

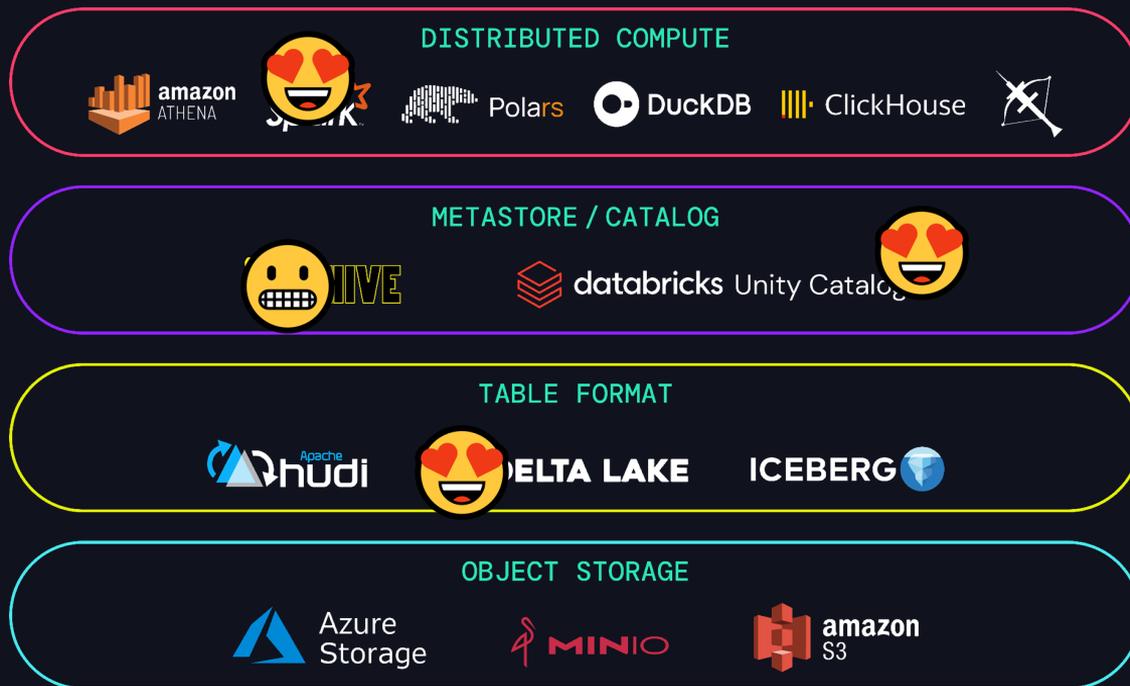
# Why Version Control Is Essential For Your Lakehouse Architecture

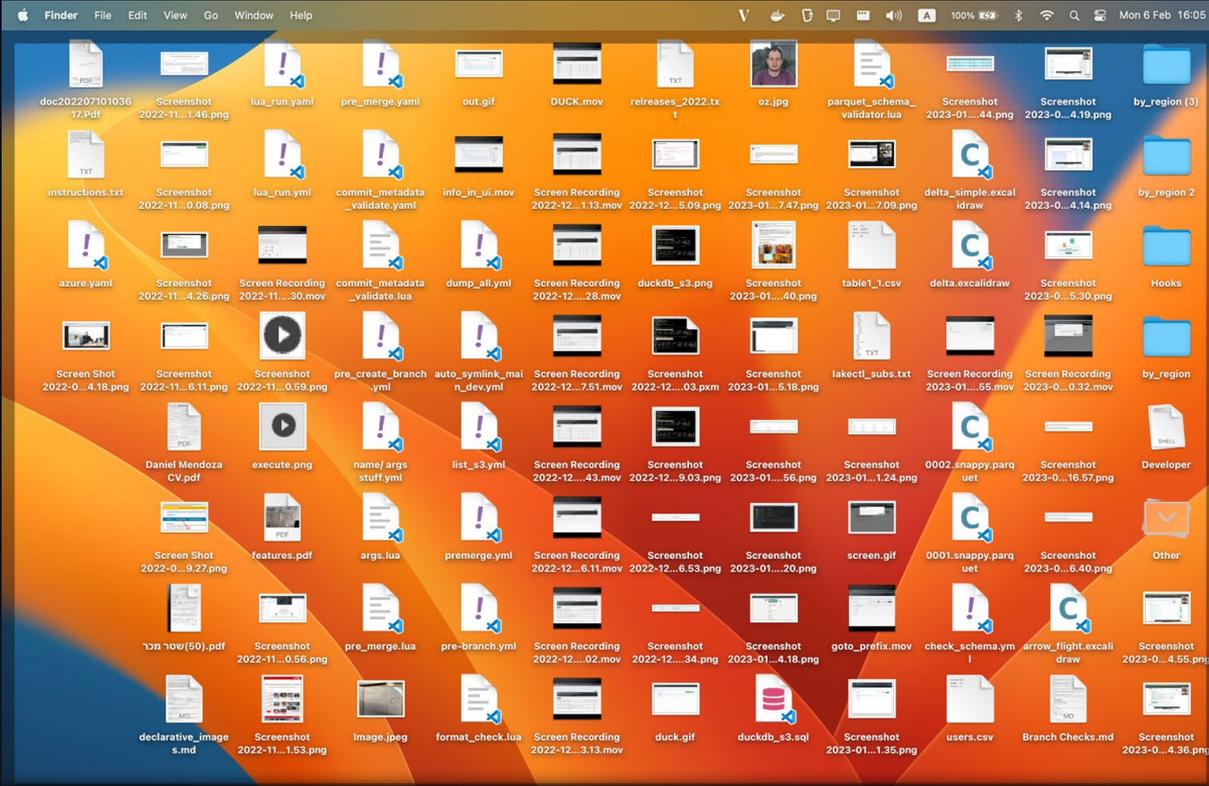
---

Oz Katz, June 2024

# DATA LAKEHOUSE

Humans, organizing things as humans tend to do.





YES,  
THIS IS MY  
ACTUAL  
DESKTOP



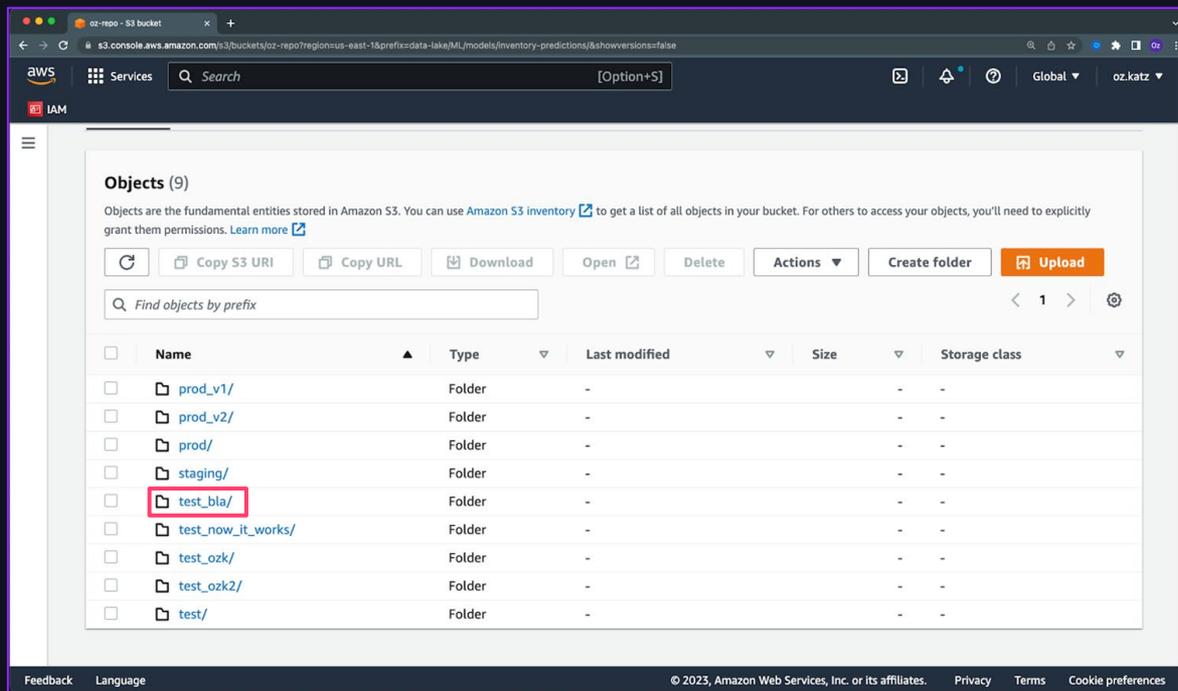
# HUMANS

ARE CREATIVE, MESSY

$$\text{mess} = \text{len}(\text{engineers}) * \text{len}(\text{data\_scientists})^2$$


# HUMANS

## NAMING THINGS



# YOUR CODE (ANOTHER HUMAN ARTIFACT)

**\*IS\* MUCH BETTER, I ASSURE YOU**

[Spark] Conditionally check for presence of delta file at checkpoint version #3080 <> Code

spark/src/test/scala/org/apache/spark/sql/delta/IdentityColumnIngestionSuite.scala

Comment on line 74  
75

**All checks have passed** Hide all checks  
8 successful checks

- ✓ **Delta Connectors Tests / Run tests (2.13.13) (pull\_request)** Successful in 16m Details
- ✓ **Delta Kernel Tests / test (pull\_request)** Successful in 11m Details
- ✓ **Delta Spark Master Tests / test (2.13.13) (pull\_request)** Successful in 171m Details
- ✓ **Delta Spark Tests / test (2.12.18) (pull\_request)** Successful in 257m Details
- ✓ **Unidoc generation / Generate unidoc (2.13.13) (pull\_request)** Successful in 5m Details
- ✓ **Delta Connectors Tests / Run tests (2.12.18) (pull\_request)** Successful in 38m Details

✓ **This branch has no conflicts with the base branch**  
Only those with [write access](#) to this repository can merge pull requests.

sumeet-db requested a review from prakharjain09 2 weeks ago

**BUT DATA IS  
HARDER**

# WHERE DO I RUN TESTS?

Objects (9)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 Inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Find objects by prefix

Name	Last modified	Size	Storage class
prod_v1/	-	-	-
prod_v2/	-	-	-
prod/	-	-	-
staging/	-	-	-
test_bla/	-	-	-
test_now_it_works/	-	-	-
test_ozk/	-	-	-
test_ozk2/	-	-	-
test/	-	-	-

Feedback Language © 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

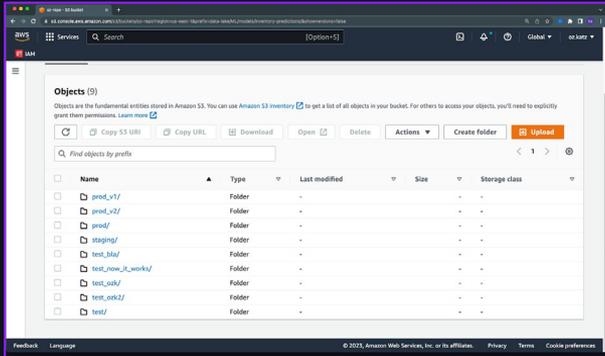
# WHEN DO I RUN TESTS?

(SPOILER: TOO LATE)

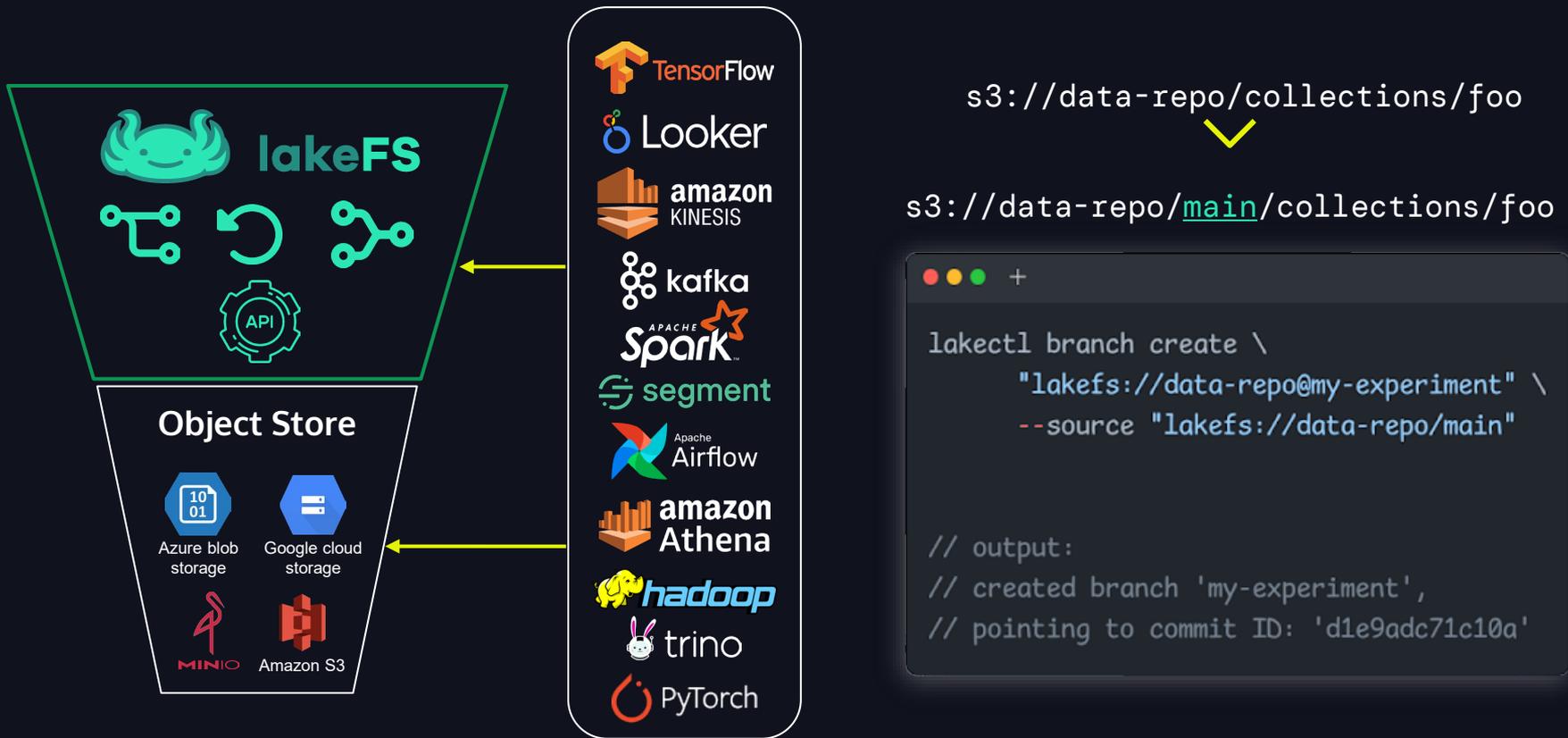


# WHAT IF...

# WE COULD TO DO THIS:



# WE CAN!



# WE CAN!

Transactional



Apache Spark

Resilient



Delta Lake

Durable



lakeFS



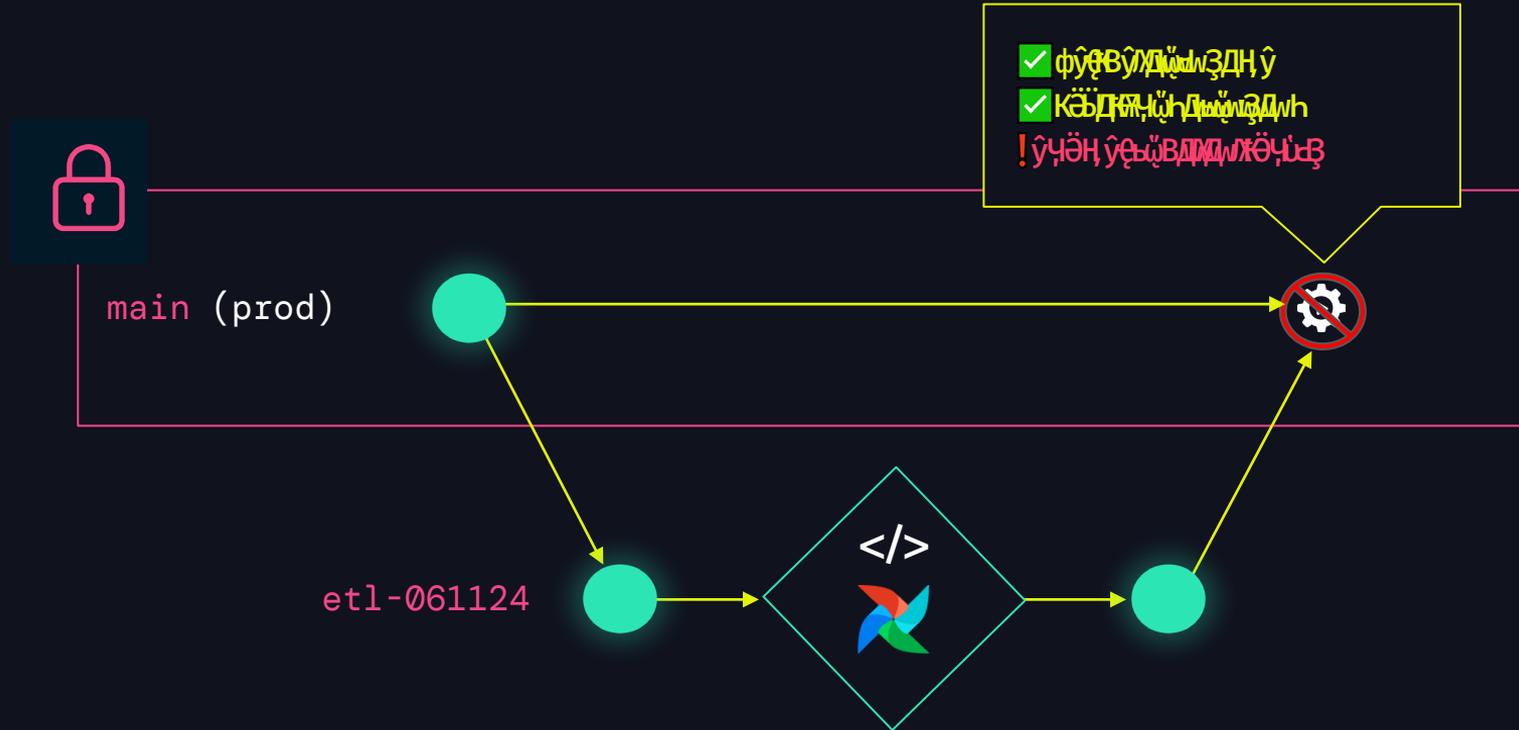
AWS S3 / ADLS / GCS / MinIO / ...

<https://docs.lakefs.io/integrations/spark.html>



# WHERE & WHEN DO I RUN TESTS?

ON A BRANCH, BEFORE MERGING



# WHERE DO I RUN TESTS?

## ON A BRANCH, BEFORE MERGING

### PYSPARK



```
repo = lakefs.repository('my-data-lake')
branch = repo.branch('etl-061124').create(source_reference='main')
```



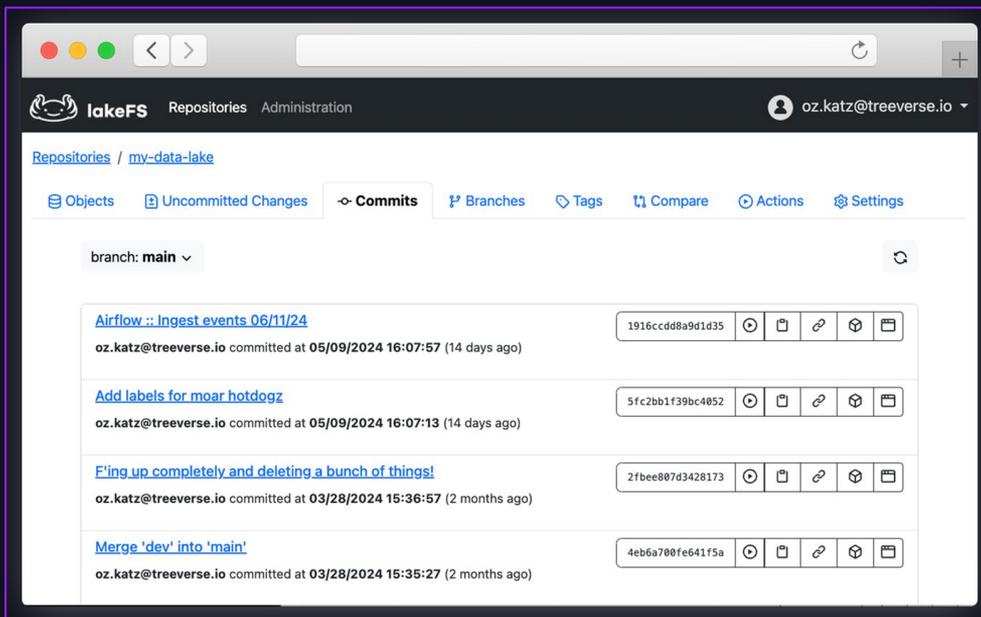
```
# write
events_path = f lakefs://my-data-lake/{branch.id}/bronze/events/'
dt = DeltaTable.forPath(spark, events_path)
dt.update(condition=expr('id > 100'), set={'id': expr('id % 100')})
...

# audit(!!)
run_checks(events_path)

# publish
branch.commit(message='Events Transformed!')
branch.merge_into('main') # ATOMIC & WILL FAIL IF TESTS DIDN'T PASS
```

# COOL! WHAT ELSE CAN IT DO?

## LOTS OF GIT-Y STUFF



■ **Reproducibility** - using immutable commits

■ **Rollback to last known good state** - lowering cost of mistakes in prod

■ **Multiple dev environments** - without copying data, using copy-on-write branching

# THANK YOU

Oz & your production DAGs

# LEARN MORE

 [lakefs.io/slack](https://lakefs.io/slack)

 [github.com/treeverse/lakeFS](https://github.com/treeverse/lakeFS)

 [docs.lakefs.io/quickstart](https://docs.lakefs.io/quickstart)



Come meet me  
at Booth #69

EXPO HALL

