



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

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.

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.

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.

Structured Query Language (SQL)

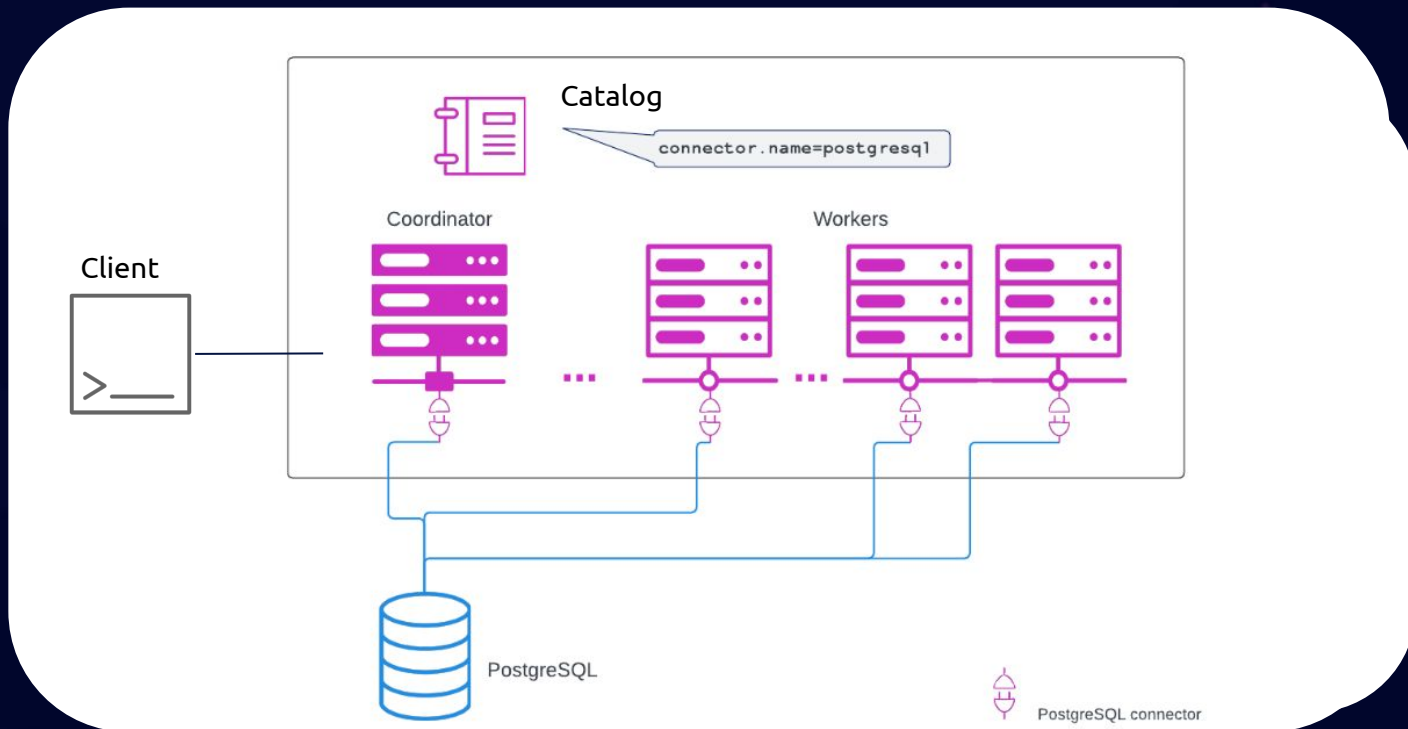
- Declarative language - specify what, not how
- Using SQL enables you to leave the heavy lift of optimizing the code to Trino

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

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

How does Trino work?



Data consumer submits a query

End users



Data scientists

Data analysts

Analytics engineers

Data engineers



Business users



ODBC



JDBC



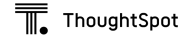
tableau



Power BI



Looker



ThoughtSpot



dbt



Apache Airflow



python



jupyter

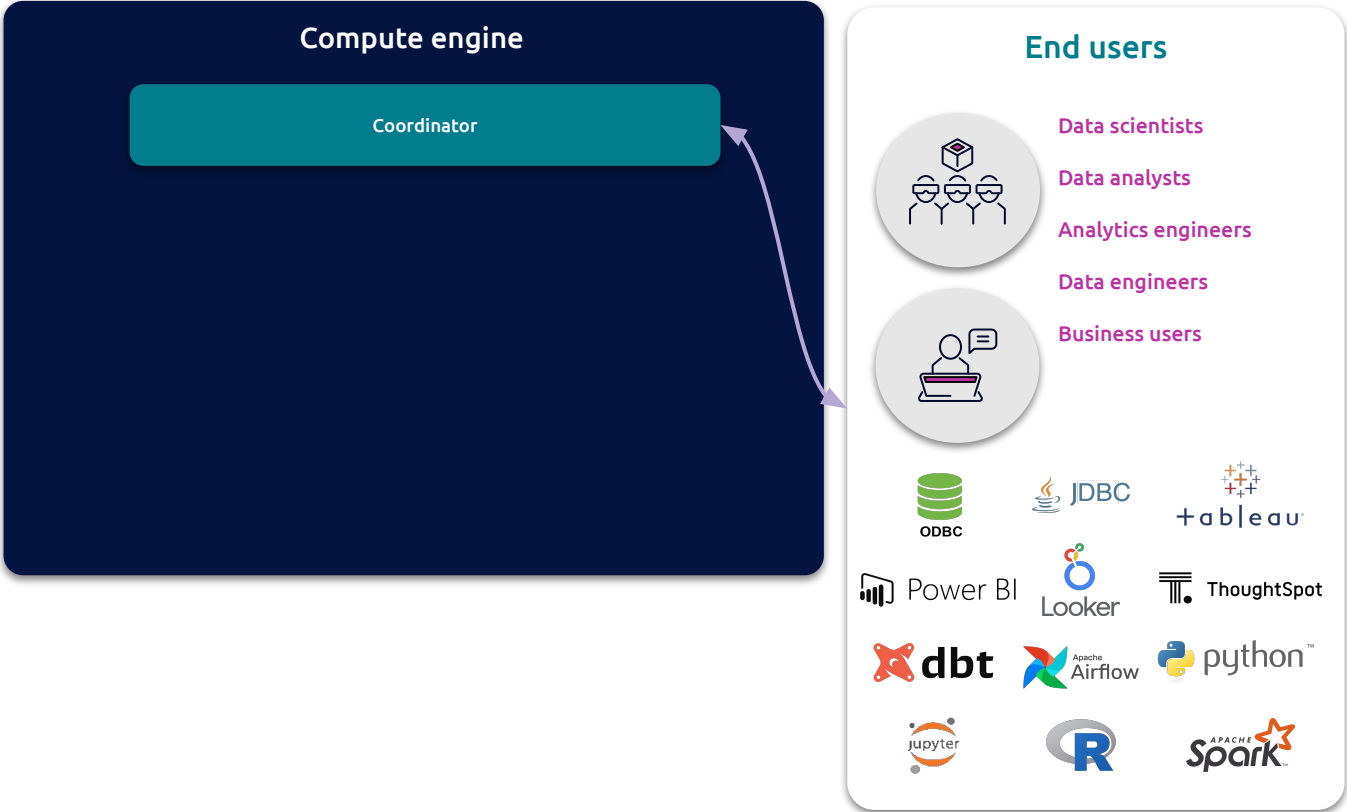


R

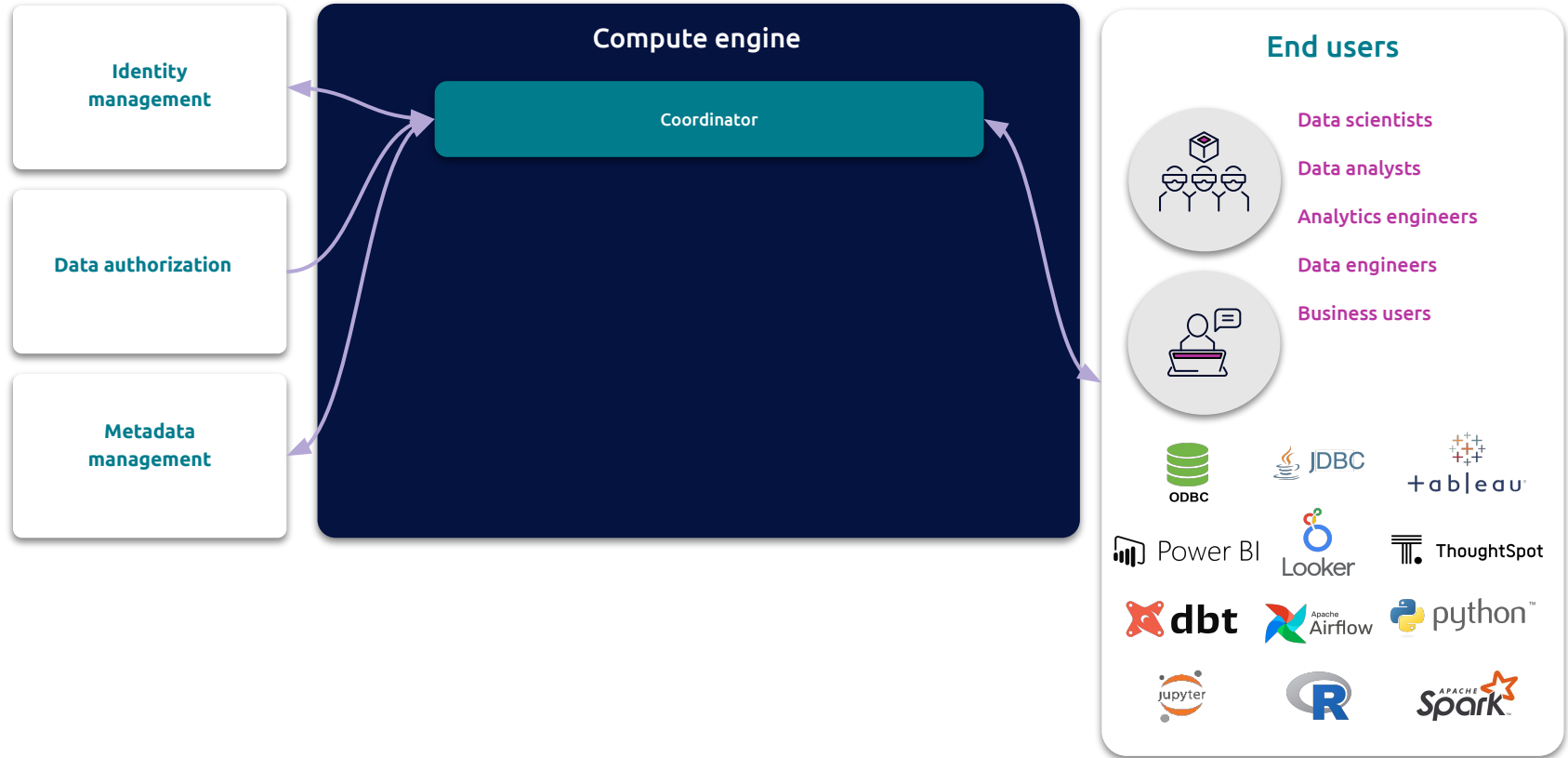


APACHE spark

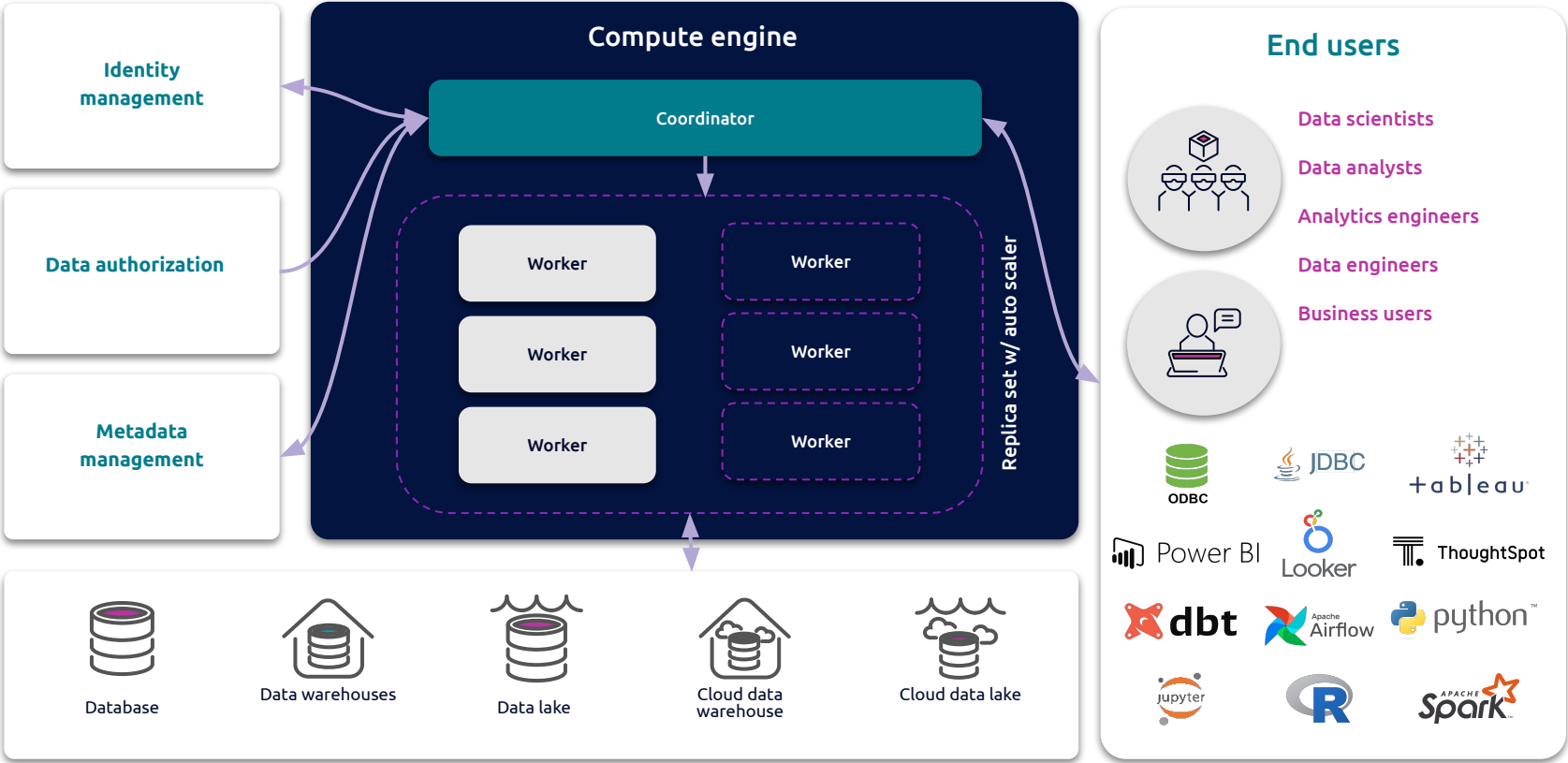
Coordinator node receives query



Parse and optimize query



Worker nodes interact with data



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



N

25PB on S3



in

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



P

600PB on S3
1000 nodes



lyft

10PB daily read data
250k queries per day



f

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



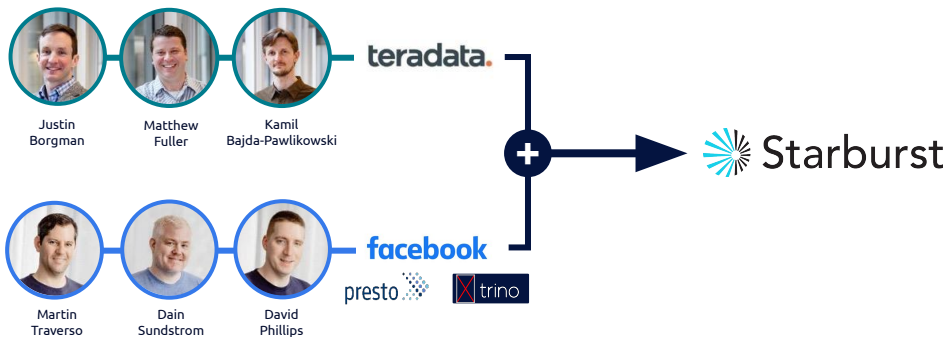
\$\$\$

(and time!)

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

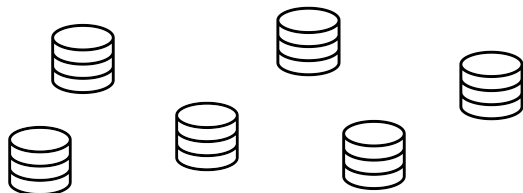


August 2012

March 2022

The Modern Data Lake

Global federated
access to data sources
beyond the lake



MPP query engine



Open table formats



Open file formats



Commodity storage
& compute



Object storage



Elastic compute



Data Lake Analytics Platform

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



90%

Faster time to
insight



53%

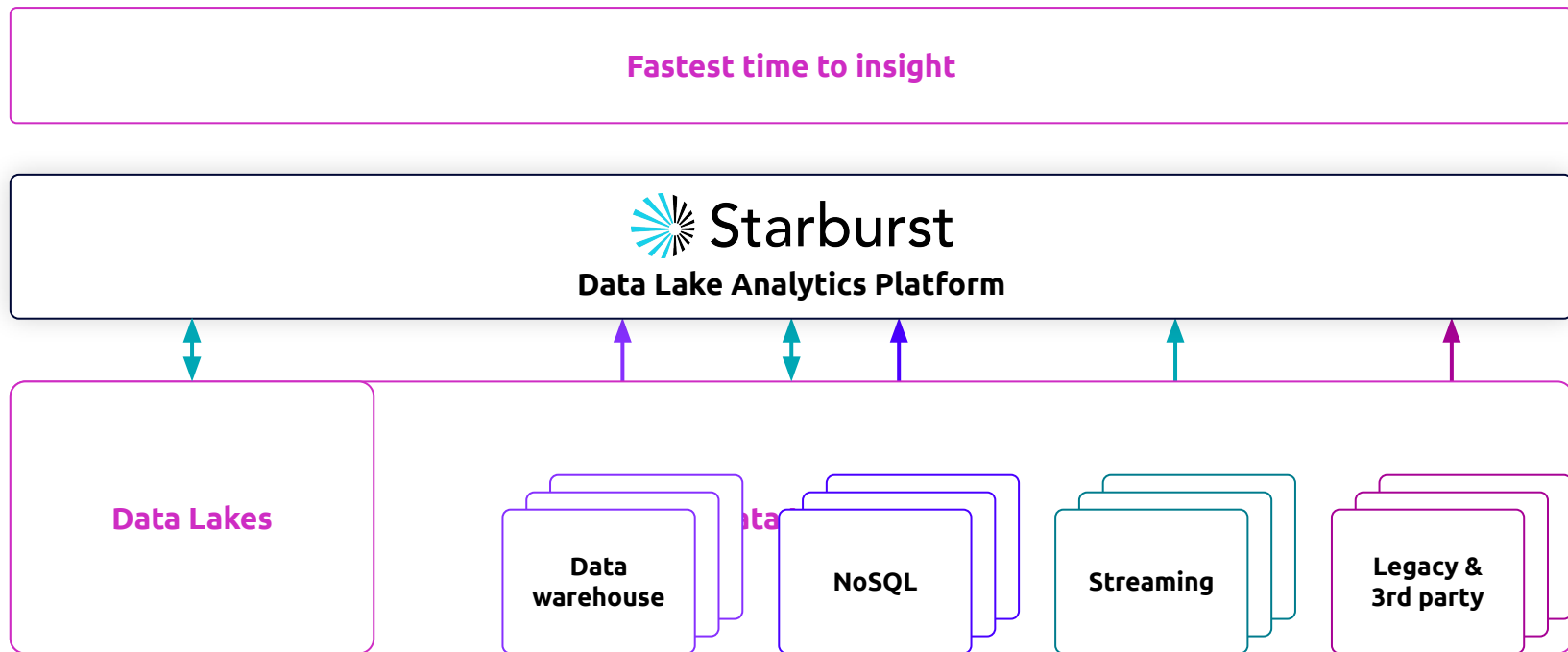
Lower TCO



100%

Future-proof
architecture

Starburst activates the data in and around your lake



Starburst Data Lake Analytics Platform

Analysis



BI Integration



Client Support

and many more...

Modeling / Semantic Layer

Data Products

Materialized Views

Data Catalog

Global Search

Data Profiling

Metastore

Query Engine

Warp Speed

Cost-based optimizer

Cluster autoscaling

Dynamic Filtering

Fault-tolerant execution

Accelerated Parquet reader

Scales to PB scale

Query Federation

Security & Governance

RBAC & ABAC

SSO & IAM

E2E Encryption

Row & Column Masking

Lineage

Monitoring & Logging

Management APIs

Service Accounts

Data Access



External data



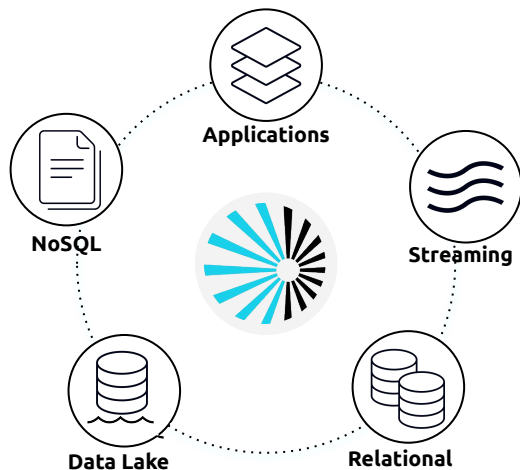
Modern data lake



On-prem,
cloud,
hybrid

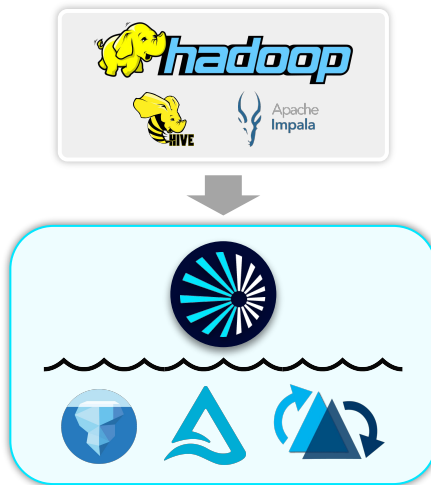
How Starburst helps

Federate and unify disparate data



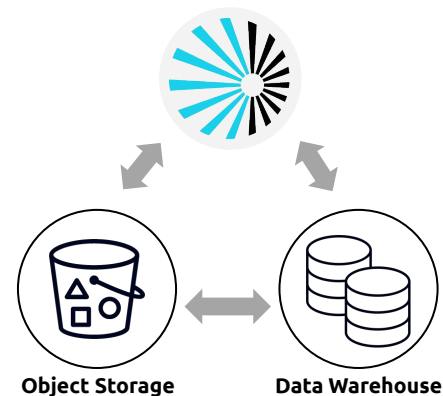
- 50+ enterprise-grade connectors
- Stargate for hybrid and cross-cloud

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



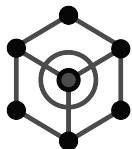
- Starburst Warp Speed (smart indexing & caching)
- Data Products
- Schema discovery

Benefits of our approach



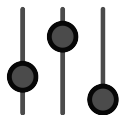
Simplicity

Take the fastest path from data to decision



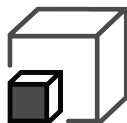
Access

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 101

Part 2: Hands-on Starburst Galaxy lab

Lab exercise

Data Engineers and Data Administrators

Part 3 agenda

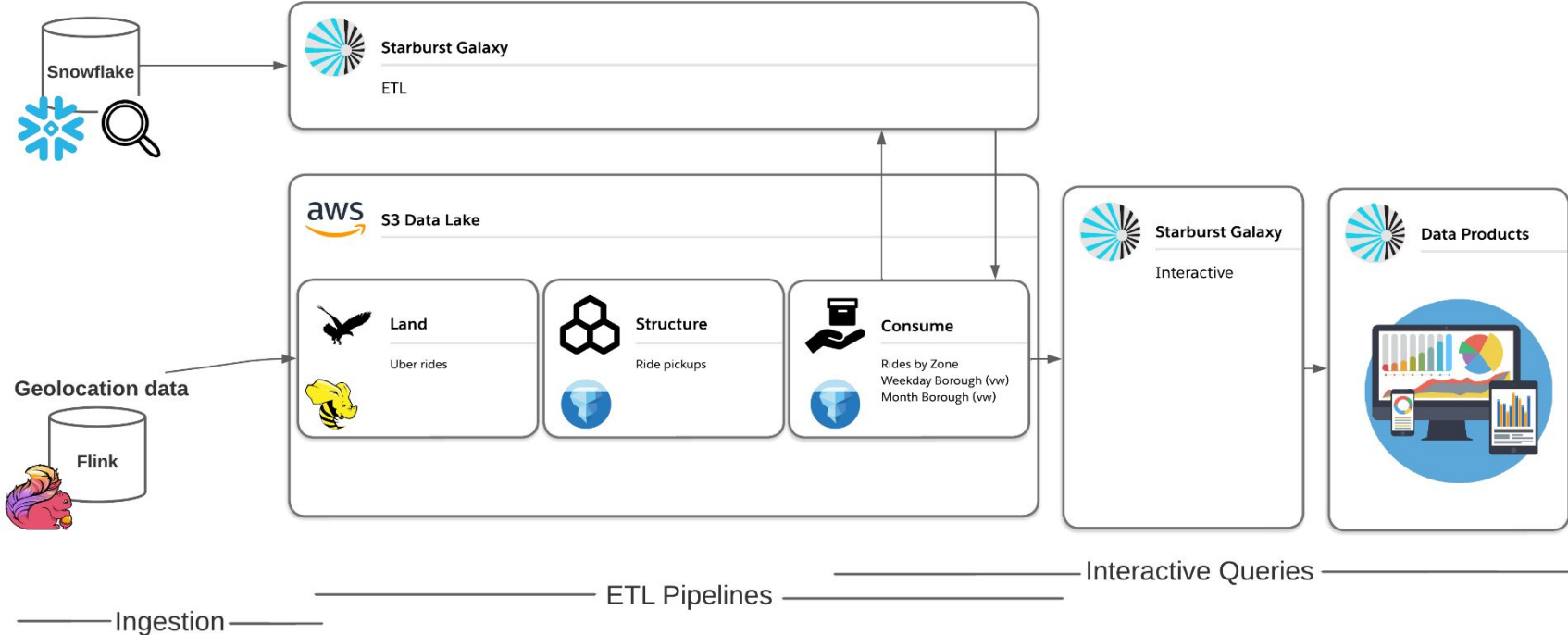
- Introduce project
- Connect to data sources
- Build within your data lake
- Create data products

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?

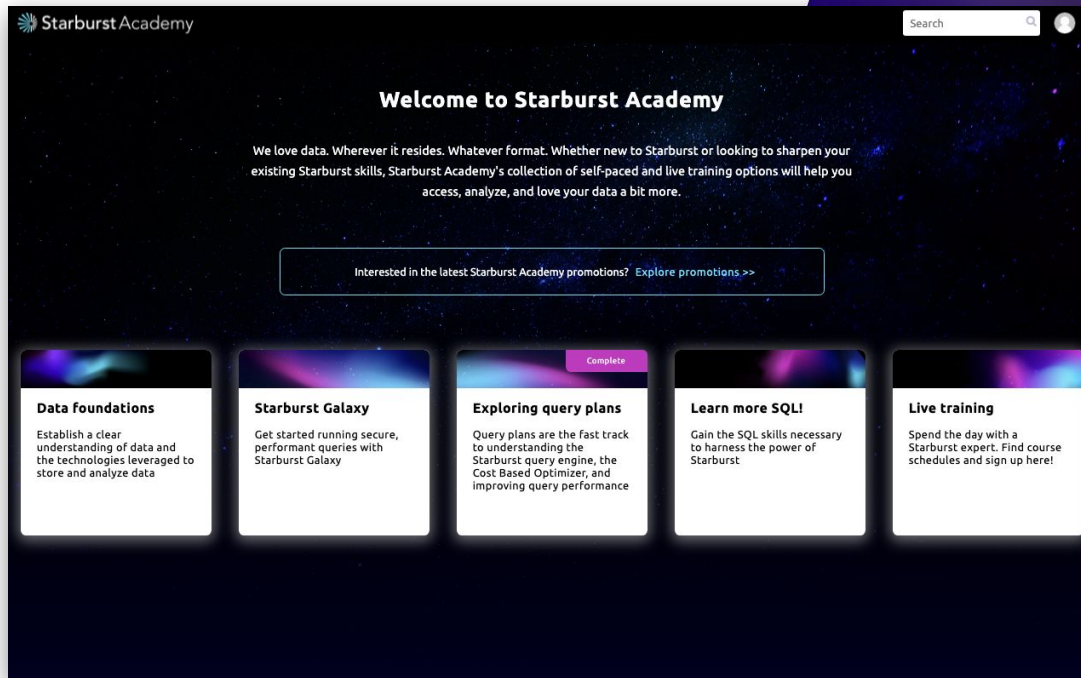
Project architecture



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



The screenshot shows the Starburst Academy website homepage. At the top left is the Starburst Academy logo. In the top right corner, there is a search bar with the text "Search" and a magnifying glass icon. The main heading is "Welcome to Starburst Academy". Below this, a paragraph reads: "We love data. Wherever it resides. Whatever format. Whether new to Starburst or looking to sharpen your existing Starburst skills, Starburst Academy's collection of self-paced and live training options will help you access, analyze, and love your data a bit more." Below the paragraph is a button that says "Interested in the latest Starburst Academy promotions? [Explore promotions >>](#)". At the bottom, there are five course cards, each with a title and a brief description:

- Data foundations**: Establish a clear understanding of data and the technologies leveraged to store and analyze data.
- Starburst Galaxy**: Get started running secure, performant queries with Starburst Galaxy.
- Exploring query plans**: Query plans are the fast track to understanding the Starburst query engine, the Cost Based Optimizer, and improving query performance. This card has a "Complete" badge in the top right corner.
- Learn more SQL!**: Gain the SQL skills necessary to harness the power of Starburst.
- Live training**: Spend the day with a Starburst expert. Find course schedules and sign up here!



Thank you!

Presenter Name

Date



Starburst

Starburst supports your evolving data strategy

Starburst technology applications

- Data lakehouse
- ETL / batch processing

Data centralization

- Query federation
- Embedded analytics engine

Connect to Data Anywhere

Business-focused Data Products

- Data Products
- Data Mesh

Starburst is the analytics engine that fits into any environment

BI Integration



Client Support



Analytics Engine

MPP Query Engine Data Products Fault-Tolerant Execution

Query Optimizer Elastic Autoscaling Smart Indexing & Caching Metrics & Logging



Global Security

End-to-End Encryption Data masking control Query auditing Access

Data Lakes



Relational DBs



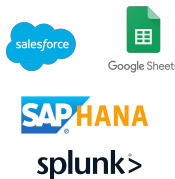
NoSQL Stores



Real-time analytics

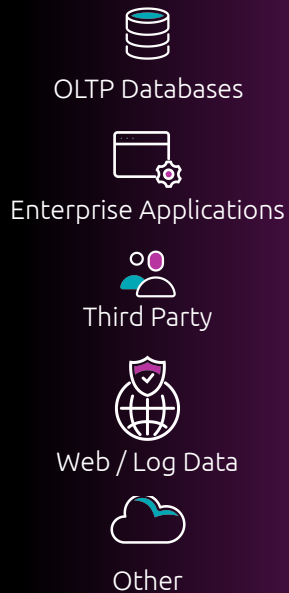


Applications



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

Data Sources



BI / Analytics



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



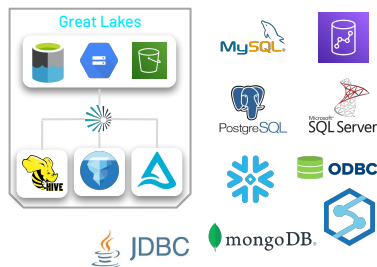
Security

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



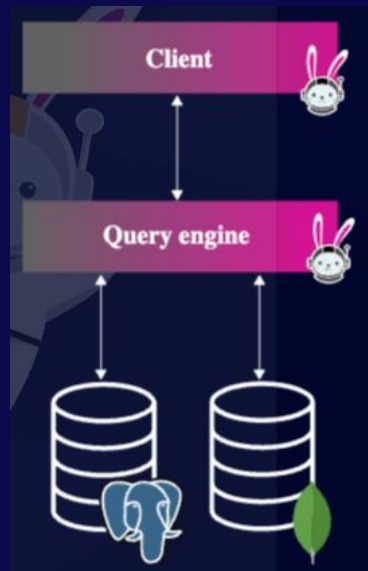
Fast, distributed, ANSI SQL processing engine

Trino is a ludicrously fast, open source, SQL query engine...

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

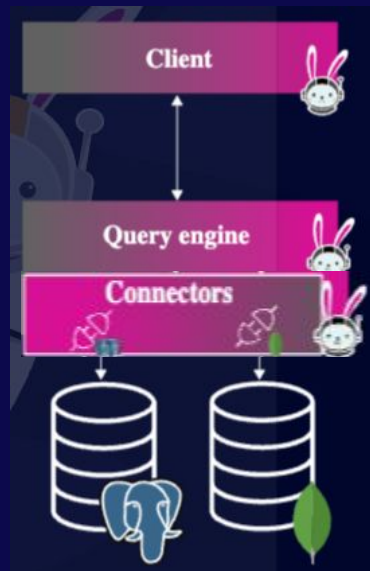
Disparate data sources

- 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



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 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.

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

