



Overview of “git for data” software

Data Versioning platform summary


Lee Tirrell

19 May 2020, Updated 13 Oct 2020

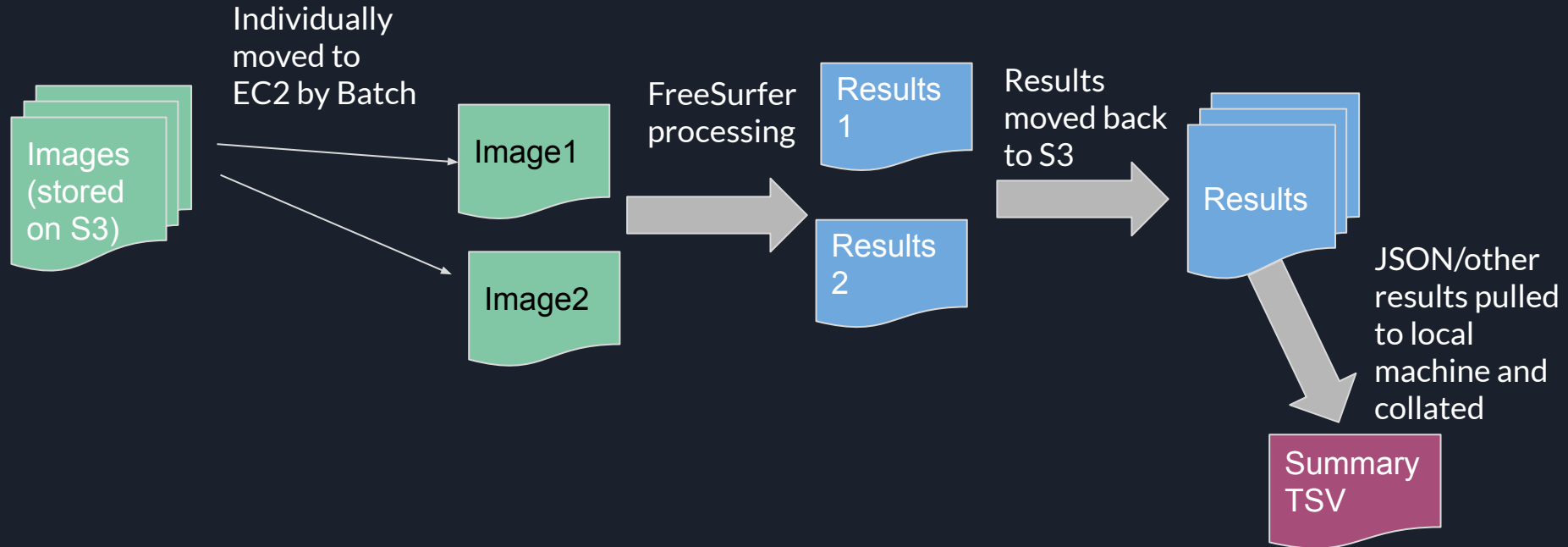


Why this overview?

- We want a clean way to manage data, and provide provenance of results, which should help with FDA/QMS requirements



Use case: Reference dataset of 1000s of images, to be processed with FreeSurfer-like code





General Requirements

- Must haves:
 - Versioned datasets
 - Match workflow and results with data used to run them
 - Private datasets and results
 - Data stored on S3, with transparent ability to access data by URI
 - Easy to use CLI
 - Self hosted service
 - Well documented deployment following best practices (Terraform, Helm, etc)
 - Access to public datasets, and capabilities to share select results
- Nice to haves (some may be out of scope)
 - Web interface for visualizing results
 - Jupyter interactivity
 - Sorting results by various metadata
 - Summary figures
 - Avoid vendor lockin
 - Logging and auditing (either forwarding to a server and/or web interface)



Service Providers

- Overview of different services, including:
 - Quilt
 - DVC
 - Pachyderm
 - Dolt
 - Bonus: “distributed data” services
 - IPFS
 - Qri - service built on IPFS
 - Dat
 - Cryptocurrency distributed file storage: StorJ, Sia, Filecoin
 - Other services (re)discovered later in my research, built with a scientific focus in mind, not discussed in further detail at this time:
 - GIN - distributed version control, flavoured for science
 - Built on Gogs (a simpler, GitLab/GitHub-like clone) to support git-annex for data storage
 - Doesn't seem like something useful to us (we'd need to spin up another GitLab-like repo for data, and git-annex isn't the most user-friendly)
 - Datalad
 - From brainhack folks, also built on git-annex. It always seems verbose/obtuse, but they built out docs more, so may be worth exploring again
 - Built on git-annex, more of a commandline driven tool on the level of DVC, worth comparing them

Quilt - Experiment faster by managing data like code

<https://quiltdata.com/>



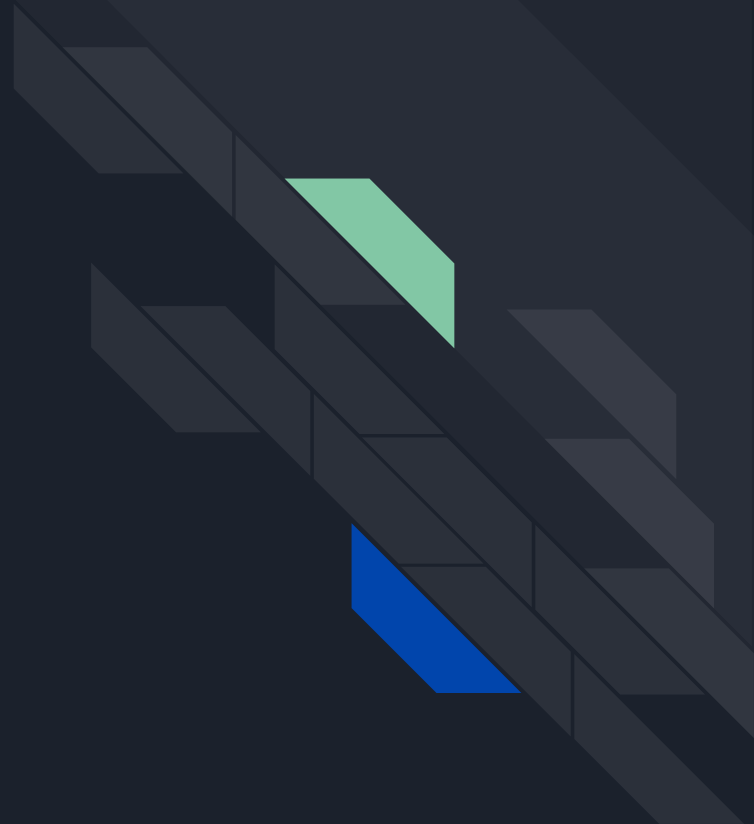


Features, pros, cons

- Built on top of an S3 bucket
- Paid backend (orchestrated by CloudFormation) for private buckets, free hosting of public data
- Displays Markdown, CSV, Jupyter, maybe PDF?, images, etc in browser
- Mark data using Packages, to commit a version (as in git)
- Overall seems like a good wrapper on top of S3 data versioning with better visualization,
 - How often is web visualization useful in our current workflows?
 - Would want to set up a trial with them to explore how to integrate into our current workflows

DVC - Open-source Version Control System for Machine Learning Projects

<https://dvc.org/>





Features, pros, cons

- More data science specific than other git-based tools for large files (git-annex or git-lfs)
- Can run on top of data stored in S3
- Need to slightly alter our processing workflow to include dvc files in repos
 - Small text file for each file/directory (depending on level of concern)

Pachyderm - Engineered to make data science

<https://www.pachyderm.com/>






Features, pros, cons

- Complete data science platform that runs on a Kubernetes cluster
- Seems a bit heavyweight for what we want
 - Can store data in S3, but doesn't seem transparent?
 - Need to run a K8s cluster at all times
 - Doesn't provide a "data explorer" except with Enterprise account (need to contact sales to see cost)
- Since it's a full service platform, includes a pipeline runner as well
 - Would be useful if we're running a lot of jobs, but our current use of GitLab CI and AWS Batch may be sufficient?
- Has a free online Hub, which may be worth exploring

Takeaways



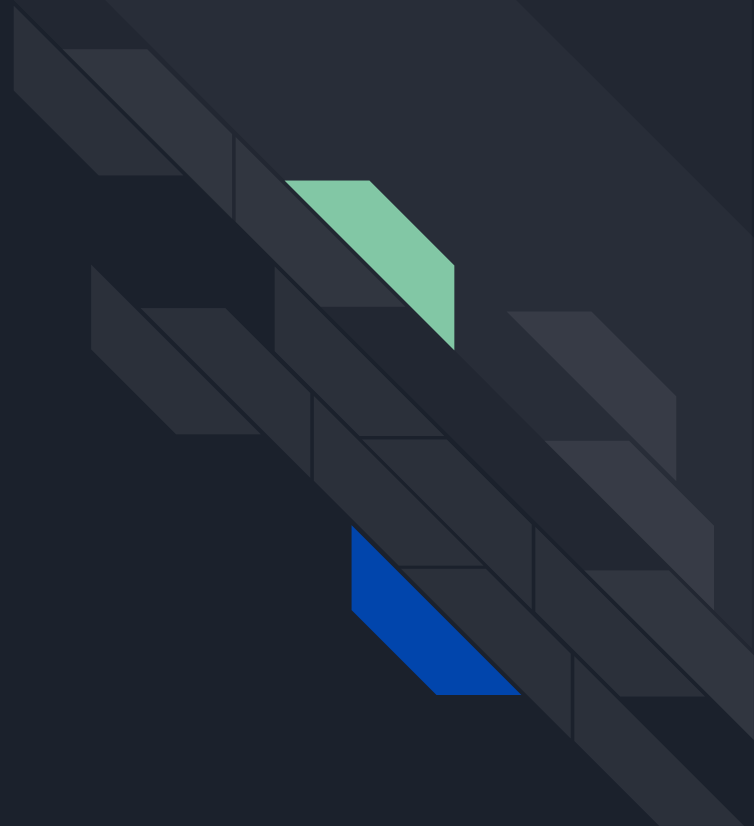
- 
- Quilt has a nice interface and would fit all our needs
 - Uses S3 directly, treats data as a separate repository from code
 - May be worth evaluating, could be useful if we wanted to QC a bunch of images or PDFs if there's quick previewing in browser
 - DVC seems like a great, simple, free solution... but would require us to change our workflows
 - Would potentially need to add directories of 100s/1000s of “fake” files to our repos
 - Pachyderm has features like DVC (versioning of data and workflows), but runs as a Kubernetes cluster, so requires a long running application. Also requires enterprise support for any visualization features
 - May be worth evaluating their Hub, to see if these features are useful for us
 - Could potentially replace our Batch usage, though this seems rare enough that a new service may not be worthwhile just for this
 - Further investigation of datalad may also be useful (on similar scope as DVC)

Cool, but not relevant for
now...



Dolt – It's Git for Data

<https://github.com/liquidata-inc/dolt>



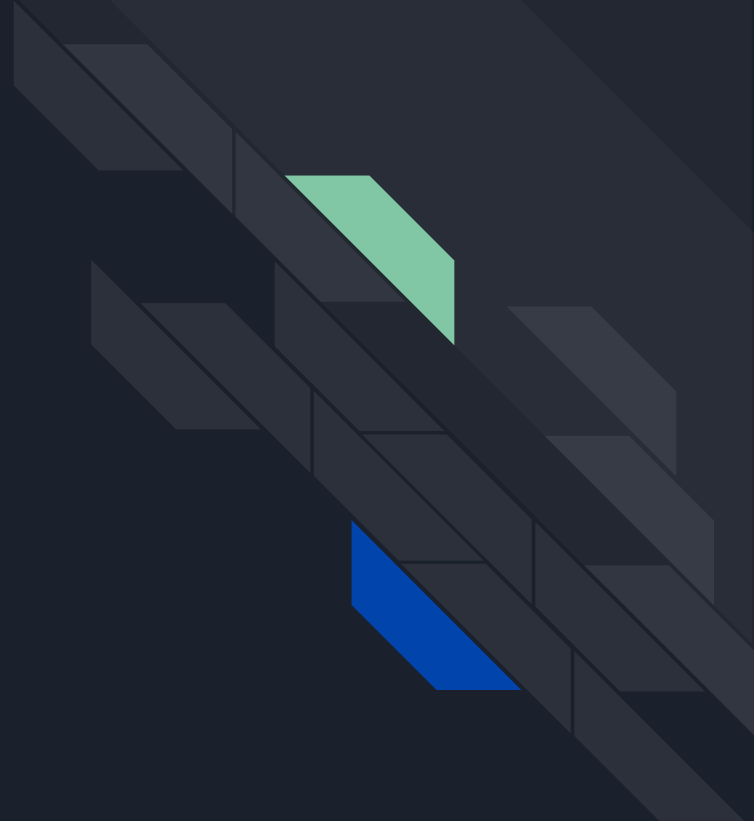


Features, pros, cons

- Version controlled relational database
- Provides [DoltHub](#), a github-like service for Dolt Tables/Views
- \$50/month to create private repos
- Doesn't fit our current "S3 filesystem" based approach of file storage - would need to use a relational database where we have a column pointing to S3 URI for images
- Has undocumented support for S3 remotes; may be able to use this to spin our own DoltHub-like situation?
- Doesn't seem to be fully implemented, and their site/docs really push towards their DoltHub product (which isn't really what we want)

IPFS - A peer-to-peer
hypermedia protocol
designed to make the
web faster, safer, and
more open.

<https://ipfs.io/>





Features, pros, cons

- Uses its own distributed network for storage, not S3 based (I'm not sure of its compatibility with S3)
- Interesting way to share/version data, and remove duplication, but may be too low level, and not mature enough of an ecosystem for our use case
- Heard of some academic interest in IPFS as potential good tech for sharing large datasets ~3 years ago. Not sure if anything has been implemented or used by our collaborators though
- Potential [small amounts of funding available](#) in 2021, for a proof of concept of medical imaging data sharing?

Qri – helps you clean,
version, organize, and
share datasets

<http://qri.io/>



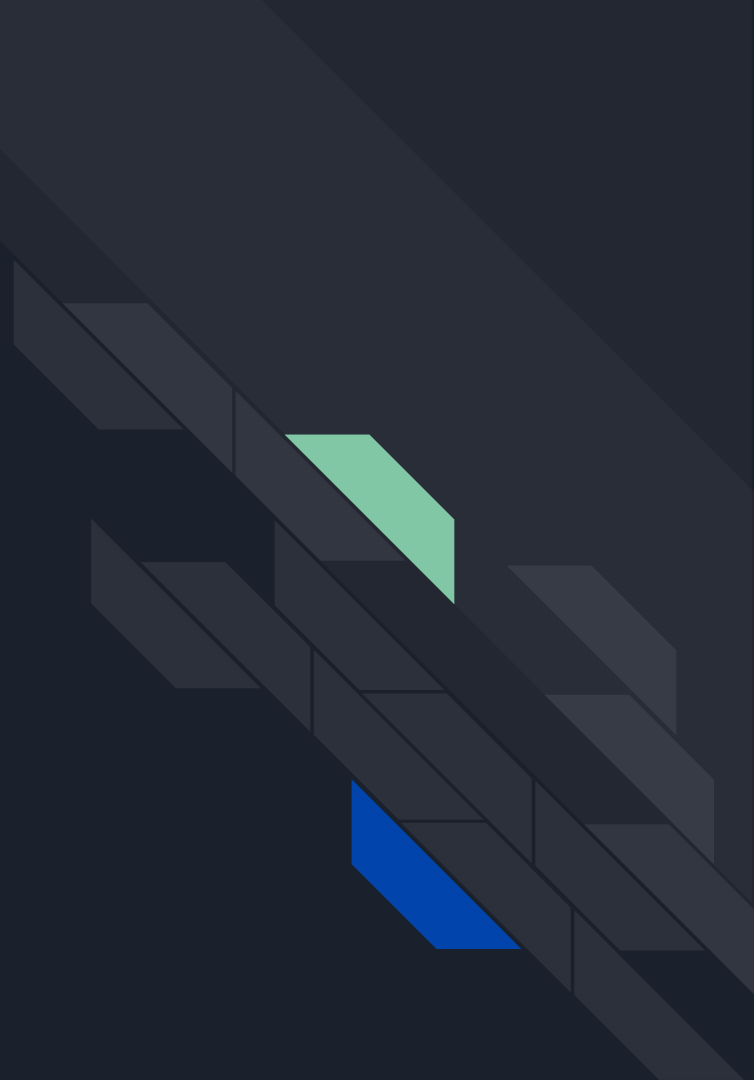


Features, pros, cons

- Built on IPFS
- Seems to have a lot of useful features, but only supports public datasets
- Their main Desktop app doesn't have a Linux download available (maybe possible to build from source?)

Dat - peer-to-peer sharing & live synchronization of files via command line

<https://github.com/datproject/dat>





Features, pros, cons

- Uses its own distributed network for storage, not S3 based (I'm not sure of its compatibility with S3)
- Designed to take “the best parts of Git, BitTorrent and Dropbox” to be able to automatically version datasets