Deduplicating Backup and Version Control for Big Binaries

Art Barnes, 2021-02-11

Agenda

- Why Deduplicating Backup
- Comparison of Deduplicating Software
- Getting Started with Duplicacy
- Git and Large Binary Files
- How Dupver Works
- An Example with Dupver

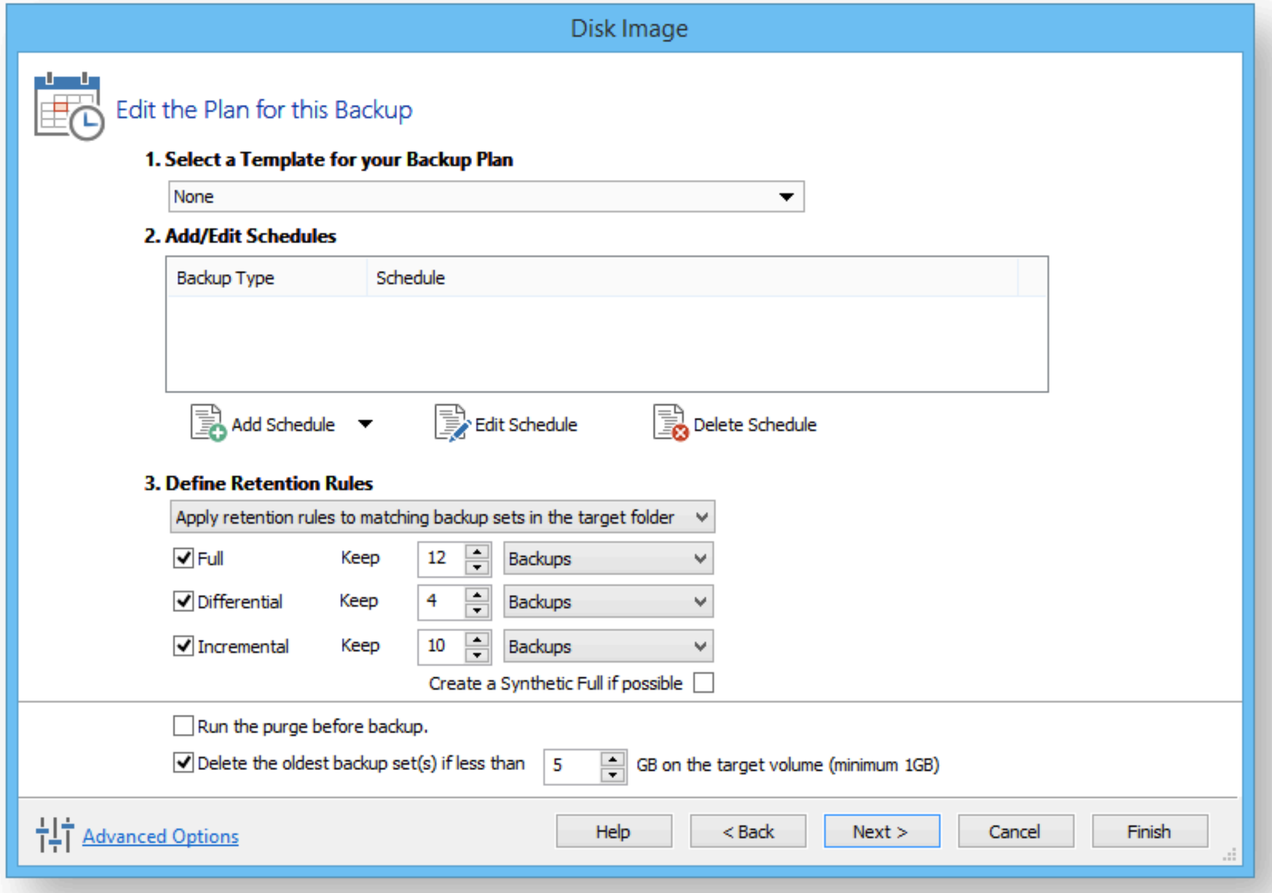
Art’s Guidelines on Backup

- Keep at least one copy of your files on a local filesystem
- Keep at least one offsite backup of your files
- Keep at least one offline backup of your files
- Keep a recent full image of your system
- Perform “thinning” on your backups

How Does Deduplicating Software Work

- Breaks up stream of files into unique chunks
- Chunks are concatenated into pack files
- Trees are indexes that map a chunk hash to a pack file and index within that pack
- Snapshots contain date stamp, a list of files along with mod time and file hash, and a list of chunk hashes
- If you delete a snapshot, the software keeps track of which chunks to keep

Why Deduplicating Backup



Traditional backup software (Duplicity, Macrium Reflect) usually uses differential backups where later backup snapshots depend on having earlier backups

Lots of snapshots can be slow & increase risk of problems restoring

Mitigate with Full/Diffential/Incremental scheme which allows “thinning” at cost of more data

https://knowledgebase.macrium.com/display/KNOW7/Creating+a+backup+image+of+your+computer%2C+drive+or+partitions

Comparison of Deduplicating Backup Software

Name	License	Speed	Compression	FUSE	Cross Platform	Cloud Storage	Multiple Hosts/Repo
Restic	BSD 2	OK	No	Yes	Yes	Yes	Yes
Borg	BSD 3	Fast	Yes	Yes	No	SSH only	Discouraged
Duplicacy	Dual	Fastest	Yes	No	Yes	Yes	Yes

An Example with Duplicacy

- For these examples you will need some large-ish files to work with
- Search for the following datasets on https://egriddata.org/ and click the “download all” link for each
  - ACTIVsg10k
  - ACTIVsg25k
  - ACTIVsg70k
- For the curious, these are synthetic electric power grid datasets

An Example with Duplicacy

```
~> mkdir power; cd power
~/power> mkdir -p ~/Backup/duplicacy_storage/power
~/power> duplicacy init power-data  ~/Backups/duplicacy_storage
/power
/Users/art/power will be backed up to /Users/art/Backups/duplic
acy_storage/power with id power-data
~/power>
```

Backup First Revision

```
~/power> cp -r ~/Sync/Data/EGridData/ActivSg10k .
~/power> duplicacy backup
Storage set to /Users/art/Backups/duplicacy_storage/power
No previous backup found
Indexing /Users/art/power
Packed ActivSg10k/ACTIVSg10k.RAW (4661029)
Packed ActivSg10k/ACTIVSg10k.aux (23924603)
Packed ActivSg10k/ACTIVSg10k.pwb (15568207)
Packed ActivSg10k/ACTIVSg10k.pwd (13587257)
Backup for /Users/art/power at revision 1 completed
```

Backup 2nd Revision

```
~/power> cp -r  ~/Sync/Data/EGridData/ACTIVSg70k/ .
~/power> duplicacy backup
Storage set to /Users/art/Backups/duplicacy_storage/power
Last backup at revision 1 found
Indexing /Users/art/power
Packed ACTIVSg70k.PWB (54576580)
Packed ACTIVSg70k.RAW (30808345)
Packed ACTIVSg70k.aux (106988707)
Packed ACTIVSg70k.pwd (91995771)
Backup for /Users/art/power at revision 2 completed
```

Backup 3rd Revision

```
~/power> cp -r ~/Sync/Data/EGridData/ACTIVSg25k/ .
~/power> duplicacy backup
Storage set to /Users/art/Backups/duplicacy_storage/power
Last backup at revision 2 found
Indexing /Users/art/power
Packed ACTIVSg25k.RAW (11148262)
Packed ACTIVSg25k.aux (51458987)
Packed ACTIVSg25k.pwb (59512394)
Packed ACTIVSg25k.pwd (32620187)
Backup for /Users/art/power at revision 3 completed
```

List the Backup Revisions

```
~/power> duplicacy list
Storage set to /Users/art/Backups/duplicacy_storage/power
Snapshot power-data revision 1 created at 2021-02-06 20:57 -has
h
Snapshot power-data revision 2 created at 2021-02-06 21:00
Snapshot power-data revision 3 created at 2021-02-06 21:00
```

Create New Storage

```
~/power> duplicacy add -copy default power02 power-data  ~/Back
up/duplicacy_storage/power02
/Users/art/power will be backed up to /Users/art/Backup/duplica
cy_storage/power02 with id power-data
```

Copy One Revision

```
~/power> duplicacy copy -id power-data -r 2 -from default -to p
ower02
Source storage set to /Users/art/Backups/duplicacy_storage/powe
r
Destination storage set to /Users/art/Backup/duplicacy_storage/
power02
Chunk 59101b4712c3e4db8a957881549133b9567494bad30b6608b6491564c
460ca96 (1/72) copied to the destination
Chunk 404a34434d3064f7be3f69d393b6800c5604b02355f87a296f530a934
0dff391 (72/72) copied to the destination
Copy complete, 72 total chunks, 72 chunks copied, 0 skipped
Copied snapshot power-data at revision 2
```

Copy Remaining Revisions

```
~/power> duplicacy copy -from default -to power02
Source storage set to /Users/art/Backups/duplicacy_storage/powe
r
Destination storage set to /Users/art/Backup/duplicacy_storage/
power02
Snapshot power-data at revision 2 already exists at the destina
tion storage
Chunk 6da879600c6fd2b76c257bbdd6befca4aea160560bf238014a5ccc3d2
Chunk dec0b5c1c22927968b9aebfe3f6cb56f26503332cb83d756d034d0d9b
befce03 (104/104) copied to the destination
Copy complete, 104 total chunks, 35 chunks copied, 69 skipped
Copied snapshot power-data at revision 1
Copied snapshot power-data at revision 3
```

Delete a File & List Files in Last Revision

```
~/power> rm ACTIVSg25*
~/power> duplicacy list -files -r 3
Storage set to /Users/art/Backups/duplicacy_storage/power
Snapshot power-data revision 3 created at 2021-02-06 21:00
Files: 12
11148262 2021-02-06 21:00:41 2f7bffd1107fc4ee220476ae684abdbd61
655b20d4485d6ba2a9b2a114b4c6a8 ACTIVSg25k.RAW
51458987 2021-02-06 21:00:41 ffebfc028fc5e77c77690006ab74efd717
71f00531dda1fa60ef22125afb0af8 ACTIVSg25k.aux
59512394 2021-02-06 21:00:41 79f2c0af19446e19ee16f255a5217626c5
20c92b71469dd12dbce1f68ad95508 ACTIVSg25k.pwb
32620187 2021-02-06 21:00:41 7fc13d39afe25a5aa1f8fb63013eb63a5d
00f6c13ad924ec9497421f04008fa1 ACTIVSg25k.pwd
```

Restore the Deleted File

```
~/power> duplicacy restore -r 3 'ACTIVSg25k.pw*'
Storage set to /Users/art/Backups/duplicacy_storage/power
Restoring /Users/art/power to revision 3
Downloaded ACTIVSg25k.pwb (59512394)
Downloaded ACTIVSg25k.pwd (32620187)
Restored /Users/art/power to revision 3
Total running time: 00:00:01
```

Summary of Commands

> mkdir power; cd power

> mkdir -p ~/Backup/duplicacy_storage/power

> duplicacy init power-data ~/Backups/
duplicacy_storage/power

> cp -r ~/Sync/Data/EGridData/ActivSg10k .

> duplicacy backup

> cp -r ~/Sync/Data/EGridData/ACTIVSg70k/ .

> duplicacy backup

> cp -r ~/Sync/Data/EGridData/ACTIVSg25k/ .

> duplicacy backup

> duplicacy list

> mkdir -p ~/Backups/duplicacy_storage/power2

> duplicacy add -copy power power2 power-data
~/Backups/duplicacy_storage/power2

> duplicacy copy -id power-data -r 2 -from default
-to power2

> duplicacy copy -from default -to power2

> rm ACTIVSg25*

> duplicacy list -files -r 3

> duplicacy restore -r 3 'ACTIVSg25k.pw*'

What’s the Problem with Git?

- Git doesn’t handle large binary files well. I create a lot of large files and datasets, both text and binary. Typically this is JSON text or Sqlite (or similar) binary database fines. Git-LFS breaks Git’s distributed model.
- Git doesn’t play well with cloud (or non-cloud) synchronization tools like Dropbox/SyncThing/GoodSync

Deduplicating Backup in a Bit More Detail

- Modern backup software has adopted a deduplicating chunk-based approach instead of incremental backups. One example is Restic, which uses content-based chunking. This includes a nice chunking library in Go based on Rabin fingerprinting, which I use here.

- Rabin fingerprinting is a hash-like algorithm that splits a byte stream into variable-length chunks. The breakpoints are dependent on the content of the stream.

Why (or Why Not) Go?

- The good
  - Very minimal language is easy to learn
  - Compiled fat binaries mean fast execution and ease of distribution
  - Nice for command-line tools
- The bad
  - Executable size is large so load time can be a bit high (100's of ms) so not great for something that needs to run in a tight loop or is frequently called interactively
  - Google has tight control of the language. It's very opinionated, eg. unused variables are errors
- Alternatives:
  - Julia is the upcoming standard for numerical computing. It's JIT compiled, so there is a fixed compile time on execution which is usually on the order of seconds, though there are now options to produce binary executables
  - Nim is pitched as "compiled statically-typed Python." I've heard great stuff about it. The only bad thing is it retains Python's meaningful whitespace

Meet Dupver

https://github.com/akbarnes/dupver

- This is similar to the earlier Boar
- It uses a centralized local repository which is shared by projects on the same computer. All state is kept in the repository and only a minimal configuration file is kept in the project working directories.
- Commits “snapshots” can be copied between repositories

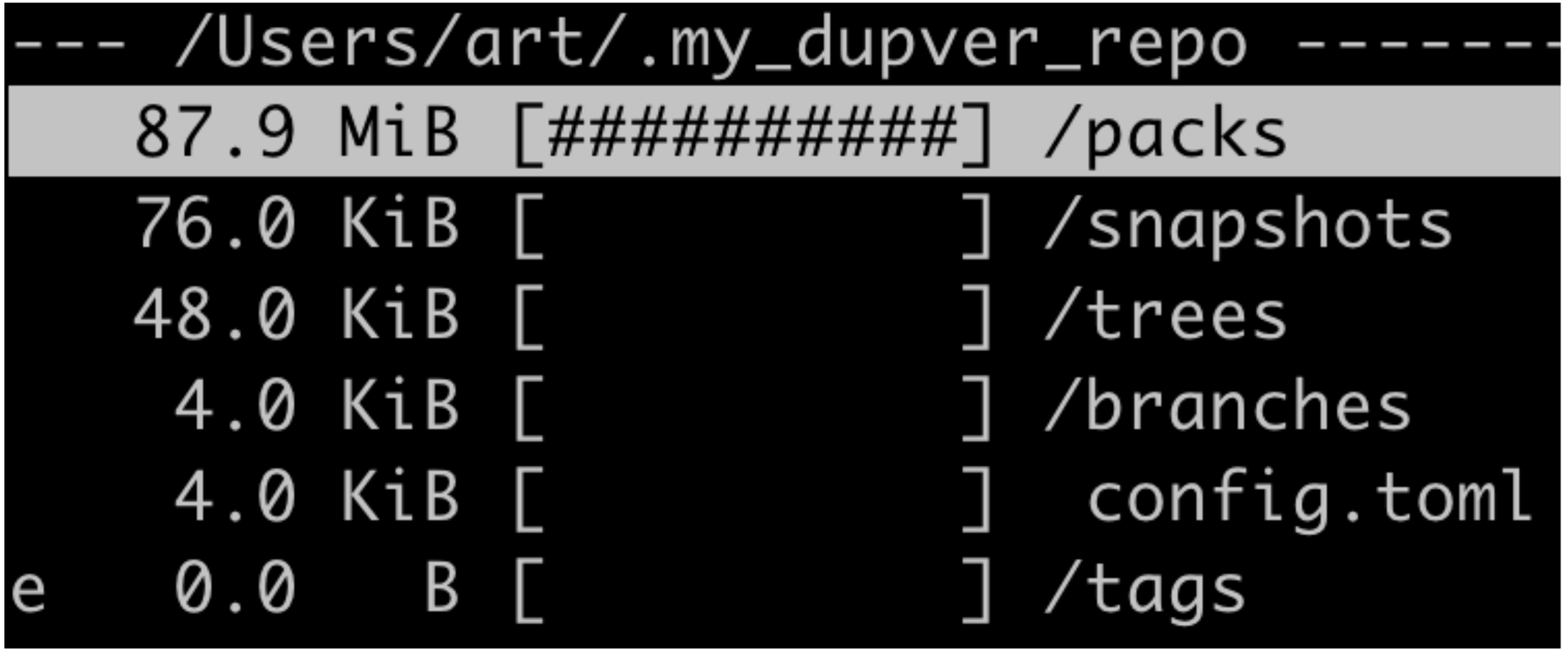
DupVer Working Directory Configuration

File: .dupver/config.toml

```
WorkDirName = "power"
Branch = "main"
DefaultRepo = "main"

[Repos]
  main = "/Users/art/.my_dupver_repo/"
```

Dupver Repository Layout



Dupver Repository Layout

```
~/.my_dupver_repo> tree
.
├── branches
│   └── power
│       └── main.toml
├── config.toml
├── packs
│   ├── 10
│   │   └── 103c9d7192a567b0ff7d159d4f68514945244d3f7c4287f069bd2e499460fb9c.zip
│   ├── 3c
│   │   └── 3cd029c5a5d155ca2e461eb422372a2bccd425b347fd64ad4eea61785512427a.zip
│   ├── 78
│   │   └── 7842e9f4a11014770fcb0ff2d8e7a3eb74fbf8d41fb2aa0ad724688d29b2af08.zip
│   ├── a7
│   │   └── a73e48f2c9704eaf4f799d71ac6cdbbe64e3a0a99659aa291fa306f8c9bc7870.zip
│   ├── d9
│   │   └── d95ba5c7ec12c40bb75bbf7adf682a8e86eaa31520709ae70fca3a540b8e1d5a.zip
│   ├── e5
│   │   └── e577fbc53fdfddcc69e10b7d8d5871c5c0eec6fffb99851fe5a3605515fa2df8.zip
│   └── ea
│       └── eab99e2d3dd37755c8357a7baa265649f4e5e87e61a1a2ec35f58c34dfb8d2d0.zip
├── snapshots
│   └── power
│       ├── 026a8023d68b05aebe8440c4f877b4eb9c3ec053.json
│       ├── 173c8688f0ada7a0f6b5c2ab02346ec09439b265.json
│       ├── 91dd929503f666a2834cdd17b944cc269f42c2b3.json
│       └── d7e68064b19f62105b67582d5830dc93e8f0f3e7.json
├── tags
└── trees
    ├── 026a8023d68b05aebe8440c4f877b4eb9c3ec053.json
    ├── 173c8688f0ada7a0f6b5c2ab02346ec09439b265.json
    ├── 91dd929503f666a2834cdd17b944cc269f42c2b3.json
    └── d7e68064b19f62105b67582d5830dc93e8f0f3e7.json
```

An Example with Dupver

```
~/power> dupver repo init ~/.my_dupver_repo/
Creating folder /Users/art/.my_dupver_repo/
Creating folder /Users/art/.my_dupver_repo/tags
Creating folder /Users/art/.my_dupver_repo/branches
Creating folder /Users/art/.my_dupver_repo/snapshots
Creating folder /Users/art/.my_dupver_repo/trees
Creating folder /Users/art/.my_dupver_repo/packs
Creating folder /Users/art/.my_dupver_repo/snapshots
Creating folder /Users/art/.my_dupver_repo/trees
Chunker Polynomial:
Chunker polynomial: 14935761938367219
Creating config /Users/art/.my_dupver_repo/config.toml
```

Initialize the Working Directory

```
~/power> dupver init -r ~/.my_dupver_repo/
Creating folder .dupver
Workdir name not specified, setting to power
Repo name: [main]
2021/02/11 16:36:20 Refusing to write existing project wo
rkdir config .dupver/config.toml
~/power> bat .dupver/config.toml
```

File: .dupver/config.toml

```
1   WorkDirName = "power"
2   Branch = "main"
3   DefaultRepo = "main"
4
5   [Repos]
6     main = "/Users/art/.my_dupver_repo/"
```

Add Some Files

```
~/power> cp -r ~/Sync/Data/EGridData/ACTIVSg10k .
~/power> dupver commit
/Users/art/power -> /Users/art, power
Message not specified, setting to: power
Tar path: /Users/art/temp/25489a53e80ebca1d18f439aefe648e
f630222bb.tar
Creating folder /Users/art/temp
2021/02/11 16:38:10 Running tar cfv /Users/art/temp/25489
a53e80ebca1d18f439aefe648ef630222bb.tar power
Files:
 1: power/
 2: power/.duplicacy/
 3: power/ACTIVSg10k/
 4: power/.dupver/
```

Add More Files

```
~/power> cp -r ~/Sync/Data/EGridData/ACTIVSg70k/ .
~/power> dupver status
+ power/ACTIVSg70k.PWB
+ power/ACTIVSg70k.RAW
+ power/ACTIVSg70k.aux
+ power/ACTIVSg70k.pwd
~/power> dupver commit -am 'add 70k case'
285.62 new, 56.51 duplicate, 342.13 total MB raw data sto
red
158 new, 38 duplicate, 196 total chunks
3 packs stored, 52.67 chunks/pack
Creating folder /Users/art/.my_dupver_repo/branches/power
Created snapshot d7e68064b19f6210 (d7e68064b19f62105b6758
2d5830dc93e8f0f3e7)
```

Add More More Files

```
~/power> cp -r ~/Sync/Data/EGridData/ACTIVSg25k/ .
~/power> ls
ACTIVSg10k/     ACTIVSg25k.pwb  ACTIVSg70k.RAW
ACTIVSg25k.RAW  ACTIVSg25k.pwd  ACTIVSg70k.aux
ACTIVSg25k.aux  ACTIVSg70k.PWB  ACTIVSg70k.pwd
~/power> rm ACTIVSg70k.PWB
~/power> dupver status
+ power/ACTIVSg25k.RAW
+ power/ACTIVSg25k.aux
+ power/ACTIVSg25k.pwb
+ power/ACTIVSg25k.pwd
- power/ACTIVSg70k.PWB
~/power> dupver commit -m 'add 25k case'
157.91 new, 284.38 duplicate, 442.30 total MB raw data st
ored
82 new, 147 duplicate, 229 total chunks
2 packs stored, 41.00 chunks/pack
Creating folder /Users/art/.my_dupver_repo/branches/power
Created snapshot 173c8688f0ada7a0 (173c8688f0ada7a0f6b5c2
ab02346ec09439b265)
```

Edit ACTIVSg70.RAW

69992,'COLSTRIP 4 1', 500.0000,1, 52, 1, 1,1.06375885, 3.091464, 1.10000, 0.90000, 1.10000, 0.90000

69993,'COLSTRIP 4 2', 161.0000,1, 52, 1, 1,1.05972719, 2.365558, 1.10000, 0.90000, 1.10000, 0.90000

69994,'COLSTRIP 4 3', 69.0000,1, 52, 1, 1,1.05814064, 1.584992, 1.10000, 0.90000, 1.10000, 0.90000

69995,'COLSTRIP 4 4', 13.8000,2, 52, 1, 1,1.04240942, 4.180727, 1.10000, 0.90000, 1.10000, 0.90000

69996,'COLSTRIP 4 5', 18.0000,2, 52, 1, 1,1.04452062, 3.899833, 1.10000, 0.90000, 1.10000, 0.90000

69997,'COLSTRIP 4 6', 13.8000,2, 52, 1, 1,1.04412520, 3.822852, 1.10000, 0.90000, 1.10000, 0.90000

69998,'COLSTRIP 4 7', 13.8000,2, 52, 1, 1,1.04379857, 4.845811, 1.10000, 0.90000, 1.10000, 0.90000

69999,'COLSTRIP 4 8', 500.0000,1, 52, 1, 1,1.06541932, 2.741757, 1.10000, 0.90000, 1.10000, 0.90000

70000,'COLSTRIP 4~1', 161.0000,1, 52, 1, 1,1.05973113, 2.302492, 1.10000, 0.90000, 1.10000, 0.90000

0 / END OF BUS DATA, BEGIN LOAD DATA

69992,'CONSTRIP 4 1', 500.0000,1, 52, 1, 1,1.06375885, 3.091464, 1.10000, 0.90000, 1.10000, 0.90000

69993,'CONSTRIP 4 2', 161.0000,1, 52, 1, 1,1.05972719, 2.365558, 1.10000, 0.90000, 1.10000, 0.90000

69994,'CONSTRIP 4 3', 69.0000,1, 52, 1, 1,1.05814064, 1.584992, 1.10000, 0.90000, 1.10000, 0.90000

69995,'CONSTRIP 4 4', 13.8000,2, 52, 1, 1,1.04240942, 4.180727, 1.10000, 0.90000, 1.10000, 0.90000

69996,'CONSTRIP 4 5', 18.0000,2, 52, 1, 1,1.04452062, 3.899833, 1.10000, 0.90000, 1.10000, 0.90000

69997,'CONSTRIP 4 6', 13.8000,2, 52, 1, 1,1.04412520, 3.822852, 1.10000, 0.90000, 1.10000, 0.90000

69998,'CONSTRIP 4 7', 13.8000,2, 52, 1, 1,1.04379857, 4.845811, 1.10000, 0.90000, 1.10000, 0.90000

69999,'CONSTRIP 4 8', 500.0000,1, 52, 1, 1,1.06541932, 2.741757, 1.10000, 0.90000, 1.10000, 0.90000

70000,'CONSTRIP 4~1', 161.0000,1, 52, 1, 1,1.05973113, 2.302492, 1.10000, 0.90000, 1.10000, 0.90000

0 / END OF BUS DATA, BEGIN LOAD DATA

Now Commit the Changes

```
~/power> dupver status
M power/ACTIVSg70k.RAW
~/power> dupver commit -am 'change a bit of a large file'

/Users/art/power -> /Users/art, power
Tar path: /Users/art/temp/6f704cb01463fb41cb69acd4f1df7efd655db7ac.tar
Creating pack number:   1, ID: e577fbc53fdfddcc
1.17 new, 441.13 duplicate, 442.30 total MB raw data stored
2 new, 227 duplicate, 229 total chunks
1 packs stored, 2.00 chunks/pack
Creating folder /Users/art/.my_dupver_repo/branches/power
Created snapshot 026a8023d68b05ae (026a8023d68b05aebe8440c4f877b4eb9c3ec053)
```

The Commit History

```
~/power> dupver log
Branch: main
Snapshot History
ID: 026a8023 (026a8023d68b05aebe8440c4f877b4eb9c3ec053)
Time: 2021/02/11 16:49:04
Message: change a bit of a large file

ID: 173c8688 (173c8688f0ada7a0f6b5c2ab02346ec09439b265)
Time: 2021/02/11 16:44:08
Message: add 25k case

ID: 91dd9295 (91dd929503f666a2834cdd17b944cc269f42c2b3)
Time: 2021/02/11 16:38:10
Message: power

ID: d7e68064 (d7e68064b19f62105b67582d5830dc93e8f0f3e7)
Time: 2021/02/11 16:39:58
Message: add 70k case
```

Let's Restore From a Commit

```
~/power> dupver checkout 91dd9295
Wrote to power-2021-02-11T16-38-10-91dd929503f666a2.tar
~/power> tar tf power-2021-02-11T16-38-10-91dd929503f666a
2.tar
power/
power/.duplicacy/
power/ACTIVSg10k/
power/.dupver/
power/.dupver/config.toml
power/ACTIVSg10k/ACTIVSg10k.RAW
power/ACTIVSg10k/ACTIVSg10k.aux
power/ACTIVSg10k/ACTIVSg10k.pwb
power/ACTIVSg10k/ACTIVSg10k.pwd
```

Summary of Dupver Commands

> dupver repo init ~/.my_dupver_repo/

> dupver init -r ~/.my_dupver_repo/

> cp -r ~/Sync/Data/EGridData/ActivSg10k .

> dupver commit

> cp -r ~/Sync/Data/EGridData/ACTIVSg70k/ .

> dupver status

> dupver commit -am 'add 70k case'

> cp -r ~/Sync/Data/EGridData/ACTIVSg25k/ .

> dupver commit -am 'add 25k case'

> dupver log

> cp -r ~/Sync/Data/EGridData/ACTIVSg70k/ .

> duper status

> dupver commit -am 'change a bit of a large file'

> dupver log

> dupver checkout afd22583

> tar tf power-2021-02-11T16-11-56-afd2258315fad397.tar

Some Concluding Thoughts about Dupver

- Dupver works best for structured files like Sqlite3 databases
  - This works great for QGIS .gpkg files, which are just Sqlite3 databases
  - For your own data, the best tradeoff between storing working files and versions efficiently is structured binary formats like MessagePack, HDF5, ProtoBuf, etc.
- A lot of modern software use compressed file formats that will hurt performance in terms of storing small changes efficiently
  - MS Office documents are collections of XML files stored in a .zip file
  - paint.net is the worst offender, and just wraps the entire stream in gzip

PS: What about REALLY BIG Repos?

- Git-Annex is a really slick piece of software written in Haskell that handles the use case of repos that are very large compared to the available amount of disk space and in general (someone on r/DataHoarders was keeping ~~18TB~~ 100TB in one repo!)
  - Replaces git packfiles with hashed links to files
  - Remotes keep track of all copies of a file that exist, including on remote or offline media
  - File metadata including tags is supported
  - Files can be locked and replaced with symlinks