DVC (DATA VERSION CONTROL)

→ STORE AND PROCESS DATA FILES (or) DATASETS TO PRODUCE OTHER DATA (or) MACHINE LEARNING MODELS

- ↳ TRACK AND SAVE DATA AND MACHINE LEARNING MODELS (THE SAME WAY YOU CAPTURE CODE)
- ↳ COMPARE MODEL METRICS
- ↳ CREATE AND SWITCH B/W VERSIONS OF DATA AND ML MODELS EASILY
  - ↳ IN DATA INGESTION PHASE YOU GET DIFFERENT TYPE OF DATA.

→ YOU CAN MANAGE ENTIRE VERSIONS OF DATA & MODELS USING DVC

→ DVC WILL ONLY TRACK THE DATA FILES

→ GIT WILL BE USED TO TRACK CODE INFORMATION AND CONFIGURATION FILE THAT IS REQUIRED BY

DVC

→ FIRST DO THE GIT INIT IN THE VSCODE TERMINAL

→ INSTALL DVC

↳ PIP INSTALL -R REQUIREMENTS.TXT FILE

(or)

↳ PIP INSTALL DVC

→ CREATE A DATA FOLDER INSIDE THAT

↳ CREATE A SAMPLE DATA.TXT FILE

→ DVC INIT IN THE TERMINAL

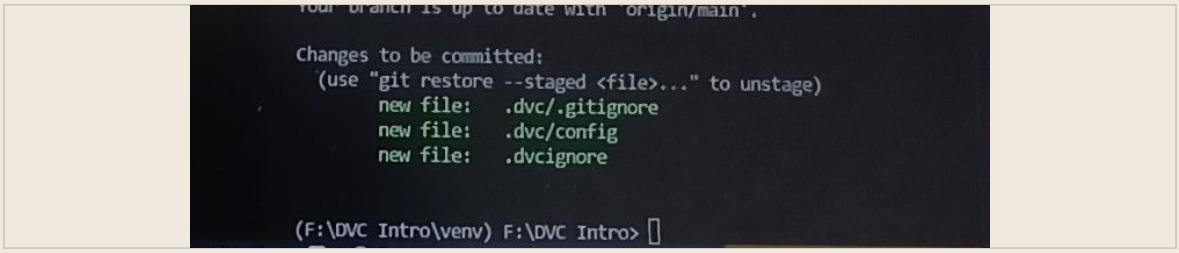
↳ ONCE YOU PRESS ENTER .DVC FOLDER WILL BE

CREATED WITH IN TMP .GITIGNORE, CONFIG FILE WILL BE CREATED

→ IN CONFIG FILE YOU CAN PROVIDE DETAILS ABOUT THE REMOTE REPOSITORY WHERE YOU WANT TO STORE THE DATA (YOU CAN GIVE THOSE DETAILS)

→ NEXT USING GIT STATUS COMMAND IN THE TERMINAL YOU CAN SEE UPDATED DVC FILES

IN IT

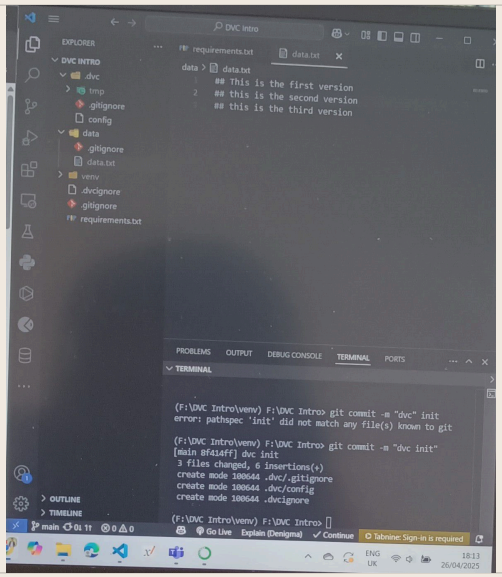


→ ABOVE PROVIDED FILES IN THE IMAGE NEED TO BE TRACKED

↳ .DVC IGNORE (IF YOU WANT IGNORE ANY FILES YOU CAN MENTION HERE SAME AS .GIT IGNORE

↳ COMMIT CODE MORE FREQUENTLY.

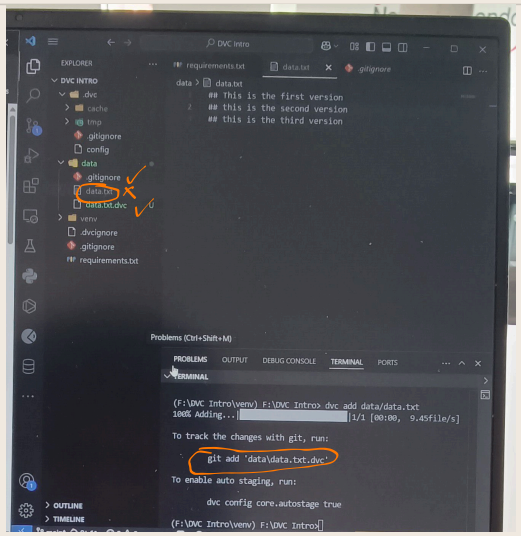
→ DVC ADD DATA/DATA.TXT (THIS IS THE FILE YOU NEED TO TRACK. (IT CAN BE CSV FILE, MODEL FILE IT CAN BE ANY FILE)

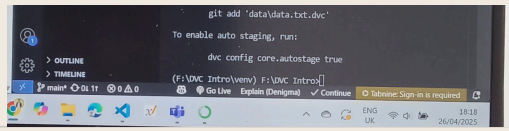


→ ONCE AFTER ENTERING THE ABOVE COMMAND

IT SAYS TO TRACK CHANGES WITH GIT RUN:

GIT ADD 'DATA\DATA.TXT'



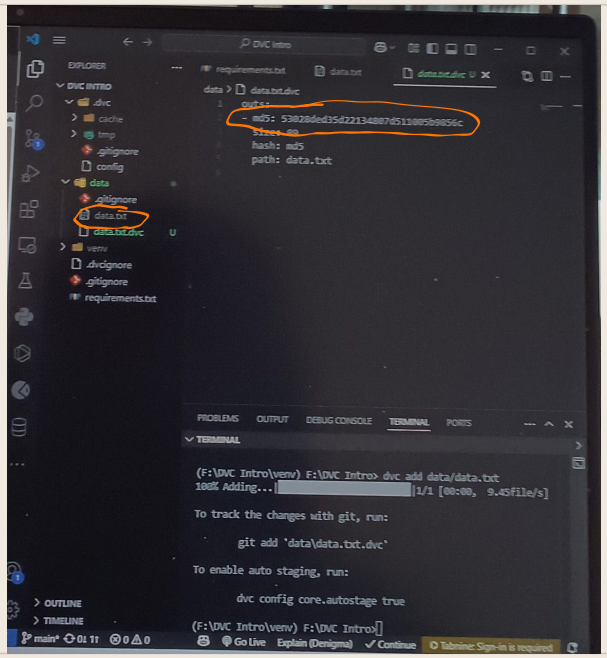


→ YOU CAN SEE .GITIGNORE IN DATA FOLDER WHERE

- GITIGNORE FILE IS CREATED SO THAT IT WILL

SAY GIT THAT IT DONT NEED TO TRACK DATA FILE

→ BUT .GITIGNORE & DATA.TEXT.DVC WILL BE TRACKED



→ MD5 IS THE HASH KEY OF DATA.TXT FILE

↳ WHAT EVER CONTENT IN THE ABOVE FILE. DVC IS CREATING HASH KEY (MD5)

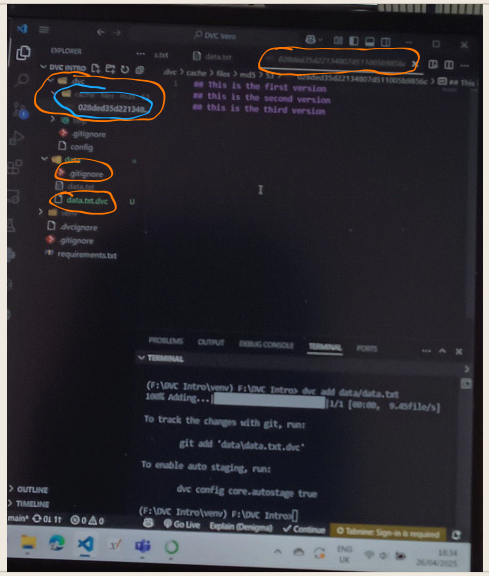
↳ THIS HASH MAPS TO A SPECIFIC DATA (IN OUR DATA.TXT FILE)

↳ IF YOU CHANGE THE CONTENT IN DATA.TXT FILE AGAIN HASH KEY (MD5) WILL BE GENERATED

HOW THIS IS GETTING MAPPED ?

IS IT STORED IN SPECIFIC LOCATION?

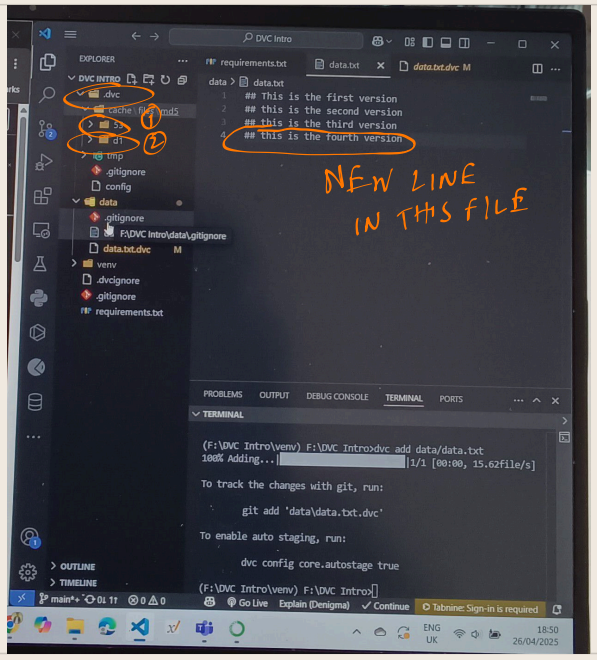
→ YOU CAN FIND THAT IN THE .DVC FOLDER INSIDE IT .DVC/CACHE



→ .GITIGNORE & DATA.TXT.DVC WILL BE TRACKED BY GIT

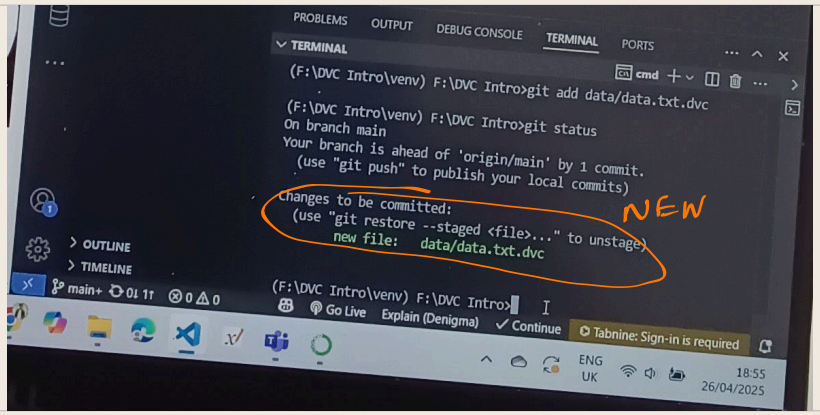
→ SO THAT YOU CAN STORE YOUR CACHE FOLDER ANYWHERE (GOOGLE CLOUD, AWS, AZURE ETC)

→ BECAUSE DATA.TXT.DVC HAS REFERENCE INFORMATION (MD5) OF (DATA.TXT (WHICH IS NOT GETTING TRACKED BY GIT)



→ ONCE YOU ADD THE DATA TO DATA.TXT.DVC & COMMIT WITH DVC ADD DATA/DATA.TXT COMMAND

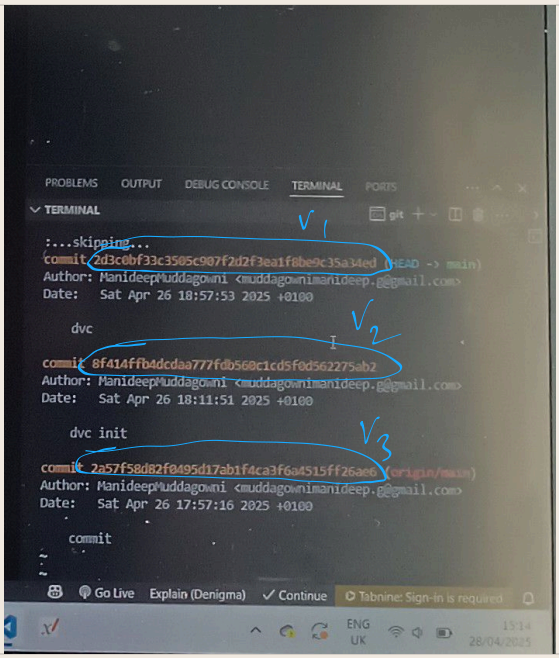
→ MD5 HASH KEY WILL BE CHANGED (YOU WILL HAVE ONE MORE FOLDER



→ NOW PUSH DATA/DATA.TXT.DVC TO GIT

→ ENTER GIT LOG IN CMD

→ YOU CAN SEE ALL THE VERISON COMMITS



→ IF YOU WANT SEE THE DATA VERSION1(V1) COPY AND PASTE IT IN NEW CMD

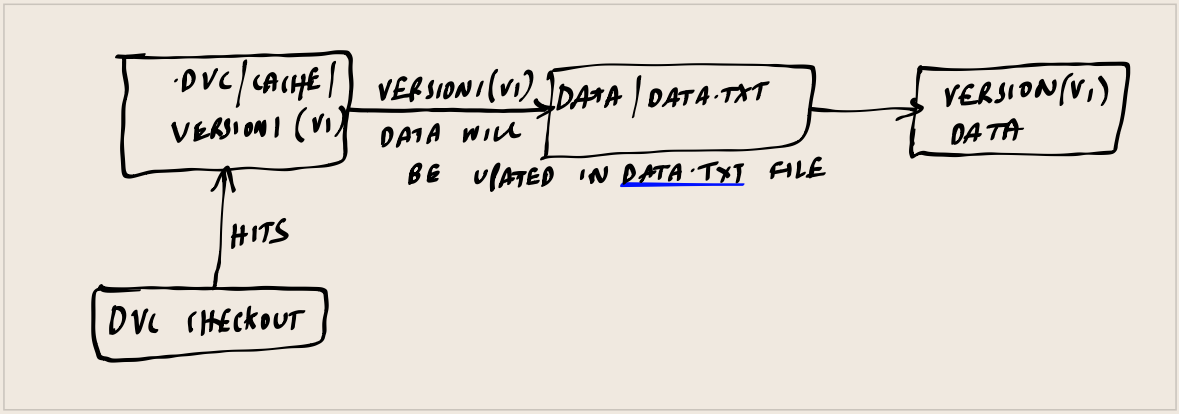
→ ACTIVATE VENV & ENTER GIT CHECKOUT (PASTE THE VERSION) WHICH IS VERSION1(V1)

→ NOW IT WILL BE SWITCHED TO VERSION1(V1) DATA IF YOU OBSERVE DATA.TXT.DVC FILE YOU CAN SEE THE UPDATED (MD5)

→ BUT DATA.TXT FILE WILL SHOW THE BEFORE

DATA UNTIL YOU DO DVC CHECKOUT

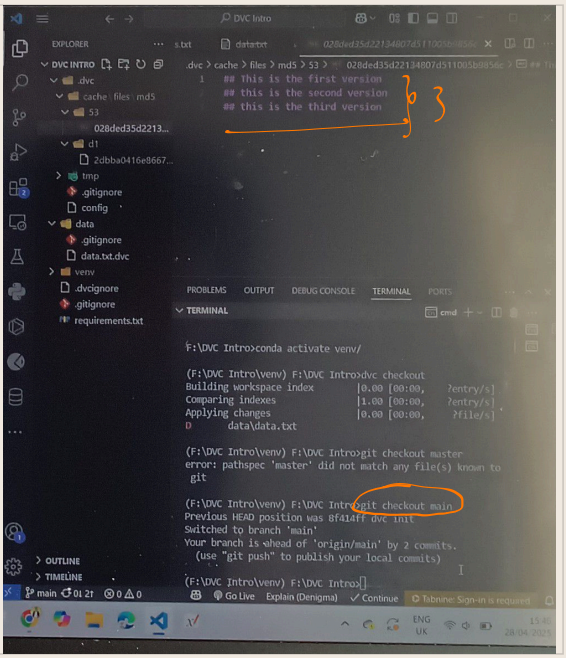
→ HERE ONCE YOU DO DVC CHECK OUT IT WILL GO BACK TO THE VERSION (V_1) DATA IN THE CACHE FOLDER & CHECK FOR THAT VERSION & DISPLAYS THE DATA IN DATA/DATA.TXT FILE

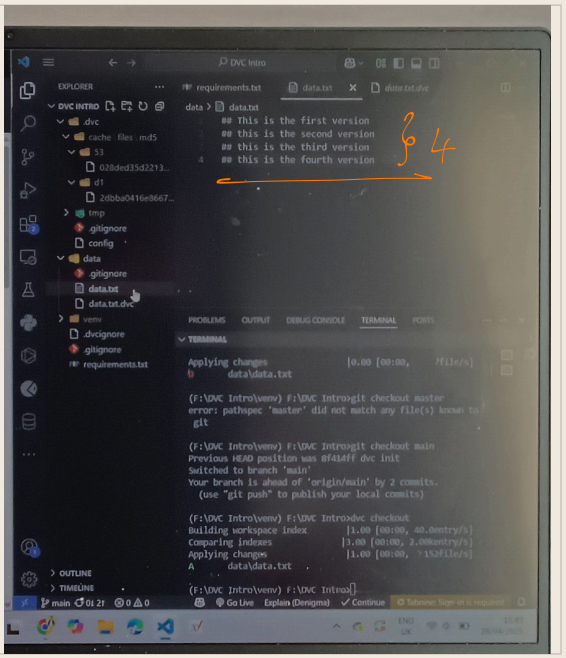


→ IF YOU JUST WANT TO GO BACK

→ ENTER GIT CHECKOUT YOUR_BRANCHNAME (MAIN)

→ AND ENTER DVC CHECKOUT


DVC Intro
EXPLORER
DVC INTRO
.dvc
cache files md5
53
028ded35d2213...
d1
2dbba0416e8667...
tmp
.gitignore
config
data
.gitignore
data.txt.dvc
venv
.dvcignore
.gitignore
requirements.txt
data.txt
This is the first version
this is the second version
this is the third version
PROBLEMS
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS
TERMINAL
F:\DVC Intro>conda activate venv/
(F:\DVC Intro\venv) F:\DVC Intro>dvc checkout
Building workspace index |0.00 [00:00, ?entry/s]
Comparing indexes |1.00 [00:00, ?entry/s]
Applying changes |0.00 [00:00, ?file/s]
D data\data.txt
(F:\DVC Intro\venv) F:\DVC Intro>git checkout master
error: pathspec 'master' did not match any file(s) known to git
(F:\DVC Intro\venv) F:\DVC Intro>git checkout main
Previous HEAD position was 8f414ff dvc init
Switched to branch 'main'
Your branch is ahead of 'origin/main' by 2 commits.
(use "git push" to publish your local commits)
(F:\DVC Intro\venv) F:\DVC Intro>
OUTLINE
TIMELINE
main
Go Live
Explain (Denigma)
Continue
Tabnine: Sign-in is required
3


DVC Intro
EXPLORER
DVC INTRO
.dvc
cache files md5
53
028ded35d2213...
d1
2dbba0416e8667...
tmp
.gitignore
config
data
.gitignore
data.txt
data.txt.dvc
venv
.dvcignore
.gitignore
requirements.txt
data.txt
This is the first version
this is the second version
this is the third version
this is the fourth version
PROBLEMS
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS
TERMINAL
Applying changes |0.00 [00:00, ?file/s]
D data\data.txt
(F:\DVC Intro\venv) F:\DVC Intro>git checkout master
error: pathspec 'master' did not match any file(s) known to git
(F:\DVC Intro\venv) F:\DVC Intro>git checkout main
Previous HEAD position was 8f414ff dvc init
Switched to branch 'main'
Your branch is ahead of 'origin/main' by 2 commits.
(use "git push" to publish your local commits)
(F:\DVC Intro\venv) F:\DVC Intro>dvc checkout
Building workspace index |1.00 [00:00, 40.0entry/s]
Comparing indexes |3.00 [00:00, 2.00kentry/s]
Applying changes |1.00 [00:00, ?15?file/s]
A data\data.txt
(F:\DVC Intro\venv) F:\DVC Intro>
OUTLINE
TIMELINE
main
Go Live
Explain (Denigma)
Continue
4