GNARLY Data_Waves

PRESENTED BY dremio

EPISODE XX

What's New in Dremio: Improved Automation, Performance + Catalog for Iceberg Lakehouses

Today's Agenda

Mark Shainman
&
Maeve Donovan

- Dremio Release Focus
- What's New in Dremio
  - Improved Performance & Automation
  - Improved Data Ingestion
  - Accelerated Cross-Database Access Control and Workload Management
  - Integrated Observability
  - Google BigQuery Connector
  - Graviton Support
- Iceberg Catalog

We're making Iceberg Lakehouses more manageable and higher performing than ever before

ICEBERG

PERFORMANCE

TCO

Decrease TCO and Improve Time to Value

Apache Iceberg

An Open Table Format for Enterprise Data Lakes

High-performance queries - Purpose-built for high performance queries on massive datasets.

Data warehouse functionality on the data lake - ACID transactions, time travel, and schema evolution enable more data warehouse workloads directly on data lake storage.

Easy data operations - Reduce overhead costs with table optimization, garbage cleanup, and more.

The Largest Open Source Community

More individual companies with **contributions** than any other open table format

More OSS integrations than any other open table format.

Enterprise Companies Using Iceberg

NETFLIX Expedia stripe

airbnb twilio

LinkedIn Adobe Tencent

Commercial Support for Iceberg

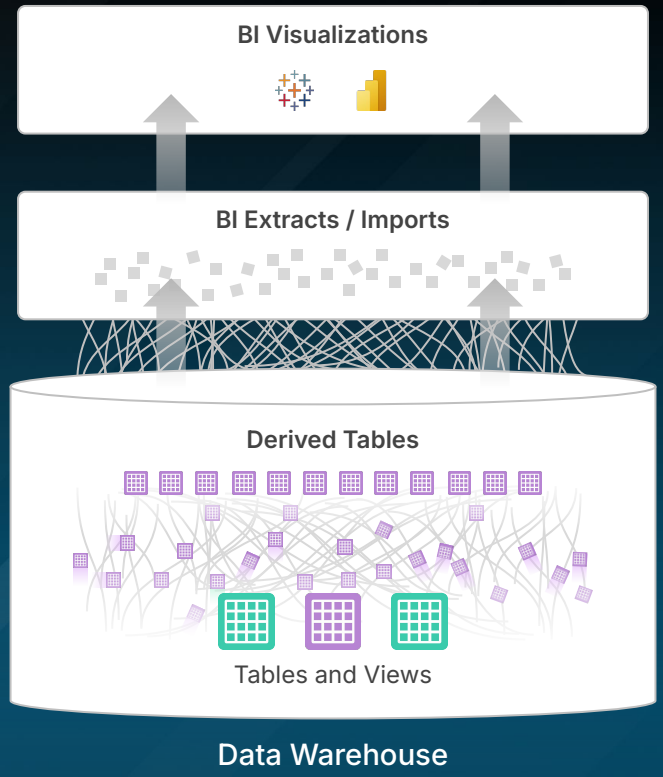
aws dremio snowflake

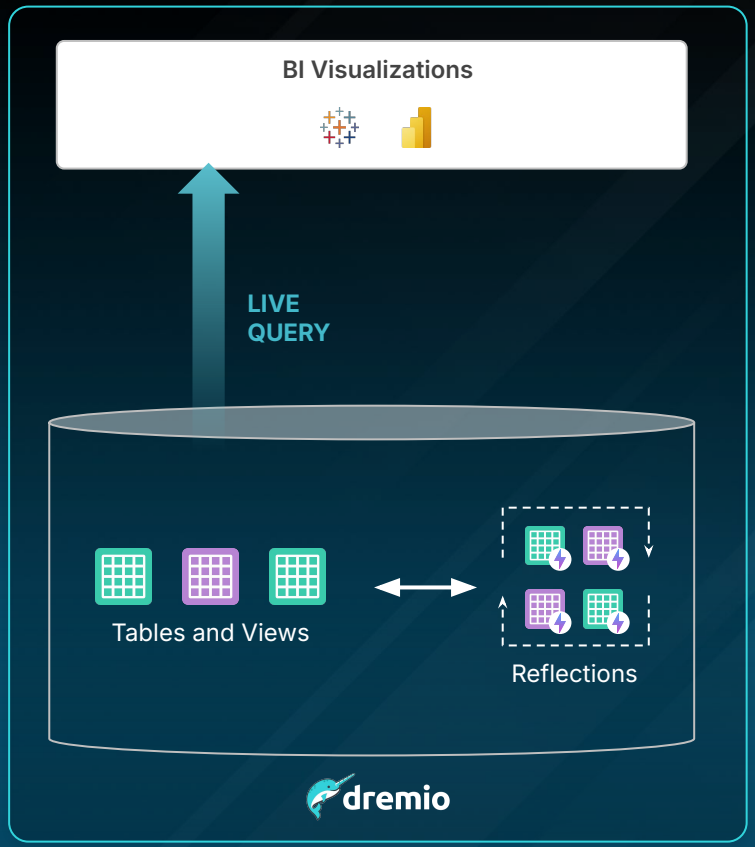
Google Cloud CLOUDERA

Starburst

Improved Performance and Automation

Reflections Eliminate the Need for BI Extracts & Imports

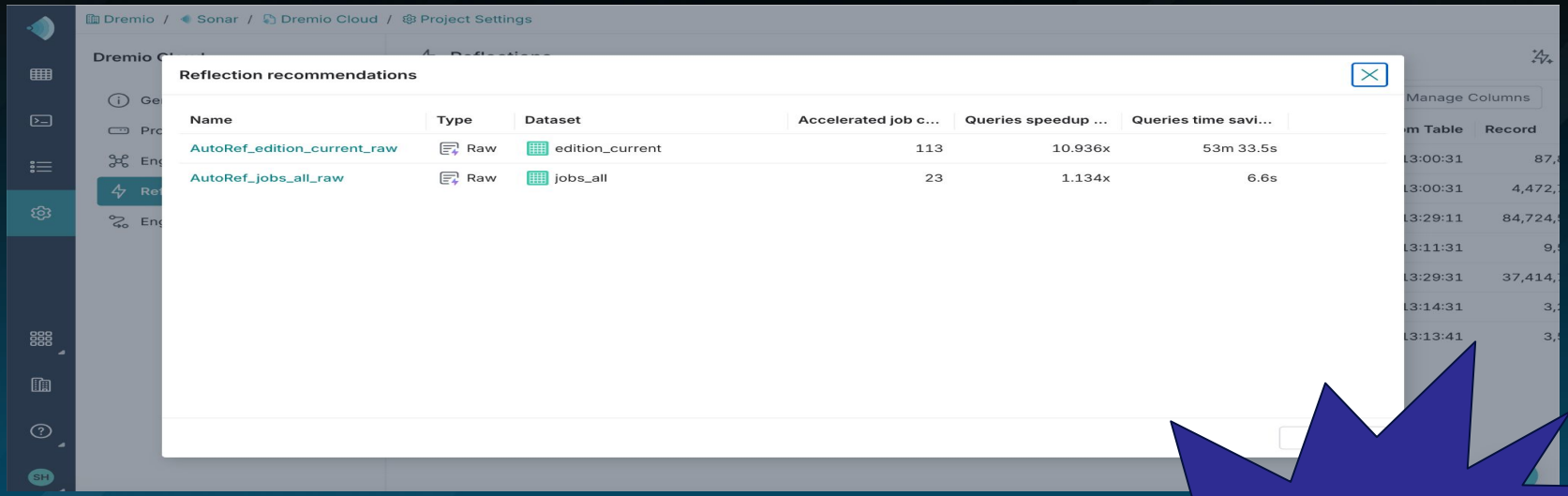




Reflection Recommendations

- Simplifies the workload of data engineers and enhance the performance of frequently executed queries on views by recommending raw Reflections with high ROI.
- Recommendations are generated daily, and are based on the query history of the preceding seven days.
- ROI is calculated based on the number of prior jobs that could be accelerated and the expected average improvement in performance.

Reflection Recommendations



To access the list of recommendations for your project and add new Reflections:

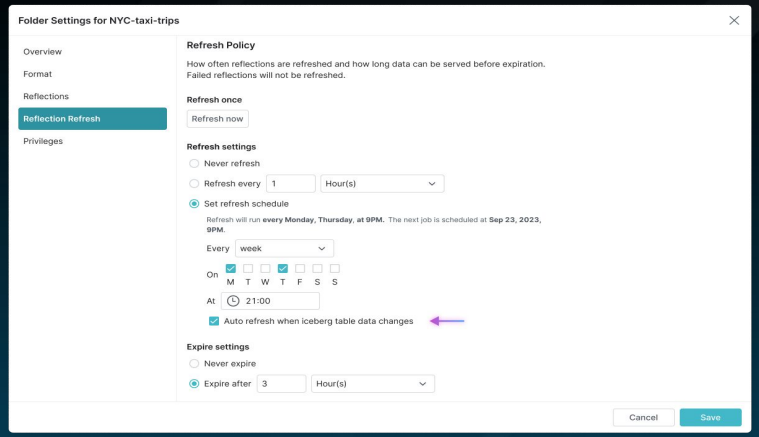
- Click the Project Settings icon in the side navigation bar.
- On your Project Settings page, click **Reflections**.
- To access your list of recommendations Click

To add a Reflection to your Reflections list, click at the end of the row.

dremio

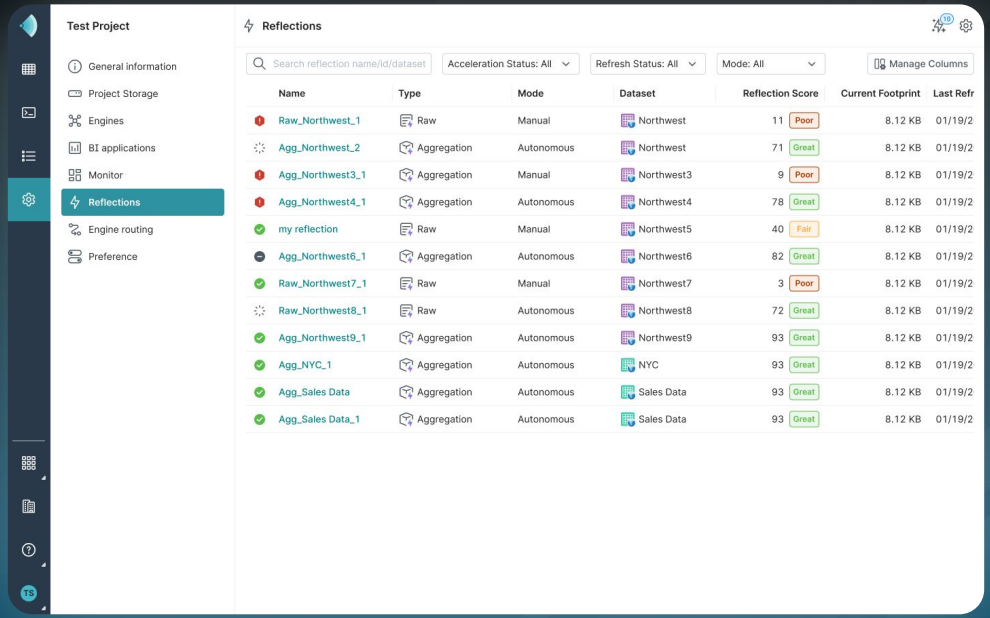
Live Reflections

- Simplifies the process of Reflection management and reduce administrative burden for data engineers.
- Reflections based on Iceberg tables are now automatically updated when their underlying tables are updated using the Dremio engine or other external engines.
- Live Reflections on Iceberg table can be enabled at source or at table level.



Reflection Scores

- Makes it easier for platform engineers to identify and take action on Reflections whose value is low or has dropped.
- Scores are based on the usage and acceleration speed up in the last 7 days.



Seamless Metadata Refresh

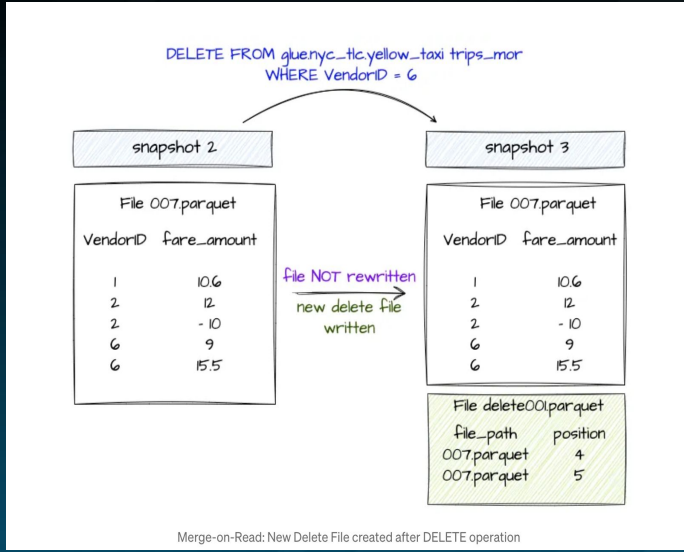
- Queries on Iceberg tables, will now do inline metadata refresh to provide users with the most current results at no performance penalty.
- It relieves data and platform engineers from the burden of setting appropriate metadata intervals and/or triggering metadata refresh.
- Reduces compute cost by not performing redundant metadata refresh on tables that are not frequently queried or updated.

Results Cache

- Accelerates the performance of queries that have been previously executed.
- Caches result sets for JDBC, ODBC and Flight type queries.
- Caches up to 20MB result sets and lasts for 24hrs.
- Stores the data in the customers distributed storage (S3 etc.)
- Benchmarking has shown a 28x increase in performance.
- If the cache is leveraged the plan and visual profile will reflect this.

Merge on Read

- New Iceberg V2 table property
- Increases write speed by writing deletes separately.
- Deleted records are tracked via delete files.
- On read, Data files and deleted records are … merged.



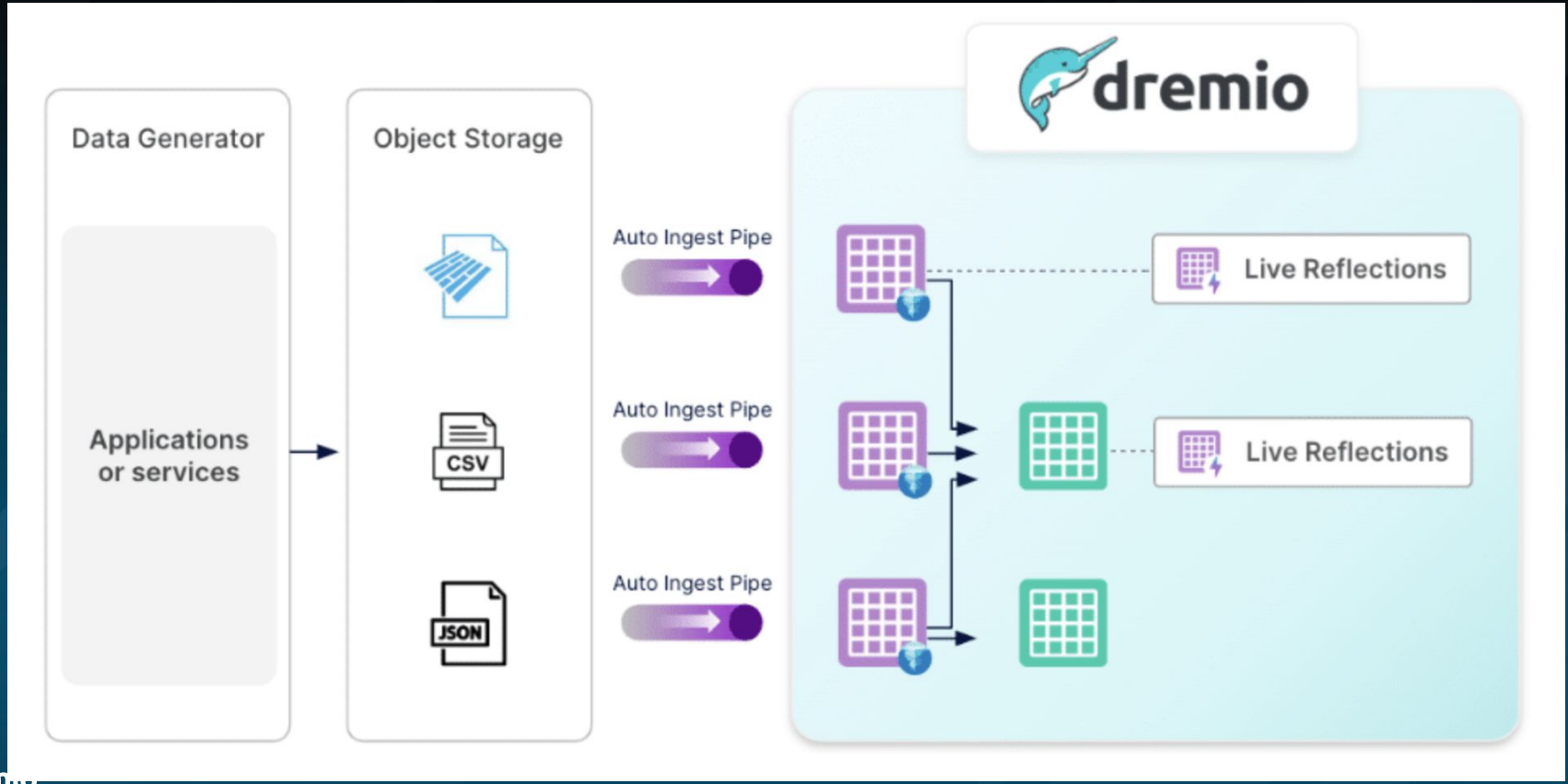
Merge-on-Read: New Delete File created after DELETE operation

Auto Ingestion for Iceberg Tables

Auto Ingest Pipes

- Reduces time and effort to build/operate/maintain ingestion pipelines for the Data Engineer.
- Quicker access to date reducing latency to near real time (~10min from file creation) for the Data Analyst.
- Automatically ingests data from S3 into Iceberg tables.
- Available for both Dremio Cloud and Dremio Software.
- Pipes have an auto deduplication capability and a default deduplication lookback period of 14 days.
- Auto Ingest Pipes leverage the cloud providers event driven architecture for providing immediate processing and ingestion.

Auto Ingest Pipes



Accelerated Cross-Database Access Control and Workload Management

Oracle DB, Teradata and Microsoft SQL Server Impersonation

- Queries submitted by Dremio can be configured to run under the users own username.
- Allows for existing access control rules to apply to queries coming from Dremio.
- More visibility of who is running what query for Administrators of respective systems.
- Allows for database native workload management capabilities to be leveraged more easily i.e. Teradata's TASM, Oracle's Automatic Workload Management, SQL Server's Resource Governor.

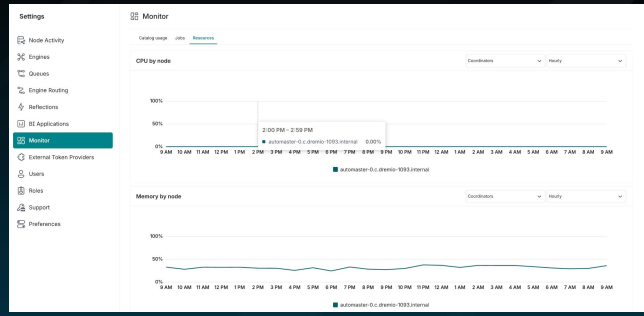
Integrated Observability

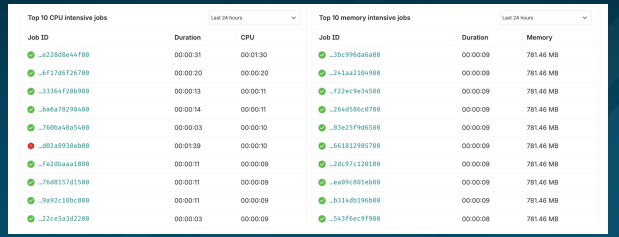
Administration and Monitoring with Integrated Observability

Out-of-the-box metrics for Resources

- Admins can now see the fluctuation of cpu and memory of engines on Software to understand peak times, need for scaling, etc.
- They also have a view of the top 10 queries that are most memory or cpu intensive.

This is in addition to the Catalog and Jobs observability that was shipped in 25.0.





Google BigQuery Connector

Google BigQuery Connector now in Preview

Enabled via a support key

- JDBC based connector for Google BigQuery
- As with other non-lake sources allows for the federation of datasets in a single query
- Where possible it pushes down the query to BigQuery to minimize data pulled across the wire
- JDBC driver must be installed by the customer, like with the Teradata connector

Graviton Support

Graviton Support

- AWS Graviton is a family of ARM based CPU's designed by Amazon
- Delivered in Amazon EC2
- Provide a lower energy use for workloads and a lower cost
- Dremio now supports Graviton instances in Dremio Software
- Dremio customers now has more choices when it comes to the instance type for their workload profile

Apache Polaris (Incubating)

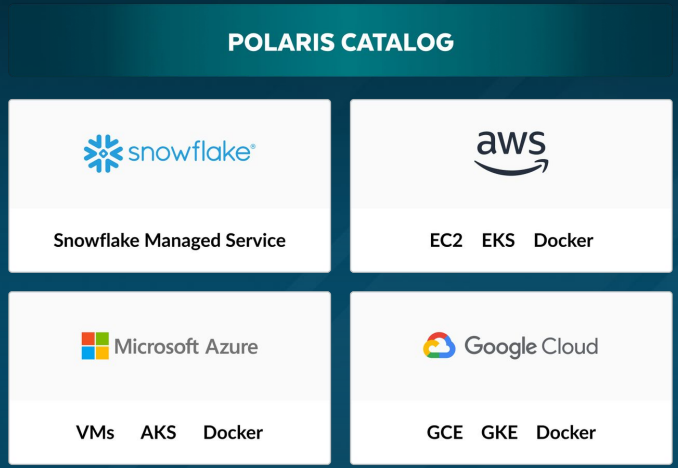
What is Apache Polaris (incubating)?

Apache Polaris (incubating) is a catalog implementation for Apache Iceberg built on the open source Apache Iceberg REST protocol that allows for:

- Cross-engine read and write interoperability
- Centralized access across engines
- Vendor-agnostic flexibility

What this means for customers:

- Maintain one data copy that many engines can query and write to
- Streamline data management with centralized access
- Run anywhere without lock-in



Apache Polaris (incubating) Impact

A stronger, converged, community model

Dremio is excited to help bring the various functions and capabilities of Nessie into the Apache Polaris (incubating) project.

- Inclusive community
- Robust open source catalog for open lakehouse architectures
- Reduce catalog sprawl
- Broader group of contributors

This partnership not only accelerates technical progress but also brings more contributors into the Nessie community, further strengthening the growing ecosystem around Polaris

Thank You

dremio