# Assignment 4 Submission

```
[Name :] Immadisetty Anvith ( 2302536 )
```

## Table of Contents



---

## Step 1: Log in to VM and Create Remote Storage Directory

To begin, log in to the virtual machine where the assignment will be set up. Use the `ssh` command to access the server, and then create a directory for storing remote data using the `mkdir` command. This step ensures that the necessary environment is prepared for the subsequent tasks.

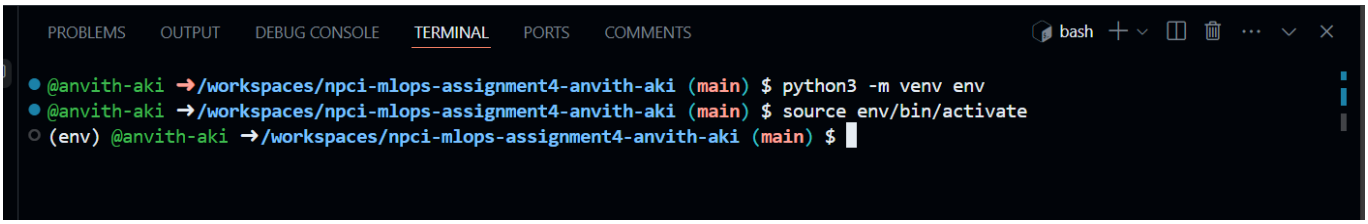
```
ssh anvit@npci-sandbox.talentsprint.com
mkdir -p /home/anvit/assignment-4
```

---

## Step 2: Set Up Python Virtual Environment

Setting up a Python virtual environment is critical for managing dependencies in an isolated environment. The following commands create and activate a virtual environment using Python's built-in `venv` module.

```
python3 -m venv env
source env/bin/activate
```




---

## Step 3: Install DVC and Initialize Repository

To enable version control for your data, install the Data Version Control (DVC) tool and initialize a DVC repository. These commands also include steps to track the initialization changes with Git.

```
pip install dvc==3.55.2 dvc-ssh==4.1.1 asyncssh==2.18.0
dvc init
git add .dvc .dvcignore
git commit -m "Initialize DVC"
```

```
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ pip install dvc==3.55.2 dvc-ssh==4.1.1 asyncssh==2.18.0
Collecting dvc==3.55.2
```

```
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ dvc init
Initialized DVC repository.

You can now commit the changes to git.

+---------------------------------------------------------------------+
|                                                                     |
|        DVC has enabled anonymous aggregate usage analytics.         |
|     Read the analytics documentation (and how to opt-out) here:     |
|             <https://dvc.org/doc/user-guide/analytics>              |
|                                                                     |
+---------------------------------------------------------------------+

What's next?
------------
- Check out the documentation: <https://dvc.org/doc>
- Get help and share ideas: <https://dvc.org/chat>
- Star us on GitHub: <https://github.com/iterative/dvc>
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $
```

---

## Step 4: Configure Remote Storage

Configure a remote storage location for DVC to store large data files. This setup uses SSH as the remote storage protocol. Replace `<your-vm-password>` with the actual password of your virtual machine.

```
dvc remote add -d myremote ssh://anvit@npci-
sandbox.talentsprint.com:22/home/anvit/assignment-4
dvc remote modify --local myremote password <your-vm-password>
git add .dvc/config
git commit -m "Remote storage configured"
git push
```

```
positional arguments:
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ dvc remote add -d myremote ssh://anvit@npci-sandbox.talentsprint
.com:22/home/anvit/assignment-4
Setting 'myremote' as a default remote.
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ dvc remote modify --local myremote password se5ooK8a
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ cat .dvc
.dvc/       .dvcignore
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ cat .dvc/config
[core]
    remote = myremote
['remote "myremote"']
    url = ssh://anvit@npci-sandbox.talentsprint.com:22/home/anvit/assignment-4
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ cat .dvc/config.local
['remote "myremote"']
    password = se5ooK8a
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $
```

```
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git add .dvc/config
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git commit -m "Remote storage configured"
[main b5a3072] Remote storage configured
 1 file changed, 4 insertions(+)
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git push
Enumerating objects: 7, done.
Counting objects: 100% (7/7), done.
Delta compression using up to 2 threads
Compressing objects: 100% (4/4), done.
Writing objects: 100% (4/4), 473 bytes | 473.00 KiB/s, done.
Total 4 (delta 1), reused 0 (delta 0), pack-reused 0 (from 0)
remote: Resolving deltas: 100% (1/1), completed with 1 local object.
To https://github.com/TSChallenges/npci-mlops-assignment4-anvith-aki
   3156e8a..b5a3072  main -> main
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $
```

---

## Step 5: Track Main Data Using DVC

The main dataset is tracked using DVC to enable versioning and collaboration. After adding the dataset, commit the changes to Git and push them to both the remote storage and the Git repository.

```
dvc add data/data_main.csv
git add data/data_main.csv.dvc data/.gitignore
git commit -m "Add main dataset to DVC"
dvc push
git tag v1
git push origin v1
```

```
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ dvc add data/data_main.csv
csv.dvc data/.gitignore
git commit -m "Add main dataset to DVC"
dvc push
git tag v1
100% Adding...|                                                                                |1/1 [00:00, 43.35file/s]

To track the changes with git, run:

        git add data/data_main.csv.dvc

To enable auto staging, run:

        dvc config core.autostage true
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git add data/data_main.csv.dvc data/.gitignore
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git commit -m "Add main dataset to DVC"
On branch main
Your branch is ahead of 'origin/main' by 2 commits.
  (use "git push" to publish your local commits)

nothing to commit, working tree clean
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ dvc push
Collecting                                                                                     |0.00 [00:00,    ?entry/s]
Pushing
Everything is up to date.
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git tag v1
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git push origin v1
```

---

## Step 6: Add Monthly Data Progressively

To update the dataset with new monthly data, run the data aggregator script, track the updated data with DVC, and push the changes to the remote storage and Git repository. The example below demonstrates the steps for updating with new data.

![Note :] Use below command to install pandas as the `data_aggregator.py` uses pandas to aggregate the data.

```
pip install pandas
```

Now, use below commands to aggregate data and tag/update dvc.

```
python data_aggregator.py
dvc add data/data_main.csv
git add data/data_main.csv.dvc
git commit -m "Update main dataset with new monthly data"
dvc push
git tag v2
git push origin v2
```

```
● (env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ python data_aggregator.py
● (env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ dvc add data/data_main.csv
  csv.dvc
  git commit -m "Update main dataset with new monthly data"
  dvc push
  git tag v2
  100% Adding...|                                                                                    |1/1 [00:00, 40.06file/s]

  To track the changes with git, run:

          git add data/data_main.csv.dvc

  To enable auto staging, run:

          dvc config core.autostage true
● (env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git add data/data_main.csv.dvc
● (env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git commit -m "Update main dataset with new monthly data"
  [main e338a6e] Update main dataset with new monthly data
   1 file changed, 2 insertions(+), 2 deletions(-)
● (env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ dvc push
  Collecting                                                                                          |1.00 [00:00,  202entry/s]
  Pushing
  1 file pushed
● (env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git tag v2
● (env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git push origin v2
  Enumerating objects: 11, done.
  Counting objects: 100% (11/11), done.
  Delta compression using up to 2 threads
  Compressing objects: 100% (8/8), done.
  Writing objects: 100% (9/9), 920 bytes | 920.00 KiB/s, done.
  Total 9 (delta 2), reused 0 (delta 0), pack-reused 0 (from 0)
  remote: Resolving deltas: 100% (2/2), completed with 1 local object.
  To https://github.com/TSChallenges/npci-mlops-assignment4-anvith-aki
   * [new tag]         v2 -> v2
○ (env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $
```

Repeat the above steps for additional monthly data updates. For example, for Month 3:

```
python data_aggregator.py
dvc add data/data_main.csv
```

```
git add data/data_main.csv.dvc
git commit -m "Update main dataset with new monthly data"
dvc push
git tag v3
git push origin v3
```

data_aggregator.py M

data_aggregator.py

```python
import pandas as pd

dataset1 = pd.read_csv("data/data_main.csv")     # Dataset1 is the main data

# Dataset2 is the data you want to add
dataset2 = pd.read_csv("month3_data.csv")        # Change the path to add month3_data.csv

# Append rows of dataset2 to dataset1
dataset1 = pd.concat([dataset1, dataset2],axis=0, ignore_index=True)

# Overwrite data_main.csv file
dataset1.to_csv("data/data_main.csv", index=False)

```

```
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ python data_aggregator.py
sv
git add data/data_main.csv.dvc
git commit -m "Update main dataset with new monthly data"
dvc push
git tag v3
git push origin v3(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ dvc add data/data_main.csv
100% Adding...|                                                                                |1/1 [00:00, 39.48file/s]

To track the changes with git, run:

        git add data/data_main.csv.dvc

To enable auto staging, run:

        dvc config core.autostage true
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git add data/data_main.csv.dvc
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git commit -m "Update main dataset with new monthly data"
[main de65a93] Update main dataset with new monthly data
 1 file changed, 2 insertions(+), 2 deletions(-)
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ dvc push
Collecting                                                                                     |1.00 [00:00,  123entry/s]
Pushing
1 file pushed
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git tag v3
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $ git push origin v3
Enumerating objects: 7, done.
Counting objects: 100% (7/7), done.
Delta compression using up to 2 threads
Compressing objects: 100% (4/4), done.
Writing objects: 100% (4/4), 468 bytes | 468.00 KiB/s, done.
Total 4 (delta 1), reused 0 (delta 0), pack-reused 0 (from 0)
remote: Resolving deltas: 100% (1/1), completed with 1 local object.
To https://github.com/TSChallenges/npci-mlops-assignment4-anvith-aki
 * [new tag]         v3 -> v3
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (main) $
```

Screenshot for the Final files in VM

```
anvit@localhost:~/assignment-4$ ls -l -R files/
files/:
total 4
drwxrwxr-x 5 anvit anvit 4096 Jan 11 12:04 md5

files/md5:
total 12
drwxrwxr-x 2 anvit anvit 4096 Jan 11 12:02 51
drwxrwxr-x 2 anvit anvit 4096 Jan 11 12:00 65
drwxrwxr-x 2 anvit anvit 4096 Jan 11 12:04 d5

files/md5/51:
total 4
-rw-rw-r-- 1 anvit anvit 3400 Jan 11 12:02 dac83d777e5d561546ab863299ac55

files/md5/65:
total 4
-rw-rw-r-- 1 anvit anvit 1761 Jan 11 12:00 0bedbddcfc3e521f78206d283ffbc0

files/md5/d5:
total 8
-rw-rw-r-- 1 anvit anvit 5048 Jan 11 12:04 d3a2ed56d24d8b6609fcae4dfdfd6a
anvit@localhost:~/assignment-4$
```

---

## Step 7: Checkout and Verify Specific Data Versions

To retrieve a specific version of the dataset, clone the repository, checkout to the desired Git tag, and pull the associated data files using DVC. This process ensures the correct dataset version is retrieved for verification or further processing.

1. Clone the repository:

Replace the `<git-repository-url>` and `<repository-directory>` with your actual values

```
# git clone <git-repository-url>
git clone https://github.com/TSChallenges/npci-mlops-assignment4-anvith-aki.git

# cd <repository-directory>
cd npci-mlops-assignment4-anvith-aki
```

2. Configure a remote storage location password using below command for security purpose not storing actual credentials in git. Replace `<your-vm-password>` with the actual password of your virtual machine.

```
# dvc remote modify --local myremote password <your-vm-password>

# for-user: anvit
# vm-password: se5ooK8a
```

```
dvc remote modify --local myremote password se5ooK8a
```

3. Checkout to the desired tag (e.g., `v1`):

```
git checkout v1
```

4. Pull the data files associated with the tag:

```
dvc pull
```

5. Verify the dataset version:

Check the dataset content or metadata to confirm the retrieved version matches expectations.

```
ls data/
cat data/data_main.csv
```

By following these steps, you can manage and verify specific data versions efficiently.

```
git commit -m "Update main dataset with new monthly data"
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (71139af) $ git checkout v1
M       data_aggregator.py
Previous HEAD position was 71139af Update main dataset with new monthly data
HEAD is now at 8d53afb Add main dataset to DVC
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (8d53afb) $ dvc pull
Collecting                                                                                                    |0.00 [00:00,    ?entry/s]
Fetching
Building workspace index                                                                                      |2.00 [00:00,  791entry/s]
Comparing indexes                                                                                            |3.00 [00:00, 1.37kentry/s]
Applying changes                                                                                              |1.00 [00:00,   346file/s]
M       data/data_main.csv
1 file modified
(env) @anvith-aki ➜/workspaces/npci-mlops-assignment4-anvith-aki (8d53afb) $ cat data/data_main.csv
age,job,marital,education,balance,default,housing,loan,contact,day,month,duration,campaign,pdays,previous,poutcome,y
32,technician,married,secondary,1500,no,yes,no,cellular,12,may,300,2,-1,0,unknown,no
45,services,single,tertiary,4000,no,no,yes,cellular,15,jun,250,1,-1,0,unknown,no
29,admin,married,secondary,2000,no,no,no,unknown,20,jul,150,3,-1,0,unknown,yes
50,management,divorced,tertiary,5000,no,yes,no,cellular,3,aug,450,2,-1,0,unknown,no
35,blue-collar,married,primary,1200,no,no,no,unknown,7,sep,100,1,-1,0,unknown,yes
31,services,married,primary,1060,no,yes,yes,cellular,7,jul,130,3,-1,0,unknown,no
32,admin,single,secondary,1070,no,no,no,unknown,8,aug,135,4,-1,1,unknown,no
33,management,divorced,tertiary,1080,no,yes,no,cellular,9,sep,140,1,-1,0,unknown,yes
34,blue-collar,married,primary,1090,no,no,yes,unknown,10,oct,145,2,-1,1,unknown,no
35,technician,single,secondary,1100,no,yes,no,cellular,11,nov,150,3,-1,0,unknown,no
36,services,divorced,tertiary,1110,no,no,no,unknown,12,dec,155,4,-1,1,unknown,no
37,admin,married,primary,1120,no,yes,yes,cellular,13,jan,160,1,-1,0,unknown,yes
38,management,single,secondary,1130,no,no,no,unknown,14,feb,165,2,-1,1,unknown,no
39,blue-collar,divorced,tertiary,1140,no,yes,no,cellular,15,mar,170,3,-1,0,unknown,no
40,technician,married,primary,1150,no,no,yes,unknown,16,apr,175,4,-1,1,unknown,no
41,services,single,secondary,1160,no,yes,no,cellular,17,may,180,1,-1,0,unknown,yes
42,admin,divorced,tertiary,1170,no,no,no,unknown,18,jun,185,2,-1,1,unknown,no
43,management,married,primary,1180,no,yes,yes,cellular,19,jul,190,3,-1,0,unknown,no
```