支付宝 Alipay.com 知托付

OceanBase
支持ACID的可扩展关系数据库

qushan@alipay.com

2013年04月

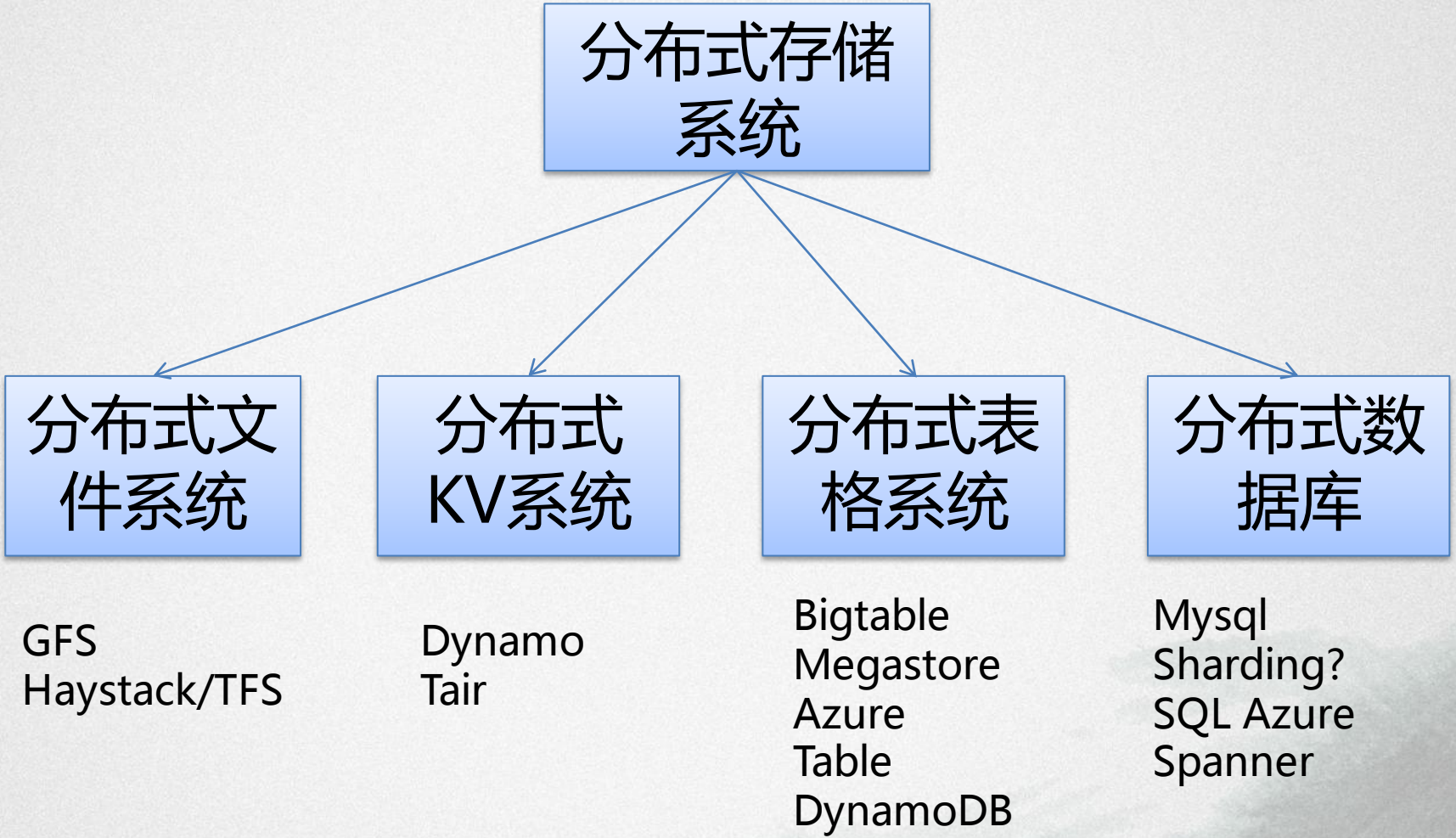
关系数据库发展

- 1970-72：E.F.Codd 数据库关系模式
- 20世纪80年代
  - 第一个商业数据库Oracle V2
  - SQL成为“数据库行业标准”
- 可扩展性
  - Mainframe：小型机 => 中型机 => 大型机
  - Sharding：全局索引? 事务? 跨库查询？
- 性能
  - Disk-based design：SSD? Memory?
- 开源数据库：主备同步？锁? Schema变更？...

互联网：云存储系统

- Google系列
  - GFS + Bigtable + Megastore
  - GFS + Bigtable + Percolator
  - GFS + Spanner
  - OLAP：Dremel、PowerDrill
- Microsoft
  - Azure Storage (Blob/Table)
  - SQL Azure
- Amazon
  - Dynamo
  - EBS、S3、DynamoDB、RDS

互联网：云存储系统分类

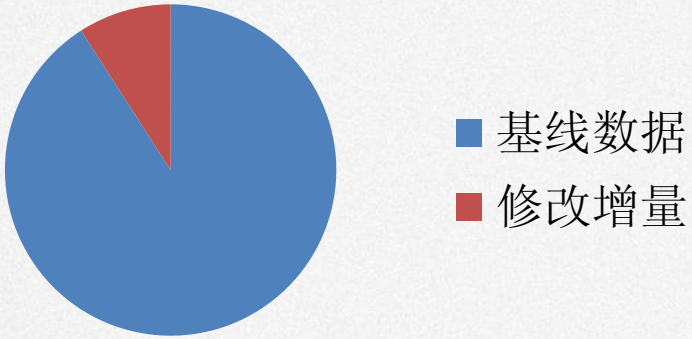

分布式存储系统
分布式文件系统
分布式KV系统
分布式表格系统
分布式数据库
GFS
Haystack/TFS
Dynamo
Tair
Bigtable
Megastore
Azure Table
DynamoDB
Mysql
Sharding?
SQL Azure
Spanner

互联网：内存数据库

- MemSQL
  - Mysql protocol compliant
  - In Memory MVCC
  - Concurrent skiplist, lock free
  - Code Gen
  - 64 Core：150W transaction per second
- VoltDB
  - Single thread, multiple process
  - Stored procedure based transaction
  - Timestamp based distributed transaction
- OLAP：SAP HANA

架构权衡：单机 or 分布式？

- Spanner vs MemSQL
- OceanBase = Bigtable + MemSQL
  - 基线数据 + 修改增量



  - 写事务(平均)：10000TPS，100Byte/事务，每天：
    - 写事务数：10000*24*60*60 = 8.64亿
    - 修改增量：8.64亿*100B = 86.4GB

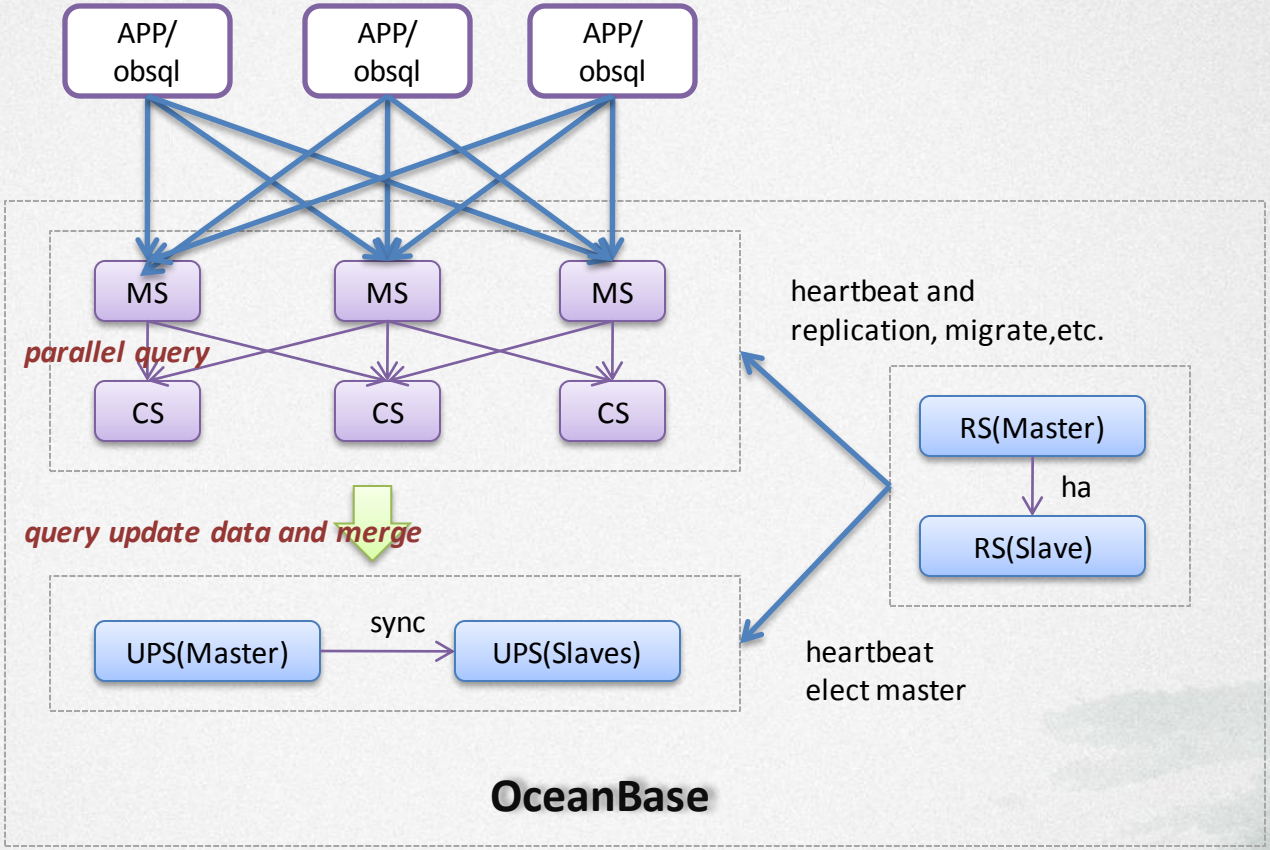
接口选择：SQL or NOSQL?

- SQL vs NOSQL
  - SQL：生态系统完善、统一标准、表达能力强、易用
  - NOSQL：无统一标准、表达能力弱
  - SQL性能差？
- OceanBase的选择
  - 标准SQL + Mysql协议
  - SQL解析开销? prepared statement
  - SQL子集：单表SQL为主，简单的多表操作，如等值连接

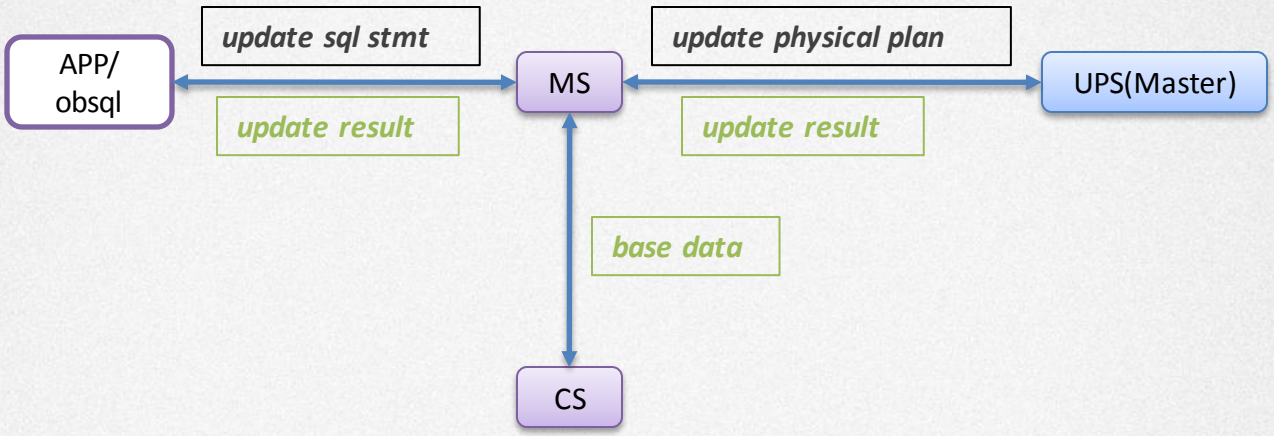
业务范畴：OLTP or OLAP?

- OLTP
  - SSD + Memory
  - 基于主键的查询 + 二级索引
  - 行式存储
- OLAP
  - SATA + Memory
  - 多机、多线程并发执行
  - 列式存储：压缩、基于压缩数据做运算
  - 数据快速导入、批量删除
- OceanBase：OLTP为主，OLAP为辅

系统架构

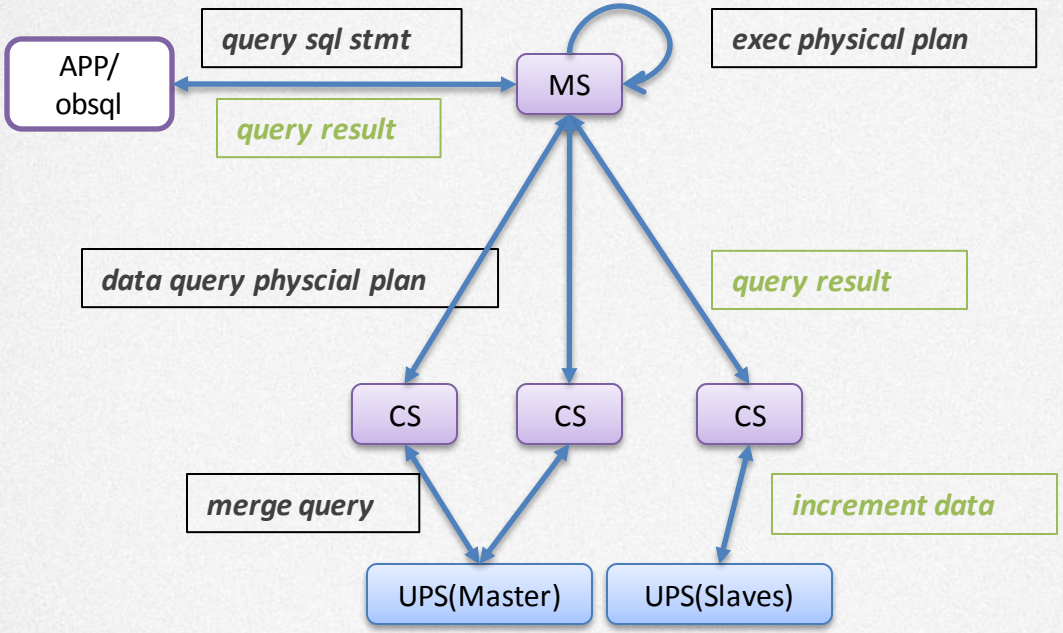

APP/
obsql
APP/
obsql
APP/
obsql
MS
MS
MS
parallel query
CS
CS
CS
query update data and merge
UPS(Master)
sync
UPS(Slaves)
heartbeat and
replication, migrate,etc.
RS(Master)
ha
RS(Slave)
heartbeat
elect master
OceanBase

写事务



MergeServer作为更新的代理接受所有的更新请求，如果需要， 会从ChunkServer查询基准数据，并将基准数据附带给UpdateServer执行更新请求。UpdateServer会将基准数据和自己的memtable中的动态数据合并以后进行判断，最终执行更新。

读事务

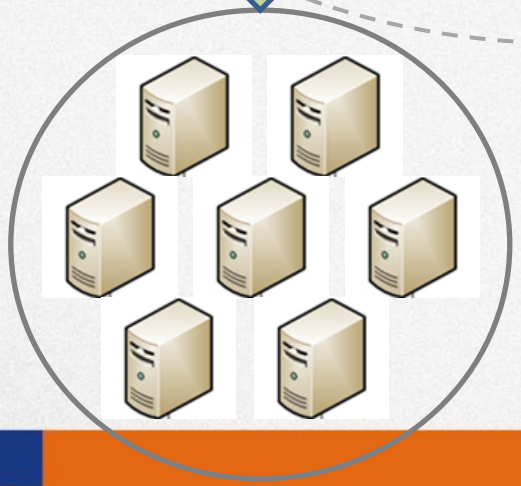


mergeserver进行sql解析，生成物理执行计划，将其中的数据操作部分计划（包括部分操作符）根据tablet划分并发下压到ChunkServer执行。

数据合并

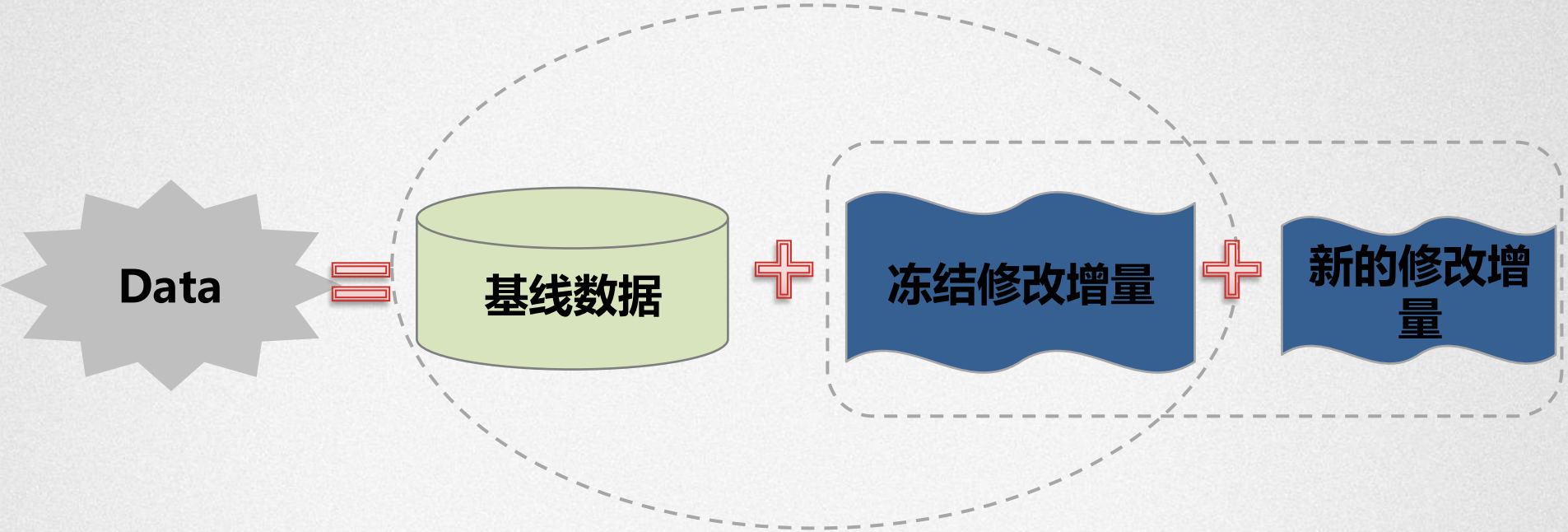
读写事务照常进行

Data = 基线数据 + 冻结修改增量 + 新的修改增量



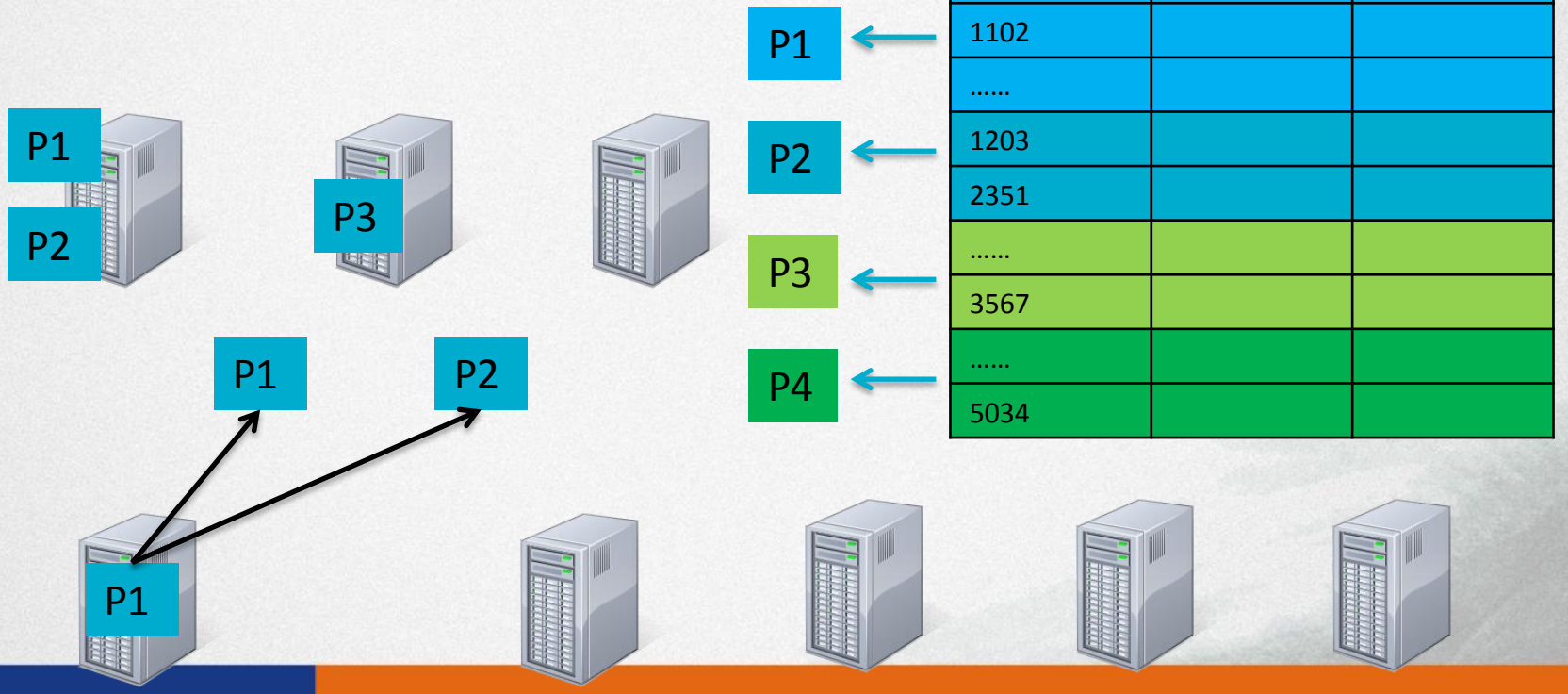
很多数据库每天有明显的访问低谷

数据合并期间的读事务

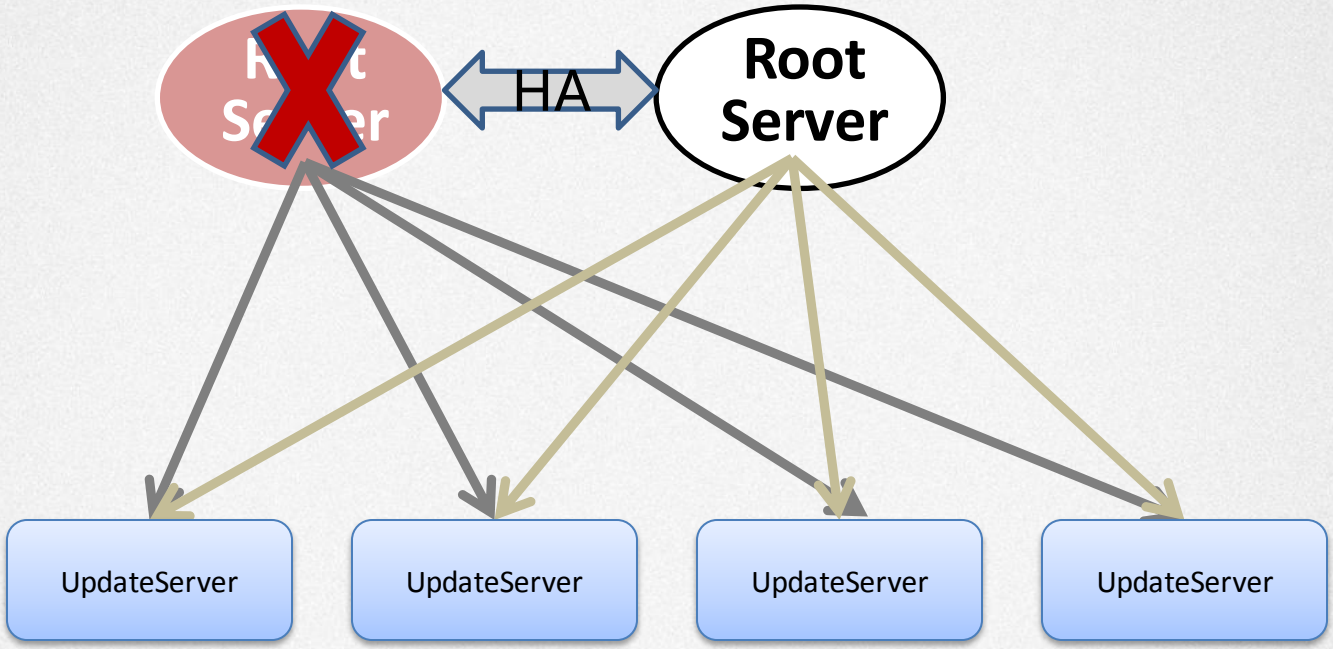


分布式存储引擎：CS/MS水平扩展

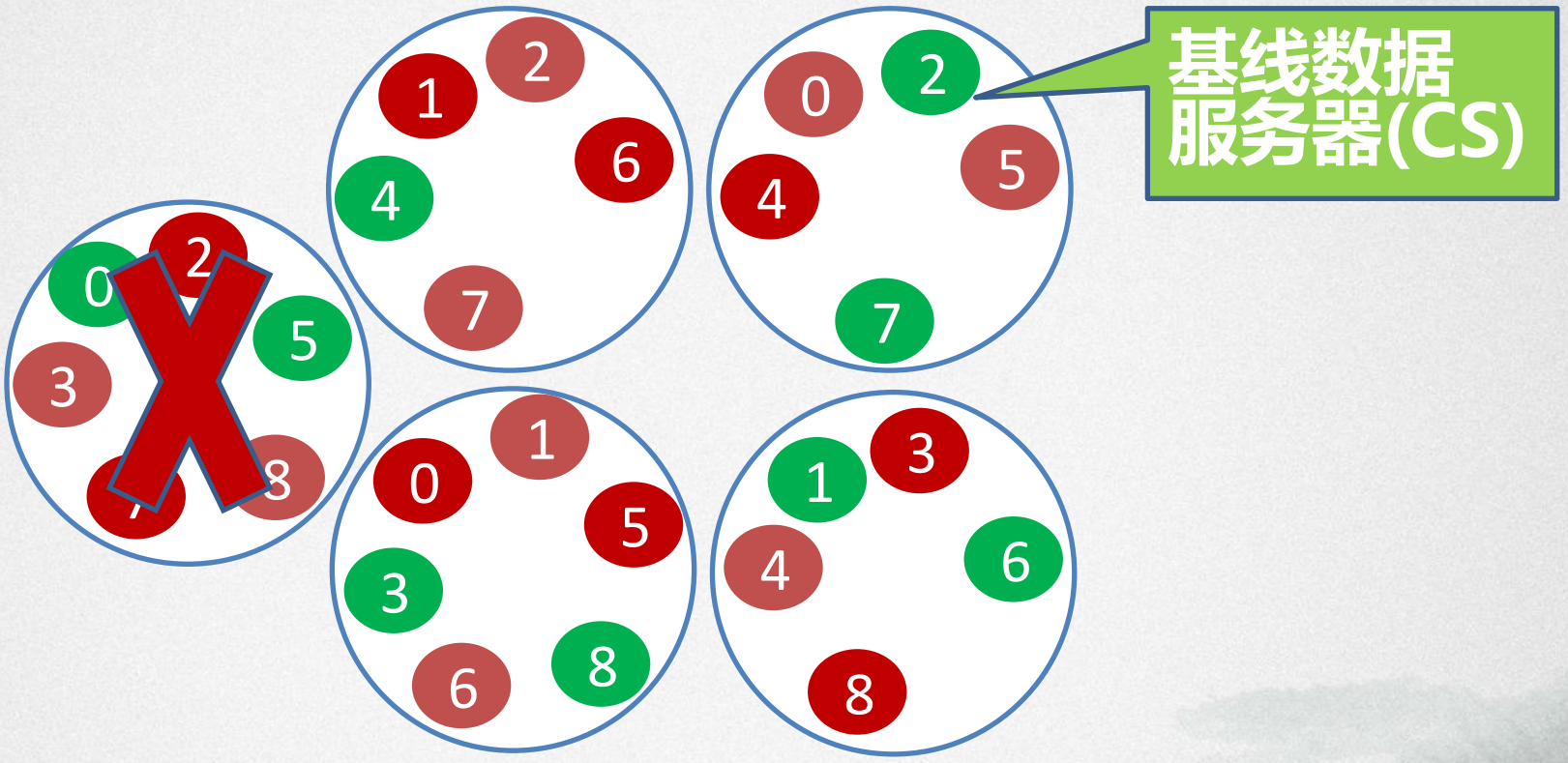
- 应用无需分库分表
  - 大表按照主键顺序自动划分为多个数据分区
  - 自动增加/减少服务器
  - 数据分区的分裂与合并



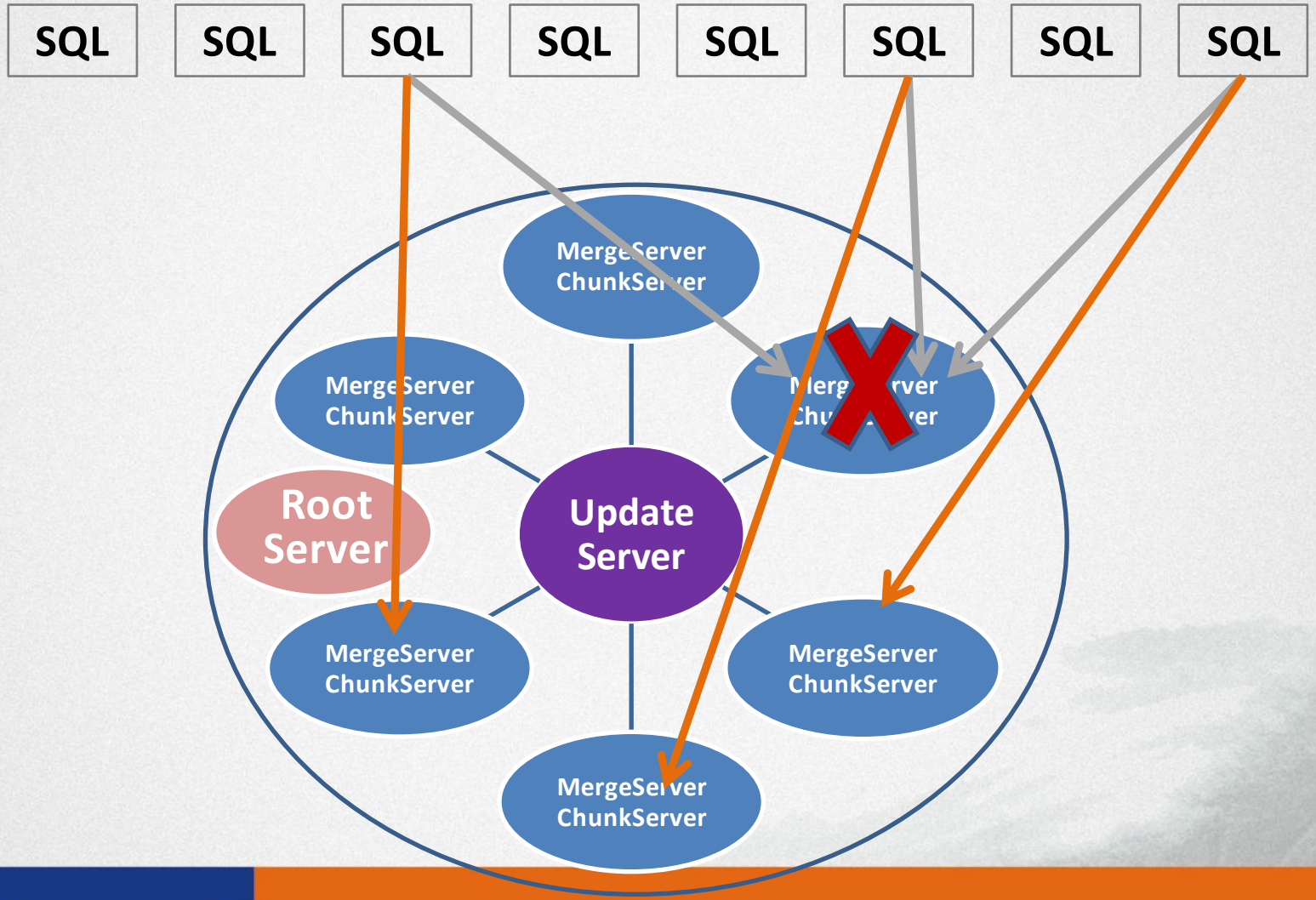
分布式存储引擎：RS/UPS容错



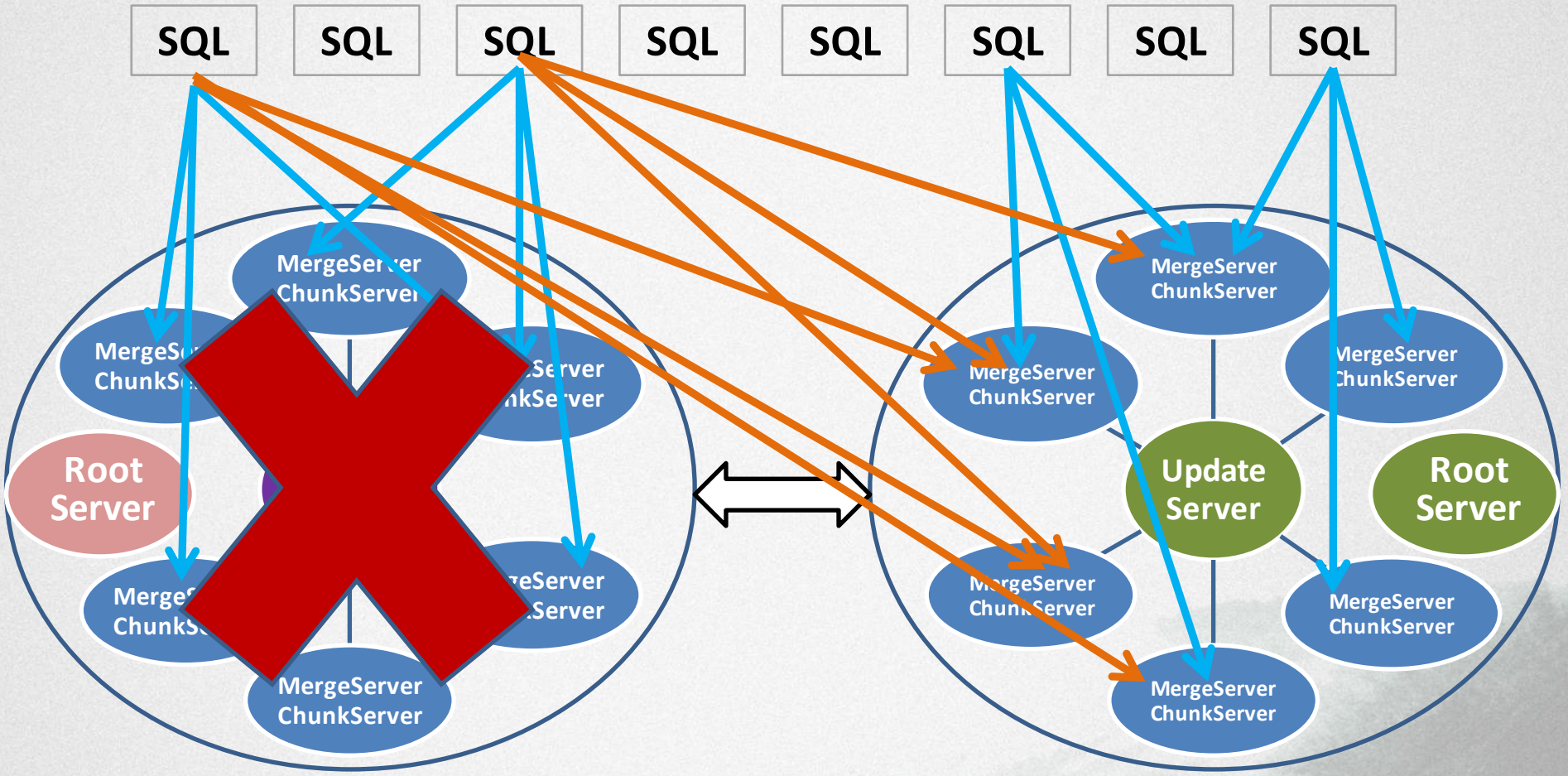
分布式存储引擎：CS容错



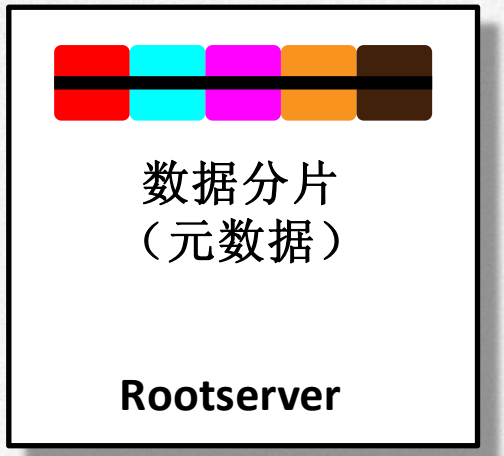
分布式存储引擎：MS容错

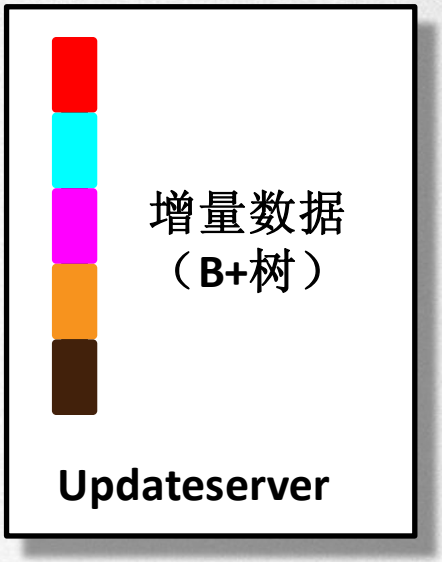


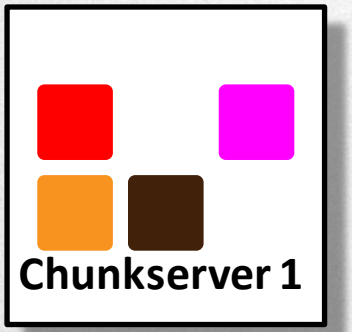
分布式存储引擎：IDC整体故障



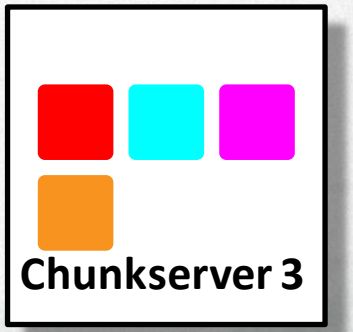
数据分布


数据分片
（元数据）
Rootserver


增量数据
（B+树）
Updateserver

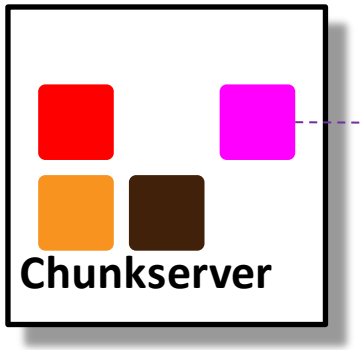

Chunkserver 1

Chunkserver 2


Chunkserver 3

Chunkserver 4

sstable存储引擎

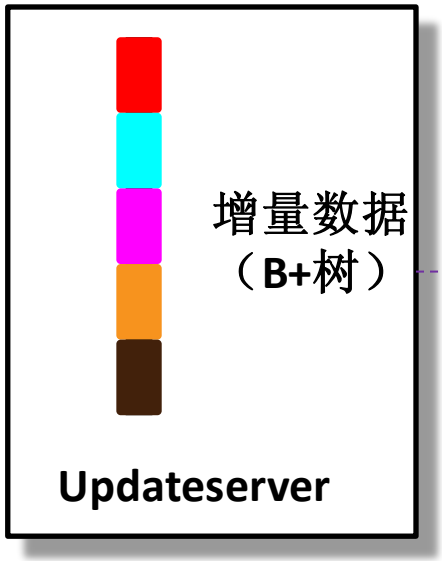


所有的基准数据都以SSTable稠密格式存储,每一片数据存储为一个SSTable

SSTable

dense format

single table



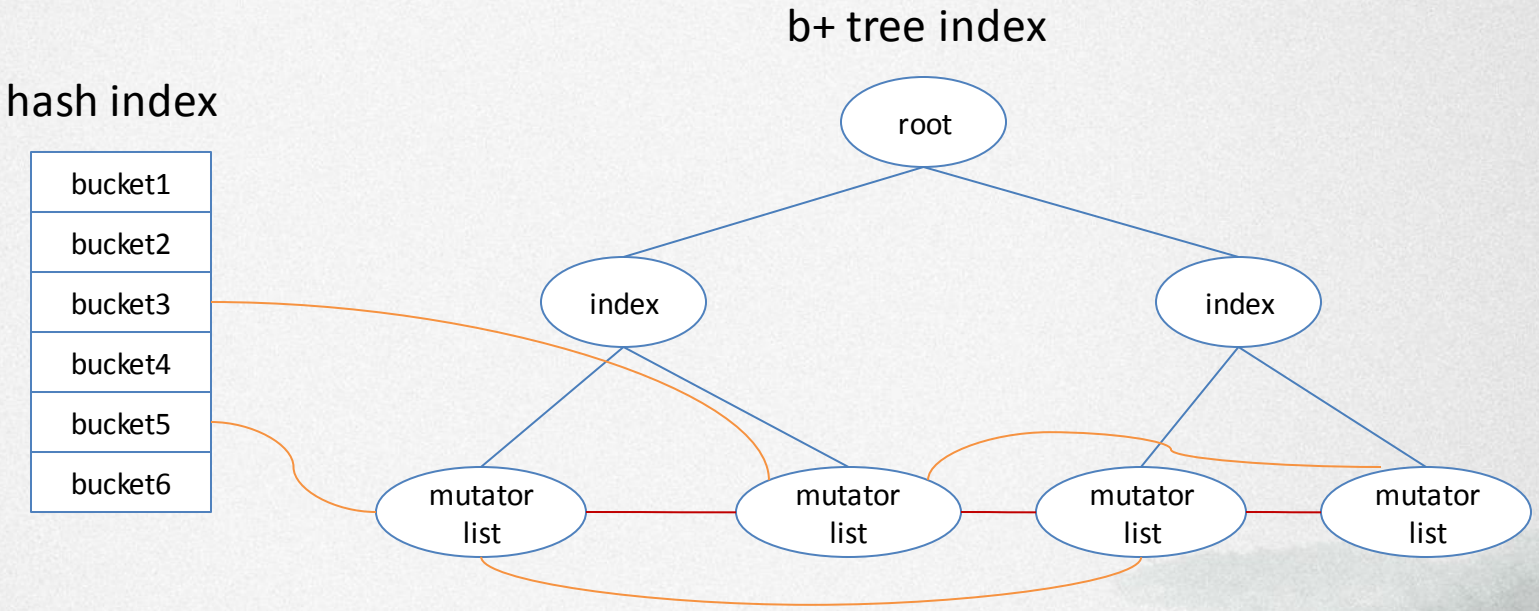
UPS的转储数据按SSTable稀疏格式存储,所有的Table的数据都存放在一个SSTable当中

SSTable

sparse format

multi table

UPS内存数据模型


b+ tree index
hash index
bucket1
bucket2
bucket3
bucket4
bucket5
bucket6
root
index
index
mutator list
mutator list
mutator list
mutator list

并发事务管理

- 使用MVCC保证写事务不阻塞只读事务
  - 每行为每次事务保存修改历史，根据事务ID读取到指定数据
  - 修改历史的合并，不在被读取的多个历史版本将被定期合并
- 写事务
  - 0.3版，单线程写+repalce语义，在MemTable层实现SnapShot
  - 0.4版，多线程读写事务+完备的DB语义，使用Snapshot+两阶段行互斥锁控制读写事务并发
- 并发日志回放
  - 以单个事务日志作为并行的最小单位
  - 并行回放的事务隔离性、一致性、原子性保证，并行提交的排序算法

支付宝 知托付 Alipay.com

数据安全

- 主备checksum校验：保证增量数据的完全一致
- tablet副本checksum 每次合并由rs做校验，保证基线数据的一致
- 主备集群按行进行checksum，保证主备集群一致

一致性选择

- 用户可以通过sql hint选择一致性
  - 写事务必须是一致的，选择主ups
  - 读事务可以选择不一致读，容忍一定数据延时获得更好的性能
- 对于一次导入，全天查询的olap应用
  - 只读基准数据提高性能

数据导入导出

- 数据导入
  - 直接写UPS(可以只更新列，提高速度)
  - UPS旁路导入(适用量相对大的导入，无事务)
  - CS旁路导入(适用超大数据量导入)
  - 集群复制
- 数据导出
  - 增量dump
  - 全量dump

优化技术举例

- 高性能网络框架：万兆网卡、减少上下文切换
- 无锁队列：push, pop每秒达到600万 ~ 1000万次
- 避免Linux gettimeofday()调用
- 定制化内存池：绝大部分事务执行过程中无需动态内存分配
- 完全避免随机写，适合SSD
- 多种IO机制（预读+异步IO）
- 优化数据结构，cache友好

UPS容量

- 修改增量的扩展
  - 转储到固态盘
  - 主备集群错峰进行“每日合并”
  - 分发到CS内存

OceanBase使用

- Java用户
  - 标准JDBC DataSource

```
OBGroupDataSource groupSource = new OBGroupDataSource();
groupSource.setUserName(“ob"); groupSource.setPasswd("test");
groupSource.setDbName("test");
groupSource.setConfigURL("http://10.232.102.182:8080/diamond-server/...") ;
```

  - Spring配置

```
<bean id="groupDataSource" class="com.alipay.oceanbase.OBGroupDataSource" init-method="init">
    <property name="userName" value=“ob" />
    <property name="passwd" value="test"/>
    <property name="dbName" value="test" />
    <property name="configURL" value="http://10.232.102.182:8080/diamond-server/.../>
</bean>
```

- C用户
  - 使用方式与libmysql相同

线上集群概况

- 30+应用，最大单个应用80台服务器
- 线上数据无丢失，无影响业务故障
- OLTP业务
  - 最大表格：收藏夹 10031274509
  - 最大qps/tps：45000 / 2500（一次scan 60行）
  - 单日更新数据量：5亿条（约120GB redo log）
- OLAP业务
  - 最大表格：P4P 4941204189
  - 单次请求最多行数：约500万，RT < 3秒
  - 最大导入数据量：SNS每天2TB（4台机器）

群集名称%	机房位置	机器数量	业务描述
收藏夹cm6	cm6	29	收藏夹
收藏夹cm4	cm4	29	收藏夹
P4P cm6	cm6	[illegible]	广告直通车报表
P4P cm4	cm4	12	广告直通车报表
TMALL评价 cm4	cm4	18	商城商品评价-cm4
TMALL评价cm6	cm6	[illegible]	商城商品评价-cm6
TMALL会员分级	cm6	6	天猫会员分级cm6
直播间 cm6	cm6	8	双11直播间cm6
直播间 cm4	cm4	8	双11直播间cm4
snsfeed	[illegible]	5	淘江湖
猜你喜欢	cm6	6	猜你喜欢
sns关系分析	[illegible]	[illegible]	sns关系分析
图片元数据管理	cm4	7	集团图片元数据管理
公用群集	cm4	[illegible]	[illegible]
…	…	…	…

支付宝 知托付 Alipay.com

读写性能

category	item	ops	average response time(us)
read	select 10 rows, block cache hit 100%	47,000	2,183
	select 10rows, block cache miss 100%	16,000	7,322
	select 1 row, row cache hit 100%	63,000	903
write	insert/update/delete(1 row)	120,000	150

开源

获取更多信息，请访问：

http://oceanbase.alibaba-inc.com/

支付宝 Alipay.com 知托付

2013年04月