Running Trino at Exabyte-Scale Data Warehouse

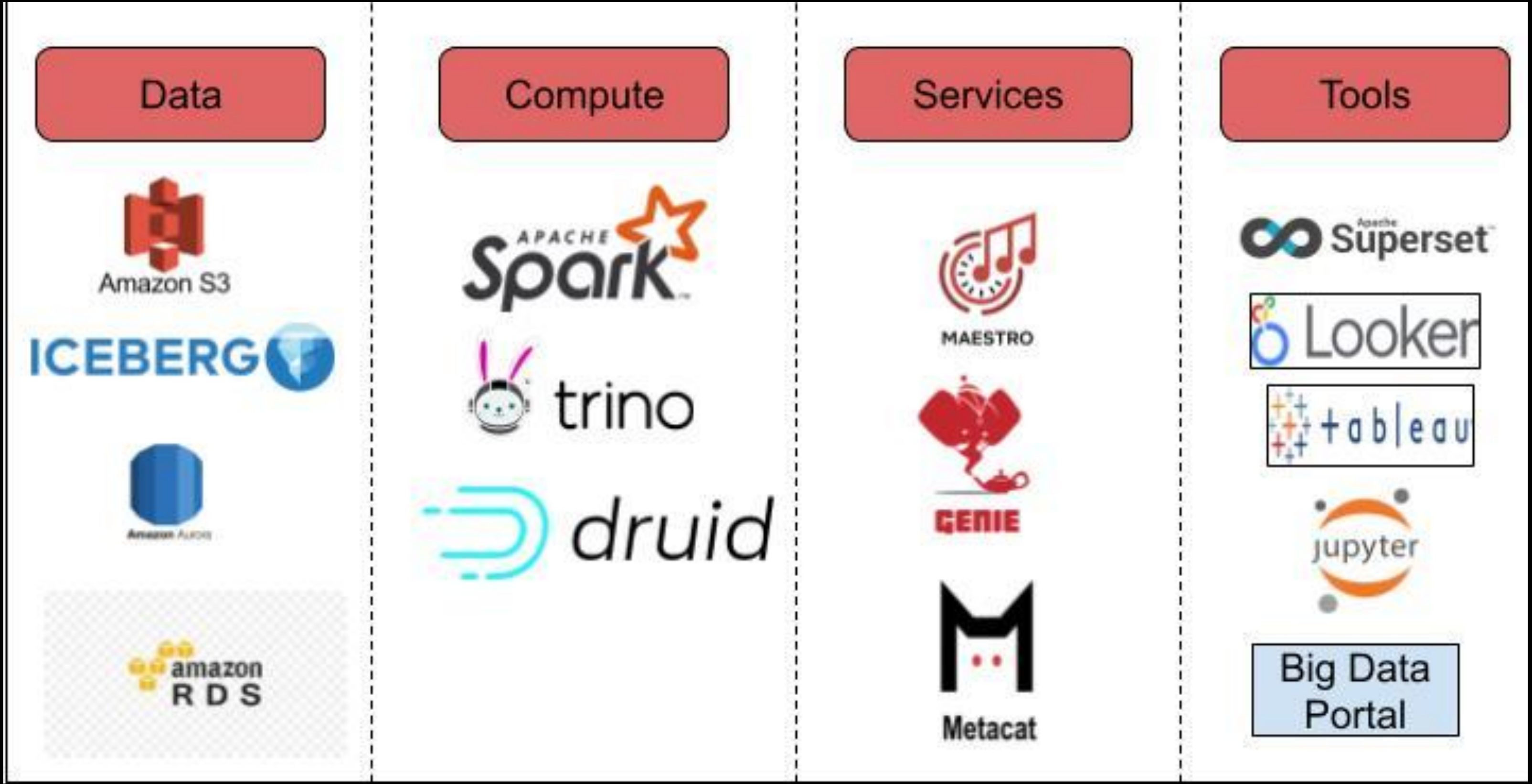
Alagappan Maruthappan

Trino Summit - 2024

Agenda

- Data Platform Architecture
- Iceberg @ Netflix
- Trino Architecture
- Internal Trino Features
- Trino-Iceberg connector
- Future

Data Platform Architecture (Analytics)



Data	Compute	Services	Tools
Amazon S3	APACHE Spark	MAESTRO	Apache Superset
ICEBERG	trino	GENIE	Looker
Amazon Aurora	druid	Metacat	tableau
amazon RDS			jupyter
			Big Data Portal

Iceberg @ Netflix

1+EB
Total Warehouse Size

3m+
Iceberg Tables

99.5%+
Iceberg adoption

36PB
Largest Iceberg Table

10+PB
Data Ingested per day

9+PB
Data Deleted per day

2PB
Data replicated per day

600
Peak commits per second

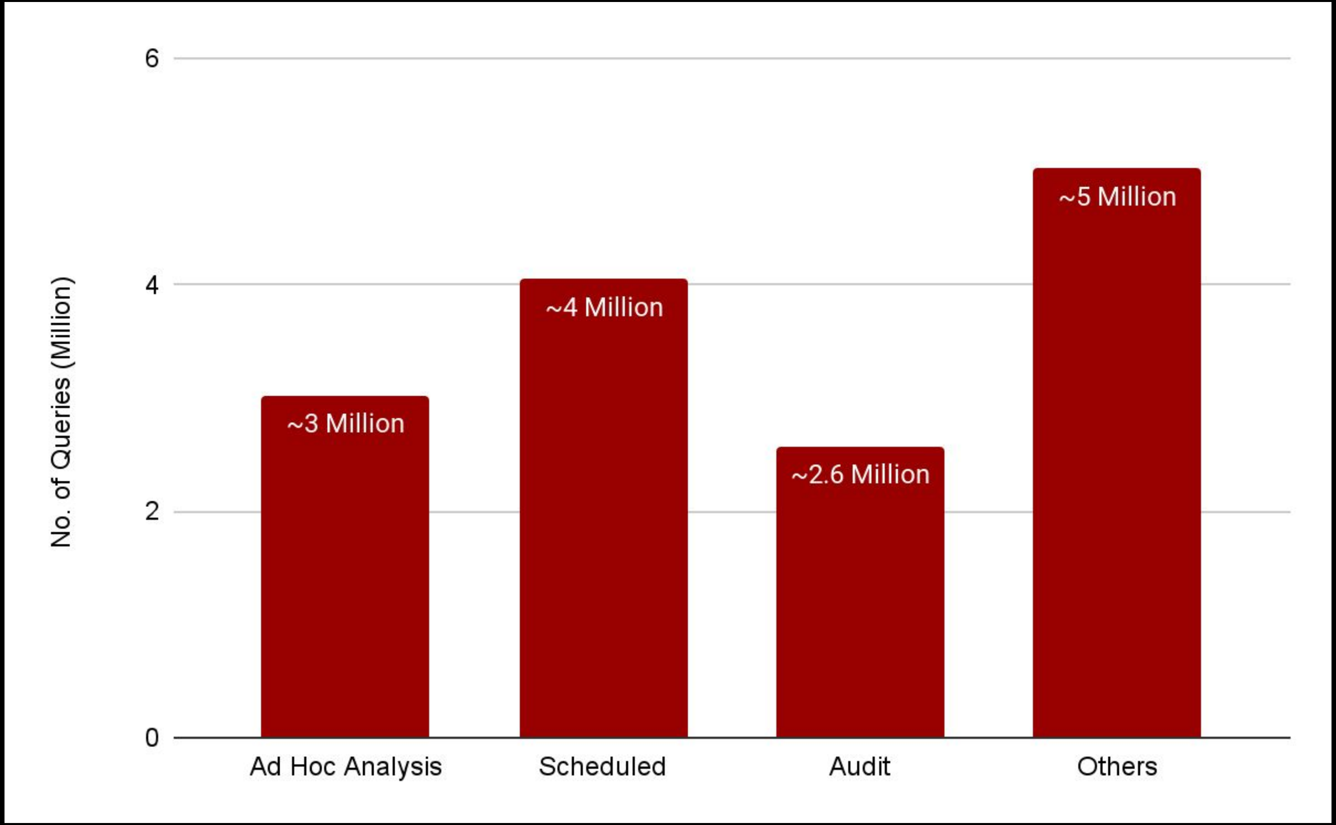
12K
Peak table loads per second

Trino @ Netflix

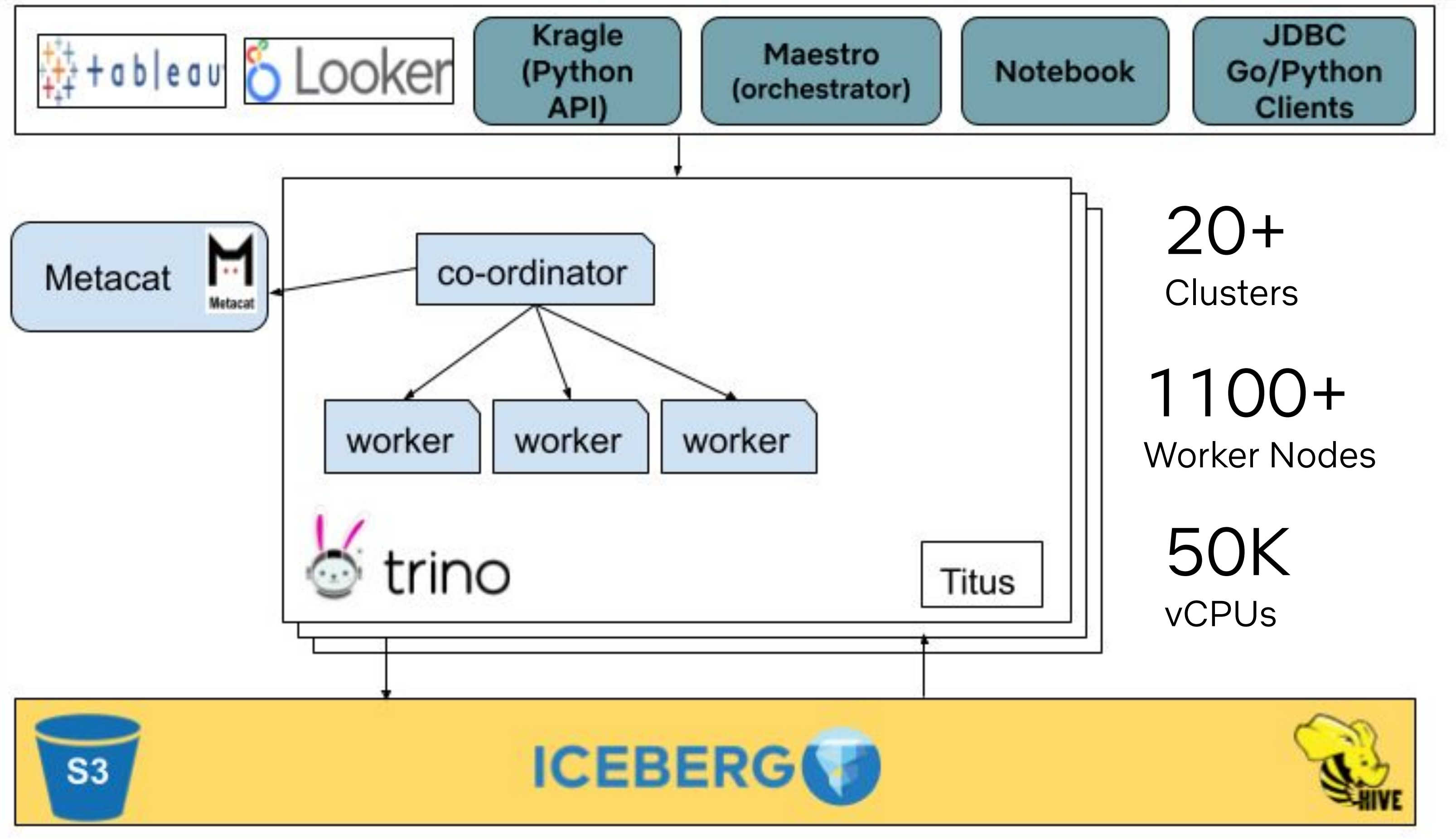
~500K
Queries/Day

~15 Million
Queries
[November, 2024]

2500+
Unique Users



Trino Architecture

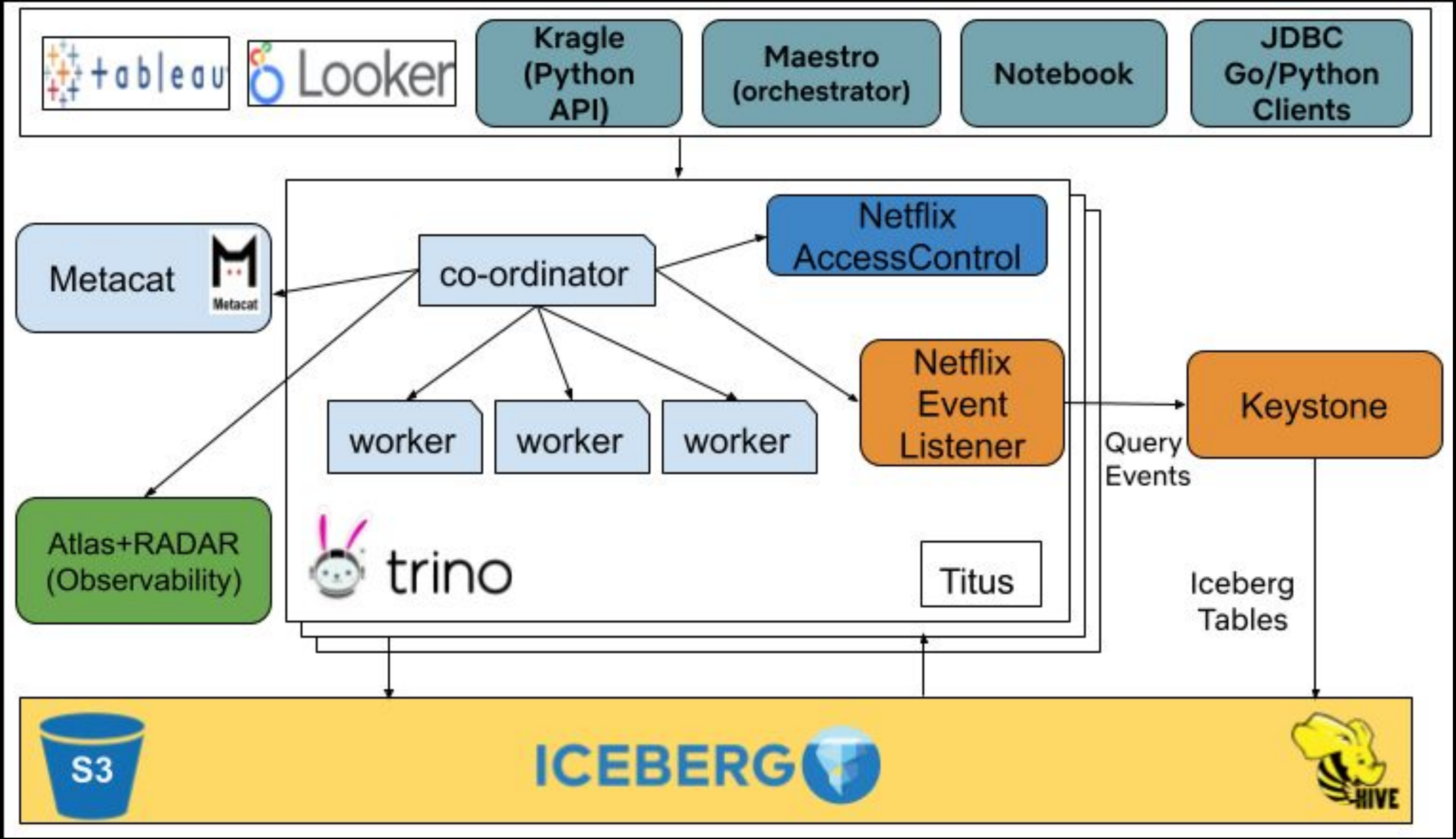


20+
Clusters

1100+
Worker Nodes

50K
vCPUs

Trino Architecture



Trino General Internal Features

- NetflixAccessControl, Metacat Connector, Netflix Event Listener plugin
- Lineage Logging
- Rich Netflix Internal UDFs - common input/output types on all supported query engines
- Aggregation pushdown for Druid connector
- Experimental:
    - HDFS based caching solution
    - Autoscaler

Trino-Iceberg Connector Features

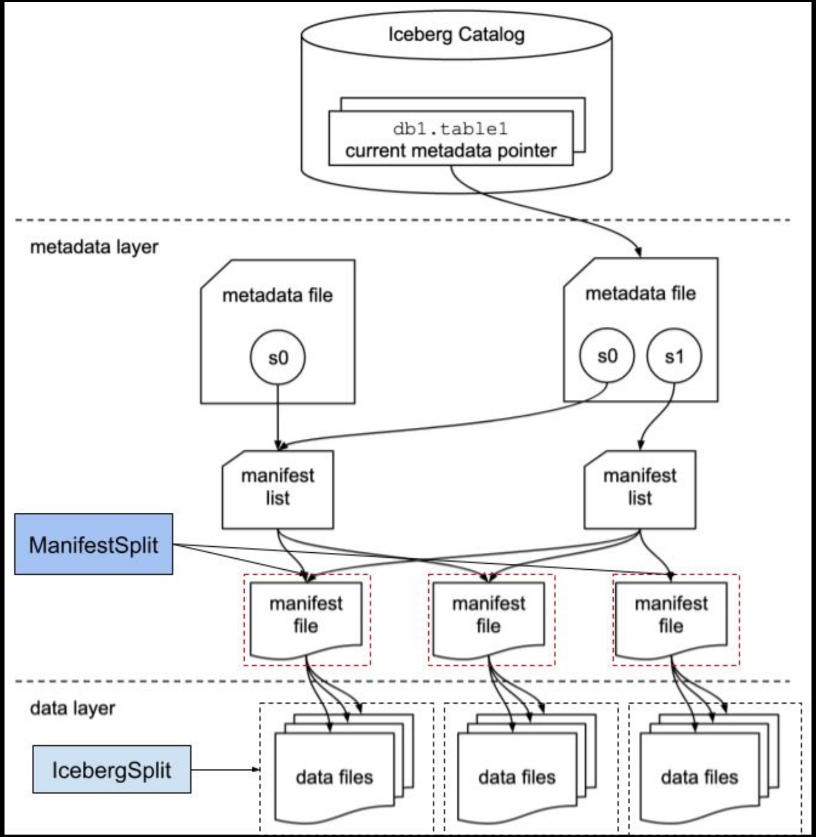
- Materialized View feature - Contributed to OSS
- ‘@’ keyword support for selecting a specific snapshot/timestamp
- Additional metadata tables - $entries, $irc_metadata
- Distributed Metadata Table Scan
- Incremental Read

Distributed Metadata Table Scan

- Metadata Tables are considered as SystemTable:
  - ALL NODES (i.e, nodes table)
  - ALL_COORDINATORS (i.e, queries table)
  - SINGLE_COORDINATOR (Iceberg Metadata Tables)
- Tables with more than 100 million files

Distributed Metadata Table Scan

- $files table:
  - Each ManifestFile as a Split
- $partitions table:
  - View on top of $files table - Group files on partition key
- 2-1000x perf improvement



Incremental Read

SELECT tables with READ options (i.e, start/end snapshot-id, start/end timestamp)

```
SELECT <columns>
FROM <table>
WITH (
 "start-snapshot-id"='1234',
 "end-snapshot-id"='5678'
);
```

```
SELECT <columns>
FROM <table>
WITH (
 "start-timestamp"='12345678',
 "end-timestamp"='987654321'
);
```

Future Work

- Trino Nightly Build - Stay as close as possible to Open Source
- Integration with Iceberg Rest Catalog (IRC)
- Trino Gateway
- Trino Caching
- ETL - Project Tardigrade

Open Source Projects

- ➔ Maestro: https://github.com/Netflix/maestro
- ➔ Genie: https://netflix.github.io/genie/
- ➔ Atlas DB: https://github.com/Netflix/atlas
- ➔ Metacat: https://github.com/Netflix/metacat

We’re hiring!

netflix.com/jobs

Thank You

Q&A...