

Implementing the medallion architecture with Starburst

Lester Martin

Michael Markieta



Starburst

Workshop approach – mostly HANDS-ON!

Talk a tiny bit, do some exercises, do it all over two more times!

Getting started

Lab 1: Create student account (5 mins)

Lab 2: Execute queries in Starburst Galaxy (15 mins)

Lab 3: Create and populate Iceberg tables (10 mins)

OPTIONAL Lab: Data modifications and snapshots with Iceberg (20 mins)

OPTIONAL Lab: Exercise advanced features of Iceberg (15 mins)

Data pipelines & data products

Lab 1: Construct a pipeline with insert-only transactions (45 mins)

OPTIONAL Lab : Construct a pipeline with the MERGE statement (20 mins)

Lab 2: Produce and consume a data product (15 mins)

Managed Iceberg pipelines

Lab 1: Explore the file ingestion service (10 mins)

Lab 2: Explore the streaming ingestion service (10 mins)



Trino & Starburst

The rise of big data

Querying large volumes of data was difficult and time consuming

Early 2000s: Data generation and collection has skyrocketed due to the rise of the Internet

2006: Apache Hadoop was designed to meet the needs of large datasets on a scale previously unimaginable

2008: Facebook created Apache Hive to query terabytes of data in Hadoop using a SQL-like interface. Data consumers were limited by the number of queries they could run — often fewer than 10/day

The birth of Trino

A new query engine designed to solve the data accessibility problem

2012: Trino (*formerly known as Presto*) is created by Martin Traverso, Dain Sundstrom, David Phillips and Eric Hwang at Facebook

Trino is an open source query engine that:

- *Harnesses the power of distributed computing*
- *Separates compute from storage*
- *Super fast and performant*
- *Supports pluggable connectors to a variety of data sources*
- *ANSI-SQL BASED!!!! Which means... SQL on anything!*



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

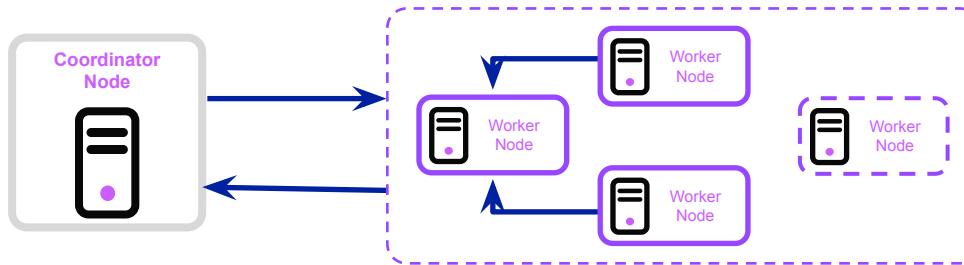
Server stereotypes

Coordinator node

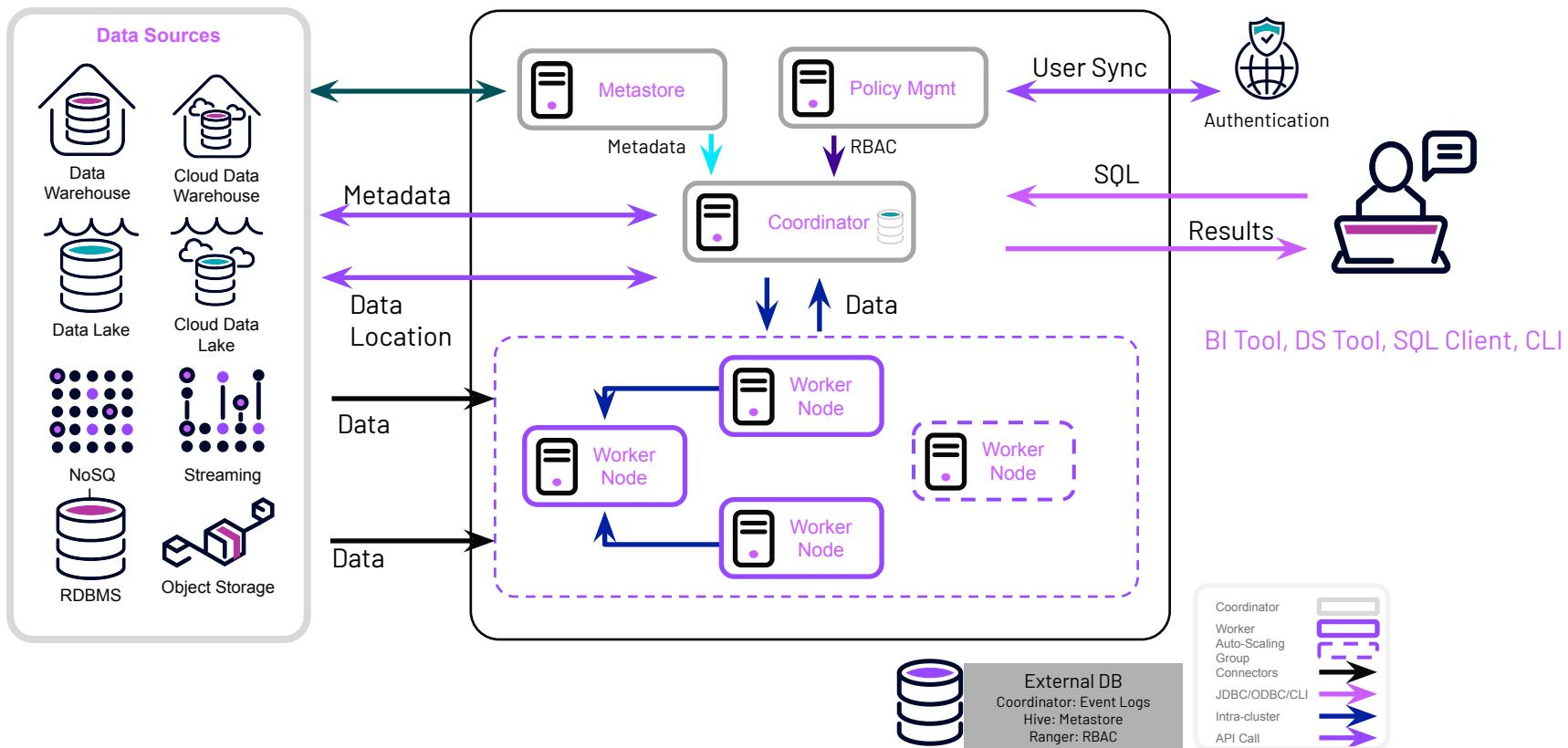
Server that is responsible for parsing statements, planning queries, and managing Trino worker nodes.

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



Rich ecosystem of data source connectors

Open-source & Starburst Proprietary

Data Source
Connectors

Real-time Analytics



Data Lakes



NoSQL Stores



Applications



Relational DBs





Apache Iceberg

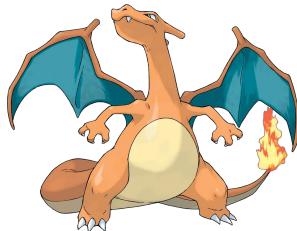


Apache Iceberg

Applying data warehouse principles to the data lake

- Created by Ryan Blue & Daniel Weeks at **Netflix** in 2017.
- Solve the challenges of performance, data modification and schema evolution in the lake.
- Uses open data concepts (orc, parquet, avro) and architecture.
- Seen enormous interest and adoption over the last few years.
- Applies SQL behavior like hidden partitioning and schema evolution the the lake also offering modern warehouse SQL such as MERGE, UPDATE, DELETE, and Time Travel.

The Open Data Lakehouse



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

Hands-on exercises

Getting started

- **Lab 1: Create student account (5 mins)**
- **Lab 2: Execute queries in Starburst Galaxy (15 mins)**
- **Lab 3: Create and populate Iceberg tables (10 mins)**
- **OPTIONAL Lab: Data modifications and snapshots with Iceberg (20 mins)**
- **OPTIONAL Lab: Exercise advanced features of Iceberg (15 mins)**





Data pipelines

History of Trino - ETL processing

From purely interactive use-cases to multiple workloads

2013: Released into production at Facebook for interactive use cases

2014: Users start scheduling batch/ETL queries with Trino instead of Hive

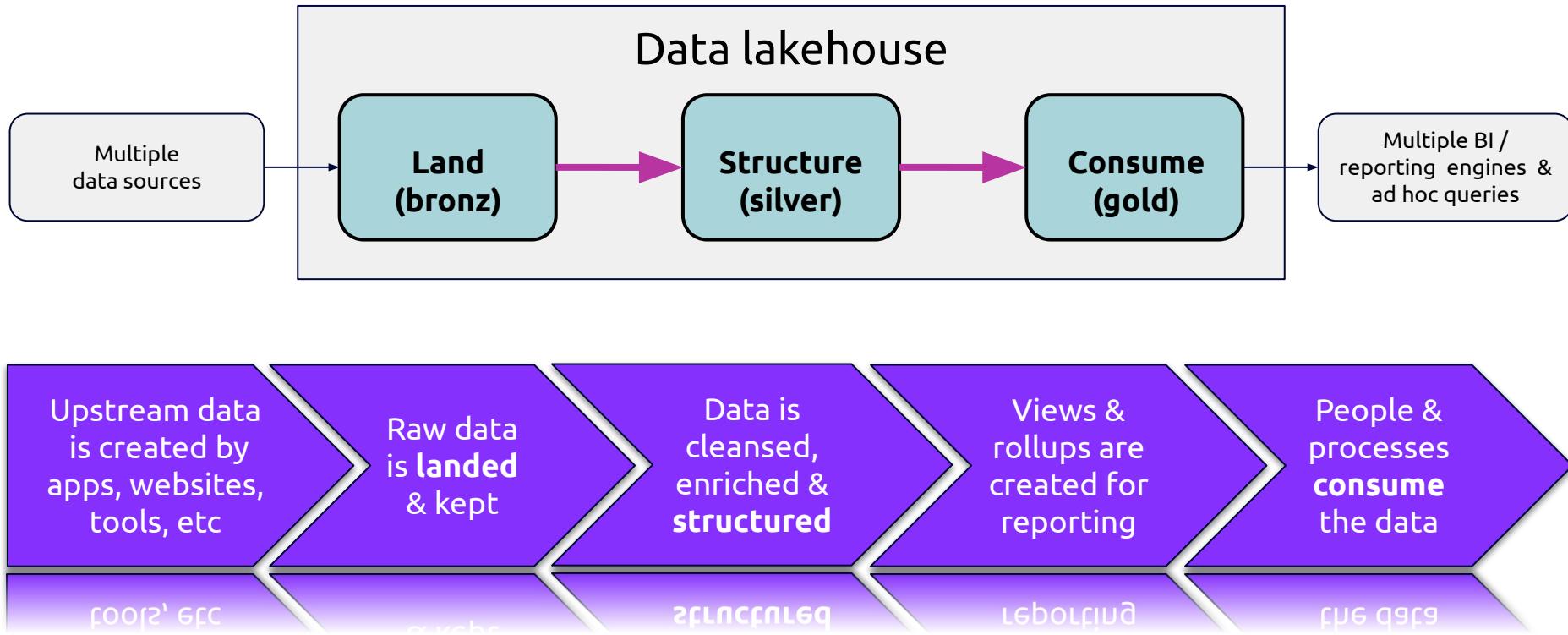
2018: 50% of existing ETL workloads and 85% of new workloads on Trino

Why?

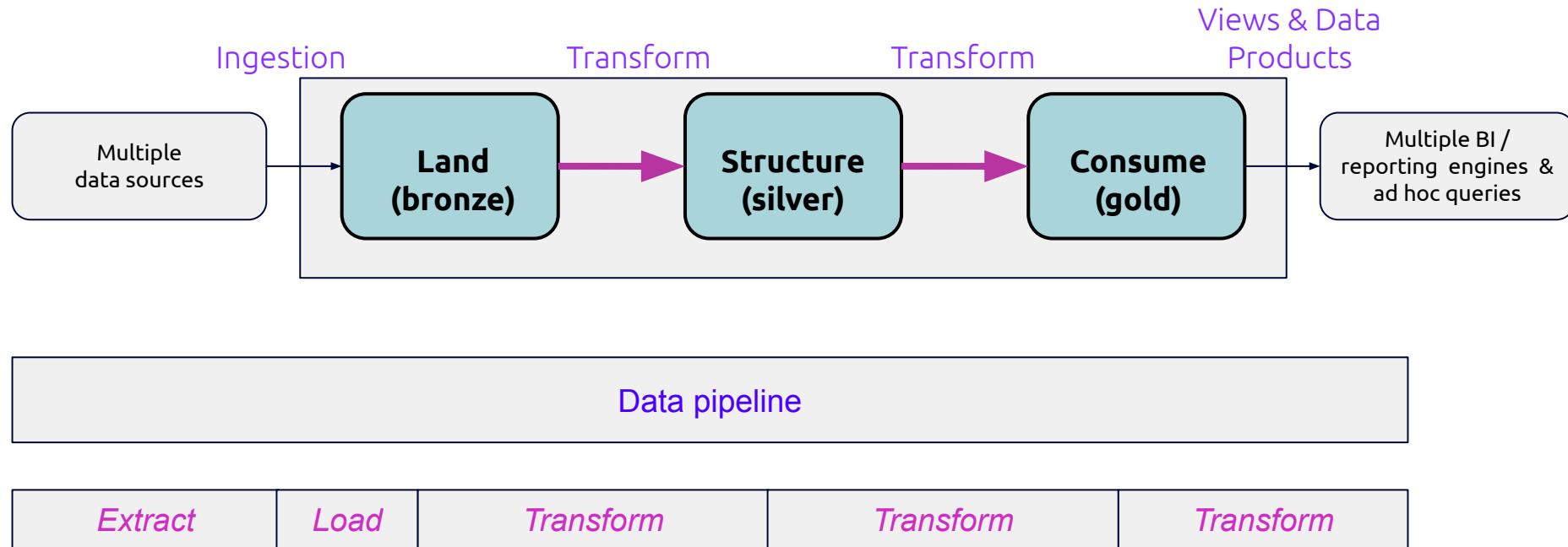
- *Trino can communicate with disparate data sources to federate data*
- *Trino is a distributed, massively parallel processing system*
- *Faster, Cheaper and ANSI-SQL BASED!*

Soon others caught on, and teams like [Salesforce](#) and [Lyft](#) started utilizing Trino for Batch/ETL capabilities.

Activities across the medallion architecture



Data pipelines across the medallion architecture





Data products

Data's biggest challenge

There's a great divide between data producers and data consumers

- Data consumers fail to **accurately** convey their needs to data producers
- Data producers struggle to understand the **business value** attached to various requests

Why is everything so grey?

- Data exchanges hands too many times to count
- Consumers will then manipulate said data themselves inaccurately

What are data products?

Data products are ***curated datasets*** packaged to ***create value*** for downstream consumers

- ***Curated datasets:*** Data products are demand-driven and built for a specific need
- ***Create value:*** Data products create value by presenting data in a way that makes it more useful and more accessible

The components of data products

Metadata

The most important part of any data product is the data inside it taking the form of:

- Tables
- Views
- Materialized views

Abstracted data

The table definition associated with the data, including:

- Business context
- Tags
- Lineage information
- Statistics
- Data samples
- Ownership

Access Patterns

Intended access plan for the end user, including:

- Who has access to specific data
- Compute
- How the data is accessed

Defining qualities of data products

Minimum Viable Data Product

Not all data products have to have each component in this list to be defined as a data product.

The minimum viable data product is a curated dataset created for a specific use case to add value by allowing others to self-serve the dataset for insight.

Independent entities

Each data product has all of the structural components to do its job as a discrete object.

Access to the data product should give you all the information you need to gain insights.

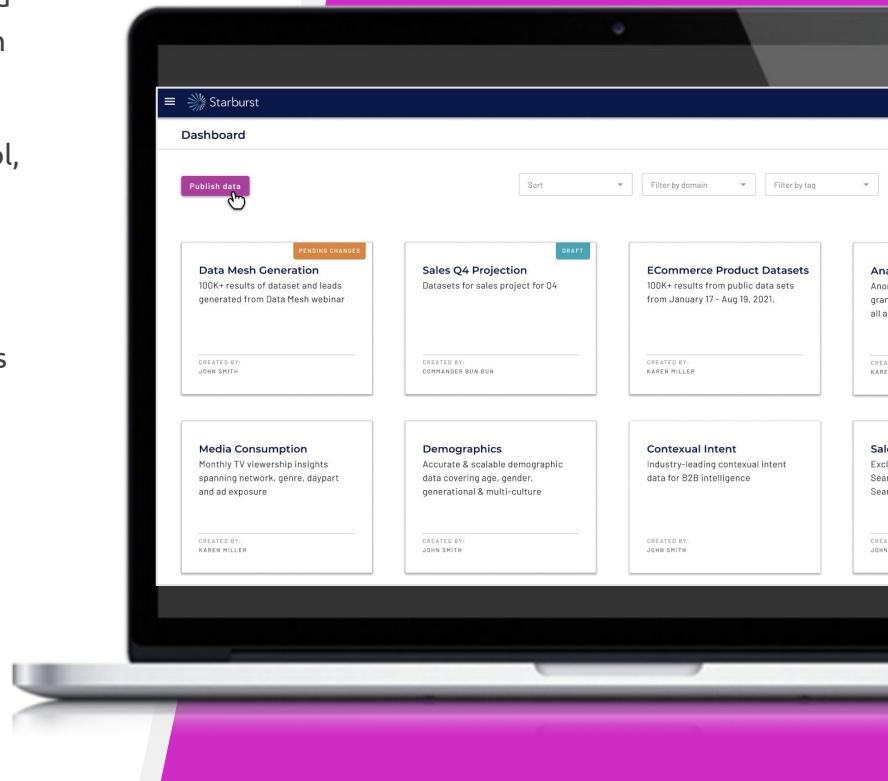
Social dimension

There is a social dimension to data products. They are typically created for others, shared widely, and used across teams.

As such, the collaborative way in which we create them, deploy them, and interact with them is one of their defining characteristics.

Discover, create, publish, manage, and share data products based on multiple datasets

- **Streamlined visibility:** quickly understand the usage metrics, and create, publish, find, and manage curated data products based on multiple data sets
- **Consistent governance:** secure data products with access control, ensuring consistent governance from source level to data products
- **Ultimate accessibility:** query data products that are trusted and approved for frequent business use, rate and share data products internally for use across the organization



Hands-on exercises

Data pipelines & data products

- **Lab 1: Construct a pipeline with insert-only transactions (45 mins)**
- **OPTIONAL Lab: Construct a pipeline with the MERGE statement (20 mins)**
- **Lab 2: Produce and consume a data product (15 mins)**





Managed Iceberg pipelines



Managed Iceberg pipelines

Creates, loads, and maintains Live tables in Iceberg table format

- Data ingestion source options
 - Kafka
 - S3
- Continuously loads **Live** tables of two types
 - **Raw** tables are exact replicas of the incoming data
 - **Transform** tables are enhancements/refinements from Raw ones
- Table maintenance activities, such as compaction and snapshot expiration, are automatically handled for all Live tables



No coding required

Just config via our wizard

Data ingest

Connected sources

lester_tx

S3 file ingest
s3://starburst101-hands-on

Select a source

Choose from streaming data or continuous file loading from an S3 bucket.

Apache Kafka

Confluent Cloud

Amazon MSK

Amazon S3

Connect to a source

Land your data in a raw table

Create transform table and schematize data

Connect new source

Search sources

Create live table

...

...
2

Hands-on exercises

Managed Iceberg pipelines

- Lab 1: Explore the file ingestion service (10 mins)
- Lab 2: Explore the streaming ingestion service (10 mins)



Thank You



starburst.ai