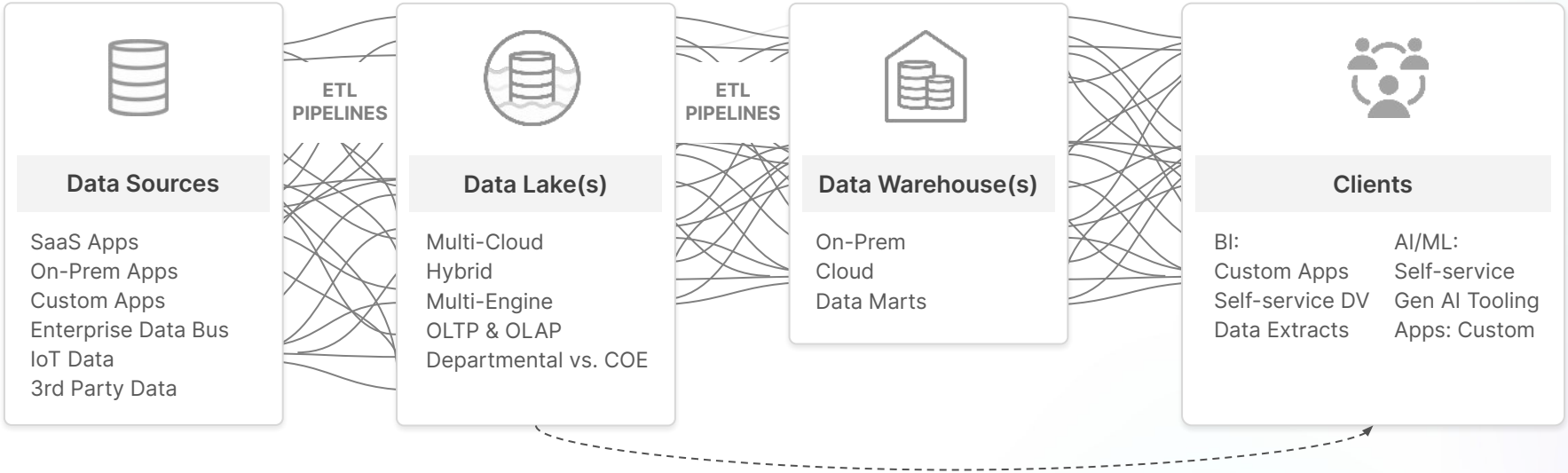
What's New in Dremio?

Mark Shainman, Dremio Product Marketing
Colleen Quinn, Dremio Product Marketing

Agenda

- Dremio Overview
- What’s New in Dremio?
- Get Started

Data lifecycle remains complex, brittle, and expensive



Data lifecycle and management remains complex, especially for large organizations
Duplicative copies, ‘expert’ ETL, “dark data”, governance complexity, not self service

Enterprises are moving to a lakehouse to simplify

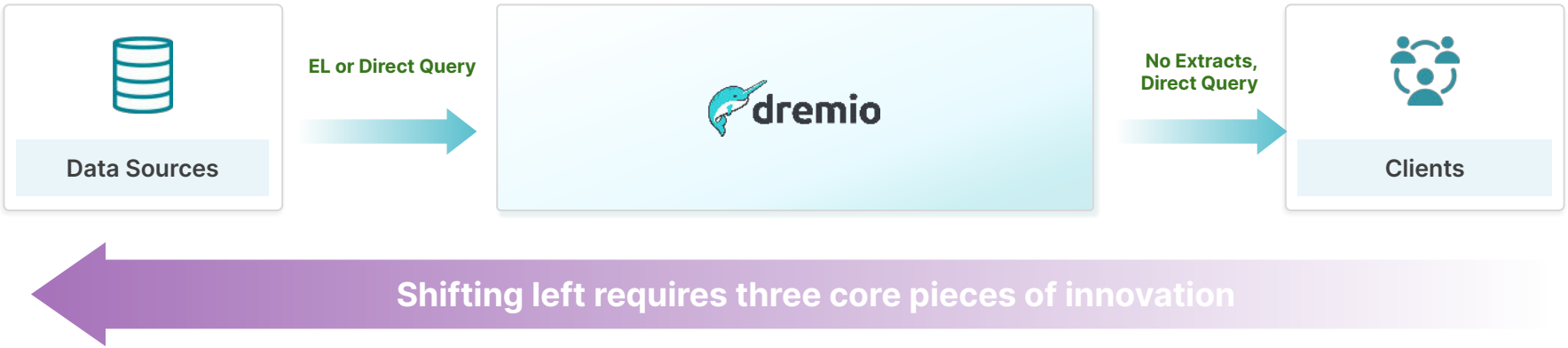
ETL to ELT

- Reduce complex transform pipelines in Java / Scala / Python (e.g., Spark)
- Move to SQL-based Transforms (DBT)
- Full transform lifecycle lives in the lake

Lakehouse Advantages

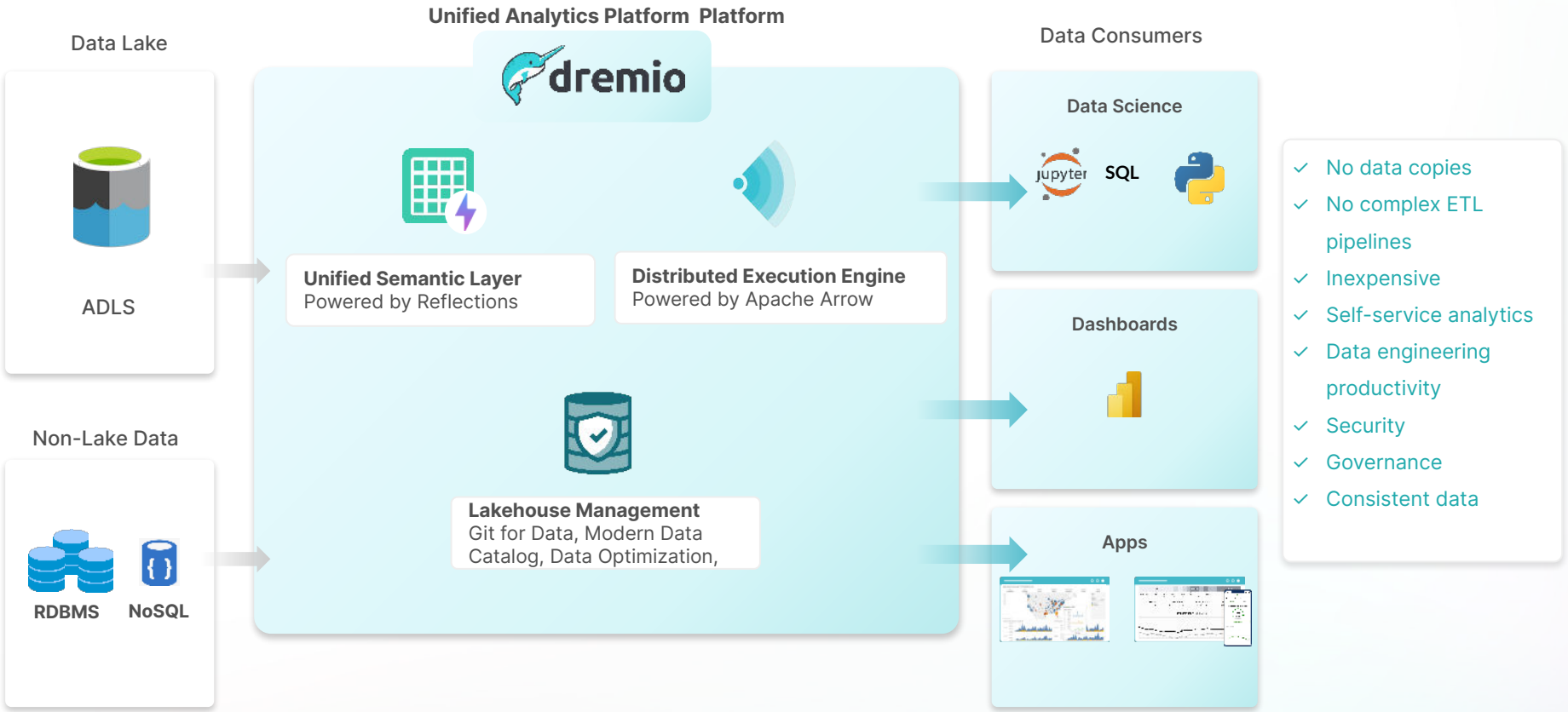
- Open data and table formats
- Storage / compute separated, elastic SQL engine
- No Copy Architecture
- Full ACID Transactions, Time-Travel, Schema / Partitioning Evolution
- Compelling Economics

Shifting Left Reduces MTTI and Shortens ETL Pipelines



Dremio Cloud

Now Available: Dremio Cloud Data lakehouse on Azure



Dremio Cloud in Azure: Eliminates the pain of managing infrastructure

Dremio Cloud

Split Plane architecture, control plane hosted by Dremio, execution hosted in customer tenant.

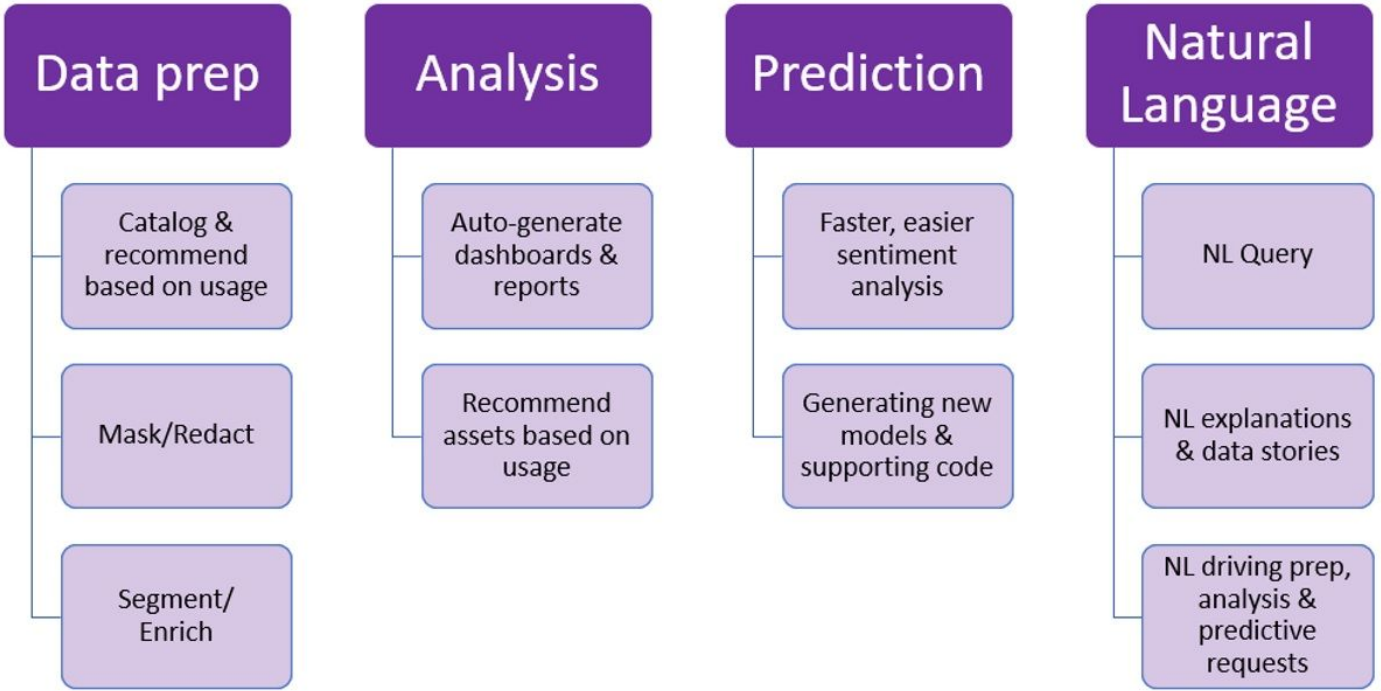
Cloud Features

- Always the latest functionality
- No management overhead
- Zero downtime automatic upgrades
- Automatic scalability
- Passwordless experience integrating with Enterprise SSO IdPs
- Managed Dremio environments with end-to-end encryption
- Capacity pricing

Dremio
Generative AI

GenAI to simplify data curation and analytics

Promising Use Cases for Generative AI to Advance Analytics/BI



Source: Constellation Research

New: Easy data curation with **GenAI**

Automatically Generate Wiki

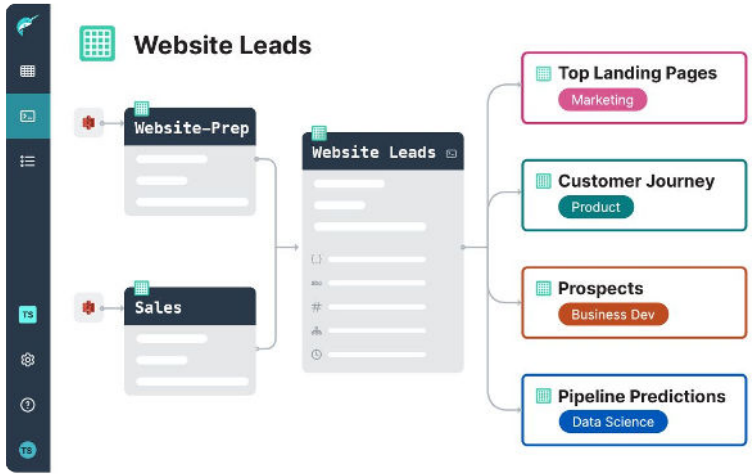
Automatic generation of dataset descriptions, SQL examples, and more.

Automatically Generate Labels

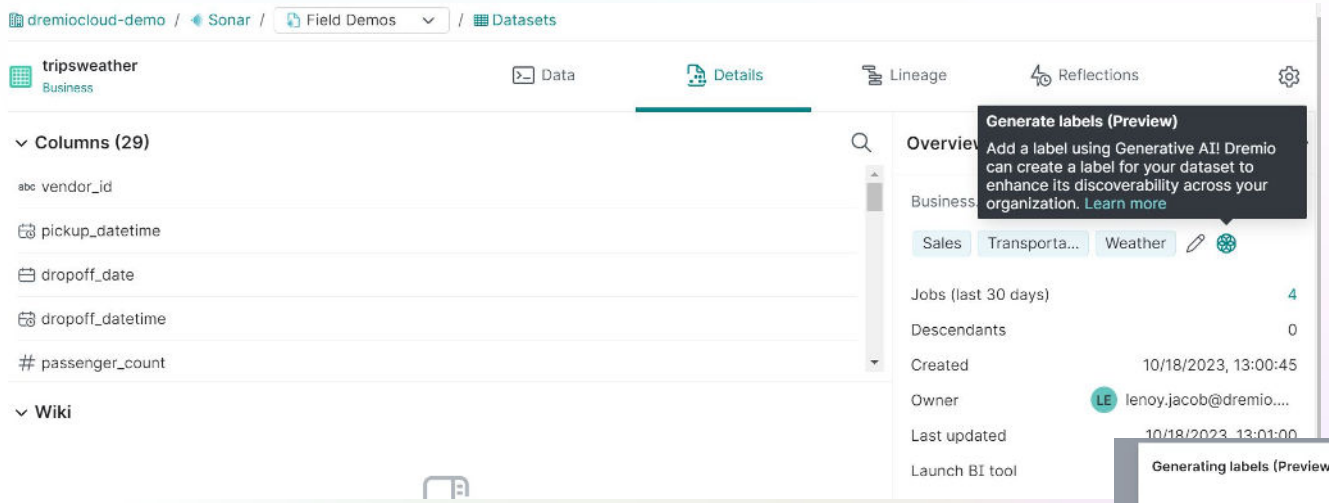
Automatic generation of data tags for tables and views.

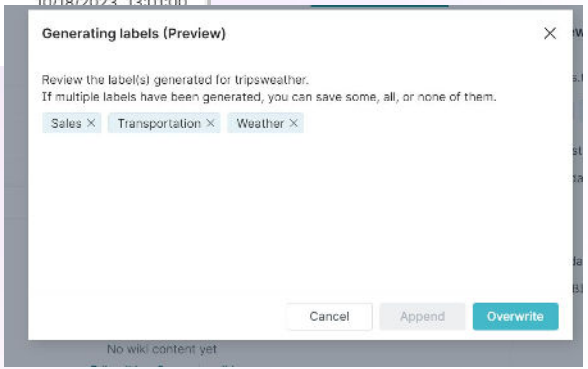
Autonomous Semantic Layer

- Easy data exploration for analysts and data scientists
- No manual enrichment

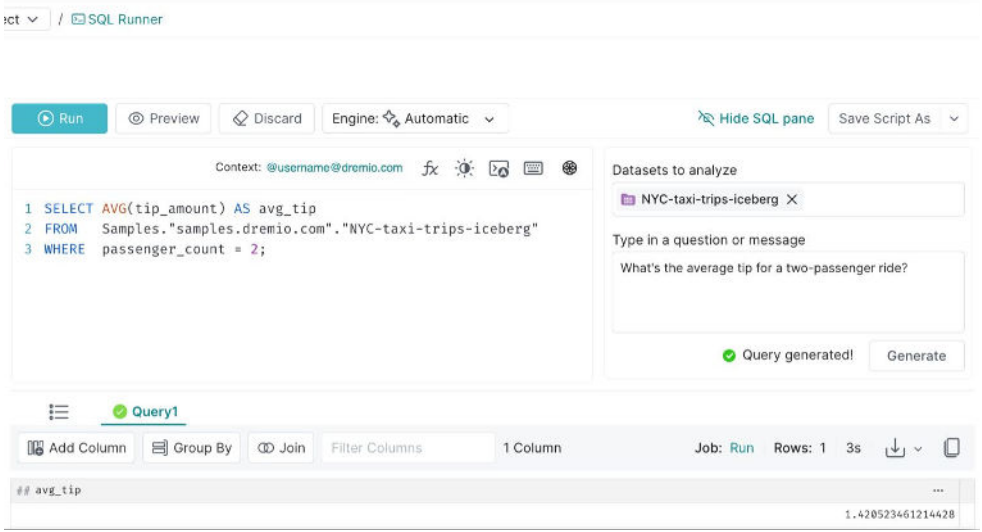


Look for the GenAI symbol





GenAI for Data Engineers and Analysts



TEXT-TO-SQL

- Generate SQL from natural language (Available now!)
- Refine generated SQL using Generative AI (Q1 2024)
- Ask questions on entire sources and catalogs (Q1 2024)

A Unified Path to Apache Iceberg

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 twillo

LinkedIn Adobe Tencent

Commercial Support for Iceberg

aws dremio snowflake

Google Cloud CLOUDERA

Starburst

New: Unified path to Iceberg

- Seamless **conversion to Apache Iceberg**
- New: Support for **Parquet** conversion
- Expands previous support **CSV and JSON**

Two-step conversion (and some housekeeping!)

Step 1: **Create Your Table** using CREATE TABLE

Step 2: **Copy data into Iceberg** using COPY INTO

Step 3: **Optimize your tables** using OPTIMIZE TABLE

Step 4: **Clean up your tables** using VACUUM CATALOG

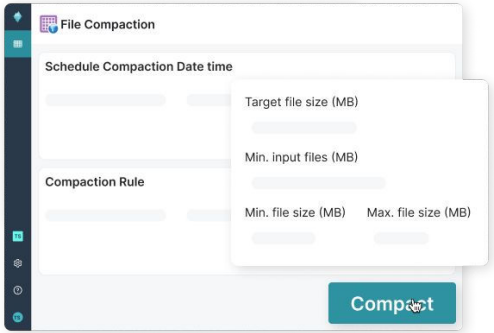
```
-- CREATE AN ICEBERG TABLE FROM OUR SALES TABLE FROM A
LEGACY VERSION OF THE TABLE IN POSTGRES
-- PARTITIONED BY SALES MONTH, SORTED BY SALES TIMESTAMP
CREATE TABLE arctic.db.sales
AS SELECT * FROM postgres.legacy_sales_table
PARTITIONED BY (month(sales_ts))
LOCALSORT BY (sales_ts);

-- COPY INTO SALES DATA FOR DECEMBER 2023 INTO TABLE
-- SALES DATA STORED AS CSV FILES ON OBJECT STORAGE
COPY INTO arctic.db.sales
FROM '@SOURCE/sales/2023/december'
FILE_FORMAT 'csv';

-- OPTIMIZE ALL NEW DATA INGESTED FOR THE LAST MONTH OF
SALES
OPTIMIZE TABLE arctic.db.sales
   REWRITE DATA USING BIN_PACK
   FOR PARTITIONS (sales_ts BETWEEN TIMESTAMP '2023-12-01
00:00:00' AND TIMESTAMP '2023-12-31 00:00:00');

-- VACUUM ALL DATA FROM BEFORE THE 90 DAY DATA RETENTION
POLICY
VACUUM TABLE arctic.db.sales
    EXPIRE SNAPSHOTS older_than '2023-10-03
00:00:00.000';  -- 90 days from January 1st, 2024
```

Ingest and optimize data automatically



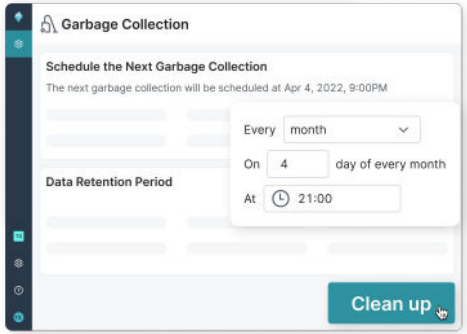


TABLE OPTIMIZATION

- Automatically **compact small files** and **group similar rows** together
- Table optimization significantly accelerates query performance

GARBAGE COLLECTION

- Automatically **remove unused data files, manifest files, and manifest lists** (Q4)
- Background cleanup ensures efficient use of data lake storage

INGESTION

- **Event-driven pipelines:** Automatically ingest from Amazon S3, ADLS, GCS (Q3)
- **Continuous ingestion:** Automatically write from Kafka topics into Arctic (Q4)

Expanded SQL Functions

New SQL array functions available now!

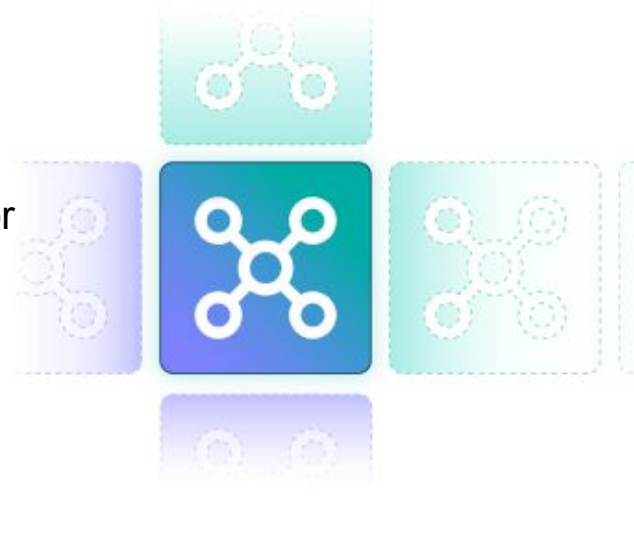
Signature	Description
`array_agg(expr)`	Returns an array consisting of all values in expr.
`array_append(A, E)`	Returns a new array with E at the end of A.
`array_distinct(A)`	Returns a new array with only the distinct elements from A.
`array_prepend(E, A)`	Returns a new array with E at the beginning of A.
`arrays_overlap(X, Y)`	Returns whether X and Y have any elements in common.
`array_to_string(A, S)`	Returns A converted to a string by casting all values to strings and concatenating them using S to separate the elements.
`set_union(X, Y, ...)`	Returns an array of all the distinct values contained in each array of the input.

Even faster!

Faster and More Performant

Query Engine

- **Query Execution -** Improve query performance and system resource utilization by 15% using TPC-DS benchmark
- **Query Planner -** using Iceberg statistics and optimization for superior out-of-the-box query performance



General Performance

- **Parquet 2.0 support using vectorized reader** - Improved performance **by up to 70%**

- **Updated Tableau connector using Flight JDBC** - Built-in Arrow Flight connector will enable ~30% improvement in Tableau query performance

Dremio Self-Managed Software

Dremio Self-Managed: Kubernetes elasticity

Dremio Self-Managed

K8S-Based Elasticity

Elasticity Features

- Infrastructure and Concurrency based elasticity rules
- Works with WLM (query routing, concurrency)
- Engine Pilot-Light Model
- Metrics reporting for scalability
- Works with CSP K8S managed offerings

Requires K8S-based deployment (not AMI-based)

Get Started

Ready to get started?

- Current Dremio Software customer? Visit Dremio Support Portal to download.
- Current Dremio Cloud customer? It’s live!
- New to Dremio? Try it for free using Dremio Cloud or the Self-Managed Community Edition

Thank you!

How it works