ALLUXIO

Presto+Alluxio加速Iceberg数据湖访问

Beinan

Sep 11, 2022

ALLUXIO

Agenda

01

Presto & Alluxio

Presto overview and Presto + Alluxio overview

02

Alluxio & Iceberg

Alluxio and Iceberg Architecture

03

Best Practices

Data Consistency and Privacy

04

Future Work

Future work of the open-source community

01

Presto & Alluxio

Presto Overview

- **Distributed SQL Query Engine**
  - ANSI SQL on Hive data warehouse, Hudi, Iceberg, Kafka, Druid and etc.
  - Designed to be interactive
  - Access to petabytes of data

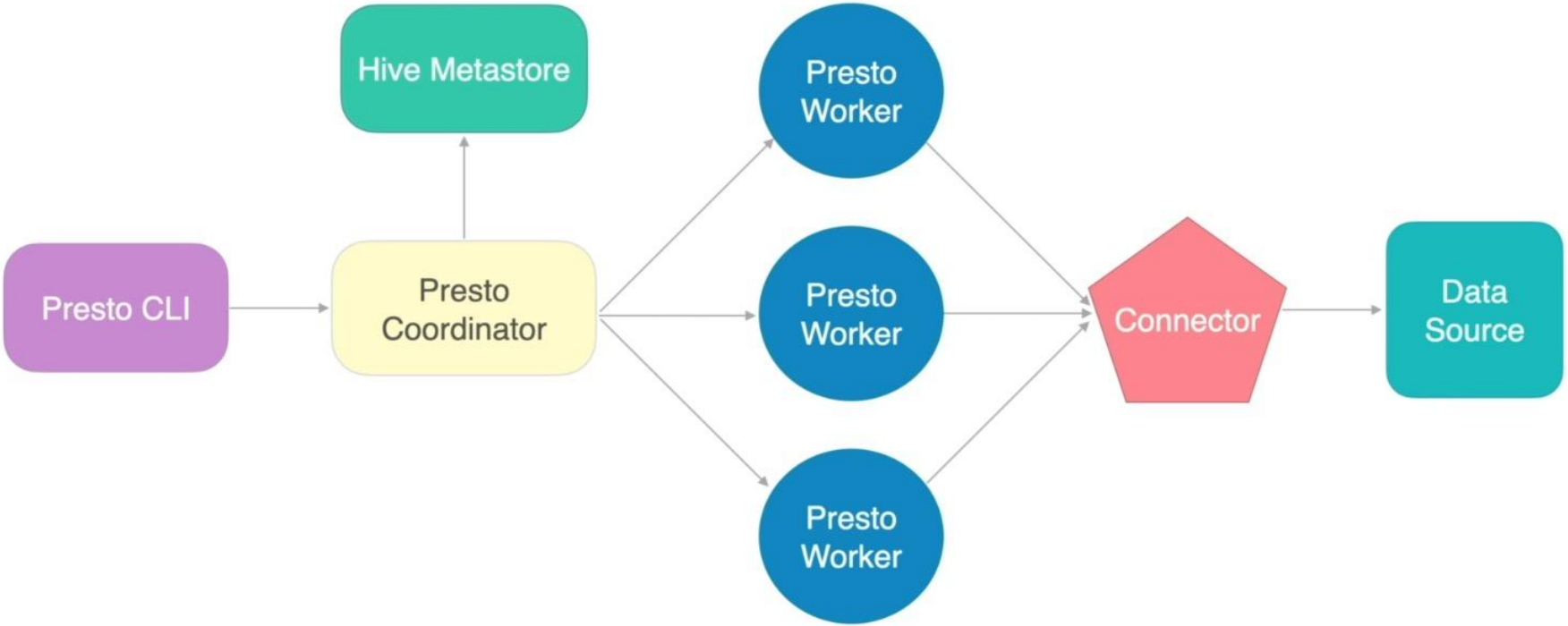
- **Open-source**
  - github.com/prestodb
  - github.com/trinodb

- **Use Cases**
  - Ad-hoc
  - BI tools
  - Dashboard
  - A/B testing
  - ETL

Presto's History

Source: [varada]

Presto Architecture



Presto + Alluxio Overview

- **Why Presto + Alluxio?**

Unified Namespace

- Alluxio is the only data source
- No code change to applications

Modernize Data Platform

- Hybrid- and multi-cloud data lake
- Data movement

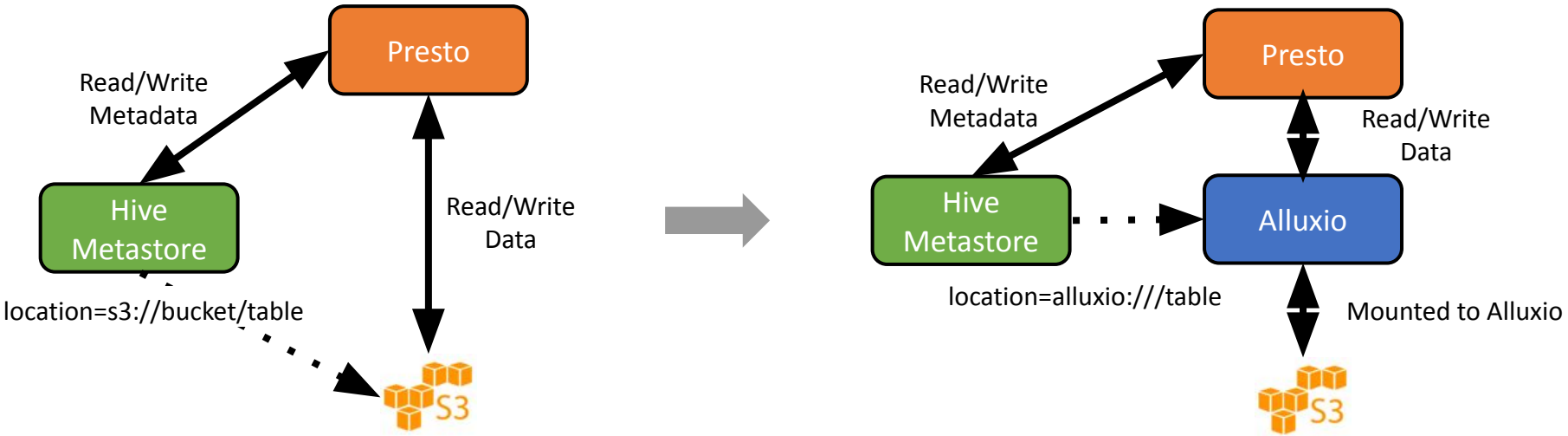
Data Sharing Between Compute

- Between compute engines
- Across entire data pipeline

*Performance

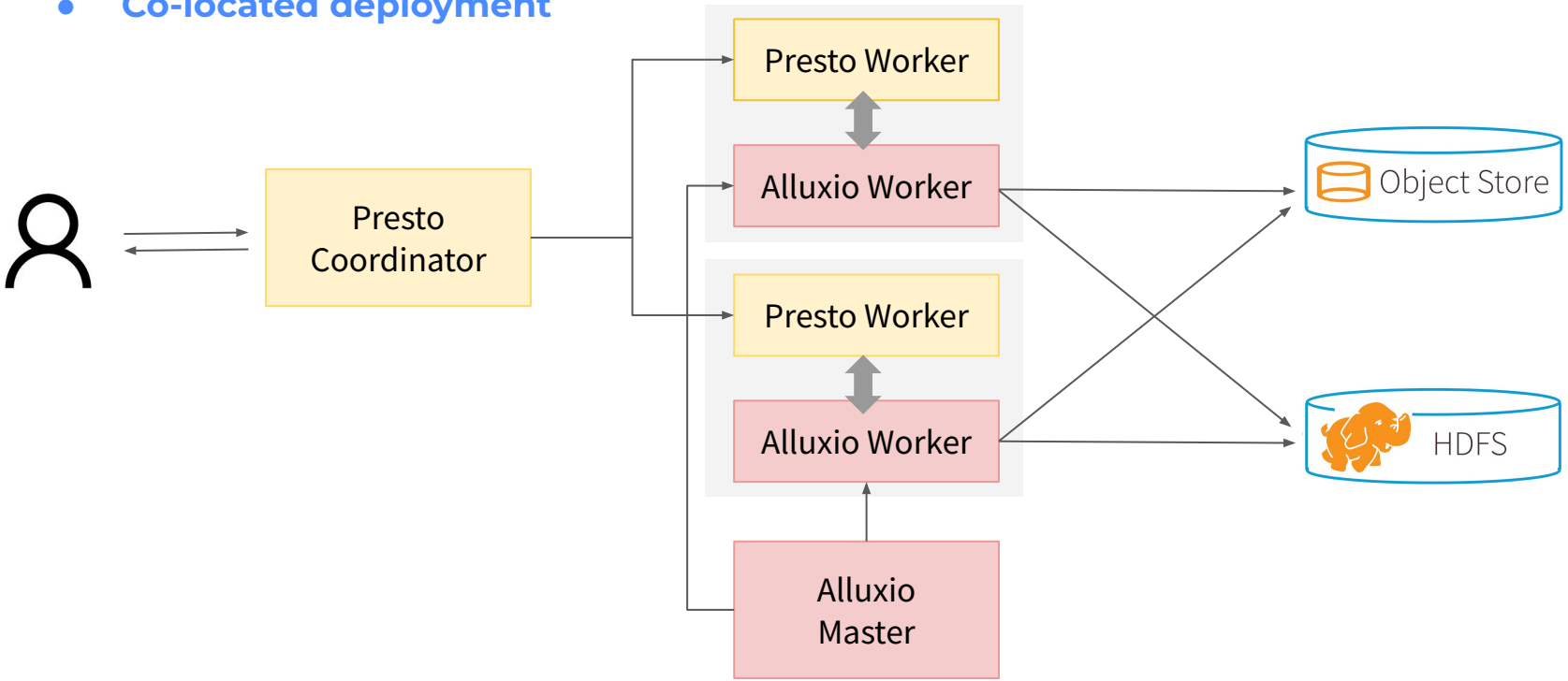
- *Accelerate queries (no guarantee, maybe yes in some queries, especially when co-located)
- Alluxio speeds up the I/O

How Presto Works with Alluxio



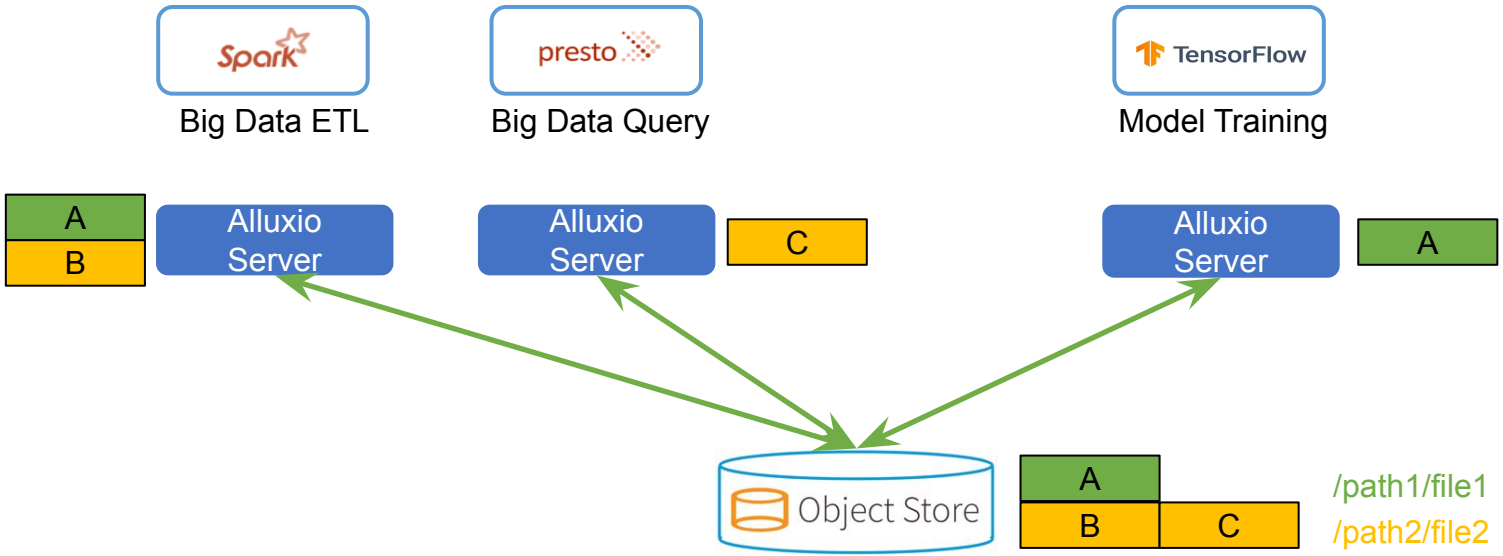
Presto + Alluxio Architecture

- **Co-located deployment**



Presto + Alluxio Architecture

- **Disaggregated deployment**



02

Alluxio & Iceberg

Apache Iceberg

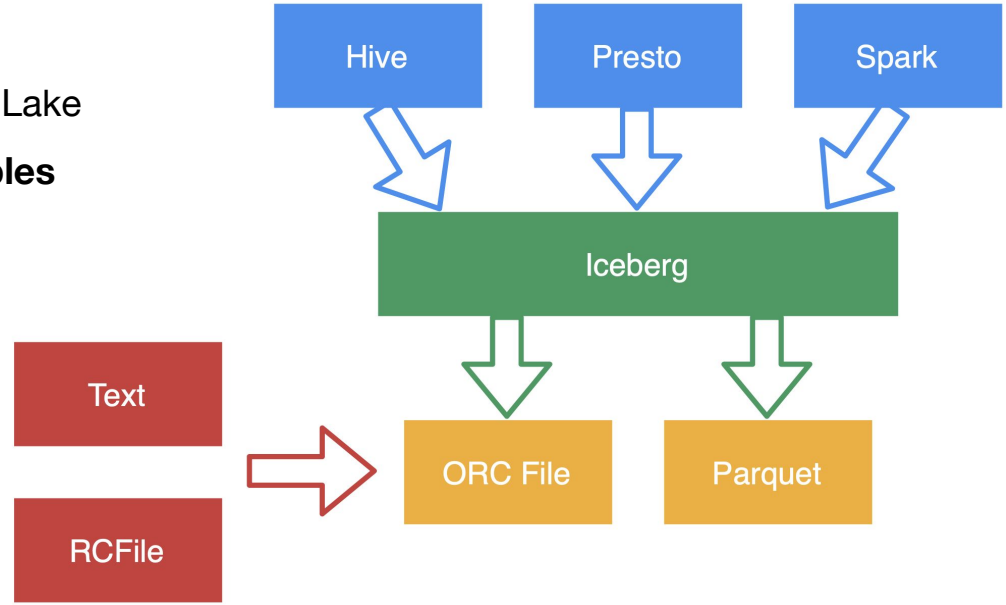
An open table format for huge analytic datasets

- ❏ Schema evolution
- ❏ Hidden partitioning
- ❏ Partition layout evolution
- ❏ Time travel
- ❏ Version rollback

Data Layers

Transitioning the Lake

from **Files** to **Tables**

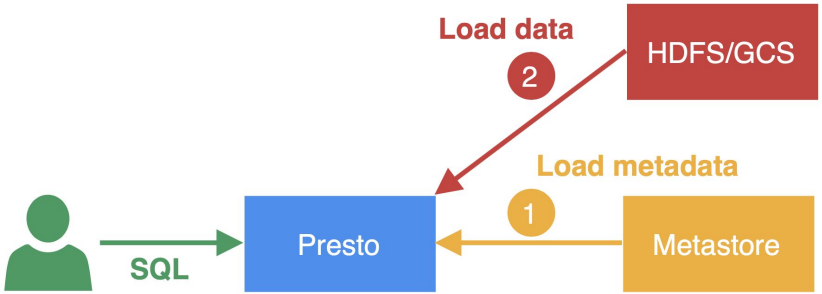


Schema Evolution

- **Add** - add a new column
- **Drop** - remove an existing column
- **Rename** - rename an existing column
- **Update** - widen the type of a column, struct field, map key, map value, or list element
- **Reorder** - change the order of columns

Design

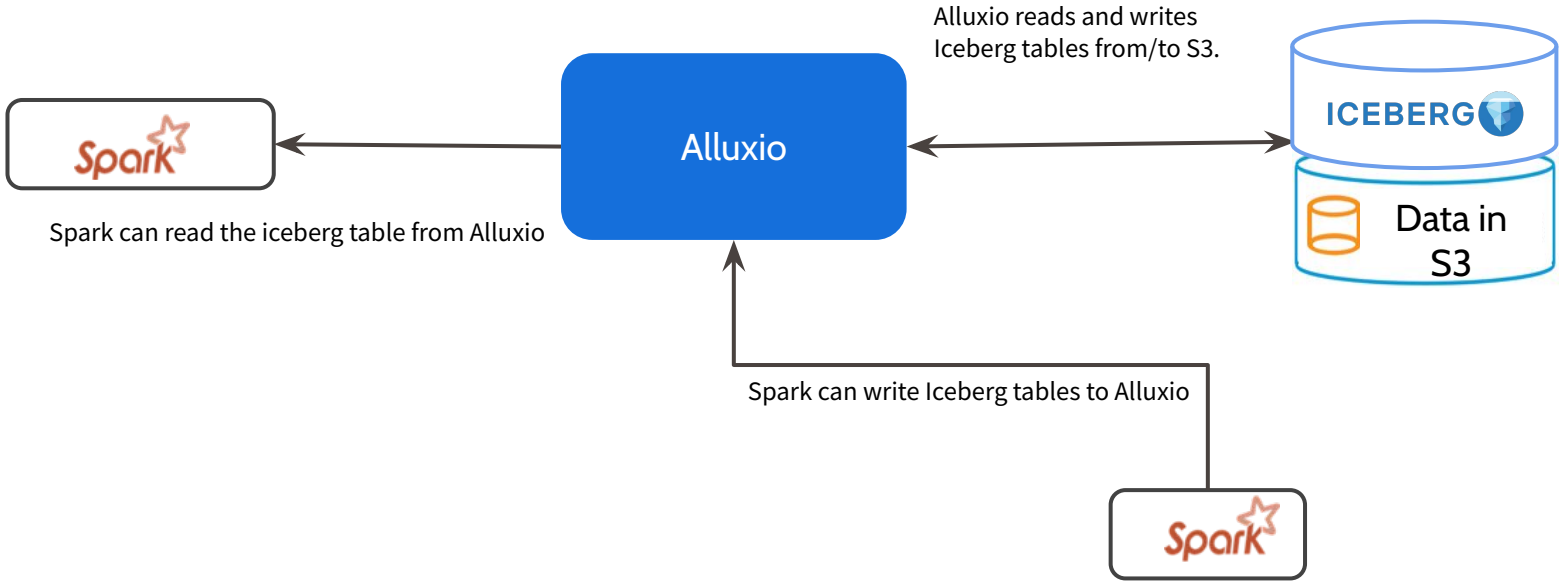
Write path



Read path

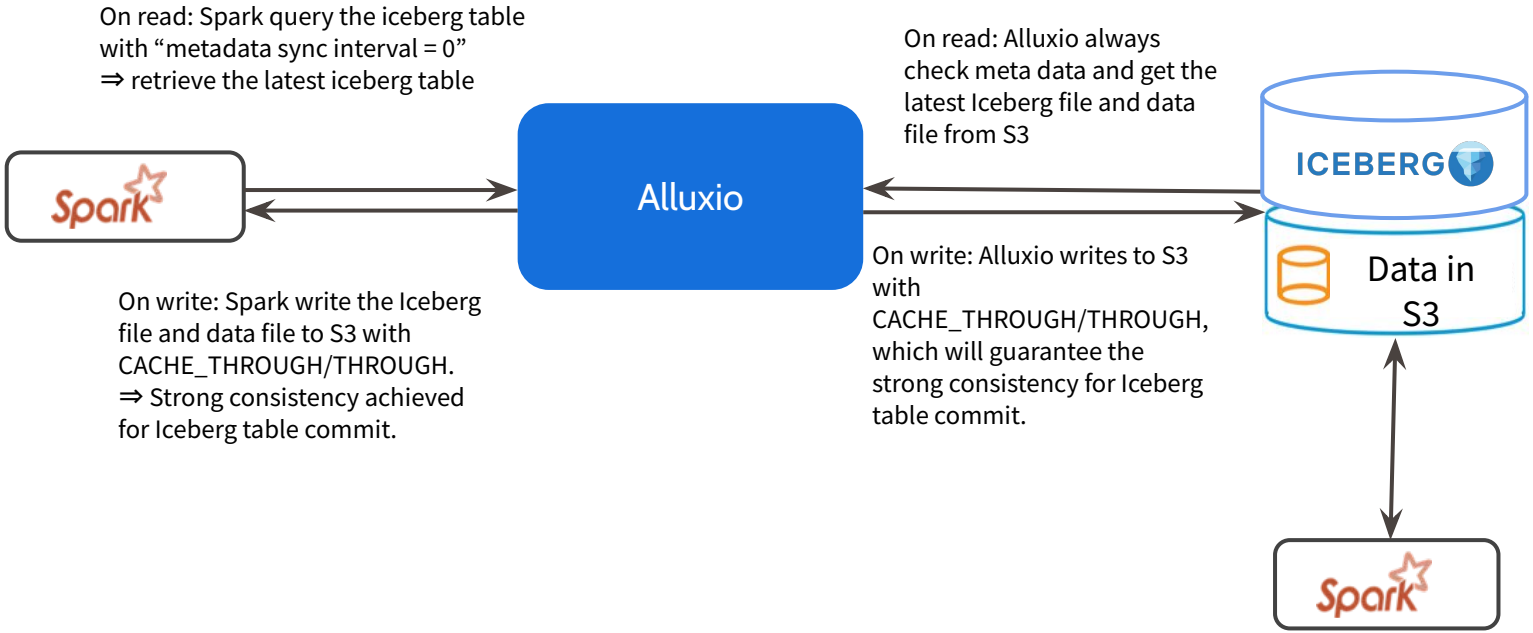
Alluxio + Iceberg Architecture: Option 1

When all accesses go through Alluxio (S3 mounted as under storage with Iceberg tables are stored)



Alluxio + Iceberg Architecture: Option 2

When Iceberg tables stored on under storage (e.g. S3 here) can be updated out side Alluxio, how to avoid reading broken table?



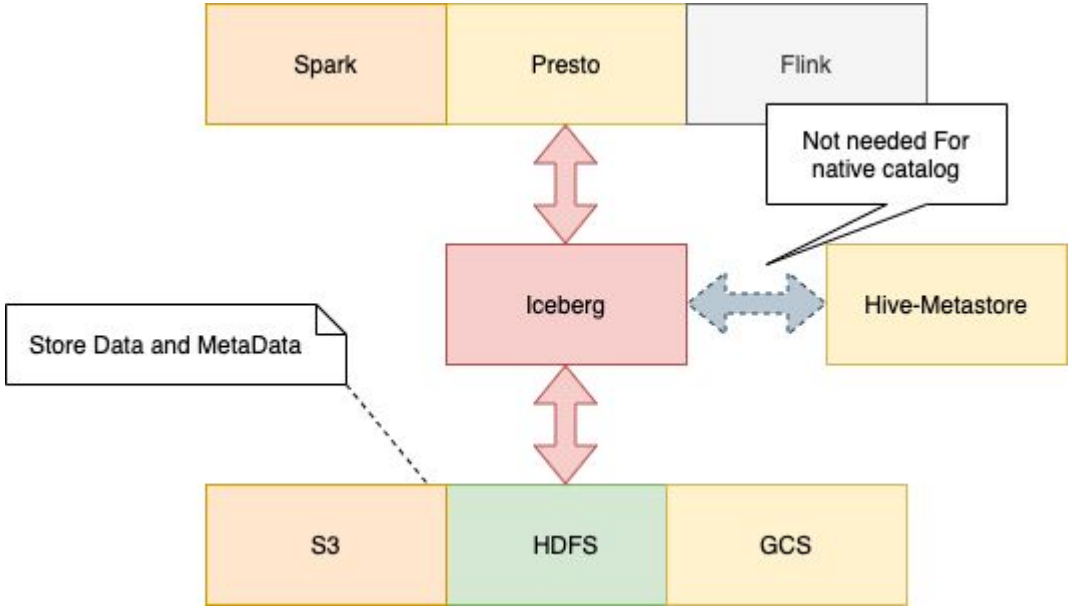
03

Best Practices

Recommendations to Users

Iceberg Native Catalog

Native folder for metadata storage (Jack Ye, AWS)



Iceberg Local Cache

Enable Iceberg Local Cache (Baolong, Tencent)

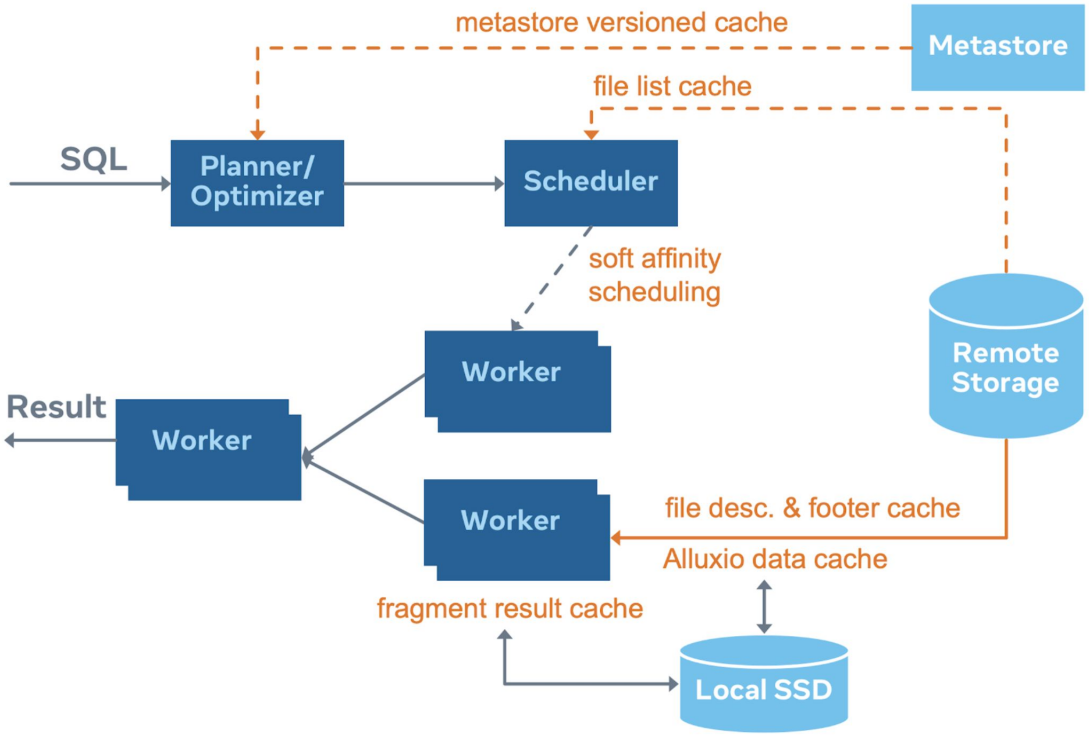
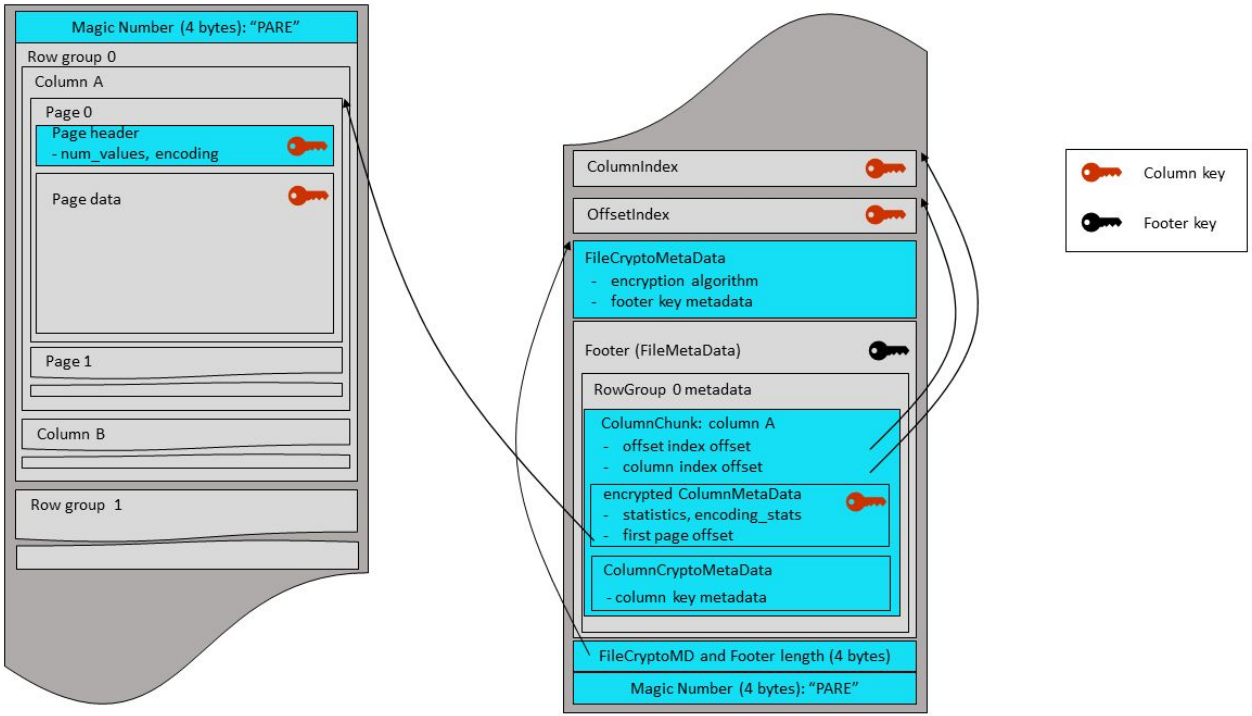


Diagram is from: https://prestodb.io/blog/2021/02/04/raptorx

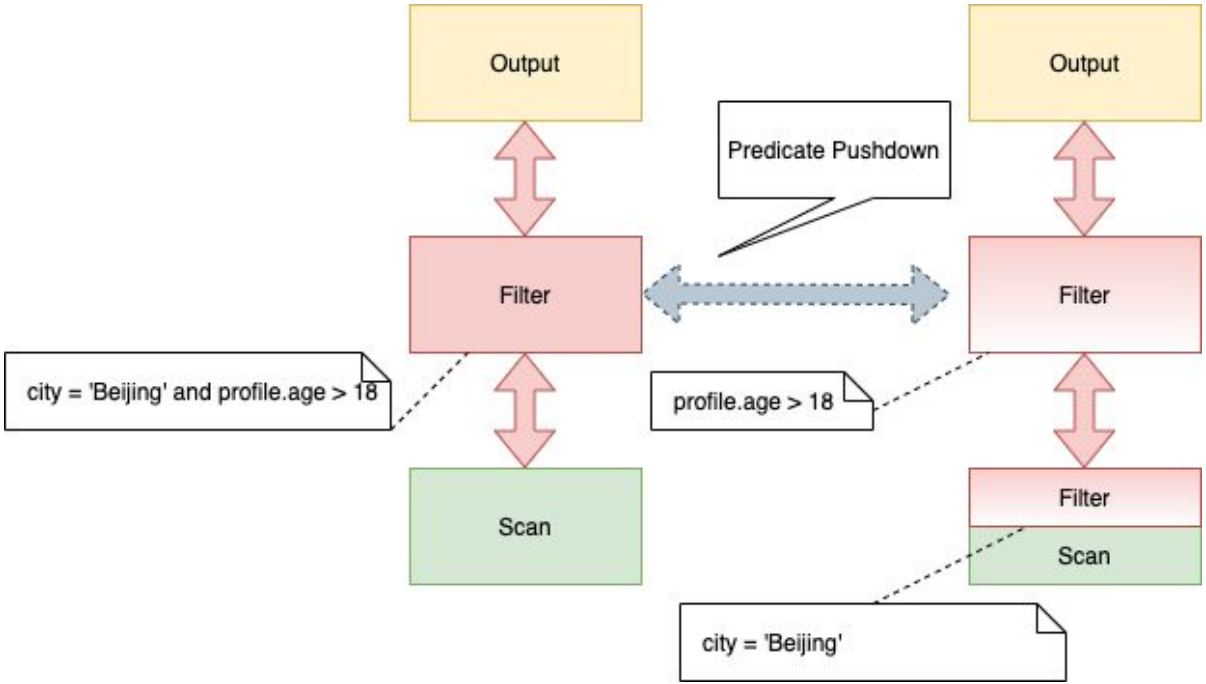
Parquet Data Encryption



From https://github.com/apache/parquet-format/blob/master/Encryption.md

Predicate Pushdown

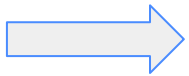
Reduce the number of partitions scanned by presto



Predicate Pushdown Resource Usage

Reduce the number of partitions scanned by presto

Resource Utilization Summary	
CPU Time	62.00ms
Scheduled Time	123.00ms
Blocked Time	289.00ms
Input Rows	2.00M
Input Data	208.55kB
Raw Input Rows	2.00M
Raw Input Data	208.55kB



Resource Utilization Summary	
CPU Time	16.00ms
Scheduled Time	100.00ms
Blocked Time	452.00ms
Input Rows	1.00
Input Data	559B
Raw Input Rows	1.00
Raw Input Data	559B

```
presto:test> select * from test1 where c_birth_month=13;
 c_customer_sk | c_birth_day | c_birth_month
---------------+-------------+---------------
          1000 |          40 |            13
(1 row)
```

ALLUXIO

04

What’s Next

Cached data transformation

- Parquet -> Arrow
  - Metadata(footer) -> flatBuffers
  - SIMD Vectorization
  - Native / Off-heap caching solution
- Computing pushdown
  - Integrate native operators with alluxio workers

Questions?

Join Alluxio & Presto slack channel

THANK YOU