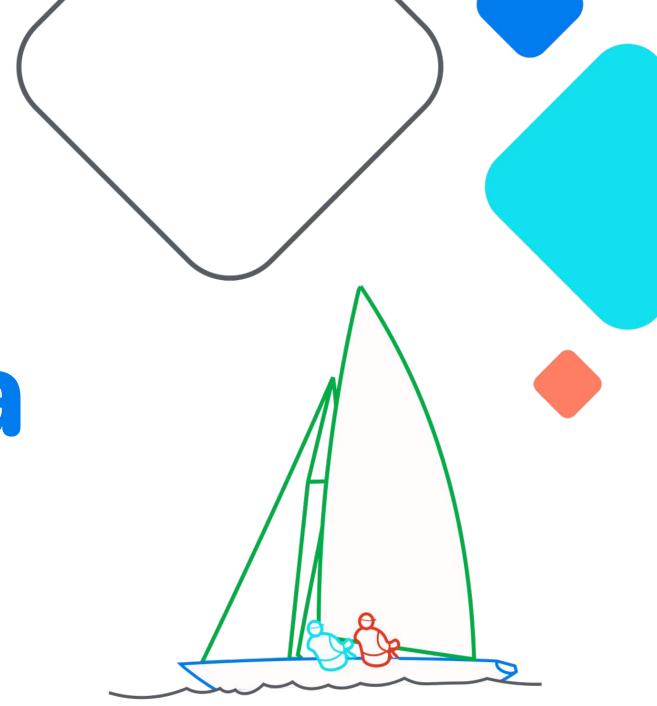


Supercharge Your Data Testing with a Fully Open Stack

Iddo Avneri, lakeFS 



 **Airflow Summit**

Let's flow together

September 19-21, 2023,
Toronto, Canada

Data VS

That Grounded US Flights

The FAA said it has made necessary repairs to the system "and has taken steps to make the pilot message system "more resilient."

World News | Reuters | Updated: January 20, 2023 7:19 am IST



WE'RE TRYING THIS IN P

Code:

CI/CD in place.

Promote from UAT to P

Data:

Hope



TRENDING



"Playing For Country...": Gavaskar Blasts Gill For Calling Physio Mid-Ove



"Eknath Shinde Could Not Have Become Chief Minister If...": Supreme Court



"Mockery Of Test Cricket": India Legend's Scathing Verdict On Indore Pitch



Prince Harry, Meghan "Evicted" From Their Home On Windsor Estate: Report



In Telangana Death By Suicide, Student's Heart-Breaking



The nationwide groundstop on January 11 that disrupted more than 11,000 flights.



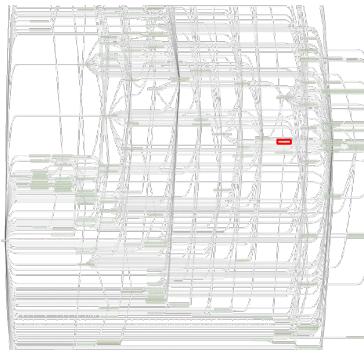
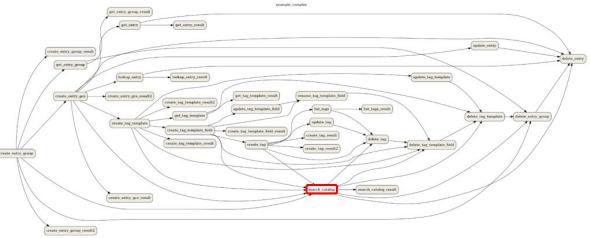
? Roll back the
sion

unit testing

ffort

Washington: The Federal Aviation Administration (FAA) said on Thursday a preliminary review found that contract personnel "unintentionally deleted files" disrupting a key computer system and prompting a nationwide groundstop on Jan. 11 that disrupted more than 11,000 flights.

It Gets Worse - A (not so fun) Day in a Life



Examine
Task Logs

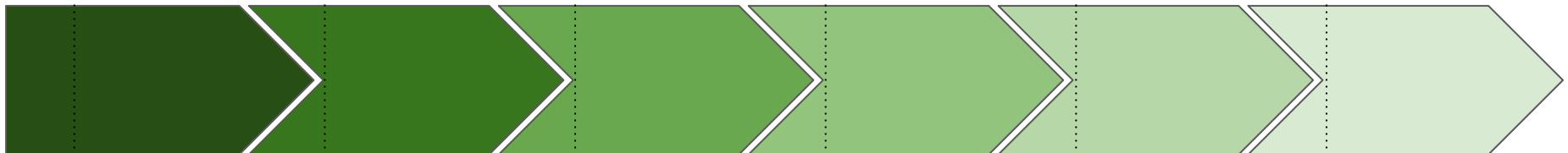
Identify The
Issue

Review
Data on
DAG's start

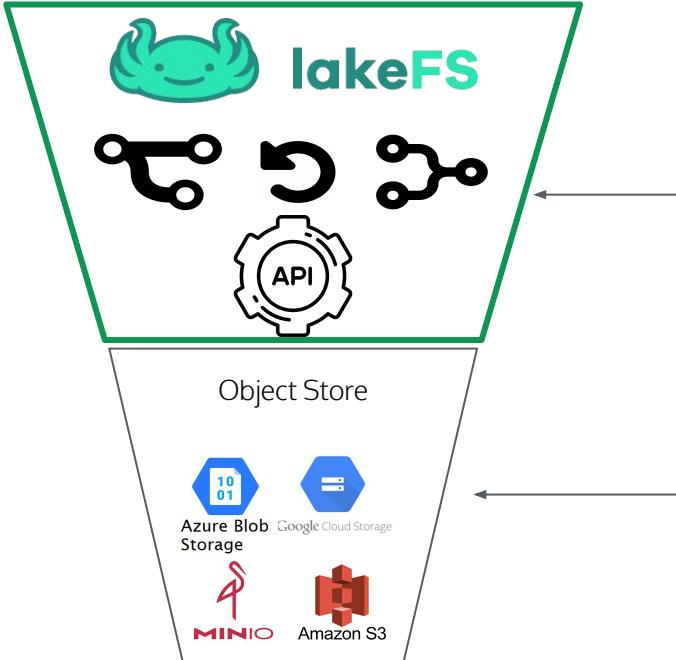
Troubleshoot
Data Evolution

Fix &
Re-Run
From Start

Hope & pray



Data as Code << In 20 Minutes >>

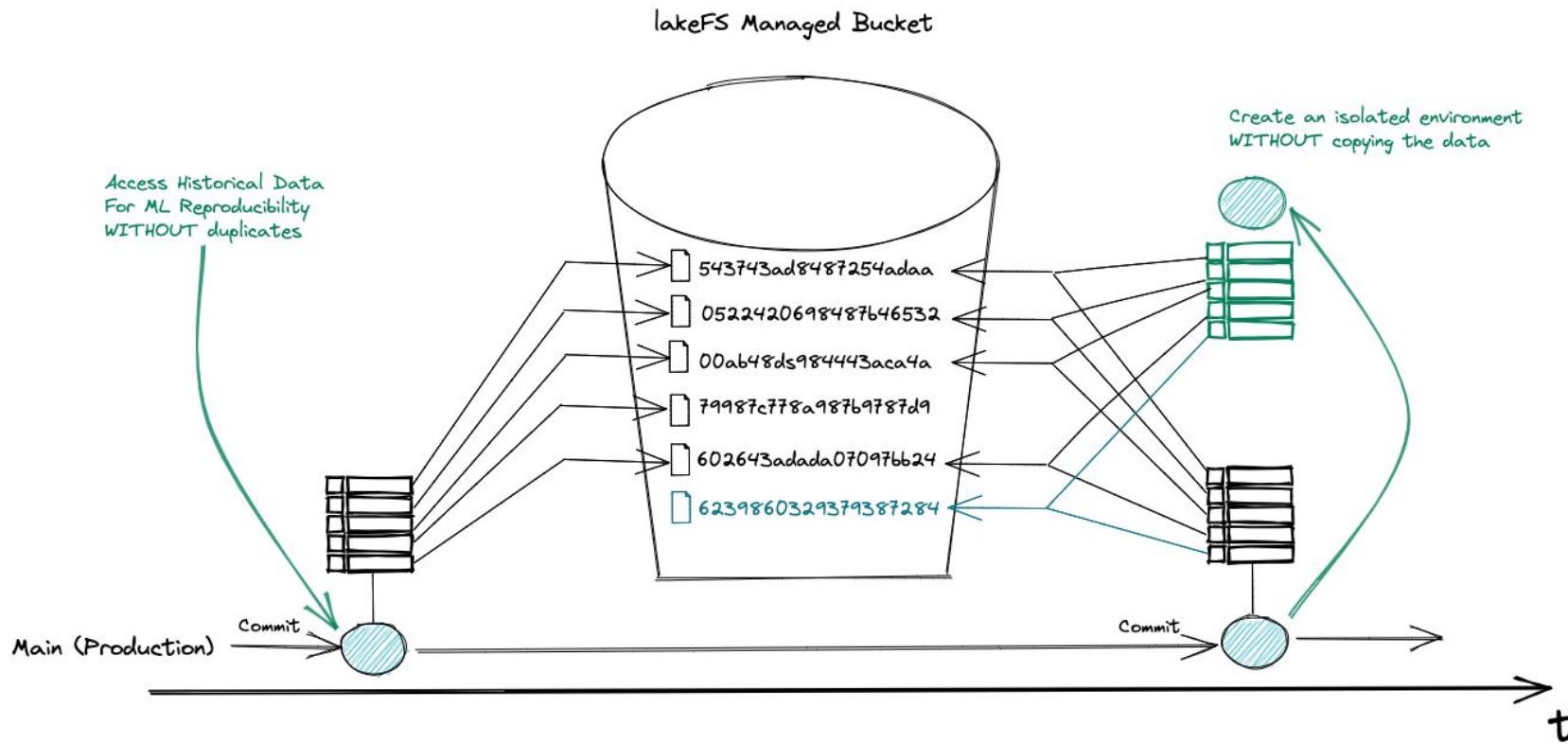


s3://data-repo/collections/foo
▼
s3://data-repo/main/collections/foo

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

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

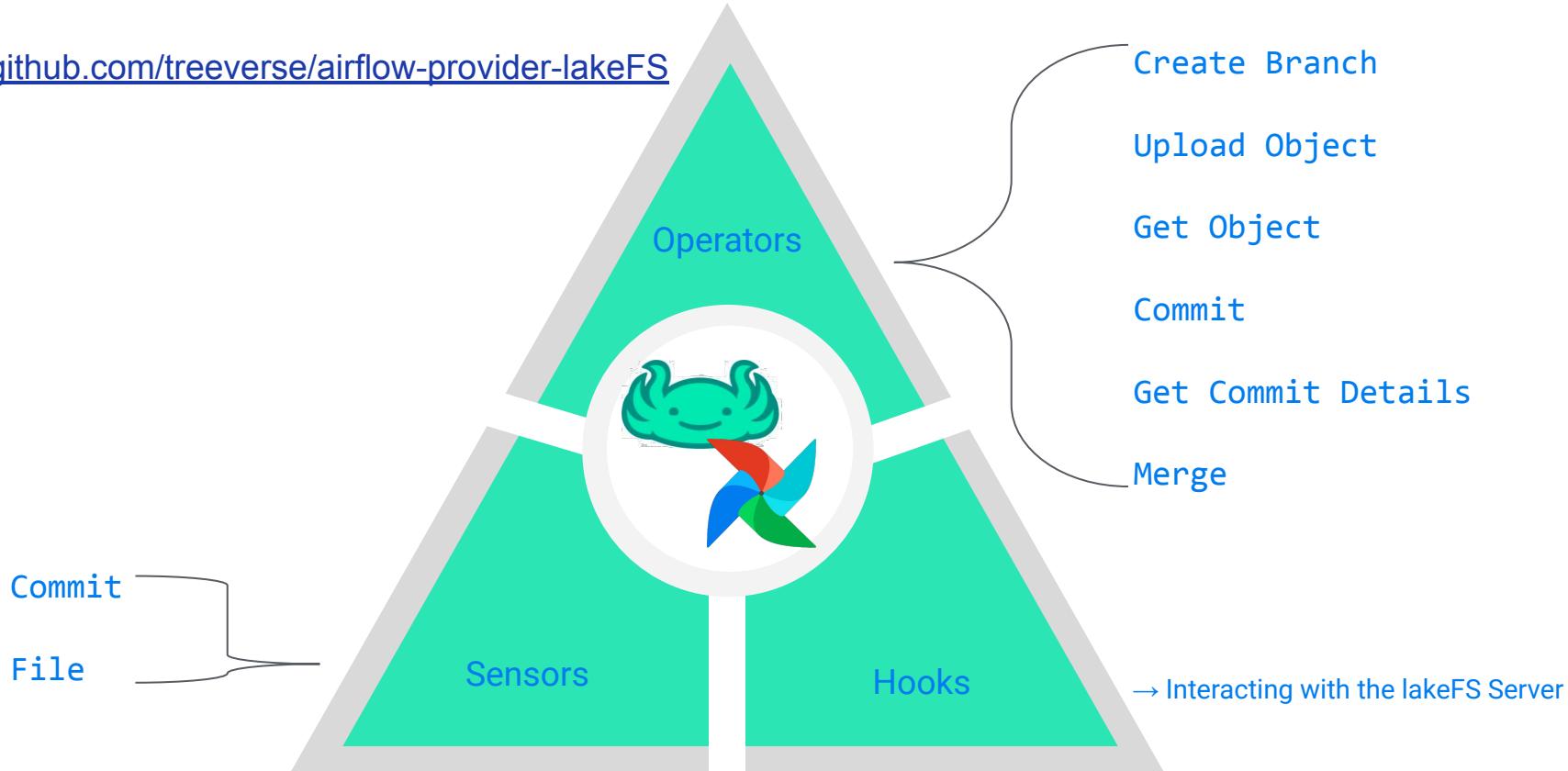
Data Version Control at Scale for Data Lakes



Demo 1: Create an isolated Dev Environment

Even better with airflow

<https://github.com/treeverse/airflow-provider-lakeFS>



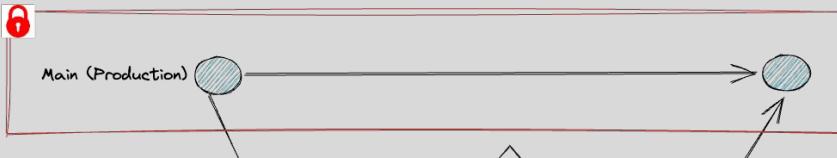
lakeFS's Python package: <https://pydocs.lakefs.io>

Quick Start, No Heavy Lift



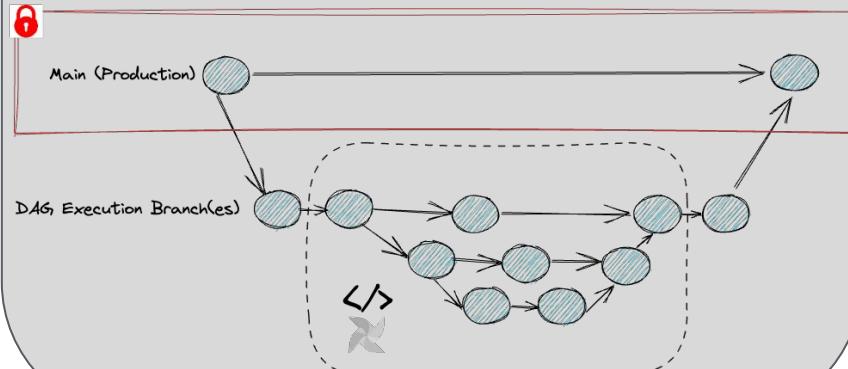
Existing DAGs

Branch before execution,
Commit & Merge after execution



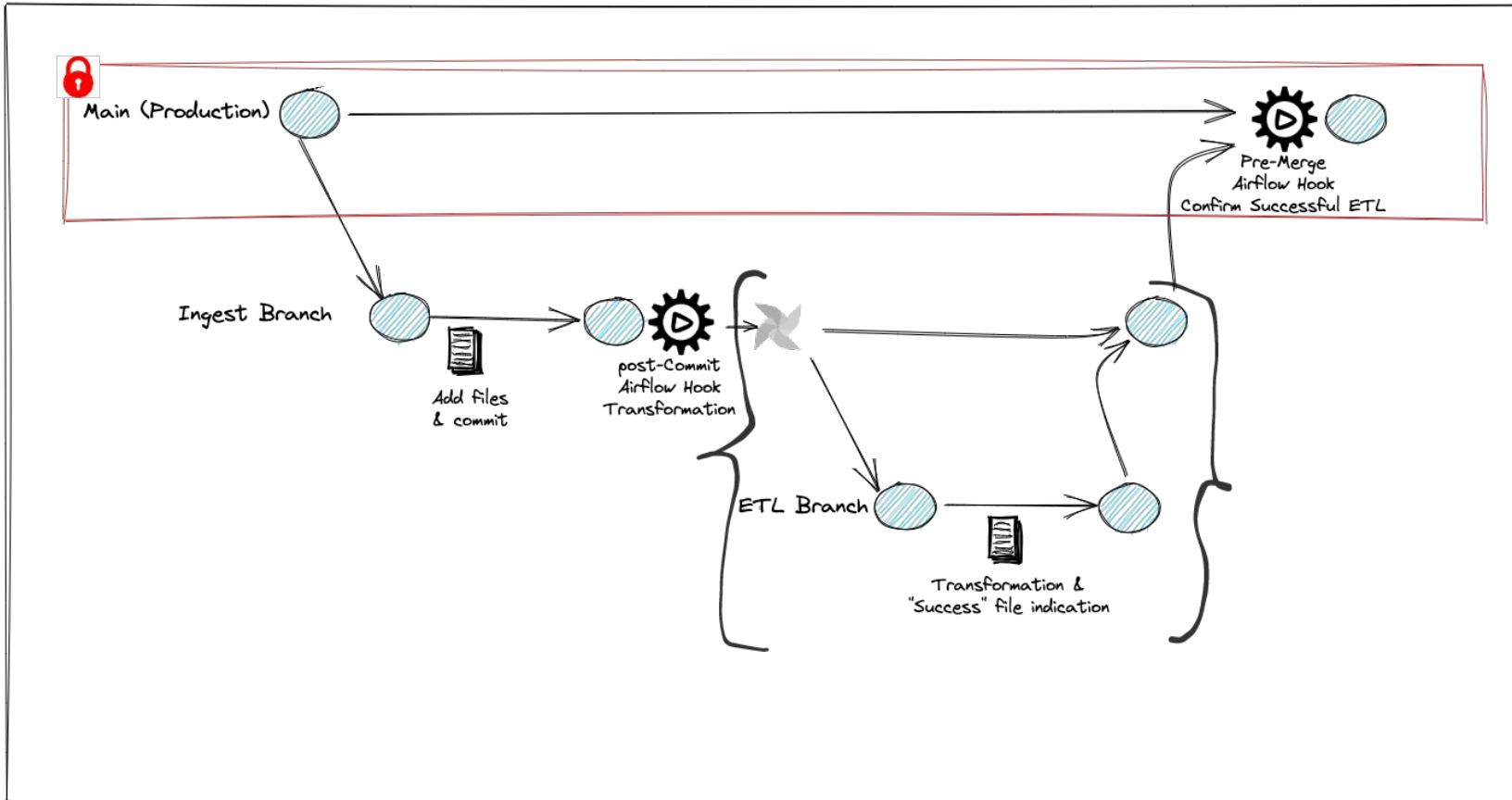
New DAGs

Branch & Commit inside the DAG
Utilizing lakeFS Airflow Operator

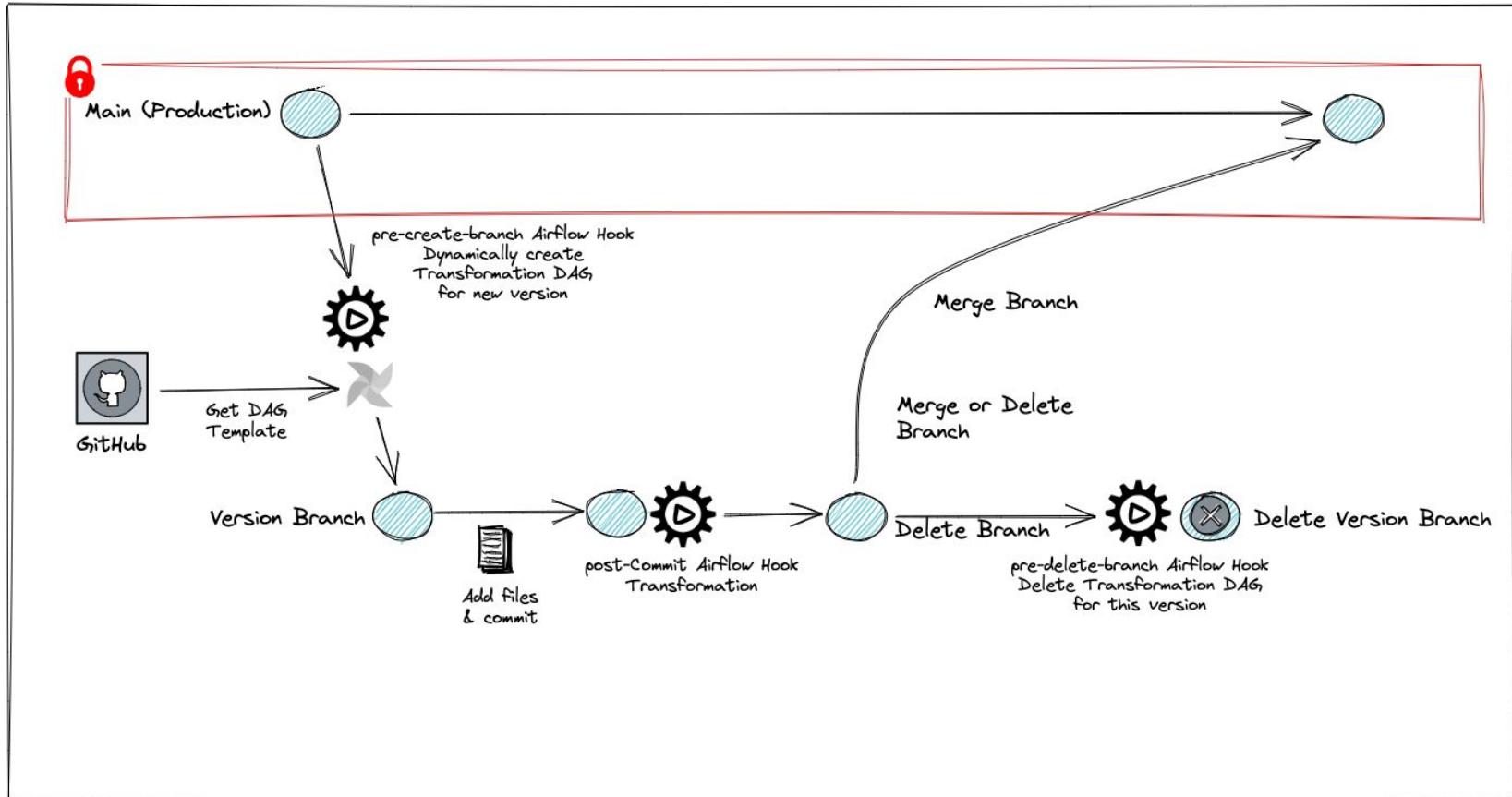


Demo 2: Troubleshoot & Reproduce Data with Apache Airflow

Ingest validation via lakeFS Hooks



Code + Data Versioning



Outcomes

20%-80%

Storage Cost
Reduction



X 2

Double Data
Engineers Efficiency



99%

Faster Recovery From
Production Outage



Complete Open-Source Stack

	Open Source	Cloud	Enterprise
Format-agnostic data version control	✓	✓	✓
Cloud-agnostic	✓	✓	✓
Zero Clone copy for isolated environment (via branches)	✓	✓	✓
Atomic Data Promotion (via merges)	✓	✓	✓
Data stays in place	✓	✓	✓
Configurable Garbage Collection	✓	✓	✓
Data CI/CD using lakeFS hooks	✓	✓	✓
Integrates with your data stack	✓	✓	✓
Role-Based Access Control	✗	✓	✓
Single Sign-On	✗	✓	✓
Unity Catalog Integration	✗	✓	✗
Audit logs	✗	✓	✗
Managed Service (Auto-updates, Auto-scaling, Disaster Recovery, etc.)	✗	✓	✗
Managed Garbage Collection	✗	✓	✗
SOC 2 Compliant	✗	✓	✗
Support SLA	✗	✓	✓



Join Our Community



Trusted by more than **1K** Companies



Liked by more than **4K** members



lakefs.io/slack



Please ask-alotl questions!



 Join our slack: <https://lakefs.io/slack>

