Storing a few versions of a 5GB file in a data science project

SciPy 2019

DVC

About the Speaker

Aman Sharma

Junior, Indian Institute of Technology Roorkee

algomaster99

- Web Development, Version Control System, Open Source <3
- Selected for Google Summer of Code 2019
- Part-time Collaborator

DVC (Data Version Control)

Built on: Python

Uses: Versioning a data science project

Speciality: Open-Source, Responsive community

Stargazers: 4K + (GitHub)

Contributing to DVC

#1963: Add `dvc version` command +62 -0

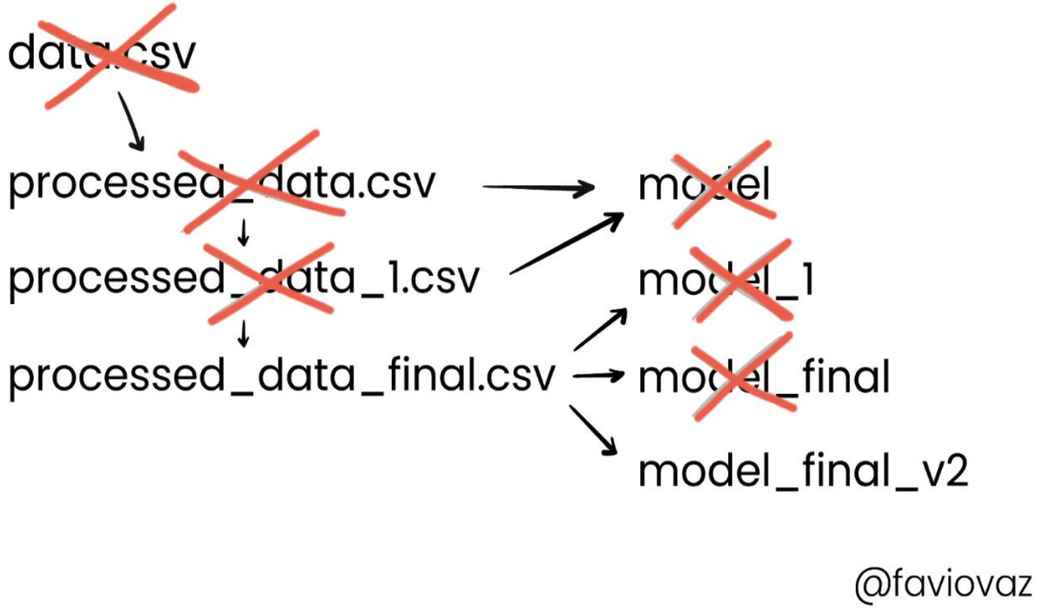
```
usage: dvc version [-h] [-q | -v]
```

```
DVC Version: 0.70.0
Python Version: 3.6.8
```

Meet some data scientists

“Too many tools for software development but none for data science :(”

Versioning by copying



~~data.csv~~

~~processed_data.csv~~

~~processed_data_1.csv~~

processed_data_final.csv

~~model~~

~~model_1~~

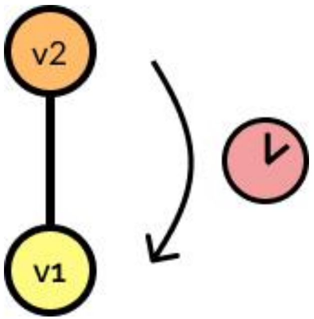
~~model_final~~

model_final_v2

@faviovaz

Are we similar?

Git to the rescue?



Increased checkout time

>>> 2GB

Allowance: 2GB/file

LFS is supported on lesser platforms

Summarising the problems

Versioning large files	?
Storing and sharing large files	?
Increased time of executing version control commands (eg. `git checkout`)	?

Initializing DVC

> *Installation*

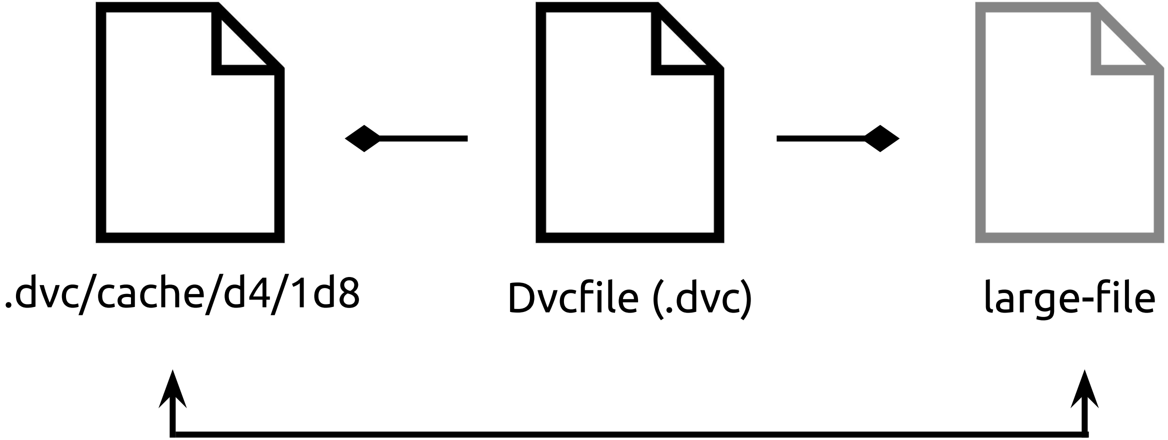
```
$ pip3 install dvc
$ conda install -c conda-forge dvc
```

> *Initialize*

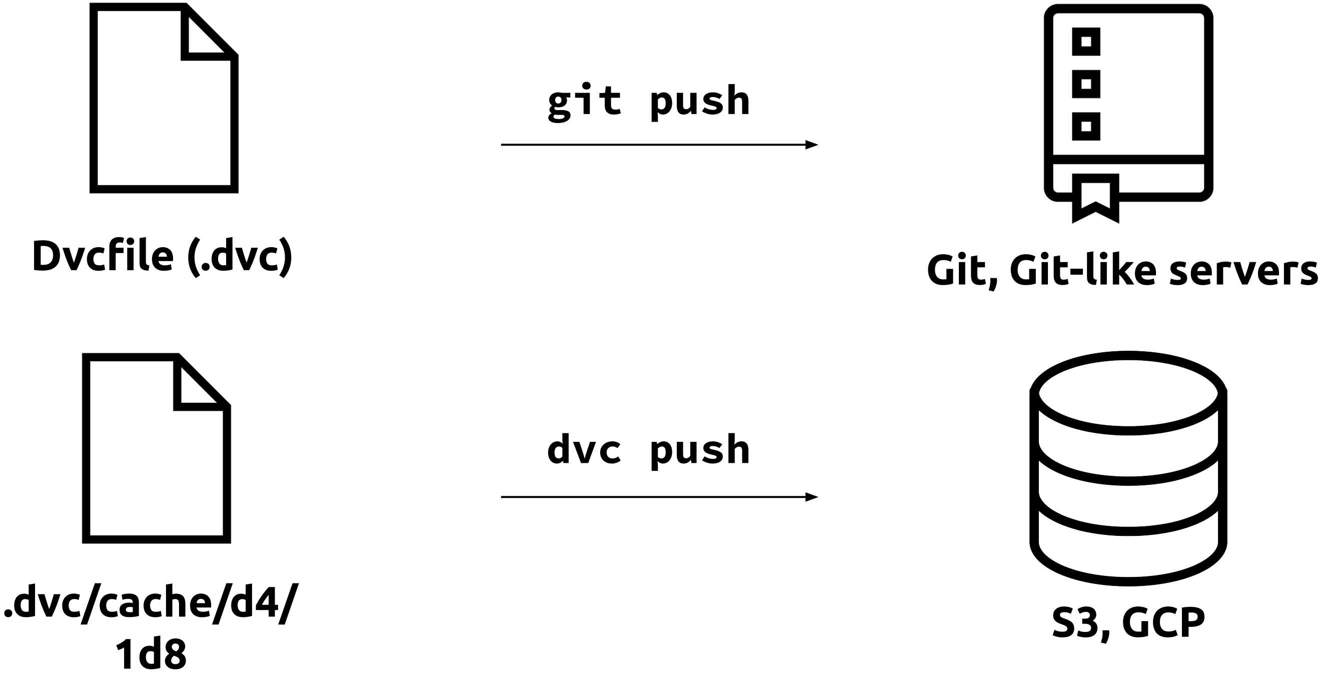
```
$ dvc init
```

Versioning a 5GB file

```
$ dvc add large-file
```



Sharing the file with others



Dvcfile (.dvc)

`git push`

Git, Git-like servers

.dvc/cache/d4/1d8

`dvc push`

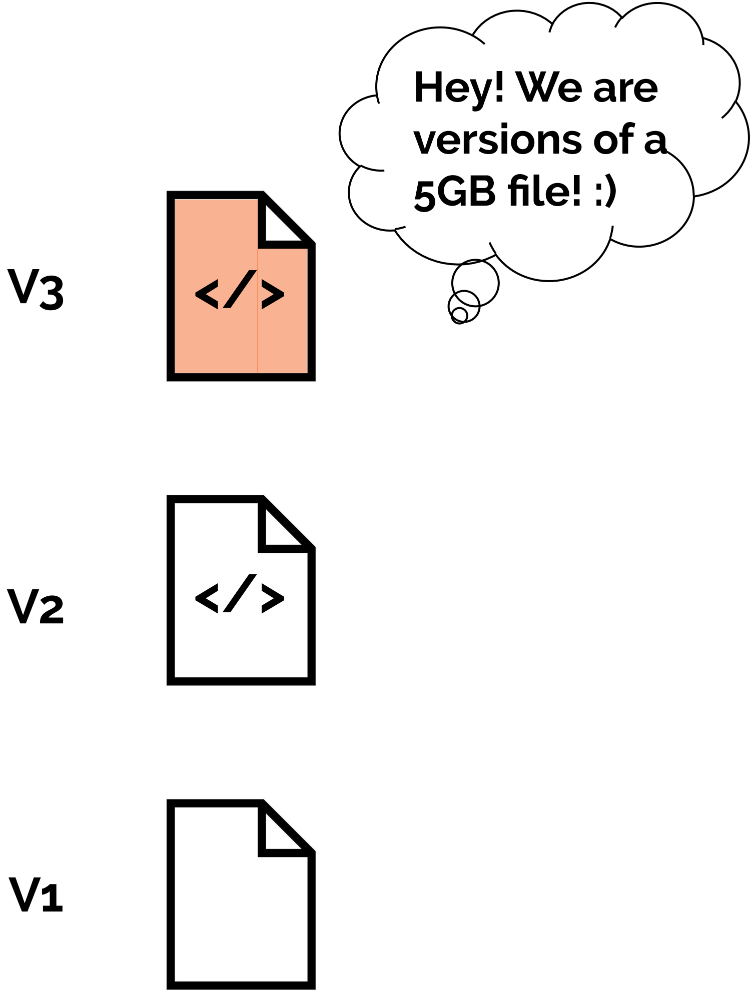
S3, GCP

What happens to ‘large-file’?

Git to large-file:

Summarising the solutions

Versioning large files	✓
Storing and sharing large files	✓
Increased time of executing version control commands (eg. `git checkout`)	✓



Hey! We are versions of a 5GB file! :)

V3

V2

V1

Thank you!

> *Questions*

Email: aman@iterative.ai

Community: dvc.org/chat

> *Actions*

Visit: dvc.org

Star: github.com/iterative/dvc