淘宝网
Taobao.com

OceanBase
结构化数据海量存储系统

杨传辉（日照）
weibo.com：淘宝日照

淘宝网 Taobao.com

Agenda

- OceanBase介绍
- OceanBase架构
- OceanBase应用
- 后续发展

淘宝网 Taobao.com

数据模型

- 数据模型
  - 主键（单列或者多列联合） + 普通列
  - 数据类型：整型，字符串，日期时间，高精度浮点数

主键		普通列		
user_id	item_id	price	collect_time	collect_count
100	10	200.0	2011-11-11	6473
100	11	180.0	2011-12-12	2198
101	9	100.0	2011-11-11	8888

- 支持的基本操作
  - 随机读取（指定主键的所有列）
  - 范围查询（指定主键的前缀）
  - 写操作（单行，多行，保证分区内事务）
  - filter, like, group by, order by, limit, offset，etc

淘宝网 Taobao.com

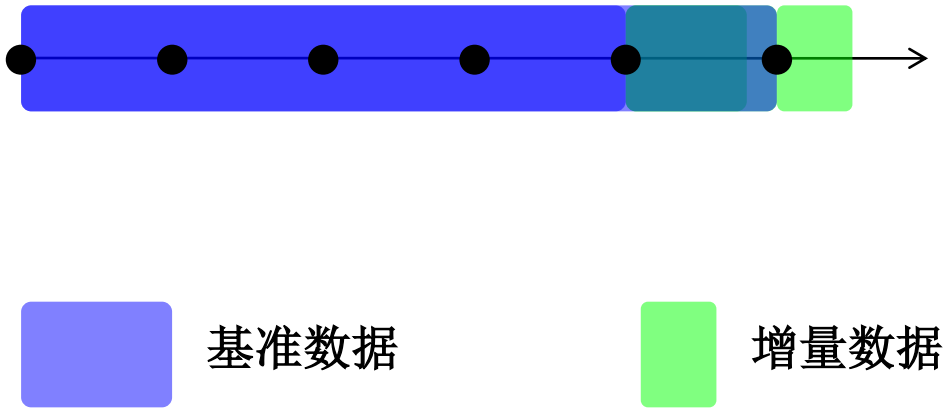
适用场景

- 特点
  - 结构化数据（核心商业数据，Hadoop，Streaming计算结果）
  - 在线存储：随机读取请求最多一次磁盘IO
  - 架构简单，强同步，宕机不停服务也不丢数据
  - 扩展性好，无需分库分表
  - 单机服务1TB ~ 4TB，节省成本
  - 特色功能：大表Join支持，千万级数据秒级实时分析
- 适用场景
  - 业务需要大表Join或者千万级数据秒级在线统计 => 考虑OB
  - 数据库性能不好，需要迁移到NOSQL => 考虑OB
  - 分库分表麻烦，数据增长快 => 考虑OB
  - NOSQL系统遇到问题，考虑其它NOSQL系统 => 考虑OB
  - 不适用场景：线下分析，网页库，淘宝图片存储等非结构化数据

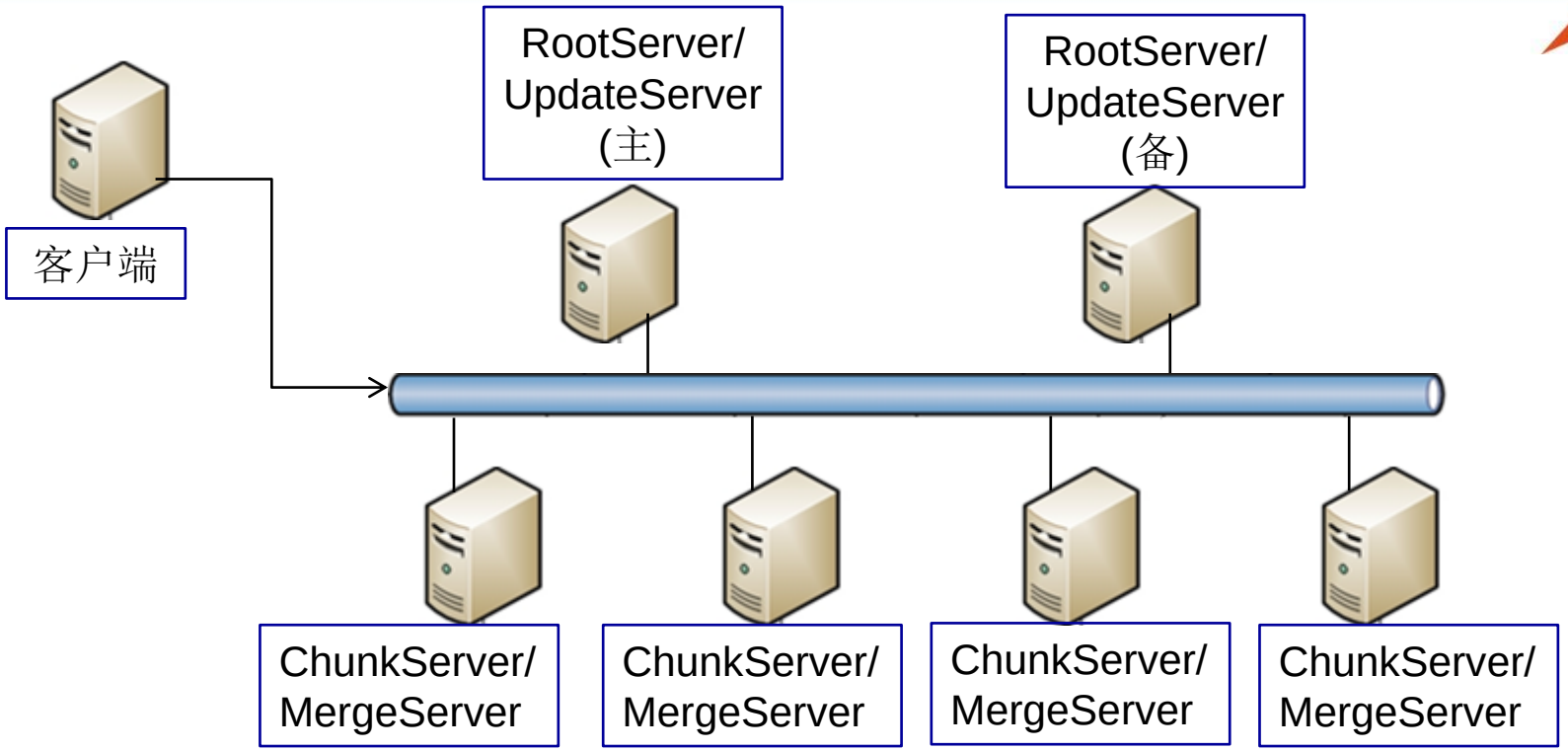
淘宝网 Taobao.com

在线存储特点

- 在线存储特点：数据量大但最近一段时间修改数据量不大
  - 基准数据和增量数据分离
  - 增量数据不断地合并到基准数据
  - 基准数据：数据量大，一般采用SATA或者SSD存储（多机）；
  - 增量数据：数据量小，一般采用内存或者SSD服务（单机）；



OceanBase系统架构



- 主控服务器RootServer：主+备，Schema/Tablet位置信息/机器管理
- 增量数据服务器UpdateServer：主+备，实时修改(内存+SSD)
- 基准数据服务器ChunkServer：多台，Tablet数据节点(磁盘或SSD)
- 查询合并服务器MergeServer：多台，基准，增量数据合并…

淘宝网 Taobao.com

设计要点

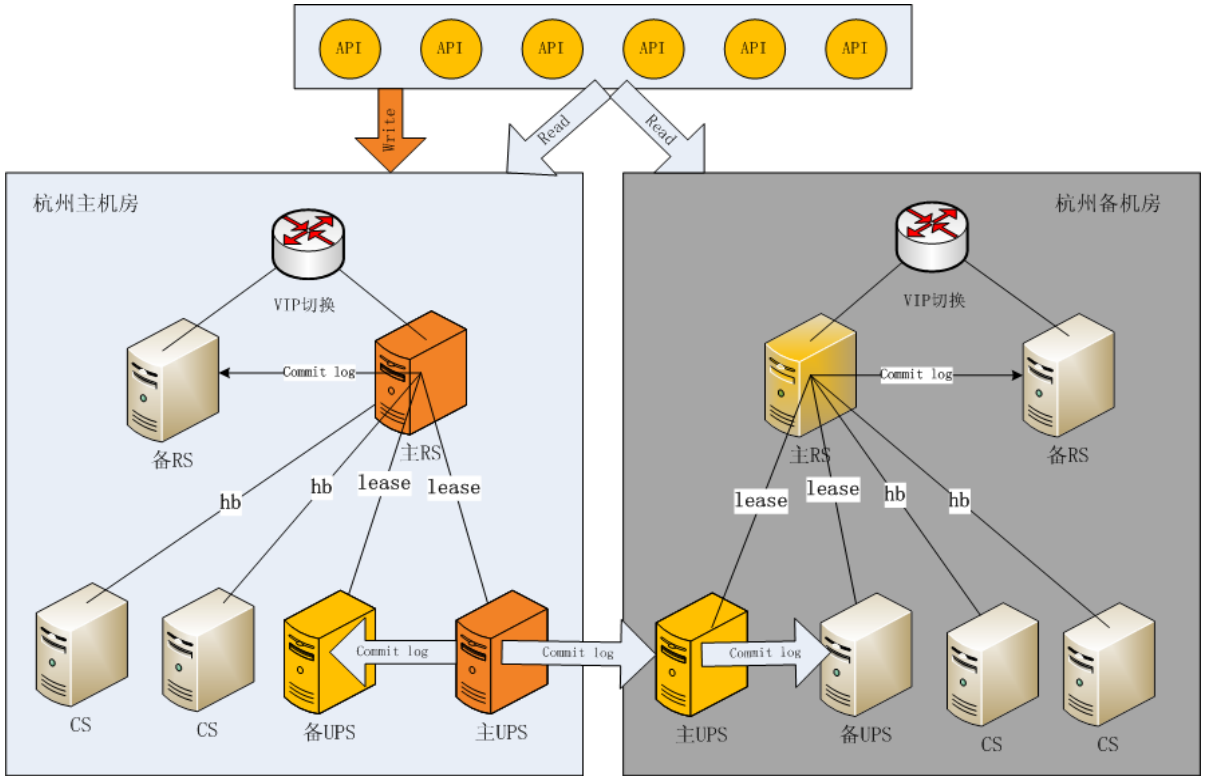
- RootServer
  - 自动负载均衡：按照Tablet个数均衡
  - 无状态，负载低，宕机无影响
- UpdateServer
  - 双机热备，强同步，宕机不丢数据；
  - Group Commit，操作日志RAID1，RAID卡带电池
  - 内存Copy-on-write B+ Tree，随机读QPS 100W，写TPS 20W
  - 内存数据定期dump到SSD（转储），支持热插拔
- ChunkServer
  - 数据一般存储2~3份；
  - SSTable数据结构：划分block，block有序，block内行有序
  - Block index缓存在内存中，一次读盘；
  - 异步IO，磁盘与CPU操作并行，列式存储；

UpdateServer是性能瓶颈?

- 内存容量
  - 每天更新一般不超过40G;
  - 大内存 + 转储到SSD;
- 磁盘
  - 操作日志：顺序写，group commit
- 网络
  - 网络出口带宽一般不超过50MB/s;
  - 普通网络收发包框架13w/s，优化过的框架30w~50w/s（千兆网卡）;
  - 万兆网卡，多网卡;
- CPU
  - 普通Get/Scan请求CPU占用很少;

多机房支持

- 多机房
  - 同城机房（强同步），异地机房（准实时同步）
  - 客户端配置多集群地址，按比例分配流量
  - 主集群脚本切换对外透明，服务器端程序版本无缝升级



淘宝网 Taobao.com

海量数据实时分析

- 海量数据实时分析（类似Google Dremel）
  - 支持千万级记录实时计算
    1. 请求拆分
    2. 请求分发到多个ChunkServer
    3. 每个ChunkServer计算局部结果
    4. 合并汇总，计算top N等
  - 支持按列存储；
  - 千万级数据实时统计时间控制在秒级，简单统计操作两秒内；

淘宝网 Taobao.com

收藏夹(2010年)

- 收藏夹需求
  - 收藏表保存收藏信息条目，40亿+
  - 商品表保存收藏的宝贝详细信息，4亿+
  - 收藏夹展示：收藏表和商品表两张大表join
- 收藏夹挑战
  - 一个用户可以收藏数千商品
  - 一件商品可被数十万用户收藏
  - 商品的属性实时变化
  - 单次查询响应时间<50ms
- 实验效果
  - 解决了大用户无法按照价格或者人气全排序的业务难题；
  - Mysql 16 * 2减少为Oceanbase 12 + 2
  - Load值更低，短期无扩容需求；
  - 平均响应时间30~50ms

收藏夹(2011.11.11)

- 数据条数：160亿（带有Join关系）
- 访问量
  - 11.10：修改量1.72亿次，scan和get分别为2.01亿次和2.57亿次；
  - 11.11：修改量1.85亿次，scan和get分别为2.30亿次和2.90亿次；
  - 最高峰：scan 9000QPS，get 5000QPS（操作需要join）
- 集群规模：44台，单机数据量600G~1.2T
- 高峰期合并：11.11手工触发合并操作，不影响服务
- 热门收藏：大店铺收藏商品数几万到几十万，实时计算top N人气收藏
  - 业务方加计算结果缓存
  - OB加强应用访问模式监控

CTU aCookie统计

- 应用情况
  - 存储每个用户的登录，交易等行为日志；
  - 每天写入25亿条，2.5T，至少保存3天数据；
  - 每天读取量100W左右，每次扫描一个用户的数据，几百KB；
  - 要求读取延时1s内；
- OB使用
  - Mongodb => OB
  - 分5个集群，每个集群一天写入500G，共25台
  - UPS的内存，SSD + CS的磁盘构成天然的多级混合存储，保证延时200~300ms内；

淘宝网 Taobao.com

OceanBase后续发展

- 可用性
  - 索引支持
  - 简单SQL支持
  - RESTful服务
  - 可运维性：内部表支持，运维工具开发
- 可扩展性
  - 大集群服务化支持
- 工作方向
  - OLAP优化
  - 性能优化
  - 定期合并操作实时化
  - Tablet合并
  - 开源支持：Postgre SQL代理开源（量子统计团队）

谢谢

- 个人博客：http://nosqlnotes.net