GNARLY Data_Waves

PRESENTED BY dremio

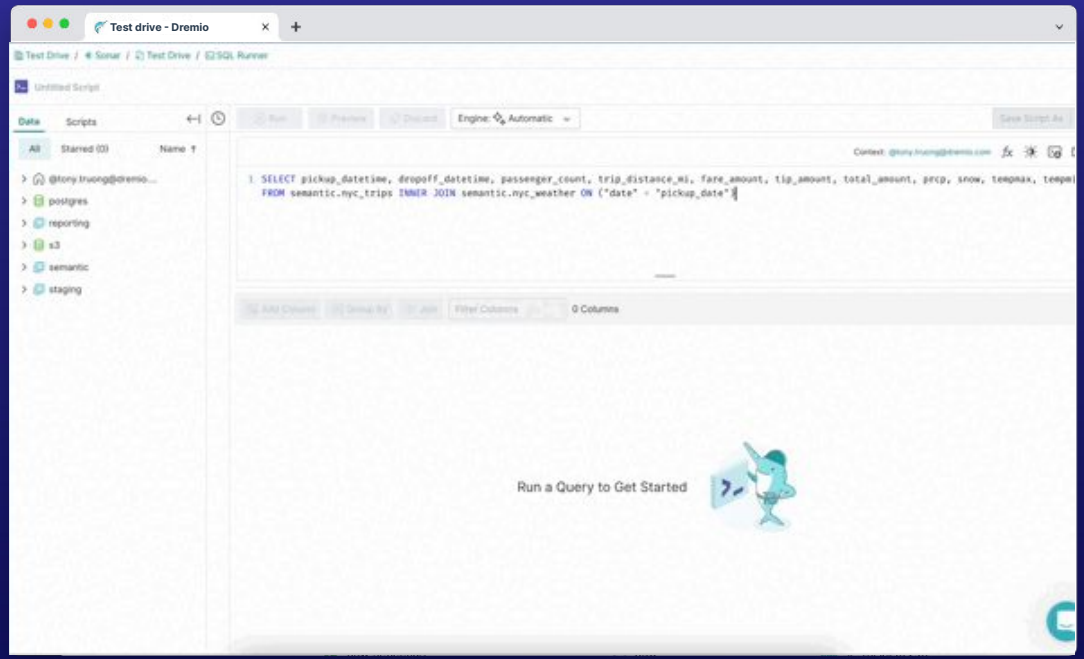
EPISODE 22

Dremio and Data Lakehouse Table Formats (Apache Iceberg, Delta Lake and Apache Hudi & Dremio)

Experience the data lakehouse with Dremio Test Drive

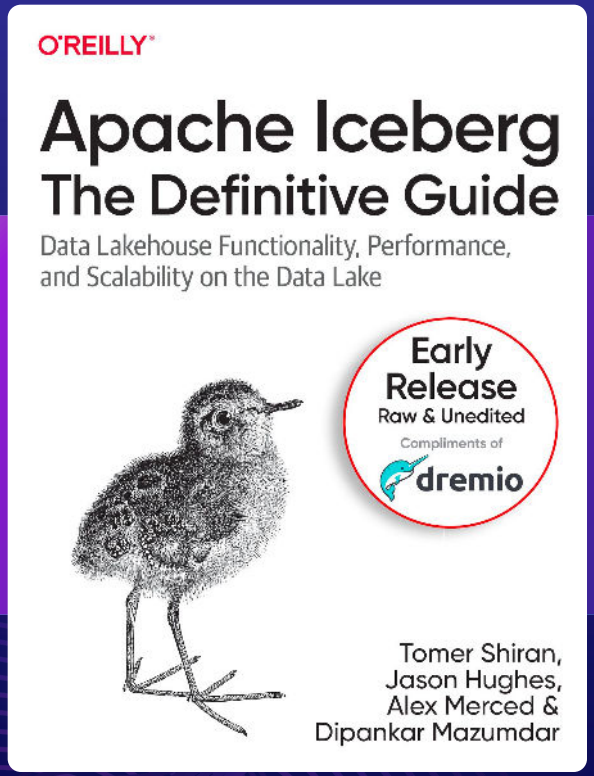
- Sub-second query on 1 billion rows of data joining Amazon S3 with a Postgres database
- Connect to Tableau or Power BI and build a dashboard with this dataset
- Everything hosted by Dremio - 100% free for you

Start Test Drive



GNARLY Data_Waves

Apache Iceberg: The Definitive Guide



Upcoming shows

Register now

EPISODE 21

Machine Learning Experimentation/Reproducibility on a Lakehouse

June 13, 2023 at 8AM PST | 11AM EST | 4PM GMT

EPISODE 22

Dremio and Data Lakehouse Table Formats (Iceberg, Delta, Hudi)

June 20, 2023 at 8AM PST | 11AM EST | 4PM GMT

EPISODE 23

Getting Started With Dremio Data Reflections

June 27, 2023 at 8AM PST | 11AM EST | 4PM GMT

EPISODE 24

Simplifying Data Mesh with Dremio's Open Data Lakehouse

July 11, 2023 at 8AM PST | 11AM EST | 4PM GMT

EPISODE 25

Best Practices for Building a Data Lakehouse on ADLS

July 18 2023 at 8AM PST | 11AM EST | 4PM GMT

EPISODE 26

Versioning Data in the Data Lakehouse (File, Table and Catalog Versioning)

July 25, 2023 at 8AM PST | 11AM EST | 4PM GMT

dremio | aws | +ableau

AWS Dev Day: Chicago

Experience Dremio Cloud and Tableau on Amazon S3

May 17th, 10am CST-12pm | Lunch to follow

dremio | aws

AWS Summit Toronto

June 14th, 2023
Metro Toronto Convention Centre

dremio | aws | +ableau

AWS Dev Day: New York

Experience Dremio Cloud

July 25th, 10:00 am to 1:00 pm
Endeavor 1 | Courtyard - 461 W 34th St.
New York, NY 10001

dremio | aws

AWS Summit New York

July 26th, 2023
Javits Center

BIG DATA & AI WORLD | BARC

dremio

Coalesce by dbt

Oct 16-20, 2023
Hilton Bayfront San Diego

BIGDATA & AI by Xcorp PARIS

TIME TO ACCELERATE

September 25 & 26, 2023

Paris Convention Center

GNARLY Data_Waves

PRESENTED BY dremio

EPISODE 22

Dremio and Data Lakehouse Table Formats

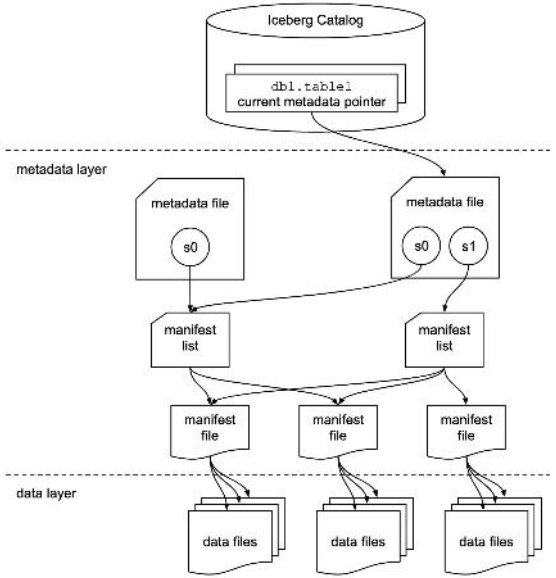
(Apache Iceberg, Delta Lake and Apache Hudi & Dremio)

June 20 at 8AM PST | 11AM EST | 4PM GMT

Alex Merced
Developer Advocate, Dremio

Table Format Overview

Apache Iceberg

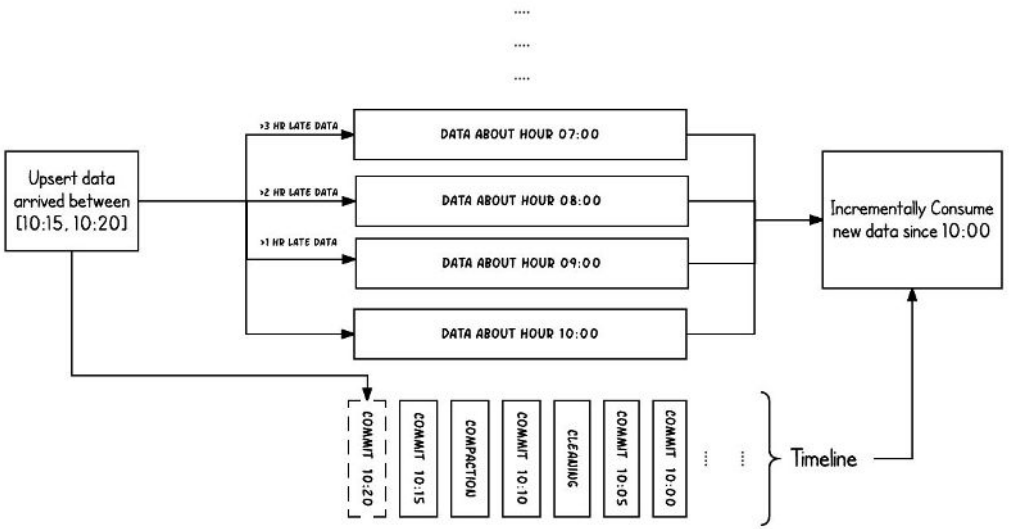


Apache Iceberg's approach is to define the table through three layers of metadata. These categories are:

- **metadata files** that define the table
- **manifest lists** that define a snapshot of the table, with a list of manifests that make up the snapshot and metadata about their data
- **manifests** is a list of data files along with metadata on those data files for file pruning

GNARLY Data_Waves

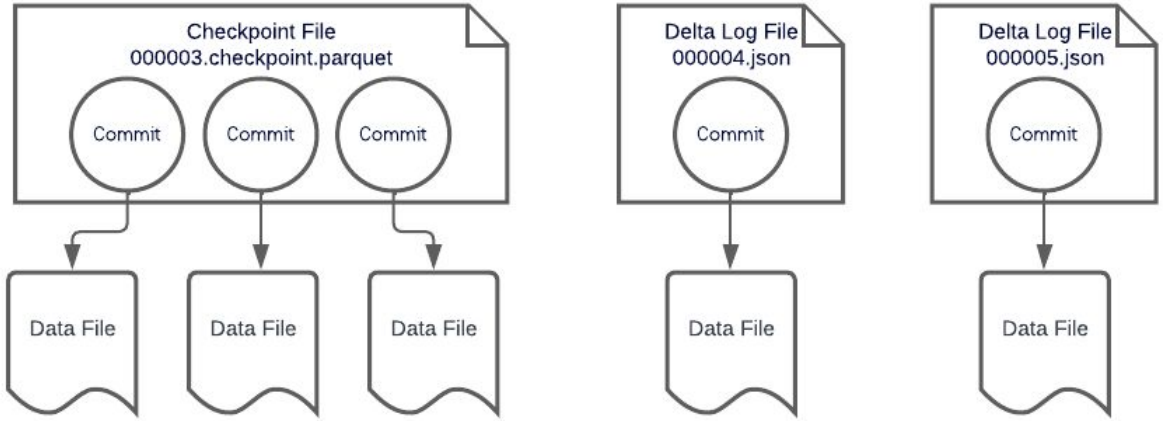
Apache Hudi



- Apache Hudi's approach is to group all transactions into different types of actions that occur along a timeline.
- Directory-based approach with timestamped files and log files that track changes.
- An optional metadata table for additional file pruning.

GNARLY Data_Waves

Delta Lake



Delta Lake's approach is to track metadata in two types of files:

- Delta Logs sequentially track changes to the table.
- Checkpoints summarize all changes to the table up to that point.

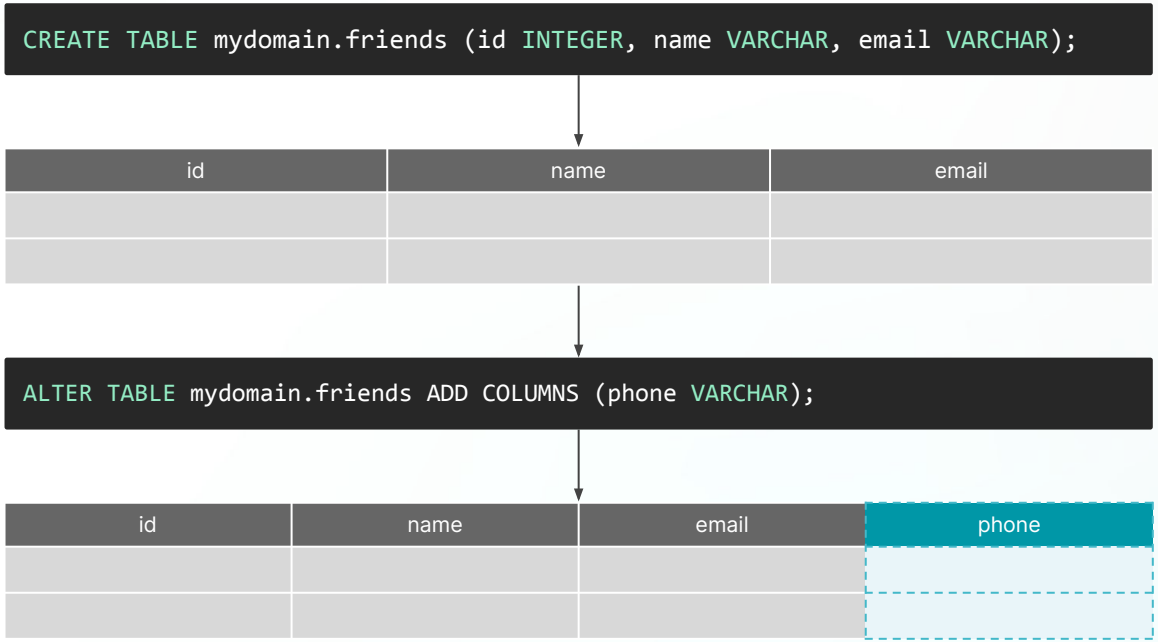
In these files are indexes of columns used for file pruning

Dremio Makes Iceberg Easy

Create and Alter Iceberg Tables

DDL

Create and alter Iceberg tables



Ingest Data into Iceberg Tables

COPY INTO

Ingest existing data into an Iceberg table

```
{"id": 1, "name": "Bob", "age":46}
{"id": 2, "name": "Josie", "age":65}
{"id": 3, "name": "Gene", "age":30}
```



```
COPY INTO mydomain.mytable
       FROM @SOURCE/bucket/path/folder]
       [ FILES ('foo.json');
```


id	name	age
1	Bob	46
2	Josie	65
3	Gene	30

Manipulate Iceberg Tables

id	name	email
1	Alex Merced	alex.merced@dremio.com
2	Bob Jones	

+

id	name	email
2	Bob Jones	Bob@SomeDomain.xyz
3	Gina Somebody	GSomebody@Domain.xyz

DML

Insert, update, delete and merge records in an Iceberg table

```
MERGE INTO names n
USING (SELECT * FROM names_staging) s
ON n.id = s.id
WHEN MATCHED THEN UPDATE SET name = s.name, age, s.email
WHEN NOT MATCHED THEN INSERT (id, name, email) VALUES (s.id, s.name,
s.email)
```


id	name	email
1	Alex Merced	alex.merced@dremio.com
2	Bob Jones	Bob@SomeDomain.xyz
3	Gina Somebody	GSomebody@Domain.xyz

Fast Queries based on Iceberg Partitions and Statistics

SELECT

Leveraging Iceberg partitions and statistics to maximize performance

```
CREATE TABLE sales (id INTEGER, sales_date DATE, total FLOAT, department
VARCHAR)
PARTITION BY (MONTH (sales_date)) LOCALSORT BY (department);
```

id	sales_date	total	department
1	2023-02-01 09:00:00.000	$10,000	Dept A
2	2023-02-15 09:00:00.000	$20,000	Dept A
3	2023-02-08 09:00:00.000	$15,000	Dept B
4	2023-02-16 09:00:00.000	$18,000	Dept B

February 2023 Partition

id	sales_date	total	department
1	2023-03-04 09:00:00.000	$10,000	Dept A
2	2023-03-18 09:00:00.000	$30,000	Dept A
3	2023-03-09 09:00:00.000	$85,000	Dept B
4	2023-03-11 09:00:00.000	$28,000	Dept B

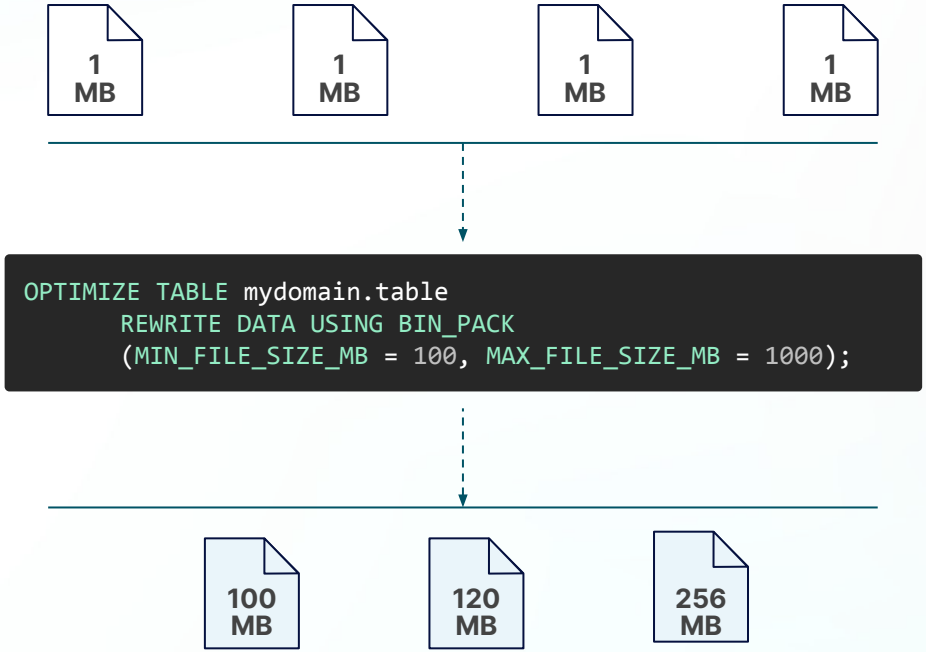
March 2023 Partition

```
SELECT * FROM sales
WHERE sales_data BETWEEN '2023-02-1 00:00:00.000' and '2023-02-28
00:00:00.000' AND department = 'Dept A';
```

Optimize Iceberg Tables

OPTIMIZE

Compact and optimize the data in an Iceberg table

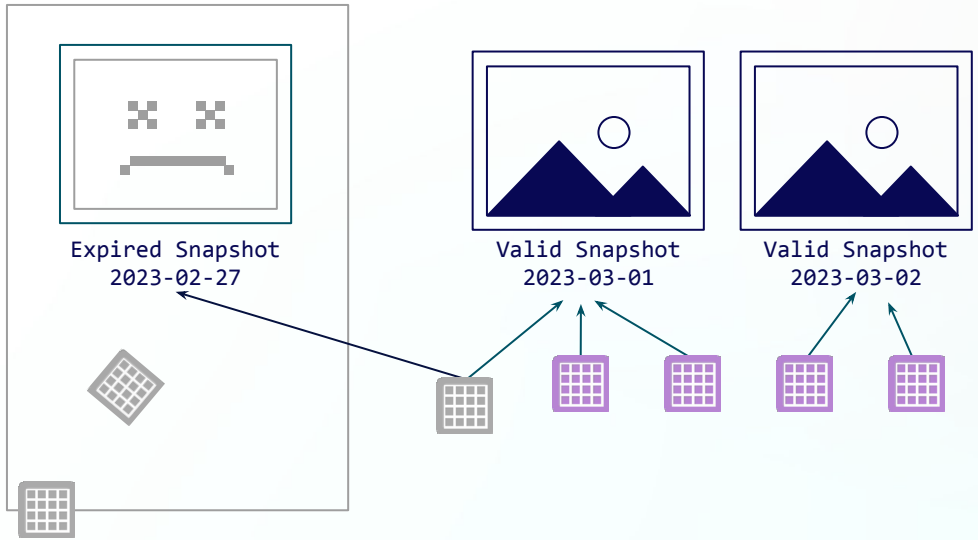


Vacuum Iceberg Tables

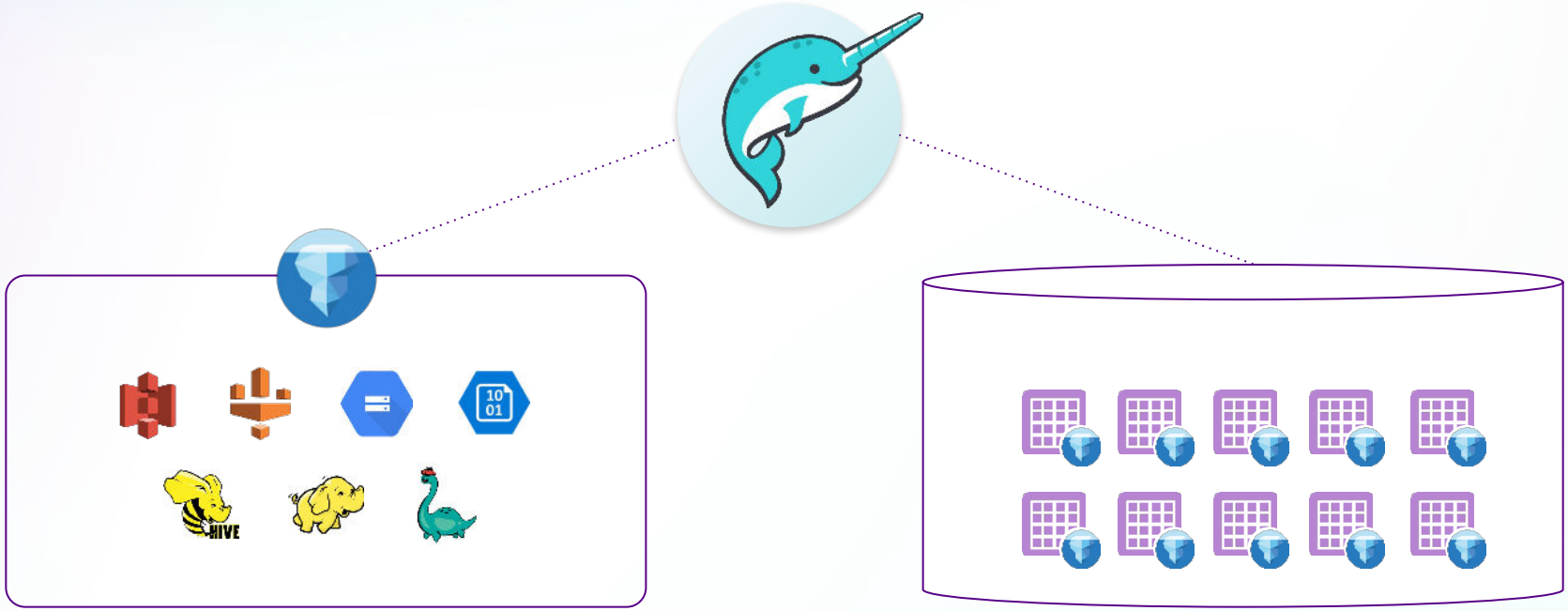
```
VACUUM TABLE names EXPIRE SNAPSHOTS older_than = '2023-28
09:01:51.741';
```

VACUUM

Garbage collect old artifacts and versions



Dremio is Open And Works with a Range of Catalogs

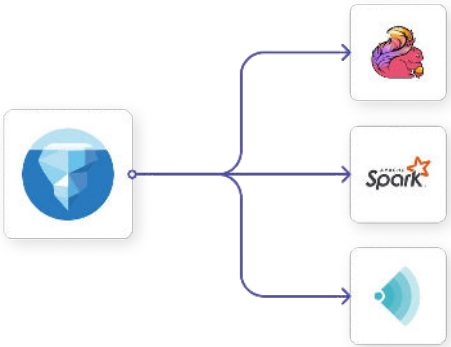


Iceberg Catalogs

Iceberg Tables in Object Storage
(S3, Azure Storage, GCS)

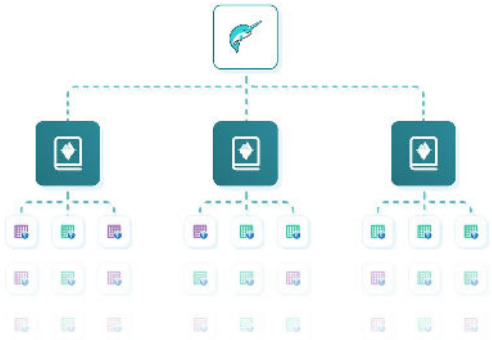
Dremio Arctic

A Lakehouse Management Service Powering the Data Mesh



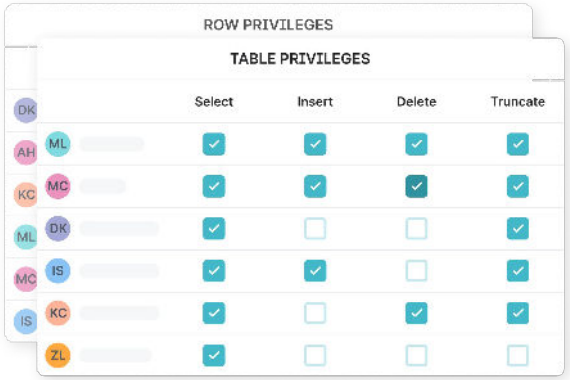
ICEBERG-NATIVE

- 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



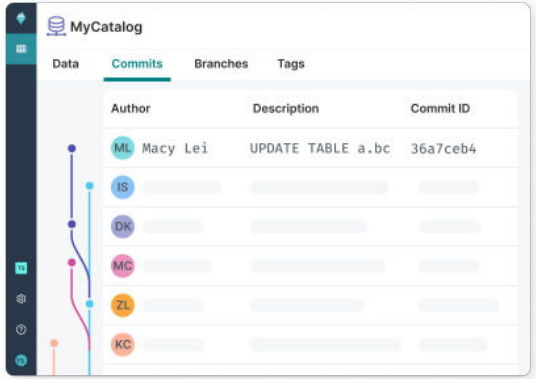
MULTIPLE DOMAINS

- 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)



ACCESS CONTROL

- Table, column- and row-based access control, and custom integration with existing user/group directories (AAD, Okta, etc.)



GOVERNANCE

- All changes to the data and metadata are audited: who accessed what data and when

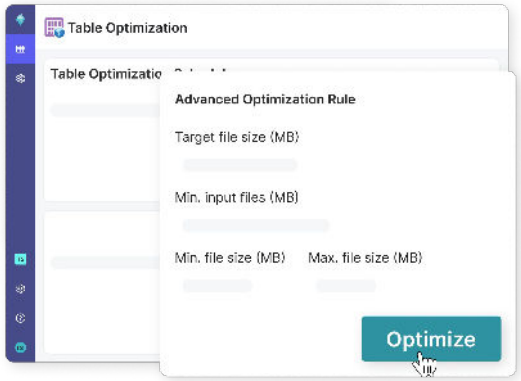


TABLE OPTIMIZATION

- Dremio Arctic automatically rewrites smaller files into larger files and groups similar rows in a table together
- Table optimization significantly accelerates query performance

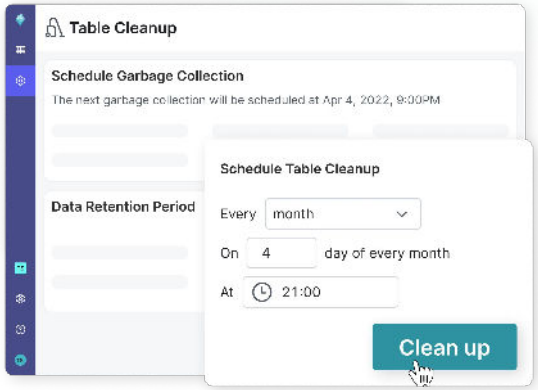
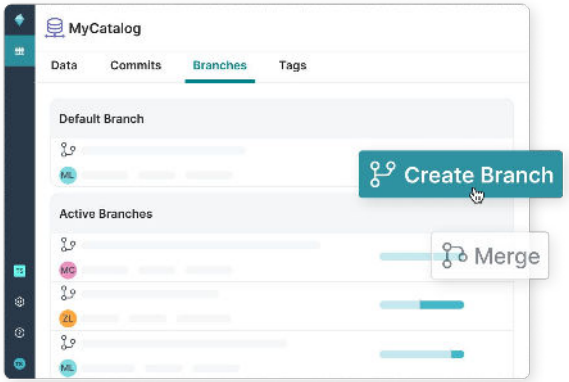


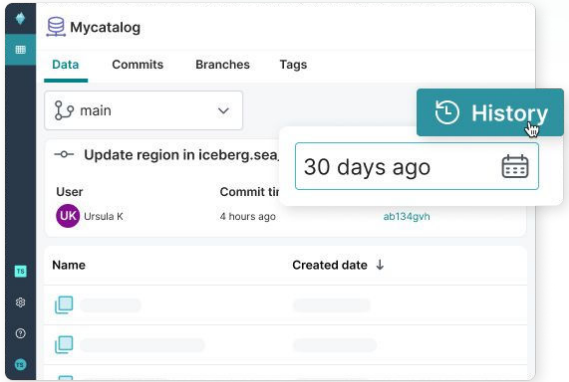
TABLE CLEANUP

- Dremio Arctic automatically removes unused manifest files, manifest lists, and data files
- Cleanup runs in the background and ensures efficient use of data lake storage



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

OTHER DREMIO AND ICEBERG NOTES

- Dremio’s Reflections are Powered by Apache Iceberg
- Tabular can currently be used with Dremio via AWS Glue

Delta Lake & Dremio

- Dremio can Read Delta Lake tables
- Reflections can be used to accelerate Delta Lake Tables

Hudi & Dremio

- Unsupported
- Unofficial Jars for Dremio Software
- Onehouse tables may be queryable via Onehouse