ATL Iceberg Meetup

12 Nov 2025

Lester Martin

Trino Developer Advocate @ Starburst

Lakeside AI with SQL & Apache Iceberg

Connection before content

Lester Martin – https://linktr.ee/lestermartin

- Developer Relations @ Starburst
  - Blogging & forums
  - Webinars & videos
  - User groups & events
  - Training & tutorials
- 30+ years of technology experience
  - Started journey on TRS-80 Model III
  - Played most roles, but a programmer at my core
  - ½ career in OLTP and ½ in data analytics
  - Decade+ of "big data" experience to include
    - Trino/Starburst, Hadoop, Hive, Spark
    - NiFi, Kafka, Storm, Flink
    - HBase, MongoDB

lester.martin@gmail.com

Starburst

Trino ? https://trino.io

Ludicrously fast, open source , distributed , massively parallel processing , SQL query engine designed to query large data sets from one or more multiple data sources .

trino

Trino trusted by industry leaders at PB scale

- Open-source query engine.
- Separates compute and storage.
- Queries across all data sources.
- Iceberg was designed for Trino.

Proven at exabyte scale/high concurrency:

25PB on S3

1 Exabyte of Data
100PB weekly data
1200 nodes
2.5M queries/week

600PB on S3
1000 nodes

10PB daily read data
250k queries per day

300PB data lake

Trino open source users

Starburst is the Trino company:

- **Bringing Trino to the enterprise**
- **Cofounded by Trino creators**
- **#1 Trino committer**
- **Largest team of Trino experts in the world**

Thriving open source community:

11300+
SLACK MEMBERS

10,000+
GITHUB STARS

750+
CONTRIBUTORS

Starburst

Server stereotypes

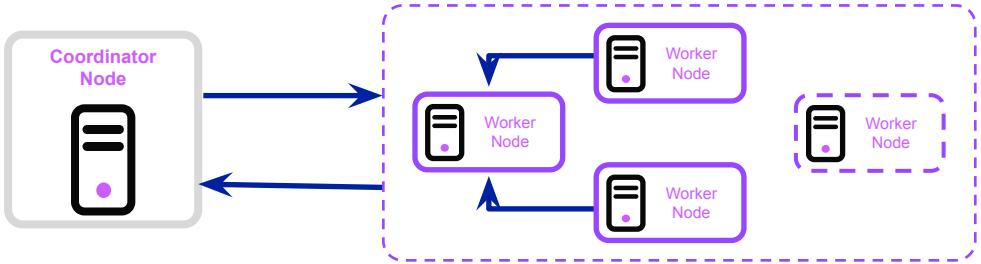
Coordinator node

Server that is responsible for parsing statements, planning queries, and managing Trino worker nodes.

Worker nodes

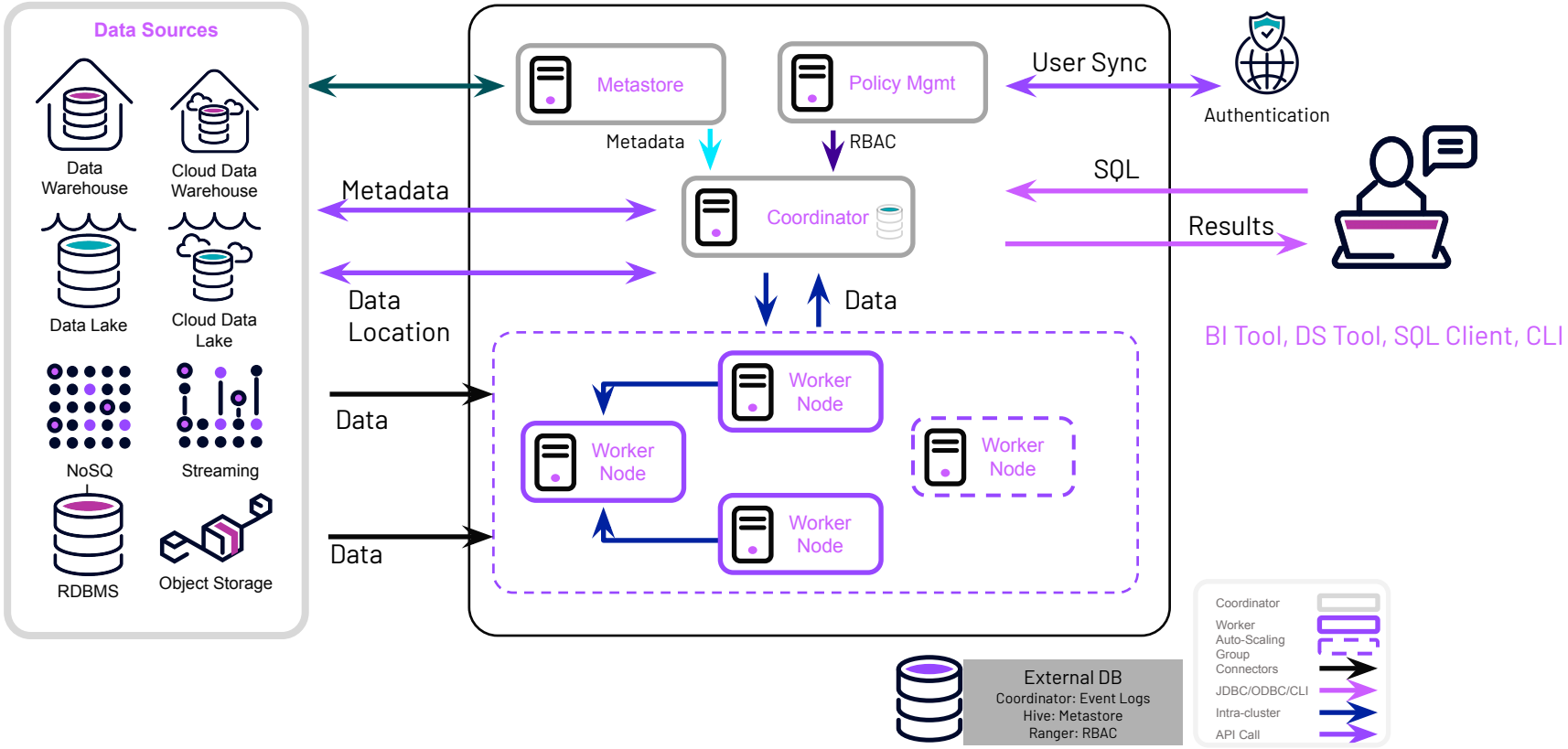
Server which is responsible for executing tasks and processing data. Worker nodes fetch data from connectors and exchange intermediate data with each other.

trino



Starburst

Logical architecture



Connectors

- SQL queries on **Data Lake / Data Lakehouse (HDFS, Object Storage)**
- **Single point of access** centralizes security and governance
- **Federation** between different data sources

Official data lake components #

Other data lake components #

Official data sources #

Other data sources #


```
SELECT
    c.custkey,c.estimated_income,
    a.products,a.cc_number,
    cp.customer_segment
FROM
    Hive.burst_bank.customer c
    JOIN MongoDB.burst_bank.account a ON c.custkey = a.custkey
    JOIN Oracle.burst_bank_large.customer_profile cp ON c.custkey = cp.custkey
WHERE
    c.state <> 'OK'
    AND a.mortgage_id IS NOT NULL;
```

trino

Trino
&
Apache Iceberg

= Open Data Lakehouse

trino

The Open Data Lakehouse – *The* ***Icehouse***

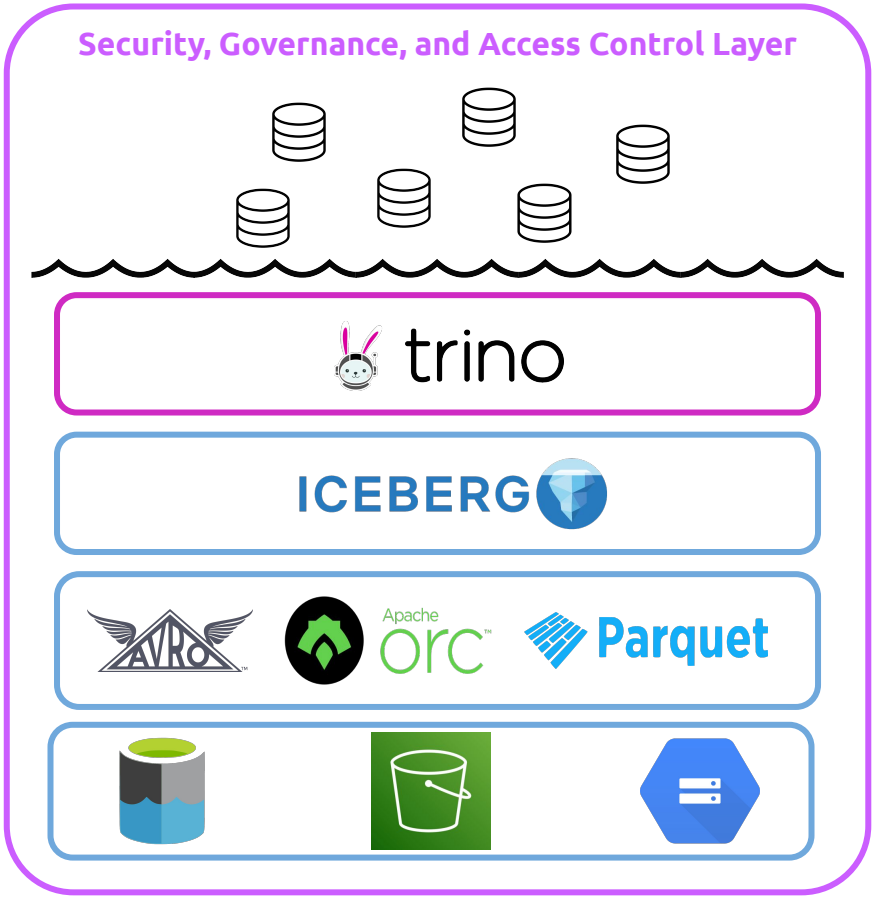
Global federated access to data sources beyond the lake

Compute engine

Table formats

Open file formats

Commodity storage



Access data in the orbit

Powers the data lakehouse

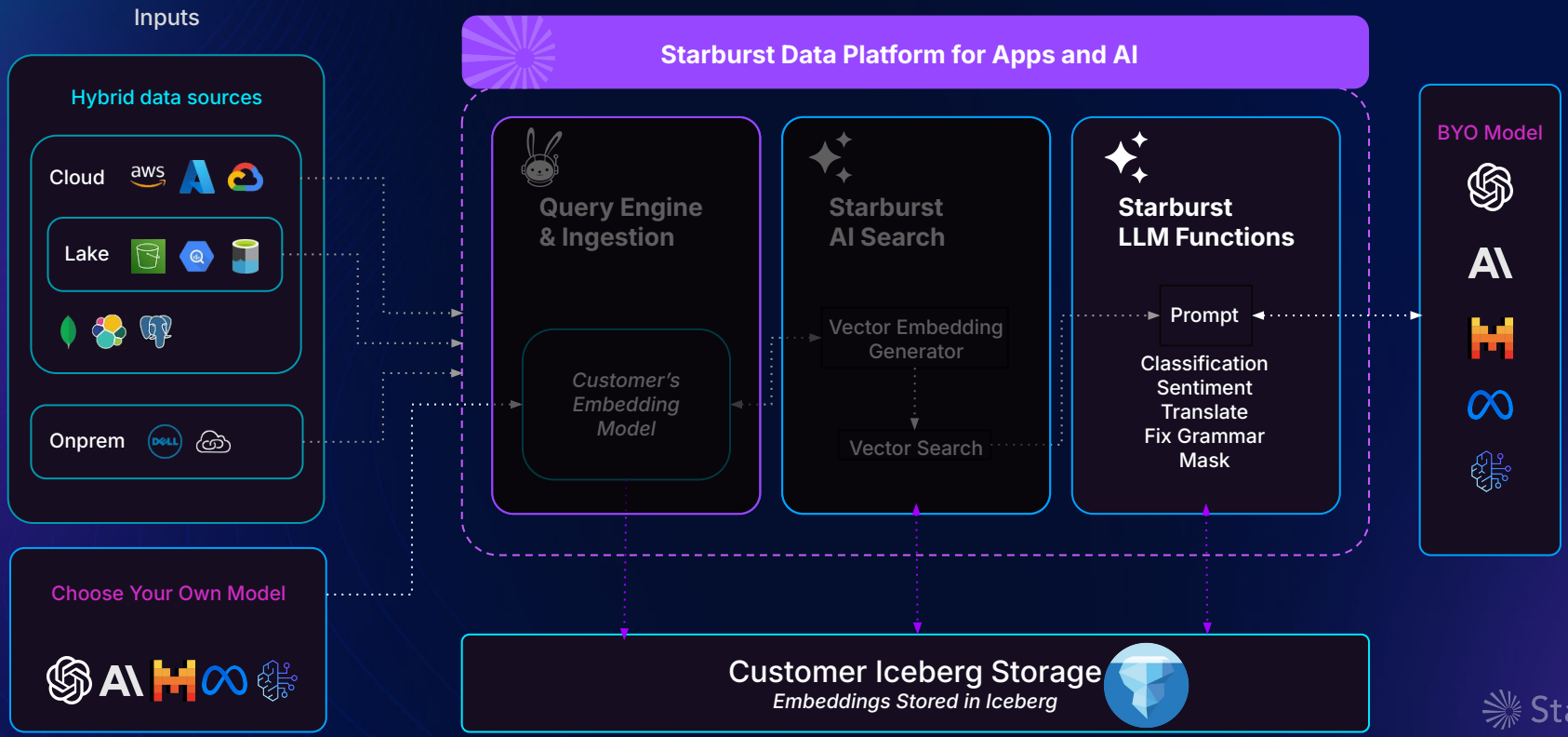
Center of gravity

Enables data lakehouses

Starburst

AI Functions

Functions for carbon-based life forms & data apps



Starburst

Empower any SQL user to move faster from raw data to enriched insight with built-in AI SQL Functions

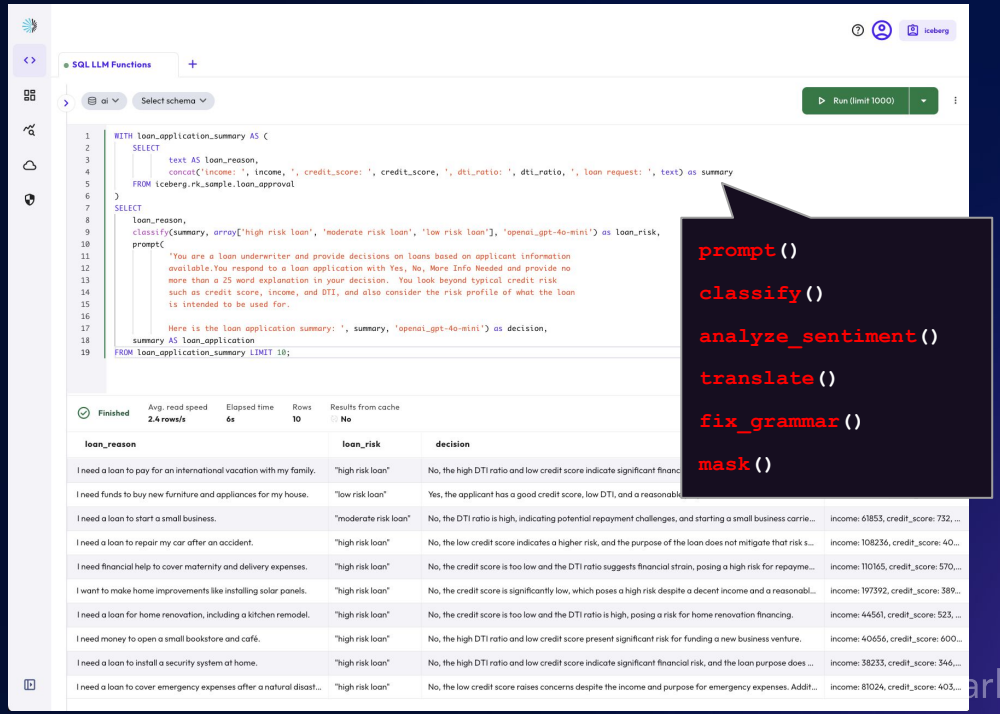
Use SQL to prompt LLMs or apply built-in AI tasks to structured and unstructured data.

LLM SQL Prompt Functions

Send context-rich prompts to your LLM of choice directly from a SQL query.

Task-Specific AI SQL Functions

Be a SQL developer, not a prompt a engineer, with a library of built-in, pre-tested functions that all work with a customer's choice of model.



Demo of AI functions

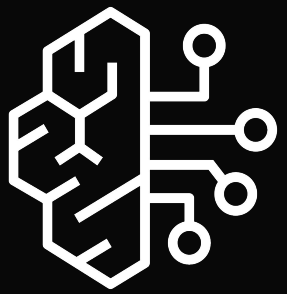
SQL available at

https://github.com/lestermartin/events/blob/main/2025-11-12_ATL-IcebergMeetup/functions.sql

trino

Augmented Generation (RAG & TAG)

What is Retrieval Augmented Generation (RAG)?



LLM Models

The Intelligence

+

Enterprise Data

The Context

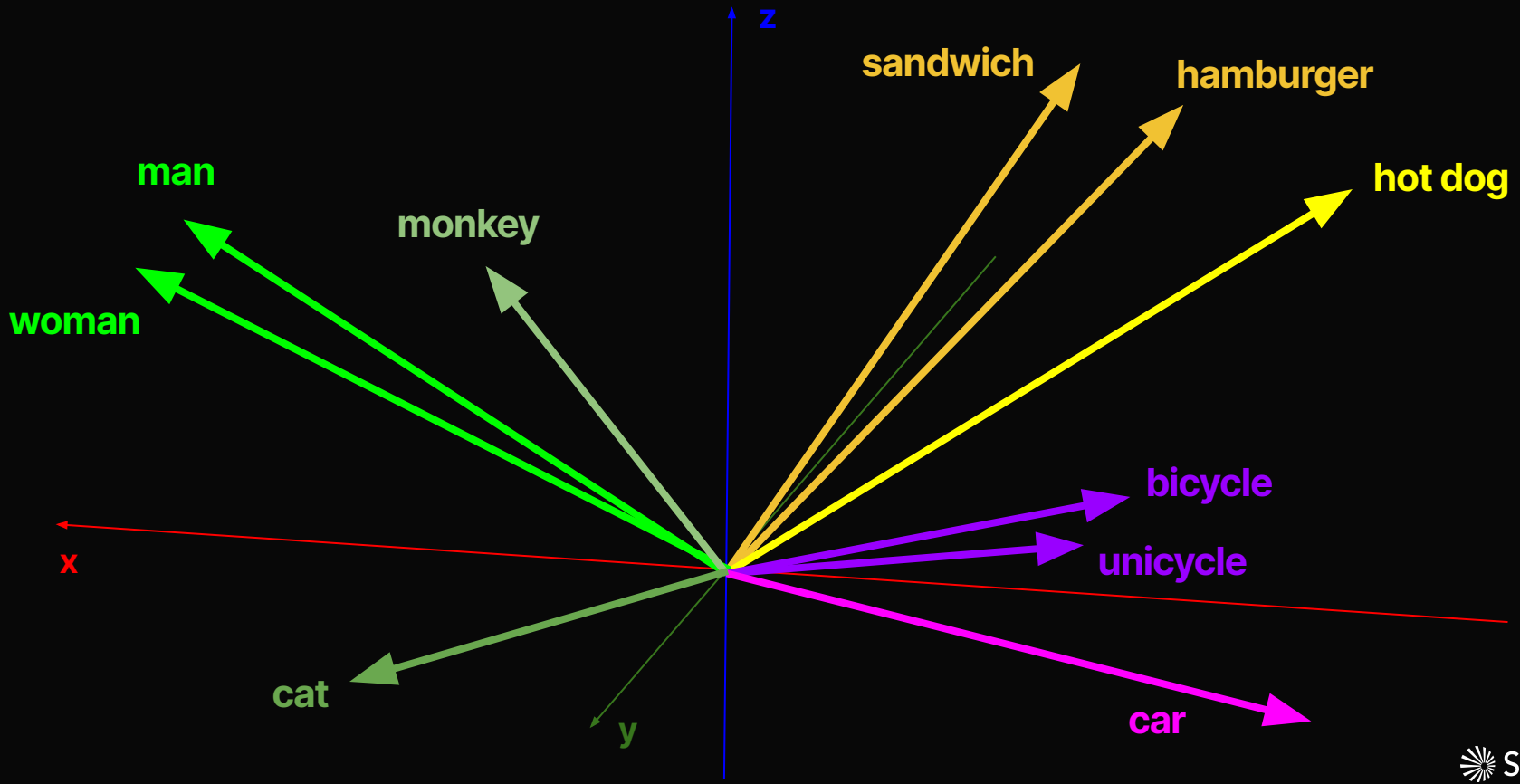
=

RAG Application

The Output

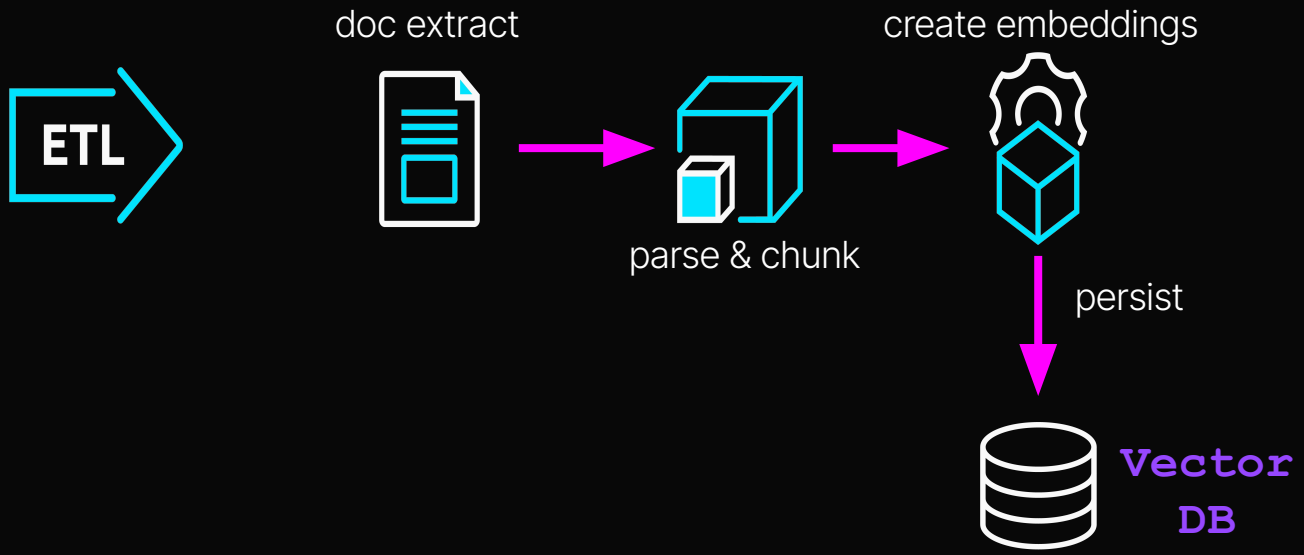
Starburst

Text chunks & vector embeddings

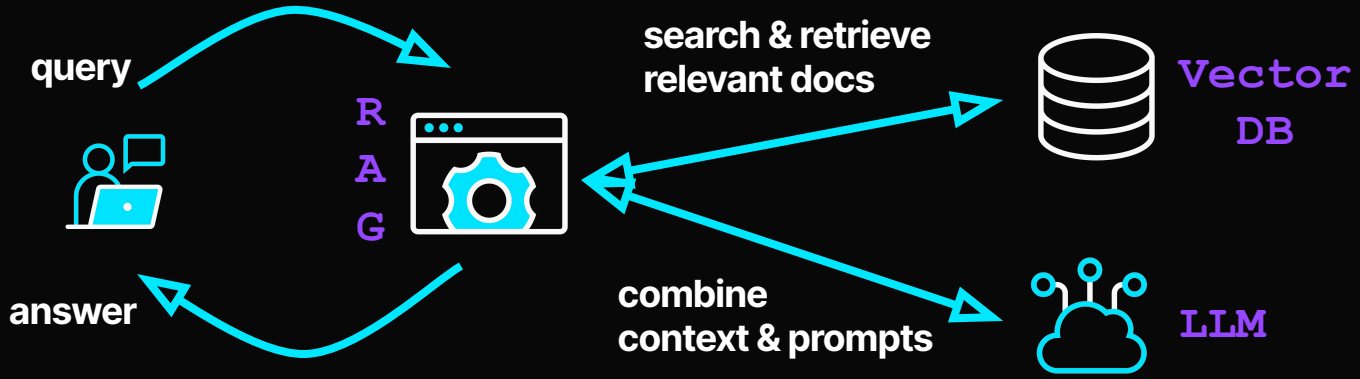


Starburst

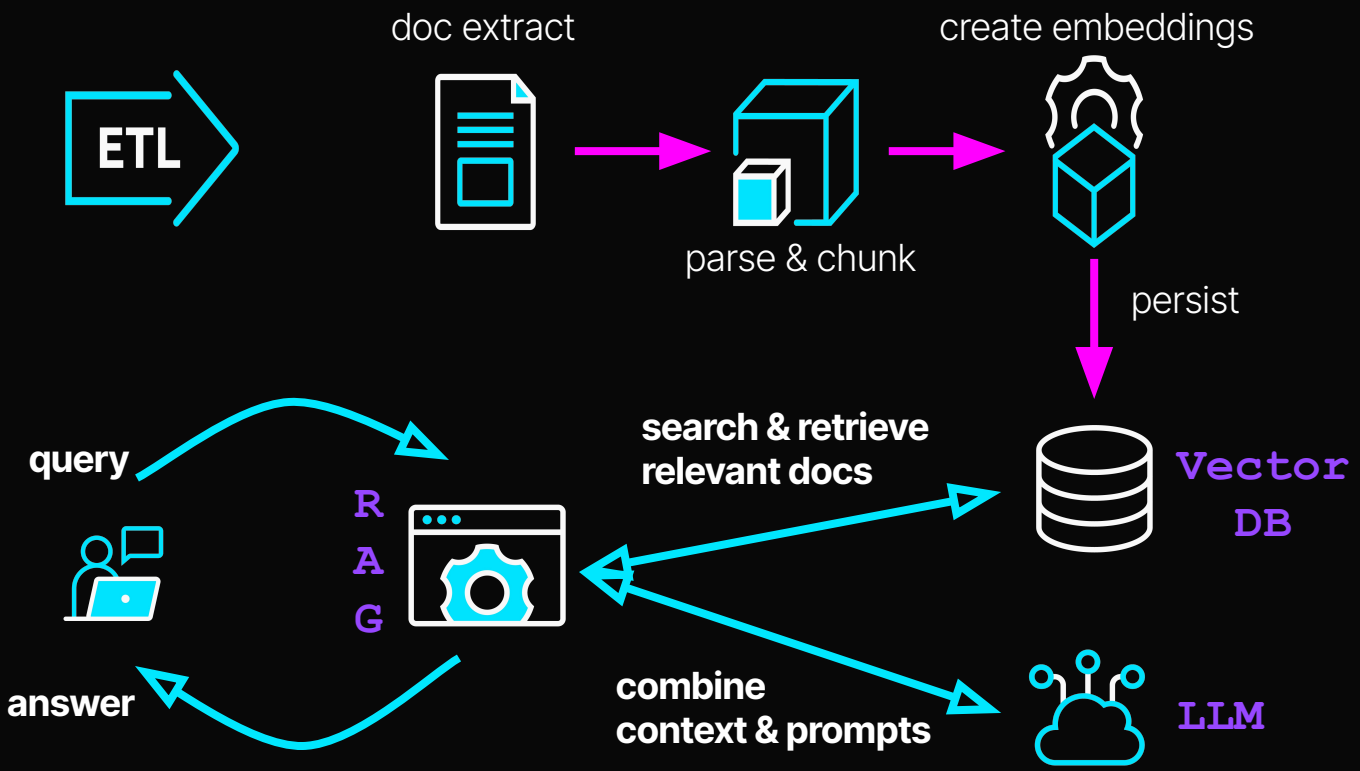
Visualization of a RAG data pipeline (unstructured)



Visualization of a RAG application

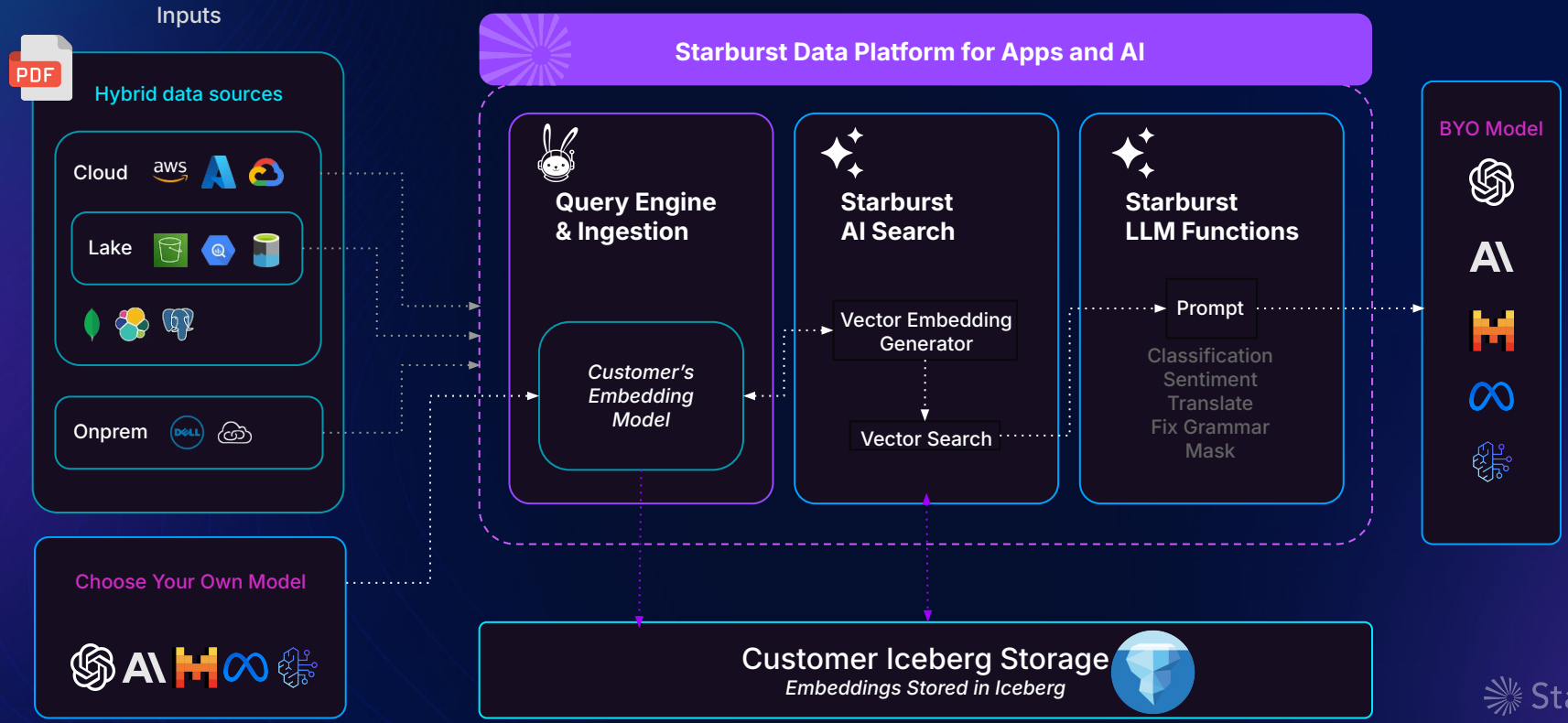


Visualization of a RAG data pipeline & application



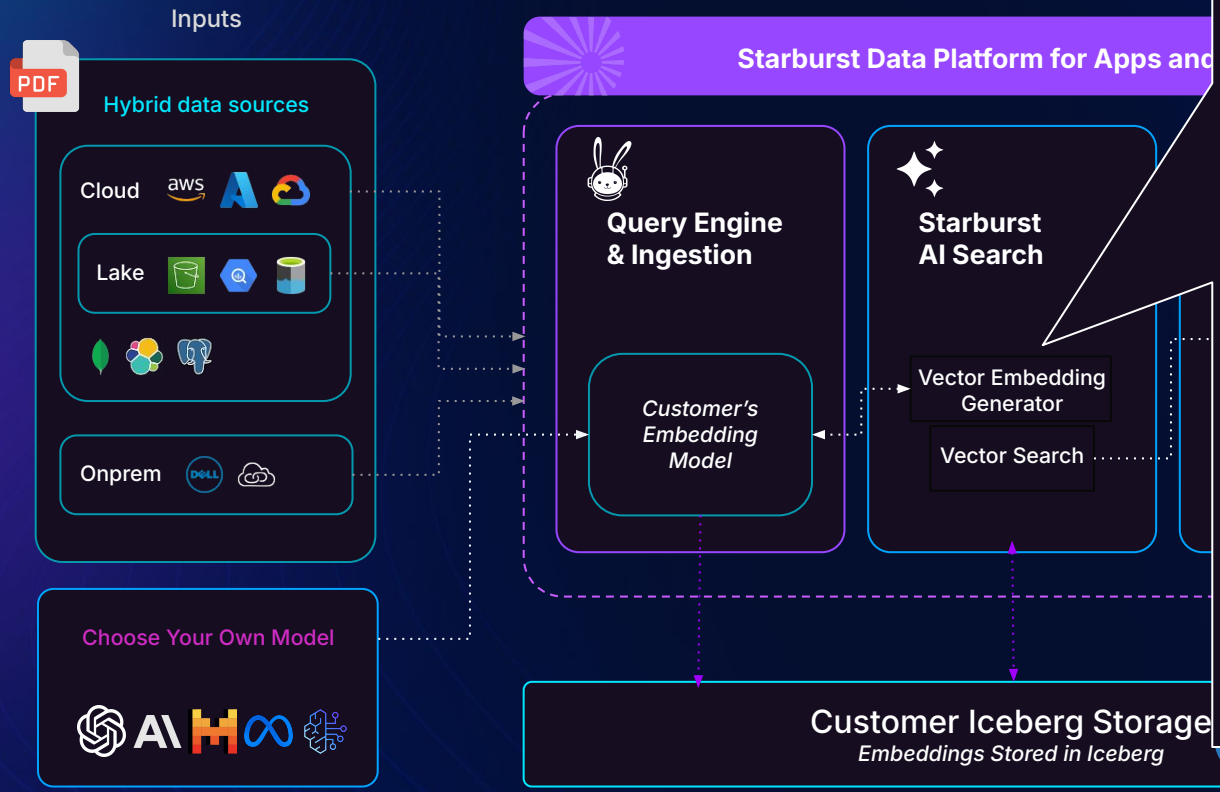
Starburst

Bring enterprise data to AI workflows & agents



Starburst

Generate embeddings and store them in the lake

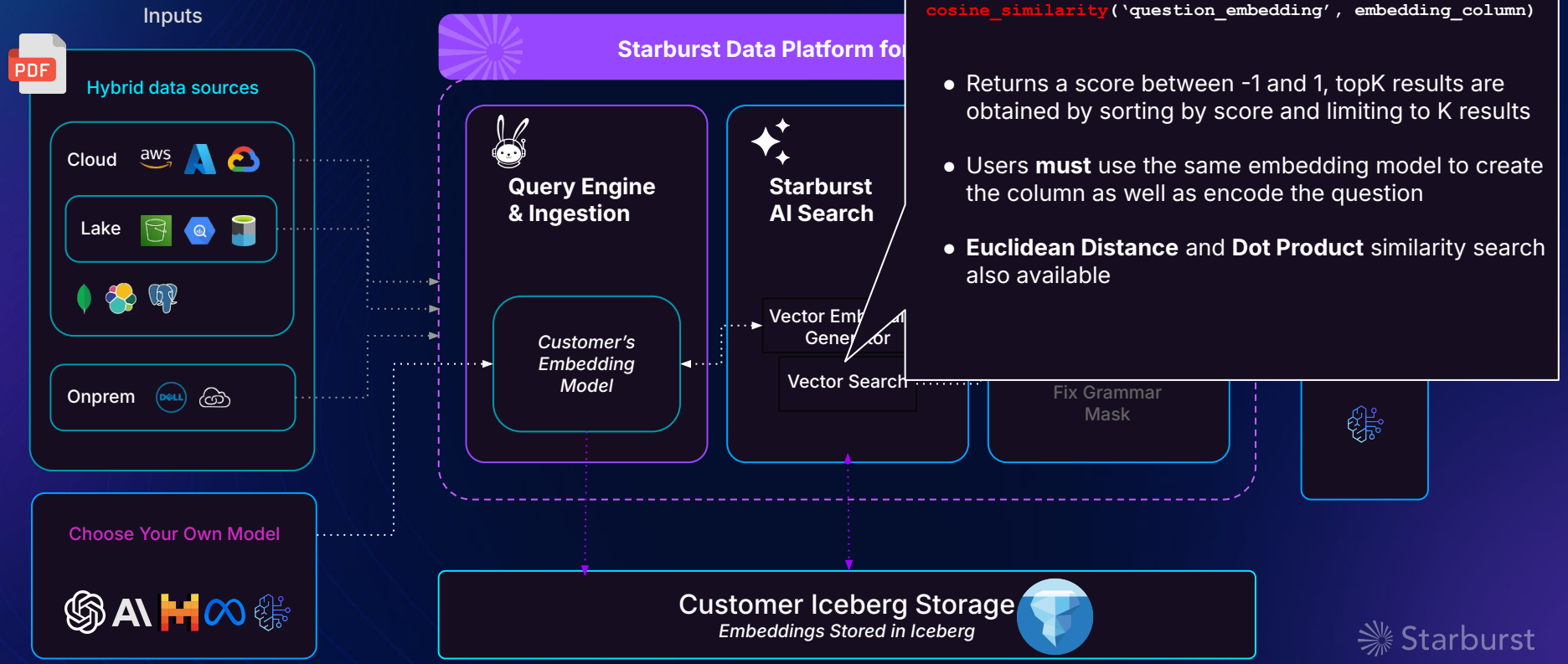



```
UPDATE iceberg.my_schema.faa_book_chapters
SET chapter_intro_embeddings =
generate_embedding(my_column, 'cohere-embed3');
```

- Writes embedding as `ARRAY(double)` into Iceberg table
- Ideal if the table already has unstructured data (e.g. `review` column in a customer support feedback table)
- PDFs, word documents, and other flat files require document prep first

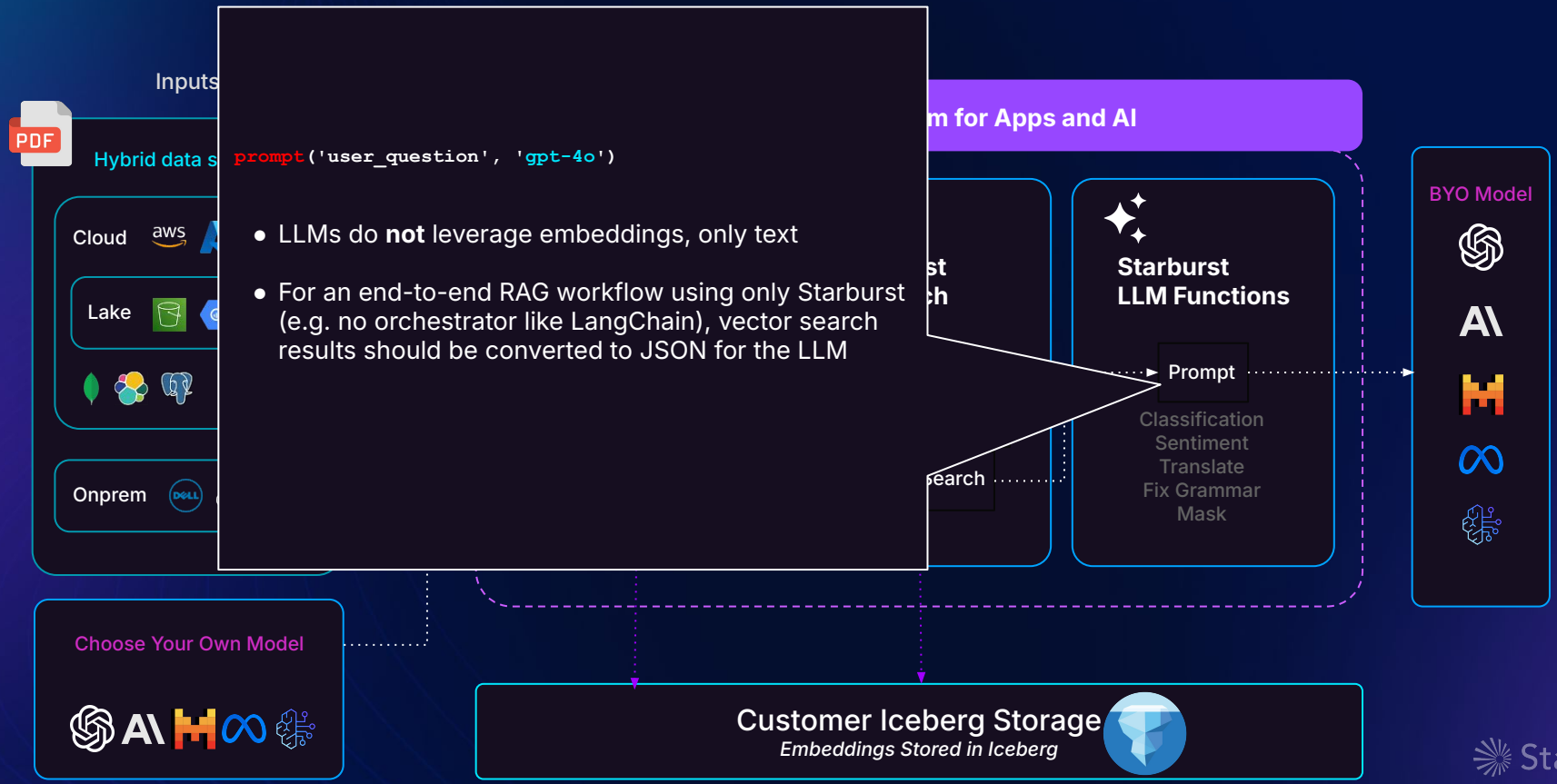
Starburst

Vector search using SQL on top of your data lake



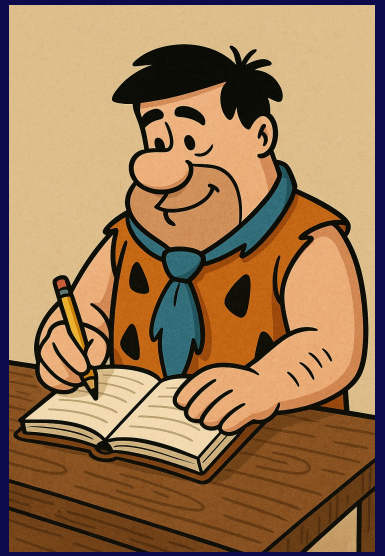
Starburst

Prompt an LLM with your search results



Starburst

Demo of RAG workflow



SQL available at
https://github.com/lestermartin/events/blob/main/2025-11-12_ATL-IcebergMeetup/workflow.sql

trino

Thanks | धन्यवाद | Grazie | 谢谢 | Merci | ありがとう | Gracias | 감사합니다 | Danke

Questions ?

Test with Trino managed service (Start Free)

Starburst Galaxy

devrel@starburst.io

Build RAG Apps with Starburst AI Workflows Webinar (on demand)

Exploring Data Products and Starburst AI Agent Webinar (on demand)

Starburst

Sign up for our newsletter and enter to win a limited-edition Starburst Community T-shirt!

Starburst | trino

trino