



Experiencing Data as Code with Dremio Arctic and Apache Iceberg

Data Council • March 2023

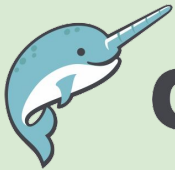


Alex Merced

Developer Advocate, Dremio

Alex Merced is a developer advocate for Dremio and has worked as a developer and instructor for companies like GenEd Systems, Crossfield Digital, CampusGuard and General Assembly.

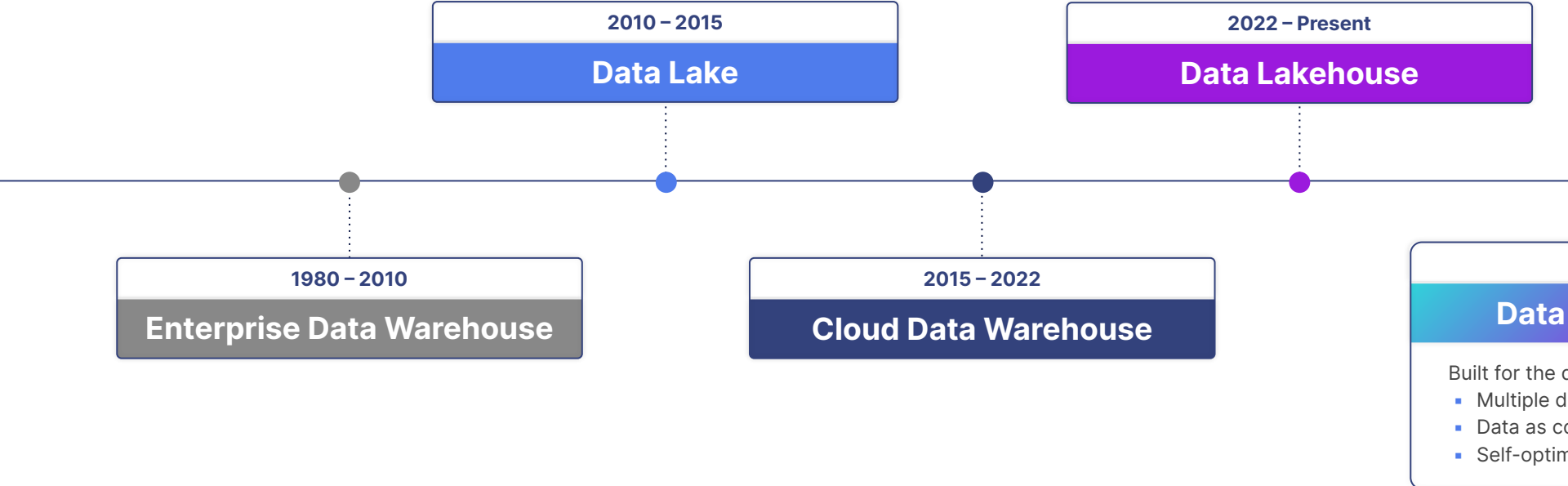
Alex is passionate about technology and has put out tech content on outlets such as blogs, videos and his podcasts Datanation and Web Dev 101. Alex Merced has contributed a variety of libraries in the Javascript & Python worlds including SencilloDB, CoquitoJS, dremio-simple-query and more.



dremio

Free T-Shirt @ Dremio Table

From Data Warehouse to Data Lakehouse



Data Lakehouse 2.0

2015
Data Lake

2022 – Present
Data Lakehouse

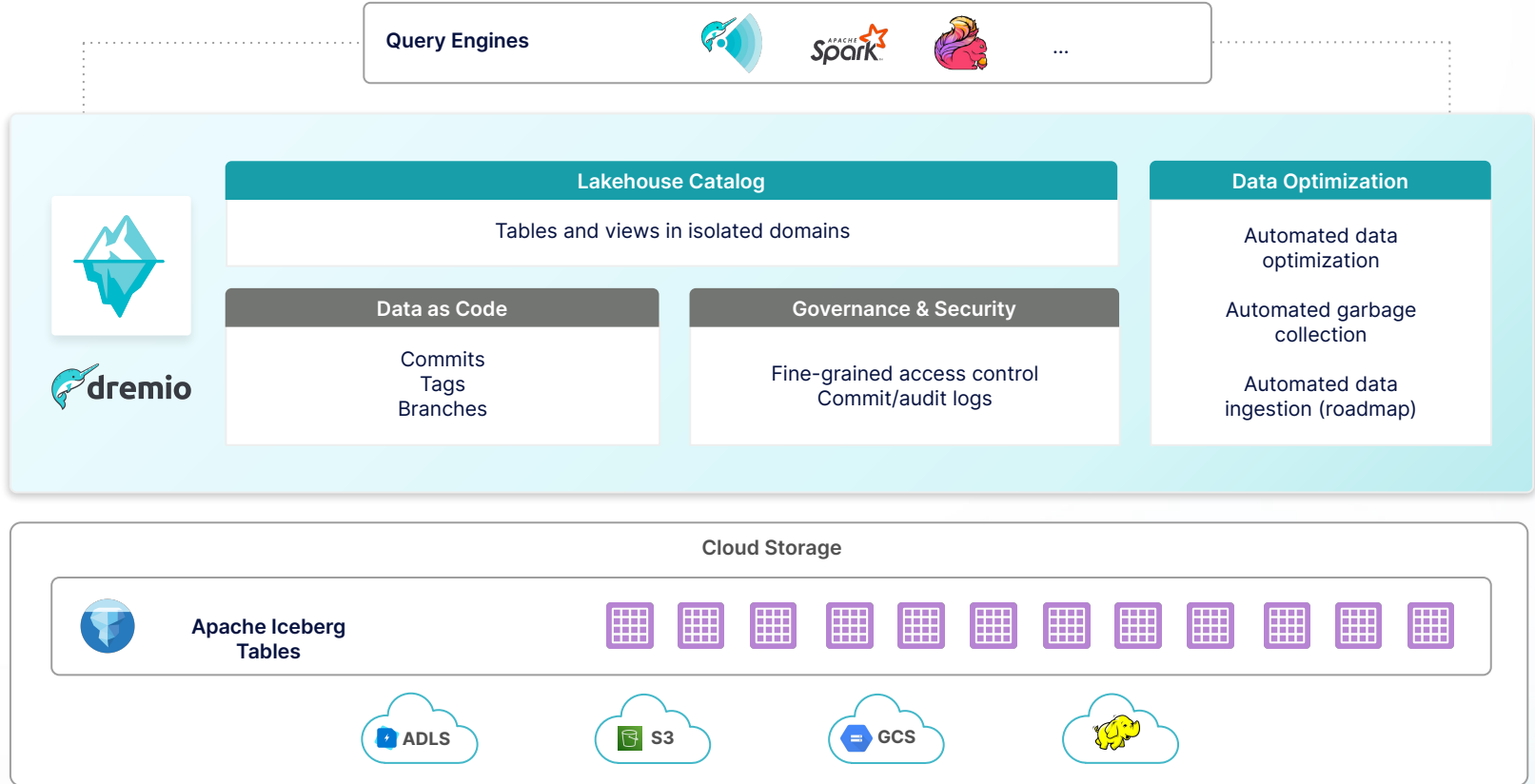
2015 – 2022
Cloud Data Warehouse

What's Next?
Data Lakehouse 2.0

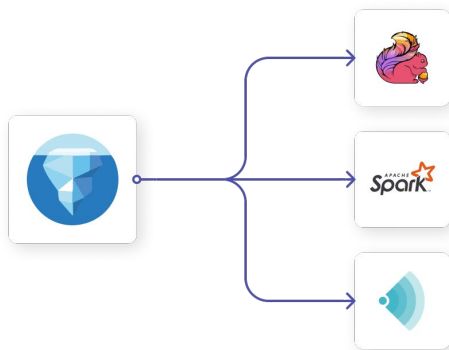
Built for the data mesh era:

- Multiple domains
- Data as code
- Self-optimizing

Dremio Arctic is a Data Lakehouse Management Service

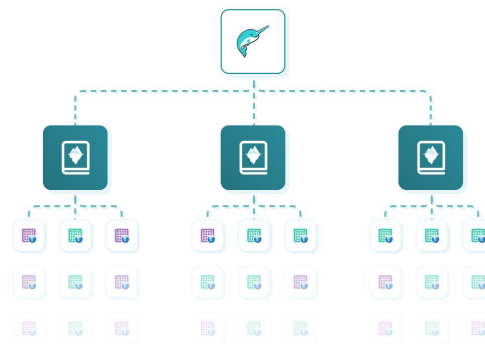


A Modern Lakehouse Catalog



ICEBERG-NATIVE

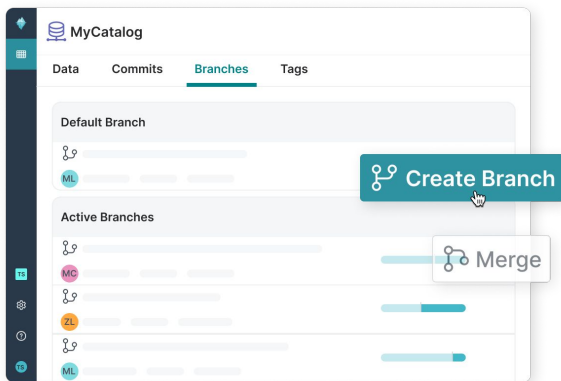
- Nessie (the Arctic catalog) is built into the open source Apache Iceberg project
- Use a variety of Iceberg-compatible engines including Dremio Sonar, Spark and Flink



MULTIPLE DOMAINS

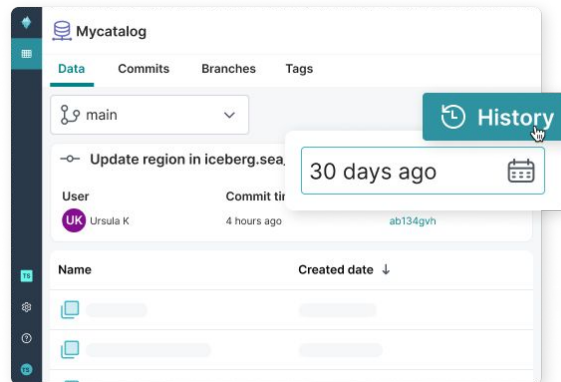
- Multiple isolated domains/catalogs in an organization, each containing a folder hierarchy of tables and views
- Designed to enable data mesh (including federated ownership and data sharing)

Data as Code



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

5 Use Cases for Data as Code

1: Ensure data quality with ETL branches

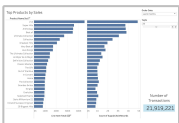
Create an ETL branch and ingest the data with COPY INTO, CTAS or Spark:

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

Run queries to test data quality:

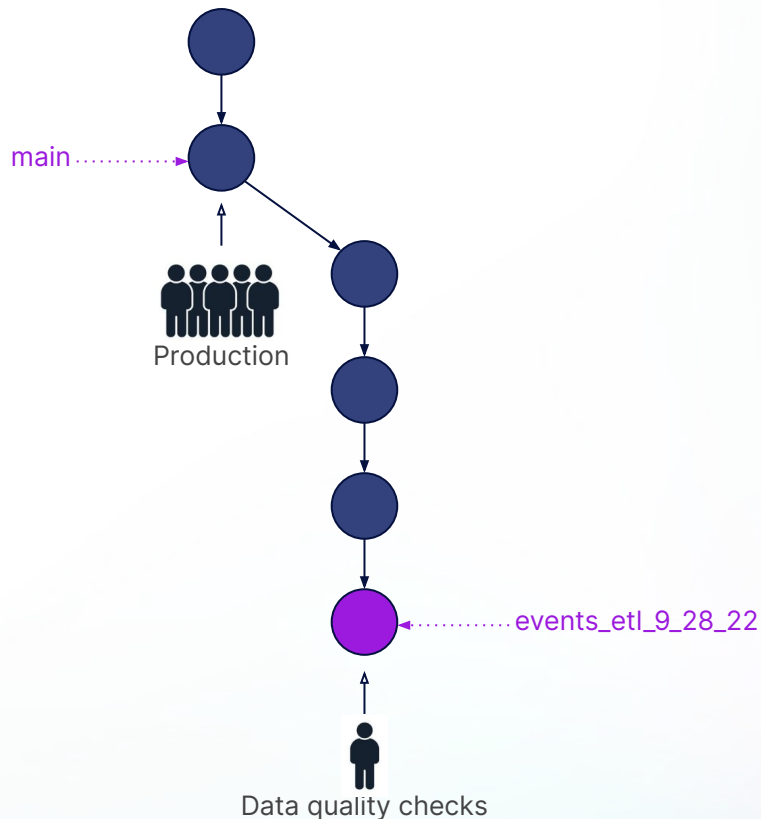
```
SELECT COUNT(*) FROM web.events WHERE
length(ip_address) >= 7
```

Test the dashboard to see that it looks okay:



Fix the problems and merge into main:

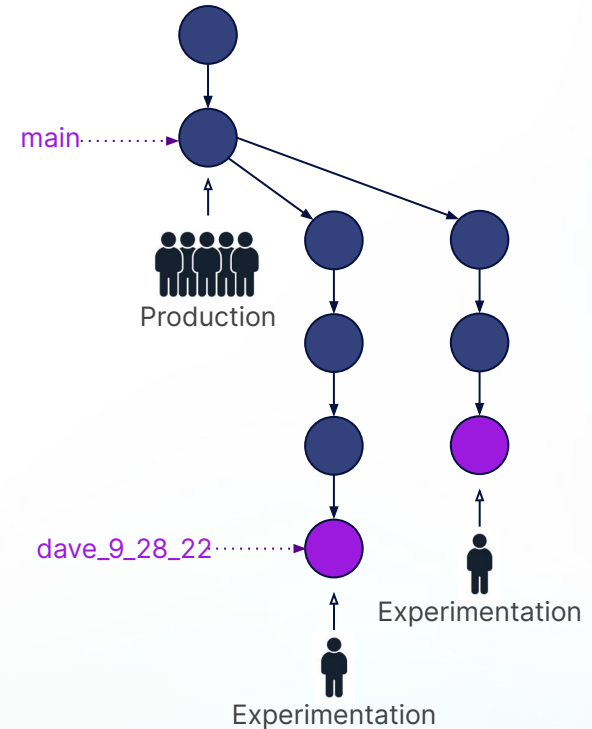
```
DELETE FROM web.events WHERE length(ip_address) >= 7
USE BRANCH main
MERGE BRANCH events_etl_9_28_22
```



2: 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 ...
```



3: Reproduce models or analysis

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)
```

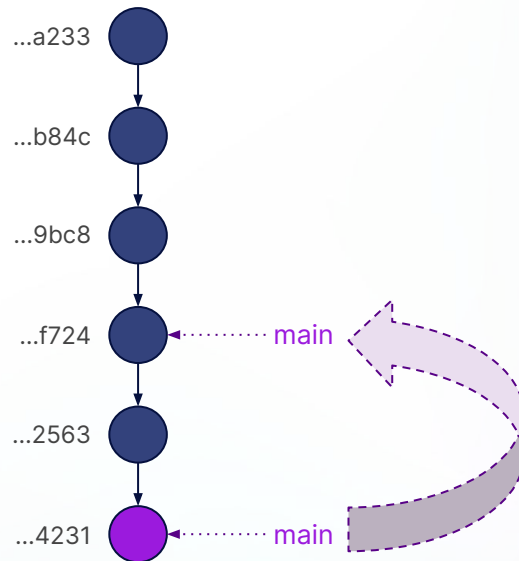
4: Recover from mistakes

If you accidentally mess up the data or schemas in your lakehouse:

```
INSERT INTO sales
  (SELECT * FROM
   sales_last_quarter_unaudited)
DROP TABLE customers
```

Move the branch head to a historical commit:

```
ALTER BRANCH main ASSIGN COMMIT ...f724
```



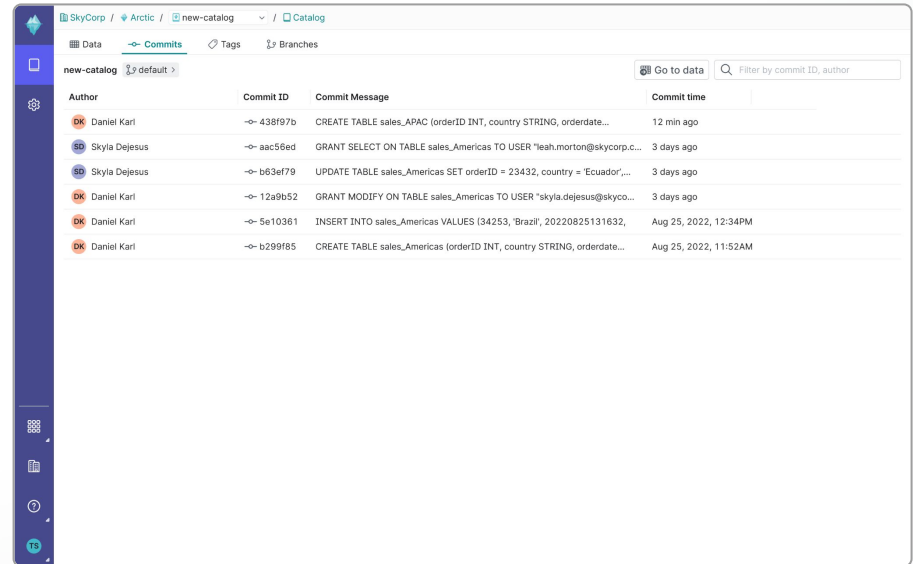
5: Troubleshooting (see who changed the data)

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>')"
```



Author	Commit ID	Commit Message	Commit time
Daniel Karl	438f97b	CREATE TABLE sales_APAC (orderID INT, country STRING, orderdate...	12 min ago
Skyia Dejesus	aac56ed	GRANT SELECT ON TABLE sales_Americas TO USER 'teah.morton@skycorp.c...	3 days ago
Skyia Dejesus	b63ef79	UPDATE TABLE sales_Americas SET orderID = 23432, country = 'Ecuador',...	3 days ago
Daniel Karl	12a9b52	GRANT MODIFY ON TABLE sales_Americas TO USER 'skyia.dejesus@skycor...	3 days ago
Daniel Karl	5e10361	INSERT INTO sales_Americas VALUES (34253, 'Brazil', 20220825131632,	Aug 25, 2022, 12:34PM
Daniel Karl	b299f85	CREATE TABLE sales_Americas (orderID INT, country STRING, orderdate...	Aug 25, 2022, 11:52AM

EASY STEP BY STEP
ACTIVITY
TRY AT HOME



DATA AS CODE WORKSHOP

DEMO TIME