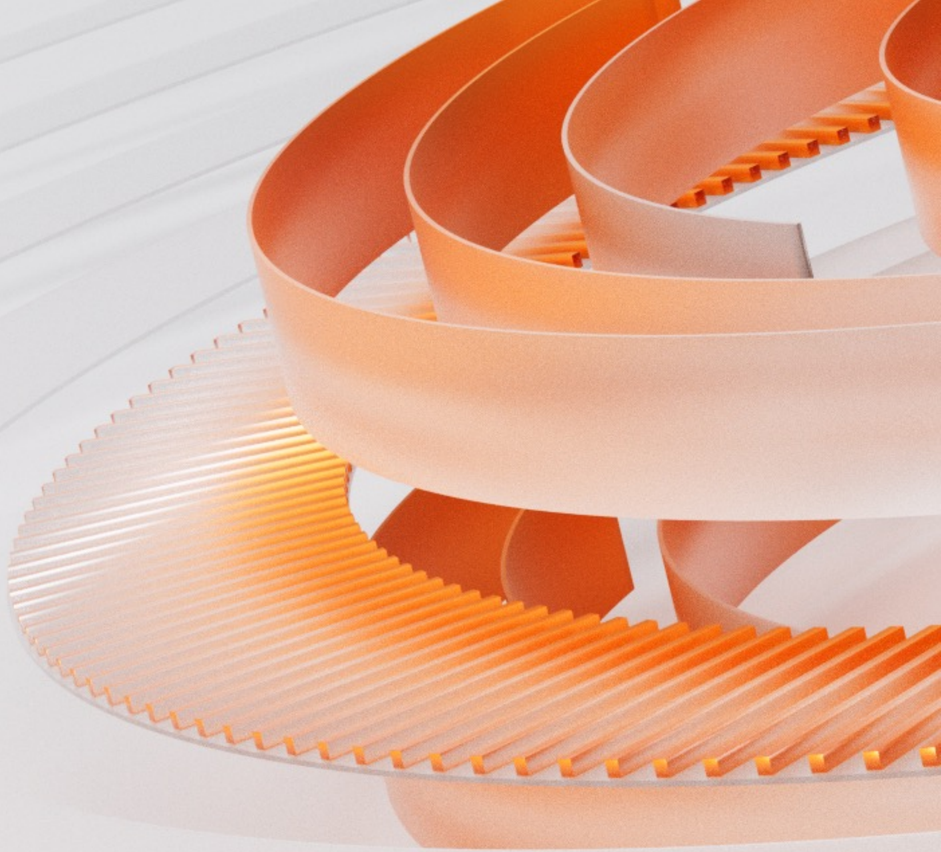
阿里云瑶池

阿里云ClickHouse最新特性介绍

分层存储、数据实时迁移



陈逸喆（白兔）

阿里云数据库OLAP ClickHouse组 开发

阿里云瑶池

分层存储

云盘多盘，本地盘，分层

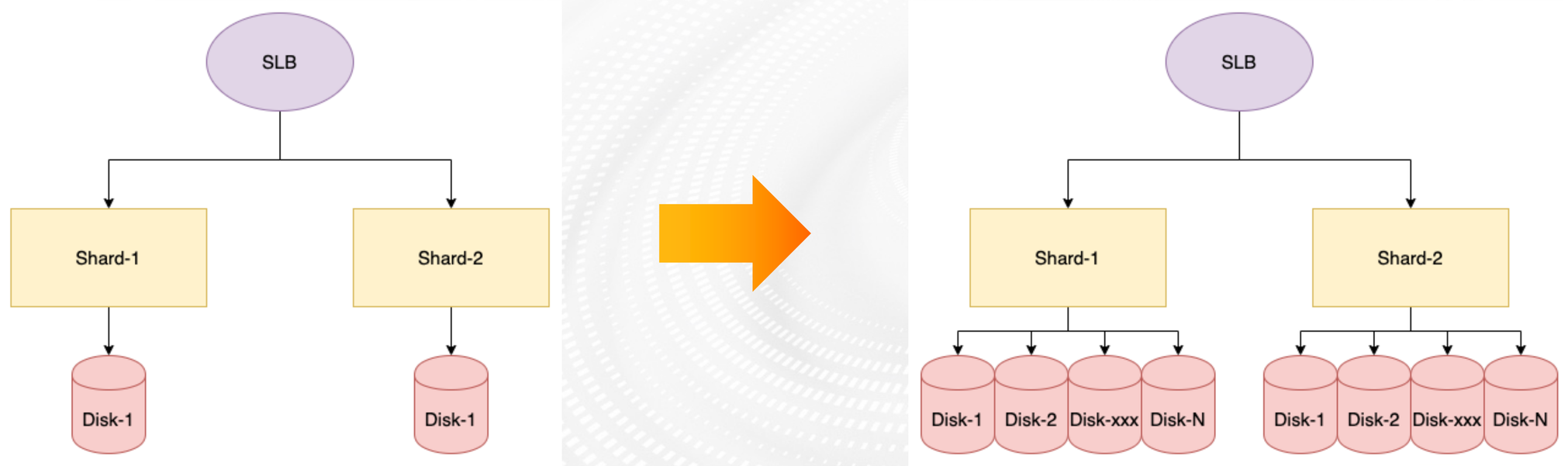
云盘多盘

解决单盘IO吞吐限制

云盘	PL3	PL2	PL1	PL0
单盘容量范围（GiB）	1,261~32,768	461~32,768	20~32,768	40~32,768
最大IOPS	1,000,000	100,000	50,000	10,000
最大吞吐（MB/s）	4,000	750	350	180

云盘多盘

根据初始购买大小，默认分配N块盘



阿里云瑶池

云盘多盘

JBOD，默认策略

```
clickhouse :) select * from system.disks;

SELECT *
FROM system.disks

Query id: 267d33a8-9fc5-4820-b113-571ffa70d5e0

┌─name────┬─path──────────────────────────┬─free_space──┬─total_space─┬─used_space─┬─keep_free_space─┬─type──┬─is_encrypted─┬─cache_path─┐
│ cdisk0  │ /clickhouse/cloud_disks/cdisk0/ │ 527278530560 │ 527295578112 │   17047552 │               0 │ local │            0 │            │
│ cdisk1  │ /clickhouse/cloud_disks/cdisk1/ │ 527278444544 │ 527295578112 │   17133568 │               0 │ local │            0 │            │
│ cdisk2  │ /clickhouse/cloud_disks/cdisk2/ │ 527278575616 │ 527295578112 │   17002496 │               0 │ local │            0 │            │
│ cdisk3  │ /clickhouse/cloud_disks/cdisk3/ │ 527278600192 │ 527295578112 │   16977920 │               0 │ local │            0 │            │
│ default │ /clickhouse/data/data/        │  20939509760 │  20957446144 │   17936384 │               0 │ local │            0 │            │
└─────────┴───────────────────────────────┴──────────────┴──────────────┴────────────┴─────────────────┴───────┴──────────────┴────────────┘

5 rows in set. Elapsed: 0.001 sec.

clickhouse :) select * from system.storage_policies;

SELECT *
FROM system.storage_policies

Query id: af9f711b-1ea8-4a99-b92a-ebb0abaa2912

┌─policy_name─┬─volume_name──────┬─volume_priority─┬─disks─────────────────────────────┬─volume_type─┬─max_data_part_size─┬─move_factor─┬─prefer_not_to_merge─┐
│ default     │ cloud_disk_volume │               1 │ ['cdisk0','cdisk1','cdisk2','cdisk3'] │ JBOD        │                  0 │           0 │                   0 │
└─────────────┴──────────────────┴─────────────────┴───────────────────────────────────┴─────────────┴────────────────────┴─────────────┴─────────────────────┘

1 row in set. Elapsed: 0.001 sec.
```

system.disks
system.storage_policies

N块盘JBOD，作为default策略

本地盘

大容量存储，性价比

性价比	HDD本地盘	高效云盘	ESSD PL1云盘
规格	20核88GB	8核32GB	8核32GB
存储	58TB	32TB	32TB
价格￥（参考）	5,452/月	11,573/月	32,373/月

本地盘

计算存储强绑定

可运维	本地盘	云盘
扩容磁盘	○	●
垂直升降规格	○	●
横向扩缩节点	●	●
数据可靠性	○	●

阿里云瑶池

本地盘

一小块云盘存放库表元数据

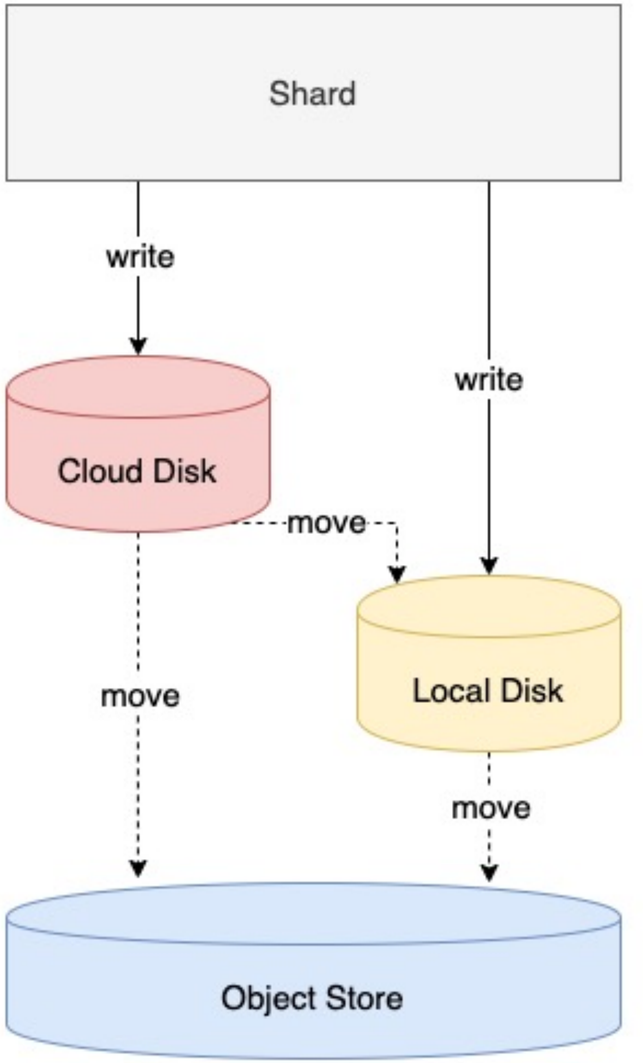
```
SELECT *
FROM system.disks

Query id: ef0b87a9-bcec-4bb6-a307-f6b227661f6f

┌─name────┬─path──────────────────────────┬─free_space─┬─total_space─┬─used_space─┬─keep_free_space─┬─type──┬─is_encrypted─┬─cache_path─┐
│ default │ /clickhouse/data/data/        │ 20711395328 │ 21003583488 │  292188160 │               0 │ local │            0 │            │
│ ldisk0  │ /clickhouse/local_disks/ldisk0/ │ 7696921047040 │ 7775441862656 │ 78520815616 │               0 │ local │            0 │            │
│ ldisk1  │ /clickhouse/local_disks/ldisk1/ │ 7696879992832 │ 7775441862656 │ 78561869824 │               0 │ local │            0 │            │
│ ldisk10 │ /clickhouse/local_disks/ldisk10/ │ 7696933740544 │ 7775441862656 │ 78508122112 │               0 │ local │            0 │            │
│ ldisk11 │ /clickhouse/local_disks/ldisk11/ │ 7696900304896 │ 7775441862656 │ 78541557760 │               0 │ local │            0 │            │
│ ldisk12 │ /clickhouse/local_disks/ldisk12/ │ 7696924774400 │ 7775441862656 │ 78517088256 │               0 │ local │            0 │            │
│ ldisk13 │ /clickhouse/local_disks/ldisk13/ │ 7696921538560 │ 7775441862656 │ 78520324096 │               0 │ local │            0 │            │
│ ldisk14 │ /clickhouse/local_disks/ldisk14/ │ 7696885157888 │ 7775441862656 │ 78556704768 │               0 │ local │            0 │            │
│ ldisk2  │ /clickhouse/local_disks/ldisk2/ │ 7696930594816 │ 7775441862656 │ 78511267840 │               0 │ local │            0 │            │
│ ldisk3  │ /clickhouse/local_disks/ldisk3/ │ 7696936280064 │ 7775441862656 │ 78505582592 │               0 │ local │            0 │            │
│ ldisk4  │ /clickhouse/local_disks/ldisk4/ │ 7696897437696 │ 7775441862656 │ 78544424960 │               0 │ local │            0 │            │
│ ldisk5  │ /clickhouse/local_disks/ldisk5/ │ 7696910282752 │ 7775441862656 │ 78531579904 │               0 │ local │            0 │            │
│ ldisk6  │ /clickhouse/local_disks/ldisk6/ │ 7696874749952 │ 7775441862656 │ 78567112704 │               0 │ local │            0 │            │
│ ldisk7  │ /clickhouse/local_disks/ldisk7/ │ 7696908640256 │ 7775441862656 │ 78533222400 │               0 │ local │            0 │            │
│ ldisk8  │ /clickhouse/local_disks/ldisk8/ │ 7696882249728 │ 7775441862656 │ 78559612928 │               0 │ local │            0 │            │
│ ldisk9  │ /clickhouse/local_disks/ldisk9/ │ 7696926117888 │ 7775441862656 │ 78515744768 │               0 │ local │            0 │            │
└─────────┴───────────────────────────────┴────────────┴─────────────┴────────────┴─────────────────┴───────┴──────────────┴────────────┘

16 rows in set. Elapsed: 0.002 sec.

clickhouse :) select * from system.storage_policies\G

SELECT *
FROM system.storage_policies

Query id: 198f67e5-0309-4cb6-88b5-08b3a6885577

Row 1:
──────
policy_name:          default
volume_name:          local_disk_volume
volume_priority:      1
disks:                ['ldisk0','ldisk1','ldisk2','ldisk3','ldisk4','ldisk5','ldisk6','ldisk7','ldisk8','ldisk9','ldisk10','ldisk11','ldisk12','ldisk13','ldisk14']
volume_type:          JBOD
max_data_part_size:   0
move_factor:          0
prefer_not_to_merge: 0

1 row in set. Elapsed: 0.001 sec.
```

system.disks
system.storage_policies

N块盘JBOD，作为default策略

分层

热、温、冷



分层组合	场景
云盘+OSS	冷热，一般TTL归档进冷存
云盘+本地盘	冷温，最近N天数据频繁查询，TTL进本地盘
云盘+本地盘+OSS	冷热温全组合

阿里云瑶池

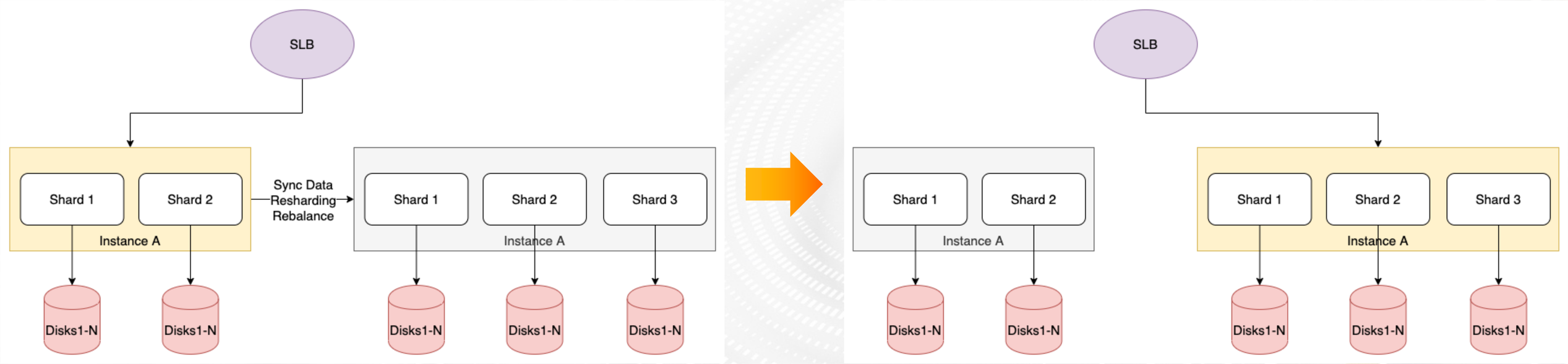
数据实时迁移

横向扩缩容节点，大版本升级，实例迁移， 单双副本切换，cksync

阿里云瑶池

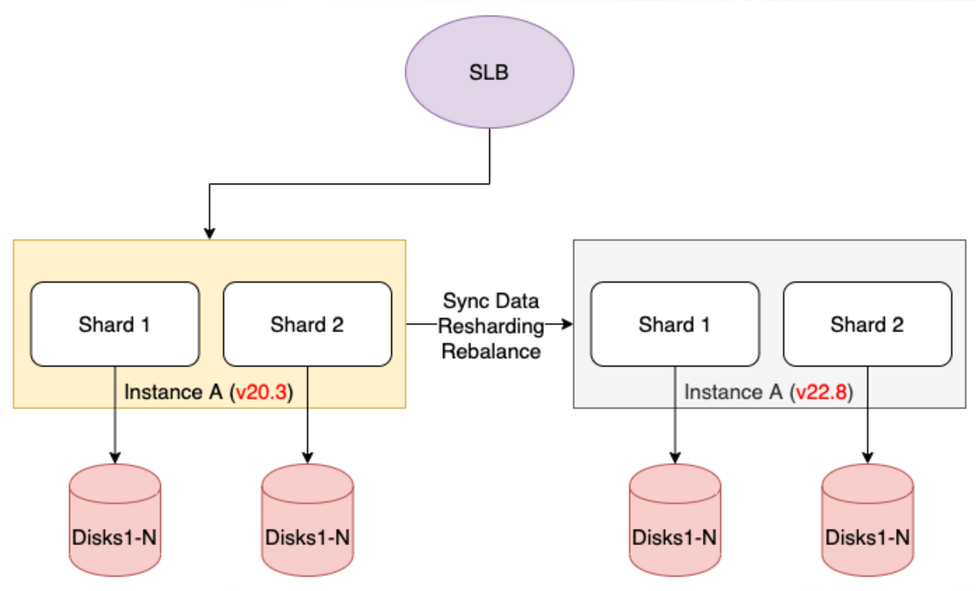
横向节点扩缩容

创建临时实例，数据同步后，切换SLB

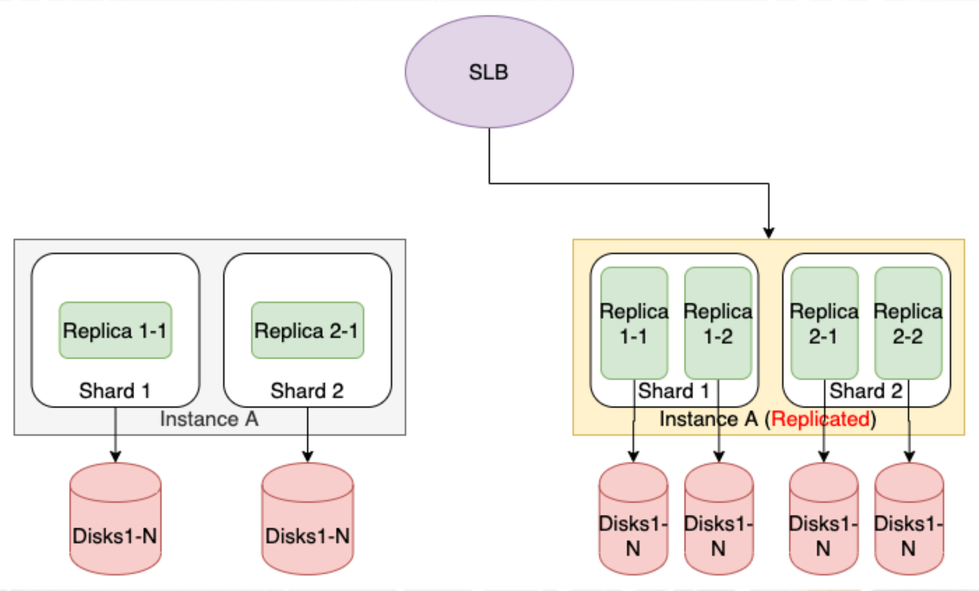


大版本升级，单双副本切换

同一套流程



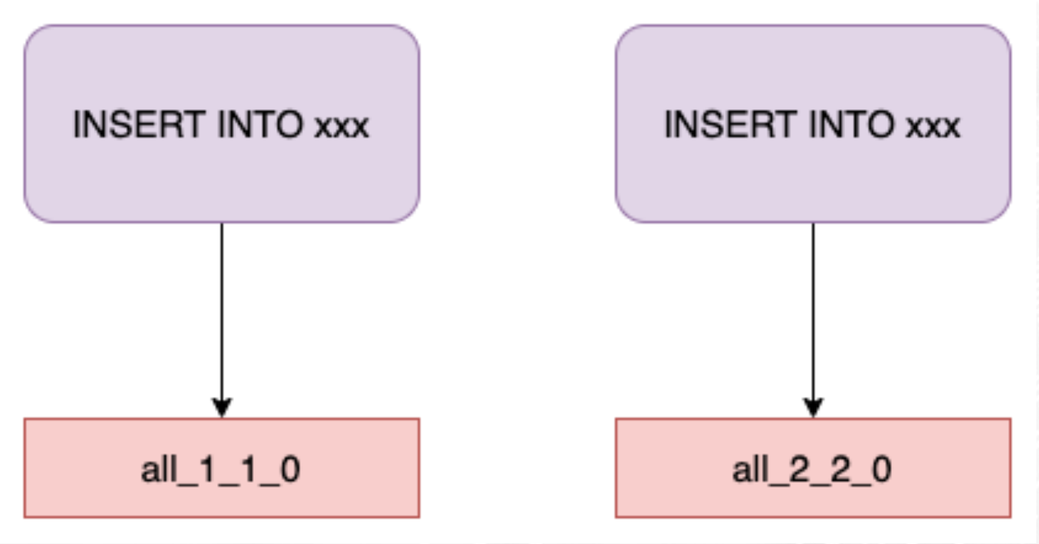
兼容性考虑，通过数据迁移方式升级大版本风险较小



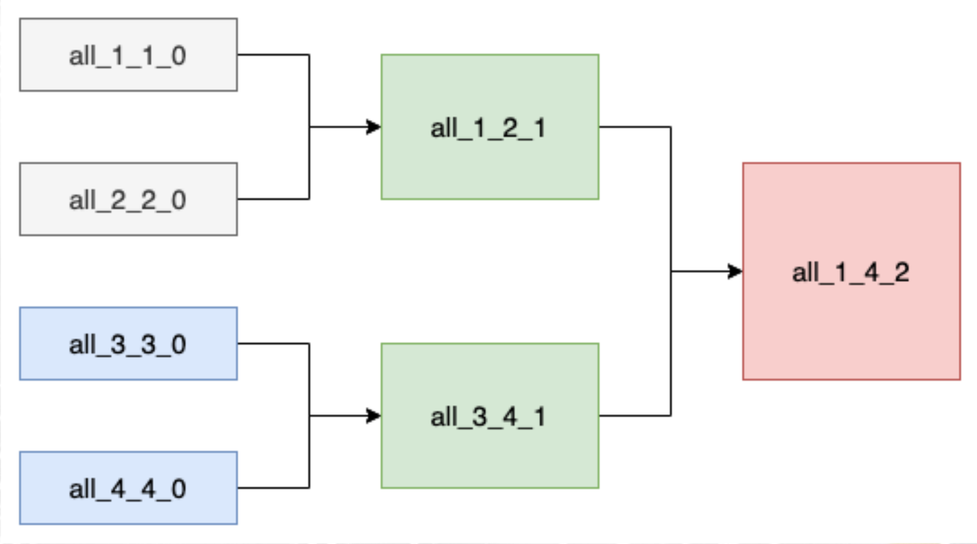
表引擎XxxMergeTree -> ReplicatedXxxMergeTree，自动变换

数据实时同步难点

Data Part不定时Merge



每次 INSERT 产生一个Data Part

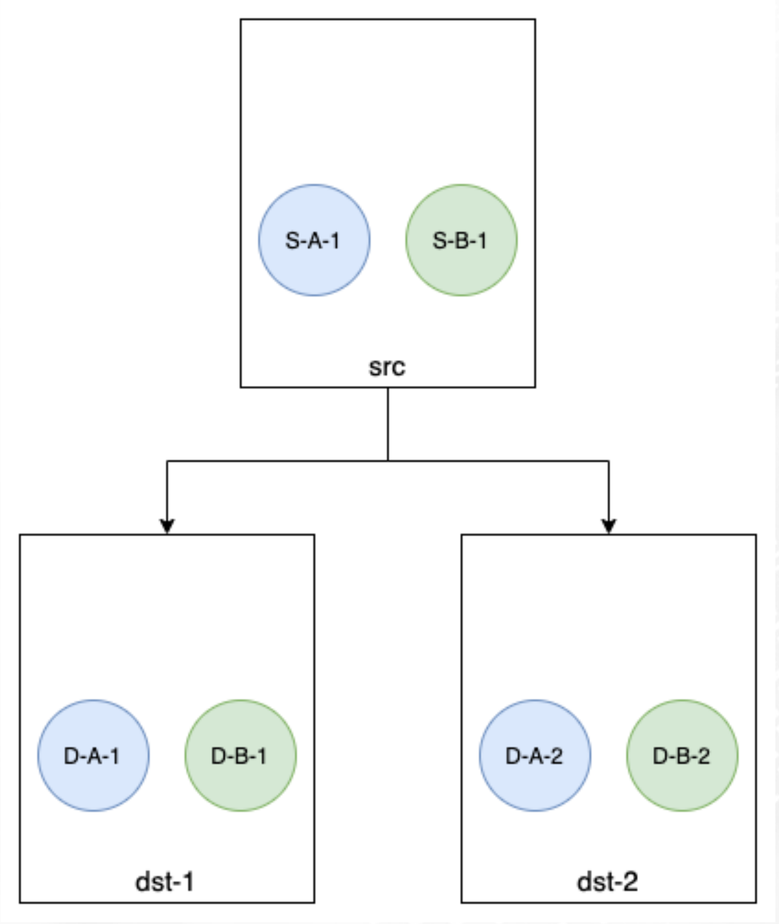


Data Part在持续Merge

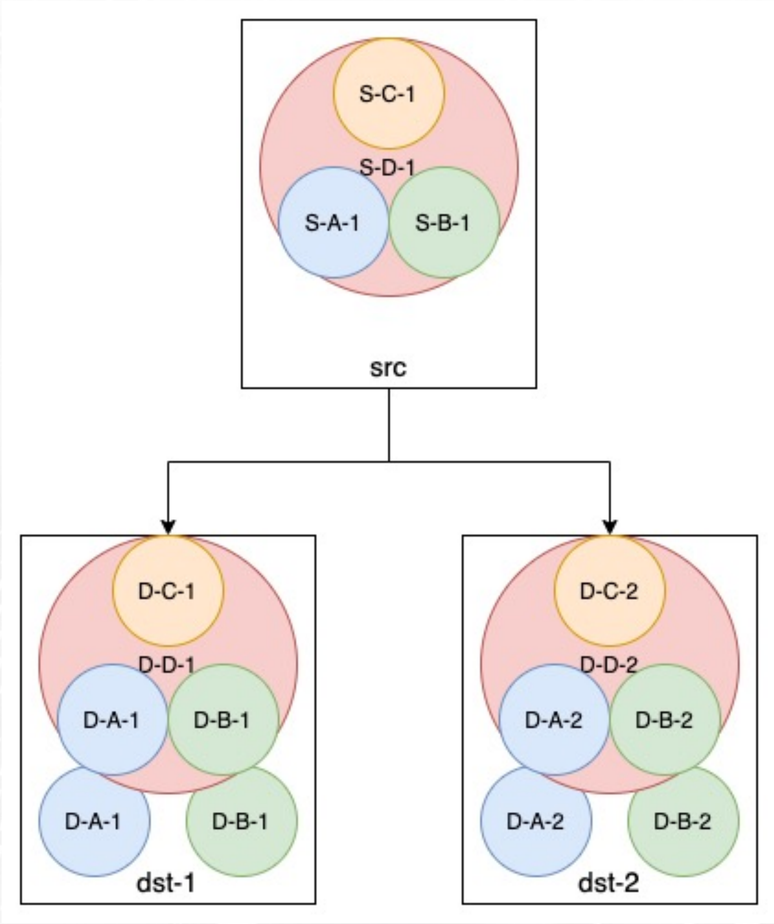
阿里云瑶池

一种数据实时同步方案

追踪Data Part轨迹



- 从1个src节点同步到2个dst节点
- src一个Data Part经过Resharding后，分别落到2个dst节点上

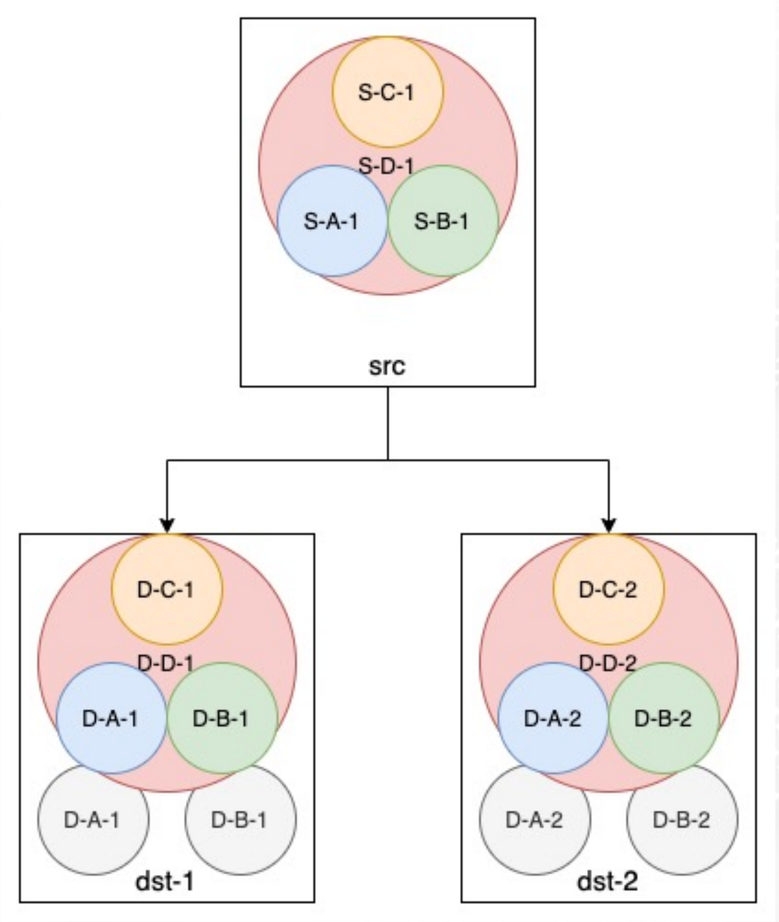


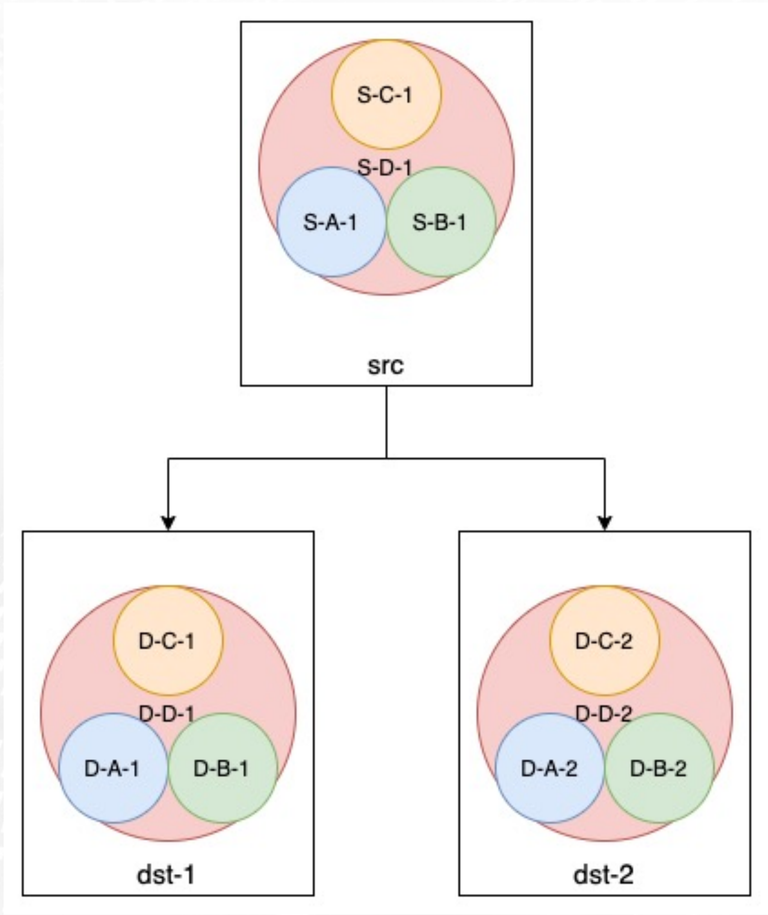
- 新的S-C-1写入后，和原来2个Data Part合并成一个全新的S-D-1
- 同步后，2个dst节点上，出现数据重复

阿里云瑶池

一种数据实时同步方案

Rollback操作，让数据一致





S-D-1 ← [S-A-1, S-B-1, S-C-1]

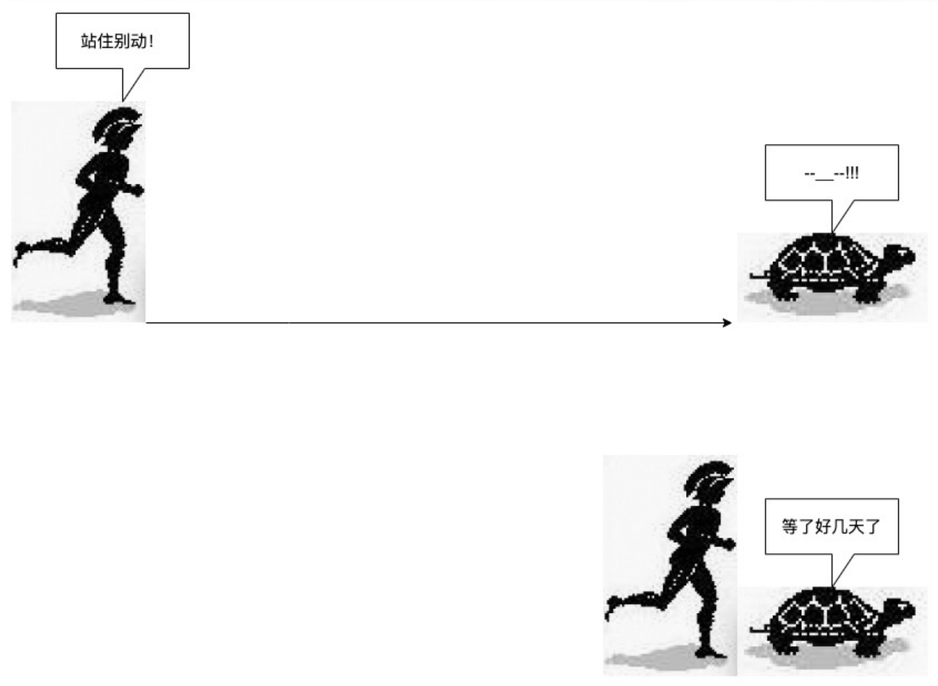
S-A-1 → [D-A-1, D-B-1]

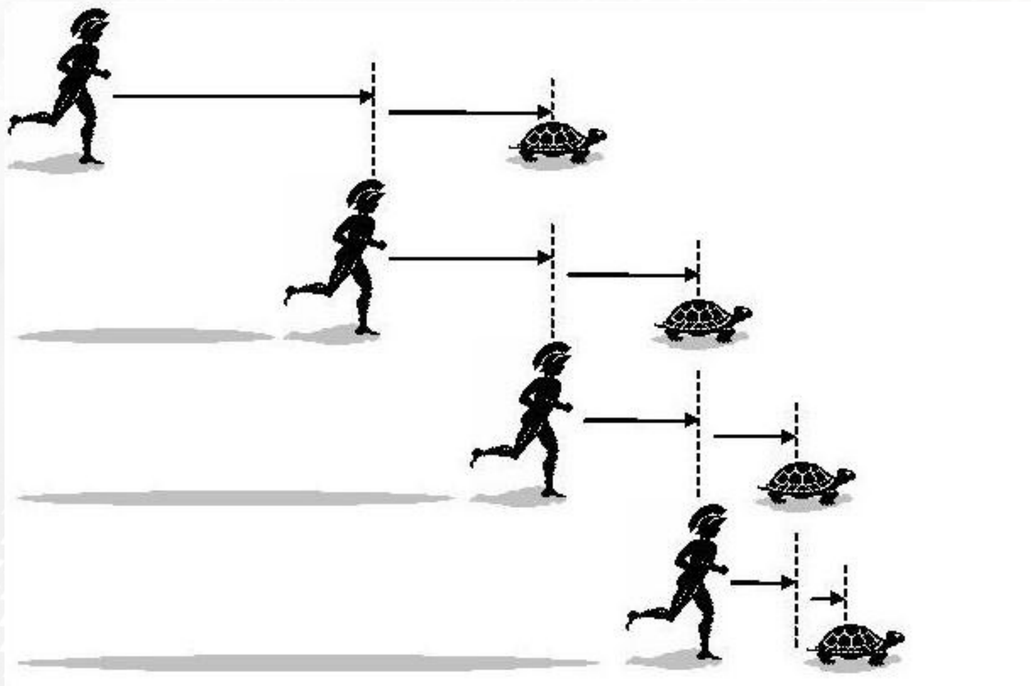
S-B-1 → [D-A-2, D-B-2]

通过追踪记录Data Part的Merge轨迹，以及Data Part的Resharding轨迹，进行Rollback操作，对数据去重，让数据一致

一种数据实时同步方案

不完美，有限制，但能解决完全锁写的业务痛点





完全锁写，根据数据量大小，可能需要几小时甚至几天

Tradeoff：

- 目标实例必须暂停Merge
- Rollback操作可能让同步效率降低
- 最后可能还是需要短暂锁写，让数据同步追上

阿里云瑶池

一种数据实时同步方案

Data Part追踪

```
query_id:
event_type:                MergeParts
event_date:                2022-09-09
event_time:                2022-09-09 16:55:55
event_time_microseconds:   2022-09-09 16:55:55.948686
duration_ms:               267
database:                  cksync_test
table:                     lineorder
part_name:                 1992_0_5_1
partition_id:              1992
path_on_disk:              /clickhouse/data/data/store/3f0/3f0daec1-1ce5-47ae-bf0d-aec11ce537ae/1992_0_5_1/
rows:                      474648
size_in_bytes:             13971450
merged_from:               ['1992_0_0_0','1992_1_1_0','1992_2_2_0','1992_3_3_0','1992_4_4_0','1992_5_5_0']
bytes_uncompressed:        20421348
read_rows:                 474648
read_bytes:                20421744
peak_memory_usage:         25505882
error:                     0
exception:
```

system.part_log

利用part_log的MergeParts，递归追踪Data Part合并轨迹

```
SELECT
    _shard_num,
    groupArray(part_name) AS parts,
    sum(rows) AS rows
FROM clusterAllReplicas(default, system, part_log)
WHERE query_id IN (
    SELECT query_id
    FROM clusterAllReplicas(default, system, query_log)
    WHERE (initial_query_id = '545e26cf-bb2f-4339-82bb-a8f91b8e93cc') AND (type = 'QueryFinish') AND (written_rows > 0)
)
GROUP BY _shard_num

Query id: 2b0db6ef-b831-4cee-99d8-6af6da2626e2

┌─_shard_num─┬─parts──────────┬─rows─┐
│          2 │ ['1992_3_3_0'] │ 474648 │
└────────────┴────────────────┴──────┘
```

qury_log + part_log

利用query_log获取一次INSERT的所有query_id，结合part_log获取该query_id在所有节点上的Data Part记录

阿里云瑶池

cksync

ClickHouse数据实时同步工具

```
NAME:
   cksync - clickhouse sync tool

USAGE:
   cmd [global options] command [command options] [arguments...]

VERSION:
   1.0.0

COMMANDS:
   check-before       check before sync start
   check-after        check after sync end
   sync-ddl-before    sync ddl before sync start
   sync-ddl-after     sync ddl after sync end
   sync-ddl           sync all ddl
   metadata-clear     clear metadata
   metadata-progress  get metadata progress
   sync               sync data
   help, h            Shows a list of commands or help for one command

GLOBAL OPTIONS:
   --debug, -d                 debug (default: false) [$CKSYNC_DEBUG]
   --log-file value, -l value  log file [$CKSYNC_LOG_FILE]
   --help, -h                  show help (default: false)
   --version, -v               print the version (default: false)
```

特性	cksync	clickhouse-copier
语言	Go	C++
资源占用	很小	相对较大
上手使用	相对简单	相对复杂
并发模型	相对简单	相对复杂
write-lock-free	●	
临时表机制		●

- 完全依赖ClickHouse SQL能力，无侵入实现
- 同步效率取决于新老实例负载

INFO[2022-11-07 11:27:12] cksync stats: shards[5] tables[295] parts[6304] rows[28920872741] bytes[1348936058976] elapsed[28m59.097303732s] rate[739.72MB/s]

阿里云瑶池