Day - 19 (14th June)

14 June 2024 10:08

Apache Hudi :- Vinod → ex-employee Uber.

↓

free/open Source

① Streamline the real time data.

② Manage SQL Table on your datalakes to build the multi-stage incremental pipeline.

Hudi

↓

Hadoop Upsert Delete & incrementality.

Hudi Feature :-

① Upsert, Delete with fast pluggable index.

② Transaction, Rollbacks with concurrency control good.

③ Automatic file sizing, data clustering.

④ Built-in Metadata tracking for scalable Storage Access.

⑤ SQL can read/write from Hive, Spark, Presto

⑥ Evolving Schema Comparison.

Hudi → COW → Copy on Write

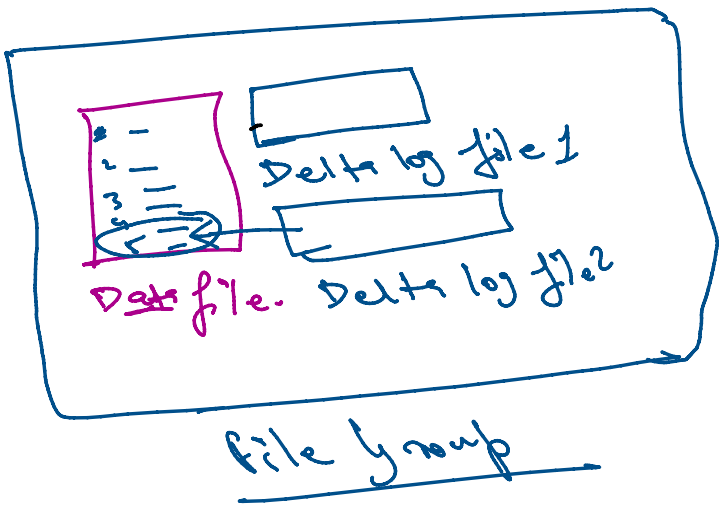
Hudi → MOR → Merge on Read.

Data file / Base file :- Hudi store data in columnar Parquet format. It is called

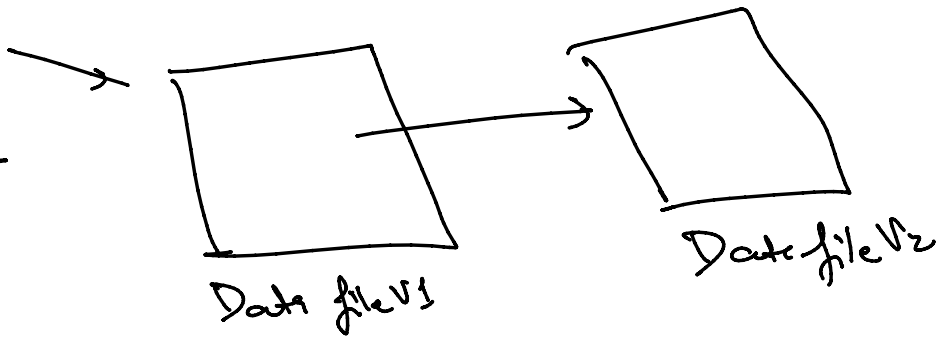
Data file / Base file.

Data file / Base file.

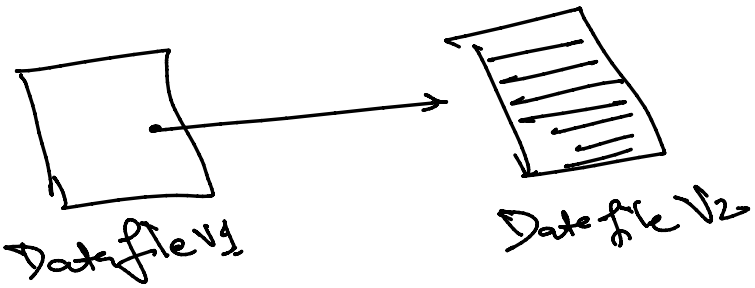
Delta log file :- MOR Table format, updates are sent to delta log file, which is stored in Avro format. This delta log file is always associated with the base / Delta log file.



File Version :-



COW :- Copy on write.



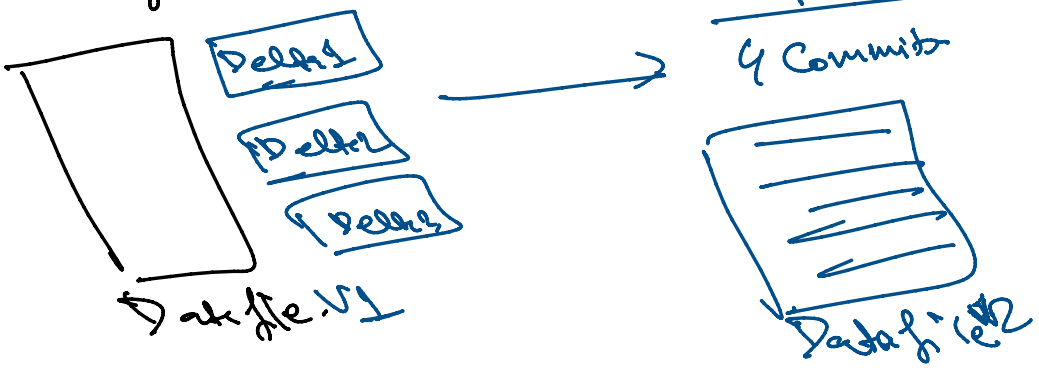
MOR :- Merge on Read.

Delta

Compaction strategy :-

- Commits

MOR:-

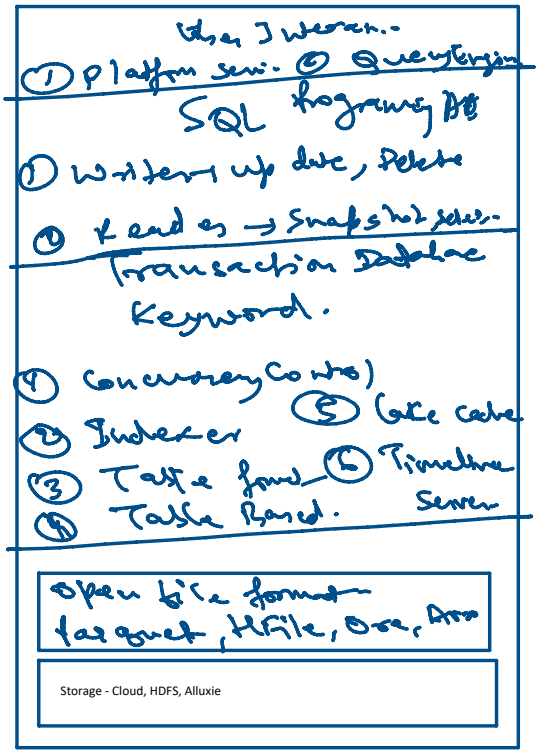


COW

1. Higher Write latency
2. Low Latency Read
3. Cost high
   I/O ↑

MOR

1. less COW, Synchronous Merging
2. High latency Reads
3. Cost less
   I/O ↓



Lab:-

1. S3 Bucket -

Database

(1) S3 Bucket.

(2) Glue → Spark Query. → Database Table.

(3.) Athena → Query the database.

rt → real time → Commit, it will be available.

ro → read optimized view ↳ Query