DATAVISION
by soda

Level Up your Data Lake – to ML and Beyond

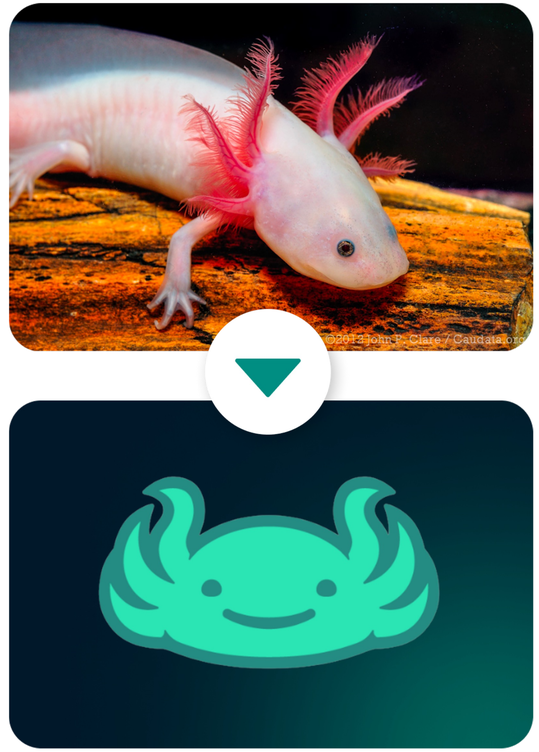
Oz Katz

Hi 👋

- Oz Katz
- CTO, Co-Creator @ lakeFS
- Prev: VP R&D, SimilarWeb (NYSE: SMWB)
- > 15 years in software + data engineering

Hi #2 👋

- Lottie
- I am an Axolotl and VP Cuteness @ lakeFS
- I am critically endangered
- ~200m years of lake experience 🌊

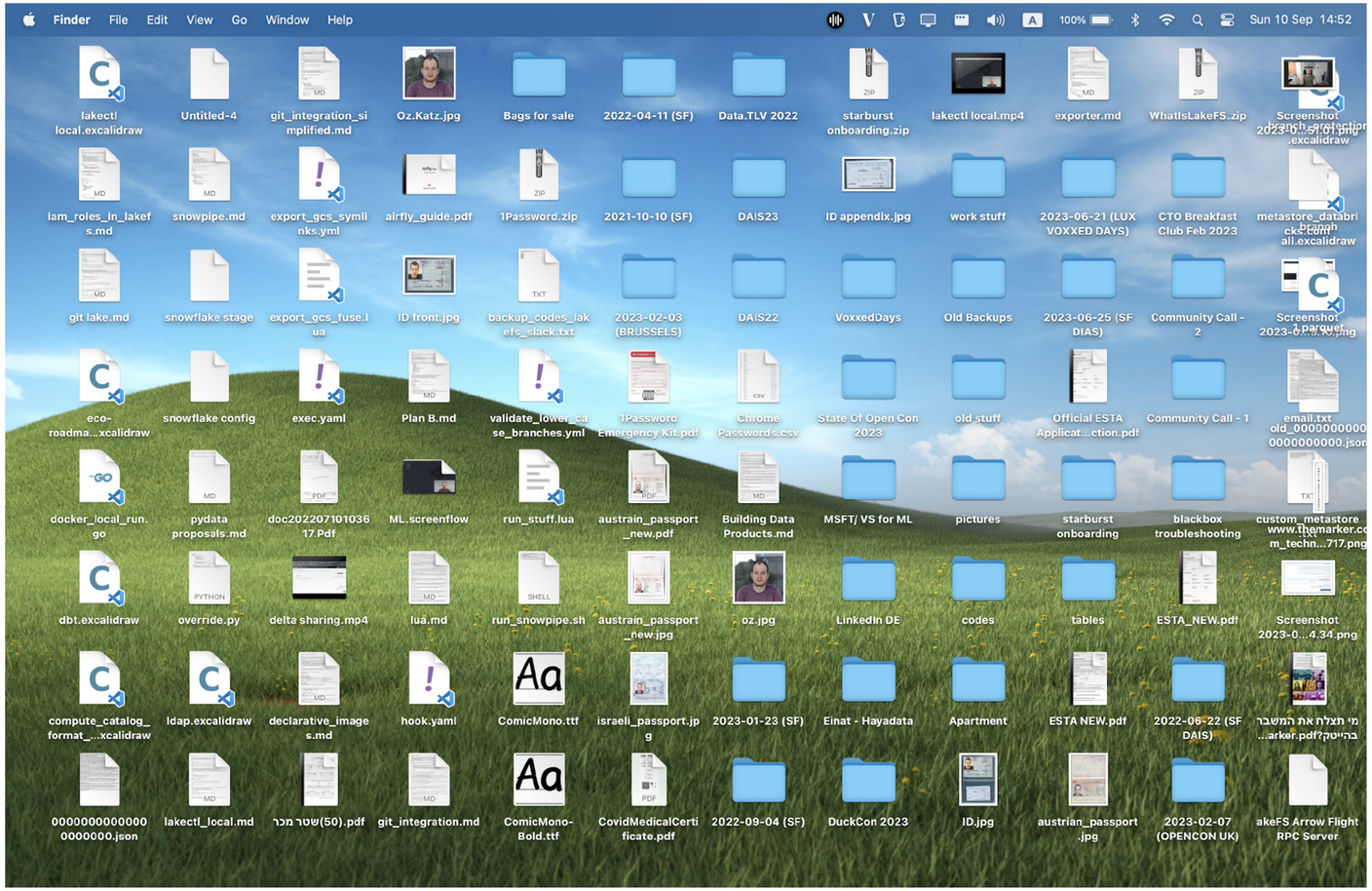


DATAVISION by soda

What we'll talk about

- Data Lakes today
- Common Challenges (no, you're not alone 💚)
- Tools and processes that can help

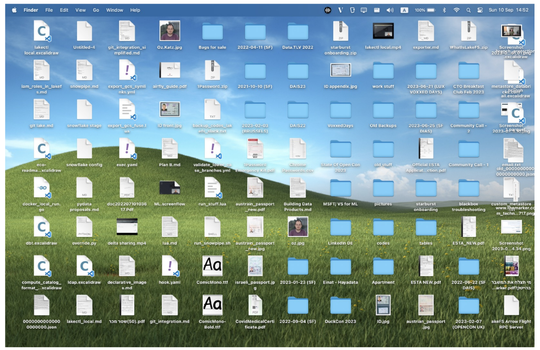
DATAVISION
by soda



Yes, This is my Actual Desktop

DATAVISION by soda

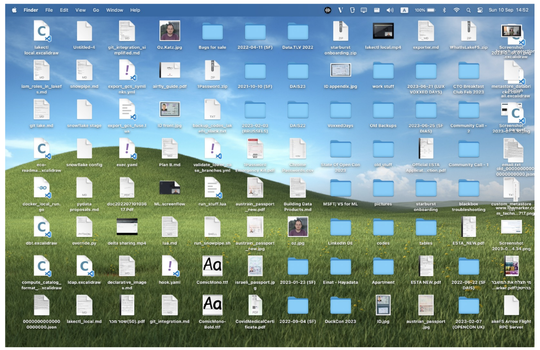
Data Lakes: a helpful illustration

Complexity =  * len(engineers) * len(scientists)^2

DATAVISION by soda

Data Lakes: a helpful illustration

`Complexity =`

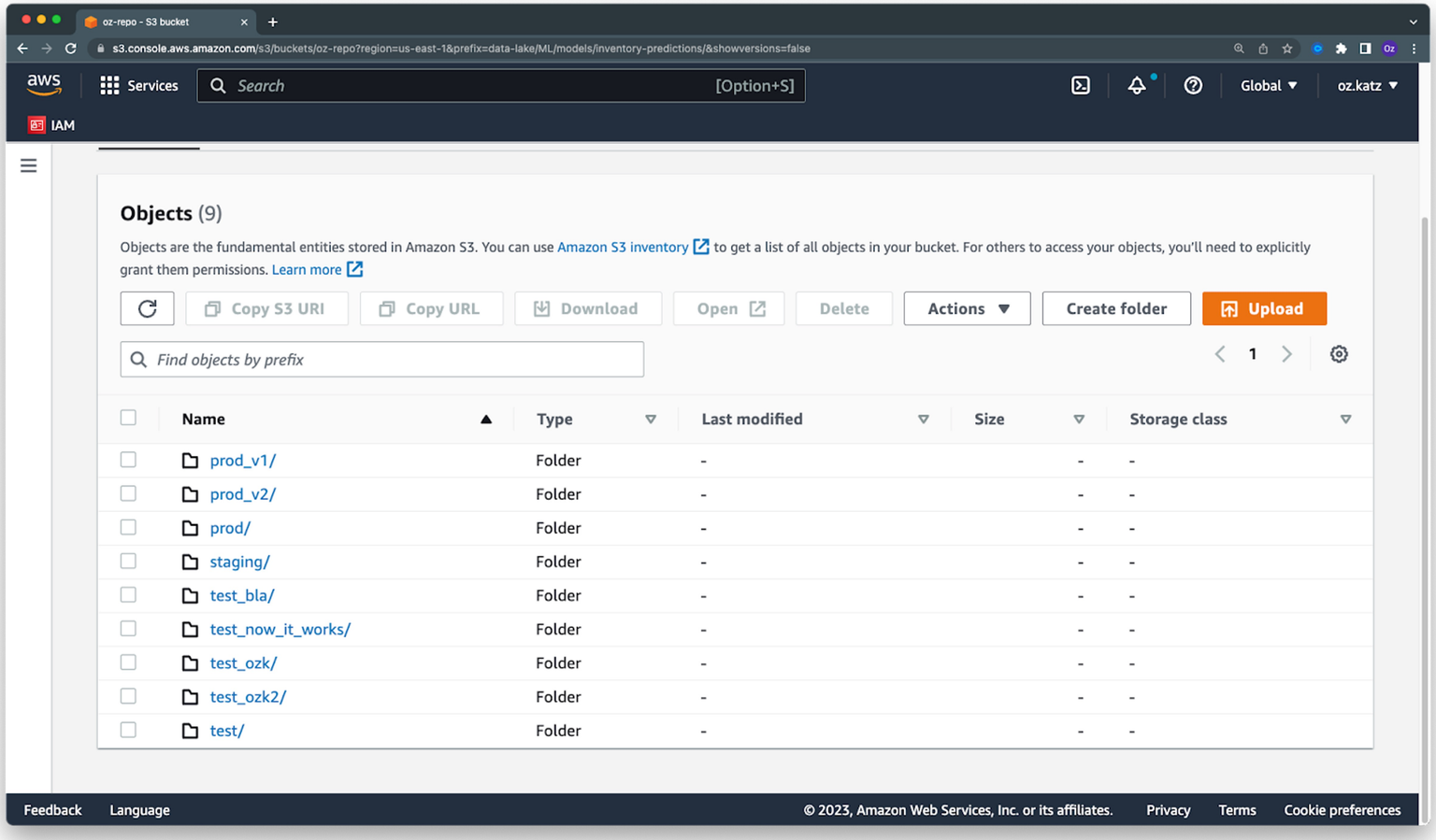


`* len(engineers) * len(scientists)²`

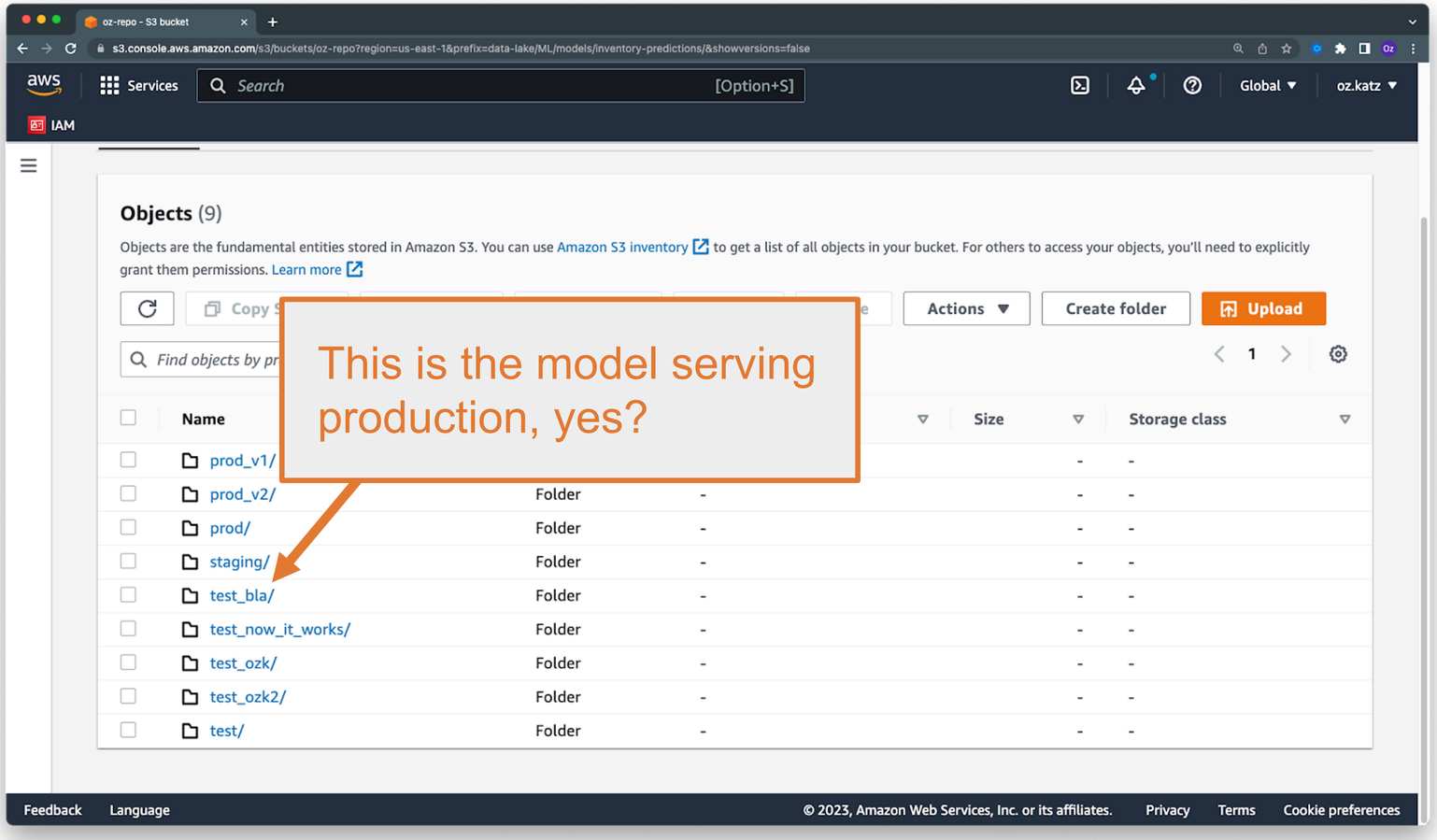
You folks are the messiest!

DATAVISION by soda

So we end up with this:



So we end up with this:

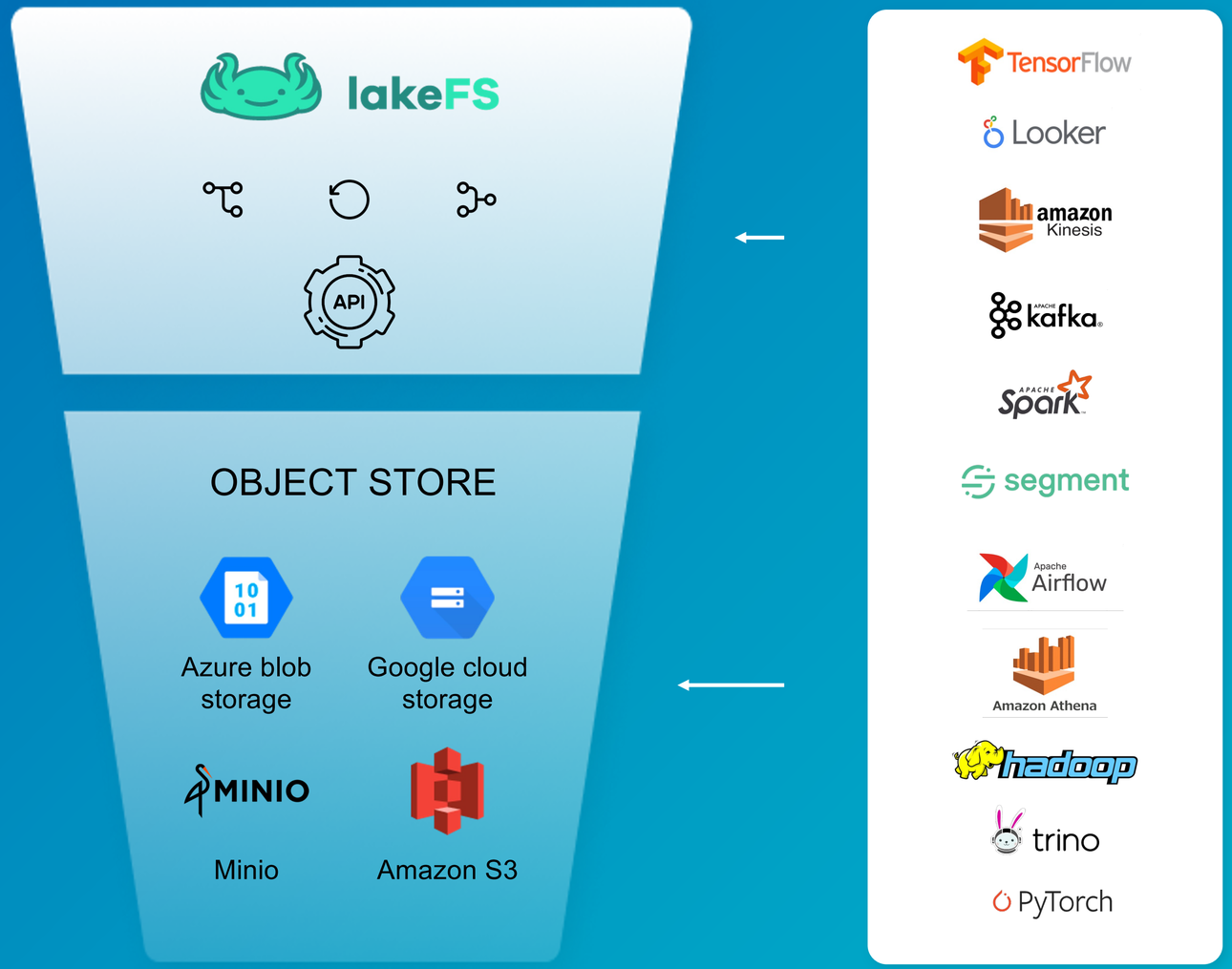


Challenges

- **Reproducibility** in an ever changing and error-prone world
- **Dev/Test** is notoriously ~~difficult~~ impossible
- As a result, so is ensuring **production safety**

DATAVISION by soda

Enter, lakeFS



s3://data-repo/collections/foo

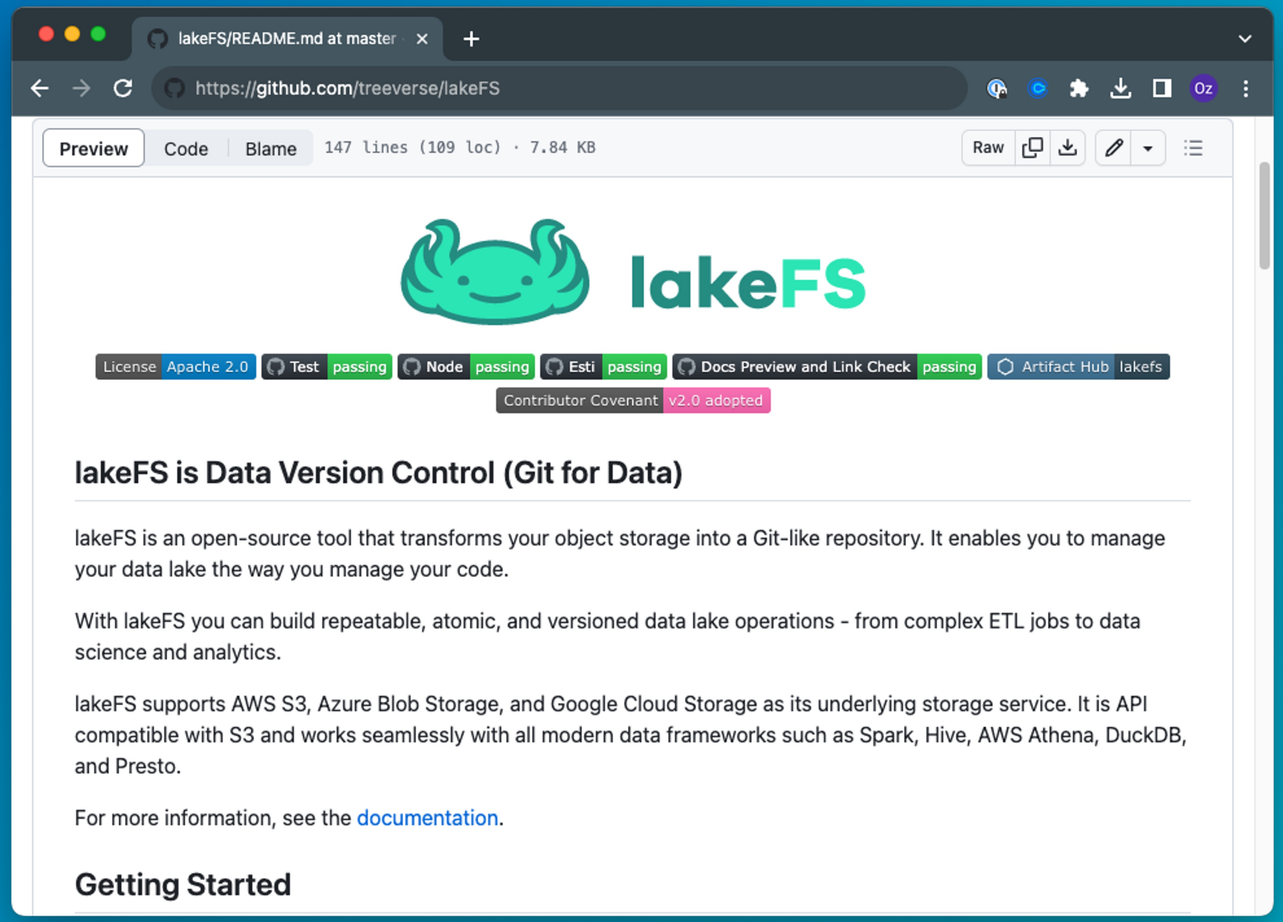
s3://data-repo/main/collections/foo

```
lakectl branch create \
      "lakefs://data-repo@my-experiment" \
      --source "lakefs://data-repo/main"


// output:
// created branch 'my-experiment',
// pointing to commit ID: 'd1e9adc71c10a'
```

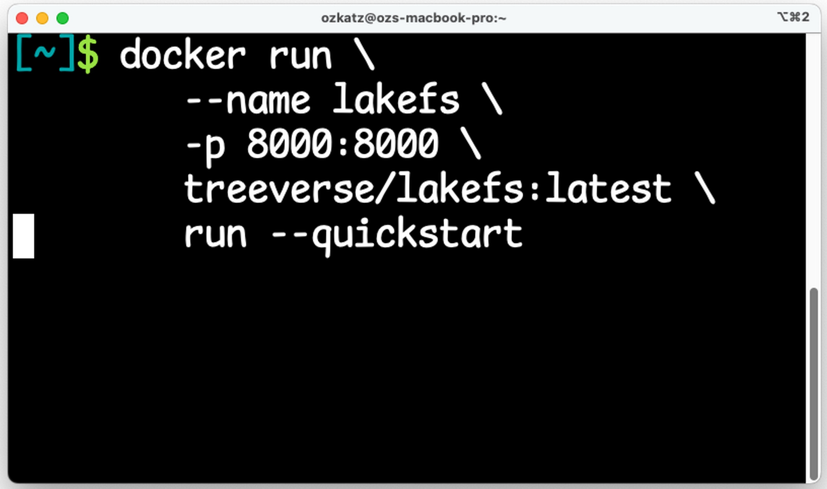
DATAVISION by soda

Enter, lakeFS

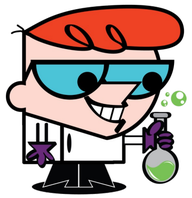


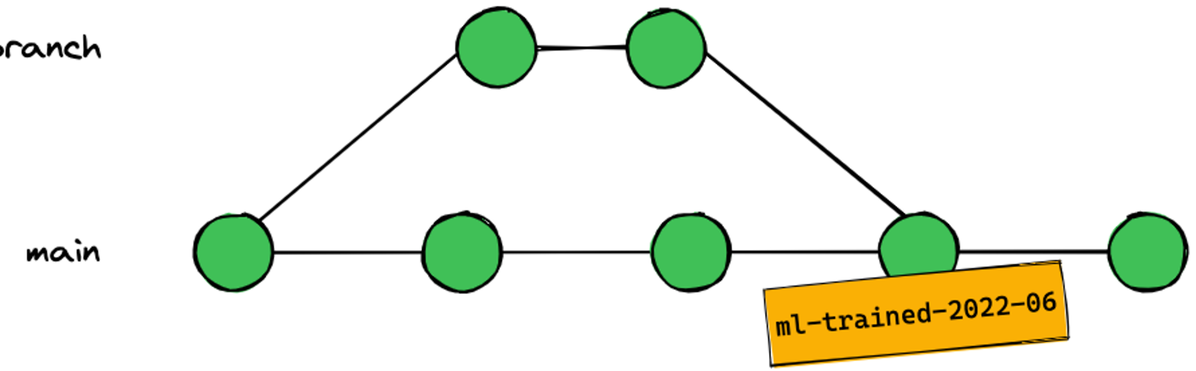
https://github.com/treeverse/lakeFS

License: Apache 2.0



Reproducibility in an ever changing and error-prone world





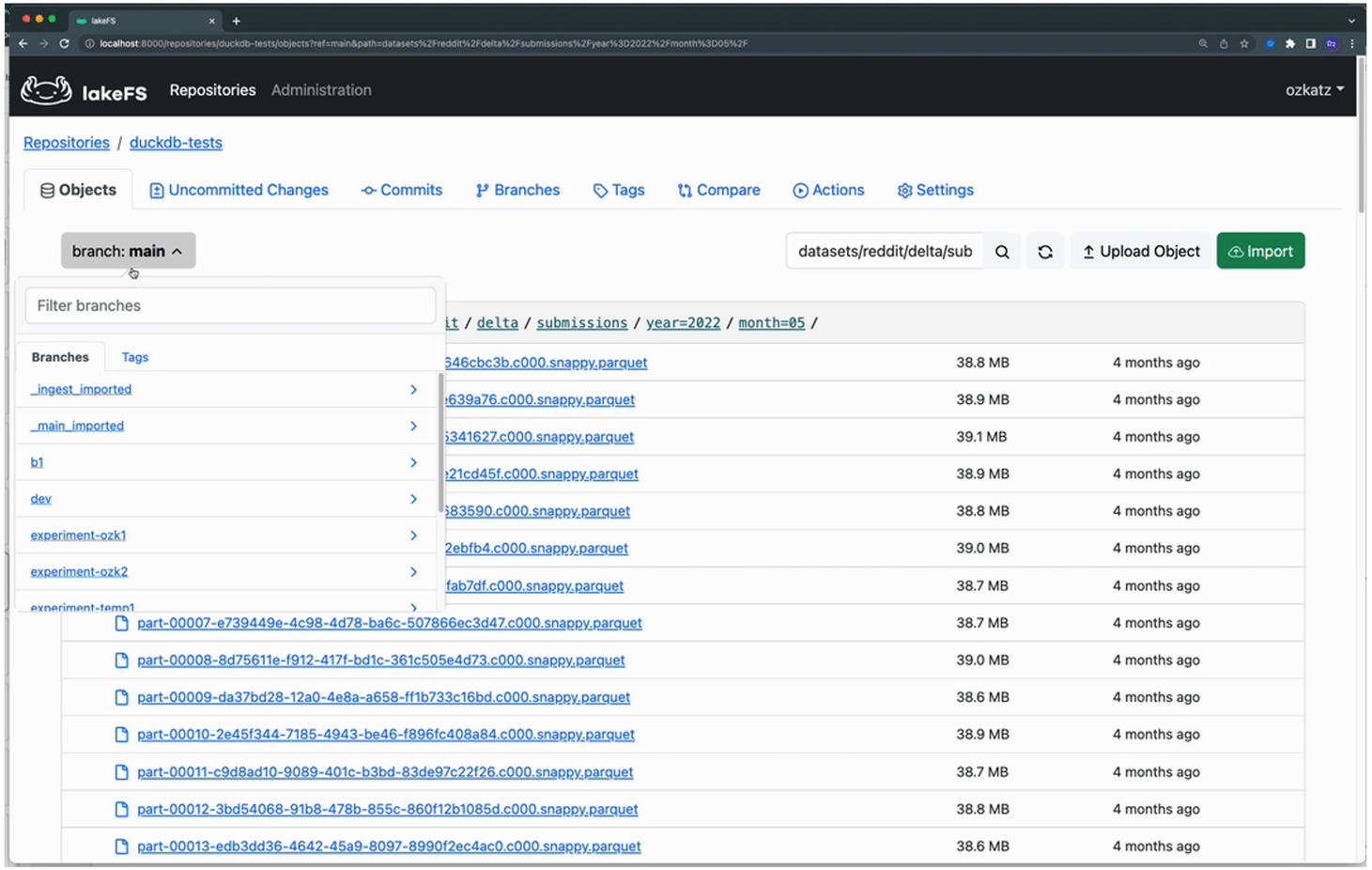

```
df = spark.read.parquet(
    'lakefs://my-repo/ml-trained-2022-06/inputs/vists/')
```

Merged estimation model using new learning set

hyper-params	{ ... }
training-job-uri	github.com/my-ml/a34902
infra-uri	github.com/my-terraform/..

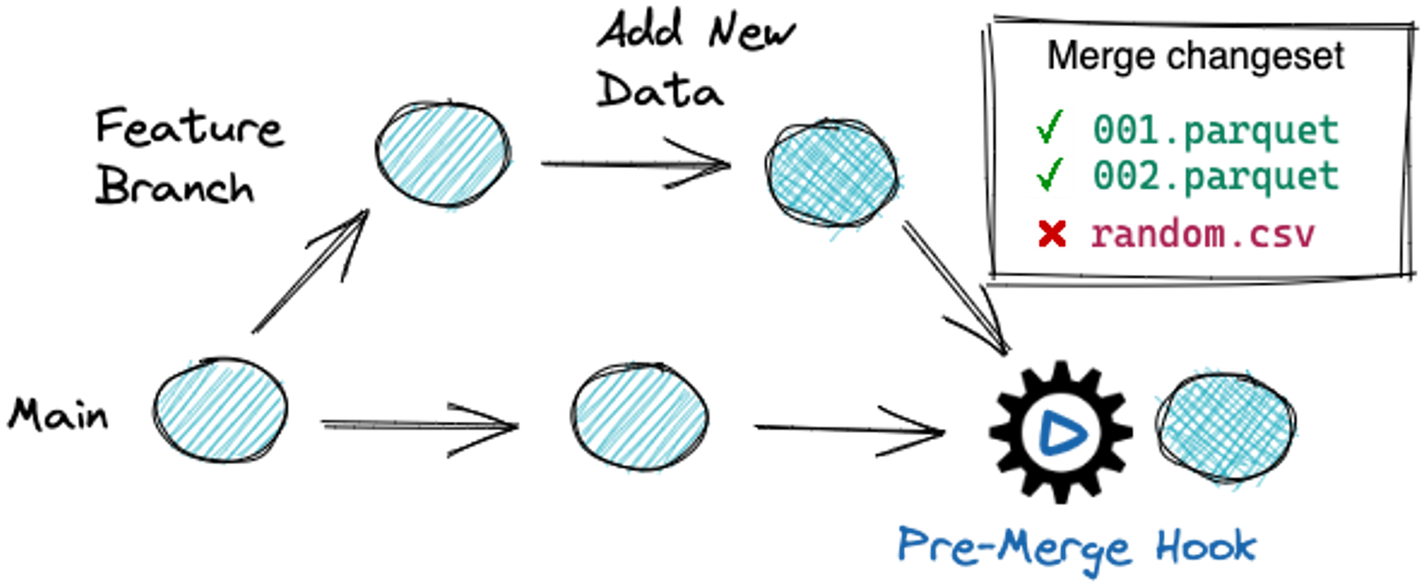
DATAVISION by soda

Dev/Test is notoriously ~~difficult~~ impossible



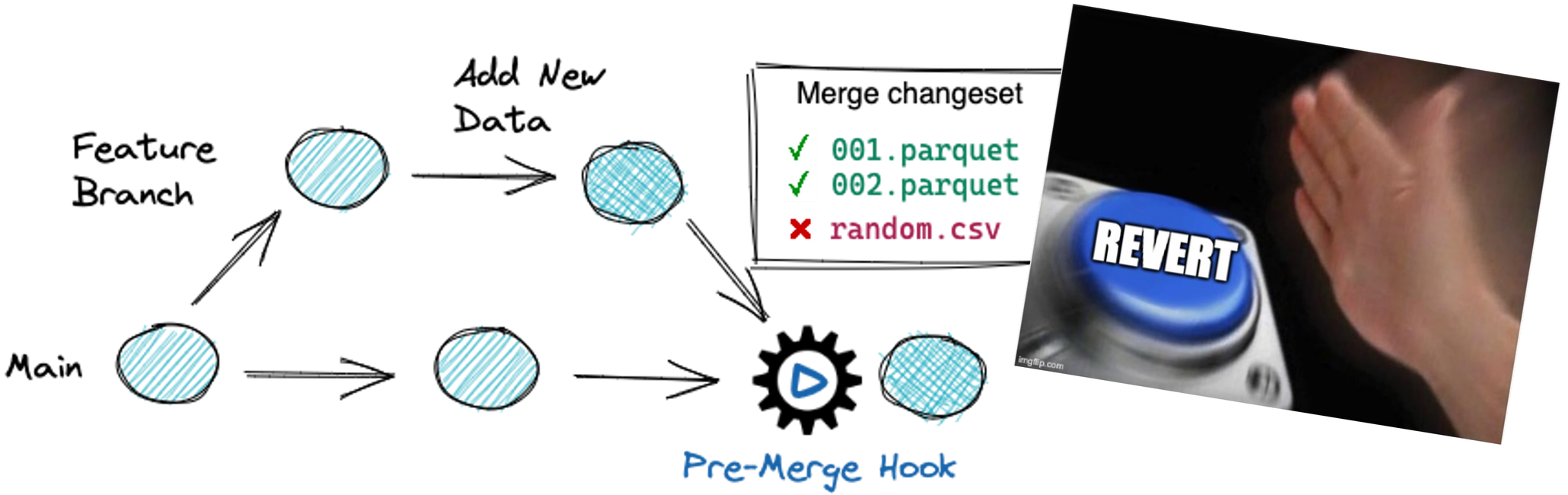
DATAVISION
by soda

As a result, so is ensuring production safety



DATAVISION
by soda

As a result, so is ensuring production safety



lakeFS enabled Data Lakes

- **Allow Reproducibility** using commits and named tags
- **Dev/Test** are easy with copy-on-write branching
- **Production safety** is achieved using hooks and a revert button 🙏

DATAVISION
by soda

Learn More

lakefs.io

https://lakefs.io/community

github.com/treeverse/lakeFS

DATAVISION
by soda

THANK YOU!

Oz & Lottie the Axolotl