Skip Rocks and Files: Turbocharge Trino queries with Hudi's multi-modal indexing subsystem

June 14th 2023

Presenters:

- Sagar Sumit {sagars@onehouse.ai}
- Nadine Farah {nadine@onehouse.ai}

Speaker Bio

Nadine Farah

- Dev Rel @Onehouse
- Contributor @Apache Hudi
- Former @Rockset, @Bose

in/nadinefarah/

@nfarah86

Sagar Sumit

- Software Engineer@Onehouse
- Committer@Apache Hudi
- Software Engineer@AWS (Amazon Aurora)
- Member Technical Staff@Oracle (Oracle GoldenGate)

Agenda

- Challenges of writing and querying data at low latency with data lakes
- How multi-modal indexing and the metadata table operate in Hudi
- Trino unlocks orders of magnitudes faster queries by leveraging Hudi's metadata table and multi-modal index
- Roadmap and community

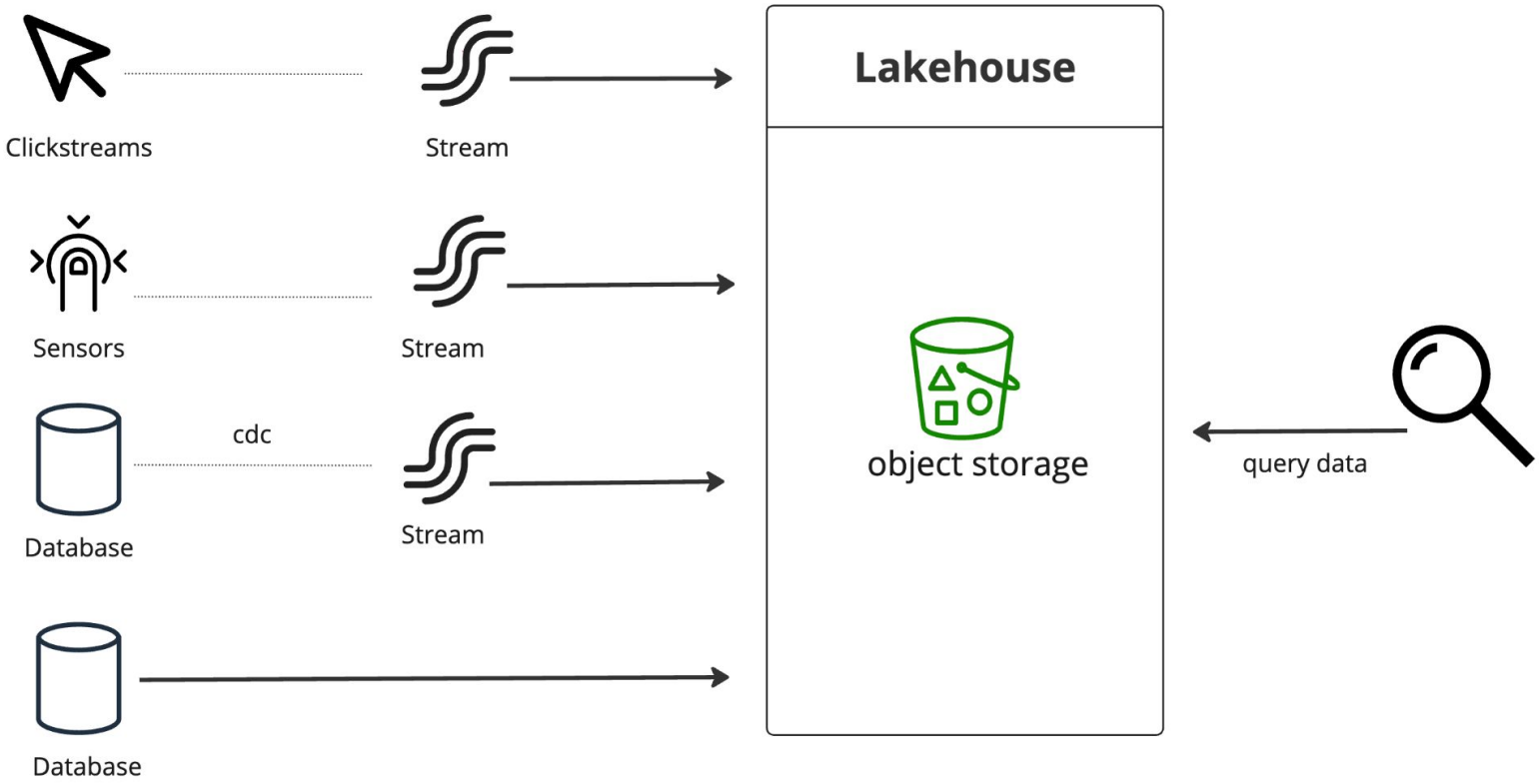
Challenges of writing and querying data at low latency with data lakes

Challenges with some lakehouse technologies

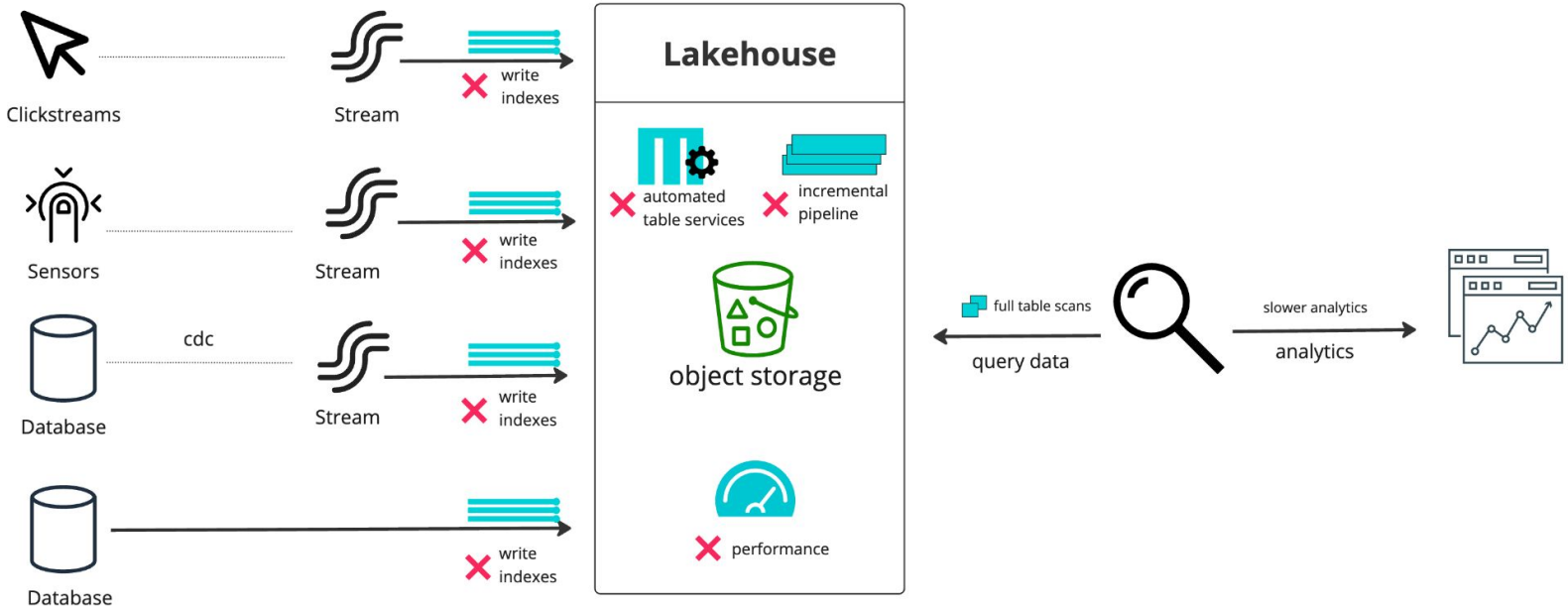
- Lack of index support
- Full table scans

Scaling applications to accommodate the vast volumes of petabyte and exabyte scale data is an ongoing challenge.

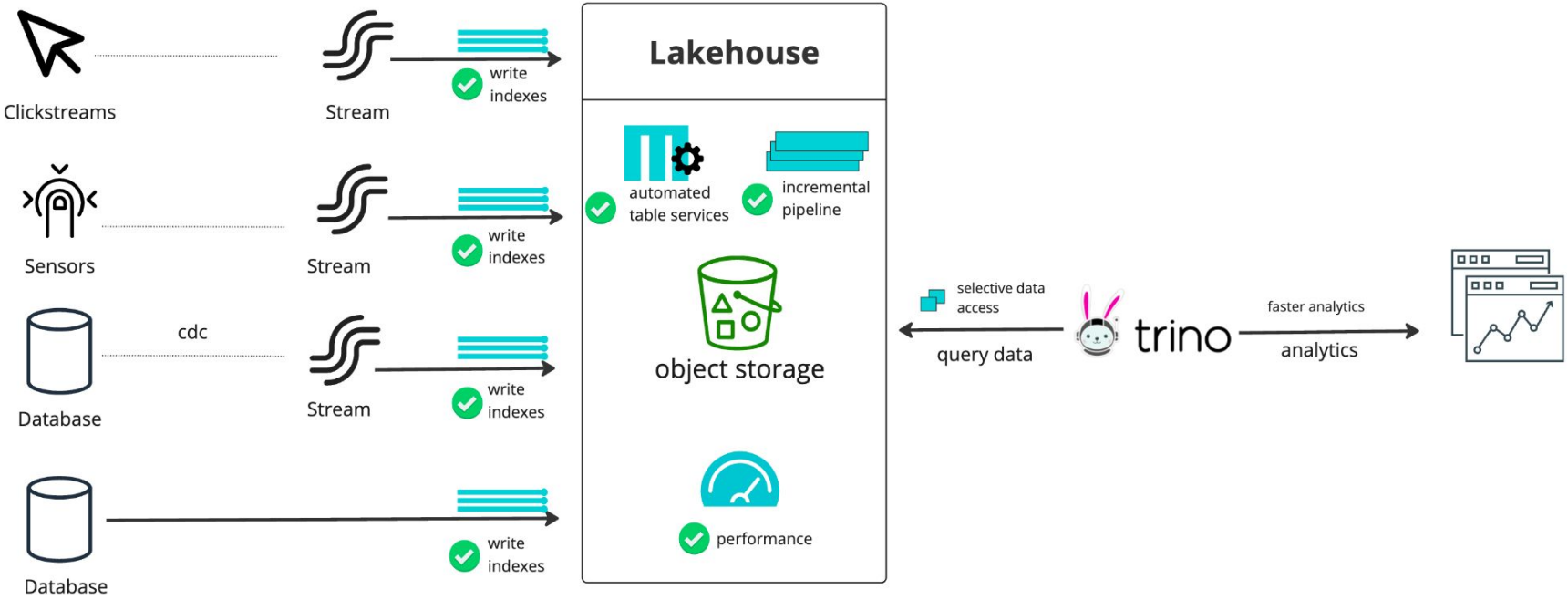
The lakehouse is the epicenter for data



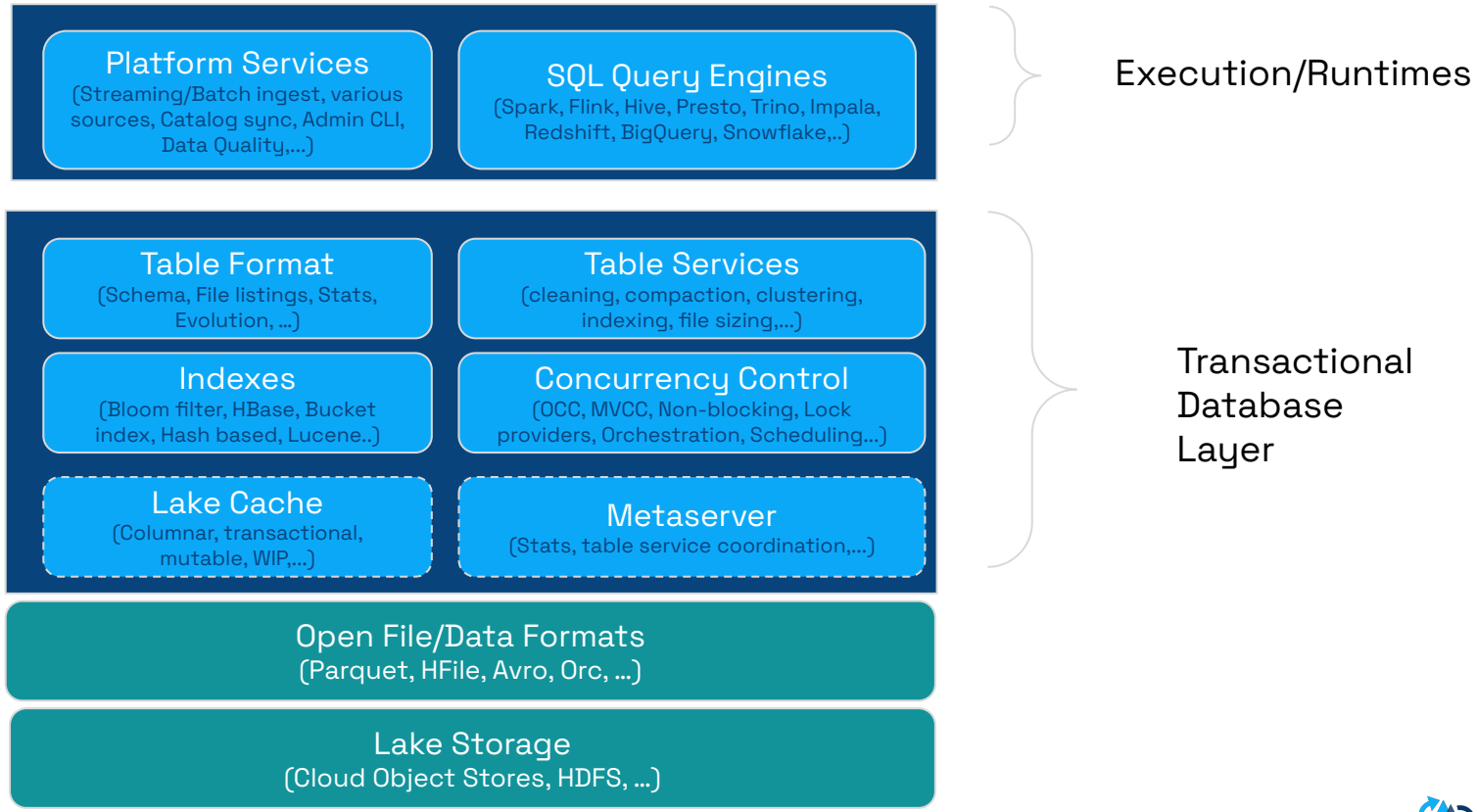
Bottlenecks in how some lakehouses process data



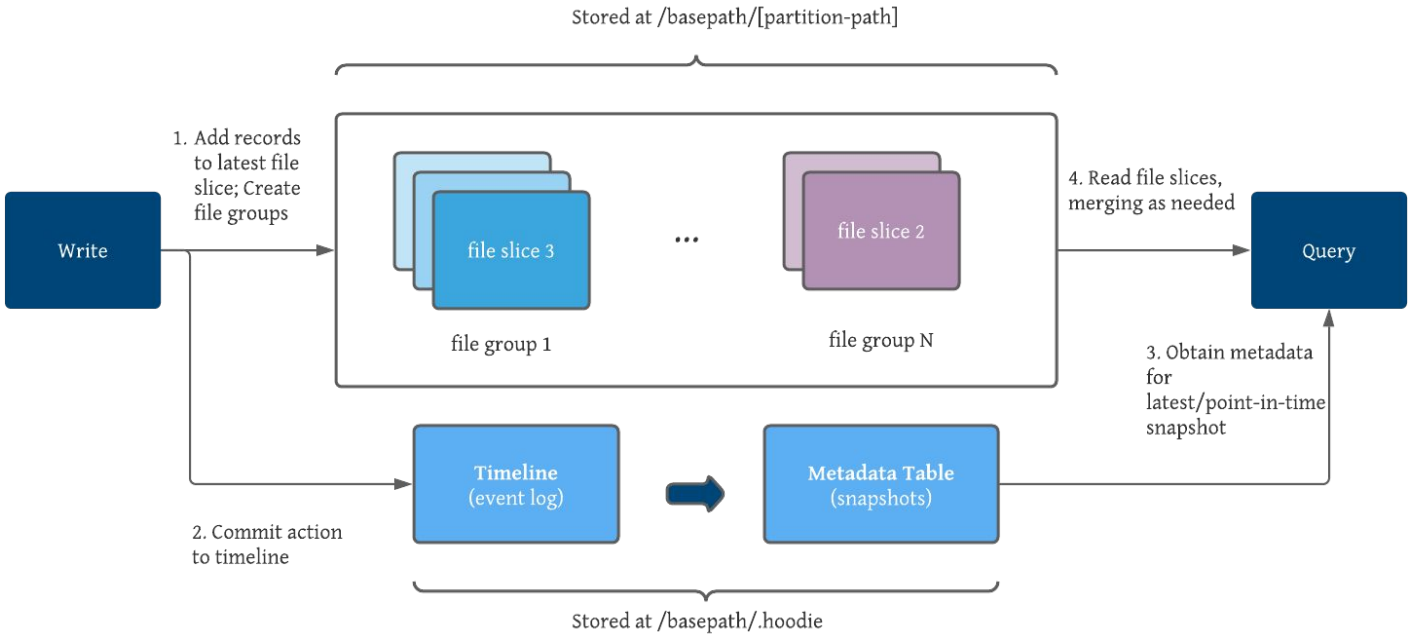
Build compute-efficient apps with Hudi and Trino



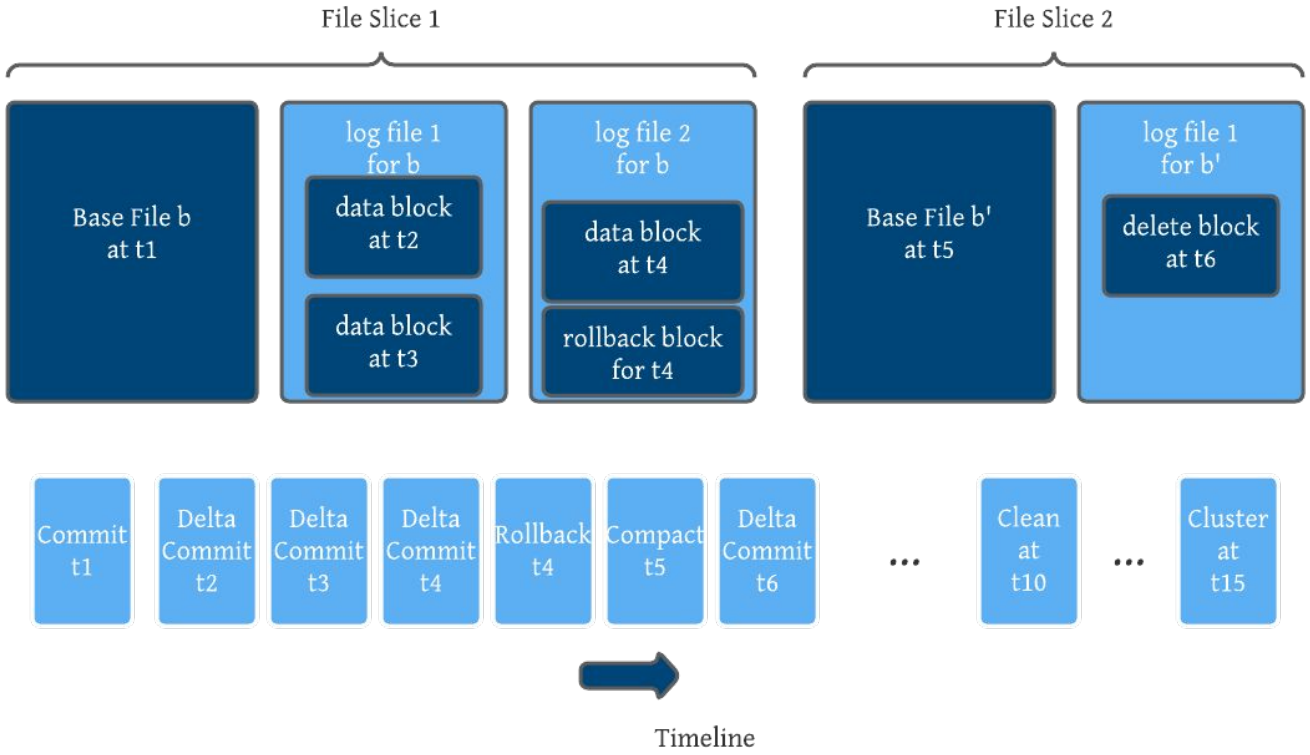
Hudi platform overview



Hudi Table Format



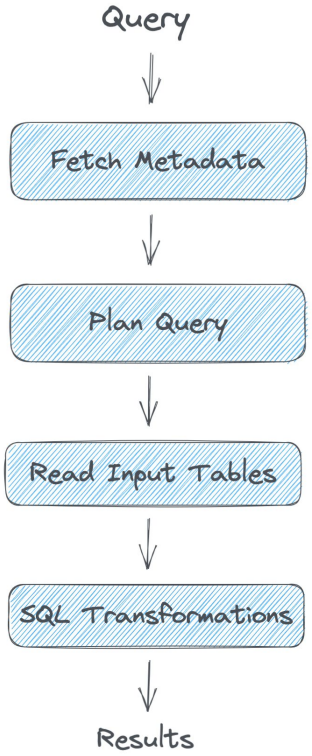
File Group Structure for a MOR table



How multi-modal indexing and the metadata table operate in Hudi

Factors affecting Query Performance

- Efficient metadata fetching -> *Table Formats* (file listings, column stats) +Metastores
- Quality of plans -> SQL optimizers
- Speed of SQL -> Engine specific (vectorized reading, serialization, shuffle algorithms..)
- Can result in orders of magnitude speed-up when implemented right.



Multi-modal indexing sub-system

Scalable metadata table

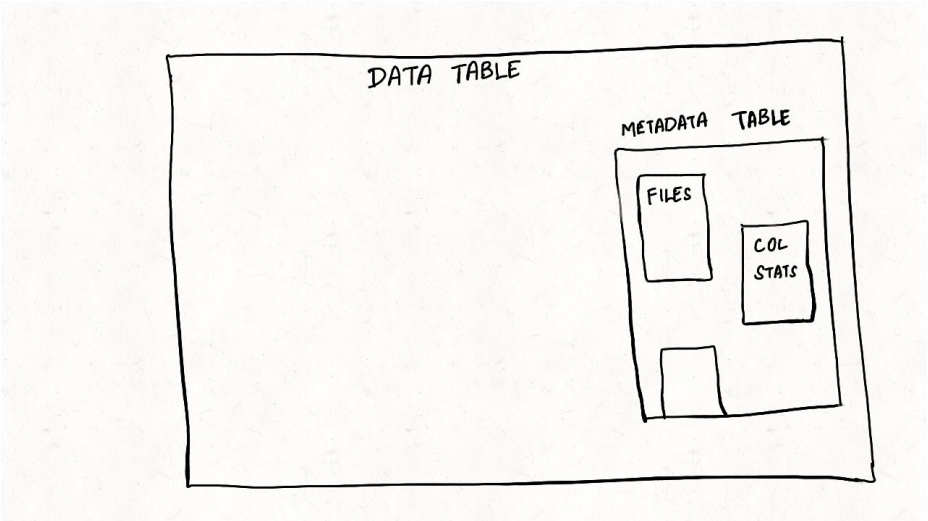
- Internal MoR table
- Different partitions store different stats, indexes

Many types of indexes

- Files, Column Stats, Bloom Filters, Record Index, secondary indexes, etc

Async Indexer

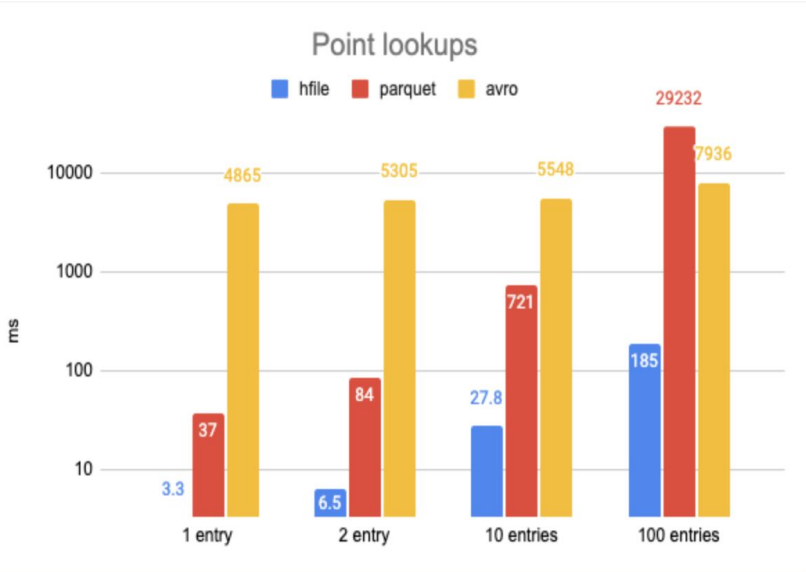
- Concurrently build index partitions
- 0-downtime operation



Design Choices

File Format

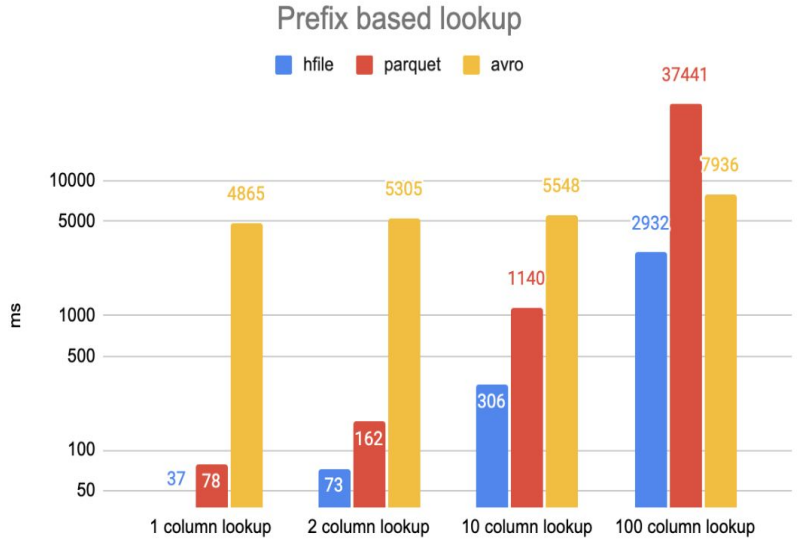
- Avro vs Parquet vs HFile
- Processing layers need point lookups within the index
- HFile 10x - 100x better than Parquet/Avro.



Design Choices

Key Format

- Ability to perform prefix lookup for range reads in column stats
- Key is composed by concatenating column name, partition name, file name
- #Entries to lookup in the index grows by the order of number of columns in the query, and not the table



Trino unlocks faster 🏃‍♀️ queries with Hudi's metadata table and multi-modal index

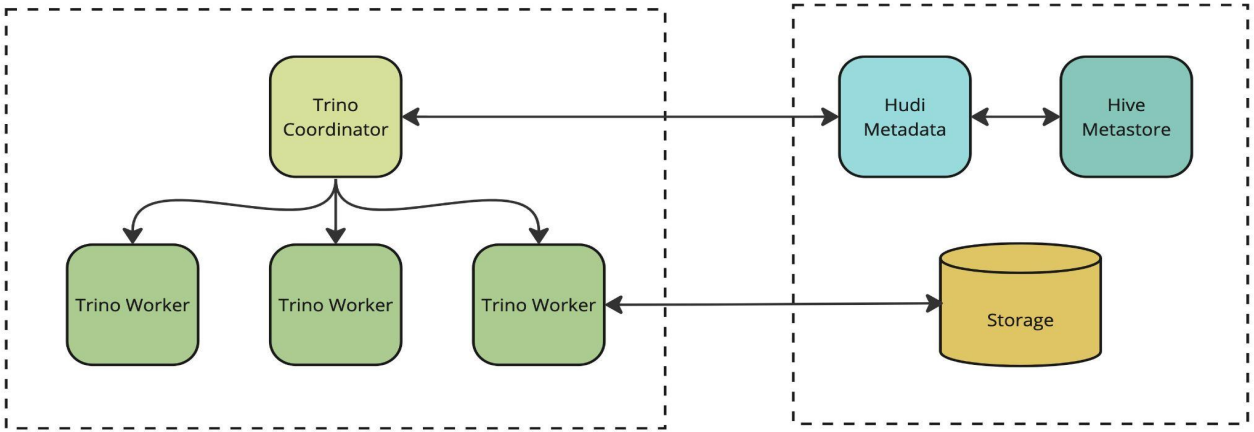
Hudi - Trino Integration

Hudi

- Rich set of FileSystem view APIs
- Fast Merge-On-Read
- Metadata indexes for data skipping

Trino

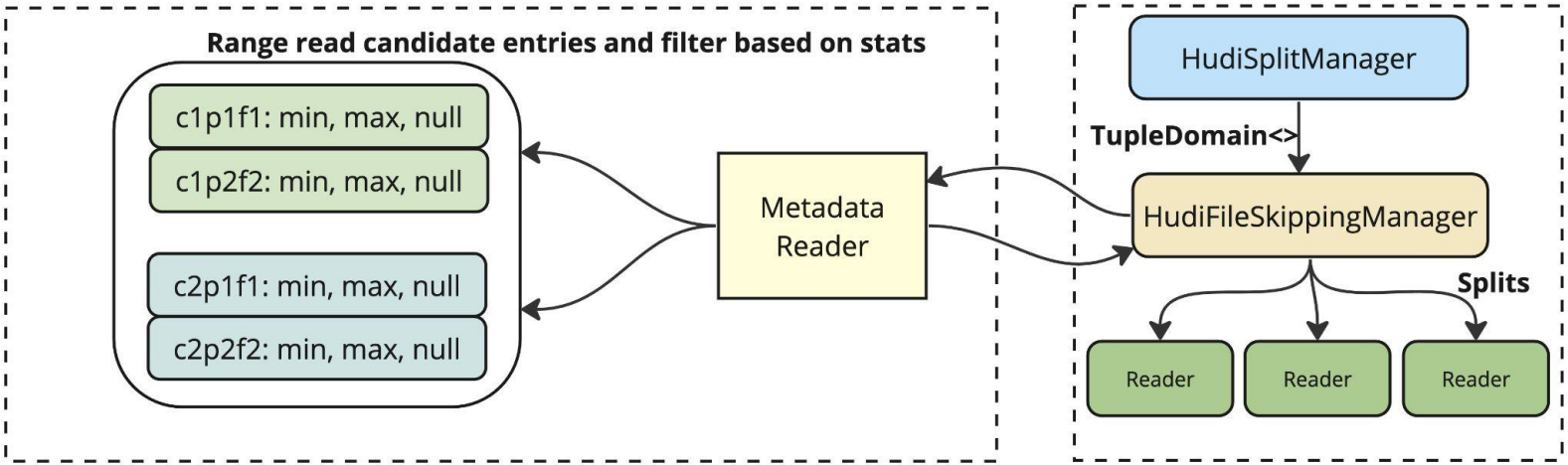
- Distributed query engine
- Seamless integration via Connector SPIs
- Highly scalable to 1000s of workers



Data Skipping with Hudi Connector

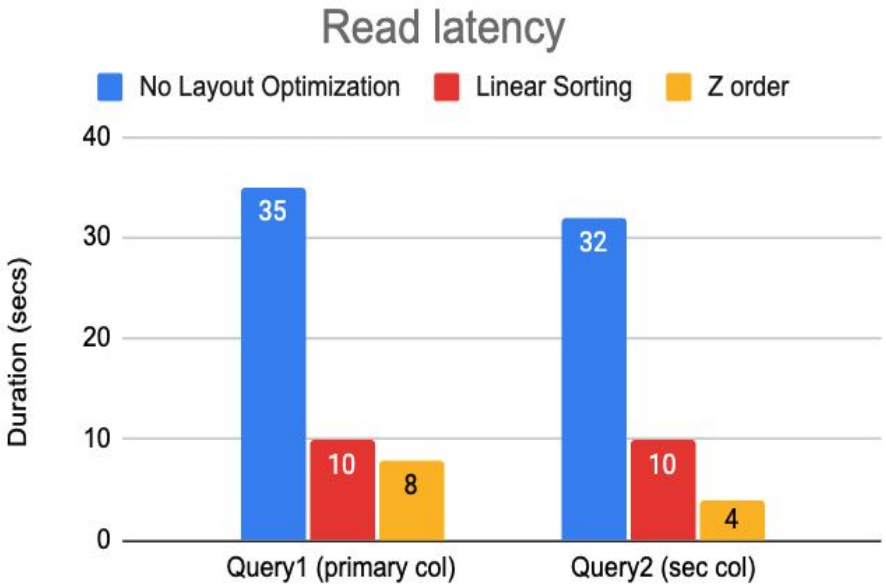
- Files index prunes partitions
- Column stats index skips data within partitions
- Set `hudi.metadata_enabled`
- Data Skipping PR

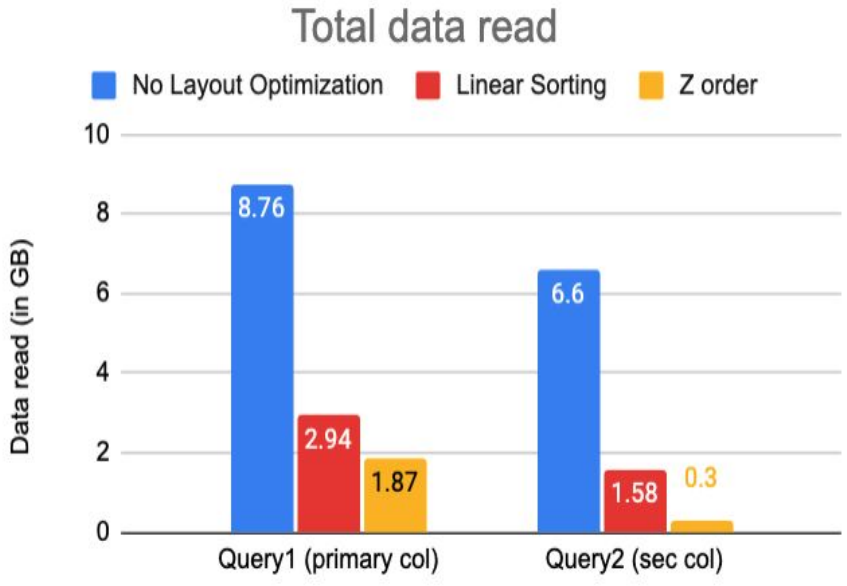
select a, b from tbl where c1 < 10 and partition_column = 'p1';
select a, b from tbl where c2 > 100 and partition_column = 'p2';



Benchmark

- Github archive data set for 6 months (220GB, 450M records)
- Sorting based on 3 diff fields





Roadmap and Community

Roadmap

- First class support for CDC data
  - Incremental queries
- Record level index
  - Global index
  - Performs better for random updates
- New Table + merge APIs
  - Easier Reader/Writer integrations
  - Engine specific merge implementations
- Write Support in Hudi connector
  - DDL/ DML
  - Storage layout optimization

Resources

https://trino.io/docs/current/connector/hudi.html

https://github.com/apache/hudi/blob/master/rfc/rfc-40/rfc-40.md

https://trino.io/episodes/41.html

https://www.onehouse.ai/blog/introducing-multi-modal-index-for-the-lakehouse-in-apache-hudi

Apache hudi The Community

Pre-installed on 5 cloud providers

aws Google Cloud Alibaba Cloud Tencent Cloud HUAWEI

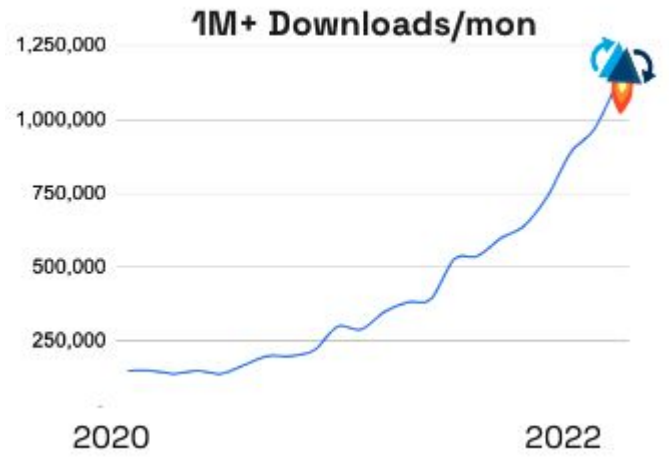
Diverse PMC/Committers

Uber aws lyft Robinhood zendesk ByteDance Tencent Cloud 蚂蚁金服 ANT FINANCIAL DV DoubleVerify

Rich community of participants

ebay Walmart udemy Disney+ Baidu 百度 Roku amazon zoom shopify SK telecom SAMSUNG Morgan Stanley HSBC Analytics Vidhya Alipay H3C 新华三集团 zomato

800B+ Records/Day (from even just 1 customer!)	3000+ Slack Members	1M DLs/month (400% YoY)
3000+ GH Engagers	300+ Contributors	30+ Committers



Come Build With The Community!

Docs : https://hudi.apache.org

Blogs : https://hudi.apache.org/blog

Slack : https://join.slack.com/t/apache-hudi/shared_invite/zt-1e94d3xro-JvlNO1kSeIHJBTVfLPlI5w

Twitter : https://twitter.com/apachehudi

Github: https://github.com/apache/hudi/ Give us a star ⭐!

Mailing list(s) :

dev-subscribe@hudi.apache.org (send an empty email to subscribe)

Join Hudi Slack

Thanks

Questions?