Big Data Laboratory – Assignment 4 Report

GitHub Repository: https://github.com/Ramesh-031102/Assignment_4_BDL.git

params.yaml basically has parameters **year** and **nlocs**.

download.py file will take these params as input and will download the files in that year based on the params.

prepare.py which takes these download files as input and extract the monthly aggregates which will be ground truth values and also it returns the list of fields.

process.py file finds the monthly aggregates (predicted values).

evaluate.py file basically finds the r2_score between the ground truth values and predicted values.

Stages of the pipeline:

First stage - Command: dvc stage add -- run -v -f -n download -p year -p nlocs -o downloaded_files/ python download.py

Second stage - Command: dvc stage add -- run -v -f -n prepare -p year -p nlocs -d downloaded_files/ -o groundtruth/ -o fieldlist/ python prepare.py

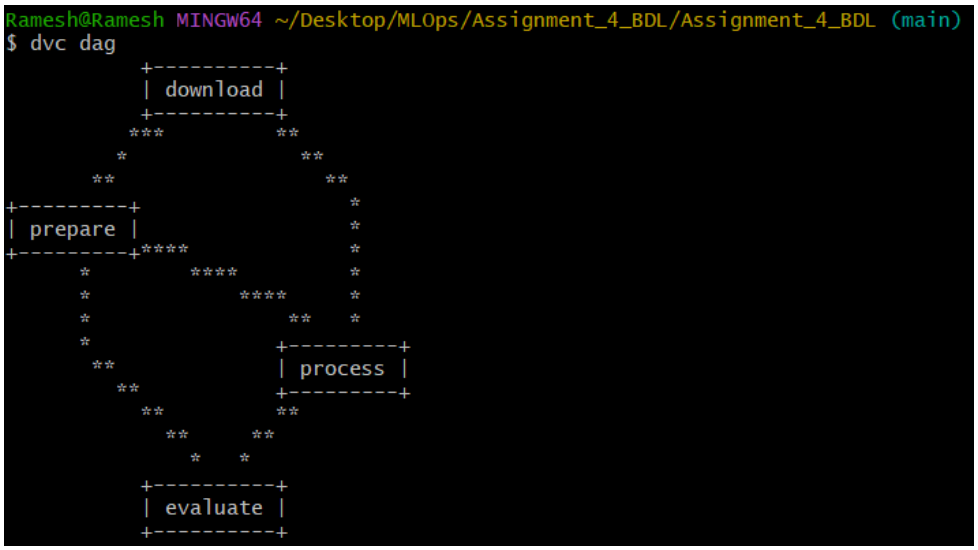
Third stage - Command: dvc stage add -- run -v -f -n process -p year -p nlocs -d downloaded_files/ -d fieldlist/ -o predicted/ python process.py

Fourth stage - Command: dvc stage add –run -v -f -n evaluate -p year -p nlocs -d predicted/ -d groundtruth/ -d fieldlist/ -o r2score/ python evaluate.py

By running all these stages, we had created the pipeline.

While running all these commands, **dvc.yaml** is created in the folder and it is also updated after each stage. **dvc.lock** is also created and updated .

To visualise the DAG of the pipeline, we must run command : **dvc dag**



dvc repro will run the pipeline again.

```
Ramesh@Ramesh MINGW64 ~/Desktop/MLOps/Assignment_4_BDL/Assignment_4_BDL (main)
$ dvc repro
Running stage 'download':
> python download.py
{'nlocs': 10, 'year': 2003}
Downloaded: 99999994996.csv
Downloaded: 99999994995.csv
Downloaded: 99999994994.csv
Downloaded: 99999994993.csv
Downloaded: 99999994992.csv
Downloaded: 99999994991.csv
Downloaded: 99999994989.csv
Downloaded: 99999994988.csv
Downloaded: 99999994985.csv
Downloaded: 99999994978.csv
Updating lock file 'dvc.lock'
```

dvc params diff will compare the experiments

```
Path         Param    HEAD    workspace
params.yaml  n_locs   15      25
```

After changing the nlocs multiple times and running multiple times, **dvc exp show** will give the list of runs