DVC Tutorial

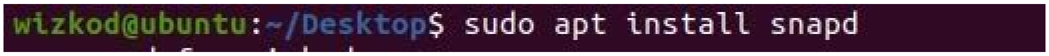
Objective : How is data filtered, transformed, or used to train ML models? We will use DVC mechanisms to capture data pipelines — a series of data processes that produce a final result for Data Pipelines (Versioning large data files and directories).

You can simply follow this well explained tutorial : Get Started: Data Pipelines | Data Version Control · DVC

Get Started on Ubuntu

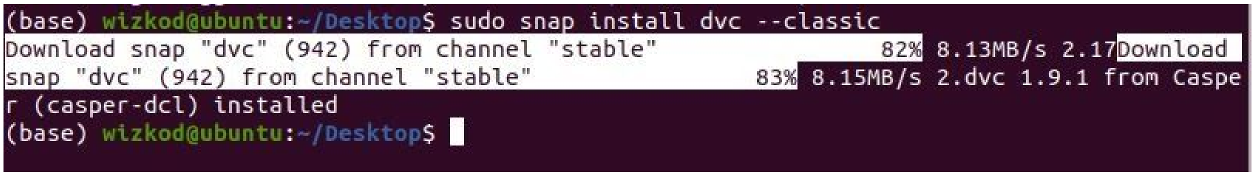
1. Install DVC

   1.install snapd first



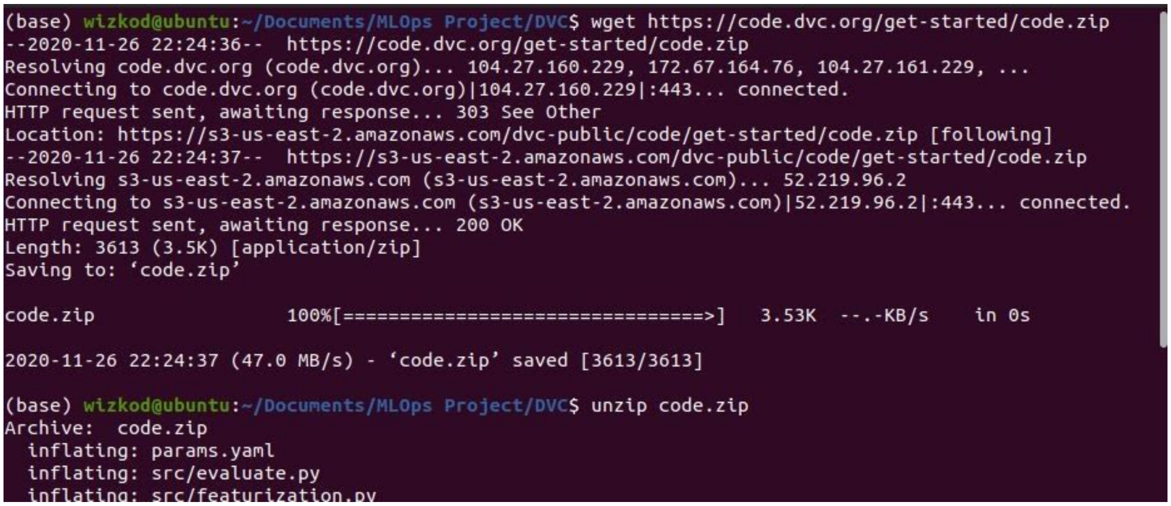
```
wizkod@ubuntu:~/Desktop$ sudo apt install snapd
```

   2.install dvc



```
(base) wizkod@ubuntu:~/Desktop$ sudo snap install dvc --classic
Download snap "dvc" (942) from channel "stable"                82% 8.13MB/s 2.17Download
snap "dvc" (942) from channel "stable"                 83% 8.15MB/s 2.dvc 1.9.1 from Caspe
r (casper-dcl) installed
(base) wizkod@ubuntu:~/Desktop$ 
```

2. Get a the example code on DVC and unzip the file



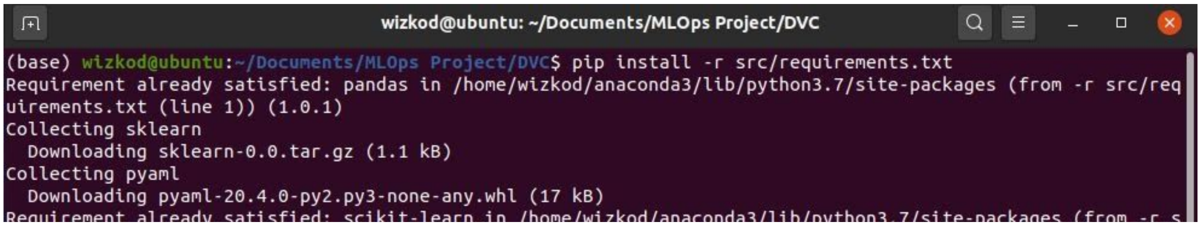
```
(base) wizkod@ubuntu:~/Documents/MLOps Project/DVC$ wget https://code.dvc.org/get-started/code.zip
--2020-11-26 22:24:36--  https://code.dvc.org/get-started/code.zip
Resolving code.dvc.org (code.dvc.org)... 104.27.160.229, 172.67.164.76, 104.27.161.229, ...
Connecting to code.dvc.org (code.dvc.org)|104.27.160.229|:443... connected.
HTTP request sent, awaiting response... 303 See Other
Location: https://s3-us-east-2.amazonaws.com/dvc-public/code/get-started/code.zip [following]
--2020-11-26 22:24:37--  https://s3-us-east-2.amazonaws.com/dvc-public/code/get-started/code.zip
Resolving s3-us-east-2.amazonaws.com (s3-us-east-2.amazonaws.com)... 52.219.96.2
Connecting to s3-us-east-2.amazonaws.com (s3-us-east-2.amazonaws.com)|52.219.96.2|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3613 (3.5K) [application/zip]
Saving to: ‘code.zip’

code.zip                100%[================================>]   3.53K  --.-KB/s    in 0s

2020-11-26 22:24:37 (47.0 MB/s) - ‘code.zip’ saved [3613/3613]

(base) wizkod@ubuntu:~/Documents/MLOps Project/DVC$ unzip code.zip
Archive:  code.zip
  inflating: params.yaml
  inflating: src/evaluate.py
  inflating: src/featurization.py
```

3. Create a virtual environment first. take some minutes



```
(base) wizkod@ubuntu:~/Documents/MLOps Project/DVC$ pip install -r src/requirements.txt
Requirement already satisfied: pandas in /home/wizkod/anaconda3/lib/python3.7/site-packages (from -r src/req
uirements.txt (line 1)) (1.0.1)
Collecting sklearn
  Downloading sklearn-0.0.tar.gz (1.1 kB)
Collecting pyaml
  Downloading pyaml-20.4.0-py2.py3-none-any.whl (17 kB)
Requirement already satisfied: scikit-learn in /home/wizkod/anaconda3/lib/python3.7/site-packages (from -r s
```

4. Use dvc run to create stages. These represent processes (source code tracked with Git) that form the steps of a pipeline. Stages also connect code to its data input and output. Let's transform a Python script into a stage:

3. Install docker engine

after setup docker repository you can install docker with this command

```
$ sudo apt-get install docker-ce docker-ce-cli containerd.io
```

This tutorial shows step by step how to install docker engine on Ubuntu.

4. Clone mlflow projects examples