DTCC

第十四届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA

数智赋能 共筑未来

IT168.com ChinaUnix ITPUB

北京国际会议中心 | 2023/8/16-18

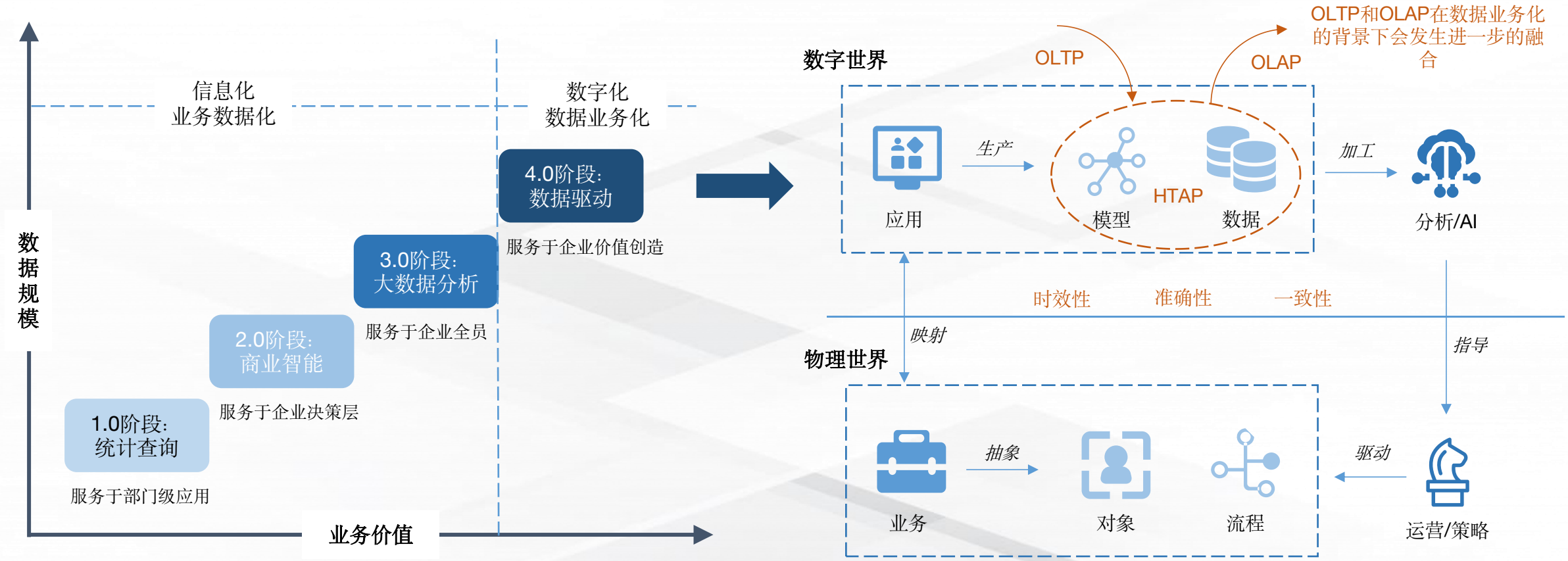
分布式数仓的TP能力探索——HashData UnionStore

HashData 资深解决方案架构师

陈义贤

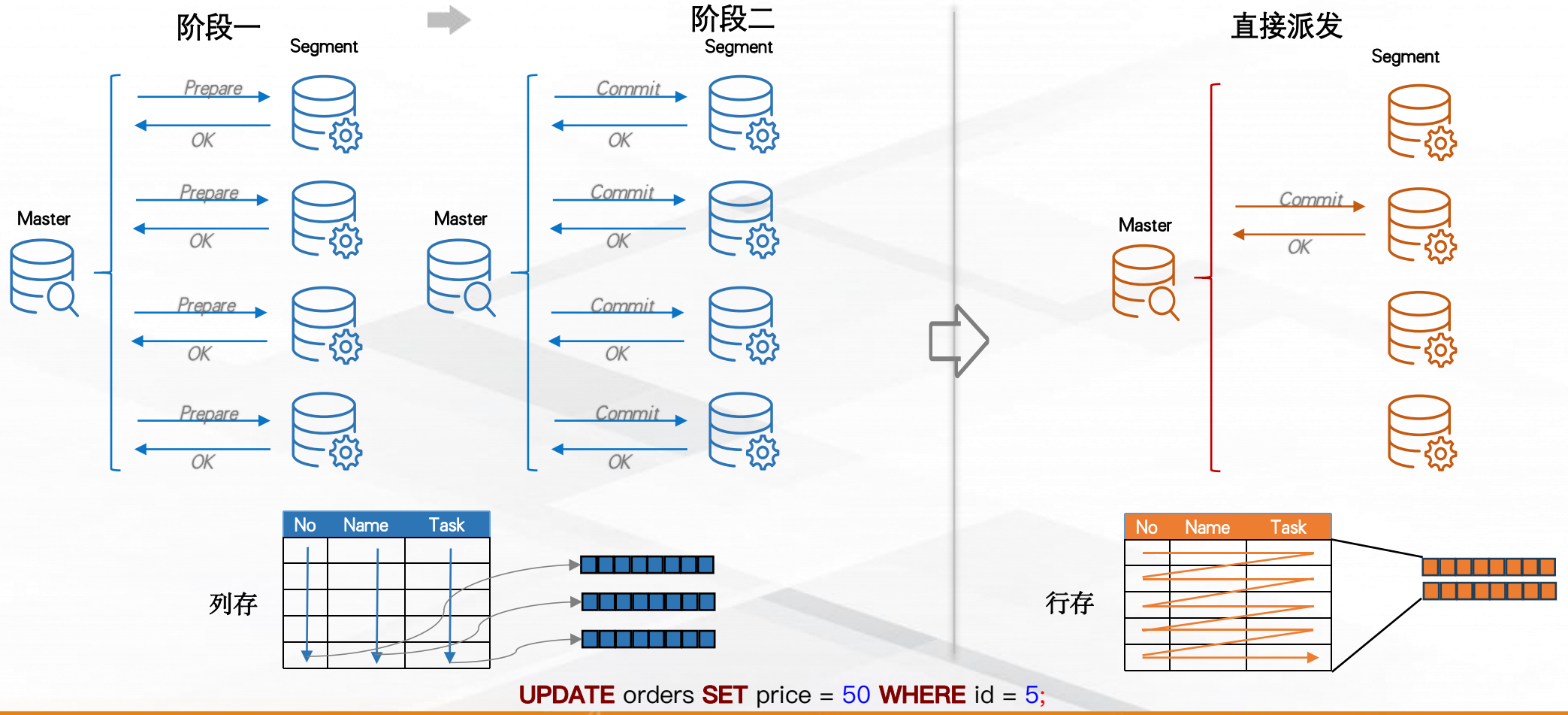
01 数据驱动——围绕数据资产构建应用的趋势发展

背景：企业IT会从信息化转向数字化，形成IT即业务的新格局，数据会成为实体世界的业务在数字世界的投影；由此，应用系统在建设之初，就需要考虑数据应用，并将数据资产描述出来，整体规划数据技术架构，避免数据孤岛，提升数据资产运营效率



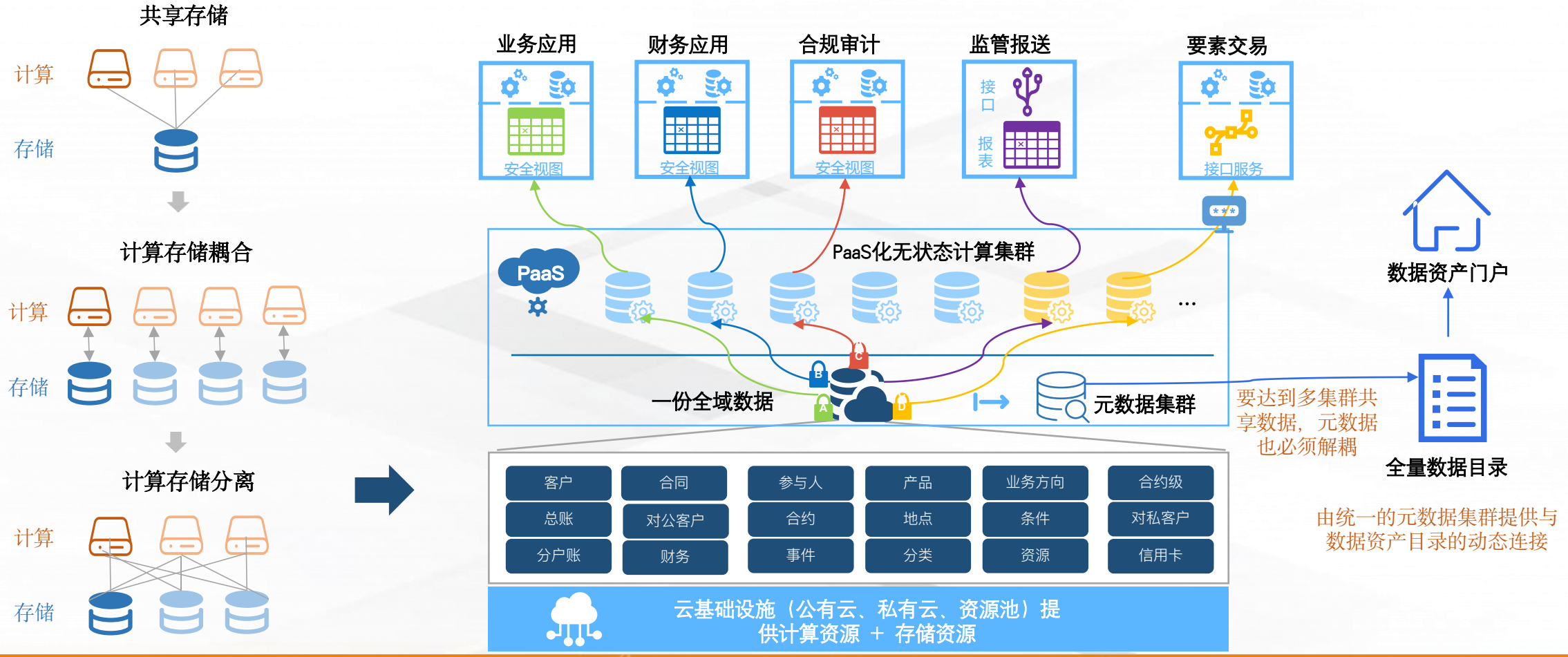
02 传统MPP分布式数仓提升OLTP能力的尝试

TP应用场景，由2PC转为直接派发，减少prepare阶段的实例等待，从而提升单条数据增删改的性能；同时，针对单条记录增删改的数据使用行存表提升性能



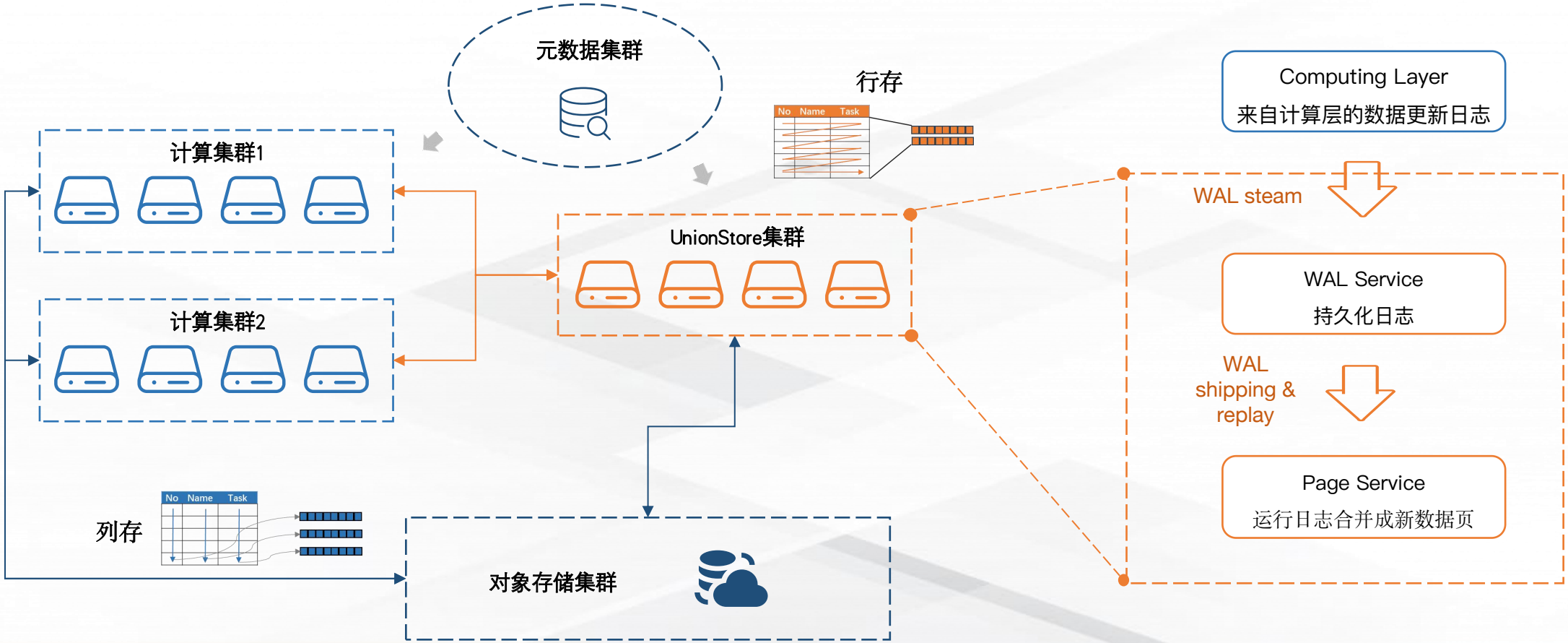
03 存算分离架构将成为未来数据架构的基本要求

云原生架构的核心理念要将存算分离，使用对象存储来保存一份全域数据，所有计算集群均为无状态，按需申请使用，也可以兼容各种不同计算引擎，满足各类不同业务的要求



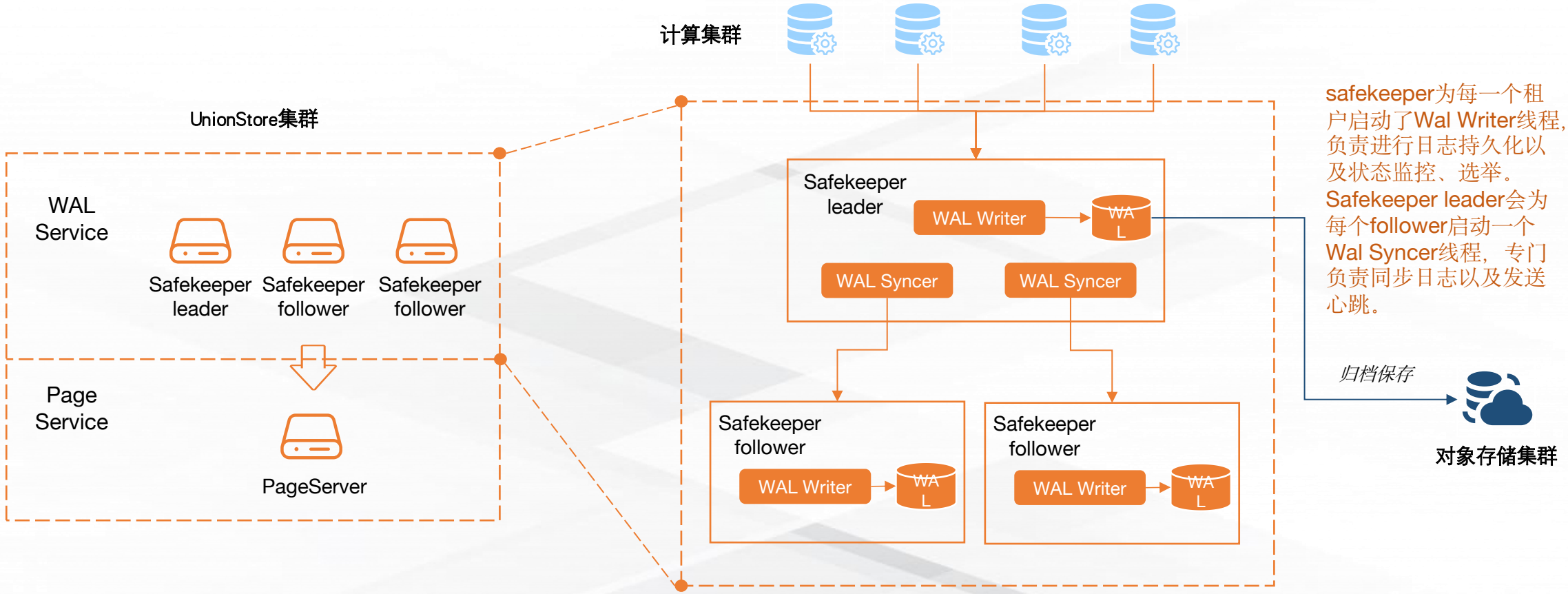
04 存算解耦后，使用不同引擎分开处理数据成为可能

Log is database的理念可以大幅优化OLTP能力，通过将数据随机写入的操作剥离，计算集群只将WAL日志提交至UnionStore集群，由UnionStore集群处理日志数据，并重放生成新的页数据，这样减少了复杂的锁定和同步操作，可以大幅提升并发能力，同时也减少随机访问



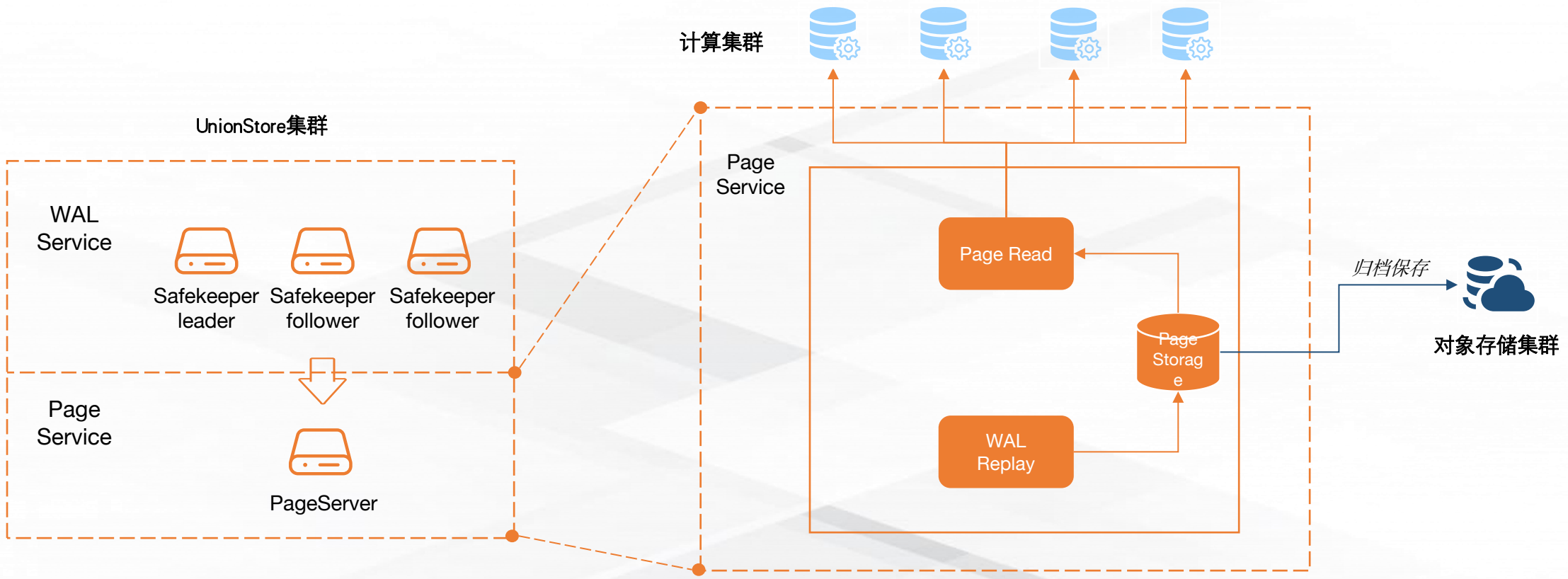
05 WAL Service

Wal Service为了保证日志持久化后的可靠性，日志通常会保存3副本。由leader节点负责接收计算集群请求，本地持久化同时将日志发送到follower节点，当所有节点都完成日志持久化之后，leader节点才会返回给计算集群。



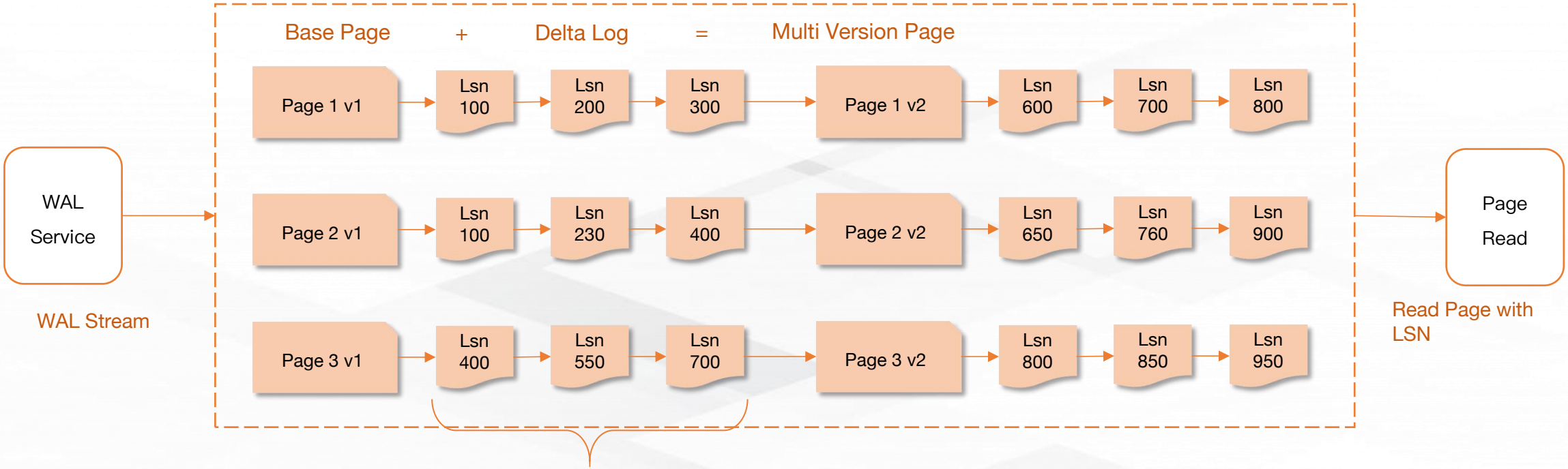
06 Page Service

Page Service主要负责从Wal Service(safekeeper leader)获取已经持久化日志并进行解析，通过重放日志去修改page数据；此外还会对计算集群提供更新后的Page读取服务。



07 Page存储形式

Page Service将当前日志作为page的delta log，通过base page + delta log方式来构建对应page的多版本。



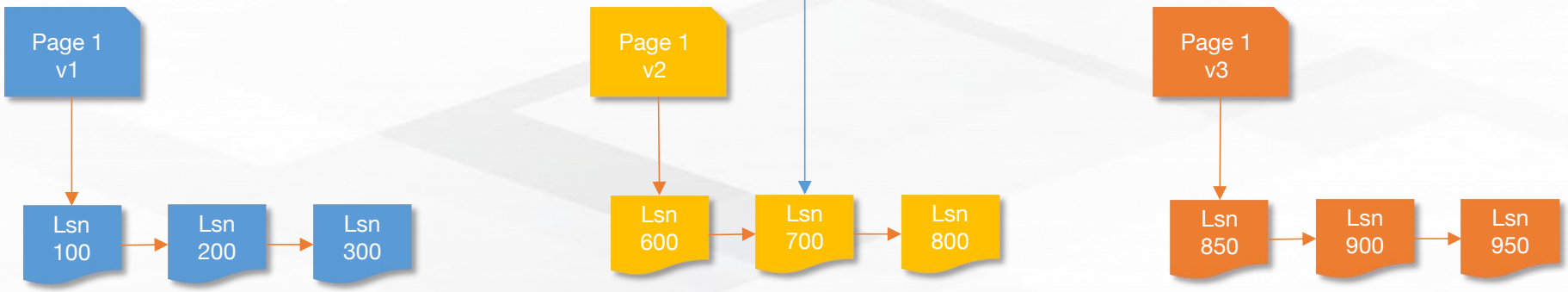
Delta log积累到一定量（默认64MB，可自定义），会checkpoint生成新的Page版本，作为新的Base page，从而提升Page的访问性能，减少延迟。
旧版本的Base page和Delta log的空间会被回收，回收的条件取决于是否已经上传至对象存储进行归档，且不在Time travel设置的周期范围内。

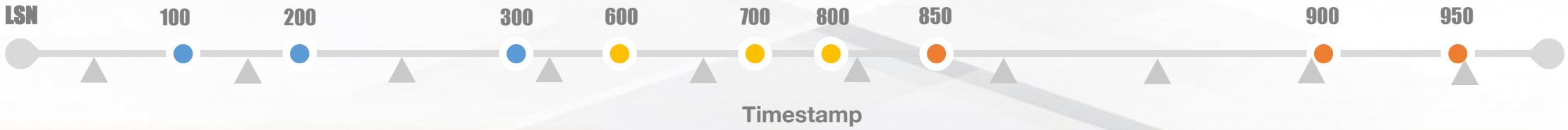
08 Time Travel

读取Page时需要带上对应的LSN，这里LSN就是一个快照。Page Service会根据请求的LSN来确定对应的Page版本。比如使用LSN 700读取Page1，则Page Service会先获取base page，然后根据LSN 700确定delta log范围，这里就是lsn600和lsn700两个delta log，然后将日志按顺序apply到base page v2，会成对应page版本返回

搜索Timestamp时间并找到对应的LSN为700

xid	commit_time	lsn	operation
15331847	2023-05-30 06:57:08.508	300	WRITE
15332107	2023-05-31 11:28:11.301	600	WRITE
15332113	2023-05-31 11:29:58.323	700	WRITE
15332166	2023-05-31 11:31:02.756	800	WRITE





09 Time Travel使用场景

恢复数据库对象

通过追溯Page版本和LSN，可以将数据恢复到任意时间点。误删除的表，Schema和库，可以直接将数据恢复到误操作之前时间点。

```
UNDROP DATABASE
mydatabase;
UNDROP SCHEMA
myschema;
UNDROP TABLE mytable;
```

查询历史数据

可以查询任意时间点的数据。获取数据在某个时间段的变更历史、增量统计用于决策分析；例如通过CDC数据入库，可以在不制作拉链表的情况下，直接选择统计数据的时间点

```
SELECT * FROM employee
at(timestamp => '2023-05-31
11:29:58.323');
```

历史数据克隆

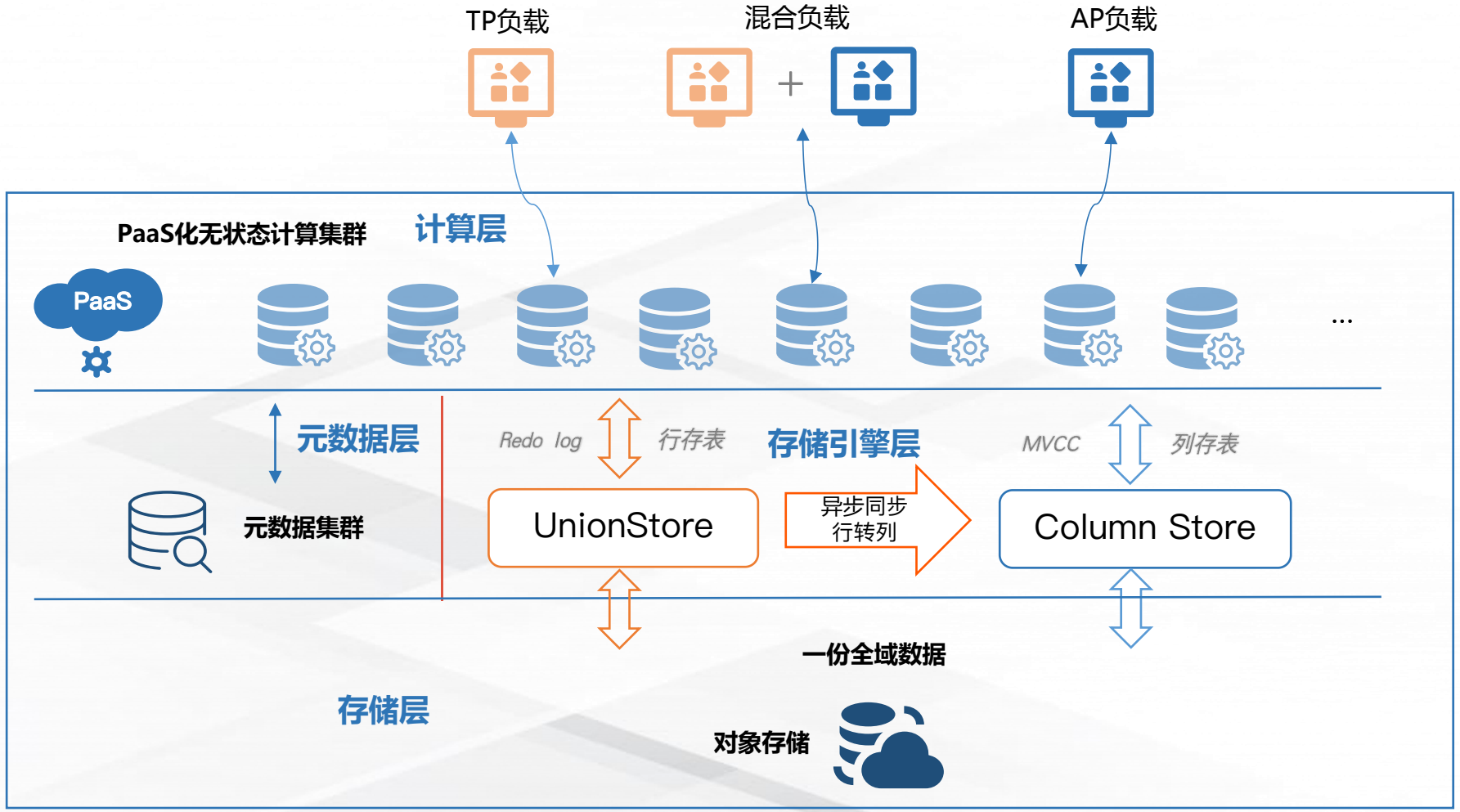
创建任意时间点数据的拷贝。辅助数据模型训练，基于某个时间点训练结果创建多份数据拷贝，使用不同参数进行训练，对比训练结果

```
CREATE TABLE
restored_employee CLONE
employee
at(timestamp => '2023-05-31
11:29:58.323');
```

HashData云原生统一架构HTAP数据平台

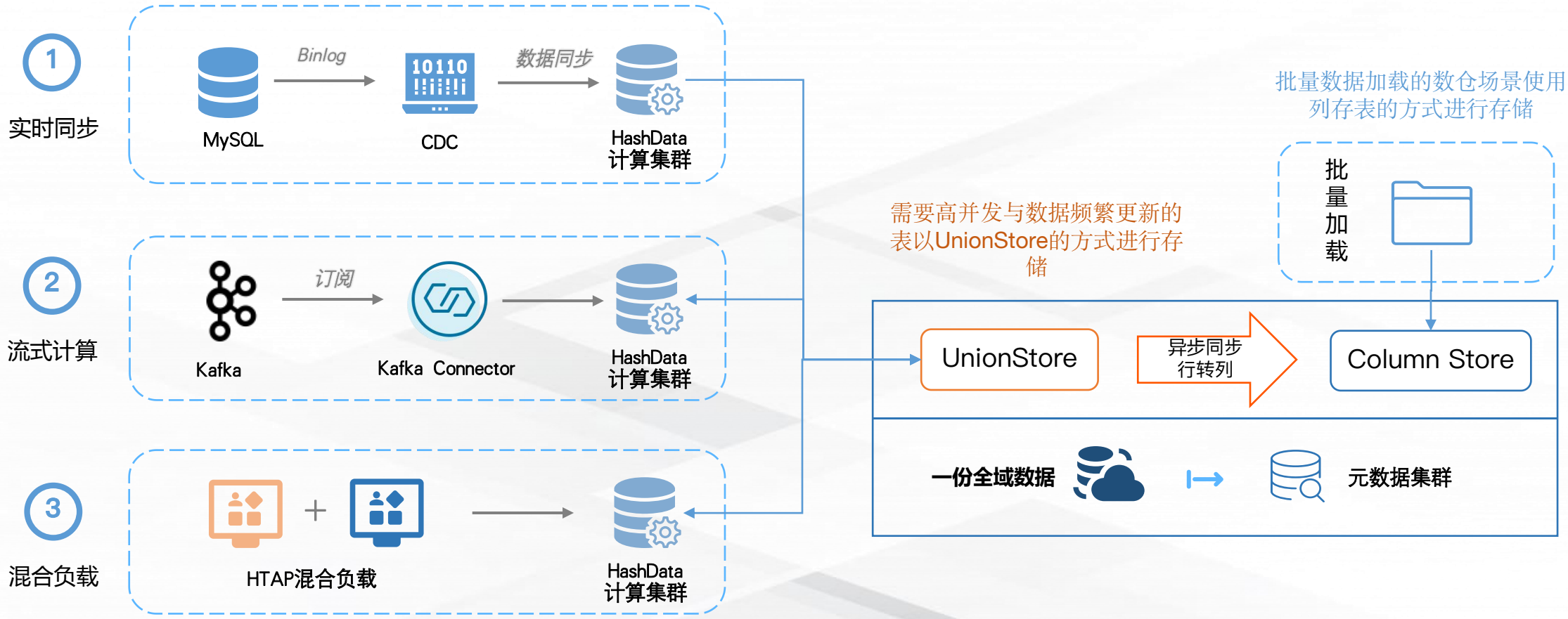
分钟↘毫秒

把传统数仓中分钟级响应的
OLTP负载提速至毫秒级



11 UnionStore应用场景

UnionStore表与列存表都存在对象存储中，可以在任一计算集群中关联查询，灵活应对各类数据应用场景



THANKS