互联网金融行业HBase实践与创新

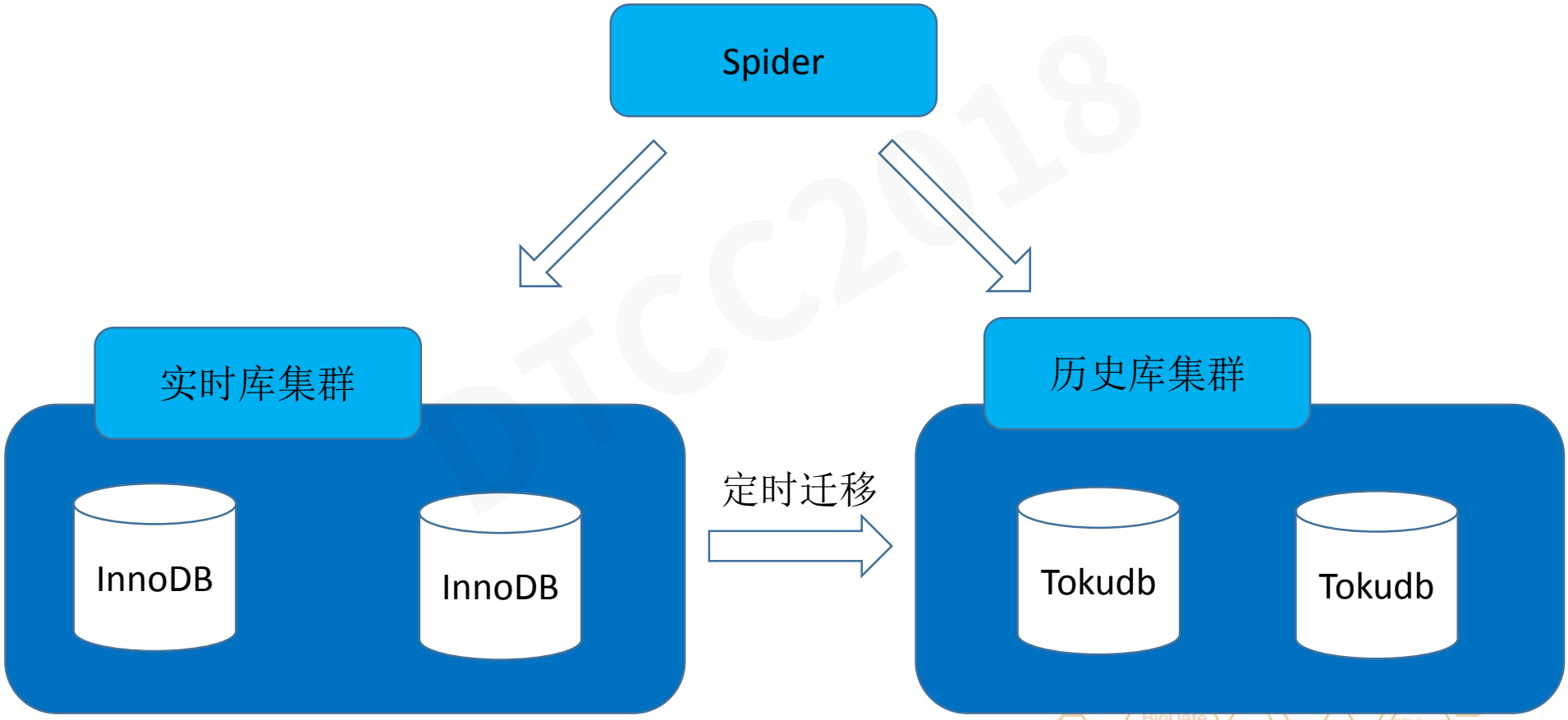
徐春明

目录

1 关系数据库挑战	扩展性、可维护性、易用性

背景

Spider + Tokudb架构



系统挑战

水平扩展	数据与日俱增，单机性能存储达瓶颈，架构难适应目前数据规模存储，需考虑数据拆分，传统DB拆分困难，工作量大且复杂，周期长
数据迁移	传统DB会定期将线上数据迁移至历史DB，需建立复杂的迁移策略，易出错；历史库备机写入性能低，延迟大
故障处理	历史DB容量大，计算资源配置低、性能差、易宕机，一旦故障迁移数据周期长，成功率低
批量查询	传统历史DB采用TokuDB引擎，数据高度压缩，磁盘IO、内存性能较差，大批量数据扫描对单机性能影响很大，易拖垮历史库
业务查询	按时间分区，跨分区查询数据不准，易错乱，批量扫描性能差易拖垮spider实例，影响查询

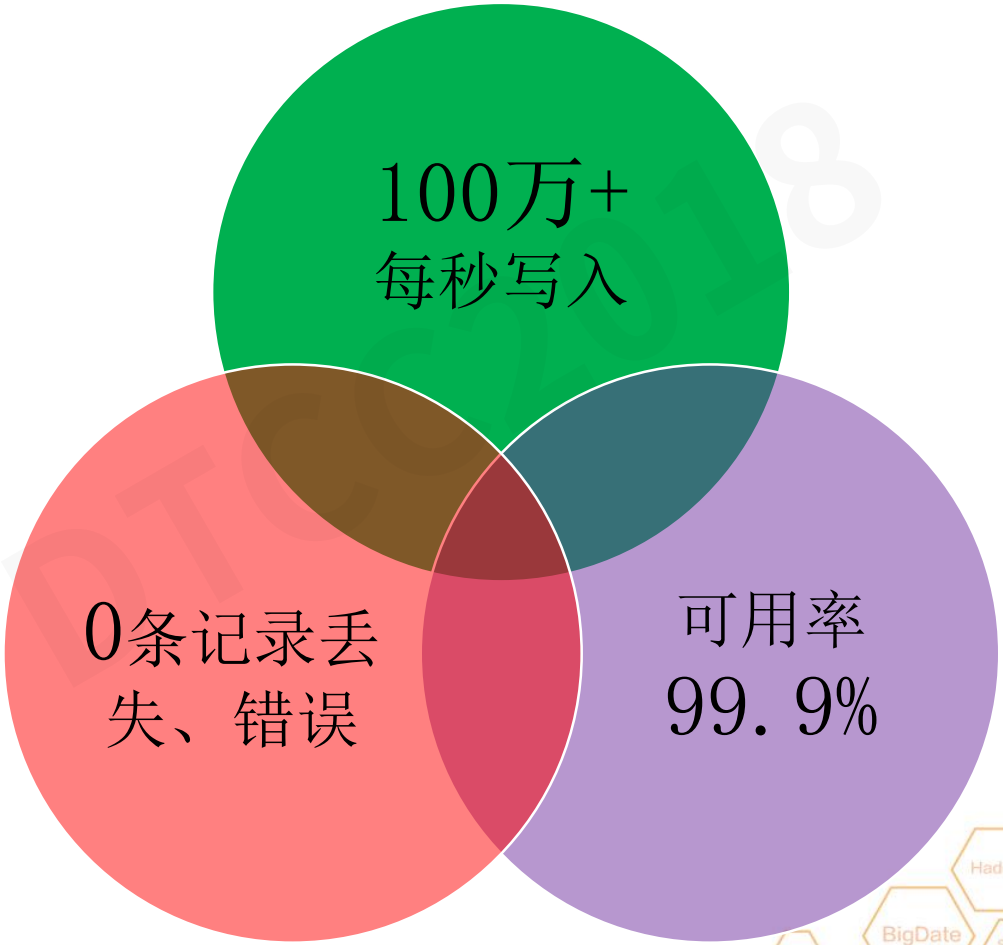
方案对比

特性	Redis	Mongodb	HBase
数据存储方式	内存型	文档型	列式存储
数据存储格式	KV	BSon	KV
数据读写性能	高	读高写低	写高读低
数据完整性	易丢失	高	高
数据一致性	弱一致	弱一致	强一致
数据安全性	低	中	高
扩展性	不灵活	中等	灵活

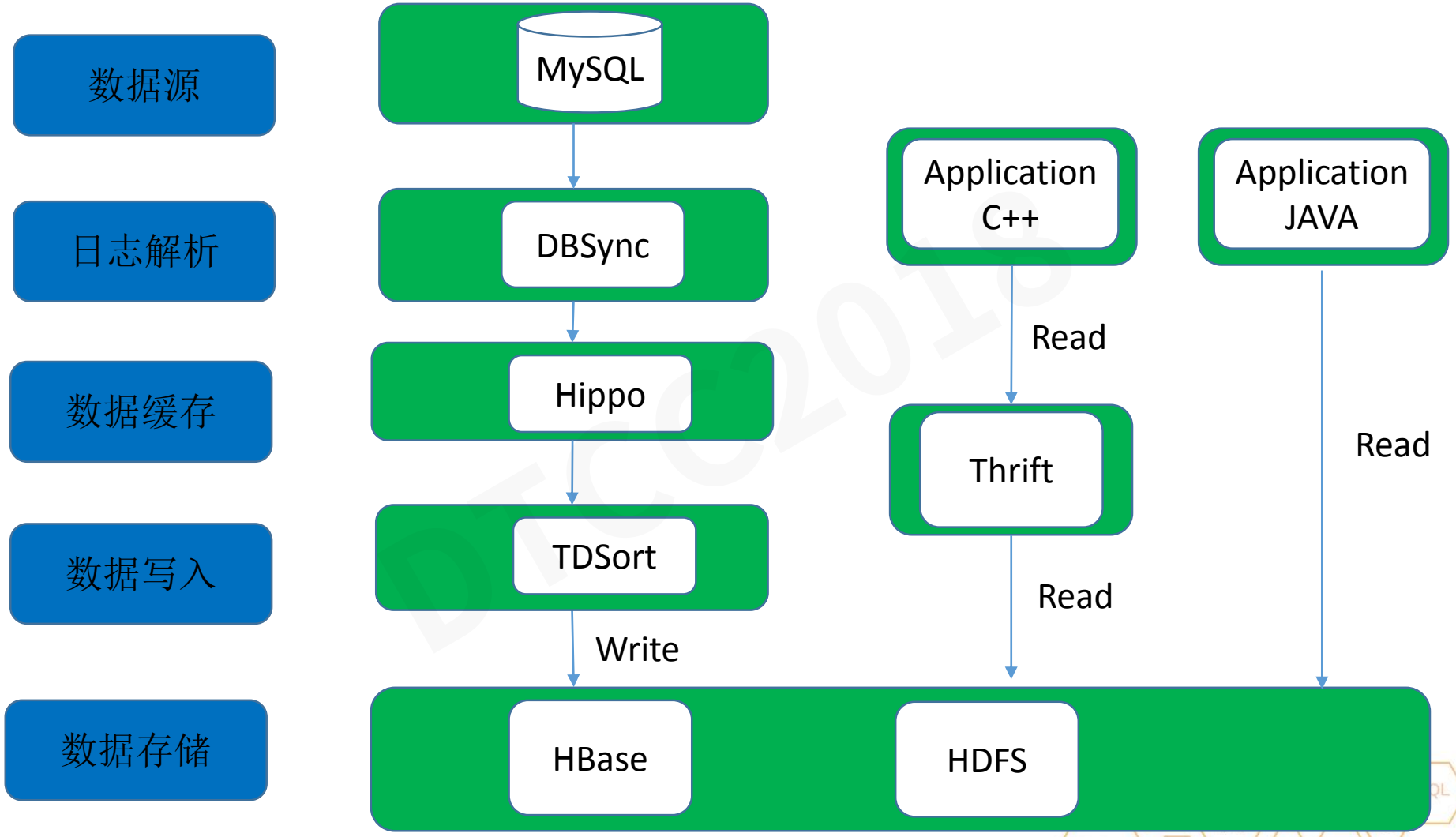
2 如何打造HBase中心	高性能、可靠、高可用

目标

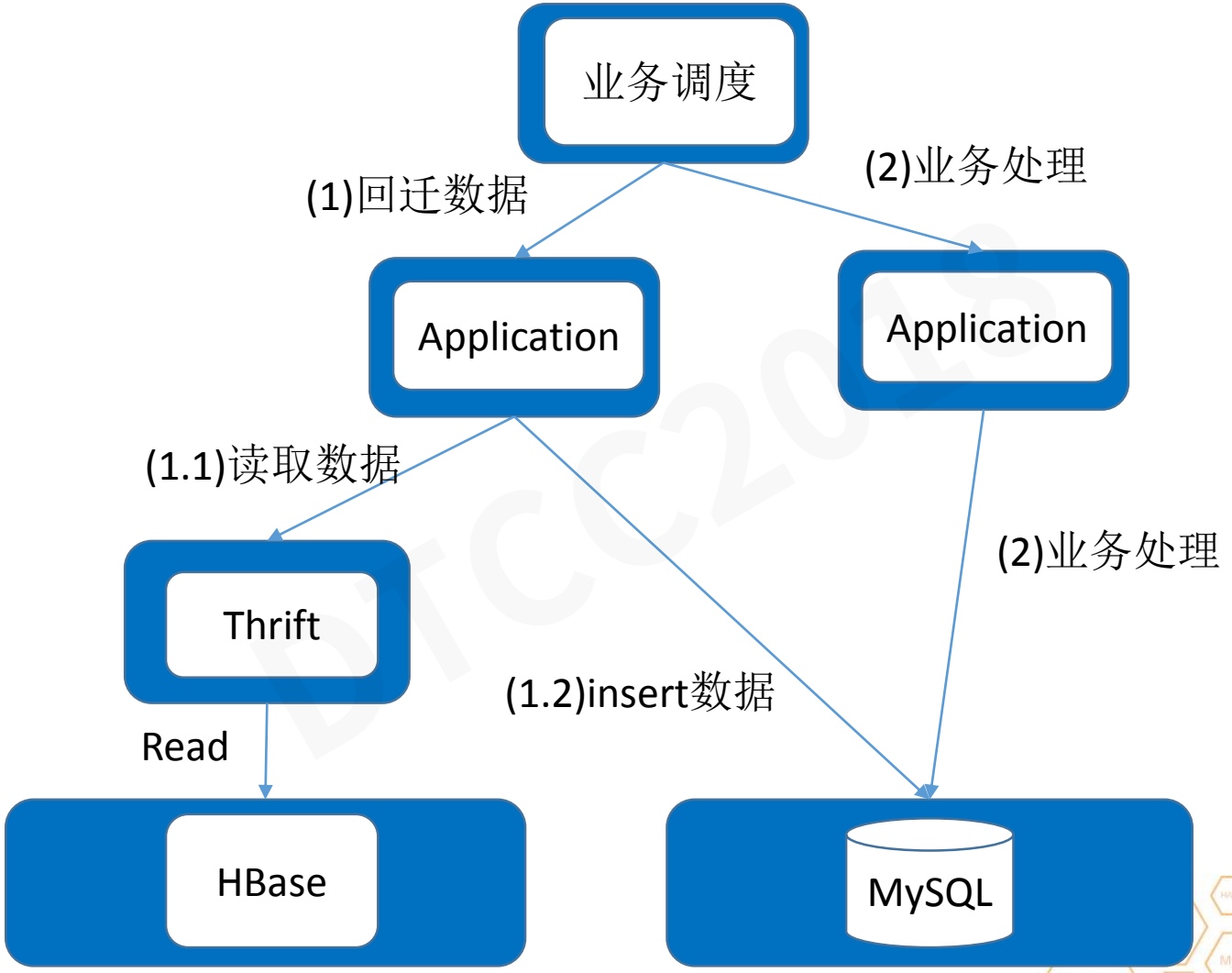
打造高性能、可靠、高可用的HBase数据中心



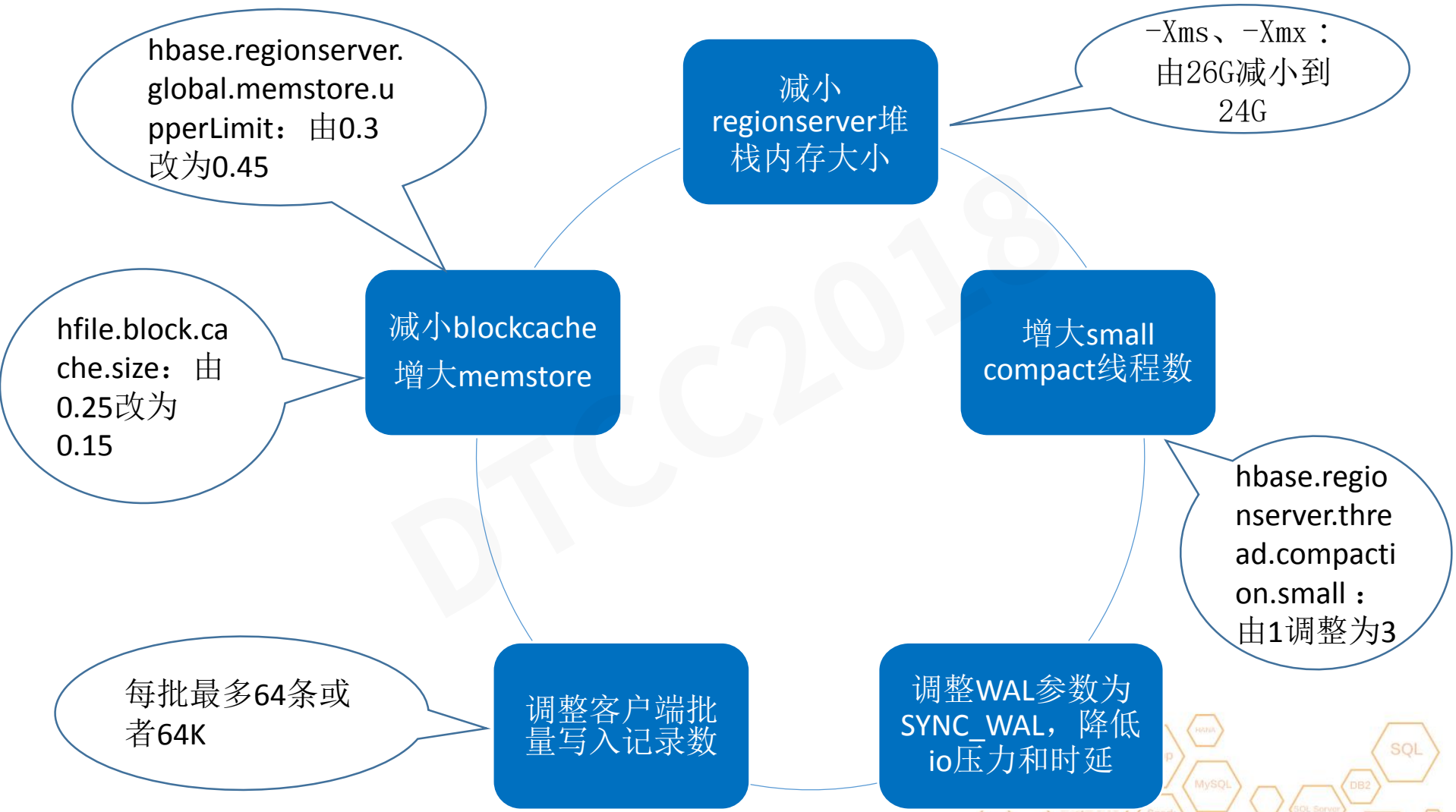
系统架构: (一)数据读写流程



系统架构: (二)业务流程MySQL+HBase



高性能-写优化



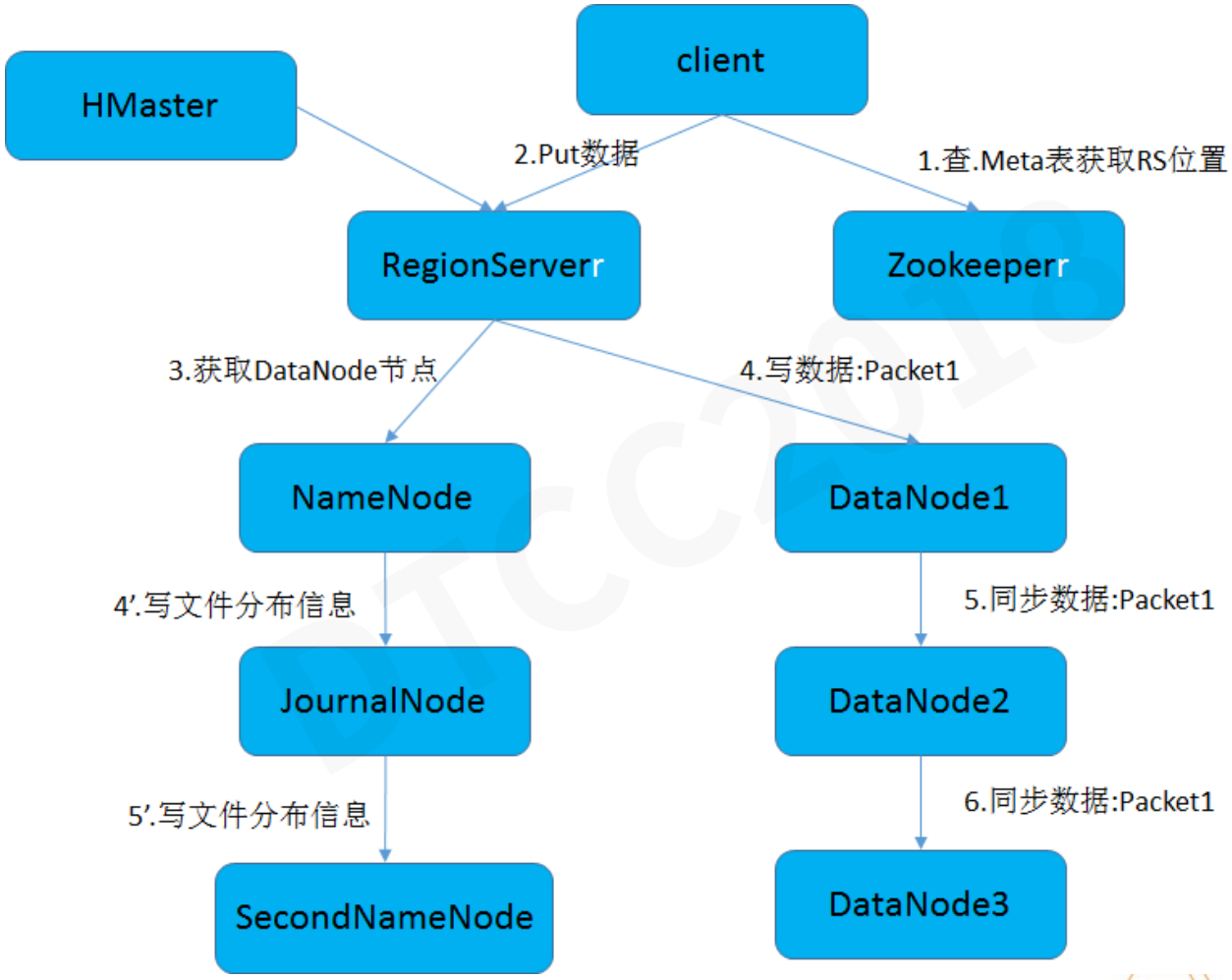
高性能-读优化

1 定期合并	制定合并优化策略，月初针对前一个月表作合并优化，超大表每隔7天进行合并优化
2 重点隔离	重点表单独隔离组，组内机器只允许指定表读写请求，避免表之间相互影响
3 参数优化	调整建表参数，调整Region数量，限制单个region最大文件数不超过3个
4 业务优化	充分了解业务的访问特点，采取不同的建表方式，如索引表按照年或月建表
5 硬件升级	扩容和更换新机器，将重点组机器替换为配置更高的机器

高可用

软件可用	避免Full GC，避免Region Server、单个Region阻塞不可写
主机可用	快速剔除机制，避免单台服务器负载高影响集群
网络可用	建立主、备集群
避免雪崩	缩短Thrift Server超时时间，减少重试次数
减小耦合	按业务重要级别分组

高可用-单点保障


client
HMaster
2.Put数据
1.查.Meta表获取RS位置
RegionServerr
Zookeeperr
3.获取DataNode节点
4.写数据:Packet1
NameNode
DataNode1
4'.写文件分布信息
5.同步数据:Packet1
JournalNode
DataNode2
5'.写文件分布信息
6.同步数据:Packet1
SecondNameNode
DataNode3

可靠性-数据准确

数据对账

- 10分钟增量数据对账
- 冗余数据对账
- 数据写入来源IP与核对IP分开

流程规范

- 制定操作规范，降低人为失误
- 系统平台建设，自动化运维

监控效率

- 秒级监控
- 对账异常上报神盾安全系统

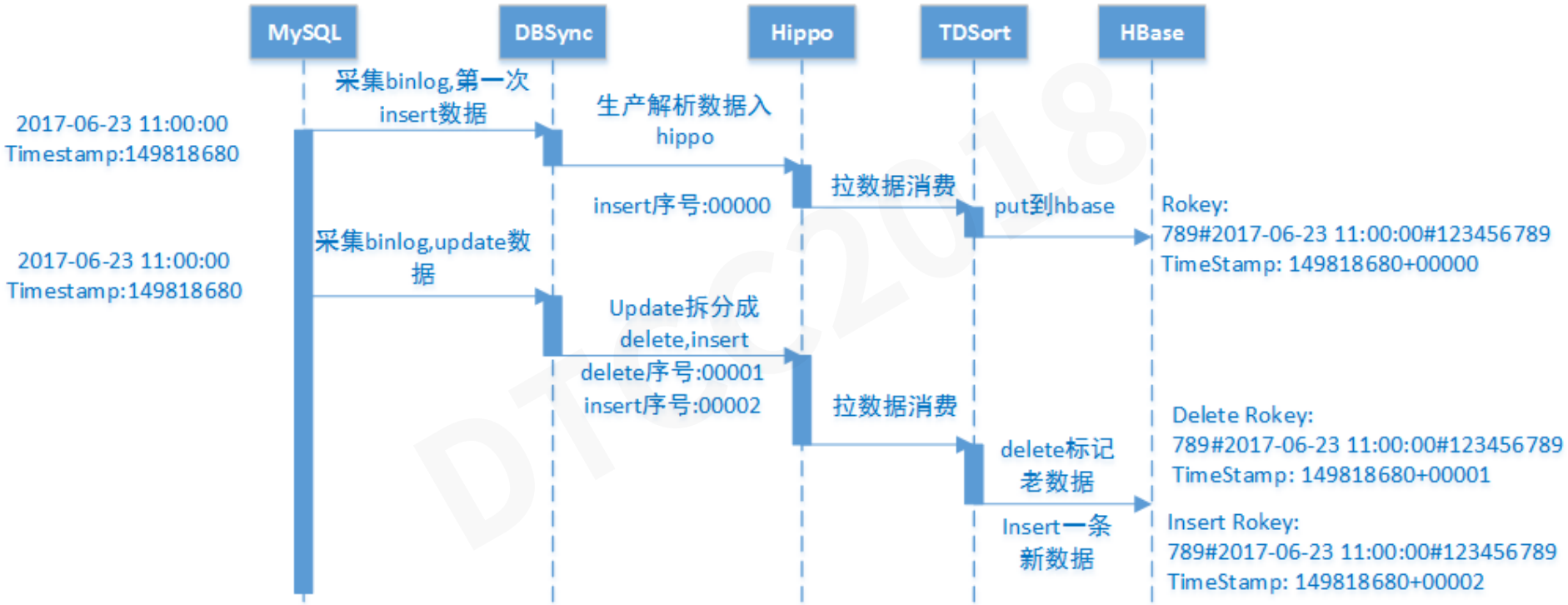
可靠性-update的特殊处理

按时间批量对账： rowkey
包括修改时间

MySQL binlog中update拆分成delete + insert两条消息

HBase保留一条记录

可靠性-update处理流程



实践问题

隔离性

- HBase读写线程队列分开，hbase.ipc.server.callqueue.read.share=0.3
- 某一台DataNode负载高导致整个集群写入量下降70%，快速迁移region解决

可维护性

- 不能动态修改配置
- 不支持命令查看HBase集群运营情况，例如compaction队列明细
- 缺少汇总统计功能

容量均衡

- 迁移数据不要从HBase层操作，可以从hdfs层停datanode节点，让namenode迁移
- 迁移数据没有流量控制

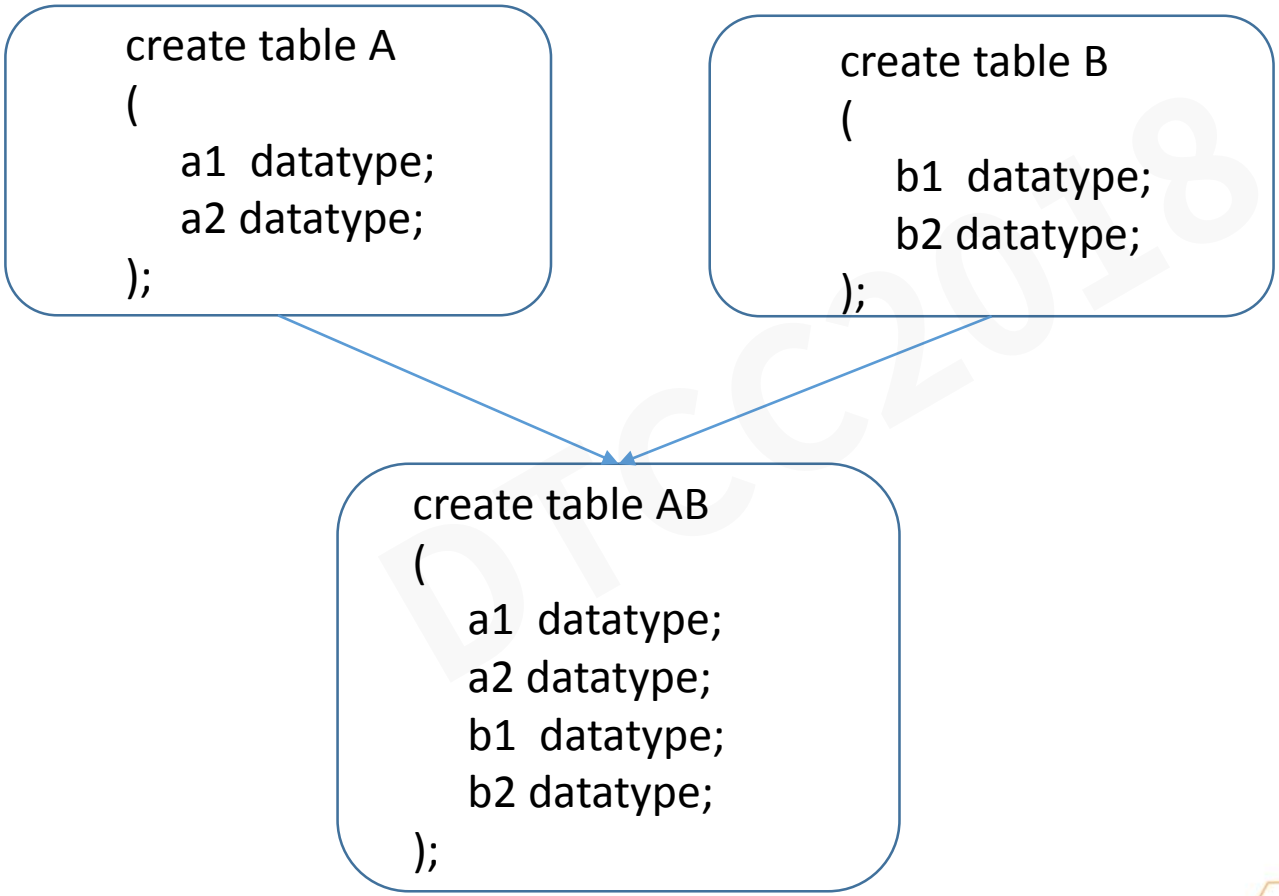
3 HBase挑战和应对	场景局限、二级索引、集群复制

应用场景局限

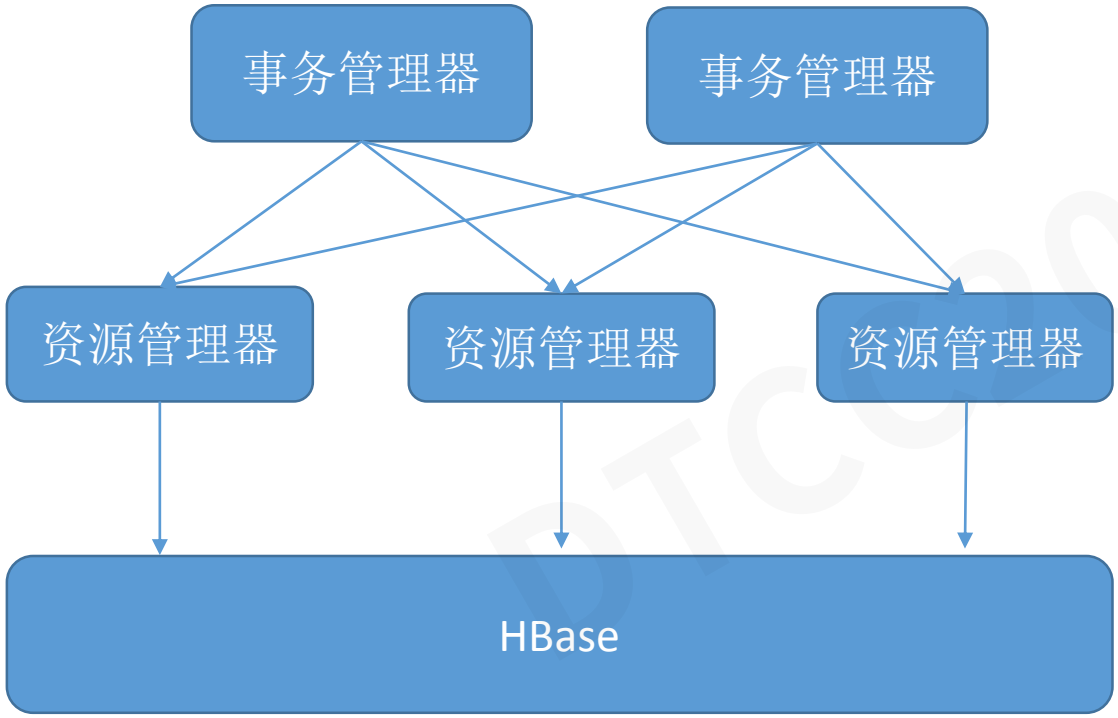
- 不能用在OLTP业务，比如支付业务的核心流程
- 适合存放历史数据，处理历史数据的对账、历史数据的回溯等需求

多表变单表

单表设计+HBase单行事务



BDT-业务分布式事务

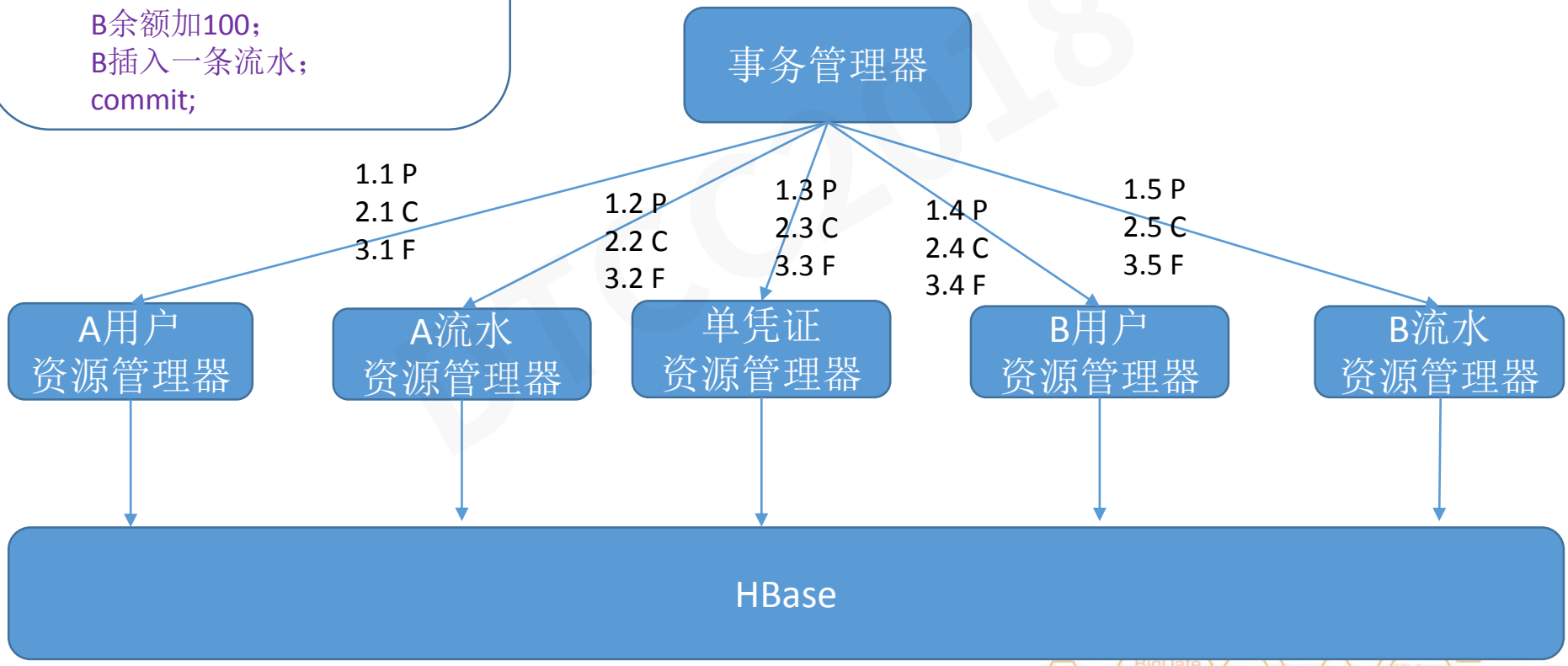


- 采用两阶段提交协议，分成P,C,F阶段
- 事务管理器控制事务的提交或者回滚
- 资源管理器对应表，一张表拆分多个资源管理器，采用有锁方案
- 事务管理器无状态，高可用方案比较成熟
- 资源管理器有状态，可以是M-S方案，但也可以设计成M-M方案，利用HBase的单行事务功能控制并发

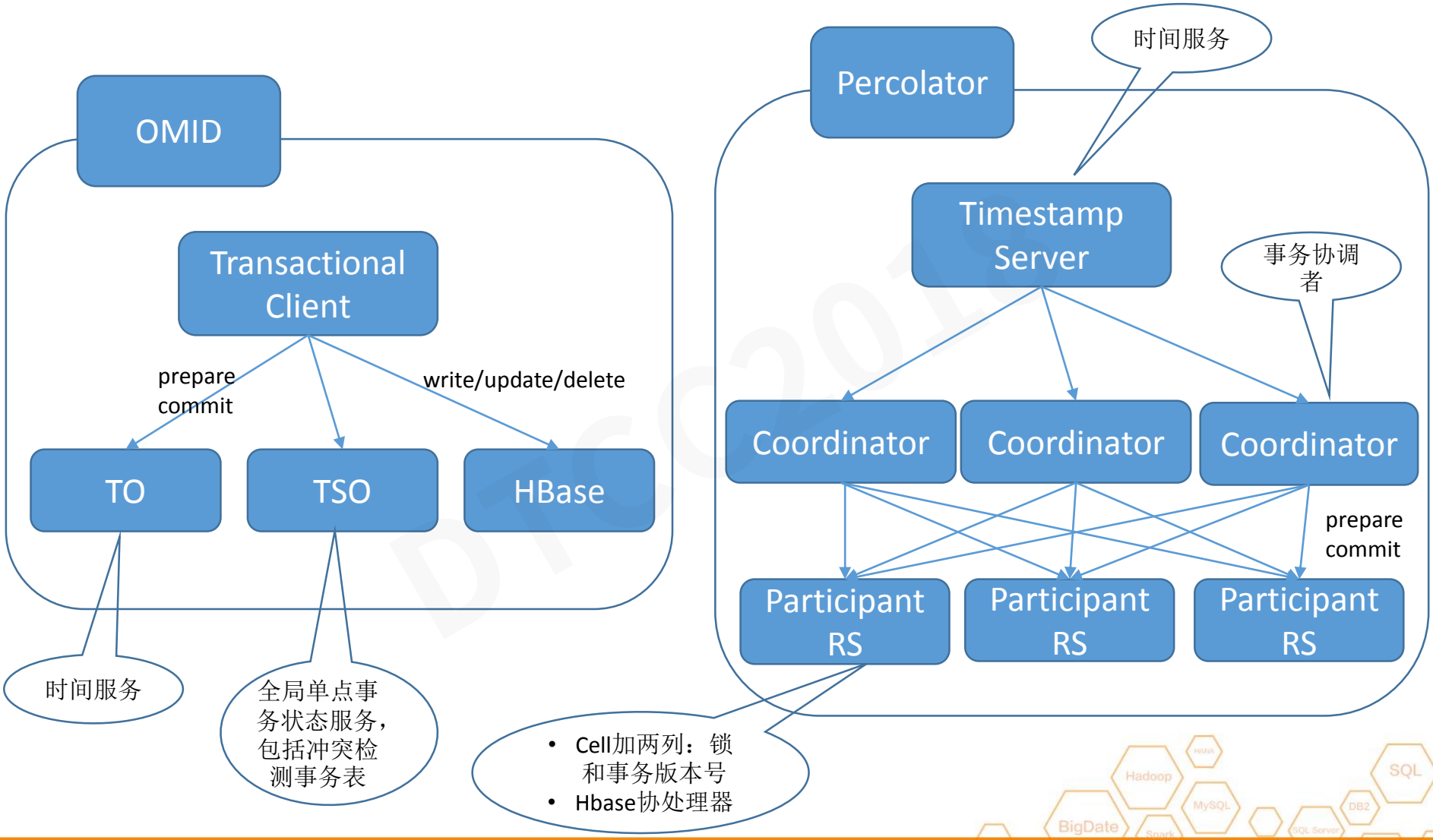
BDT-业务分布式事务

一个典型的支付场景，A转账100元给B，那么事务就是：

begin;
A余额减100;
A插入一条流水;
插入一条单凭证;
B余额加100;
B插入一条流水;
commit;



开源分布式事务框架

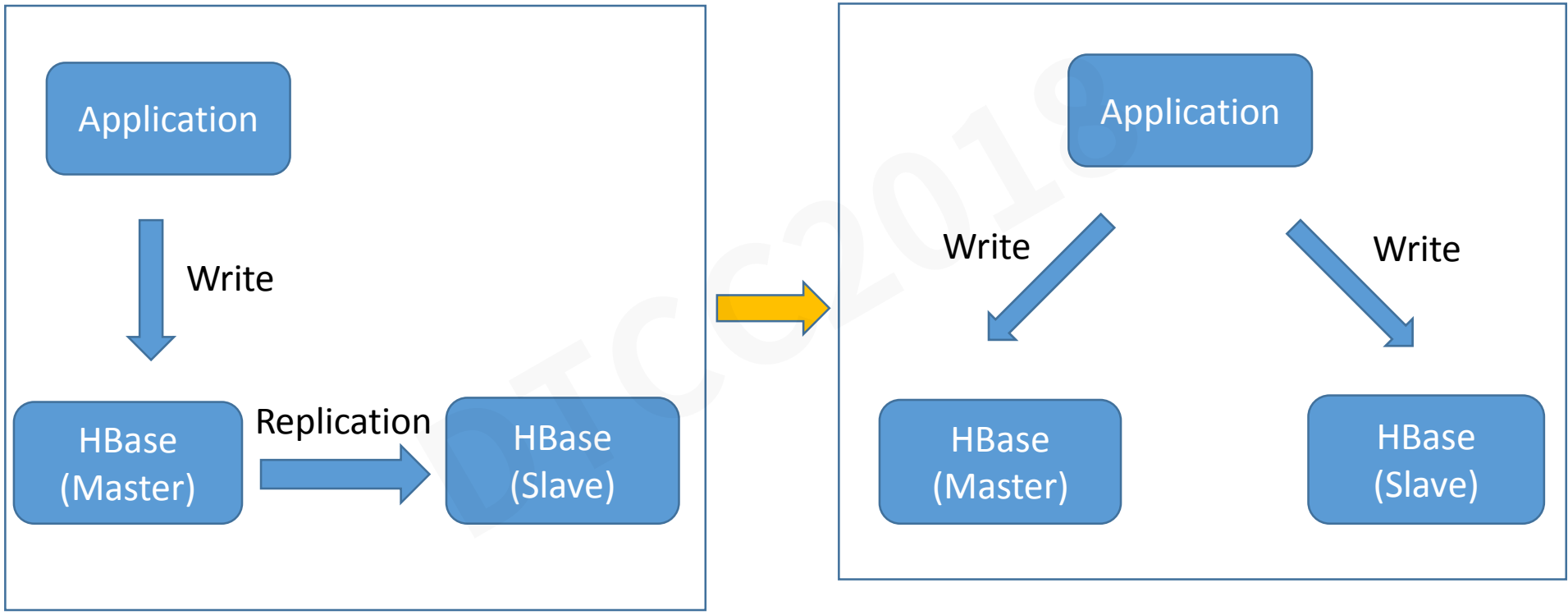

OMID
Transactional Client
prepare
commit
write/update/delete
TO
TSO
HBase
时间服务
全局单点事务状态服务，包括冲突检测事务表
Percolator
时间服务
Timestamp Server
事务协调者
Coordinator
Coordinator
Coordinator
prepare
commit
Participant RS
Participant RS
Participant RS
• Cell加两列：锁和事务版本号
• Hbase协处理器

二级索引

Phoenix	• 支持sql，二级索引，通过Coprocessor实现 • 不稳定，容易造成HBase崩溃
coprocessor实现	• 用Observer实现在写主表前先写索引表
时间库表	• 表设计成按月或者按日分表 • 新建索引时不用copy历史数据

集群容灾

Replication实验性项目，业务双写

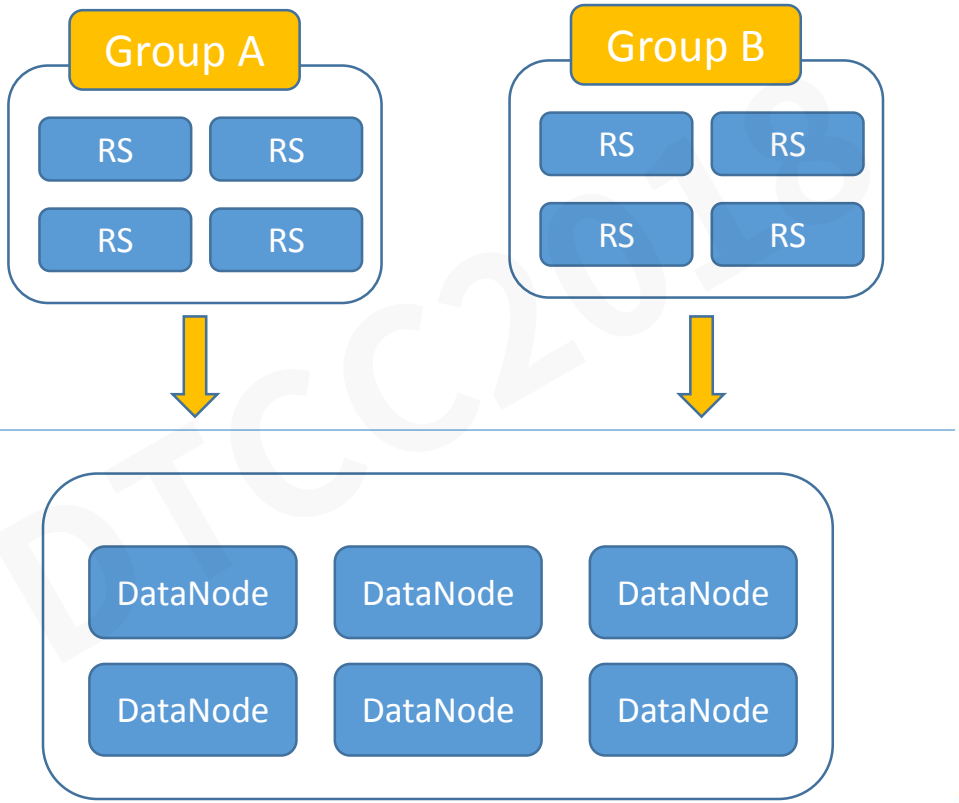


存储隔离

分布式存储不能按组隔离，耦合比较强，定位问题比较复杂

HBase层

HDFS层

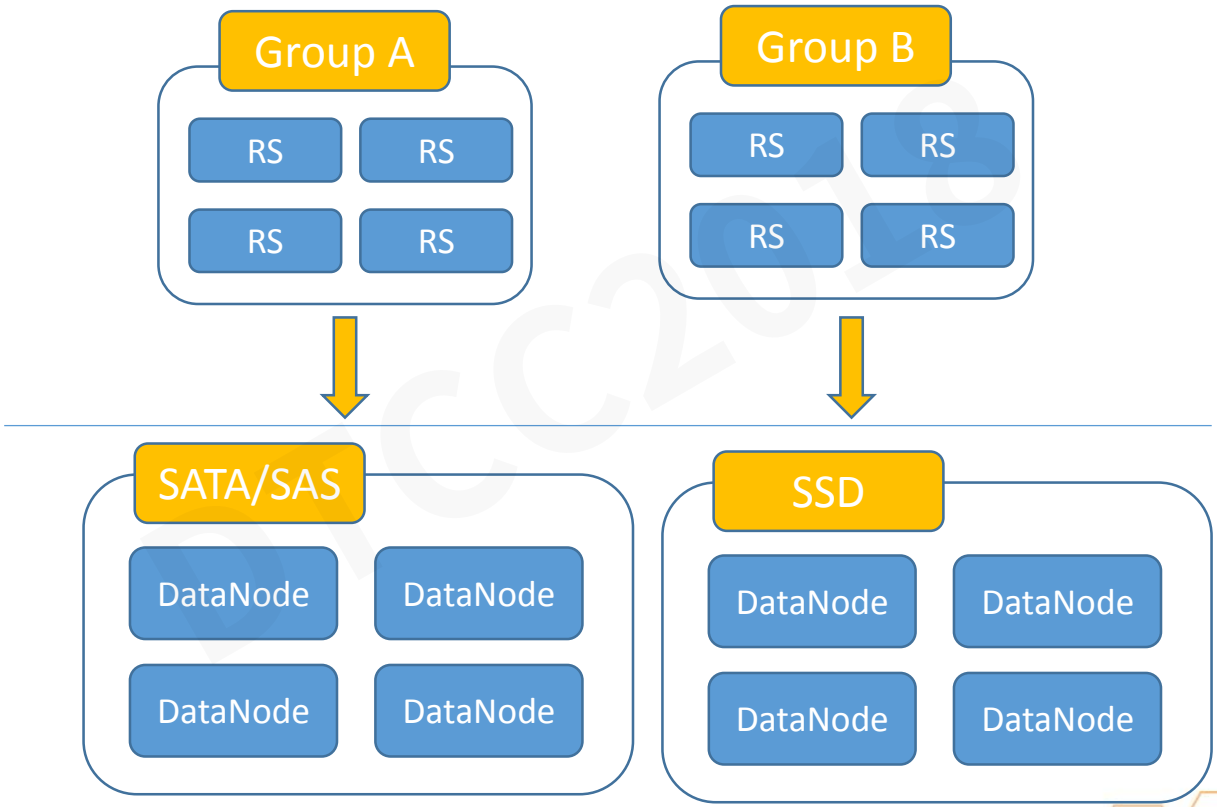


存储隔离

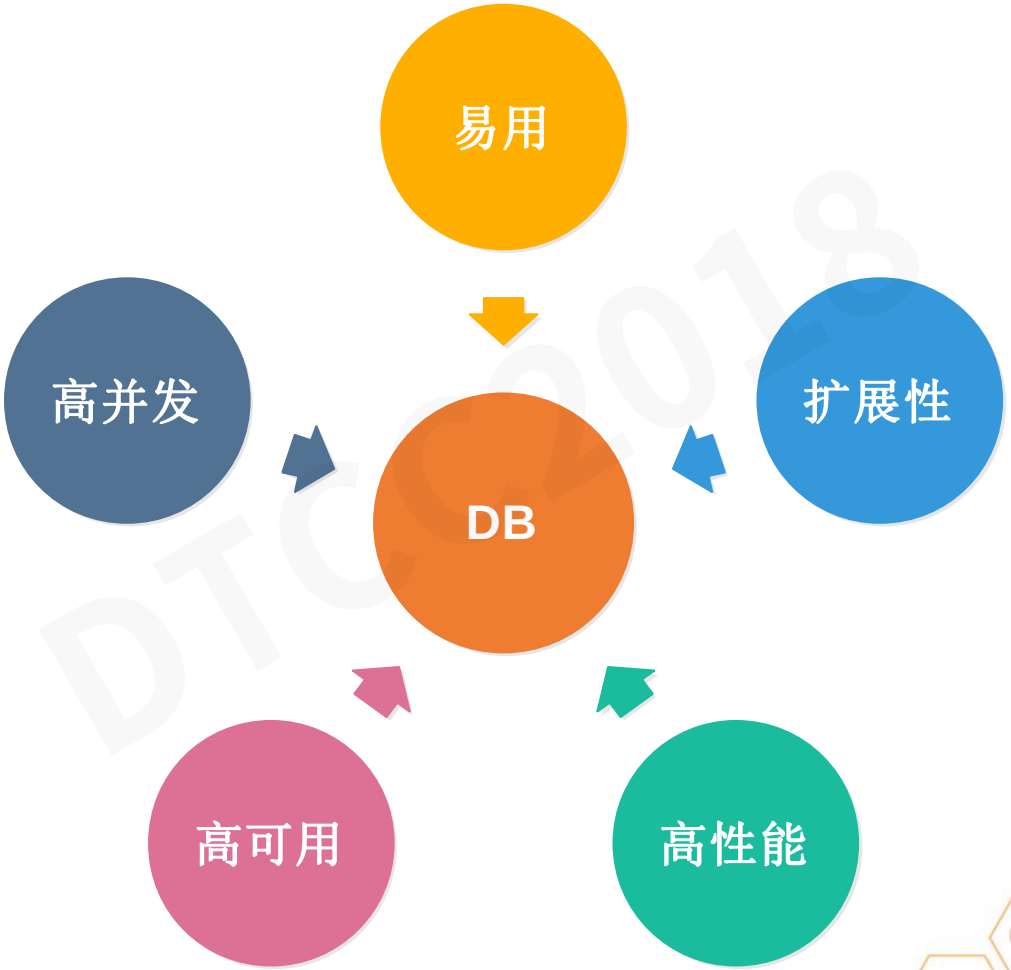
异构存储：SSD与SATA/SAS隔离

HBase层

HDFS层



发展方向



THANKS


HANA
Hadoop
SQL
MySQL
DB2
BigDate
Spark
SQL Server
NewSQL
Oracle
MPP
NoSQL
OLAP

讲师申请
联系电话（微信号）：18612470168

关注“ITPUB”更多
技术干货等你来拿~

与百度外卖、京东、魅族等先后合作系列分享活动

微学堂

让学习更简单

微学堂是以ChinaUnix、ITPUB所组建的微信群为载体，定期邀请嘉宾对热点话题、技术难题、新产品发布等进行移动端的在线直播活动。

截至目前，累计举办活动期数60+，参与人次40000+。

ITPUB学院

ITPUB学院是盛拓传媒IT168企业事业部（ITPUB）旗下
企业级在线学习咨询平台
历经18年技术社区平台发展
汇聚5000万技术用户
紧随企业一线IT技术需求
打造全方式技术培训与技术咨询服务
提供包括企业应用方案培训咨询（包括企业内训）
个人实战技能培训（包括认证培训）
在内的全方位IT技术培训咨询服务

ITPUB学院讲师均来自于企业
一些工程师、架构师、技术经理和CTO
大会演讲专家1800+
社区版主和博客专家500+

培训特色

无限次免费播放
随时随地在线观看
碎片化时间集中学习
聚焦知识点详细解读
讲师在线答疑
强大的技术人脉圈

八大课程体系

基础架构设计与建设
大数据平台
应用架构设计与开发
系统运维与数据库
传统企业数字化转型
人工智能
区块链
移动开发与SEO

联系我们

联系人：黄老师
电　话：010-59127187
邮　箱：edu@itpub.net
网　址：edu.itpub.net
培训微信号：18500940168