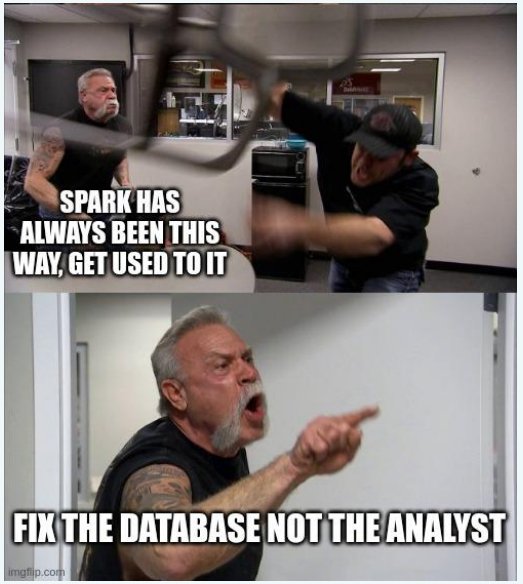
Unsolved Challenges in Data Infrastructure

Ryan Blue
Subsurface Winter 2022

Disclaimer:

There are more questions than answers in here

6 years ago . . .



Tabular

5-year challenges

- Smarter processing engines
  - CBO data, join optimizations
  - Result set caching, materialized views
- Reduce manual maintenance
  - Librarian services
  - Declarative instead of imperative

. . . as remembered in Nov 2019

Tabular

Problem whack-a-mole

- Unsafe operations everywhere
  - Writing to multiple partitions
  - Renaming a column
- Object stores cause headaches
  - Eventual consistency
  - Performance problems (latency)
  - Output committers are insufficient
- Endless scale challenges

. . . as remembered in Nov 2019

Tabular

This is why we built Iceberg

Tabular

Then you might ask . . .

Tabular

Tabular

Problem whack-a-mole

- Unsafe operations everywhere — Fixed!
  - Writing to multiple partitions
  - Renaming a column
- Object stores cause headaches — Fixed!
  - Eventual consistency
  - Performance problems (latency)
  - Output committers are insufficient
- Endless scale challenges — Fixed! *or at least delayed*

Tabular

5-year challenges

- Smarter processing engines
  - CBO data, join optimizations
  - Result set caching, materialized views

Work in progress across engines

- Reduce manual maintenance
  - Librarian services
  - Declarative instead of imperative

Work in progress across engines

Tabular

5-year challenges

- Smarter processing engines
  - CBO data, join optimizations
  - Result set caching, materialized views

Work in progress **across engines**

- Reduce manual maintenance
  - Librarian services
  - Declarative instead of imperative

Work in progress **across engines**

Tabular

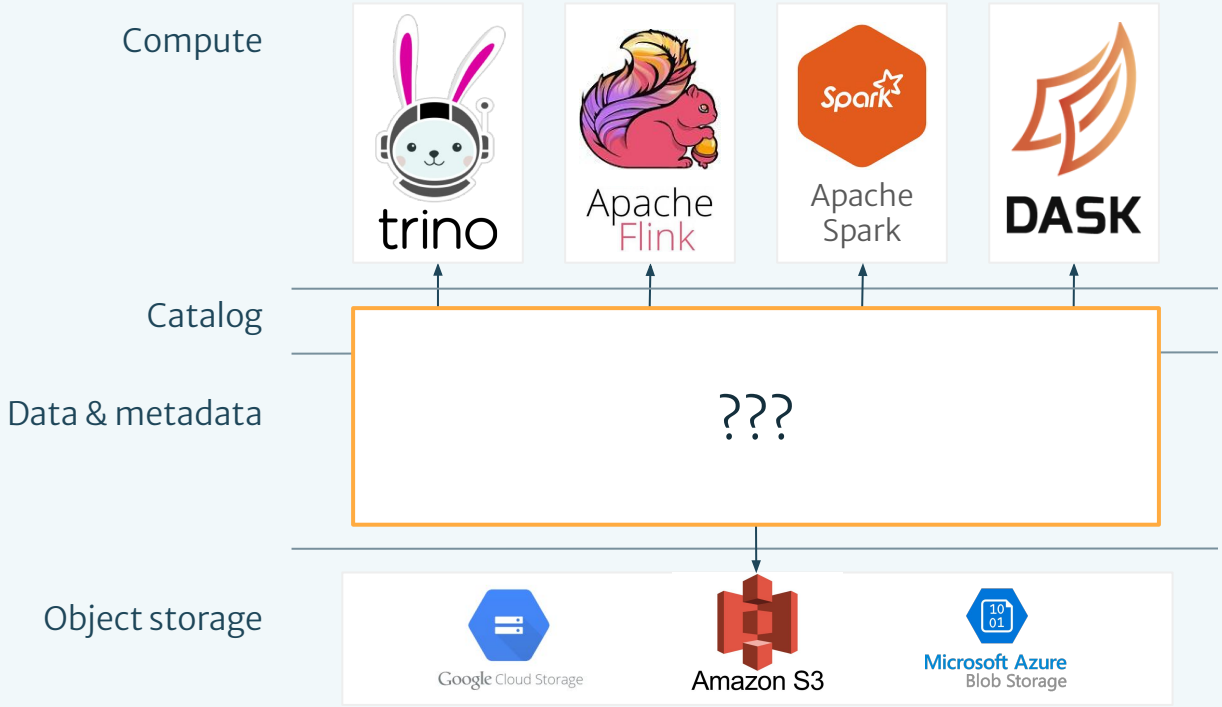
We've moved the goal posts

Tabular

Or maybe, we can see the challenge clearly now

Tabular

The multi-engine platform is the next challenge



Tabular

What's already in?

- Central table store
  - Share data instead of copying
- Iceberg
  - Open standard for huge tables
  - SQL abstraction *and* behavior
  - Data warehouse fundamentals
- Data services
  - Automate tasks, don't make humans babysit
- Declarative data engineering
  - Vastly different engines require better ways to work

Tabular

New 5-year challenges

- Access control
  - Consistent authorization policy
  - Enforced across engines
- Portable compute
  - Multi-engine is the new normal
  - SQL translation – Substrait
- New catalogs
  - Purpose built or generic?
  - Warehouse catalogs and business catalogs
- Stop losing structure
  - Most “unstructured” data didn’t start that way
  - Consistent schemas for streams

Tabular

And we need a name for this box

Tabular

Thanks!

Tabular