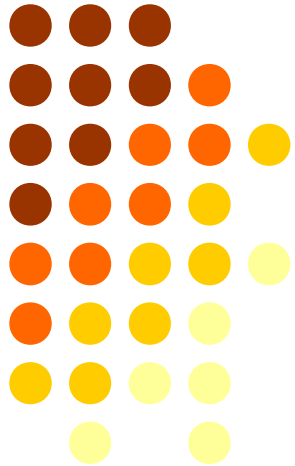
淘宝数据库架构演进历程



丹臣 / 赵林
数据架构师
2010-11-19

提纲

- 淘宝数据库发展的三个阶段
- 用户，商品，交易现在的架构
- 2010 双 11 大促的挑战
- MySQL 源代码研究的一些思路
- 淘宝自主数据库 Oceanbase 原理介绍

淘宝的数据很美丽

淘宝数据库发展三阶段

淘宝网 Taobao.com

第一阶段

- 整个网站采用LAMP架构
- 数据库采用几台MySQL
- 应用系统分为前台，后台两大系统

第二阶段

- MySQL迁到Oracle
- Pc server升级到IBM小型机
- 低端存储升级到高端存储

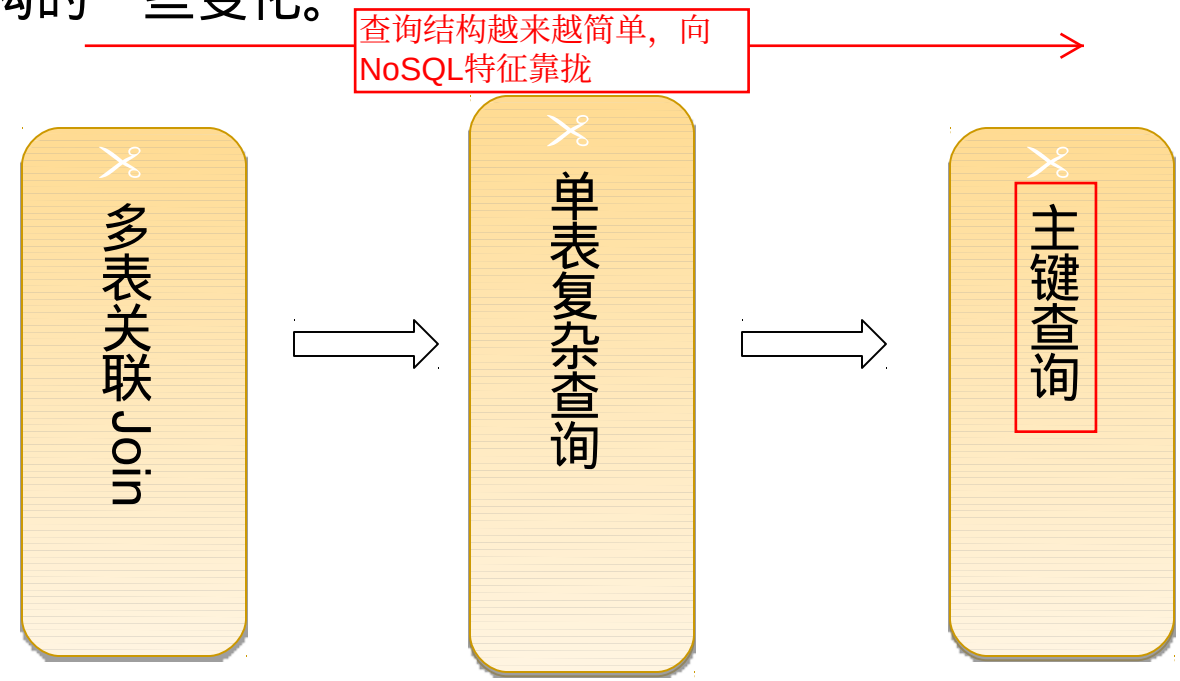
第三阶段

- 核心业务从Oracle逐步迁到分布式MySQL集群中
- 大量采用pc server,采用本地硬盘

SQL 语句变化

淘宝网 Taobao.com

SQL 语句复杂程度由繁到简的过程，折射出淘宝数据架构的一些变化。



淘宝电子商务网站的特点

- 高并发， PV13 亿，光棍节促销 PV 达到了 17 亿
- 数据实时性要求高
- 数据准确性要求高
- 大多数页面属于动态网页
- 网站需要大量商品图片展示
- 用户通过搜索引擎，广告，类目导航寻找商品
- 网站读多写少，比例超过 10:1
- 卖家相关的数据量较大，比如商品数，评价数
- 业务量快速增长

不同的时期，不同的策略

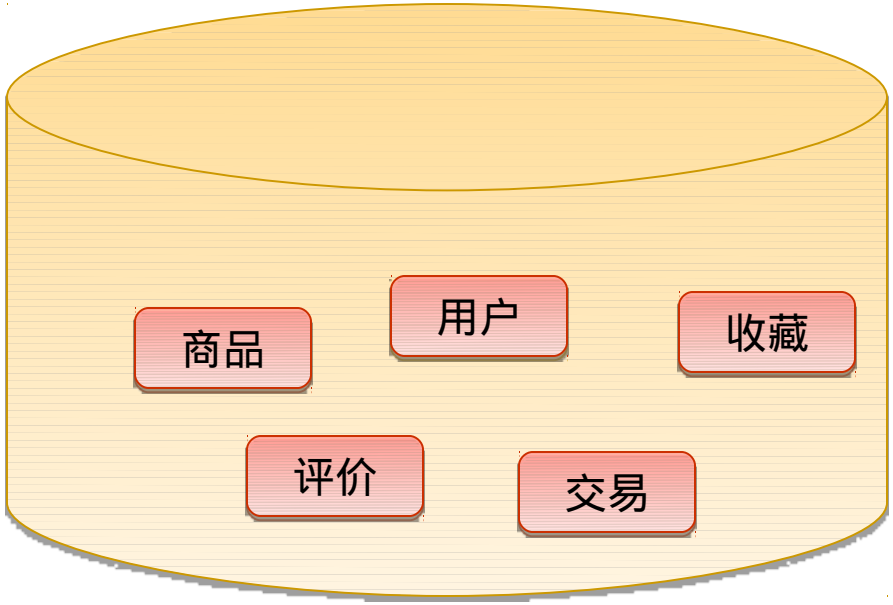
正是因为如上的业务特点：

- 早期的淘宝前端应用系统，严重依赖于数据库系统
- 早期单机式的 mysql 的使用方式，在业务的高速发展下，很快达到瓶颈
- Mysql 迁移到 Oracle ，并升级到小型机，高端存储后，几年的时间里，满足了淘宝业务快速变化发展的需要。
- 我们的业务发展很快，但我们的技术没有成长

数据库里的数据

淘宝网 Taobao.com

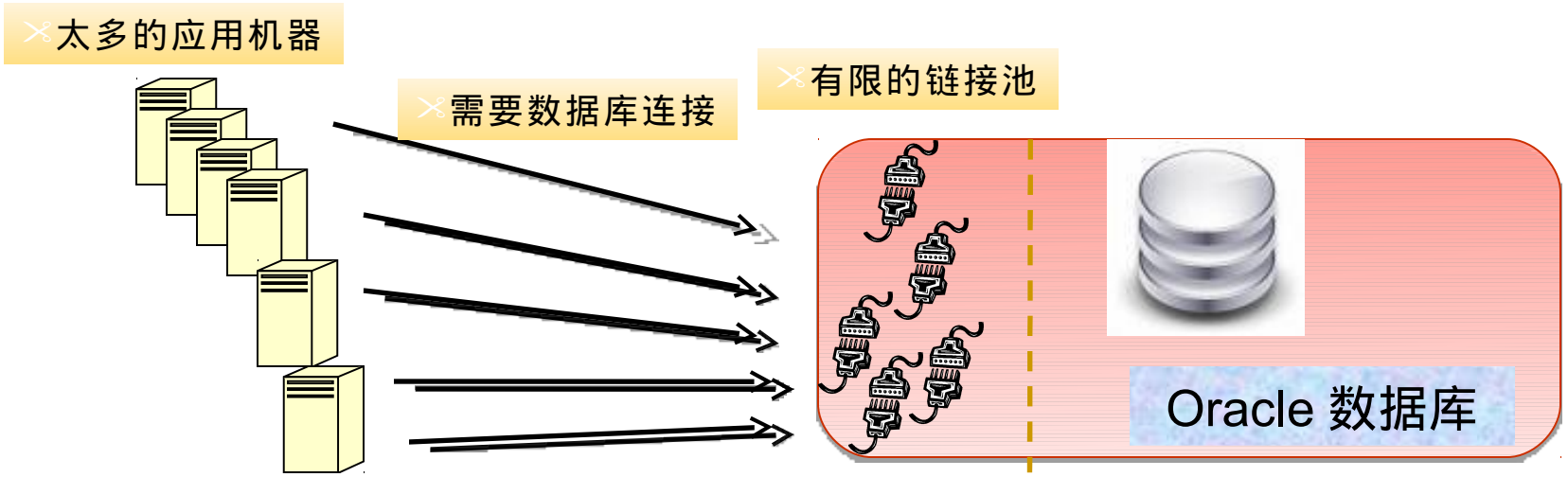
第一，二阶段的单台数据库里，用户，商品，交易等数据都在一起，存在许多的关联查询，应用完全耦合



连接数问题

淘宝网 Taobao.com

小型机的内存有限，发现了 Oracle 数据库有连接数瓶颈， 5000 个以后相当吃力。



中心化，服务化

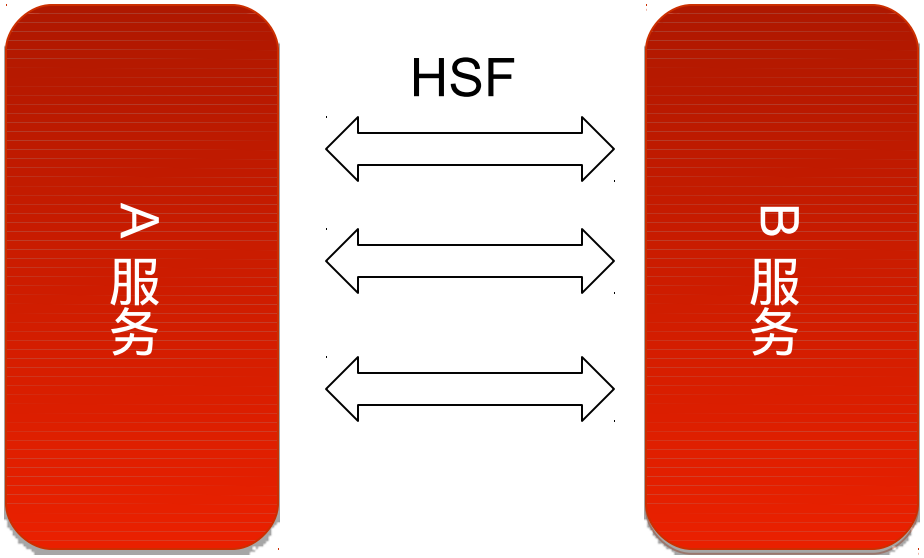
淘宝网 Taobao.com

- 用户，商品，交易三大中心的建设

HSF 的诞生

淘宝网 Taobao.com

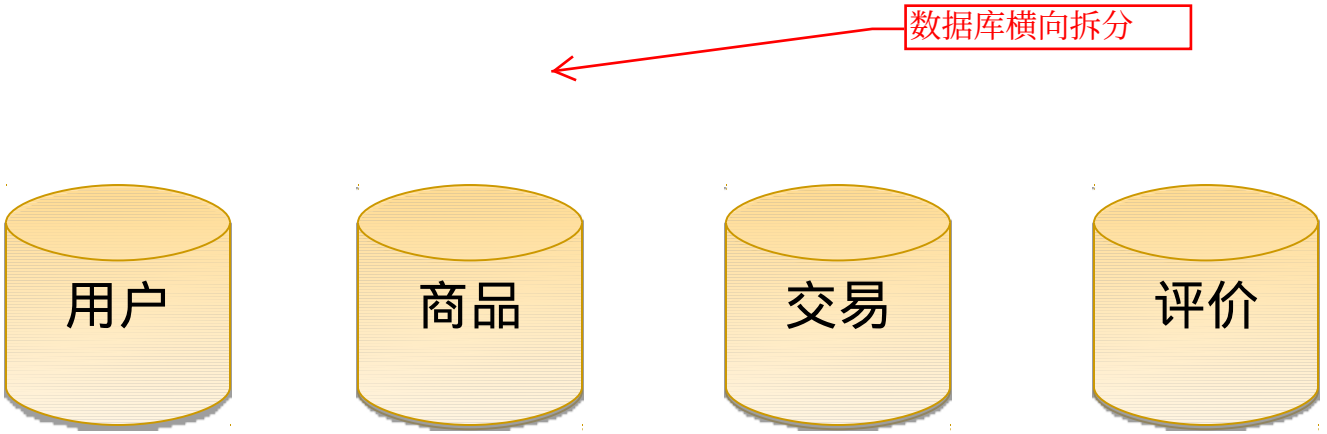
- 中心化后面临另一个问题，服务调用者，与服务者之间如何进行远程通信，淘宝 HSF 诞生



数据垂直化

淘宝网 Taobao.com

- 应用中心化之后，底层数据库系统按照不同的业务数据进行了一系列的垂直拆分 . 此类拆分方式具有如下的特点：
  a. 拆分方式简单，只需要把不同的业务数据进行分离
  b. 避免了不同的业务数据读写操作时的相互影响
  c. 该业务内部及其所导致的问题依旧



问题

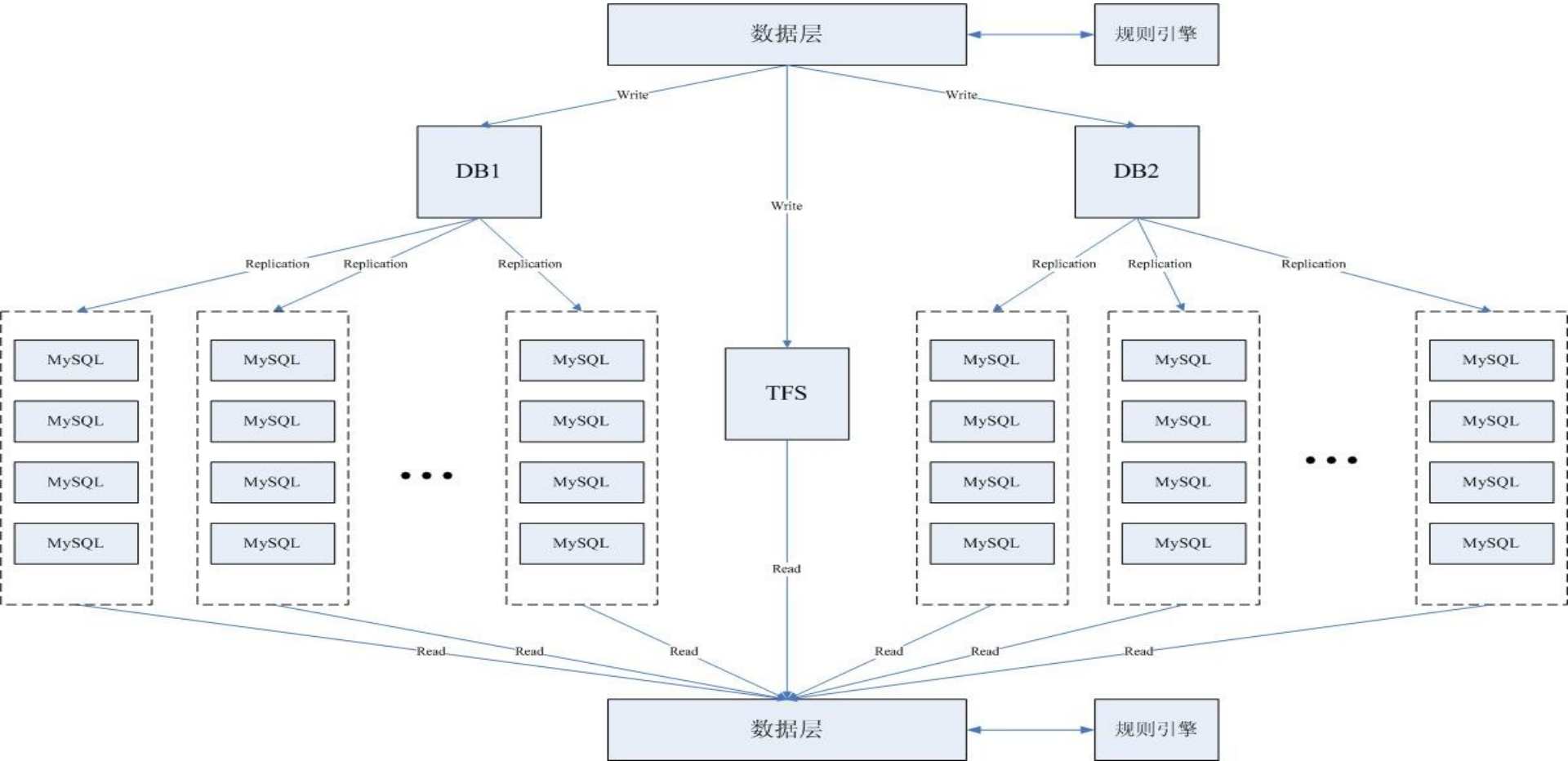
淘宝网 Taobao.com

- 单库 IOPS 3w
- 单库连接数已经 4k 个了，应用还在不断加机器？
- 单库每秒 SQL 执行次数到 4w 次
- Oracle 单库事务数 2k 个
- 搜索 dump 数据缓慢， DW ETL 缓慢

数据库架构发展新思路

淘宝网 Taobao.com

异构数据库读写分离原始架构图（08年8月份）：



异构的读写分离

a. 写库为集中式的 oracle 环境，提供数据安全性保障

b. 读库使用 mysql，采用数据分片，分库分表，每台 mysql 放少量的数据，单个数据分片内部采用 mysql 复制机制

c. 读库的超大 memory 容量，起到了很好的 cache 作用，在内存中的数据查询性能远远高于在硬盘上的性能

d. oracle 到多台 mysql 按规则复制

e. 分区键的选择至关重要，尽量让数据访问落在单台数据库上

g. 利用好当前的高端硬件，保护好自己的投资

构建数据查询的高速公路

淘宝网 Taobao.com

- 应用到 DB 的数据写入与查询从双向通行变成了单向通行，通行效率更高，大大避免了相互影响。“借道行驶”的情况不再出现。

跨不过去的坎

淘宝网 Taobao.com

为什么不直接迁到 MySQL 上面去呢？

a. 对于核心业务，停机时间有限，宠大的数据无法短时间内迁移

b. 无法在短时间内完成项目发布过程中的测试

c. 没有搞过 mysql 分布式系统，对完全使用 MySQL 还没有信心

大数据量核心业务数据迁移思路

采用两步走战略，不仅走得稳，而且走得好：

- 先采用异构的数据库读写分离，将数据复制到目标 mysql 各结点，不断切换应用相关的读服务到 mysql 结点上，验证可靠性，机器压力，服务响应时间

- 将写压力从 oracle 结点迁移到 mysql 各结点， oracle 停止写

对于一些不太核心，业务不太复杂，相关影响点不多的数据，可以直接进行迁移。

水库模型

你的系统可以撑多少？系统余量还有多少？

数据库系统余量

淘宝网 Taobao.com

两轮测试过程，确保上线稳定：

- 底层数据库环境性能，稳定性的基础测试，常用的工具可以采用 sysbench, orion, supersmack

- 选择不同的硬件，软件组合，模拟应用的压力测试，要超越当前业务压力的几倍进行，这个压力的幅度可以根据自己的业务增长设计一个合理的值。

我们如何做到用数据来说话？靠测试拿数据，不靠经验

用数据来说话

现有数据库压力，还能支撑多久，什么时候开始扩容

DB性能情况	2010-11-19数据	光棍节大促数据	扩容点数据信息
iostat-await	1.53~2.12	2.33~2.91	2.8~3.78
iostat-Uitl	56%~63%	63%~77.70%	93%~98%
IOstat-svctm	0.87~1.36	1.0~1.65	2.76
cpu-iowait	4~10%	9%~14.4%wa	16.5%wa
cpu-sys	0.4%sy	1.1%sy	3.7%sy
cpu-user	2.1%us	1.7%us~2.5%us	9%us
load	2	2.4	4.23
com_select	1000.60/S	830~1200	3600~3700/S
com_insert	62/S	28~50	150~160/S
com_update	110.40/S	50~100	324~336/S
com_delete	35/S	4~11	42~46/S
BPR 物理/逻辑/命中率	1116 / 57440 /98.15%	1293 / 53380 97.56%	1847 / 195519/99.06%
BPW 物理/逻辑/命中率	215 / 6665/ 94.63%	534 / 5830 /90.84%	1401 / 36550/96.17%
备注		光棍节大促数据	基于线上日常压力模拟(90%相似度)的4倍压力

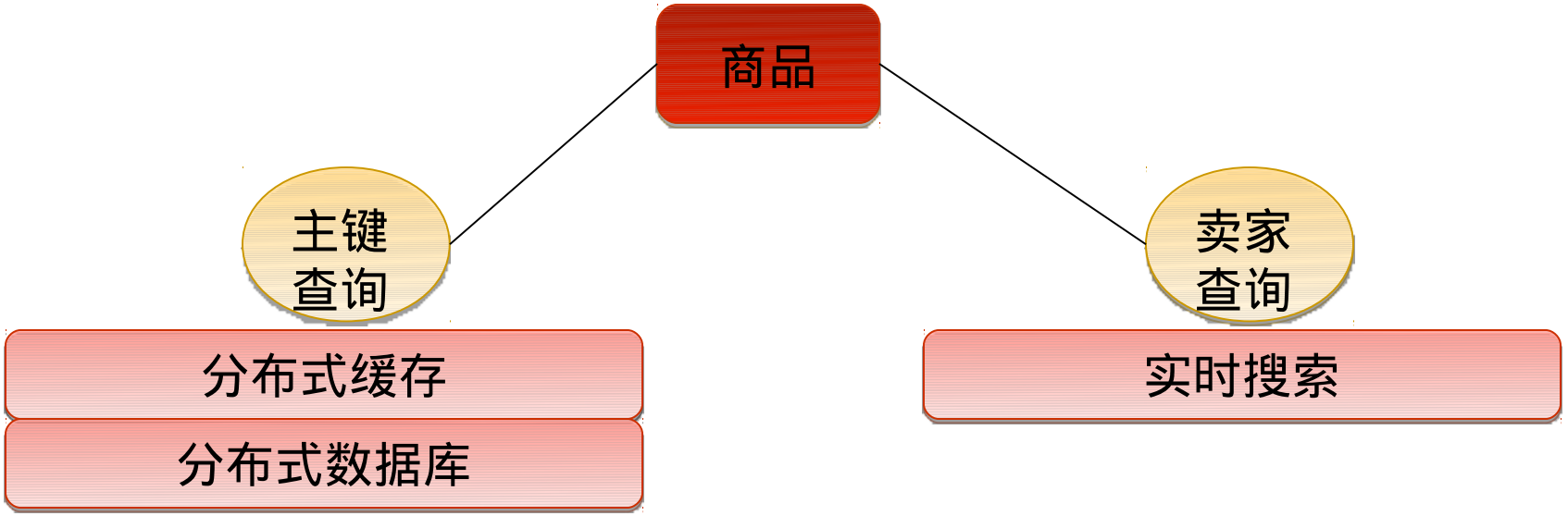
不同的读服务，不同的读载体

淘宝网 Taobao.com

淘宝商品的几个主要的查询：

a. 主键查询通过分布式数据库，以及分布式缓存系统解决

b. 卖家商品管理类查询，这一类的查询数据量大，并且还有 like 查询的需求，通过实时搜索解决

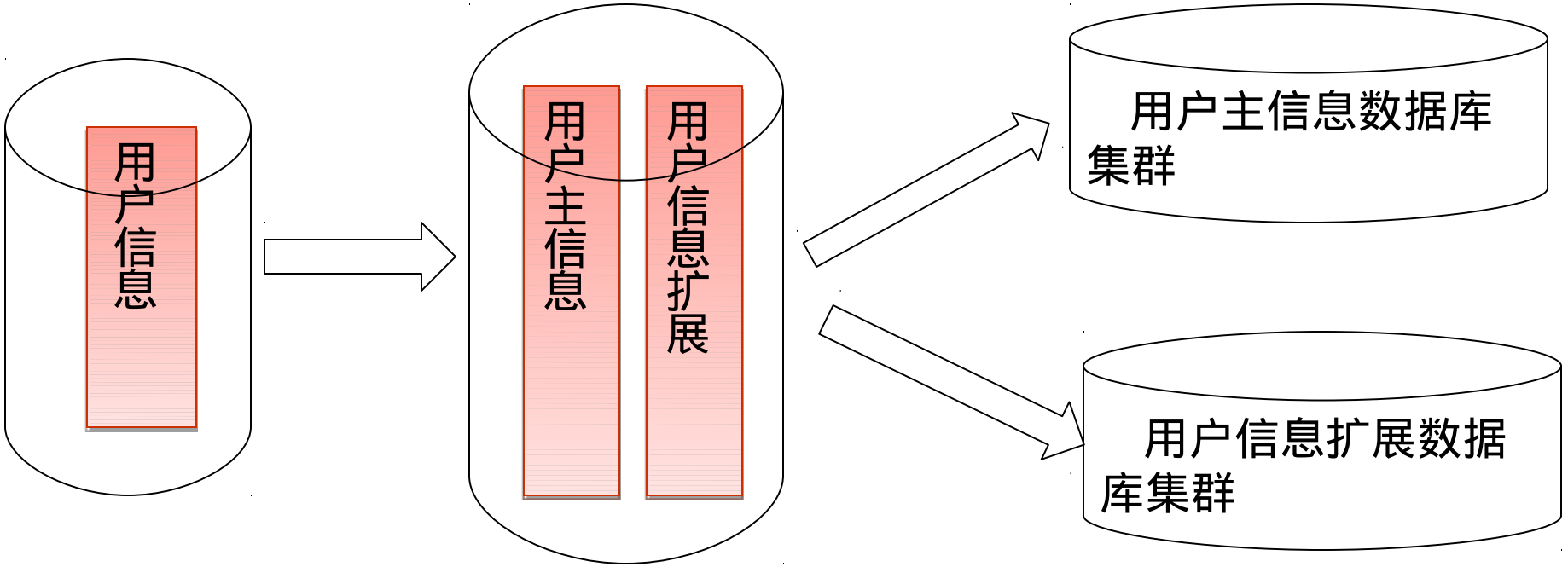


注：考虑不同的读载体的技术实现，性能，成本

数据修改频率

淘宝网 Taobao.com

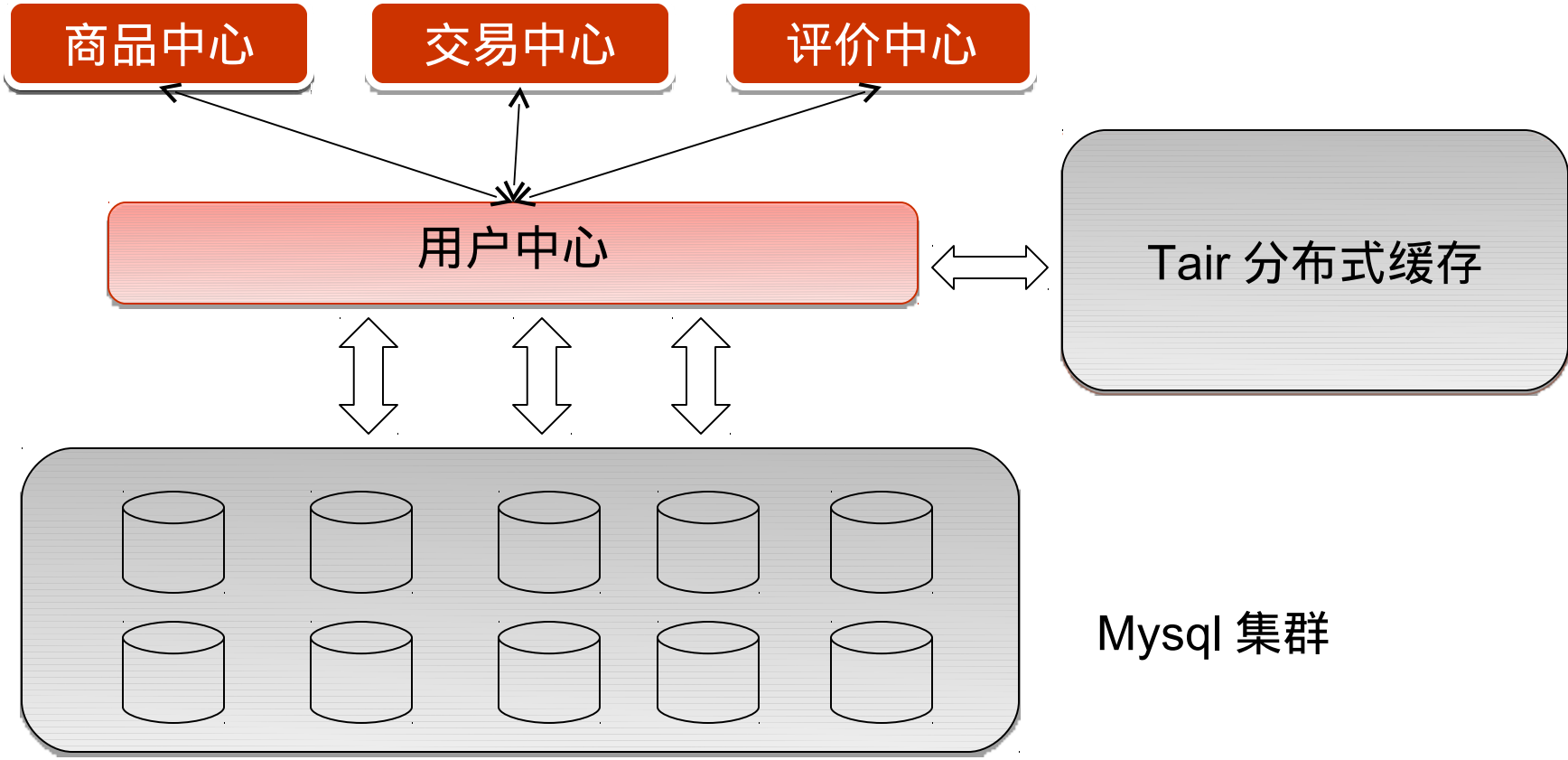
- 用户登陆事件数据 (日志量 90%) 与用户主数据 (日志量 10%) 分离，不仅要分表，而且要放到不同的数据库集群中 , 并且作好不同数据等级的容灾处理。



过度中心化

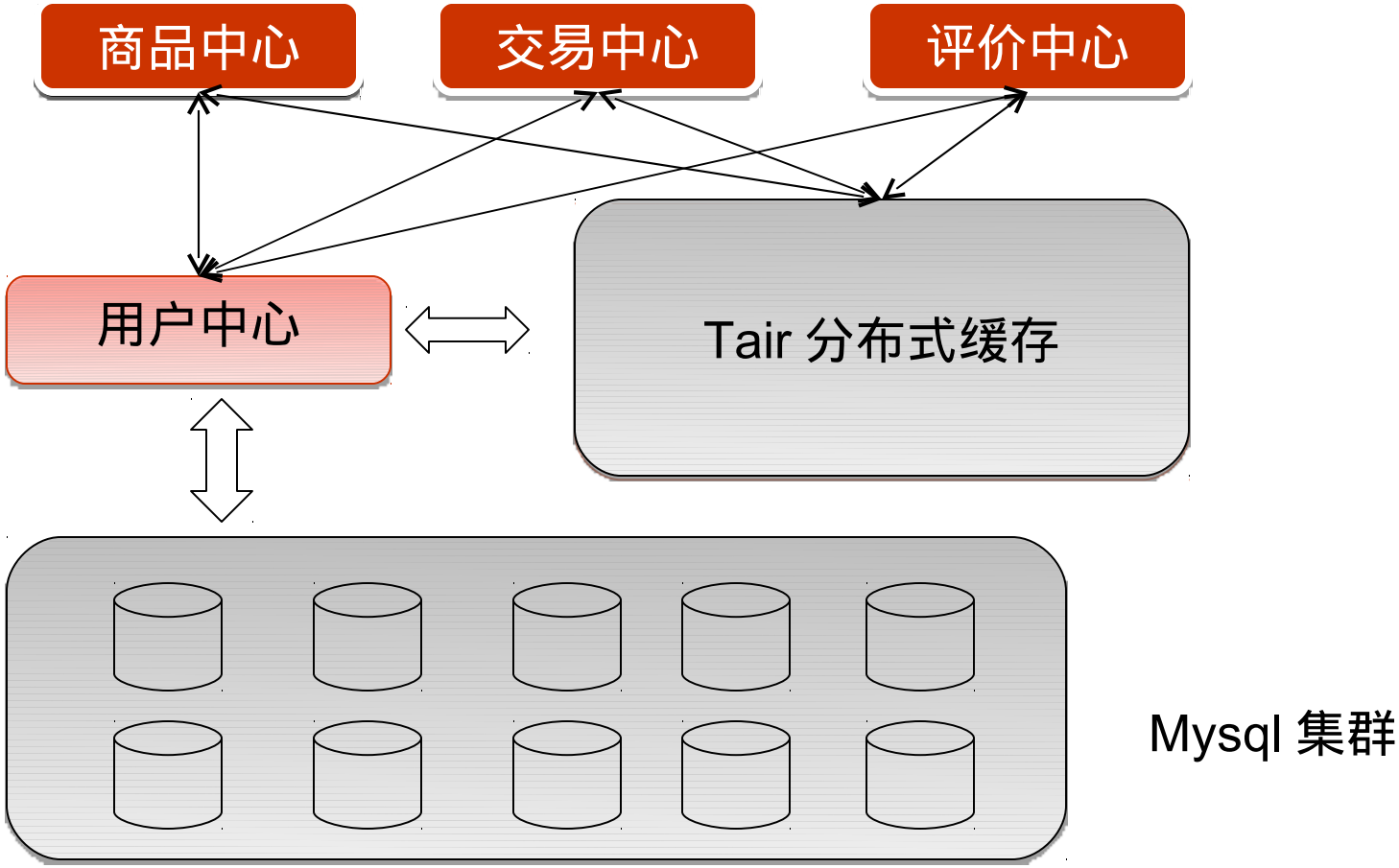
淘宝网 Taobao.com

用户中心调用次数，高峰时期达到了每天 60 亿次，用户中心的过度中心化问题越来越显著，成为各种操作的关键路径。



用户中心中的读写分离

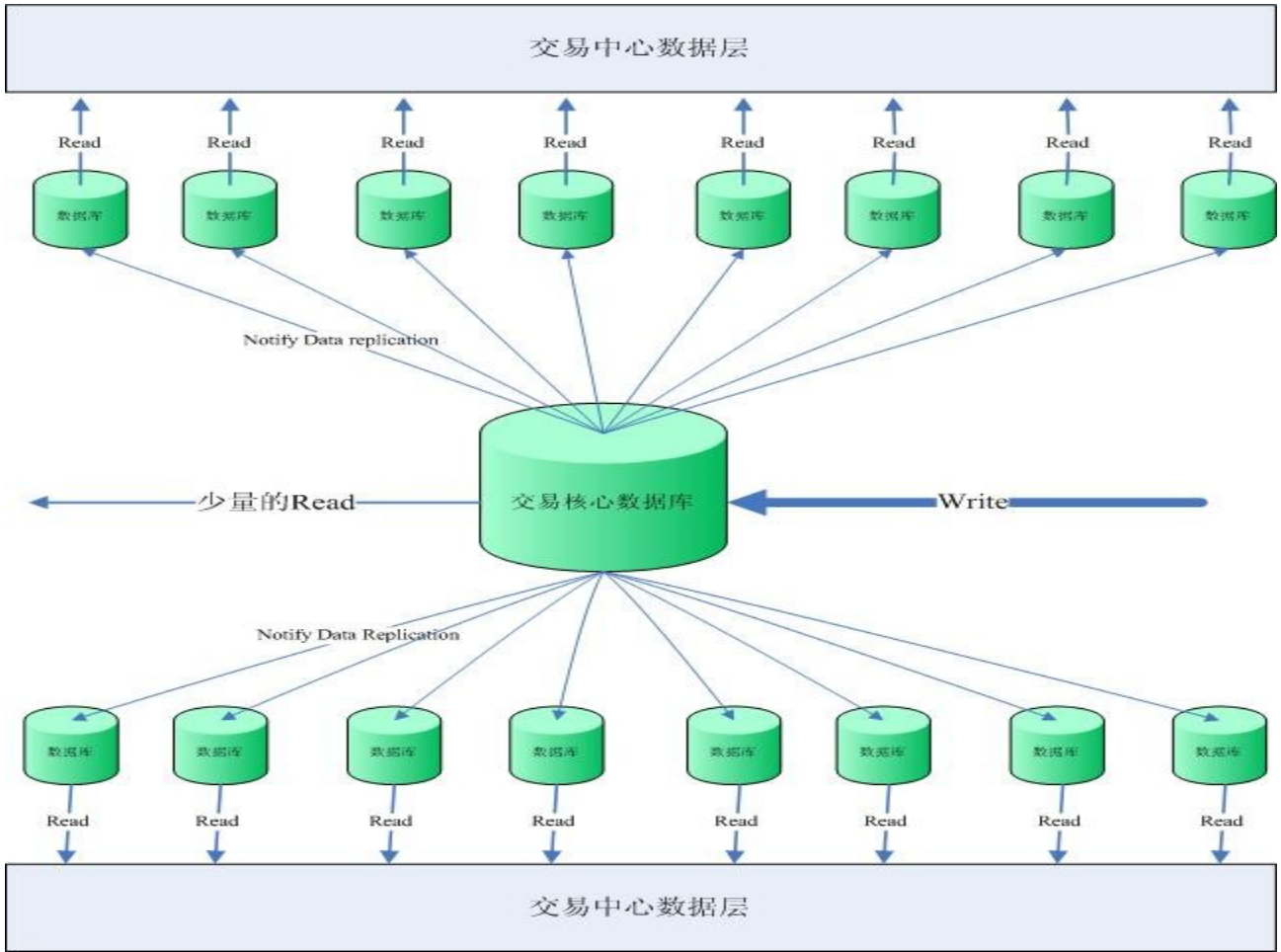
在其它中心中内置可以访问 tair 的客户端，大部份的读不需要经过用户中心，直接读 tair ，写需要经过用户中心。



交易的读写分离框架

淘宝网 Taobao.com

- 主库按照买家拆分，读库按照卖家拆分。



一些难题

- 数据库集群自动扩展仍然是个难题，但是是可以忍受的，底层数据库集群经过评估，扩展的频率并不高。
- MySQL DDL 操作不便，锁表，对写操作影响较大，为了减少影响，分了比较多的表，进一步加重了维护的负担。
- 其它。。。

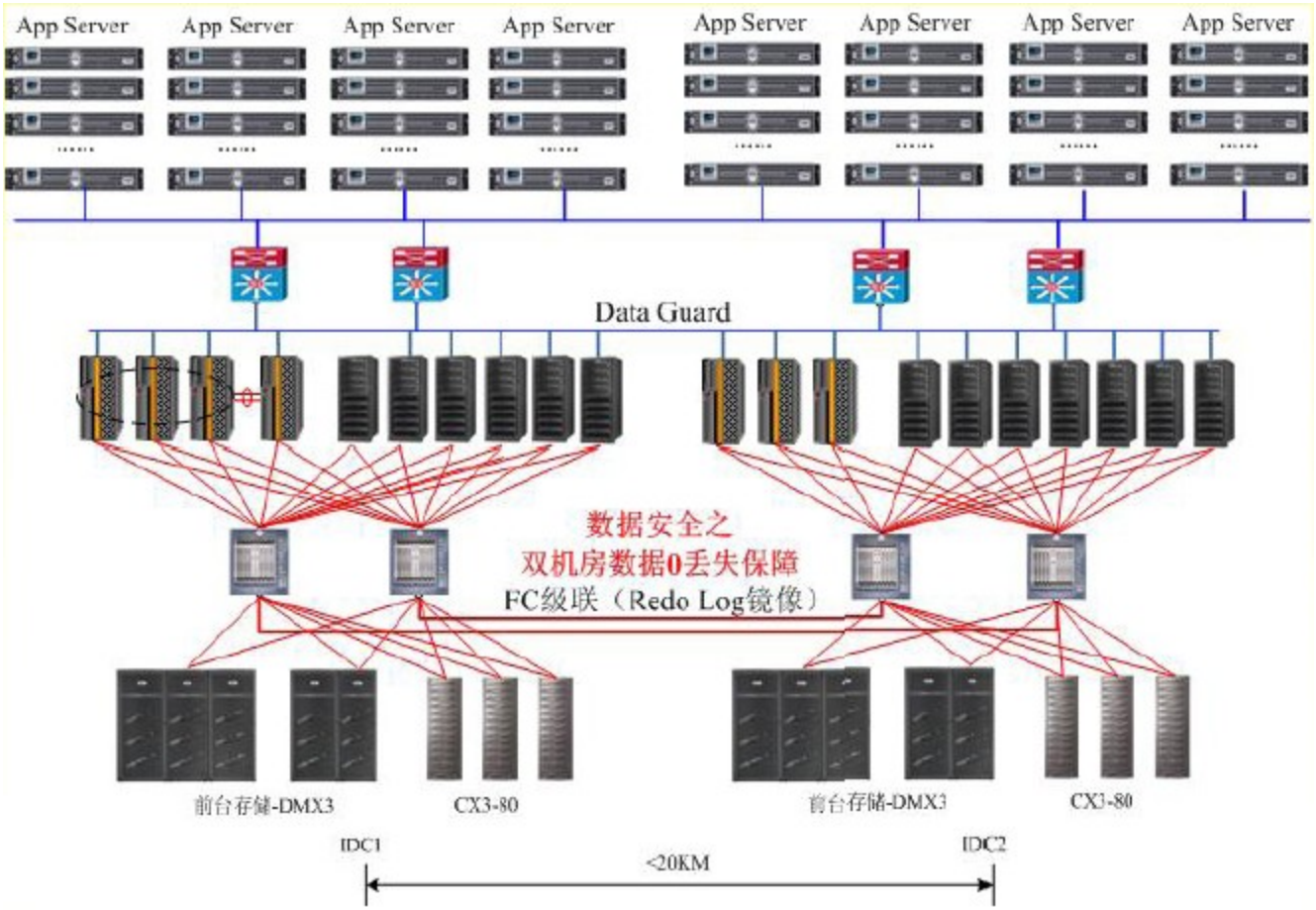
光棍节大促

淘宝网 Taobao.com

活动前，经过了充分的准备与系统评估工作： CDN 面临的压力最大，预估流量将会达到 280G 左右，准备了各个层面的系统降级方案。

一个小意外

- Dataguard+mirror redo 对写的影响比较大，临时删除远程的 redo member 解决这个问题



MySQL 源代码研究

淘宝网 Taobao.com

我们主要从两方面着手：

- MySQL 内部，源代码熟悉，性能优化，新增功能
- MySQL 外部，比如利用 binlog 做数据复制

MySQL 源代码研究

淘宝网 Taobao.com

内部新增的一些功能：

a. 给 innodb 动态加数据文件

b. 禁止新连接

c. 表的访问统计

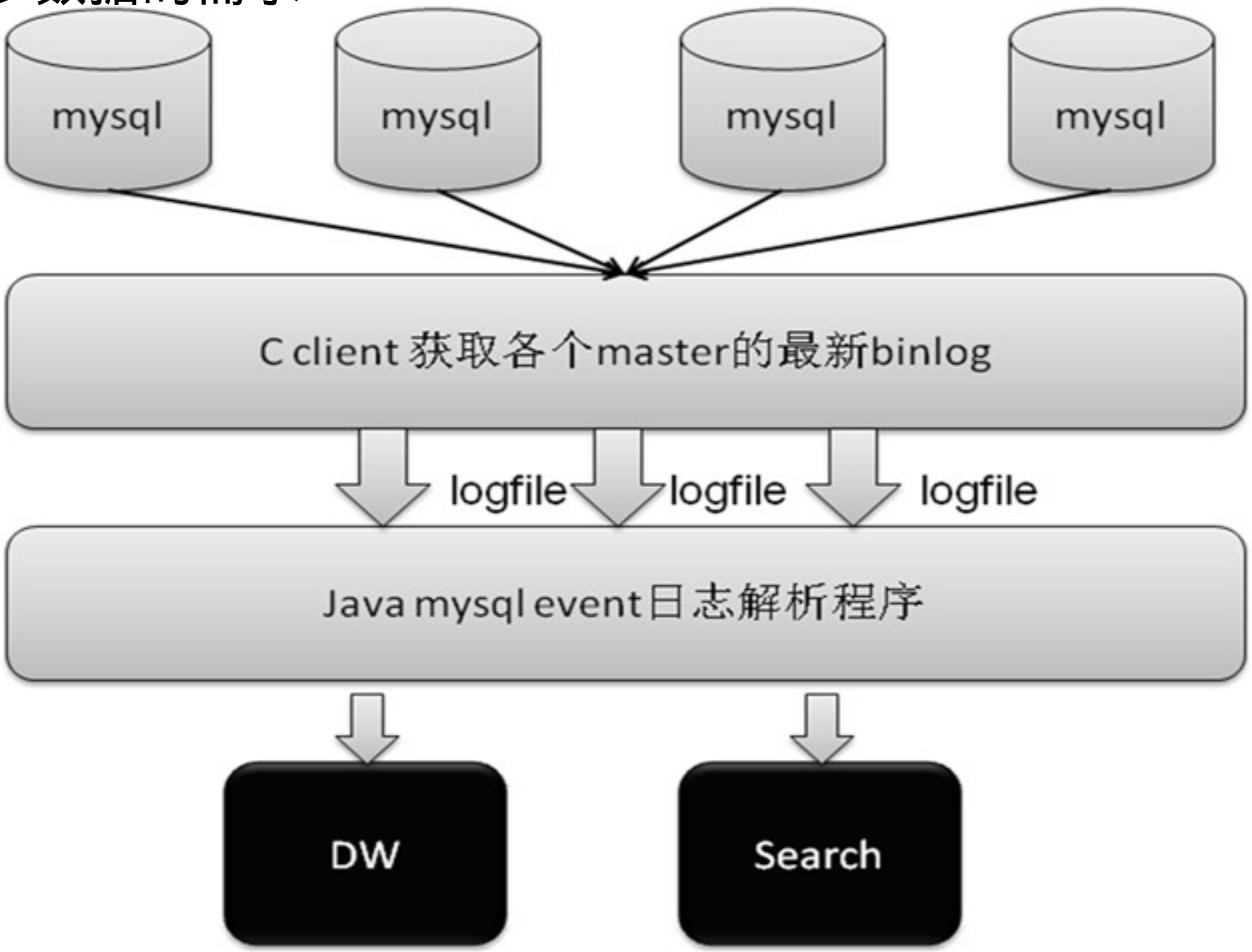
d.Innodb ssd 加速

e.Mysql replication 并行复制

MySQL Binlog 解析数据复制中心

淘宝网 Taobao.com

解决商品，用户，评价，收藏夹等应用向数据仓库，搜索增量同步数据的需求

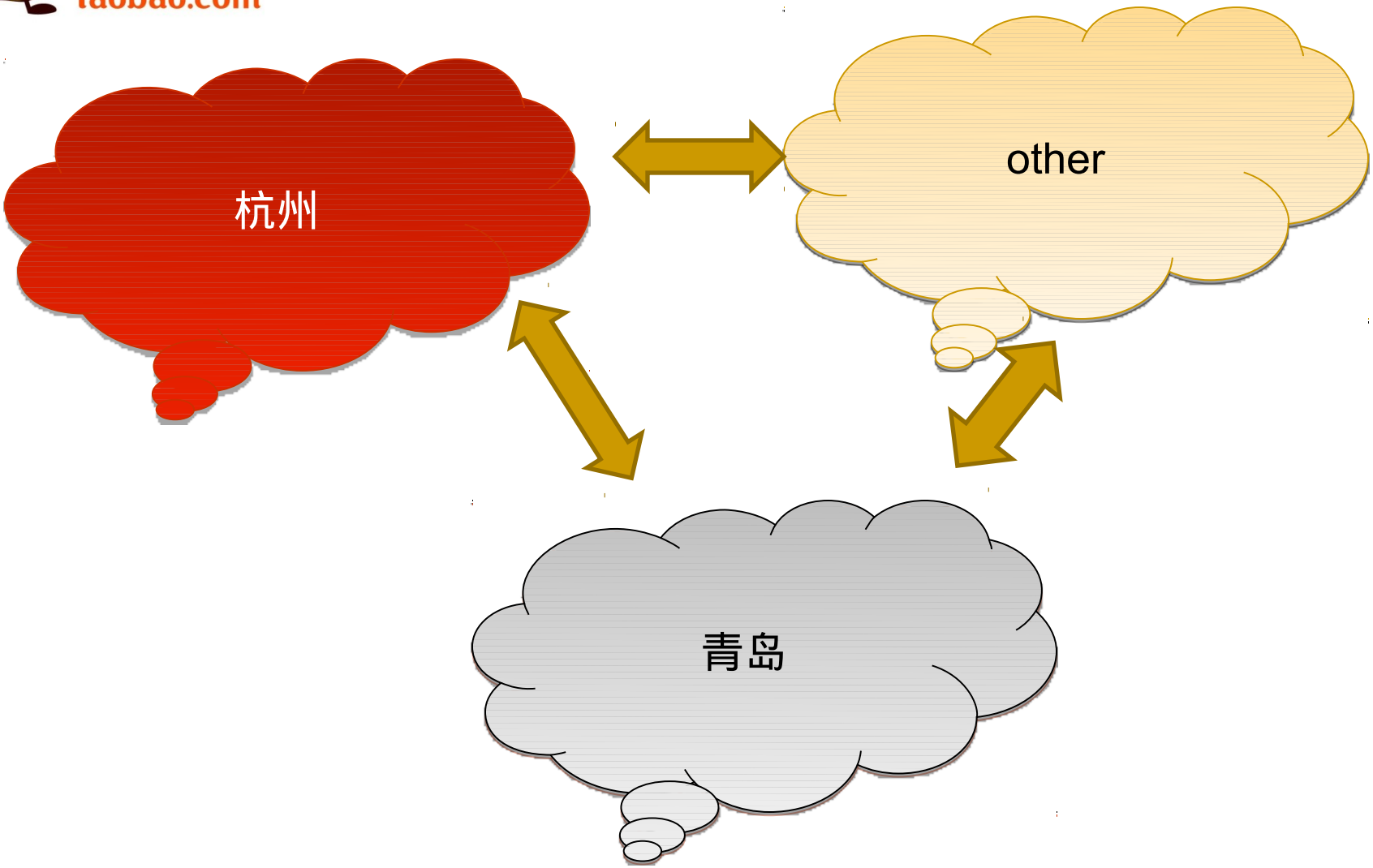


MySQL Binlog 解析数据复制中心

- C client 端特性：
  a. 支持 mysql master,slave 主备切换，获取 binlog 不受影响
  b. 自动重连主机
  c. 支持 checkpoint，支持断点续传 binlog

- Java 端复制代码特性：
  a. 支持 statement, row 两种复制模式
  b. 支持按规则复制
  c. 支持一定条件下的并行复制
  c. 支持 checkpoint

异地多数据中心的数据同步



异地多数据中心的数据同步

- 除了 oracle dataguard,master-slave replication 数据复制，我们还有其它哪些可选方案？

淘宝自主数据库 Oceanbase

- 动态数据与静态数据进行分离，动态数据采用集中式，静态数据存放与服务采用分布式
- 设计了一个宽表，冗余数据，将离散型 IO 合并成连续型 IO
- 每晚动态数据，与静态数据合并一次
- 将首先在收藏夹应用上试点

Questions ?