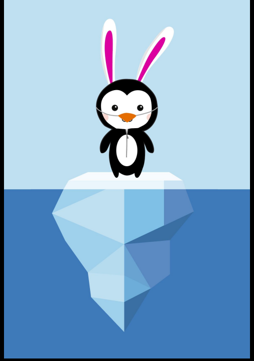
Rewriting History

Migrating petabytes of data to Apache Iceberg using Trino



Marc Laforet

Senior Data Engineer
@ Shopify

whoami

- 📍 Toronto, Canada
- 📊 Worked in data space for ~ 6 years
- 🏊 Swim in my free time

Agenda

1 Problem statement

2 Why Iceberg

3 Migration strategies

4 Trino’s Iceberg integration

5 Benchmarks

6 Reflections

Agenda

1 Problem statement

Why Iceberg

Migration strategy

Trinos Iceberg integration

Benchmarks

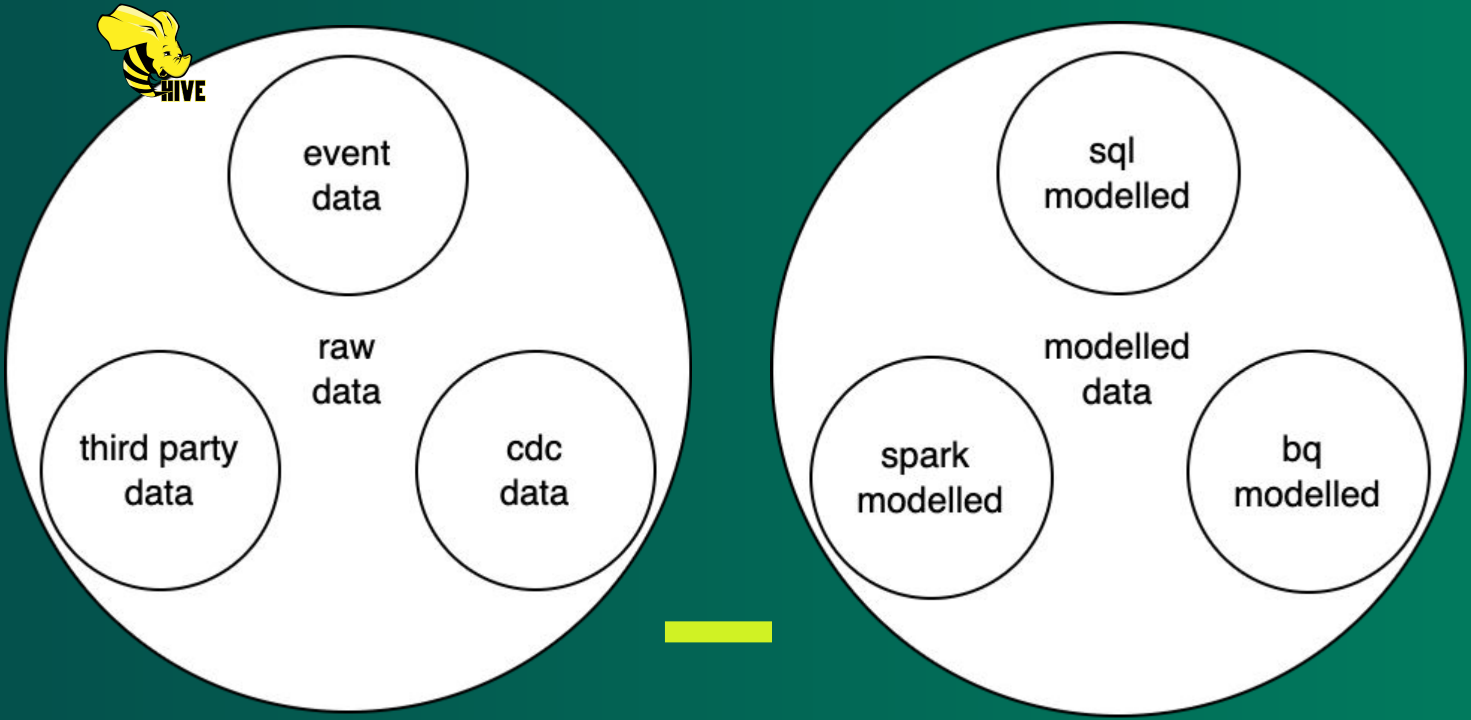
Reflections

Lakehouse Team

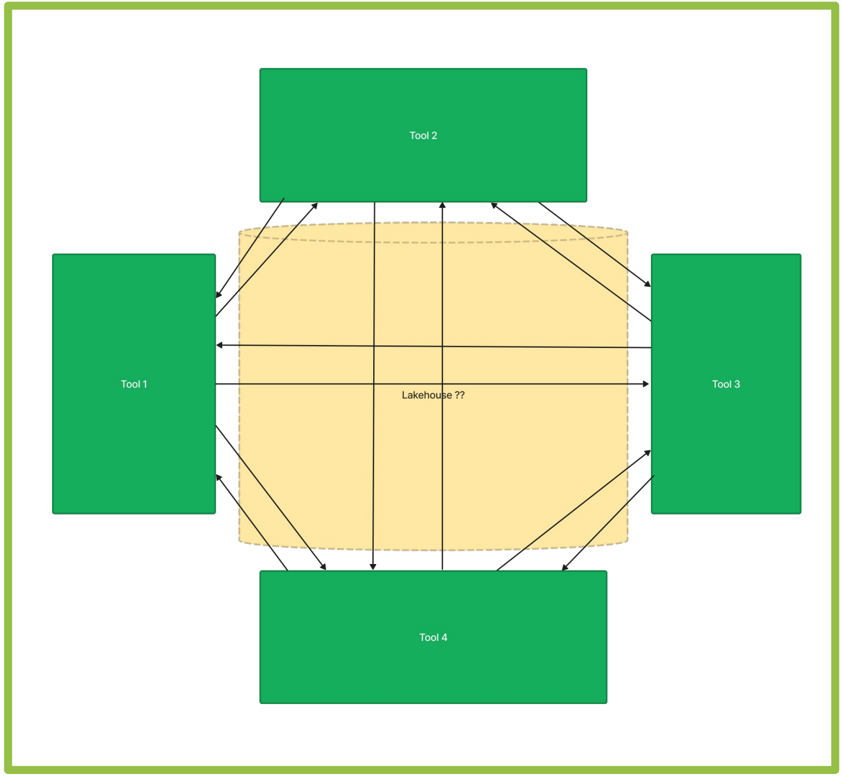
Ingest 1st party data

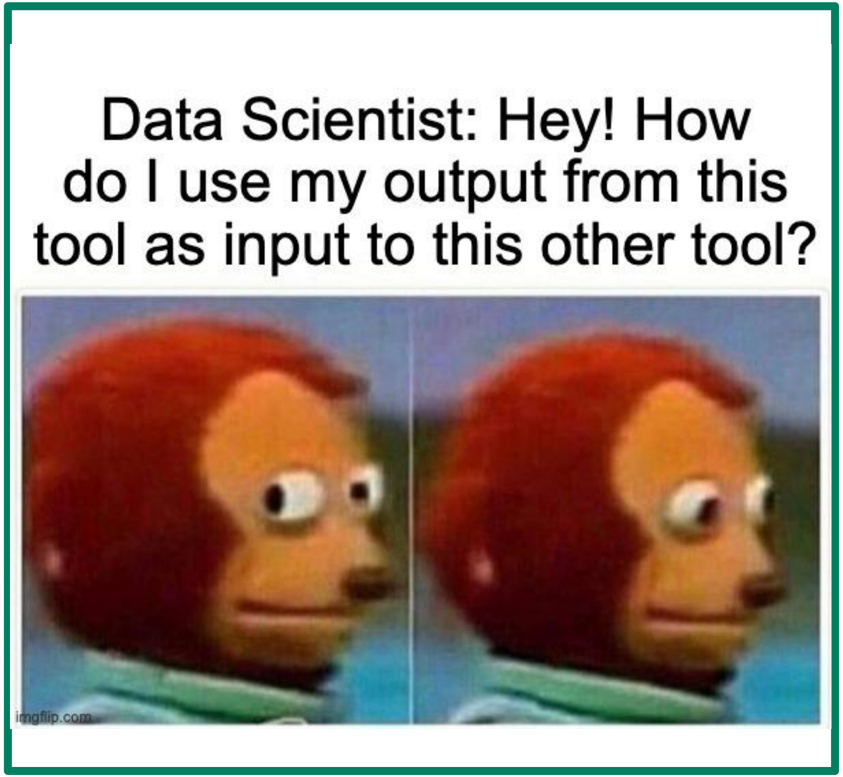
Storage and retrieval of 1st party data

Lakehouse Datasets



Lakehouse Architecture

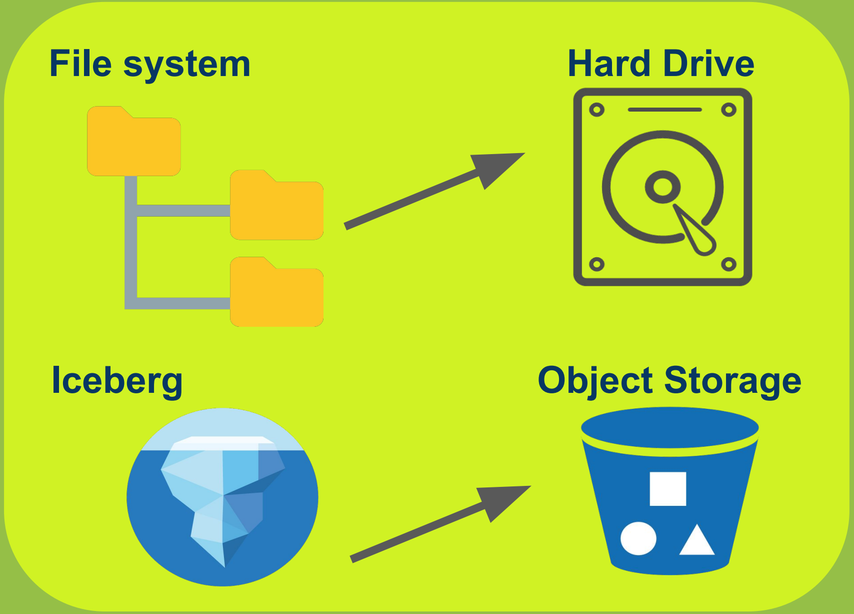



Data Scientist: Hey! How do I use my output from this tool as input to this other tool?
imgflip.com

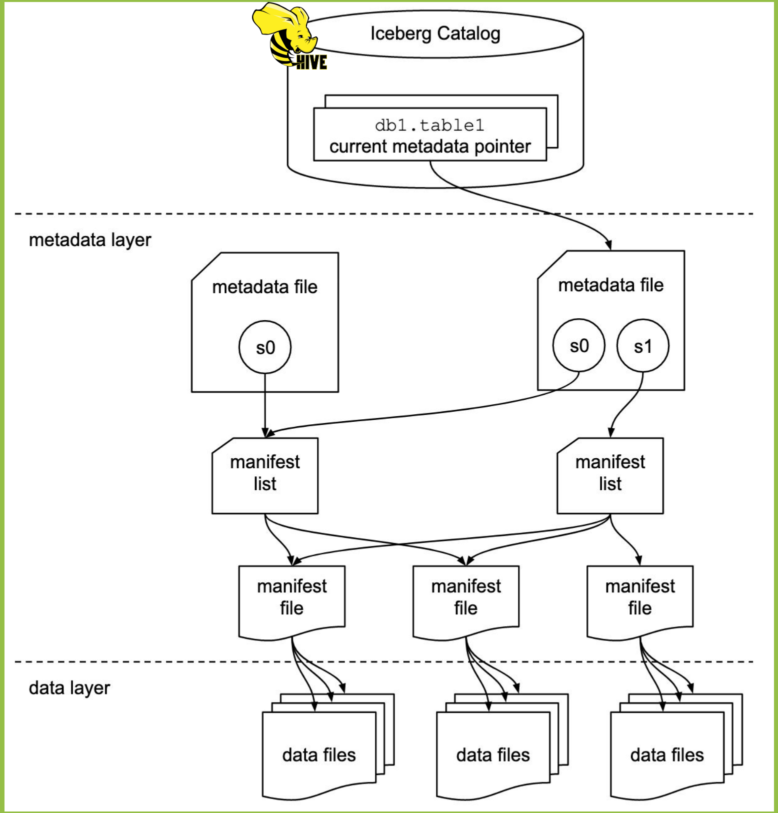
ICEBERG
imgflip.com

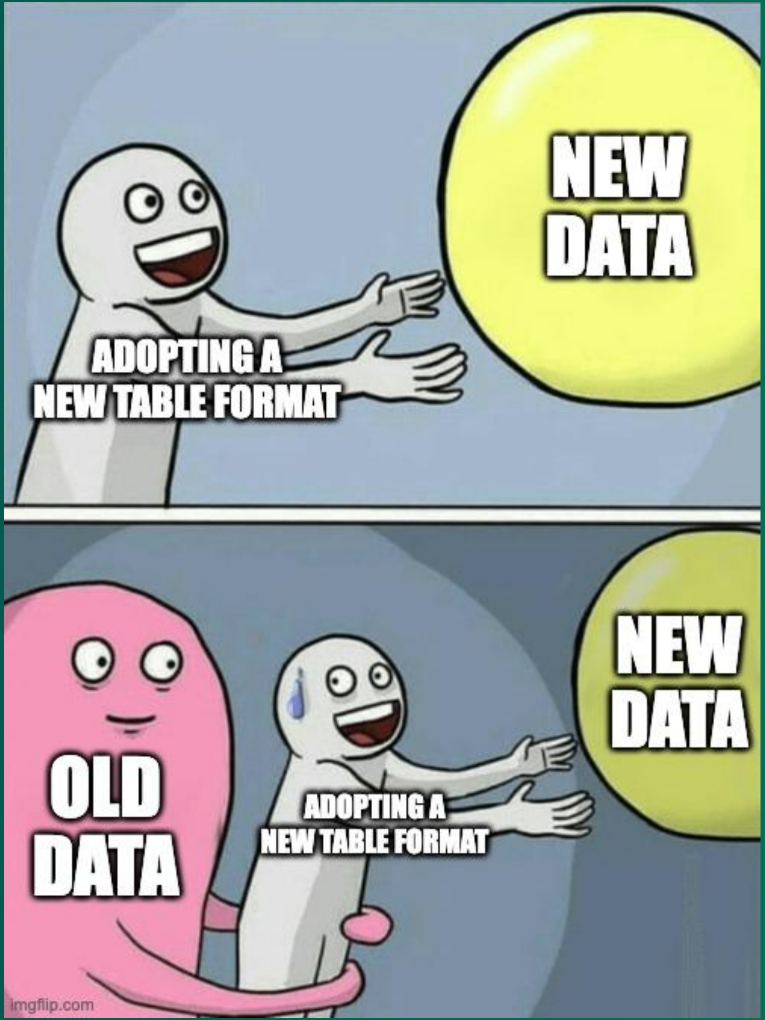
Open Table Formats

Libraries that keep track of files making up a table

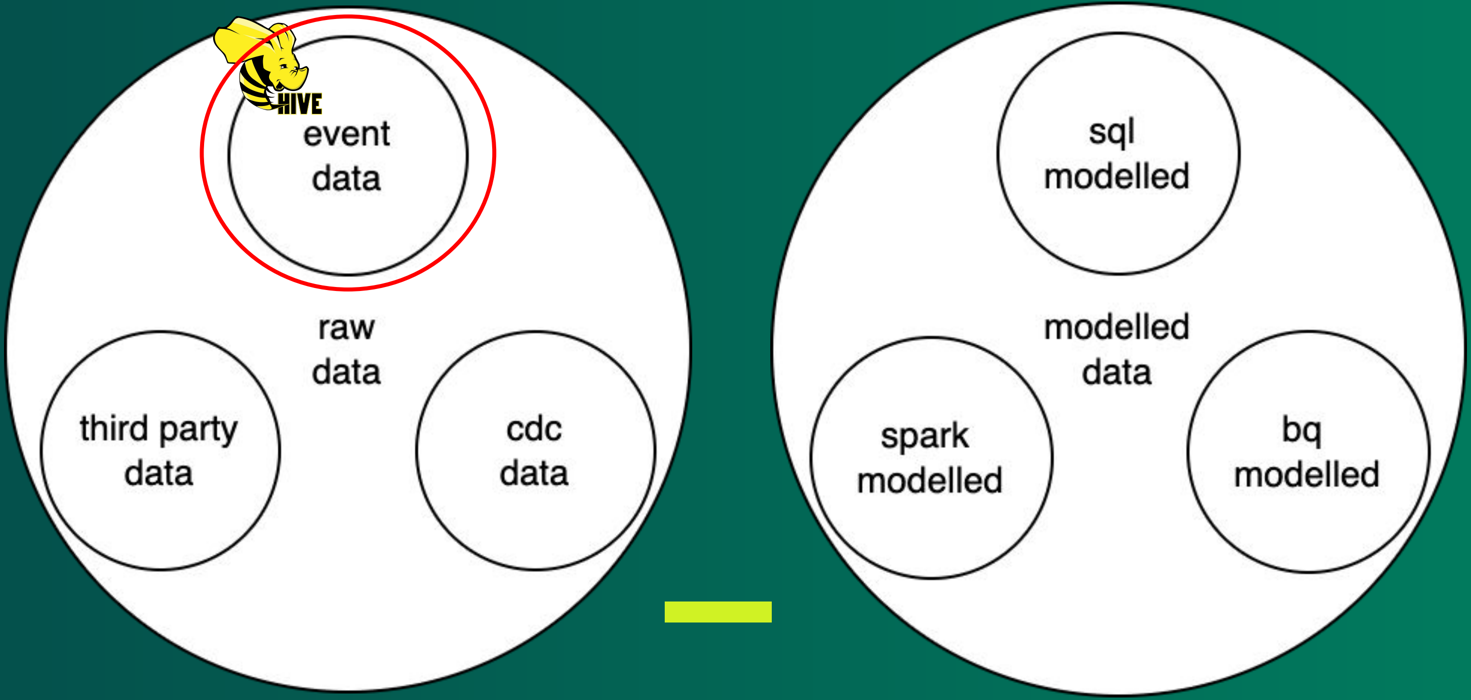


Iceberg


Iceberg Catalog
HIVE
db1.table1
current metadata pointer
metadata layer
metadata file
s0
metadata file
s0
s1
manifest list
manifest list
manifest file
manifest file
manifest file
data layer
data files
data files
data files


NEW DATA
ADOPTING A NEW TABLE FORMAT
OLD DATA
ADOPTING A NEW TABLE FORMAT
NEW DATA
imgflip.com

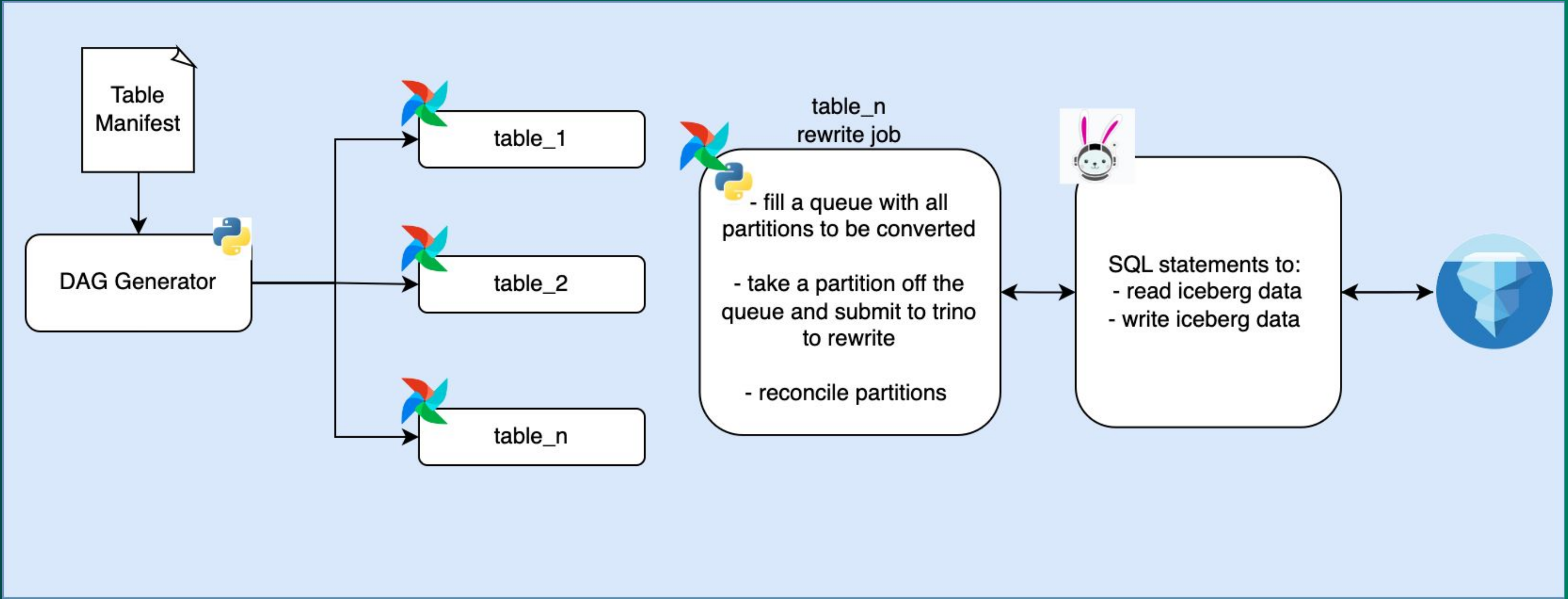
Lakehouse Datasets



Migration Strategies

Current State	Strategy
Data is not in a supported file format	Rewrite everything
Data is in a supported file format but you want to change the partioning strategy	Rewrite everything
Data is in a supported file format and properly partitioned	Generate manifest files only

System Design



- fill a queue with all partitions to be converted

- take a partition off the queue and submit to trino to rewrite

- reconcile partitions

SQL statements to:
- read iceberg data
- write iceberg data

Agenda

1 Problem statement

2 Why Iceberg

3 Migration strategy

4 Trinos Iceberg integration

Benchmarks

Reflections

Iceberg tables partitioned on timestamp with time zone

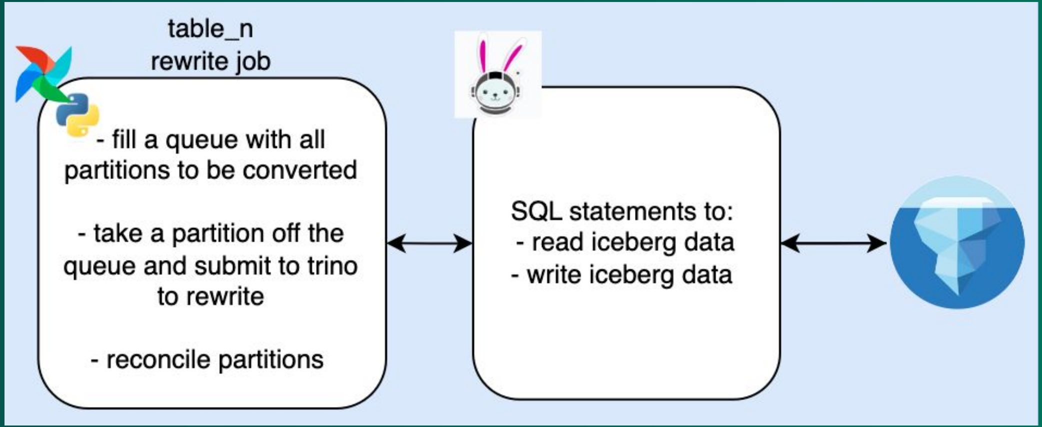
Unknown type: class io.trino.spi.type.LongTimestampWithTimeZone java.lang.IllegalArgumentException:

Unknown type: class io.trino.spi.type.LongTimestampWithTimeZone

PR-9757

Return timestamp time zone partitions in UTC

Scale Iceberg Writes



table_n
rewrite job

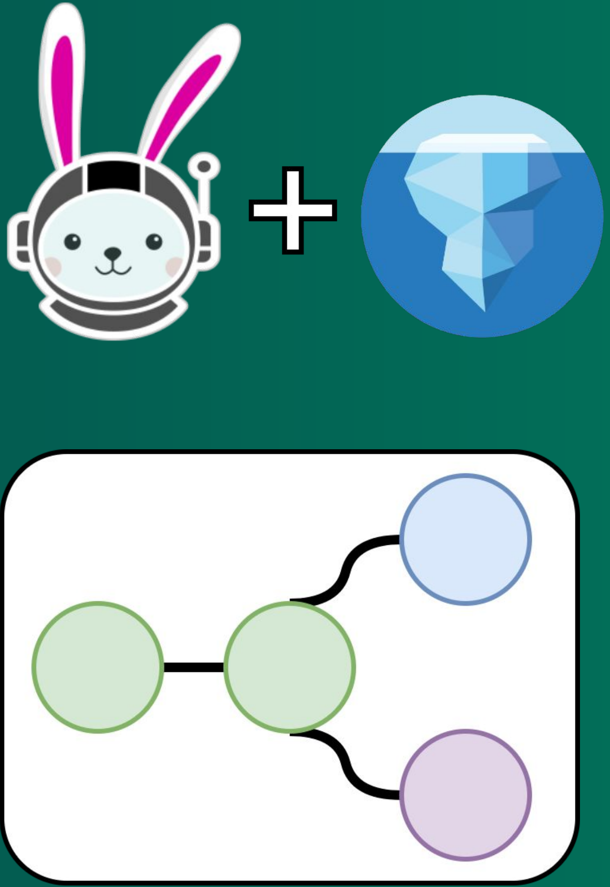
- fill a queue with all partitions to be converted

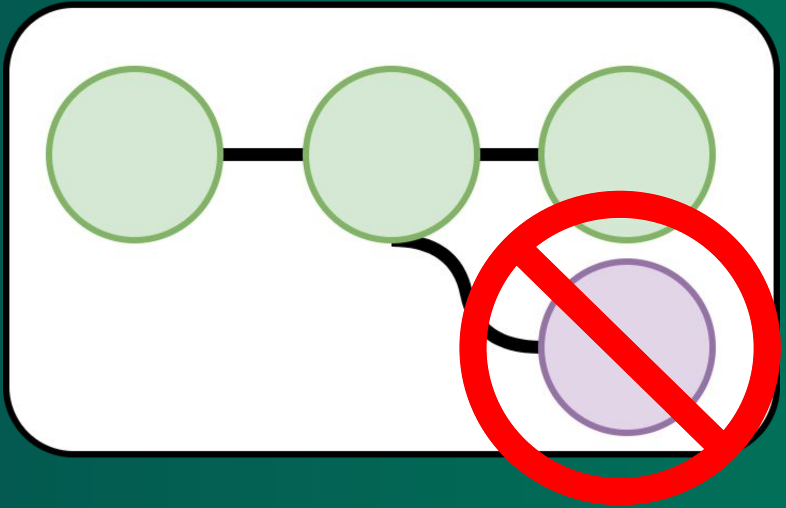
- take a partition off the queue and submit to trino to rewrite

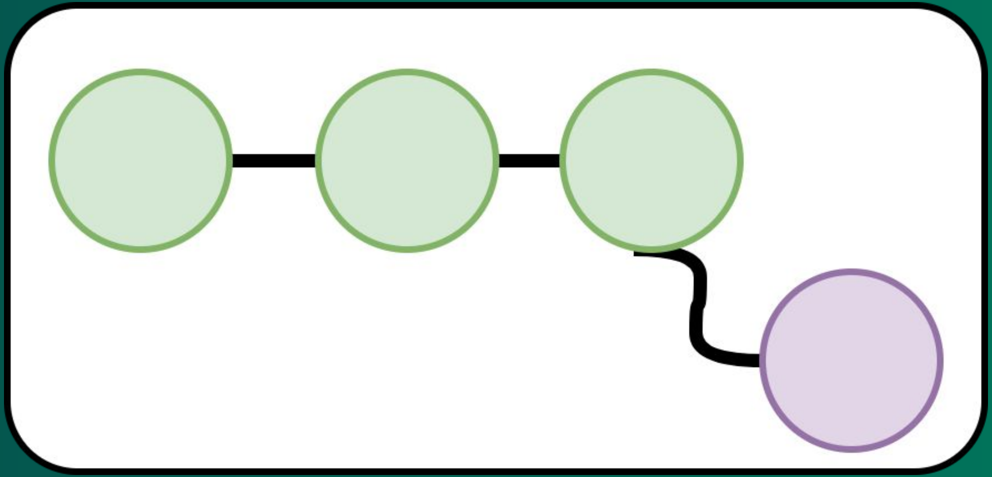
- reconcile partitions

SQL statements to:
- read iceberg data
- write iceberg data

```
org.apache.iceberg.exceptions.CommitFailedException:
Metadata location
[<bucket-name>/<namespace>/<table-name>/metadata/203942-8936d
432-7ef0-4bfe-8098-bf9ca06580fe.metadata.json] is not same as table
metadata location
[<bucket-name>/<namespace>/<table-name>/metadata/203943-cf4a61
94-1ee7-4483-846c-f495ab237e78.metadata.json] for
<namespace>.<table_name>
```







PR-11886

Throw proper error to make use of iceberg’s commit retries

Missing File Problem

```
org.apache.iceberg.exceptions.NotFoundException:
Failed to open input stream for file:
[<bucket-name>/<namespace>/<table-name>/metadata/snap-8501054
657342128291-2-8eeb7ee8-54c7-4e08-a2db-f5cbdd855835.avro
```

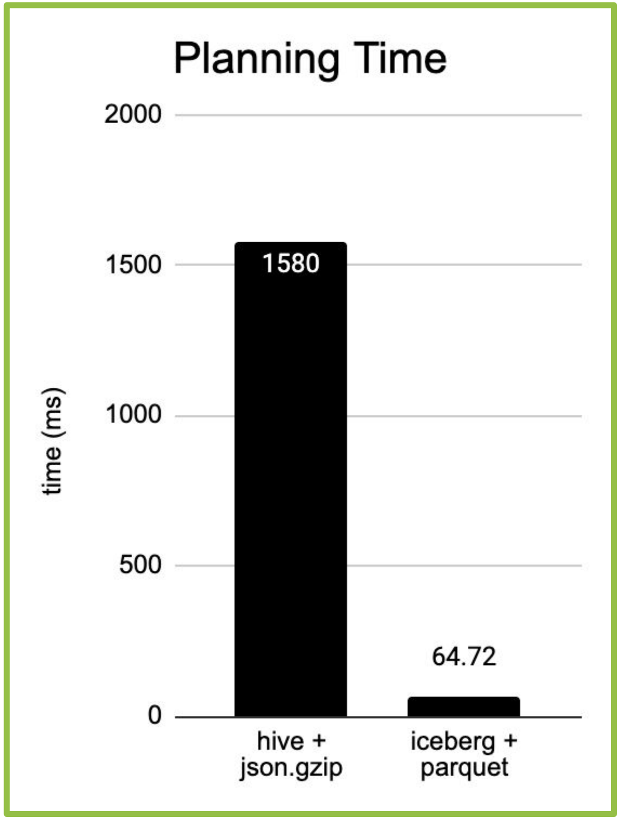
PR-14118

Avoid pro-active file cleanup on HMS time-out

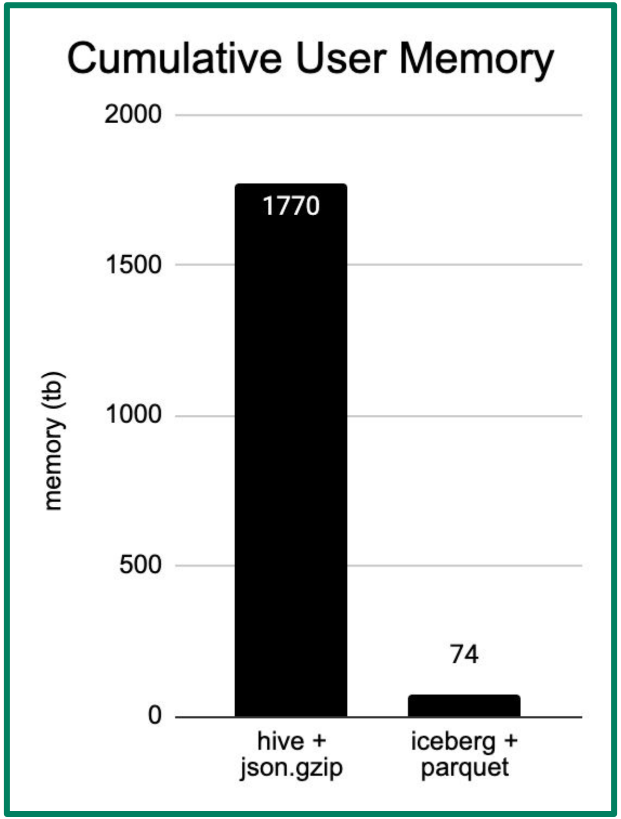
Agenda

1	Problem statement
2	Why Iceberg
3	Migration strategy
4	Trinos Iceberg integration
5	Benchmarks
	Reflections

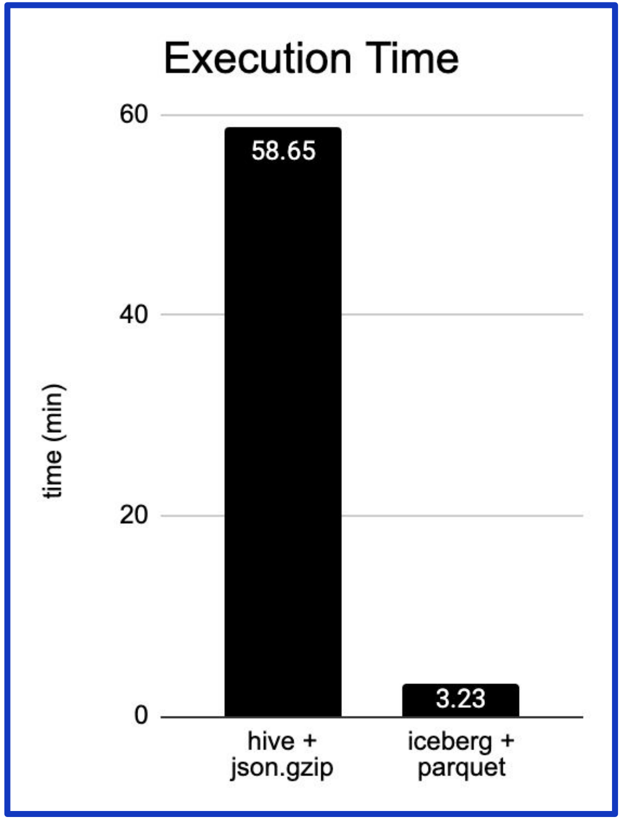
Benchmarks



Planning Time

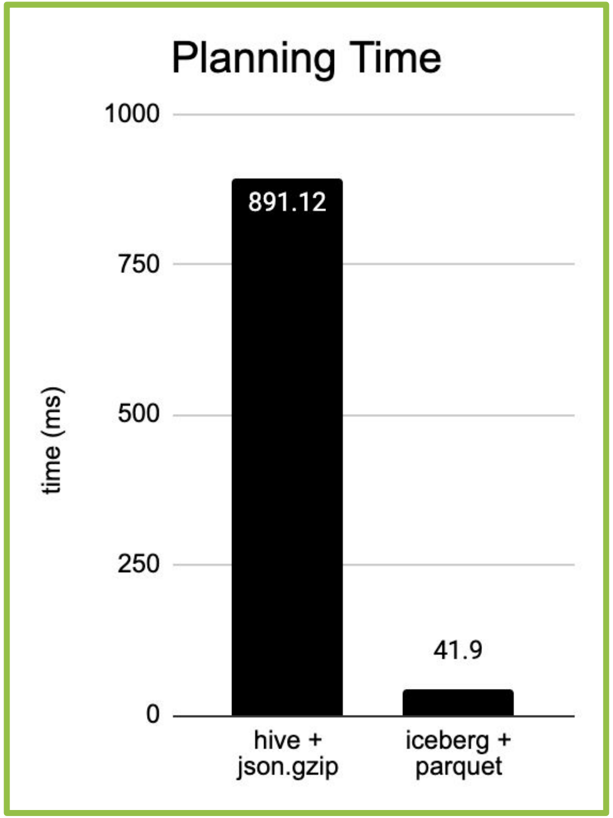


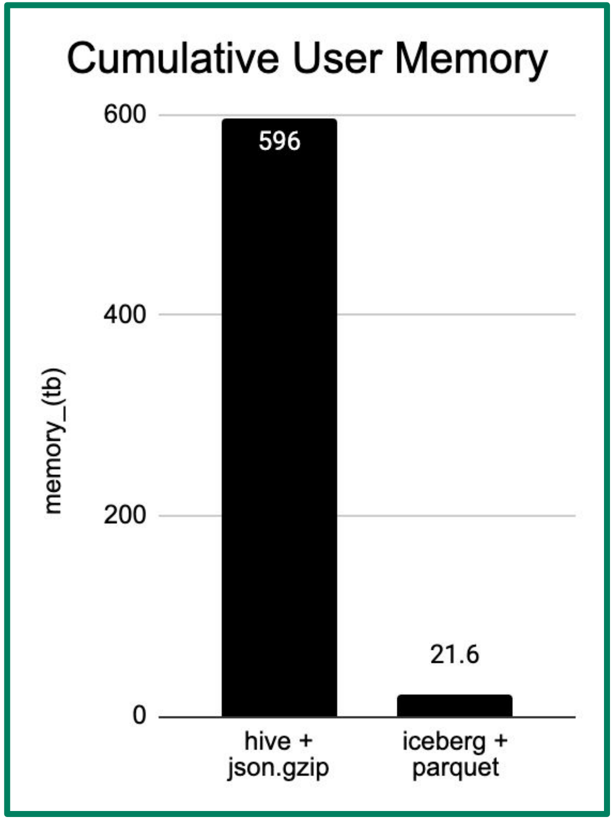
Cumulative User Memory

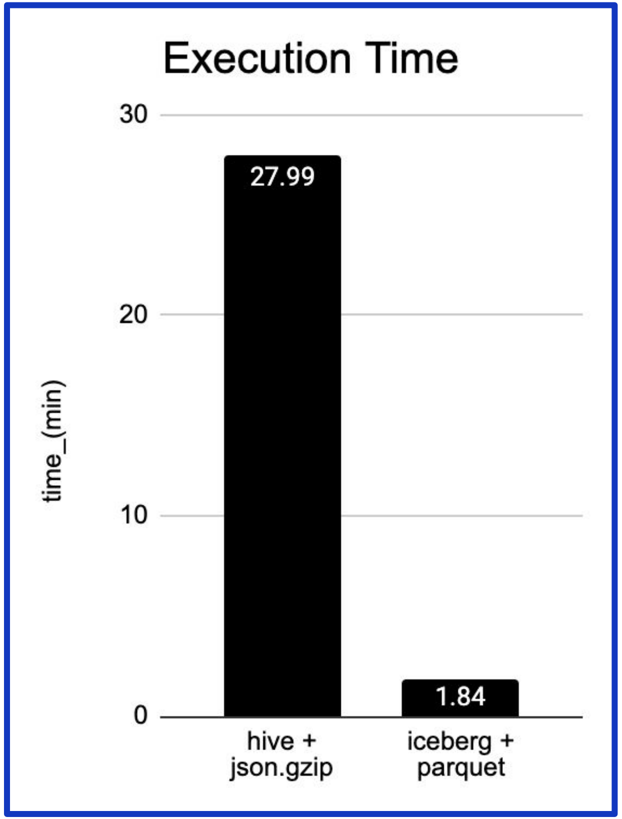


Execution Time

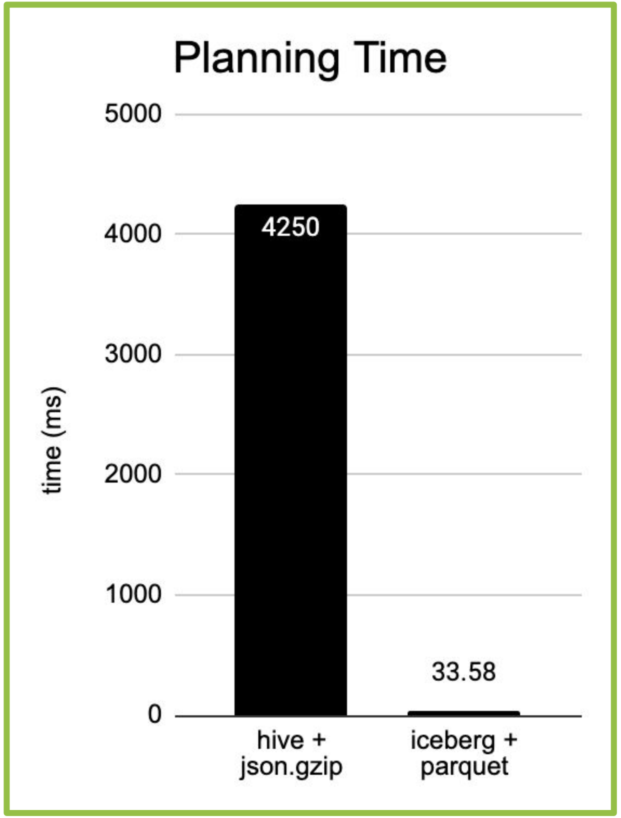
Benchmarks

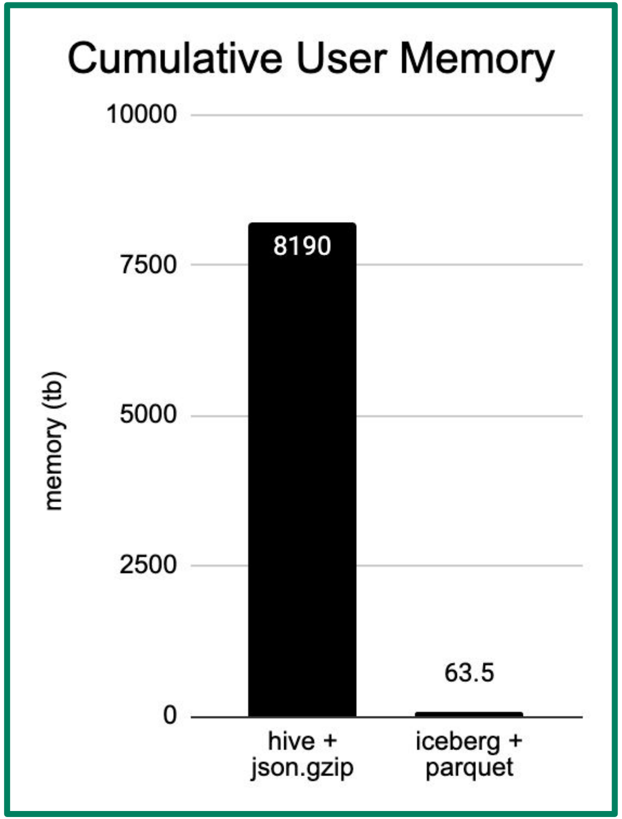

Planning Time

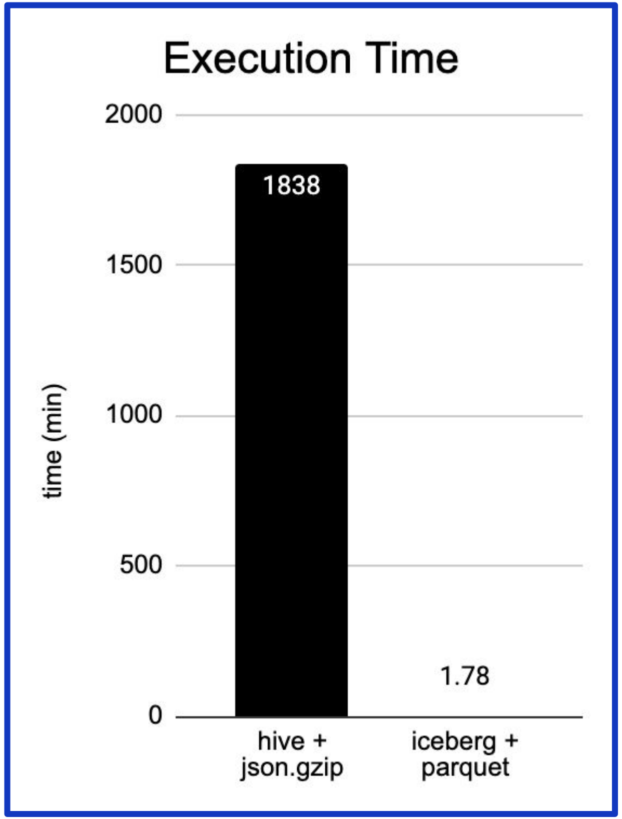

Cumulative User Memory

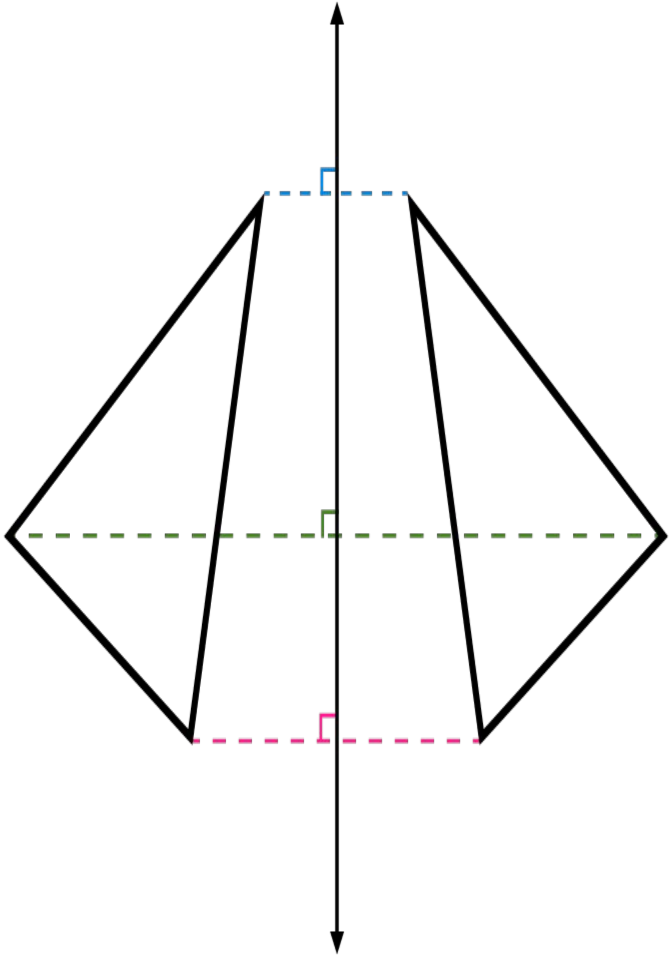

Execution Time

Benchmarks


Planning Time
time (ms)
5000
4000
3000
2000
1000
0
4250
33.58
hive + json.gzip
iceberg + parquet


Cumulative User Memory
memory (tb)
10000
7500
5000
2500
0
8190
63.5
hive + json.gzip
iceberg + parquet


Execution Time
time (min)
2000
1500
1000
500
0
1838
1.78
hive + json.gzip
iceberg + parquet



Reflections & Advice

- Using a modern table and file format can really improve query performance
- Huge benefit using open source software
- Write new and old data to the same table
- Construct a migration priority list
- Airflow offered nice out of the box task scheduling but wasn’t great at managing long-running tasks

Project Contributors

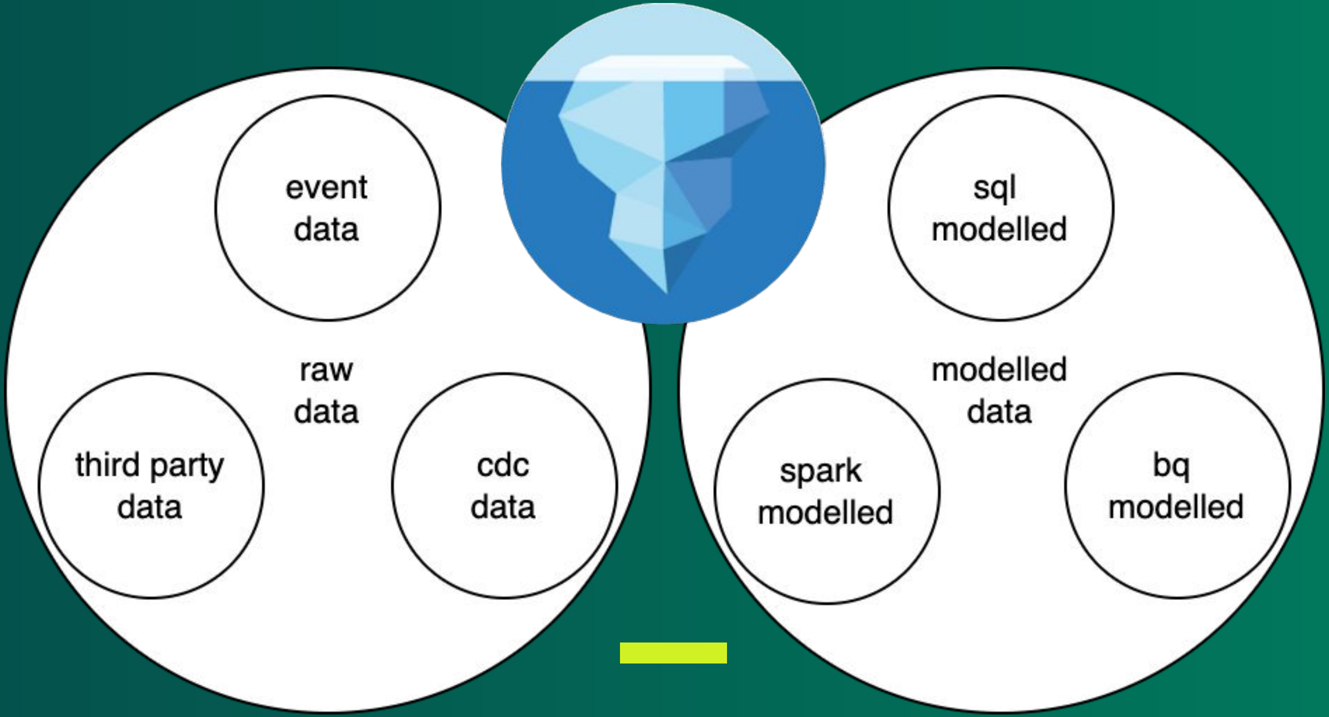
Sam Wheating

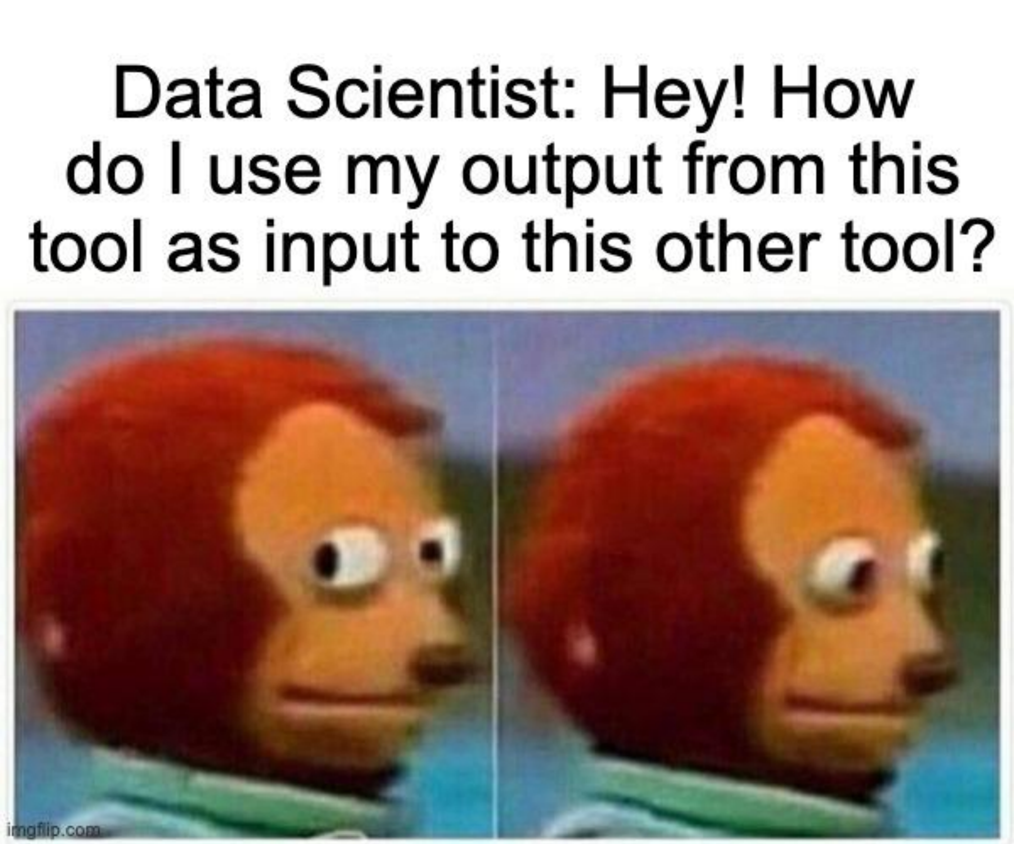
Victoria Bukta

Thank you

POWERED BY SHOPIFY POWERED BY SHOPIFY

The Iceberg Lakehouse





Data Scientist: Hey! How do I use my output from this tool as input to this other tool?

How many table formats are in use?

Hive Tables

- Standard hive tables
- Synthetic partition key column
- Columnar data being stored in json.gzip format

Proprietary Table Formats

- Datasets migrated into bigquery

Custom Table Formats

- Teams invented their own table formats for their tool
- No ACID properties
- Full drop every time

Dynamic Writer Scaling

https://github.com/trinodb/trino/pull/10614

File pruning using table statistics

https://github.com/trinodb/trino/pull/9326
https://github.com/trinodb/trino/issues/4115

Stitching with a view

https://github.com/trinodb/trino/pull/8621
https://github.com/trinodb/trino/pull/8540

Utilities

https://github.com/trinodb/trino/pull/10480 (vicky)
https://github.com/trinodb/trino/pull/10459 (austen)

Monitoring

https://github.com/trinodb/trino/pull/12026 - update
https://github.com/trinodb/trino/issues/7933 - merge