Tencent 腾讯 | DataFun.

SPARK读写ICEBERG在腾讯的实践和优化

刘献杨 腾讯 高级工程师

目录 CONTENT

Tencent 腾讯 | DataFun.

01

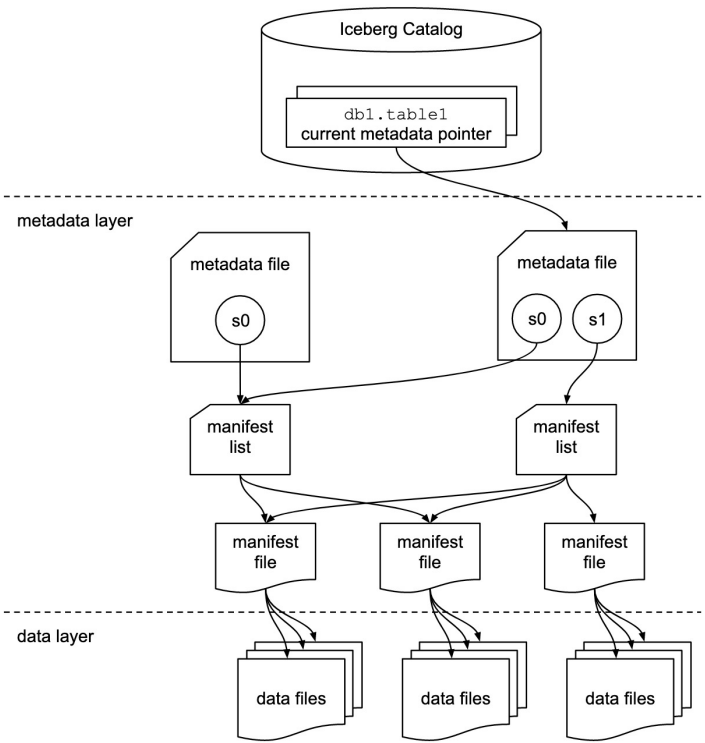
Apache Iceberg介绍

Apache Iceberg – 表格式(Table Format)

Iceberg table format that is designed to manage a large, slow-changing collection of files in a distributed file system or key-value store as a table.

- ACID
- Scale
- Evaluation
- Storage separation

Iceberg 表



Catalog: HiveCatalog, HadoopCatalog, JDBCCatalog等。
Catalog 的具体实现需要提供原子性能力，可以根据需要实现SupportNamespace接口

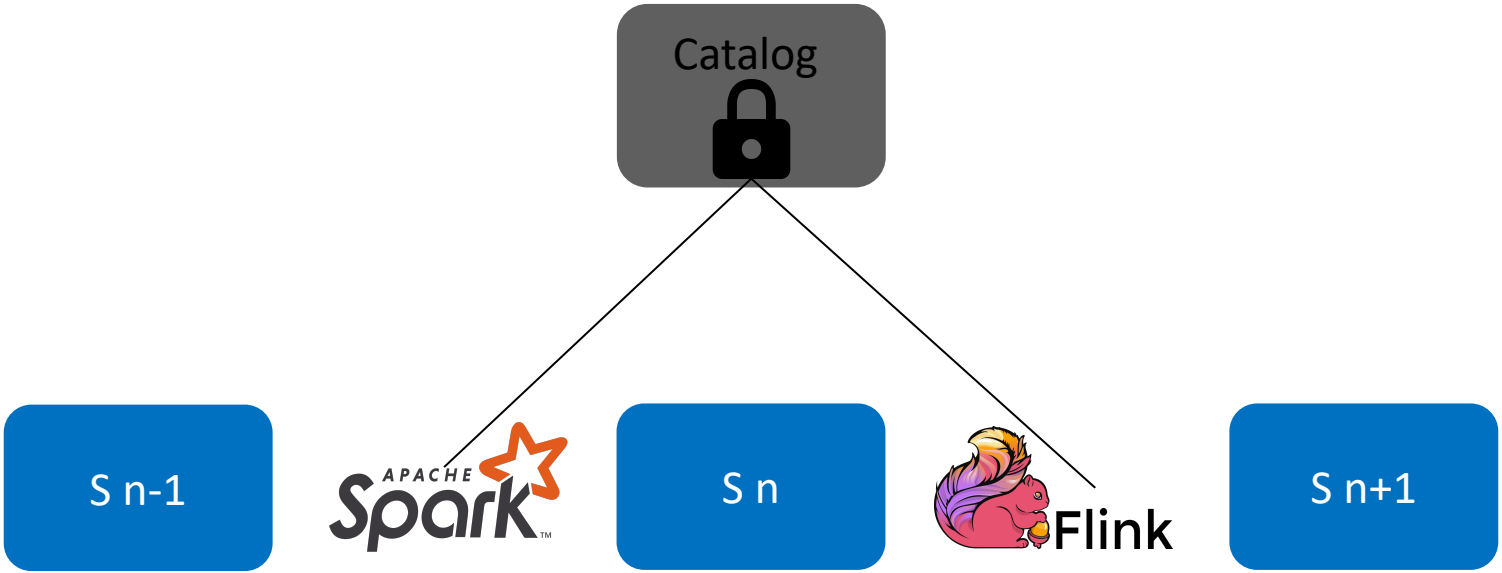
TableMetadata: 对表的更新会生成新的TableMetadata

Snapshot(ManifestList): 对表的数据更新会生成新的Snapshot

ManifestFile: DataFile的集合， Min-Max信息用于文件过滤

DataFile/DeleteFile: 存储在分布式文件系统的文件

Iceberg 表 ACID



Iceberg Evaluation

Schema Evaluation

可以并发的修改表的Schema，如增加/删除/修改列。

Partition Evaluation

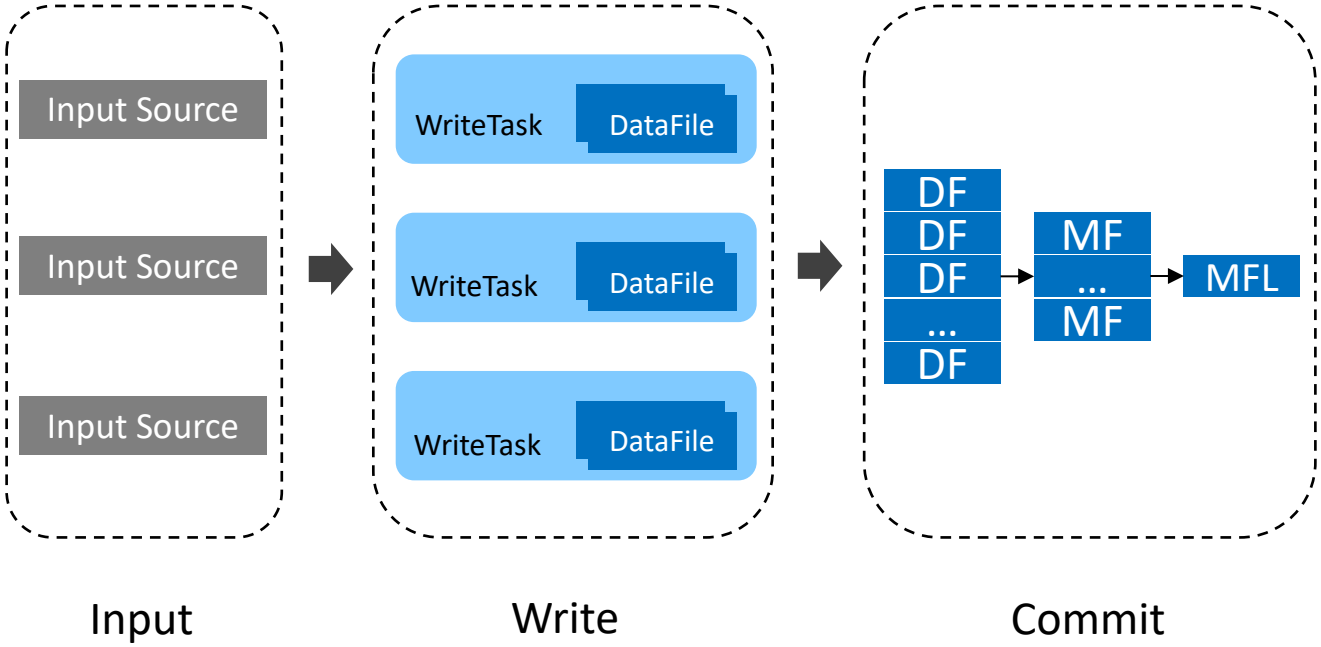
```
// 建表，partition是基于列transform得到的
CREATE TABLE iceberg.db.table (id BIGINT, created_t TIMESTAMP) USING iceberg
PARTITIONED BY (months(created_t ));
// 写入的数据只需要有表的列，不需要加上partition列
INSERT INTO TABLE iceberg.db.table VALUES …;
// 可以更新表的partition
ALTER TABLE iceberg.db.table ADD PARTITION FIELD days(created_t);
// 更新后不需要改变写入数据的格式
INSERT INTO TABLE iceberg.db.table VALUES …;
```

Tencent 腾讯 | DataFun.

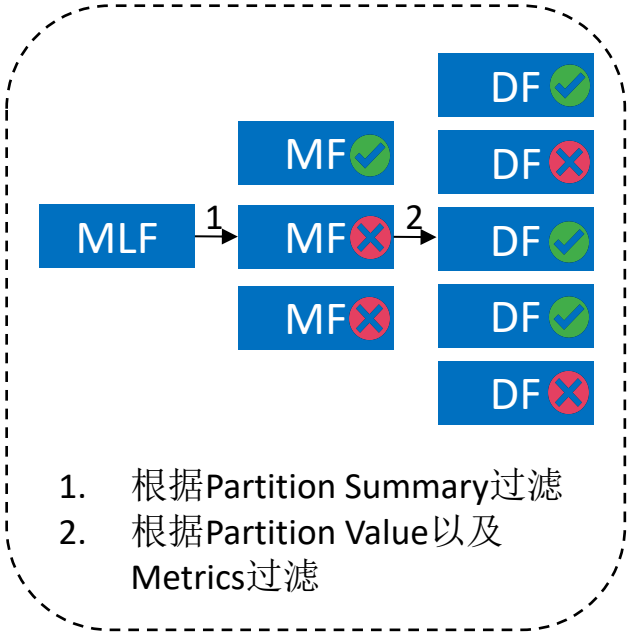
02

Spark读写Iceberg

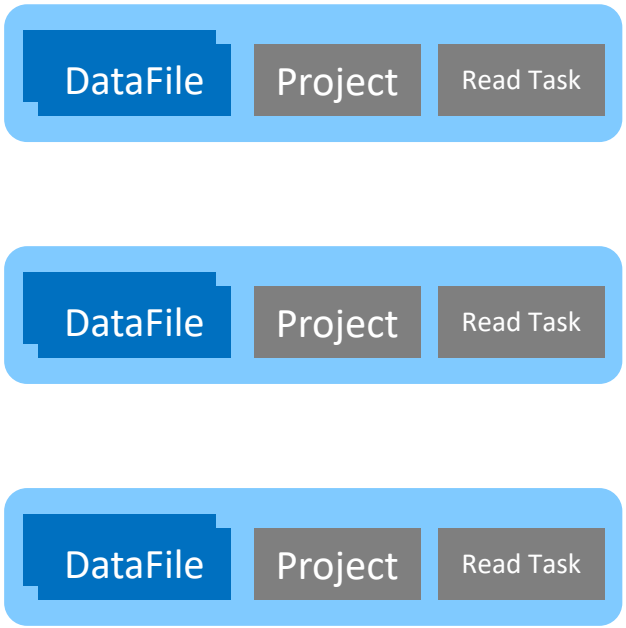
Spark 写 Iceberg 表



Spark 读 Iceberg 表



Plan Task



Execute Task

Iceberg 文件过滤

```
// 创建表，当前表的partition spec ID为0
CREATE TABLE iceberg.db.table (id BIGINT, created_t TIMESTAMP) USING iceberg PARTITIONED BY (months(created_t ));
// 写入数据，DataFile中记录文件的partition spec ID为0
INSERT INTO TABLE iceberg.db.table VALUES …;
// 更新表的partition，当前表的partition spec ID 为1
ALTER TABLE iceberg.db.table ADD PARTITION FIELD days(created_t);
// 写入数据，DataFile中记录文件的partition spec ID为1
INSERT INTO TABLE iceberg.db.table VALUES …;

SELECT * FROM iceberg.db.table where created_t = some_value;
对于 partition spec ID为0的文件，会生成额外的
  partition filter: months(created_t) = months(some_value);
对于 partition spec ID为1的文件，会生成额外的
  partition filter: months(created_t) = months(some_value) AND days(created_t) = days(some_value)
```

MOR - Position/Equality Delete

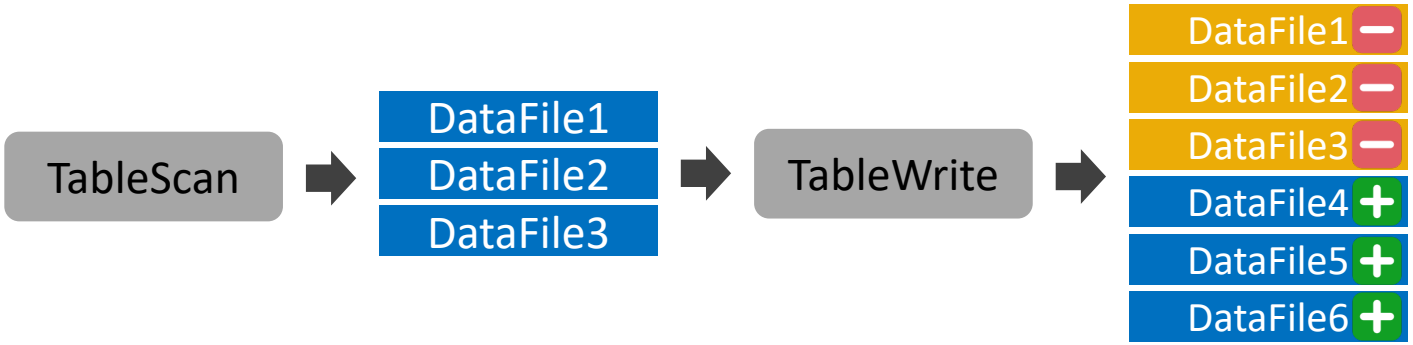
Position Delete

file_path	pos
hdfs://path/to/table/data/file1	0
hdfs://path/to/table/data/file1	19
hdfs://path/to/table/data/file2	5
hdfs://path/to/table/data/file2	9

Equality Delete

id	name
1	name1
2	name2
3	name3
4	name4

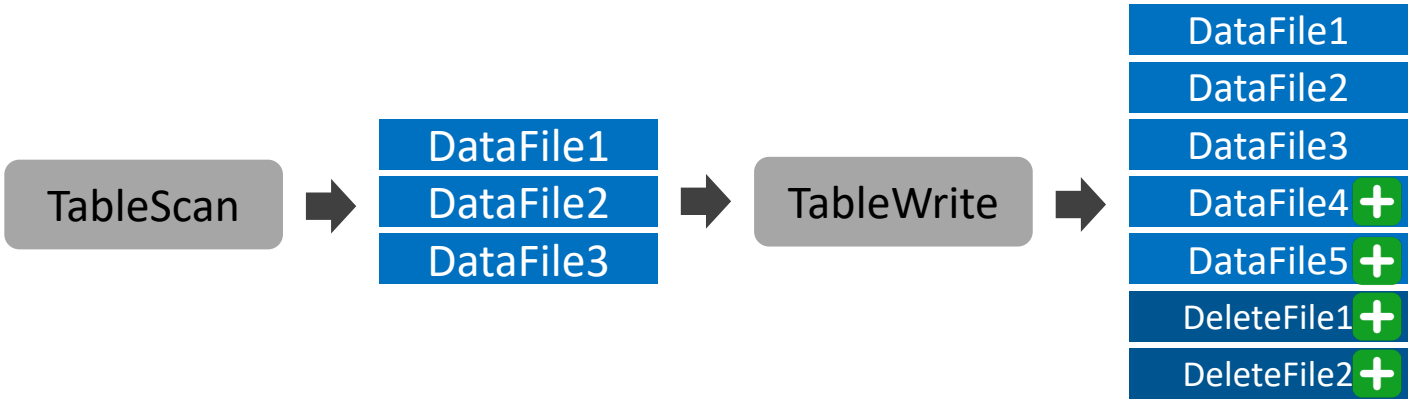
Upsert - COW



根据过滤条件找出需要更新的文件

更新生成新的DataFile

Upsert - MOR



根据过滤条件找出需要更新的文件

更新生成新的DataFile和 DeleteFile

Tencent 腾讯 | DataFun.

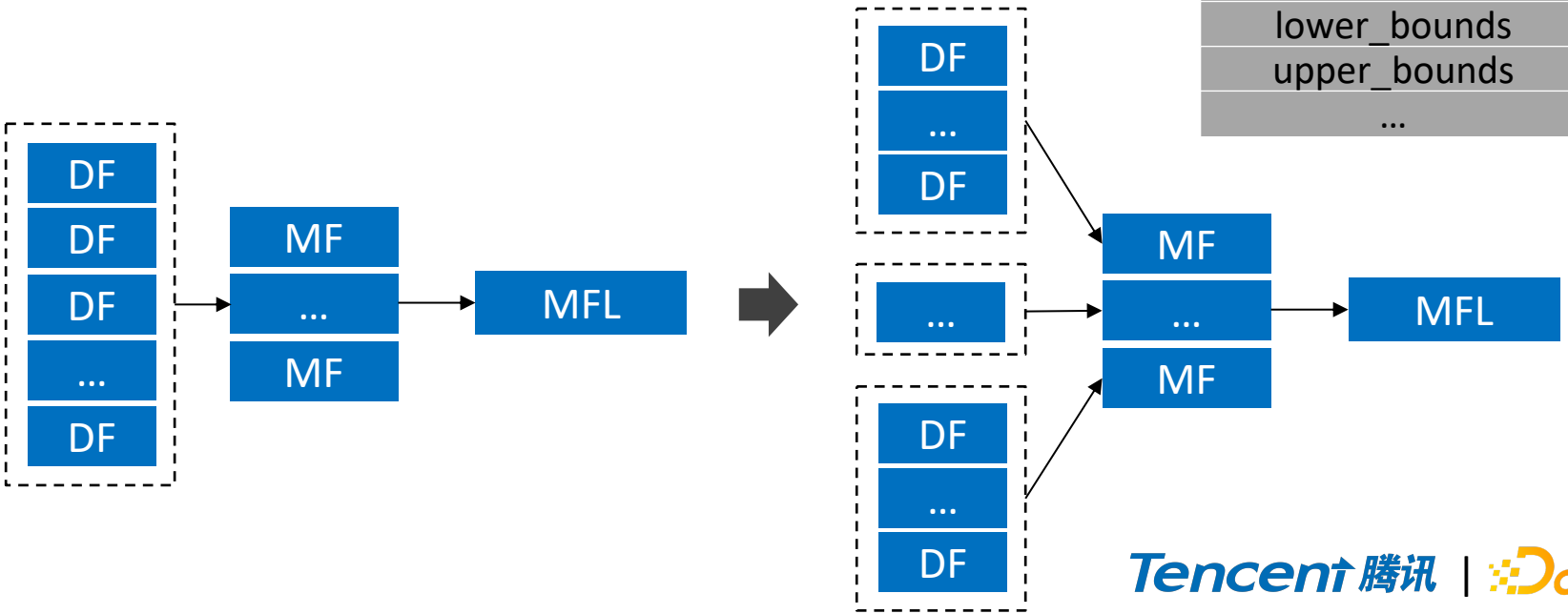
03

Iceberg生产实践

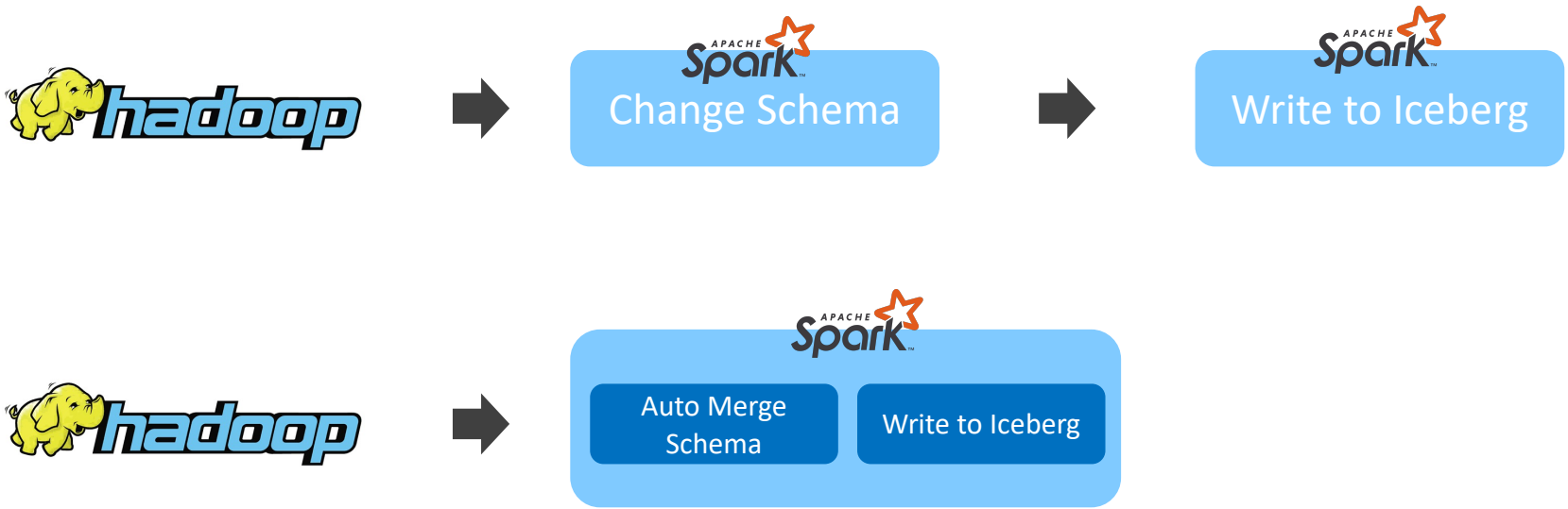
挑战1 - 宽表

1. Spark commit时会collect所有的DataFile到Driver，然后再commit到Iceberg
2. DataFile的存储空间会随着列的增加而增加
3. 可以通过以下的Table Properties来设置是否记录对应列的metrics：
   - "write.metadata.metrics.column.”
   - "write.metadata.metrics.default"

DataFile
…
column_sizes
value_counts
null_value_counts
nan_value_counts
lower_bounds
upper_bounds
…

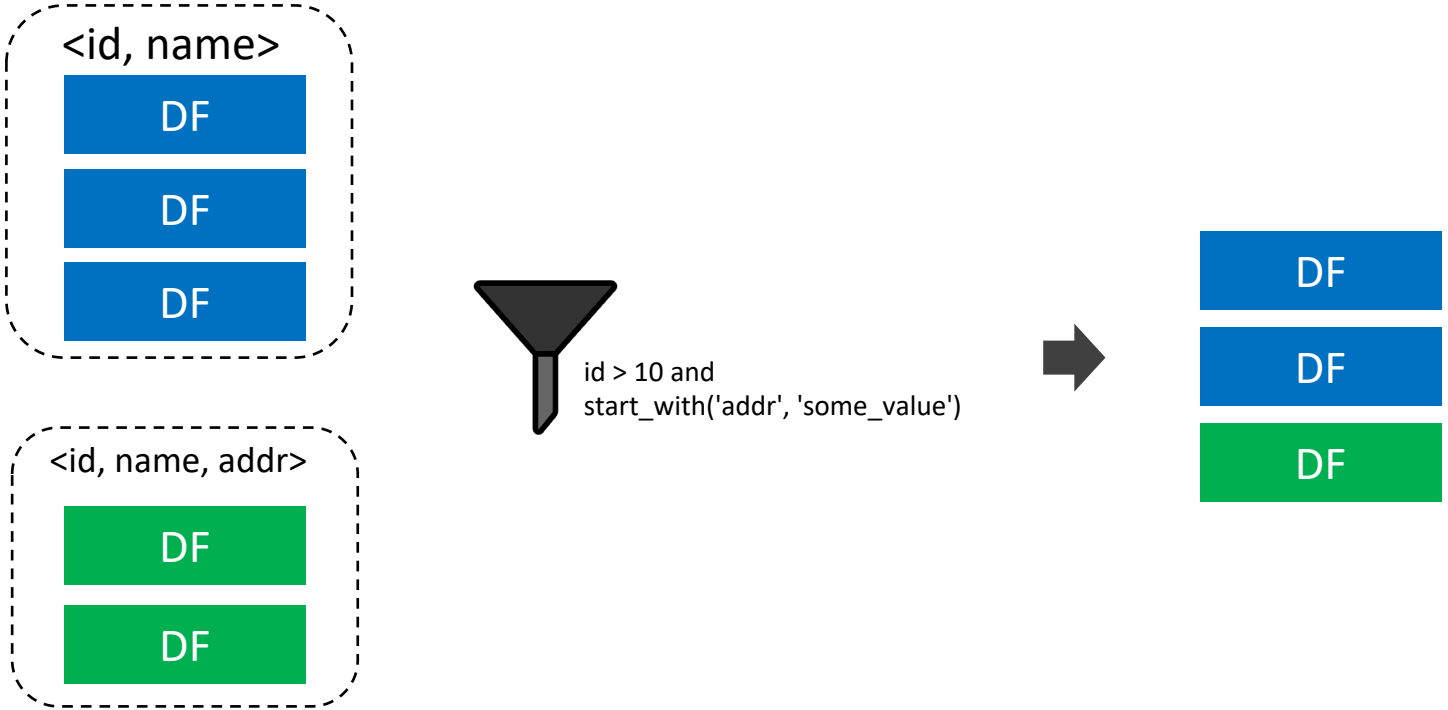


挑战2 – Schema变动频繁



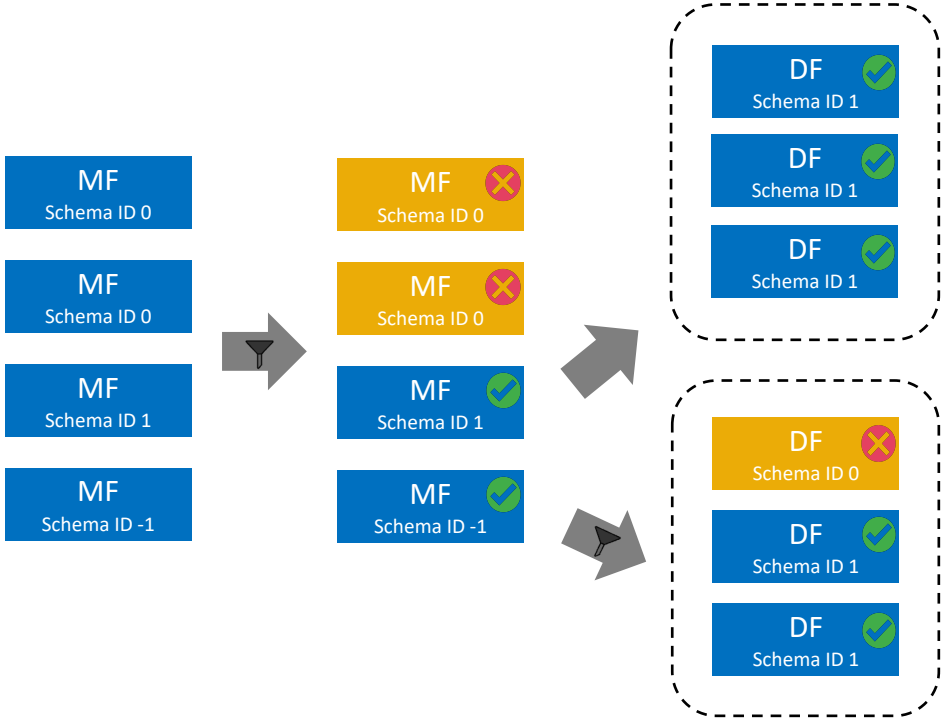
1. 设置TableProperties: “write.spark.accept-any-schema”为 true
2. df. writeTo(tableName).option("merge-schema", "true").XX

挑战3 – Schema 影响文件过滤



基于Schema过滤文件

1. 在ManifestFile和DataFile中新添加SchemaID的字段，表示写入时的Schema。
2. -1表示Schema未知
3. 首先根据SchemaID过滤ManifestFile，然后再根据需要对DataFile进行过滤



Schema ID	Schema
0	<id: long, name: str>
1	<id: long, name: str, addr: str>

一些别的优化

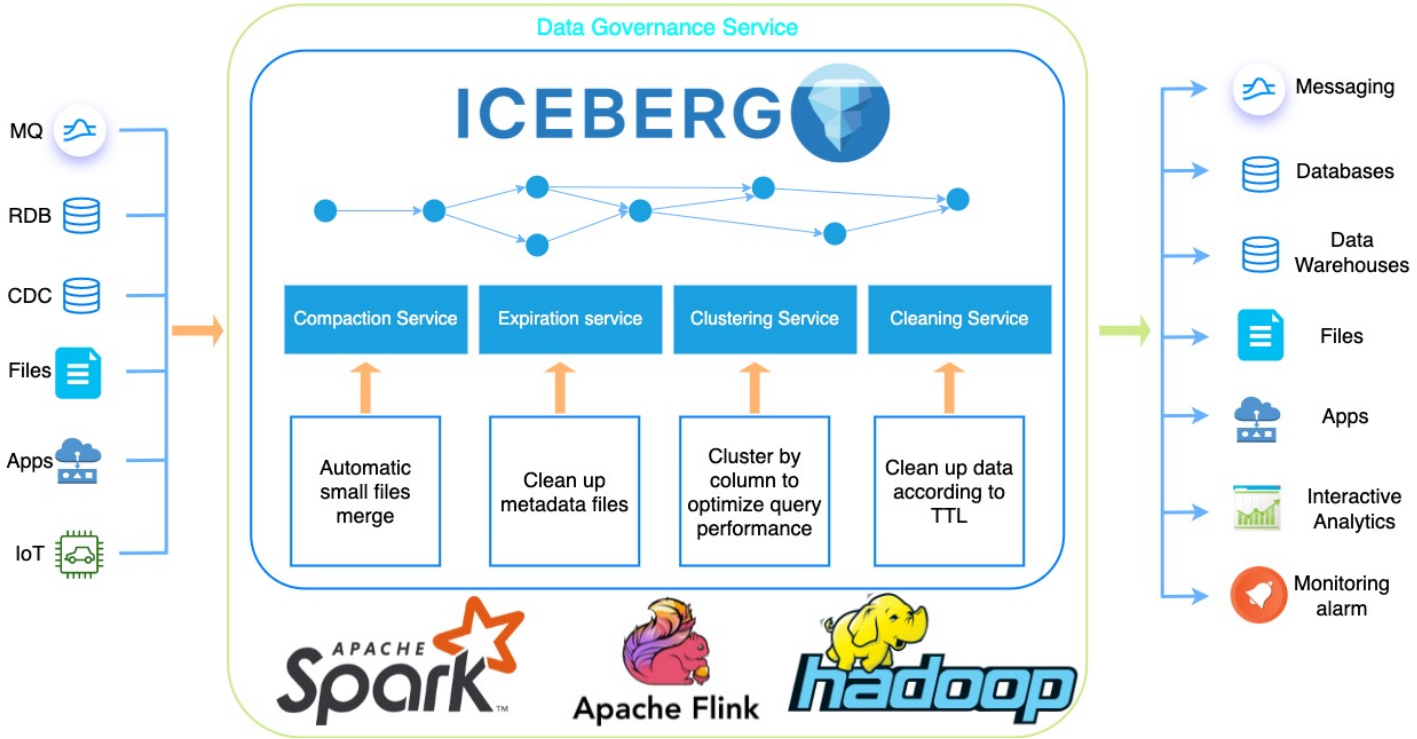
- ZOrder优化文件布局
- Parquet Bloom Filter
- Iceberg索引
- 优化Parquet Vectorized Read Decimal
- 多线程Plan Tasks，并发或者分布式的删除文件
- View的支持
- …

Tencent 腾讯 | DataFun.

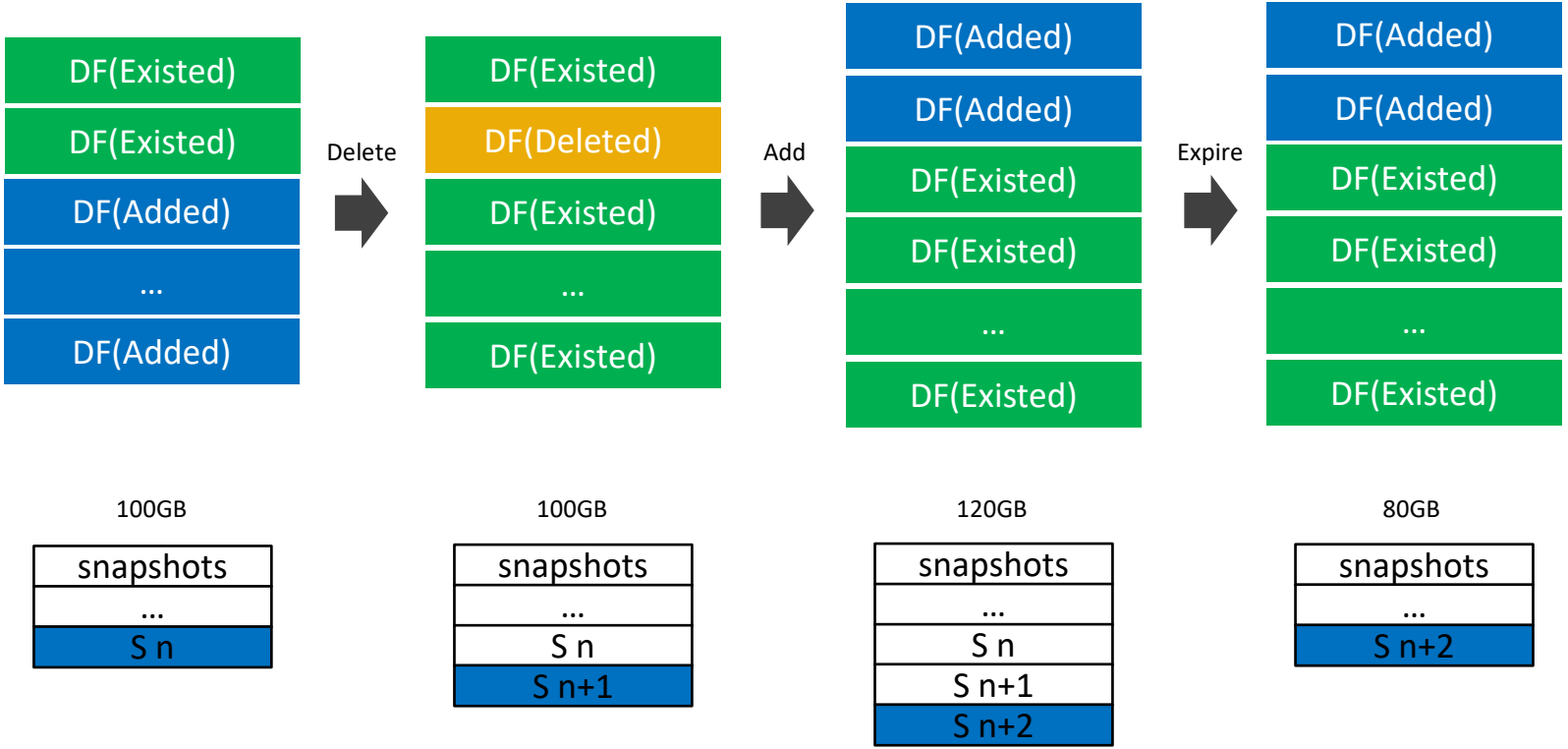
04

数据治理服务

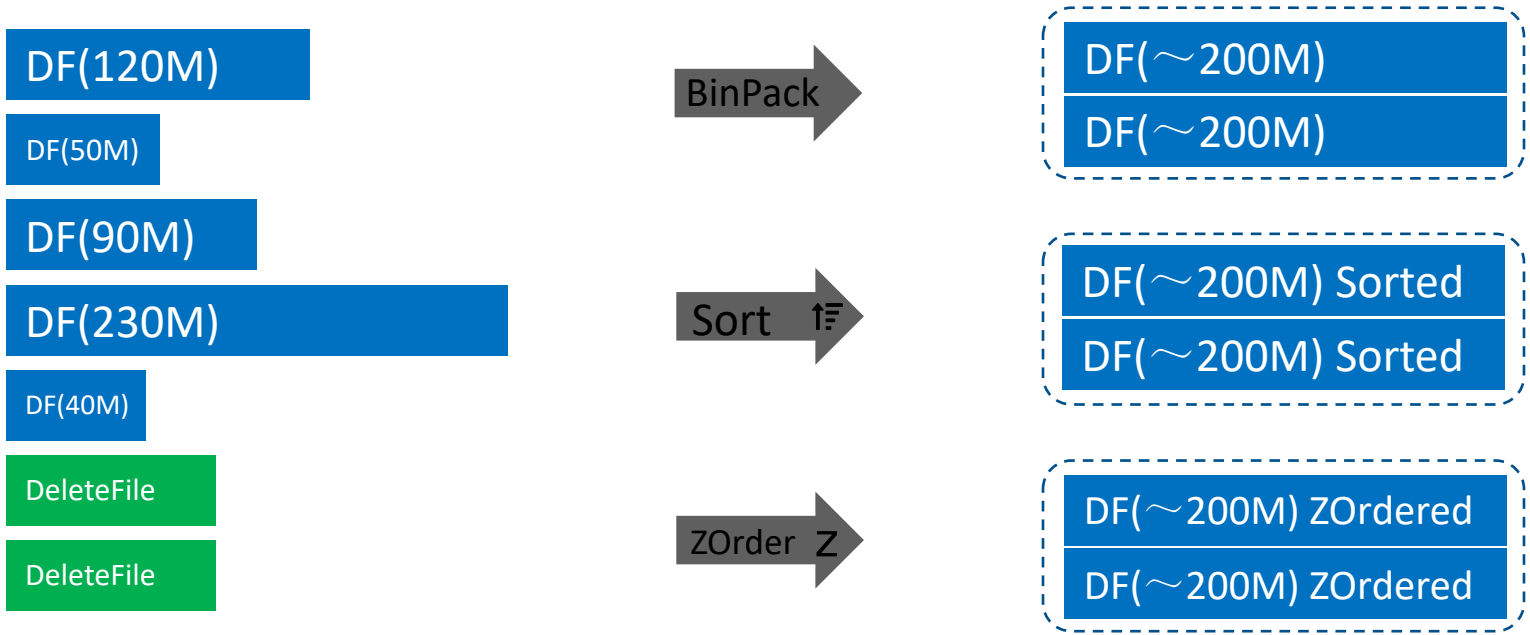
数据治理服务总览



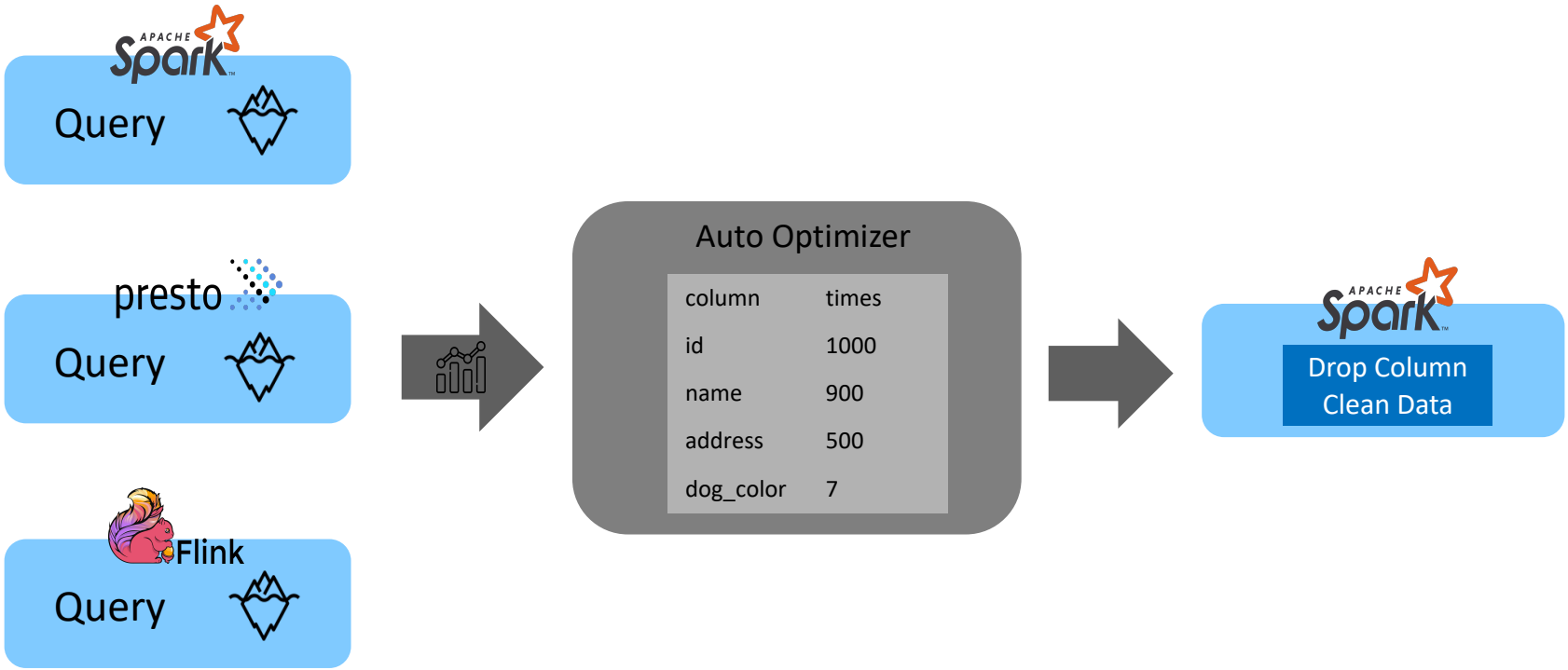
Expire Snapshots



合并小文件



列生命周期管理


Spark
Query
presto
Query
Flink
Query
Auto Optimizer
column times
id 1000
name 900
address 500
dog_color 7
Spark
Drop Column
Clean Data

非常感谢您的观看

Tencent 腾讯 | DataFun.