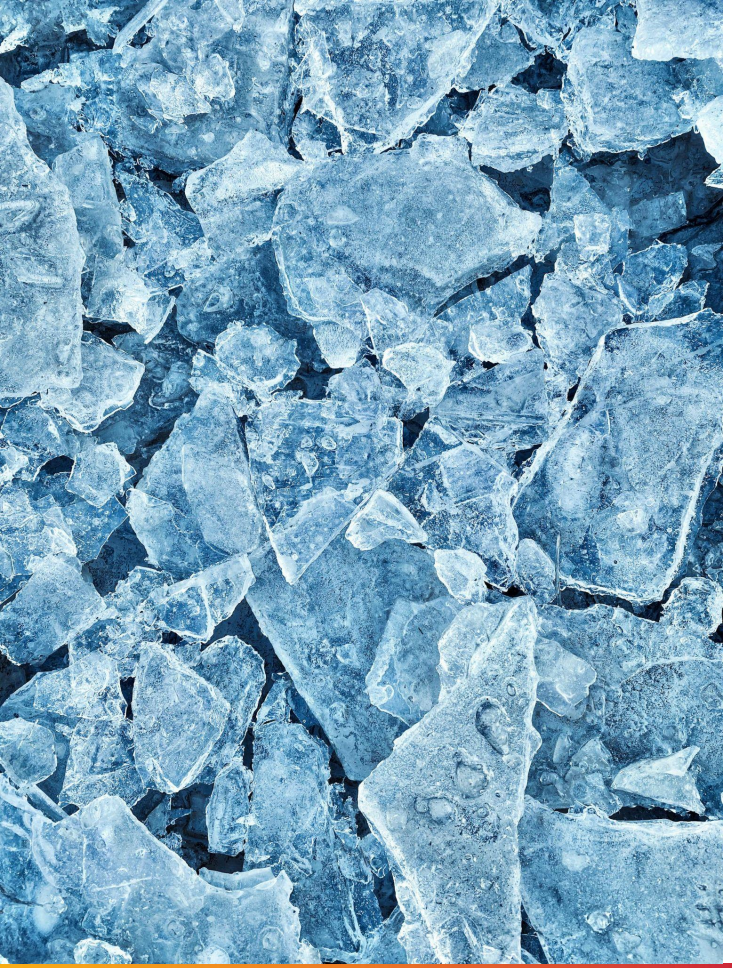
COMMUNITY
THE ASF CONFERENCE
CODE

Healing Iceberg Tables with Impala

Noémi Pap-Takács, Software Engineer at Cloudera



Agenda

Background

- **What is Apache Iceberg?**

Big Data - Big Mess

- Why Iceberg tables need maintenance

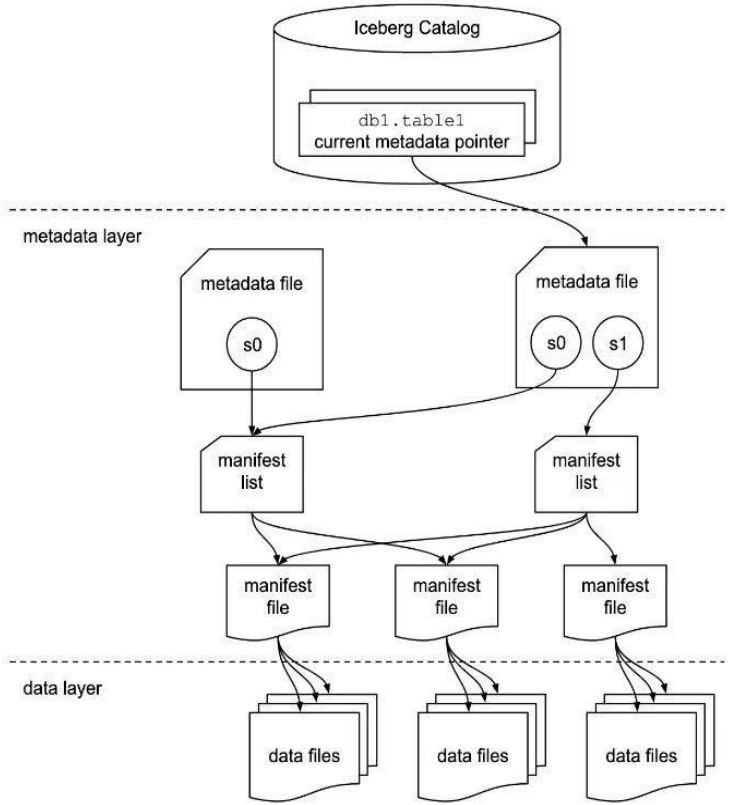
Apache Impala on Iceberg

- Introduction to Impala
- How Impala keeps Iceberg tables healthy

Apache Iceberg

- Popular **table format** for large analytic tables
- Defines how to:
  - Organize table data and metadata
  - Interact with metadata -> Spec
- Table metadata on storage
- Offers high flexibility and ACIDity

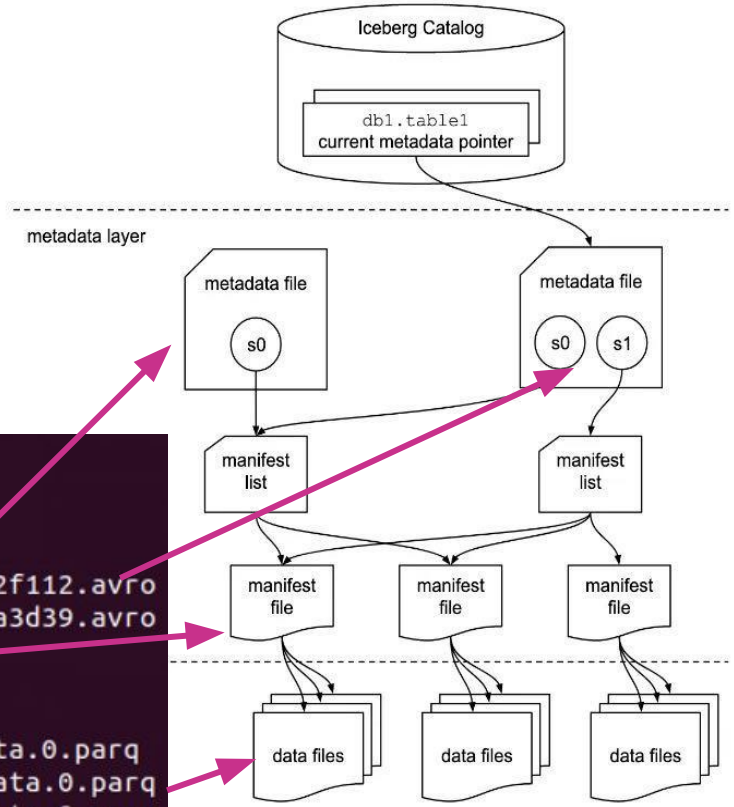
- **Library/API**
  - Clients can interact with tables
- **Catalogs**
  - HMS, Glue, JDBC Rest (Polaris)



Apache Iceberg

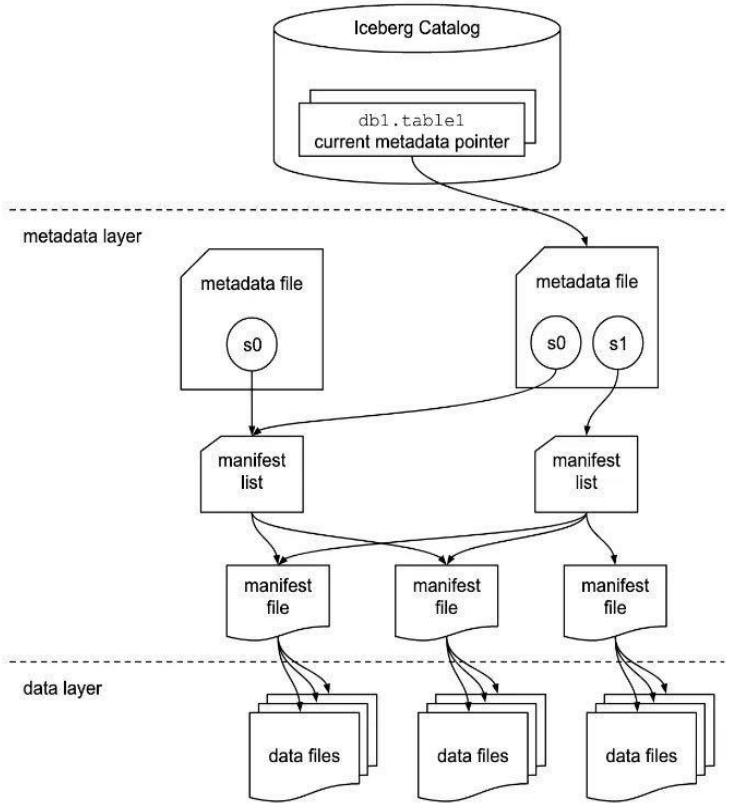
- Immutable files
- Table - metadata file
- Serialized snapshots
- Contents - snapshots
- Files in folder ≠ table content

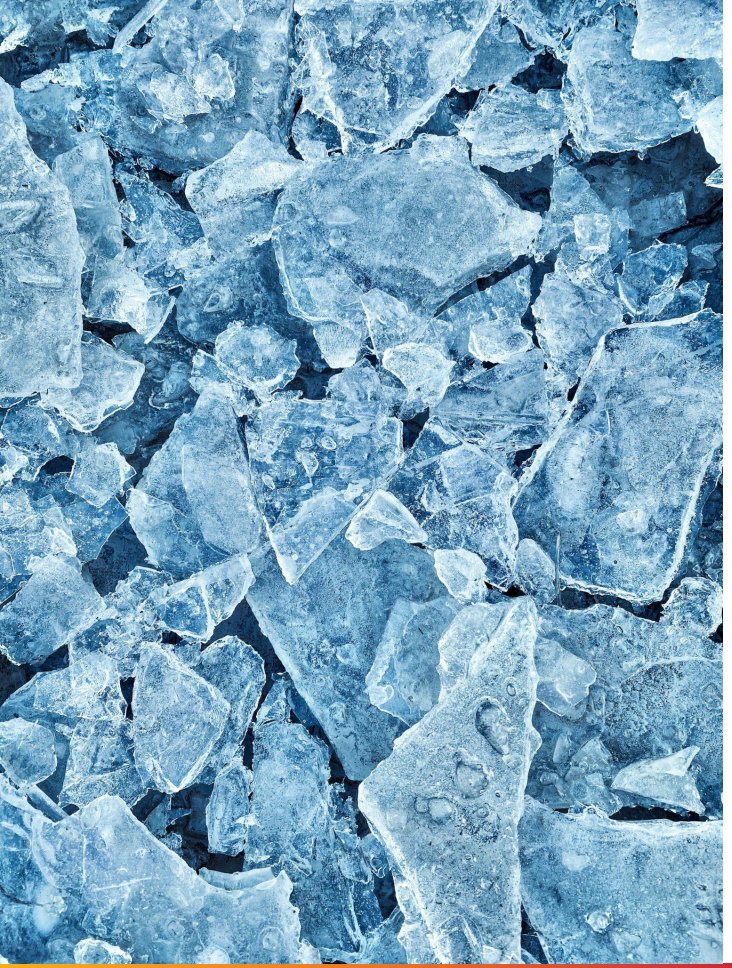
```
tbl/metadata/
  00000-7e01eda3-380a-4d83-9416-050cec97ef81.metadata.json
  00001-212fafed-3bf0-4f91-beb8-835969c4b13c.metadata.json
  00002-7f081e0a-7a0f-4aa8-aa3e-99f35e97658b.metadata.json
  snap-3990482029540480076-1-1f831dc7-16bb-4490-8354-594717d2f112.avro
  snap-8137342376748057061-1-628a9e5a-c146-4a93-96d7-cd3a546a3d39.avro
  1f831dc7-16bb-4490-8354-594717d2f112-m0.avro
  628a9e5a-c146-4a93-96d7-cd3a546a3d39-m0.avro
tbl/data/
  s_trunc=abc/1c470c37d3f7cf65-c8fbfa3800000000_558329292_data.0.parq
  s_trunc=abc/e443b0000d3885ce-57be348300000000_2116373773_data.0.parq
  s_trunc=xyz/1c470c37d3f7cf65-c8fbfa3800000000_1332670711_data.0.parq
```



Apache Iceberg

- Offers flexibility to Big Data:
  - Flexible partitioning (transforms)
  - Partition/schema evolution
    - Change partition layout without rewriting existing files
  - Time travel
  - Branching and tagging
  - Row-level modifications
    - UPDATE
    - DELETE
    - MERGE
- ...with ACID guarantees
  - Optimistic concurrency
    - No locking





Agenda

Background
- What is Apache Iceberg?

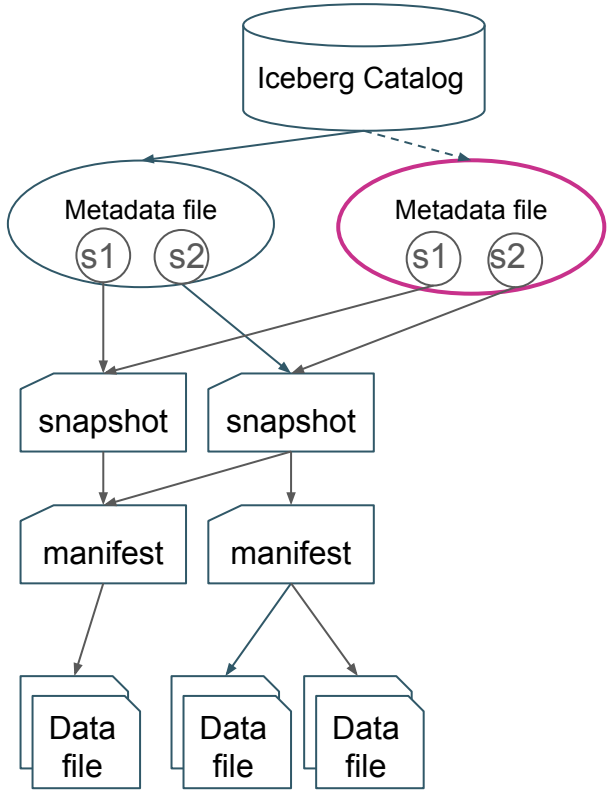
Big Data - Big Mess
- **Why Iceberg tables need maintenance**

Apache Impala on Iceberg
- Introduction to Impala
- How Impala keeps Iceberg tables healthy

Why Iceberg tables need maintenance

- Immutable files
  - Accumulate

- Metadata operations
  - Change schema, partitioning
- Effect on the table:
  - New metadata file

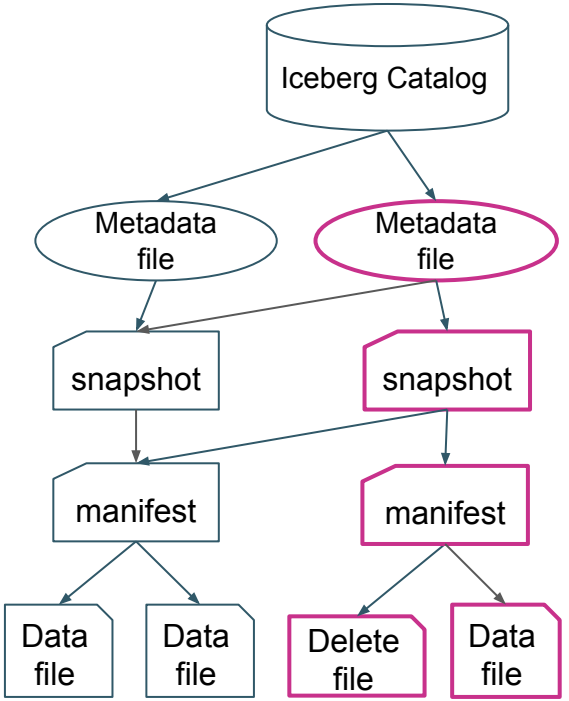


Why Iceberg tables need maintenance

Row-level modifications

- DELETE
- UPDATE
- MERGE

- Delete strategies:
  - Copy-on-write
  - **Merge-on-read**
    - Positional deletes
    - Equality deletes



Why Iceberg tables need maintenance

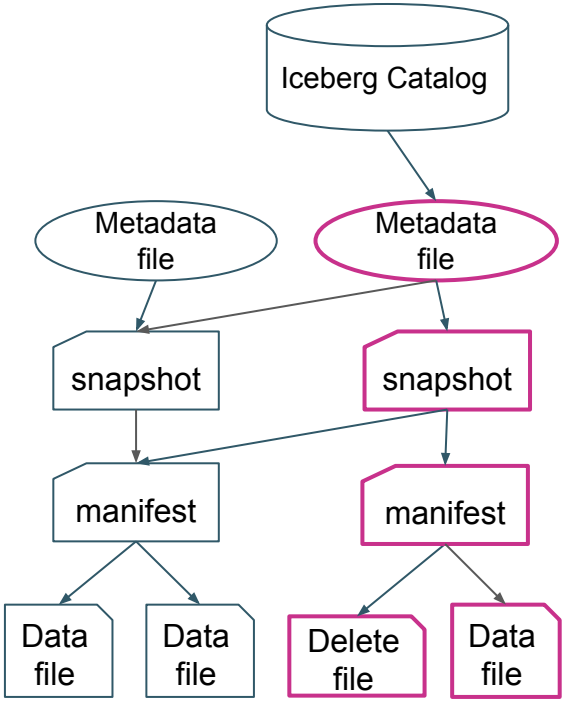
Row-level modifications

Strategy		Write speed	Read speed	Ideal use case
Copy-on-write		slowest	fastest	Infrequent updates
Merge-on-read	positional	fast	**slower**	Frequent updates
	equality	fastest	**slowest**	Frequent updates, streaming

Why Iceberg tables need maintenance

DML operations

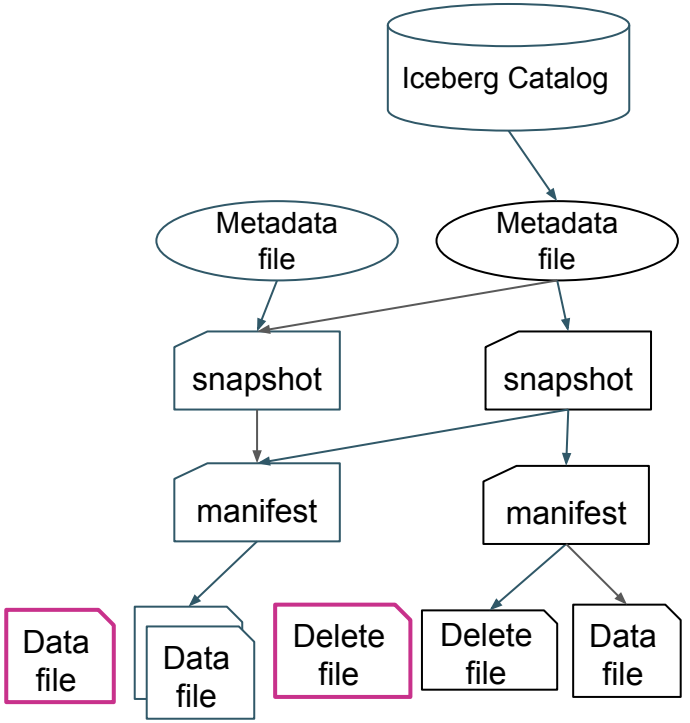
- DML
  - INSERT
  - UPDATE
  - DELETE
  - MERGE
- Effect on the table:
  - New metadata file
  - New snapshot
  - New data and delete files
    - Small files problem
  - **Performance regression**
    - 1 trillion row challenge:
      - 1 trillion row data
      - 68 billion deleted rows (~7%)
      - +30% read time

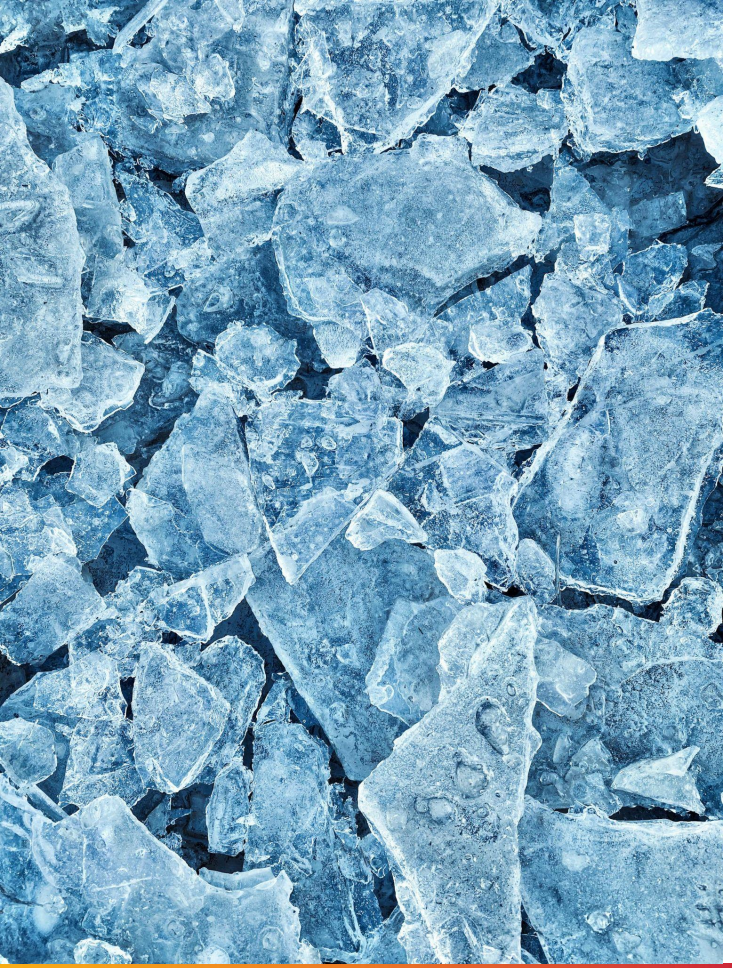


Why Iceberg tables need maintenance

Orphan files

- Not reachable by any snapshot
- Result of failures





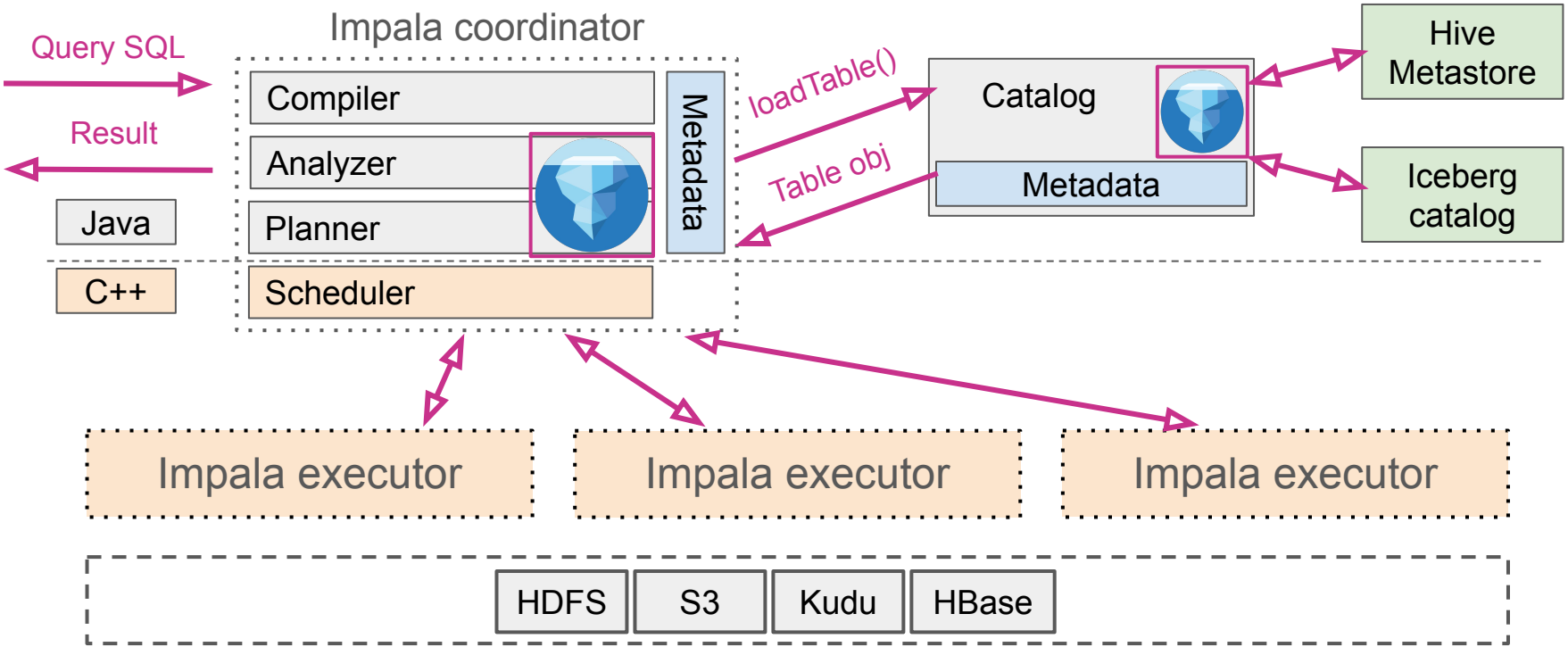
Agenda

Background
- What is Apache Iceberg?

Big Data - Big Mess
- Why Iceberg tables need maintenance

Apache Impala on Iceberg
- **Introduction to Impala**
- **How Impala keeps Iceberg tables healthy**

Apache Impala on Iceberg



Apache Impala on Iceberg

- Extensive support for Iceberg tables:
  - Row-level modifications
    - merge-on-read
  - Metadata table queries
  - **Table maintenance**

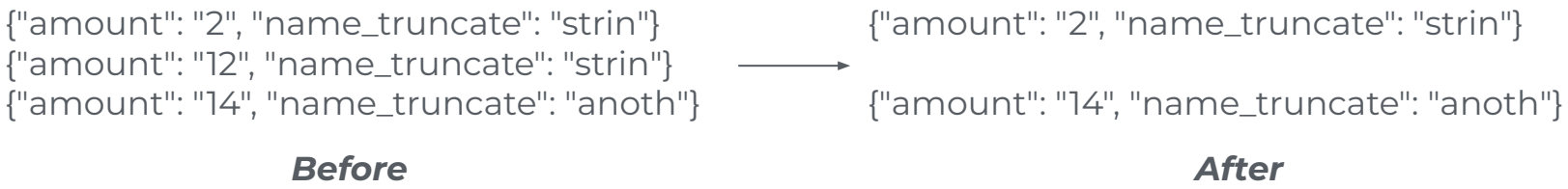
Iceberg Table Maintenance in Impala

DROP PARTITION

- Supports partition transforms
- Metadata-only operation

```
ALTER TABLE ice_table DROP PARTITION (day(d)='2024-06-03');
```

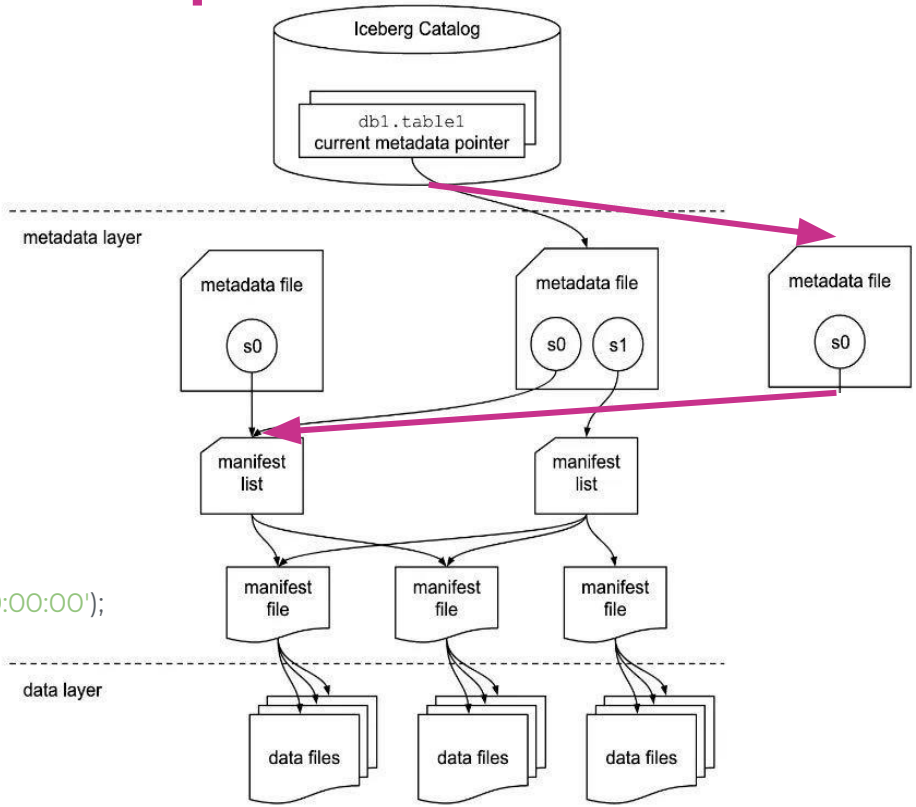
```
ALTER TABLE ice_table DROP PARTITION (amount > 10 and truncate(5, name) = 'strin');
```



Iceberg Table Maintenance in Impala

EXECUTE ROLLBACK

- Reset table to a known good state
- Roll back the table data based on a snapshot id or a timestamp
- New metadata file



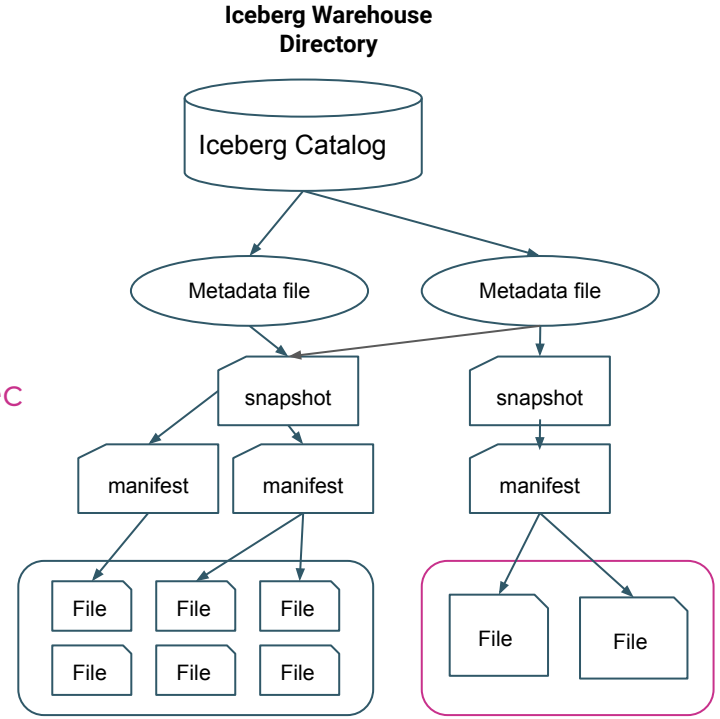

```
ALTER TABLE ice_table EXECUTE ROLLBACK('2022-08-08 00:00:00');
```

Iceberg Table Maintenance in Impala

OPTIMIZE

OPTIMIZE TABLE iceberg_table;

- Merge delete files
- Combine small files into larger ones
- Use latest schema
- Rewrite table according to the latest partition spec
- Executes compaction on entire table

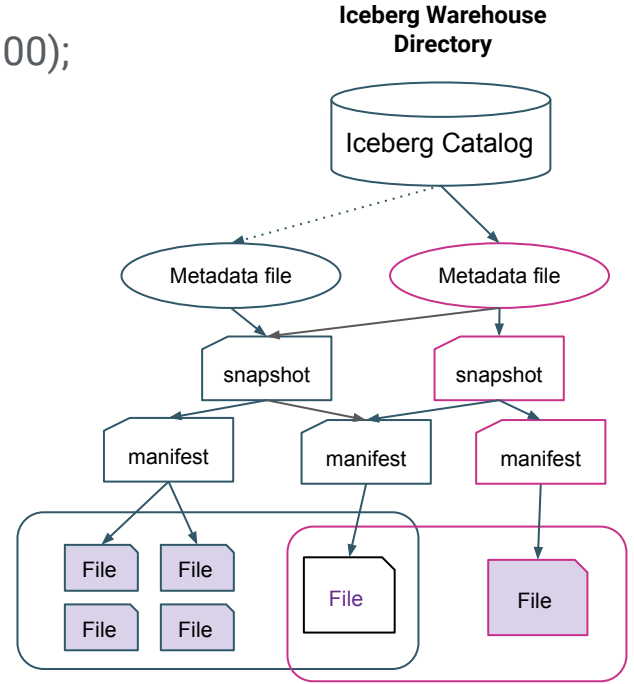


Iceberg Table Maintenance in Impala

OPTIMIZE

OPTIMIZE TABLE iceberg_table (**FILE_SIZE_THRESHOLD_MB**=100);

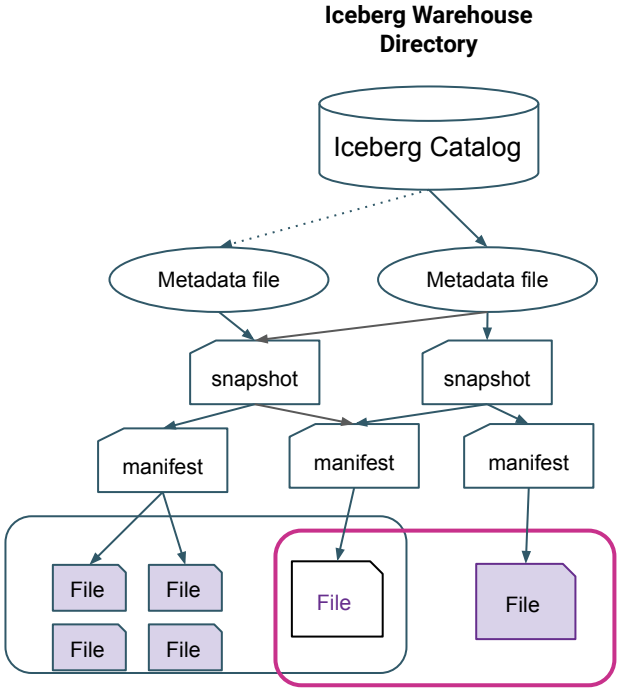
- Rewriting the entire table is expensive
- Conflict with updates
- Filter data files based on size:
  - Small files
  - Delete files



Iceberg Table Maintenance in Impala

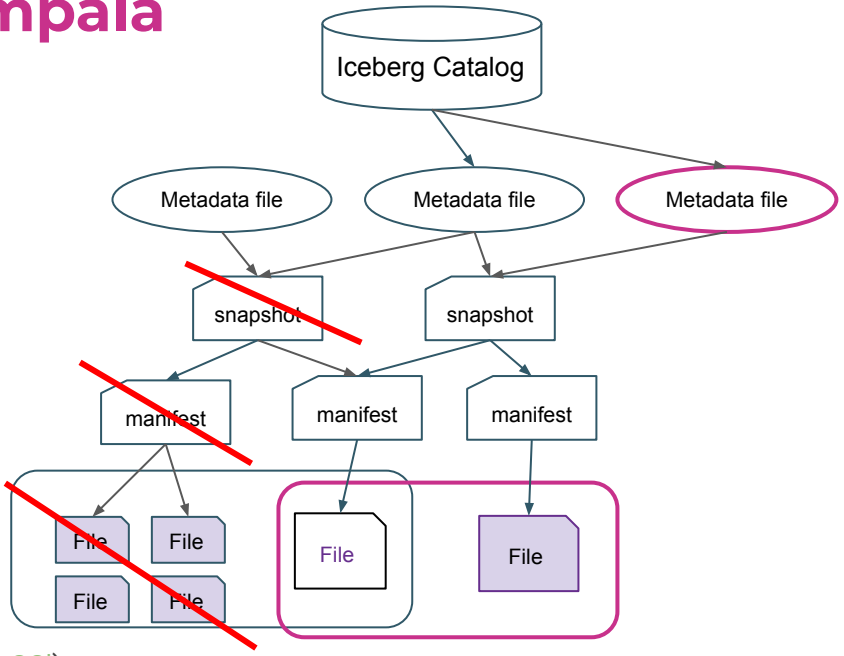
COMPACTION effect on performance

- Restoring read performance
  - 1TRC: 7% -> 30%
- Also compacts metadata layer
- Less disk usage*
  - *current table content
  - Better compression



Iceberg Table Maintenance in Impala

EXPIRE SNAPSHOTS



Snapshots accumulate until expired

- delete data files that are no longer referenced
- reduce the size of table metadata

```
ALTER TABLE ice_table EXECUTE expire_snapshots('2022-01-04 10:00:00');

ALTER TABLE ice_table EXECUTE expire_snapshots(now() - INTERVAL 5 DAYS);
```

Iceberg Table Maintenance in Impala

DELETE ORPHAN FILES

- Files that are not reachable by metadata
  - Uncommitted files
- Future work

When does the table need maintenance?

Iceberg Table Maintenance in Impala

When does the table need maintenance?

- Performance regression
- File statistics
  - Check file system
  - Query metadata table
    - Partitions, files, history

```
SELECT SUM(file_size_in_bytes) FROM db.tbl.all_files;
```

```
SELECT partition, AVG(file_size_in_bytes) FROM db.tbl.files GROUP BY partition;
```

Summary

- Use case
- Streaming source
- Chose delete strategy wisely
- Compact regularly - maintenance window
- Expire unnecessary snapshots
- Clean orphan files

COMMUNITY
THE ASF CONFERENCE
CODE

Questions?

Thank you for your attention!