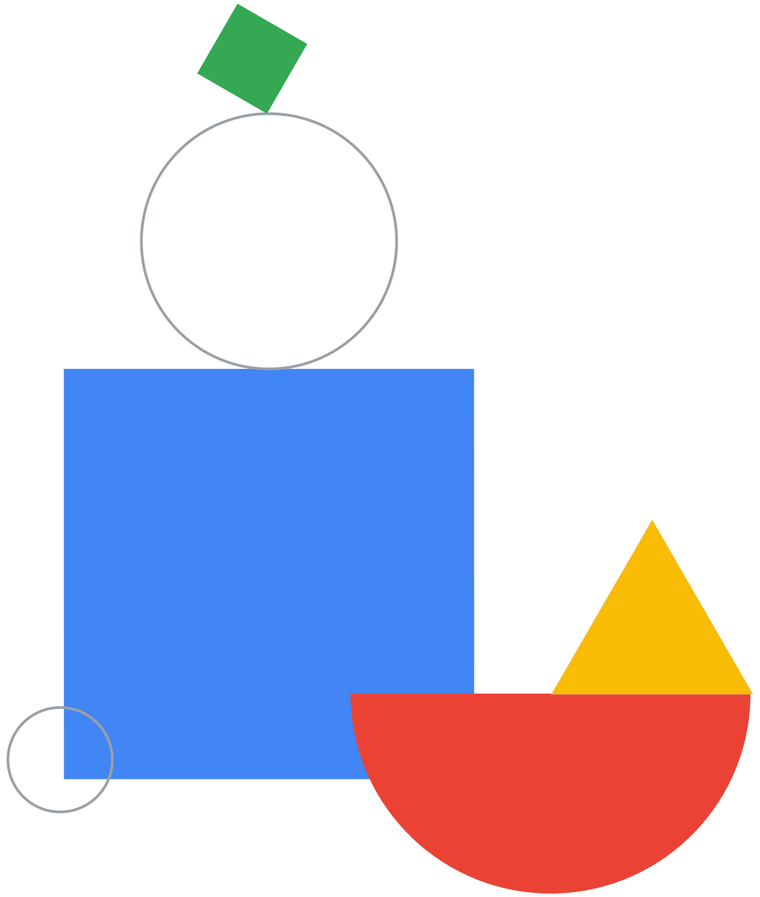
Google Cloud

Preparing for Your Professional Data Engineer Journey



Introduction

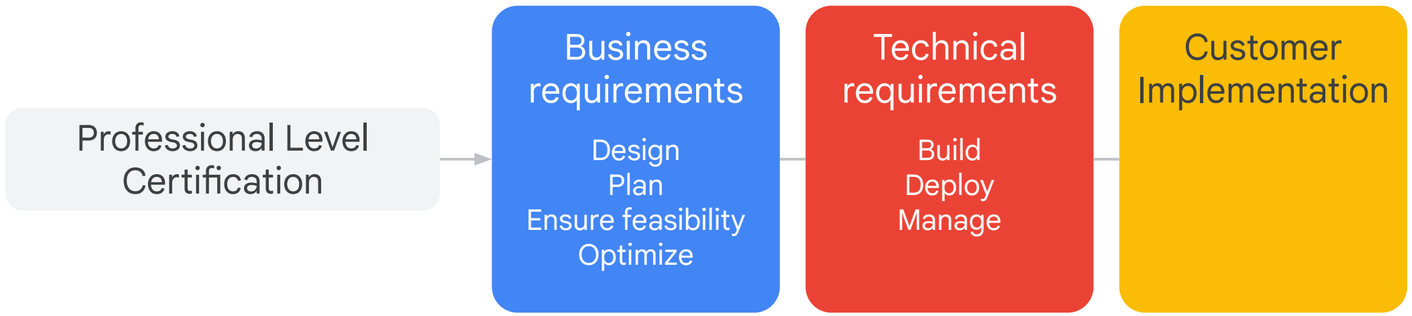
Supratim Bhattacharya

Technical mentor

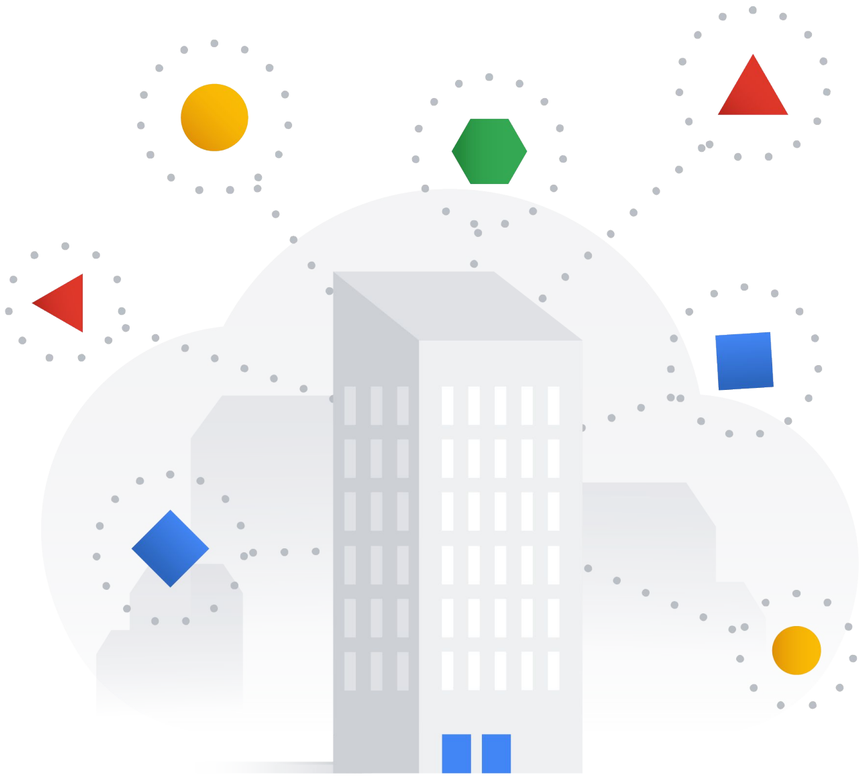
Why become Google certified?

- Gain industry recognition!
- Validate your technical expertise!
- Take your career to the next level!

What is a Professional certification?

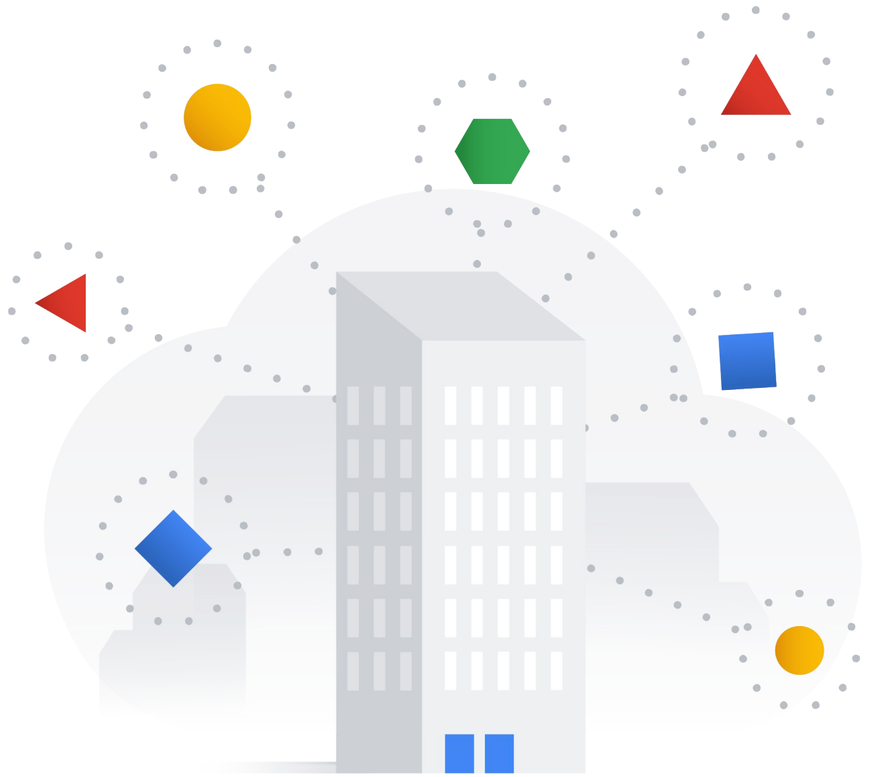


What is the role of a Professional Data Engineer?



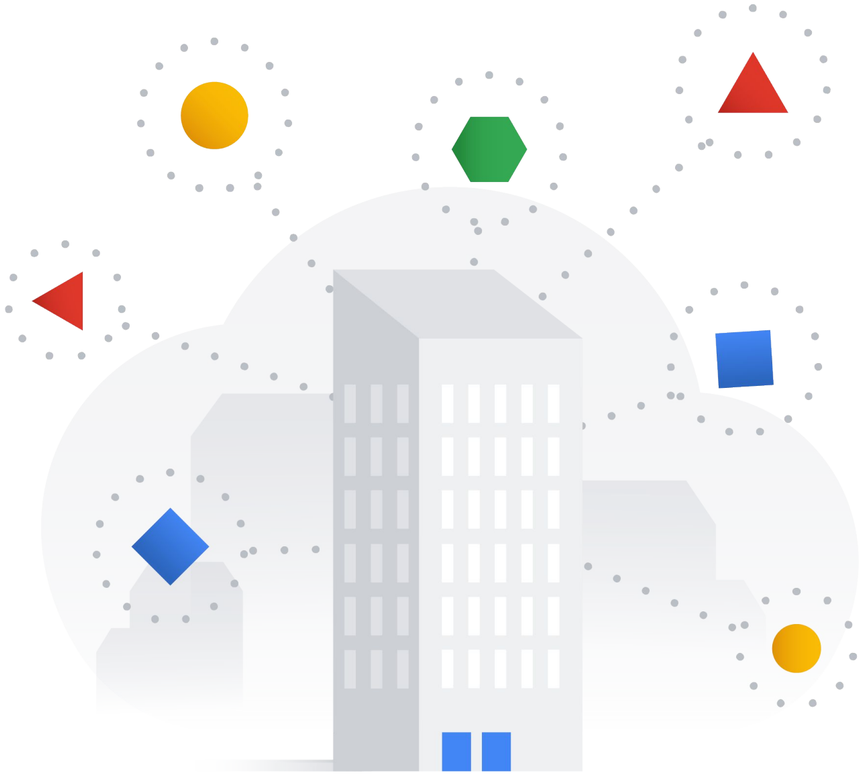
What is the role of a Professional Data Engineer?

- Makes data usable and valuable for others
- Evaluates and selects products and services



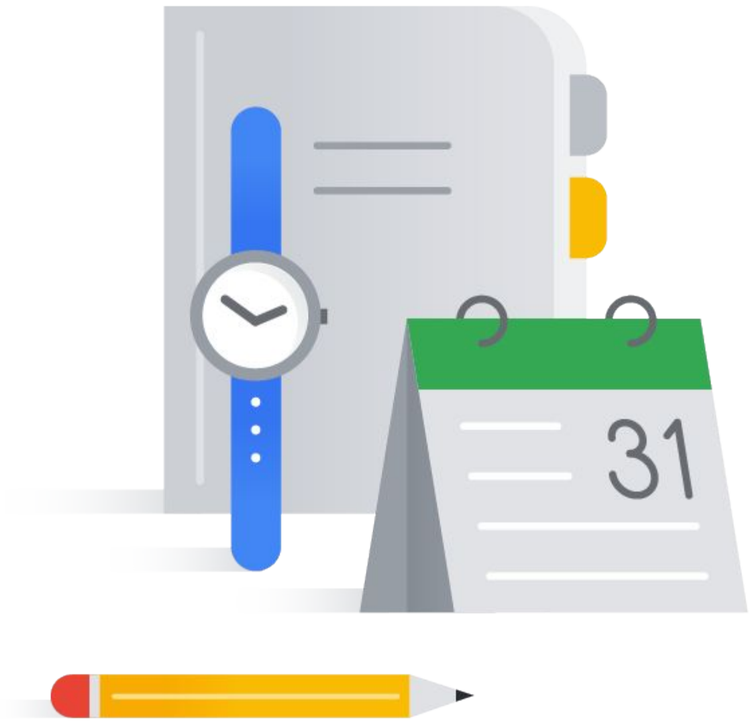
What is the role of a Professional Data Engineer?

- Creates and manages robust data processing systems
- Demonstrates these abilities with data processing workloads:
  - Design
  - Build
  - Deploy
  - Monitor
  - Maintain
  - Secure
- For more information, visit the Professional Data Engineer Certification page



Course agenda

0	Introduction
1	Designing Data Processing Systems
2	Ingesting and Processing Data
3	Storing Data
4	Preparing and Using Data for Analysis
5	Maintaining and Automating Data Workloads
6	Your Next Steps



Understand the scope of the exam based on the Professional Data Engineer [Exam Guide](#).

In short, the exam tests your skills in:

- Designing data processing systems
- Ingesting and processing the data
- Storing the data
- Preparing and using data for analysis
- Maintaining and automating data workloads

PDE Exam High-level Overview

- How to position PDE? One of 9 professional level exams (+ CDL & Associate Cloud Engineer)
- Technically, no hard prerequisites (such as other GCP certs / courses / architect & cloud experience), but...
- Very broad, logical, use case specific
- 80h-150h+ to complete. Consistency is the key! Learn for 2-3 hours a day, starting from now...
- Exam structure:
  - Multiple-choice questions, asking about “real-world” challenges. Most with a single correct answer, some with more (you will know how many). Get a feeling by doing a sample test.
  - In a lot of cases, you will need to choose OPTIMAL answer from 2-4 which are technically correct (focus would be on time / flexibility / availability / cost / automation / security etc).
  - NO labs, NO hands-on exercises on the exam (but essential when preparing!)
  - 50 questions / 2h for the exam. Pro tip: Handle the easier questions first (aim at ~1.5 min per question), mark the rest for review. Also, don’t leave any questions unanswered (no negative points).
    - English language only (no additional time for non-native speakers).
- When you submit the exam, you’ll only see PASS / FAIL.
  - You will NOT be presented with any details of accuracy / areas to improve etc.
  - It’s not clear what is the percentage needed to pass (aim at 85+% accuracy for practise questions).
- Can be taken online or at a testing center. Sign up here.
- Certificate is valid for 2 years.
- If you fail, retake policy is: 14 days / 60 days / 1 year (separate voucher needed for each attempt)

Exam question - example

Notice the business context

You work for a manufacturing company that owns a high-value machine which has several machine settings and multiple sensors. A history of the machine's hourly sensor readings and known failure event data are stored in BigQuery. You need to predict if the machine will fail within the next 3 days in order to schedule maintenance before the machine fails. Which data preparation and model training steps should you take?

A. Data preparation: Daily max value feature engineering with DataPrep; Model training: AutoML classification with BQML
B. Data preparation: Daily min value feature engineering with DataPrep; Model training: Logistic regression with BQML and AUTO_CLASS_WEIGHTS set to True
C. Data preparation: Rolling average feature engineering with DataPrep; Model training: Logistic regression with BQML and AUTO_CLASS_WEIGHTS set to False
D. Data preparation: Rolling average feature engineering with DataPrep; Model training: Logistic regression with BQML and AUTO_CLASS_WEIGHTS set to True

Sections	Approximate % Scored Questions	Section Performance
Section 1: Setting up a cloud solution environment	17.5%	**Meets**
Section 2: Planning and configuring a cloud solution	17.5%	**Does Not Meet**
Section 3: Deploying and implementing a cloud solution	25.0%	**Meets**
Section 4: Ensuring successful operation of a cloud solution	20.0%	**Meets**
Section 5: Configuring access and security	20.0%	**Meets**

Pro tip: Make notes while you learn!!!

- Choose your favourite tool that handles text **and images**.
- Copy & paste from GCP documentation / course transcripts
- Copy & paste important slides / images / decision trees / tables
- Use colours / bold
- Paste difficult quiz questions from quizzes etc with explanations / links to solutions
- Remember about 3-year validity of GCP certificates -> notes will most probably be very useful next time.
- **Switch between resources if you get bored**.

Custom roles
https://cloud.google.com/iam/docs/creating-custom-roles

Use the gcloud iam list-testable-permissions command to get a list of permissions that are available for custom roles in a specific project or organization. The response lists the permissions that you can use in custom roles for that project or organization.

To list permissions that are available in custom roles for a project or organization, run this command:

*gcloud iam list-testable-permissions **full-resource-name***
--filter="customRolesSupportLevel!=NOT_SUPPORTED"

You can create a custom role at the project or organization level. = NOT ON FOLDER LEVEL!!

To view the role metadata, use one of the methods below:

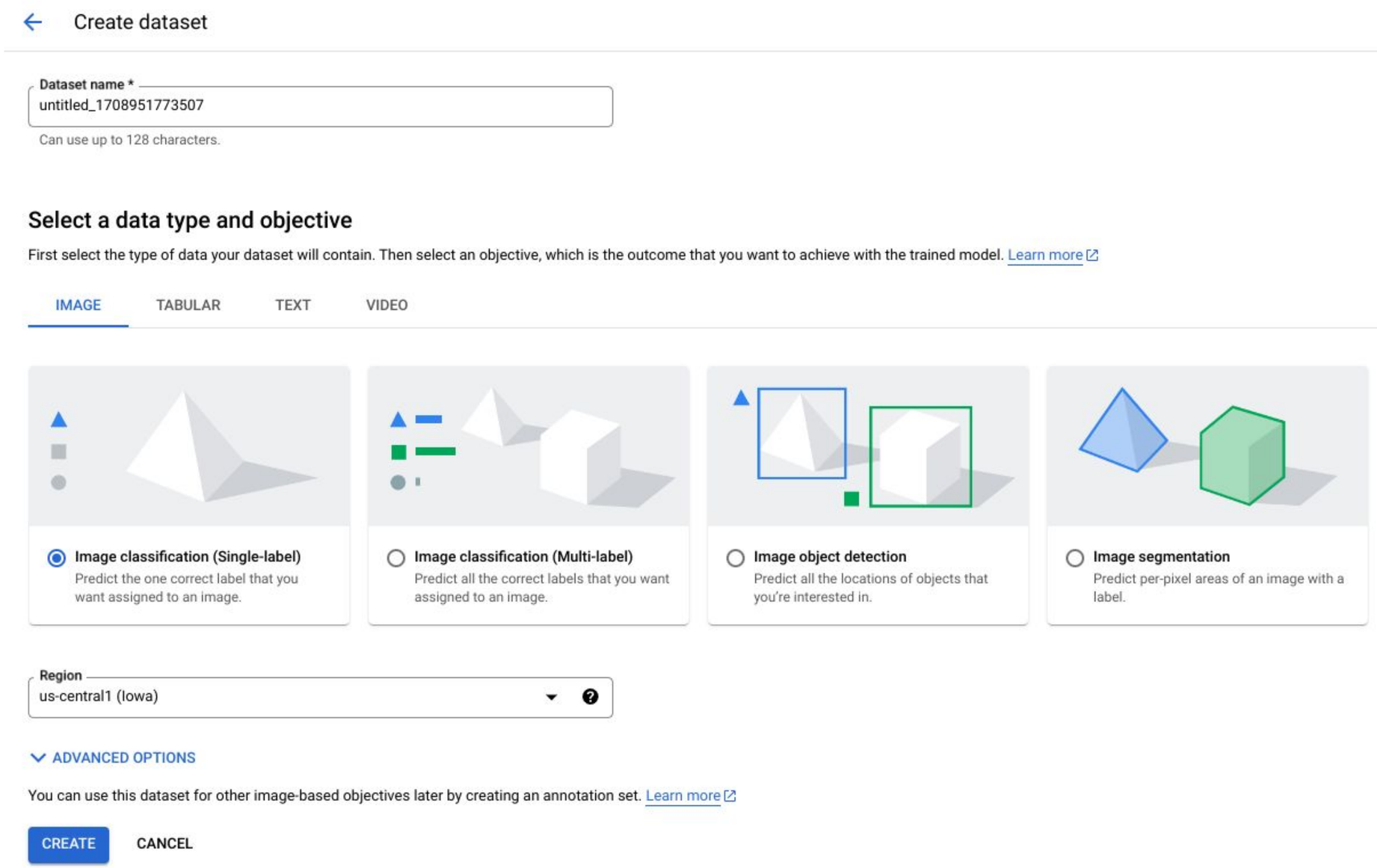
gcloud iam roles describe role-id

Example:

```
gcloud iam roles describe roles/iam.roleViewer
```

```
description: Read access to all custom roles in the project.
etag: AA==
includedPermissions:
- iam.roles.get
```

Pro tip no.2: be curious!

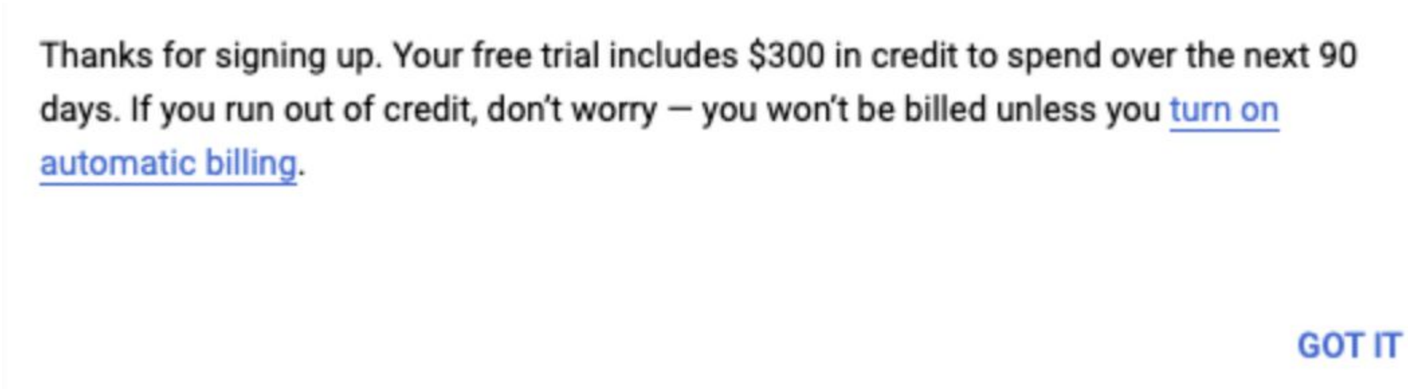


Pro tip no.3: use GCP Free Trial 300USD (*) and Free Tier

It will help you be curious :)

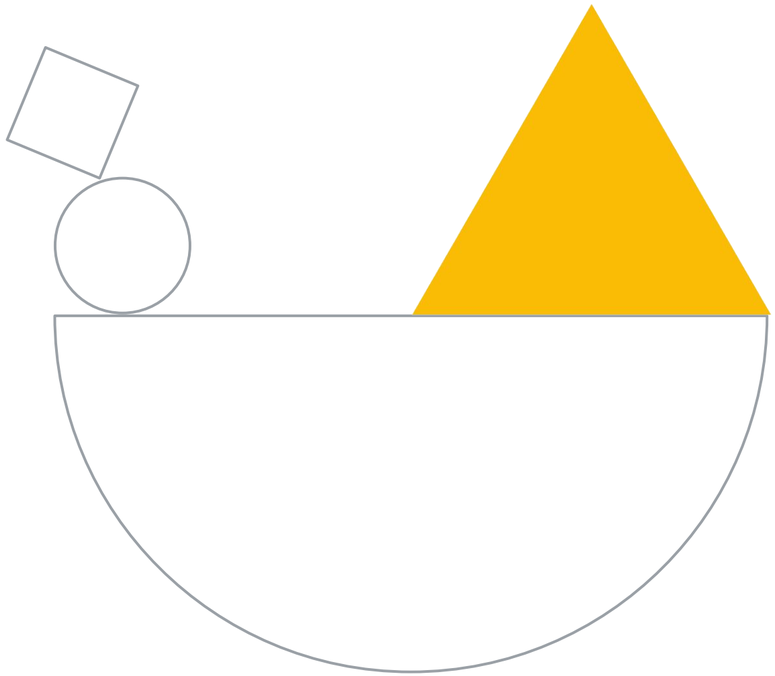
- **90-day, $300 Free Trial**: New Google Cloud and Google Maps Platform users can take advantage of a 90-day trial period that includes $300 in free Cloud Billing credits to explore and evaluate Google Cloud and Google Maps Platform products and services. You can use these credits toward one or a combination of products.

https://cloud.google.com/free



* To complete your Free Trial signup, you must provide a credit card or other payment method to set up a Cloud Billing account and verify your identity. Don't worry, setting up a Cloud Billing account does not enable us to charge you. You are not charged unless you explicitly enable billing by upgrading your Cloud Billing account to a paid account.

Resources to support your certification journey



Review diagnostic questions

1.1 Diagnostic Question 01 Discussion

Business analysts in your team need to run analysis on data that was loaded into BigQuery. You need to follow recommended practices and grant permissions.

What role should you grant the business analysts?

A. bigquery.resourceViewer and bigquery.dataViewer

B. bigquery.user and bigquery.dataViewer

C. bigquery.dataOwner

D. storage.objectViewer and bigquery.user

Quizzes - planned for weeks 2-6

How to implement back-out/rollback plan for website with 100s of VMs, when the site has frequent critical updates? *

- ◯ Create a Nearline copy of static data in Cloud Storage.
- ◯ Create a snapshot of each VM prior to update, in case of failure.
- ◯ Use managed instance groups with the “update-instances” command when starting a rolling update.
- ◯ Only deploy changes using Deployment Manager templates.

- Aim: validate **technical** knowledge (no business context)
- NOT as complex as questions on the exam

Final exam-like quiz

~30 questions; Link to be shared during our last meeting

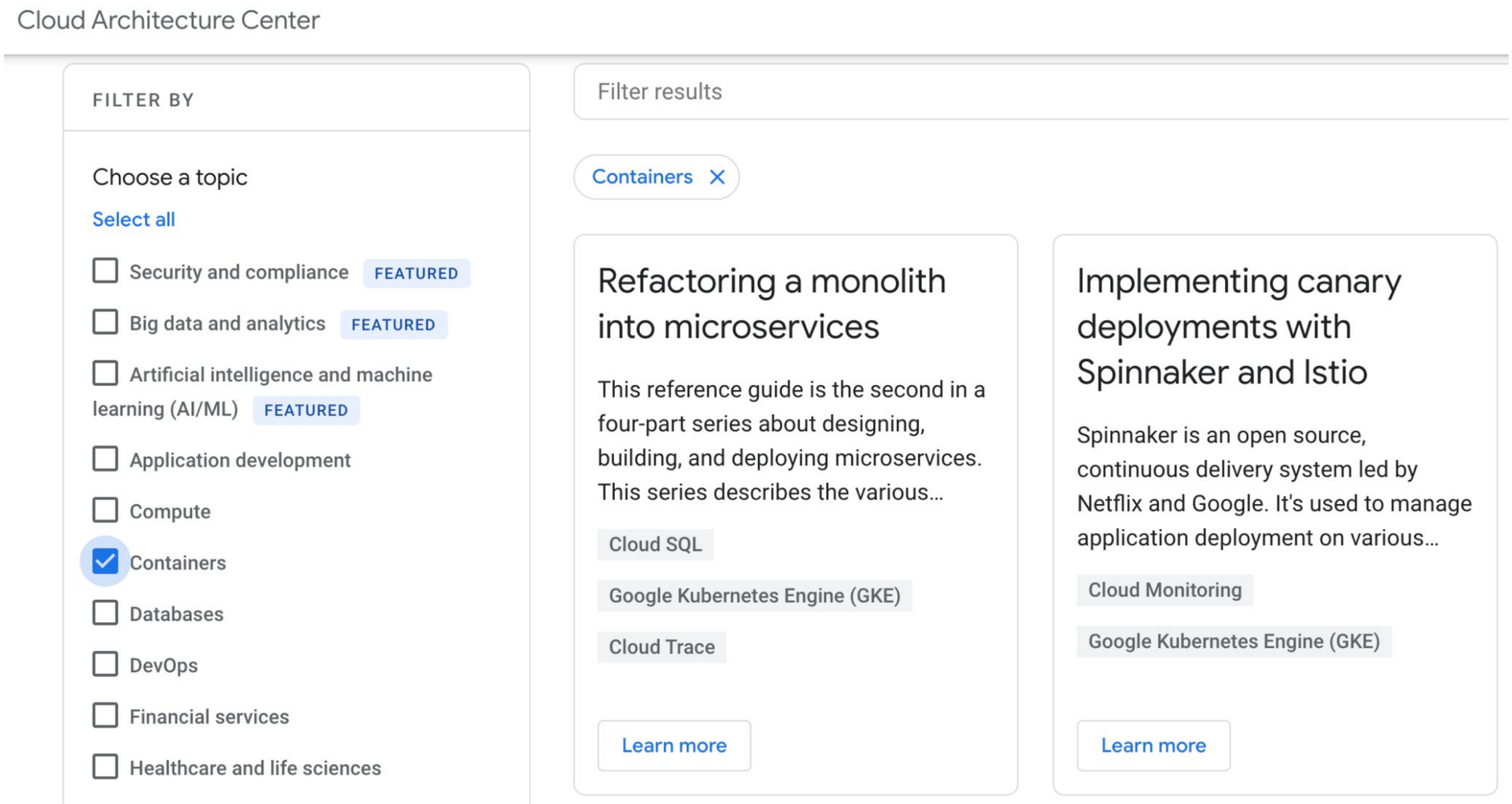
One of the developers on your team deployed their application in Google Container Engine with the Dockerfile below. They report that their application deployments are taking too long.You want to optimize this Dockerfile for faster deployment times without adversely affecting the app's functionality.Which two actions should you take? Choose 2 answers.

```
FROM ubuntu:16.04
COPY . /src
RUN apt-get update && apt-get install -y python python-pip
RUN pip install -r requirements.txt
```

- [ ] Remove Python after running pip
- [ ] Remove dependencies from requirements.txt
- [ ] Use a slimmed-down base image like Alpine Linux
- [ ] Use larger machine types for your Google Container Engine node pools
- [ ] Copy the source after the package dependencies (Python and pip) are installed

Exam Tip: *aim at 85%+ accuracy before signing up for the exam (don't know what's the passing score, remember?)*

Cloud Architecture Center

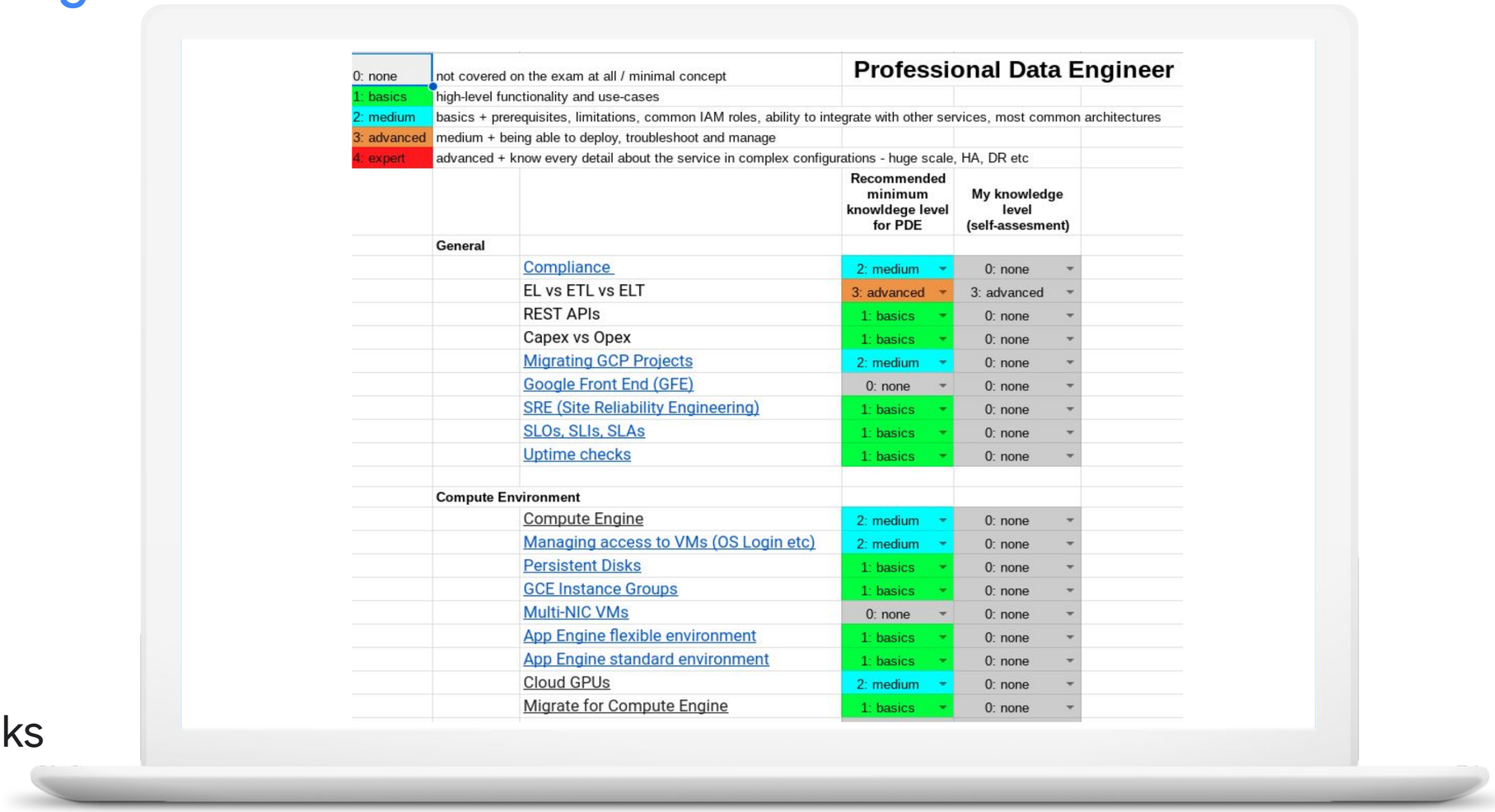


Cloud Architecture Center

GCP Services - exam relevance

= “when will I know enough?”

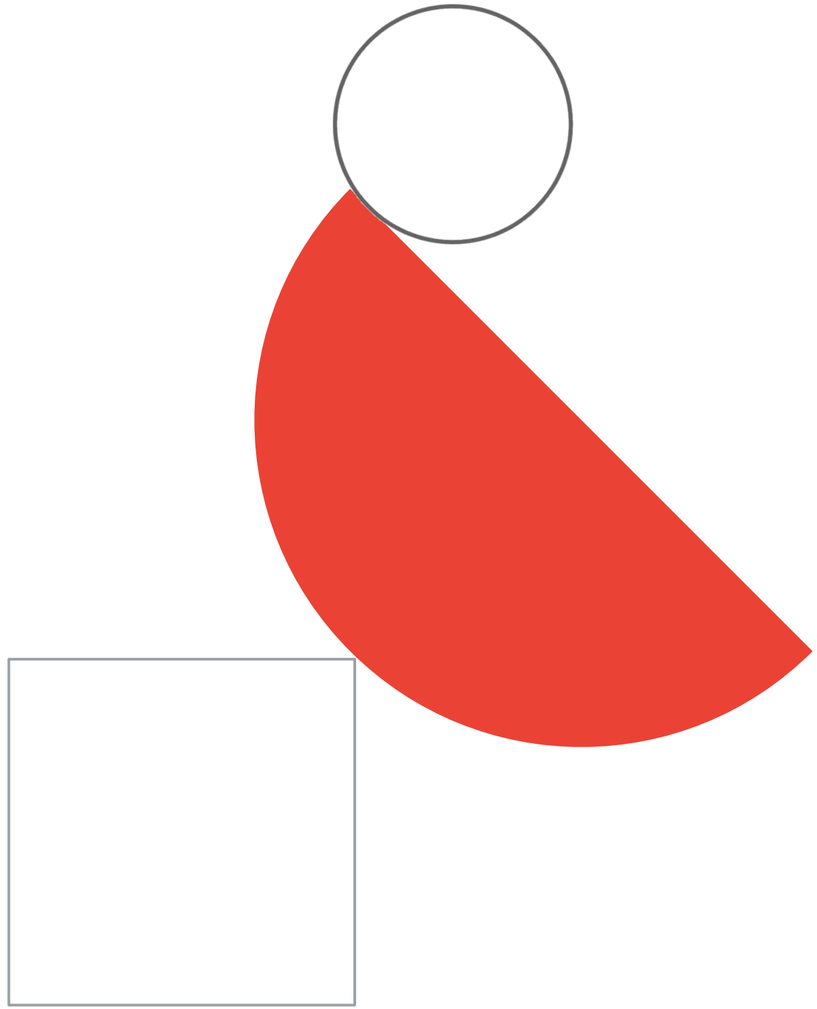
- VERY subjective, NOT a Google-owned or recommended.
- Should be treated as supplemental study materials; NOT an exhaustive list or requirements.
- If it helps -> use it. If it looks scary to you -> don’t 😉



		Professional Data Engineer	
0: none	not covered on the exam at all / minimal concept		
1: basics	high-level functionality and use-cases		
2: medium	basics + prerequisites, limitations, common IAM roles, ability to integrate with other services, most common architectures		
3: advanced	medium + being able to deploy, troubleshoot and manage		
4: expert	advanced + know every detail about the service in complex configurations - huge scale, HA, DR etc		
		Recommended minimum knowldege level for PDE	**My knowledge level (self-assesment)**
General			
	Compliance	2: medium	0: none
	EL vs ETL vs ELT	3: advanced	3: advanced
	REST APIs	1: basics	0: none
	Capex vs Opex	1: basics	0: none
	Migrating GCP Projects	2: medium	0: none
	Google Front End (GFE)	0: none	0: none
	SRE (Site Reliability Engineering)	1: basics	0: none
	SLOs, SLIs, SLAs	1: basics	0: none
	Uptime checks	1: basics	0: none
Compute Environment			
	Compute Engine	2: medium	0: none
	Managing access to VMs (OS Login etc)	2: medium	0: none
	Persistent Disks	1: basics	0: none
	GCE Instance Groups	1: basics	0: none
	Multi-NIC VMs	0: none	0: none
	App Engine flexible environment	1: basics	0: none
	App Engine standard environment	1: basics	0: none
	Cloud GPUs	2: medium	0: none
	Migrate for Compute Engine	1: basics	0: none

LINK

Case Study

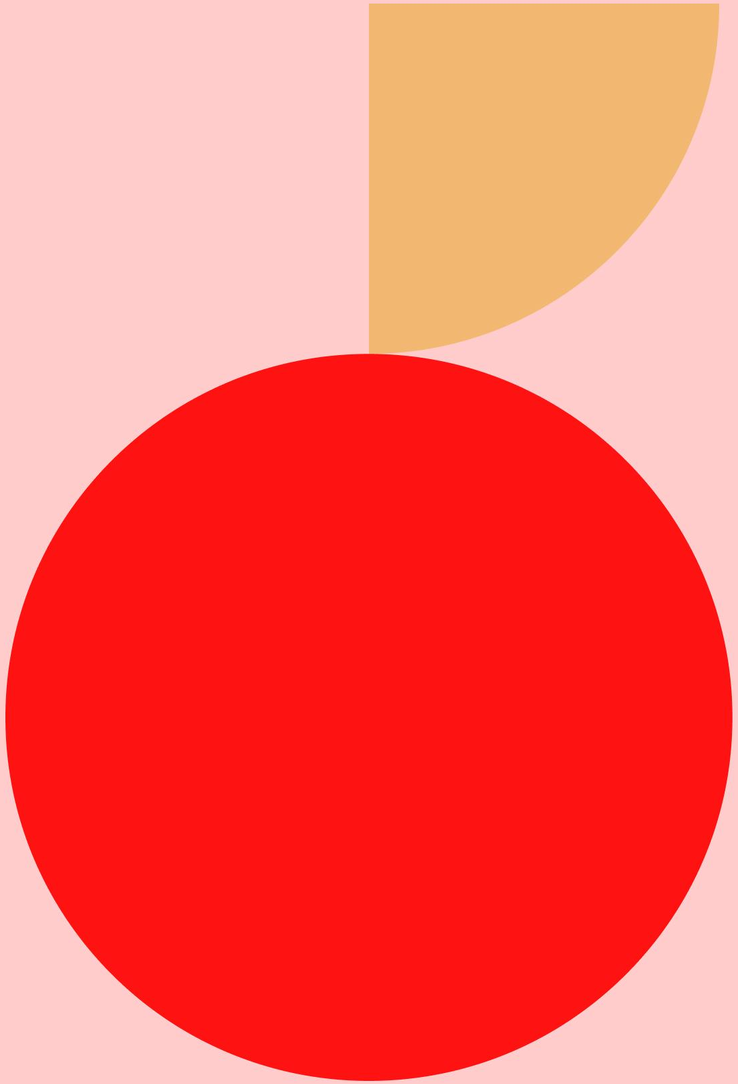


See what others think...

Posts

Published	Title/Link	Author
2022/10	How I prepared for GCP PDE Certification	Narayan Singh
2022/07	How do I prepared and passed the GCP PDE Certification Exam (2022)	Gary Yiu
2022/05	How I Passed the PDE Exam on my First Attempt	Damian Ohienmhen
2021/10	Top ten tips for passing the GCP PDE Certification exam	konrad bachusz
2021/09	How I cleared the Google Cloud Professional Data Engineer exam in 2021	Bhavesh Bhatt
2021/01	How to pass the PDE certification – as a sales guy	Rolf Siegel
2020/09	GCP Professional Data Engineer Guide	A Estevez
2020/07	preparation, tips, and my journey from failing to passing experience	Kshitij Bhadage
2020/05	Data Engineering on GCP Specialisation	Alex Papageorgiou
2020/04	How to Prepare for GCP PDE Exam	Hardik Rathod
2020/02	Topics in the GCP PDE Exam	Sabeehah Ahmed
2020/02	My personal road map and thoughts in 2020	João Vitor Baptista
2020/01	How I pass the GCP Data Engineer exam - Key points to study and personal experience	Kenny Contreras
2019/12	How to pass the Google Cloud Professional Data Engineer exam	Jonathan M
	Google Cloud - Professional Data Engineer Exam Study Materials	Leverege Developers
2019/12	Road to Success — Professional Data Engineer	Biswarup Ghoshal
2019/10	The Ultimate Hack to passing Data Engineer exam	Kelly Sun
2019/07	How to pass the Cloud Architect and Data Engineer GCP certifications	Ivam Luz

LINK

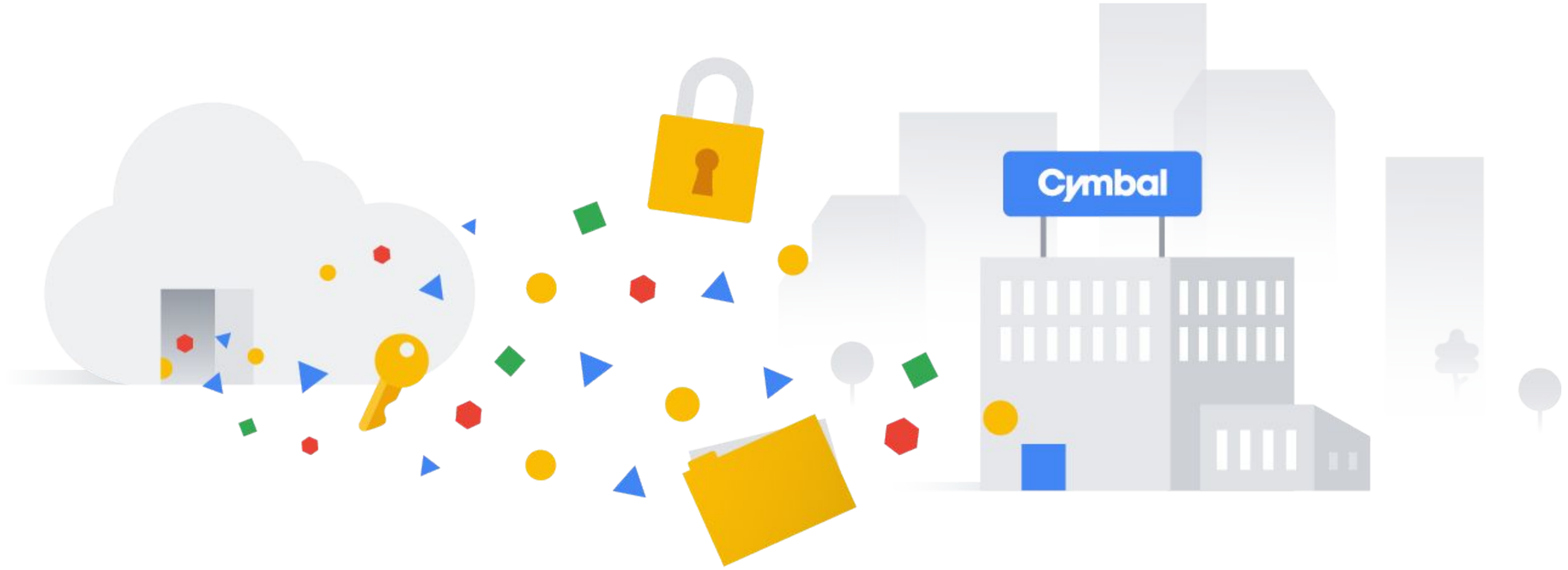


Cymbal Superstore

Cymbal Retail

Global B2C retailer serving customers
in 52 countries and territories

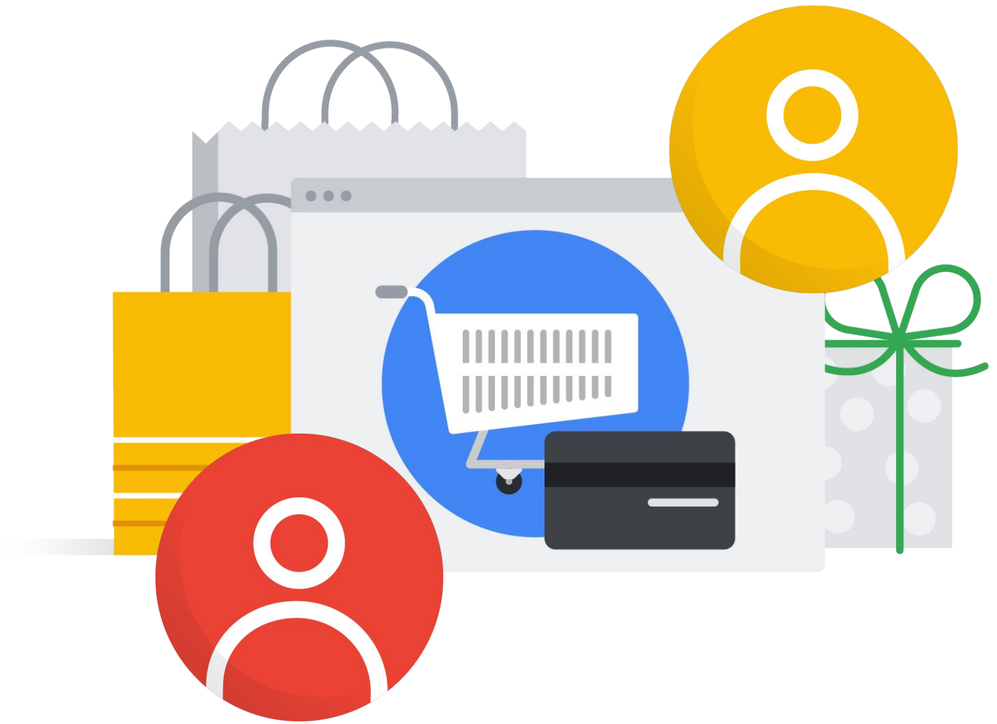
Transitioning Cymbal Retail's data to the cloud



As a Professional Data Engineer, you need to ensure that Cymbal complies with local regulations and data is secure.

Integrating solutions to delight customers

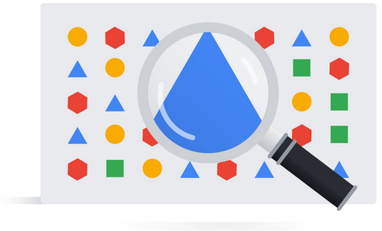
As a Professional Data Engineer, you must integrate Google Cloud products and services with the technologies used by Cymbal Retail's acquisitions.



Meeting business needs through technology

You need to balance four areas of focus to meet business needs:

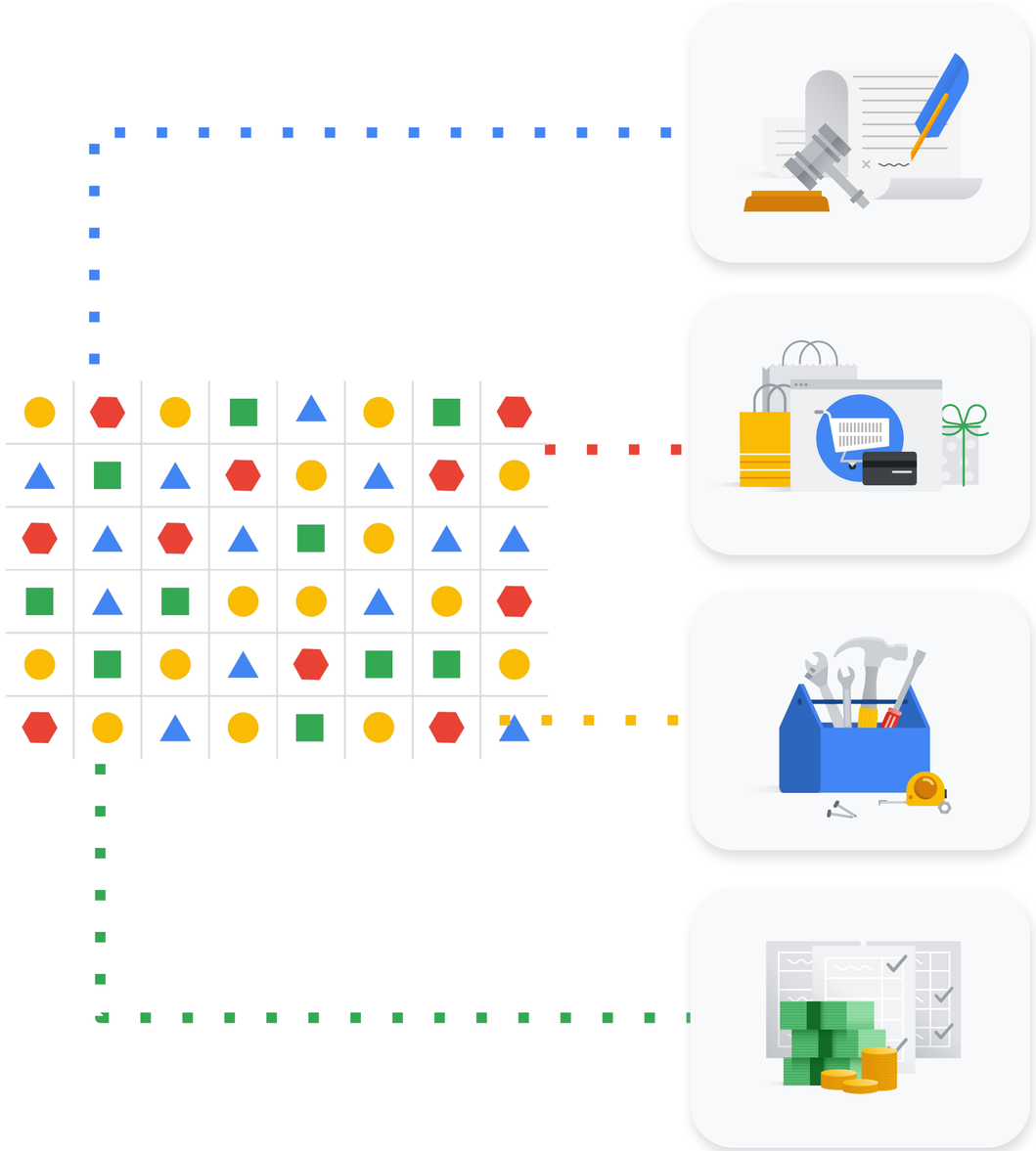
Rapidity in the supply chain



Accuracy of analytics and predictions

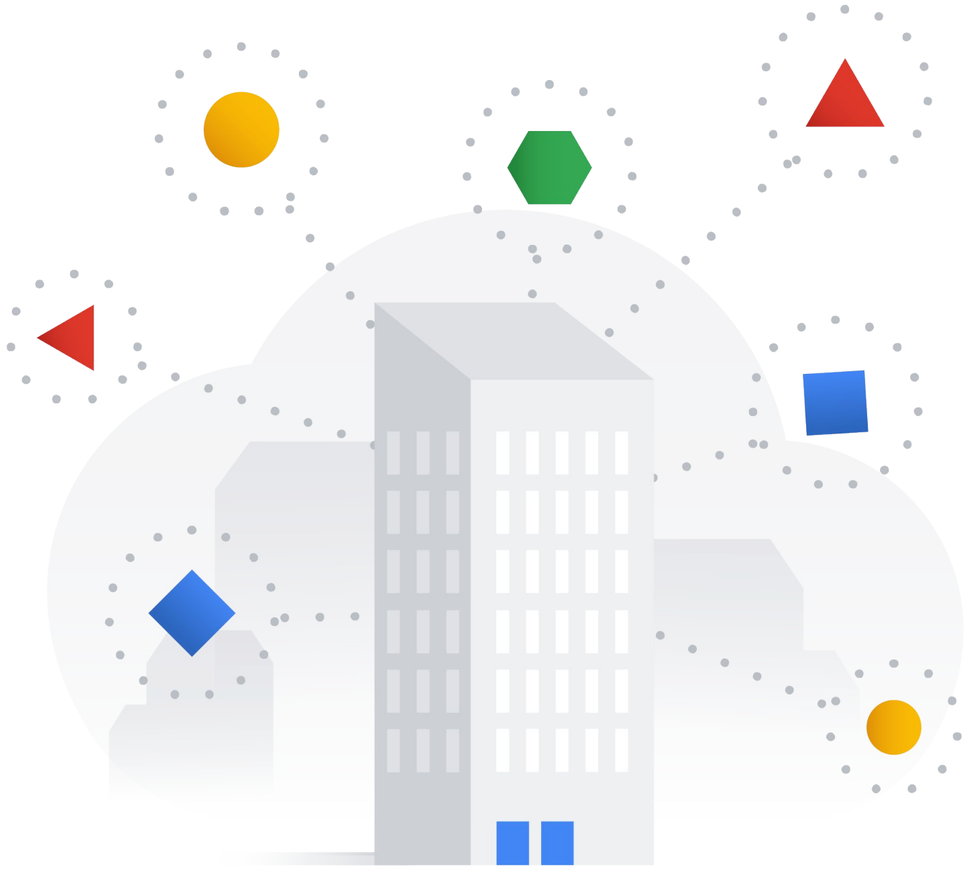
Centralized control and security

Controlled costs



Making data available to the organization

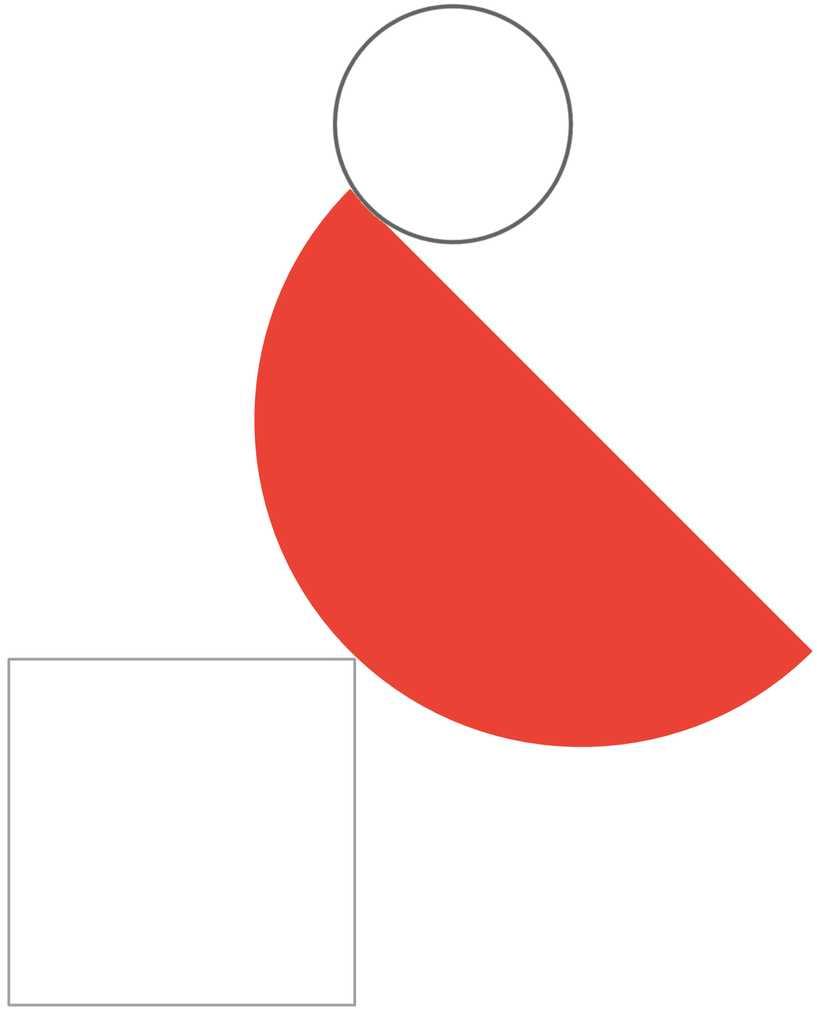
- Cymbal Retail's management wants the organization to have visibility into data for decision making.
- You need to make data easily available to all parts of the organization.



Managing data for Cymbal Retail

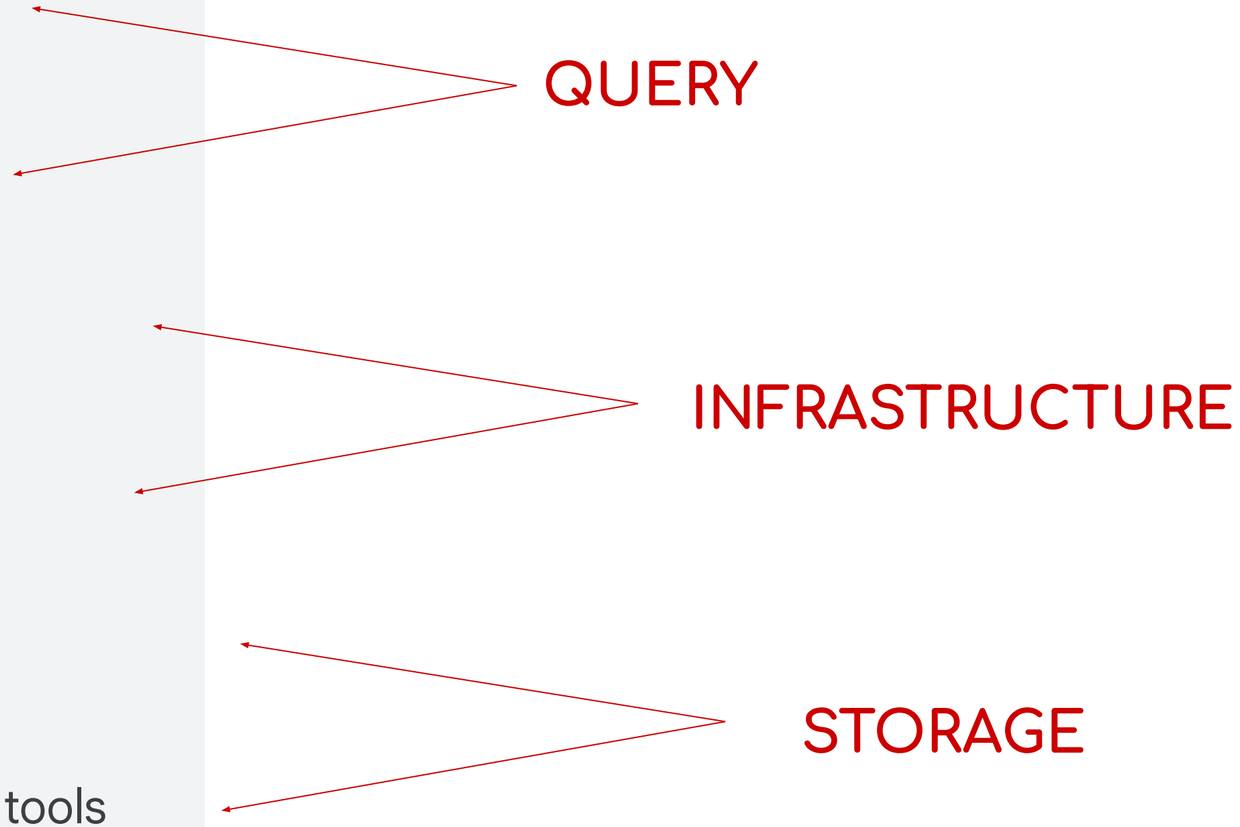
Lab Demo

Tips and comments to the course content



Analyst Challenges

- Query speed and performance
- Query complexity for available data
- Inability to manage the infrastructure
- Lack sufficient scaling capabilities
- Insufficient storage space
- No central data repository and unified tools



If you had to prioritize improving one aspect of Data Analytics project, which would have the most impact?

- Using the latest optimizing algorithm
- The quality and size of your data
- A deeper network
- Evaluation matrix

Garbage in...
garbage out

On-premises investments

01

Storage

02

Compute

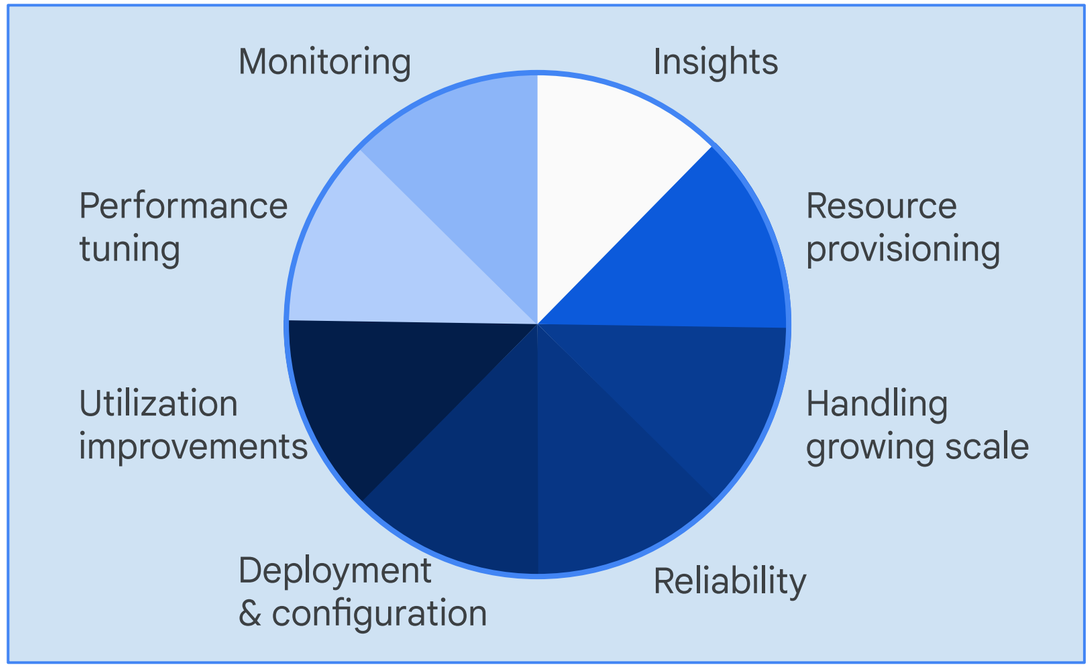
03

Network

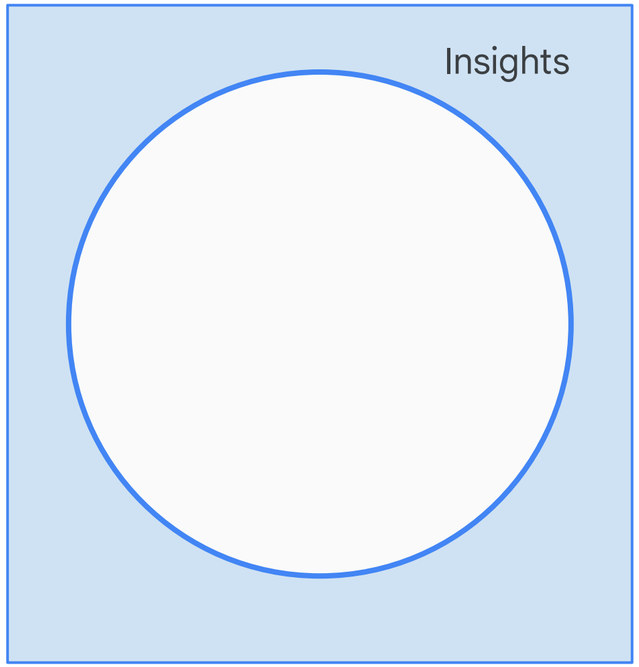
04

Maintenance

Cloud allows data engineers to spend less time managing hardware and enabling scale. Let Google do that for you

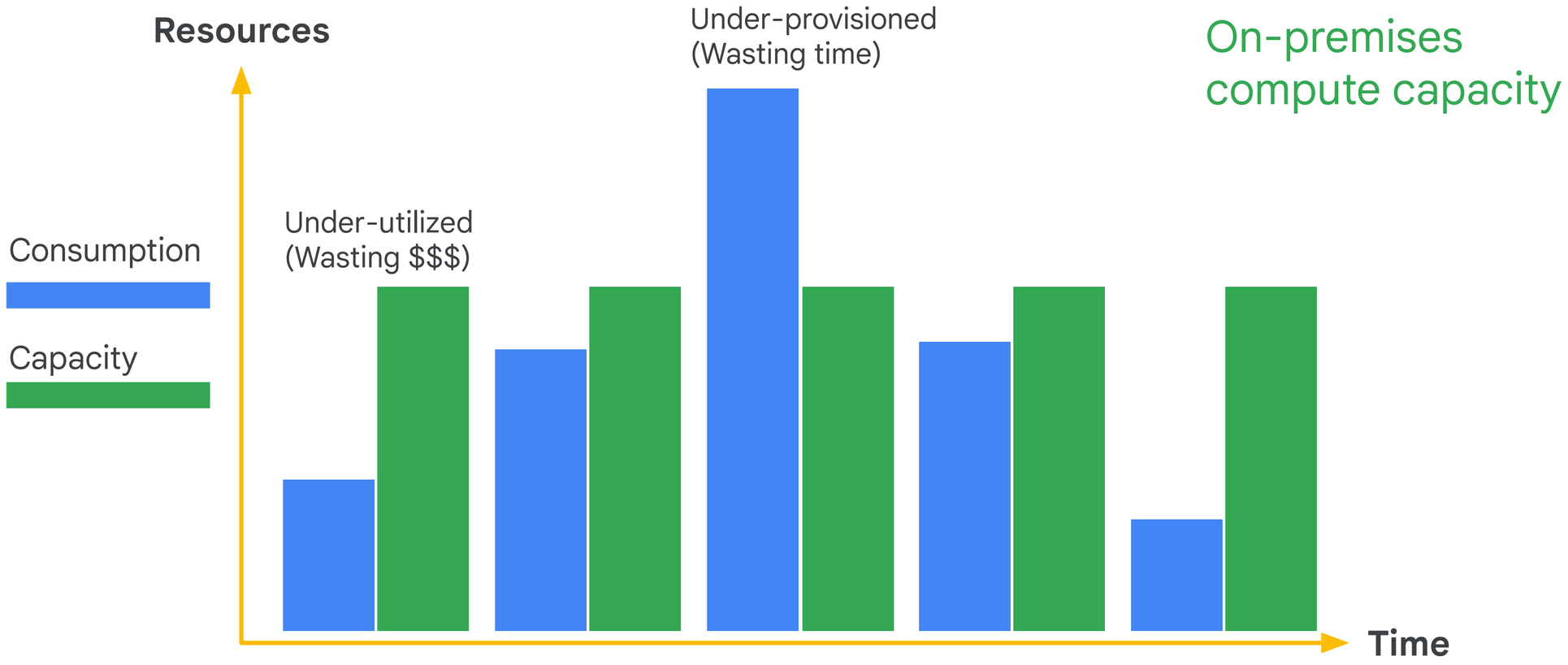


Typical Big Data Processing

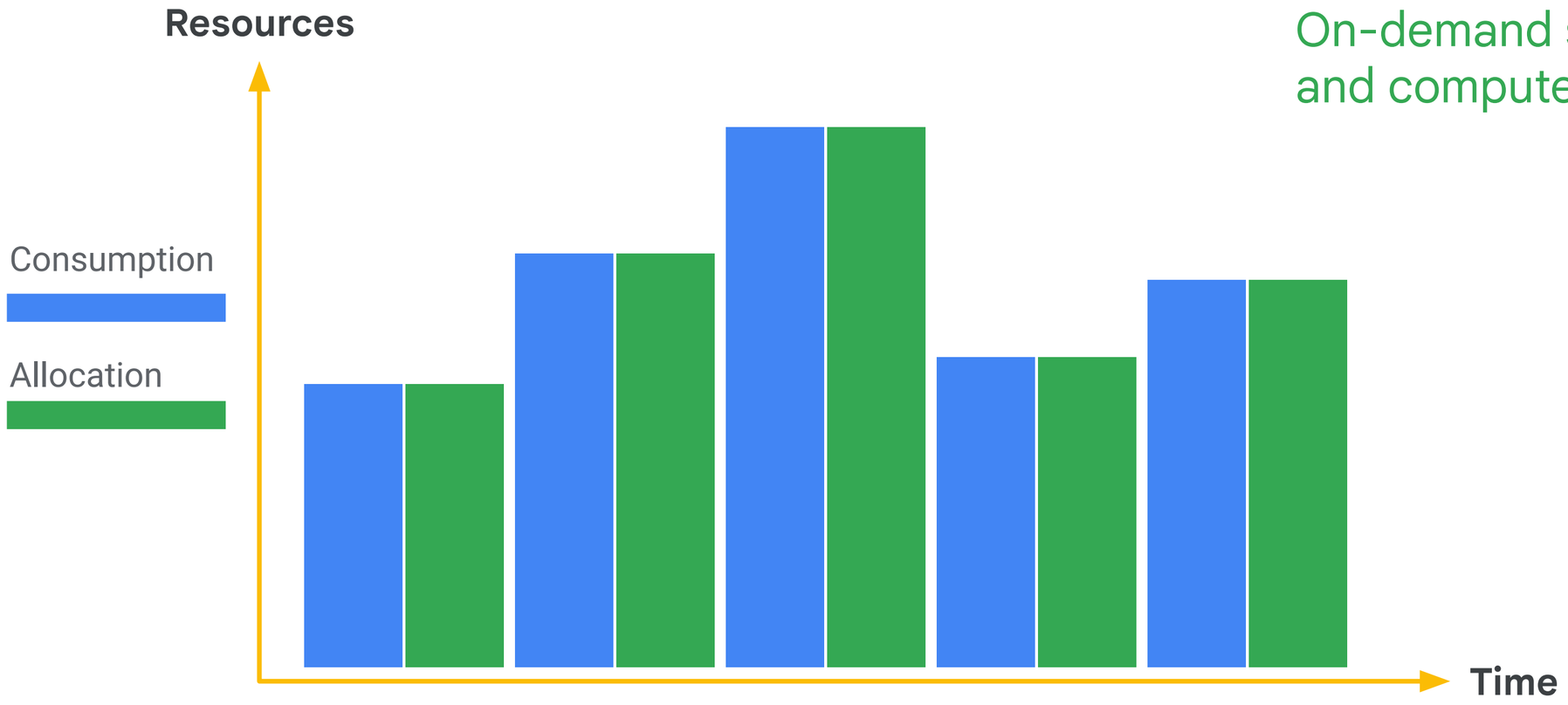


With Google

Challenge: Data Engineers need to manage server and cluster capacity if using on-premise

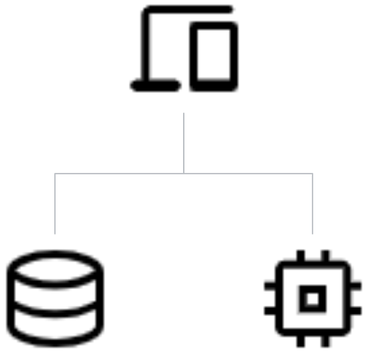


You don't need to provision resources before using BigQuery



On-demand storage and compute

Separate storage and computing for efficient resource allocation



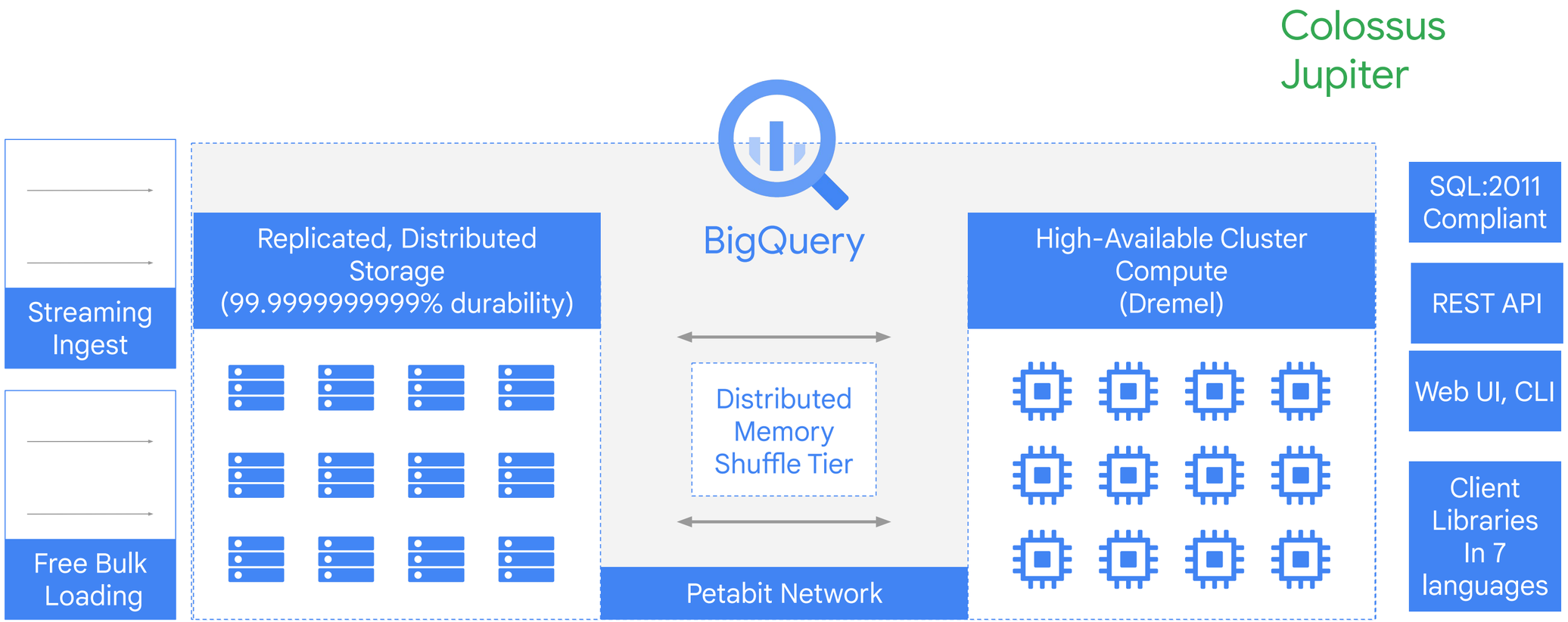
On-premises

- Storage and compute are bounded
- Lower scalability
- Pay for resource allocation

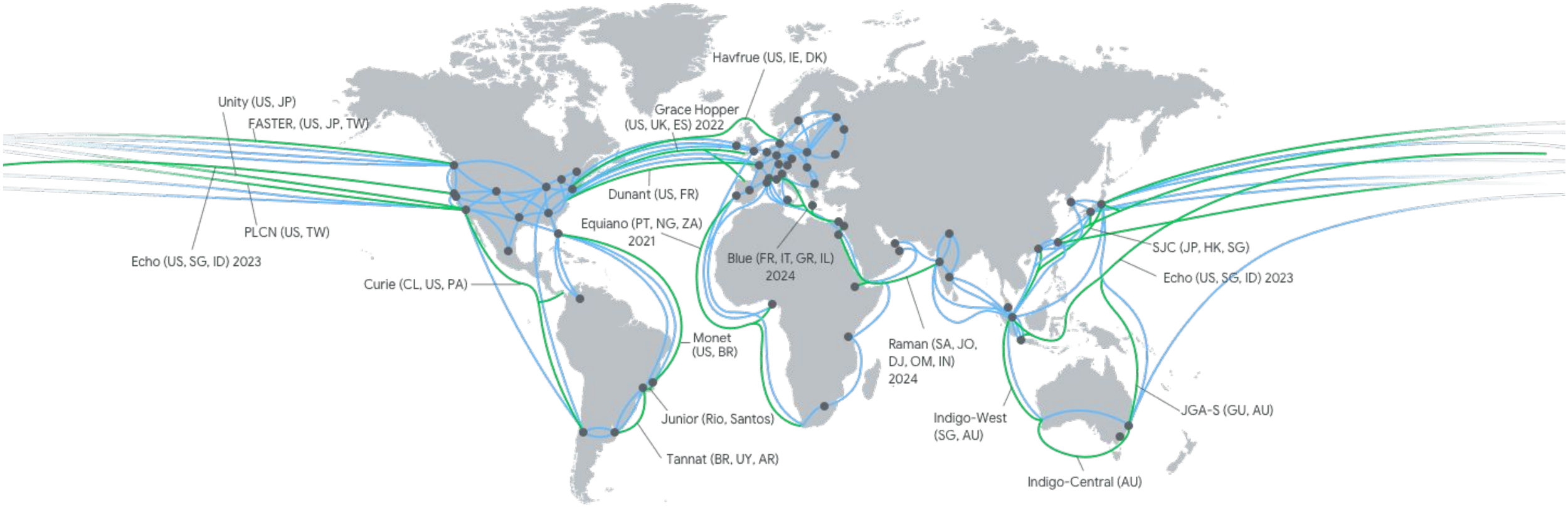
Google Cloud

- Separate services for compute and storage
- High scalability and flexibility
- Pay as you go

The data is physically stored in a redundant way separate from the compute cluster



Leveraging the petabyte scale global network



High quality datasets conform to strict integrity rules

01 Validity

Data conforms to your business rules

Challenges

- Out of range
- Empty fields
- Data mismatch

02 Accuracy

Data conforms to an objective true value

Challenges

- Lookup datasets
- Do not exist

03 Completeness

Create, save, and store datasets

Challenges

- Missing data

04 Consistency

Derive insights from data

Challenges

- Duplicate records
- Concurrency issues

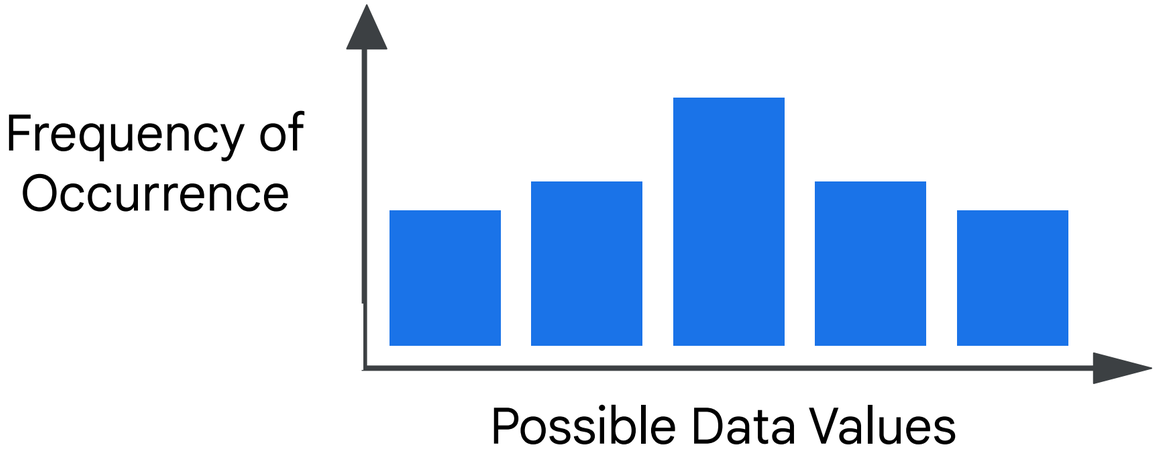
05 Uniformity

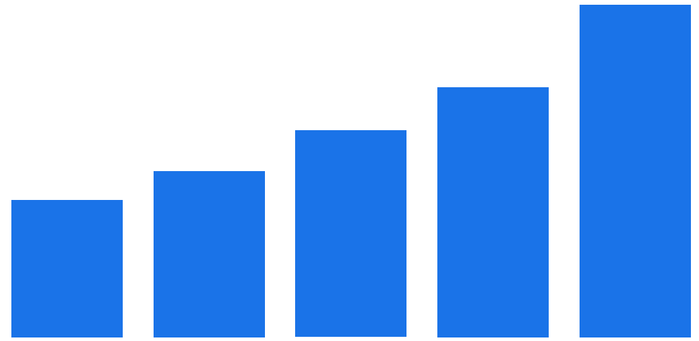
Explore and present data

Challenges

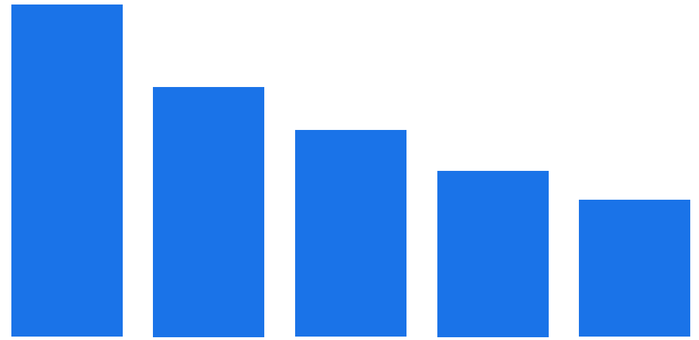
- Same units of measurement

Understanding dataset skew (distribution of values)




Skewed Left


Heavy Skew


Skewed Right

Different ways to clean and prepare your data

01

BigQuery

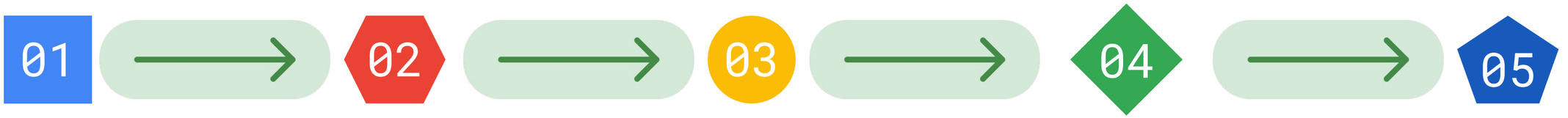
02

Dataprep

03

Data Fusion

Analysing data pipeline



Ingest	Transform	Store	Analyze	Visualize
Get data in	Prepare, clean, apply transformation	Save and store dataset	Derive insights from data	Explore and present results

Analysing data pipeline (ELT)

Ingest

Get data in

BigQuery Storage (import)

BigQuery Transfer

Transform

Prepare, clean, apply transformation

BigQuery Analysis (SQL)

Dataprep

Data Fusion Wrangler

Store

Save and store dataset

BigQuery Storage (tables)

Cloud Storage (buckets)

BigQuery External Tables

Analyze

Derive insights from data

BigQuery Analysis (SQL)

Visualize

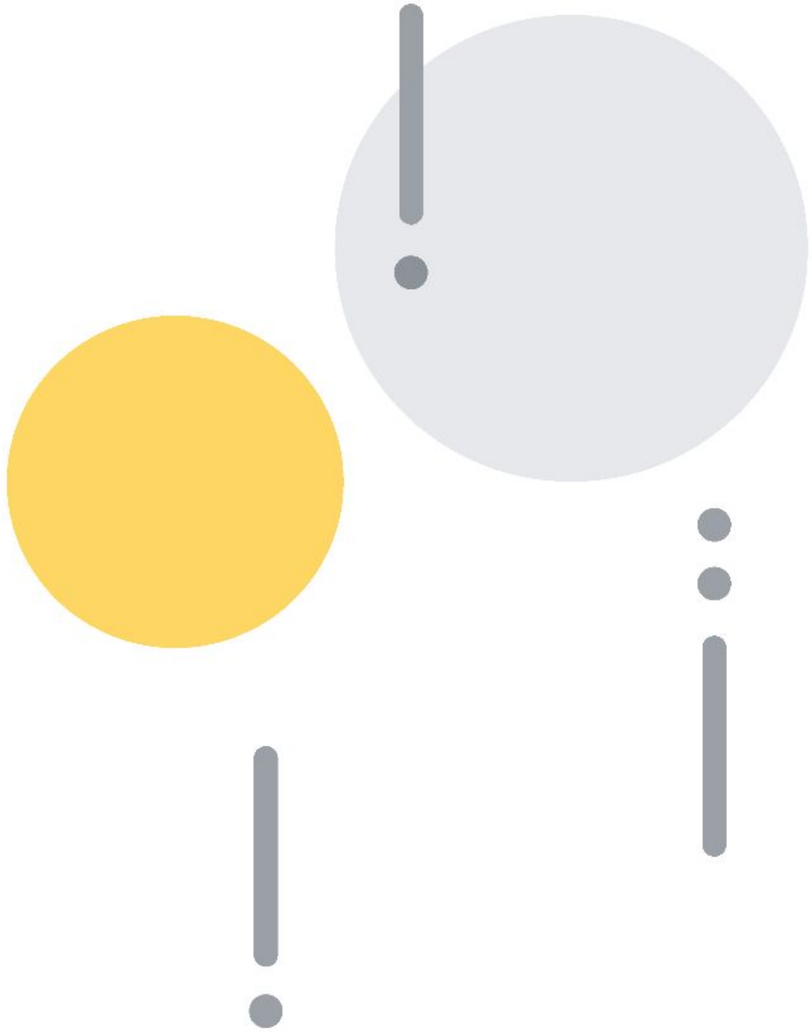
Explore and present results

Looker

Looker Studio (Data Studio)

* 3rd Party Tools (e.g.Tableau, Qlik)

Sample Questions Review



Sample Question 1

Your company is loading comma-separated values (CSV) files into Google BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file. What is the most likely cause of this problem?

Sample Question 1

Your company is **loading comma-separated values (CSV) files into Google BigQuery**. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file. What is the most likely cause of this problem?

Sample Question 1

Your company is **loading comma-separated values (CSV) files into Google BigQuery**. The data is **fully imported successfully**; however, the imported data is not matching byte-to-byte to the source file. What is the most likely cause of this problem?

Sample Question 1

Your company is **loading comma-separated values (CSV) files into Google BigQuery**. The data is **fully imported successfully**; however, the imported data is **not matching byte-to-byte** to the source file. What is the most likely cause of this problem?

Sample Question 1

Your company is **loading comma-separated values (CSV) files into Google BigQuery**. The data is **fully imported successfully**; however, the imported data is **not matching byte-to-byte** to the source file. What is the most likely cause of this problem?

A. The CSV data loaded in BigQuery is not flagged as CSV.

Sample Question 1

Your company is **loading comma-separated values (CSV) files into Google BigQuery**. The data is **fully imported successfully**; however, the imported data is **not matching byte-to-byte** to the source file. What is the most likely cause of this problem?

A. The CSV data loaded in BigQuery is not flagged as CSV. ✕

Sample Question 1

Your company is **loading comma-separated values (CSV) files into Google BigQuery**. The data is **fully imported successfully**; however, the imported data is **not matching byte-to-byte** to the source file. What is the most likely cause of this problem?

A. The CSV data loaded in BigQuery is not flagged as CSV. ✕
B. The CSV data has invalid rows that were skipped on import.

Sample Question 1

Your company is **loading comma-separated values (CSV) files into Google BigQuery**. The data is **fully imported successfully**; however, the imported data is **not matching byte-to-byte** to the source file. What is the most likely cause of this problem?

A. The CSV data loaded in BigQuery is not flagged as CSV. ✗
B. The CSV data has invalid rows that were skipped on import. ✗

Sample Question 1

Your company is **loading comma-separated values (CSV) files into Google BigQuery**. The data is **fully imported successfully**; however, the imported data is **not matching byte-to-byte** to the source file. What is the most likely cause of this problem?

A. The CSV data loaded in BigQuery is not flagged as CSV.
B. The CSV data has invalid rows that were skipped on import.
C. The CSV data has not gone through an ETL phase before loading into BigQuery.

Sample Question 1

Your company is **loading comma-separated values (CSV) files into Google BigQuery**. The data is **fully imported successfully**; however, the imported data is **not matching byte-to-byte** to the source file. What is the most likely cause of this problem?

A. The CSV data loaded in BigQuery is not flagged as CSV. ✗
B. The CSV data has invalid rows that were skipped on import. ✗
C. The CSV data has not gone through an ETL phase before loading into BigQuery. ✗

Sample Question 1

Your company is **loading comma-separated values (CSV) files into Google BigQuery**. The data is **fully imported successfully**; however, the imported data is **not matching byte-to-byte** to the source file. What is the most likely cause of this problem?

A. The CSV data loaded in BigQuery is not flagged as CSV.
B. The CSV data has invalid rows that were skipped on import.
C. The CSV data has not gone through an ETL phase before loading into BigQuery.
D. The CSV data loaded in BigQuery is not using BigQuery's default encoding.

Sample Question 1

Your company is **loading comma-separated values (CSV) files into Google BigQuery**. The data is **fully imported successfully**; however, the imported data is **not matching byte-to-byte** to the source file. What is the most likely cause of this problem?

A. The CSV data loaded in BigQuery is not flagged as CSV. ✗
B. The CSV data has invalid rows that were skipped on import. ✗
C. The CSV data has not gone through an ETL phase before loading into BigQuery. ✗
D. The CSV data loaded in BigQuery is not using BigQuery's default encoding. ✓

OPTIONAL study materials:

[READING]

- Google Cloud Data Analytics
- What is Big Data?
- Data science with R on Google Cloud: Exploratory data analysis tutorial
- Analyzing FHIR data in BigQuery
- Genomic data processing reference architecture
- Geospatial analytics architecture
- Set up a regulatory reporting architecture with BigQuery

[VIDEOS]

- Analyzing Big Data in less time with Google BigQuery
- BigQuery Overview
- Tableau's Visual Analytics Brings Big Data in Google's cloud to Life
- Big Data for Training Models in the Cloud
- Exporting Data from Google Analytics

Make sure to...

Enjoy the journey as much as the destination!