TBase数据治理与应用

嘉宾：胡森
公司：腾讯

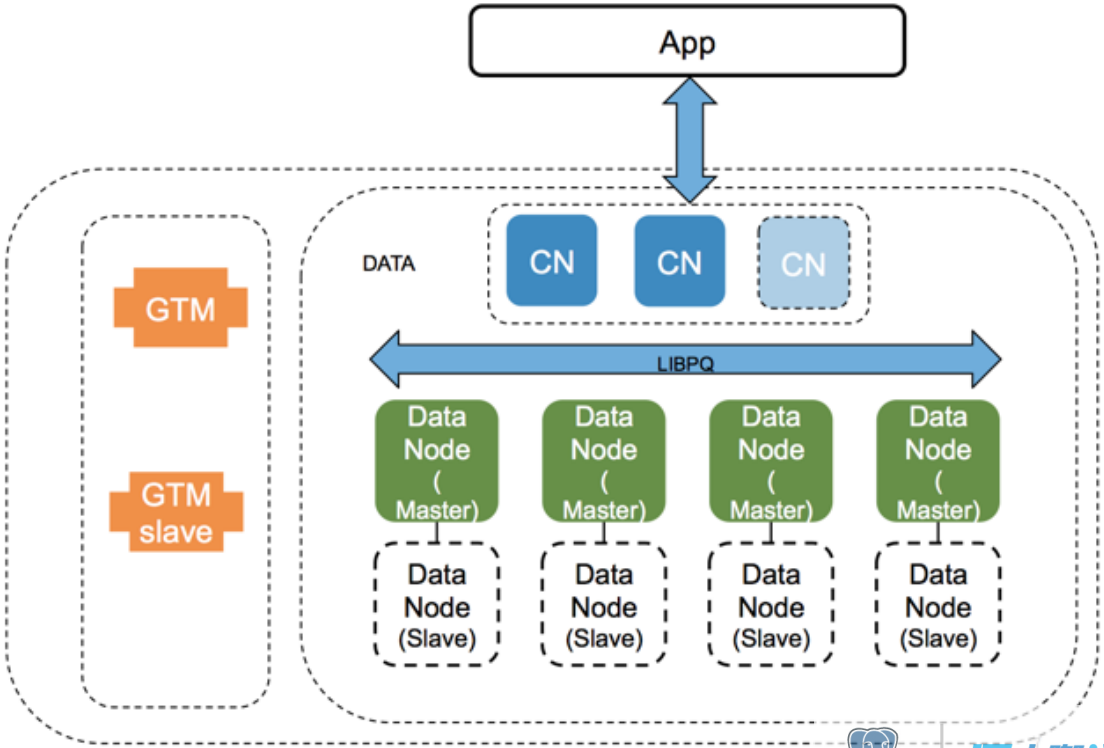
腾讯·互联网+

- TBase背景和架构
- TBase数据倾斜治理
- TBase冷热数据分离
- TBase在线扩容

TBase架构

架构特点：

- share-nothing
- 关系型数据库
- OLTP业务
- 分布式事务
- 跨节点join



TBase遇到的问题

- 数据倾斜

  数据集中在某几个节点，导致资源使用集中，影响整个集群。

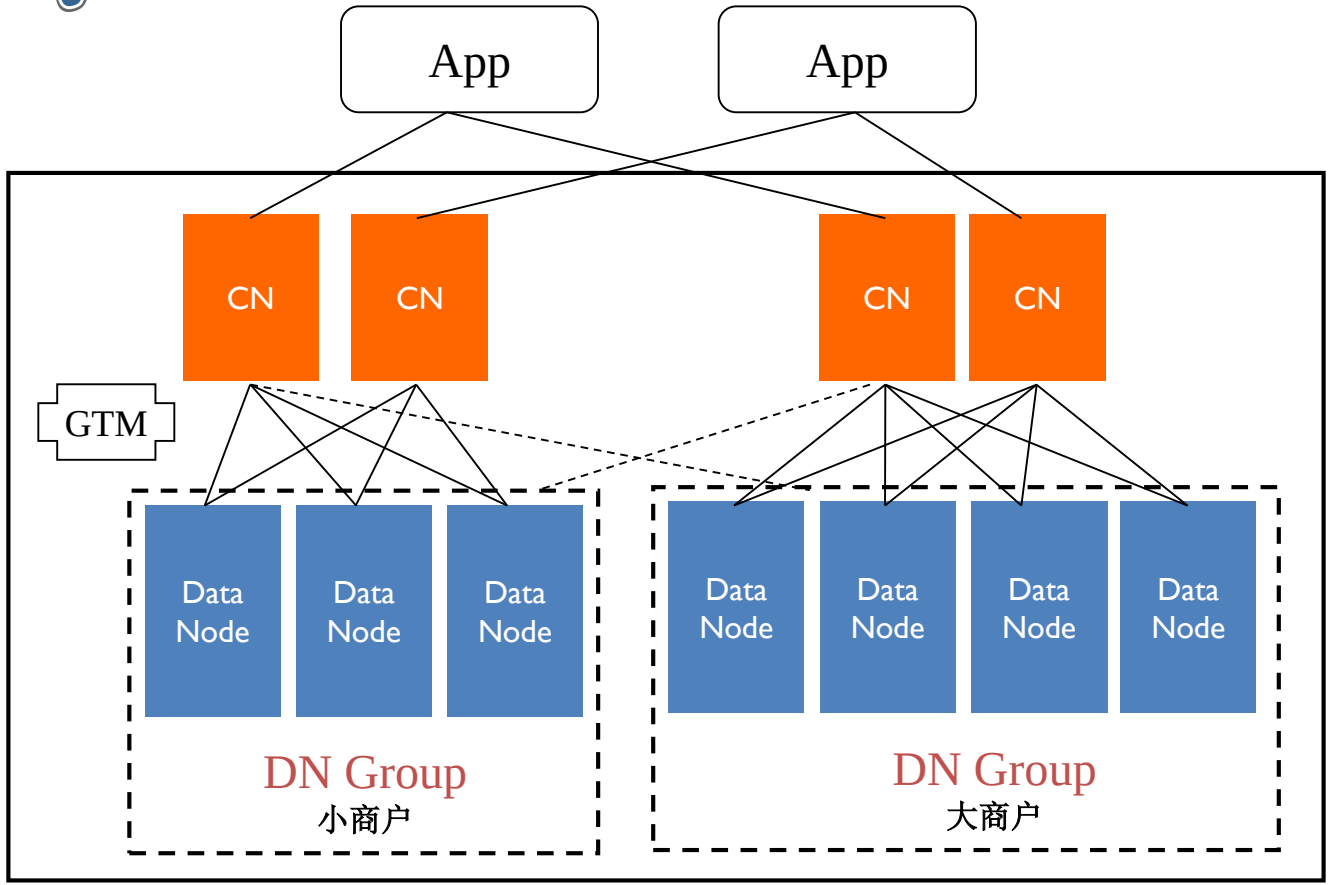
- 业务成本

  业务主要访问近一段时间热数据，但是冷热数据都放在一起，成本过高。

- 在线扩容

  业务增长过快，存储到达瓶颈。

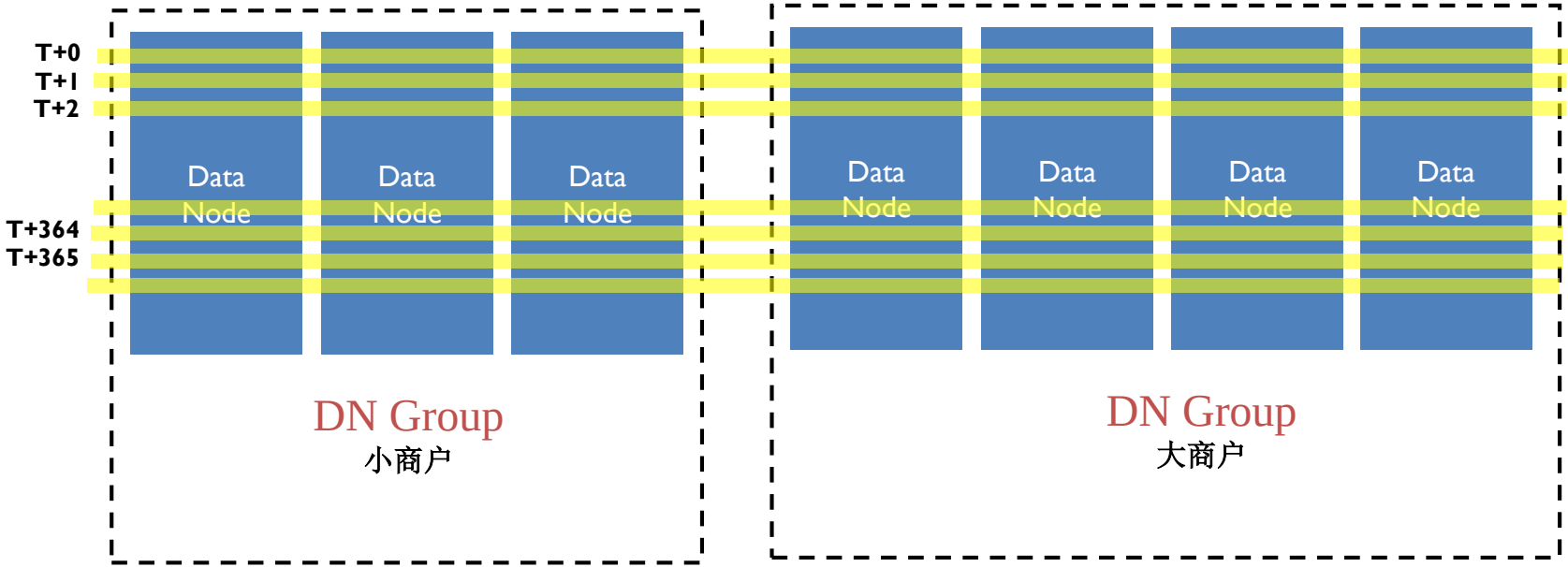
TBase数据倾斜治理—节点分组



业务透明：数据库内部对数据进行管理，用户无感知。

分组管理：数据库内部对DN节点进行分组，不同规模用户可以放到不同的DN GROUP，业务可指定用户规模。

TBase数据倾斜治理—分表

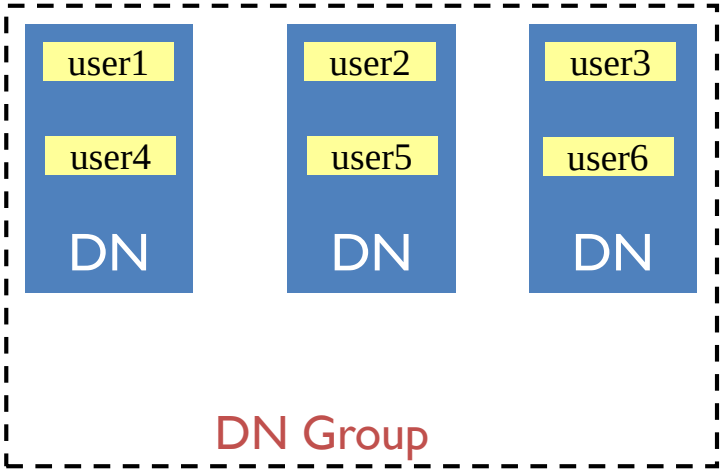


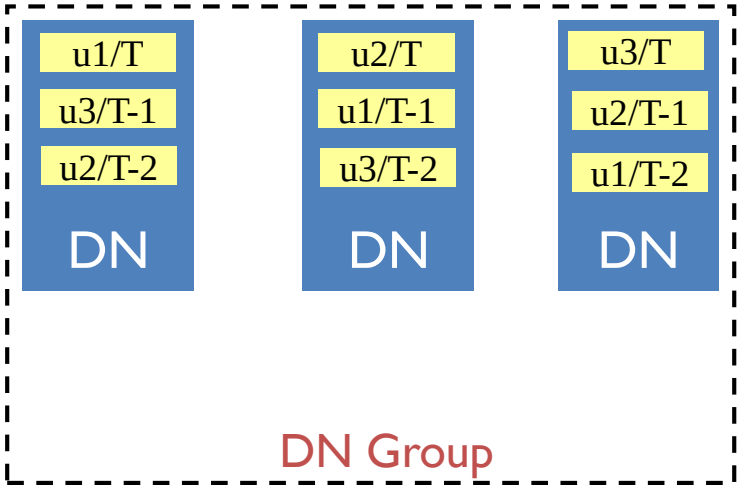
按天分表：1张逻辑表1年有366个子表(或者按其他纬度分表)，每个子表在每个DN上都有一个物理表

便于管理：删除，搬迁

性能提升：查询

TBase数据倾斜治理—分库





小商户

商户特征：每个DN可容纳万级~10万级商户

分布方法：按用户ID直接Hash

方案：

1. 数据以用户ID聚集:同一个用户的数据在同一台DN上。

2. 性能：每个小散户数据量较小，负载模型偏OLTP，以用户Hash保证一次操作只与一台DN相关，降低网络消耗

3. 扩容：以Hash桶为最小粒度进行数据迁移。当前已经实现节点一分为二，未来实现节点的线性扩容

大商户

商户特征：

1. 单个商户数据量比小散户大很多
2. 每个DN能够容纳10个~1000个商户

分布方法：

(负载均衡与数据按用户ID聚集的平衡)

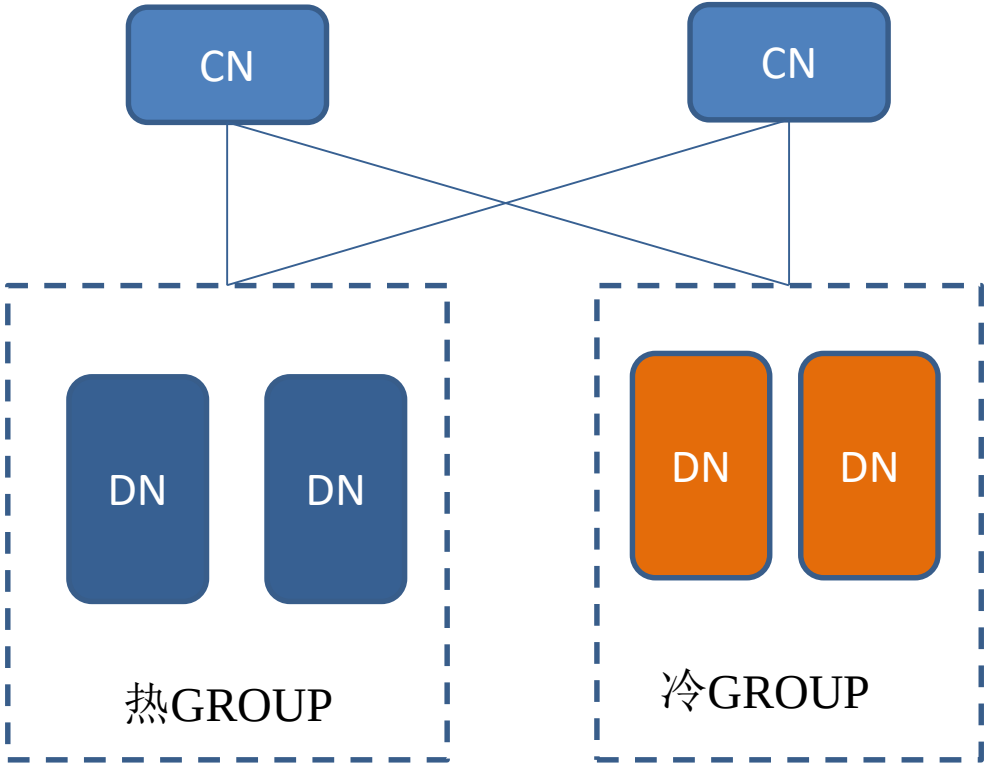
1. 一个用户的数据可以散列到多个DN
2. 一个用户某一天的数据必须在一个DN

方案：

1. 在一天以内保证数据以用户ID聚集：某用户查询一天的数据只需一个DN参与

2.多用户的数据按天交替存放保证热点数据均匀分布

TBase冷热数据分离



热数据（最近几个月的数据）

冷数据（历史数据）

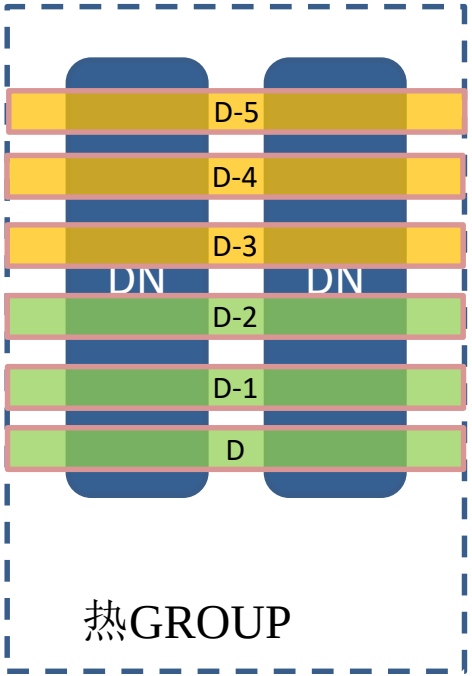
1、冷热数据使用不同的节点group存储，这些节点组内部使用的机型不同，从而达到冷热分离节省成本的目的。

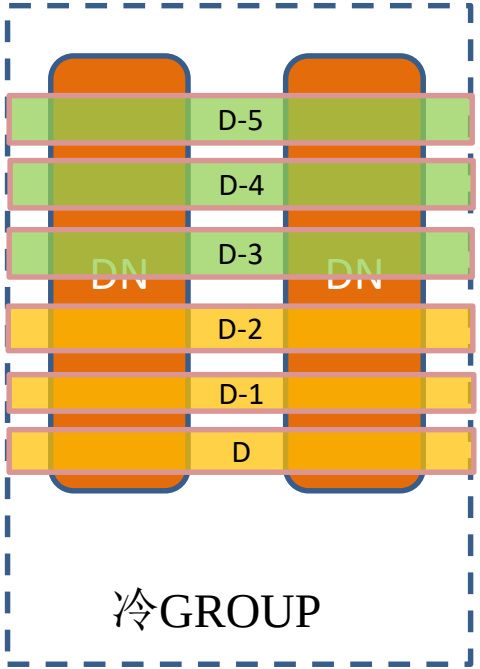
2、集群通过CN对外提供统一的数据视图，业务对冷热分离的状态无感知。

TBase冷热数据分离

冷数据阈值为3天时系统的数据分布情况

绿色为有数据，黄色为空表

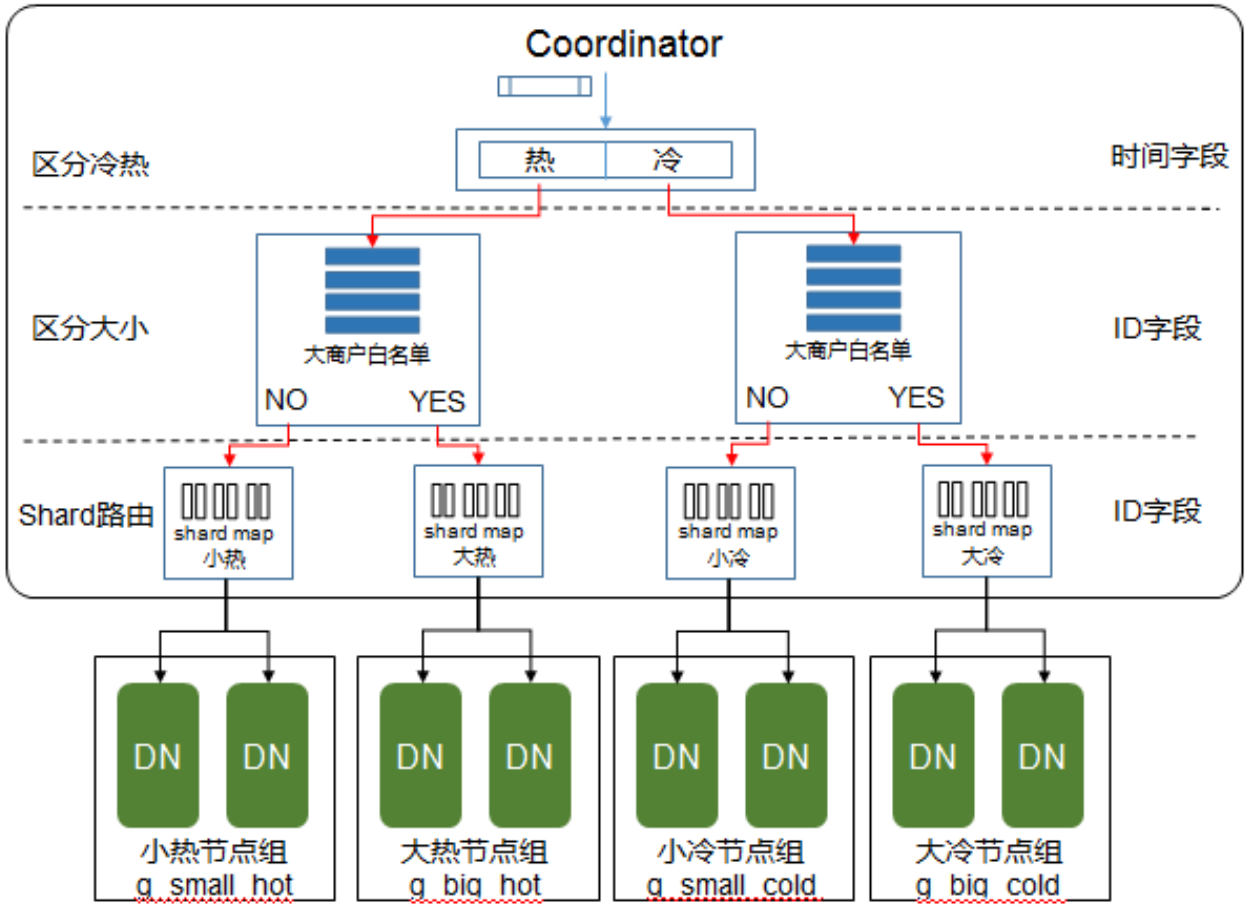


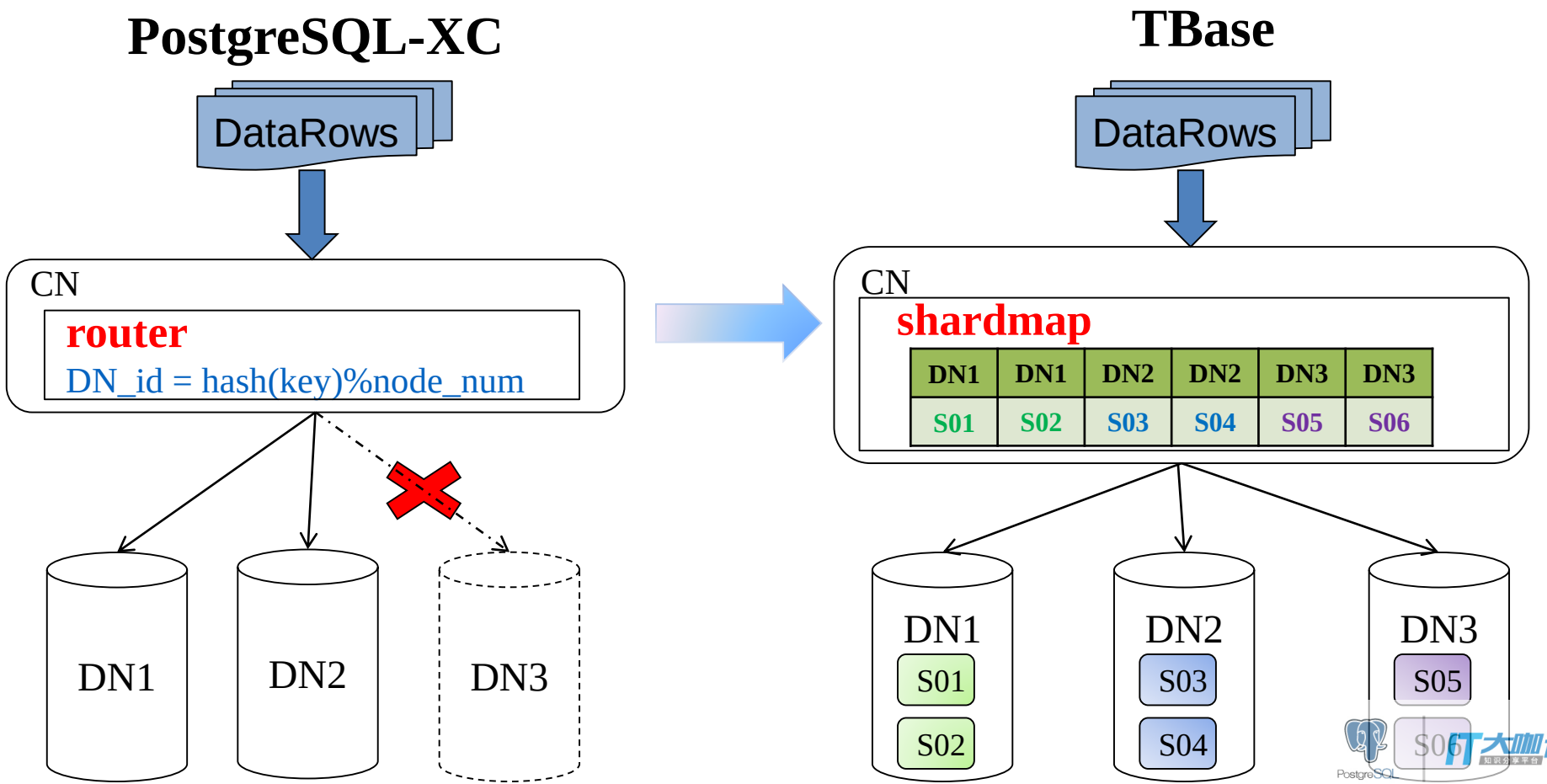


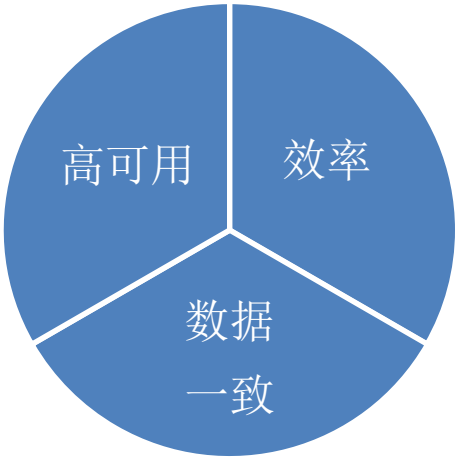
1、表按天分区，每张表在冷数据组和热数据组节点中都有一张物理表。
2、数据写入时都写入到热数据部分，对应的冷数据表为空表。
3、冷热数据转换的阈值是集群公用的值，是距离当前日期多少天的数值。
4、当冷数据阈值达到时，后台搬迁变冷的当天数据。
5、热数据搬迁到冷数据集群后，热数据部分的表truncate为空表。

热数据（最近几个月的数据）　　冷数据（历史数据）

TBase路由逻辑




PostgreSQL-XC
DataRows
CN
router
DN_id = hash(key)%node_num
DN1
DN2
DN3
TBase
DataRows
CN
shardmap
DN1 DN1 DN2 DN2 DN3 DN3
S01 S02 S03 S04 S05 S06
DN1
S01
S02
DN2
S03
S04
DN3
S05
S06

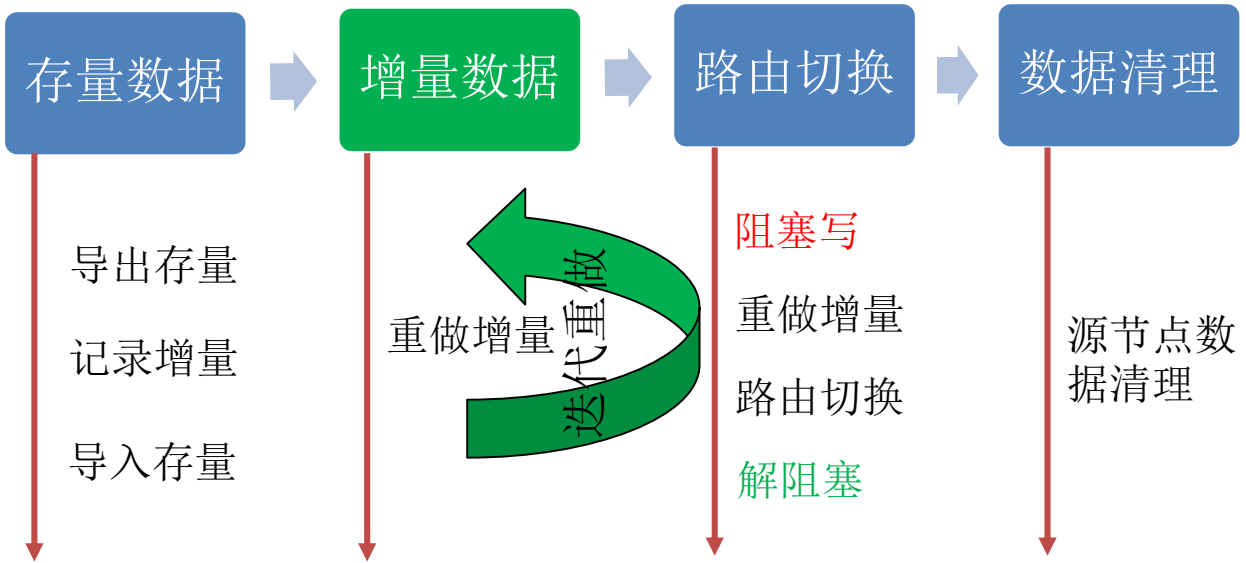


- **高可用**

  数据搬迁过程中，不影响业务的正常读写
- **效率**

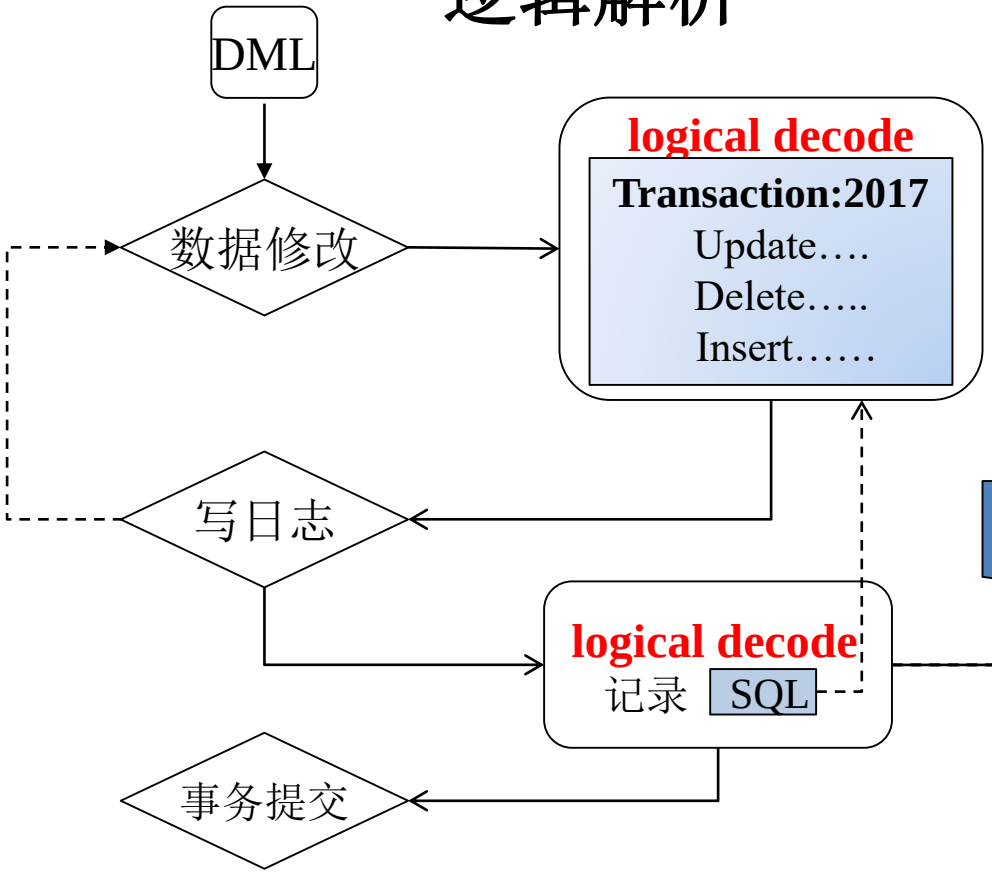
  高效地完成数据搬迁
- **数据一致**

  一条不多，一条不少，一条不错

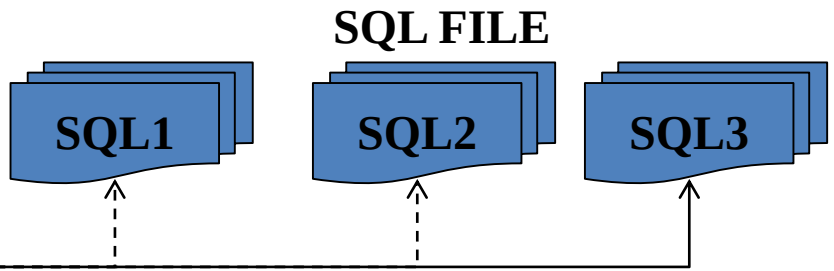


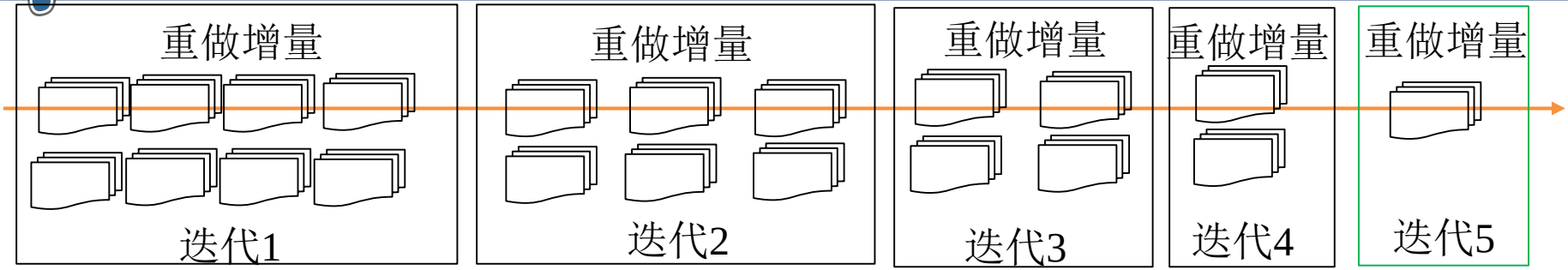
- **增量数据**
  对数据一致性起着至关重要的作用
- **路由切换**
  保证业务数据读取一致

逻辑解析



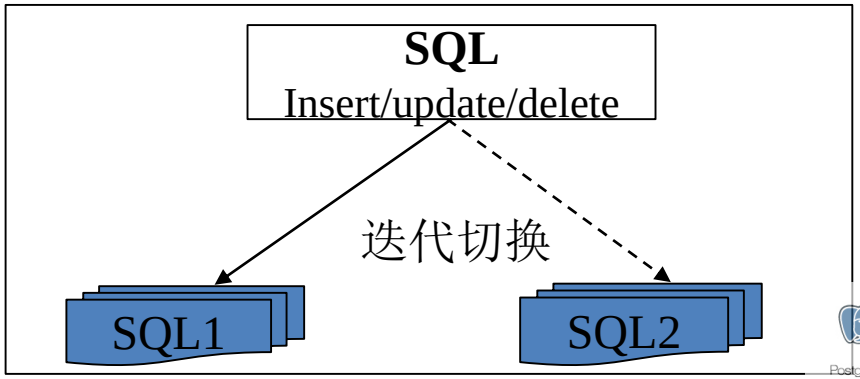
- SQL形式呈现增量数据
- 节点物理文件信息解耦
- 跨平台





增量数据小于阈值，路由切换

迭代之间如何完美衔接，保证数据既不重复，也不遗漏？



串行重做增量

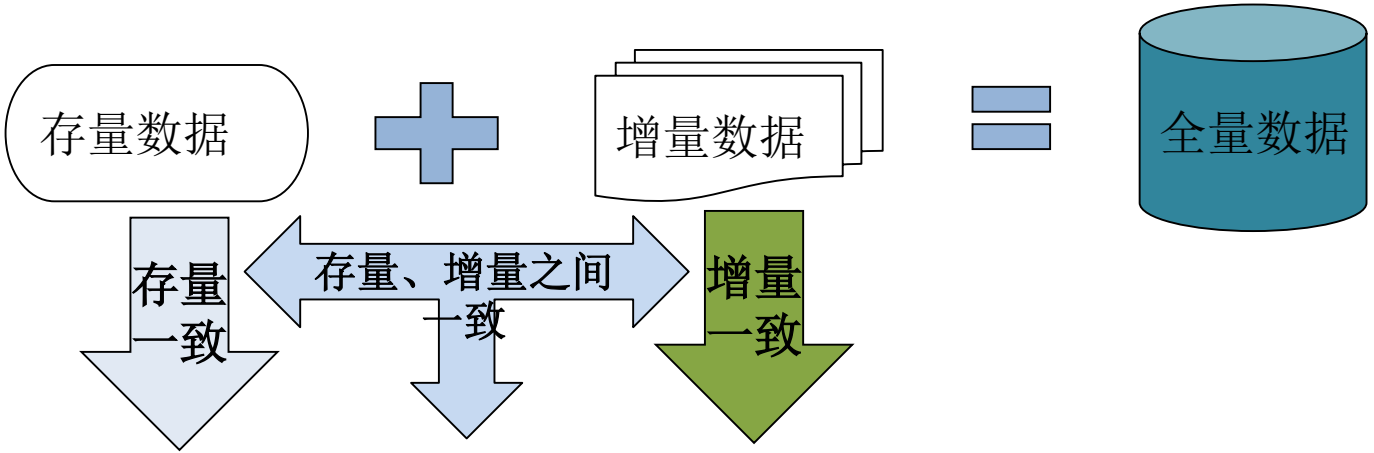
- 重做速度追赶不上产生速度
- 增量数据堆积

并行重做

并行粒度：

- 行级并行
  表有主键
- 表级并行
  表没有主键

微信现网8个并发： 17min/G -> 2.8min/G，
速度提升6倍



校验：

- 条数
- 内容

相同数据库快照

逻辑解析：

- 事务提交时记录修改，保证事务修改原子性
- 迭代切换保证增量重做之间数据无缝衔接