

Level Up your Data Lake – to ML and Beyond

Oz Katz

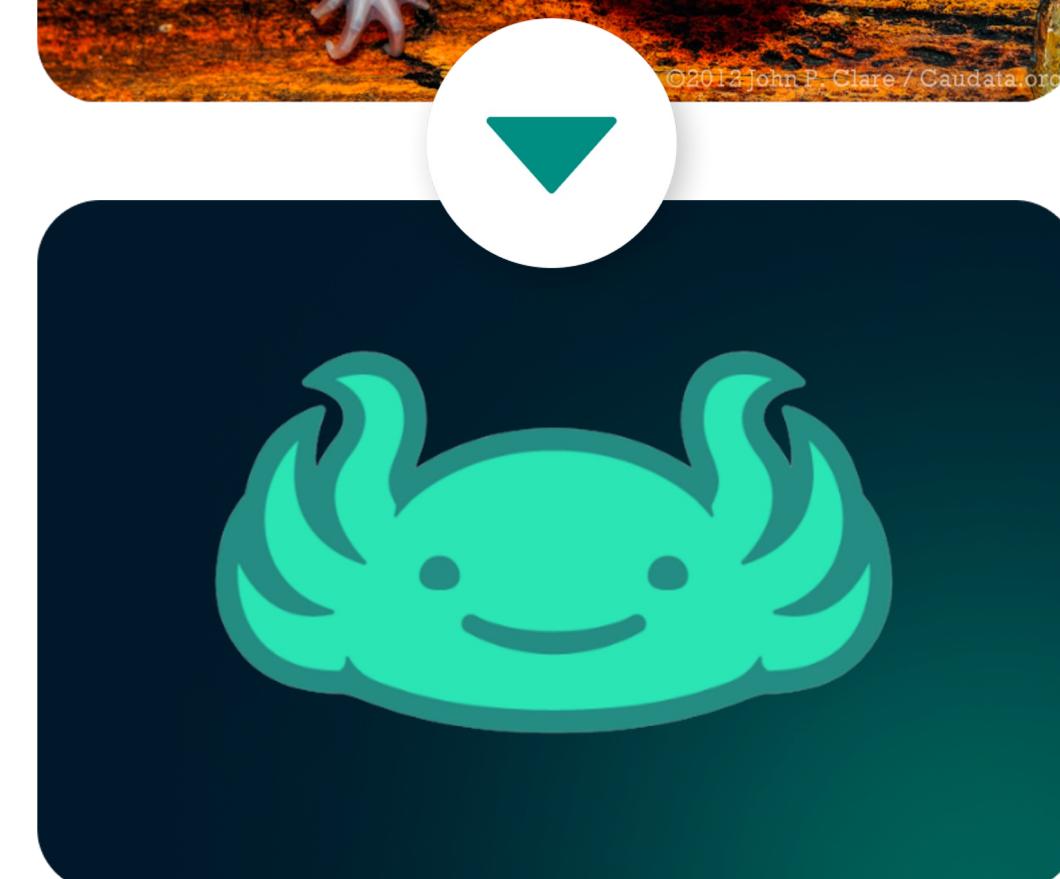
Hi 🙌

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



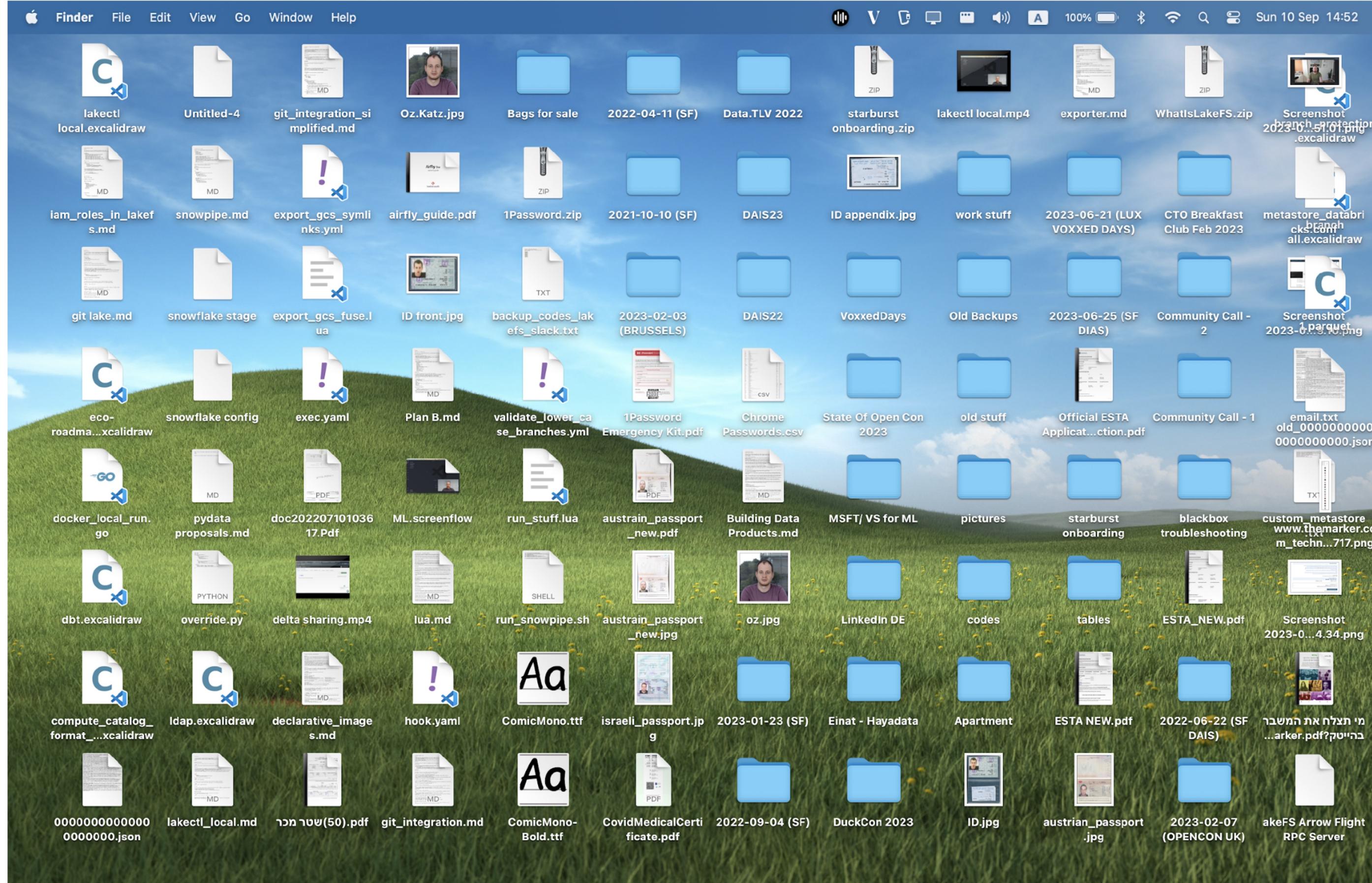
Hi #2 🙌

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



What we'll talk about

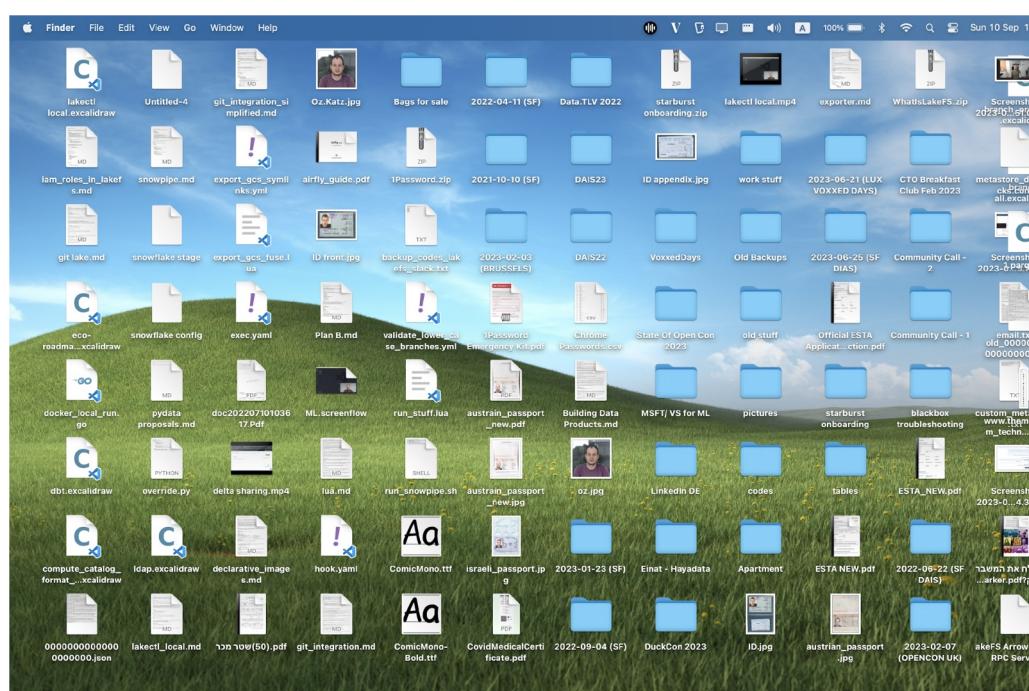
- Data Lakes today
- Common Challenges (no, you're not alone ❤️)
- Tools and processes that can help



**Yes,
This is my
Actual
Desktop**

Data Lakes: a helpful illustration

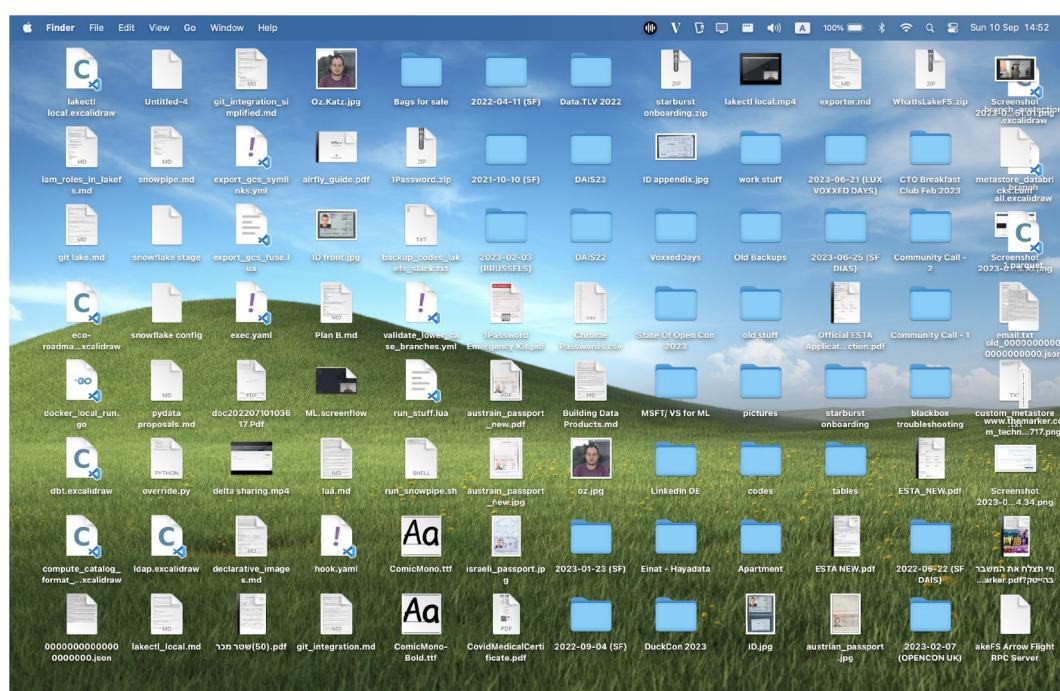
Complexity =



$* \text{len(engineers)} * \text{len(scientists)}^2$

Data Lakes: a helpful illustration

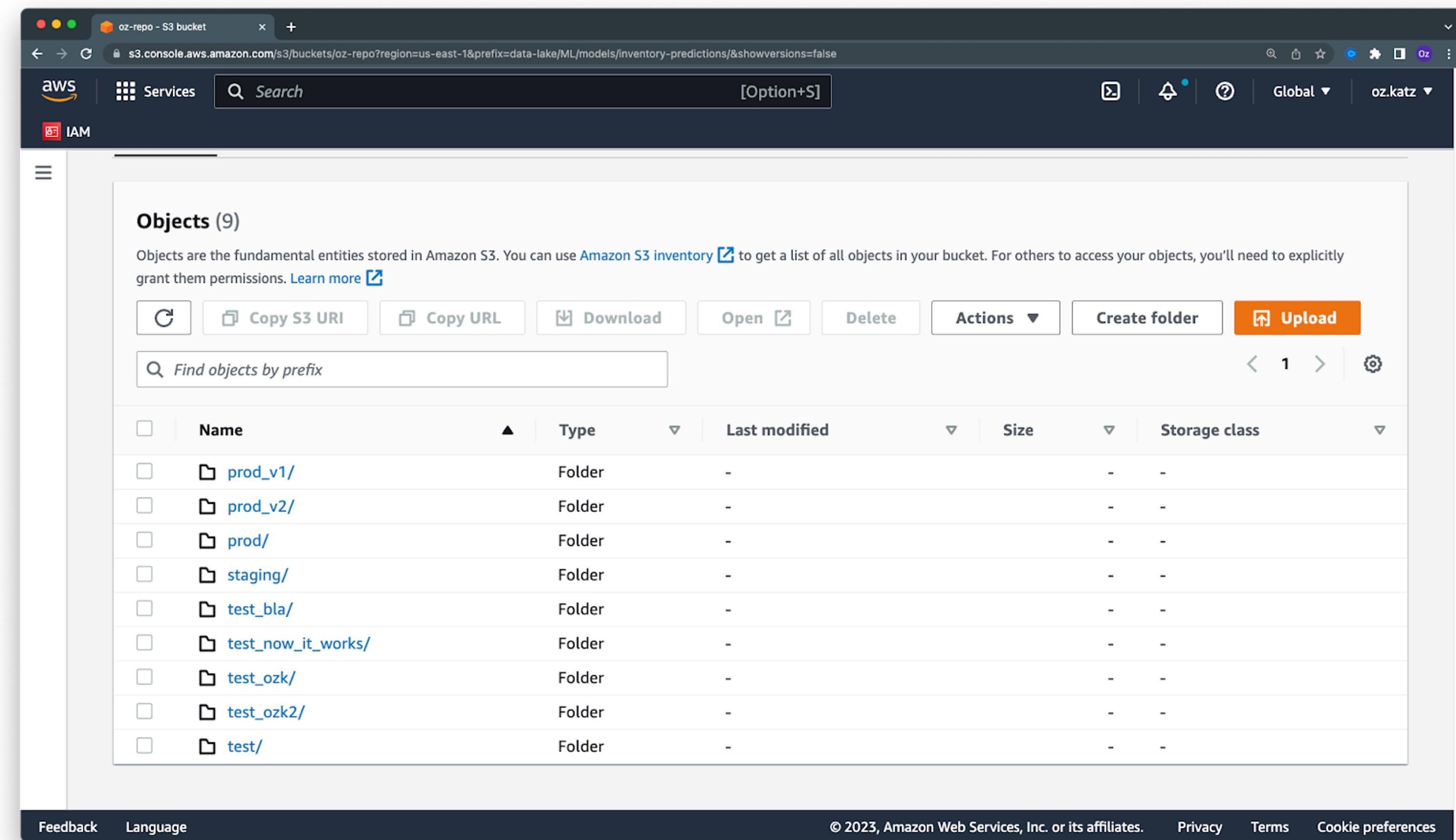
Complexity =



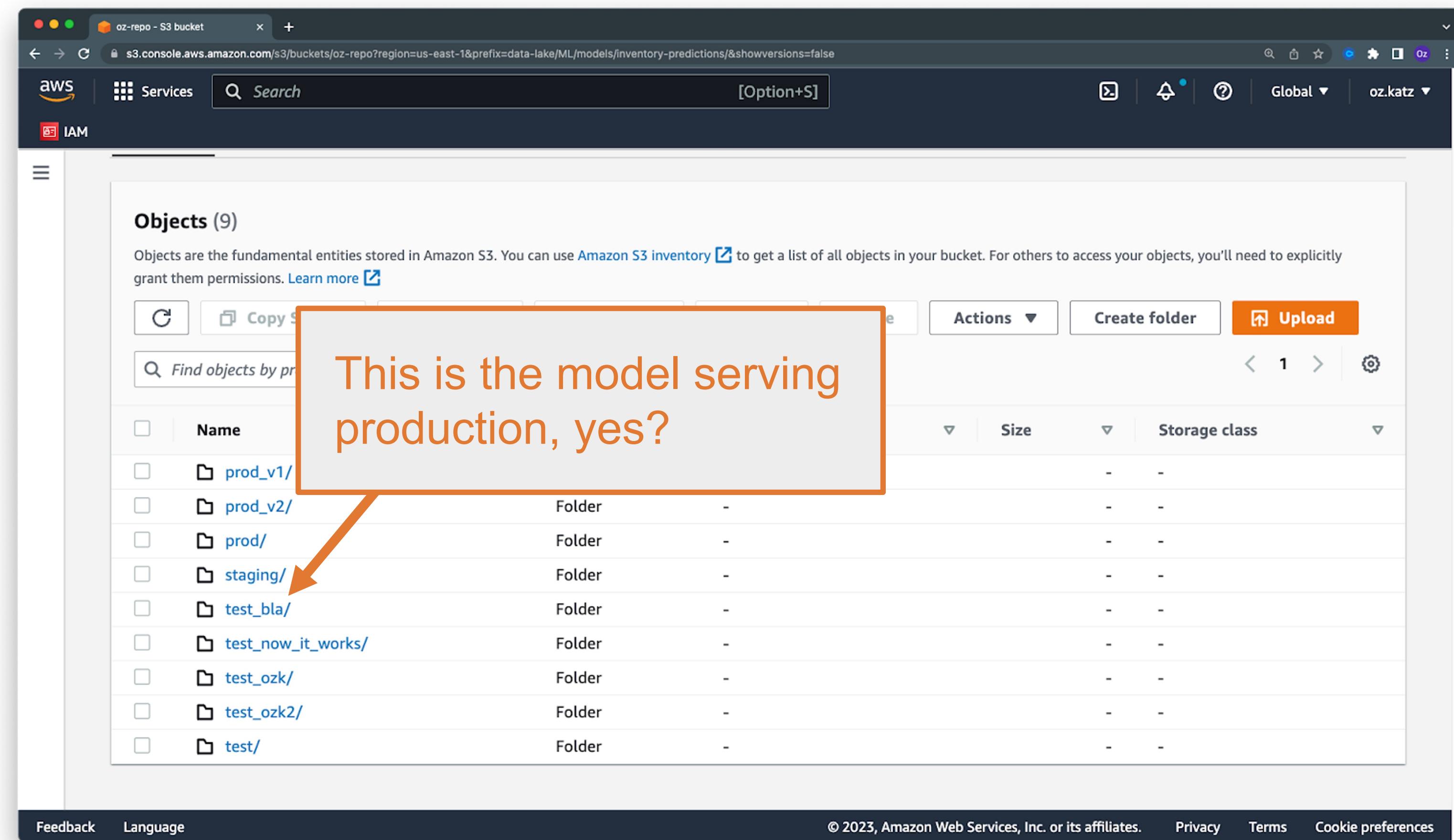
You folks are the messiest!

* len(engineers) * len(scientists)²

So we end up with this:



So we end up with this:





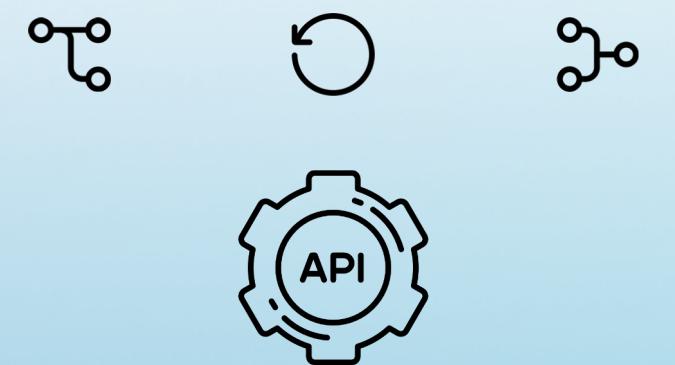
Challenges

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

Enter, lakeFS



lakeFS



OBJECT STORE



Azure blob
storage



Google cloud
storage



Minio



Amazon S3



Looker



amazon
Kinesis



APACHE
kafka



segment



Apache
Airflow



Amazon Athena



trino



PyTorch

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'
```

Enter, lakeFS

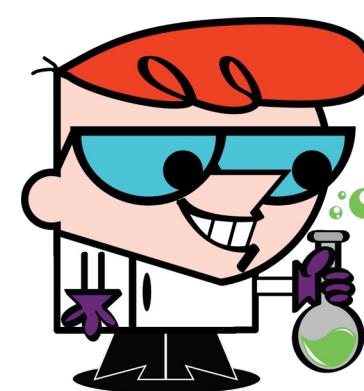
The screenshot shows a GitHub README page for the lakeFS repository. At the top, there's a green octocat icon followed by the text "lakeFS". Below the logo, the word "lakeFS" is written in a large, bold, teal font. Underneath the title, there's a row of status badges: "License Apache 2.0" (green), "Test passing" (green), "Node passing" (green), "Esti passing" (green), "Docs Preview and Link Check passing" (green), and "Artifact Hub lakefs" (blue). A pink button below the badges says "Contributor Covenant v2.0 adopted". The main content starts with a section titled "lakeFS is Data Version Control (Git for Data)". It describes lakeFS as an open-source tool that transforms object storage into a Git-like repository, enabling data lake management. It highlights the ability to build repeatable, atomic, and versioned operations. lakeFS supports AWS S3, Azure Blob Storage, and Google Cloud Storage, and is API compatible with S3, working with frameworks like Spark, Hive, AWS Athena, DuckDB, and Presto. For more information, a blue link points to the documentation.

<https://github.com/treeverse/lakeFS>

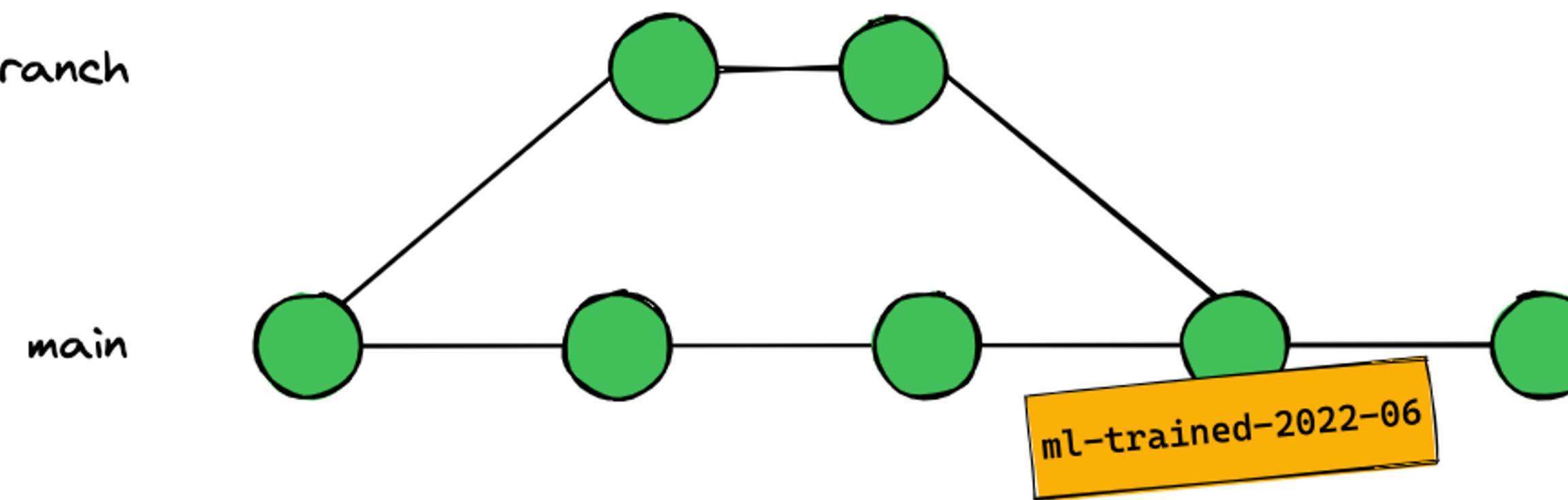
License: [Apache 2.0](#)

```
[~]$ docker run \
--name lakefs \
-p 8000:8000 \
treeverse/lakefs:latest \
run --quickstart
```

Reproducibility in an ever changing and error-prone world



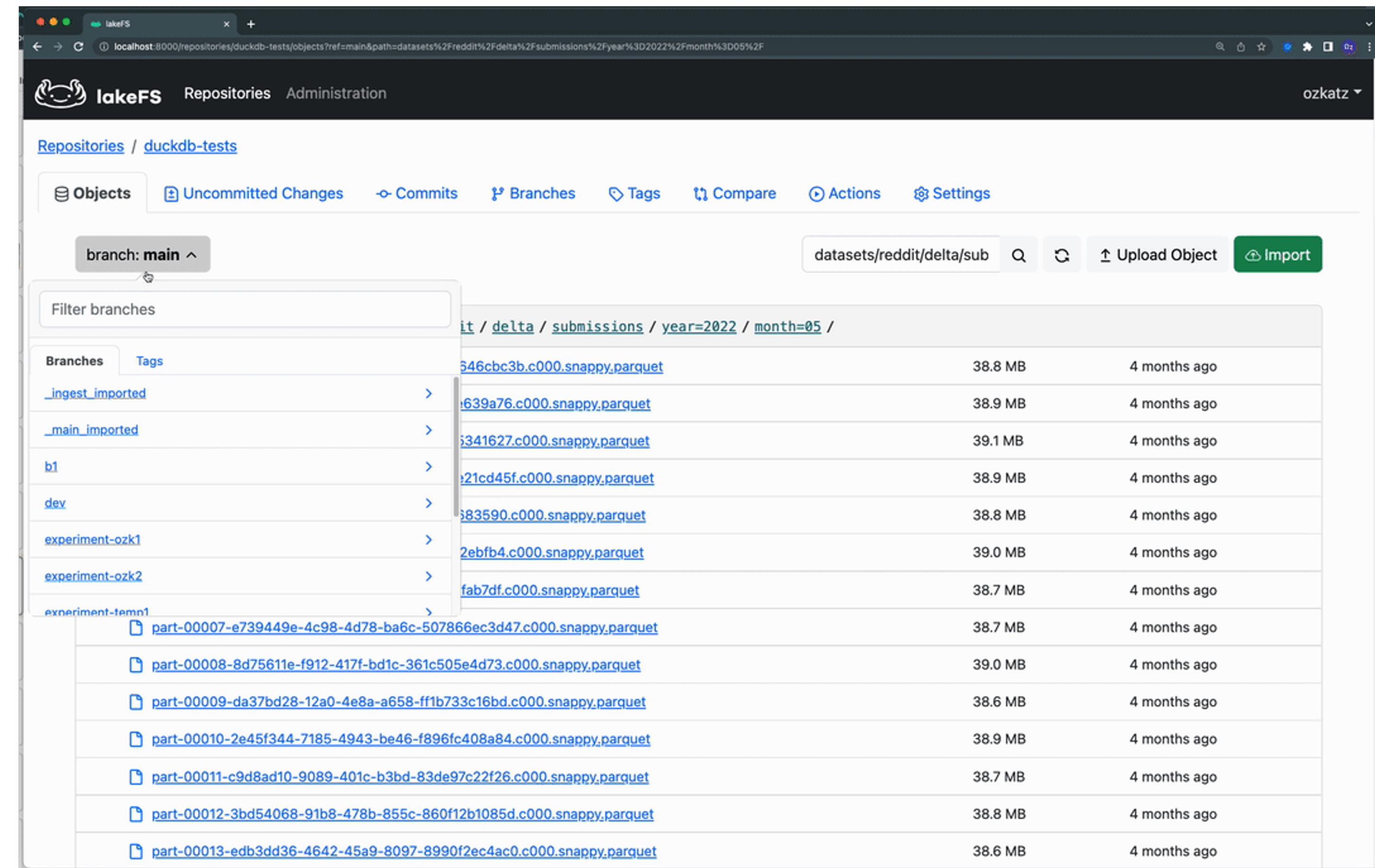
experiment branch



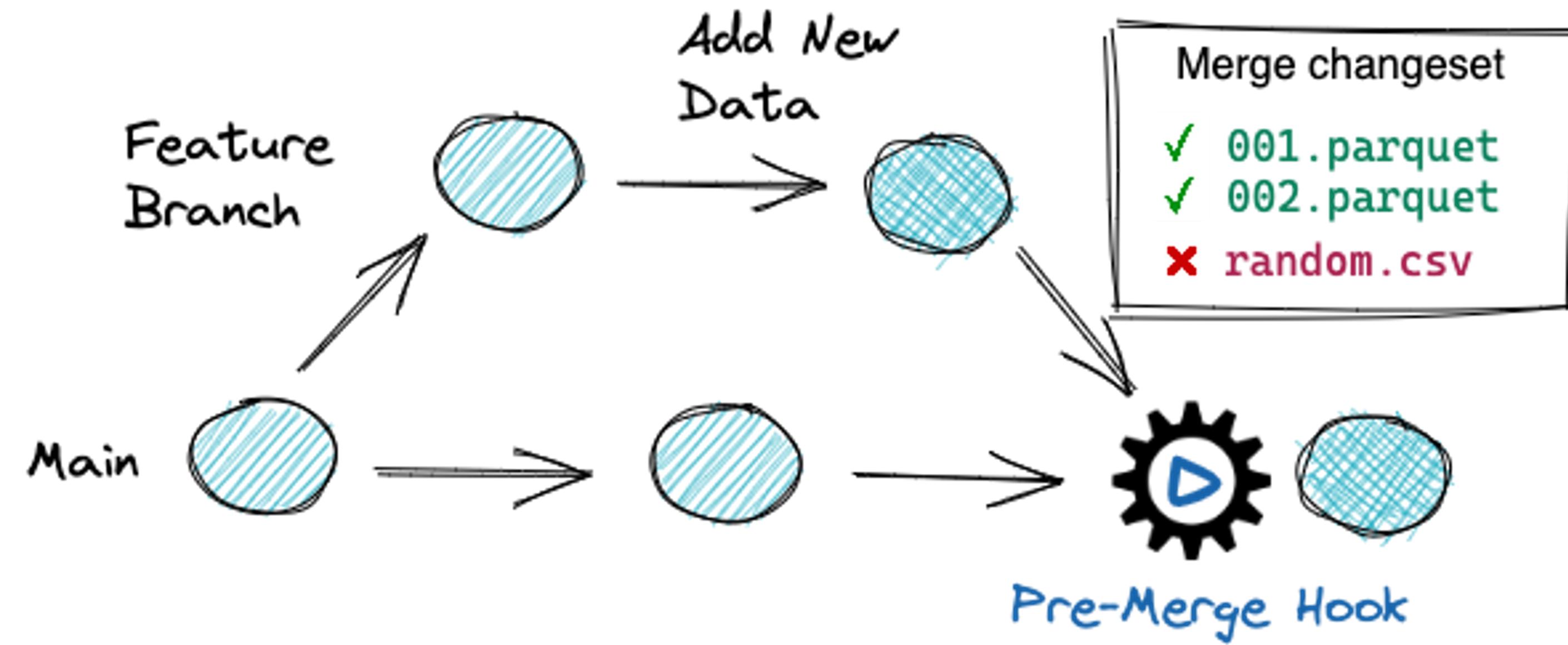
```
df = spark.read.parquet(  
    'lakefs://my-repo/ml-trained-2022-06/inputs/vists/')
```

Merged estimation model using new learning set	
hyper-params	{ ... }
training-job-uri	github.com/my-ml/a34902
infra-uri	github.com/my-terraform/..

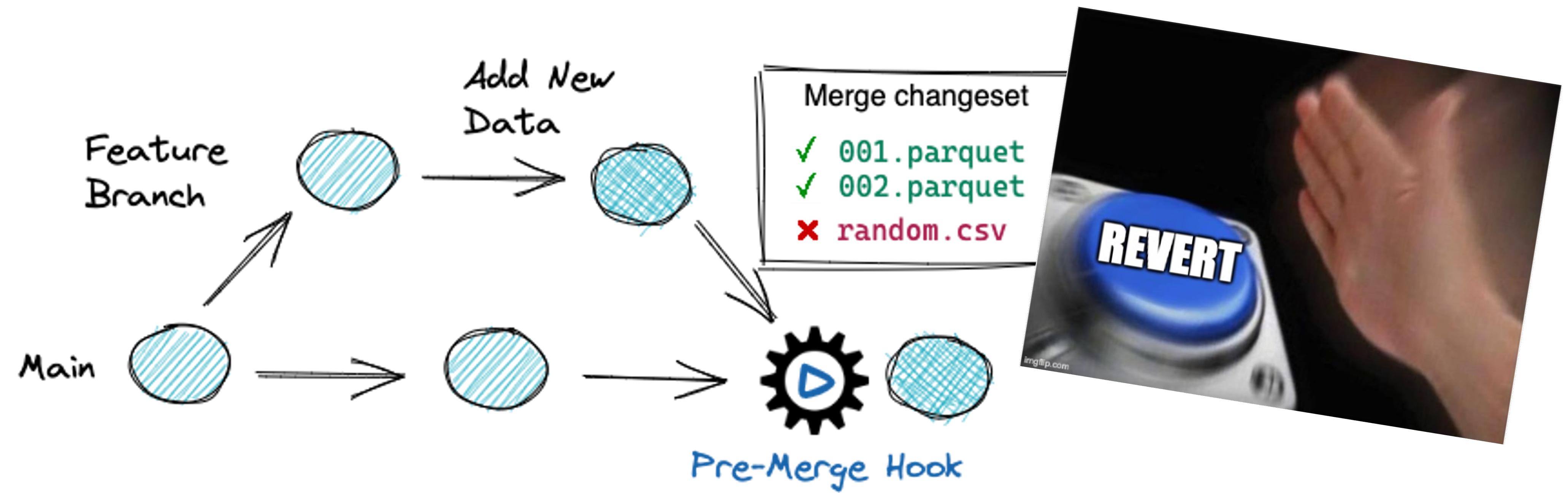
Dev/Test is notoriously difficult impossible



As a result, so is ensuring production safety



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 🛡️

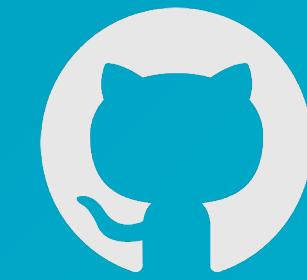
Learn More



lakefs.io



<https://lakefs.io/community>



github.com/treeverse/lakeFS



THANK YOU!

Oz & Lottie the Axolotl