



# Transitioning to Trino @ Lyft

Charles Song ([charless@lyft.com](mailto:charless@lyft.com))

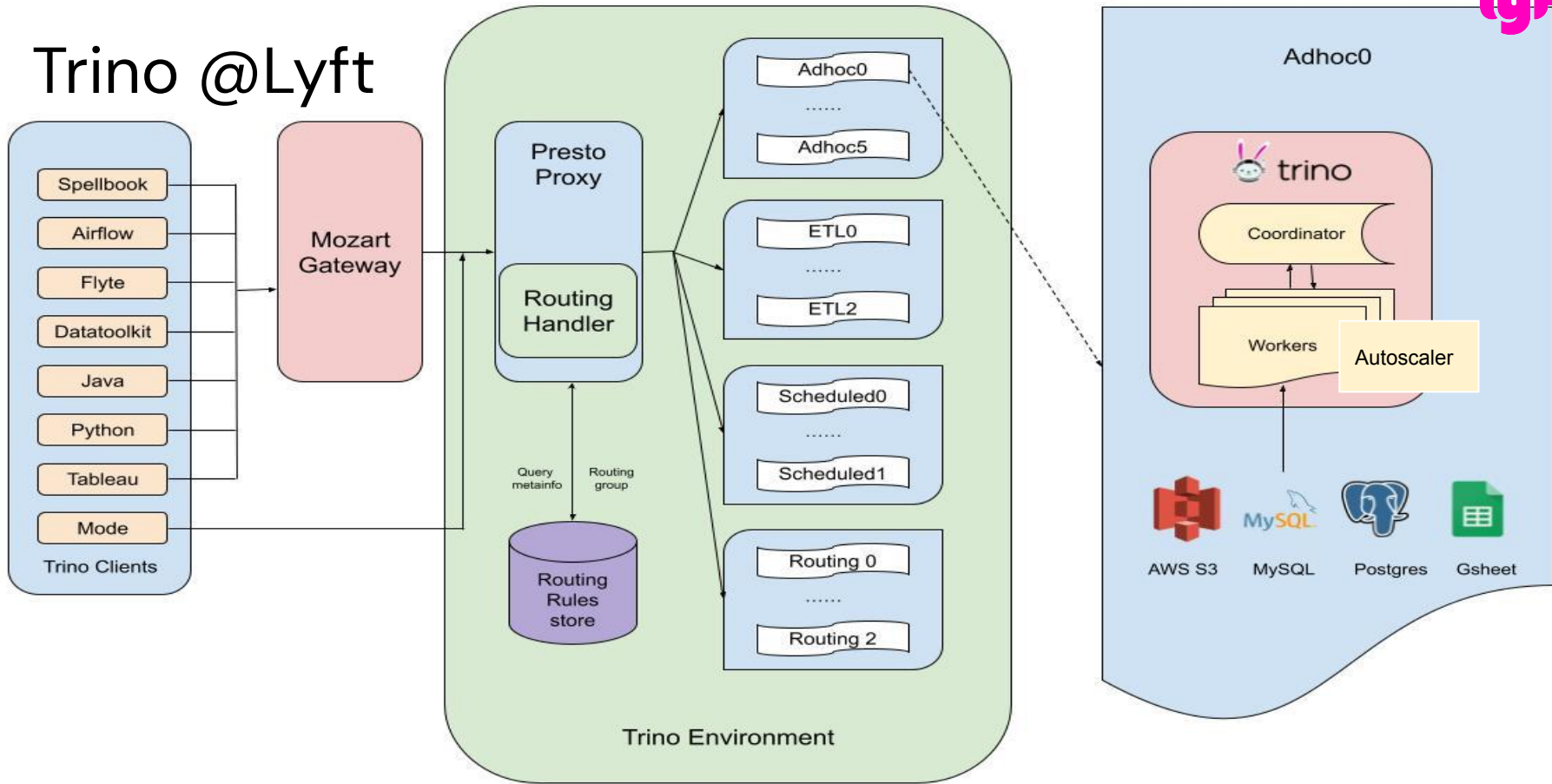
Data Platform @ Lyft



# Contents

- **Trino @Lyft**
- **Trino & Spark (Hive) – When to Use**
- **What Faster Query Means**
- **Key Numbers & Factors**
- **Future Work**

# Trino @Lyft



Trino Architecture

# When to Use Trino



ChatGPT 4 >



You

When Shall i use trino over spark and why?



ChatGPT

Choose Trino over Spark for interactive, ad-hoc querying and data exploration, particularly when speed is crucial. Trino excels in low-latency responses, making it ideal for querying large datasets in real-time. Conversely, Spark is better suited for complex data processing tasks, like ETL operations and machine learning, due to its robust analytics capabilities and support for a wider range of data processing tasks.

# When to Use Trino

## Great For Interactive Queries

- Easy to use
- Low latency



ChatGPT 4 >



You

When Shall i use trino over spark and why?



ChatGPT

Choose Trino over Spark for interactive, ad-hoc querying and data exploration, particularly when speed is crucial. Trino excels in low-latency responses, making it ideal for querying large datasets in real-time. Conversely, Spark is better suited for complex data processing tasks, like ETL operations and machine learning, due to its robust analytics capabilities and support for a wider range of data processing tasks.

# When to Use Spark

## Great For ETL

- Complexity
- Reliability
- Large volume

## And Machine Learning

- Built-in query functions
- Faster ML integration



You

When Shall i use trino over spark and why?



ChatGPT

Choose Trino over Spark for interactive, ad-hoc querying and data exploration, particularly when speed is crucial. Trino excels in low-latency responses, making it ideal for querying large datasets in real-time. Conversely, Spark is better suited for complex data processing tasks, like ETL operations and machine learning, due to its robust analytics capabilities and support for a wider range of data processing tasks.

# When to Use ~~Spark~~ Trino

## Great For ETL

- ~~Complexity~~
- Queries can be complex
- Reliability
- Highly reliably + Tardigrade
- ~~Large volume~~
- 10 TB+ Level Volume

## And Machine Learning

- ~~Built in query functions~~
- Customized query functions
- ~~Faster ML integration~~
- Faster ML data aggregation



ChatGPT 4 >



You

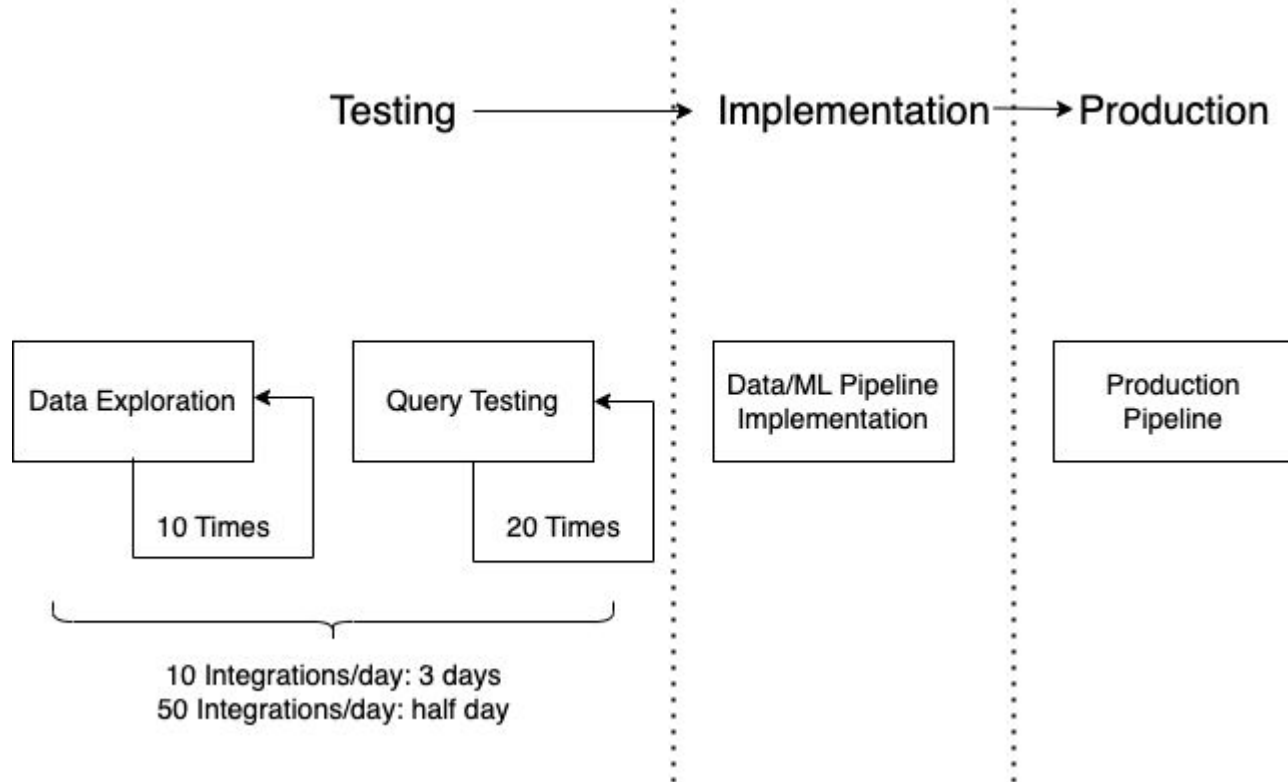
When Shall i use trino over spark and why?



ChatGPT

Choose Trino over Spark for interactive, ad-hoc querying and data exploration, particularly when speed is crucial. Trino excels in low-latency responses, making it ideal for querying large datasets in real-time. Conversely, Spark is better suited for complex data processing tasks, like ETL operations and machine learning, due to its robust analytics capabilities and support for a wider range of data processing tasks.

# For Data/ML Engineer, Faster Query Means...







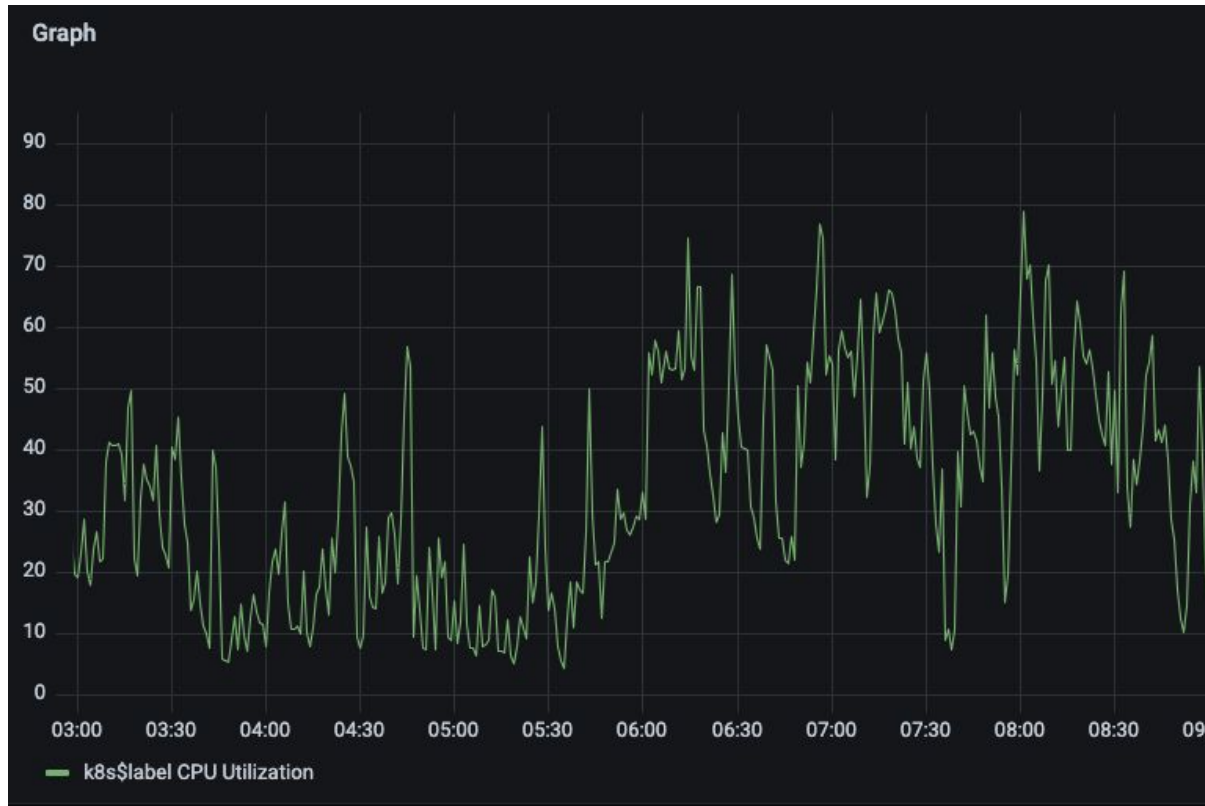
# Key Numbers

- 100PB level data volume
  - 10x volume with similar infra costs vs Spark SQL
- 0.01%–0.1% infra level reliability issues
- Consistent seconds level latency for small queries
- Consistent high success rate for long-running ETL queries
- 10TB level queries through organic growth
- Low occurrence of “noise neighbor” problem over the year

# Key Factors

- Keep Trino Version up-to-date (Replayer)
- Autoscaling
- Load Balancing (Prestoproxy/Gateway)
- Fault-Tolerant Feature (Tardigrade)
- Graviton Instances

# Without Autoscaling



# With Autoscaling





# What's Next for Trino @Lyft

- Better ML pipeline integration
- Keep pushing Trino growth and simplify potential migration from other engines
- Try on new connectors
- Caching
- Faster Trino upgrades

# Thank you

Q&A

Charles Song ([charless@lyft.com](mailto:charless@lyft.com))

Data Platform @ Lyft

