DVC Tutorial: Track and Share Data with Google Cloud Storage

Prerequisites

- DVC installed with GCS support: `pip install "dvc[gcs]"`
- Google Cloud SDK (`gcloud`) installed and configured
- A Google Cloud Storage (GCS) bucket
- DVC initialized in your Git repo: `dvc init`

Step-by-Step Guide

Step 1: Authenticate with GCP

Run the following to authenticate:

gcloud auth application-default login

Step 2: Add a Remote in DVC

dvc remote add -d gcsremote gs://your-dvc-bucket-name

dvc remote modify gcsremote credentialpath ~/.config/gcloud/application_default_credentials.json

Step 3: Track a File or Folder

Example using a CSV:

mkdir data

echo "name,age

Alice,30

Bob,25" > data/people.csv

dvc add data/people.csv

git add data/people.csv.dvc .gitignore

git commit -m "Track data with DVC"

Step 4: Push to GCS

dvc push

Step 5: Simulate Collaboration

Delete your file:

rm data/people.csv

Then pull it back from GCS:

dvc pull

Step 6: Version Your Data

Update the file, re-add with DVC, commit with Git, and push again:

echo "Charlie,40" >> data/people.csv

dvc add data/people.csv

git commit -am "Add Charlie"

dvc push

Step 7: Restore Old Version

Use Git and DVC together to go back:

git checkout HEAD~1 data/people.csv.dvc

dvc checkout

Benefits of DVC with GCS

- Version control for large data files
- Collaborate without sending files over email
- Push/pull from cloud storage easily
- Integrates with Git and CI/CD