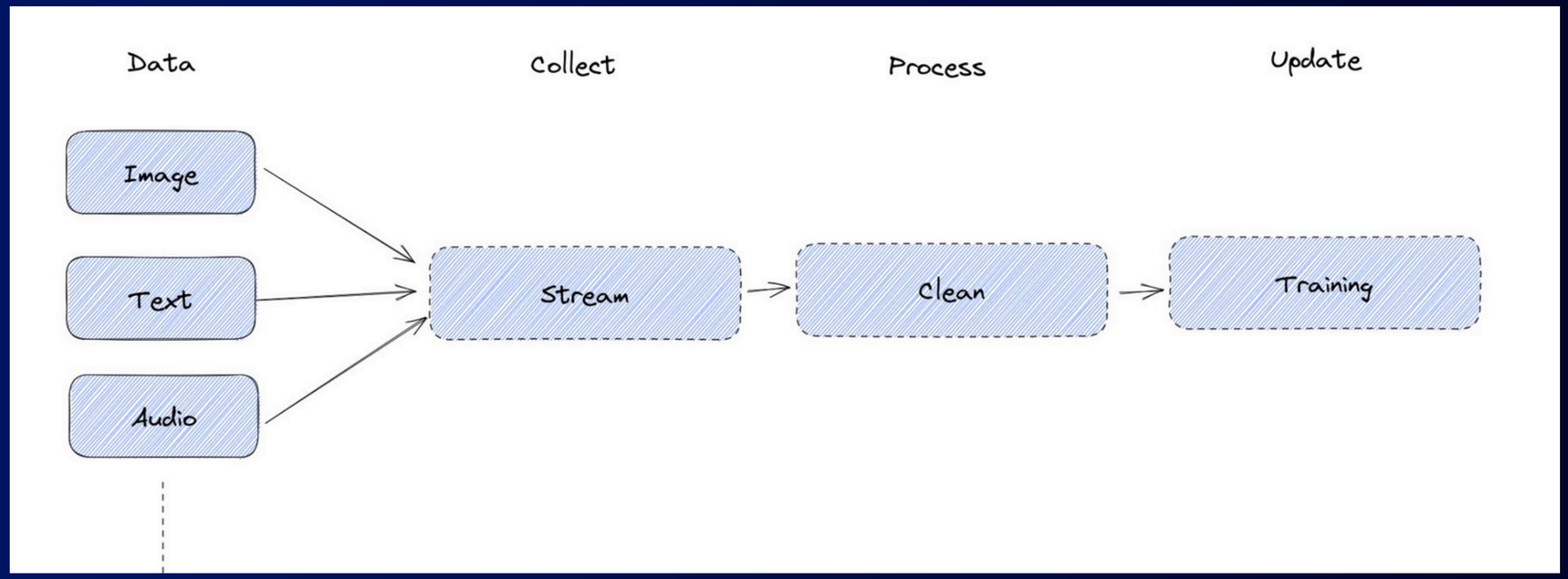
MACHINE LEARNING OPERATIONS

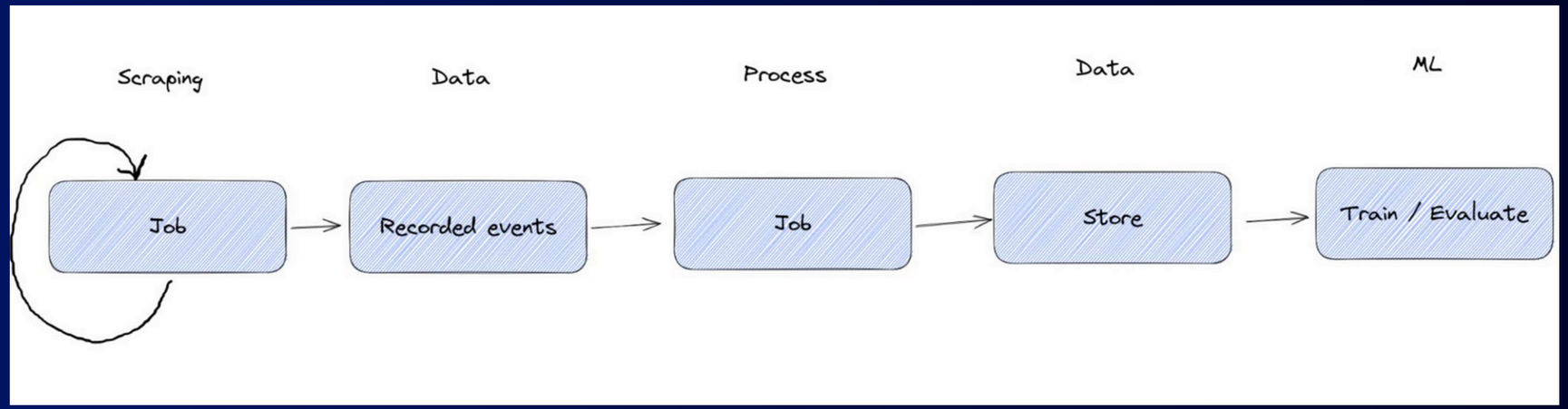
WEEK 6

UNLIMITED

Presented by **Asst. Prof. Dr. Tuchsanai Ploysuwan**

Data Version Control : DVC


Data
Collect
Process
Update
Image
Text
Audio
Stream
Clean
Training

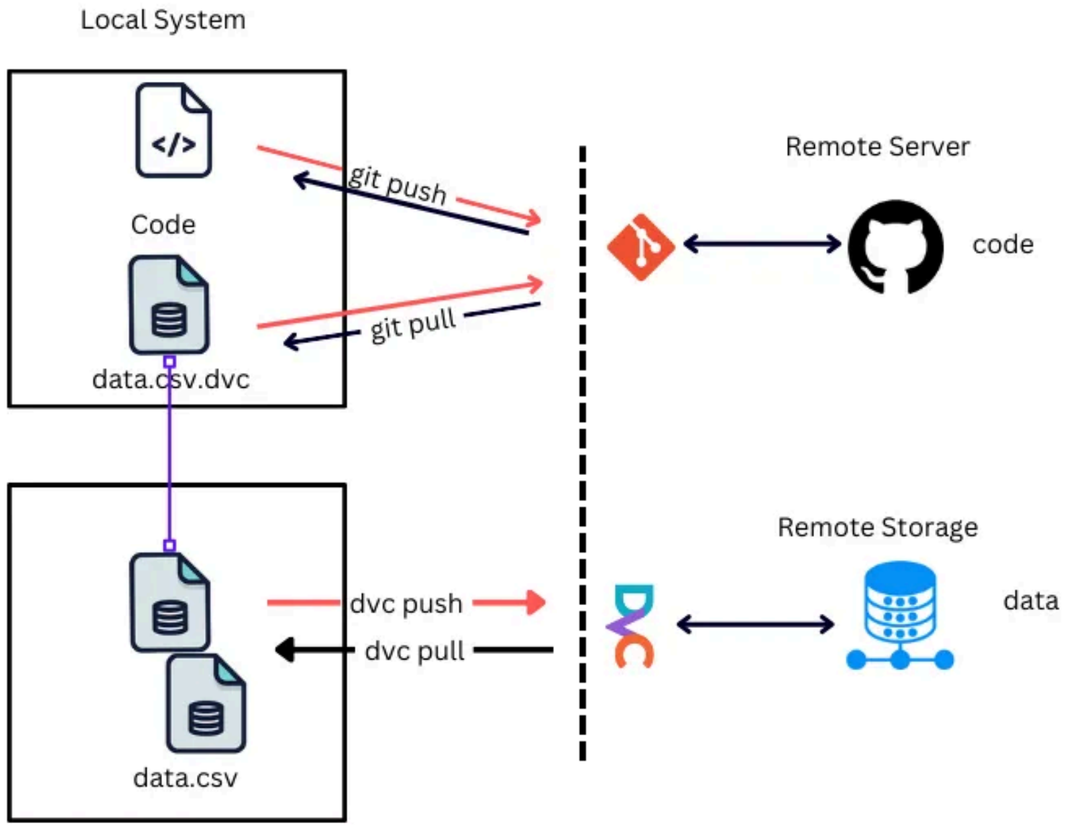

Scraping
Data
Process
Data
ML
Job
Recorded events
Job
Store
Train / Evaluate

DVC

Gitcannot be used for datasets versioning, specially with large datasets. **For dataset versioning** we have **DVC**, which integrates with Git. It is an open-source command-line tool that mimics Git flows and commands.

DVC Characteristics:

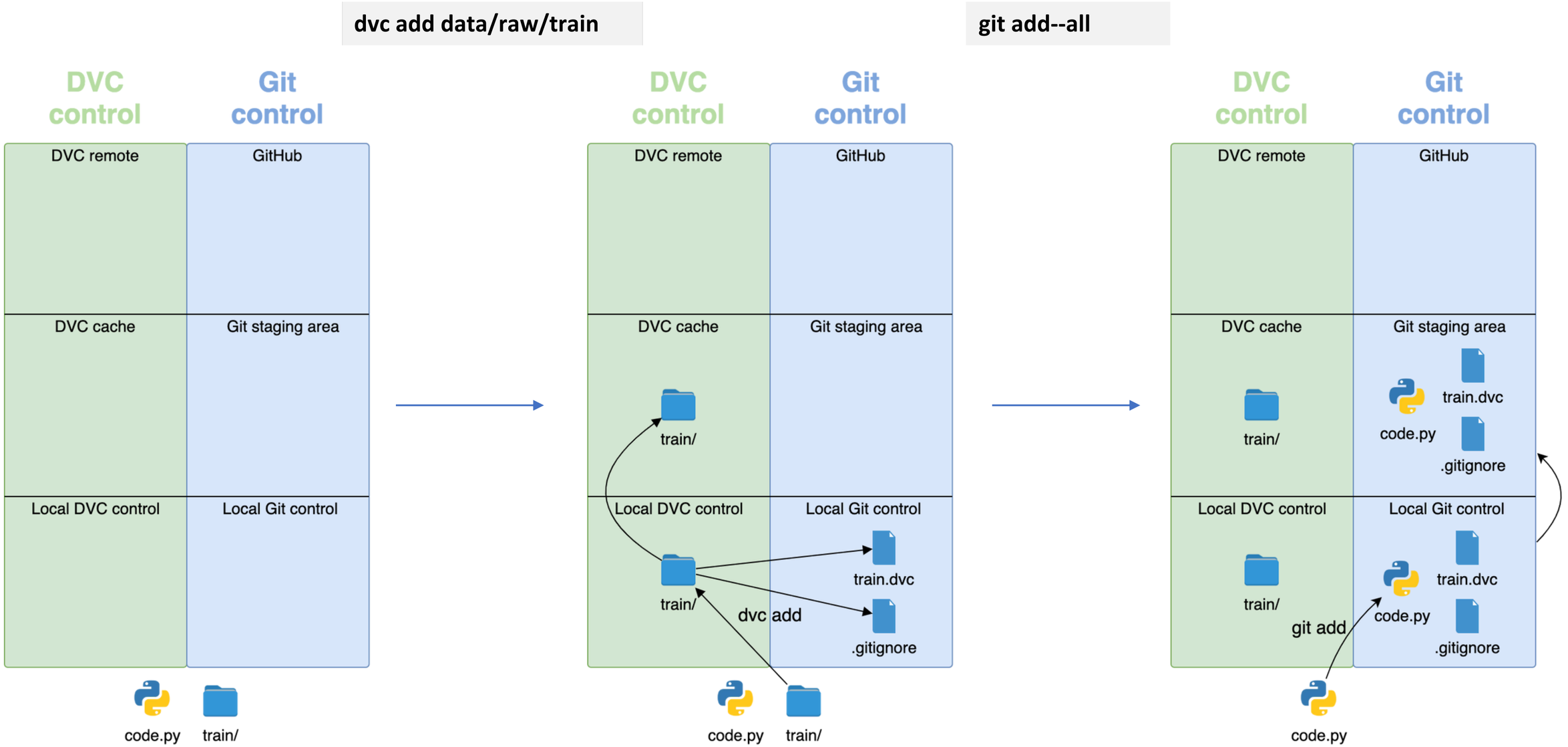
- **Git-compliant**
- Easy data **version control**
- **Storage**independent
- Reproducible
- Language and **framework independent**
- Low friction **branching**
- Easy to use



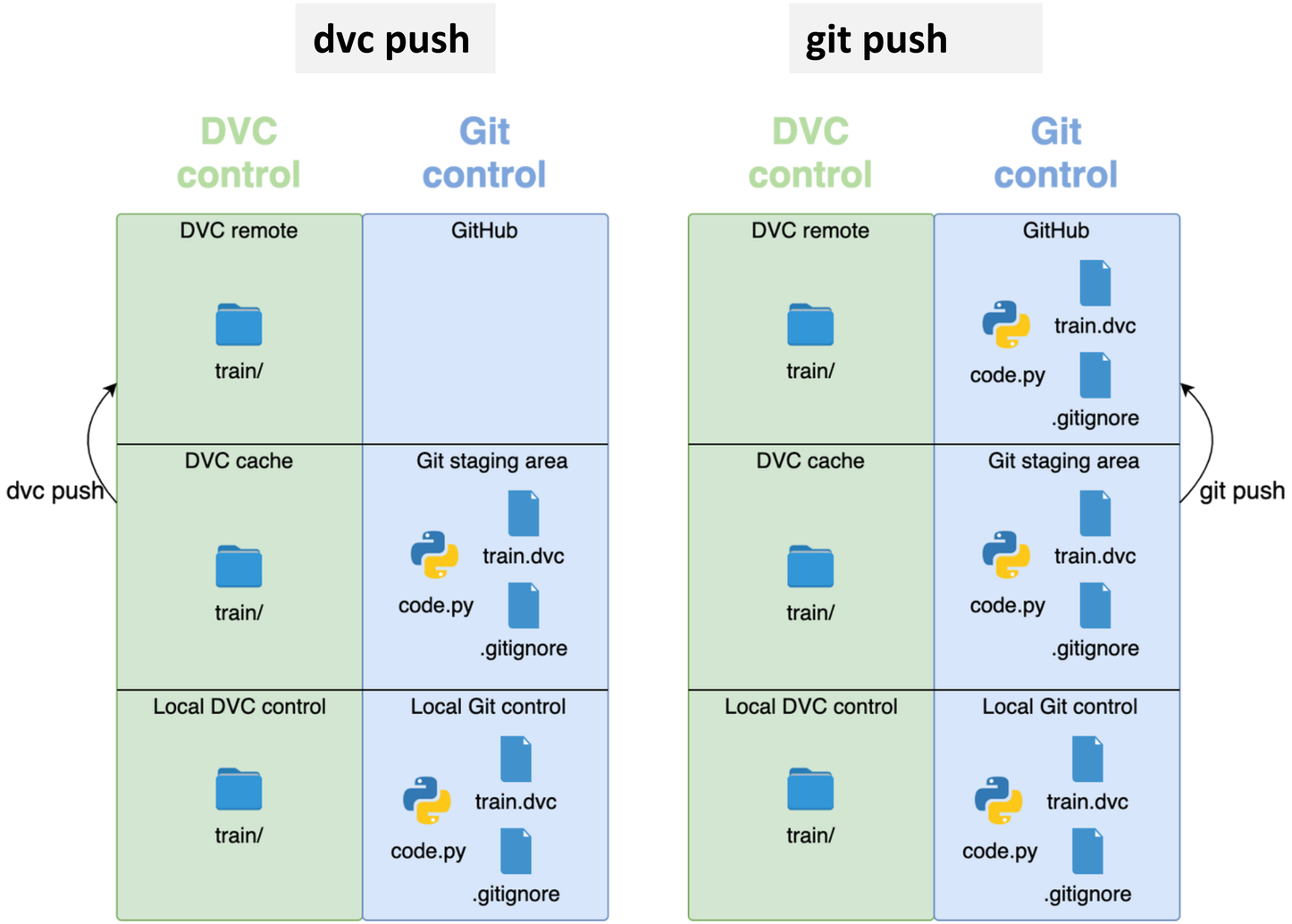
Data versioning is crucial because it allows for tracking changes to data over time, ensuring that users have access to accurate and reliable information. By maintaining a history of data revisions, data versioning enables reproducibility, collaboration, and auditing in various fields such as software development, scientific research, and financial analysis. It helps prevent errors, facilitates troubleshooting, and promotes transparency, ultimately improving decision-making processes and reducing the risk of costly mistakes.

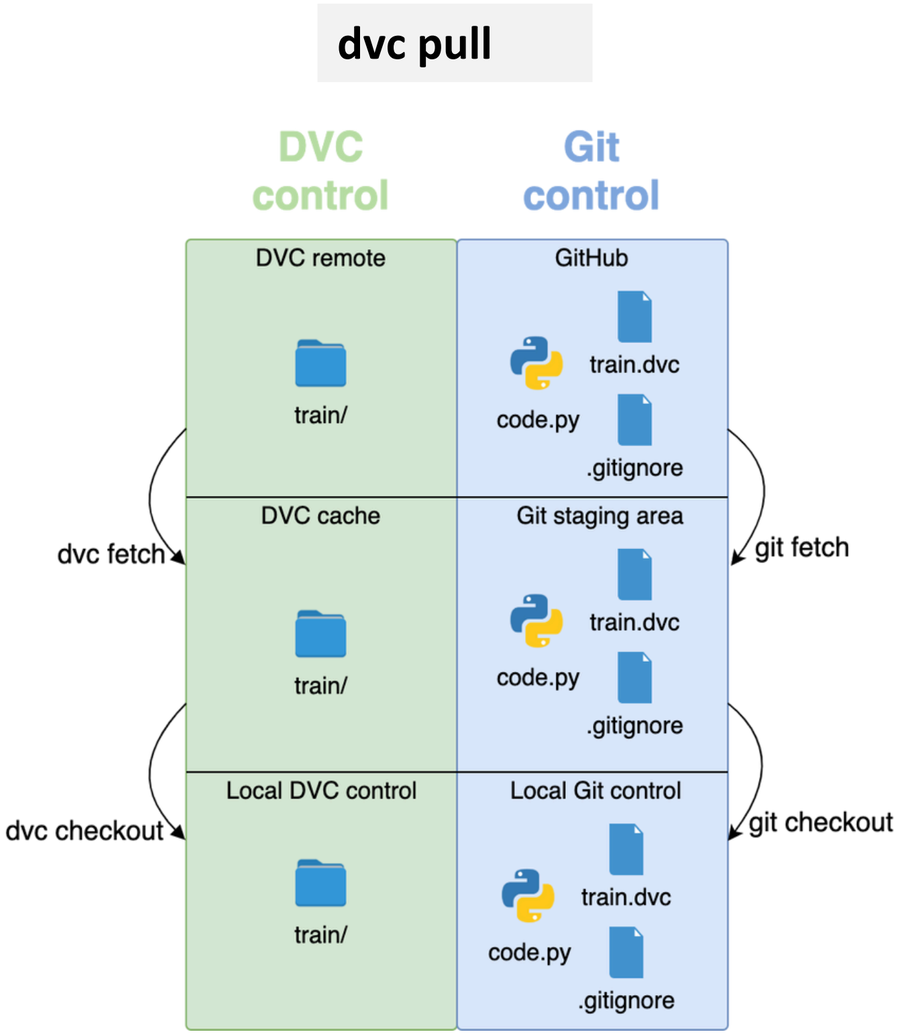
File tracking withDVC

Largedata files and folders go to **DVC remote storage**, but **small .dvc**files go to **GitHub**



Upload and download files with DVC





DVC Commands

DVC main commands to upload data to a remote environment.

```
$ dvc init # initialize the repo
$ dvc add . # add the files that have been changed
$ dvc commit -m "making some changes" # commit the updates with a message
$ dvc remote add newremote s3://bucket/path # point the repo to an S3
bucket for storage
$ dvc push # push the changes to the DVC repo hosted in the default S3
bucket
$ dvc pull # pull the latest changes from the DVC repo hosted in the
default S3 bucket
```

DVC Files

DVC files are **YAML**files. Information is stored in **key-value**pairs and lists. The first key is **md5**followed by a string of characters. MD5 is a hash function. Two files that are exactly the same, will produce the same hash.

YAML

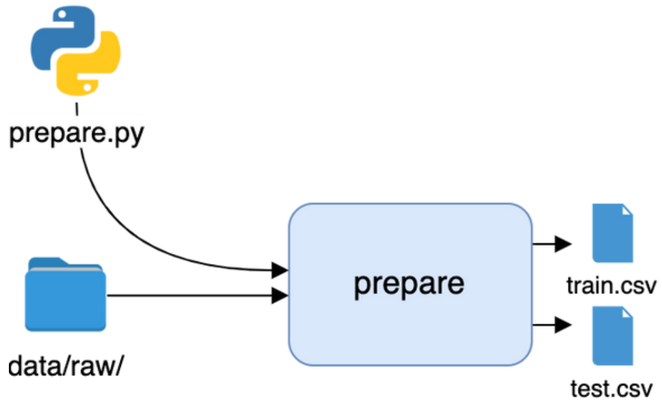
```yaml
md5: 62bdac455a6574ed68a1744da1505745
outs:
  - md5: 96652bd680f9b8bd7c223488ac97f151
    path: model.joblib
    cache: true
    metric: false
    persist: false
```

DVC pipelines

DVCallows to **chain the files** of the **entire process** in a single execution called the **DVC pipeline** that requires a single command to be executed: **dvcrepro**.

A pipeline consists of multiple stages, and each stage has three components:

- Dependencies
- Outputs
- Command



Shell

```
dvc run -n prepare \
      -d src/prepare.py -d data/raw \
      -o data/prepared/train.csv -o data/prepared/test.csv \
      python src/prepare.py
```

DVC Commands

DVC will create two files:

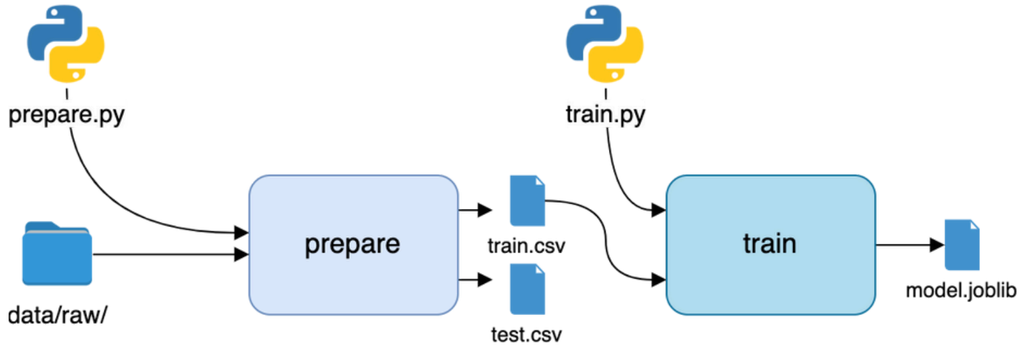
- **dvc.yaml**
- **dvc.lock**

The internal information of both files is similar, with the addition of MD5 hashes in dvc.lock

YAML

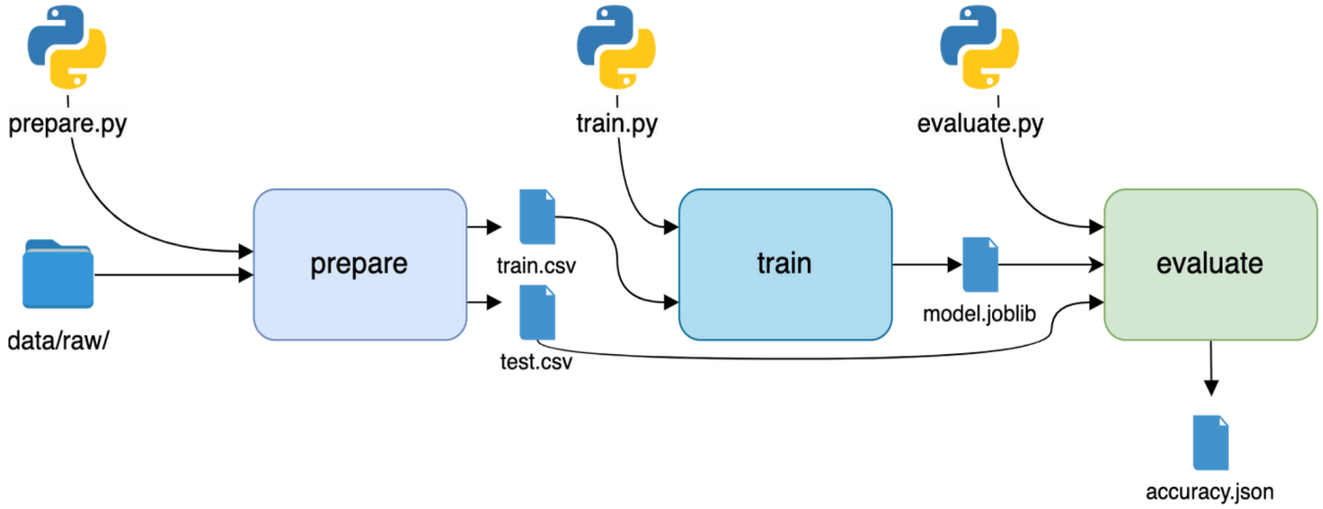
```yaml
prepare:
  cmd: python src/prepare.py
  deps:
    - path: data/raw
      md5: a8a5252d9b14ab2c1be283822a86981a.dir
    - path: src/prepare.py
      md5: 0e29f075d51efc6d280851d66f8943fe
  outs:
    - path: data/prepared/test.csv
      md5: d4a8cdf527c2c58d8cc4464c48f2b5c5
    - path: data/prepared/train.csv
      md5: 50cbdb38dbf0121a6314c4ad9ff786fe
```

ExamplesofDVC pipelines



Shell

```
$ dvc run -n train \
        -d src/train.py -d data/prepared/train.csv \
        -o model/model.joblib \
        python src/train.py
```

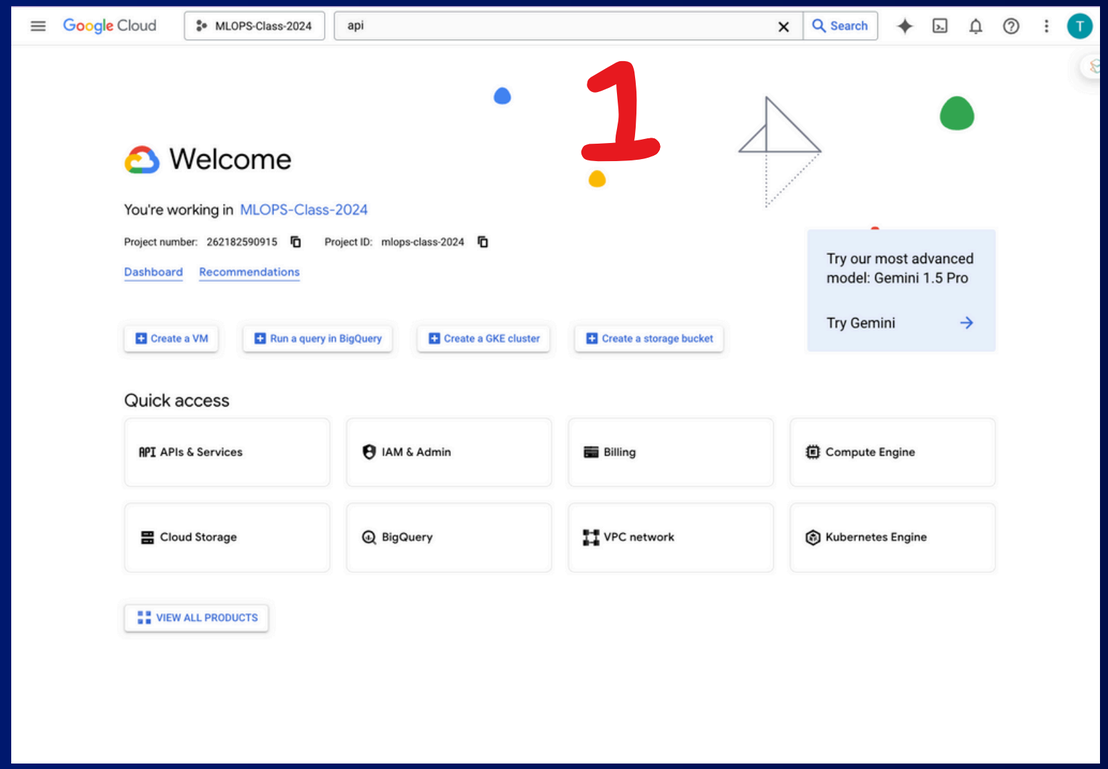


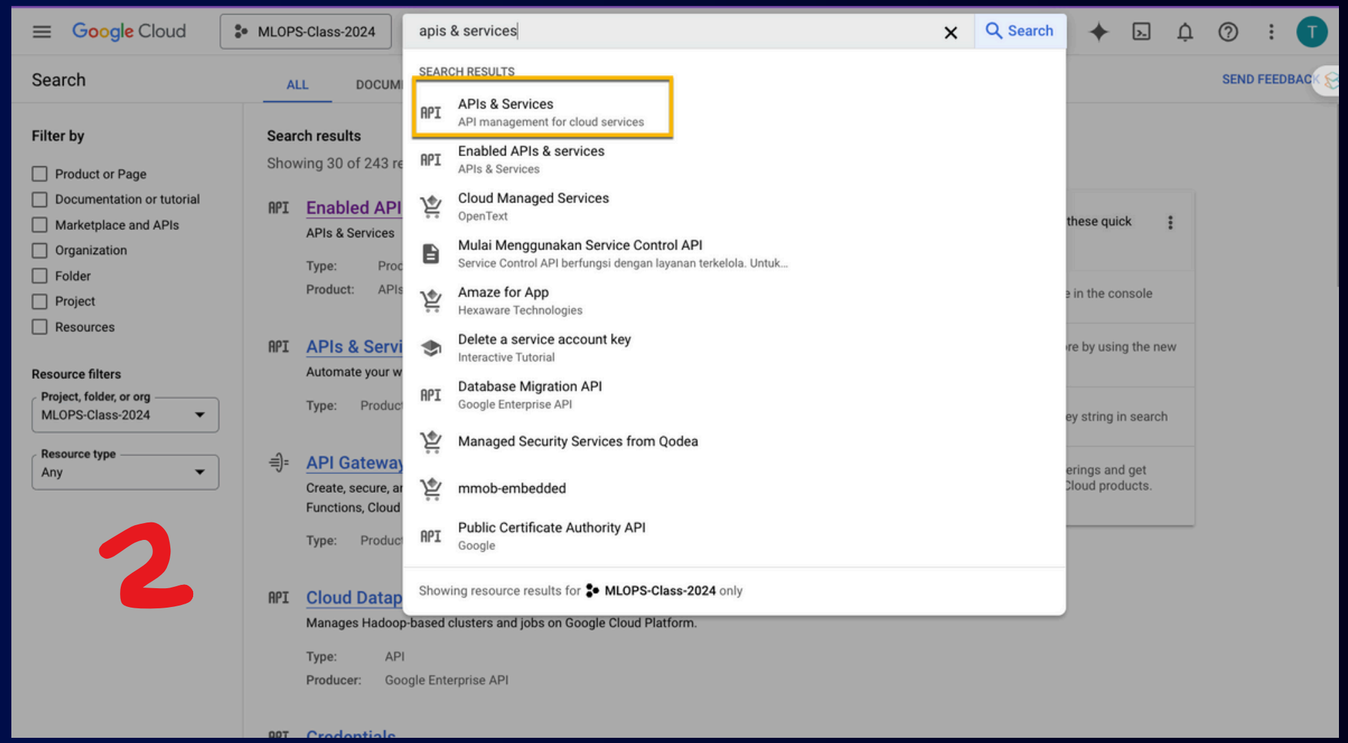
Shell

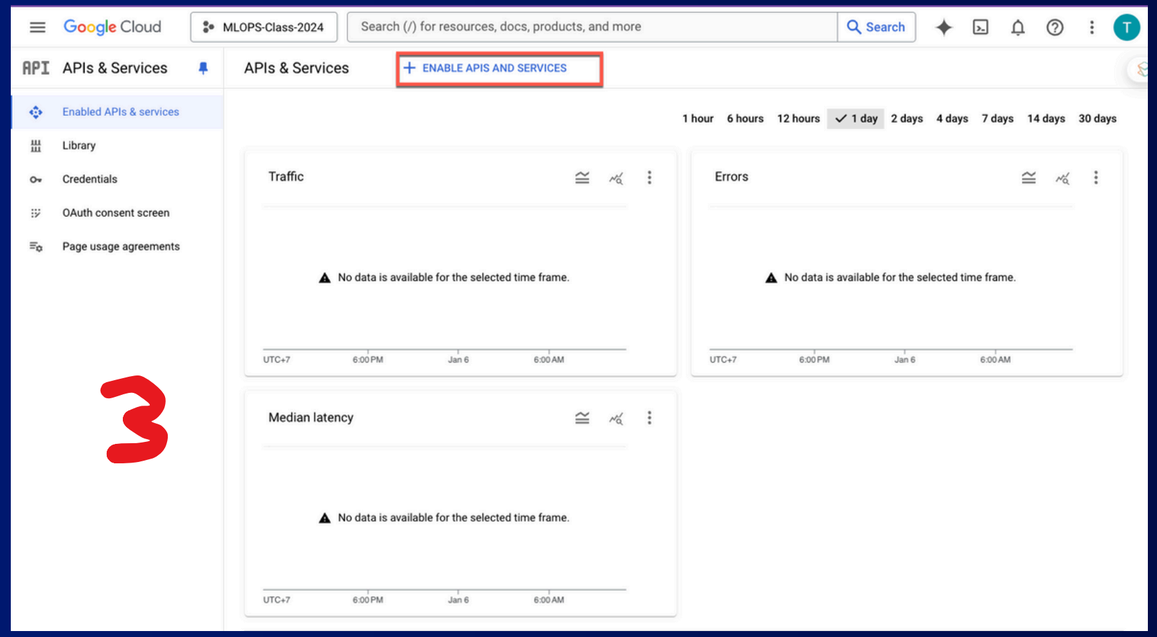
```
$ dvc run -n evaluate \
        -d src/evaluate.py -d model/model.joblib \
        -M metrics/accuracy.json \
        python src/evaluate.py
```

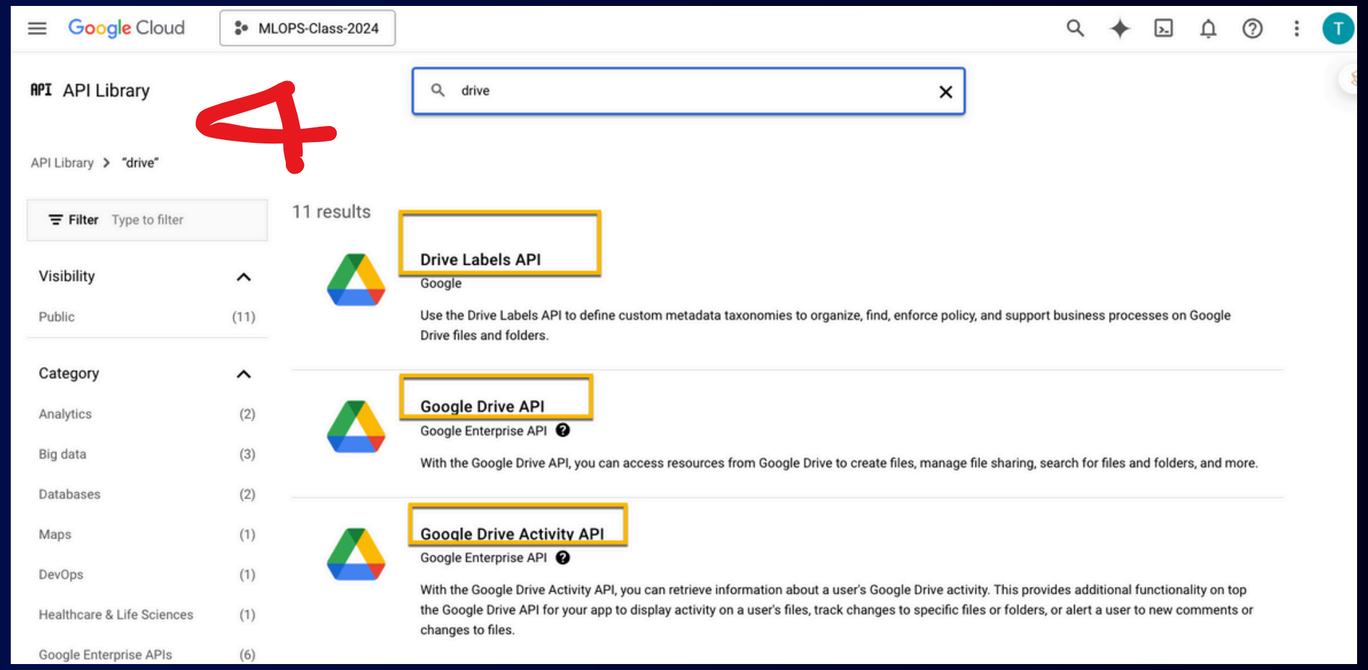
LAB 1. Setting Up a Workflow with DVC Google Cloud Storage(GCS)

Go to Google cloud console –> Click APIs & Services –> Click Enable APIs and Services

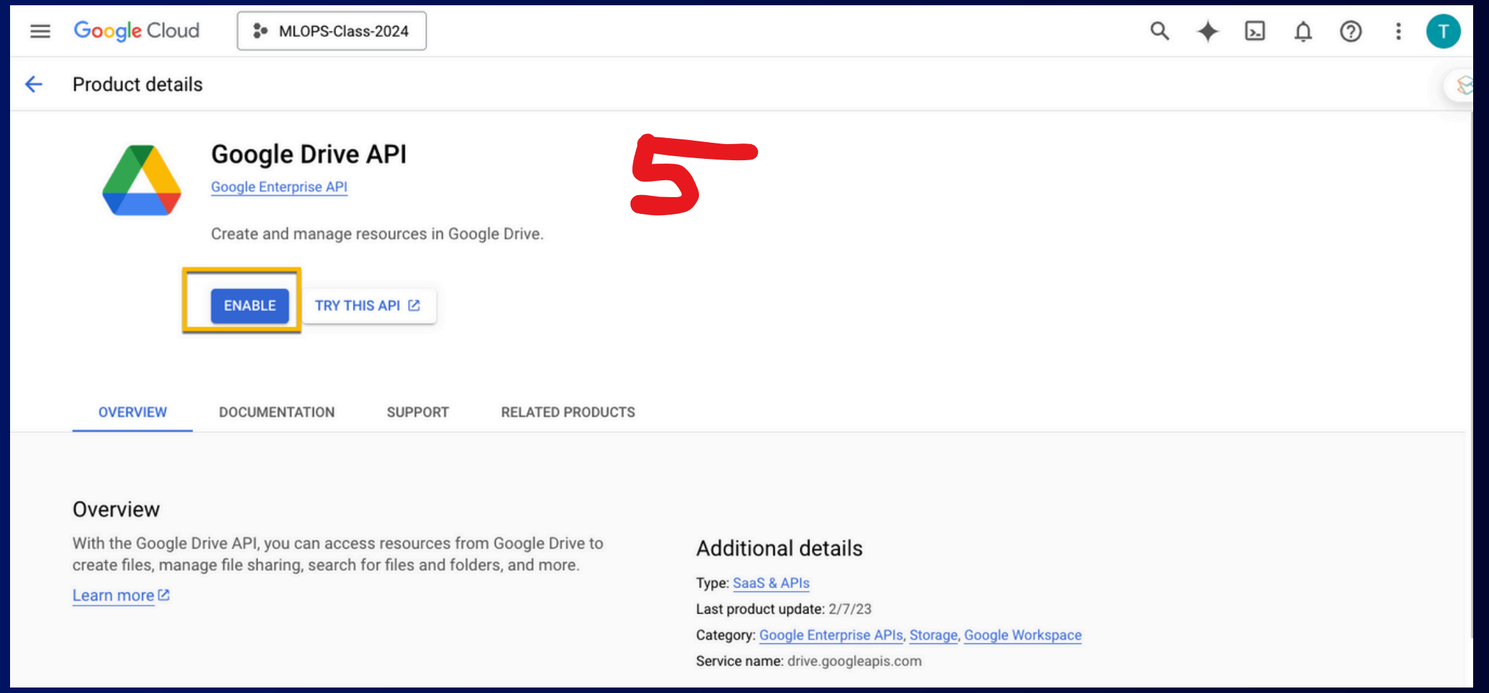


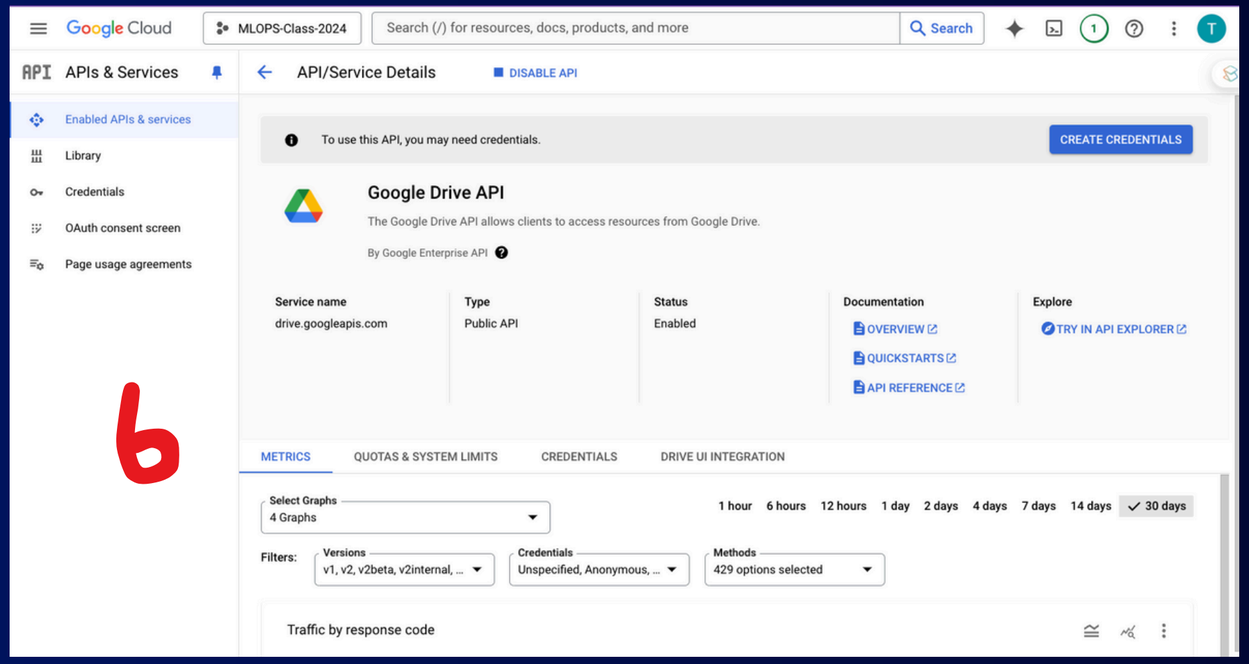


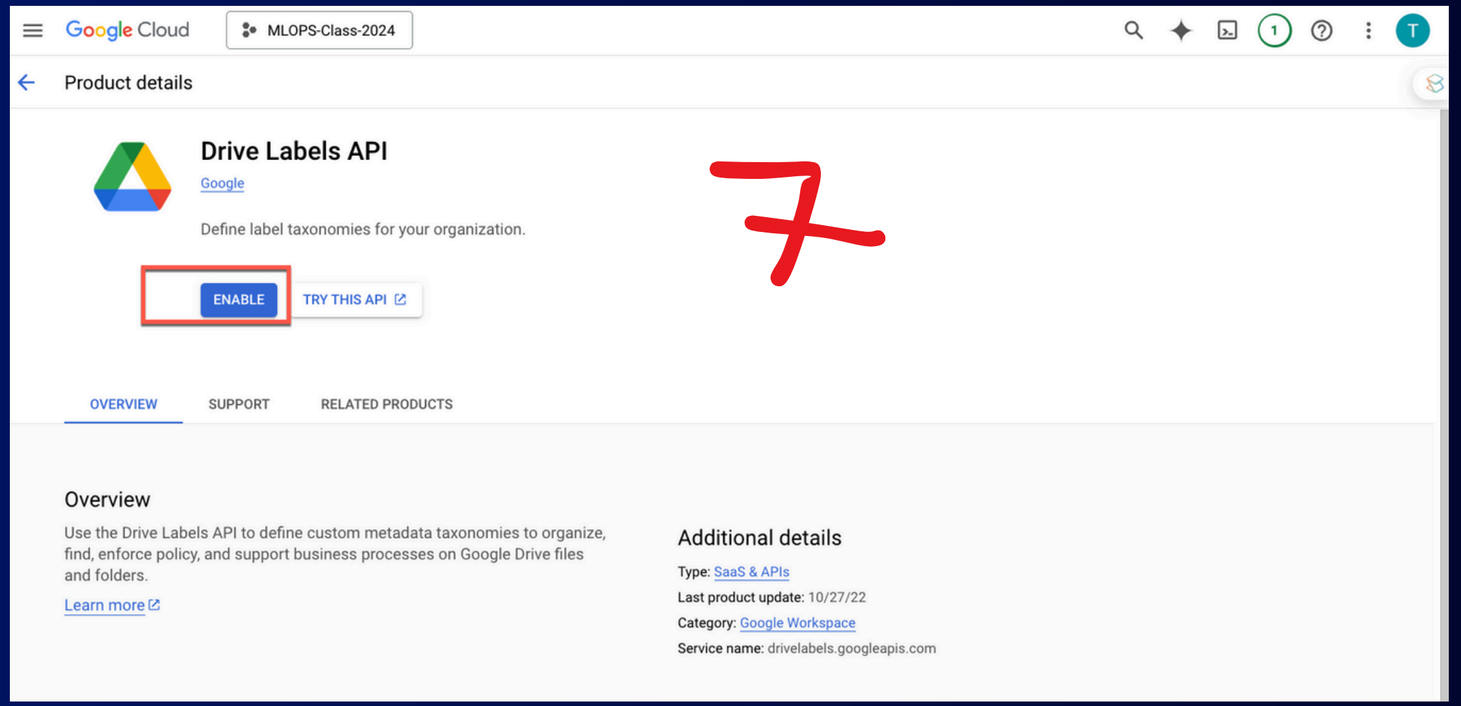


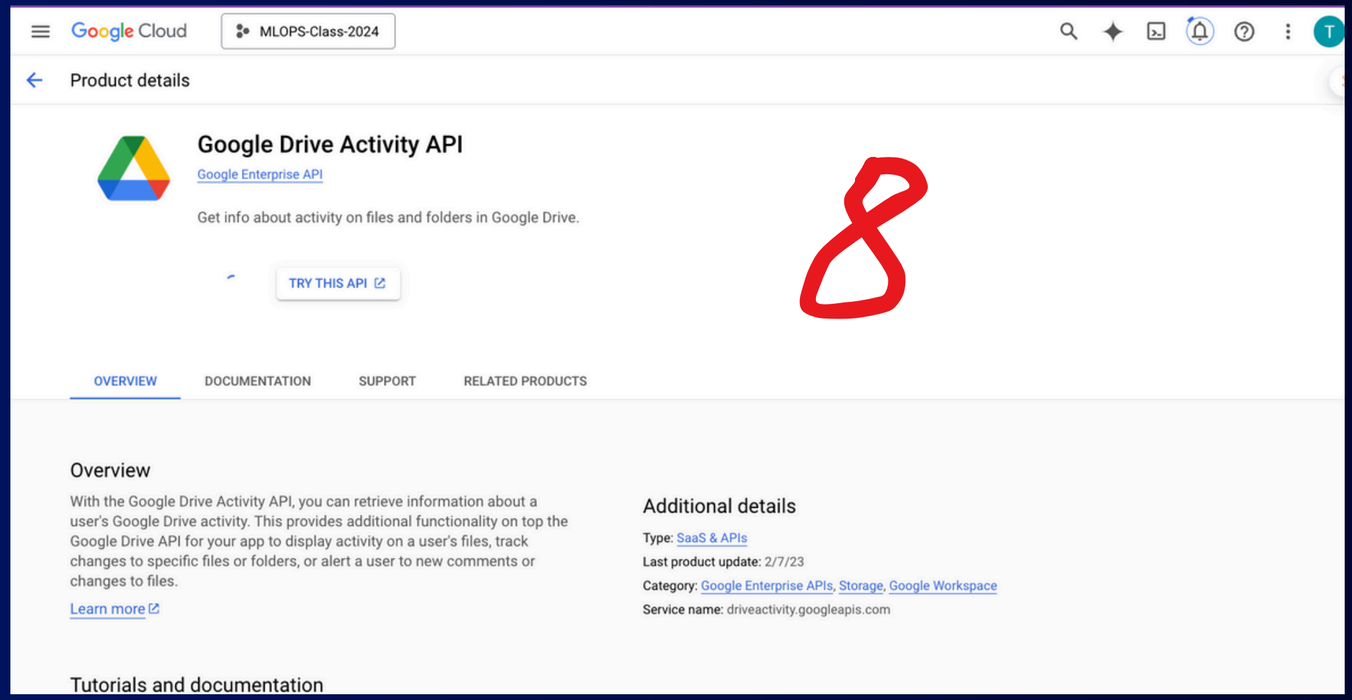


In this method, we can store data in a Google Cloud Storage (GCS) bucket and fetch it using service account authentication.

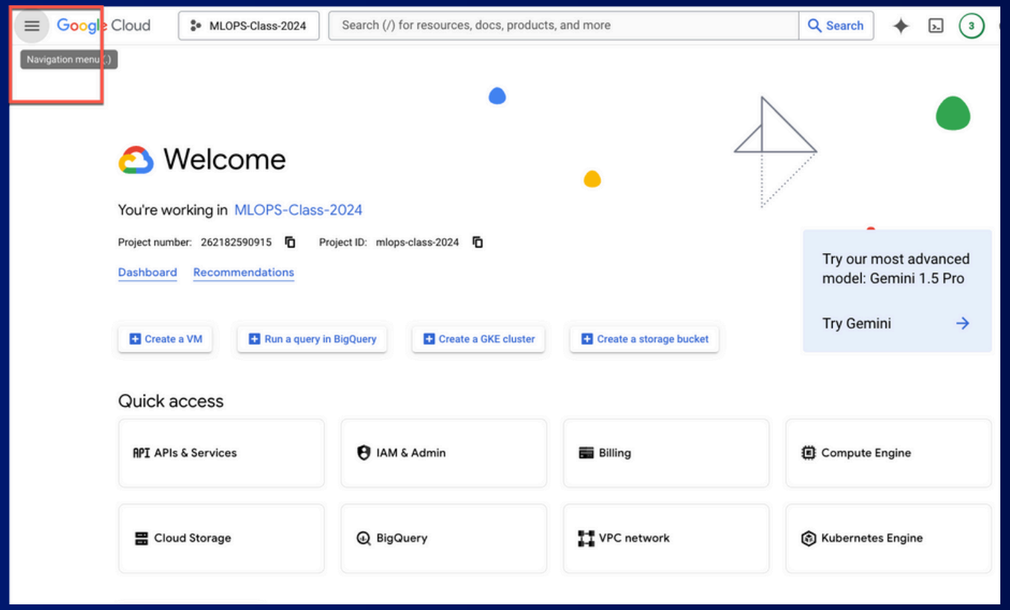

Google Cloud
MLOPS-Class-2024
Product details
Google Drive API
Google Enterprise API
Create and manage resources in Google Drive.
5
ENABLE
TRY THIS API
OVERVIEW
DOCUMENTATION
SUPPORT
RELATED PRODUCTS
Overview
With the Google Drive API, you can access resources from Google Drive to create files, manage file sharing, search for files and folders, and more.
Learn more
Additional details
Type: SaaS & APIs
Last product update: 2/7/23
Category: Google Enterprise APIs, Storage, Google Workspace
Service name: drive.googleapis.com


Google Cloud
MLOPS-Class-2024
Search (/) for resources, docs, products, and more
Search
APIs & Services
Enabled APIs & services
Library
Credentials
OAuth consent screen
Page usage agreements
6
API/Service Details
DISABLE API
To use this API, you may need credentials.
CREATE CREDENTIALS
Google Drive API
The Google Drive API allows clients to access resources from Google Drive.
By Google Enterprise API
Service name
drive.googleapis.com
Type
Public API
Status
Enabled
Documentation
OVERVIEW
QUICKSTARTS
API REFERENCE
Explore
TRY IN API EXPLORER
METRICS
QUOTAS & SYSTEM LIMITS
CREDENTIALS
DRIVE UI INTEGRATION
Select Graphs
4 Graphs
1 hour
6 hours
12 hours
1 day
2 days
4 days
7 days
14 days
30 days
Filters:
Versions
v1, v2, v2beta, v2internal, ...
Credentials
Unspecified, Anonymous, ...
Methods
429 options selected
Traffic by response code

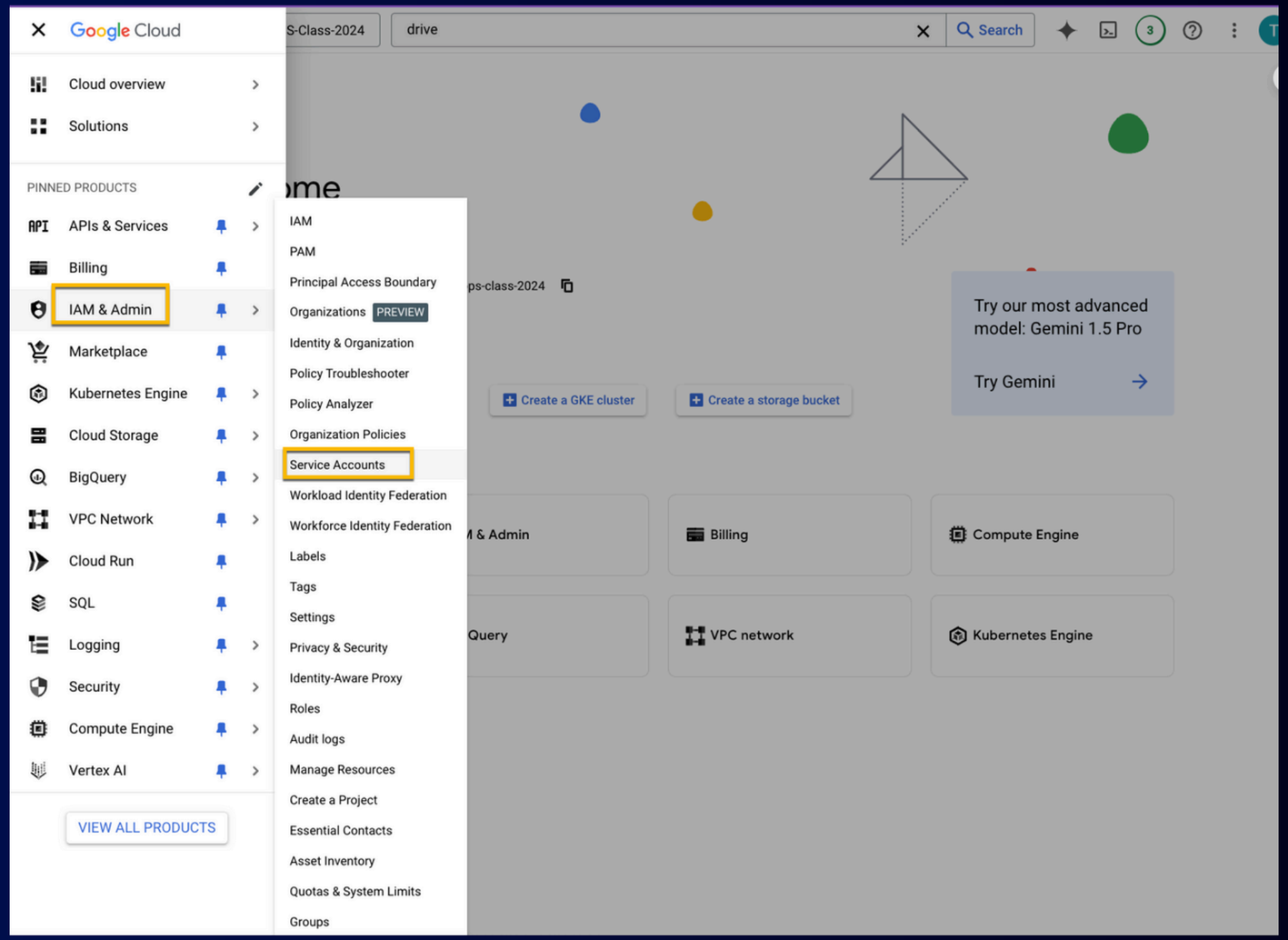

Google Cloud
MLOPS-Class-2024
Product details
Drive Labels API
Google
Define label taxonomies for your organization.
ENABLE
TRY THIS API
7
OVERVIEW
SUPPORT
RELATED PRODUCTS
Overview
Use the Drive Labels API to define custom metadata taxonomies to organize, find, enforce policy, and support business processes on Google Drive files and folders.
Learn more
Additional details
Type: SaaS & APIs
Last product update: 10/27/22
Category: Google Workspace
Service name: drivelabels.googleapis.com

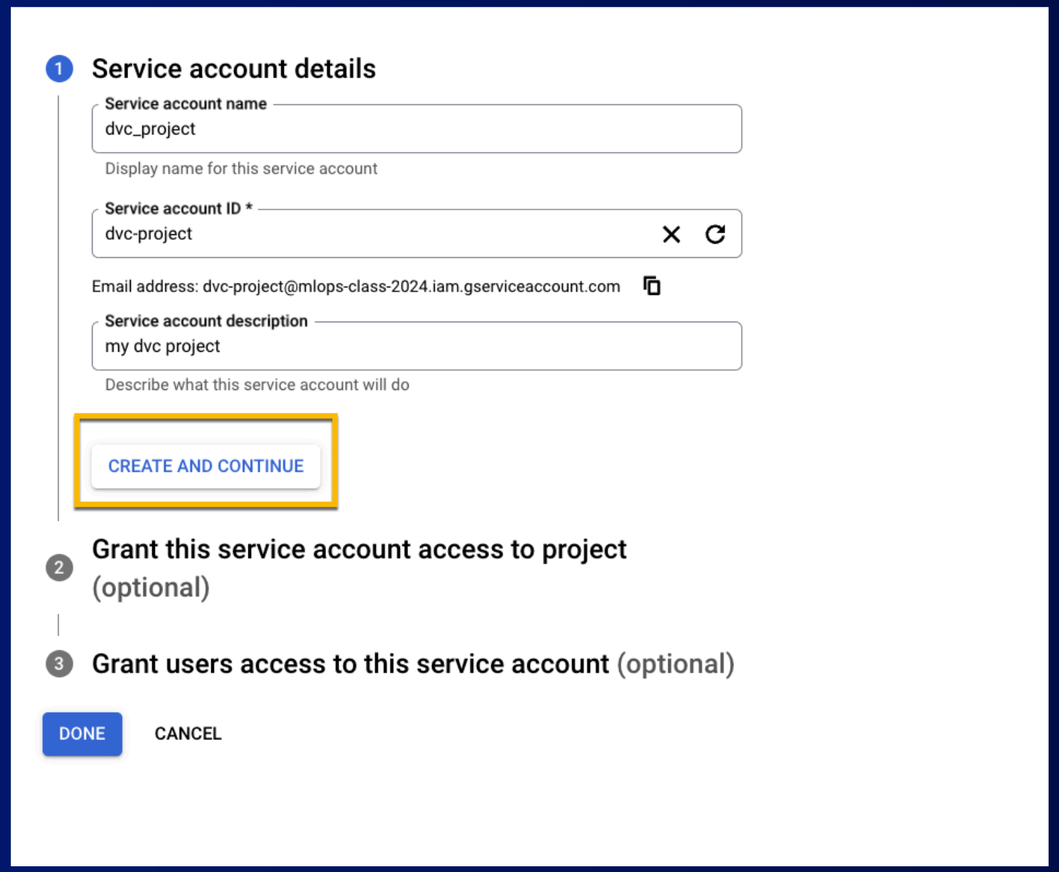

Google Cloud
MLOPS-Class-2024
Product details
Google Drive Activity API
Google Enterprise API
Get info about activity on files and folders in Google Drive.
TRY THIS API
8
OVERVIEW
DOCUMENTATION
SUPPORT
RELATED PRODUCTS
Overview
With the Google Drive Activity API, you can retrieve information about a user's Google Drive activity. This provides additional functionality on top the Google Drive API for your app to display activity on a user's files, track changes to specific files or folders, or alert a user to new comments or changes to files.
Learn more
Additional details
Type: SaaS & APIs
Last product update: 2/7/23
Category: Google Enterprise APIs, Storage, Google Workspace
Service name: driveactivity.googleapis.com
Tutorials and documentation

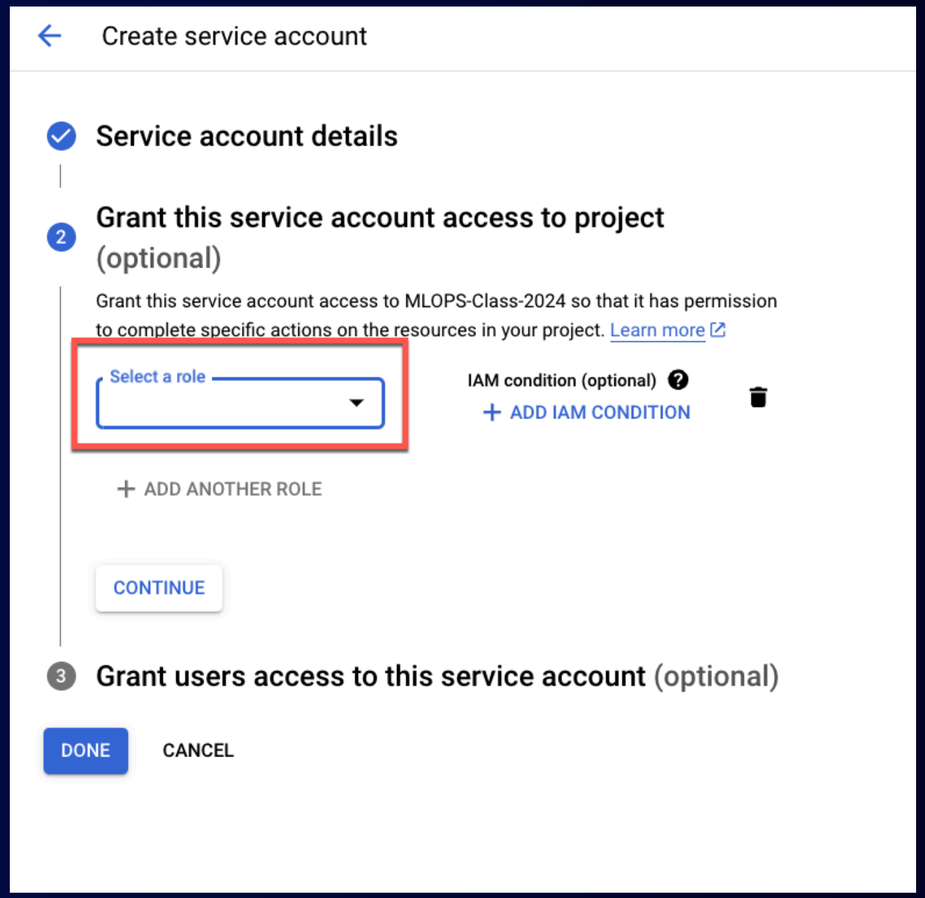
To create a service account, navigate to IAM & Admin in the left sidebar, and select Service Accounts.



10

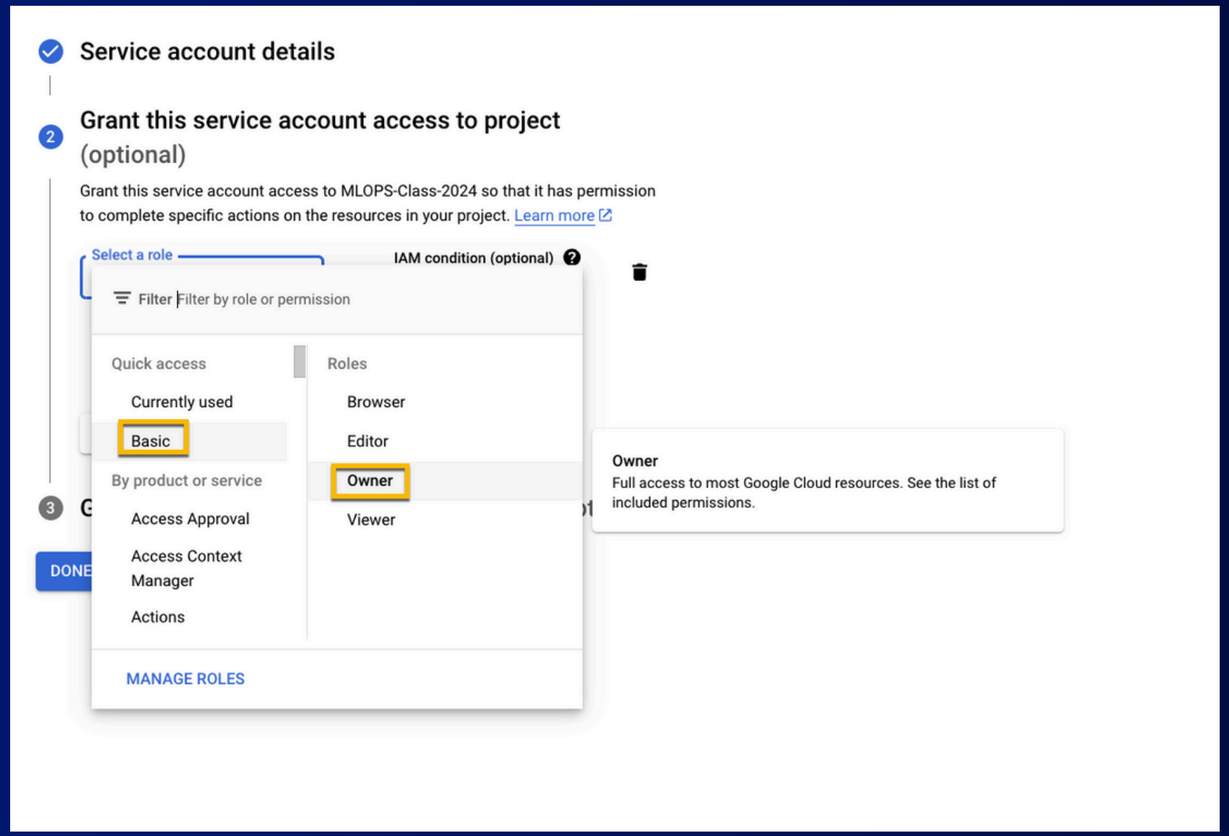



1 Service account details
Service account name
dvc_project
Display name for this service account
Service account ID *
dvc-project
Email address: dvc-project@mlops-class-2024.iam.gserviceaccount.com
Service account description
my dvc project
Describe what this service account will do
CREATE AND CONTINUE
2 Grant this service account access to project (optional)
3 Grant users access to this service account (optional)
DONE
CANCEL

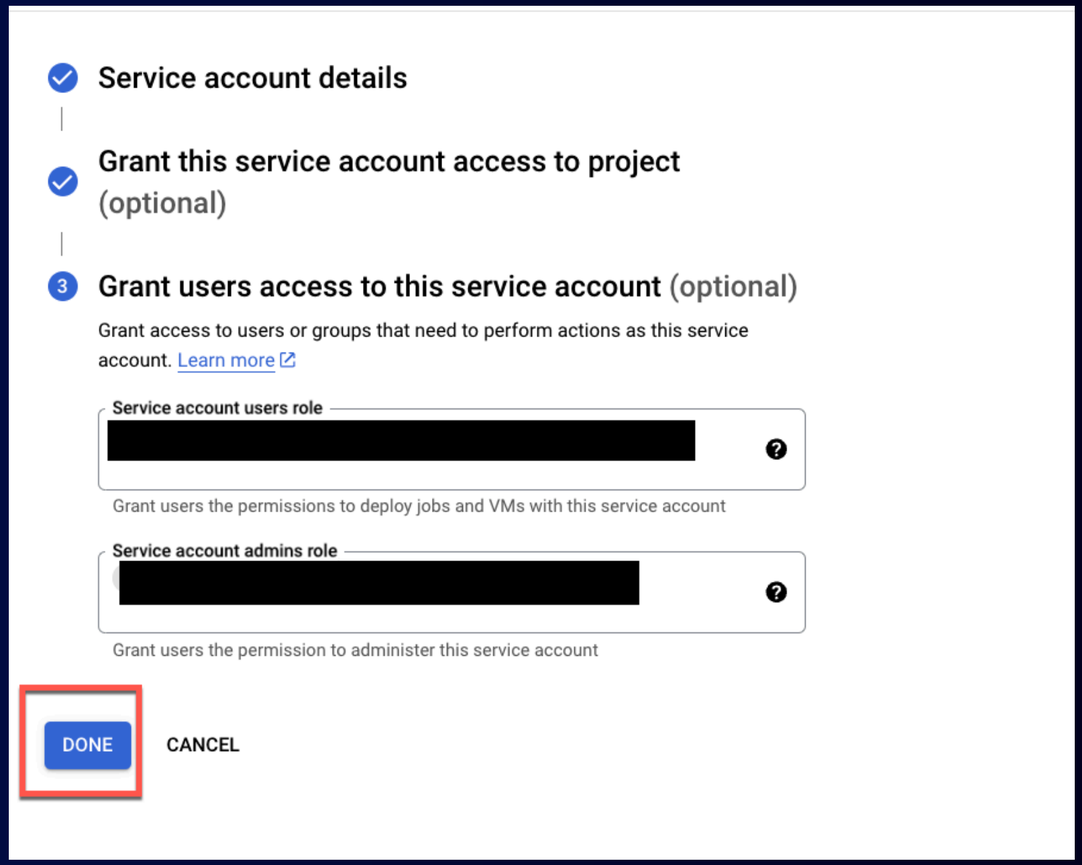

Create service account
Service account details
2 Grant this service account access to project (optional)
Grant this service account access to MLOPS-Class-2024 so that it has permission to complete specific actions on the resources in your project. Learn more
Select a role
IAM condition (optional)
ADD IAM CONDITION
ADD ANOTHER ROLE
CONTINUE
3 Grant users access to this service account (optional)
DONE
CANCEL

12

13

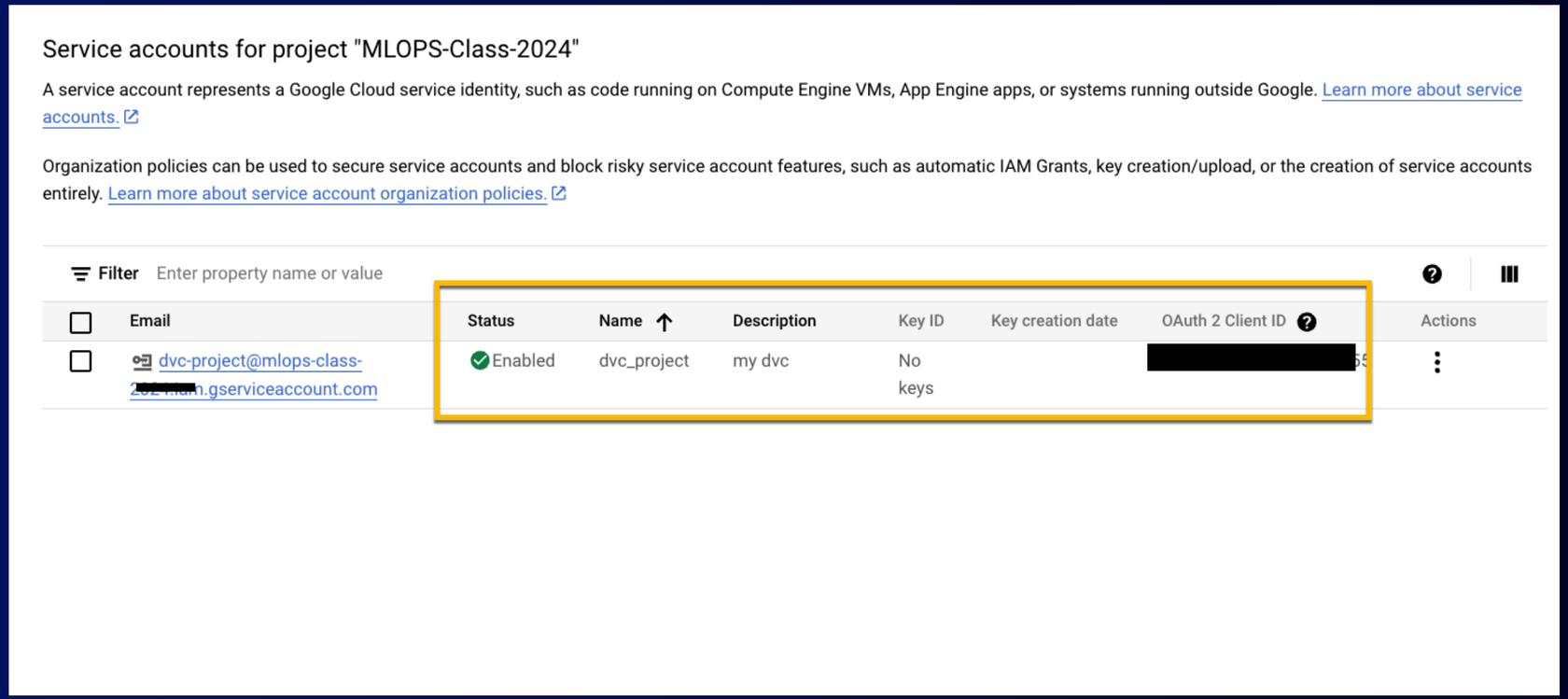


14

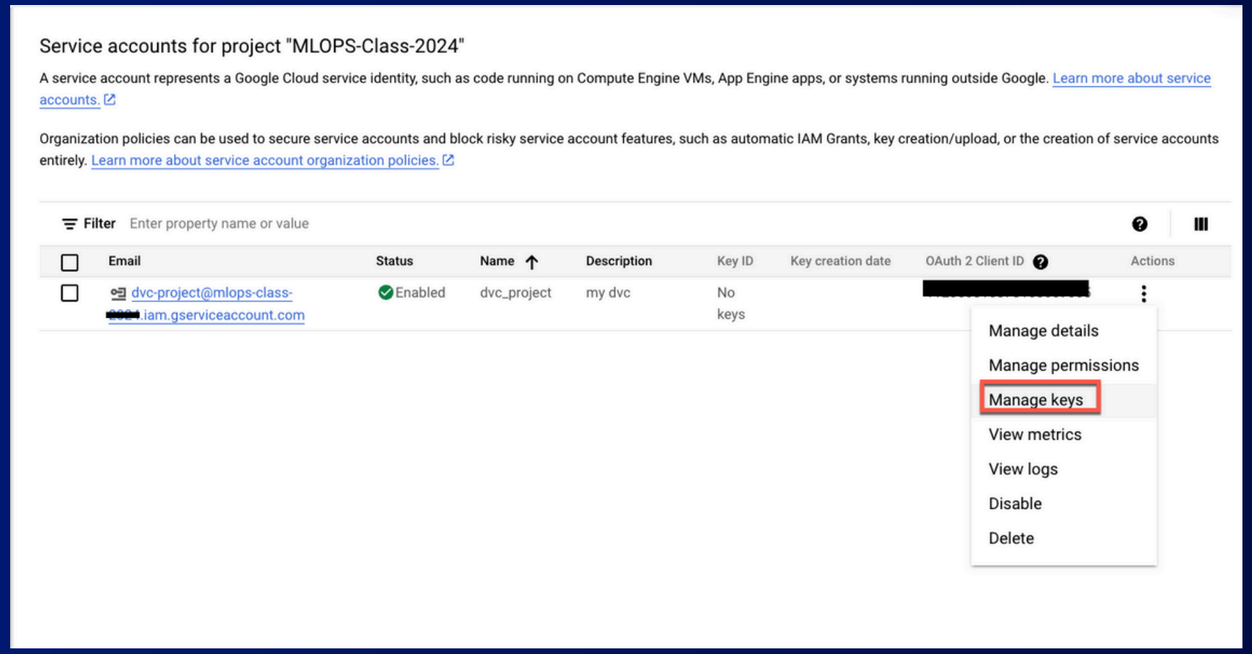


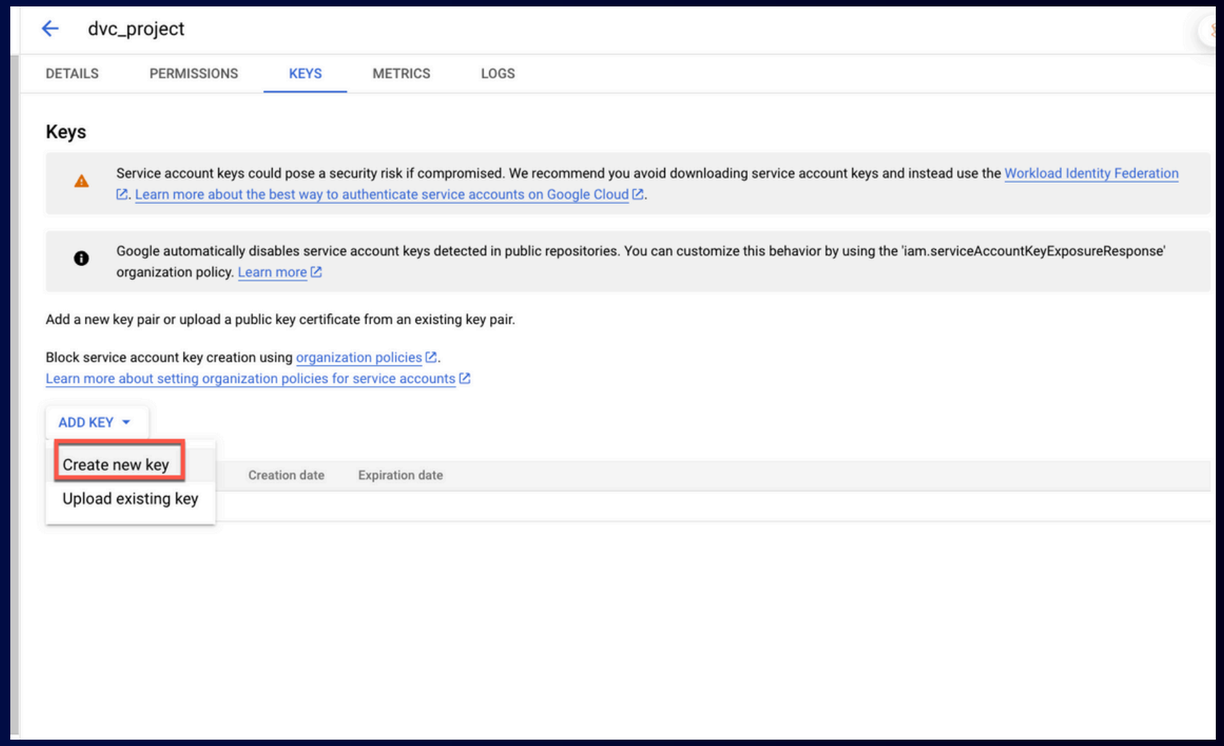
Add user accounts for which you need to grant access.

Now you can see your service account; click on it and go to the Keys tab.



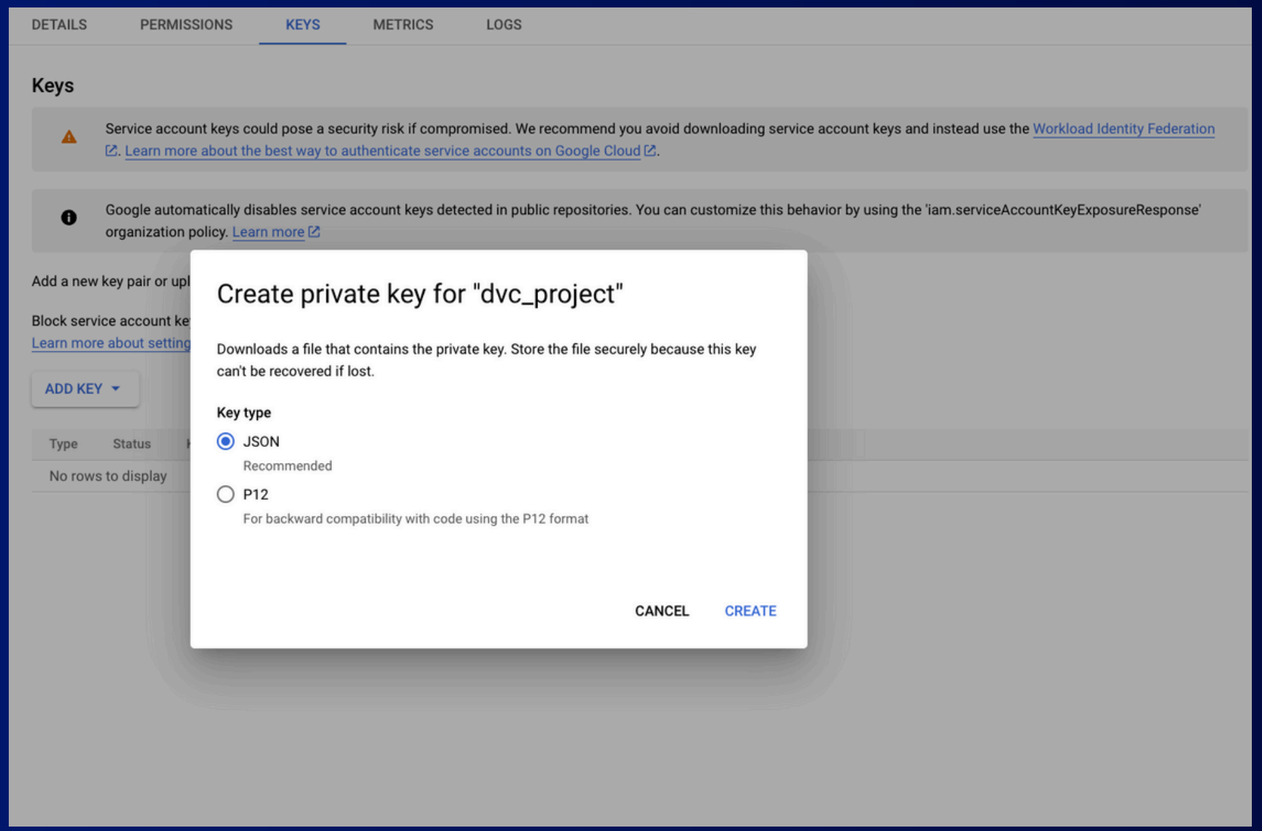
Under Add Key, select Create New Key, choose JSON, and click CREATE.





Download the generated projectname-xxxxxx.json key file to a safe location.

Download the generated projectname-xxxxxx.json key file to a safe location.



Important: Store the API key in a local folder as credentials.json, but do not commit it to GitHub. If you do so, GitHub will raise a warning, and Google will be notified, revoking the credentials.

LAB 2. DVC data application

images data : dvc push

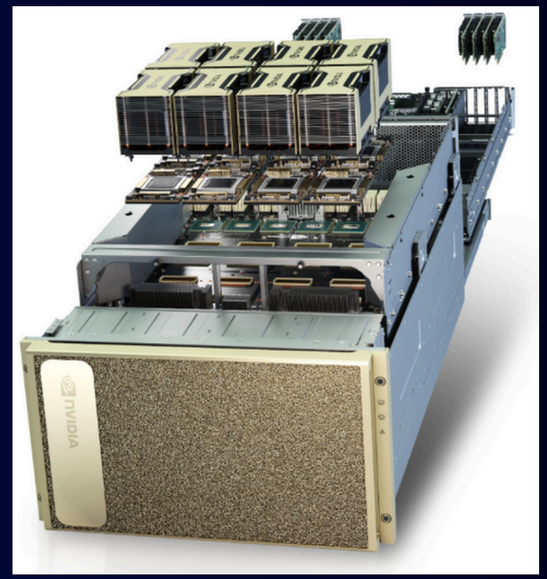
Local Com

code : git push

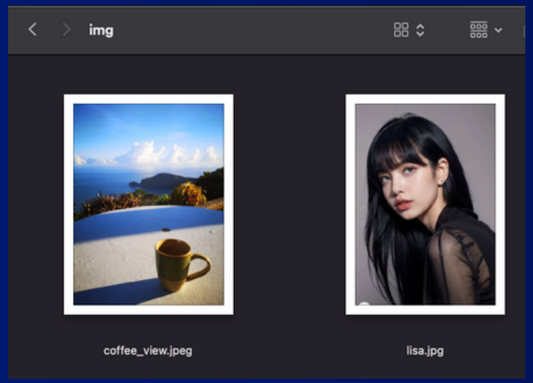
dvc pull

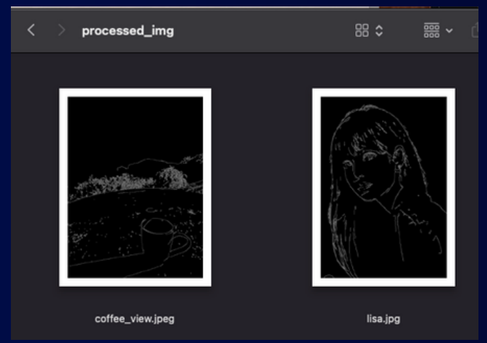
git pull

Remote server



update image data : dvc pull



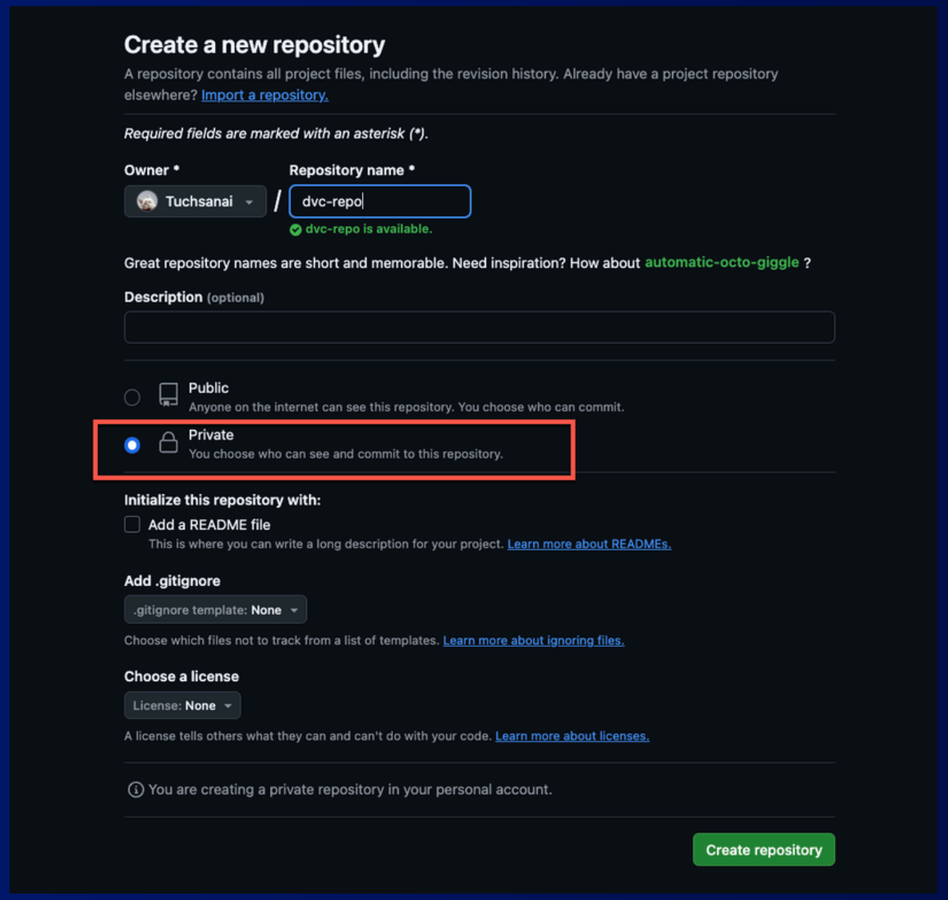


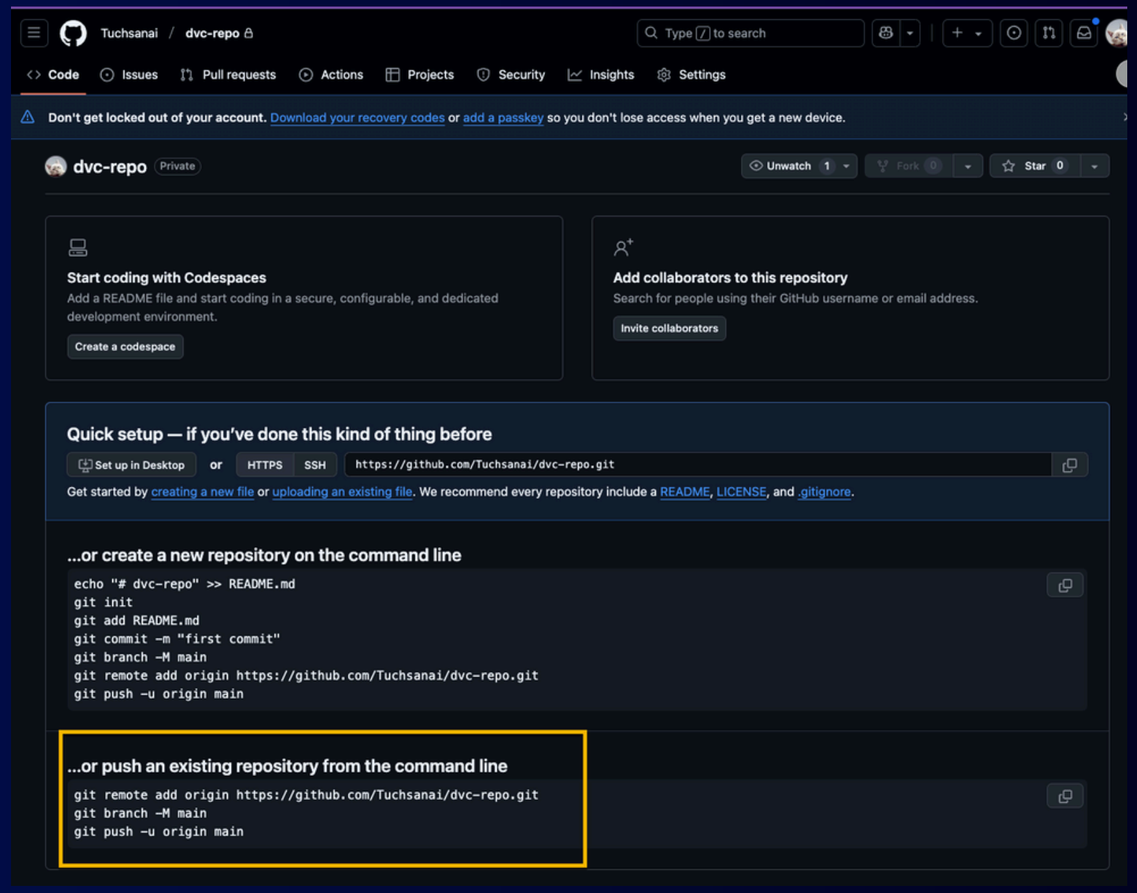

```python
import cv2
import os

input_folder = "./img"
output_folder = "./processed_img"
os.makedirs(output_folder, exist_ok=True)

for filename in os.listdir(input_folder):
    img = cv2.imread(os.path.join(input_folder, filename), cv2.IMREAD_GRAYSCALE)
    edges = cv2.Canny(img, 100, 200)
    cv2.imwrite(os.path.join(output_folder, filename), edges)
```

Create a new repository on a platform like GitHub, GitLab, or Bitbucket





gs://<your-bucket-name>