

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**



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 trusted by industry leaders at PB scale



# trino

- ✓ 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

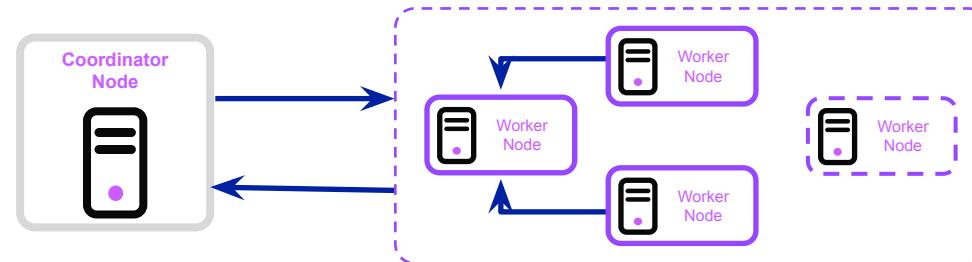
# Server stereotypes

## Coordinator node

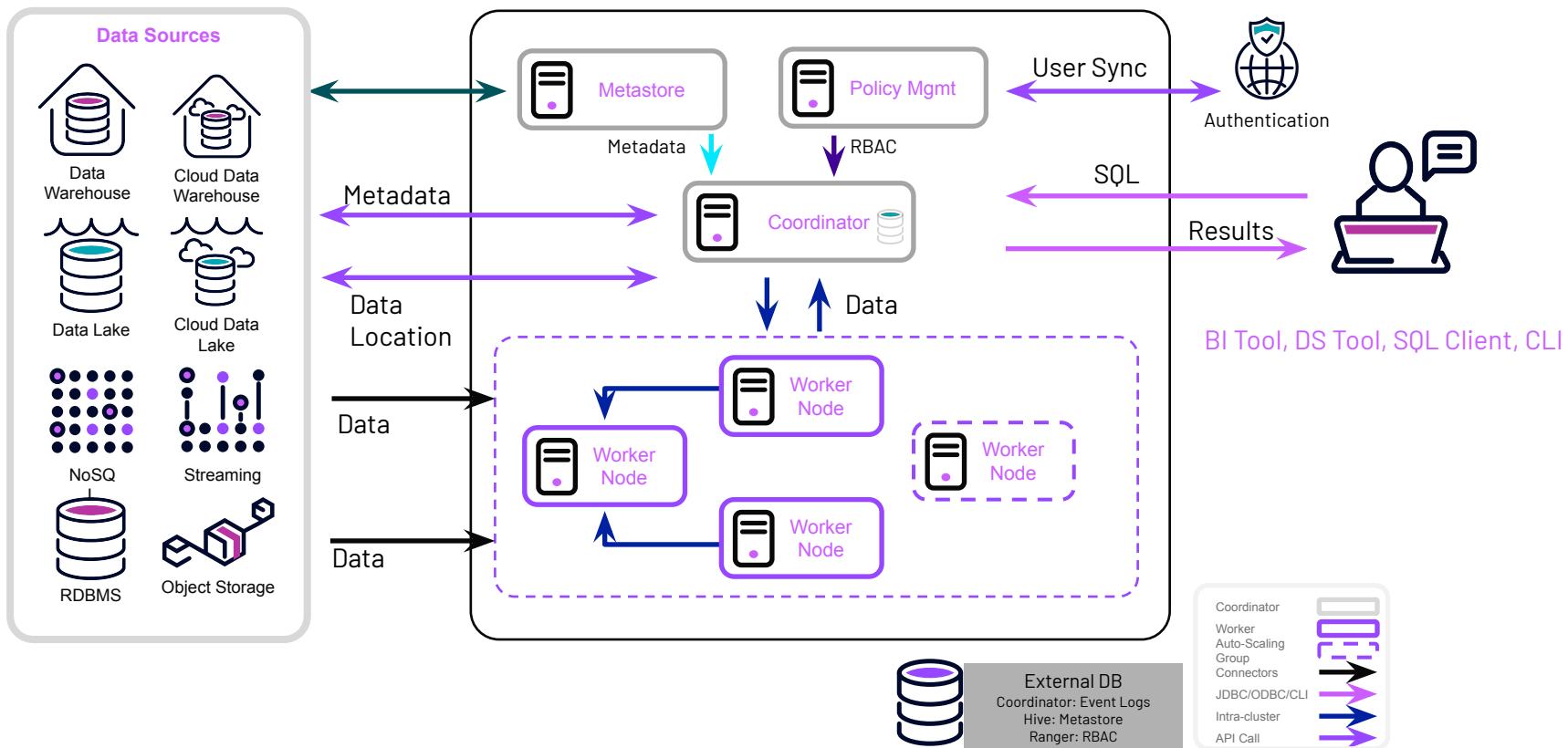
Server that is responsible for parsing statements, planning queries, and managing Trino worker nodes.

## 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.



# Logical architecture



# Connectors

## Official data lake components #



## Other data lake components #



## 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;
```

- 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 sources #



## Other data sources #



# Trino & Apache Iceberg

= Open Data Lakehouse



# The Open Data Lakehouse – The *Icehouse*



Global federated access to data sources beyond the lake

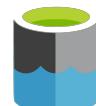
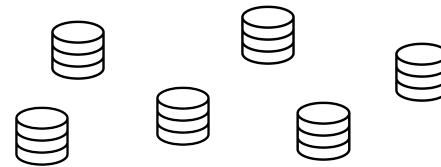
Compute engine

Table formats

Open file formats

Commodity storage

Security, Governance, and Access Control Layer



Access data in the orbit

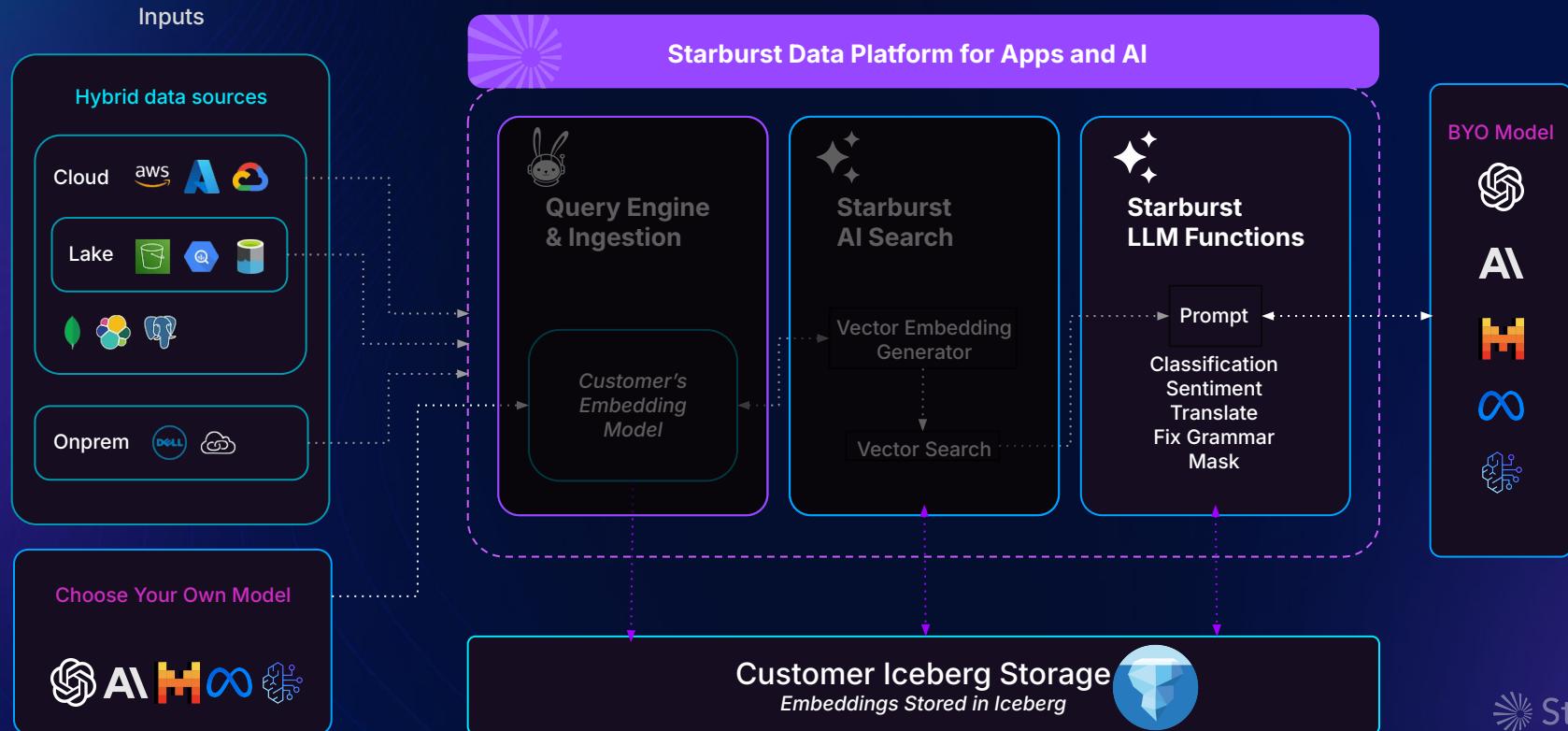
Powers the data lakehouse

Enables data lakehouses

Center of gravity

# AI Functions

# Functions for carbon-based life forms & data apps



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

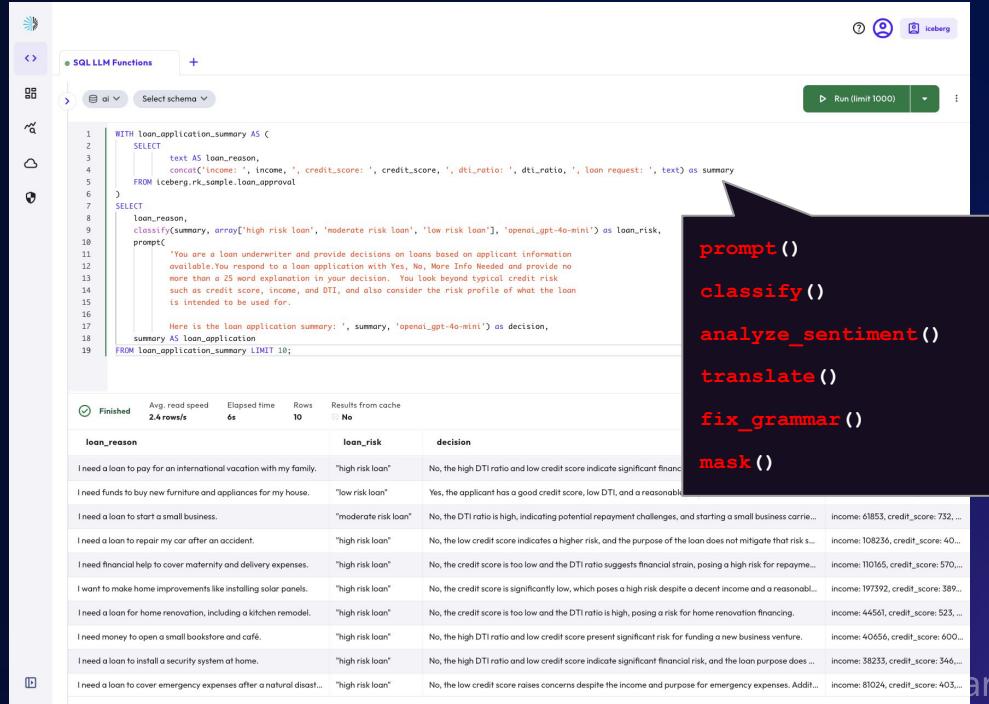
Use SQL to prompt LLMs or apply built-in AI tasks to structured and unstructured data.

## 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 engineer, with a library of built-in, pre-tested functions that all work with a customer's choice of model.



The screenshot shows a SQL editor interface with a code editor and a results table. The code editor contains the following SQL query:

```
WITH loan_application_summary AS
  SELECT
    text AS loan_reason,
    concat('income: ', income, ', credit_score: ', credit_score, ', dti_ratio: ', dti_ratio, ', loan request: ', text) AS summary
  FROM iceberg.rk_sample.loan_approval
)
SELECT
  loan_reason,
  classify(summary, array['high risk loan', 'moderate risk loan', 'low risk loan']), 'openai.gpt-4o-mini') AS loan_risk,
  prompt(
    "You are a loan underwriter and provide decisions on loans based on applicant information available. You respond to a loan application with Yes, No, More Info Needed or provide no more than a 25 word explanation in your decision. You look beyond typical credit risk such as credit score, income, and DTI, and also consider the risk profile of what the loan is intended to be used for.

    Here is the loan application summary: ", summary, "'openai.gpt-4o-mini') AS decision,
  summary AS loan_application
  FROM loan_application_summary LIMIT 10;
```

A callout box highlights the AI functions used in the query:

- `prompt()`
- `classify()`
- `analyze_sentiment()`
- `translate()`
- `fix_grammar()`
- `mask()`

The results table shows the output of the query, displaying columns for `loan_reason`, `loan_risk`, and `decision`. The results are as follows:

loan_reason	loan_risk	decision
I need a loan to pay for an international vacation with my family.	"high risk loan"	No, the high DTI ratio and low credit score indicate significant financial risk.
I need funds to buy new furniture and appliances for my house.	"low risk loan"	Yes, the applicant has a good credit score, low DTI, and a reasonable purpose.
I need a loan to start a small business.	"moderate risk loan"	No, the DTI ratio is high, indicating potential repayment challenges, and starting a small business carries additional risks.
I need a loan to repair my car after an accident.	"high risk loan"	No, the low credit score indicates a higher risk, and the purpose of the loan does not mitigate that risk significantly.
I need financial help to cover maternity and delivery expenses.	"high risk loan"	No, the credit score is too low and the DTI ratio suggests financial strain, posing a high risk for repayment.
I want to make home improvements like installing solar panels.	"high risk loan"	No, the credit score is significantly low, which poses a high risk despite a decent income and a reasonable purpose.
I need a loan for home renovation, including a kitchen remodel.	"high risk loan"	No, the credit score is too low and the DTI ratio is high, posing a risk for home renovation financing.
I need money to open a small bookstore and cafe.	"high risk loan"	No, the high DTI ratio and low credit score present significant risk for funding a new business venture.
I need a loan to install a security system at home.	"high risk loan"	No, the high DTI ratio and low credit score indicate significant financial risk, and the loan purpose does not justify the risk.
I need a loan to cover emergency expenses after a natural disaster.	"high risk loan"	No, the low credit score raises concerns despite the income and purpose for emergency expenses. Additional investigation is recommended.

# Demo of AI functions

SQL available at

[https://github.com/lestermartin/events/blob/main/2025-II-12\\_ATL-IcebergMeetup/functions.sql](https://github.com/lestermartin/events/blob/main/2025-II-12_ATL-IcebergMeetup/functions.sql)

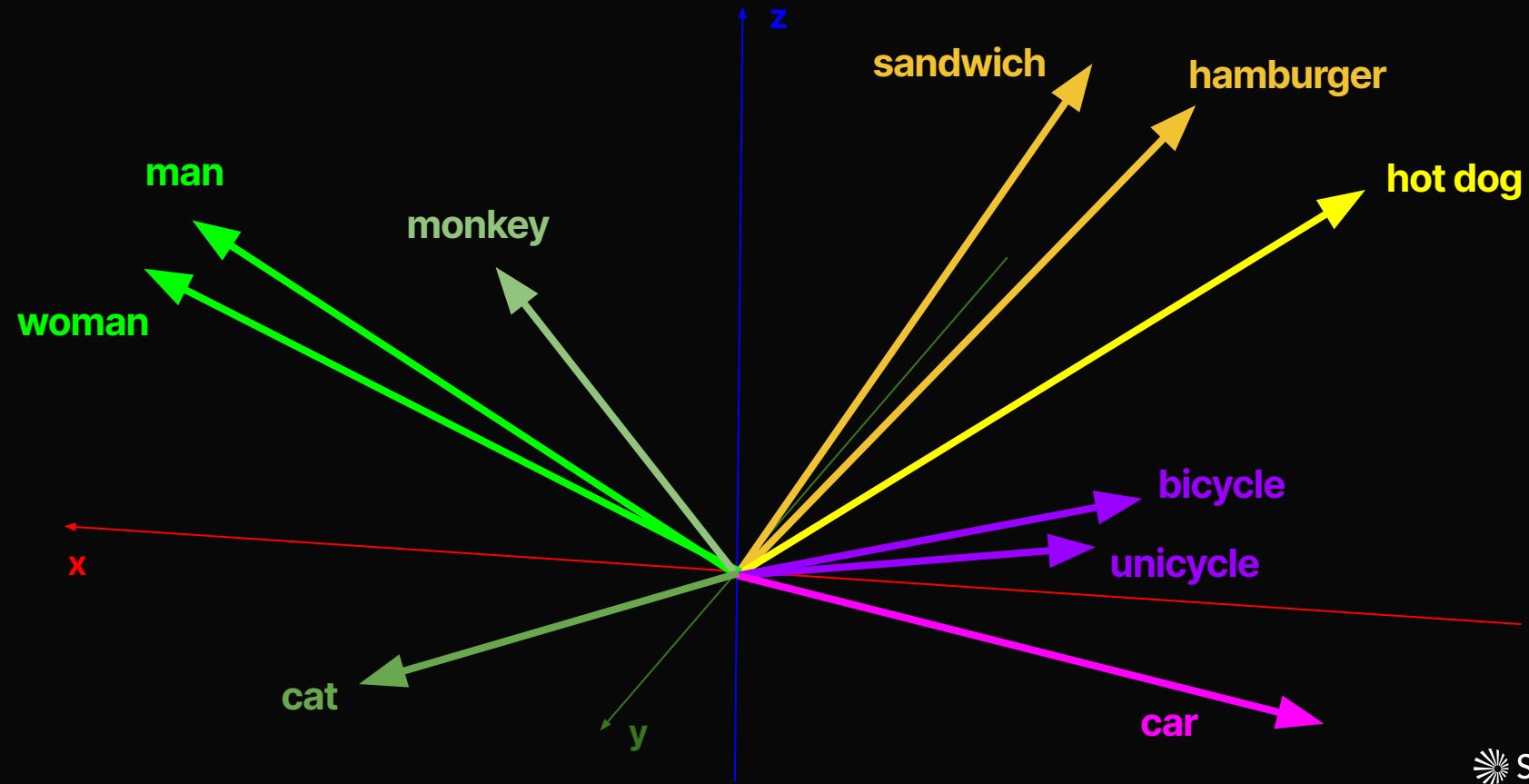


# **Augmented Generation (RAG & TAG)**

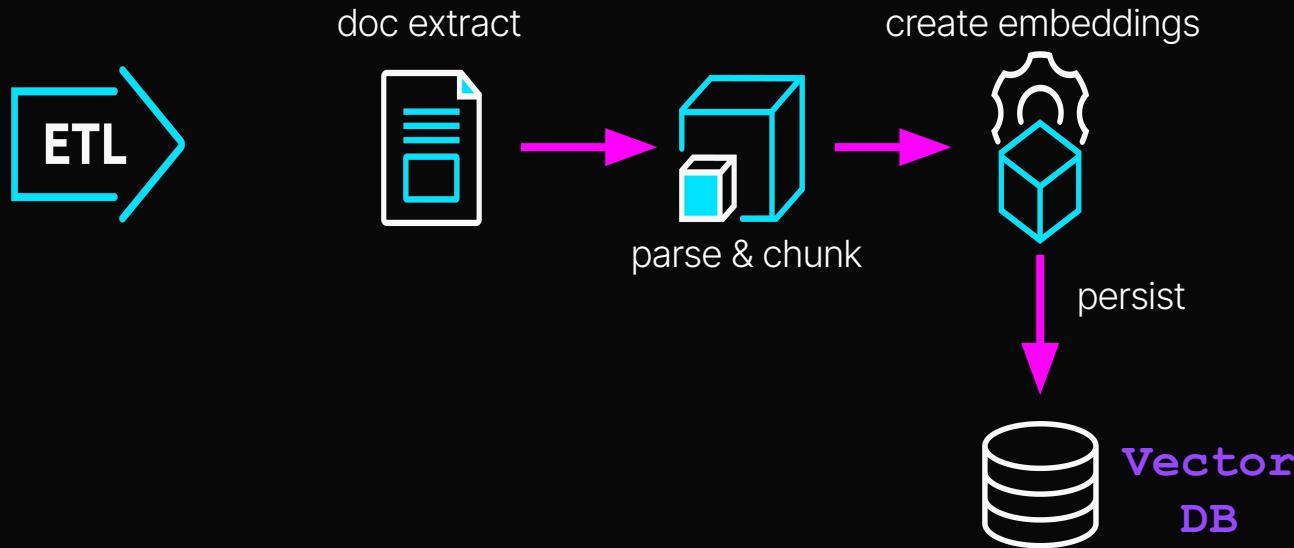
# What is Retrieval Augmented Generation (RAG)?



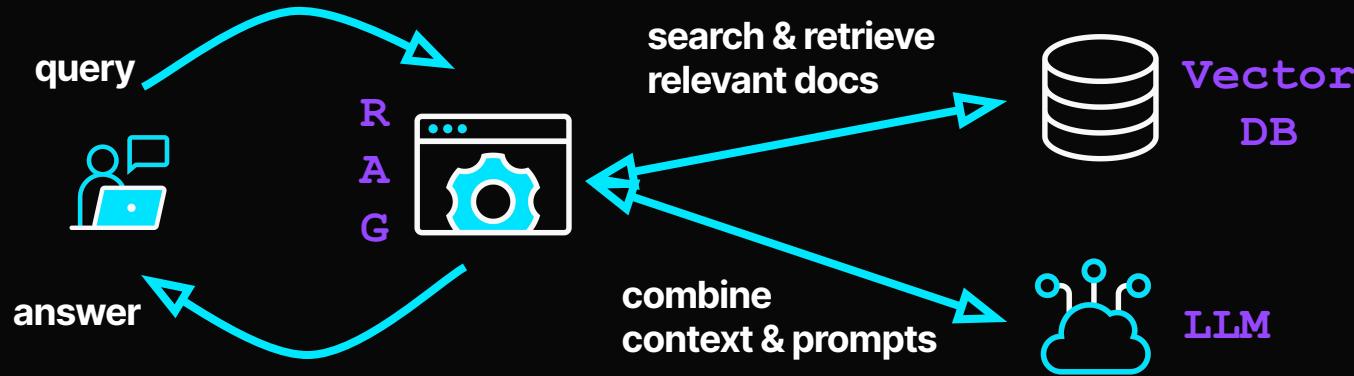
# Text chunks & vector embeddings



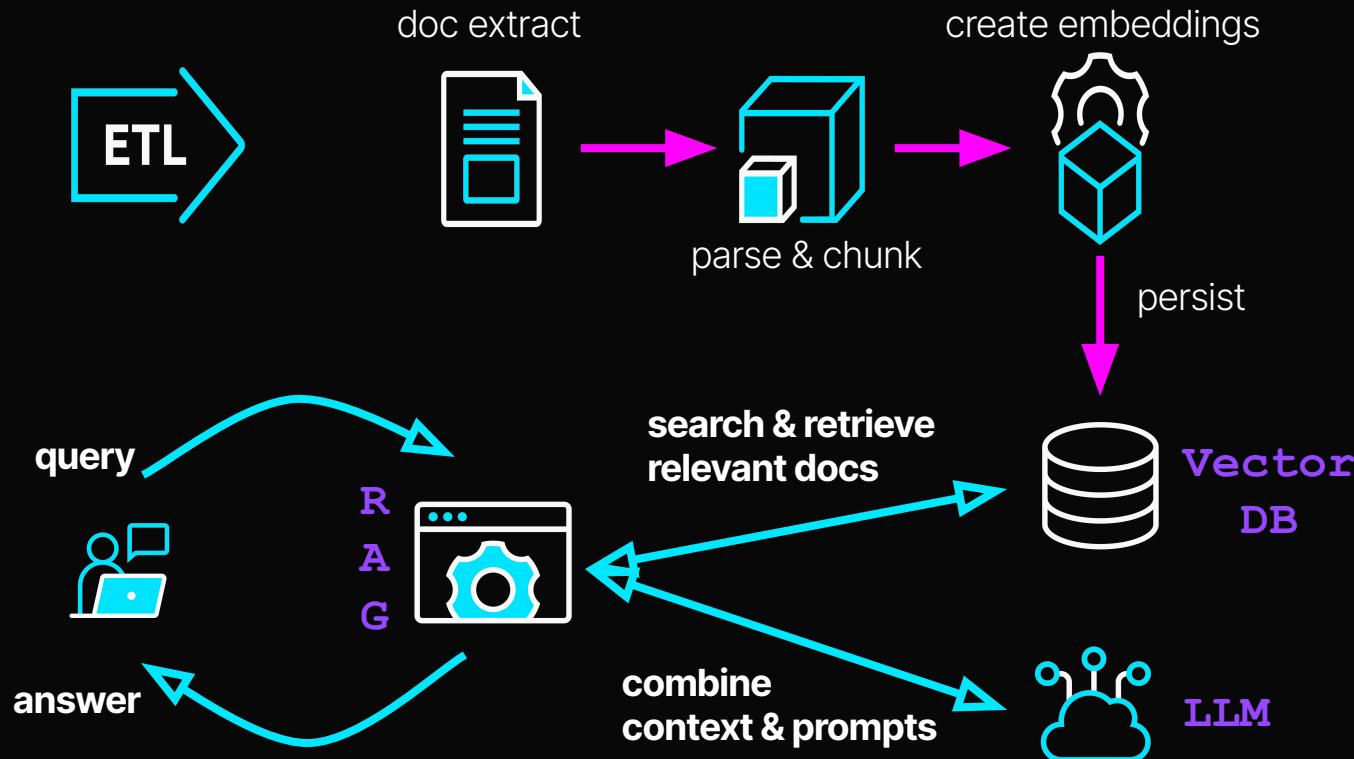
# Visualization of a RAG data pipeline (unstructured)



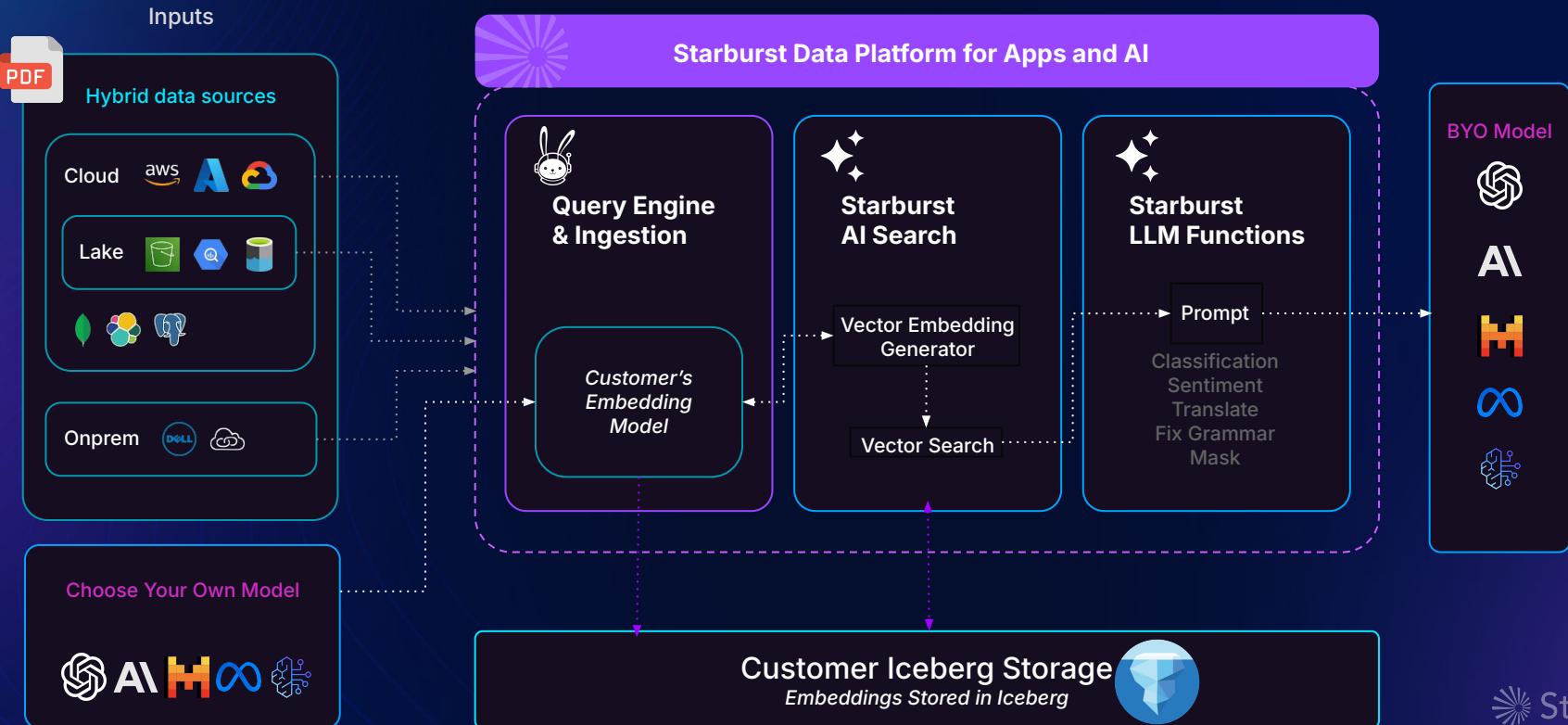
# Visualization of a RAG application



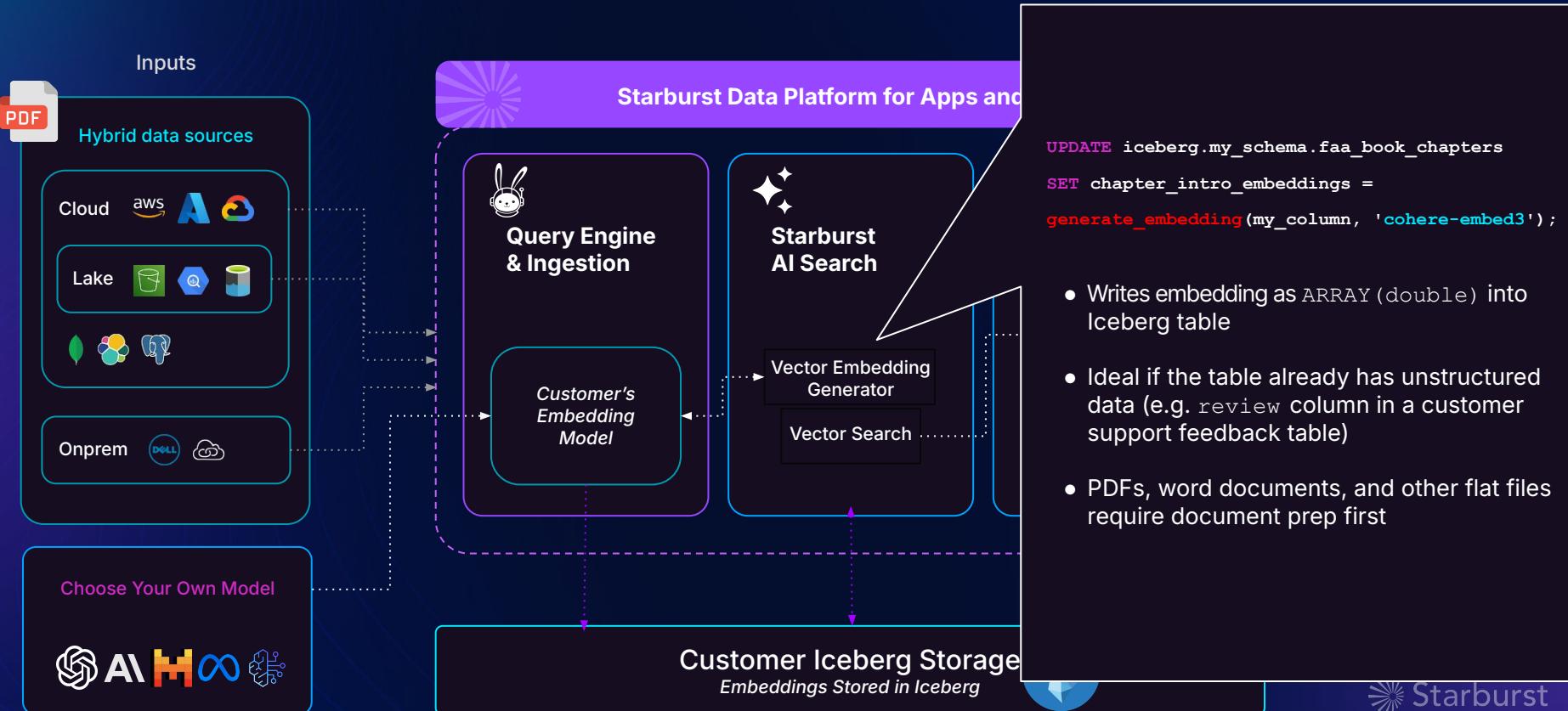
# Visualization of a RAG data pipeline & application



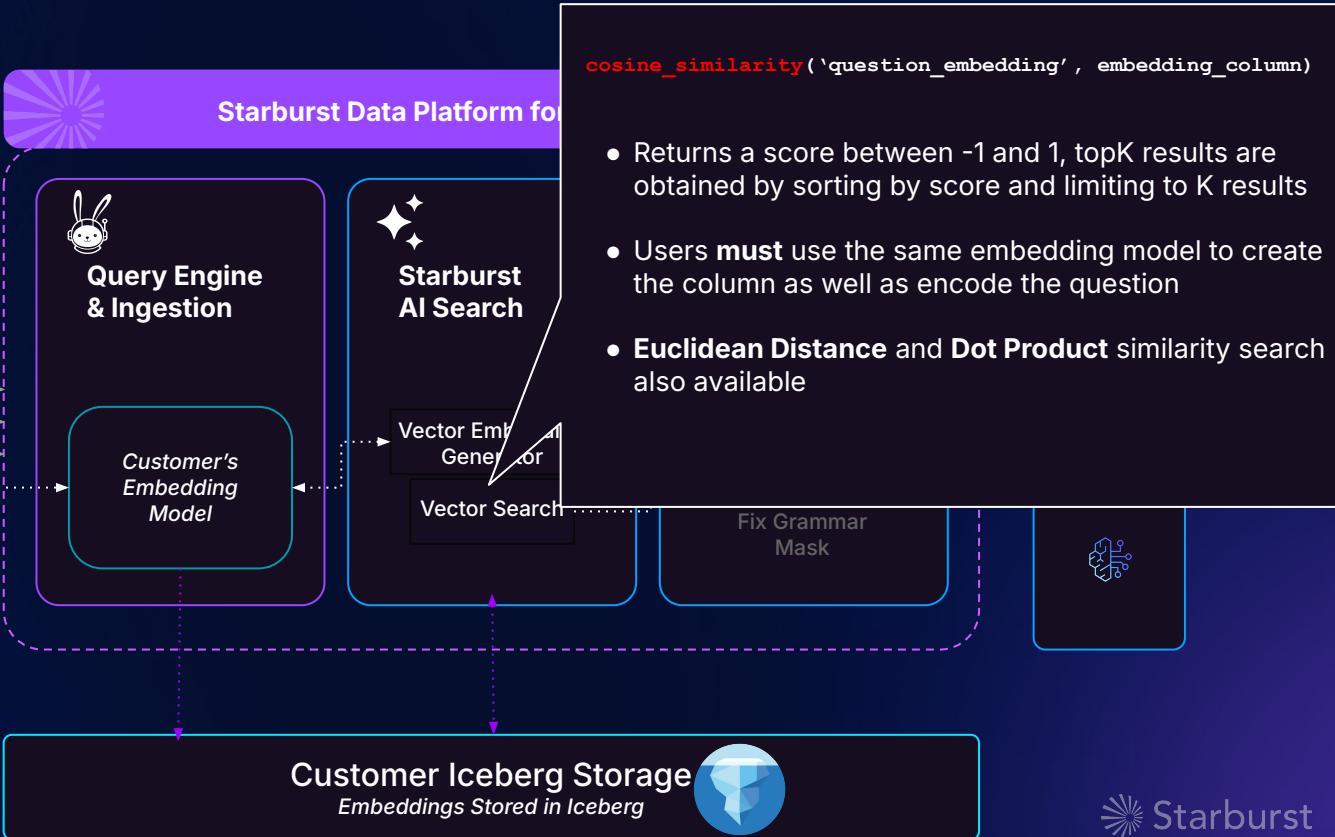
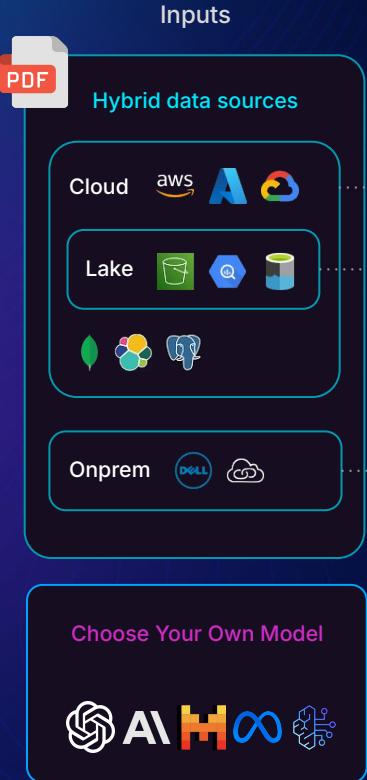
# Bring enterprise data to AI workflows & agents



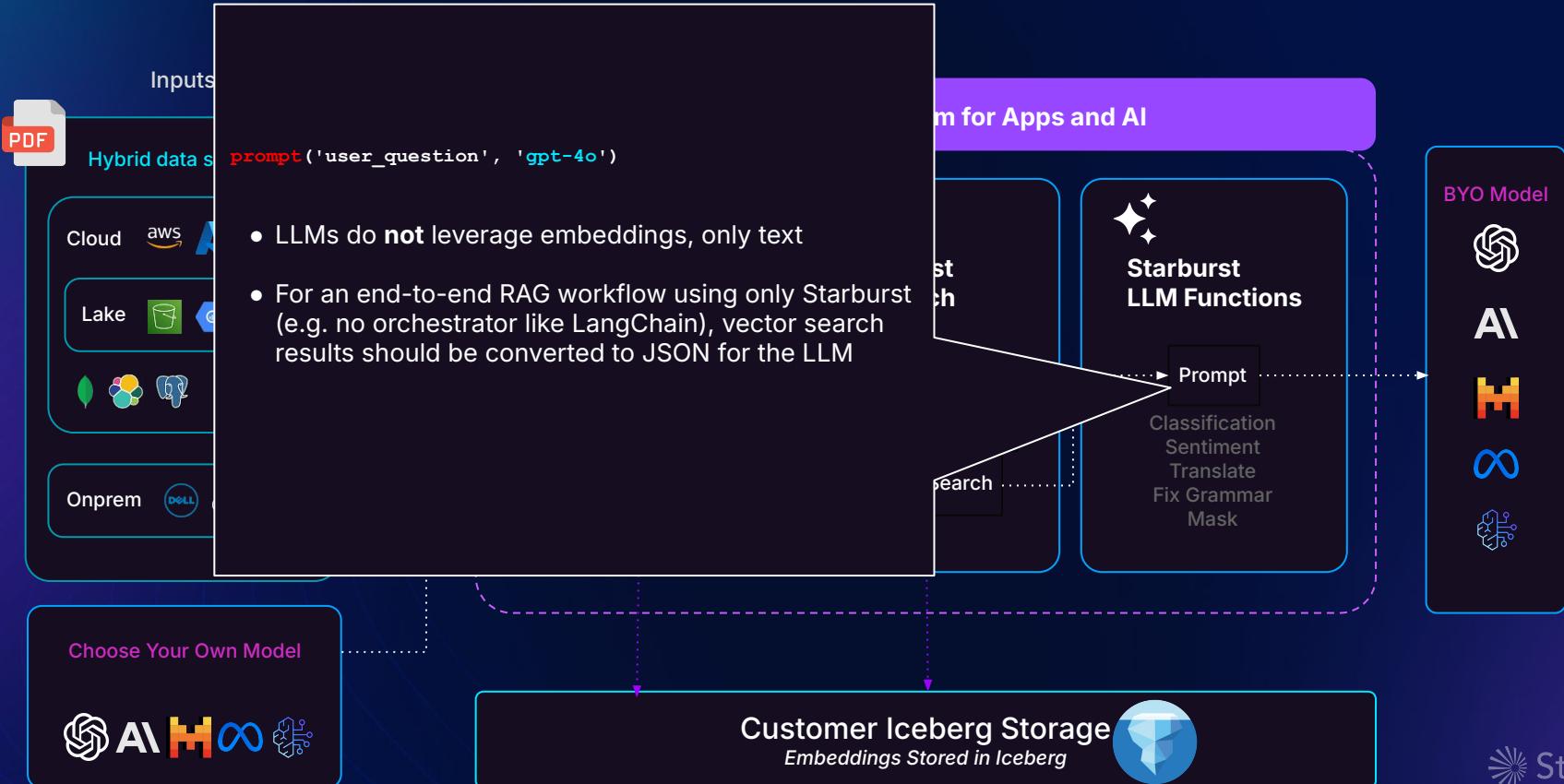
# Generate embeddings and store them in the lake



# Vector search using SQL on top of your data lake



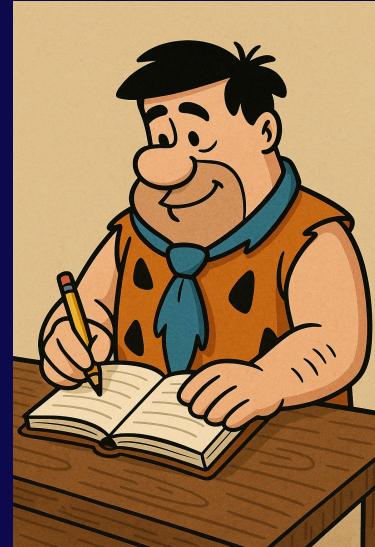
# Prompt an LLM with your search results



# Demo of RAG workflow

SQL available at

[https://github.com/lestermartin/events/blob/main/2025-II-I2\\_ATL-IcebergMeetup/workflow.sql](https://github.com/lestermartin/events/blob/main/2025-II-I2_ATL-IcebergMeetup/workflow.sql)



Thanks. | ດັນຍວາດ | 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](mailto:devrel@starburst.io)



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



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





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

