Data Acquisition

Wk1. ETL, Virtual Environment, GCP

Diane Woodbridge, PH.D

UNIVERSITY OF SAN FRANCISCO

CHANGE THE WORLD FROM HERE

About Data Acquisition

This hands-on course to build end-to-end data acquisition pipelines using ETL (Extract, Transform, Load) principles.

- Collect, clean, and store data from a variety of sources and formats, including PDF, Excel (xlsx), Word (docx), HTML, JSON, and databases with wget, API integration, and web scraping from both static and dynamic websites.
- Students will develop a web application, containerize it with Docker, and deploy it to a cloud environment.

Learning Outcomes

By the end of the class, students will be able to

- Design and implement data acquisition pipelines following ETL (Extract, Transform, Load) principles using Python.
- Collect data from diverse sources including files (PDF, Excel, Word), web APIs, and websites (both static and dynamic).
- Interact with RESTful APIs for programmatic data access and integration.
- Automate data extraction from dynamic web pages using tools like Playwright.
- Apply data cleaning and transformation techniques to prepare raw data for analysis or storage.
- Develop a web application using Python frameworks.
- Containerize and deploy applications using Docker and cloud platforms.

Class Schedule

Class Schedule

- Session 1 : 10 AM - 12 PM #527
- Session 2 : 1 PM - 3 PM #529/527 (Do not enter the room until 12:55)
- Please come to your assigned session.

Office Hours

- Mon/Thu 12:00 - 12:45 PM
- Location: #606

Class Git Repo: https://github.com/dianewoodbridge/msds692_data_acquisition_2025

Class Schedule

Week	Topic	Assignment	Quiz
1	Introduction to ETL, Virtual Environment, GCP Setup		
2	Git Branch File-based Acquisition and Cloud Storage Streamlit Application		
3	APIs for Programmatic Access	HW1	Quiz1
4	Creating APIs using FastAPI GCP Cloud Run Containerization and Docker	Group Project 1	
5	Static Web Scraping Dynamic Web Scraping	HW2	Quiz2
6	Dynamic Web Scraping Retrieval from databases using SQLAlchemy	HW3 Group Project 2	
7	Chatbot Development using LLMs	HW4	Quiz3
8	Final Project	Group Project 3	

Grades Breakdown

Individual Assignments : 16% (4% Each)

- Grader : Helen Lin (hlin65@usfca.edu)

Group Assignments : 20%

Quiz: 60% (20% Each)

- Will allow cheatsheet - A single sheet of letter size (both sides), hand-written only

Attendance/Professionalism : 4%

Grading Policies

- See more details including plagiarism and other policies on the syllabus.
- The expected final score for this course is 85 ± 5 (and close to normal distributions). The following grades will be given if the class grade distribution falls within the expectation.
  - 90-100 : A-, A, and A+
  - 80-90: B+, B and B-
  - 70-80: C+, C and C-
  - Under 70: F
- However, if the grade distribution does not meet the aforementioned criteria, grades will be curved.

Any questions or requests?

Nobody has responded yet.

Hang tight! Responses are coming in.

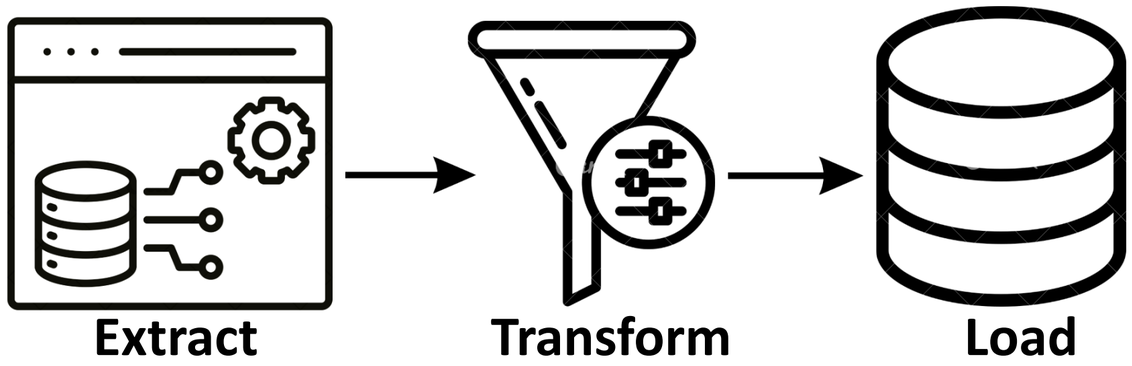
Contents

Contents

Introduction to ETL

ETL Pipeline

- **E**xtract: Retrieval of data from multiple sources including APIs, websites, databases, applications, etc.
- **T**ransform: Cleaning data in a valid data type/format, and converting into a proper format for storages, and analysis.
- **L**oad: Moving data to a data warehouse, database, or other storage system.



Introduction to ETL

Benefits of ETL Pipeline

- Improved Data Integration
  - Collect and combine data from various sources to a single consistent format, providing a unified view.
- Enhanced Data Quality
  - Through transformation, it eliminates inaccuracies, duplicates, and inconsistencies.
- Enhanced Data Accessibility
  - Consolidate data into a centralized database or storage.
- Increased Efficiency
  - Automating an ETL pipeline can reduce manual efforts and time

Introduction to ETL

Time and Effort for ETL as a Data Scientist

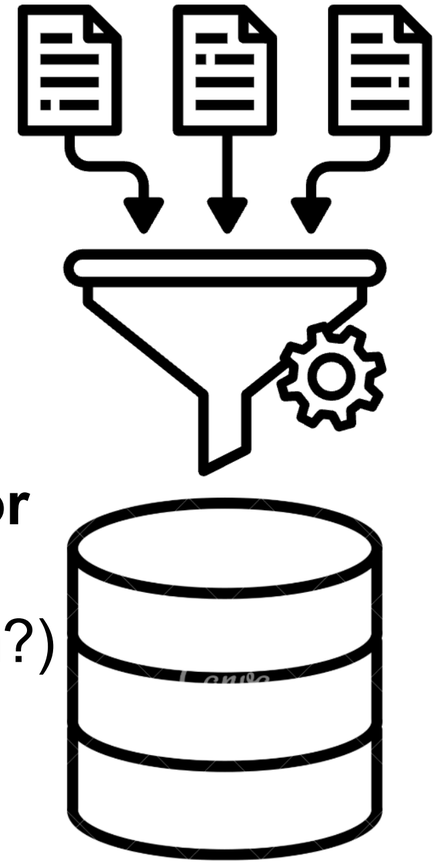
- "**Our respondents reported that almost half of their time is spent on the combined tasks of data loading and cleansing.** Data visualization tasks come second, taking about 21% of time. Modeling tasks consume the remaining third of a data professional's time, with selection comprising 11%, training and scoring 12%, and deployment 11%." - Anaconda, State of Data Science - Moving from hype toward maturity, 2020

- "**Practitioners continue to struggle with data preparation and cleaning.**" - Anaconda, State of Data Science - AI takes center stage, 2023

Exercise

Choose one of the following domain and come up with an interesting problem to investigate with your partners.

- Customer behavior data on a retail store website
- Social media user data
- Healthcare patient records
- Student performance and retention data



How would you design your ETL pipeline for your data analysis for the chosen domain?

- Extract (Which data would be useful? Where does data come from?)
- Transform (How do we clean/prepare it?)
- Load (Where does it go?)

Which one is not "Transformation" in ETL?

Collecting data from the web

Imputing missing data

Converting data to numerical format

Deleting duplicate data

Which one is not "Transformation" in ETL?

Collecting data from the web

0%

Imputing missing data

0%

Converting data to numerical format

0%

Deleting duplicate data

0%

Which one is not "Transformation" in ETL?

Collecting data from the web

0%

Imputing missing data

0%

Converting data to numerical format

0%

Deleting duplicate data

0%

Contents

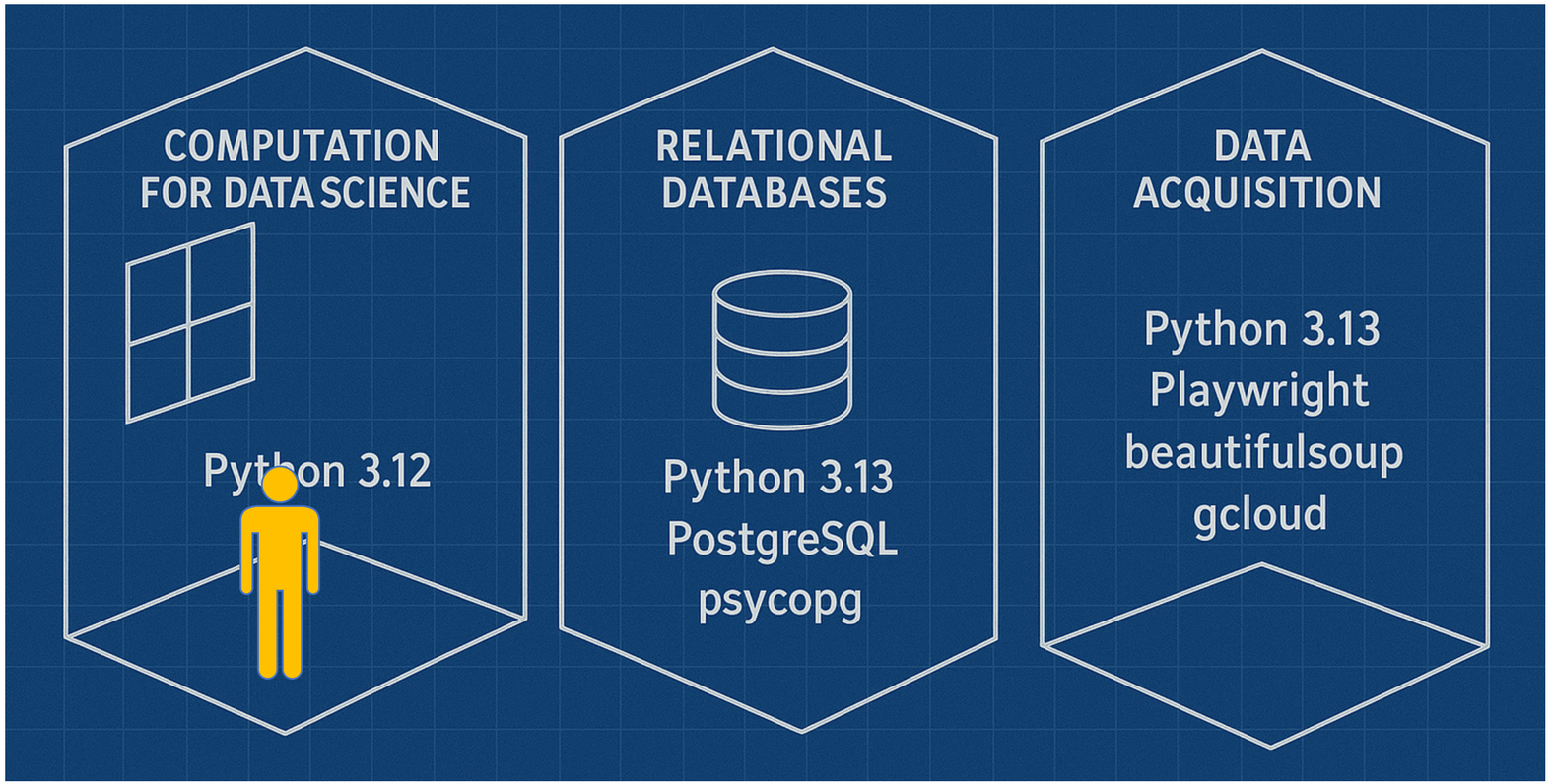
Virtual Environment

A project team member might work on several projects requiring different versions of Python or other packages.

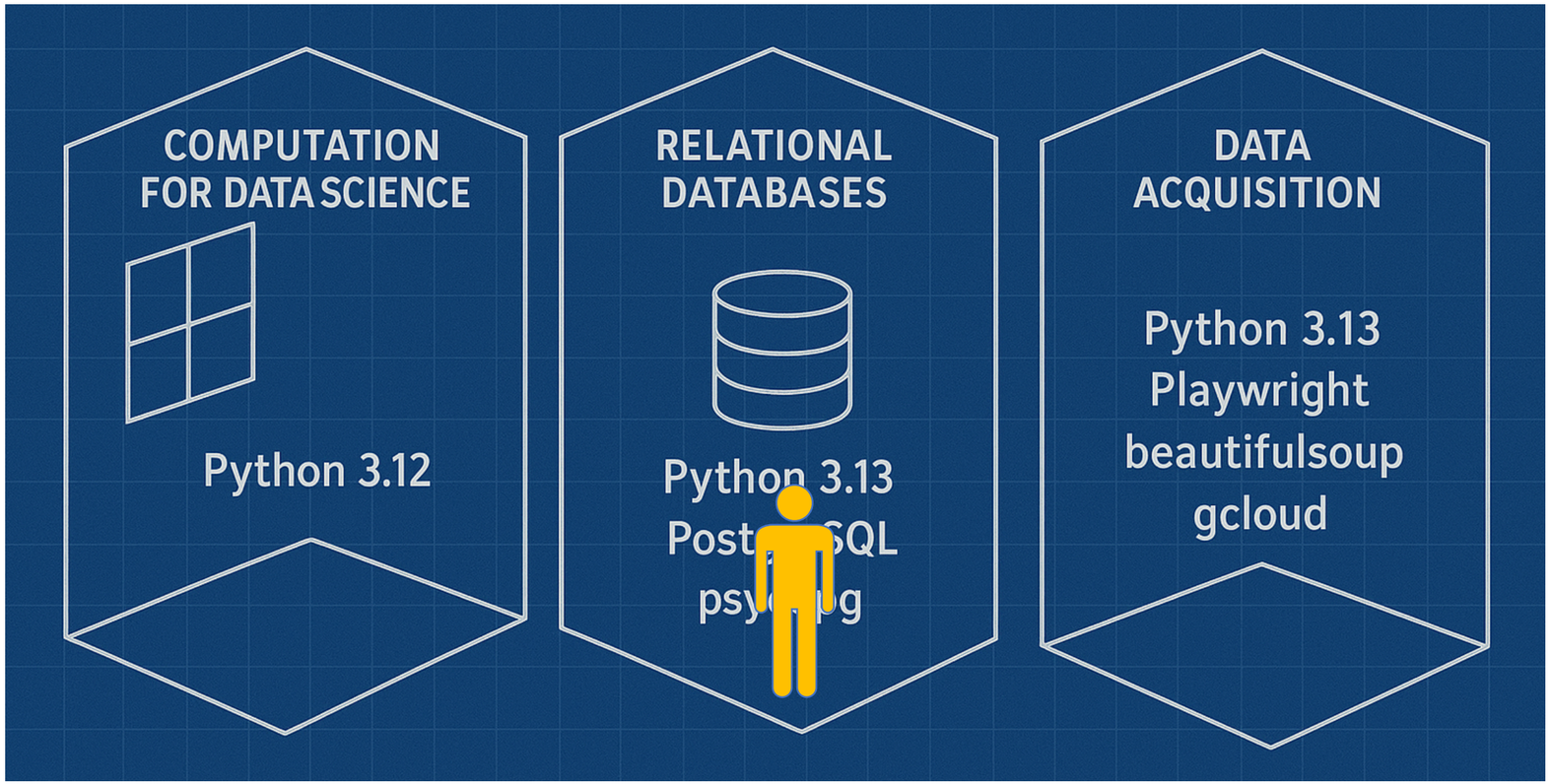
Virtual environment

- A self-contained/isolated environment that houses a specific Python interpreter and its associated libraries/packages, creating an isolated space for a project
- Benefits
  - Reproducibility and Portability
    - Makes it easy to reproduce the exact environment on another machine, ensuring that the project runs consistently
  - Dependency Management and Conflict Prevention
    - Create isolated environments for each project, allowing you to install specific library versions without affecting other projects

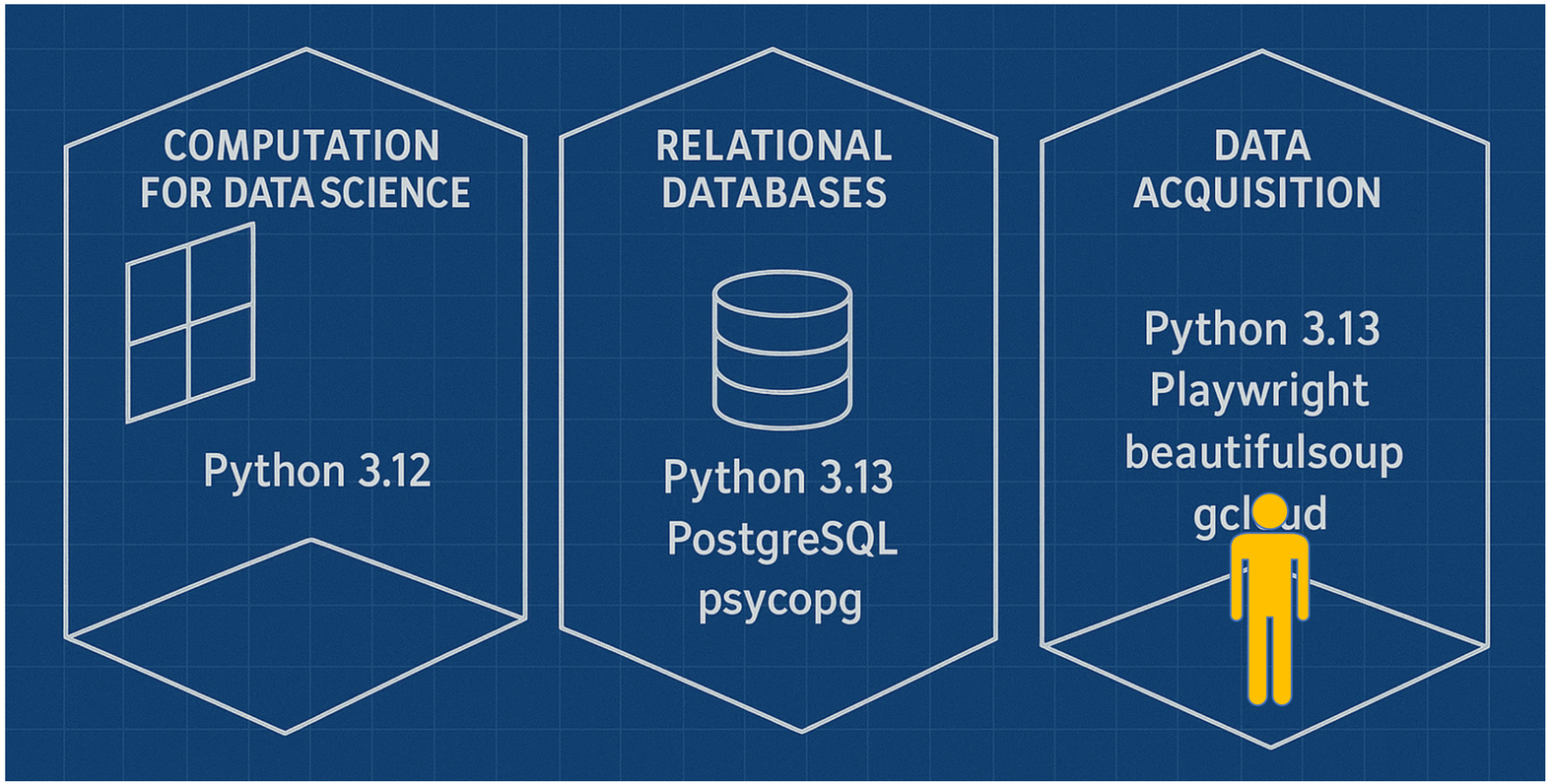
Virtual Environment



Virtual Environment



Virtual Environment



Virtual Environment

Create a virtual environment

```
$ conda create --name msds692 python=3.13 -y
```

- `-y` : yes to all the questions.
- It will create ”msds692” directory under /Users/USER_ID/anaconda3/envs/

```
dwoodbridge@ML-ITS-210588 HW % conda create --name msds692 python=3.13 -y
2 channel Terms of Service accepted
Retrieving notices: done
Channels:
 - defaults
Platform: osx-arm64
Collecting package metadata (repodata.json): done
Solving environment: done


==> WARNING: A newer version of conda exists. <==
    current version: 25.5.1
    latest version: 25.7.0

Please update conda by running
```

Create an Environment

Activate/Deactivate Virtual Environment

- Activating the environment
  `$ conda activate msds692`
- See available environments
  `$ conda env list`
- Determining the current environment
  `$ conda info --envs`
- Deactivating the environment
  `$ conda deactivate`

```
dwoodbridge@ML-ITS-210588 HW % conda activate msds692
(msds692) dwoodbridge@ML-ITS-210588 HW % conda info --envs

# conda environments:
#
                         /opt/anaconda3/envs/Admin
                         /opt/anaconda3/envs/Dicom
                         /opt/anaconda3/envs/DistributedComputing
                         /opt/anaconda3/envs/metformin
                         /opt/anaconda3/envs/ylabs
base                     /opt/homebrew/anaconda3
msds692                * /opt/homebrew/anaconda3/envs/msds692

(msds692) dwoodbridge@ML-ITS-210588 HW % conda deactivate
dwoodbridge@ML-ITS-210588 HW %
```

Create an Environment

Activate/Deactivate Virtual Environment

- Activating the environment
  `$ conda activate msds692`
- See available environments
  `$ conda env list`
- Determining the current environment
  `$ conda info --envs`
- Deactivating the environment
  `$ conda deactivate`

```
(base) dwoodbridge@ML-ITS-210588 HW % conda activate msds692
(msds692) dwoodbridge@ML-ITS-210588 HW % conda info --envs

# conda environments:
#
                         /opt/anaconda3/envs/Admin
                         /opt/anaconda3/envs/Dicom
                         /opt/anaconda3/envs/DistributedComputing
                         /opt/anaconda3/envs/metformin
                         /opt/anaconda3/envs/ylabs
base                     /opt/homebrew/anaconda3
msds692                * /opt/homebrew/anaconda3/envs/msds692

(msds692) dwoodbridge@ML-ITS-210588 HW % conda deactivate
dwoodbridge@ML-ITS-210588 HW %
```

Indicating the current environment

Create an Environment

Activate/Deactivate Virtual Environment

- Activating the environment
  `$ conda activate msds692`
- See available environments
  `$ conda env list`
- Determining the current environment
  `$ conda info --envs`
- Deactivating the environment
  `$ conda deactivate`

```
dwoodbridge@ML-ITS-210588 HW % conda activate msds692
(msds692) dwoodbridge@ML-ITS-210588 HW % conda info --envs

# conda environments:
#
                         /opt/anaconda3/envs/Admin
                         /opt/anaconda3/envs/Dicom
                         /opt/anaconda3/envs/DistributedComputing
                         /opt/anaconda3/envs/metformin
                         /opt/anaconda3/envs/ylabs
base                     /opt/homebrew/anaconda3
msds692                * /opt/homebrew/anaconda3/envs/msds692

(msds692) dwoodbridge@ML-ITS-210588 HW % conda deactivate
dwoodbridge@ML-ITS-210588 HW %
```

Exercise

What is your env after deactivating the environment?

Let’s activate msds692 again.

- `conda activate msds692`

Create an Environment

Virtual Environment

- Export the environment file to share and reproduce the environments including all the packages with corresponding versions.

```
$ conda env export > msds692_environment.yml
```

```
name: msds692
channels:
  - defaults
dependencies:
  - bzip2=1.0.8=h80987f9_6
  - ca-certificates=2025.7.15=hca03da5_0
  - expat=2.7.1=h313beb8_0
  - libcxx=19.1.7=hb09ecce_3
  - libffi=3.4.4=hca03da5_1
  - libmpdec=4.0.0=h80987f9_0
  - ncurses=6.5=hee39554_0
  - openssl=3.0.17=h4ee41c1_0
  - pip=25.1=pyhc872135_2
  - python=3.13.5=h2eb94d5_100_cp313
  - python_abi=3.13=0_cp313
  - readline=8.3=h0b18652_0
  - setuptools=78.1.1=py313hca03da5_0
```

Create an Environment

Virtual Environment

- Export the environment file to share and reproduce the environments including all the packages with corresponding versions.

```
$ conda env export > msds692_environment.yml
```

  - YAML (YAML Ain't Markup Language) file
    - Human-readable data serialization format.
    - Uses indentation and key-value pairs to represent data structures.
    - Commonly used for configuration files and in applications where data is being stored or transmitted.

```yaml
name: msds692
channels:
  - defaults
dependencies:
  - bzip2=1.0.8=h80987f9_6
  - ca-certificates=2025.7.15=hca03da5_0
  - expat=2.7.1=h313beb8_0
  - libcxx=19.1.7=hb09ecce_3
  - libffi=3.4.4=hca03da5_1
  - libmpdec=4.0.0=h80987f9_0
  - ncurses=6.5=hee39554_0
  - openssl=3.0.17=h4ee41c1_0
  - pip=25.1=pyhc872135_2
  - python=3.13.5=h2eb94d5_100_cp313
  - python_abi=3.13=0_cp313
  - readline=8.3=h0b18652_0
  - setuptools=78.1.1=py313hca03da5_0
```

https://en.wikipedia.org/wiki/YAML

Virtual Environment

Remove Virtual Environment & Create/Updated from a File

- Remove the environment

```
$  conda remove --name msds692 --all
```

- Create or update an environment using .yml.

```
$  conda env create -f msds692_environment.yml -n msds692
$  conda env update -f msds692_environment.yml -n msds692
```

Example 1

1. Export your environment into msds692_environment.yml.
2. Deactivate your environment.
3. Delete your environment.
4. Create an environment using msds692_environment.yml
5. Update the environment with msds692_environment.yml in the class git hub.

Sharing an exported virtual environment requirements file (e.g., requirements.yml) helps ensure all team members can recreate the same development environment.

True

False

Sharing an exported virtual environment requirements file (e.g., requirements.yml) helps ensure all team members can recreate the same development environment.

Sharing an exported virtual environment requirements file (e.g., requirements.yml) helps ensure all team members can recreate the same development environment.

Contents

Google Cloud Computing (GCP)

Cloud Computing

- Cloud computing provides computing resources (such as storage and infrastructure) on demand, as services over the internet.
  - It manages cloud resources, providing scalability, availability, security, and cost efficiency.

Google Cloud Computing (GCP)

Cloud Computing

- Cloud computing providers offer various services including
  - **Storage** - Store and manage files, objects, and backups at scale. (Ex. GCP Cloud Storage, AWS S3)
  - **Compute** - Run applications on virtual machines, containers, or serverless platforms. (Ex. GCP Compute Engine, GCP Cloud Run, AWS EC2)
  - **Databases** - Managed relational (SQL) and non-relational (NoSQL) databases. (Ex. GCP BigQuery, GCP Datastore, AWS RDS, AWS DynamoDB)
  - **Networking & Security** - Global networking, load balancing, firewalls, identity & access management. (Ex. GCP VPC, GCP Cloud Armor, AWS Shield)
  - **Data Analysis** - Tools for querying, warehousing, and visualizing large datasets. (Ex. GCP BigQuery, GCP Looker, AWS Redshift)
  - **ML & AI** - Pre-trained APIs, AutoML, and custom model training. (Ex. GCP VertexAI, AWS SageMaker)
  - And many others…

https://cloud.google.com/terms/services?hl=en

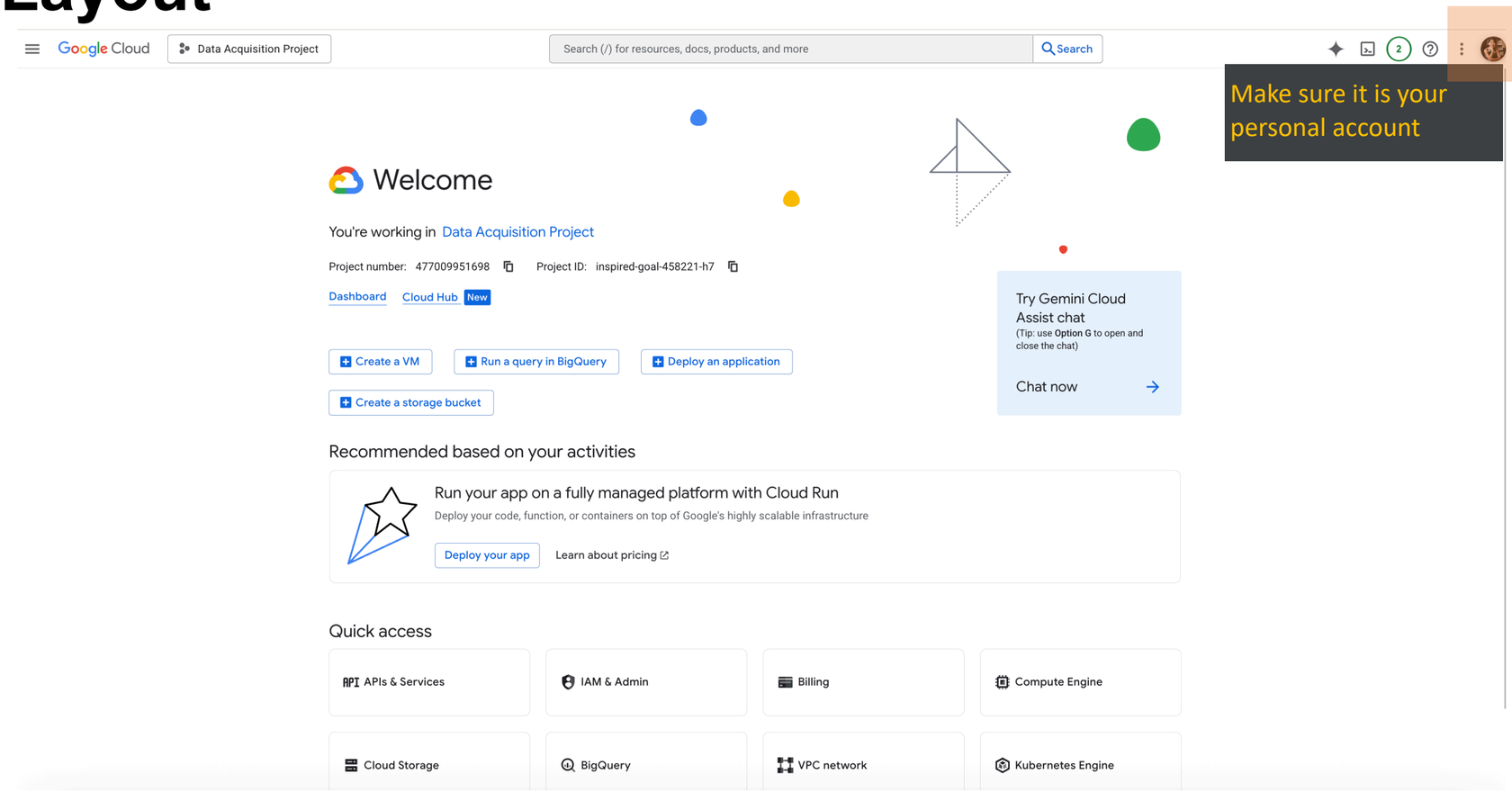
Google Cloud Computing (GCP)

Let’s create a GCP account

- Go to https://console.cloud.google.com/ with your personal google account.
  - USF email often doesn't work with GCP, and **I encourage you to use your personal Gmail account for GCP account and attach the coupon to your personal account.** (USF email is for verifying your student status)
- Student Coupon Retrieval Link ($50)
  - You will be asked for a name and email address, which needs to match your school domain. A confirmation email will be sent to you with a coupon code.

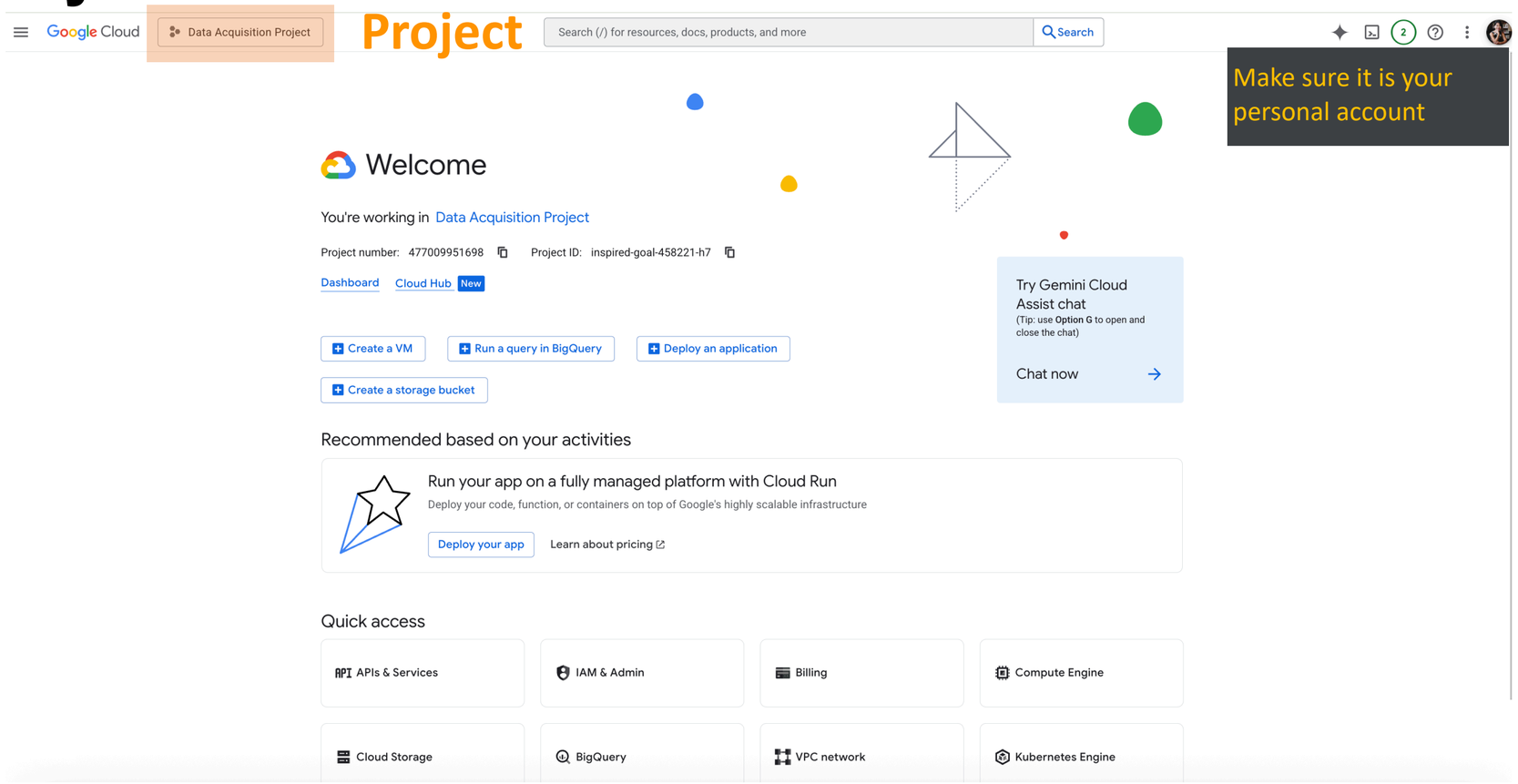
Google Cloud Computing (GCP)

Console Layout



Google Cloud Computing (GCP)

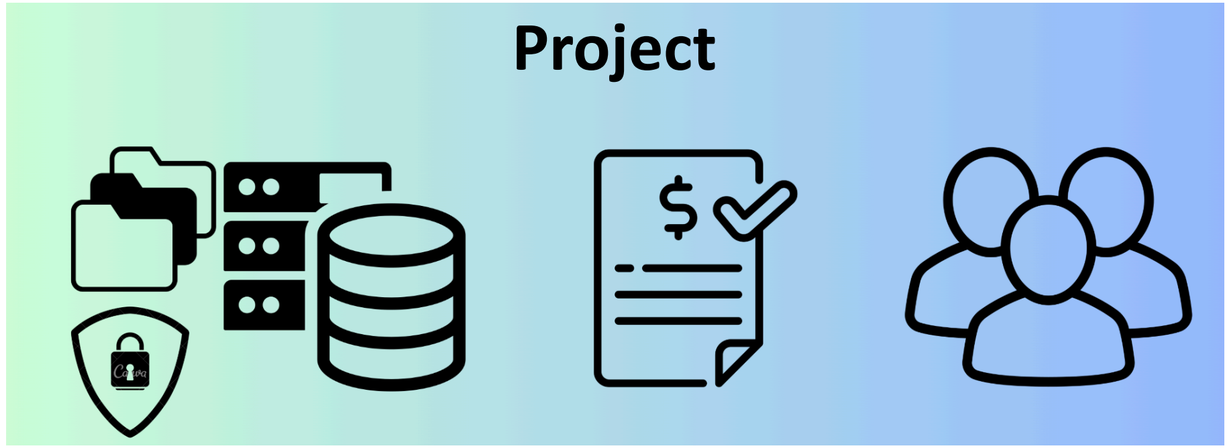
Console Layout



Google Cloud Computing (GCP)

“Project” in GCP

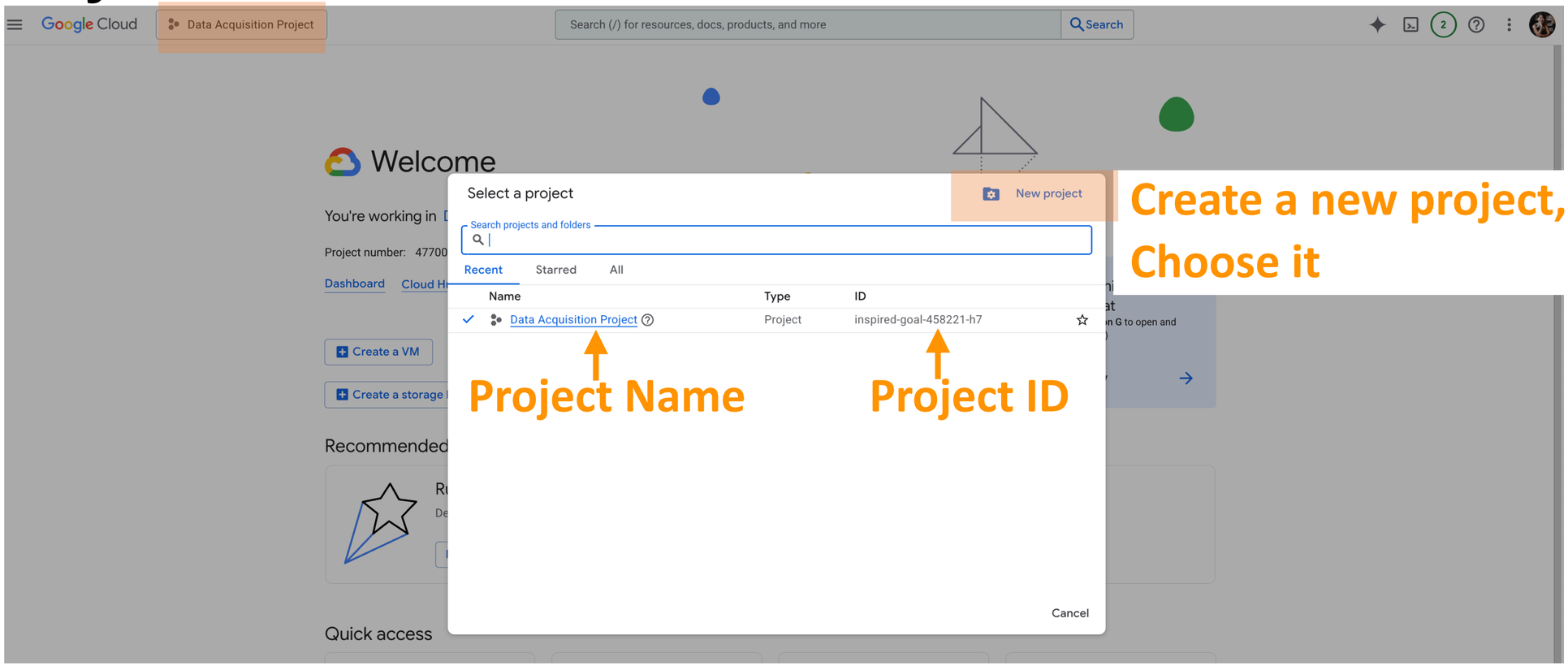
- A project is an organizing entity for what you’re building.
  - A project can have a set of users, a set of services, and billing, authentication, and monitoring settings for those services.
    - Resources within a project can work together easily.



https://cloud.google.com/docs/overview#projects

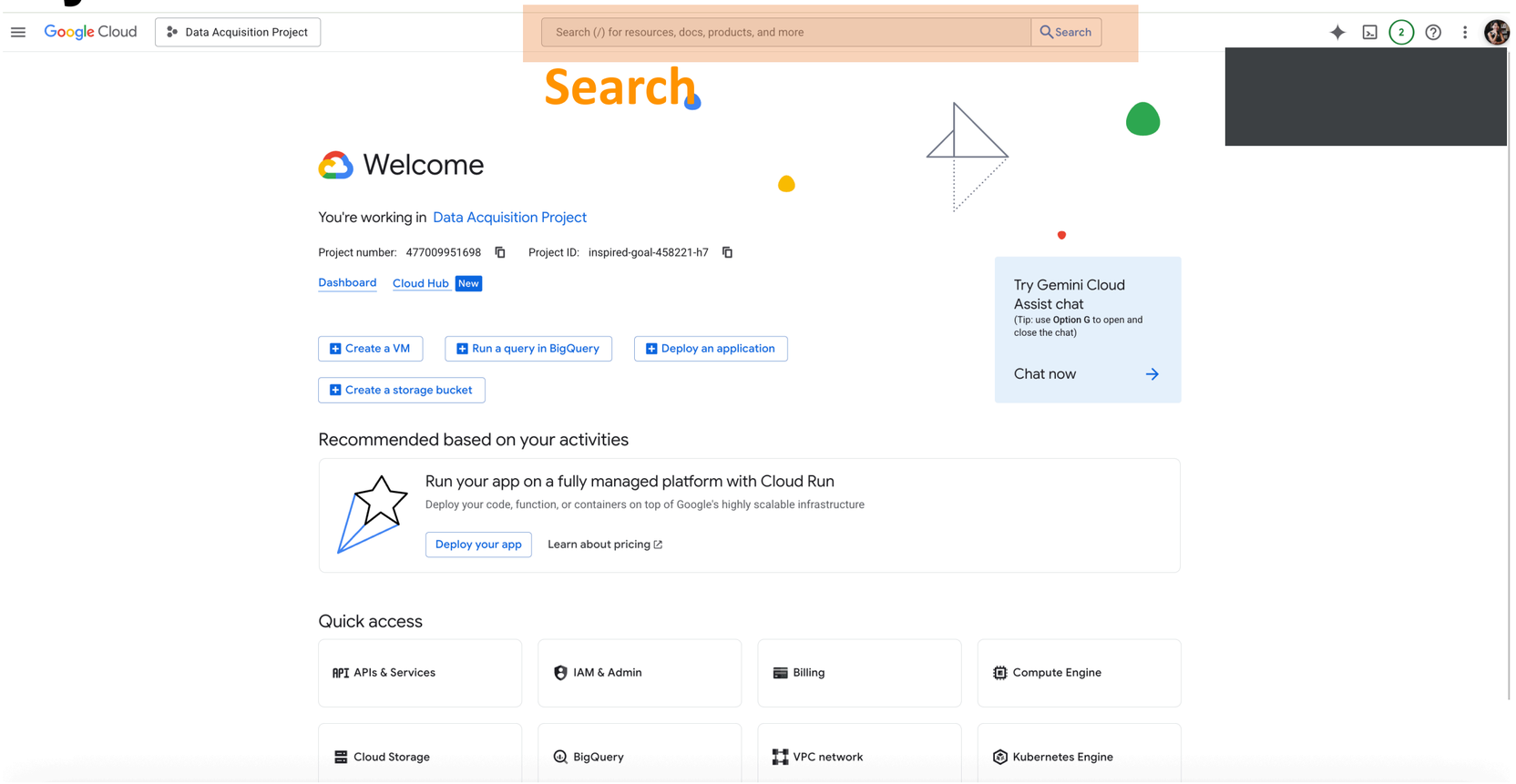
Google Cloud Computing (GCP)

Console Layout



Google Cloud Computing (GCP)

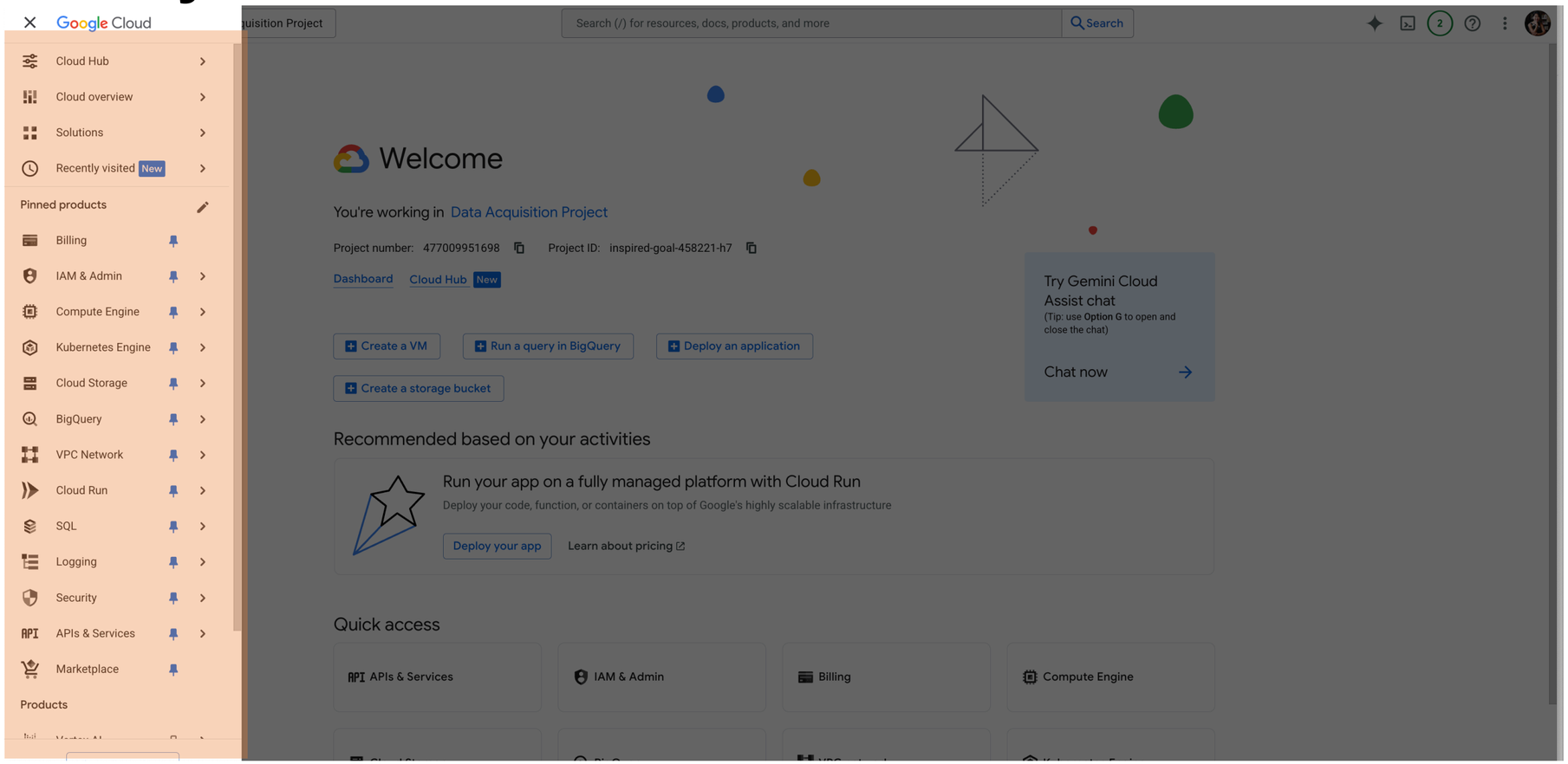
Console Layout



Google Cloud Computing (GCP)

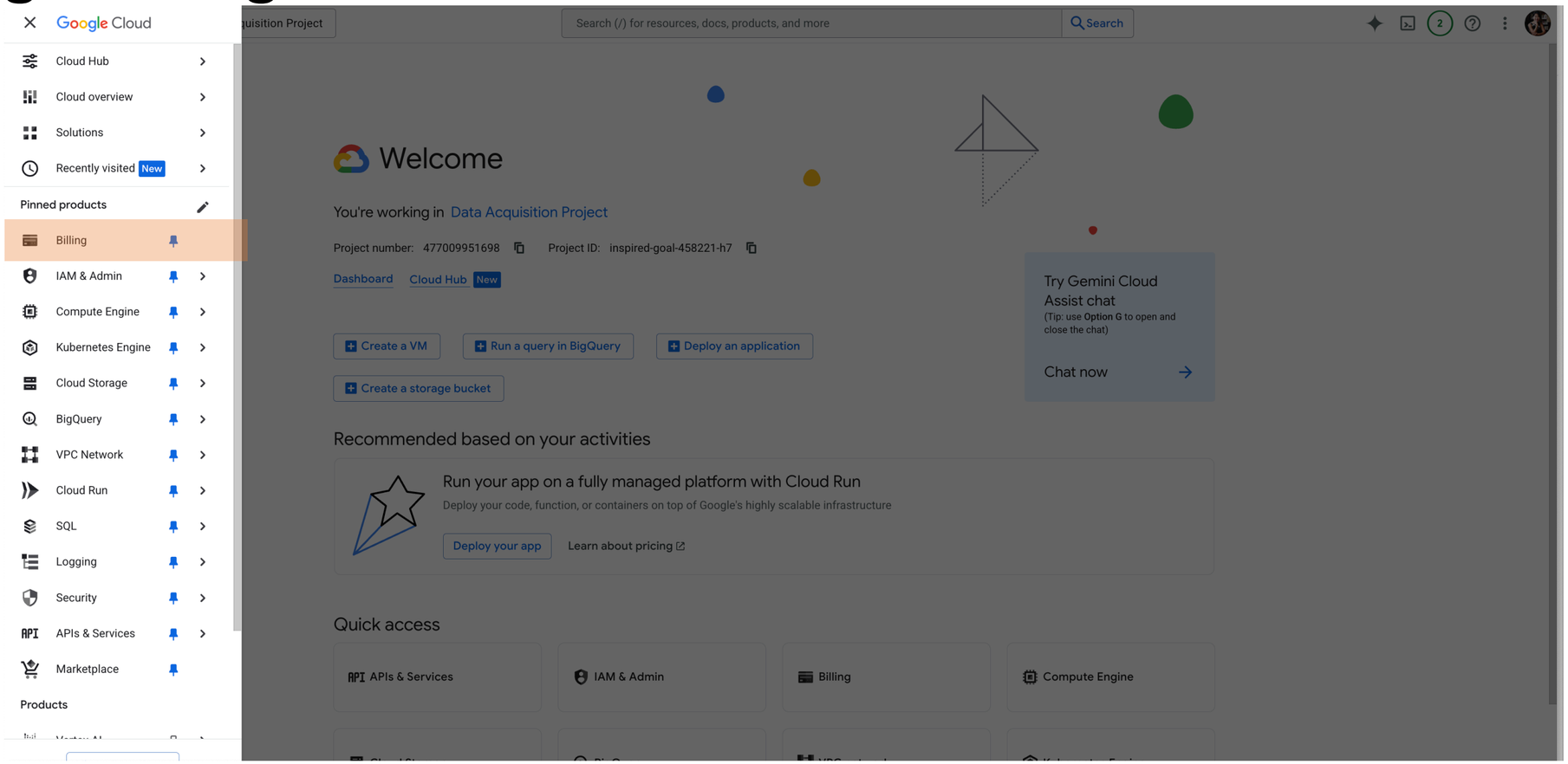
Console Layout

Navigation



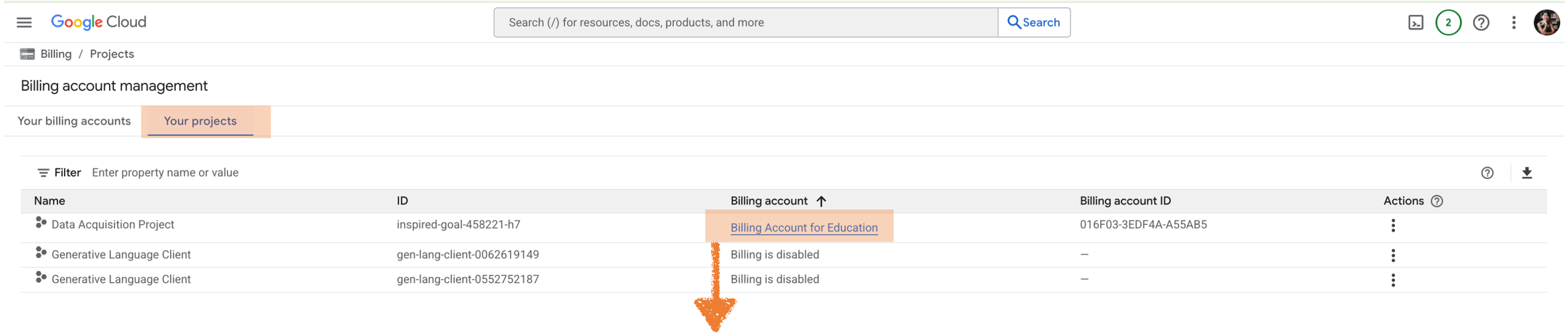
Google Cloud Computing (GCP)

Manage Billing



Google Cloud Computing (GCP)

Manage Billing

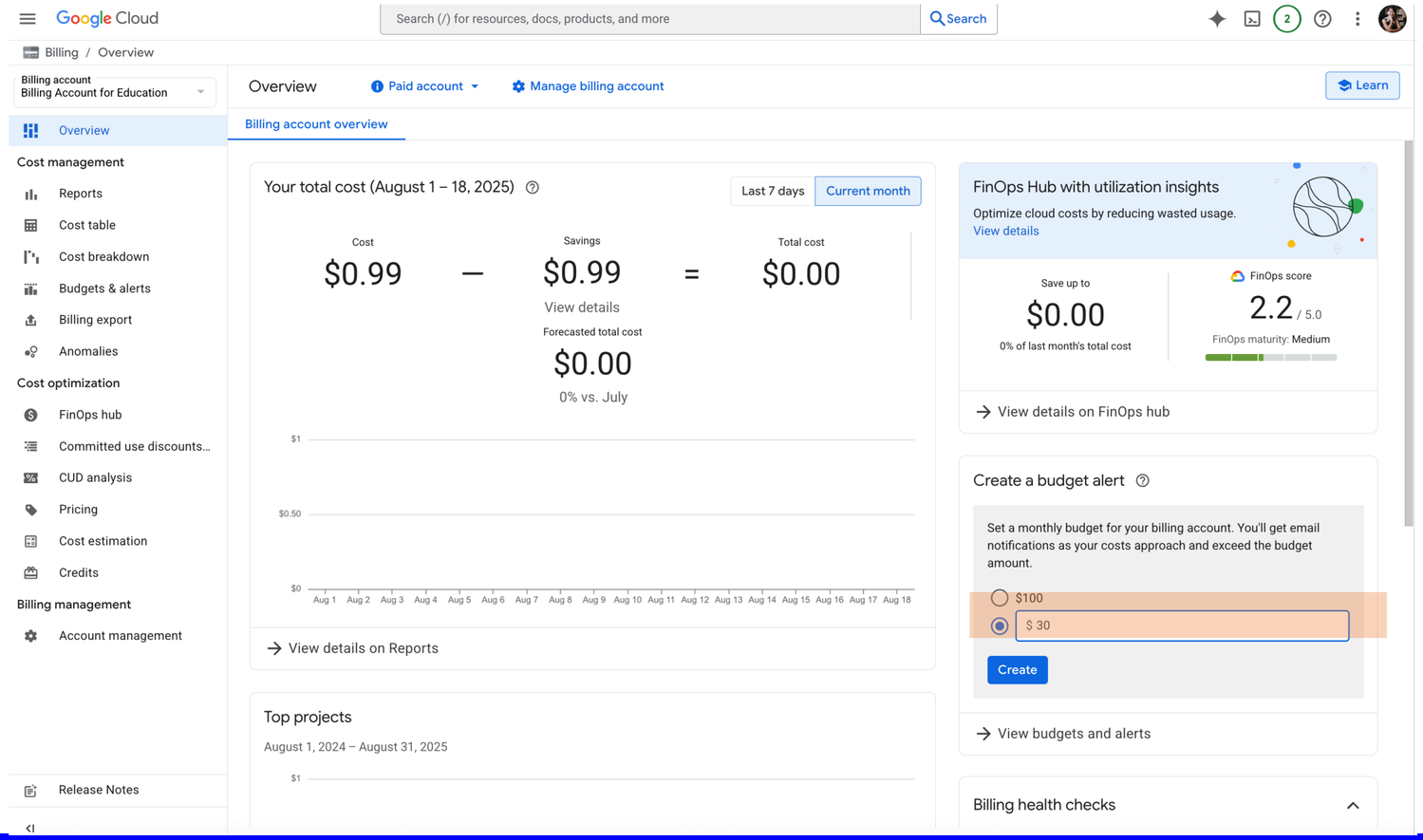


Make sure that your project is attached to “Billing Account for Education”

Click this to set a budget alert.

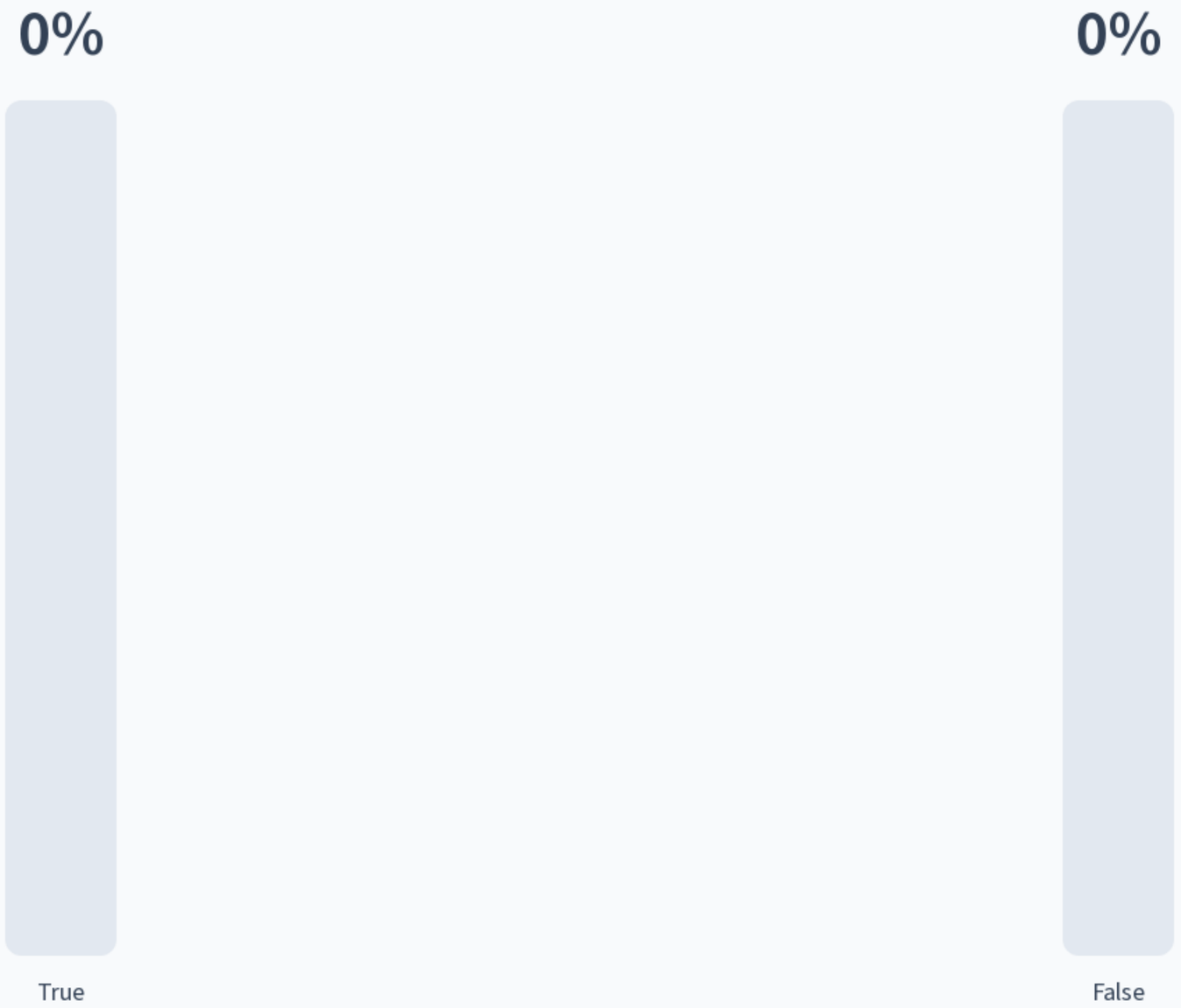
Google Cloud Computing (GCP)

Manage Billing



Set the monthly budget account small enough to avoid excessive charges.

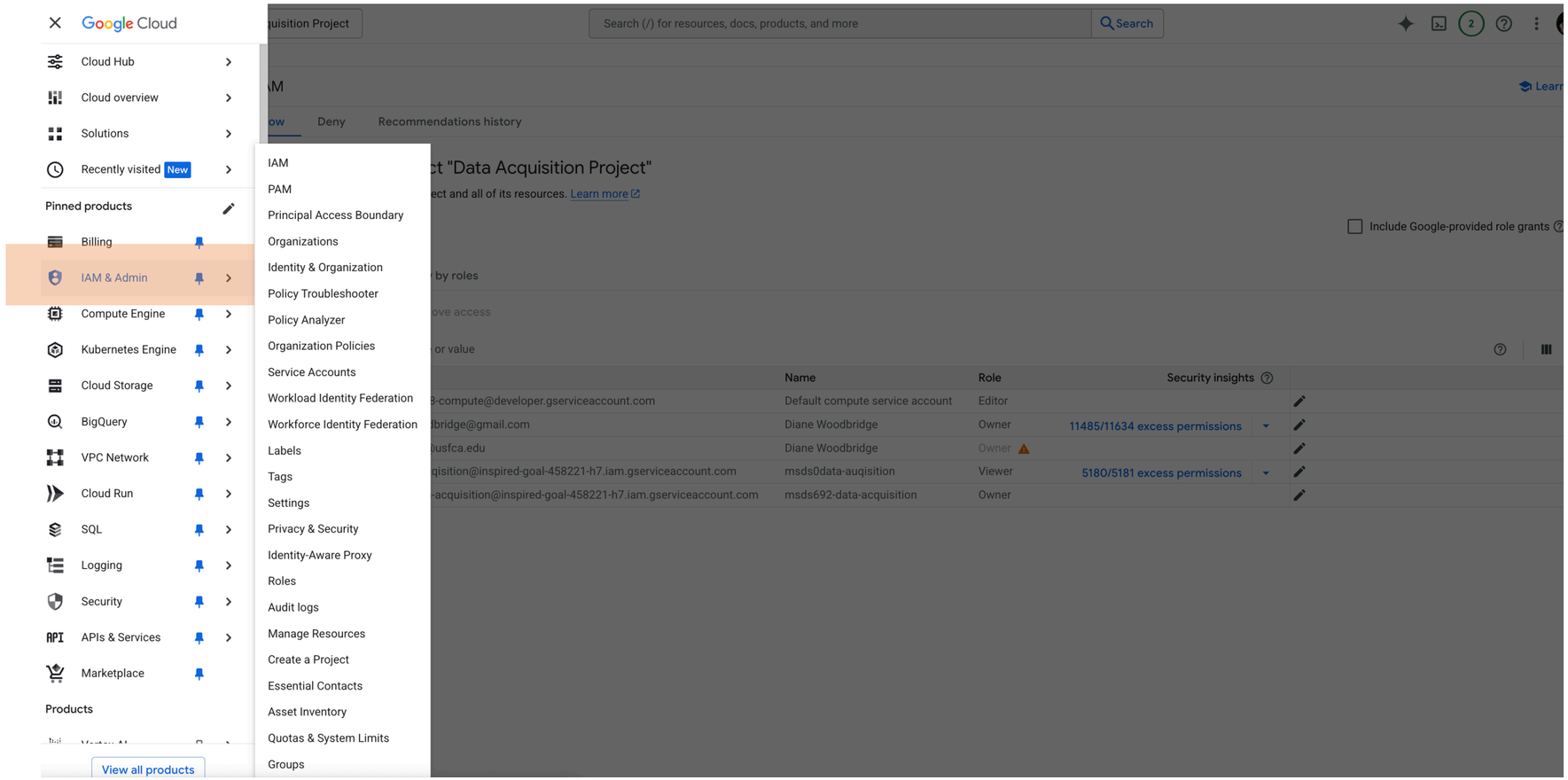
I have 1) created a GCP account, 2) created a project and 3) applied the coupon



Google Cloud Computing (GCP)

Services that we will use…

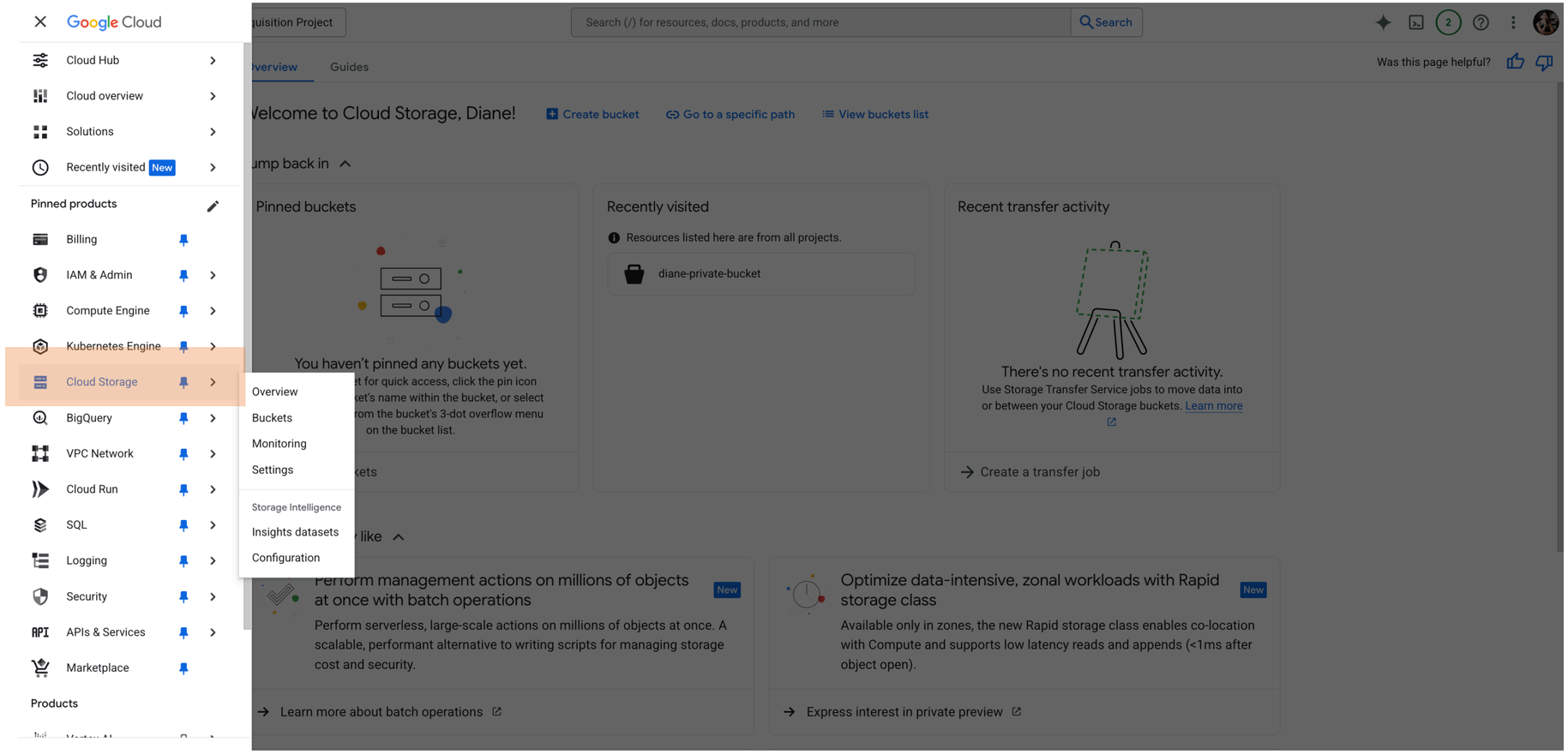
IAM & Admin



Google Cloud Computing (GCP)

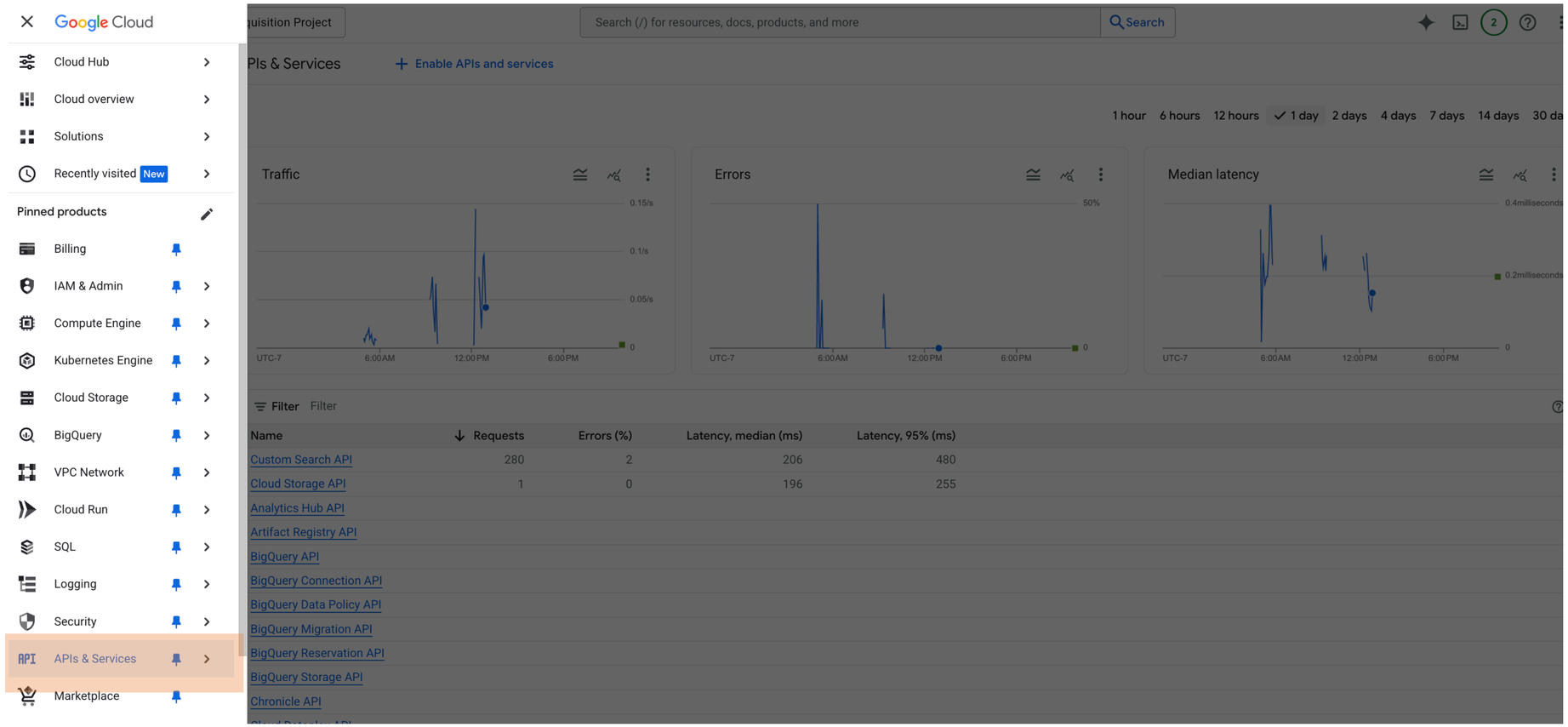
Services that we will use…

Cloud Storage



Google Cloud Computing (GCP)

Services that we will use…

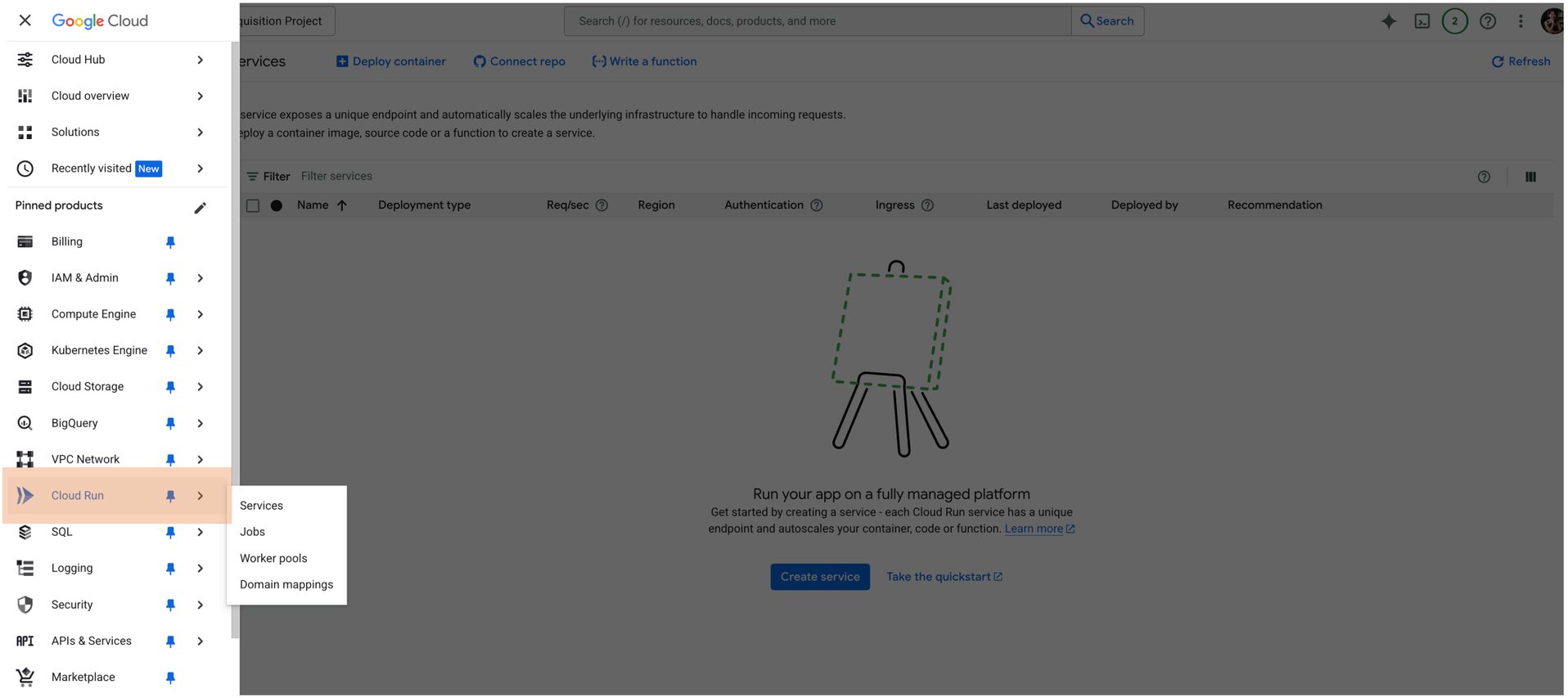


API & Services

Google Cloud Computing (GCP)

Services that we will use…

Cloud Run



GCP vs AWS (Amazon Web Services)

AWS is the oldest and largest cloud computing provider (2006), whereas GCP launched later (2011) but has been growing rapidly.

- They offer similar services, but GCP is more developer friendly and used more for data-driven workloads including databases (BigQuery), and AI/ML.
- GCP is also cheaper, and provides transparent billing.
- GCP is easier to learn and manage resources.

https://cloud.google.com/docs/get-started/aws-azure-gcp-service-comparison

Google Cloud Computing (GCP)

Google Cloud Command Line Interface (gcloud CLI)

- You can automate launching and managing GCP resources via gcloud CLI without logging in GCP console.
  - You can write a script and/or automate commands to manage and operate your cloud resources.
  - Installation Guide
    - https://cloud.google.com/sdk/docs/install-sdk

Contents

Your feedback is important and will help shape the rest of the course. What aspects of the course have been most helpful? Do you have suggestions for changes that could make the course more engaging or effective for you?

Nobody has responded yet.

Hang tight! Responses are coming in.

References

Managing environments, Conda, https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html

Google Cloud Documentation, Google Cloud, https://cloud.google.com/docs