Starburst

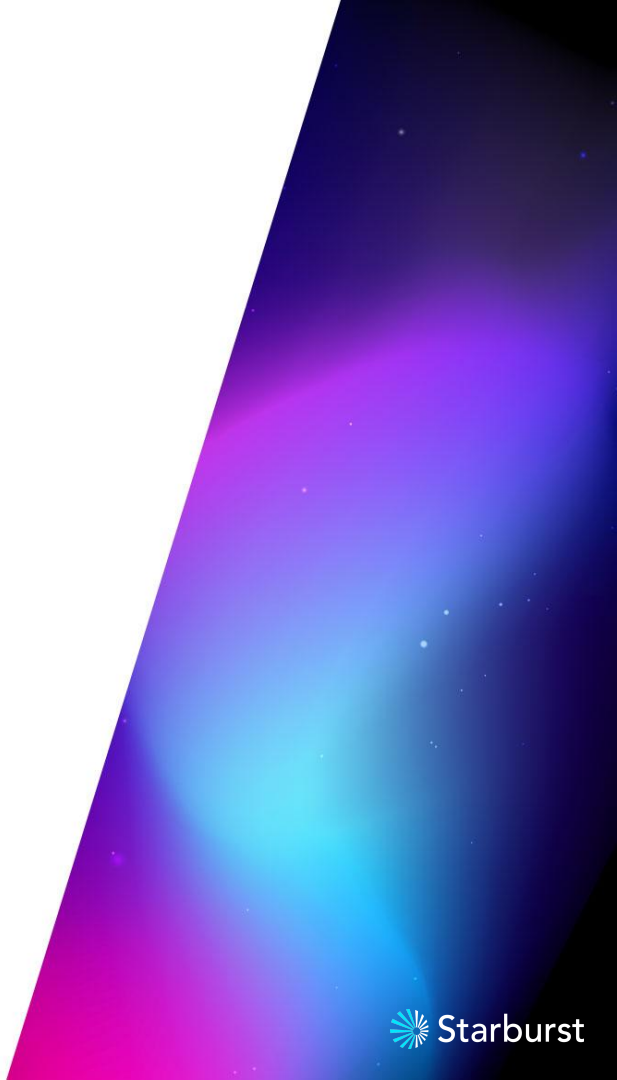
Starburst 101

05.29.2024

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

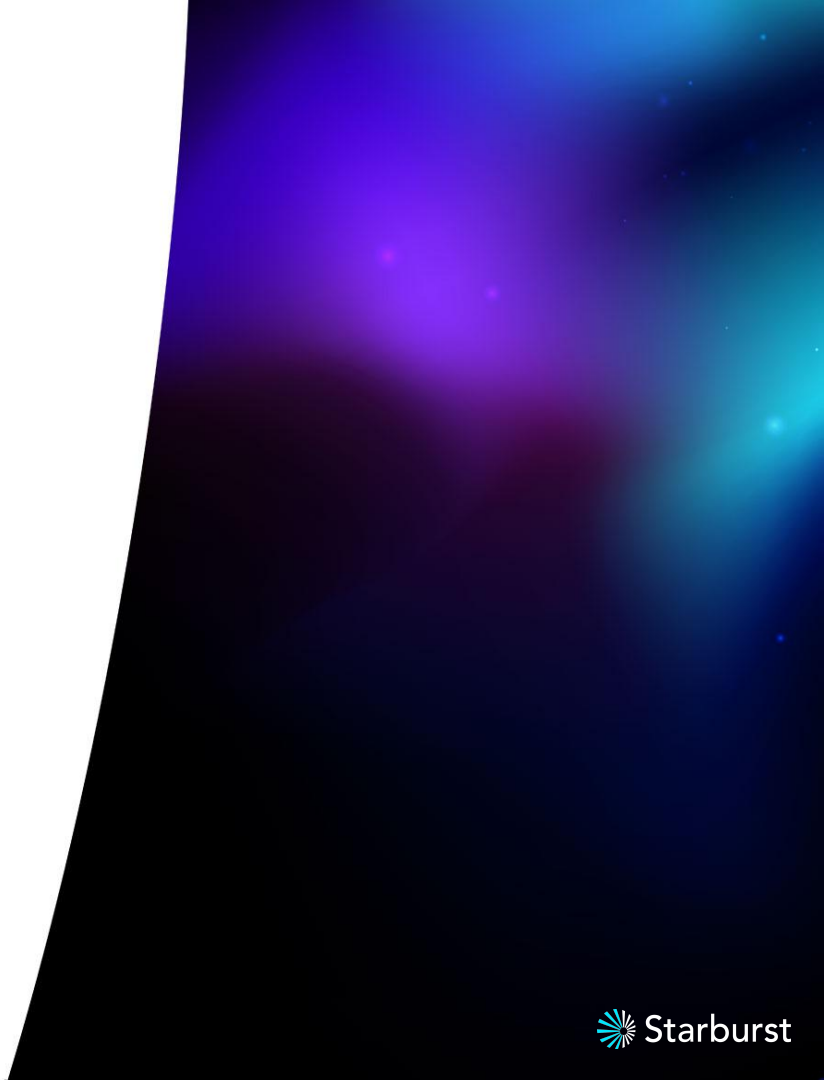
Part 1 agenda

- 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

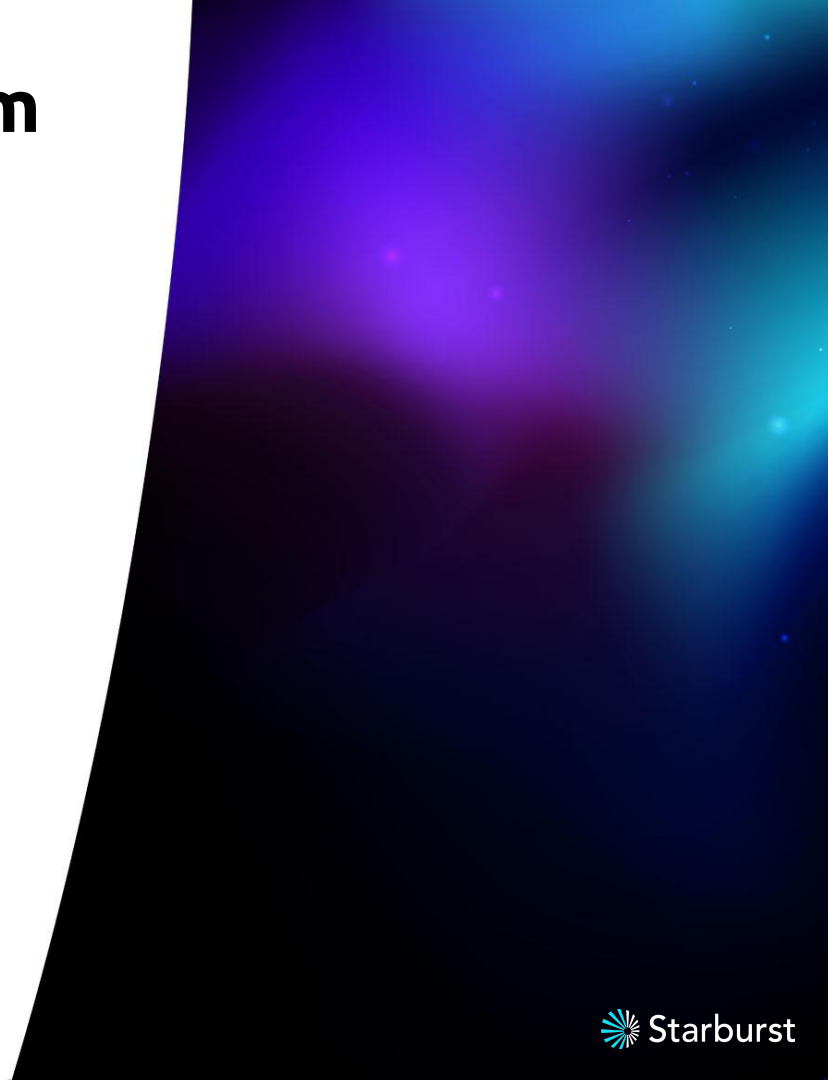
- Since the early 2000s, data generation and collection has skyrocketed due to the rise of the Internet.
- In 2005, Roger Magoulas referred to a large dataset that was almost impossible to manage and process using traditional BI tools as ***Big Data.***
- In 2006, Hadoop was designed to meet the needs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

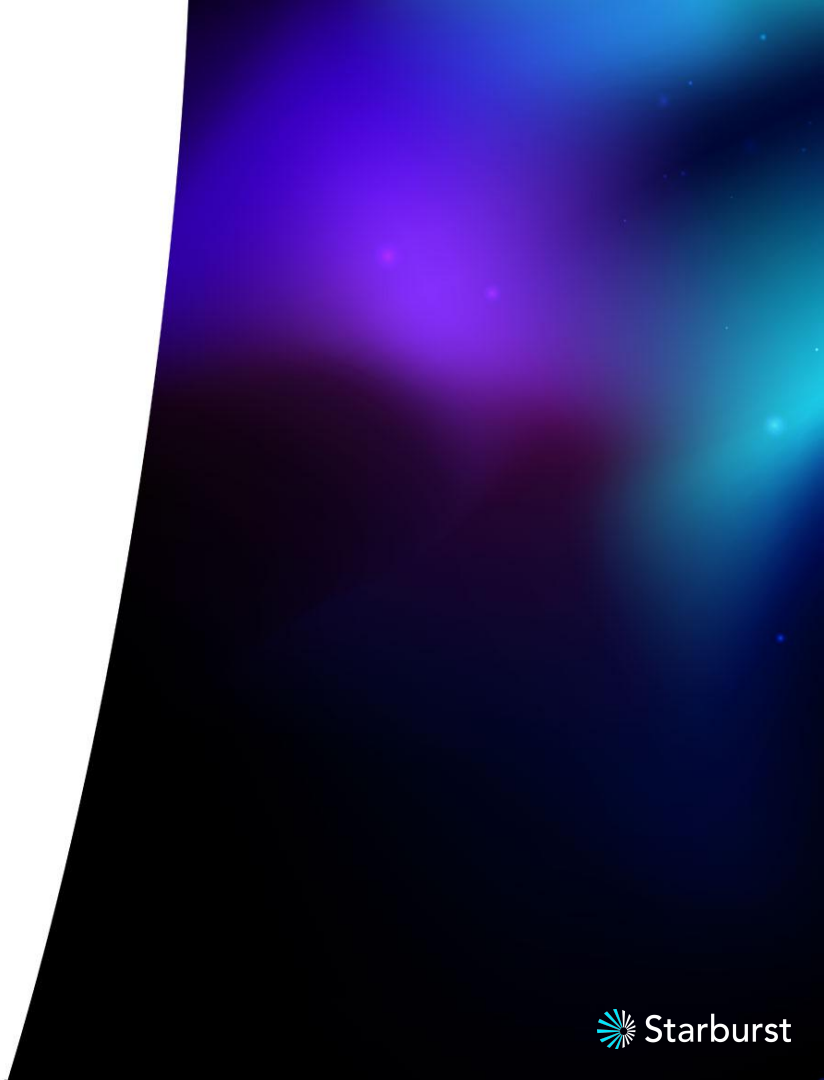


Starburst

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;
```

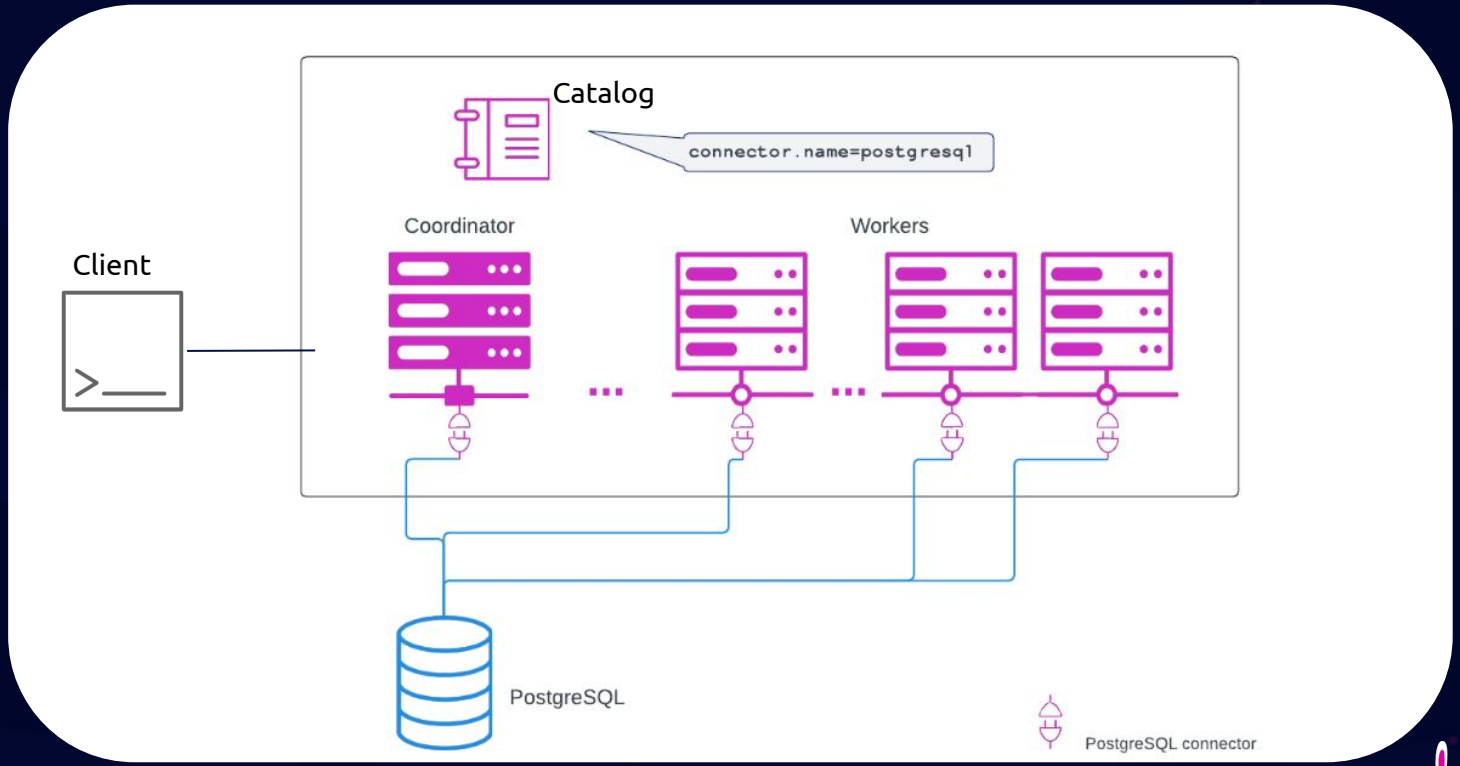
trino

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

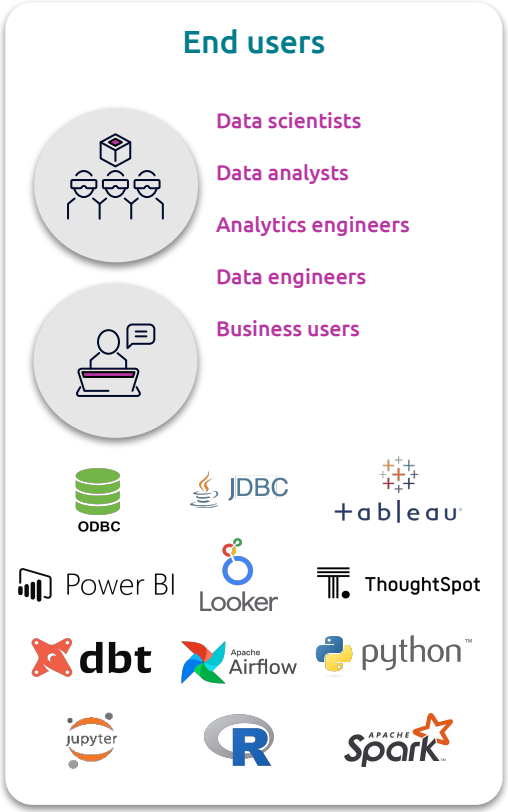
trino

How does Trino work?

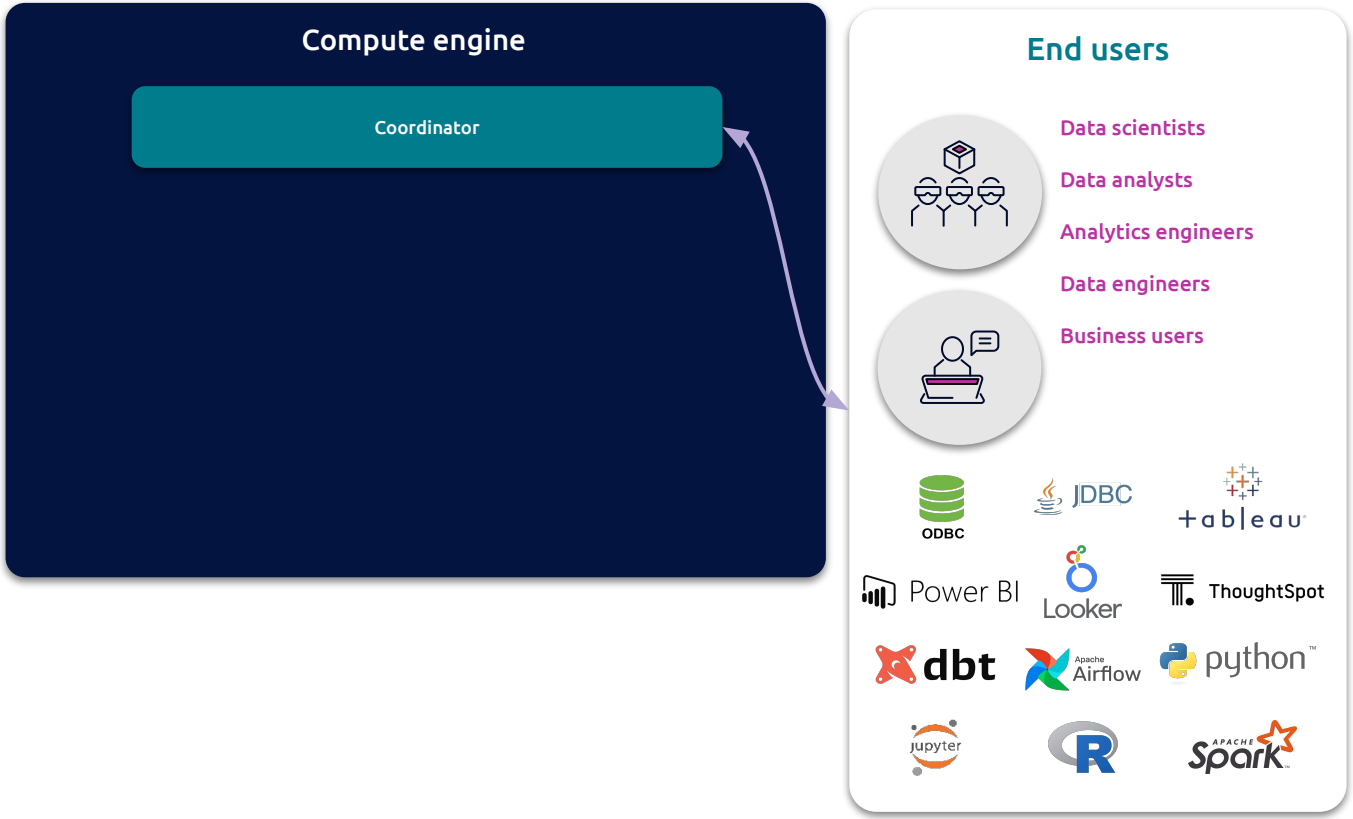


trino

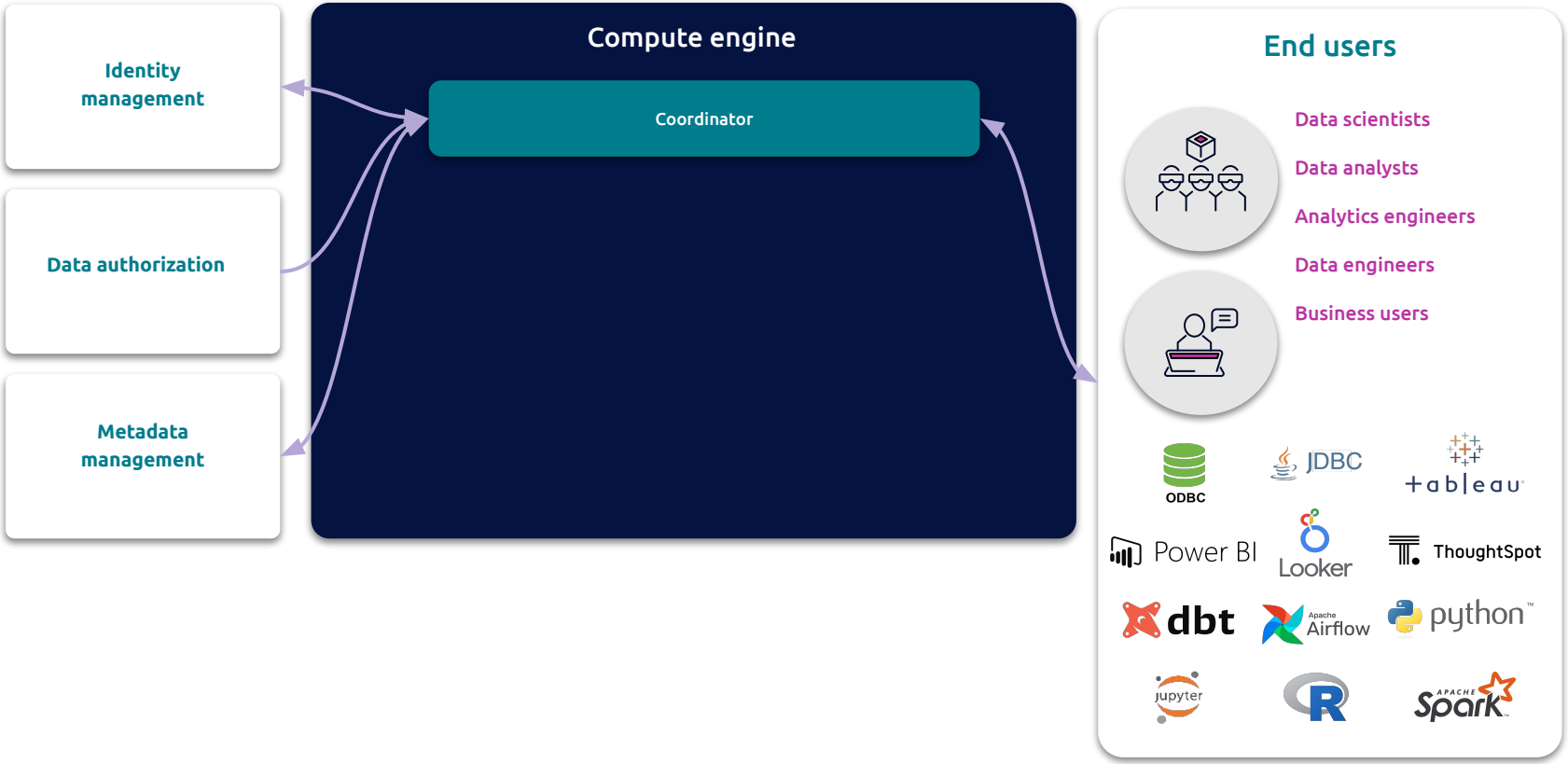
Data consumer submits a query



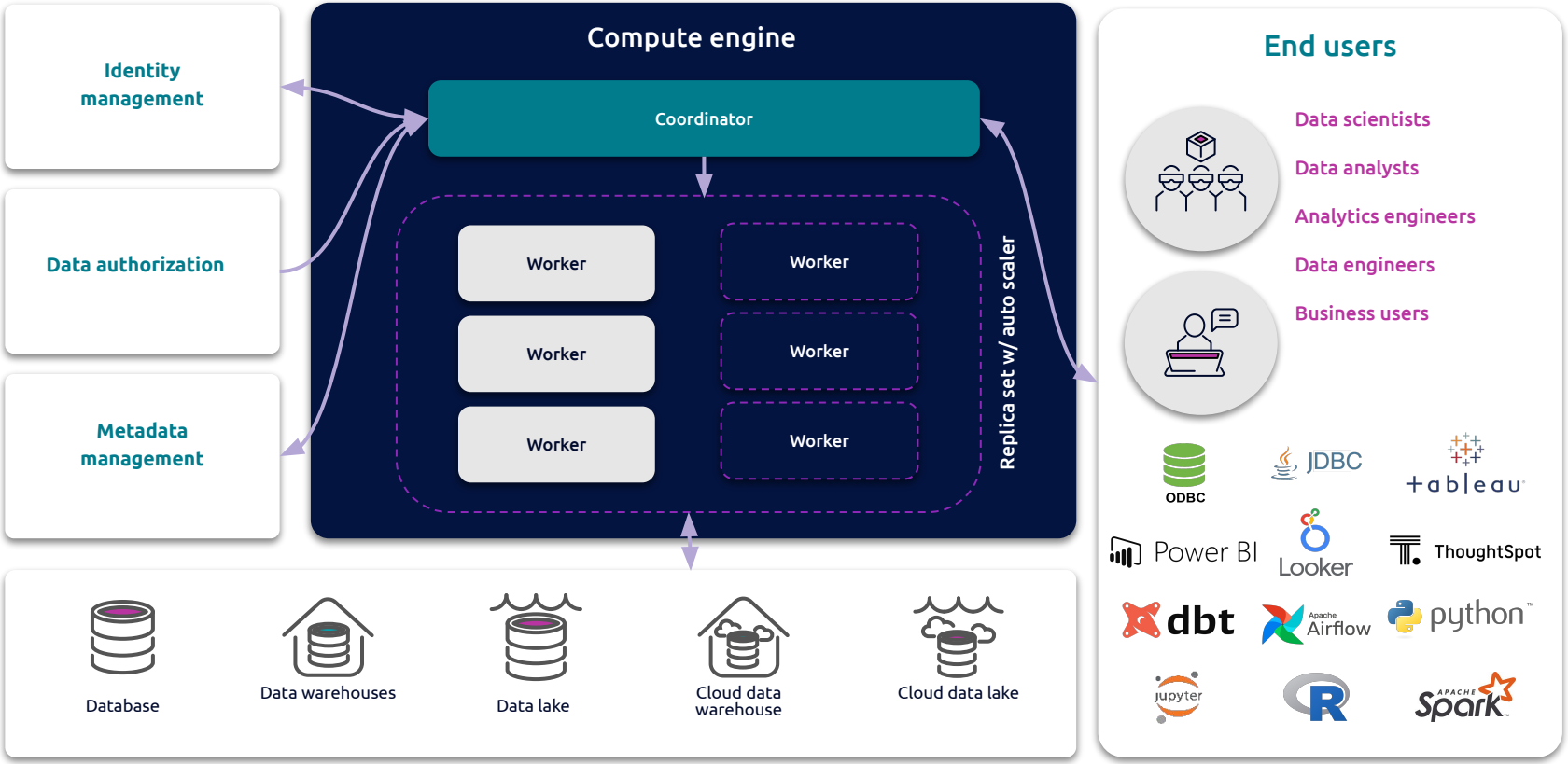
Coordinator node receives query



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

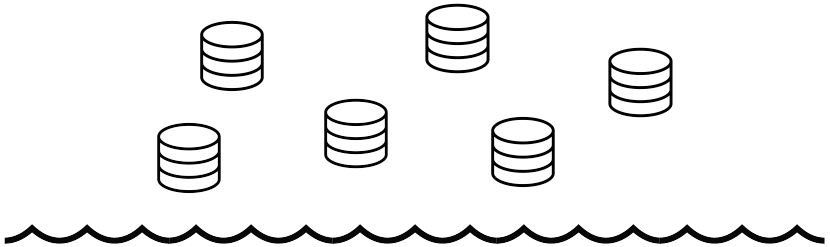
Getting to know Starburst

A data lakehouse platform

Starburst

The Open Data Lakehouse

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 Lakehouse Platform

The easiest way to *build and manage* your Open Data Lakehouse

90%
Faster time to insight

53%
Lower TCO

100%
Future-proof architecture

https://www.starburst.io/blog/icehouse-open-lakehouse/

Introducing Icehouse

Fully managed end-to-end open lakehouse platform

trino + ICEBERG

Open source trusted by

NETFLIX shopify stripe

SK telecom

Managed Icehouse

yello BESTSECRET kovi

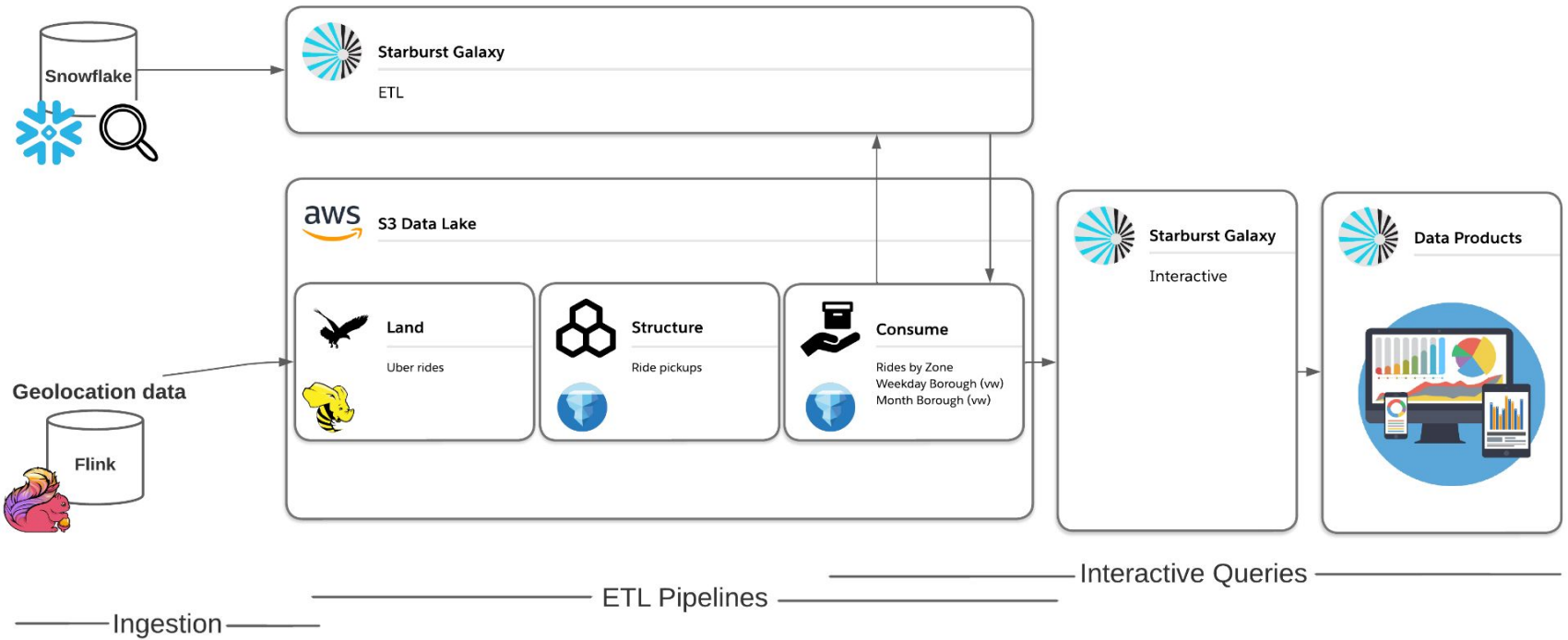
Starburst

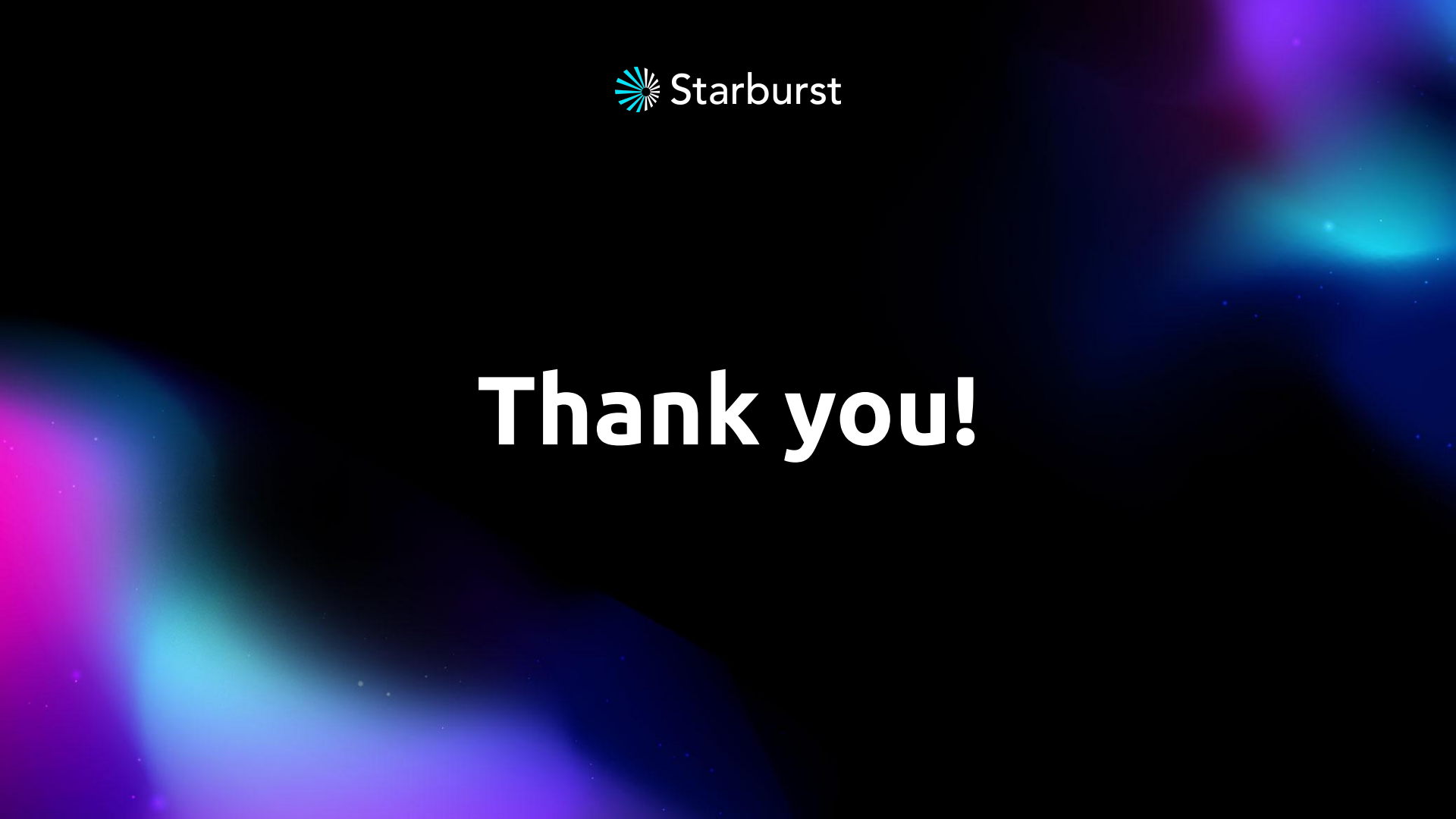
Starburst 101

Part 2: Hands-on Starburst Galaxy lab

Starburst

Project architecture




Starburst
Thank you!


Starburst