GNARLY
Data_Waves

PRESENTED BY dremio

EPISODE 8

Managing Your Data as Code with Dremio Arctic

Jeremiah Morrow
Director, Product Marketing for Dremio Arctic
jeremiah.morrow@dremio.com
www.linkedin.com/in/jeremiahmorrow

Agenda

1. The need for data as code
2. Dremio Arctic
3. Use Cases for Data as Code
4. Demonstration

The need for data as code

The Data Explosion

More data

175 ZB of data worldwide by 2025, with as much residing in the cloud as on-prem

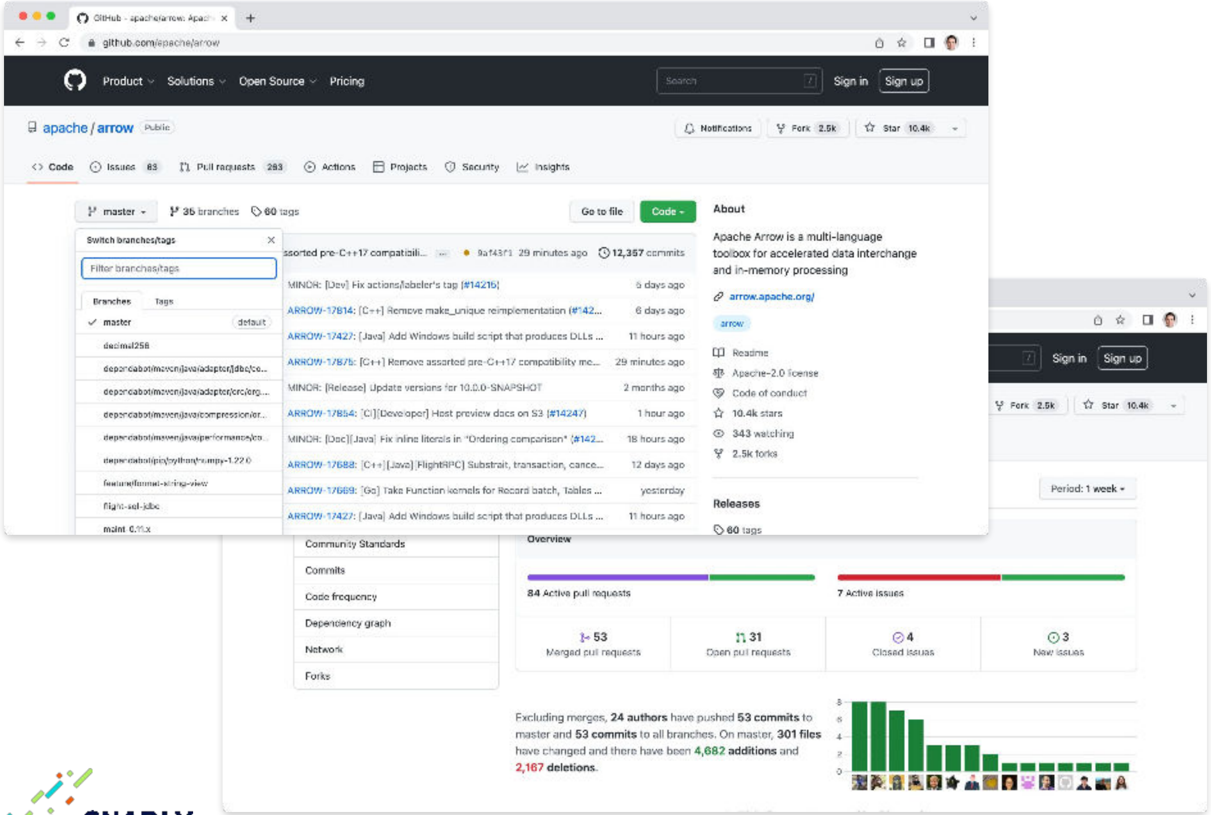
More consumers

Growing number of data requests from technical & non-technical users

More complexity

New sources, users and use cases mean we need to change the way we work with data.

GitHub/GitLab is Central to How We Build Products



...ssorted pre-C++17 compatibili... · 9af43f1 29 minutes ago · 12,357 commits

MINOR: [Dev] Fix actions/labeler's tag (#14215)	5 days ago
ARROW-17814: [C++] Remove make_unique reimplementation (#142...	6 days ago
ARROW-17427: [Java] Add Windows build script that produces DLLs ...	11 hours ago
ARROW-17875: [C++] Remove assorted pre-C++17 compatibility me...	29 minutes ago
MINOR: [Release] Update versions for 10.0.0-SNAPSHOT	2 months ago
ARROW-17854: [CI][Developer] Host preview docs on S3 (#14247)	1 hour ago
MINOR: [Doc][Java] Fix inline literals in "Ordering comparison" (#142...	18 hours ago
ARROW-17688: [C++][Java][FlightRPC] Substrait, transaction, cance...	12 days ago
ARROW-17669: [Go] Take Function kernels for Record batch, Tables ...	yesterday
ARROW-17427: [Java] Add Windows build script that produces DLLs ...	11 hours ago

About

Apache Arrow is a multi-language toolbox for accelerated data interchange and in-memory processing

arrow.apache.org/

arrow

Readme

Apache-2.0 license

Code of conduct

10.4k stars

343 watching

2.5k forks

Releases

60 tags

84 Active pull requests

7 Active issues

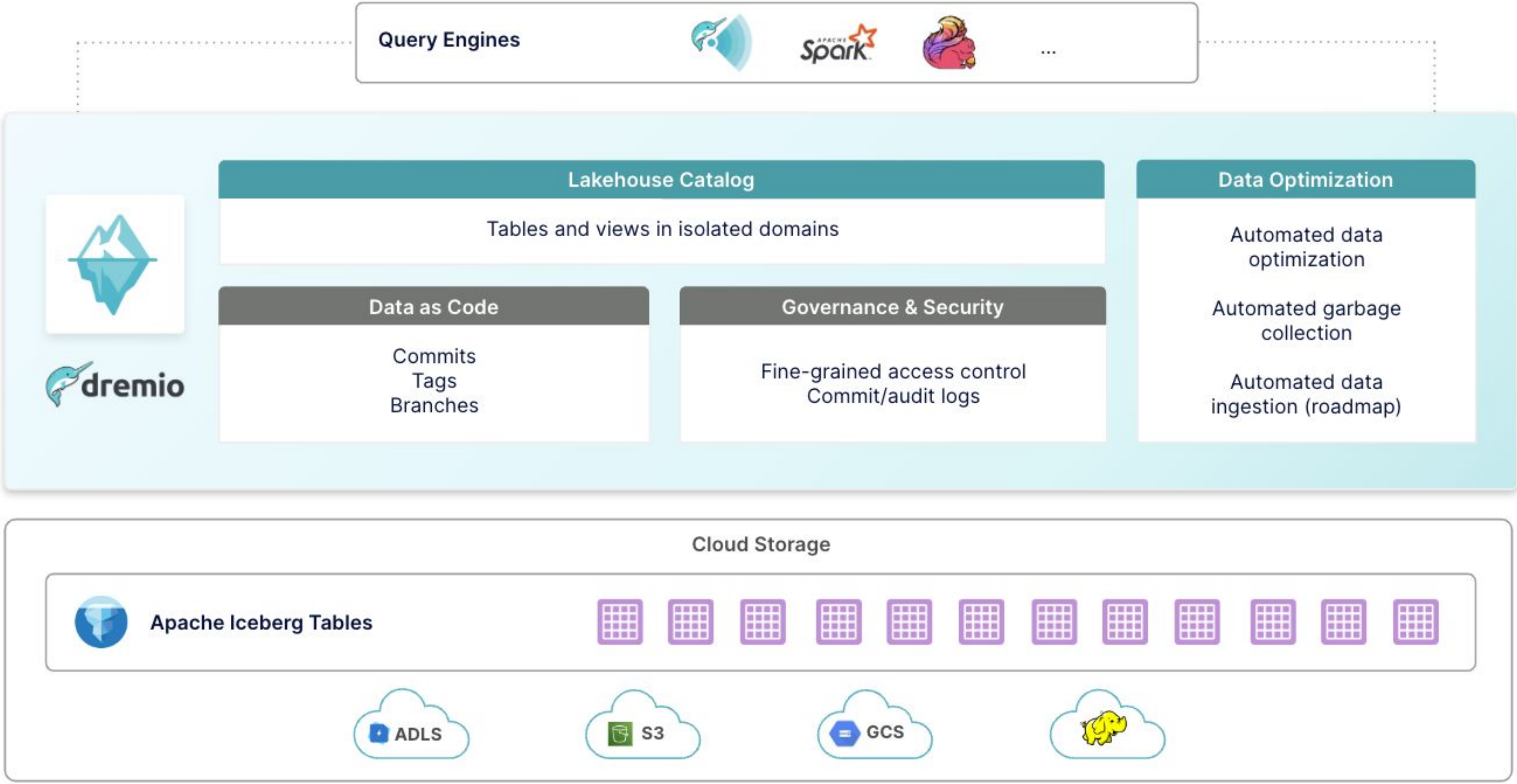
53	31	4	3
Merged pull requests	Open pull requests	Closed issues	New issues

Excluding merges, **24 authors** have pushed **53 commits** to master and **53 commits** to all branches. On master, **301 files** have changed and there have been **4,682 additions** and **2,167 deletions**.

- ✓ Code storage
- ✓ Atomic changes
- ✓ Documentation
- ✓ Access old versions
- ✓ Recover from mistakes
- ✓ Trace history
- ✓ Isolated development
- ✓ CI/CD
- ✓ Collaboration

Dremio Arctic

Dremio Arctic is a Data Lakehouse Management Service



Query Engines

Spark

...

Lakehouse Catalog

Tables and views in isolated domains

Data as Code

Commits
Tags
Branches

Governance & Security

Fine-grained access control
Commit/audit logs

Data Optimization

Automated data optimization

Automated garbage collection

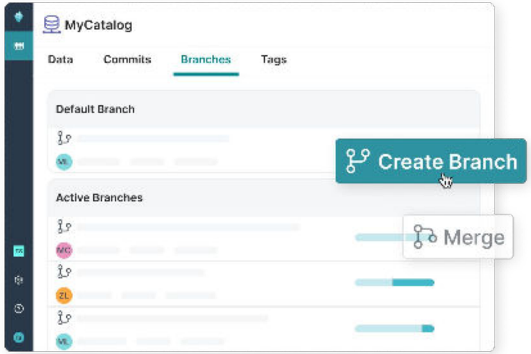
Automated data ingestion (roadmap)

dremio

Cloud Storage

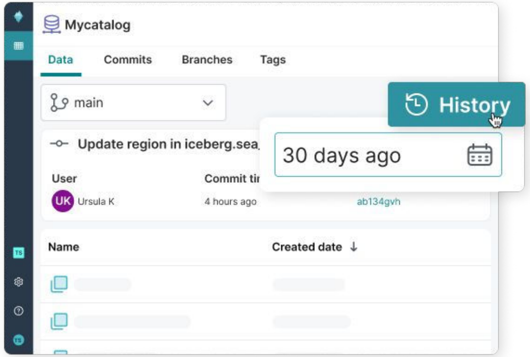
Apache Iceberg Tables

Dremio Arctic Enables Data as Code Management



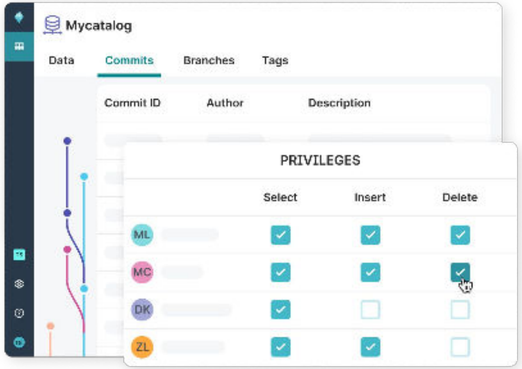
ISOLATION

- Experiment with data without impacting other users
- Ingest, transform and test data before exposing it to other users in an atomic merge



VERSION CONTROL

- Reproduce models and dashboards from historical data based on time or tags
- Recover from any mistake by instantly undoing accidental data or metadata changes



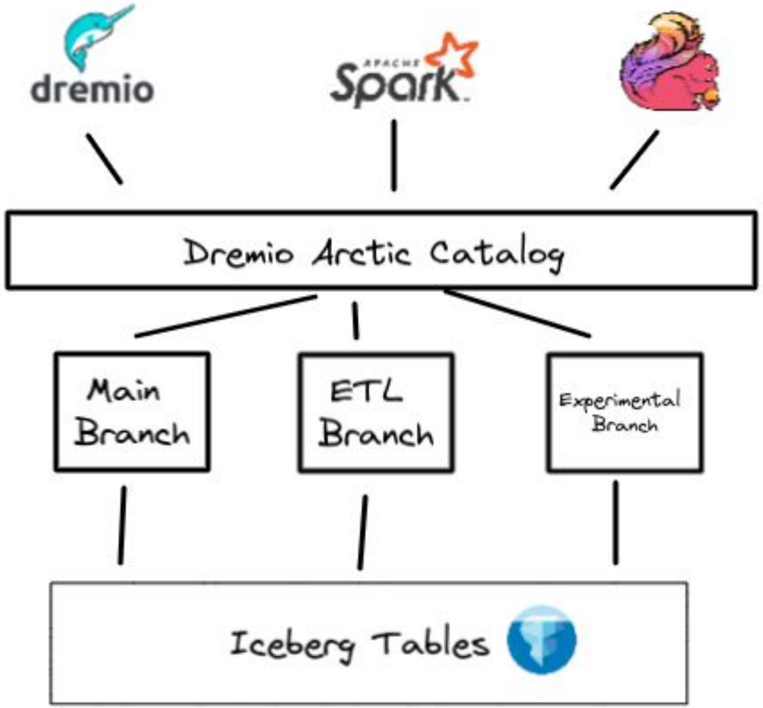
PRIVILEGES

	Select	Insert	Delete
ML	☑	☑	☑
MC	☑	☑	☑
DK	☑	□	□
ZL	☑	☑	□

GOVERNANCE

- All changes to the data and metadata are tracked: who accessed what data and when
- Fine-grained privileges to control access to the data at the table, column and row level

Branches, Tags, and Commits: A No-Copy Solution for Data Management



Benefits of Data as Code

Consistency

Experimentation

Collaboration

Governance

Reproducibility

5 Use Cases for Data as Code

Ensure Data Quality with ETL Branches

Create an ETL branch and ingest the data with COPY INTO, CTAS or Spark:

```
CREATE BRANCH events_etl_9_28_22
USE BRANCH events_etl_9_28_22
COPY INTO web.events ...
```

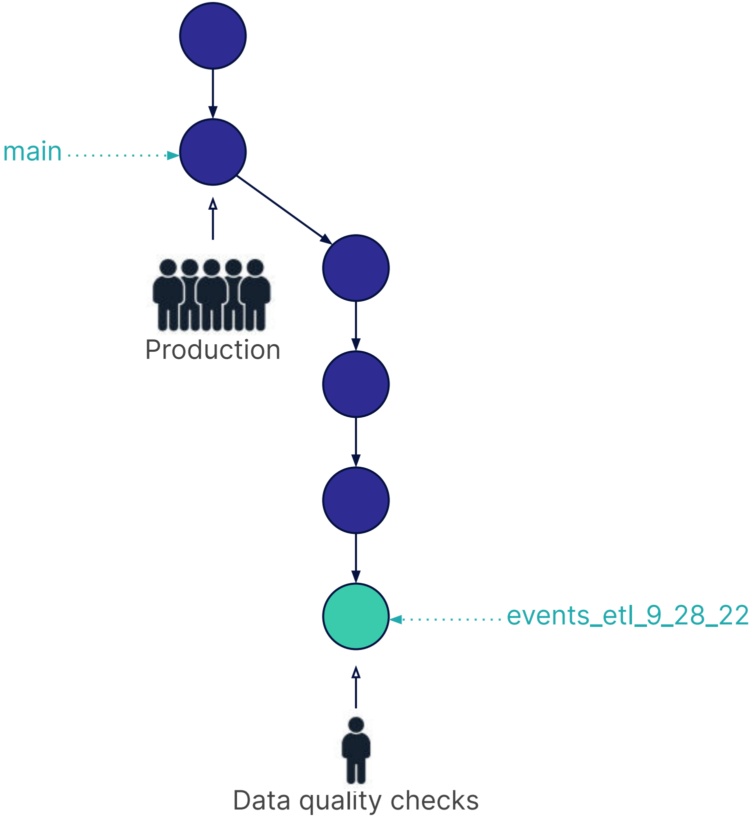
Run queries to test data quality:

```
SELECT COUNT(*) FROM web.events WHERE
length(ip_address) >= 7
```

Test the dashboard to see that it looks ok:

Fix the problems and merge into main:

```
DELETE FROM web.events WHERE length(ip_address) >= 7
USE BRANCH main
MERGE BRANCH events_etl_9_28_22
```



main

Production

events_etl_9_28_22

Data quality checks

Experiment with Data in Transient Branches

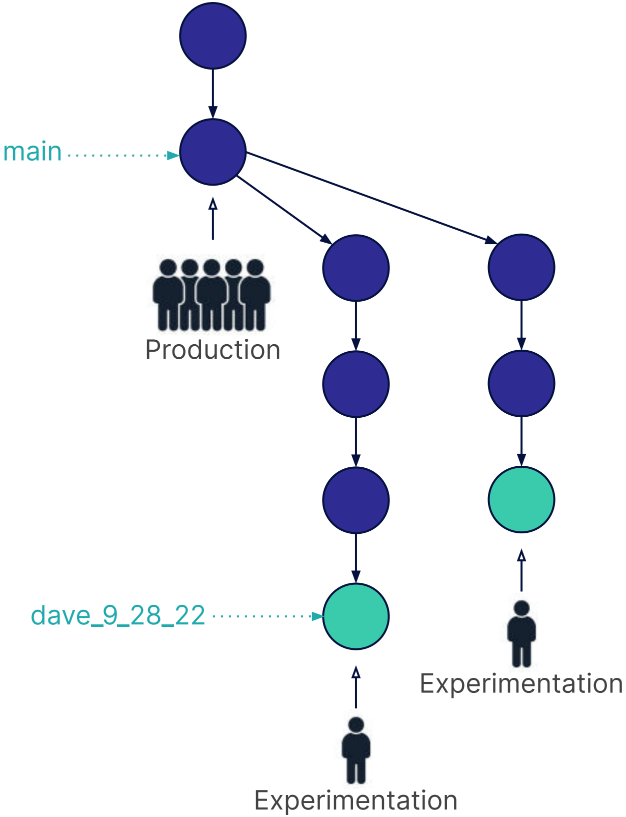
Create a transient branch and perform data explorations and transformations in it:

```
CREATE BRANCH dave_9_28_22
USE BRANCH dave_9_28_22
CREATE TABLE t AS SELECT ...
UPDATE t ... SET ...
```

Create ad-hoc visualizations on the branch via a Notebook:

Delete the branch or merge it when experimentation is complete:

```
DROP BRANCH dave_9_28_22
```



Reproduce Models

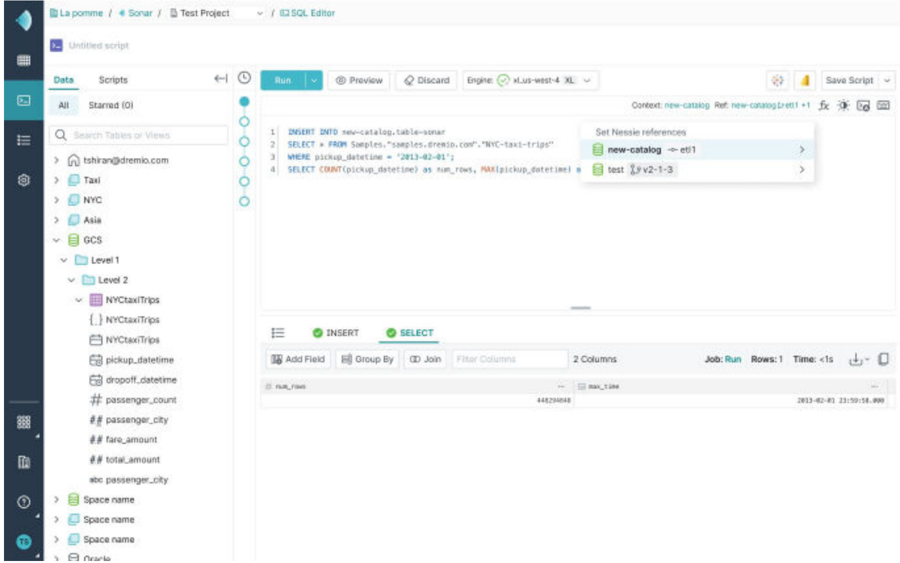
Change context to a named tag:

```
spark.sql("USE REFERENCE modelA in arctic;")
```

Create ML model based on historic data:

```
val trainingData = spark.read.table("arctic.t")
val lr = new LogisticRegression()
// configure logistic regression...
val paramMap = ParamMap(...)
val model = lr.fit(trainingData, paramMap)
```

Select a tag, commit or branch to query in SQL Runner:



```
INSERT INTO new-catalog.table-sonar
SELECT * FROM Samples."samples.dremio.com"."NYC-taxi-trips"
WHERE pickup_datetime = '2013-02-01';
SELECT COUNT(pickup_datetime) as num_rows, MAX(pickup_datetime) a
```

Context: new-catalog Ref: new-catalog ⌥ etl1 +1

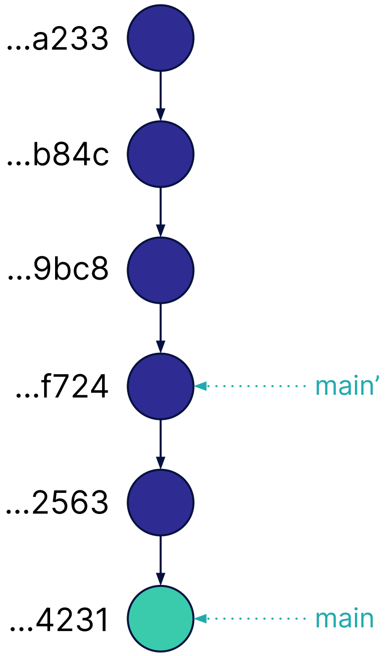
Set Nessie references
- new-catalog ⊸ etl1
- test v2-1-3

num_rows	max_time
448294848	2013-02-01 23:59:58.000

Recover From Mistakes

Move the branch head to a historical commit:

```
ALTER BRANCH main ASSIGN COMMIT …f724
```



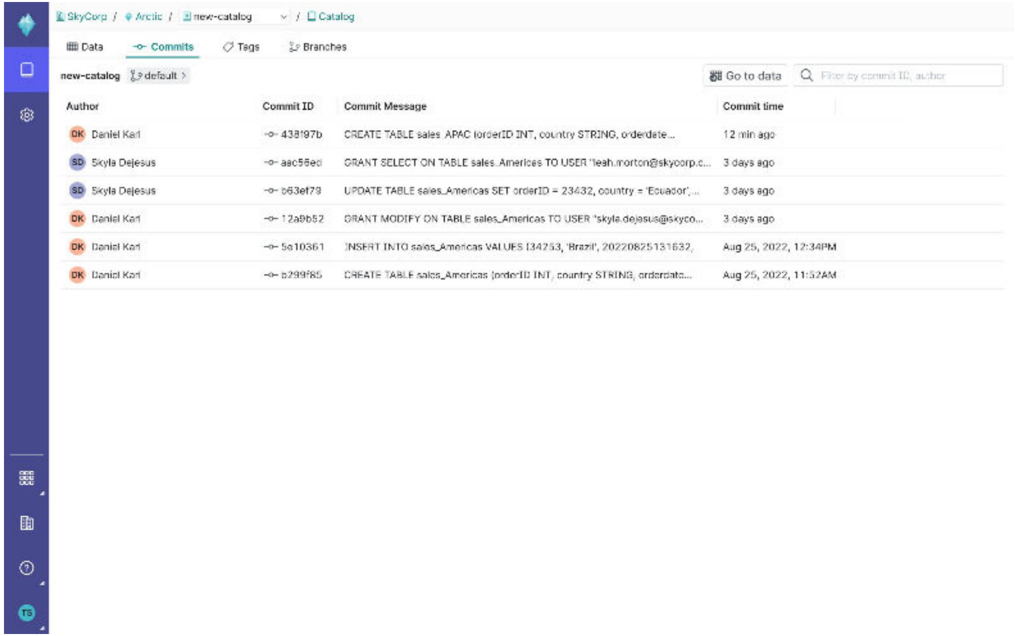
Troubleshoot

Get the commit history for a branch:

```
SHOW LOGS AT REFERENCE etl;
```

Get the commit history for a specific table:

```
curl -X GET -H 'Authorization: Bearer
<PAT>' <Catalog API
Endpoint>/trees/tree/<reference
name>/log\?filter="operations.exists(op,op.
key=='<table name>')"
```


SkyCorp / Arctic / new-catalog / Catalog
Data
Commits
Tags
Branches
new-catalog
default
Go to data
Filter by commit ID, author

Author	Commit ID	Commit Message	Commit time
DK Daniel Karl	438197b	CREATE TABLE sales_APAC (orderID INT, country STRING, orderdate...	12 min ago
SD Skyla Dejesus	aac56ed	GRANT SELECT ON TABLE sales_Americas TO USER 'leah.morton@skycorp.c...	3 days ago
SD Skyla Dejesus	b63ef79	UPDATE TABLE sales_Americas SET orderID = 23432, country = 'Ecuador',...	3 days ago
DK Daniel Karl	12a9b52	GRANT MODIFY ON TABLE sales_Americas TO USER 'skyla.dejesus@skyco...	3 days ago
DK Daniel Karl	5a10361	INSERT INTO sales_Americas VALUES (34253, 'Brazil', 20220825131632,	Aug 25, 2022, 12:34PM
DK Daniel Karl	b299f85	CREATE TABLE sales_Americas (orderID INT, country STRING, orderdate...	Aug 25, 2022, 11:52AM

Dremio Arctic Demo