Created and Maintained by Joseph Jacks (OSS Capital) and Justin Cormack (Docker) in early December 2020

Criteria: This sheet lists FOSS technologies (and COSS companies if they exist) which focus on creating "git for data" solutions in various domains.

OSS Company	*Known Funding (M)*	*OSS Technology*	*Stars*	*Launched*	*Status*	*Focus*	*Description*	*Lead Creator/PM*
-	-	https://github.com/dolthub/dolthub	250+	September 2013	Dormant	data collaboration	Data collaboration platform	https://twitter.com/aircrash
Datahub	[illegible]	https://github.com/datajuicer/datajuicer	1,400+	August 2018	Dormant	lsdfs	Branch and merge SQLite	https://twitter.com/datajuicer_io
Nile-table	[illegible]	https://github.com/nile-table/nile-table	7,300+	May 2018	Dormant	Database	Declarative content addressed database	https://twitter.com/dbcodetales
Dolthub	[illegible]	https://github.com/dolthub/dolthub	2,000+	December 2018	Active	SQL	Version SQL tables, merge, branch, Hosted hub for public data.	https://twitter.com/liquidata
Datahub.io	[illegible]	https://github.com/datahub-io/datahub	500+	February 2019	Dormant	ML	Originally general purpose versioned data, pivoted to replicatable experiments	https://twitter.com/rgrossman
Orbit.ml	[illegible]	https://github.com/orbitml/orbitml	400+	July 2018	Active	ETL	Orchestration of ELT pipelines	https://twitter.com/DataKind
Gretel.ai	[illegible]	https://github.com/gretelai/gretel-synthetics	100+	March 2020	Active	data generation	Synthetic privacy preserving data generation	https://twitter.com/AlexWaldman54
Qeel.xlsx	-	https://github.com/qeel/qeel	550+	January 2013	Dormant	Diff	Data diff tool	https://twitter.com/BenVanRoy
Qeel.xlsx	-	https://github.com/qeel/qeel-core	20+	May 2020	Active	Spreadsheet	Versioned spreadsheet	https://twitter.com/BenVanRoy
Dolthub	[illegible]	https://github.com/dolthub/dolt	6,800+	March 2017	Active	ML	SWGit LFS and Kitsafilles for ML and data science	https://twitter.com/liquidata
Datachefhq	[illegible]	https://github.com/datachefhq/datachef	4700+	October 2014	Active	data science	Version controlled data ingestion and processing pipeline	https://twitter.com/datachef
Dvc	-	https://github.com/iterative/dvc	7,000+	October 2016	Active	data management	Dataset version control	https://twitter.com/b_ff_fluid
Pachyderm	[illegible]	https://github.com/pachyderm/pachyderm	5,000+	February 2017	Active	ML/data	Versioning for small and large data that don't fit in git eg ML models. S3/AWS based	https://twitter.com/dwhitena
Quilt.data	-	https://github.com/quiltdata/quilt	500+	August 2020	Active	ML	Version ML models, focus on simpler workflows and introducing people to version control	https://twitter.com/fiskus
Dagster	-	https://github.com/dagster-io/dagster	1,400+	August 2017	Active	Blockchain/general	Git for merging distributed data models. OCaml. Used by Tezos.	https://twitter.com/dsiegelman
Dgraph.io/DB	[illegible]	https://github.com/dgraph-io/dgraph	5,000+	May 2019	Active	Database	Revision controlled graph database	https://twitter.com/DamanSingh
Datagitlab	-	https://github.com/datagitlab/datagitlab	500+	September 2019	Active	data management	Versioned data lake for ETL and data science	https://twitter.com/datagitlab
AskHub	$7.5M	-	0	-	Active	data management		https://twitter.com/