Supercharge Your Data Testing with a Fully Open Stack

Iddo Avneri, lakeFS

Airflow Summit
Let’s flow together
September 19-21, 2023, Toronto, Canada
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

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 >>

```
lakeclt 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

Access Historical Data For ML Reproducibility WITHOUT duplicates

Create an isolated environment WITHOUT copying the data

lakeFS Managed Bucket

Main (Production) Commit

Commit

t
Demo 1: Create an isolated Dev Environment
Even better with airflow

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

Operators
- Create Branch
- Upload Object
- Get Object
- Commit
- Get Commit Details
- Merge

Sensors
- Commit Sensor
- File Sensor

Hooks
- lakeFS Hook (to interact with a lakeFS server)

Interacting with the lakeFS Server

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

Main (Production) -> Ingest Branch
  - Add files & commit
  - post-Commit Airflow Hook Transformation

ETL Branch
  - Transformation & "Success" file indication

Pre-Merge Airflow Hook
  - Confirm Successful ETL
Code + Data Versioning

1. **Main (Production)**
   - **pre-create-branch Airflow Hook**
   - **Dynamically create Transformation DAGs for new version**

2. **GitHub**
   - **Get Data Template**

3. **Version Branch**
   - **Add files & commit**
   - **post-Commit Airflow Hook Transformation**

4. **Delete Branch**
   - **Delete Transformation DAGs for this version**

5. **Merge Branch**
   - **Merge or Delete Branch**

6. **Delete Version Branch**
Outcomes

- 20%-80% Storage Cost Reduction
- X 2 Double Data Engineers Efficiency
- 99% Faster Recovery From Production Outage
<table>
<thead>
<tr>
<th>Feature</th>
<th>Open Source</th>
<th>Cloud</th>
<th>Enterprise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format-agnostic data version control</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Cloud-agnostic</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Zero Clone copy for isolated environment (via branches)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Atomic Data Promotion (via merges)</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Data stays in place</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Configurable Garbage Collection</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Data CI/CD using lakeFS hooks</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Integrates with your data stack</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Role-Based Access Control</td>
<td>✗</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Single Sign-On</td>
<td>✗</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Unity Catalog Integration</td>
<td>✗</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>Audit logs</td>
<td>✗</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>Managed Service (Auto-updates, Auto-scaling, Disaster Recovery, etc.)</td>
<td>✗</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>Managed Garbage Collection</td>
<td>✗</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>SOC 2 Compliant</td>
<td>✗</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>Support SLA</td>
<td>✗</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
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