Unsolved Challenges in Data Infrastructure



Ryan Blue Subsurface Winter 2022

Disclaimer: There are more questions than answers in here



6 years ago...





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



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



This is why we built lceberg





Then you might ask . . .







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

Fixed!

Fixed!

Fixed! or at least delayed



5-year challenges

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

Work in progress across engines

Work in progress across engines



5-year challenges

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

Work in progress across engines

Work in progress across engines



We've moved the goal posts





Or maybe, we can see the challenge clearly now



The multi-engine platform is the next challenge



What's already in?

- Central table store
 - Share data instead of copying

- Data services
 - Automate tasks, don't make humans babysit

• Iceberg

- Open standard for huge tables
- SQL abstraction *and* behavior
- Data warehouse fundamentals

- Declarative data engineering
 - Vastly different engines require better ways to work



New 5-year challenges

- Access control
 - Consistent authorization policy
 - Enforced across engines

- New catalogs
 - Purpose built or generic?
 - Warehouse catalogs and business catalogs

- Portable compute
 - Multi-engine is the new normal
 - SQL translation Substrait

- Stop losing structure
 - Most "unstructured" data didn't start that way
 - Consistent schemas for streams



And we need a name for this box





Thanks!

