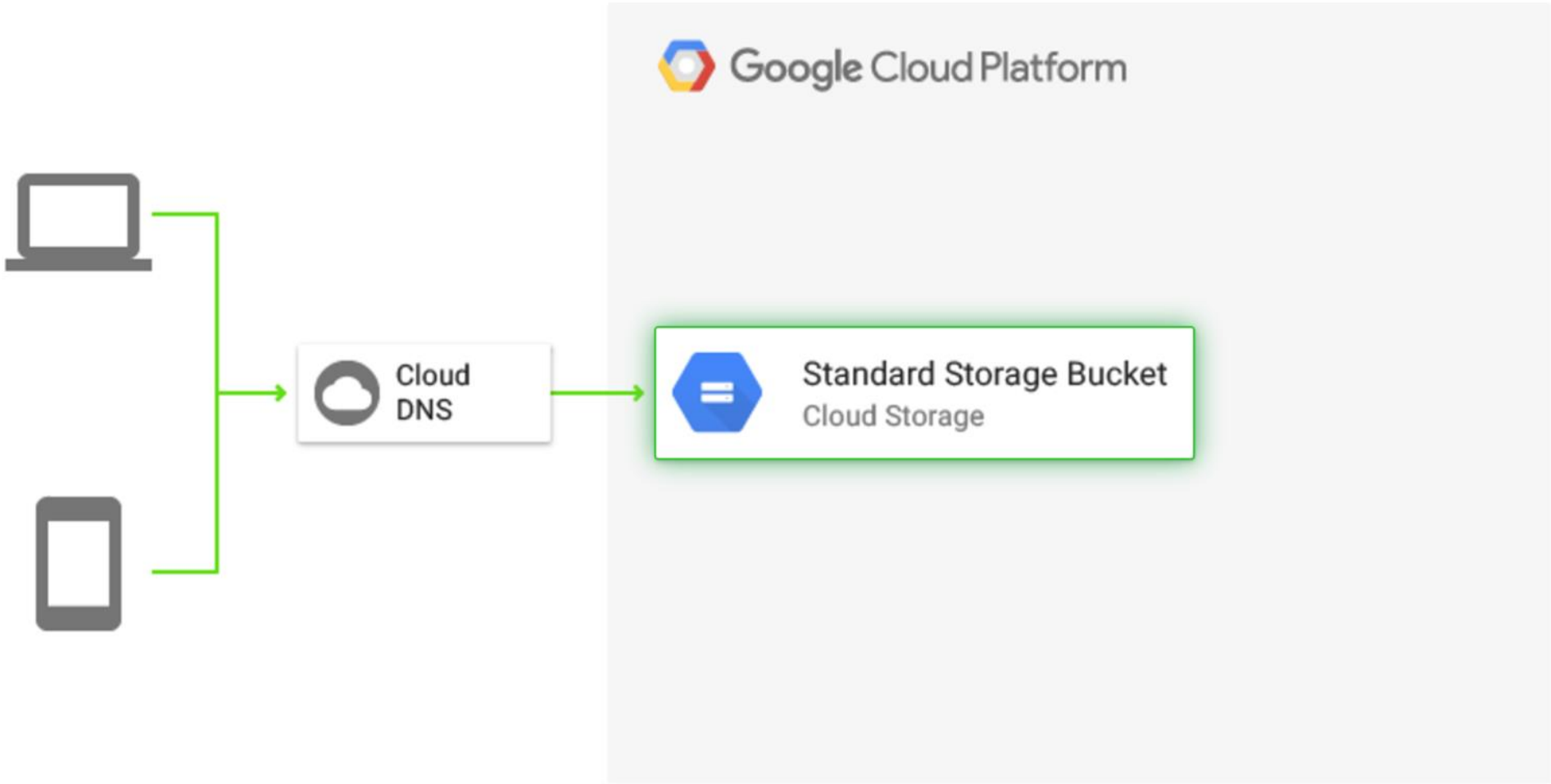
Storage Overview

GCP Storage Options

Architecting Data Processing Solutions

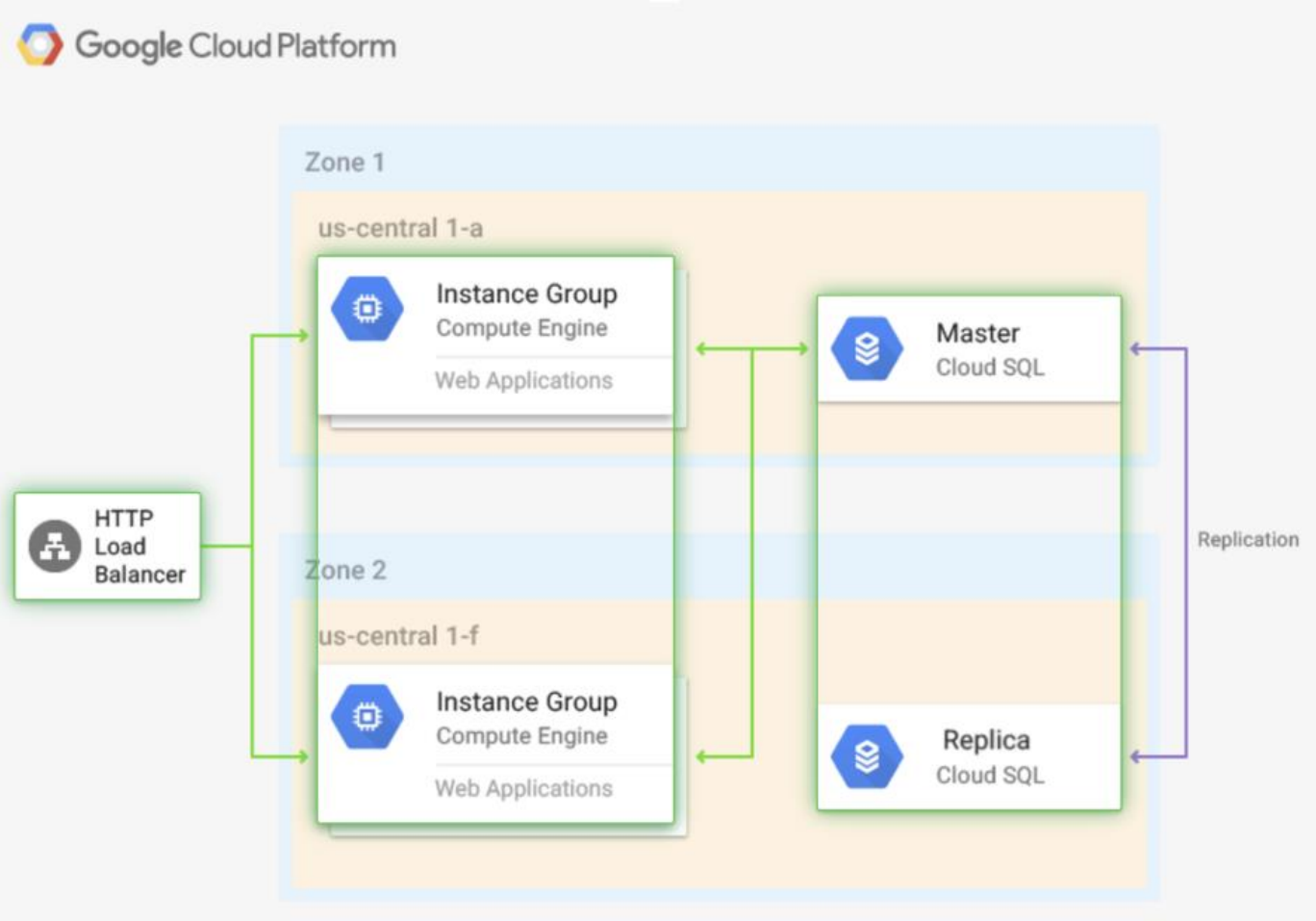
Exam Prep

Static Websites



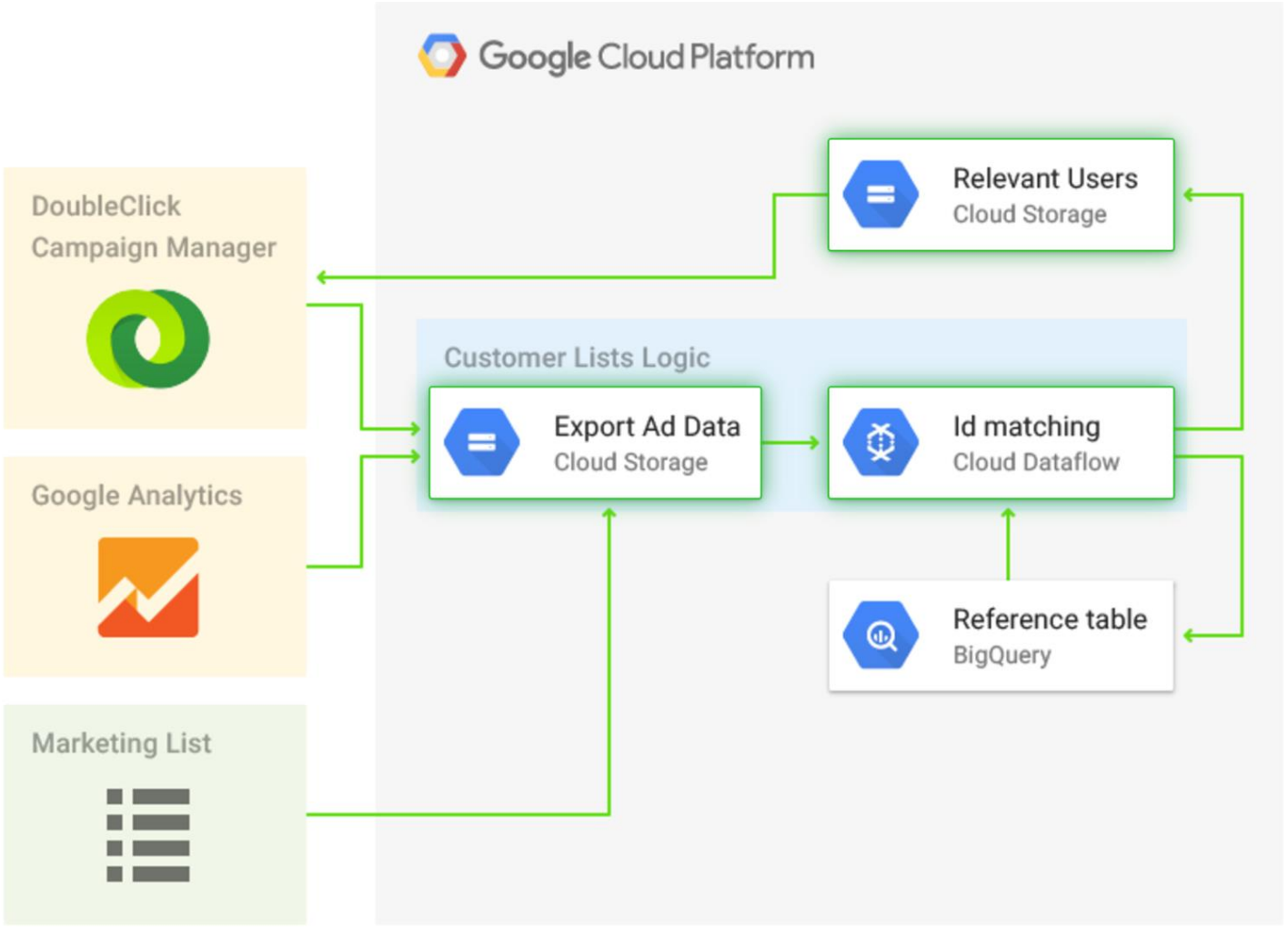
http://gcp.solutions/diagram/Static%20Hosting

Dynamic Web Applications



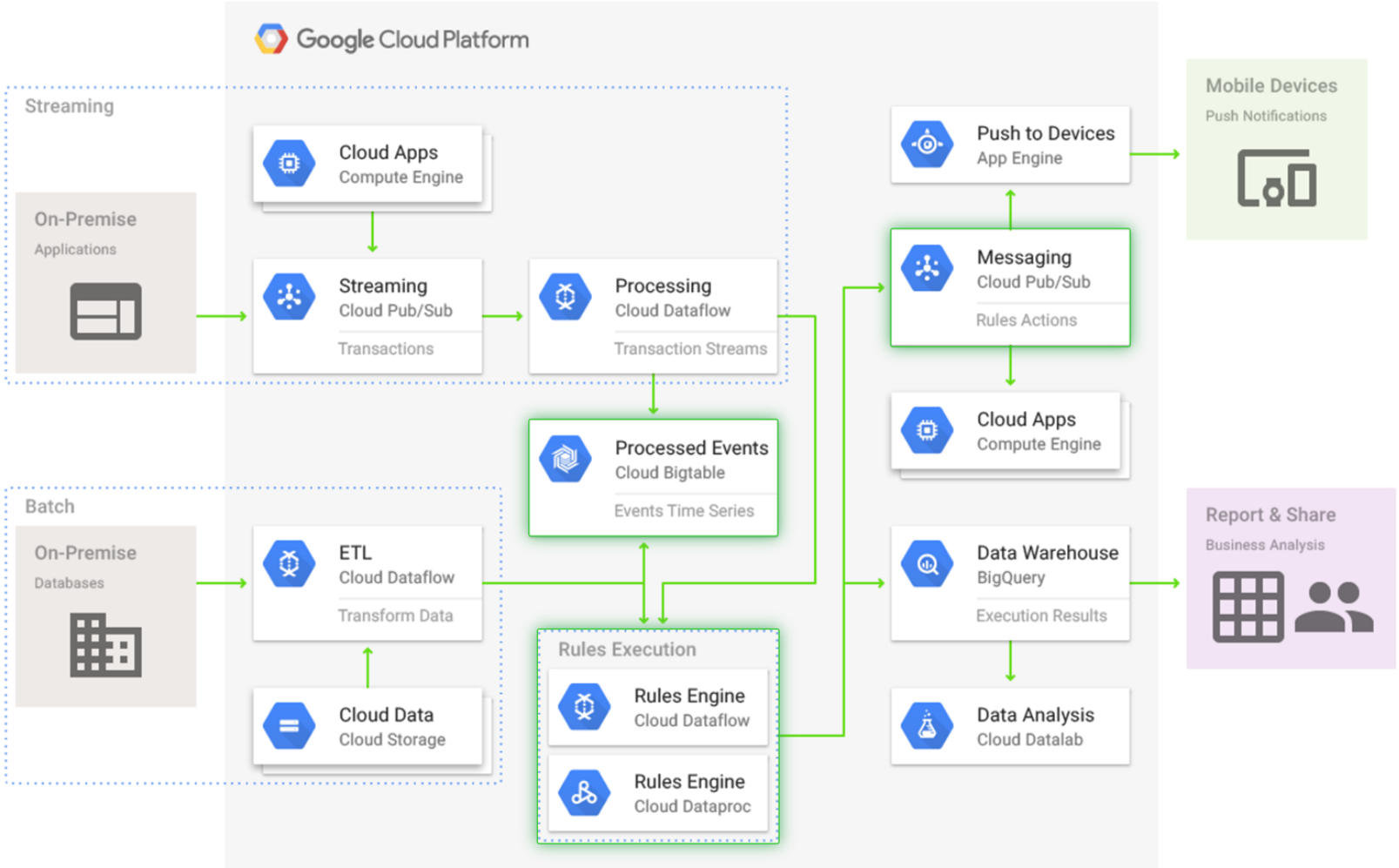
http://gcp.solutions/diagram/Dynamic%20Hosting

Data Warehousing



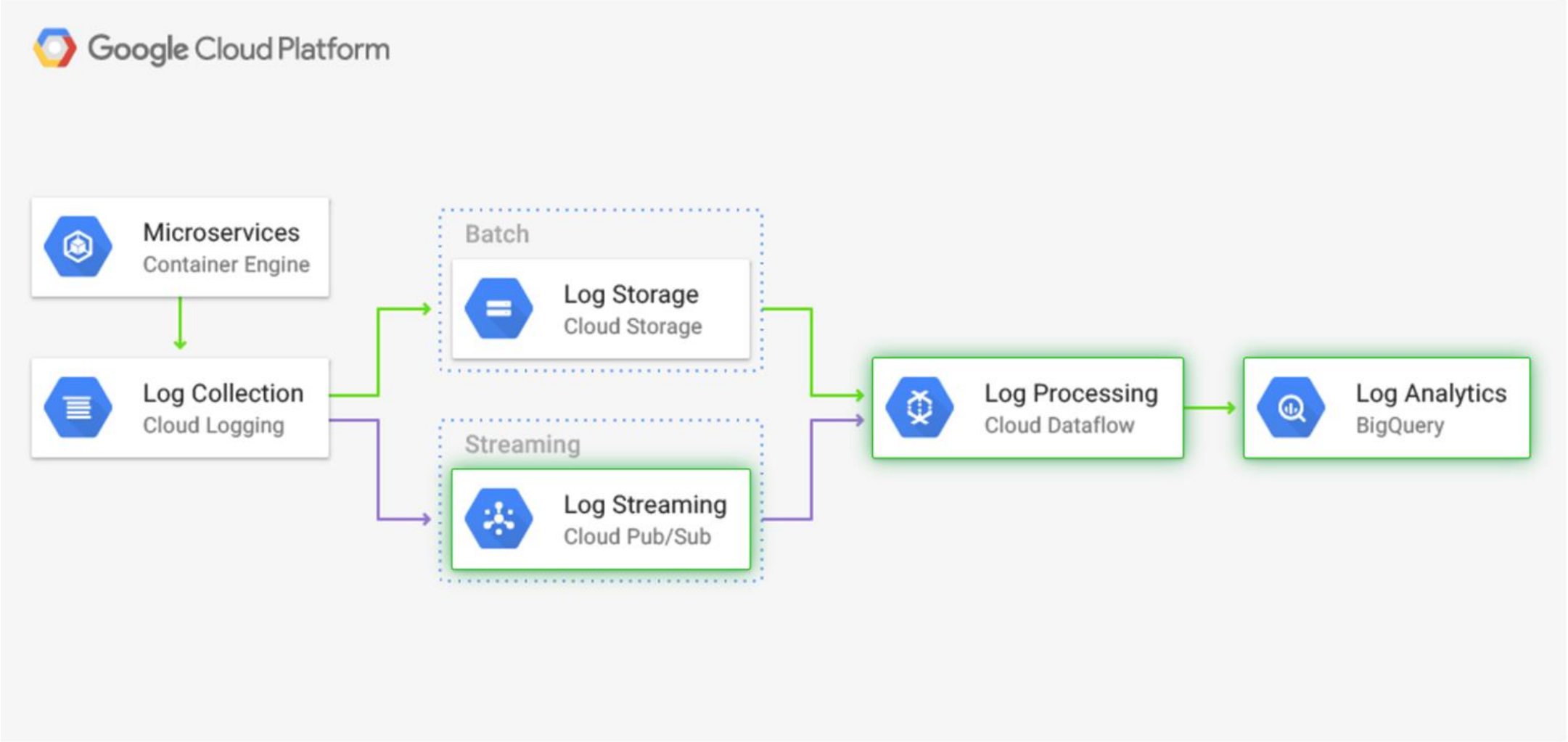
http://gcp.solutions/diagram/DMP%20%2F%20Data%20Warehouse

Event Processing



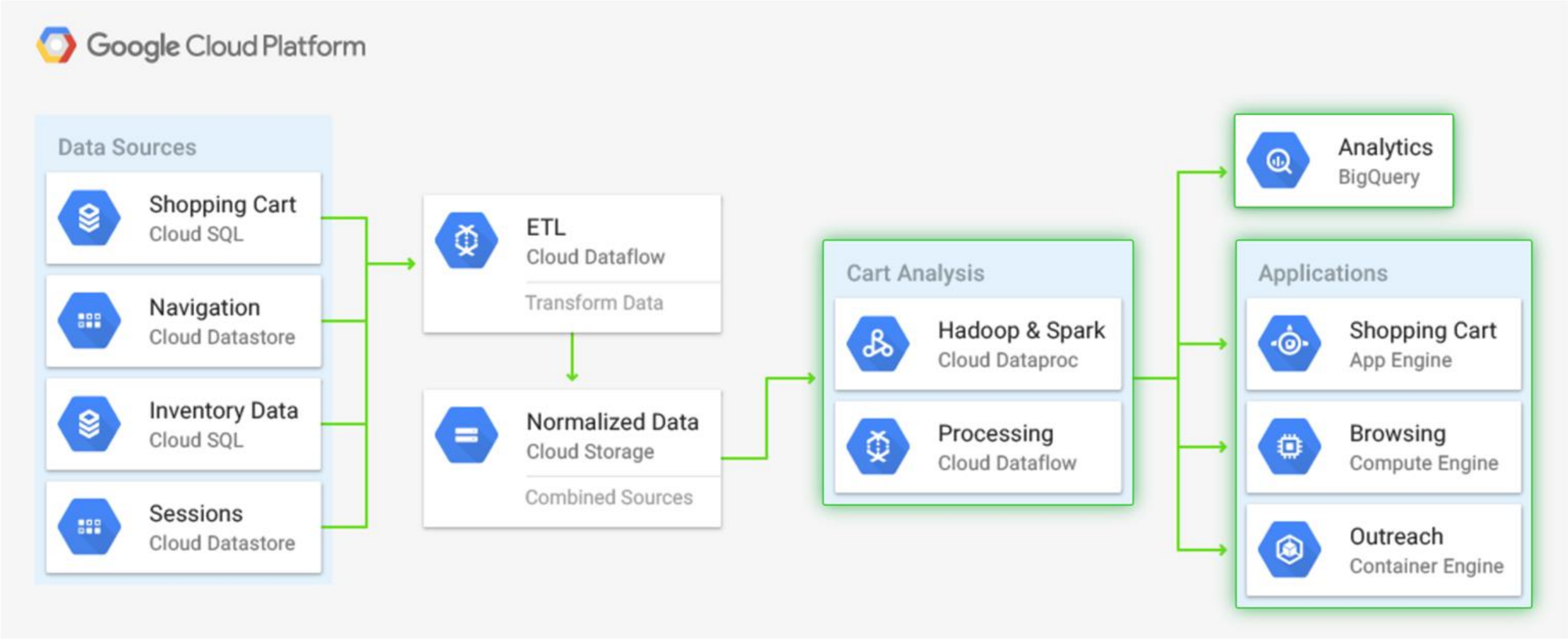
http://gcp.solutions/diagram/Complex%20Event%20Processing

Log Processing



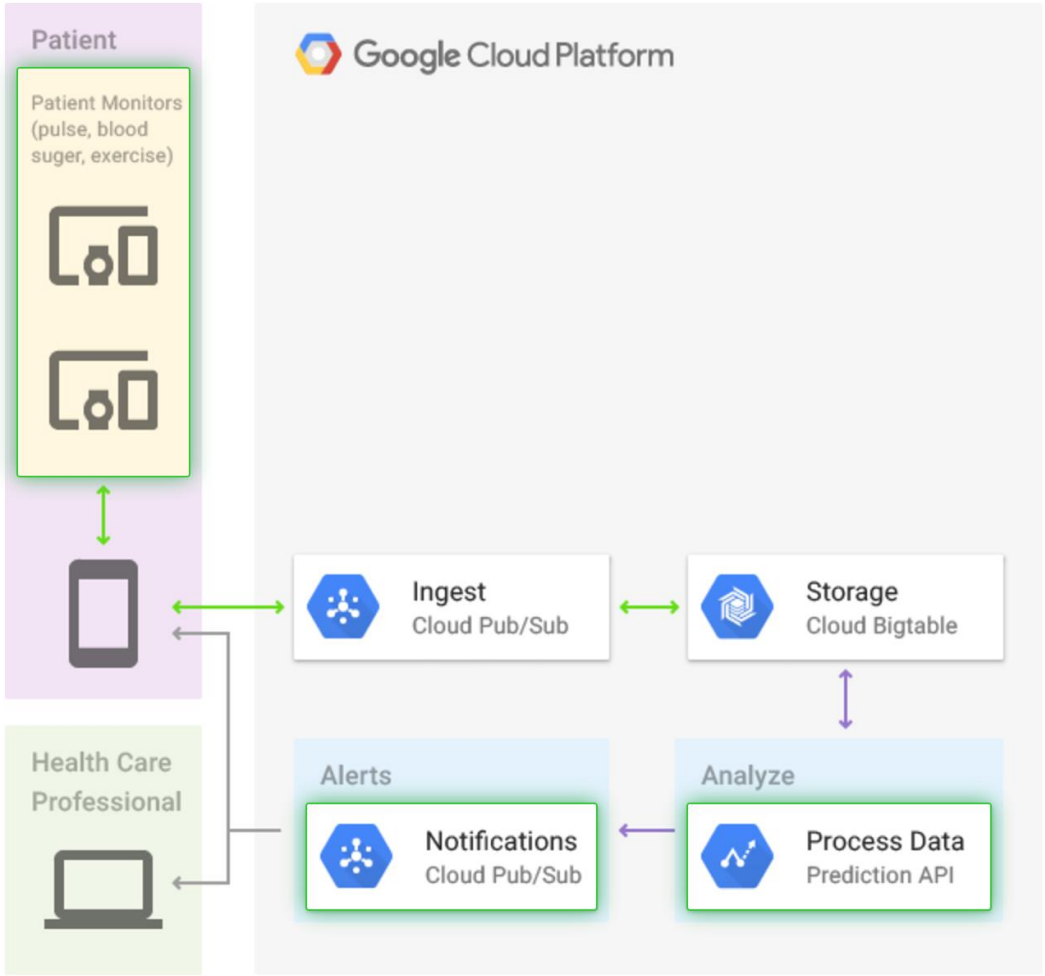
http://gcp.solutions/diagram/Log%20Processing

Shopping Cart Analysis



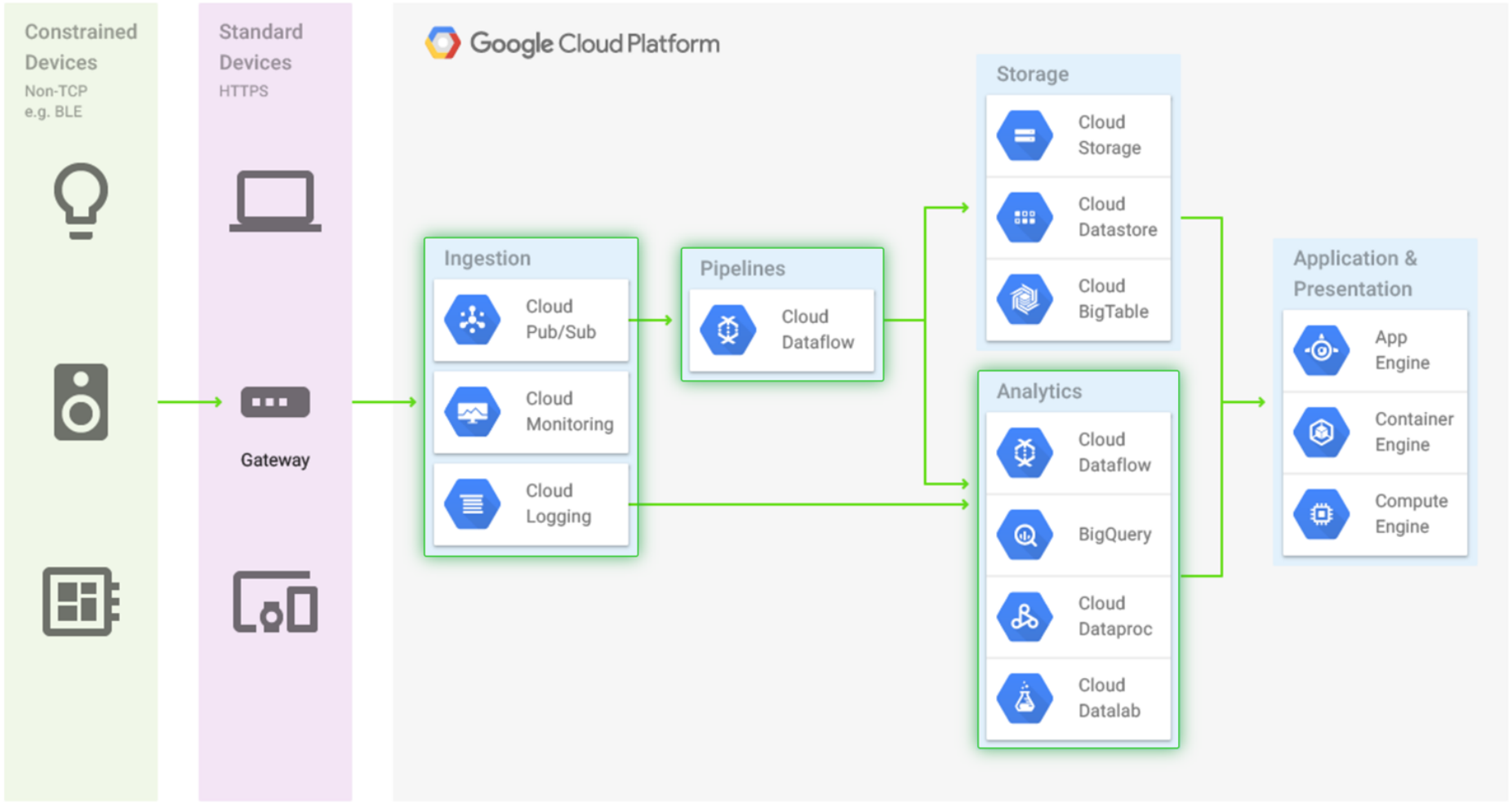
http://gcp.solutions/diagram/Shopping%20Cart%20Analysis

Patient Monitoring



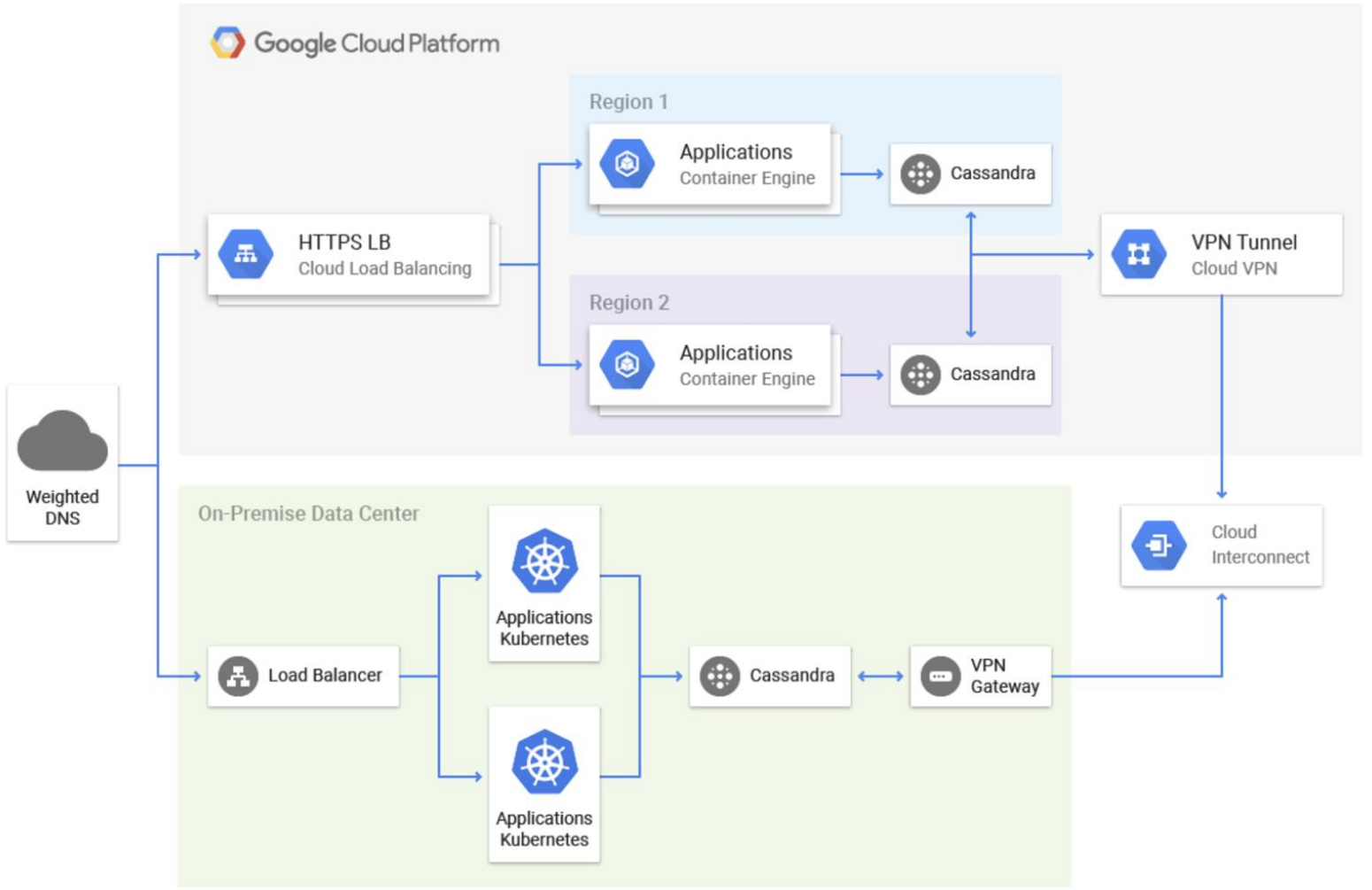
http://gcp.solutions/diagram/Patient%20Monitoring

Sensor Stream Ingest and Processing



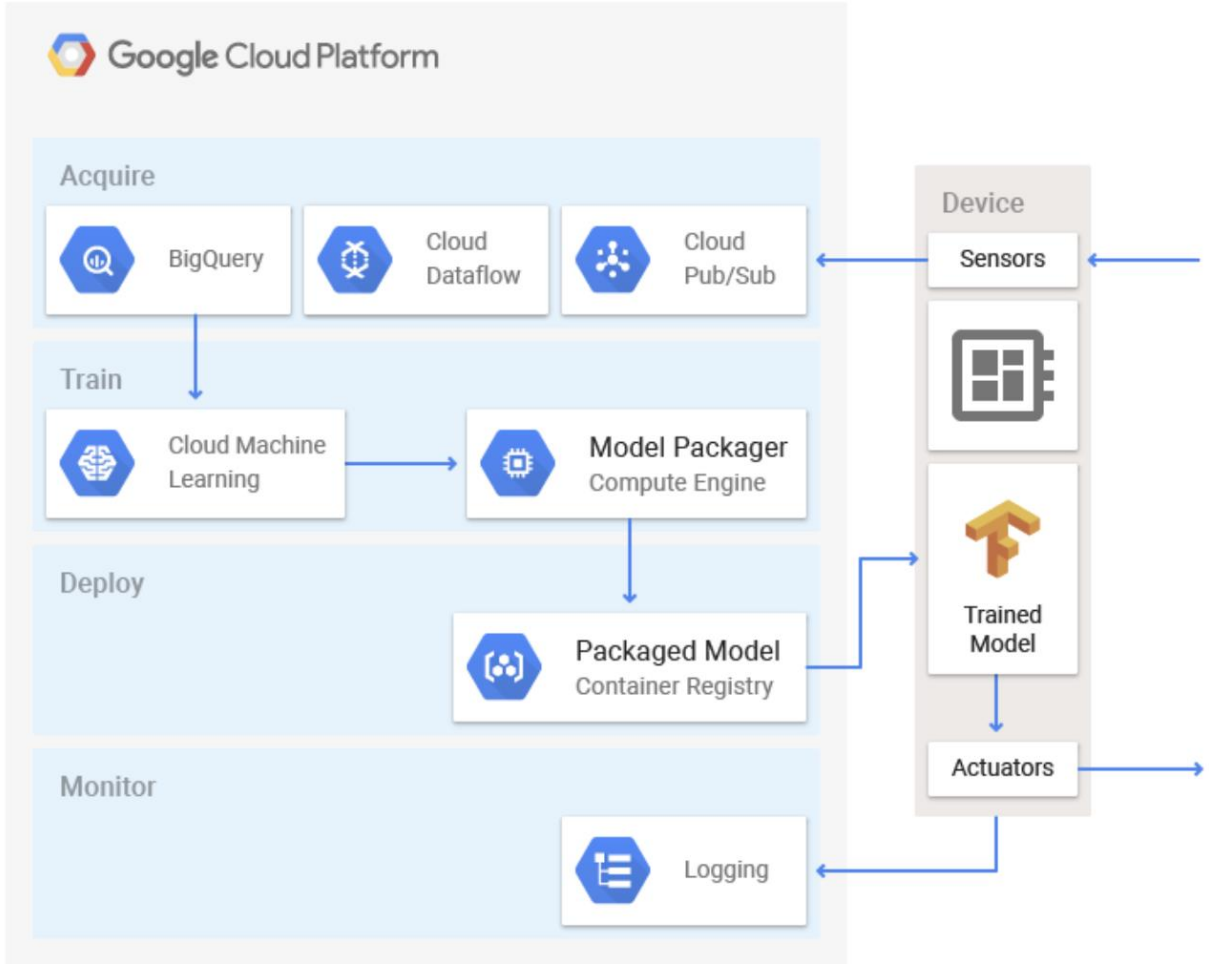
http://gcp.solutions/diagram/Sensor%20stream%20ingest%20and%20processing

Hybrid Cloud Solutions



http://gcp.solutions/diagram/Hybrid%20Federated%20Kubernetes%20with%20Shared%20Services

Edge Computing



http://gcp.solutions/diagram/Cloud%20to%20Edge%20ML

Chapter Concepts

Storage Overview

GCP Storage Options

Architecting Data Processing Solutions

Exam Prep

Do Now: Practice Quiz

- Take this practice quiz

ROITRAINING
MAXIMIZE YOUR TRAINING INVESTMENT

Google Cloud Certification Workshop—Data Engineer

Chapter 3: Storing Binary Data

Chapter Objectives

In this chapter, we will learn how to:

- Optimize storage cost and efficiency using Google Cloud Storage
- Add storage to virtual machines using persistent disks
- Back up disks using snapshots
- Migrate data into GCP

Chapter Concepts

Google Cloud Storage

Persistent Disks

Encryption

Data Transfer Options

Exam Prep

Google Cloud Storage

- Binary storage for any type of object
- Objects are stored in buckets
  - Buckets must have names that are unique in all the world
- There are a number of different storage classes
  - When creating a bucket, a default storage class is specified, but not all objects in a bucket must share the same storage class
  - Changing a bucket's storage class does not affect existing objects
- There is no limit to the number of objects that can be stored in a bucket
- The maximum size of a single object is 5TB

Storage Classes

- Multi-regional storage copies the files ‘geo-redundantly’ across multiple data centers in a geographic area
  - Choose United States, Europe, or Asia
  - No control over which regions are chosen within the geographic area
- Dual-region storage allows you to specify which two regions you want
  - Greater storage cost per GB
- Regional storage copies data in multiple zones within a single data center
- Nearline and Coldline storage are cheaper per GB, but there is a cost to access the data and a minimum storage charge per object

Do Now: Understanding Storage Classes

- Fill in the table below; the information can be found at: https://cloud.google.com/storage/docs/storage-classes

	Multi-Regional	Regional	Nearline	Coldline
Cost/GB/Month				
Availability				
Durability				
Is there a cost to access files?				
Minimum storage duration				

Choosing Storage Classes

Choose storage class based on the use case for the data:

- Do you want the highest levels of availability?
- Is the data accessed by machines all in the same data center?
- Is the data accessed frequently?
- What is the data used for? (i.e., web content, backup, archive, disaster recovery, etc.)

Do Now: Matching Storage Class to Use Case

- Draw an arrow to select the best storage class for each use case

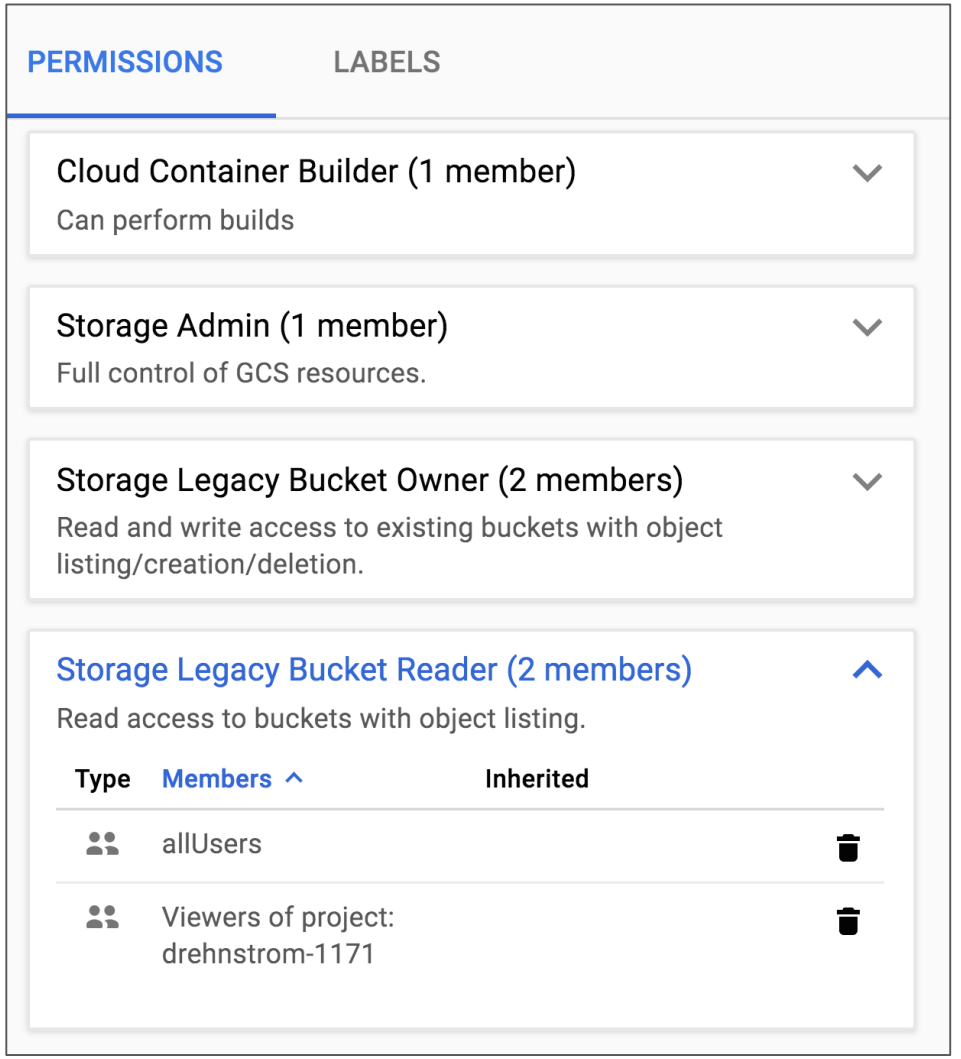
Storing CSV files for analysis using a Hadoop cluster running in Google Cloud Dataproc	Multi-Regional
Archiving old emails to conform to a regulatory requirement	
Storing static web files (.pdf, .jpg, .css, .js, etc.) for distribution all over the world	Regional
Storing data files that will be used for training in a machine learning project	Nearline
Storing a huge number of photos that will be made available via a website	
Storing snapshots of virtual machines in another region so they can be started if the main region is down	Coldline

Bucket Features

- Object metadata
- Encryption by default
- Versioning
- Change notifications and lifecycle management
- Simple static website hosting
- Retention policies

Bucket Security

- Permissions can be added to control access to buckets and objects
  - Use IAM users and groups



Predefined IAM Storage Roles

- Roles can be added to member and service accounts at the project or bucket level

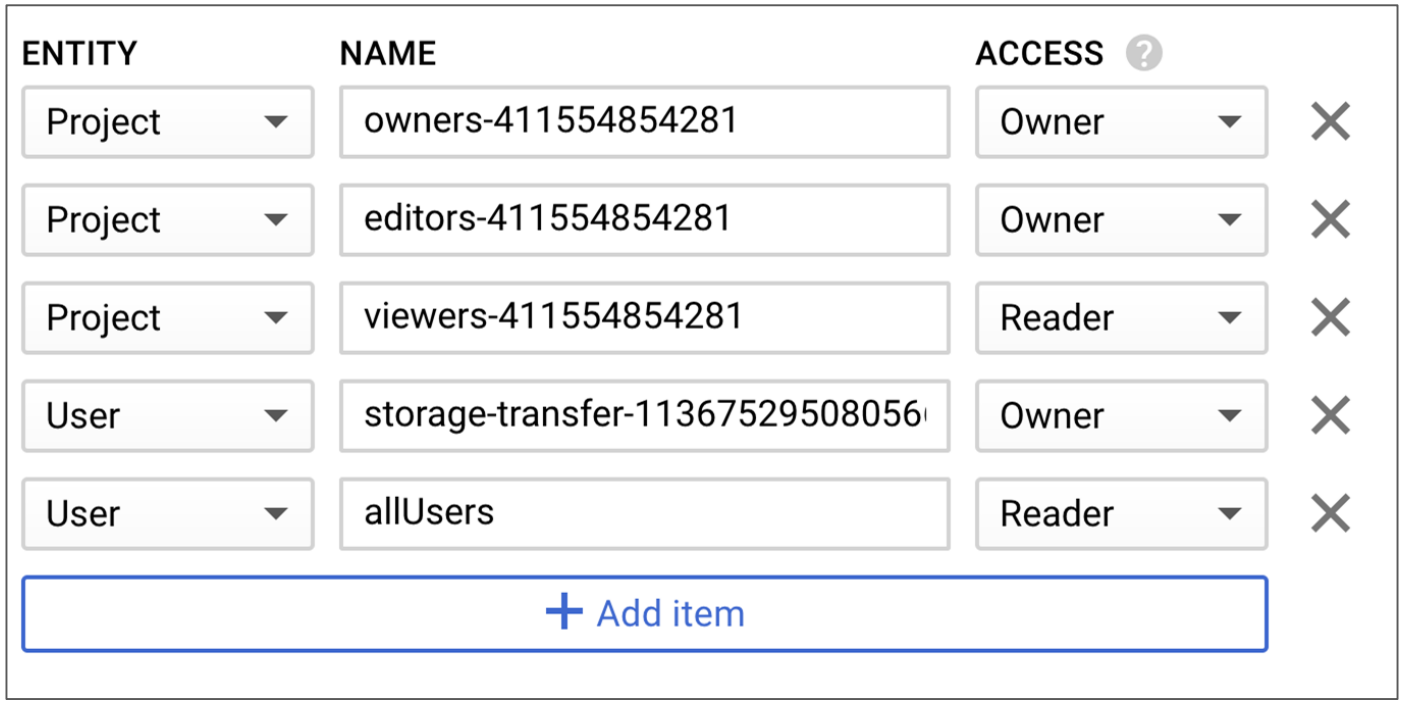
Storage Admin	Storage	Enabled	⋮
Storage Legacy Bucket Owner	Storage Legacy	Enabled	⋮
Storage Legacy Bucket Reader	Storage Legacy	Enabled	⋮
Storage Legacy Bucket Writer	Storage Legacy	Enabled	⋮
Storage Legacy Object Owner	Storage Legacy	Enabled	⋮
Storage Legacy Object Reader	Storage Legacy	Enabled	⋮
Storage Object Admin	Storage	Enabled	⋮
Storage Object Creator	Storage	Enabled	⋮
Storage Object Viewer	Storage	Enabled	⋮

Storage Role Permissions

Storage Object Admin	Storage Object Creator	Storage Object Viewer
Description Full control of storage objects. 9 assigned permissions: • resourcemanager.projects.get • resourcemanager.projects.list • storage.objects.create • storage.objects.delete • storage.objects.get • storage.objects.getIamPolicy • storage.objects.list • storage.objects.setIamPolicy • storage.objects.update	Description Access to create objects in storage. 3 assigned permissions: • resourcemanager.projects.get • resourcemanager.projects.list • storage.objects.create	Description Read access to storage objects. 4 assigned permissions: • resourcemanager.projects.get • resourcemanager.projects.list • storage.objects.get • storage.objects.list

Storage Object ACLs

- Access Control Lists (ACLs) can be used to grant access to objects in buckets



Signed URLs

- Provide temporary access to buckets
  - Create a service account with rights to storage
  - Create a service account key
  - Use signurl command to create a URL that allows access to the resource
    - -d parameter is used to specify duration

```
gcloud iam service-accounts keys create ~/key.json --iam-account
storage-admin-sa@doug-demo-project.iam.gserviceaccount.com


gsutil signurl -d 10m ~/key.json gs://super-secure-bucket/noir.jpg
```

Signed URL Example Output

```
me@doug-demo-project:~$ gsutil signurl -d 10m ~/key.json gs://super-secure-bucket/noir.jpg
URL     HTTP Method     Expiration      Signed URL
gs://super-secure-bucket/noir.jpg       GET     2018-08-31 16:29:25     https://storage.google
apis.com/super-secure-bucket/noir.jpg?x-goog-signature=107d26e38f5c962296c26f4153a1cbeb61a84ac
a905009752e849f8f890de1f9a80e482da3bae562c7796389e12a8657a70c87860700149c4b2218c81ad3d57730cd3
5ced850b266cdffd84de01898ee8c807d742a85136e56f46d83c29ceb792bdd3a22adbe2e540ba27b0f565bbf8f31a
ee6ae61d6ae20968021d5a47c8d0aada43f2d32407f2977a4c7b4c66ef64ddd68bd6f6135936f847ace3530a968d72
63ff5e70f9fc39bf16fabbd472f63584a8d8c6b24b1f81859f1c5176b8e97580a6b4a7613ad76bfcdd403e6afc9a70
90a3e1b4cf95c7fb68142416af86ef5ef6bfab93c00492b307233180df9b3dfeefe1bb9a5bf81cb441f879ecc2e57c
def&x-goog-algorithm=GOOG4-RSA-SHA256&x-goog-credential=storage-admin-sa%40doug-demo-project.i
am.gserviceaccount.com%2F20180831%2Fus%2Fstorage%2Fgoog4_request&x-goog-date=20180831T201925Z&
x-goog-expires=600&x-goog-signedheaders=host
me@doug-demo-project:~$
```

Delivering Static Content Worldwide

- Objects that are public include a public link
  - Give `allUsers` read access
- Link in the form of:
  - `https://storage.googleapis.com/{bucket}/{filename}`
- Upload all static files to a bucket (.css, .js, .html, .jpb, .png, etc.)
- Refer to these objects from dynamic web pages
  - Very scalable and inexpensive
  - No administration
- Use buckets to host web applications written with Javascript frameworks
  - Angular, React, etc.

Mapping a Bucket to a Domain

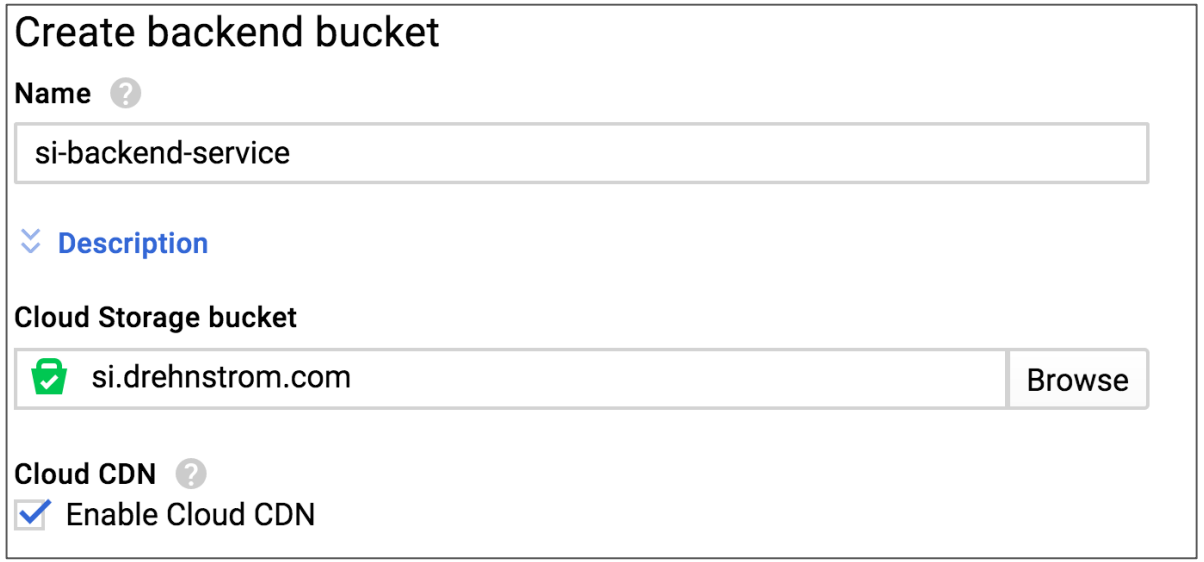
- Give a bucket a name includes a domain you own
  - Must prove you own the domain
  - Example bucket: `invaders.drehnstrom.com`
- Add a CNAME record that points to `c.storage.googleapis.com`

☐	invaders.drehnstrom.com.	CNAME	300	c.storage.googleapis.com.

- Example: go to the link: http://invaders.drehnstrom.com

Google CDN (Content Delivery Network)

- CDN copies data from a bucket and caches it all over the world
- Enabled when creating a load balancer's back end
  - Just check the box to enable it
- Content automatically delivered from a location close to the user
- Map your domain to the load balancer IP address



Do Now: CDN

- When delivering web content, you are charged for network egress
- Compare the costs of delivering content from a bucket vs. the CDN
  - https://cloud.google.com/storage/pricing#network-pricing
  - https://cloud.google.com/cdn/pricing
- Which is cheaper?

Exercise: Google Cloud Storage

- In this exercise, you will deploy a website using Google Cloud Storage
  - Google Cloud Storage

Chapter Concepts

Google Cloud Storage

Persistent Disks

Encryption

Data Transfer Options

Exam Prep

Persistent Disks

- Virtual hard drives that can be attached to virtual machines
- Three types of disks (Standard, SSD, Local SSD)
- Size from 10GB to 64TB
- Large drives have greater IOPs and throughput
- Must be in the same zone as the virtual machine they are attached to
  - Local SSDs are directly attached to the same hardware as the VM

Do Now: Persistent Disk Pricing

- Go to the following URL and compare the price of persistent disks and storage
  - https://cloud.google.com/compute/docs/disks/

Snapshots

- Snapshots are copies of disks
- Stored in Cloud Storage
- Multiple snapshots of the same disk only save incremental changes
  - When deleting earlier snapshots, all information required to rebuild the disk is retained
- Snapshots can be used when creating disks
  - Useful for moving disks to other zones or making disks larger

Prefer Cloud Storage over Persistent Disks

- Cloud Storage is less expensive per GB
- With Cloud Storage, you pay for what you use; with Disks you pay for what you allocate
- Using Cloud Storage separates compute from storage
  - Allows virtual machines to be more disposable
  - Makes autoscaling and maintenance easier

Exercise: Disks and Snapshots

- In this exercise, you will create a disk, attach it to a virtual machine, and backup the disk using a snapshot
  - Disks and Snapshots

Chapter Concepts

Google Cloud Storage

Persistent Disks

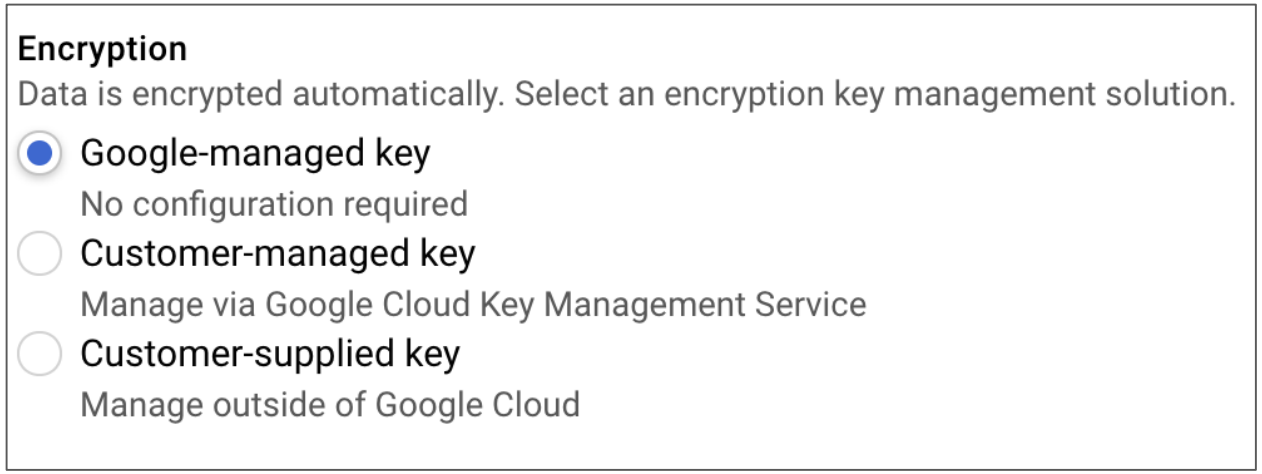
Encryption

Data Transfer Options

Exam Prep

Data Encryption

- All data stored in GCP is encrypted by default
  - Files are broken into chunks
  - Chunks are encrypted with a Data Encryption Key (DEK)
  - The DEK is encrypted with a Key Encryption Key (KEK)
  - Chunks are then stored multiple times on the physical devices
- Key management can be done 3 ways

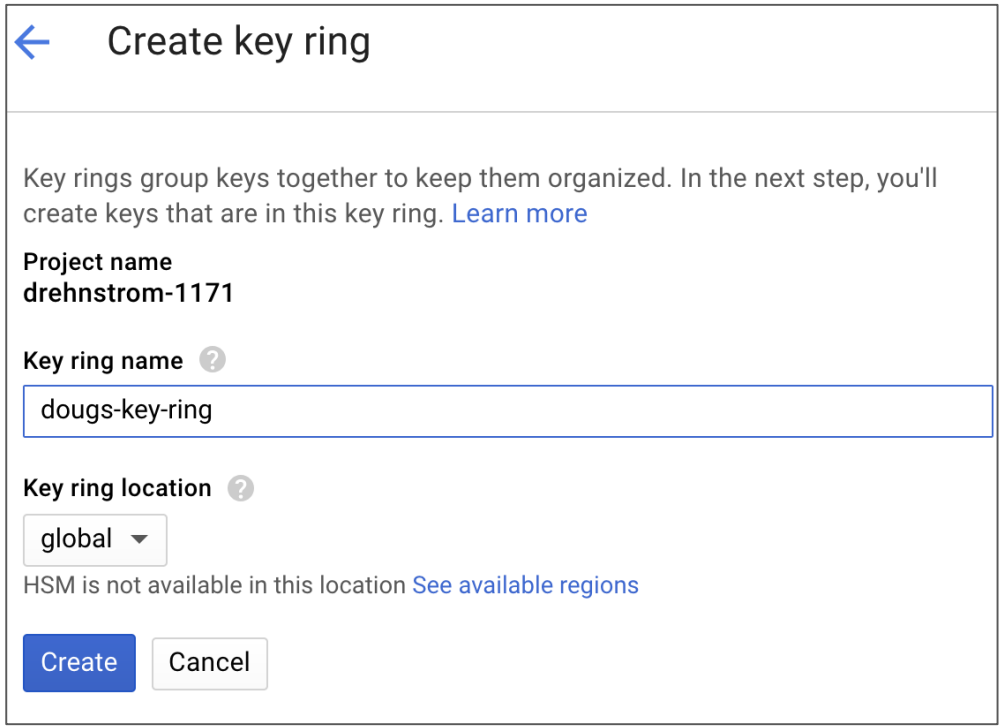


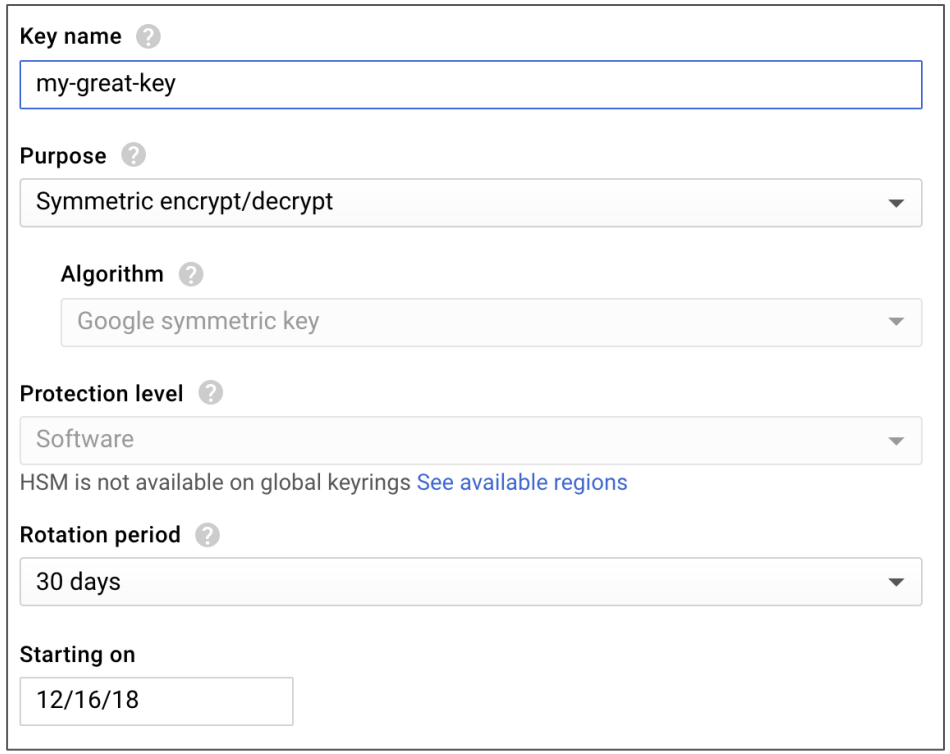
Google-Managed Keys

- Google does all the work managing and rotating the keys
- Key Encryption Keys are rotated frequently
  - Rotation schedule varies by service
- Data Encryption Keys are rotated less frequently
  - Re-encryption of data is required at least once every 5 years

Customer-Managed Keys

- Customer uses Google's Key Management Service (KMS) to manage keys
  - Create a key ring
  - You manage the key type, rotation frequency, etc.





Customer-Supplied Keys

- Create the key in your own environment
  - Upload the key to Google
- Don't lose the key; if you do, you can't recover the data

Client-Side Encryption

- It should go without saying that you can encrypt your data before sending it to Google
  - Google will then encrypt your encrypted data

Chapter Concepts

Google Cloud Storage

Persistent Disks

Encryption

Data Transfer Options

Exam Prep

Cloud Data Transfer Services

- Google provides a range of data transfer methods to get data into GCP
  - Choose based on type and volume of data
- Web console, gsutil, JSON API
  - Small amounts of data
- Storage Transfer Service
  - Bucket-to-bucket
  - Scheduled or ad hoc
- BIgQuery Data Transfer Service
  - Import into BIgQuery from GCS and selected Google applications
- Transfer Appliance
  - Offline import for large (20TB plus) amounts of data

Transfer Appliance

- If you have a lot of data, Google will send you a transfer appliance
  - Load the data locally and then ship it back
- Use when transfer times would be too long over your network

	1 Mbps	10 Mbps	100 Mbps	1 Gbps	10 Gbps	100 Gbps
1 GB	3 hrs	18 mins	2 mins	11 secs	1 sec	0.1 sec
10 GB	30 hrs	3 hrs	18 mins	2 mins	11 secs	1 sec
100 GB	12 days	30 hrs	3 hrs	18 mins	2 mins	11 secs
1 TB	124 days	12 days	30 hrs	3 hrs	18 mins	2 mins
10 TB	3 years	124 days	12 days	30 hrs	3 hrs	18 mins
100 TB	34 years	3 years	124 days	12 days	30 hrs	3 hrs
1 PB	340 years	34 years	3 years	124 days	12 days	30 hrs
10 PB	3,404 years	340 years	34 years	3 years	124 days	12 days
100 PB	34,048 years	3,404 years	340 years	34 years	3 years	124 days

Chapter Concepts

Google Cloud Storage

Persistent Disks

Encryption

Data Transfer Options

Exam Prep

Do Now: Practice Quiz

- Take this practice quiz

Links

- Google Cloud Storage
- Persistent Disks
- Cloud CDN
- Cloud Data Transfer

Google Cloud Certification Workshop—Data Engineer

Chapter 4: Storing Relational Data

Chapter Objectives

In this chapter, we will learn how to:

- Store relational data using Google Cloud SQL
- Create massively scalable relational databases using Spanner

Chapter Concepts

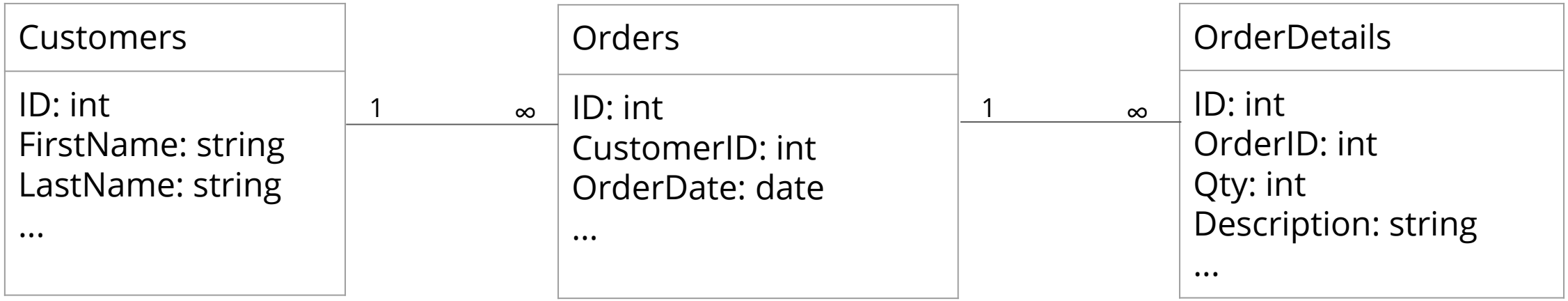
- **Understanding Relational Storage**
- Cloud SQL
- Spanner
- Exam Prep

Benefits of Relational Databases

- Strongly type schema that enforces data integrity
- Relationships ensure no orphan records between parent-child tables
- Flexible indexing for fast, efficient retrieval
- Strongly consistency, ACID transactions
- Familiar to many database admins and programmers
- Standard SQL language works across many database implementations

Modeling Relational Data

- Tables contain fields, indexes, and constraints
- Primary key ensures each row in a table should be unique
- Relationships are constraints that ensure a parent row cannot be deleted if there are child rows in another table



Chapter Concepts

Understanding Relational Storage

Cloud SQL

Spanner

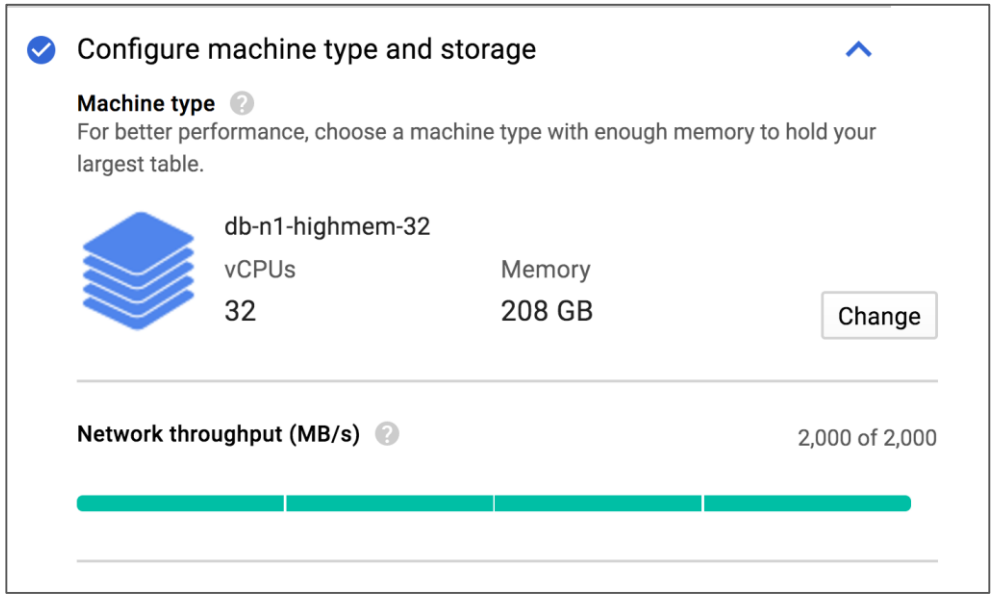
Exam Prep

Google Cloud SQL

- Managed service for relational databases running in GCP
- Supports MySQL and PostgreSQL
- Database size from 10GB to 10TB
- Automatic maintenance and backups
- Create failover replica for high availability
- Firewall blocks all traffic outside a project by default
  - Authorized networks to provide access as needed

Scaling Cloud SQL Databases

- Scale up (vertical scaling)
- Machine type determines capacity
  - Add vCPUs and memory to support more throughput
- Can also create a read replica



Exercise: Cloud SQL Quickstart

- In this exercise, you will get started using Google Cloud SQL
  - Cloud SQL Quickstart

Chapter Concepts

Understanding Relational Storage

Cloud SQL

Spanner

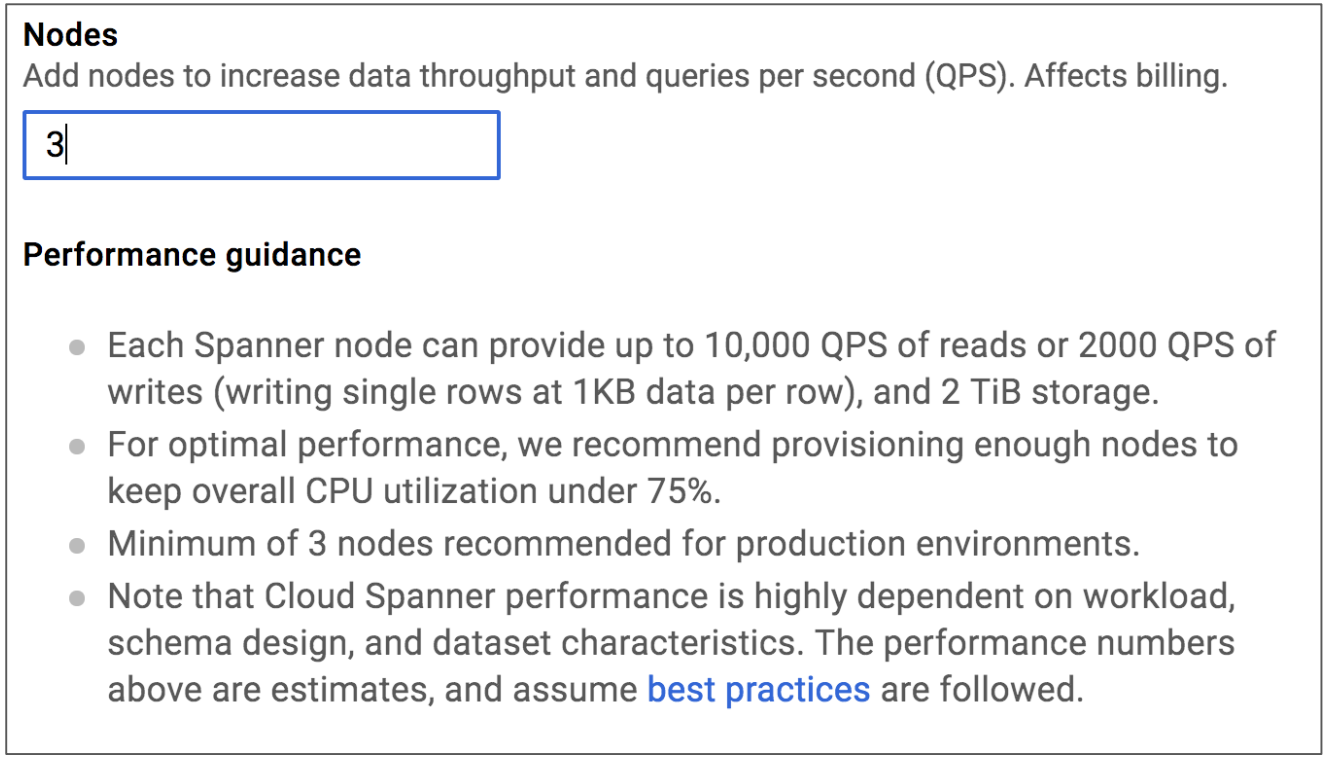
Exam Prep

Google Cloud Spanner

- Completely managed relational database
  - No need to worry about updates or maintenance
- Scales globally across regions
- Number of nodes determines capacity
- Designed for extremely large relational databases

Scaling Spanner Database

- Scale out (horizontal scaling)
  - Add nodes to support more users



What Makes Spanner Special?

- Spanner can scale horizontally by adding nodes while still maintaining strongly consistent ACID transactions
  - Uses atomic clocks attached to the servers that run Spanner
- Spanner is unique in the world and has no competitor
- Spanner is proven, having been used internally at Google for years

Do Now: Comparing Cloud SQL and Spanner

- Visit the following URLs and fill in the table below
  - https://cloud.google.com/sql/docs/
  - https://cloud.google.com/spanner/docs/

	Cloud SQL	Spanner
Cost of compute		
Cost of storage		

Exercise: Spanner Quickstart

- In this exercise, you will get started using Google Cloud Spanner
  - Google Cloud Spanner Quickstart

Chapter Concepts

Understanding Relational Storage

Cloud SQL

Spanner

Exam Prep

Do Now: Practice Quiz

- Take this practice quiz

Links

- Cloud SQL
- Spanner

ROI TRAINING
MAXIMIZE YOUR TRAINING INVESTMENT

Google Cloud Certification Workshop—Data Engineer

Chapter 5: Managed NoSQL Solutions

Chapter Objectives

In this chapter, we will learn how to:

- Consider the advantages and disadvantages of NoSQL storage
- Store massive amounts of data using Google Bigtable
- Leverage Google Cloud Datastore and Firestore for NoSQL storage
- Describe the benefits of Memorystore

Chapter Concepts

Understanding NoSQL Storage

Bigtable

Cloud Datastore/Firestore

Memorystore

Exam Prep

NoSQL Storage

- Structured storage like relational databases but with key differences
  - Tend to be schemaless
  - May or may not support transactions
  - Varying support for SQL
- Easier to use, especially when storing data in object-oriented systems
  - Often support hierarchical and nested data
- NoSQL systems tend to be highly scalable and can scale horizontally
- Tools are often less mature for NoSQL than for relational databases
  - Ad-hoc reporting sometimes harder
  - Query syntax tends to be different

Types of NoSQL Database

- Key-value stores
  - Data is stored in key-value pairs
  - Examples include Redis and SimpleDB
- Document stores
  - Data is stored in some standard format like XML or JSON
  - Nested and hierarchical data can be stored together
  - MongoDB, CouchDB, and DynamoDB are examples
- Wide-column stores
  - Key identifies a row in a table
  - Columns can be different within each row
  - Cassandra and HBase are examples

Do Now: Comparing NoSQL to Relational

- Fill in the table below:

Advantages of NoSQL Databases	Advantages of Relational Databases

Chapter Concepts

Understanding NoSQL Storage

Bigtable

Cloud Datastore/Firestore

Memorystore

Exam Prep

Google Bigtable

- Fully-managed, wide-column NoSQL datastore similar to Cassandra or HBase
- Designed for scalability—add compute nodes to meet demand
- Compute separate from storage
  - Allows nodes to be added or removed without affecting the data
- Secure
  - Data encrypted by default
  - Use IAM to assign permissions
- Uses same API as HBase
  - Allows on-premises HBase applications to be easily migrated

Bigtable Schemas

- Store data in tables
  - Tables have rows
  - Each row has a key
  - Rows have columns
  - Columns can be grouped in column families
- Store all data about an entity in a single row
  - There is no penalty for sparse columns (columns with no data)
- Updates to a single row are in implicit transactions
- Transactions across multiple rows are not supported

Bigtable Row Keys

- Bigtable does not support indexes other than on the row key
  - Rows are stored in row-key order
- When selecting data, you select it by row key, key prefix, or range of keys
- Choose keys based on how you would most likely select the data
  - Can embed multiple values in the key
- For more information on selecting row keys see: https://cloud.google.com/bigtable/docs/schema-design#row-keys

Chapter Concepts

Understanding NoSQL Storage

Bigtable

Cloud Datastore/Firestore

Memorystore

Exam Prep

Google Cloud Datastore

- Completely managed document store
  - No administration, no maintenance, nothing to provision or set up
- 1GB per month free tier
- Indexes created for every property by default
  - Secondary indexes and composite indexes are supported
- Supports ACID transactions
- Schemaless
- For pricing info see: https://cloud.google.com/datastore/pricing

Relational vs. Datastore Terminology

Relational	Datastore
Tables	Kinds
Records	Entities
Fields	Properties
Primary Key	Key
Relationship	Entity Group
Primary-Foreign Keys	Ancestor Paths

Modeling Datastore Entities

- Datastore entities have one or more properties
  - Can be scalar values like strings, dates, numbers, bools, blobs, etc.
  - Properties can be arrays
  - Properties can be embedded entities
- Properties can be indexed or not
  - Can create composite indexes
- Different entities of the same kind can have different properties

Google Cloud Firestore

- Firestore is the new and improved version of Datastore
  - Reworking of Firebase Realtime Database
  - Two modes: Native and Datastore
- Native mode supports all Firebase features
  - Uses Firebase API
  - Not supported with older App Engine runtimes
- Datastore mode does not support all Firestore features like offline support for mobile devices and synchronization
  - Compatible with Datastore API
- Older Datastore database will be migrated to Firestore in Datastore mode
- See: https://cloud.google.com/datastore/docs/

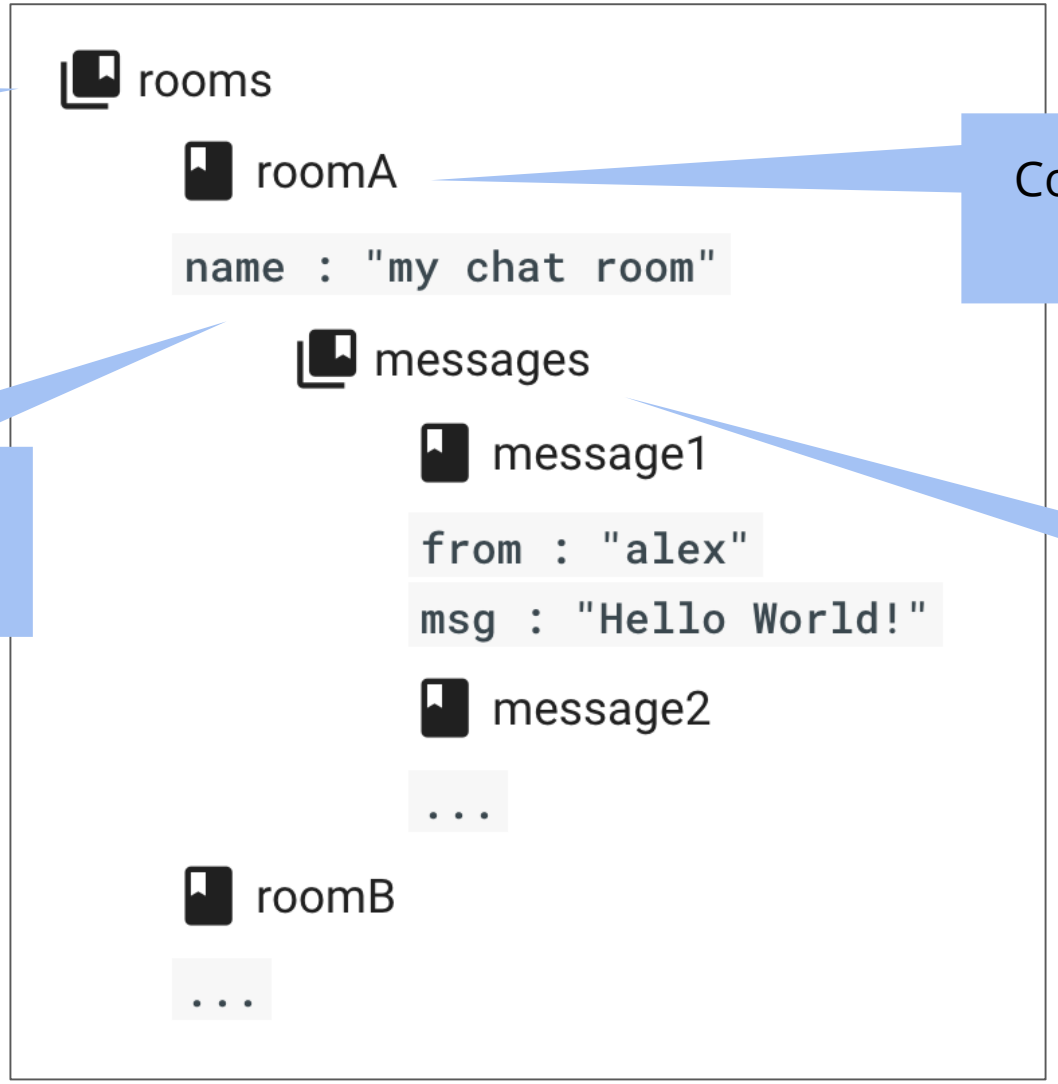
	Native mode Enable all of Cloud Firestore's features, with offline support and real-time synchronization. SELECT NATIVE MODE	Datastore mode Leverage Cloud Datastore's system behavior on top of Cloud Firestore's powerful storage layer. SELECT DATASTORE MODE
API	Firestore	Datastore
Scalability	Automatically scales to millions of concurrent clients	Automatically scales to millions of writes per second
App engine support	Not supported in the App Engine standard Python 2.7 and PHP 5.5 runtimes	All runtimes
Max writes per second	10,000	No limit
Real-time updates	✓	✗
Mobile/web client libraries with offline data persistence	✓	✗

Firestore vs. Datastore Terminology

Firestore	Datastore
Collections	Kinds
Documents	Entities
Key-Value Pairs	Properties
Document Name	Key
Sub-Documents	Entity Groups

Modeling Firestore Documents

Data is organized in Collections



Collections contain Documents

Documents have properties

Documents can also have Subcollections

Exercise: Datastore/Firestore Quickstarts

- Google Cloud Datastore Quickstart
- Google Cloud Firestore Quickstart

Chapter Concepts

Understanding NoSQL Storage

Bigtable

Cloud Datastore/Firestore

Memorystore

Exam Prep

Memorystore

- Fully-managed, in-memory Redis database solution
- Scale from 1GB to 300GB instances
- Basic tier has a single instance
- Standard tier deploys a failover replica in another zone
  - 99.9% availability SLA
- Secure with only an internal IP address
  - Available only from within the same VPC
  - Use IAM roles to control access

Redis

- Redis is a key-value store
  - `SET favorite_team "Steelers"`
  - `GET favorite_team`
    - Returns "Steelers"
- Often used to cache database data to improve website performance
  - Data is stored in memory for very fast access
  - Data is also stored on disk for fault tolerance
- Open-source and supported by many languages
  - Not altered in any way by Memorystore
- Try Redis here: https://try.redis.io/

Exercise: Memorystore Quickstart

- Memorystore Quickstart

Chapter Concepts

Understanding NoSQL Storage

Bigtable

Cloud Datastore/Firestore

Memorystore

Exam Prep

Do Now: Comparing Storage Prices

- Using the Price Calculator, estimate the cost of running a 10TB database using Cloud SQL, Spanner, Datastore, and Bigtable
  - Do your best to estimate the values for each product
  - Use the documentation to help with your estimates

Do Now: Practice Quiz

- Take this practice quiz

Links

- Datastore
- Firestore
- Bigtable
- Memorystore

Google Cloud Certification Workshop—Data Engineer

Homework

Links

- Google Cloud Storage
- Persistent Disks
- Cloud Data Transfer Service
- Cloud CDN
- Cloud SQL
- Spanner
- Datastore
- Firestore
- Bigtable
- Memorystore

Videos

- Google Cloud Platform Storage Versioning Demo
- Google Cloud Platform Enabling CORS
- Google Cloud Platform Cloud SQL Demo
- Google Cloud Platform Datastore Demo
- Google Cloud Platform Bulk Loading to Datastore Demo
- Choosing your storage and database on Google Cloud Platform
- Cloud Spanner 101
- Efficiently migrating your data into Google Cloud Platform

Tutorials

- Google Cloud Storage
- Disks and Snapshots
- Storage Transfer Service
- Cloud SQL Quickstart
- Cloud Spanner Quickstart
- Datastore Quickstart
- Firestore Quickstart
- Memorystore Quickstart

ROITRAINING
MAXIMIZE YOUR TRAINING INVESTMENT

Google Cloud Certification Workshop—Data Engineer

Chapter 6:
Big Data Processing and Analytics

Chapter Objectives

In this chapter, we will learn how to:

- Process big data using Google Cloud Platform services
- Run Hadoop and Spark Jobs in GCP using Google Cloud Dataproc
- Leverage BigQuery for NoOps data warehousing and analysis
- Write BigQuery queries

Chapter Concepts

Big Data Processing Overview

Google Cloud Dataproc

BigQuery

Choosing Big Data Strategies

Exam Prep

Google’s Original Mission Statement

“To organize the world's information and make it universally accessible and useful.”

Achieving the Mission

- Traditional technologies were not capable of collecting and processing the amount of data Google was accumulating
  - For Google to achieve its mission, it had to invent new technologies
- Needed a massively distributed file system
  - Drives are slow, so connect many together to reduce the bottleneck
- Needed a way of getting huge jobs done quickly

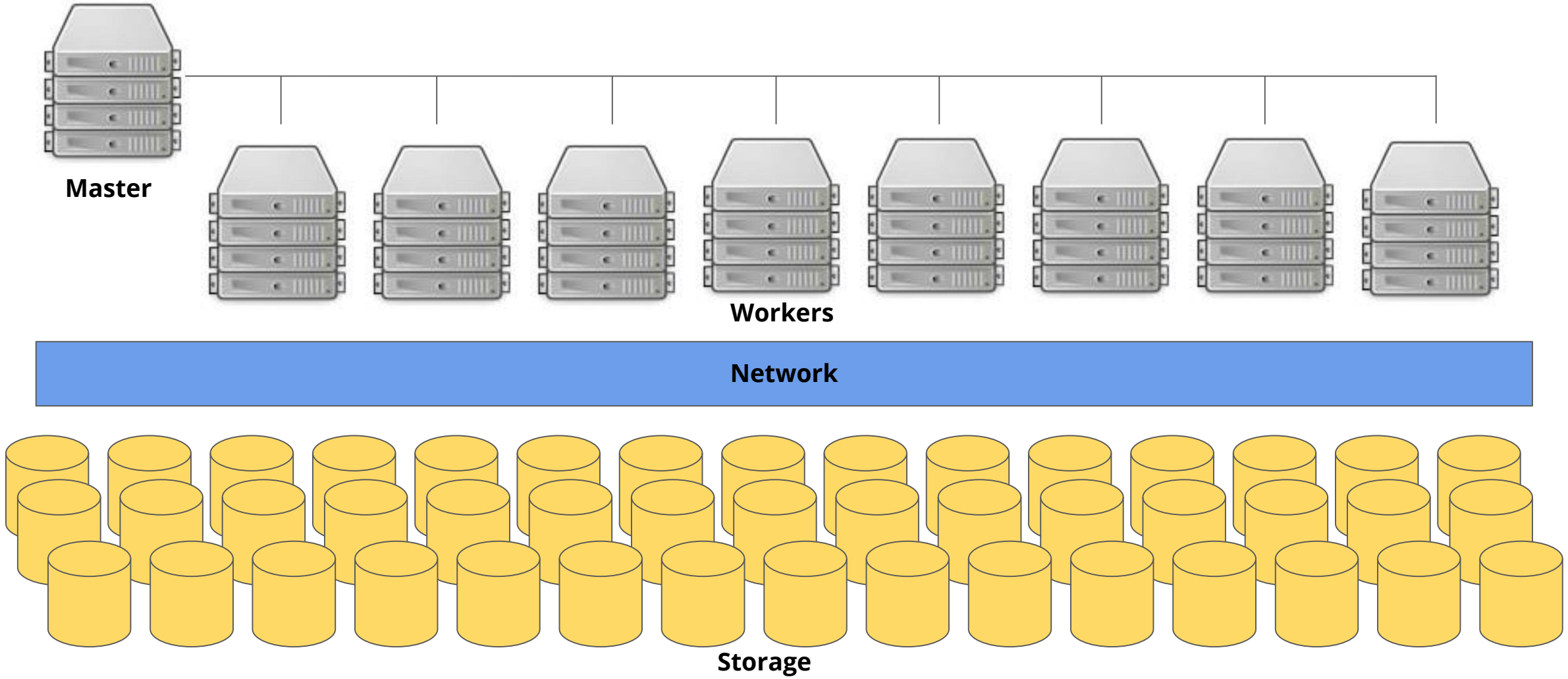
The 3 Vs of Big Data

- Volume – there is a lot of data
  - Need many drives to store the data
- Velocity – if there is a lot, then you must be collecting it at a fast rate
  - Need to be able to write to drives very quickly
  - Need to be able to get the data back very quickly
- Variety – the data is coming for many different sources
  - Web pages
  - Text files
  - Logs
  - PDFs
  - Databases
  - Etc.

Google Version 1: GFS and MapReduce

- Google created a distributed file system named GFS (Google File System)
  - Allowed data to be read and written quickly
- Google created the MapReduce algorithm for splitting a job across many machines to get it done quicker
- GFS and MapReduce were later implemented at Yahoo based on a paper released by Google
  - This became HDFS and Hadoop
  - Hadoop project was later given to the Apache Group to manage

GFS and MapReduce Illustrated



Google Version 2: Colossus and Dremel

- Colossus replaced GFS as the distributed file system
  - We know Colossus as Google Cloud Storage
- Dremel replaced MapReduce for distributed processing
  - Similar to Hive, Spark SQL, or Presto
  - We know Dremel as BigQuery

Chapter Concepts

Big Data Processing Overview

Google Cloud Dataproc

BigQuery

Choosing Big Data Strategies

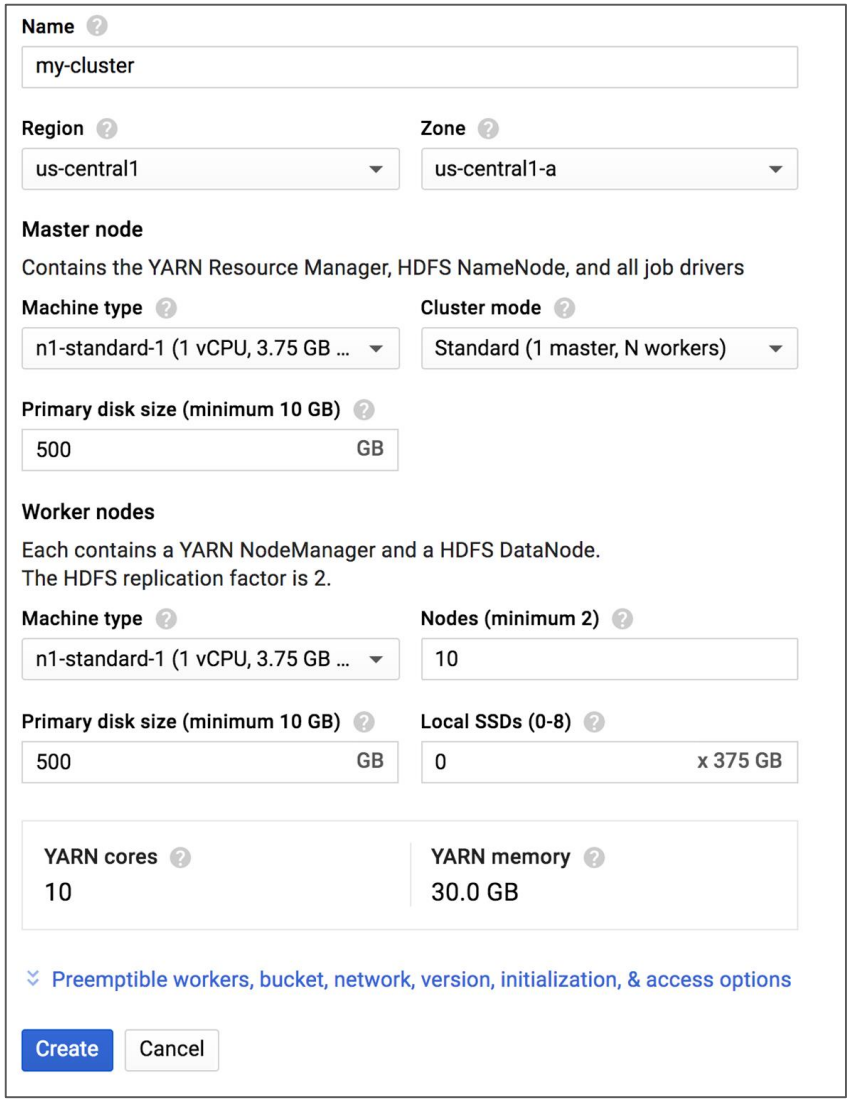
Exam Prep

Dataproc

- Service for easily creating Hadoop/Spark clusters
  - Simple configuration
  - Clusters created in just a couple minutes
  - Easy to submit jobs
- Dataproc clusters supports all common Hadoop and Spark frameworks
  - Easy to migrate on-premises workloads into the cloud
- HDFS Cluster automatically created for storage

Dataproc Supports Hadoop Ecosystem

Creating a Dataproc Cluster



Select a zone close to your data

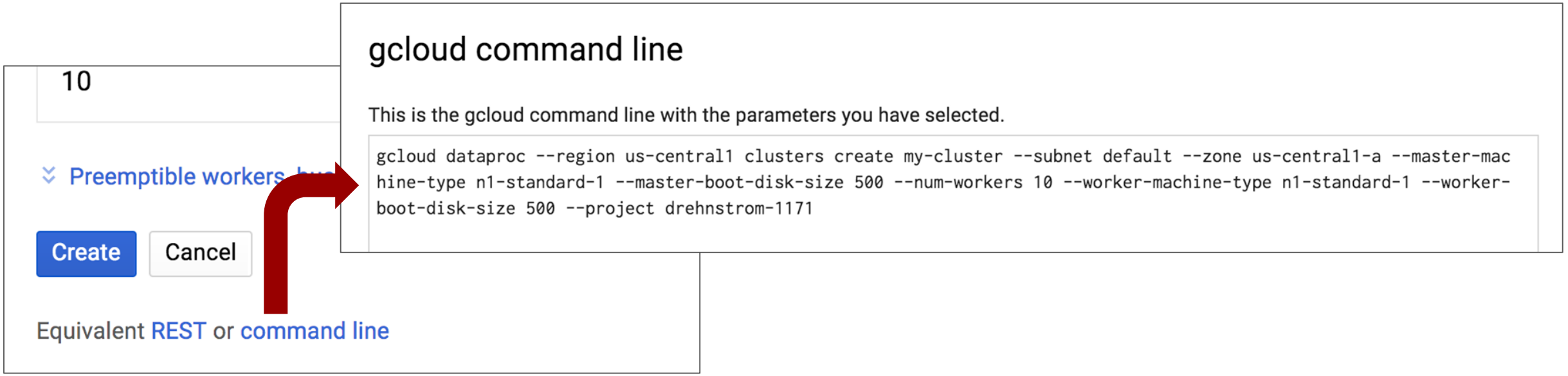
Configure the Master

Configure the Workers

The disks are used in the HDFS cluster

Scripting Dataproc Cluster Creation

- Click the command line link on the configuration page
  - Opens another window with the command pre-built
- Use this command for creating a script to automate cluster creation



Separate Compute and Storage

- Dataproc creates an HDFS cluster, but don't use it for long-lived storage
- Store the data to analyze in Google Cloud Storage
  - Cloud Storage is cheaper (HDFS cluster created on Persistent Disks)
  - Only pay for what you use, not for what you allocate
- Separating storage and compute allows the cluster to be disposable
  - Can size the cluster for specific jobs
  - Delete the cluster as soon as possible
  - Only pay for it while it is working
  - Can recreate the cluster in a couple minutes

Exercises: Dataproc Clusters

- In this exercise, you will create a Dataproc cluster and experiment with managing the cluster and running Hadoop and Spark jobs
  - Creating Dataproc Clusters
- In this exercise, you see how to run Hive and Spark jobs on Dataproc clusters and also how to scale clusters while they are running using preemptible workers
  - GCP Dataproc: Running Hive and Spark Jobs

Chapter Concepts

Big Data Processing Overview

Google Cloud Dataproc

BigQuery

Choosing Big Data Strategies

Exam Prep

BigQuery

- BigQuery consists of two main services
  - Data warehousing
  - Data analytics

BigQuery Data Warehousing

- BigQuery has its own storage system
  - Most efficient storage when running BigQuery queries
- Projects contain datasets, which contain tables
  - There is no limit to the number of tables in a dataset
  - There is no limit to the number of rows in a table
- Each field in a table is stored separately
  - Makes querying more efficient because only fields in a query are read
- All data is encrypted by default
- All data is compressed for faster processing

BigQuery Schemas

- Tables must have a schema which defines field names and data types
  - Schemas support nested, hierarchical data
  - Schemas support repeated fields
  - For example, an Orders table can have a field called Details which is an array of records which provides details about each order
- Repeated, nested fields allow querying parent-child relationships without needing to join two tables
  - Joins are expensive in BigQuery since there are no indexes

BigQuery Storage Cost

- Storage cost is:
  - 2 cents per GB per month for the first 3 months
  - After 3 months, data storage is reduced to 1 cent per GB per month
- Quiz: If you put 10TB of data in BigQuery today and left it there for a year, how much would it cost?

BigQuery Data Analytics

- Use standard ANSI SQL 2011 for building Select queries
- Can write user-defined functions to manipulate data in SQL or JavaScript
- Can query from a number of data sources:
  - BigQuery Storage
  - Google Cloud Storage
  - Google Bigtable
  - Google Drive
- Completely NoOps
  - No need to provision anything
  - No need to tune queries

BigQuery Analysis Cost

- Queries are charged one of two ways: on-demand and flat-rate
- On demand at $5/TB of data processed with 1TB free per month
- With flat-rate pricing, you pre-purchase BigQuery capacity
  - Capacity is measured in “slots”, a unit of processing in BigQuery
  - Run as many queries as you can, but you will never go over your purchased slots

Writing BigQuery Queries

- There are two dialects of BigQuery SQL, Legacy and Standard
  - Legacy was the original, replaced in 2016
- Standard SQL is ANSI 2011 SQL compliant
  - Very minor differences exist due to the platform
- Standard SQL includes data manipulation language (DML) statements
  - INSERT, UPDATE, DELETE
  - Strict quotas apply to DML statements

Standard SQL Examples

Comment instructs BigQuery to use Standard not Legacy SQL

```
#standardsql
SELECT city, count(city) as city_count FROM `drehnstrom-
1171.sales_dataset.stores`
GROUP BY city
ORDER BY city_count DESC
LIMIT 1000
```

Back ticks, not quotes, surround the table name

To refer to a table, the syntax is: `project-id.dataset.table`

```
#standardsql
SELECT store, name FROM `drehnstrom-1171.sales_dataset.stores`
WHERE city = 'DES MOINES'
LIMIT 1000
```

Standard SQL JOIN Example

```
#standardsql
SELECT name, sum(total) as sales
FROM `drehnstrom-1171.sales_dataset.stores` as stores
JOIN `drehnstrom-1171.sales_dataset.sales` as sales
ON stores.store = sales.store
GROUP BY name
LIMIT 10
```

Standard SQL WITH Example

```
#standardsql
with total_sales as
(SELECT name, sum(total) as sales
FROM `drehnstrom-1171.sales_dataset.stores` as stores
JOIN `drehnstrom-1171.sales_dataset.sales` as sales
ON stores.store = sales.store
GROUP BY name)

SELECT name, CAST(ROUND(sales, 0) as STRING) from total_sales
WHERE sales > 10000
ORDER BY sales DESC
```

Denormalization Example

```
#standardsql
SELECT stores.store, name,
    ARRAY_AGG(
    STRUCT(sales.date, sales.category,
           vendor_no, item,sales.total)) as sale
  FROM
    `drehnstrom-1171.sales_dataset.sales` as sales
JOIN

   `drehnstrom-1171.
ON
  sales.store = stor
  GROUP BY stores.st
```

`STRUCT` function creates a composite field. `ARRAY_AGG` function creates an array.

Row	store	name	sale.date	sale.category	sale.vendor_no	sale.item	sale.total
1	4965	Pronto	2014-02-12	1051120	434	55246	24.18
			2014-02-12	1051100	434	54056	24.18
			2014-02-12	1051120	434	55246	24.18
			2014-02-12	1051100	434	54056	24.18
			2014-01-08	1011100	260	67273	83.28
			2014-01-08	1011100	260	67273	83.28
			2014-01-08	1011100	260	25604	63.0

Querying Denormalized Data

```
#standardsql
SELECT name, ARRAY_LENGTH(sale) as count_sales
FROM `drehnstrom-1171.sales_dataset.store_sales_denormalized`
ORDER BY count_sales DESC
LIMIT 10
```

```
#standardsql
SELECT name, (SELECT sum(total) from UNNEST(sale)) as total_sales
FROM `drehnstrom-1171.sales_dataset.store_sales_denormalized`
ORDER BY total_sales DESC
LIMIT 10
```

UNNEST flattens the array of objects so they are queryable

User-Defined Function Example

```
CREATE TEMPORARY FUNCTION to_celsius(temp FLOAT64)
RETURNS FLOAT64
LANGUAGE js AS """
  return (temp - 32.0) * 5.0 / 9.0;
""";

WITH temps AS
  (SELECT 32 AS Fahrenheit
  UNION ALL
  SELECT -40 AS Fahrenheit
  UNION ALL
  SELECT 212 as Fahrenheit)
SELECT Fahrenheit, to_celsius(Fahrenheit) as Celsius
FROM temps;
```

Can use JavaScript or SQL when writing a UDF

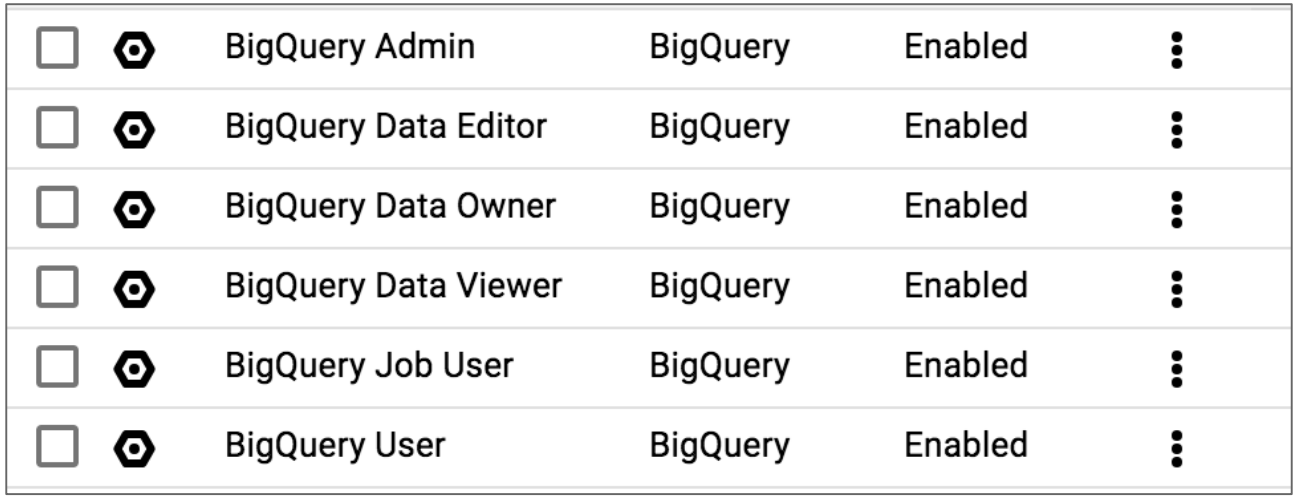
Row	Fahrenheit	Celsius
1	32	0.0
2	-40	-40.0
3	212	100.0

Do Now: Using BigQuery

- Log into GCP Management Console
- From the **Products and Services** menu, choose **BigQuery**
- Run each of the queries from the last few slides

Securing BigQuery

- Access is granted to BigQuery using IAM Members and Roles
- Table access is granted at the Dataset level
  - All Tables within a Dataset share the same permissions
  - For public Datasets, grant Viewer role to `allAuthenticatedUsers`
- Members at minimum need Job User role to run queries



BigQuery IAM Roles Described

Role	Description
BigQuery Admin	Can do everything in BigQuery. Create and read data, run jobs, set IAM policies etc.
BigQuery Data Owner	Read/write access to data, plus can grant access to other users and groups by setting IAM policies.
BigQuery Data Editor	Read/write access to data.
BigQuery Data Viewer	Read-only access to data.
BigQuery Job User	Can create and run jobs, but no access to data.
BigQuery User	Can run jobs, create datasets, list tables, save queries. But no default access to data.

BigQuery Performance Tips

- Denormalize parent-child relationships
  - Store child records as repeated records with the parent row
  - Allows querying related data without a join
- Queries that were already run are cached
  - Data returned from the cache is free
- For very large tables, create smaller temp tables when possible
- Partition tables when data is accumulated on a regular basis
  - For example, daily logs, daily sales, etc.
  - Can specify the data range of partitions to query, avoiding a table scan
- Don’t group by fields with a very large number of different values
- Prefer built-in functions to UDFs if possible

Querying External Data Sources

- BigQuery can analyze data directly from:
  - Bigtable
  - Cloud Storage
  - Google Drive
- Must define a table schema for the external source
- Not as efficient as BigQuery native storage
  - Native storage puts fields in separate tables so queries only have to scan the fields in the query, not the whole table
- Useful for ETL and denormalization jobs
  - Read data from external source, manipulate it, and load it into BigQuery
- See: https://cloud.google.com/bigquery/external-data-sources

Exercises: Query and Secure BigQuery Data

- In these exercises, you will upload data into BigQuery and then query that data, and you will also secure data with an authorized view
  - Querying Data with BigQuery
  - Creating an authorized view in BigQuery

Chapter Concepts

Big Data Processing Overview

Google Cloud Dataproc

BigQuery

Choosing Big Data Strategies

Exam Prep

Migrating from On-Premises Hadoop to Dataproc

- Can run existing Hadoop and Spark jobs on Dataproc
- Use Cloud Storage not HDFS so the Dataproc cluster can be deleted without deleting the data
  - Make the attached disks small
- Move HBase workloads to Bigtable to reduce administration
- Script creation of Dataproc clusters and jobs
  - Delete the cluster when the jobs are completed to reduce cost
  - Machines are billed in 1-minute increments with a 10-minute minimum
- Consider using preemptible instances for some of the workers

Migrating a Data Warehouse to BigQuery

- Denormalize data for better performance in BigQuery
  - Traditional data warehouses use a star schema and indexes for efficiency
  - BigQuery has no indexes, so Joins are costly
- Can use Dataflow to denormalize data prior to putting it in BigQuery
- Can alternatively leverage BigQuery for denormalization jobs
  - Write a query that joins tables and creates arrays of structs
  - Specify a destination table
- Partition tables that accumulate data over time

Do Now: Choosing Dataproc or BigQuery

- Fill in the table below:

	Advantages	Disadvantages
Dataproc		
BigQuery		

Chapter Concepts

Big Data Processing Overview

Google Cloud Dataproc

BigQuery

Choosing Big Data Strategies

Exam Prep

Do Now: Practice Quiz

- Take this practice quiz

Links

- Dataproc
- BigQuery

ROITRAINING
MAXIMIZE YOUR TRAINING INVESTMENT

Google Cloud Certification Workshop—Data Engineer

Chapter 7:
Data Processing Pipelines

Chapter Objectives

In this chapter, we will learn how to:

- Examine data processing challenges and architectures
- Leverage Google Cloud Dataflow for NoOps processing pipelines
- Process real-time, streaming pipelines with Pub/Sub
- Simplify data pipelines with Google Cloud Dataprep
- Describe the benefits of Cloud Composer
- Recognise the use cases for Data Loss Prevention (DLP)

Chapter Concepts

Google Cloud DataFlow

Pub/Sub

Cloud Dataprep

Cloud Composer

Data Loss Prevention

Exam Prep

Data Processing Pipelines

- Perform a set of actions in a chain
  - Data from one or more sources is read into the pipeline
  - Actions are performed on the data to manipulate or transform it
  - The manipulated results are sent as output from the pipeline
- Actions within a data pipeline can run at the same time (concurrently)
- At scale, multiple machines can participate to get the pipeline done faster
- MapReduce jobs are examples of pipelines at scale
  - Multiple nodes read data from disk and perform an initial map step
  - Data is then organized by key (the shuffle step)
  - Keyed data is processed separately in parallel (the reduce step)

Batch vs. Streaming Pipelines

- Batch data flows process big chunks of data at set intervals
  - Analyzing daily logs
  - Importing monthly sales
  - Periodic data conversions
- Streaming data flows process data as it is accumulated
  - Analyzing traffic to determine the quickest route
  - What tweets are trending right now
  - What products are selling today

Moving Data

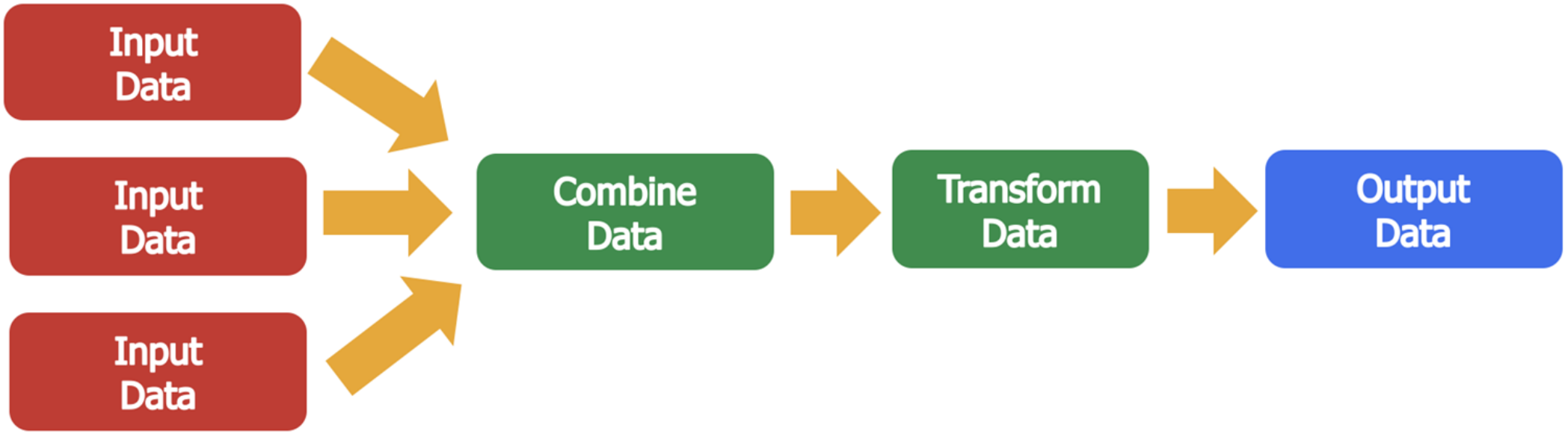
- The simplest data flow is simply moving data from one location to another
  - Hardly a difficult problem unless huge amounts of data are involved

Moving and Transforming Data

- Data is moved from one place to another, but altered along the way
  - More complicated as the amount of data increases and the transformations become more complex
- More data and more processing means more time
  - More machines can participate in the process to get the job done faster
  - This increases cost and complexity

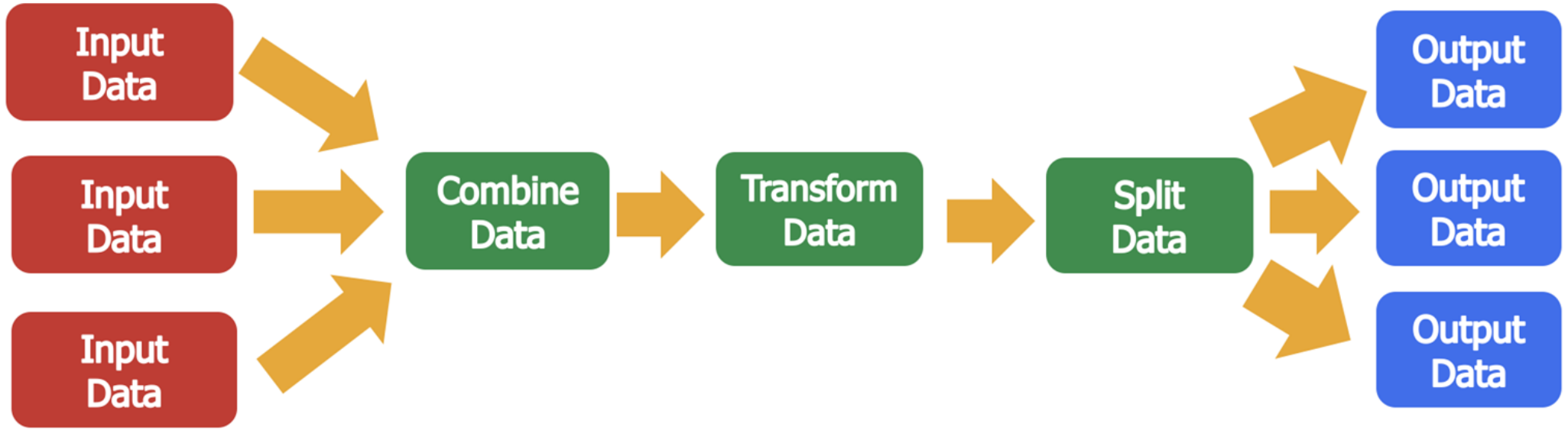
Multiple Inputs

- Sometimes a pipeline requires input from multiple data sources
  - Data from a web log might be combined with sales data from a database



Multiple Outputs

- Sometimes a pipeline needs to output the data more than one way
  - Outputting the same data in multiple, different formats can make data analysis easier and more efficient
  - Storage is cheap relative to processing power and time



Google Cloud Dataflow

- Google Cloud Dataflow consists of two major pieces
  - A managed service for running data flow pipelines
  - Apache Beam SDK for making programming of data flows easier

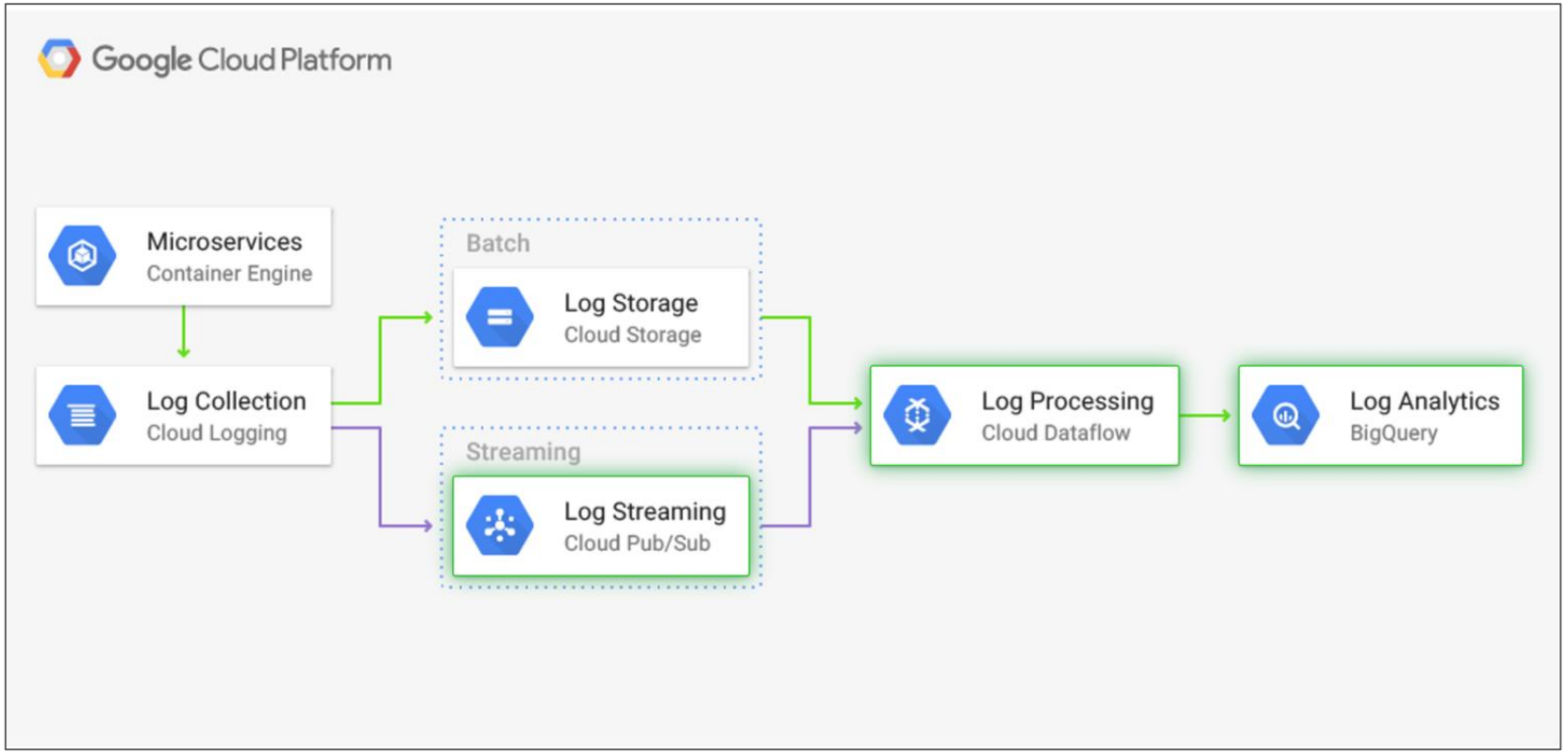
Dataflow

Dataflow Managed Service

- Dataflow pulls together a number of GCP services to run data flows
  - It is the job of the Dataflow service to optimize execution
- Cloud Storage is used as a staging area for data flow code
  - Can also be used for data input and output
- BigQuery tables can be used for input and output
  - BigQuery is frequently the preferred tool to analyze data flow output
- Cloud Compute instances are used to execute data flows
  - The Dataflow service determines how many instances are required
  - Google's high-speed network is used to move data around
- Pub/Sub is used to provide streaming data flows

Processing Batch and Streaming Pipelines

- Dataflow jobs can be used to process both batch and streaming pipelines
  - BigQuery is frequently where the output is written for later analysis



Apache Beam

- Open-source SDK for writing data processing pipelines
- Java and Python versions
- Runs on any platform, not just Google Cloud Dataflow
  - Spark clusters, for example
- See: https://beam.apache.org/

Java Word Count Example

```
Pipeline p = Pipeline.create(options);
p.apply(TextIO.read().from("gs://apache-beam-samples/shakespeare/*"))
.apply("ExtractWords", ParDo.of(new DoFn<String, String>() {
    @ProcessElement
    public void processElement(ProcessContext c) {
        for (String word : c.element().split("[^\\p{L}]+")) {
            if (!word.isEmpty()) {
                c.output(word);
            }}}}))
.apply(Count.<String>perElement())
.apply("FormatResults", MapElements.via(new SimpleFunction<KV<String, Long>, String>() {
    @Override
    public String apply(KV<String, Long> input) {
        return input.getKey() + ": " + input.getValue();
    }}))
.apply(TextIO.write().to("wordcounts"));
p.run().waitUntilFinish();
```

Python Word Count Example

```
p = beam.Pipeline(options=options)
p | beam.io.ReadFromText('gs://dataflow-samples/shakespeare/kinglear.txt')
  | 'ExtractWords' >> beam.FlatMap(lambda x: re.findall(r'[A-Za-z\']+', x))
  | beam.combiners.Count.PerElement()
  | beam.Map(lambda (word, count): '%s: %s' % (word, count))
  | beam.io.WriteToText('gs://my-bucket/counts.txt')

result = p.run()
```

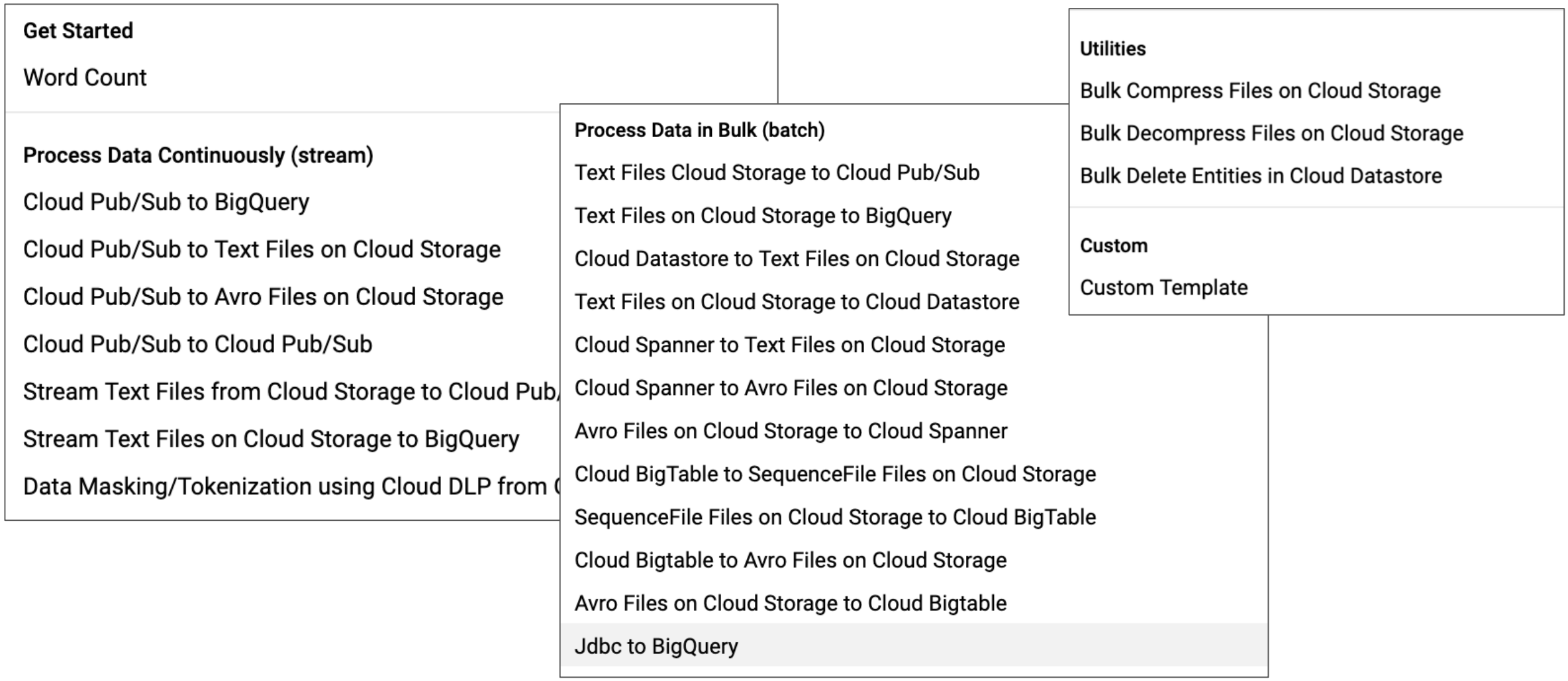
Exercise: Google Cloud Dataflow

- In this exercise, you will run an Apache Beam pipeline using Google Cloud Dataflow
  - Dataflow Quickstart using Python
  - Dataflow Quickstart using Java

Dataflow Templates

- Premade templates for creating Dataflow jobs with no coding
- Support both streaming and batch operations
- Support many different sources and sinks
  - Pub/Sub
  - Storage
  - BigQuery
  - Datastore
  - Spanner JDBC
  - Etc.

Dataflow Templates Illustrated



Exercise: Dataflow Templates

- In this exercise, you will create a Dataflow job using a template:
  - Quickstart Using Templates

Chapter Concepts

Google Cloud DataFlow

Pub/Sub

Cloud Dataprep

Cloud Composer

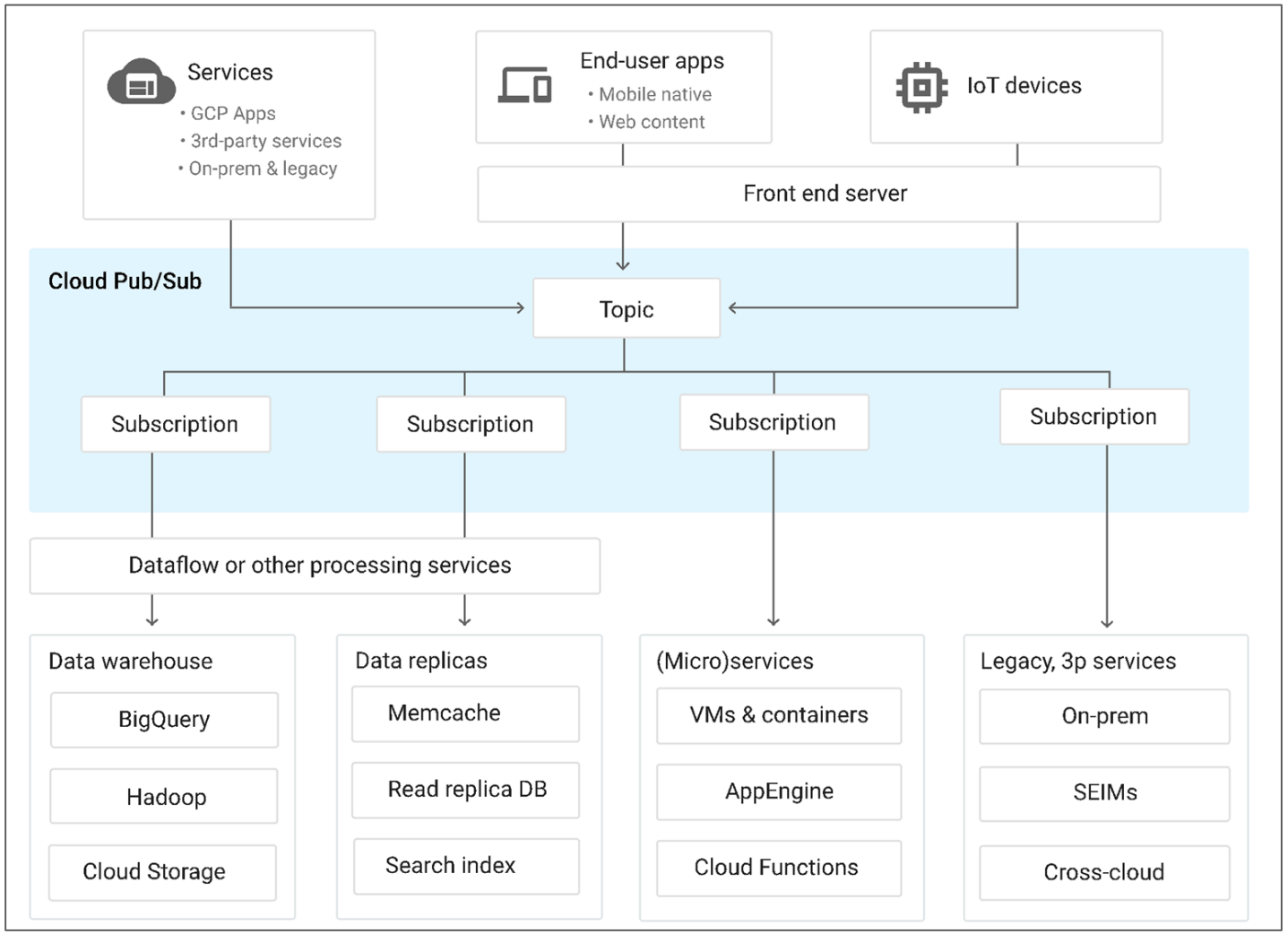
Data Loss Prevention

Exam Prep

Pub/Sub

- Pub/Sub is a fully managed, massively scalable messaging service
  - It allows messages to be sent between independent applications
  - Can scale to millions of messages per second
- Pub/Sub messages can be sent and received via HTTP(S)
- Pub/Sub supports multiple senders and receivers simultaneously
- Pub/Sub is a global service
  - Messages are copied to multiple zones for greater fault tolerance
  - Uses dedicated resources in every region for fast delivery worldwide
- Pub/Sub is secure
  - All message are encrypted at rest and in transit

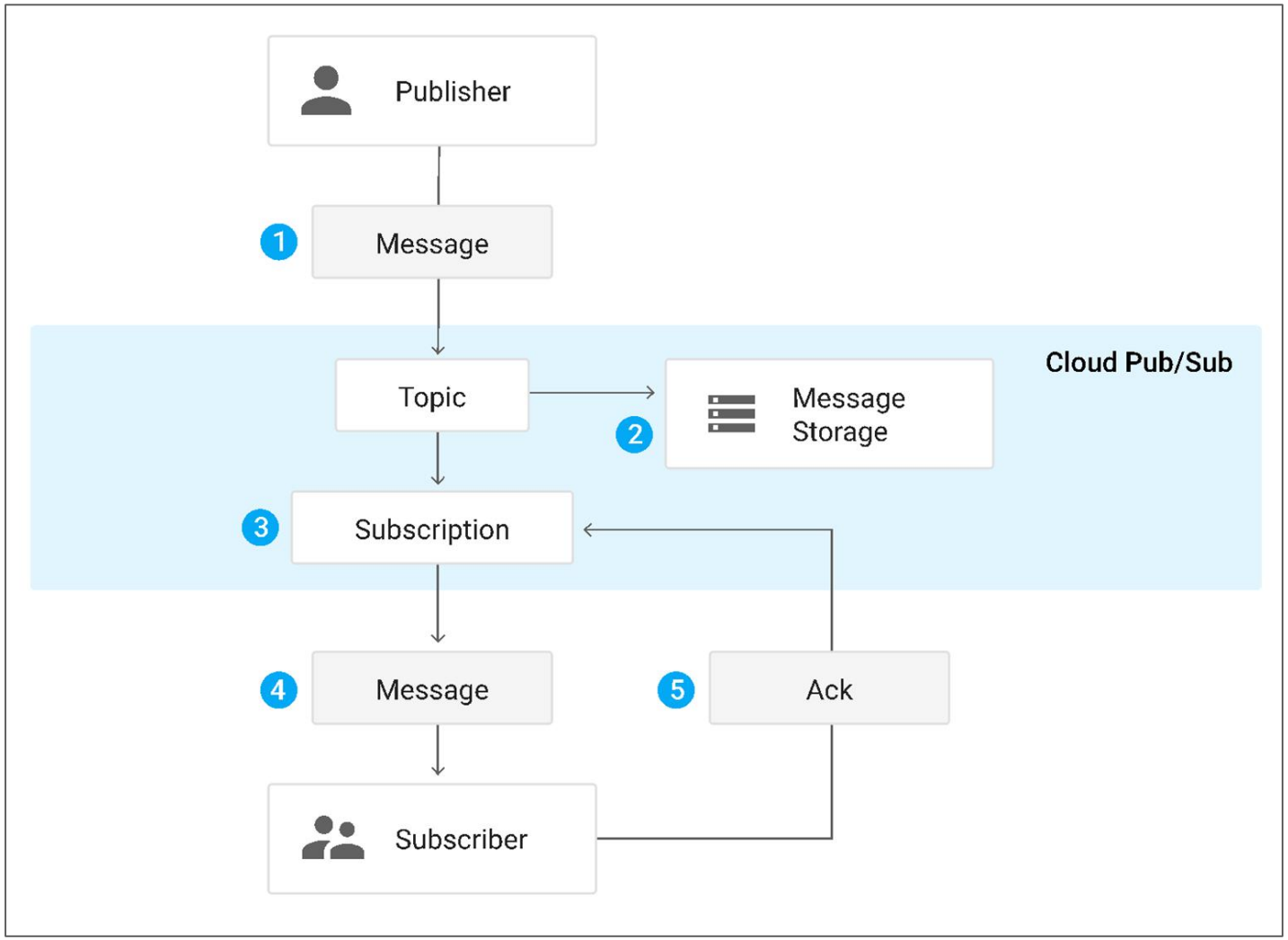
Integrating Pub/Sub with Other Services



Topics and Subscriptions

- Messages in Pub/Sub are sent to a Topic
  - Messages can contain data and attributes
- Topics are named endpoints where messages are sent
  - Topic names are in the form:
    `projects/<project-id>/topics/<topic-name>`
- Subscriptions represent a stream of messages within a topic
  - Topics can contain multiple subscriptions
  - Each subscription belongs to one topic
  - Subscribers get messages from subscriptions

Pub/Sub Message Flow



Subscribers

- Subscribers are applications that process Pub/Sub messages
  - Subscribers get messages from a subscription
- Two types of subscriptions, push and pull
- Push messages are automatically sent to the subscriber via an endpoint
  - Acknowledgement of the message is implied by a response code 200
- Pull messages must be requested by the subscriber
  - Subscriber calls the `pull()` method of the Pub/Sub API
  - If a message exists, it is sent
  - Subscriber then calls the `acknowledge()` method

Push or Pull Subscribers

- Push subscribers must be web servers that support HTTPS
  - Must expose an endpoint to receive the message (a webhook)
  - Delivery is immediate unless throttled
  - Can be load balanced
- Pull subscribers can be any type of application
  - Must be able to use the Pub/Sub REST API
- App Engine applications are ideal push subscribers
- Dataflow jobs are pull subscribers

Using Pub/Sub with Dataflow

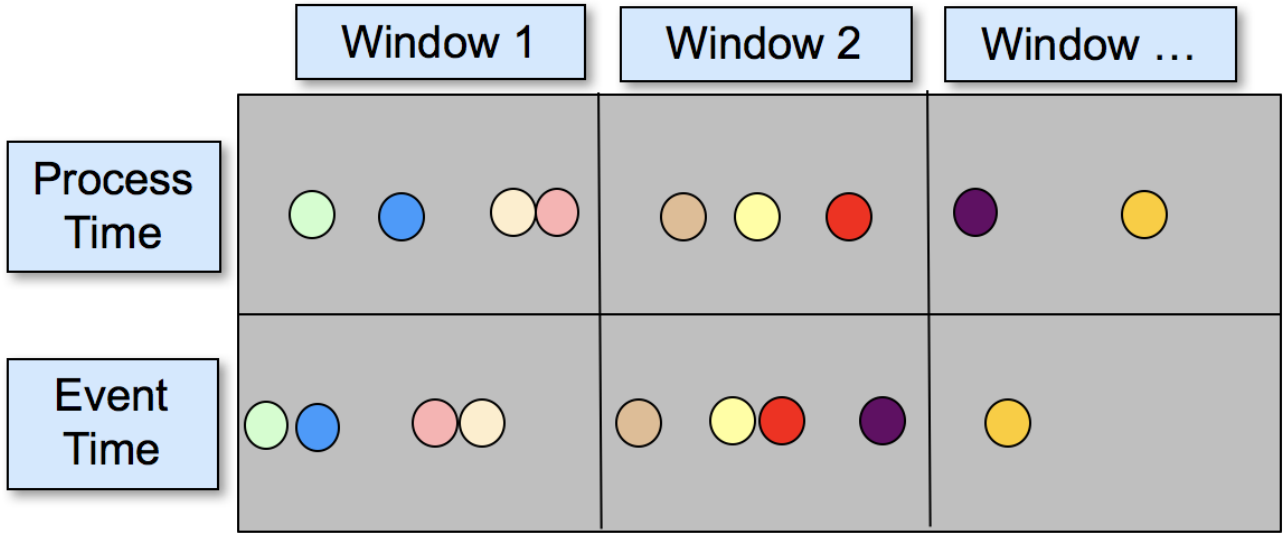
- Pub/Sub can be utilized within a Dataflow pipeline
  - PubSub.IO class can read or write to a Pub/Sub topic
- Allows for streaming data flows in real time
  - Pipeline is constantly being sent messages for processing
- Can set up windows to process groups of messages by time
- When using Pub/Sub, data sets are unbounded
  - They continue to grow over time
  - Aggregate operations must be run periodically

Event Time vs. Process Time

- Event time is when something actually occurred
  - The time an order is placed for example
  - This is when the message is published to Pub/Sub
- Process time is when the system observes the event
  - When the subscriber receives the Pub/Sub message
- Obviously, process time is always after event time
  - Usually, this is a short period of time
  - Sometimes, a system problem will delay the process time
- The difference between event and process times can vary significantly
  - Event occurs in a mobile application when the user is on an airplane

Windowing

- Groups data into chunks that aggregate functions are run against
  - Often based on time, but can also be based on session
  - Time-based windows can use either event or process time
- Sometimes the process time will fall into a different window than the event
  - What do you notice about the purple data point?



Windowing in Dataflow

- Dataflow has built-in support for three types of windows
  - Shuffles arriving messages into the correct window based on event time
- Fixed time windows are the simplest
  - Each window is given an interval
- Sliding time windows have an interval and a period
  - The interval defines how long a window collects data for
  - The period defines how often a new window starts
- Session windows define windows on areas of concentrated data
  - Uses a key to combine data into groups
  - Like sessions in a web application

Chapter Concepts

Google Cloud DataFlow

Pub/Sub

Cloud Dataprep

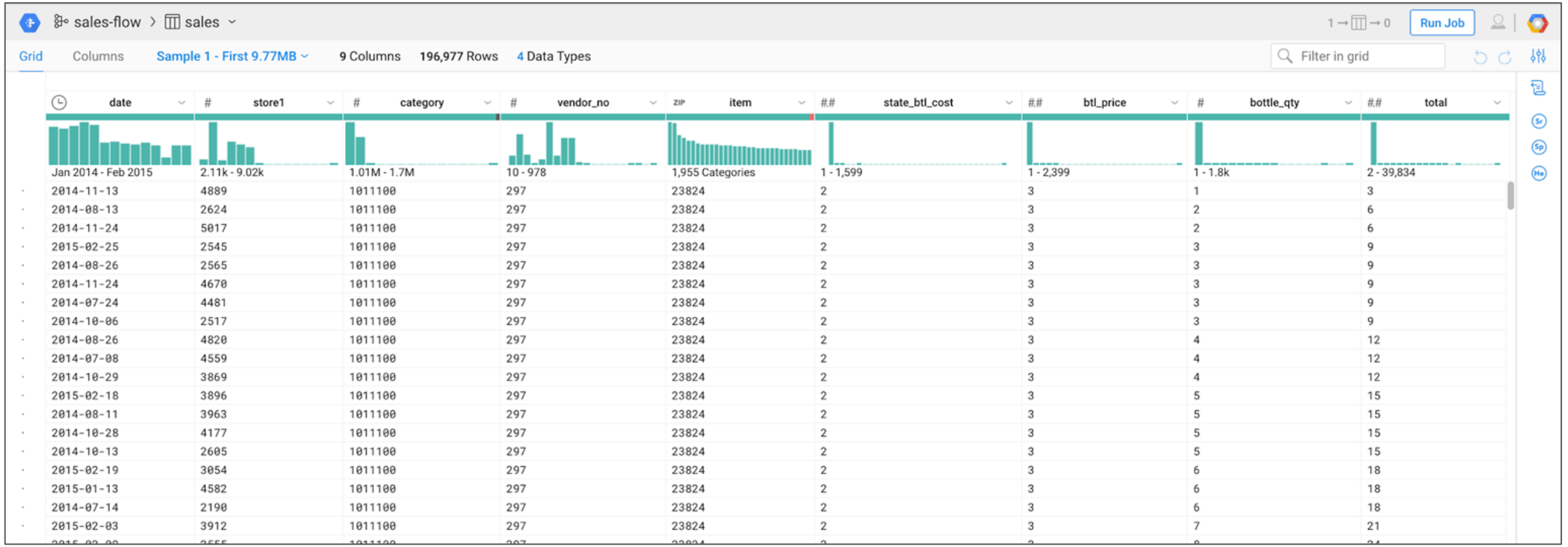
Cloud Composer

Data Loss Prevention

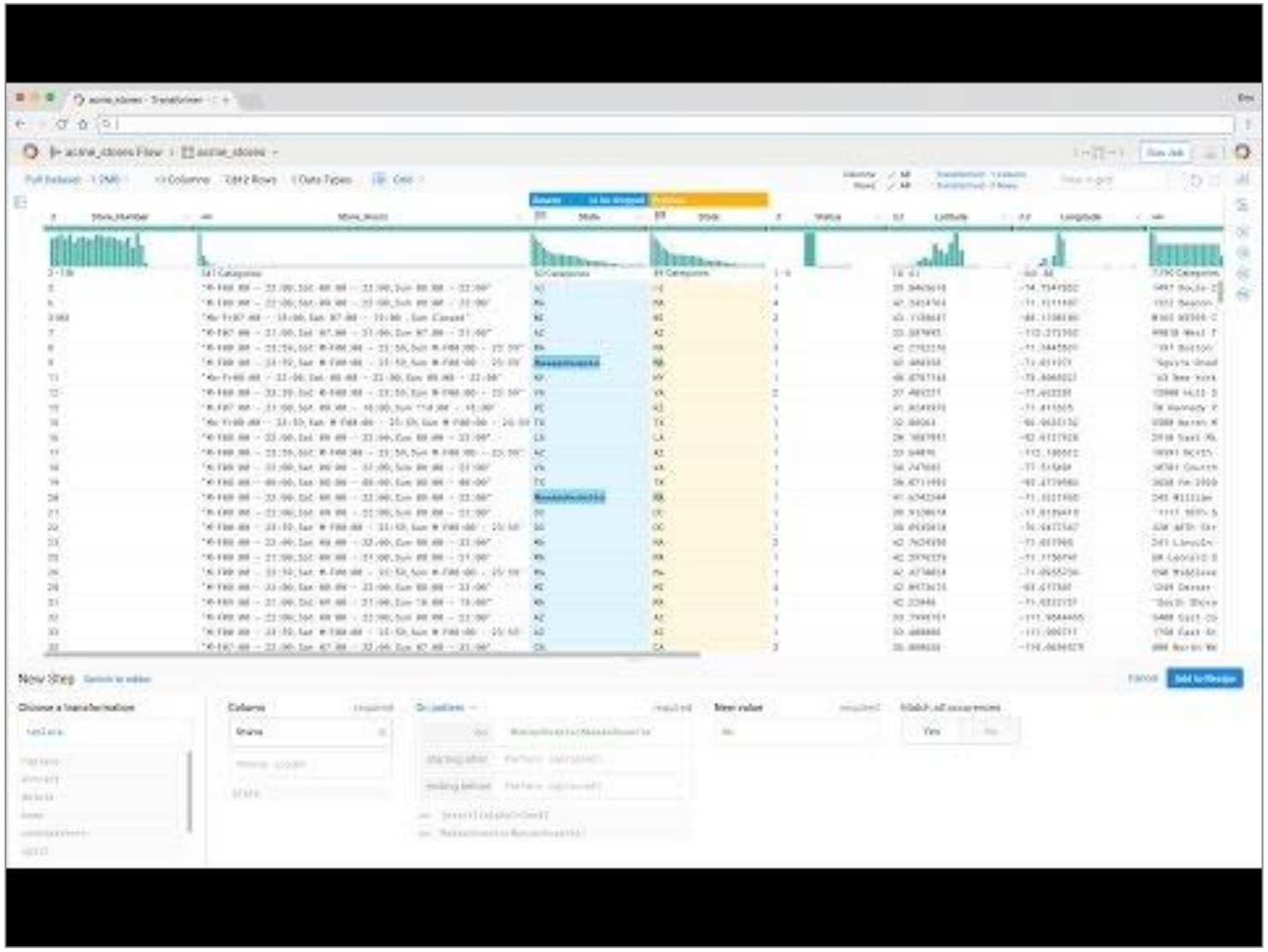
Exam Prep

Cloud Dataprep

- Visual tool for cleaning and manipulating data
- Serverless, runs on top of Dataflow
- Can process very large datasets



Demo: Dataprep Video



Chapter Concepts

Google Cloud DataFlow

Pub/Sub

Cloud Dataprep

Cloud Composer

Data Loss Prevention

Exam Prep

Cloud Composer

- Workflow orchestration service based on Apache Airflow
  - Can orchestrate workloads across GCP, on-prem, or other clouds
  - Uses Python as orchestration language
- Built-in connectors for many GCP services
  - Dataproc, Cloud MLE, GCS, Pub/Sub, BigQuery, Dataflow, etc.

Exercise: Cloud Composer Quickstart

- In this exercise, you will create and run an Apache Airflow workflow in Cloud Composer
  - Cloud Composer Quickstart

Chapter Concepts

Google Cloud DataFlow

Pub/Sub

Cloud Dataprep

Cloud Composer

Data Loss Prevention

Exam Prep

Data Loss Prevention (DLP)

- API to help classify and redact sensitive data
  - Helps customers meet their compliance obligations
  - Works with image or text data
- Data can be in GCP, other clouds, or on-prem
  - Built-in support for BigQuery, Datastore, and GCS
- 90-plus built-in detectors for common sensitive data items
  - E.g., credit card numbers
- Detectors can be customised to classify/redact new data items
  - E.g., social security numbers in a particular country's format

Exercise: DLP Quickstart

- In this exercise, you will perform basic tasks with the DLP API at the command line:
  - Cloud DLP API Quickstart

Chapter Concepts

Google Cloud DataFlow

Pub/Sub

Cloud Dataprep

Cloud Composer

Data Loss Prevention

Exam Prep

Links

- Dataflow
- Pub/Sub
- Cloud Composer
- Cloud Data Loss Prevention

ROITRAINING
MAXIMIZE YOUR TRAINING INVESTMENT

Google Cloud Certification Workshop—Data Engineer

Chapter 8: Analytics and Visualization

Chapter Objectives

In this chapter, we will learn how to:

- Analyze data using Google Cloud Datalab
- Create reports and visualizations using Data Studio

Chapter Concepts

Datalab

Data Studio

Exam Prep

Cloud Datalab

- Interactive tool for data analysis, machine learning, and many other tasks
- Based on Jupyter, an open-source project for creating iPython Notebooks
- Supports many languages: Python, JavaScript, Shell scripts, HTML, SQL, etc.
- Integrated with GCP, so access to other services like BigQuery are simple
- Integrates with Git to enable collaboration and sharing notebooks
- Runs in a Compute Engine Virtual Machine
  - No extra charge for Datalab beyond the machine cost

Managing Datalab Instances

- Manage Datalab instances using the Google Cloud SDK in Cloud Shell
  - To create a Datalab instance:

    `datalab create` *`instance-name`*
  - To stop an instance without deleting it:

    `datalab stop` *`instance-name`*
  - To restart a stopped instance:

    `datalab connect` *`instance-name`*
  - To delete an instance:

    `datalab delete` *`instance-name`*
- For more options and details on using the SDK see: https://cloud.google.com/datalab/docs/how-to/lifecycle

Do Now: Create a Datalab Instance

1. Log onto the GCP Management Console and go to a project (if you don't have a project, create one)
2. Run the Datalab Quickstart
3. Before deleting the your Datalab instance, create a new notebook and enter code as shown on the following slide
   - Click the **Run** button after every code block

Do Now: Create a Datalab Instance (continued)


```
def to_celsius(temp):
    return (temp - 32.0) * 5.0 / 9.0

print to_celsius(212)
```

```
100.0
```

```
%javascript

function to_fahrenheit(temp){
  return temp * 9.0 / 5.0 + 32.0;
}

element.text(to_fahrenheit(100));
```

```
212
```

```
%%bq query

#standardsql
SELECT name, sum(total) as sales
FROM `drehnstrom-1171.sales_dataset.stores` as stores
JOIN `drehnstrom-1171.sales_dataset.sales` as sales
ON stores.store = sales.store
GROUP BY name
ORDER BY sales DESC
LIMIT 10
```

name	sales
Hy-vee #3 / Bdi / Des Moines	13,920,087.22
Central City 2	11,942,399.97
Sam's Club 6344 / Windsor Heights	6,159,480.06
Sam's Club 8162 / Cedar Rapids	5,734,721.57
Hy-vee Wine and Spirits / I	5,665,143.7
Costco Wholesale #788	4,907,465.88

Copy-Paste code is shown in the Notes section below

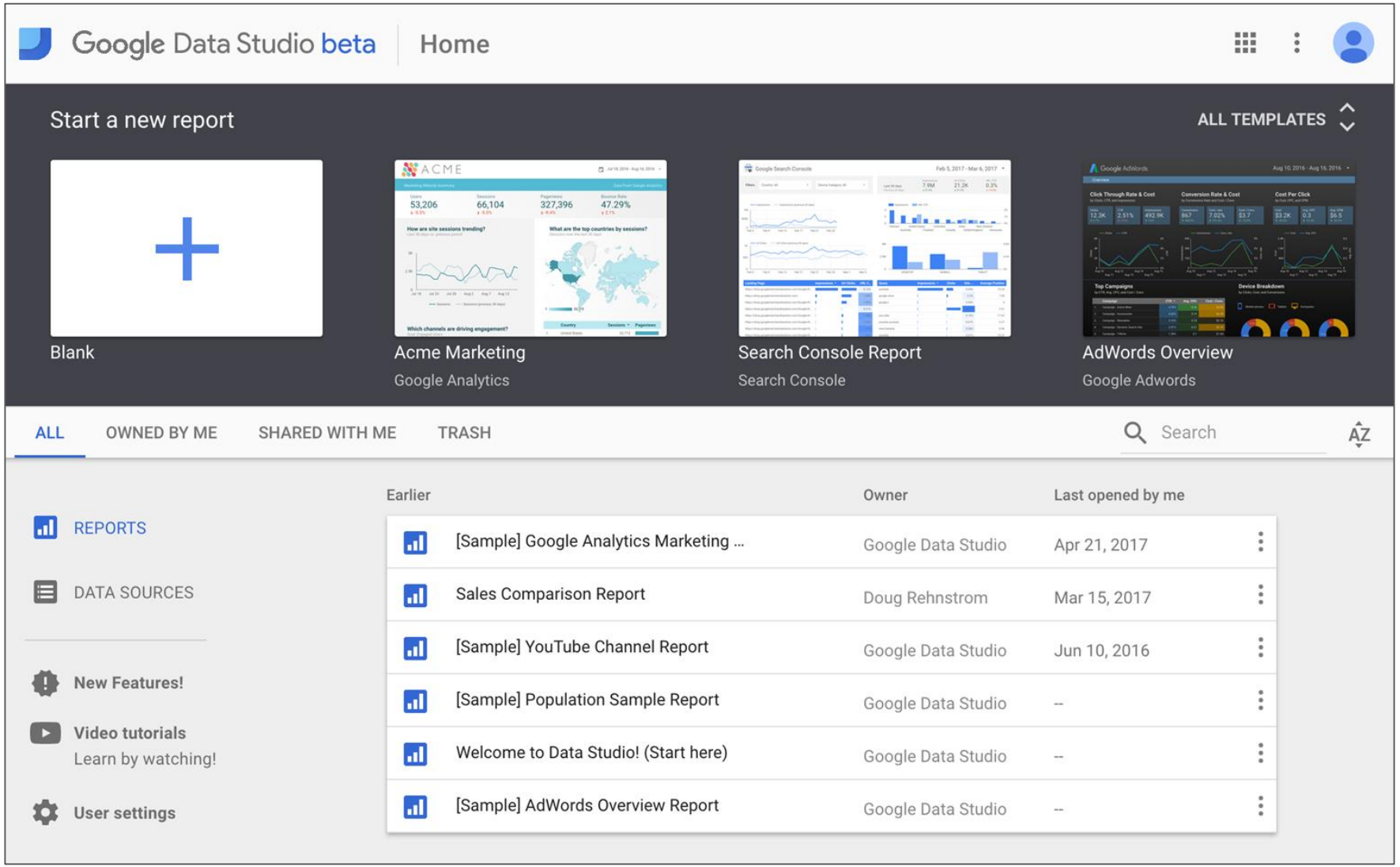
Chapter Concepts

Datalab

Data Studio

Exam Prep

Google Cloud Data Studio



Data Studio

- Reporting and visualization tool
  - Powerful chart and graphing tools built in
- Nothing to install, runs in the browser
- Supports many data sources
  - BigQuery, Cloud SQL, MySQL, Google Sheets, file uploads, and many more
- FREE!

Do Now: Data Studio Examples

- Go to the following URL and explore the examples: https://www.google.com/analytics/data-studio/gallery/

Exercise: Data Studio

- In this exercise, you will create report using Data Studio:
  - Visualizing BigQuery data with Data Studio

Chapter Concepts

Datalab

Data Studio

Exam Prep

Links

- Datalab
- Data Studio

ROITRAINING
MAXIMIZE YOUR TRAINING INVESTMENT

Google Cloud Certification Workshop—Data Engineer

Chapter 9: Machine Learning Basics

Chapter Objectives

In this chapter, we will learn how to:

- Learn basic machine learning terminology
- Choose Linear Regression, Classification, or Deep Neural Networks for different machine learning use cases

Chapter Concepts

Machine Learning Overview

Machine Learning Algorithms

Machine Learning

- Programming computers to make decisions using non-traditional techniques
  - Algorithms are built based on math and statistics
  - Not if-statements and loops
- Allows complex algorithms that would not otherwise be possible
  - Image and speech recognition
  - Sentiment analysis
- Can be used to enhance and improve traditional algorithms
  - Google Search now uses machine learning as part of its algorithm
  - Spam detection has been greatly improved using ML

Machine Learning Steps

- Gather data (*lots of it*)
- Train the model
- Calculate the accuracy of the model
- Use the model
- Repeat to improve the model accuracy

Gathering ML Data

- A significant amount of data is required to do machine learning
- Based on the data, can we find a problem we can answer?
  - Is a transaction fraudulent?
  - How much should a product cost?
  - Is an email spam?
  - Is there a cat in that picture?

Targets, Features, and Examples

- A target is what we want to train the computer to predict
  - Target is also referred to as a Label
- Features are the inputs that are used to build the model
  - If the target is the price of an apartment, sq. footage might be a feature
  - What are other features be that could predict the price of a apartment?
- Each input consisting of a target and its features is an example

Quiz

- Which are the Target, Features, and Examples?

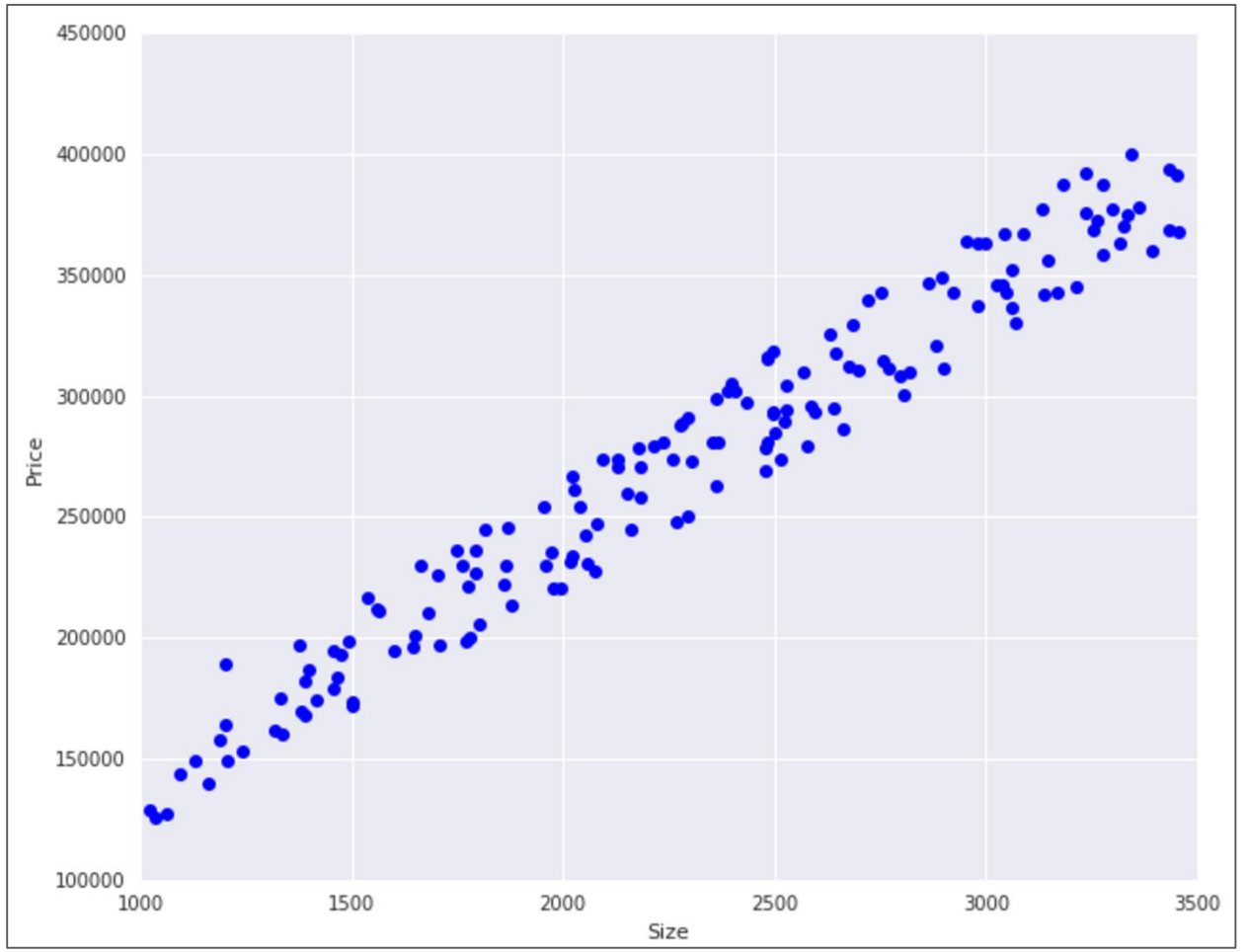
Size	Bedrooms	Price
1000	2	1000
1200	2	1400
600	1	750
1500	3	1400

Feature Tuning

- Feature tuning is the process of analyzing the data to come up with good inputs for training a model
  - Is the feature relevant to the problem?
  - Are there anomalies or errors in the data?
  - Will the feature be known at run time?
  - Etc.

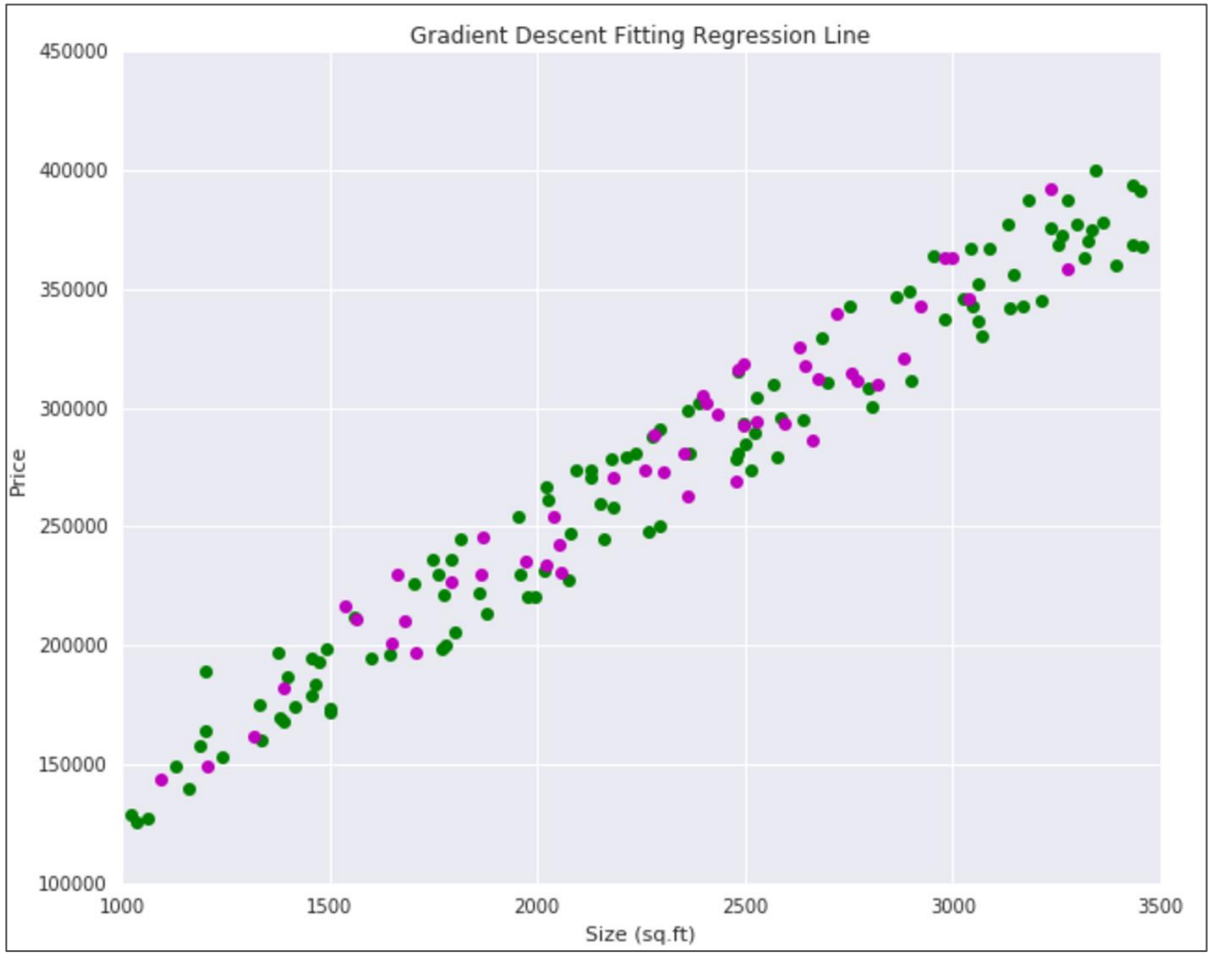
Gather Data

1. Start with some data



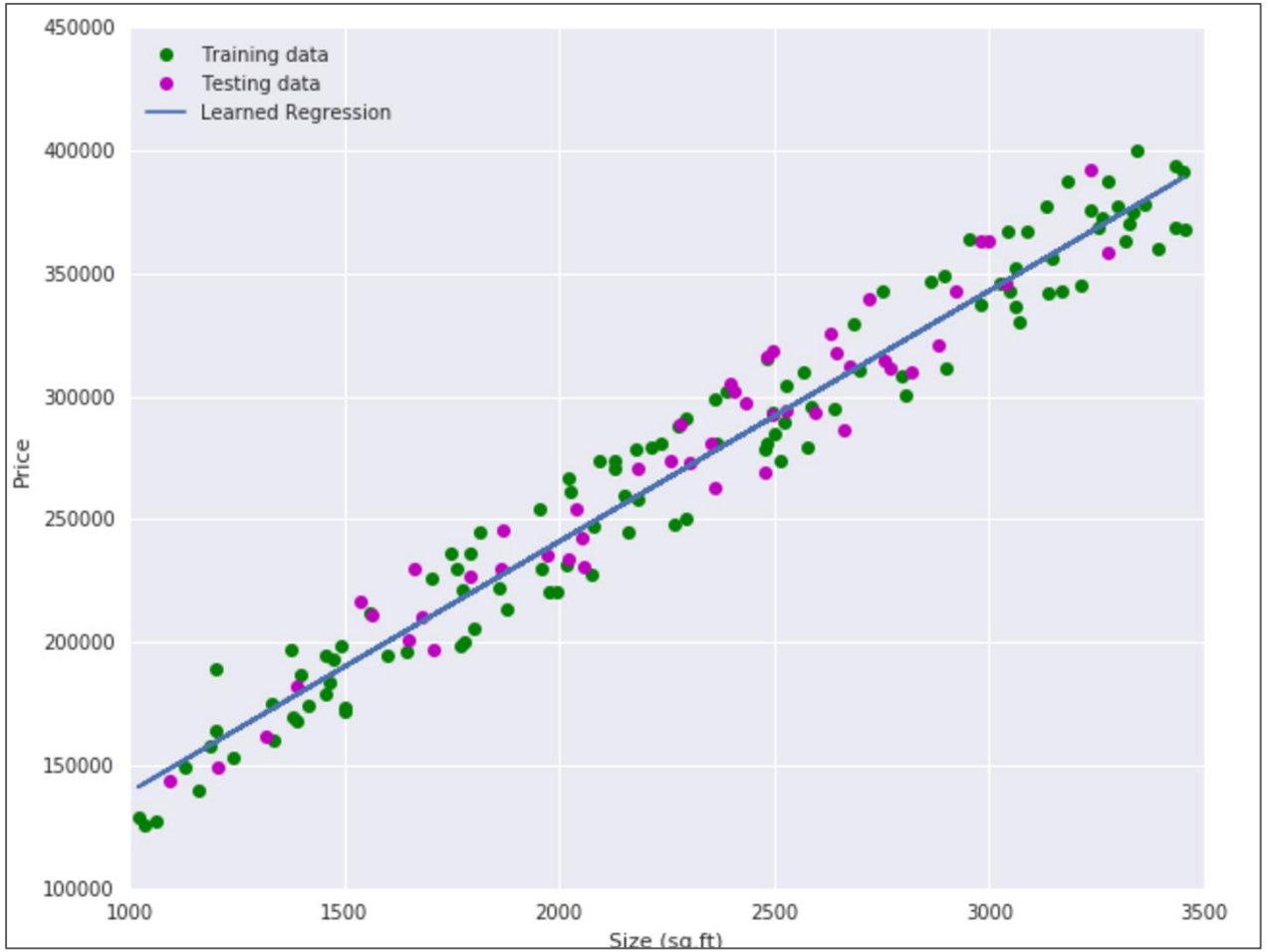
Split Data

2. Split the data into training and test data
   - The model is built with the training data
   - The model is evaluated with the test data



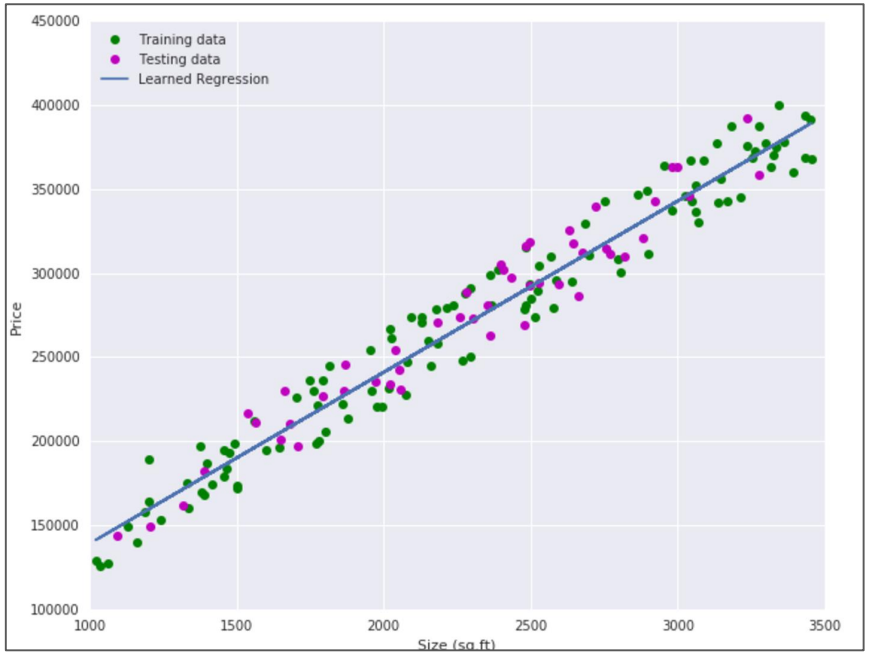
Train the Model

3. Draw a line
   - Calculate the accuracy of the line for both the training and test data
   - Draw the line over and over looking for the most accurate line



Calculating the Accuracy of Model (RMSE)

- Need to calculate how accurate a line is
- Calculate each point's distance from the line
- Some are positive and some are negative, so square them
- Sum the squares
- Divide by the number of points
  - This gives the Mean Squared Error (MSE)
- Take the square root of the MSE (*because the errors were all squared*)
- This gives Root Mean Squared Error (RMSE)



Gradient Descent

- Technique is to find a line with the best fit
- Keep moving the line slightly in one direction or the other
  - With each move, calculate the RMSE
  - If the RMSE goes down, move a little more in the same direction
  - If the RMSE goes up, move a little in the other direction

Weighing the Features

- Each feature has an optimal line that fits the target
- Different features can be given different weights
  - Some features are more important than others
  - The sum of the weighted features gives the prediction
- There are a huge number of calculations required to build the model
  - That's where the computer comes in!

Chapter Concepts

Machine Learning Overview

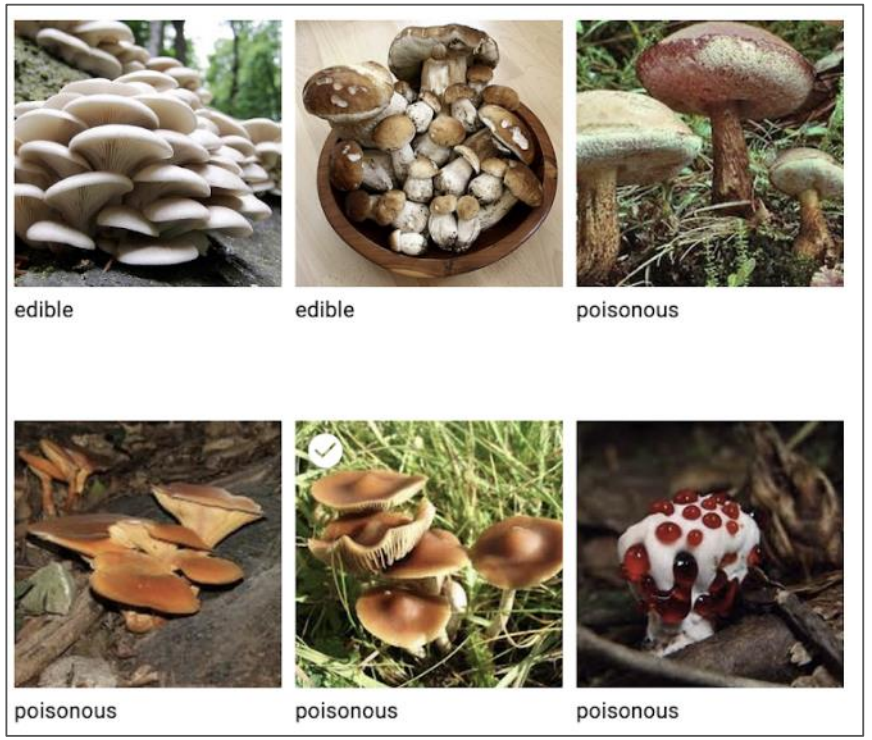
Machine Learning Algorithms

Linear Regression

- Used to predict a value along a curve
  - The apartment rental price example is linear regression
- What are other examples of linear regression?

Classification

- Used to predict whether the target is one of a finite number of values
  - Spam or not spam
  - Positive or negative comment
  - Edible or poisonous

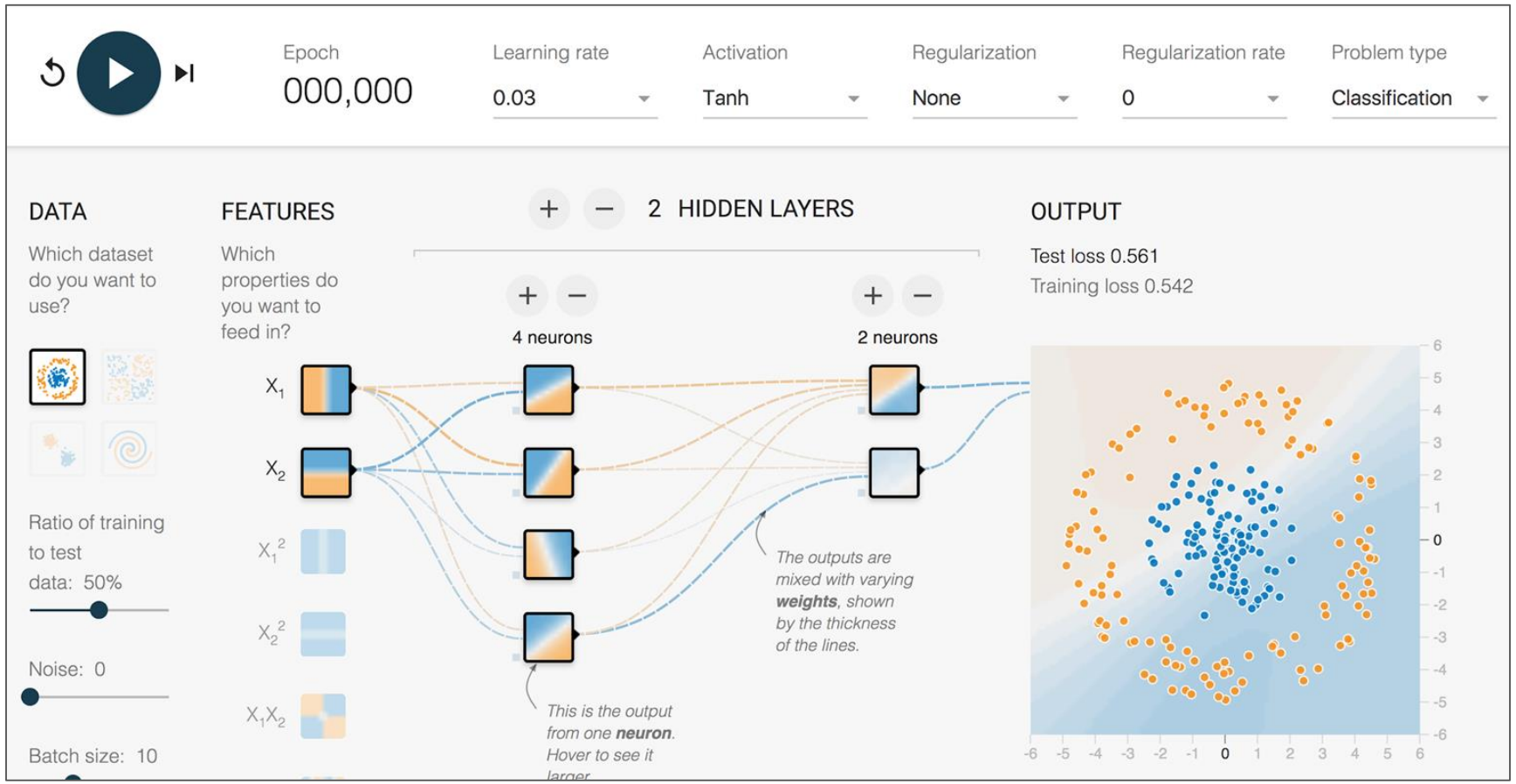


Neural Networks

- Neural networks consist of a collection of decision points
  - Each decision is called a neuron
  - Neuron are grouped into layers
- The data passes from layer to layer
  - Each neuron makes one decision
  - The model assigns weights each neuron
- Complex models can have huge numbers of layers and neurons
- Training is used to adjust each neuron and assign the weights

Do Now: Neural Networks

- Go to http://playground.tensorflow.org/



ROITRAINING
MAXIMIZE YOUR TRAINING INVESTMENT

Google Cloud Certification Workshop—Data Engineer

Chapter 10: Google Machine Learning Tools

Chapter Objectives

In this chapter, we will learn how to:

- Write machine learning code using TensorFlow
- Simplify and scale machine learning training with Google Cloud ML Engine
- Leverage Google’s pre-trained machine learning models
- Extend Google's pre-trained models with AutoML
- Use BigQuery ML to create machine learning models

Chapter Concepts

TensorFlow

Cloud MLE

Pre-Built ML Models

AutoML

BigQuery ML

Exam Prep

Machine Learning Frameworks

- Building machine learning models requires advanced math and statistics
  - Different models require different algorithms and optimizations
- Luckily, the math has been done and programmed into a number of machine learning frameworks
  - Allows machine learning models to be built without coding the math
- Common ML frameworks include:
  - Scikit-learn
  - Spark MLlib
  - TensorFlow
  - And others

TensorFlow

- Open-source library for machine learning originally created at Google
- Used to train Google's internal machine learning models
- High-level API exposed Python API
- Built on top of C++ library for speed
- Runs on any platform, including mobile devices
- The same code runs on CPUs, GPUs, or TPUs

Do Now: Basic TensorFlow Code

- Click the URL below and read through the TensorFlow code
  Getting Started with Tensorflow

Using TensorFlow to Train Models

```
import tensorflow as tf
from tensorflow import keras
```

```
model = keras.Sequential([
    keras.layers.Flatten(input_shape=(28, 28)),
    keras.layers.Dense(128, activation=tf.nn.relu),
    keras.layers.Dense(10, activation=tf.nn.softmax)
])
model.compile(optimizer='adam',
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])
model.fit(train_images, train_labels, epochs=5)
```

```
test_loss, test_acc = model.evaluate(test_images, test_labels)
```

```
predictions = model.predict(test_images)
```

Exercise: Getting Started with Tensorflow

- In these exercise, you will learn the basics of using Tensorflow for machine learning
  - Getting Started with Tensorflow

Chapter Concepts

TensorFlow

Cloud MLE

Pre-Built ML Models

AutoML

BigQuery ML

Exam Prep

Google Cloud ML Engine

- Cloud-based service used to build machine learning models
- Used by Google to build their own models
- Fully-managed NoOps service
- Supports training on CPUs, GPUs, and TPUs
- Can be used to deploy models as services
- Command-line API for submitting jobs, deploying models, and making predictions

Cloud ML Command Examples

```
gcloud ml-engine jobs submit training my_job \
          --module-name trainer.task \
          --staging-bucket gs://my-bucket \
          --package-path /my/code/path/trainer \
          --packages additional-dep1.tar.gz,dep2.whl
```

```
gcloud ml-engine jobs submit prediction JOB --data-format=DATA_FORMAT
--input-paths=INPUT_PATH,[INPUT_PATH,…] --output-path=OUTPUT_PATH --
region=REGION
```

```
gcloud ml-engine models create MODEL [--enable-logging] [--
regions=REGION,[REGION,…]] [GCLOUD_WIDE_FLAG …]
```

Homework: Learning Cloud MLE

- Run the following tutorials:
  - Quickstart using the Command Line
  - Quickstart using Datalab

Chapter Concepts

TensorFlow

Cloud MLE

Pre-Built ML Models

AutoML

BigQuery ML

Exam Prep

Using Google's Pre-Built ML Models

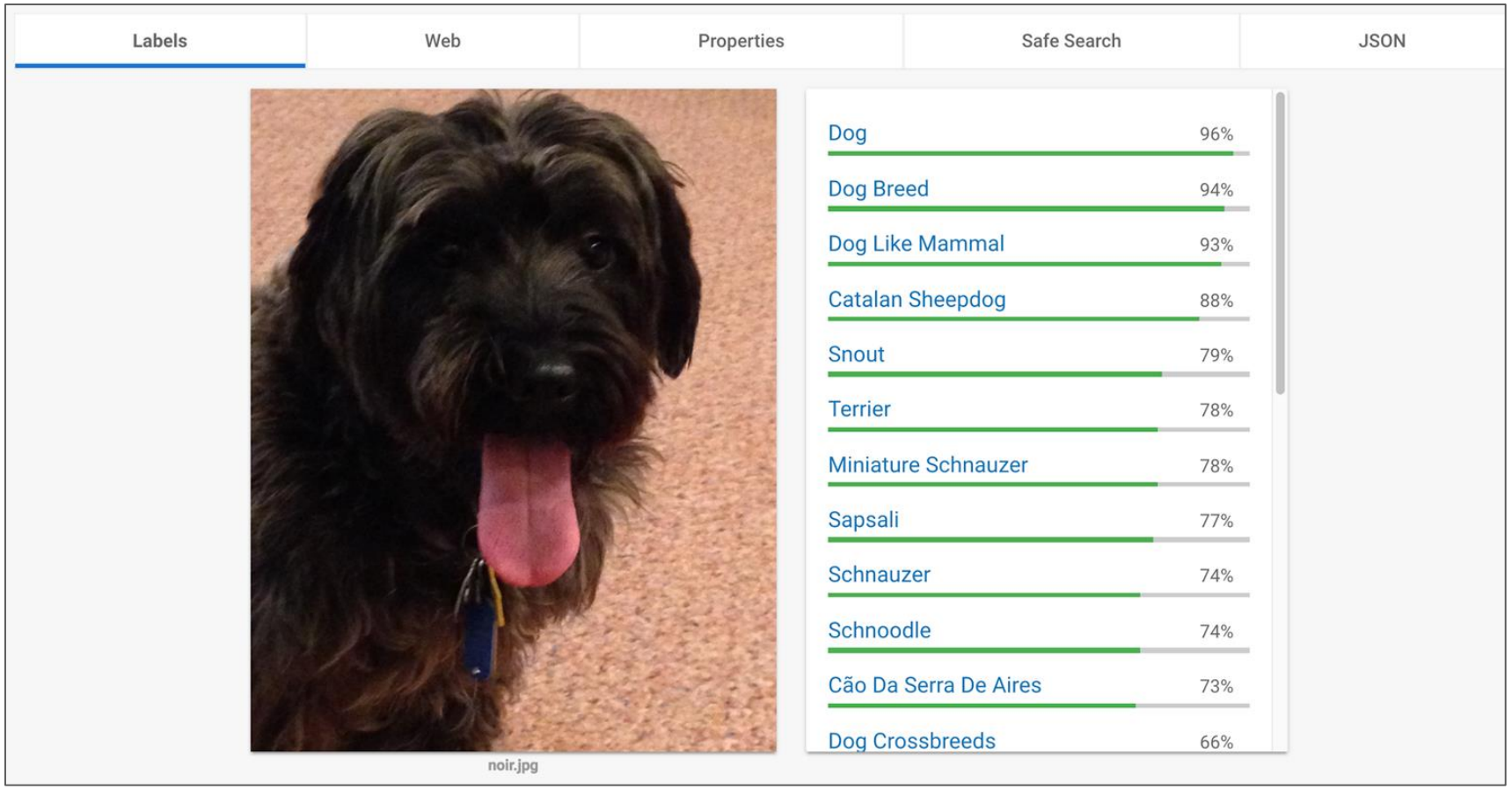
- Google has a number of already trained models that you can use today
  - Reasonably priced
  - Extremely fast
  - Simple, consistent coding
- Prebuilt models include:
  - Vision API
  - Natural Language API
  - Speech API
  - Translate API
  - And others

Vision API Sample Code

```
from google.cloud import storage, vision
vision_client = vision.ImageAnnotatorClient()

def __check_image(file_data):
    file_name = file_data['name']
    bucket_name = file_data['bucket']
    blob_uri = f'gs://{bucket_name}/{file_name}'
    blob_source = {'source': {'image_uri': blob_uri}}

    result = vision_client.safe_search_detection(blob_source)
    detected = result.safe_search_annotation

    if detected.adult == 5 or detected.violence == 5:
        return False
    else:
        return True
```

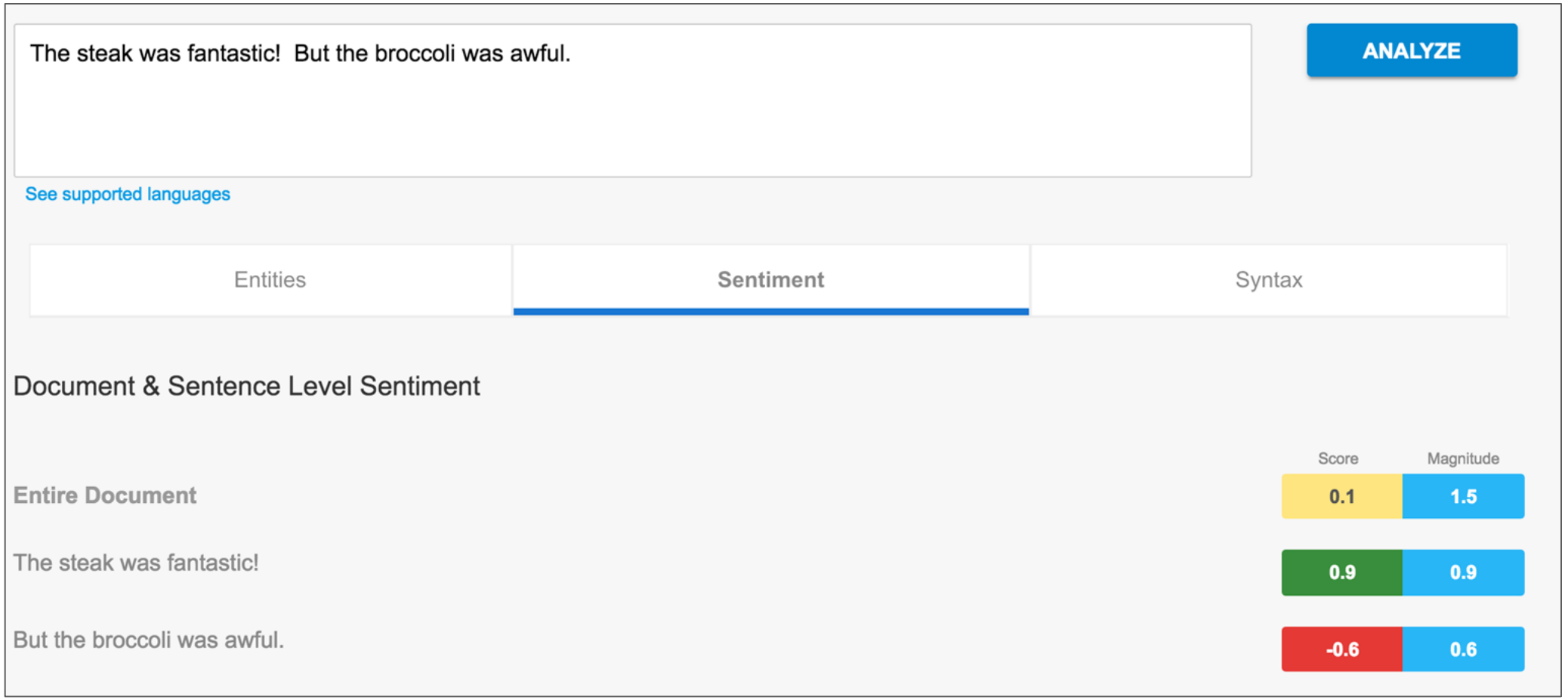
Vision API

- https://cloud.google.com/vision/



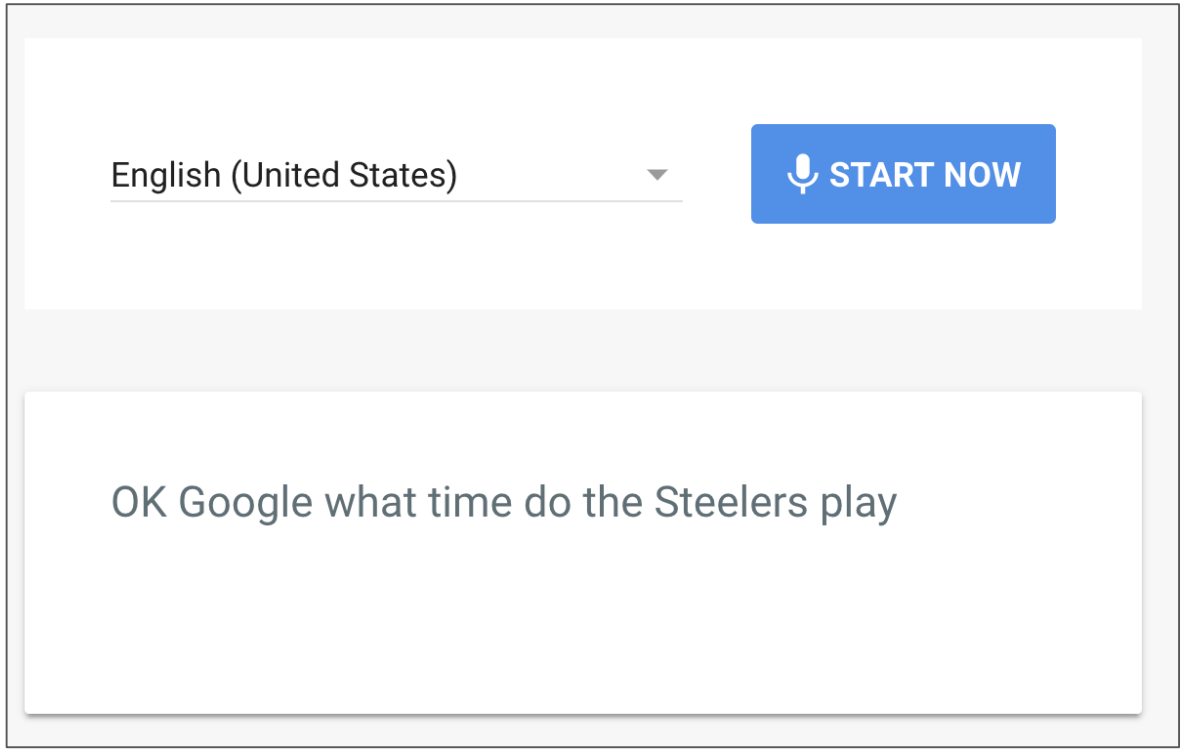
Natural Language API

- https://cloud.google.com/natural-language/



Speech API

- https://cloud.google.com/speech/



Translate API

- https://cloud.google.com/translate/

```
const Translate = require('@google-cloud/translate');

exports.translate = (event, callback) => {
  const pubsubMessage = event.data;
  const target = 'es'
  const textToTranslate = Buffer.from(pubsubMessage.data, 'base64').toString()
  translate_client.translate(textToTranslate, target)
  .then(results => {
    const translation = results[0];
    console.log(`Text: ${textToTranslate}`);
    console.log(`Translation: ${translation}`);
  });
};
```

Exercise: GCP Machine Learning

- In this exercise, you will use Google's pre-built machine learning APIs
  - GCP Machine Learning Pre-Built Models

Chapter Concepts

TensorFlow

Cloud MLE

Pre-Built ML Models

AutoML

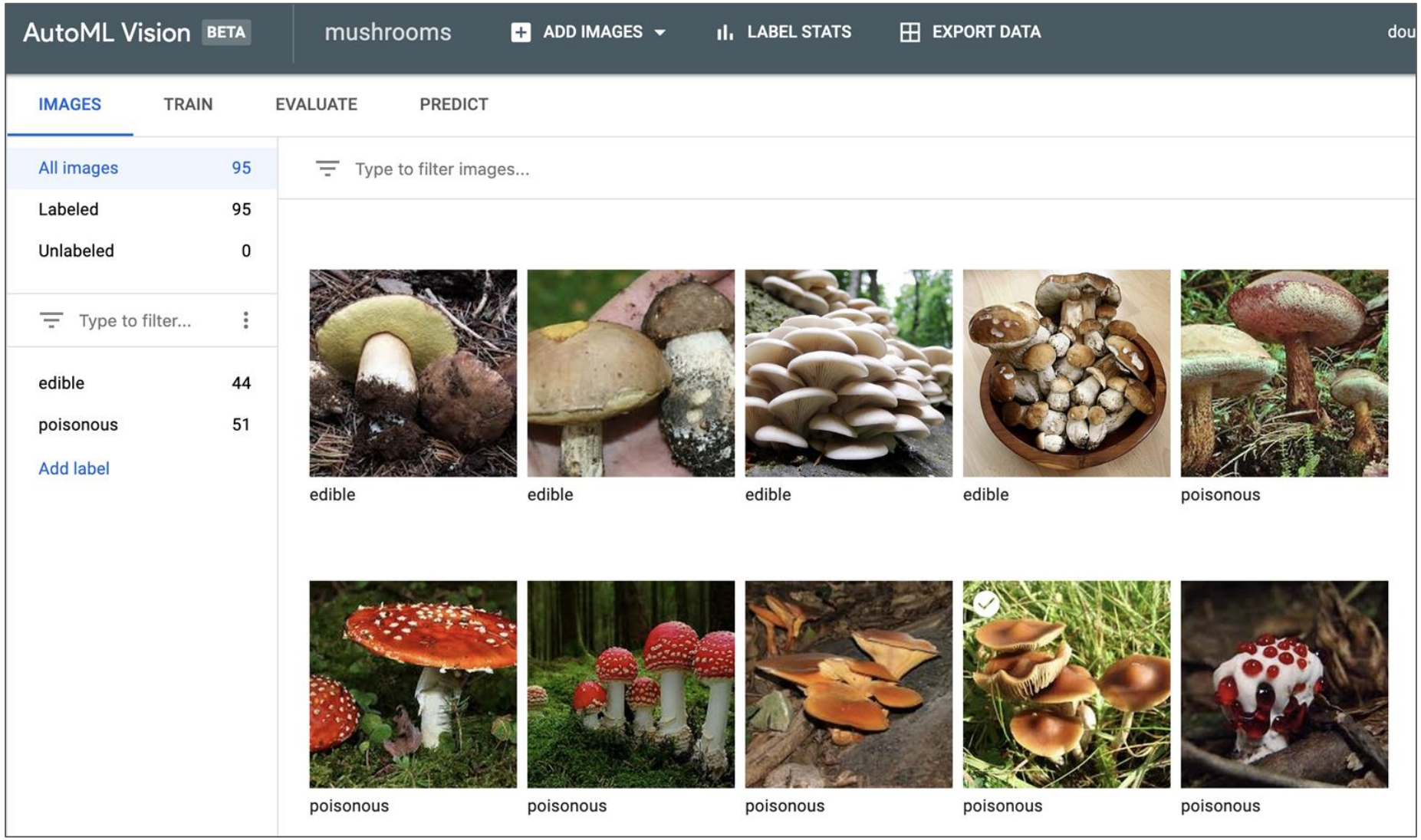
BigQuery ML

Exam Prep

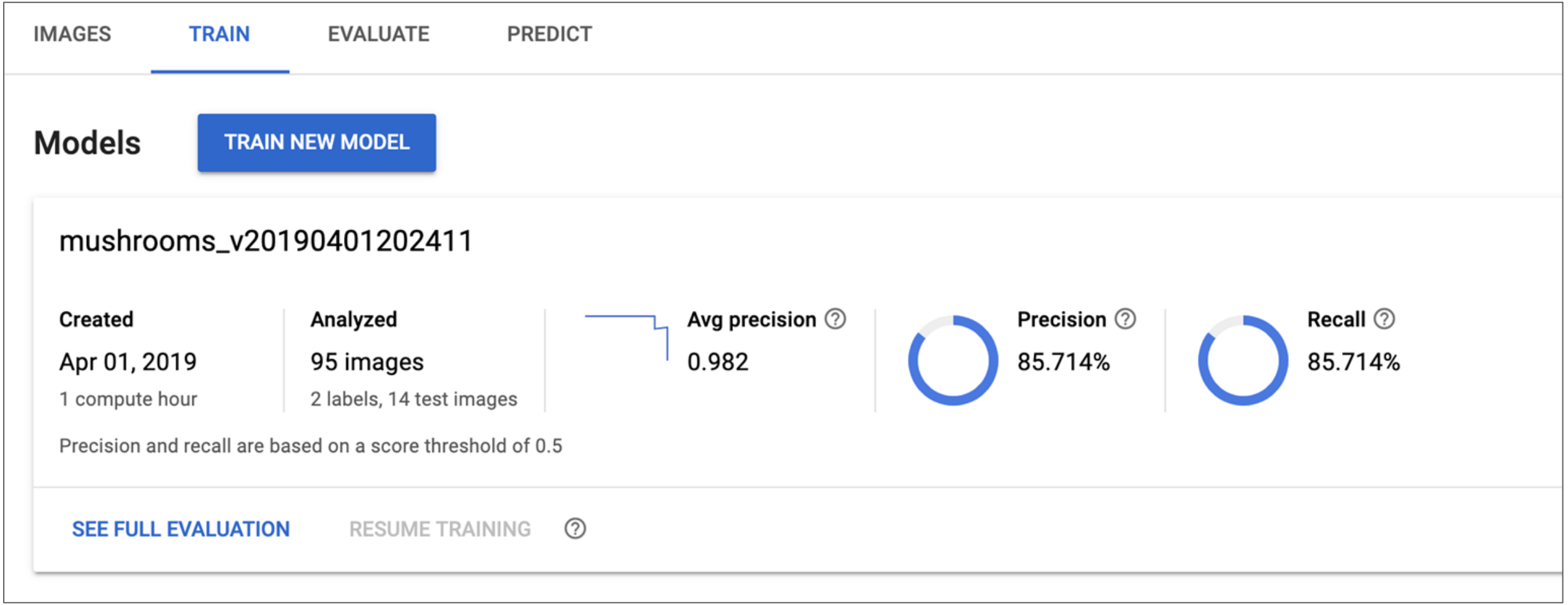
AutoML

- Allows developers to extend Google's machine learning model to work with their specific use cases
  - Known as Transfer Learning
- Provides a simple user interface for:
  - Creating datasets
  - Training and evaluating modes
  - Using your trained models to make predictions
- AutoML is available for Vision, Natural Language, and Translation

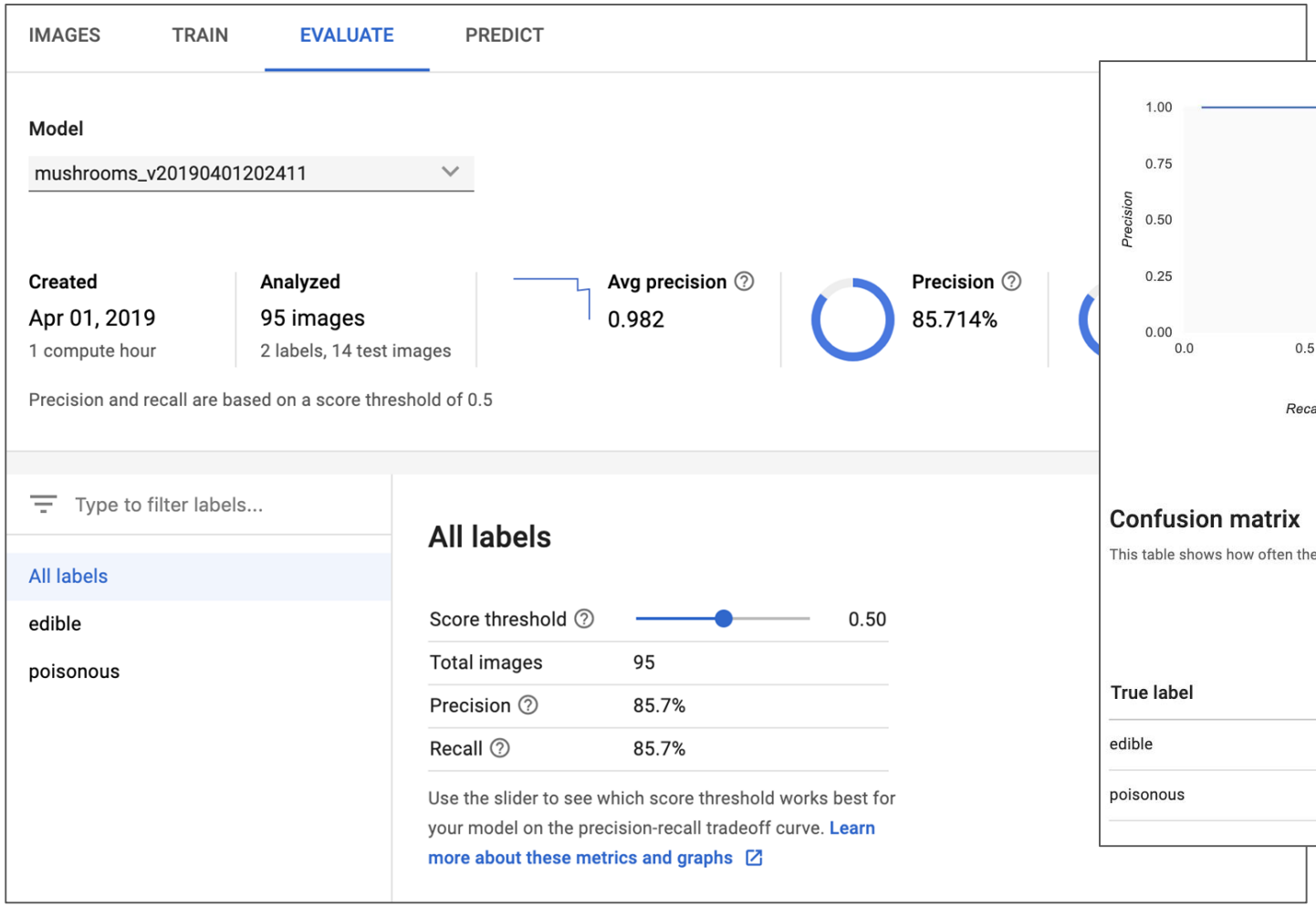
AutoML Vision Dataset

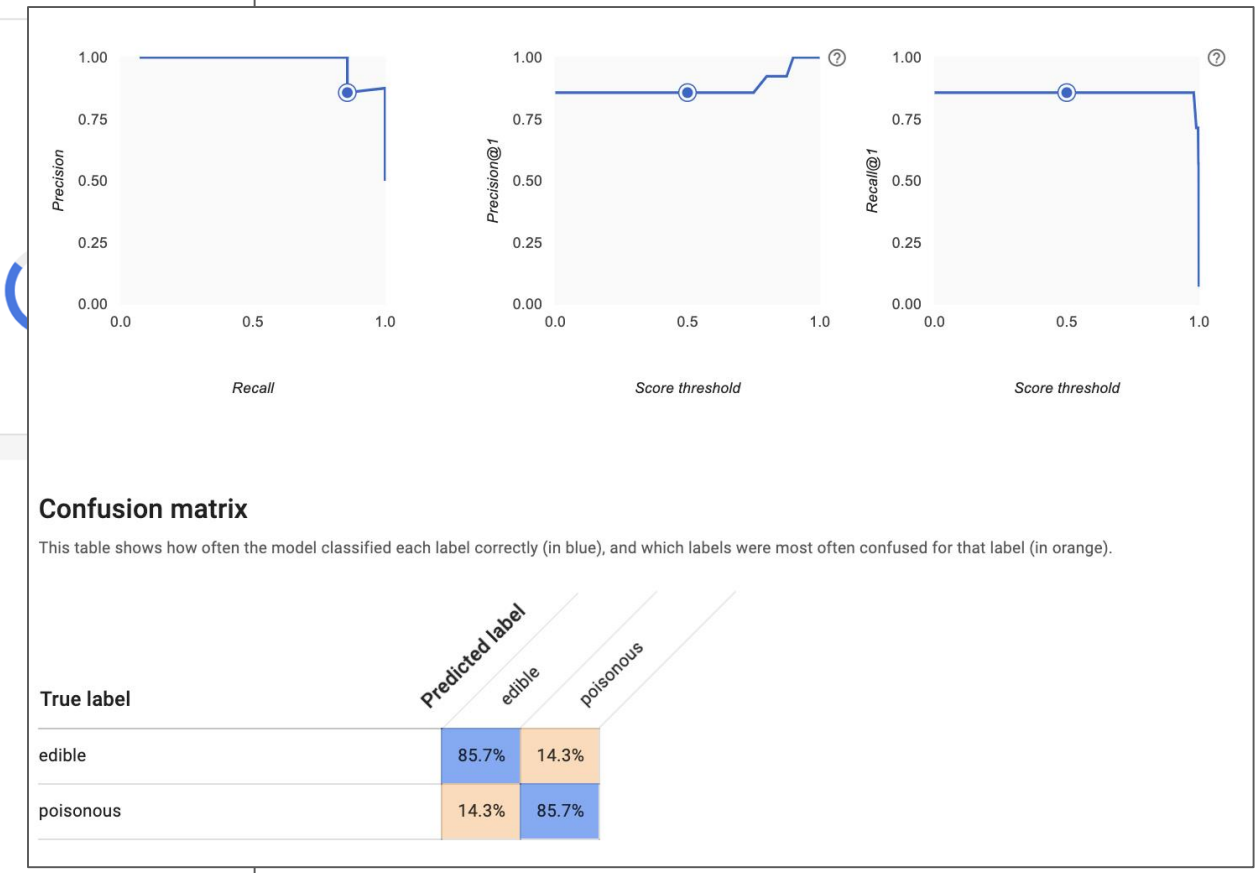


AutoML Training

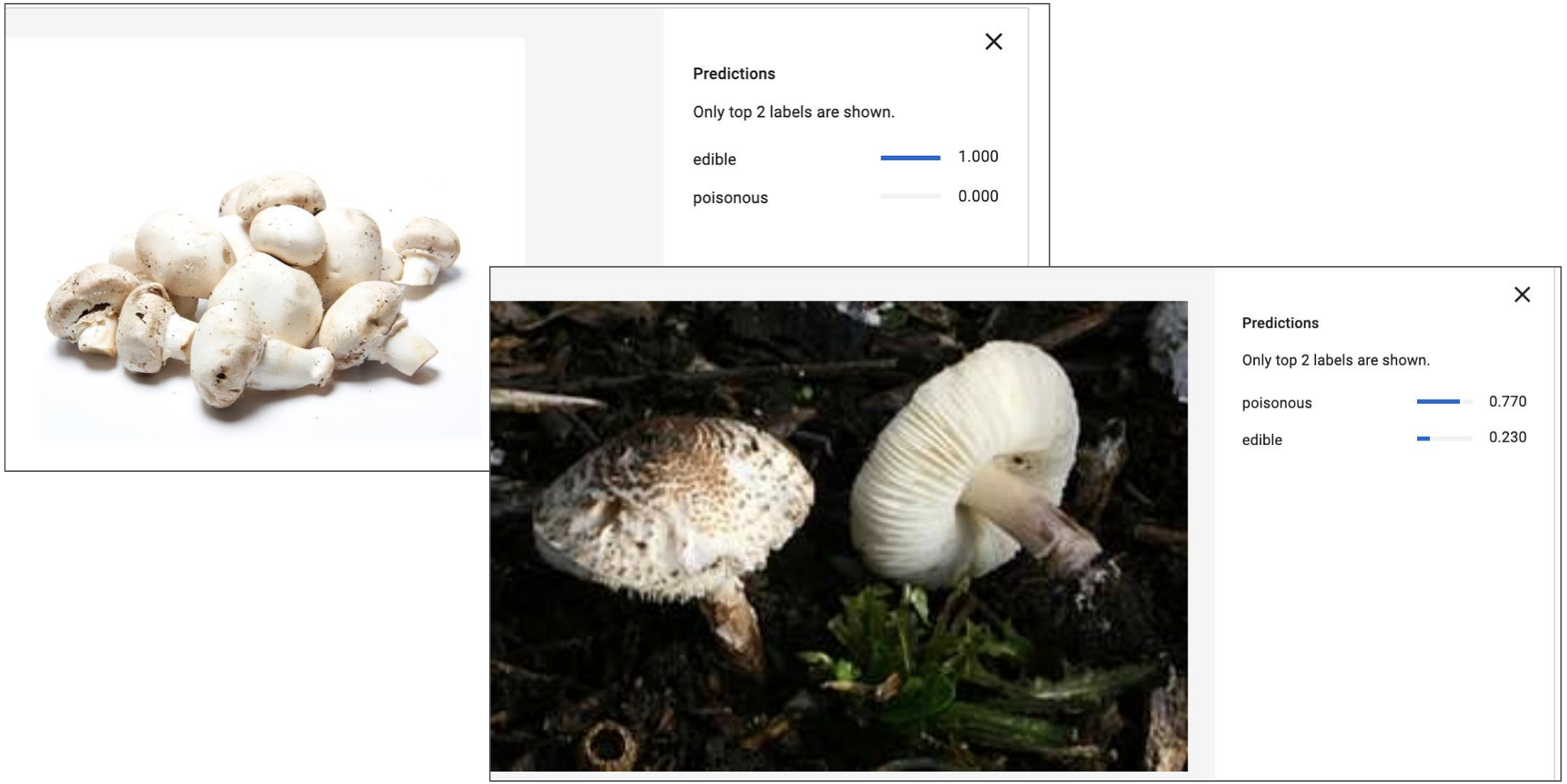


AutoML Evaluation





AutoML Prediction



AutoML Code

- AutoML provides sample code for using your model

```
import sys

from google.cloud import automl_v1beta1
from google.cloud.automl_v1beta1.proto import service_pb2


def get_prediction(content, project_id, model_id):
  prediction_client = automl_v1beta1.PredictionServiceClient()

  name = 'projects/{}/locations/us-central1/models/{}'.format(project_id, model_id)
  payload = {'image': {'image_bytes': content }}
  params = {}
  request = prediction_client.predict(name, payload, params)
  return request  # waits till request is returned
```

Chapter Concepts

TensorFlow

Cloud MLE

Pre-Built ML Models

Auto ML

BigQuery ML

Exam Prep

BigQuery ML

- Allows ML models to be built using SQL and trained using BigQuery
- Simplifies creation of models
  - No need to move data from BigQuery
  - No need to program in Python, Java, or any other language than SQL
- Supported BigQuery ML model types:

Model	Description
Linear regression	Used to predict a numerical value
Binary logistic regression	Used to predict a value from one of two classes
Multiclass logistic regression for classification	Used to predict a value from more than two classes

BigQuery ML – Training Example

```
#standardSQL
CREATE MODEL `bqml_tutorial.sample_model`
OPTIONS(model_type='logistic_reg') AS
SELECT
  IF(totals.transactions IS NULL, 0, 1) AS label,
  IFNULL(device.operatingSystem, "") AS os,
  device.isMobile AS is_mobile,
  IFNULL(geoNetwork.country, "") AS country,
  IFNULL(totals.pageviews, 0) AS pageviews
FROM
  `bigquery-public-data.google_analytics_sample.ga_sessions_*`
WHERE
  _TABLE_SUFFIX BETWEEN '20160801' AND '20170630'
```

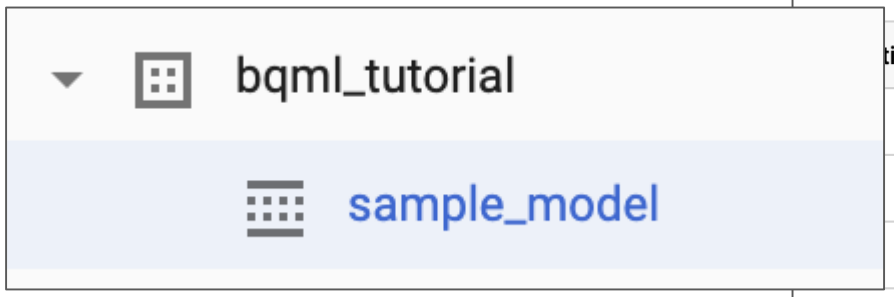
Define the type of model

Label is what you are trying to predict

The other fields are the features

BigQuery ML – The Model

- The model is stored in a dataset



Details | Training | Schema

tion	Training Data Loss	Evaluation Data Loss	Learn Rate	Duration (seconds)
	0.0439	0.0454	25.6000	18.52
	0.0447	0.0455	25.6000	20.31
	0.0473	0.0483	12.8000	20.64
5	0.0539	0.0533	6.4000	19.96
4	0.0678	0.0664	3.2000	20.04
3	0.0975	0.0962	1.6000	20.44
2	0.1698	0.1689	0.8000	22.80

BigQuery ML – Evaluation Example

```
#standardSQL
SELECT
  *
FROM
  ML.EVALUATE(MODEL `bqml_tutorial.sample_model`, (
SELECT
  IF(totals.transactions IS NULL, 0, 1) AS label,
  IFNULL(device.operatingSystem, "") AS os,
  device.isMobile AS is_mobile,
  IFNULL(geoNetwork.country, "") AS country,
  IFNULL(totals.pageviews, 0) AS pageviews
FROM
  `bigquery-public-data.google_analytics_sample.ga_sessions_*`
WHERE
  _TABLE_SUFFIX BETWEEN '20170701' AND '20170801'))
```

Use Evaluate method to see how good your model is

Use data that was not included in the training step

BigQuery ML – Prediction Example

```
#standardSQL
SELECT
  country,
  SUM(predicted_label) as total_predicted_purchases
FROM
  ML.PREDICT(MODEL `bqml_tutorial.sample_model`, (
SELECT
  IFNULL(device.operatingSystem, "") AS os,
  device.isMobile AS is_mobile,
  IFNULL(totals.pageviews, 0) AS pageviews,
  IFNULL(geoNetwork.country, "") AS country
FROM
  `bigquery-public-data.google_analytics_sample.ga_sessions_*`
WHERE
  _TABLE_SUFFIX BETWEEN '20170701' AND '20170801'))
GROUP BY country
ORDER BY total_predicted_purchases DESC
LIMIT 10
```

Use Predict method to use your model

Exercise: Getting Started with BigQuery ML

- In these exercise, you will learn the basics of using BigQuery ML for machine learning
  - Getting Started with BigQuery ML

Chapter Concepts

TensorFlow

Cloud MLE

Pre-Built ML Models

Auto ML

BigQuery ML

Exam Prep

Do Now: Practice Quiz

- Take this practice quiz

Google Cloud Certification Workshop—Data Engineer

Practice Exam

Practice Exam

- Take this practice exam

Google Cloud Certification Workshop—Data Engineer

QwikLabs Quest

Data Engineering Quest

- Complete the Data Engineering Quest provided by Qwiklabs
  - Go to: Data Engineering Quest
  - If you do not already have a QwikLabs account, Join
  - Enroll in this Quest

Google Cloud Certification Workshop—Data Engineer

Homework

Links

- The Great AI Awakening
- Dataproc
- BigQuery
- Dataflow
- Apache Beam
- TensorFlow
- Google Cloud ML Engine
- Cloud Composer
- Cloud Data Loss Prevention (DLP)

Videos

- TensorFlow and Deep Learning without a PhD
- Effective TensorFlow for Non-Experts
- Introduction to Google Cloud Machine Learning
- Coursera Machine Learning Course
- Big Data and Machine Learning Fundamentals

Tutorials

- Google Cloud Storage
- Disks and Snapshots
- Cloud SQL Quickstart
- Cloud Spanner Quickstart
- Datastore Quickstart
- Creating Dataproc Clusters
- GCP Dataproc: Running Hive and Spark Jobs
- Querying Data with BigQuery
- Dataflow Quickstart using Python
- Dataflow Quickstart using Java
- Getting Started with Tensorflow
- Quickstart using the Command Line
- Quickstart using Datalab

ROI TRAINING
MAXIMIZE YOUR TRAINING INVESTMENT

Google Cloud Certification Workshop—Data Engineer

Course Summary

Course Summary

In this course, we have learned how to:

- Prepare for the GCP Data Engineer certification exam
- Choose the appropriate GCP data storage solution
- Architect batch and streaming data processing pipelines on GCP
- Leverage GCP tools for data manipulation, analysis, and visualization
- Build machine learning models with GCP tools

Looking for more training or training for your team?

ROITRAINING
MAXIMIZE YOUR TRAINING INVESTMENT

- ➔ Visit google.roitraining.com or email us directly at GoogleOps@roitraining.com
- ➔ Be sure to check out the rest of Google's Cloud Curriculum below

Cloud Infrastructure

Google Cloud Fundamentals: Core Infrastructure

Architecting with Google Cloud Platform: Infrastructure

Architecting with Google Cloud Platform: Design and Process

Google Cloud Certification Workshop: Cloud Architect

Google Cloud Certified Professional - Cloud Architect

Data and Machine Training

Google Cloud Fundamentals: Big Data and Machine Learning

Data Engineering on Google Cloud Platform

From Data to Insights with Google Cloud Platform

Google Cloud Certification Workshop: Data Engineer

Google Cloud Certified Professional - Data Engineer

Application Development

Google Cloud Fundamentals: Core Infrastructure

Developing Applications with Google Cloud Platform

Getting Started with Google Kubernetes Engine

Other Course Offerings

Getting Started with Google Kubernetes Engine

758: Migrating to the Cloud - Sales Workshop

797: Google Cloud Platform Launchpad

Build a Business Transformation Vision with Google Cloud

Google Cloud
Premier Partner

20% off for ROI Alumni with Offer Code: **ALUMNI**
Google Cloud Student Resource Group on LinkedIn