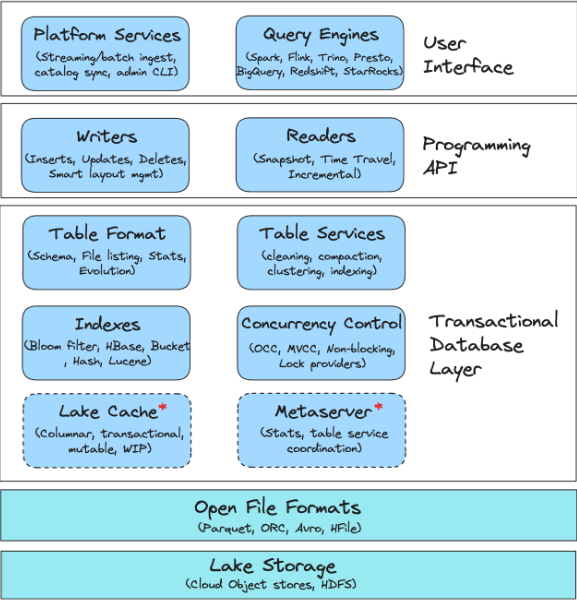
1. Getting Started
2. Design and concetps
3. Architechtural

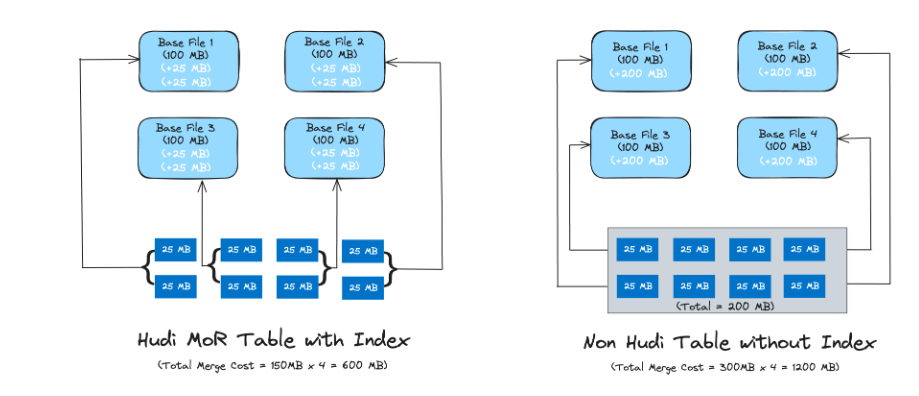


Storage: file parquet, lưu s3, gcs, azure, hdfs

2, indexing

[Copy-On-Write tables](https://hudi.apache.org/docs/next/table_types#copy-on-write-table) : cho phép các hoạt động upsert/delete nhanh chóng, bằng cách tránh nhu cầu phải join với toàn bộ tập dữ liệu để xác định các tệp nào cần ghi lại.

[Merge-On-Read](https://hudi.apache.org/docs/next/table_types#merge-on-read-table) , thiết kế này cho phép Hudi ràng buộc số lượng bản ghi mà bất kỳ tệp cơ sở nào cần được merge



Hudi cung cấp hỗ trợ sử dụng kết hợp nhiều loại chỉ mục

Các loại chỉ mục hudi hỗ trợ: **BLOOM, GLOBAL\_BLOOM, SIMPLE (default for Spark engines), GLOBAL\_SIMPLE, HBASE, INMEMORY (default for Flink and Java), BUCKET, RECORD\_INDEX**

1. Writing tables : có các giá trị upsert(default), bulk\_insert , insert, delete

1.Insert data ,

- Insert Overwrite Table

+ Xóa toàn bộ dữ liệu trước đó và ghi đè dữ liệu

option("hoodie.datasource.write.operation","insert\_overwrite\_table").  
 option("hoodie.datasource.write.precombine.field", "ts").  
 option("hoodie.datasource.write.recordkey.field", "uuid").  
 option("hoodie.datasource.write.partitionpath.field", "partitionpath").  
 option("hoodie.table.name", tableName).  
 mode(Append).  
 save(basePath)

ví dụ:

- Insert Overwrite

+ ghi đè tất cả các bản ghi có cùng partion được lựa chọn (upsert), bỏ qua indexing, precombining,

option("hoodie.datasource.write.operation","insert\_overwrite").  
 option("hoodie.datasource.write.precombine.field", "ts").  
 option("hoodie.datasource.write.recordkey.field", "uuid").  
 option("hoodie.datasource.write.partitionpath.field", "partitionpath").  
 option("hoodie.table.name", tableName).  
 mode(Append).  
 save(basePath)

ví dụ :

2.Update data

- các bảng hudi có thể được update thông qua dataframe: sẽ tạo ra các dataframe mới sau đó append lại vào table

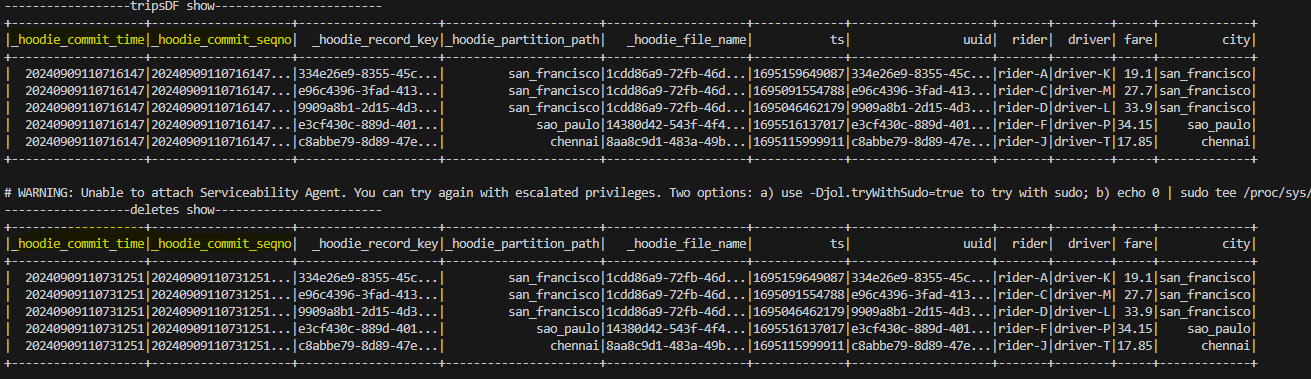
updatesDf.write.format("hudi"). options(\*\*hudi\_options). mode("append"). save(basePath)

3.Delete data

Xóa các bản ghi được chỉ định ra khỏi bảng , có 2 cách xóa

- Sort Deletes: giữ lại record key và loại bỏ các giá trị cho tất cả các trường khác, luôn lưu trữ trong bộ nhớ và không bao giờ bị xóa,

option("hoodie.datasource.write.operation", "upsert").  
 option("hoodie.datasource.write.precombine.field", "ts").  
 option("hoodie.datasource.write.recordkey.field", "uuid").  
 option("hoodie.datasource.write.partitionpath.field", "partitionpath").  
 option("hoodie.table.name", tableName).  
 mode(Append).  
 save(basePath)

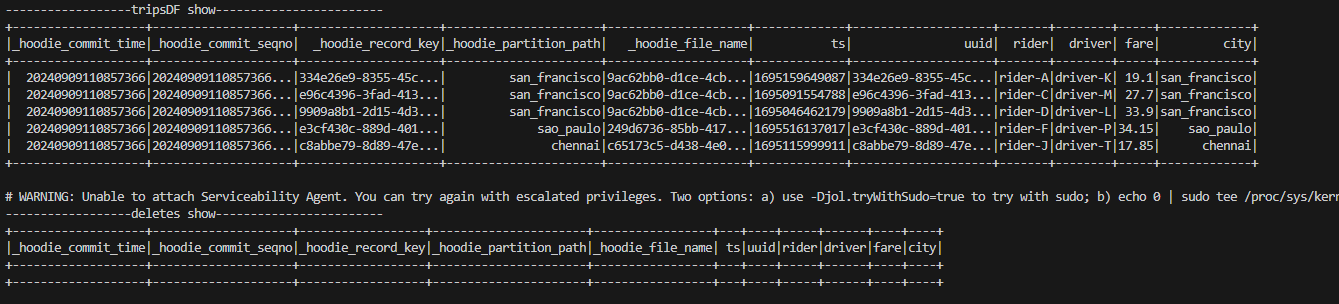


- Hard Deletes: xóa toàn bộ dữ liệu trong table, có 3 cách khác nhau:

+ sử dụng 'hoodie.datasource.write.operation': 'delete'

+ option(DataSourceWriteOptions.PAYLOAD\_CLASS\_OPT\_KEY, "org.apache.hudi.EmptyHoodieRecordPayload" )

+ "\_hoodie\_is\_deleted" : true, tạo ra 1 cột mới \_hoodie\_is\_deleted , đặt giá trị là true cho các giá trị cần xóa, và false cho các giá trị giữ lại



4.Incremental query

Tìm ra các record thay đổi trong 1 khoảng thời gian bắt đầu đến kết thúc, cung cấp mới nhất trạng thái các bản ghi trong khoảng thời gian.

'hoodie.datasource.query.type': 'incremental',

'hoodie.datasource.read.begin.instanttime': beginTime,

5.CDC query

Hỗ trợ cho cdc :

hudi\_insert\_options = {  
 'hoodie.table.name': tableName,  
 'hoodie.datasource.write.partitionpath.field': 'city',  
 'hoodie.table.cdc.enabled': 'true'  
}

cdc\_read\_options = {  
 'hoodie.datasource.query.incremental.format': 'cdc',  
 'hoodie.datasource.query.type': 'incremental',  
 'hoodie.datasource.read.begin.instanttime': 0  
}