NAME : NAVIYA DHARSHINI A S
ROLL NO : 23AD083
DAY -07 - 10/07/2025

DVC (Data Version Control) - Full Notes from Basics to Advanced

---

🔰 1. What is DVC?

- **DVC = Git for Data + ML Pipelines**
- Manages **data**, **models**, and **experiments** in ML projects.
- Works with Git but tracks **large files**, **directories**, and **ML stages**.
- Makes projects **reproducible**, **shareable**, and **collaborative**.

⚙️ 2. Why DVC?

🔥 **Git Problem**	✅ **DVC Solution**
Can't track large datasets	DVC tracks large files via `.dvc` metadata
Git lacks pipeline tracking	DVC handles pipelines + stages like `make`
No reproducibility for ML	DVC ensures same input = same output
Can't share data easily	DVC supports remote storage (S3, GDrive etc)

📁 4. DVC Project Structure

```
my-ml-project/
├── data/            # Raw datasets (not tracked by Git)
├── data.dvc         # DVC metadata for dataset
├── model.pkl        # Trained model (tracked by DVC)
├── model.pkl.dvc
├── dvc.yaml         # ML pipeline stages
├── dvc.lock         # Exact command runs (hashes etc.)
├── .dvc/            # DVC internal files
└── .git/            # Git repo
```

📦 5. DVC Core Commands (Basic Level)

🔹 Initialize DVC in your project

```
dvc init
```

Tack a data file or folder

```
dvc add data/
```

> This creates `data.dvc` and adds `data/` to `.gitignore`.

🔹 Push data to remote (e.g. Google Drive, S3)

```
dvc remote add -d myremote s3://mybucket/data
dvc push
```

🔹 Pull data in a new machine / from collaborator

```
git clone <repo>
dvc pull
```

🔹 Remove data from local (for space)

```
dvc remove data.dvc
```

⚙️ 6. DVC + Git Workflow

```
# Add large files via DVC
dvc add data/

# Git track the DVC metafiles
git add data.dvc .gitignore
git commit -m "Add dataset via DVC"

# Push code to Git + data to DVC remote
git push
dvc push
```

Step-by-Step: Push Local Code to GitHub Repo (No README) using Token

---

✅ 1. Create a GitHub Repo (without README)

1. Go to https://github.com/new
2. Fill in:
    - **Repo name**
    - Set to **Private** or **Public**
    - ❌ *Don't check README / .gitignore / License*
3. Click **Create repository**

> You'll see instructions to push from command line — we'll use that, but with a token.

📂 2. Initialize Git Locally

Go to your project folder in terminal:

```
cd your-project-folder/
git init
```

Optional (if not already set):

```
git config user.name "your-name"
git config user.email "your@email.com"
```

📄 3. Add Files + Commit

```
git add .
git commit -m "Initial commit"
```

🔑 4. Add GitHub Remote Using Personal Access Token (PAT)

Let’s say:

- Your GitHub username: `yourusername`
- Your repo name: `my-repo`
- Your token: `ghp_xxxxyyyyzzzz` (example)

```
git remote add origin
https://<TOKEN>@github.com/<USERNAME>/<REPO>.git
```

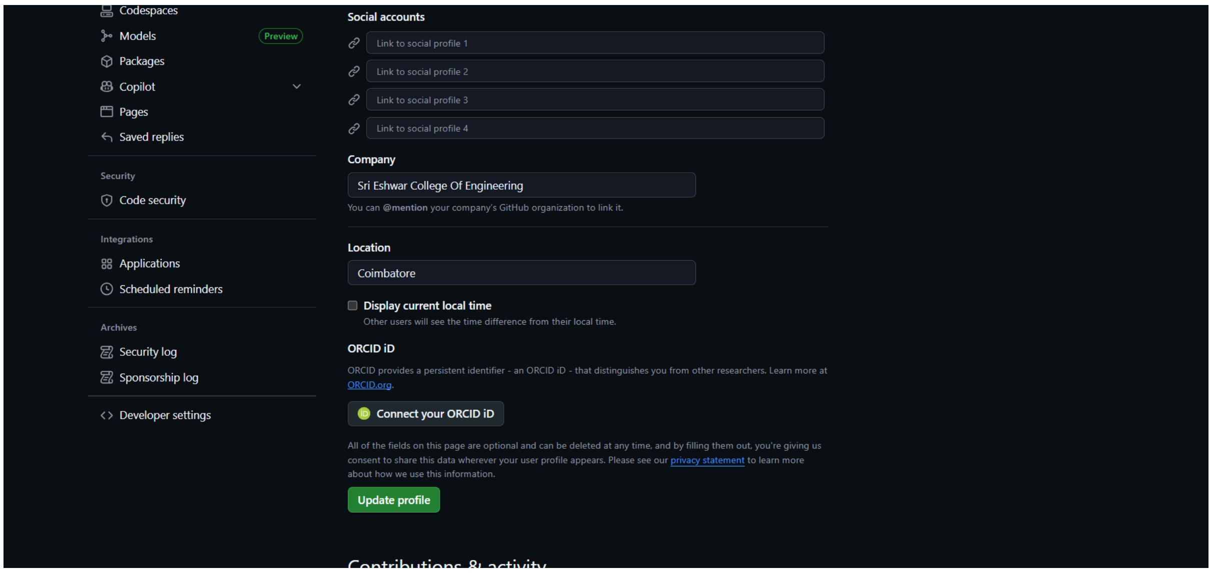
For example:

```
git remote add origin
https://ghp_xxxxyyyyzzzz@github.com/yourusername/my-repo.git
```

> ✅ Make sure token has **repo** scope if private.

🚀 5. Push to GitHub

```
git branch -M main
git push -u origin main
```



Social accounts

Link to social profile 1
Link to social profile 2
Link to social profile 3
Link to social profile 4

Company

Sri Eshwar College Of Engineering

You can **@mention** your company's GitHub organization to link it.

Location

Coimbatore

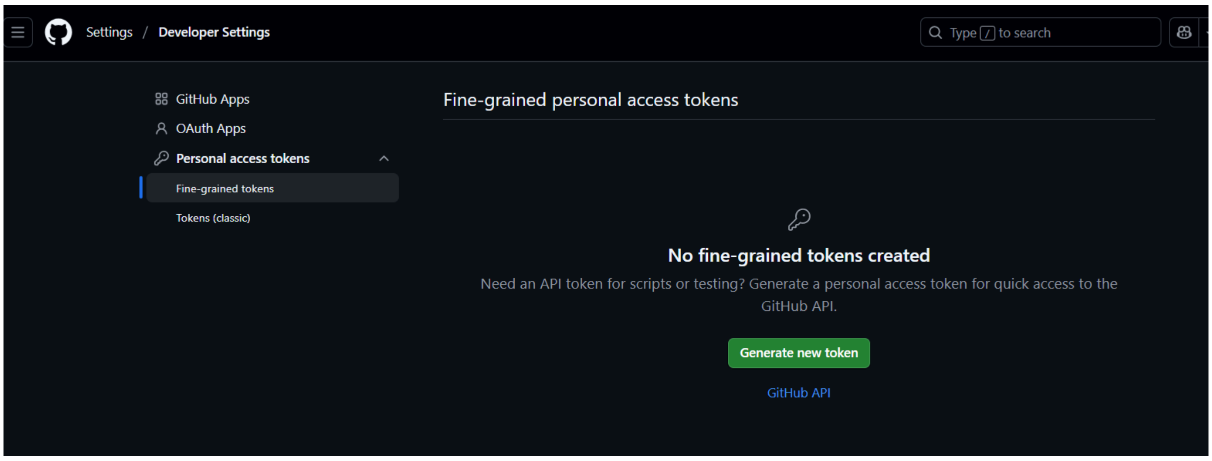
Display current local time

Other users will see the time difference from their local time.

ORCID iD

ORCID provides a persistent identifier - an ORCID iD - that distinguishes you from other researchers. Learn more at ORCID.org.

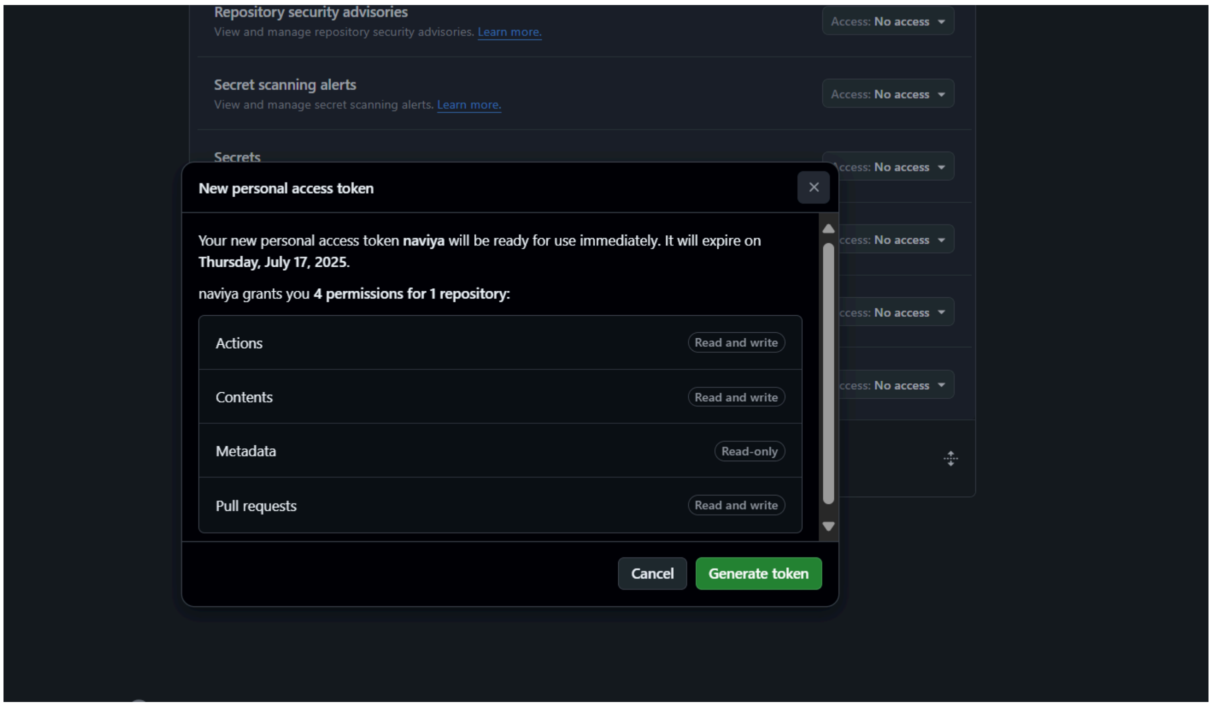
All of the fields on this page are optional and can be deleted at any time, and by filling them out, you're giving us consent to share this data wherever your user profile appears. Please see our privacy statement to learn more about how we use this information.



Fine-grained personal access tokens

No fine-grained tokens created

Need an API token for scripts or testing? Generate a personal access token for quick access to the GitHub API.


Repository security advisories
View and manage repository security advisories. Learn more.
Access: No access

Secret scanning alerts
View and manage secret scanning alerts. Learn more.
Access: No access

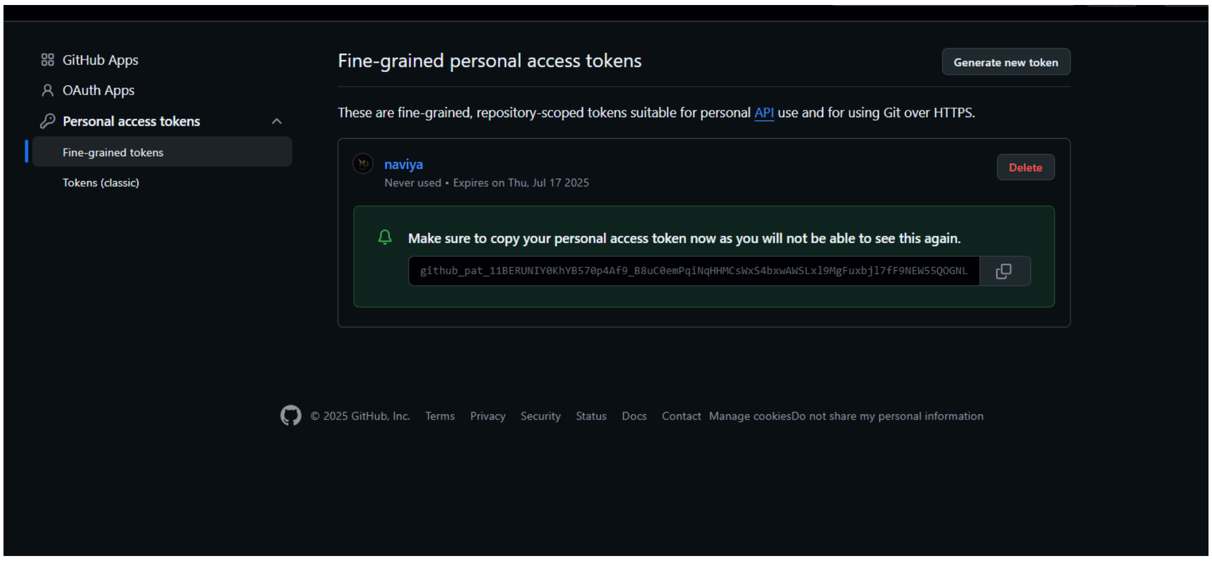
Secrets

New personal access token

Your new personal access token **naviya** will be ready for use immediately. It will expire on **Thursday, July 17, 2025**.

naviya grants you **4 permissions for 1 repository:**

Actions	Read and write
Contents	Read and write
Metadata	Read-only
Pull requests	Read and write



Fine-grained personal access tokens

These are fine-grained, repository-scoped tokens suitable for personal API use and for using Git over HTTPS.

naviya
Never used • Expires on Thu, Jul 17 2025

Make sure to copy your personal access token now as you will not be able to see this again.

```
github_pat_11BERUNIY0KhYB570p4Af9_B8uC0emPqiNqHHMCsWxS4bxwAWSLxl9MgFuxbjl7fF9NEW55QOGNL
```

© 2025 GitHub, Inc. Terms Privacy Security Status Docs Contact Manage cookiesDo not share my personal information

---

🌐 . DVC Remotes (Cloud Backends)

Supports:
✅ AWS S3
✅ Google Drive
✅ Azure Blob
✅ GCP Bucket
✅ SSH, HDFS, WebDAV

bash
CopyEdit

```
dvc remote add -d storage s3://mybucket
dvc remote modify storage access_key_id ...
dvc push
```