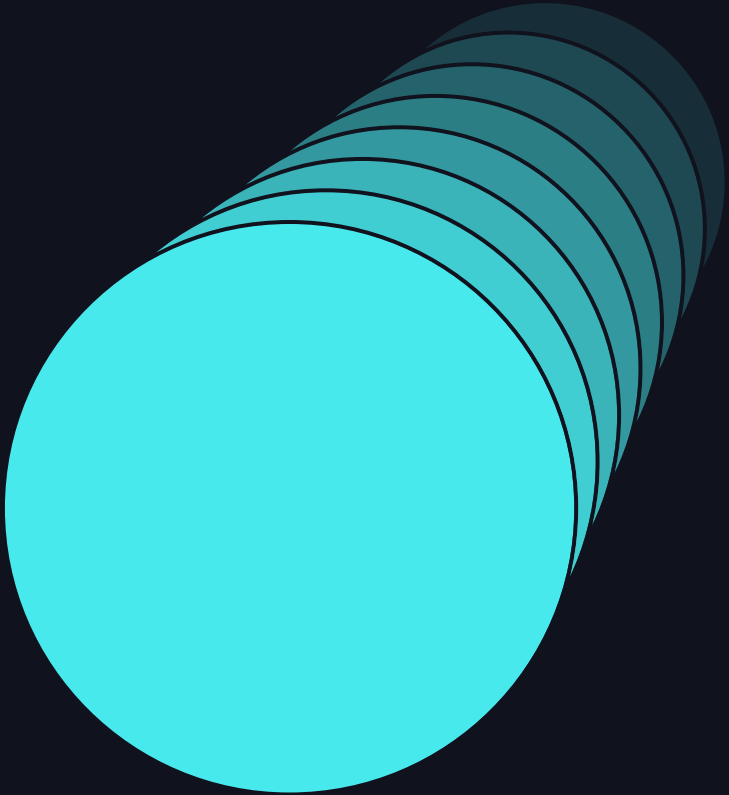
DATA+AI SUMMIT
BY databricks

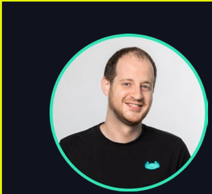
Power Up Your Lakehouse with Git Semantics & Delta Lake

Oz Katz, June 2024



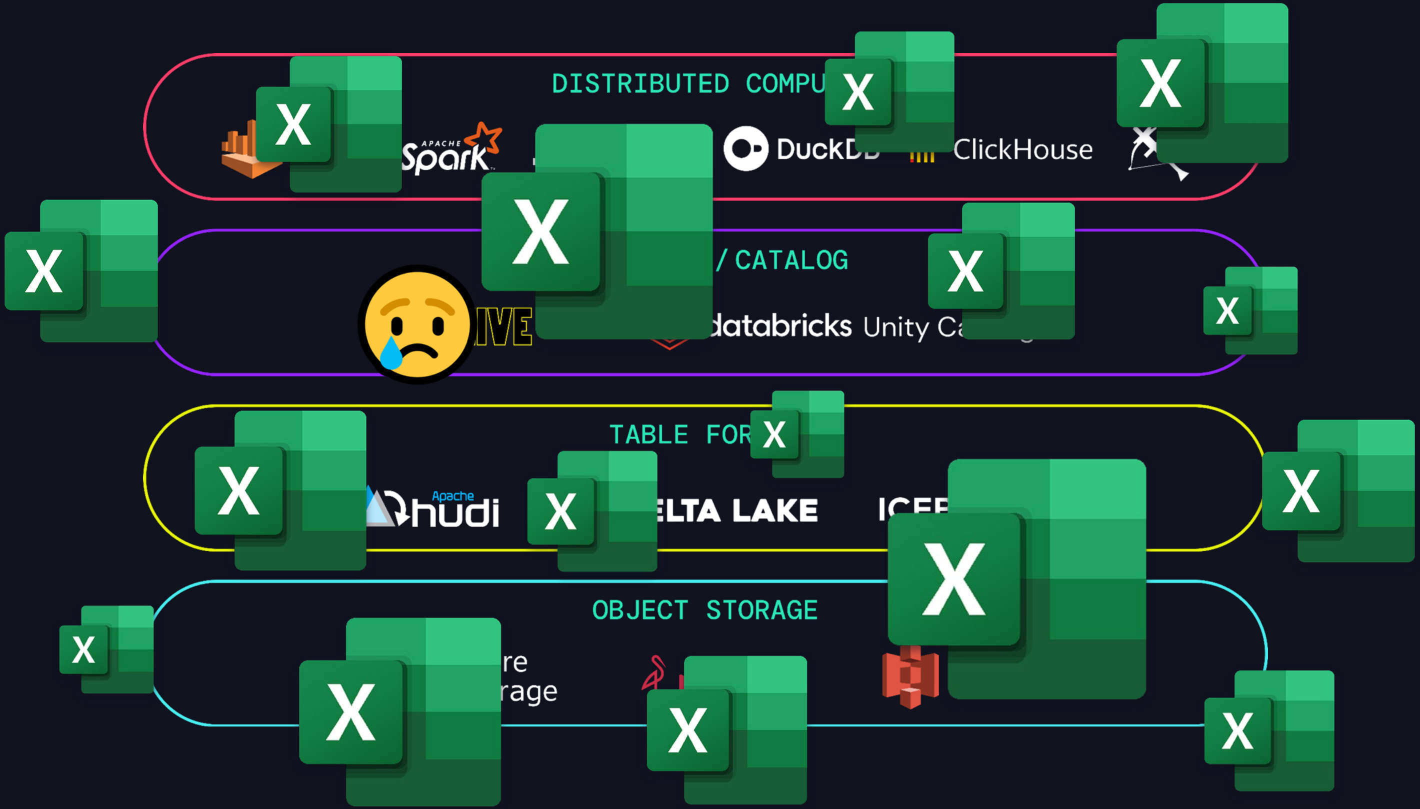
Always has been
So S3 is just a big shared folder?

HI 👋

Oz Katz	Lottie the Axolotl
 CTO, Co-Creator @ lakeFS	CCO, Mascot @ lakeFS
github.com/ozkatz	github.com/treeverse/lakeFS
𝕏 @ozkatz100	linkedin.com/company/ lakefs-treeverse/

YOUR DATA LAKE

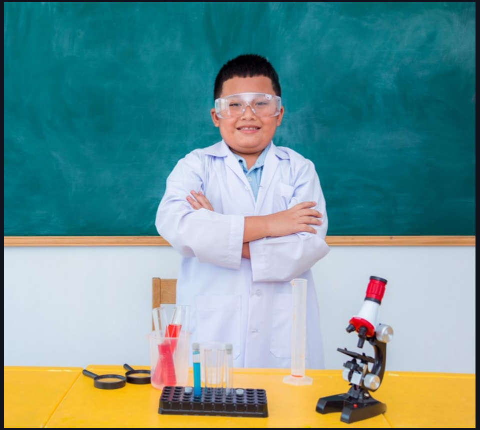
Might look something like this



Data Lake Resiliency

Ingredients

Reproducible outputs



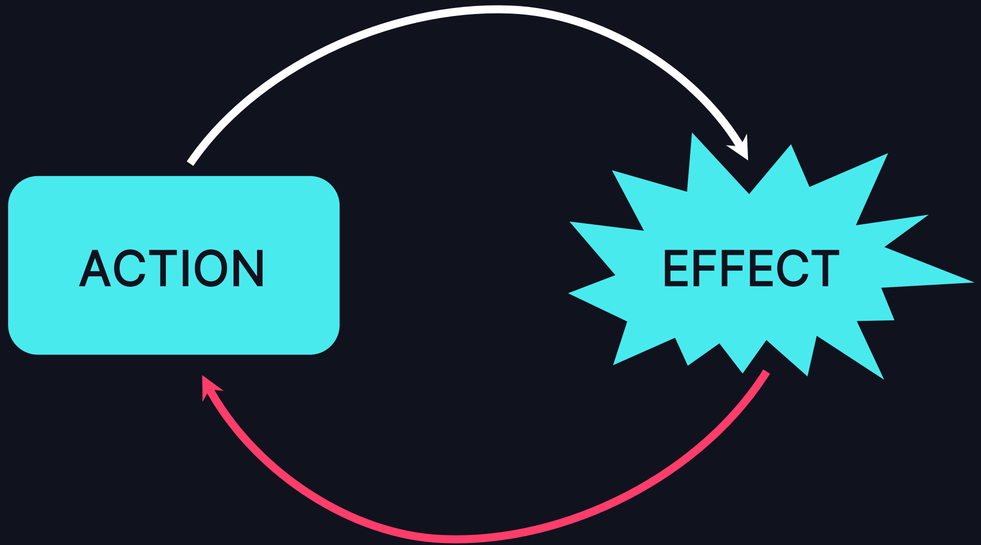
Write-Audit-Publish

A big red button

REPRODUCIBLE OUTPUTS

Why should we care?

- code + data = output
- Improving either requires a feedback loop!



But also

ACHIEVING REPRODUCIBILITY

Thank You, Delta Lake Time Travel

```
SELECT * FROM image_classification
VERSION AS OF 5238
WHERE label <> 'Hot Dog'
```

Hot dogs over time

ACHIEVING REPRODUCIBILITY

Across Tables?

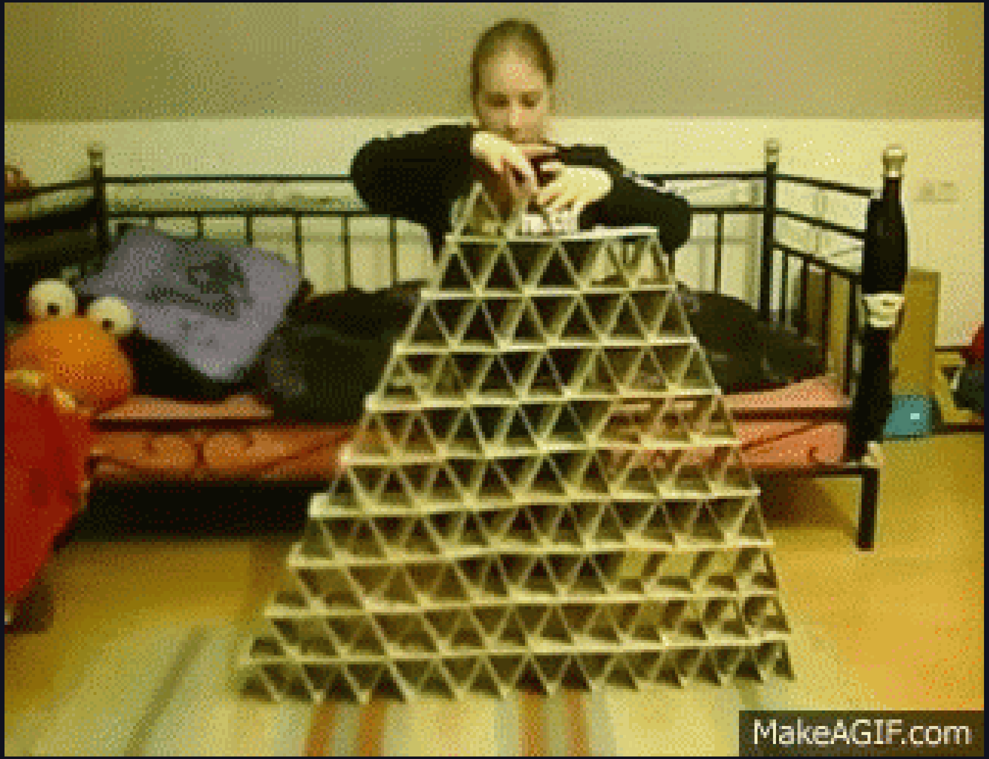
```
SELECT * FROM images VERSION AS OF 5238
JOIN labels VERSION AS OF 83111 ON …
JOIN image_licensing VERSION AS OF 3489 ON …
JOIN privacy_settings VERSION AS OF 2371 ON …
JOIN exif_info VERSION AS OF 45821 ON …
WHERE label <> 'Hot Dog'
```

Hot dogs over time

WRITE-AUDIT-PUBLISH

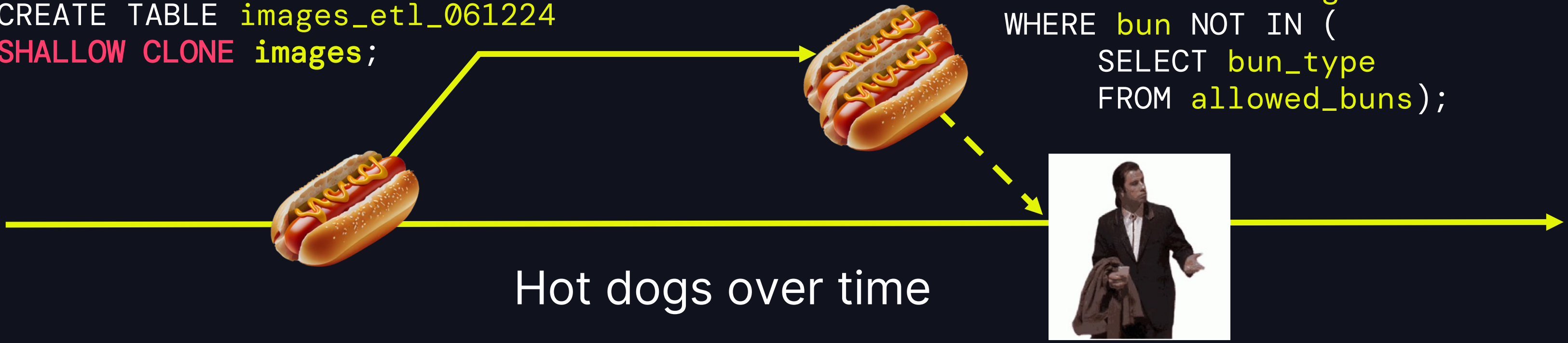
Why should we care?

- All datasets have downstream consumers
- All failures cascade
- But successes cascade *harder*



ACHIEVING WRITE-AUDIT-PUBLISH

Delta Lake Shallow Clones



```
INSERT INTO images_etl_061224 …
```

```
CREATE TABLE images_etl_061224
SHALLOW CLONE images;
```

```
SELECT 1 FROM images_etl_061224
WHERE bun NOT IN (
     SELECT bun_type
     FROM allowed_buns);
```

Hot dogs over time

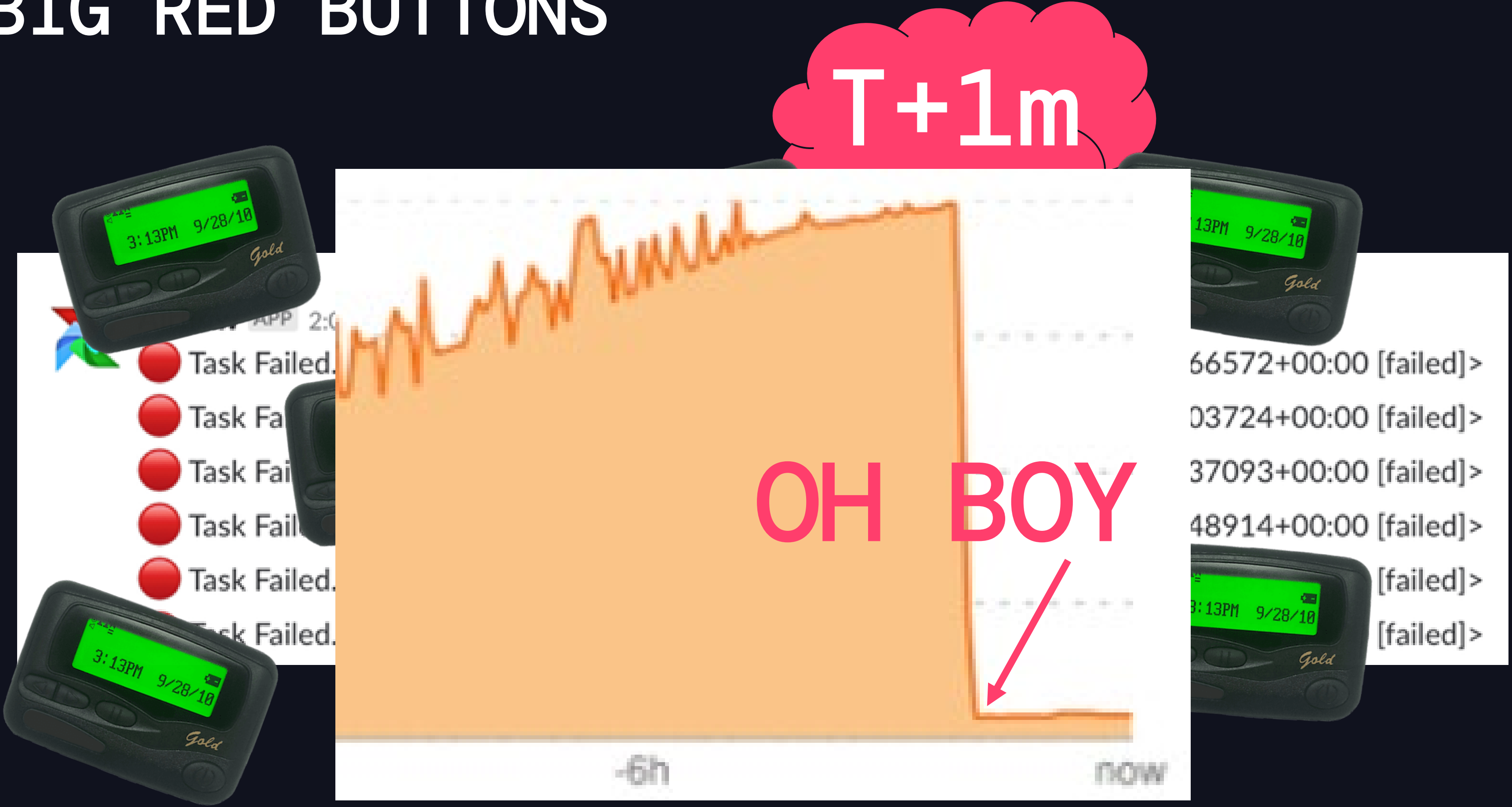
BIG RED BUTTONS

A story about how **someone** completely F***ed production

BIG RED BUTTONS

T-3h

BIG RED BUTTONS



T+1m

OH BOY

APP 2:(

Task Failed.

Task Fa

Task Fai

Task Fail

Task Failed.

...sk Failed.

66572+00:00 [failed]>

03724+00:00 [failed]>

37093+00:00 [failed]>

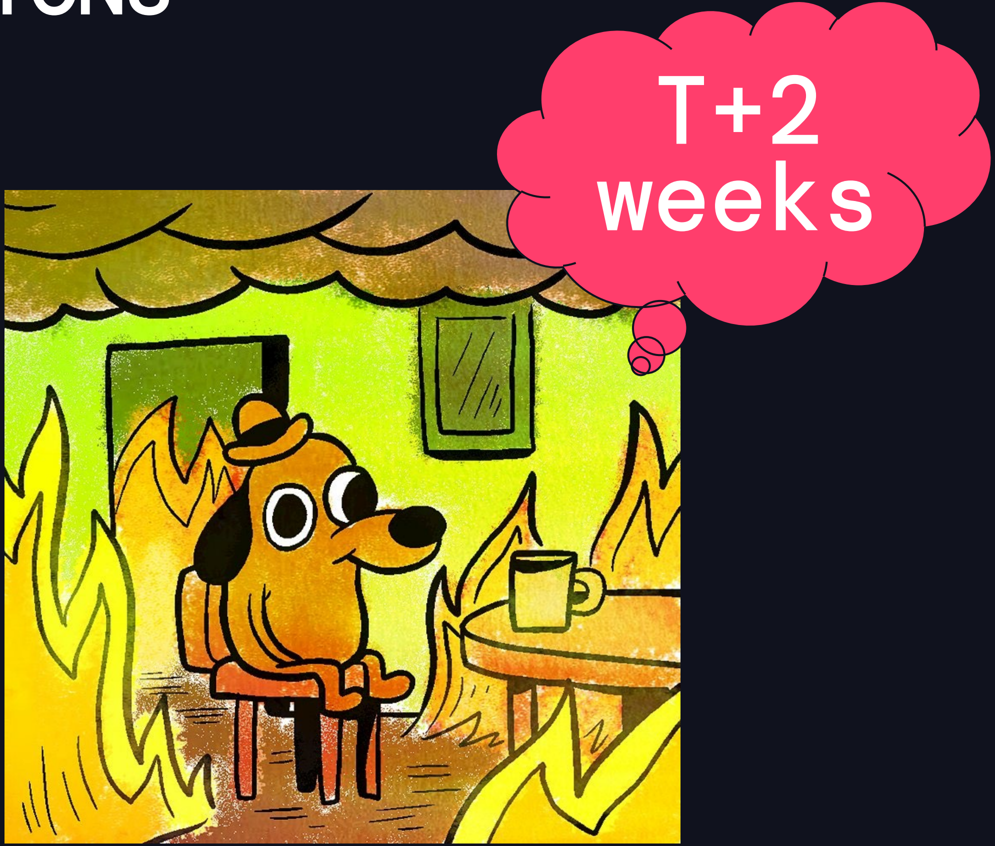
48914+00:00 [failed]>

[failed]>

[failed]>

BIG RED BUTTONS

BIG RED BUTTONS



BIG RED BUTTONS

Delta Lake’s RESTORE… AS OF

```
RESTORE TABLE images
TO VERSION AS OF 2392;
```

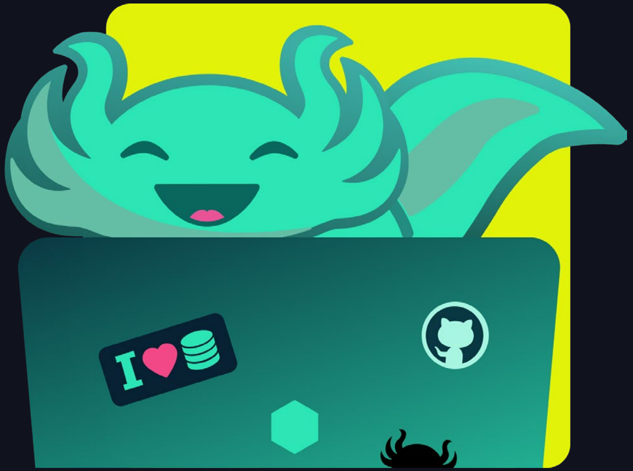
Hot dogs over time

ACHIEVING PROD ROLLBACKS

WITH 2,340 PRODUCTION TABLES

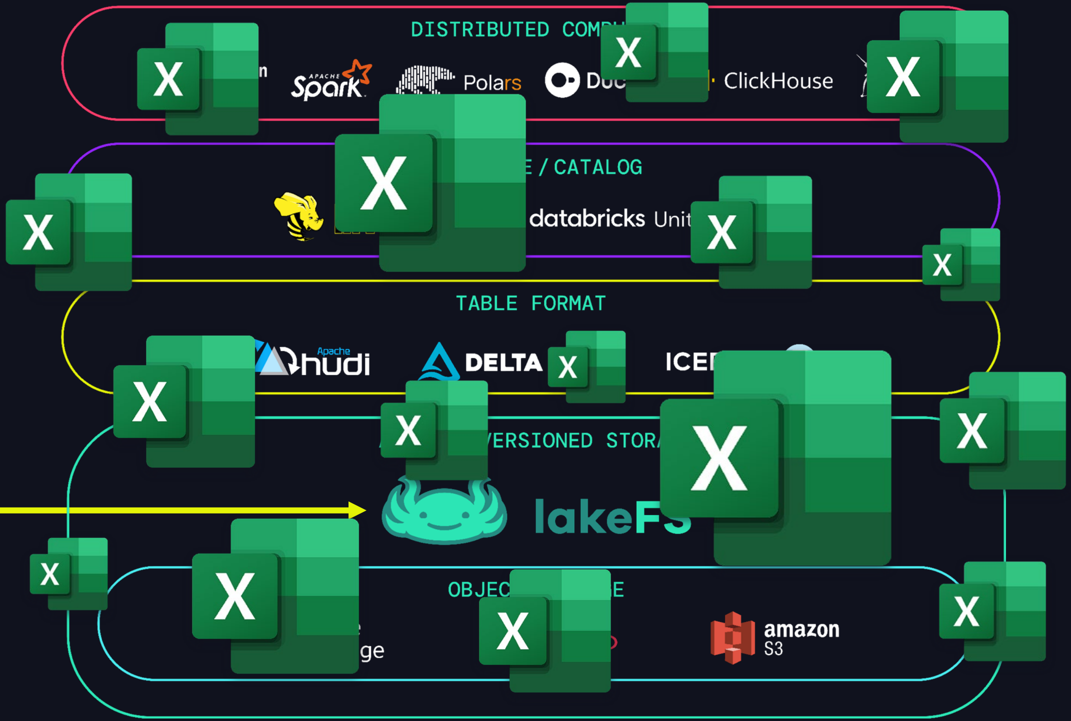
POWERING UP WITH GIT SEMANTICS

REMEMBER ME?

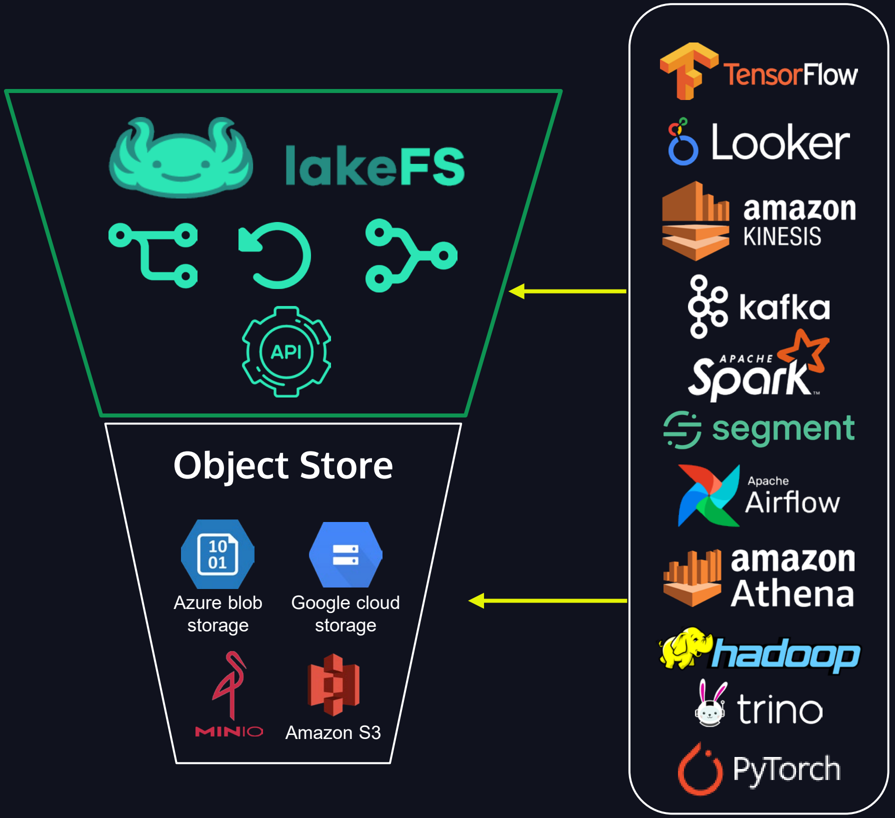


lakeFS IN 30 SECONDS OR ~~LESS~~ MORE

This is where
I fit in

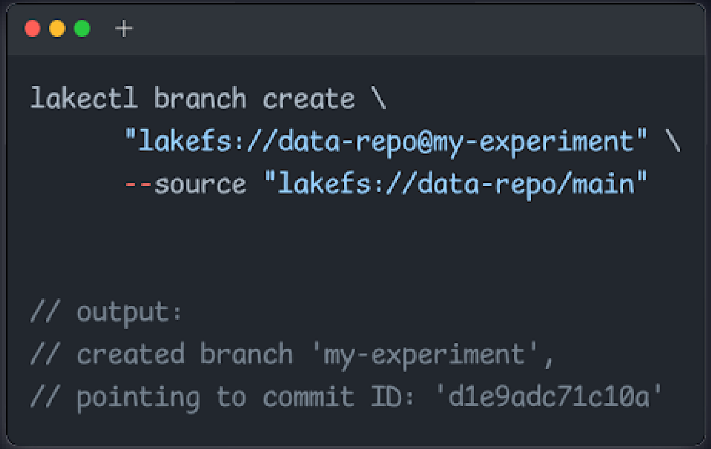


LAKEFS IN 30 SECONDS OR ~~LESS~~ MORE



s3://data-repo/collections/foo

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

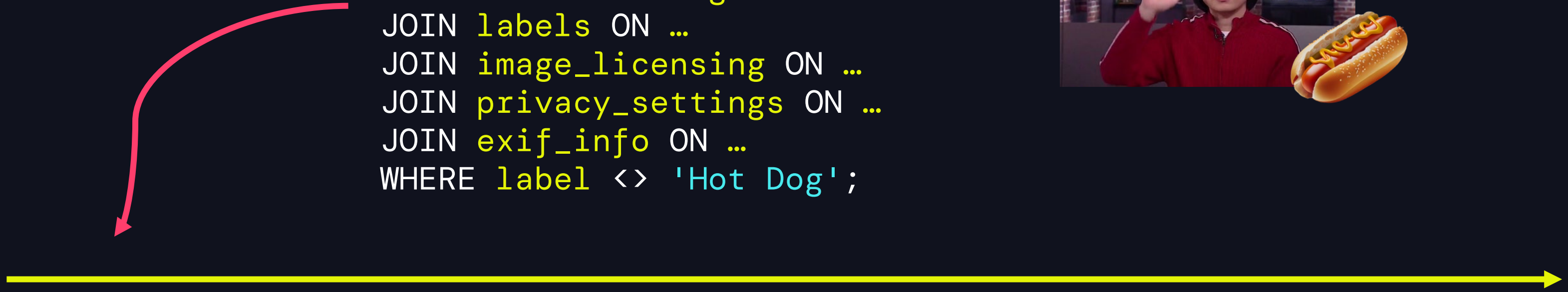


```
lakectl branch create \
        "lakefs://data-repo@my-experiment" \
        --source "lakefs://data-repo/main"


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

ACHIEVING REPRODUCIBILITY

Across Tables!

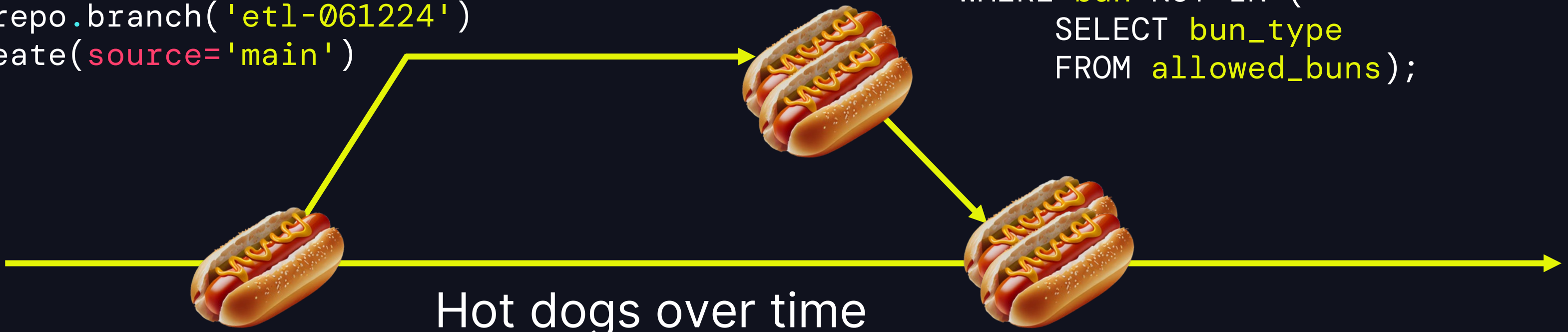
```
USE unity.model_v15_prod;

SELECT * FROM images
JOIN labels ON …
JOIN image_licensing ON …
JOIN privacy_settings ON …
JOIN exif_info ON …
WHERE label <> 'Hot Dog';
```



Hot dogs over time

ACHIEVING WRITE-AUDIT-PUBLISH

Branches



```
branch = repo.branch('etl-061224')
branch.create(source='main')
```

```
INSERT INTO images …
```

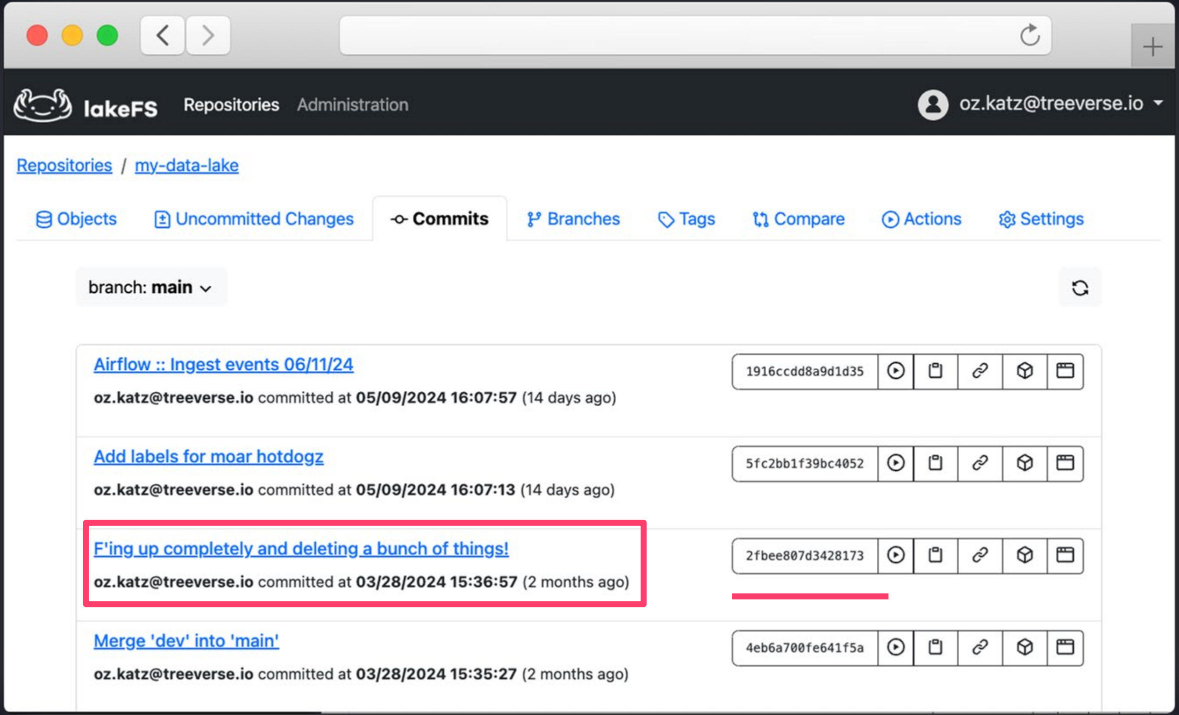
```
SELECT 1 FROM images
WHERE bun NOT IN (
      SELECT bun_type
      FROM allowed_buns);
```

Hot dogs over time

```
branch.merge(into='main')
```

BIG RED BUTTON

Revert commit!



Repositories / my-data-lake

Airflow :: Ingest events 06/11/24
oz.katz@treeverse.io committed at **05/09/2024 16:07:57** (14 days ago)
1916ccdd8a9d1d35

Add labels for moar hotdogz
oz.katz@treeverse.io committed at **05/09/2024 16:07:13** (14 days ago)
5fc2bb1f39bc4052

F'ing up completely and deleting a bunch of things!
oz.katz@treeverse.io committed at **03/28/2024 15:36:57** (2 months ago)
2fbee807d3428173

Merge 'dev' into 'main'
oz.katz@treeverse.io committed at **03/28/2024 15:35:27** (2 months ago)
4eb6a700fe641f5a

BIG RED BUTTON

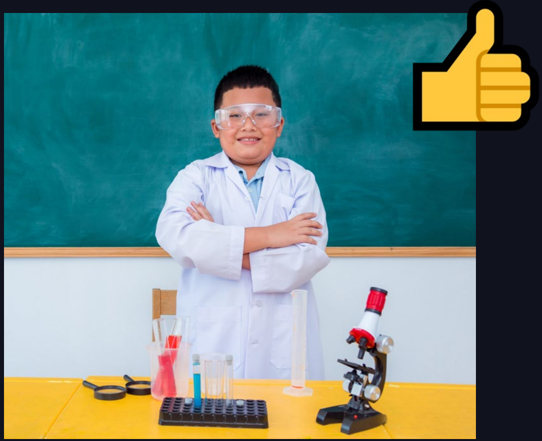
On any number of tables, atomically

PYTHON

```
import lakefs

repo = lakefs.repository('my-
main = repo.branch('main')

bad_commit_id = '2fbee807d342
main.revert(repo.ref(bad_comm
```

Data Lake Resiliency

Reproducible outputs



Write-Audit-Publish

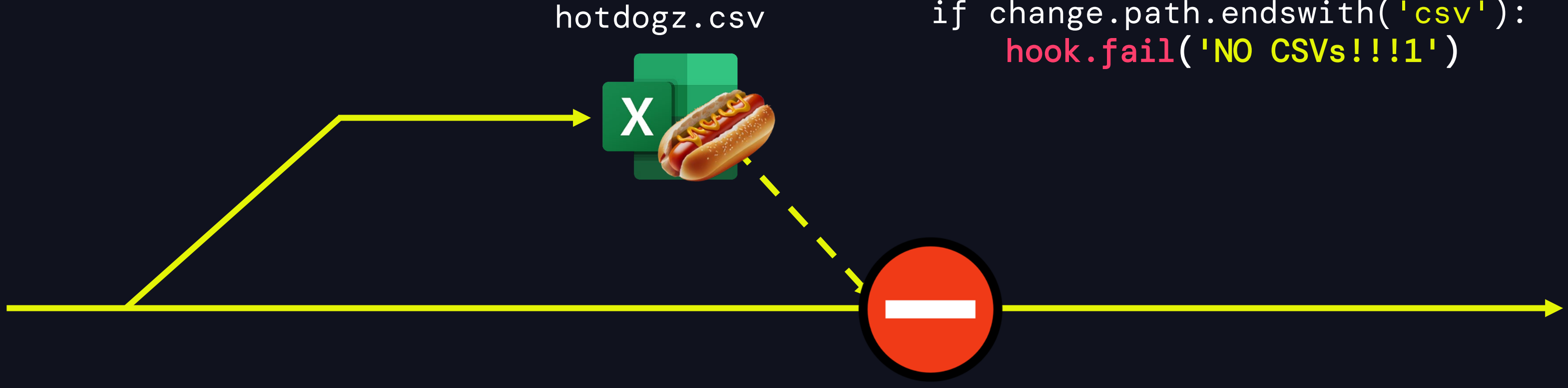
A big red button

BONUS: HOOKS

hotdogz.csv

BONUS: HOOKS

```
_lakefs_actions/pre-merge.py:

for change in source.diff('main'):
    if change.path.endswith('csv'):
        hook.fail('NO CSVs!!!1')
```



Hot dogs over time

THANK YOU

Oz & your production DAGs

LEARN MORE

lakefs.io/slack

github.com/treeverse/lakeFS

docs.lakefs.io/quickstart