

# MACHINE LEARNING OPERATIONS

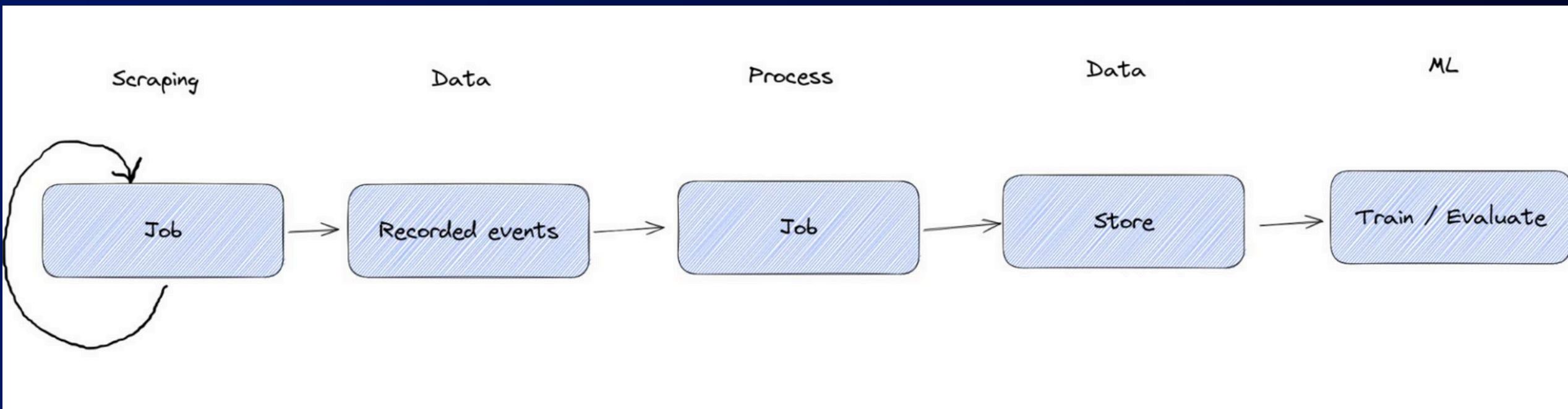
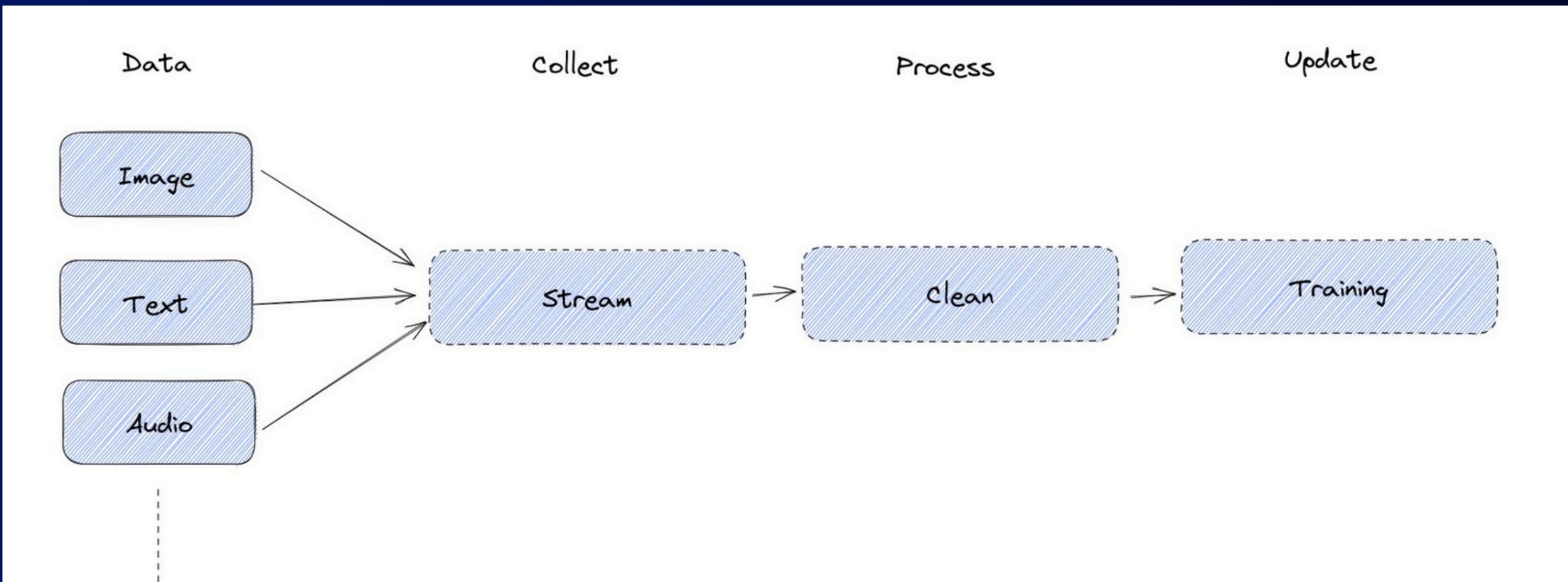


Presented by **Asst. Prof. Dr. Tuchsanai Ploysuwan**

WEEK 6



# Data Version Control : DVC

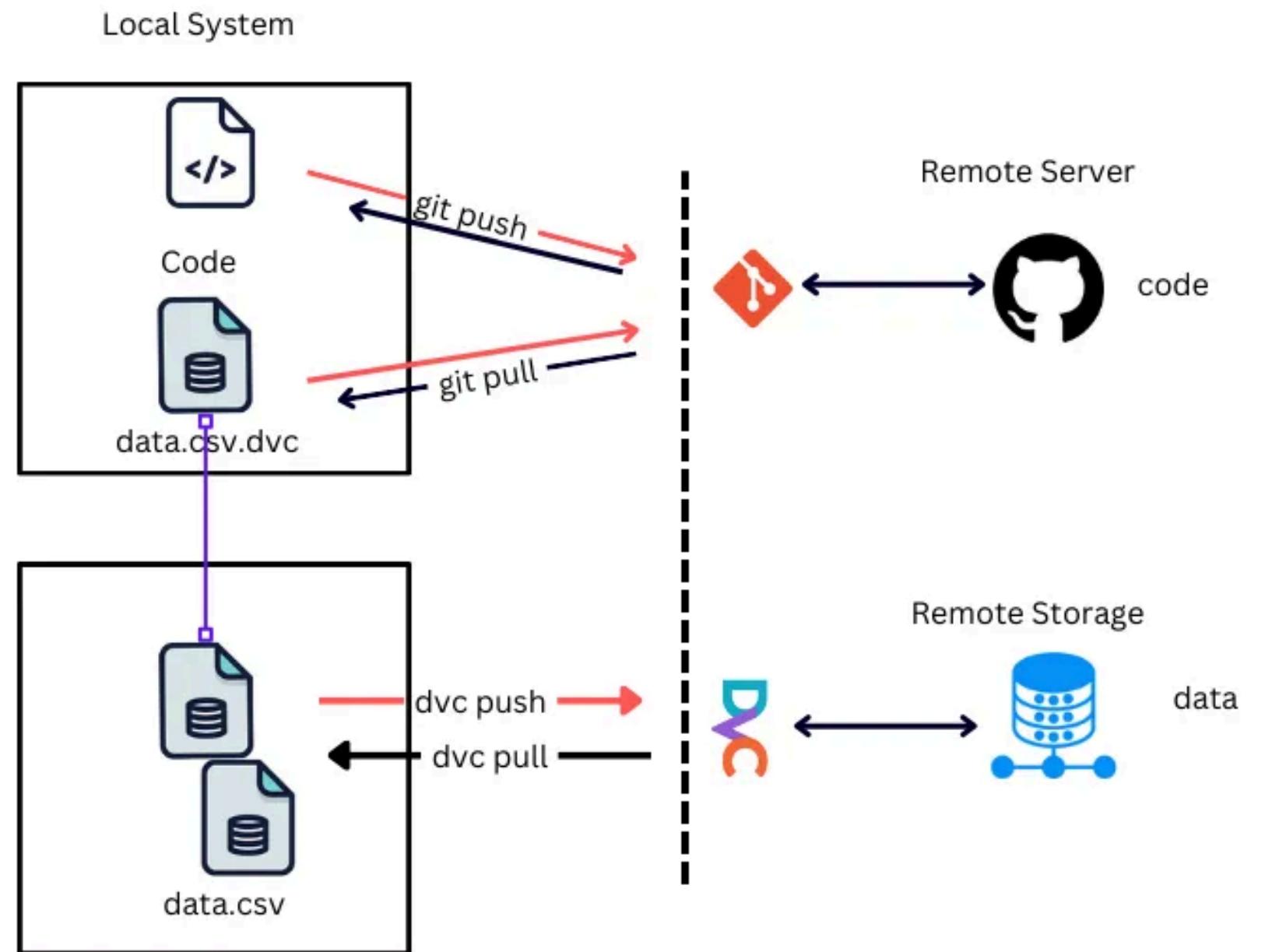


**Git** cannot be used for datasets versioning, specially with large datasets. For dataset versioning we have **DVC**, which integrates with Git. It is an open-source command-line tool that mimics Git flows and commands.

DVC Characteristics:

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

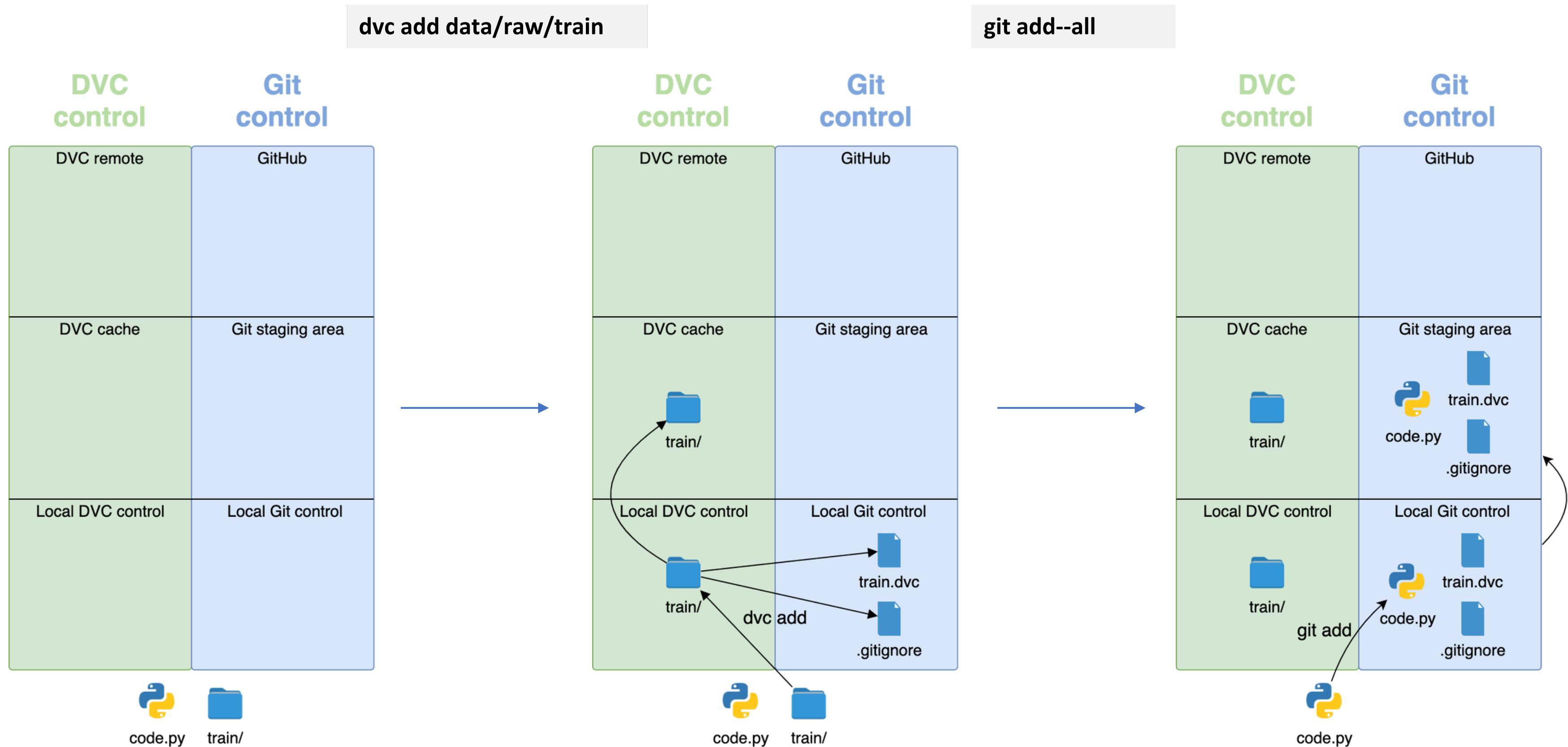




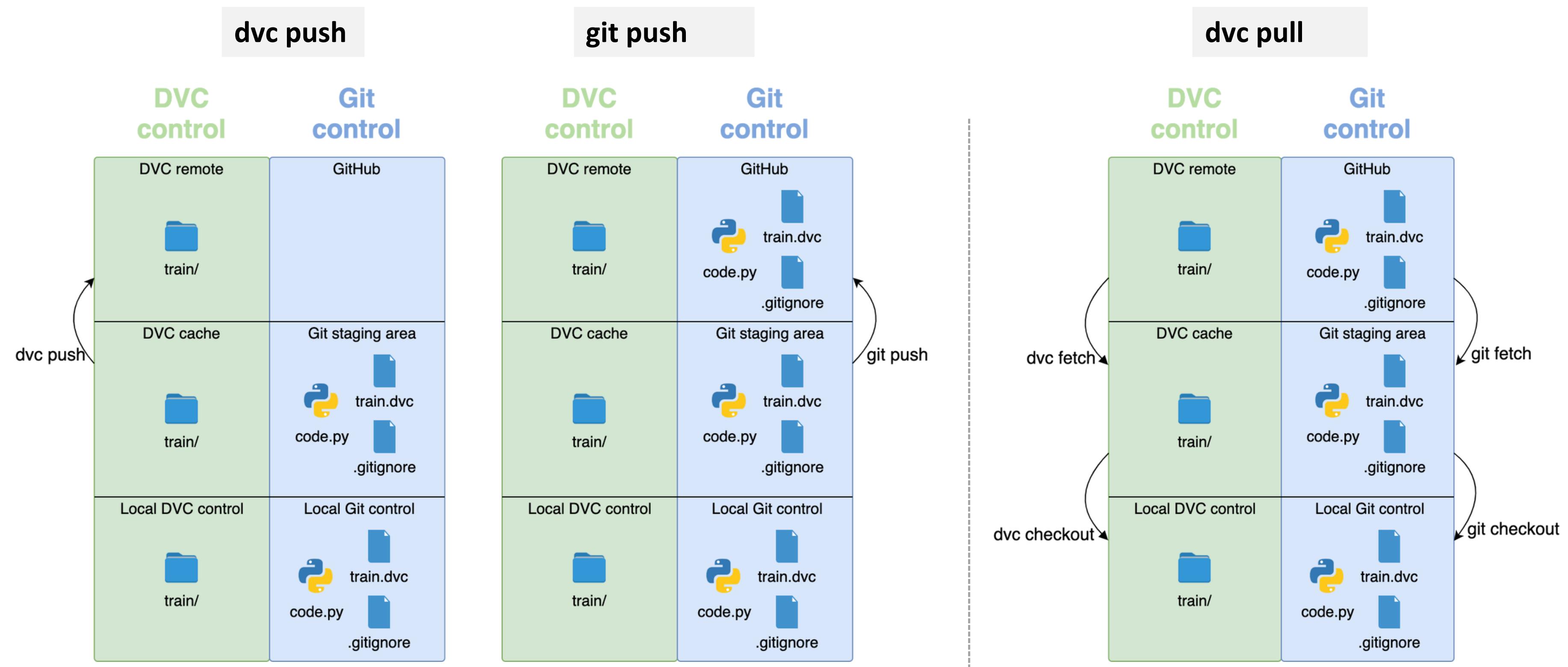
**Data versioning** is crucial because it allows for tracking changes to data over time, ensuring that users have access to accurate and reliable information. By maintaining a history of data revisions, data versioning enables reproducibility, collaboration, and auditing in various fields such as software 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 with DVC

Large data 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

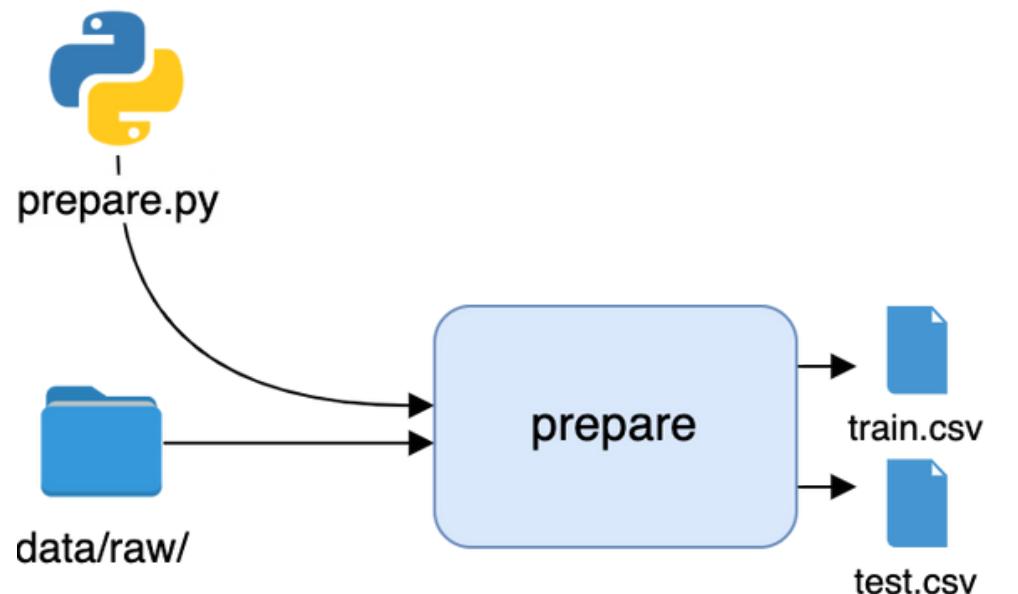
```
md5: 62bdac455a6574ed68a1744da1505745
outs:
  - md5: 96652bd680f9b8bd7c223488ac97f151
    path: model.joblib
    cache: true
    metric: false
    persist: false
```

# DVC pipelines

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

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:

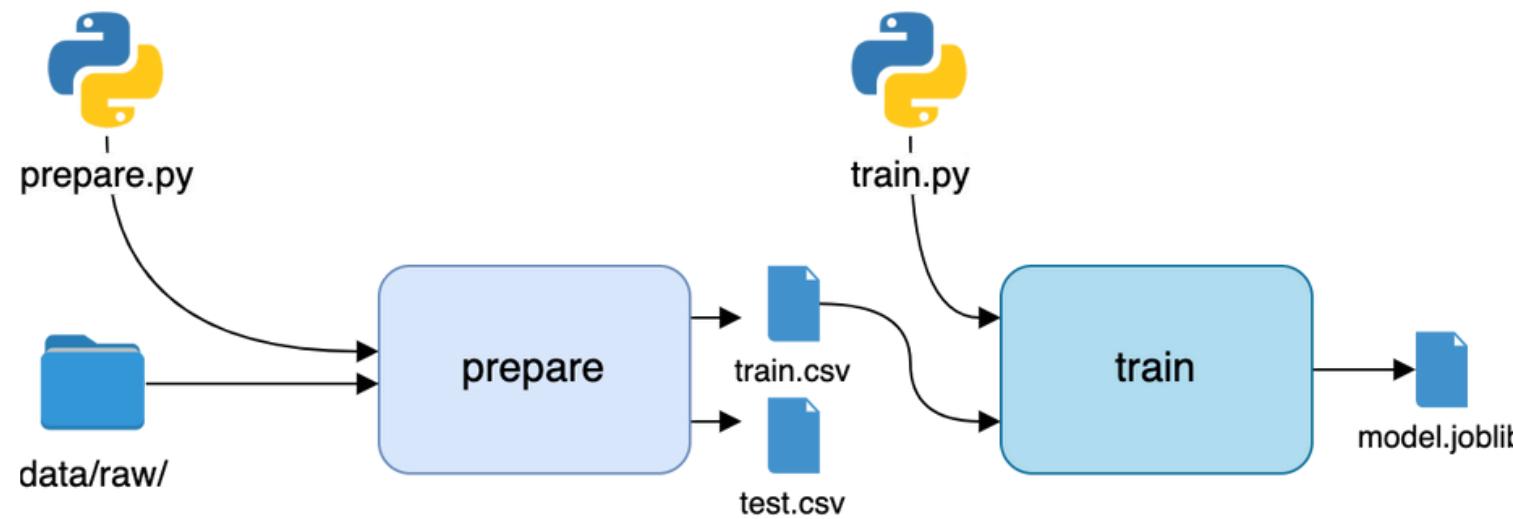
- **dvc.yaml**
- **dvc.lock**

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

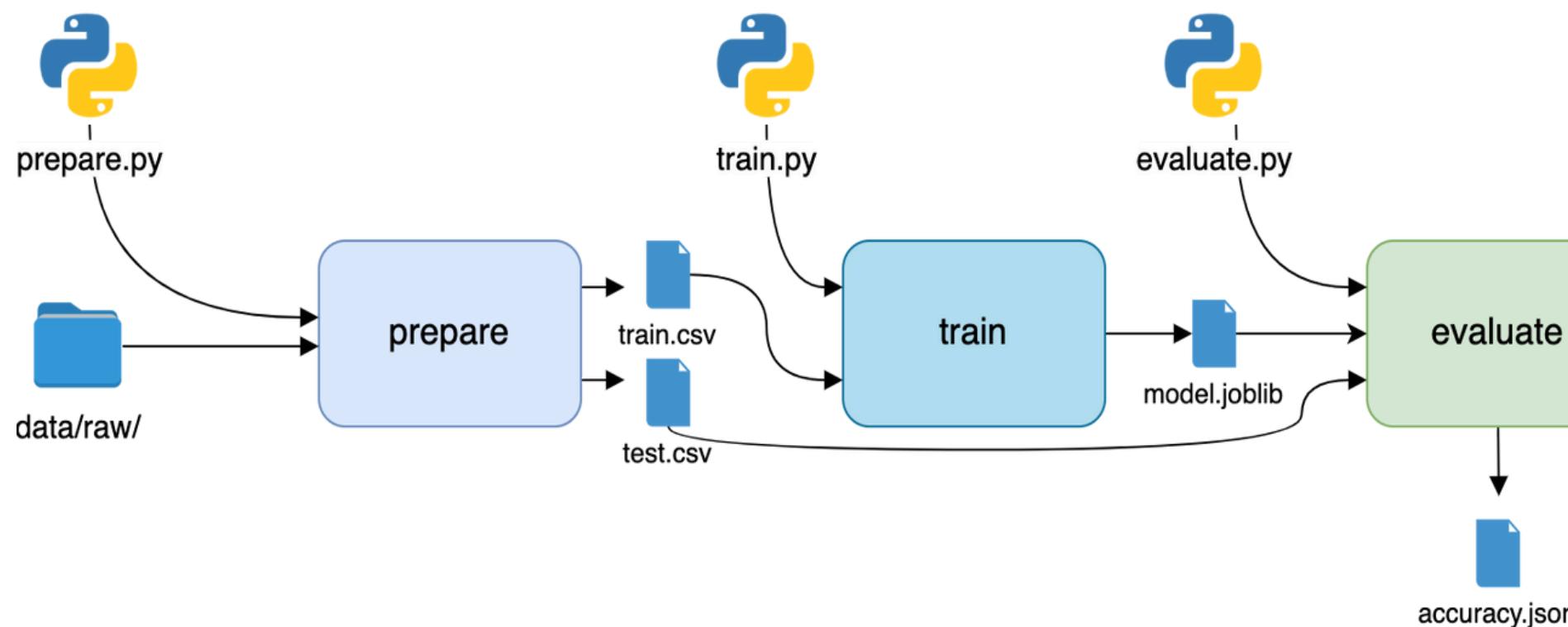
YAML

```
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
```

# Examples of DVC 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
```



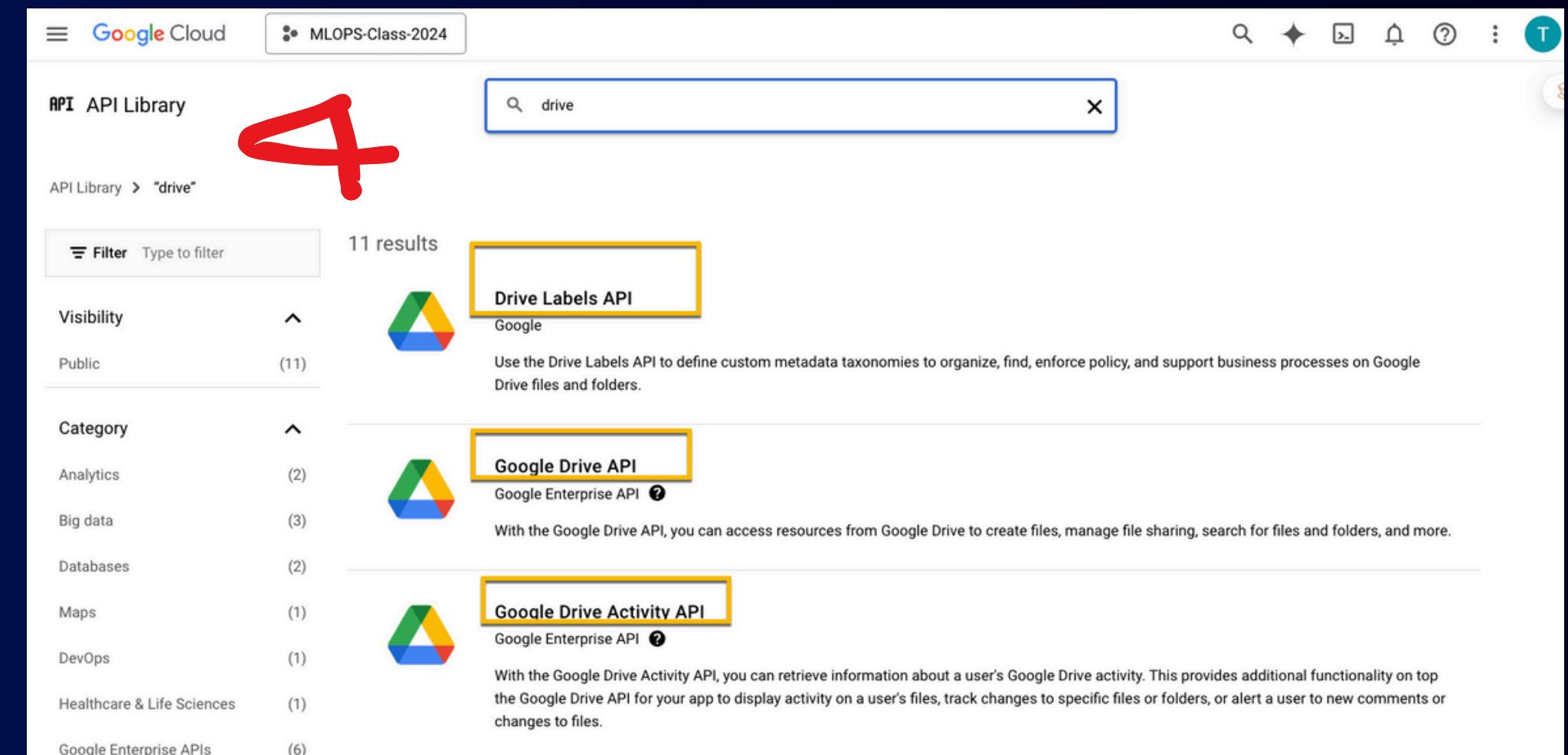
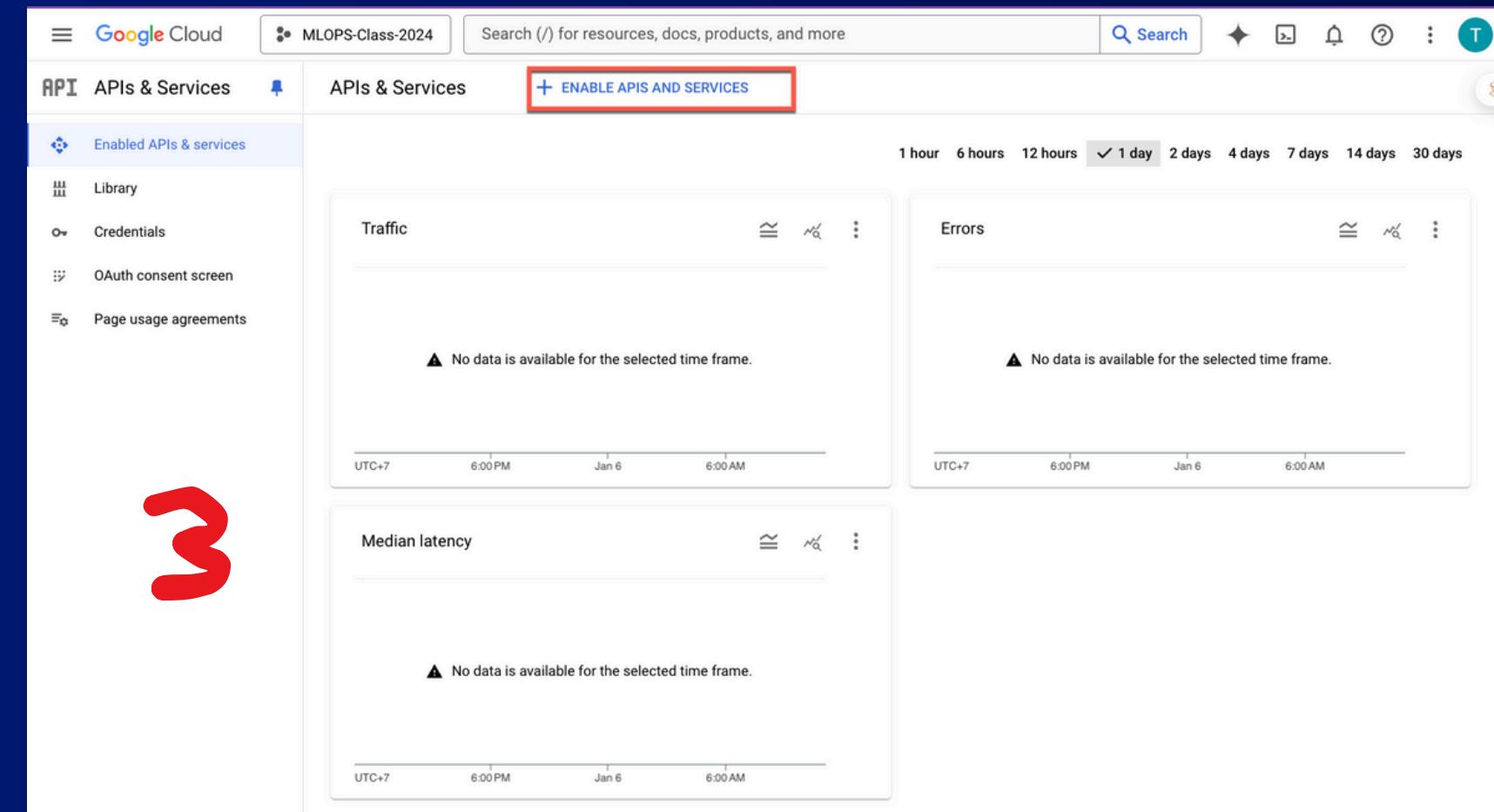
# LAB 1. Setting Up a Workflow with DVC Google Cloud Storage(GCS)



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

The screenshot shows the Google Cloud Welcome screen for the project "MLOPS-Class-2024". The interface includes a "Welcome" section with a large red number 1, a "Quick access" sidebar with various services like API APIs & Services, IAM & Admin, Billing, Compute Engine, Cloud Storage, BigQuery, VPC network, and Kubernetes Engine, and a central area with buttons for creating VMs, running queries in BigQuery, creating GKE clusters, and creating storage buckets. A "Try Gemini" button is also present.

The screenshot shows the Google Cloud APIs & Services search results page. A search bar at the top contains the text "apis & services". The results list includes several items, with the first item, "API APIs & Services", highlighted with a yellow box. A large red number 2 is overlaid on the page. The search results table has columns for Type, Producer, and Description. Other visible items include "Enabled APIs & services", "Cloud Managed Services", "Mulai Menggunakan Service Control API", "Amaze for App", "Delete a service account key", "Database Migration API", "Managed Security Services from Qodea", "mmob-embedded", and "Public Certificate Authority API".



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.

**ENABLE** TRY THIS API

OVERVIEW DOCUMENTATION SUPPORT RELATED PRODUCTS

**Overview**

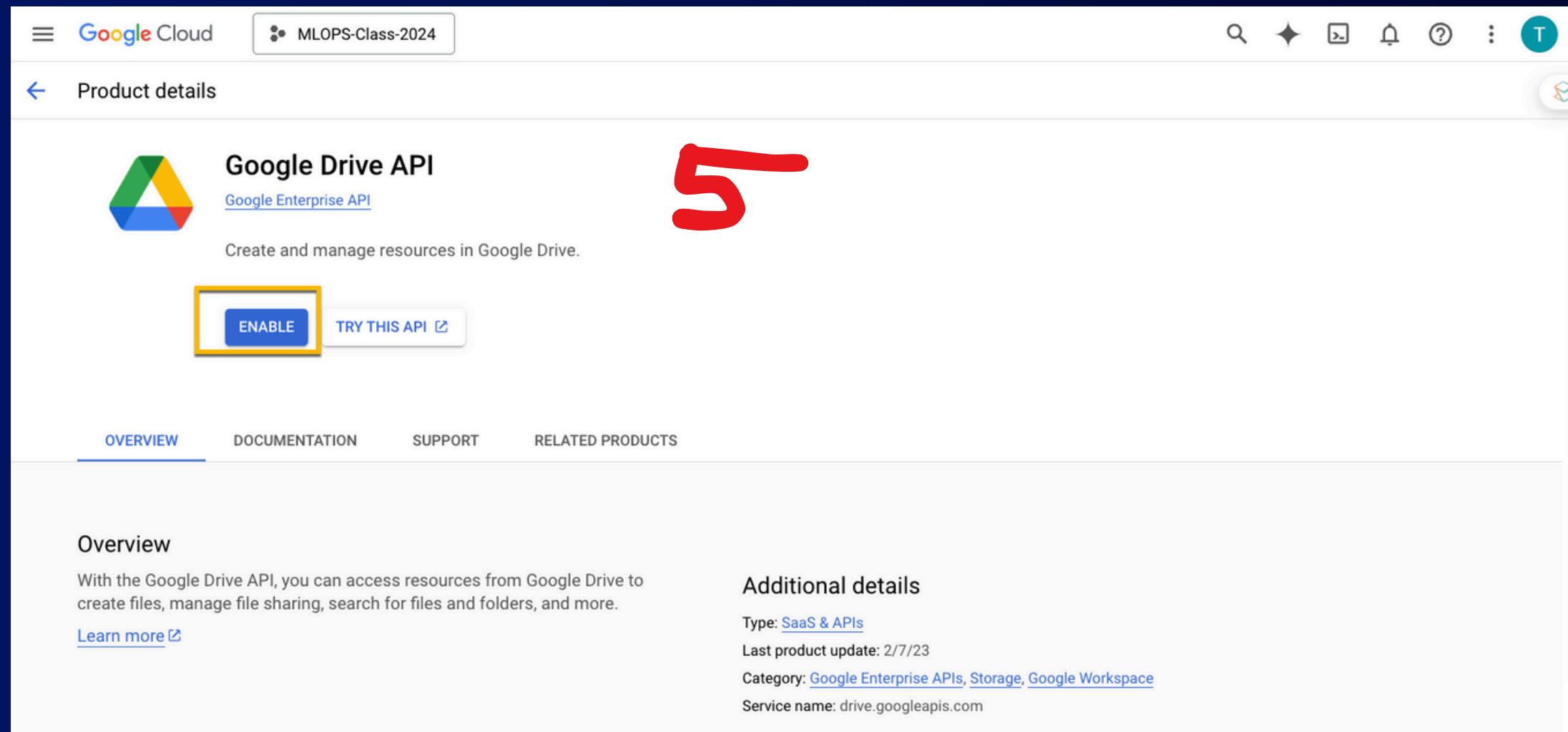
With the Google Drive API, you can access resources from Google Drive to create files, manage file sharing, search for files and folders, and more.

[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

5



Google Cloud MLOPS-Class-2024

Search (/) for resources, docs, products, and more

API APIs & Services

Enabled APIs & services

Library Credentials OAuth consent screen Page usage agreements

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](#)

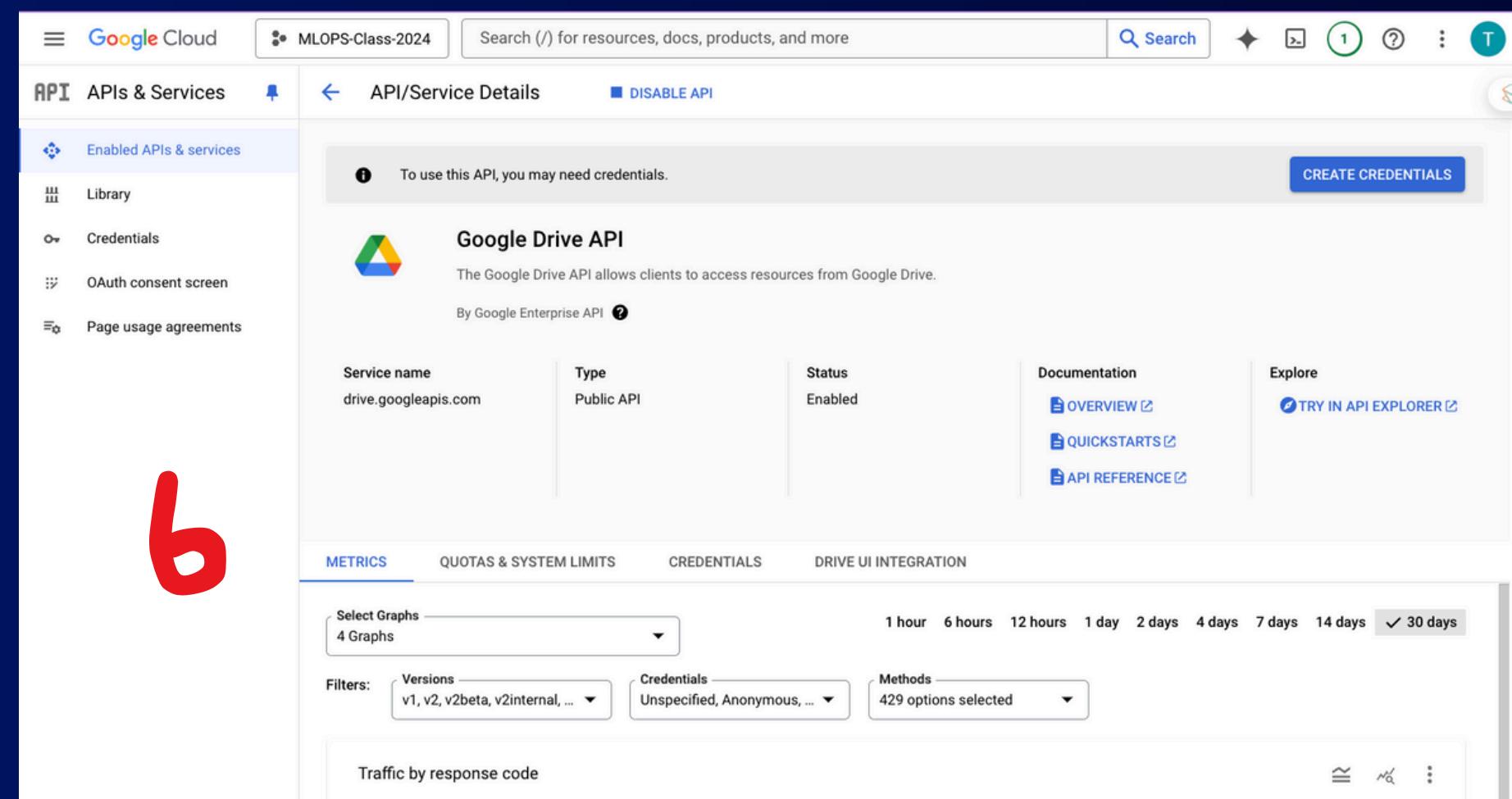
METRICS QUOTAS & SYSTEM LIMITS CREDENTIALS DRIVE UI INTEGRATION

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

6



Google Cloud MLOPS-Class-2024

Product details

## Drive Labels API



Define label taxonomies for your organization.

**ENABLE** TRY THIS API

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](#)

7

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

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 of 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](#)

8

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

The screenshot shows the Google Cloud Welcome screen for the project "MLOPS-Class-2024". The navigation bar at the top includes the Google Cloud logo, the project name "MLOPS-Class-2024", and a search bar. A red box highlights the "Navigation menu" icon in the top-left corner of the main content area. Below the header, there's a "Welcome" section with a "Cloud overview" card, followed by a "Quick access" grid containing links for API APIs & Services, IAM & Admin, Billing, Compute Engine, Cloud Storage, BigQuery, VPC network, and Kubernetes Engine. At the bottom, there are buttons for creating a VM, running a query in BigQuery, creating a GKE cluster, and creating a storage bucket. A large red number "10" is overlaid on the bottom-left of the screen.

The screenshot shows the Google Cloud IAM & Admin menu. The "Service Accounts" option is highlighted with a yellow box. The menu also lists other options like IAM, PAM, Principal Access Boundary, Organizations (PREVIEW), Marketplace, Kubernetes Engine, Cloud Storage, BigQuery, VPC Network, Cloud Run, SQL, Logging, Security, Compute Engine, and Vertex AI. At the bottom of the menu, there are buttons for "VIEW ALL PRODUCTS" and "Create a Project". The background shows a blurred view of the Google Cloud interface, including a "drive" folder and various service icons. A grey callout box on the right side of the screen says "Try our most advanced model: Gemini 1.5 Pro" and "Try Gemini".

**1 Service account details**

Service account name  X C

Display name for this service account

Service account ID \*  X C

Email address: dvc-project@mlops-class-2024.iam.gserviceaccount.com ✉

Service account description

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**

**1 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 ▼

+ ADD ANOTHER ROLE

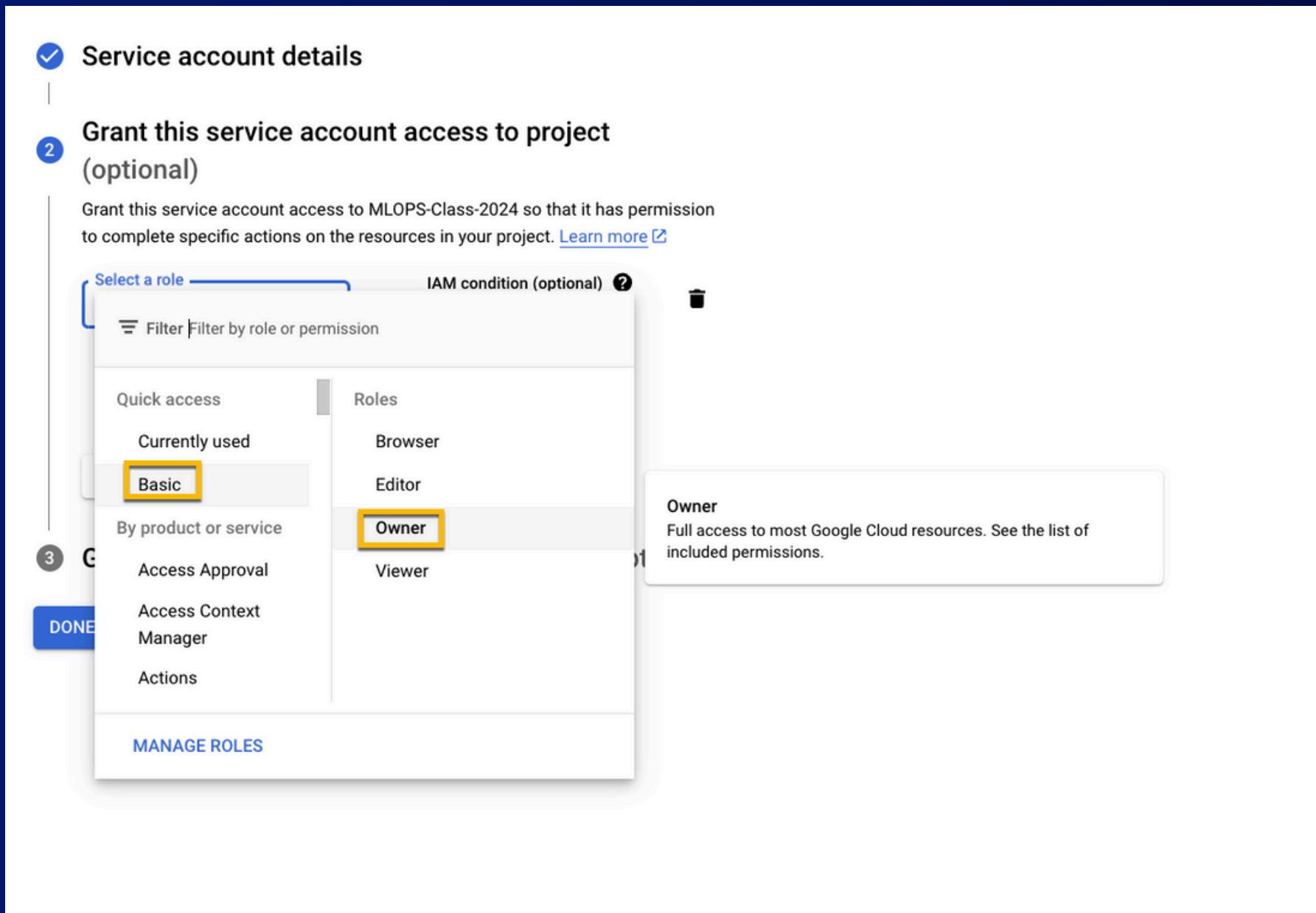
**CONTINUE**

**3 Grant users access to this service account (optional)**

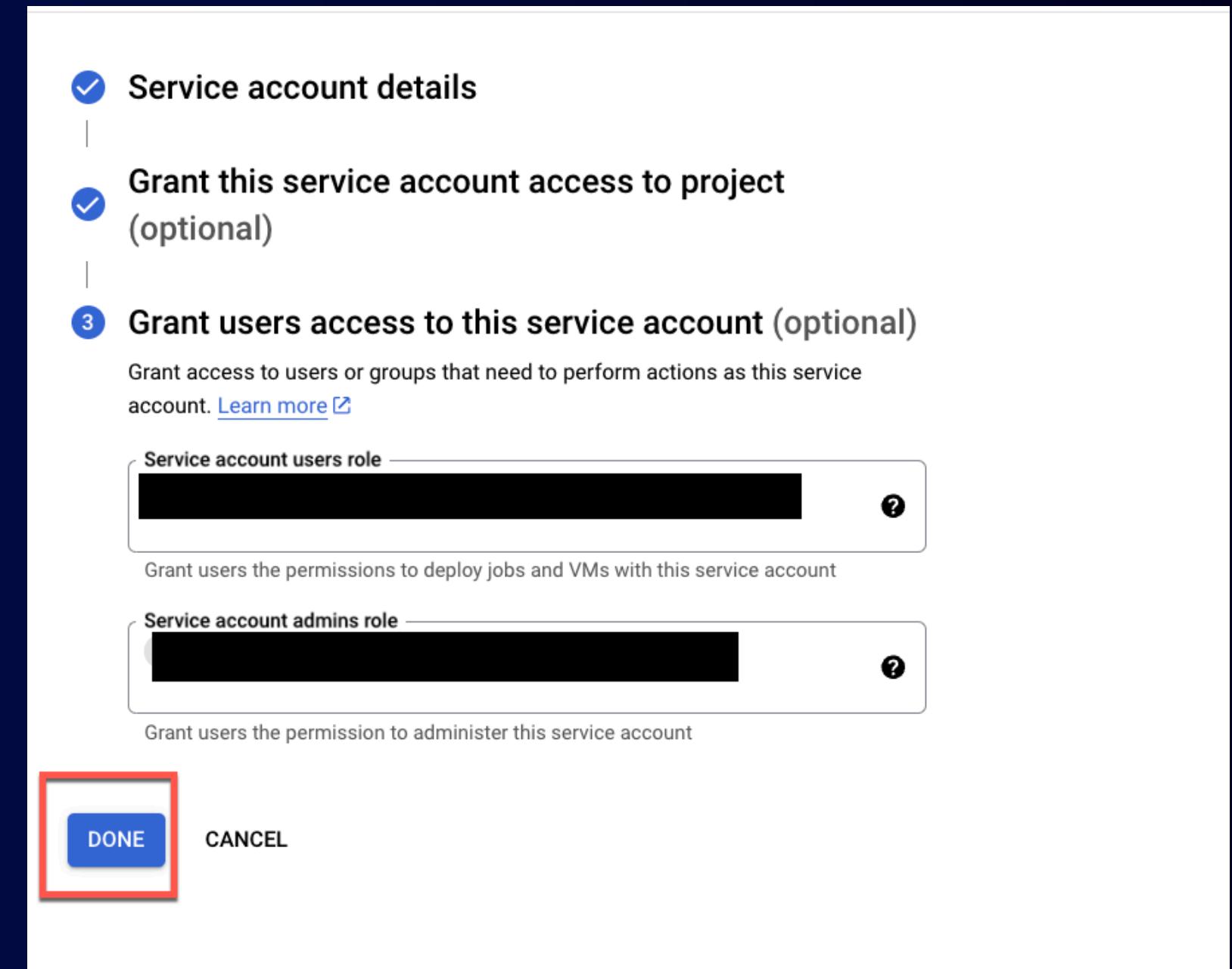
**DONE** **CANCEL**

12

| 3



| 4



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.**

Service accounts for project "MLOPS-Class-2024"

A service account represents a Google Cloud service identity, such as code running on Compute Engine VMs, App Engine apps, or systems running outside Google. [Learn more about service accounts.](#)

Organization policies can be used to secure service accounts and block risky service account features, such as automatic IAM Grants, key creation/upload, or the creation of service accounts entirely. [Learn more about service account organization policies.](#)

<input type="checkbox"/> Filter Enter property name or value	Status	Name ↑	Description	Key ID	Key creation date	OAuth 2 Client ID <a href="#">?</a>	Actions
<input type="checkbox"/> Email <input type="checkbox"/> <a href="#">dvc-project@mlops-class-2024.iam.gserviceaccount.com</a>	<input checked="" type="checkbox"/> Enabled	dvc_project	my dvc	No keys	[REDACTED]	[REDACTED]	<a href="#">⋮</a>

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

Service accounts for project "MLOPS-Class-2024"

A service account represents a Google Cloud service identity, such as code running on Compute Engine VMs, App Engine apps, or systems running outside Google. [Learn more about service accounts](#).

Organization policies can be used to secure service accounts and block risky service account features, such as automatic IAM Grants, key creation/upload, or the creation of service accounts entirely. [Learn more about service account organization policies](#).

Email	Status	Name	Description	Key ID	Key creation date	OAuth 2 Client ID	Actions
<a href="#">dvc-project@mlops-class- iam.gserviceaccount.com</a>	Enabled	dvc_project	my dvc	No keys			<ul style="list-style-type: none"><li>⋮</li><li>Manage details</li><li>Manage permissions</li><li><b>Manage keys</b></li><li>View metrics</li><li>View logs</li><li>Disable</li><li>Delete</li></ul>

dvc\_project

DETAILS PERMISSIONS KEYS METRICS LOGS

**Keys**

⚠ Service account keys could pose a security risk if compromised. We recommend you avoid downloading service account keys and instead use the [Workload Identity Federation](#). [Learn more about the best way to authenticate service accounts on Google Cloud](#).

ⓘ Google automatically disables service account keys detected in public repositories. You can customize this behavior by using the 'iam.serviceAccountKeyExposureResponse' organization policy. [Learn more](#)

Add a new key pair or upload a public key certificate from an existing key pair.

Block service account key creation using [organization policies](#). [Learn more about setting organization policies for service accounts](#)

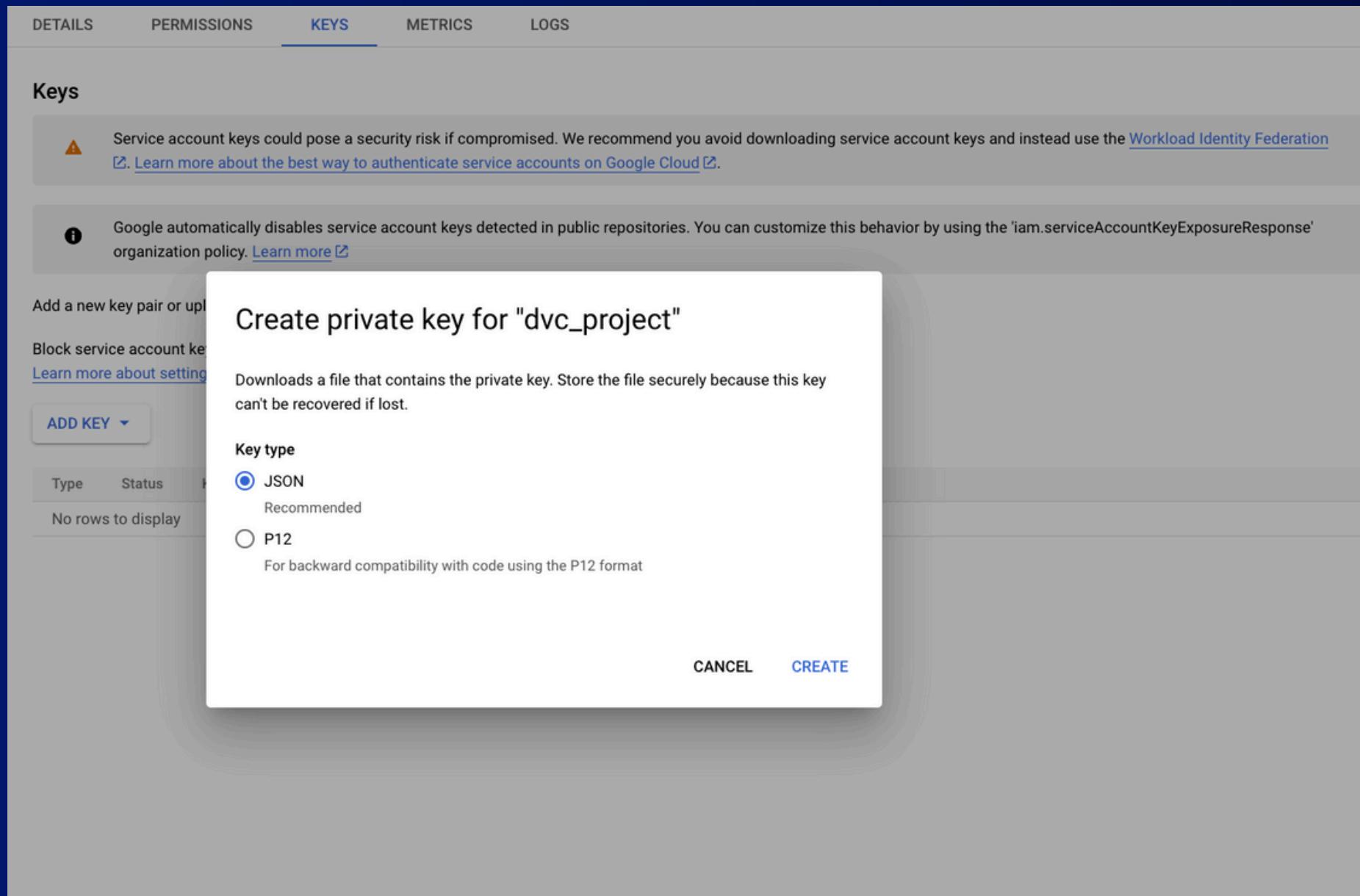
**ADD KEY**

**Create new key**

Upload existing key

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

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

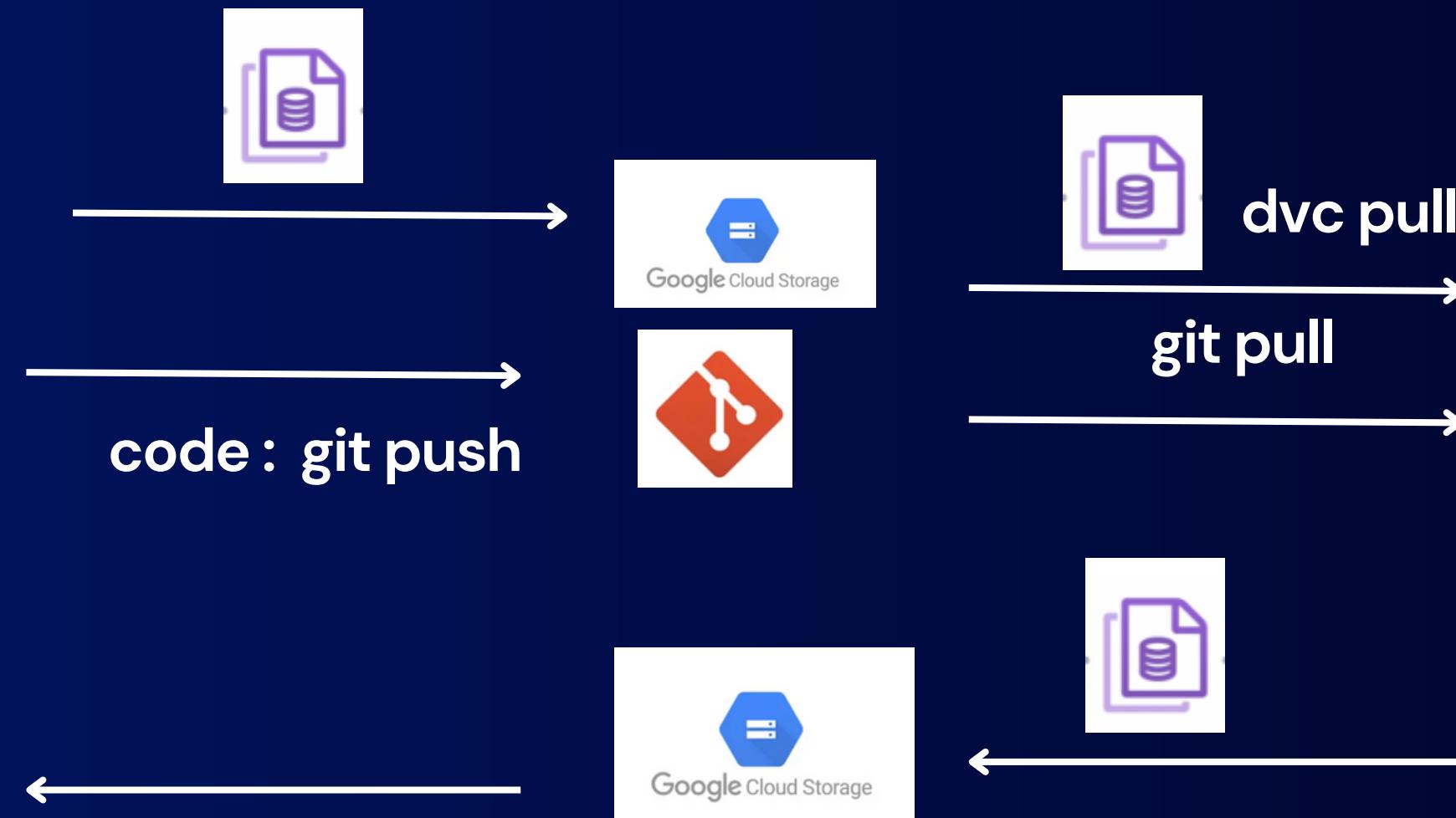


**Important: Store the API key in a local folder as `credentials.json`, but do not commit it to GitHub. If you do so, GitHub will raise a warning, and Google will be notified, revoking the credentials.**

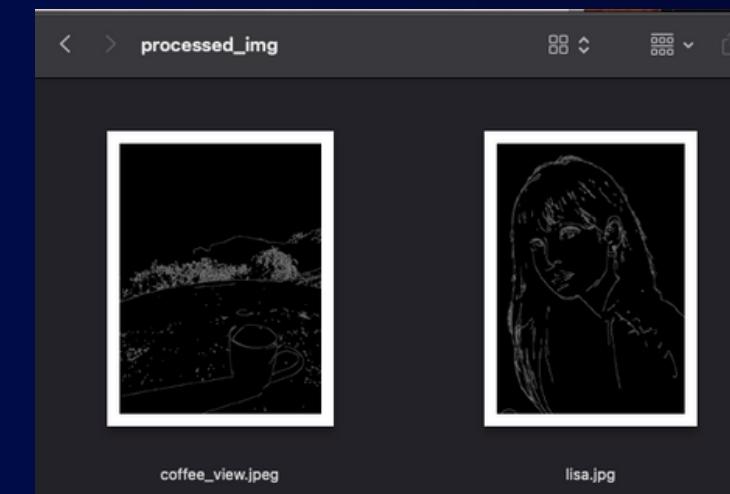
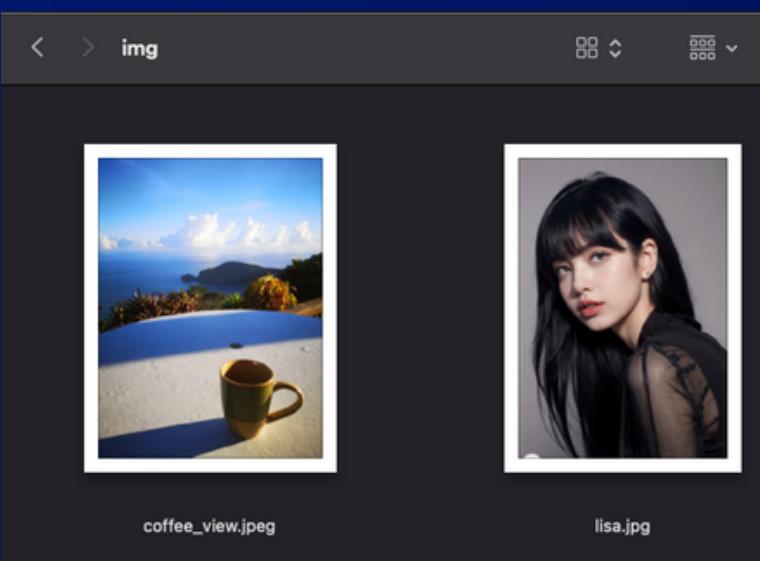


# LAB 2. DVC data application

images data : dvc push



update image data : dvc pull



```
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

**Create a new repository**

A repository contains all project files, including the revision history. Already have a project repository elsewhere? [Import a repository](#).

Required fields are marked with an asterisk (\*).

**Owner \*** Tuchsanai / **Repository name \*** dvc-repo

dvc-repo is available.

Great repository names are short and memorable. Need inspiration? How about [automatic-octo-giggle](#) ?

**Description (optional)**

**Public**  
Anyone on the internet can see this repository. You choose who can commit.

**Private**  
You choose who can see and commit to this repository.

**Initialize this repository with:**

**Add a README file**  
This is where you can write a long description for your project. [Learn more about READMEs](#).

**Add .gitignore**

.gitignore template: [None](#)

Choose which files not to track from a list of templates. [Learn more about ignoring files](#).

**Choose a license**

License: [None](#)

A license tells others what they can and can't do with your code. [Learn more about licenses](#).

ⓘ You are creating a private repository in your personal account.

**Create repository**

Tuchsanai / dvc-repo

Type ⌘ to search

Code Issues Pull requests Actions Projects Security Insights Settings

⚠ Don't get locked out of your account. Download your recovery codes or add a passkey so you don't lose access when you get a new device.

dvc-repo Private

Unwatch 1 Fork 0 Star 0

Start coding with Codespaces

Add a README file and start coding in a secure, configurable, and dedicated development environment.

Create a codespace

Add collaborators to this repository

Search for people using their GitHub username or email address.

Invite collaborators

Quick setup — if you've done this kind of thing before

Set up in Desktop or HTTPS SSH https://github.com/Tuchsanai/dvc-repo.git

Get started by [creating a new file](#) or [uploading an existing file](#). We recommend every repository include a [README](#), [LICENSE](#), and [.gitignore](#).

...or create a new repository on the command line

```
echo "# dvc-repo" >> README.md
git init
git add README.md
git commit -m "first commit"
git branch -M main
git remote add origin https://github.com/Tuchsanai/dvc-repo.git
git push -u origin main
```

...or push an existing repository from the command line

```
git remote add origin https://github.com/Tuchsanai/dvc-repo.git
git branch -M main
git push -u origin main
```

# gs://<your-bucket-name>

The screenshot shows the Google Cloud Storage Bucket details page for a bucket named 'dvc\_tp'. The bucket's name is displayed prominently at the top center. On the left, a sidebar menu includes 'Overview', 'Buckets' (which is selected and highlighted with a yellow box), 'Monitoring', and 'Settings'. The main content area displays the bucket's location as 'asia-southeast1 (Singapore)', storage class as 'Standard', public access as 'Not public', and protection as 'Soft Delete'. Below this, a navigation bar offers tabs for 'OBJECTS' (selected), 'CONFIGURATION', 'PERMISSIONS', 'PROTECTION', 'LIFECYCLE', 'OBSERVABILITY', and 'INVENTORY'. Under the 'OBJECTS' tab, a breadcrumb path shows 'Buckets > dvc\_tp'. There are buttons for 'CREATE FOLDER', 'UPLOAD', 'TRANSFER DATA', and 'OTHER SERVICES'. A filter bar allows filtering by name prefix, objects, and services, with options to show 'Live objects only'. The main table at the bottom shows columns for Name, Size, Type, Created, Storage class, Last modified, Public access, Version history, and Encryption. A message indicates 'No rows to display'.