SACC

2020 中国系统架构师大会

SYSTEM ARCHITECT CONFERENCE CHINA 2020

架构融合 云化共建

live 2020年10月22日-24日网络直播

云上会议

MarxDB金融级分布式数据库

MarxDB团队
杨轩嘉
newsql@jd.com

产品介绍

架构融合
云化共建

MarxDB金融级分布式数据库是(京东零售-基础架构部-数据库技术部)在开源分布式数据库CockroachDB的基础上兼容了MySQL协议推出的新一代弹性数据库，旨在提供透明可扩展，跨地域部署，多副本数据强一致，支持分布式事务的存储服务。

MarxDB金融级分布式数据库相比于传统的数据库中间件产品有着巨大的优势，业务不再需要分库分表，支持分布式事务，并提供更加丰富的SQL语法支持.

核心优势

架构融合
云化共建

1.兼容MySQL协议

MarxDB兼容MySQL协议,从 MySQL无缝切换到MarxDB，无需修改任何代码，迁移成本极低.
MarxDB提供了对应的迁移工具.

2.海量数据规模

MarxDB可以支持20TB单副本的有效数据，3副本后整个集群数据量在60TB的规模.
(更大数据规模后期会持续测试验证)

3.分布式事务

把 MarxDB 看作是一个单机的MySQL 。采用无锁分布式事务模型,隔离级别支持最高的SSI(serializable snapshot isolation 串行化快照隔离级别).

4.故障自动恢复

无需人工干预,节点故障自动触发故障恢复,恢复过程对用户透明,不影响服务,真正意义上的Auto-Failover. MarxDB采用去中心化设计,集群任何业务模块没有单点, 保证高可用.

核心优势

架构融合
云化共建

5.在线DDL
Online Schema Change在线表结构变更。添加新的列和索引等DDL操作，不影响在线业务.

6.数据强一致
数据分片采用多副本备份(默认3副本),使用**raft**共识算法保证副本之间数据强一致,实现了真正的高可用,高可靠,完全满足金融级OLTP业务.

7.从副本读取
支持从副本的一致性读取.获得更好的读写性能.

8.透明弹性伸缩
MarxDB 可随着业务的数据增长而无缝地水平扩展，只需要通过增加更多的机器来满足业务增长需要，应用层可以不用关心存储的容量和吞吐。
MarxDB 根据存储、网络、地域等因素，动态进行负载均衡调整，以保证更优的读写性能。

核心优势

架构融合
云化共建

9.异地多活,跨地域部署

MarxDB支持跨机房,跨地域部署,支持异地多活.

10.完备生态建设

MarxDB支持增量数据实时同步到大数据平台, 并和大数据平台打通抽数和推数服务.

适用场景

架构融合
云化共建

1.金融级OLTP业务特别适合
交易, 支付, 小账单
结算, 金融, 小理财
储蓄, 保险, 小理赔
财务, 基金, 小股票
白条, 金条, 小金库
定投, 期货, 小红包
银行, 贷款, 小彩票

2.海量数据存储和在线处理

3.跨数据中心数据业务

4.轻量级OLAP业务(重量级OLAP业务需配合SparkSQL使用)

系统架构

架构融合
云化共建

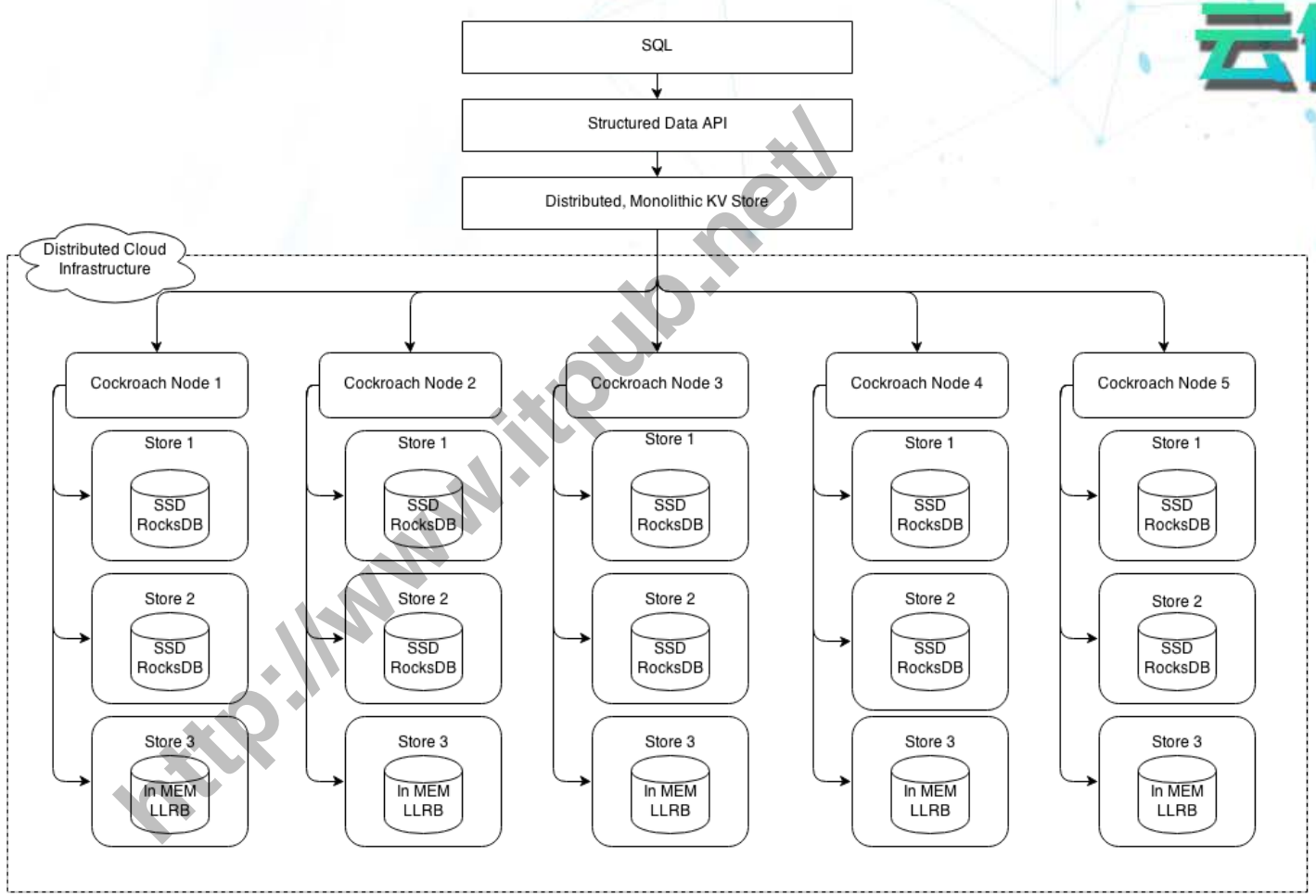
SQL

Txn

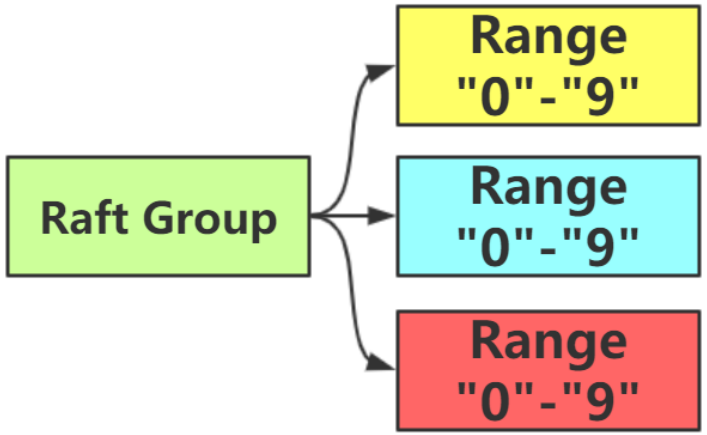
Raft

MVCC

RocksDB



系统架构



Range是数据元的最小单位, 默认512M.
Raft Group管理一个Range的多个副本.

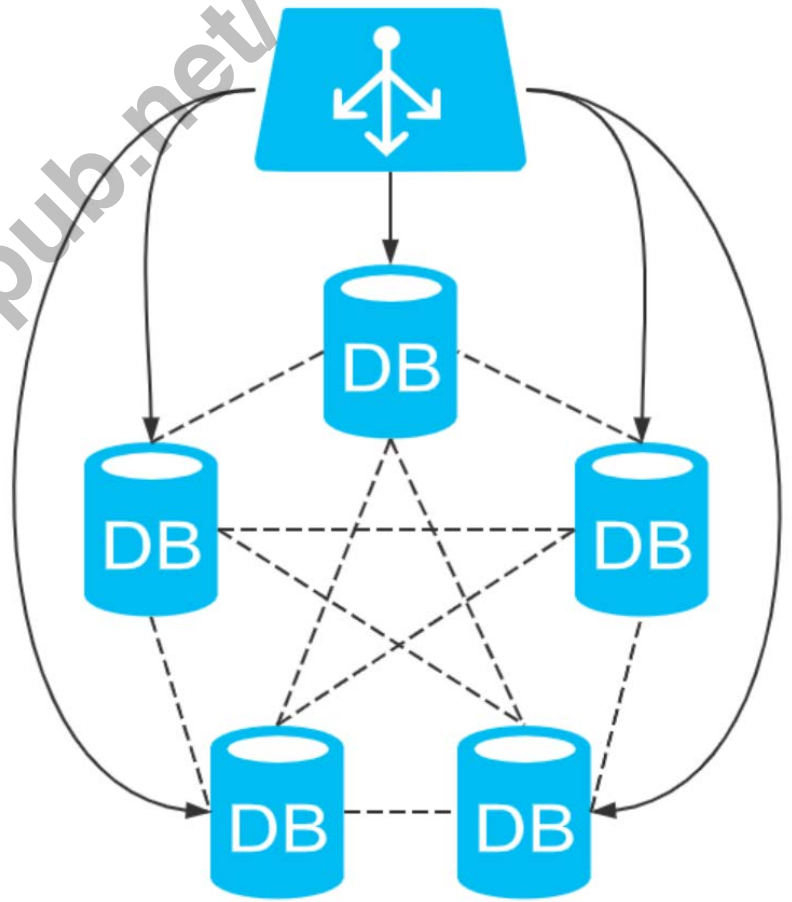
系统架构

架构融合
云化共建

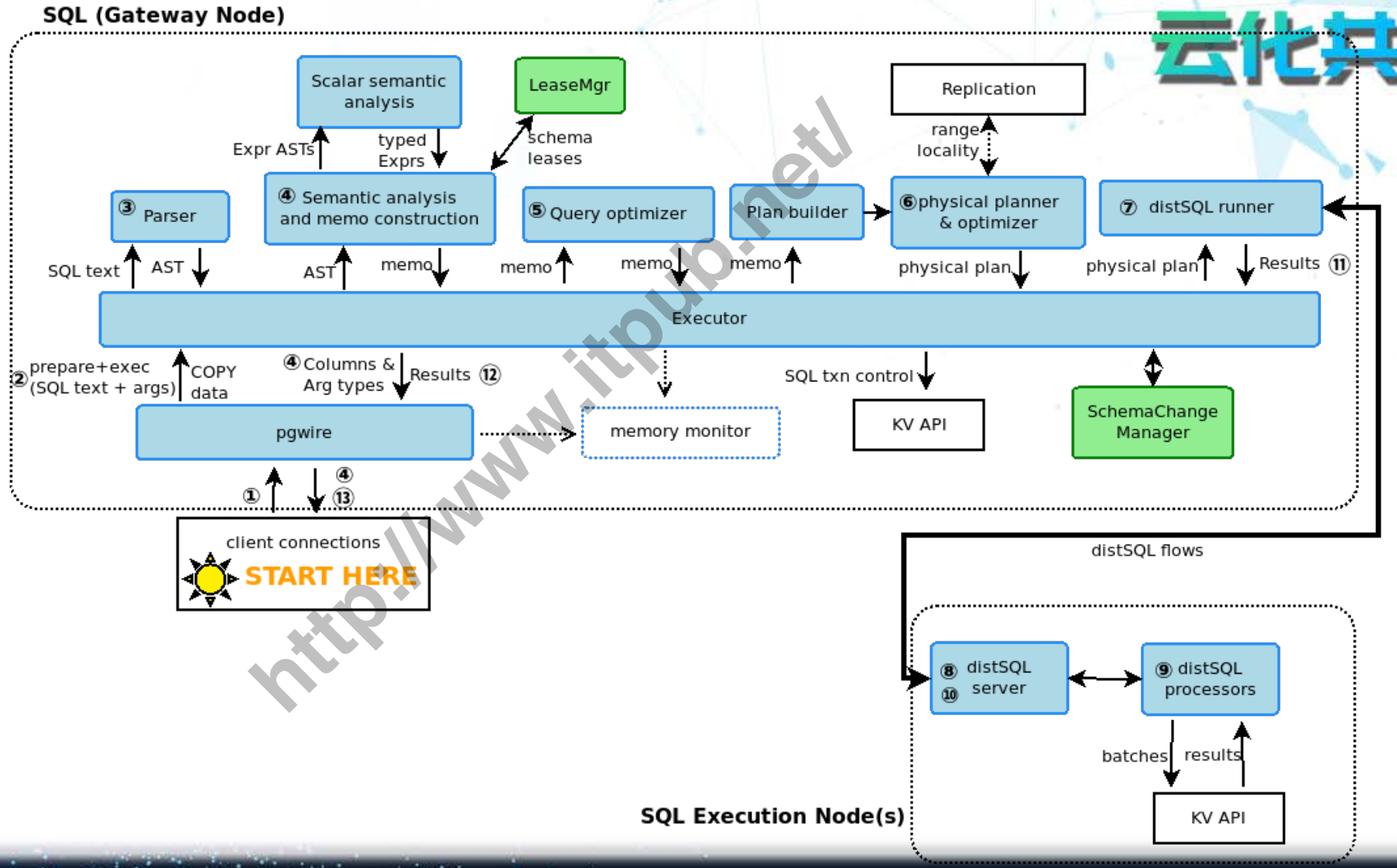
MarxDB节点间通过Gossip协议进行通信, 最大支持1万节点的集群规模.

优势是去中心化, 无单点故障, 无单点授时瓶颈.

通过LB负载均衡可以把写流量打到所有节点上.



SQL层架构



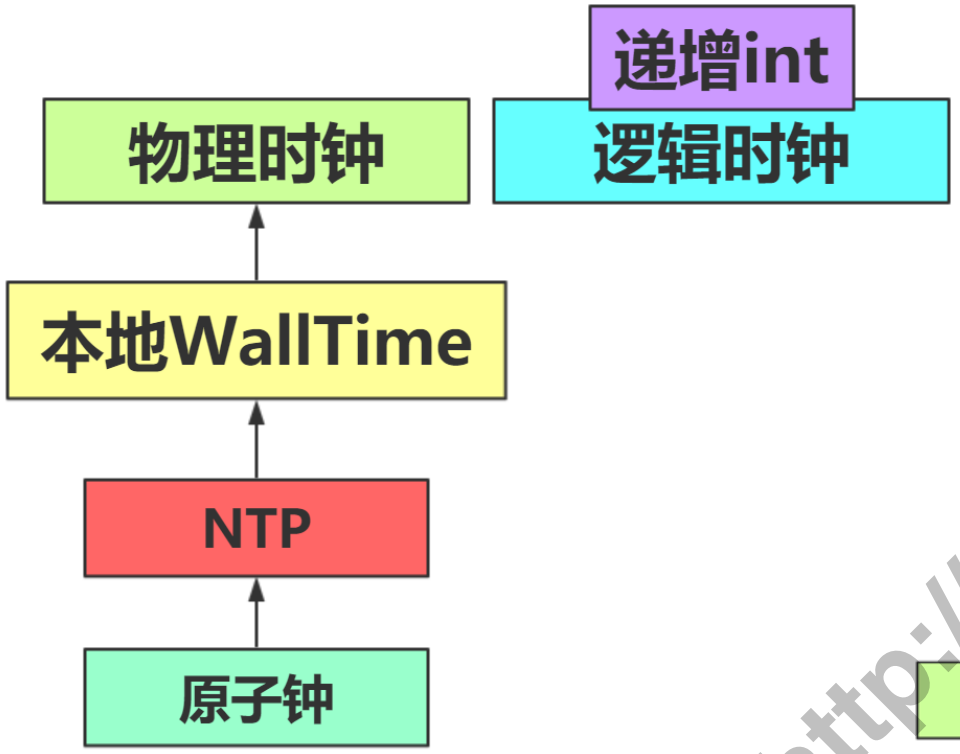
SQL转KV存储

每个table都有一个主键, 由table ID, 主键ID, 主键值拼接构成了一行数据在KV层存储时的key, 其他非主键的列值编码成value.
对于二级索引, 又分唯一索引和非唯一索引:
唯一索引：key由table ID, 唯一索引ID, 唯一索引值拼接而成.
非唯一索引：key由table ID, 非唯一索引ID, 非唯一索引值, 主键值拼接而成.

		记录Row
Table ID: 1	表名: Student	
Primary Key ID: 9	主键列: Erp	Jay
Uni Key ID: 11	唯一索引列: Mail	jay@jd.com
Key ID: 22	非唯一索引列: Name	周杰伦

Key类型	Key	Value
Primary Key	/table/1/9/Jay	jay@jd.com,周杰伦
Uni Key	/table/1/11/jay@jd.com	Jay
Non-Uni Key	/table/1/22/周杰伦/Jay	NULL

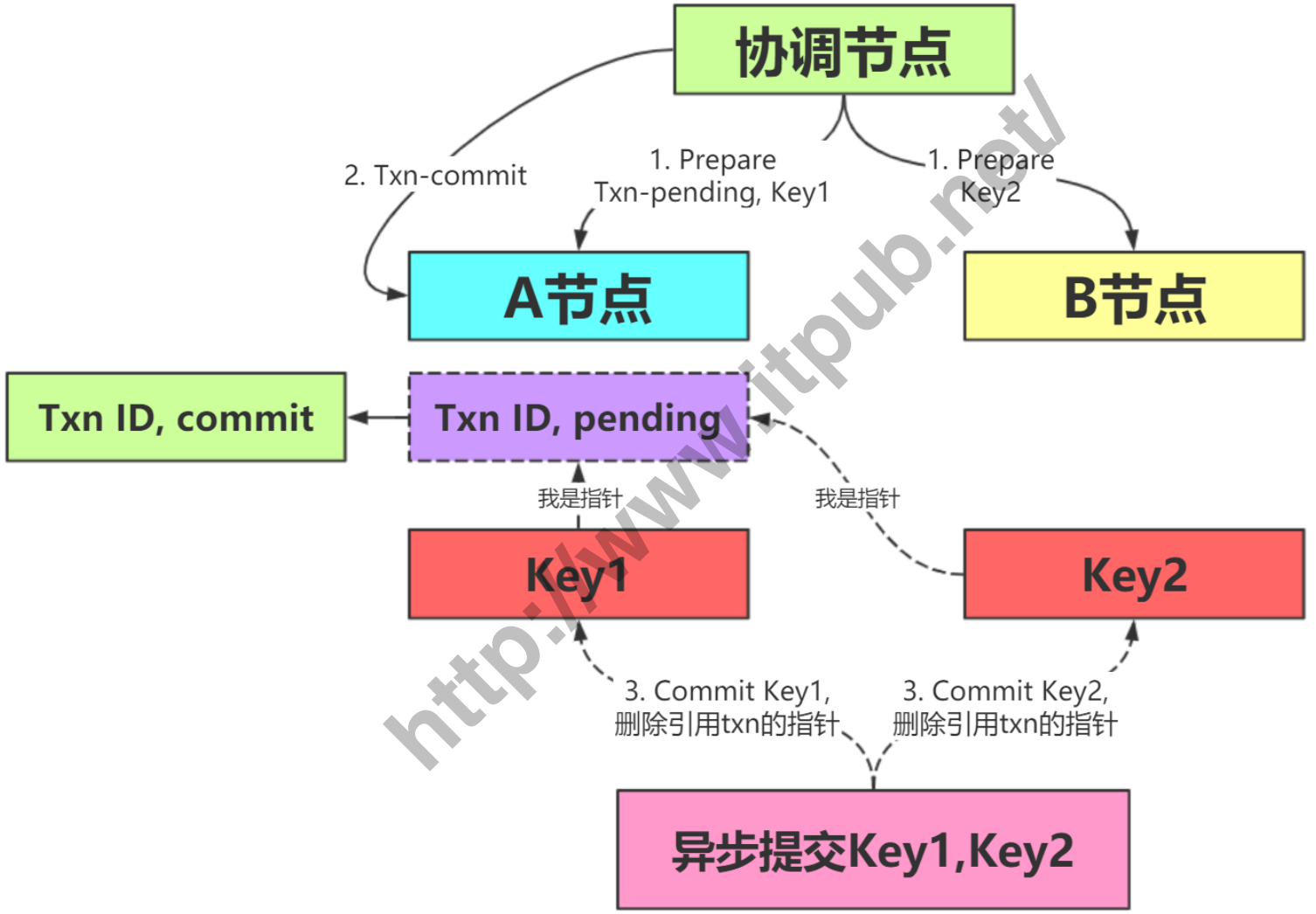
分布式事务 – HLC混合逻辑时钟




```
更新本地HLC
B.Phy=max(A.Phy, B.Phy, B.Wall)
if B.Phy 变了 {
    B.Logic = 0
} else {
    B.Logic = max(A.Logic, B.Logic) + 1
}
```

A节点 —Event→ B节点

分布式事务 – 2PC



分布式事务 – Lock Free无锁乐观模型

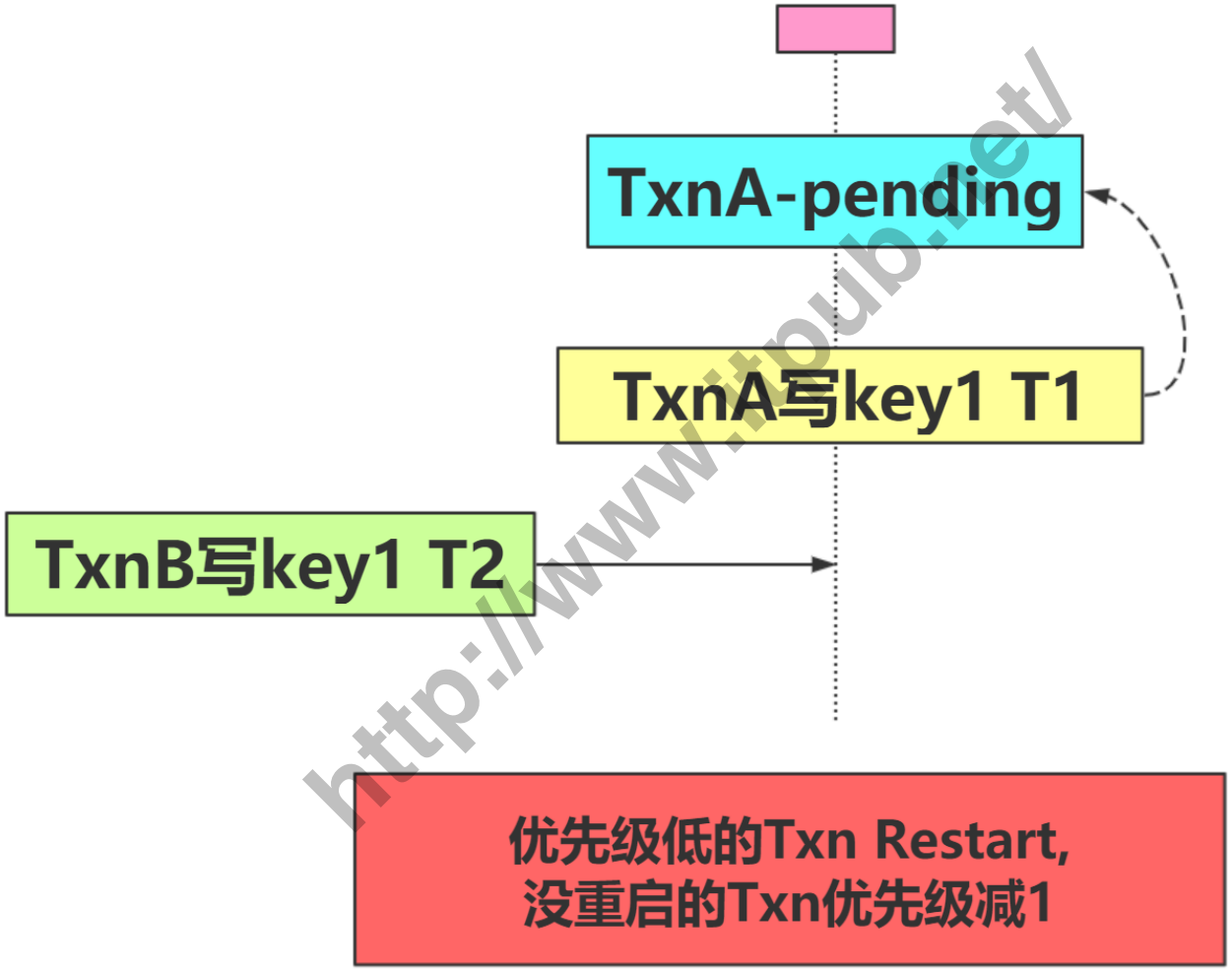
架构融合
云化共建

MySQL为了满足事务的ACID, 主键上有3种锁, 行锁, Gap锁, Next-Key锁.

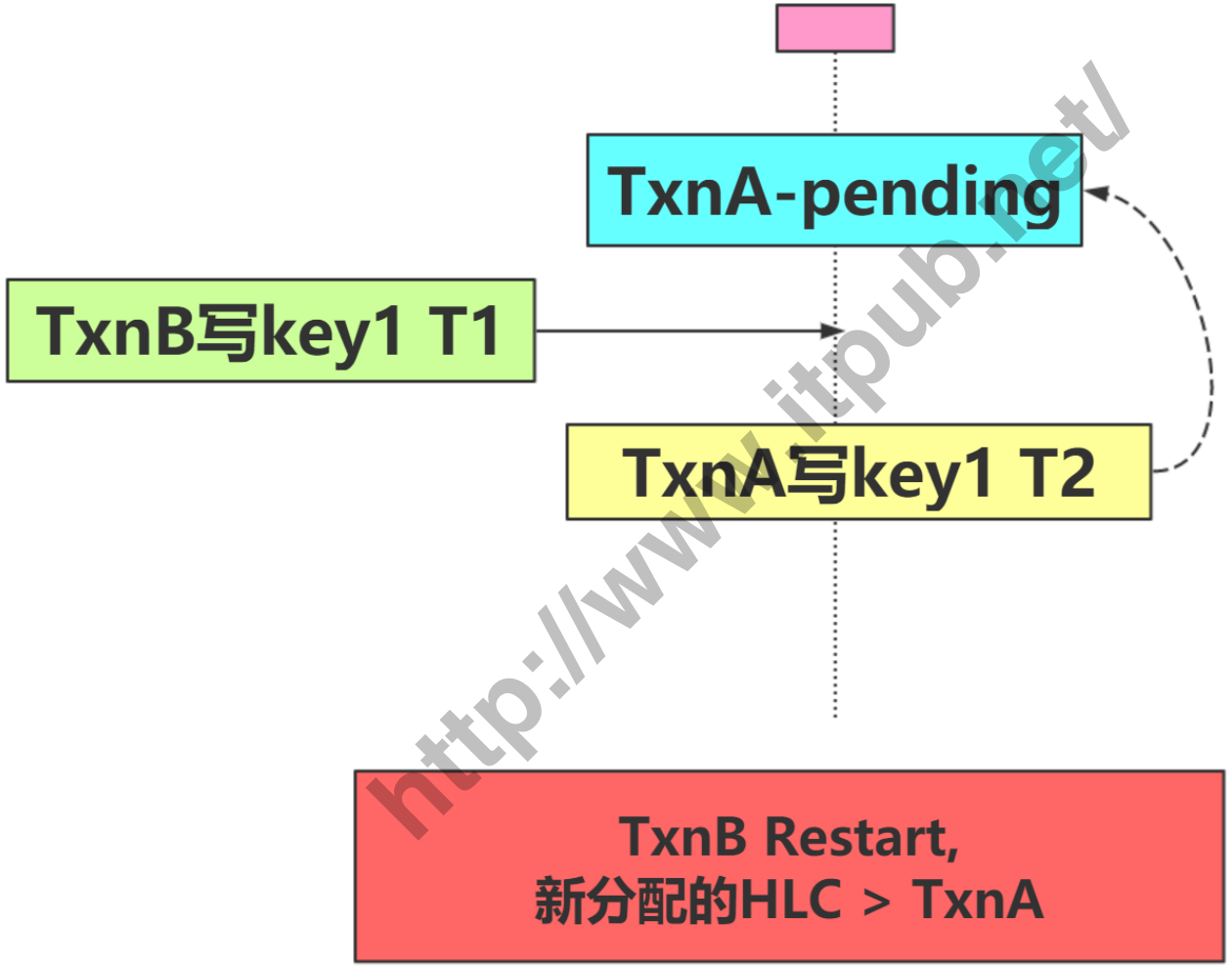
MarxDB都没有, 设计了一个Read Timestamp Cache, Key只要被读过就会记到这个Cache里.
Lock Free的代价是要解决好各种Key键冲突, 如果冲突太大, 性能就会下降, 这就是乐观的代价.
当然, 如果冲突较少, 性能就会非常好, 这就是架构的权衡.

冲突分3种: 写写冲突, 写读冲突, 读写冲突

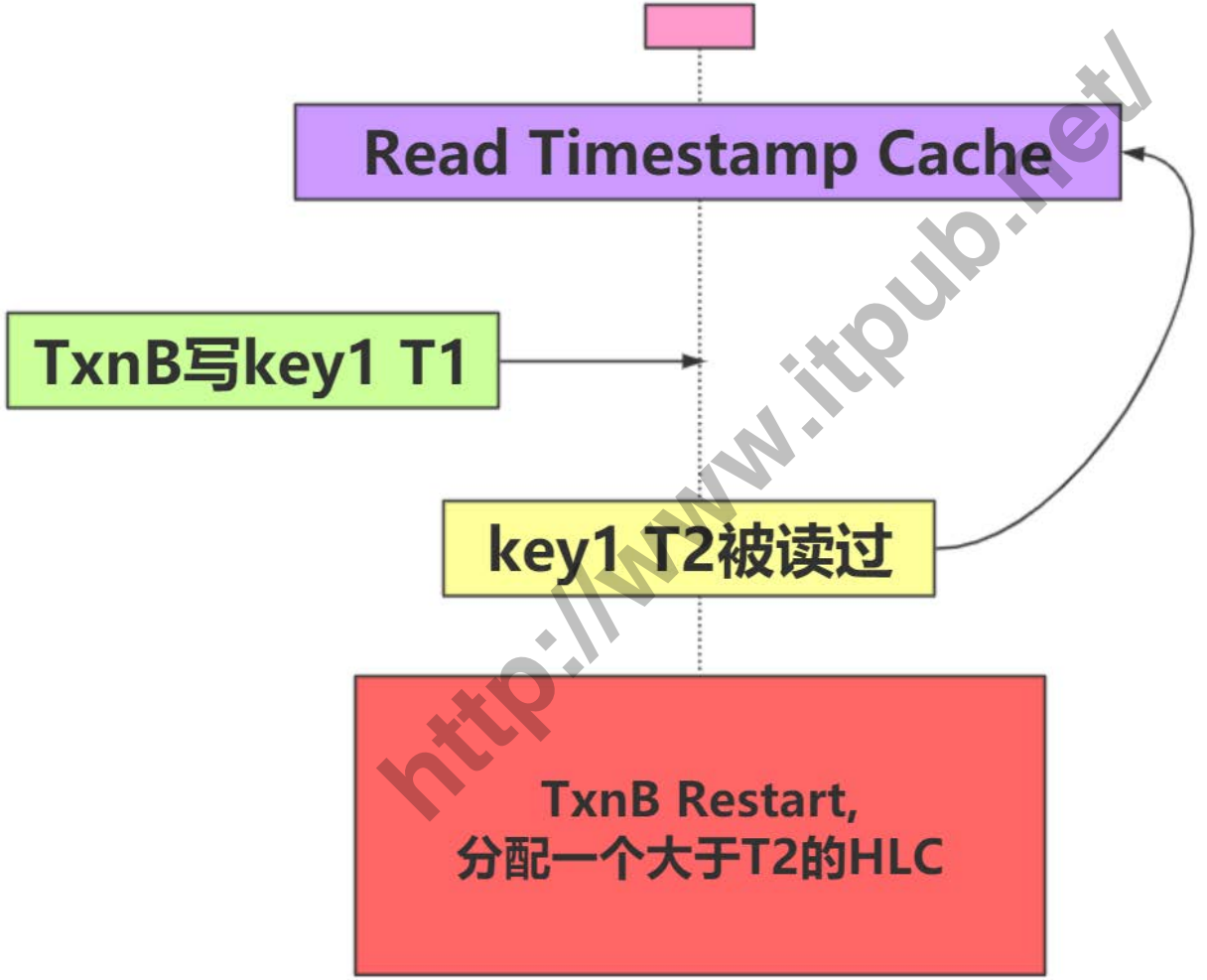
分布式事务 – 写写冲突 Case1



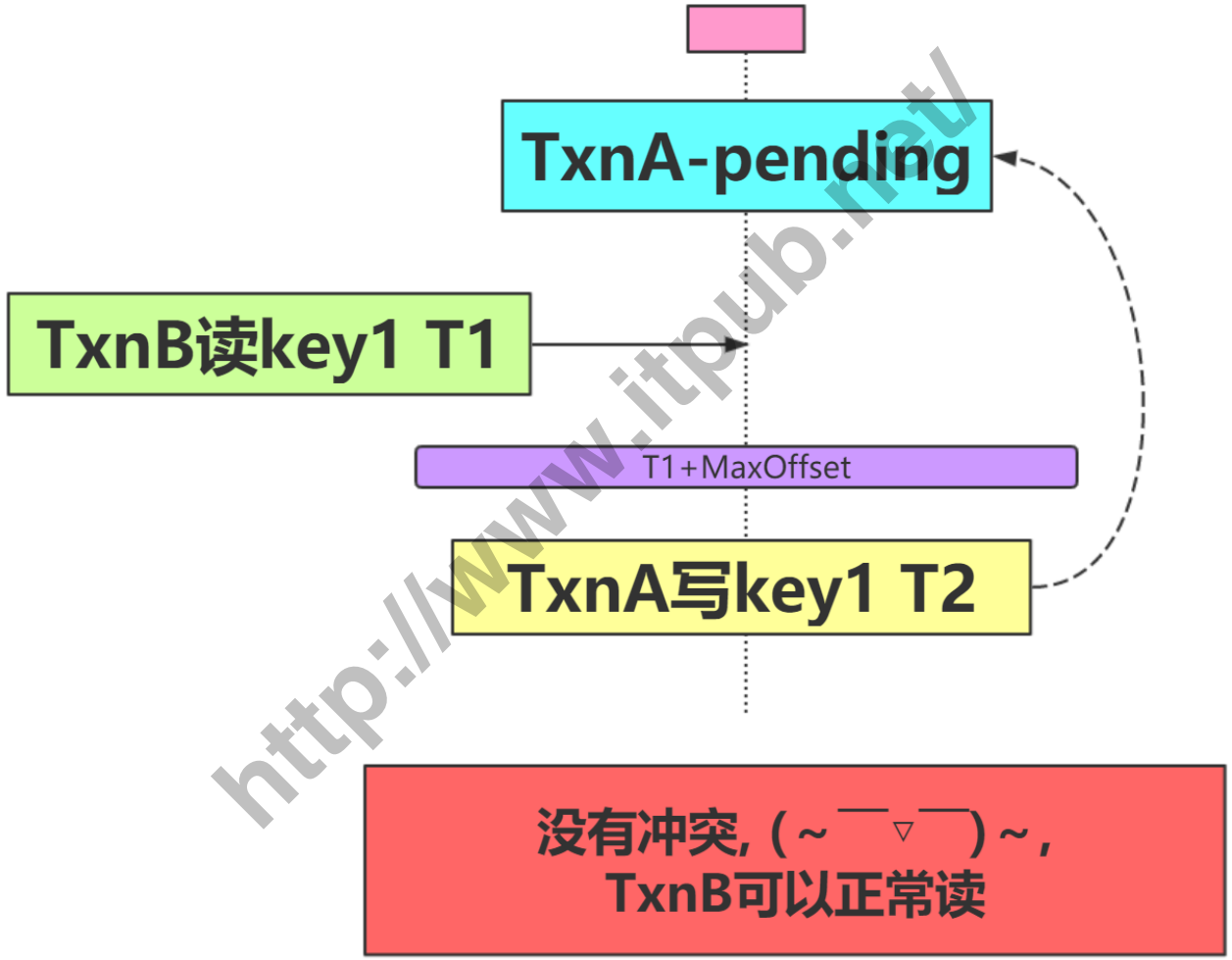
分布式事务 – 写写冲突 Case2



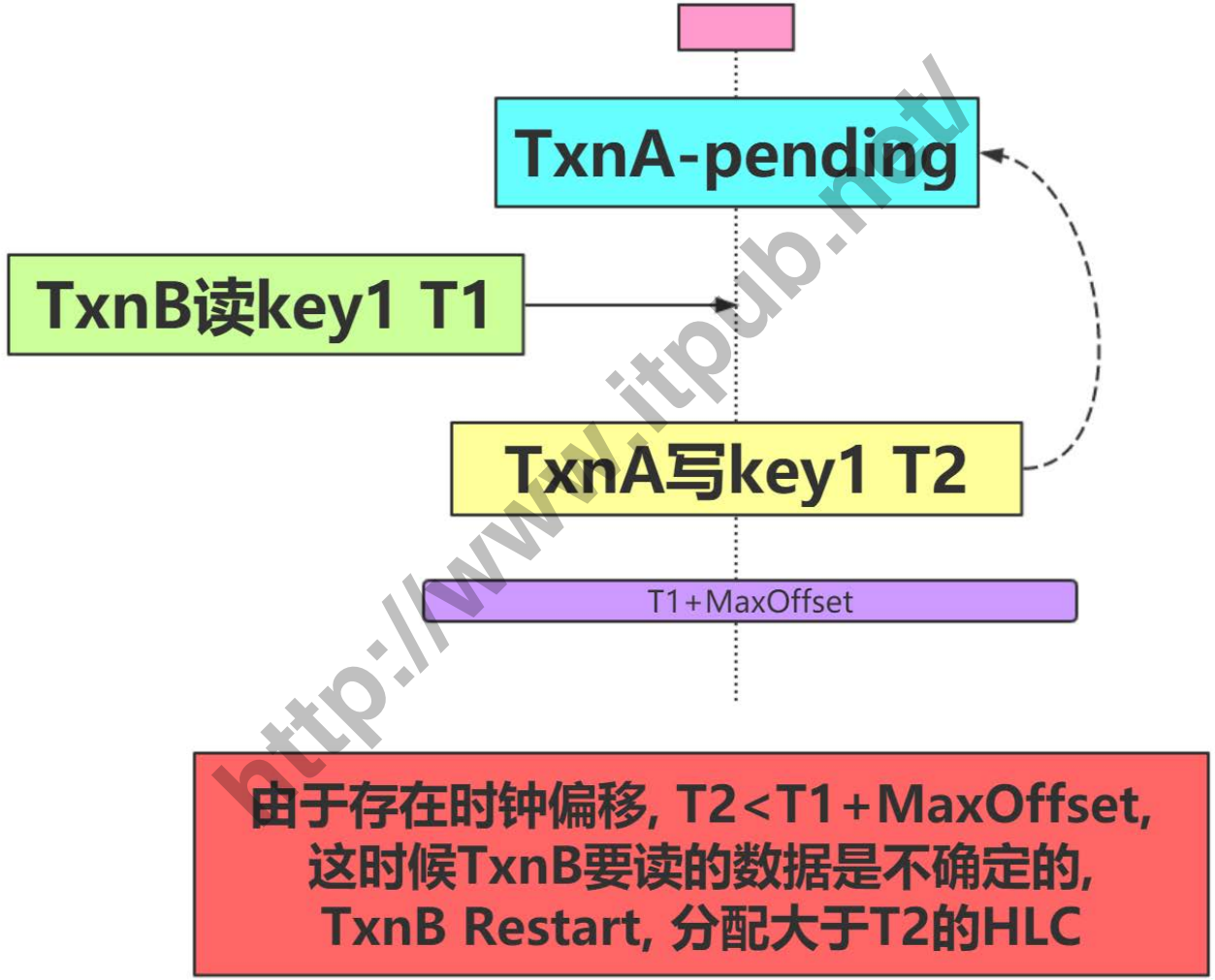
分布式事务 – 写读冲突 Case



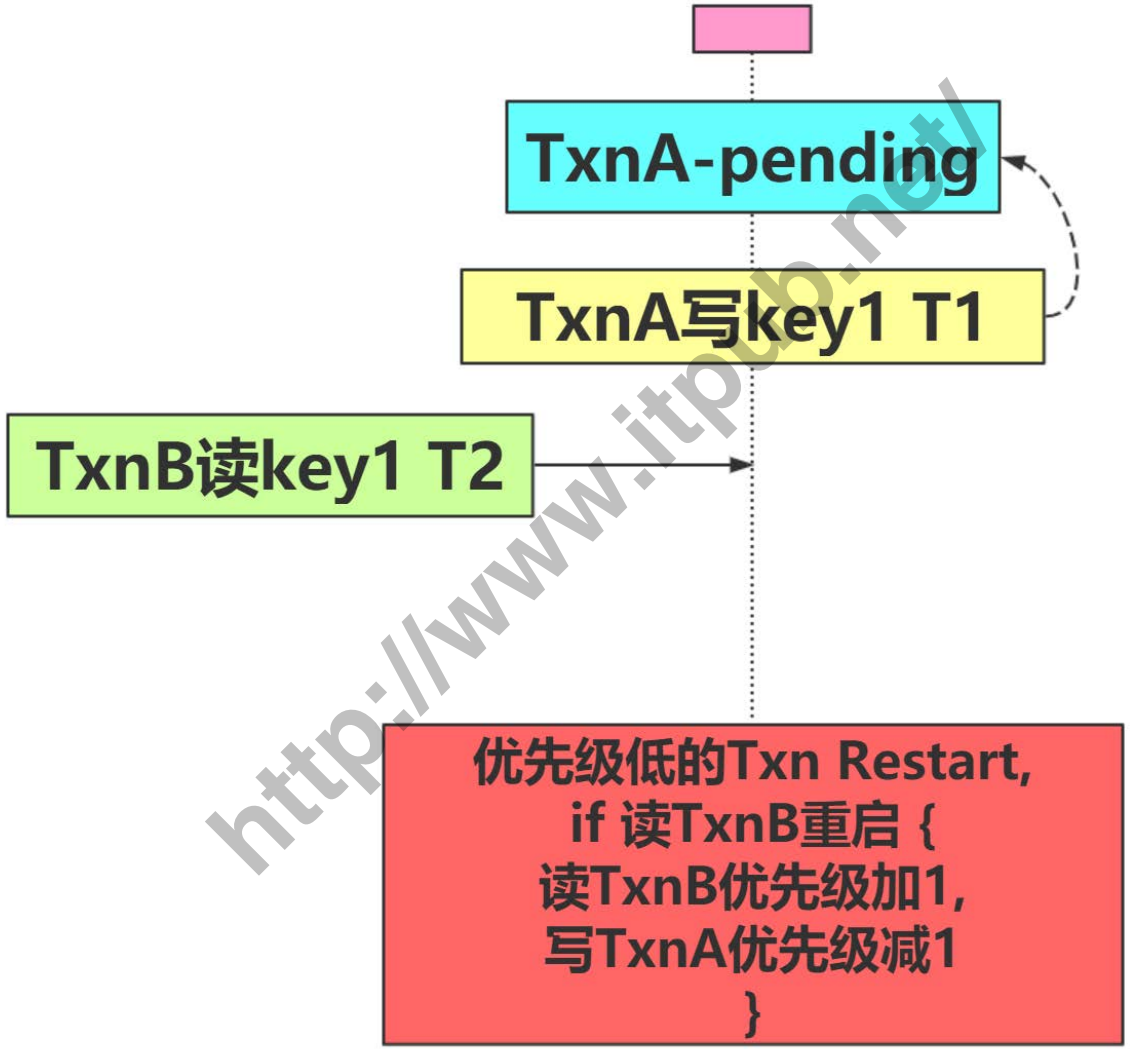
分布式事务 – 读写冲突 Case1



分布式事务 – 读写冲突 Case2



分布式事务 – 读写冲突 Case3



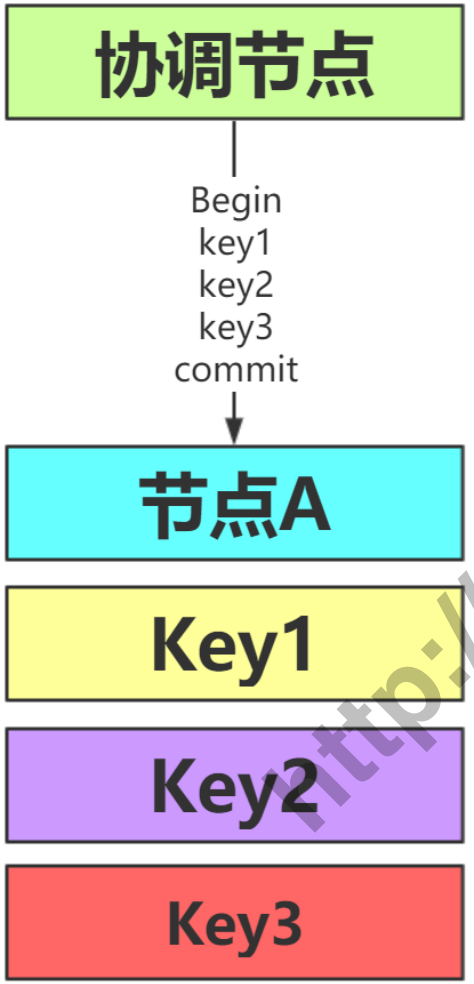
分布式事务 – 优化进阶篇

架构融合
云化共建

2PC和无锁乐观模型已经在分布式事务中成为了MarxDB的基石, 那么还有哪些架构模型可以优化分布式事务的性能和提高吞吐呢?

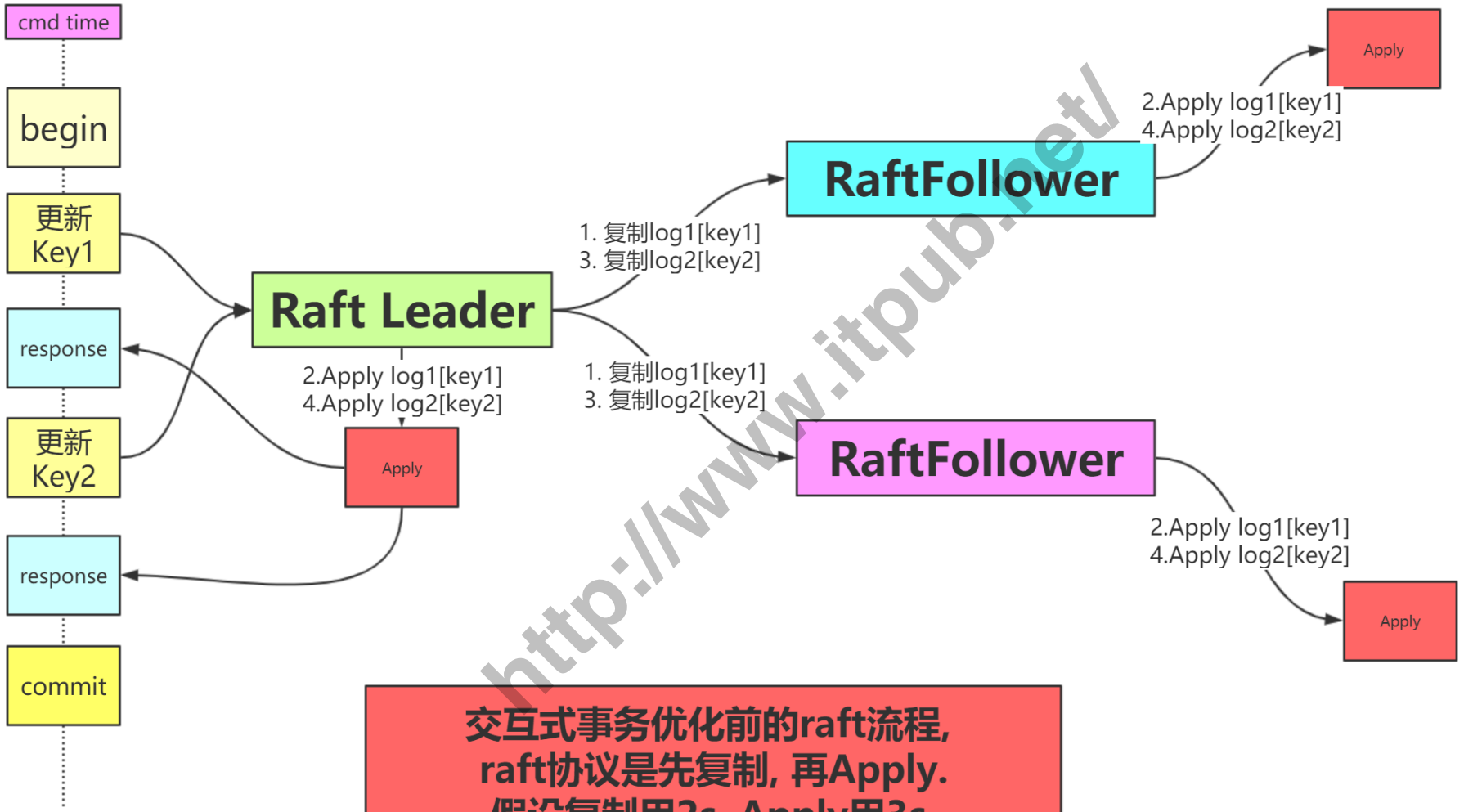
MarxDB目前支持了3种优化方案, 分别是1PC优化, 事务流水线和并行提交.

分布式事务 – 1PC优化



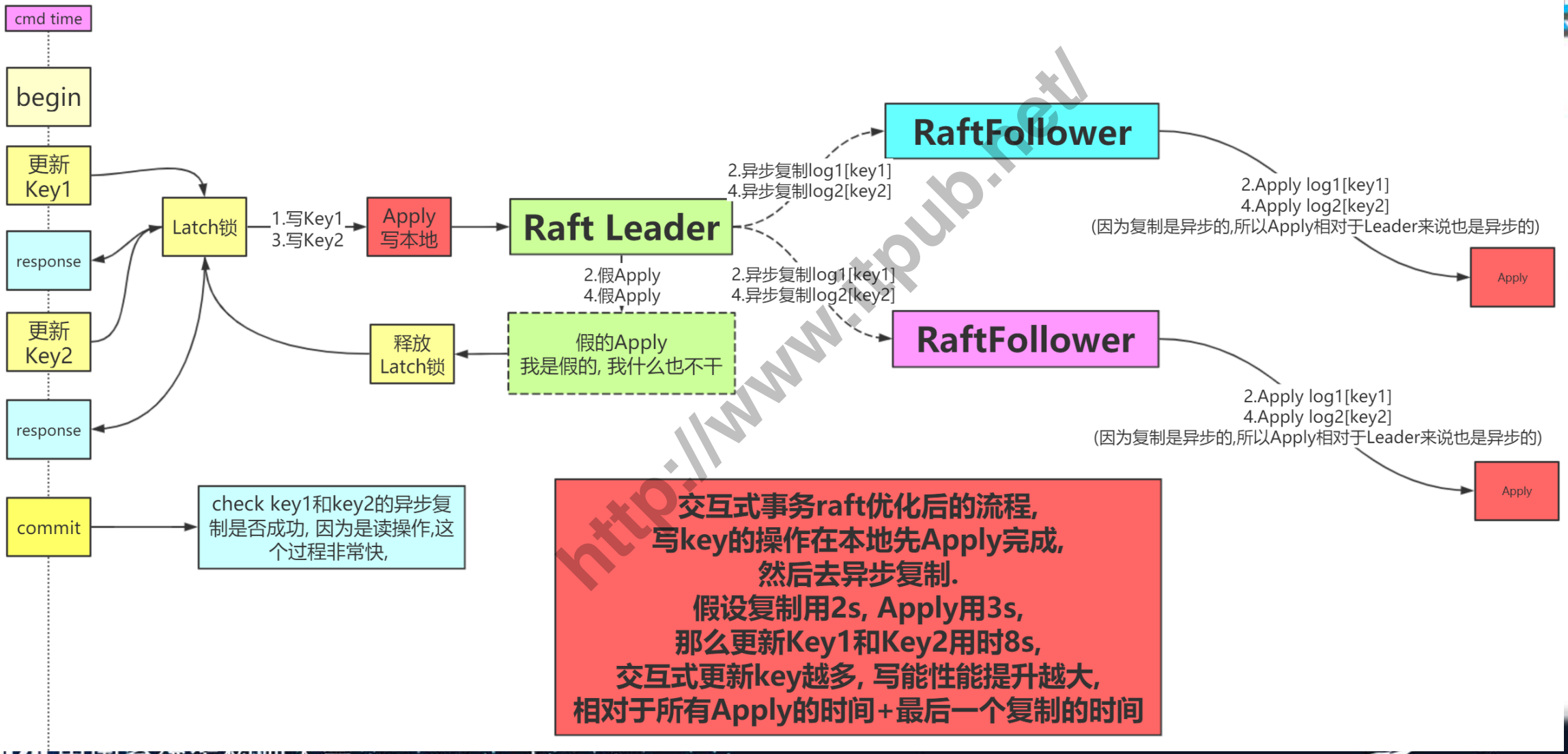
当事务发现所有要修改的数据，
都落在同一个存储节点上，
使用1PC优化.
没有Txn记录，没有prepare，没有commit

分布式事务 – 事务流水线[优化前]



交互式事务优化前的raft流程,
raft协议是先复制, 再Apply.
假设复制用2s, Apply用3s,
那么更新Key1和Key2用时10s

分布式事务 – 事务流水线[优化后]



分布式事务 – 事务流水线

架构融合
云化共建

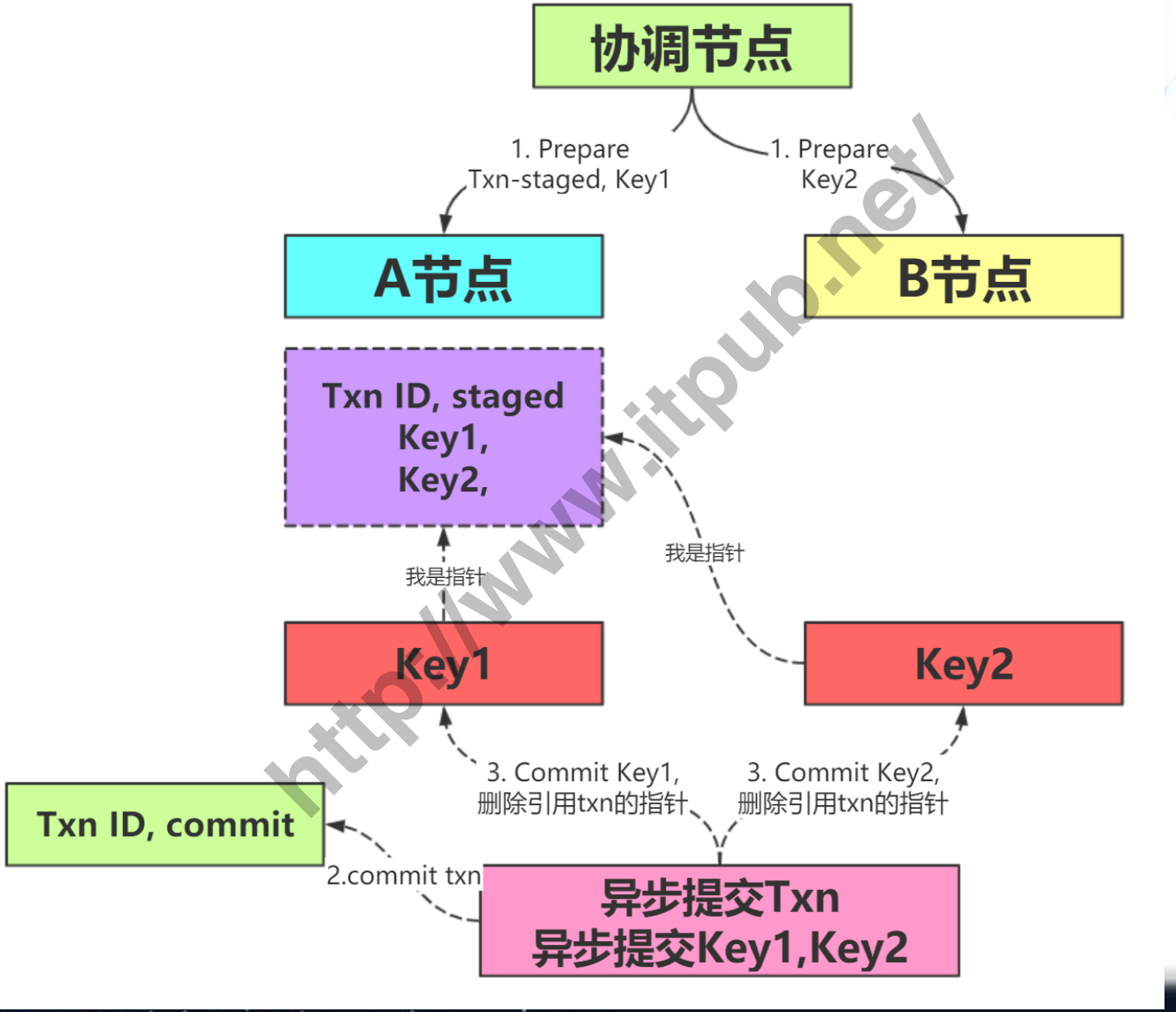
事务流水线理论的关键点:

1.只有Txn事务记录的状态是commit, 才认为事务提交成功.
也就是说如果Txn事务记录的状态不是commit, 在存储节点里写的Key都对外无效.
所以raft才可以先本地Apply, 后异步复制, 根本就不怕你复制失败啥的, 顶多就是Txn事务失败,
不会有任何脏数据.

2. Txn事务记录提交之前, 需要检查复制是否成功, 读的过程非常的快.

3.用了一个Latch锁, 防止多个事务并发Apply同一个raft group, 因为raft的Apply必须是串行的.

分布式事务 – 并行提交

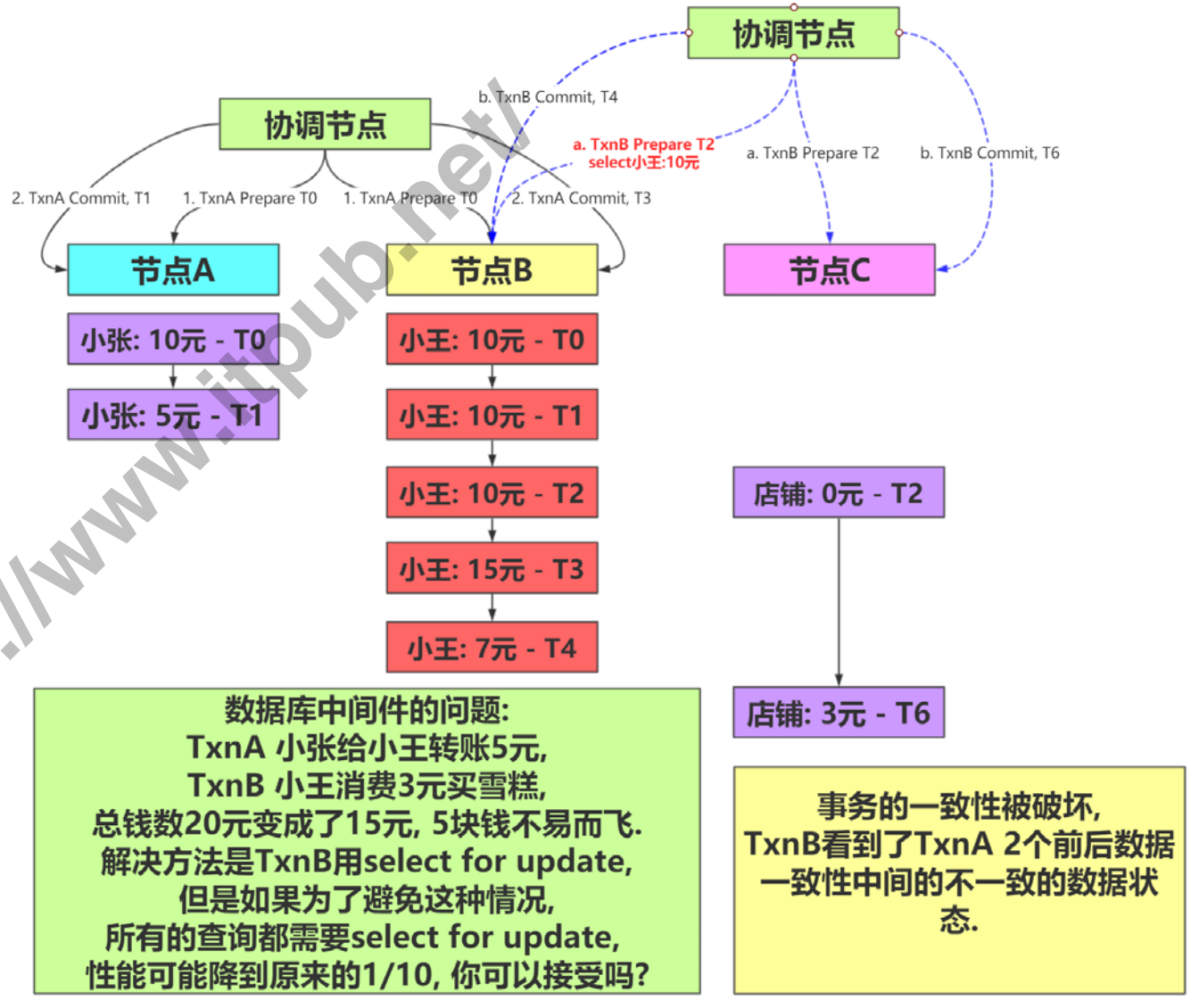


分布式事务 – 数据库中间件2PC的问题

数据库中间件分布式事务方案:
2PC
AF / XA

中间件的事务一致性是数据最终一致性
BA: Basic Availability
基本可用, 最终一致

MarxDB就没有这样的问题,
通过HLC和MVCC保证一致性.

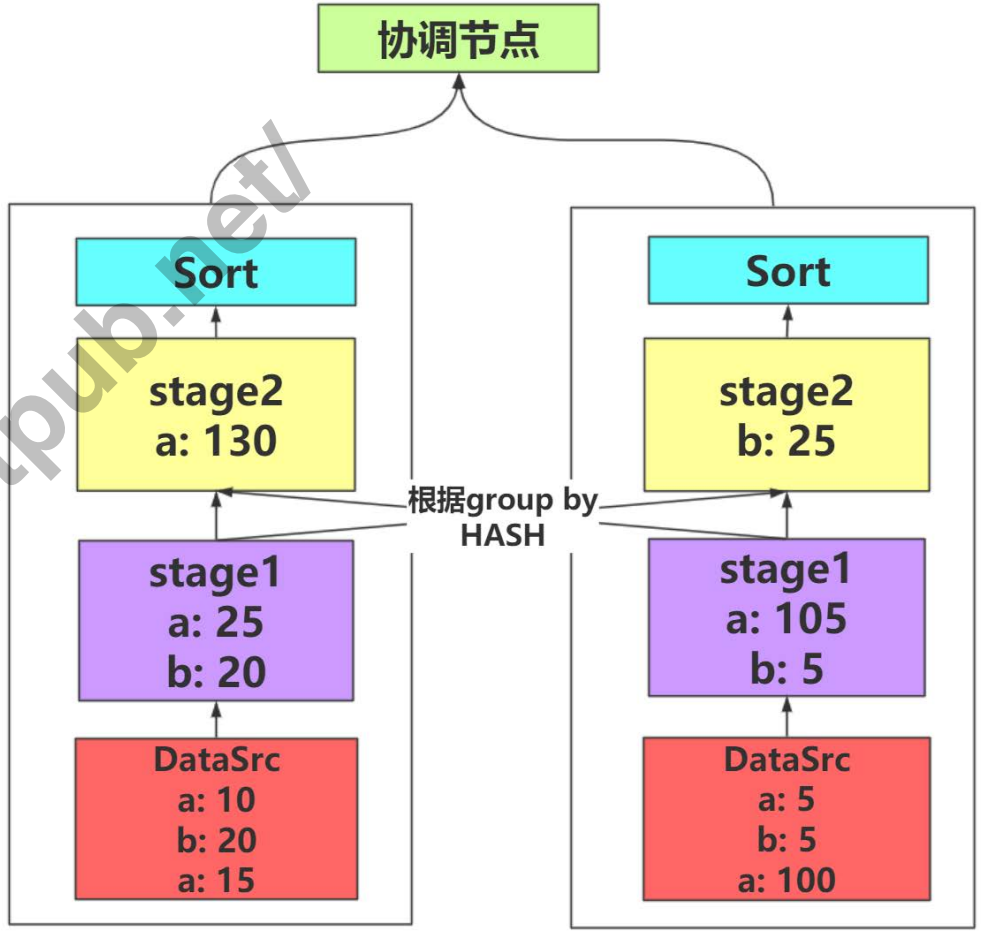


物理执行计划 – group by

火山模型

DataSrc: 数据源
Stage1: 本地先Aggr聚合, 然后根据group by的列, 把数据hash分配到不同的存储节点上.
Stage2: 收到所有节点hash到本节点的Aggr聚合数据, 在本节点再此进行Aggr聚合.
Sort: 排序

优势:
所有group by的列, 在Stage2阶段已经完成了聚合, 不再需要协调节点再次Aggr聚合.
CPU计算和内存使用均衡.

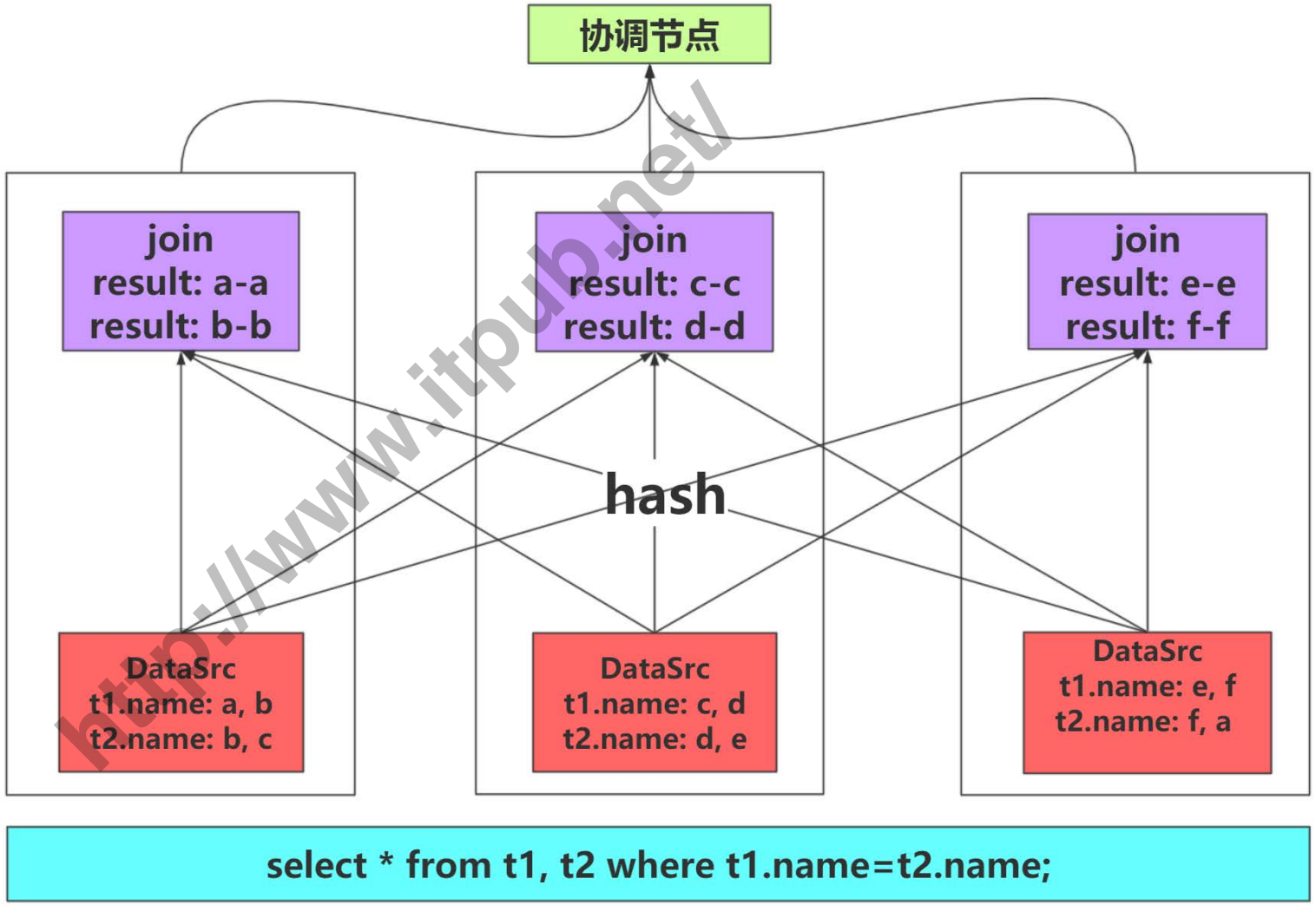


物理执行计划 – join

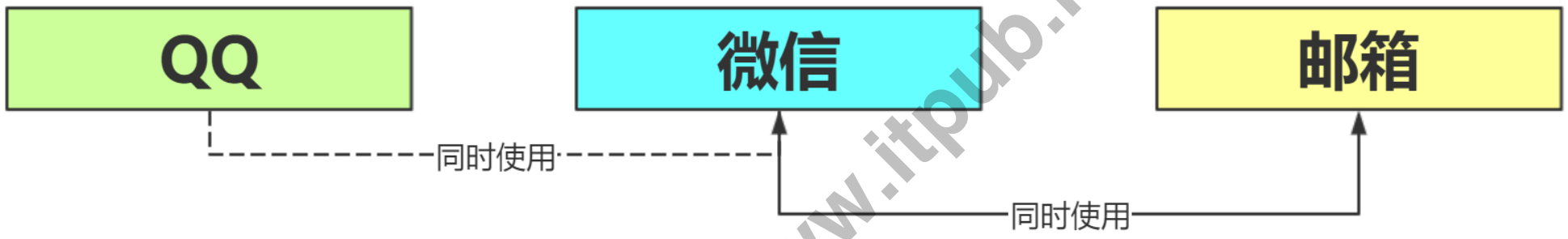
火山模型

DataSrc: 数据源
Join: join执行器

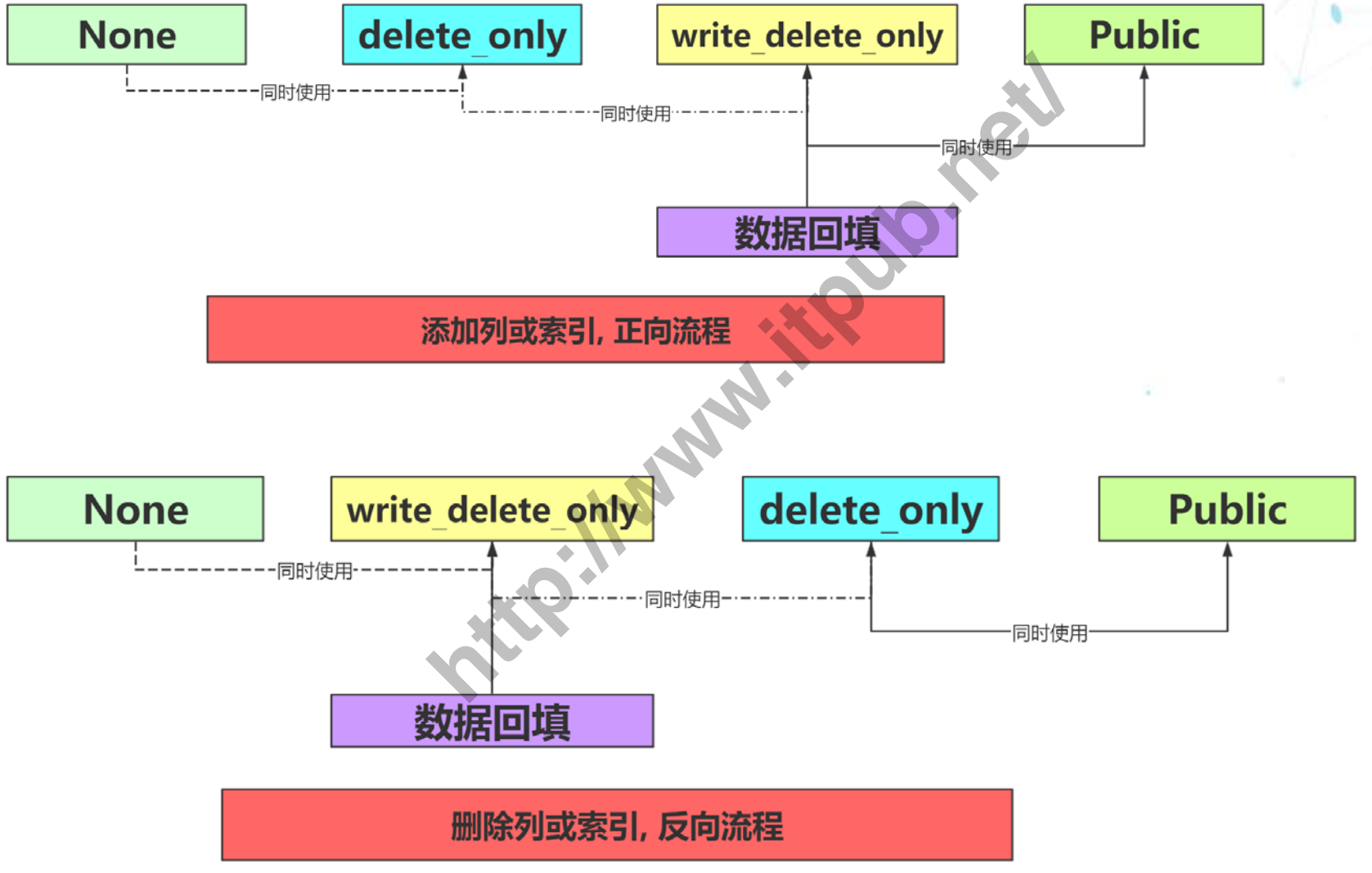
优势:
不需要**协调节点**再进行join.
CPU计算和**内存**使用**均衡**.



在线表变更



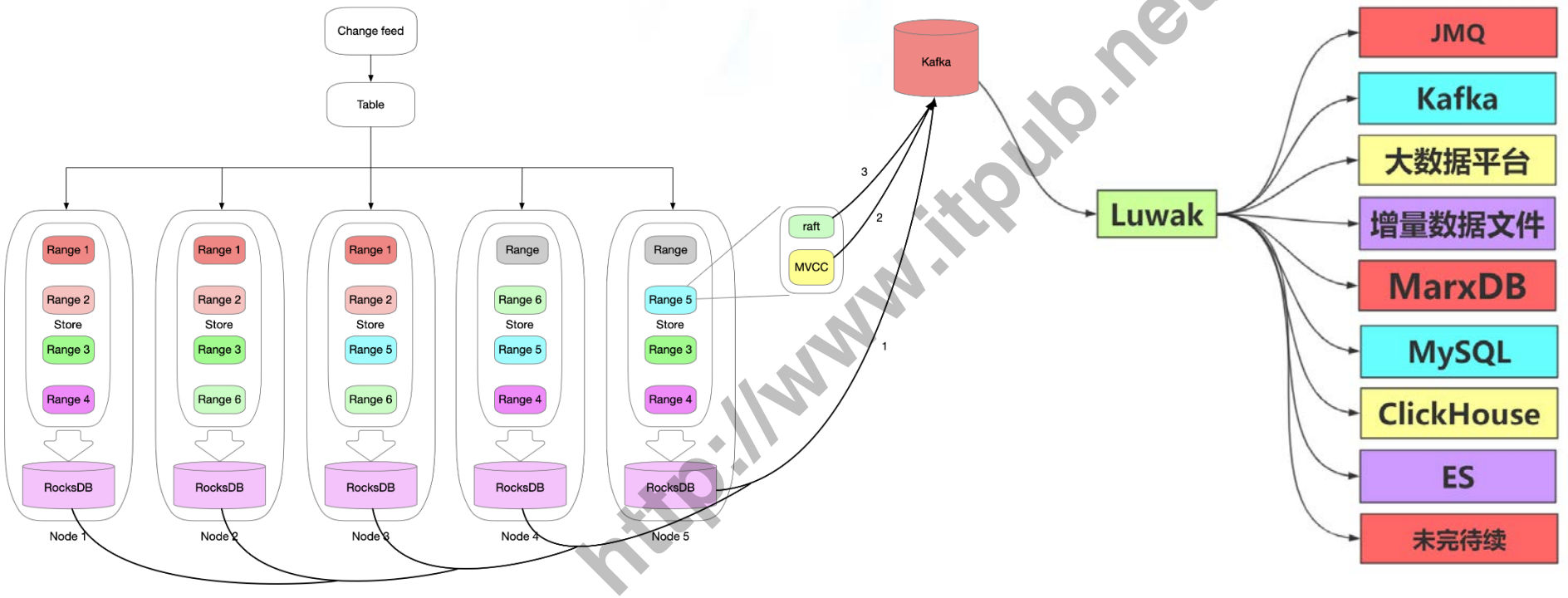
在线表变更



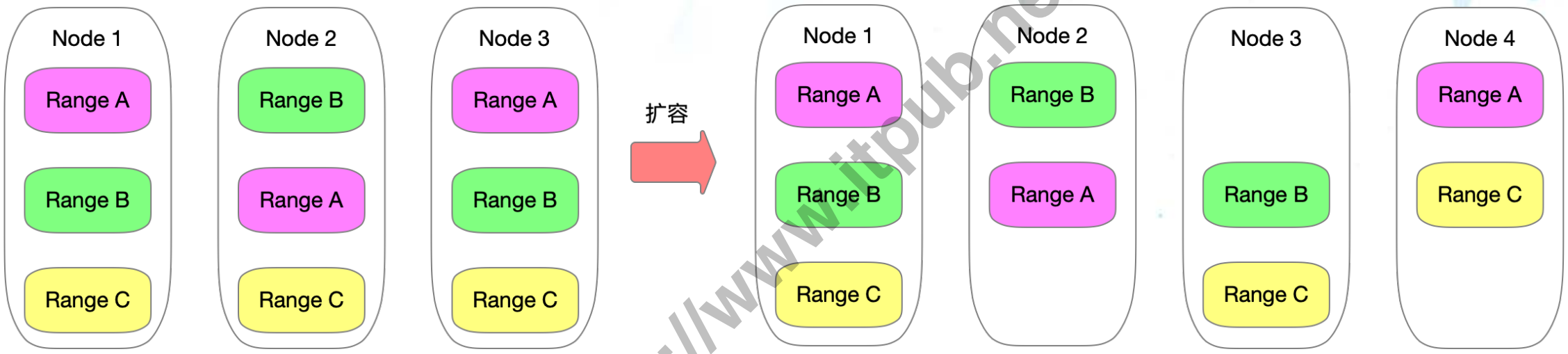
增量数据同步 – CDC捕获数据变更

架构融合
云化共建

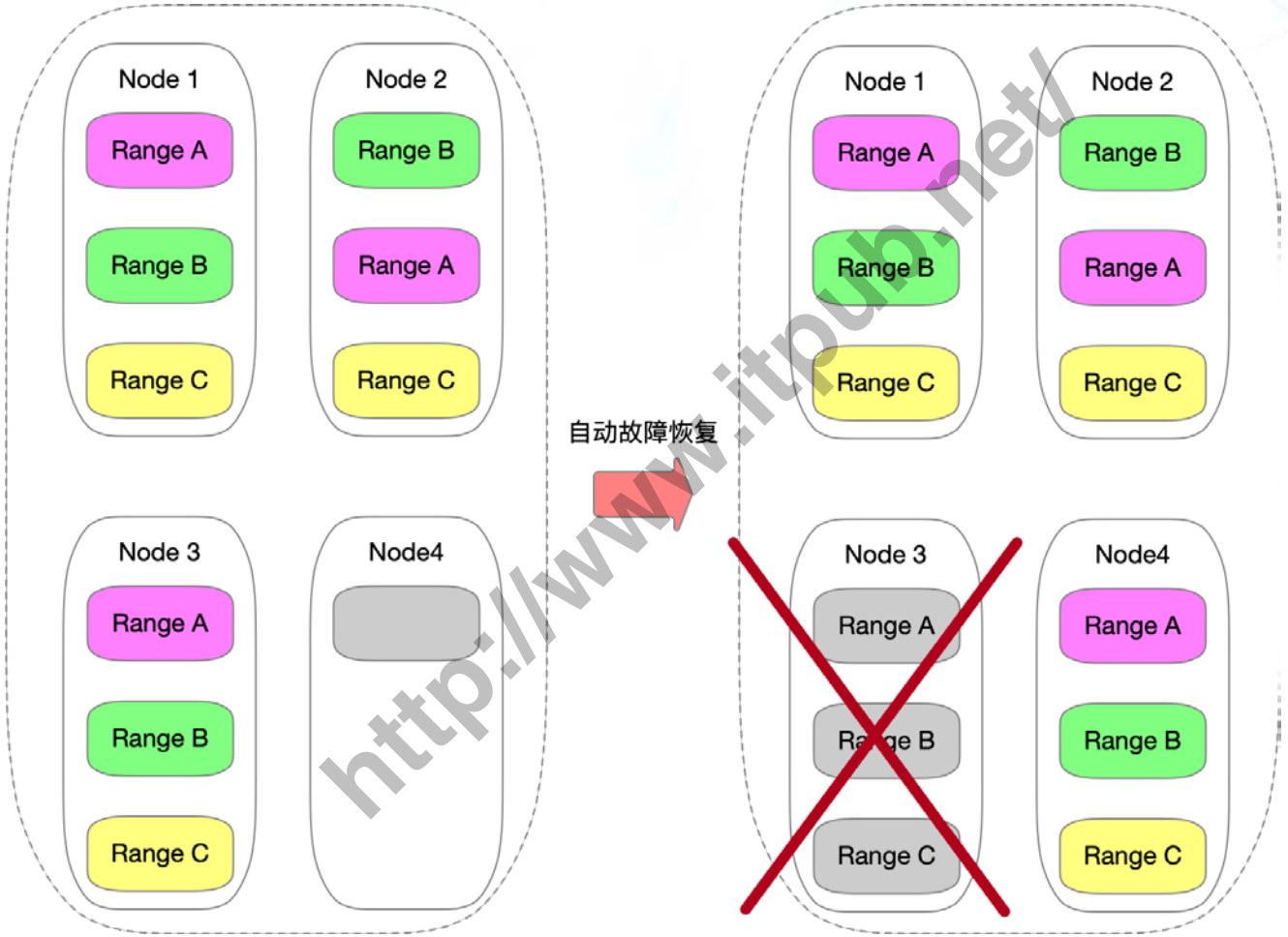
支持 DDL 和 DML的变更



弹性伸缩 – rebalance



故障自愈



跨地域部署

架构融合
云化共建

支持业务数据按地域分区存储

支持业务冷热数据分离

支持业务数据就近访问

招聘

架构融合
云化共建

MarxDB团队

newsql@jd.com

请把简历丢在我的脸上

(~￣▽￣)~

Thanks