Starburst

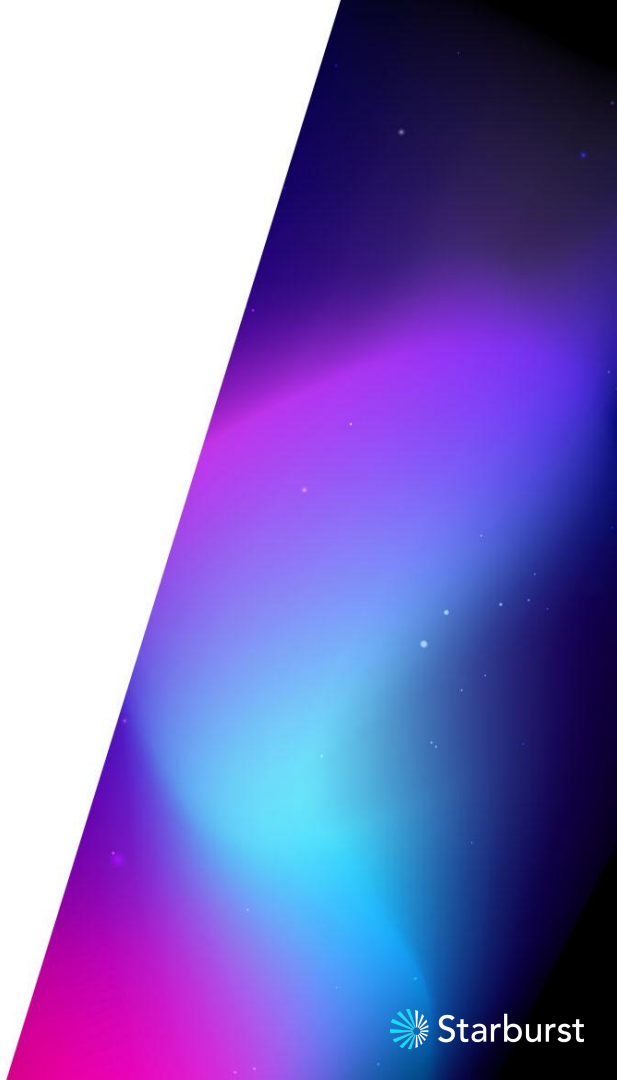
# Starburst 101

09.27.2023

# Agenda and objectives

**Part 1:** Introduction to Trino and Starburst

**Part 2:** Hands-on Starburst Galaxy lab



# Starburst 101

## Part 1: Introduction to Starburst and Trino

Starburst

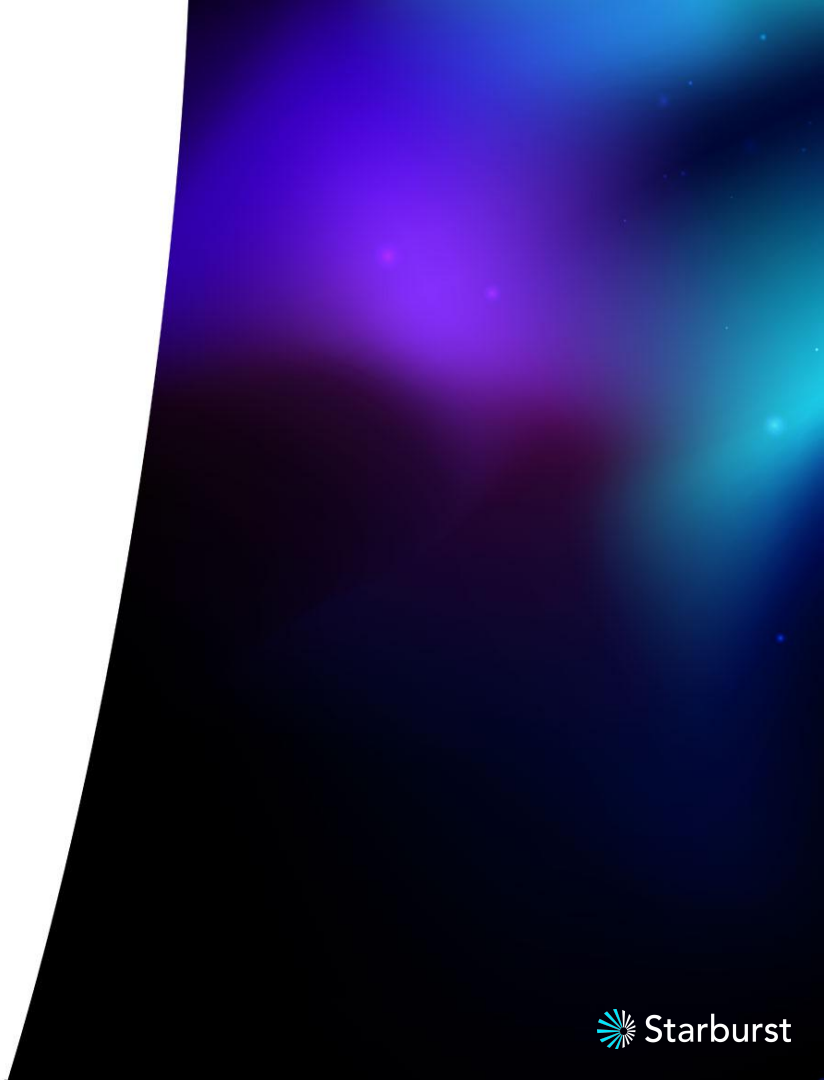
# Part 1 agenda

- The data accessibility problem
- The Trino query engine
- The Starburst data lake analytics platform

# Early challenges of big data

## Querying large volumes of data was difficult and time consuming

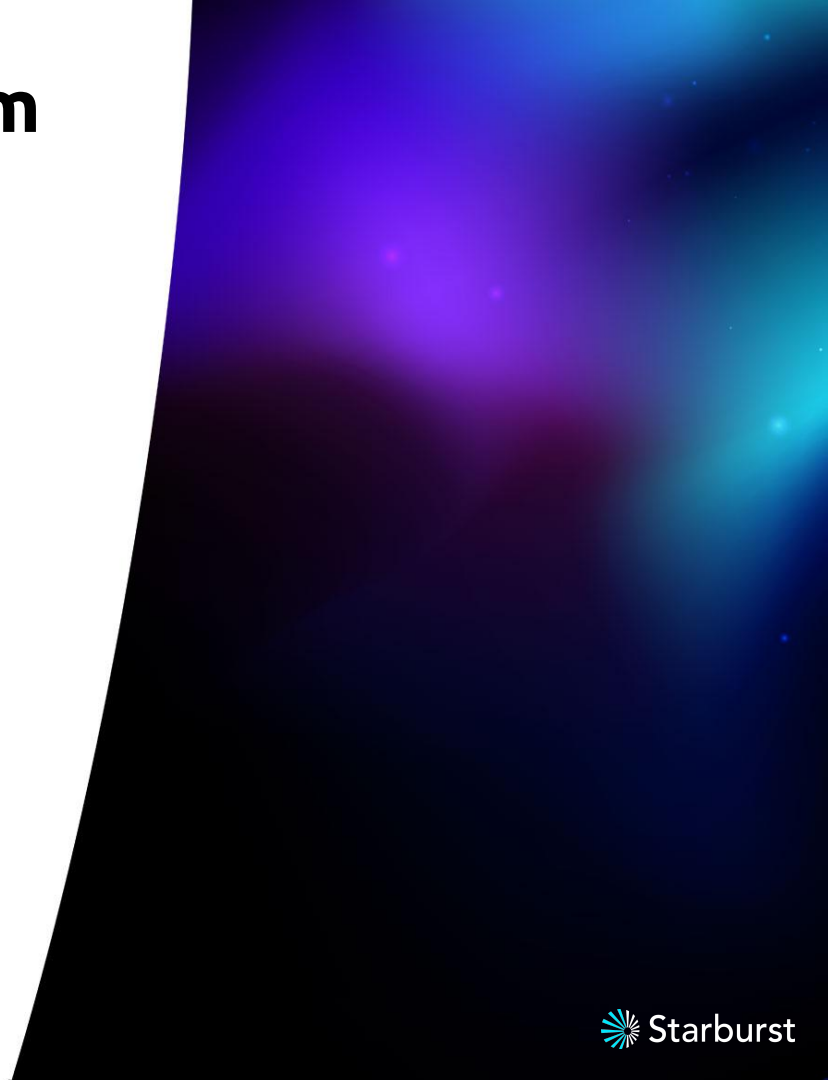
- Since the early 2000s, data generation and collection has skyrocketed due to the rise of the Internet.
- In 2005, Roger Magoulas referred to a large dataset that was almost impossible to manage and process using traditional BI tools as ***Big Data.***
- In 2006, Hadoop was designed to meet the needs of large datasets on a scale previously unimaginable.



# The data accessibility problem

## Data practitioners faced the same challenges at Facebook in 2010

- Facebook created Hive to query terabytes of data in Hadoop using SQL.
- Data scientists attempted to query massive object stores, but performance was too slow.
- Data consumers were limited by the number of queries they could run — often ***fewer than 10*** in one day.

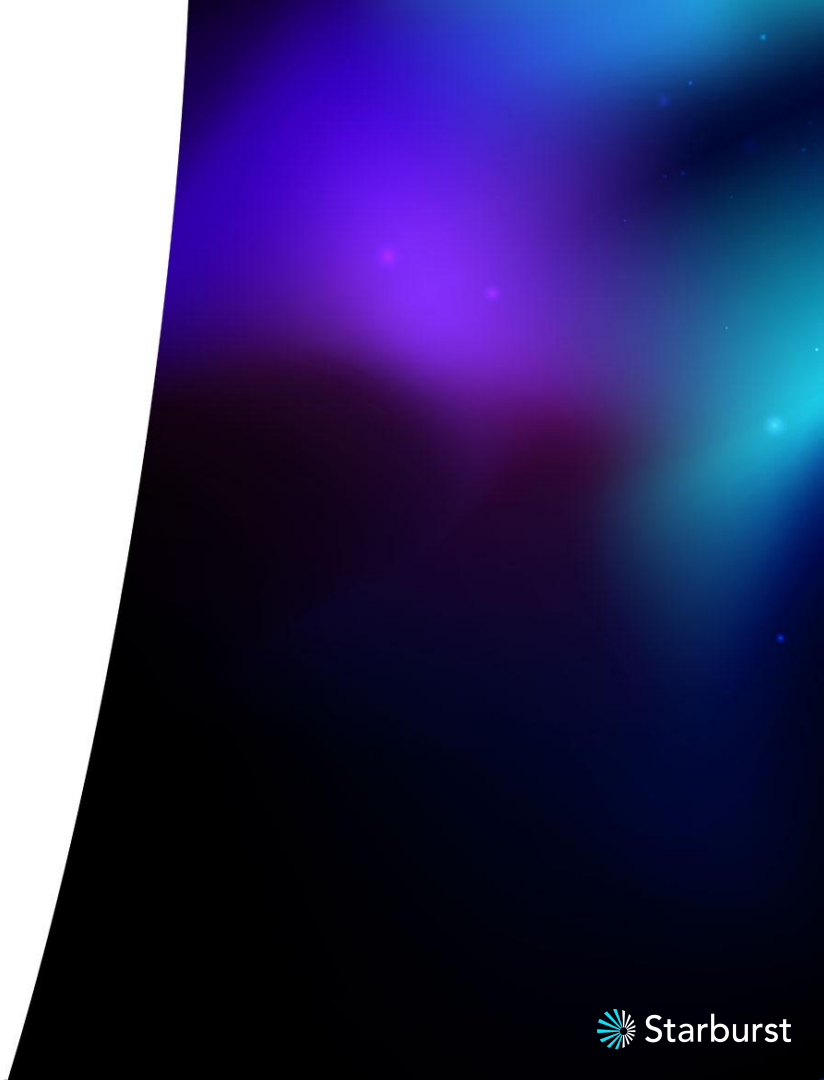


Starburst

# Enter Trino (Presto)

## A new query engine designed to solve the data accessibility problem

- **Trino** is a query engine that:
  - Harnesses the power of distributed computing
  - Separates compute from storage
- It allows fast querying on a data lake, and can federate data across data sources, helping to solve the data accessibility problem.



Starburst

# What is Trino?

- A ludicrously fast, open source, SQL query engine.
- Created and maintained by a community of contributors.
  - Licensed under the Apache license, version 2.0.

trino

# Structured Query Language (SQL)

- Declarative language - specify what, not how
- Using SQL enables you to leave the heavy lift of optimizing the code to Trino

```
SELECT nationkey, count(*) AS count
FROM tpch.tiny.customer
WHERE mktsegment='AUTOMOBILE'
GROUP BY nationkey;
```

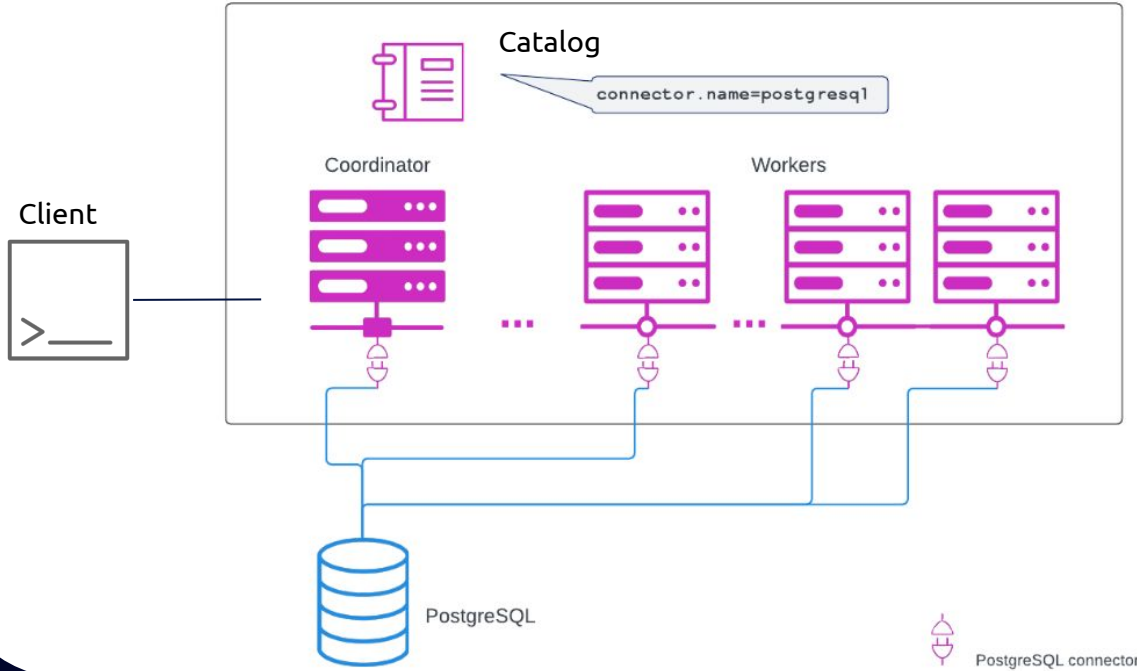
trino

# What are the benefits of a query engine?

- Trino can communicate with disparate data sources to federate data
- Trino is a distributed, massively parallel processing system

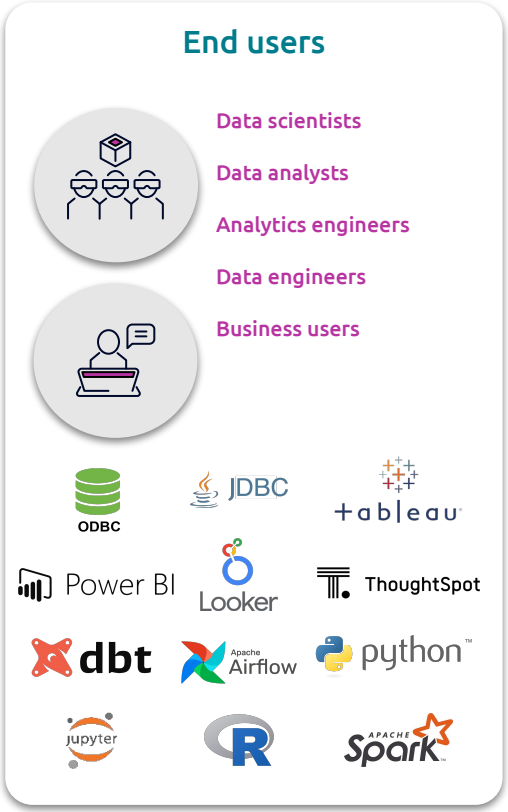
trino

# How does Trino work?

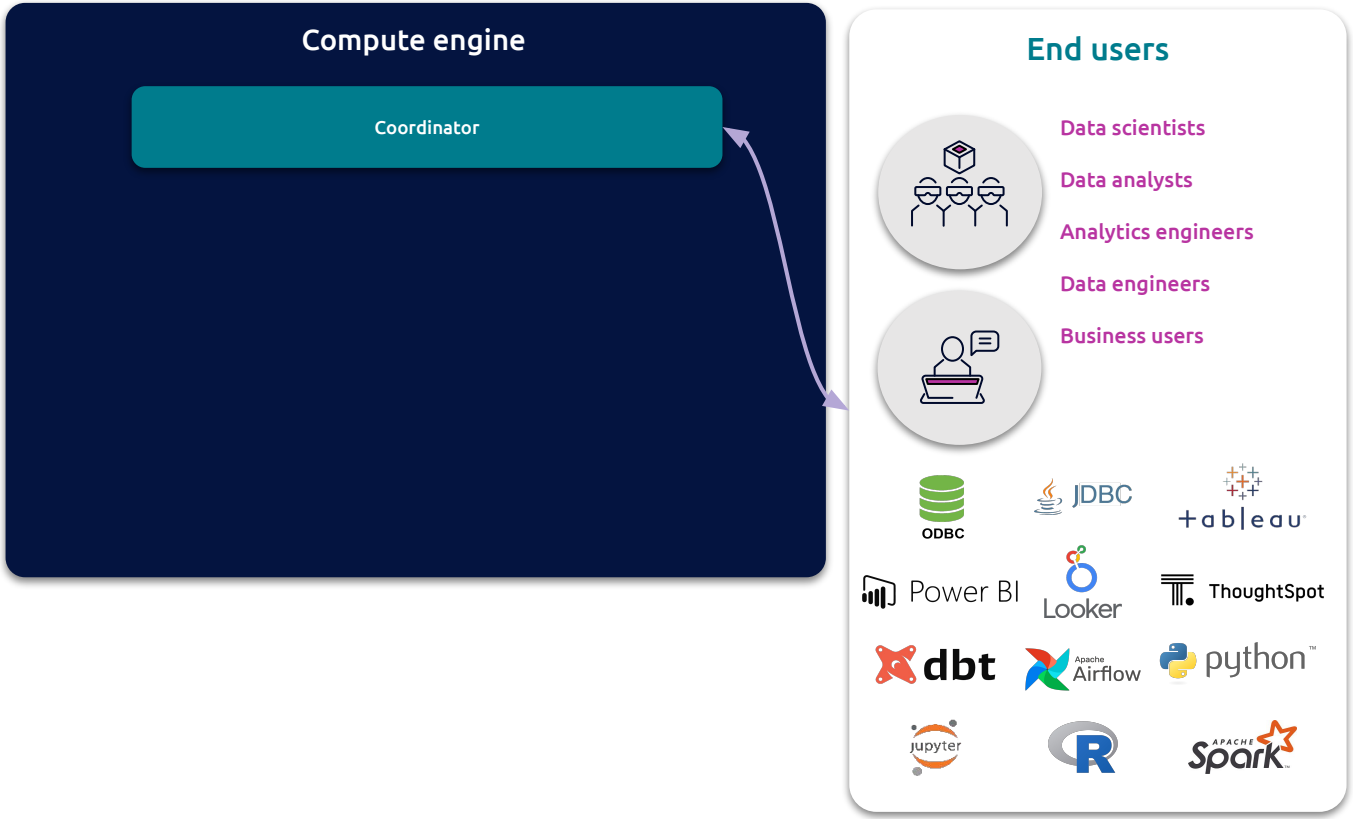



trino

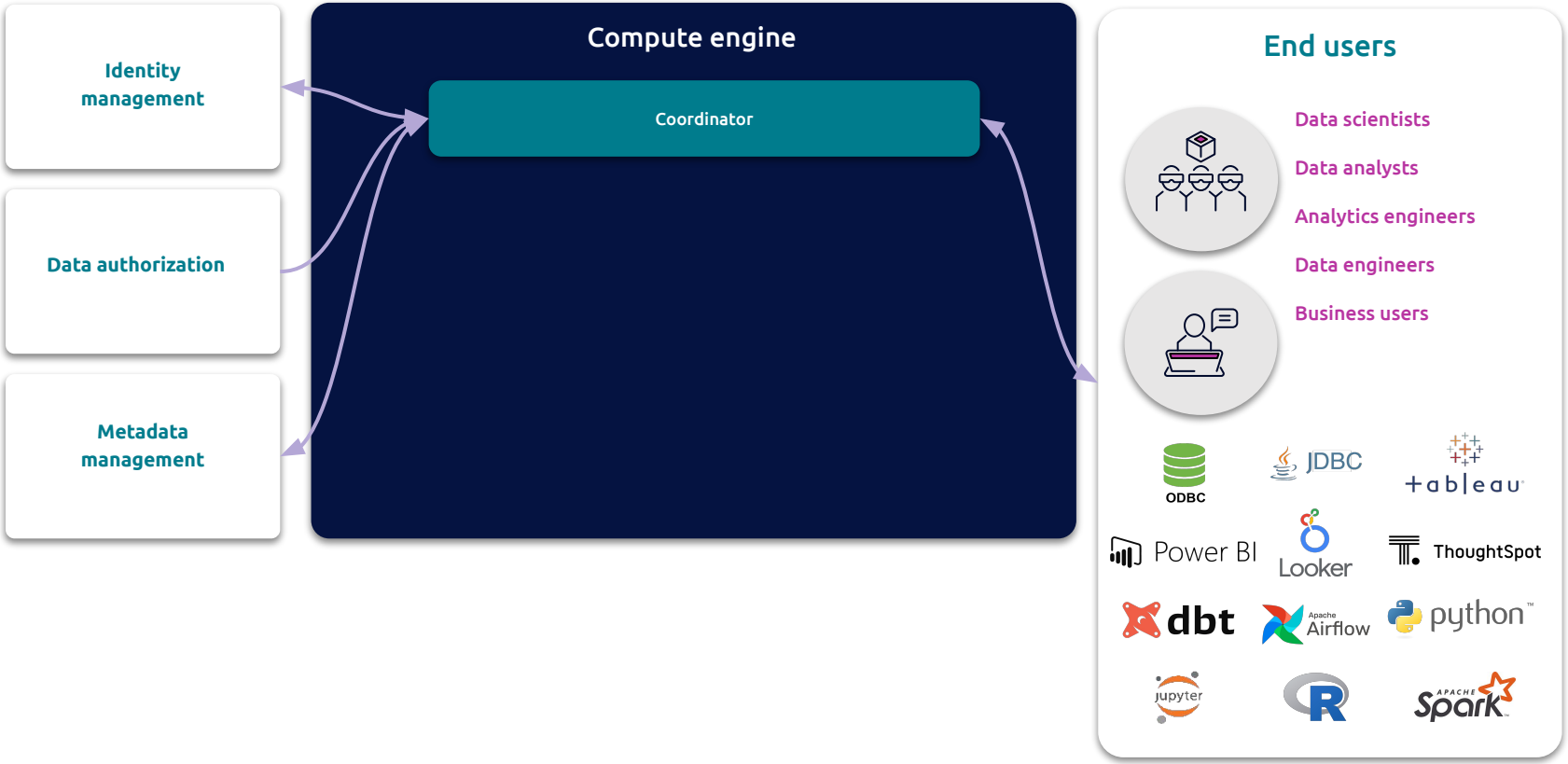
# Data consumer submits a query



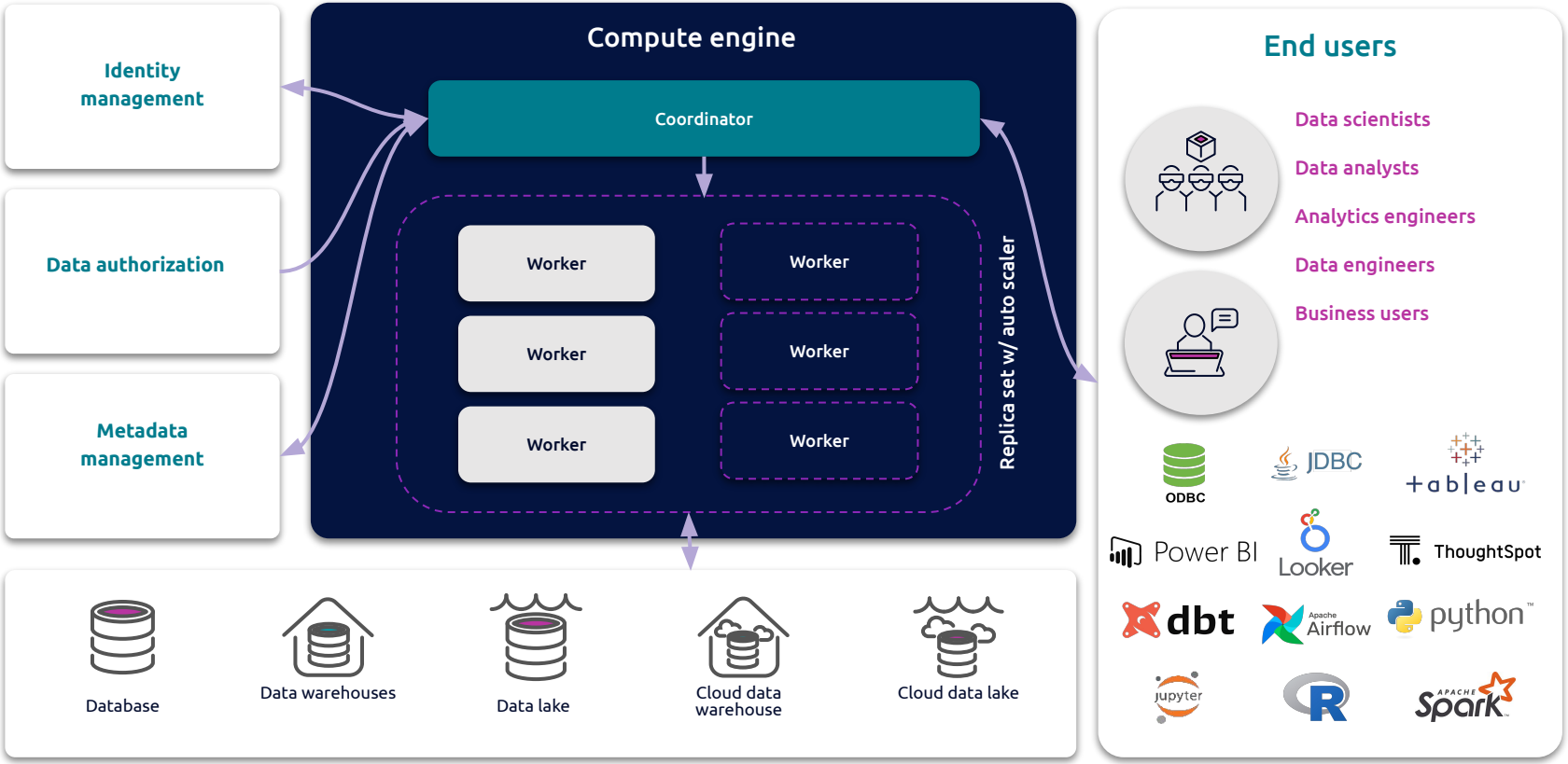
# Coordinator node receives query



# Parse and optimize query



# Worker nodes interact with data



# Trino is the query engine trusted by industry leaders at PB scale

**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

*But Trino requires extensive resources to run successfully…*

**Management:** All manual. No autoscaling

**Security:** No built-in security integrations

**Access Control:** Requires 3rd parties for RBAC

**Support:** No support team, reliant on community responsiveness

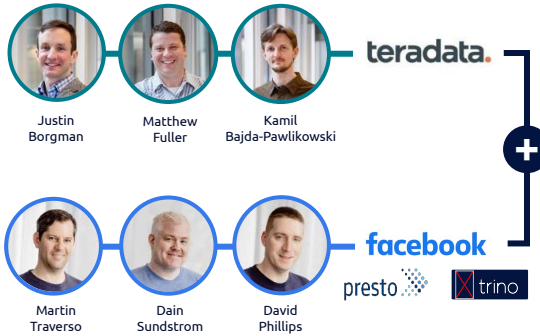
› **$$$**
*(and time!)*

Starburst

# Getting to know Starburst

A data lake analytics platform

# Our founding story




Starburst **$3.35B**
Total Valuation Mar '22

4 developers start Presto at Facebook

Facebook open sources Presto

Teradata invests heavily in community

Starburst founded by Facebook and Teradata team

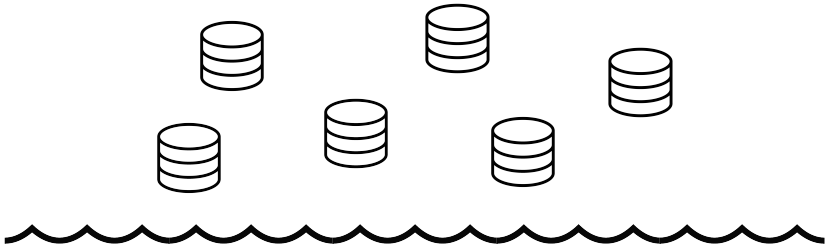
Trino (formerly Presto) reaches 25,000 commits

**August 2012**

**March 2022**

# The Modern Data Lake

**Global federated access to data sources beyond the lake**



**MPP query engine**

trino

**Open table formats**

ICEBERG

DELTA LAKE

Apache hudi

**Open file formats**

AVRO

Apache orc

Parquet

**Commodity storage & compute**

Object storage

Elastic compute

Starburst

Data Lake Analytics Platform

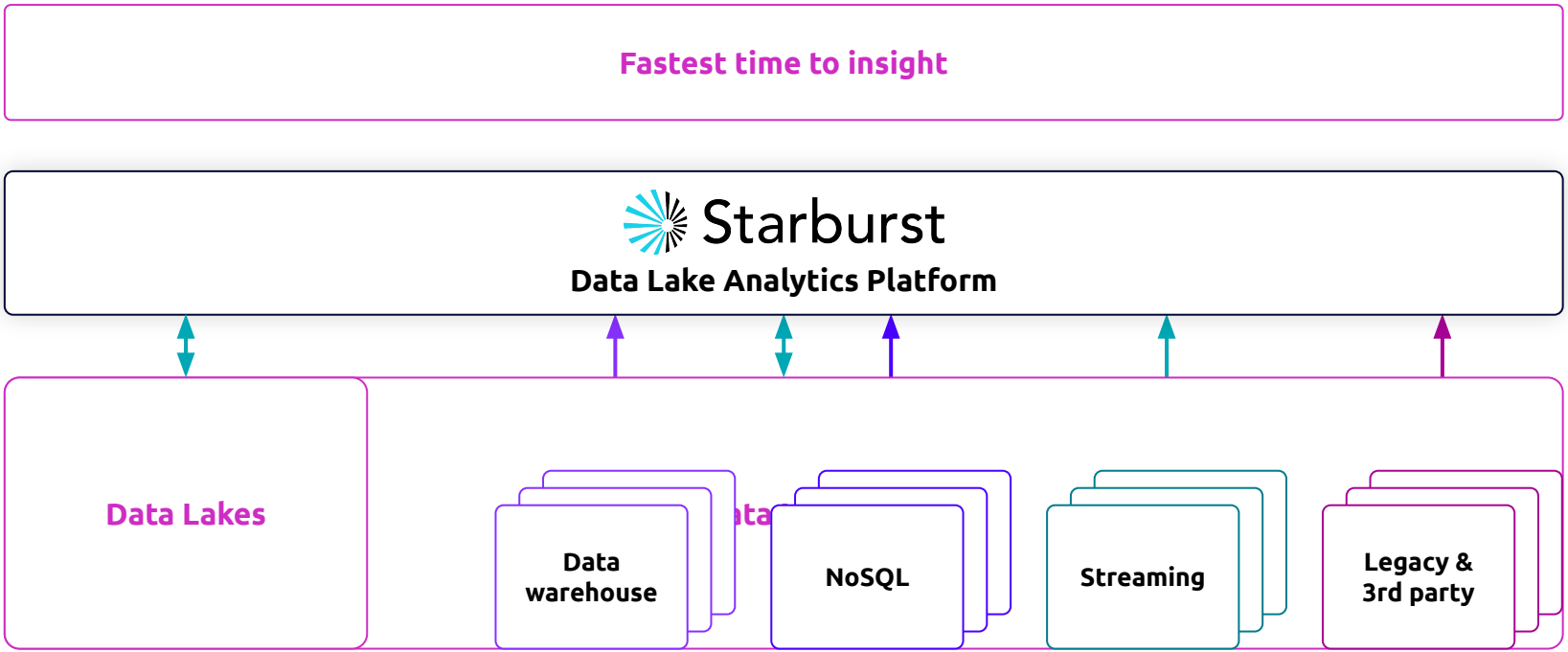
The easiest way to *build and manage* your Modern Data Lake

**90%**
Faster time to insight

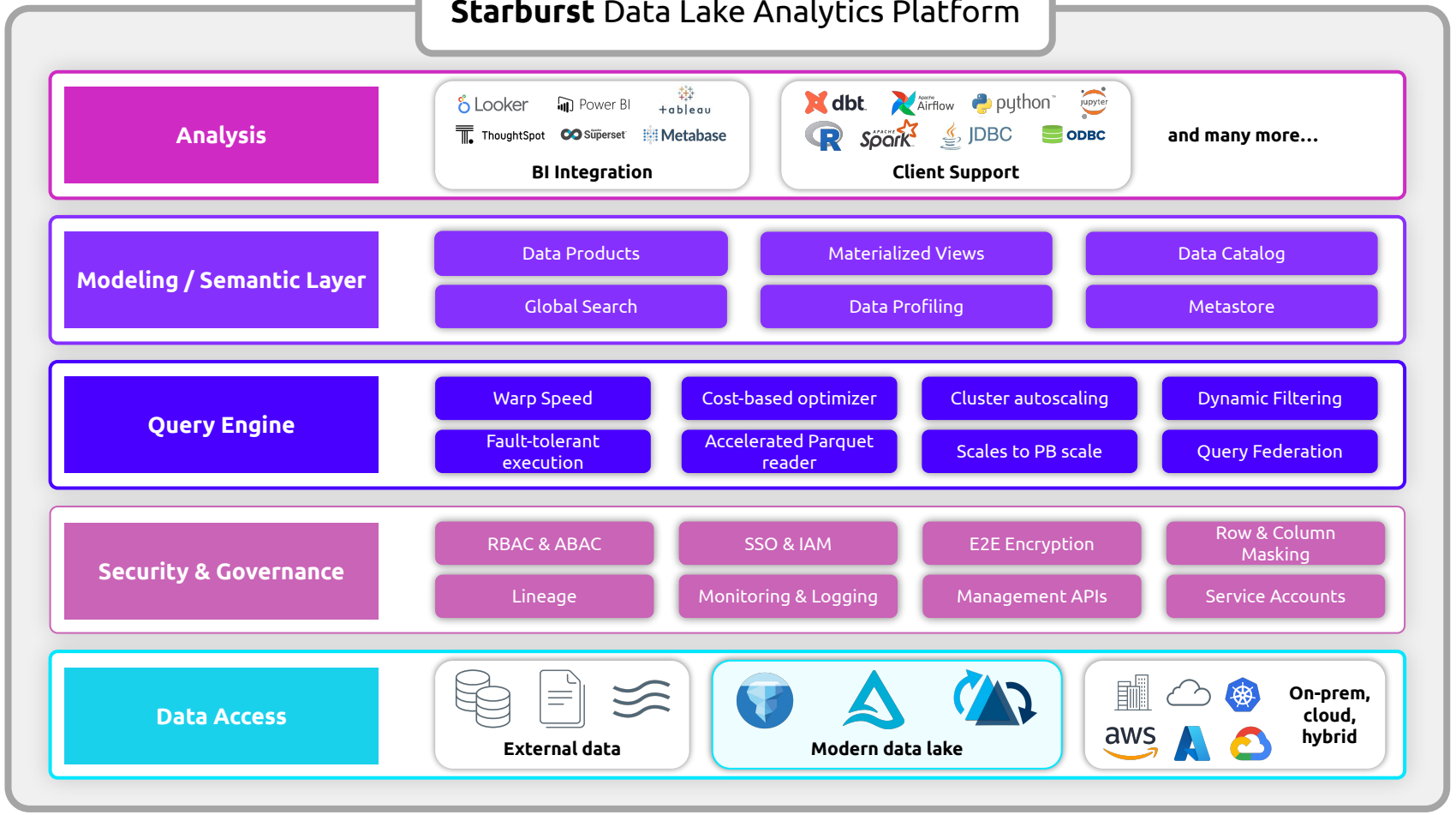
**53%**
Lower TCO

**100%**
Future-proof architecture

# Starburst activates the data in and around your lake



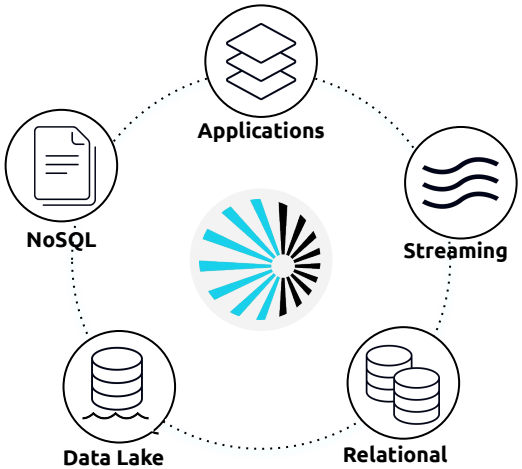
# **Starburst** Data Lake Analytics Platform




Starburst

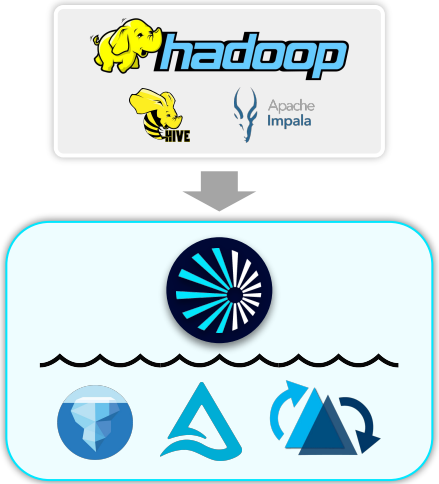
# How Starburst helps

## Federate and unify disparate data



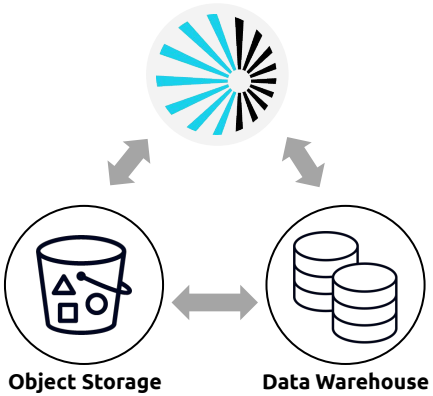

- 50+ enterprise-grade connectors
- Stargate for hybrid and cross-cloud

## Modernize legacy data lakes




- Optimized for Iceberg, Delta Lake, Hudi
- Accelerated Parquet Reader
- Transform, structure, and share directly in the data lake

## Complement the data warehouse

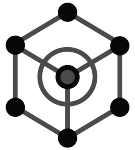



- Starburst Warp Speed (smart indexing & caching)
- Data Products
- Schema discovery

# Benefits of our approach

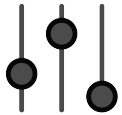
## Simplicity

Take the fastest path from data to decision



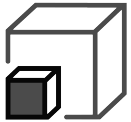
## Access

Future-proof your architecture with a single point of access



## Optionality

Never get locked into a proprietary data ecosystem



## Scalability

Operate efficiently and reliably at internet scale

Starburst

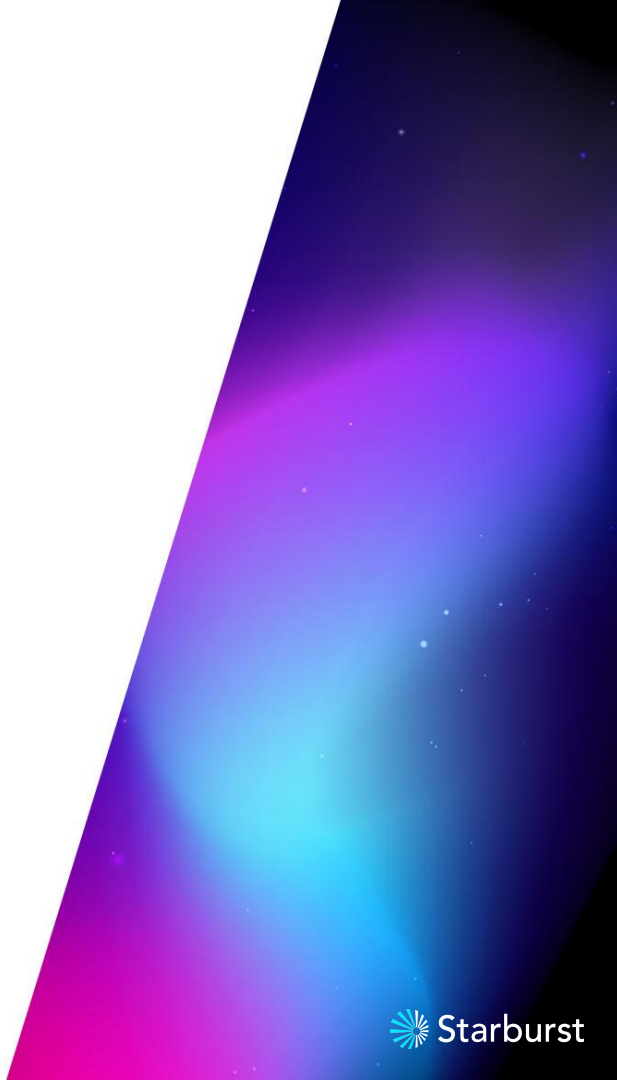
# Starburst 101

## Part 2: Hands-on Starburst Galaxy lab

Starburst

# Lab exercise

Data Engineers and Data Administrators



Starburst

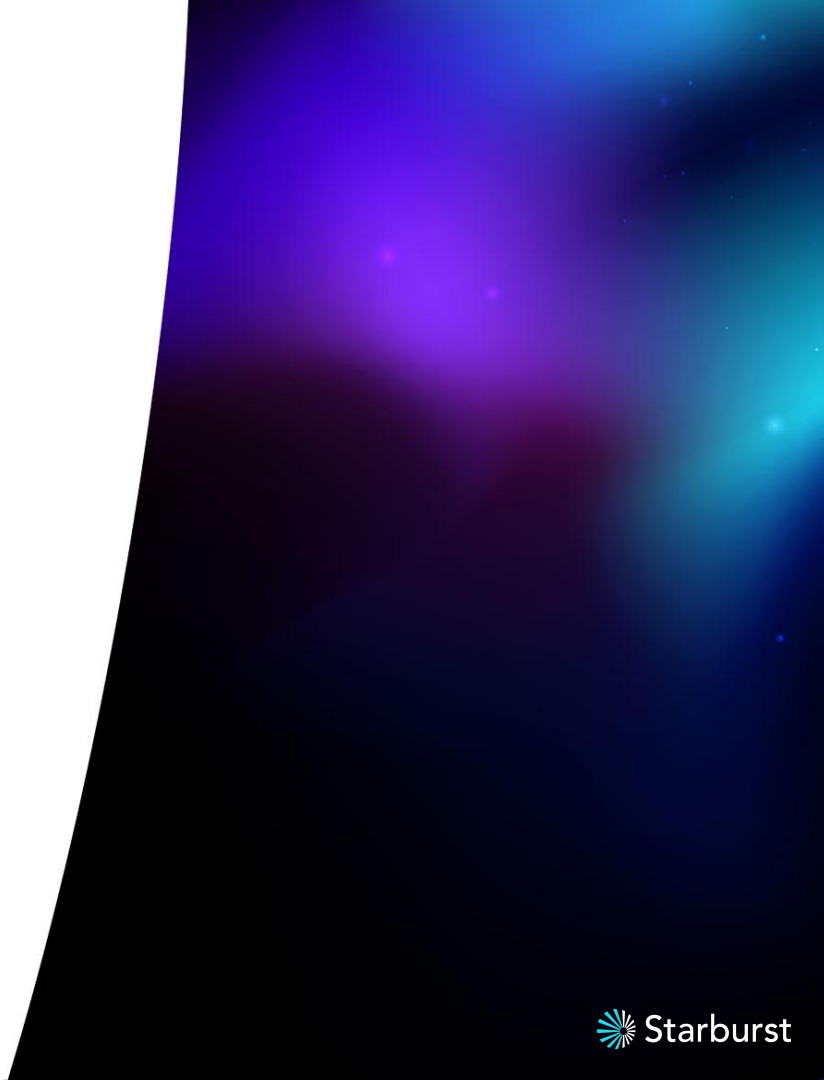
# Part 3 agenda

- Introduce project
- Connect to data sources
- Build within your data lake
- Create data products

# Introduce project

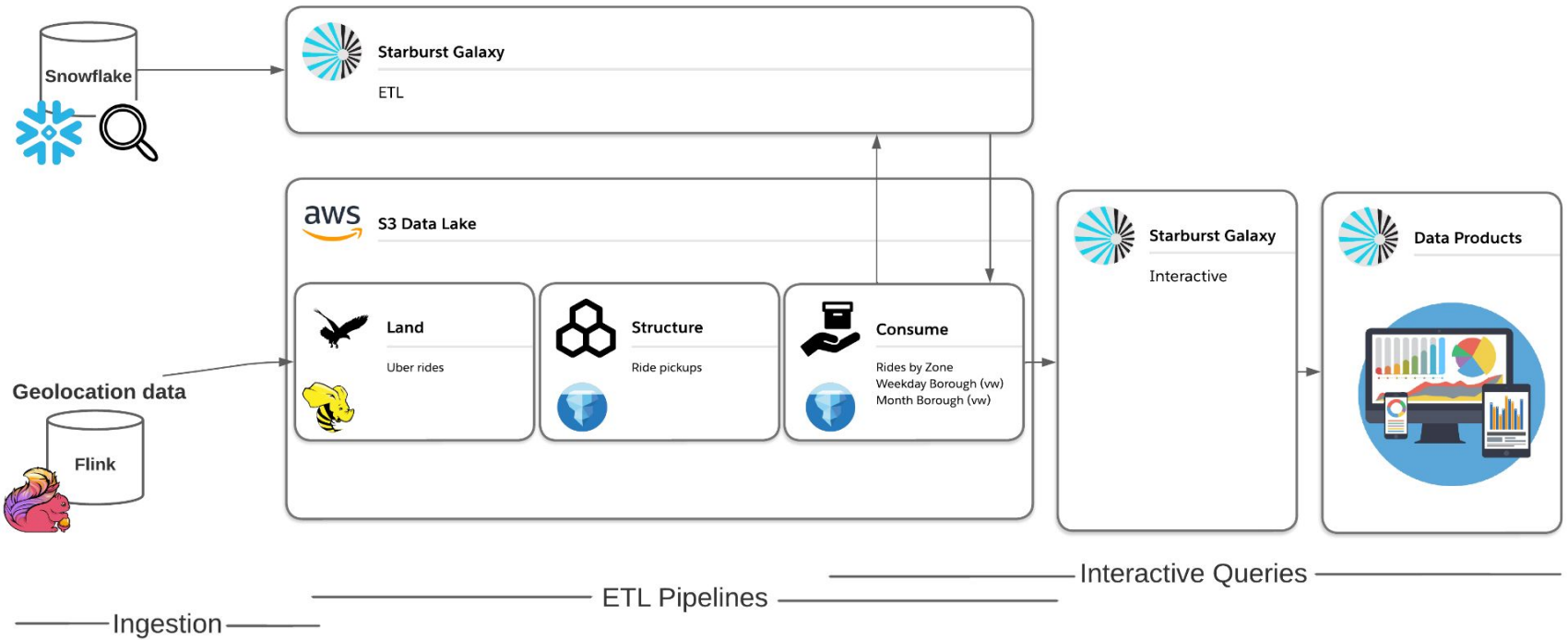
**Objectives:**

- Create a final table output and share an example query with your data science team.
- Create a data product answering specific business questions from the marketing department.
    - What is the most popular weekday for each borough?
    - What is the most popular month for each borough?



Starburst

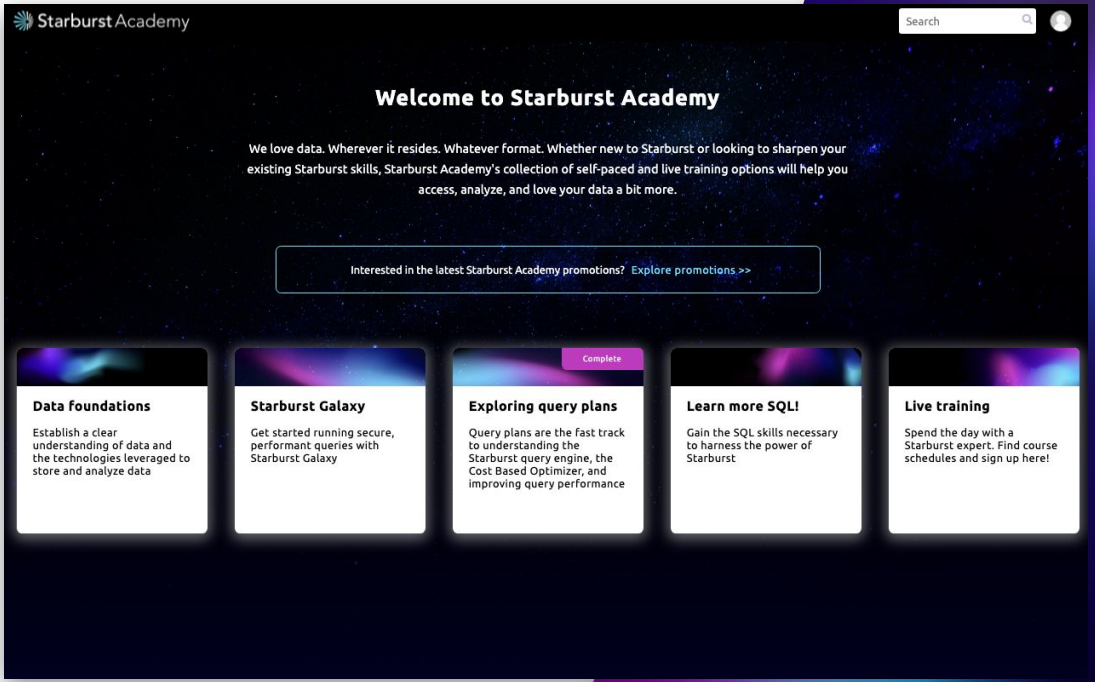
# Project architecture



# Starburst Academy

## Developing skills that lead to faster adoption and successful outcomes

- Introductory to advanced content
- Self-paced and live options
- Hands-on exercises
- **Free** and fee-based content




**academy.starburst.io**

Starburst

Starburst

# Thank you!

**Presenter Name**

Date

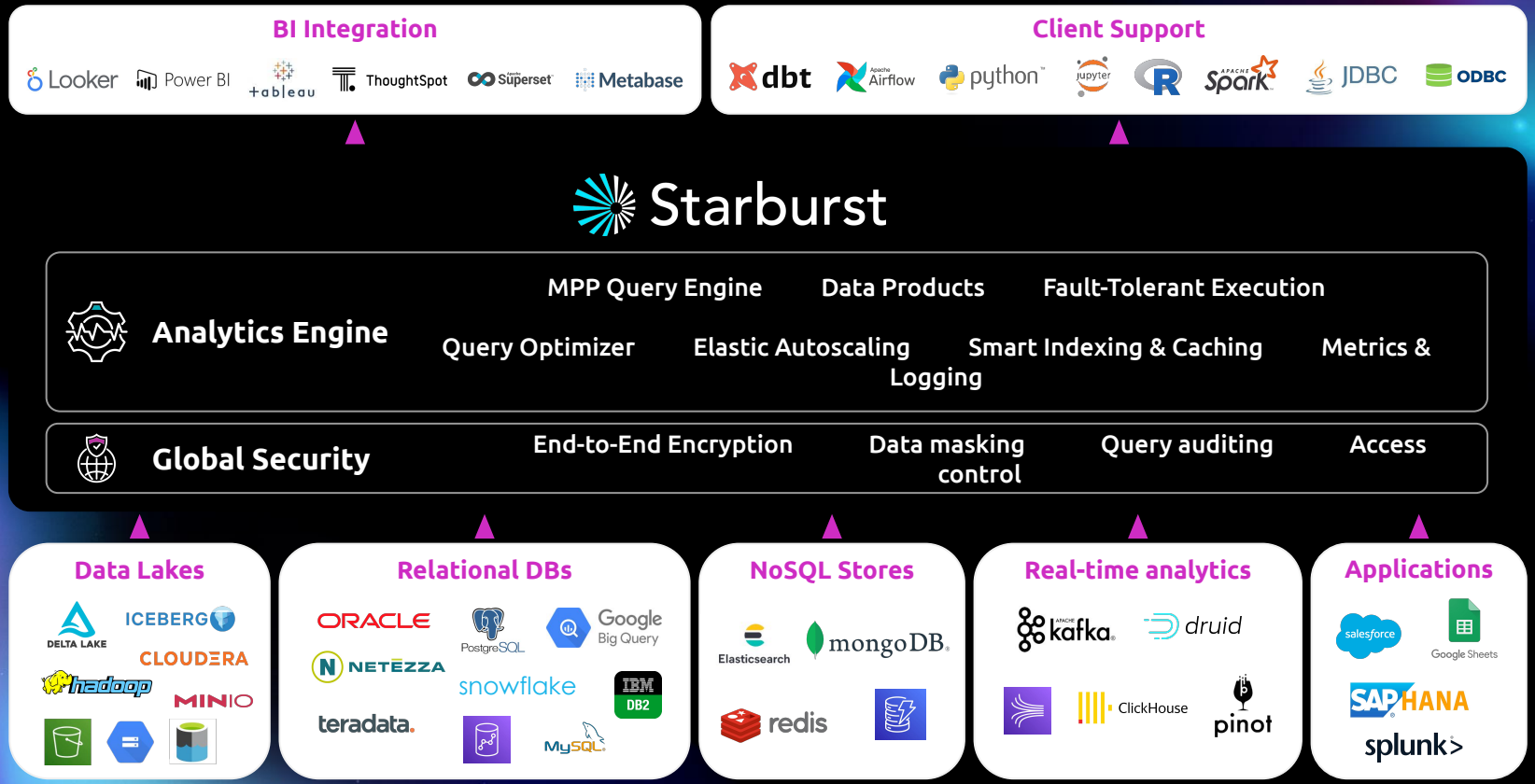

Starburst

# Starburst supports your evolving data strategy

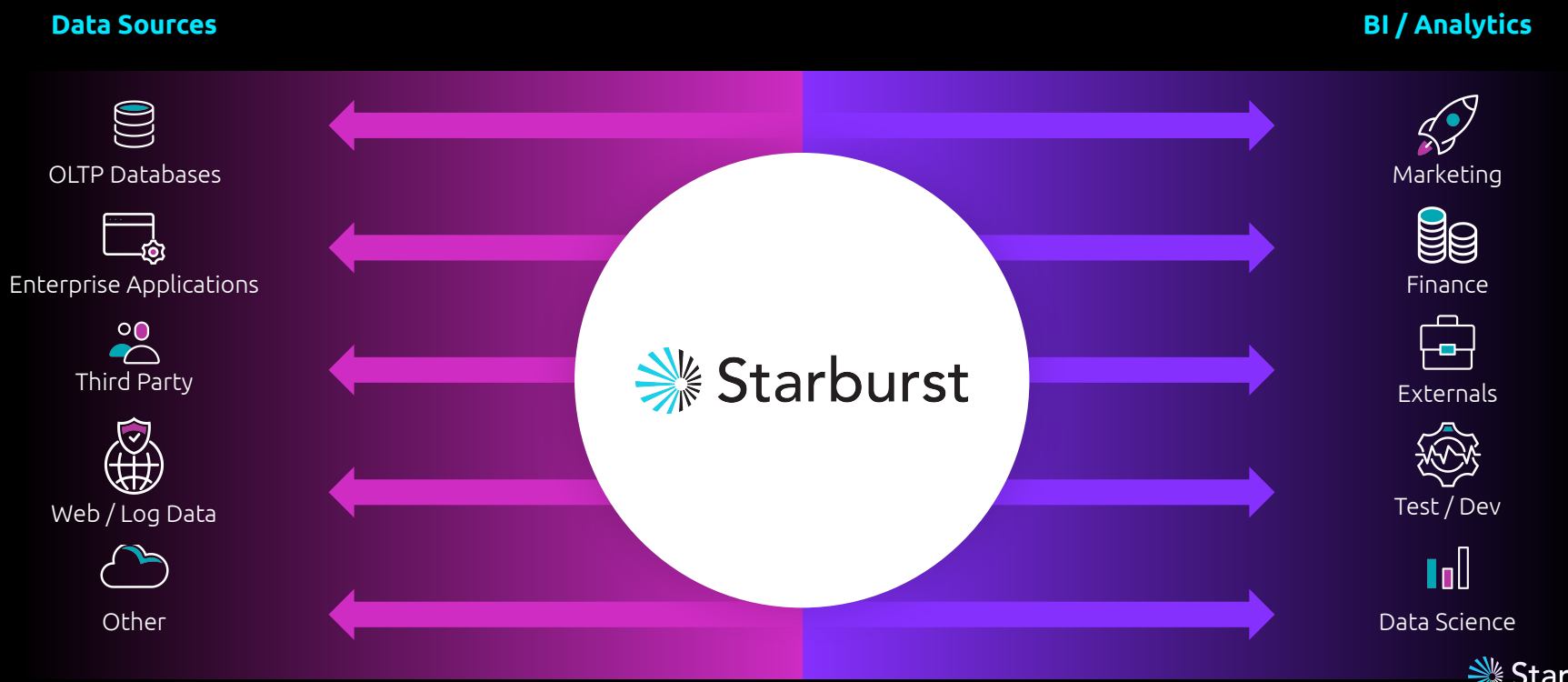

**Starburst technology applications**

Starburst

# Starburst is the analytics engine that fits into any environment



# Starburst connects to all your data and gives you the power of optionality



# Starburst Galaxy

**Cloud-native, frictionless, and fully managed.**
**The fastest path from big data to better decisions**

## Platform

Fully managed

Flexible billing

Advanced query editor

Autoscaling and auto-suspend

Audit logs and query history

## Security

End-to-end encryption

RBAC - cluster down to table level

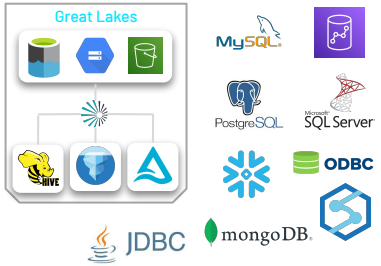
SOC 2 Type 1 and 2

SSO integration

## Connectivity

Direct connect, bastion, and cross account role

Vast BI tool support




## Support

In-app 24/7 support

Trino experts

Professional Services

Intuitive experience removes self-managed tuning and troubleshooting

trino

**Fast, distributed, ANSI SQL processing engine**

# Trino is a ludicrously fast, open source, SQL query engine…



trino

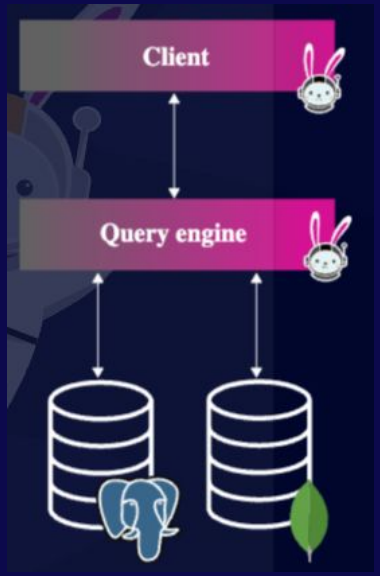
Trino is a ludicrously fast, open source, SQL query engine… designed to query disparate data sources.



trino

# Disparate data sources

- Trino can connect to disparate data sources
- These include databases, files in object storage, and any data that can be represented in a tabular format.
- For example: MySQL, Cassandra, S3 object stores

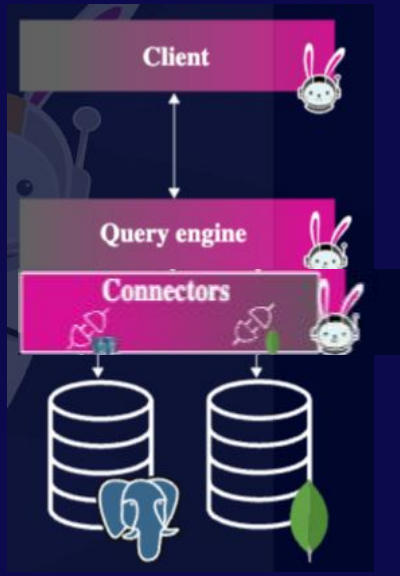





trino

# Disparate data sources

- Trino can connect to disparate data sources
- These include databases, files in object storage, and any data that can be represented in a tabular format.
- Trino's connector-based architecture allows this
  - Connectors are an interface between Trino SQL operations and the domain-specific language of various heterogeneous data sources






trino

Trino is a ludicrously fast, open source, distributed, massively parallel processing, SQL query engine designed to query large data sets from one or more disparate data sources.



trino


trino

# Starburst powers innovation across every industry…

### Financial Services

- Fraud detection
- Anti-money laundering
- Risk management

### Consumer

- Customer 360
- Marketing analytics
- Supply chain analytics

### Healthcare and Life Sciences

- Patient care optimization
- Regulatory compliance
- Health record analytics

### Technology

- Product analytics
- In-product functionality
- Security / Log analytics

### Telco

- Marketing operations
- Customer care
- Capacity planning

**Trusted by industry leaders**

BANK OF AMERICA · BARCLAYS · NIKE · CVS · Pfizer · Adobe · T-Mobile · COMCAST

Santander · citibank · Nestlé · MERCK · slack · vmware · orange

STATE STREET · NORDSTROM · COLGATE-PALMOLIVE · emis health · Roche · Snap Inc. · CISCO · verizon