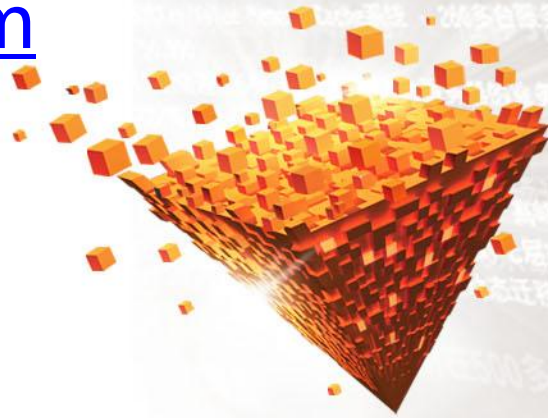
# 淘宝网

# 海量数据存储与管理

阳振坤
yangzhenkun@gmail.com
2011年10月14日



# Agenda

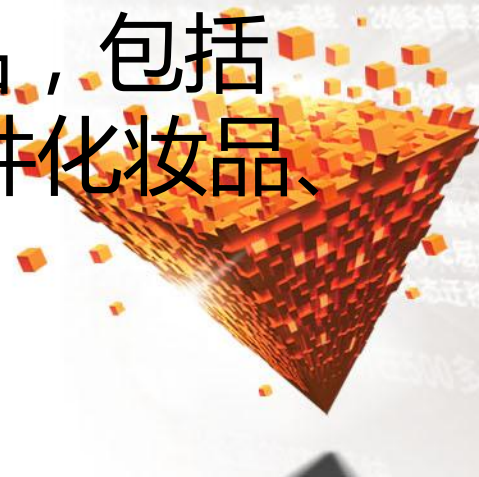
**关于淘宝** 你可能不知道的……

**存储与处理** 哪些数据？如何存储？怎么处理？

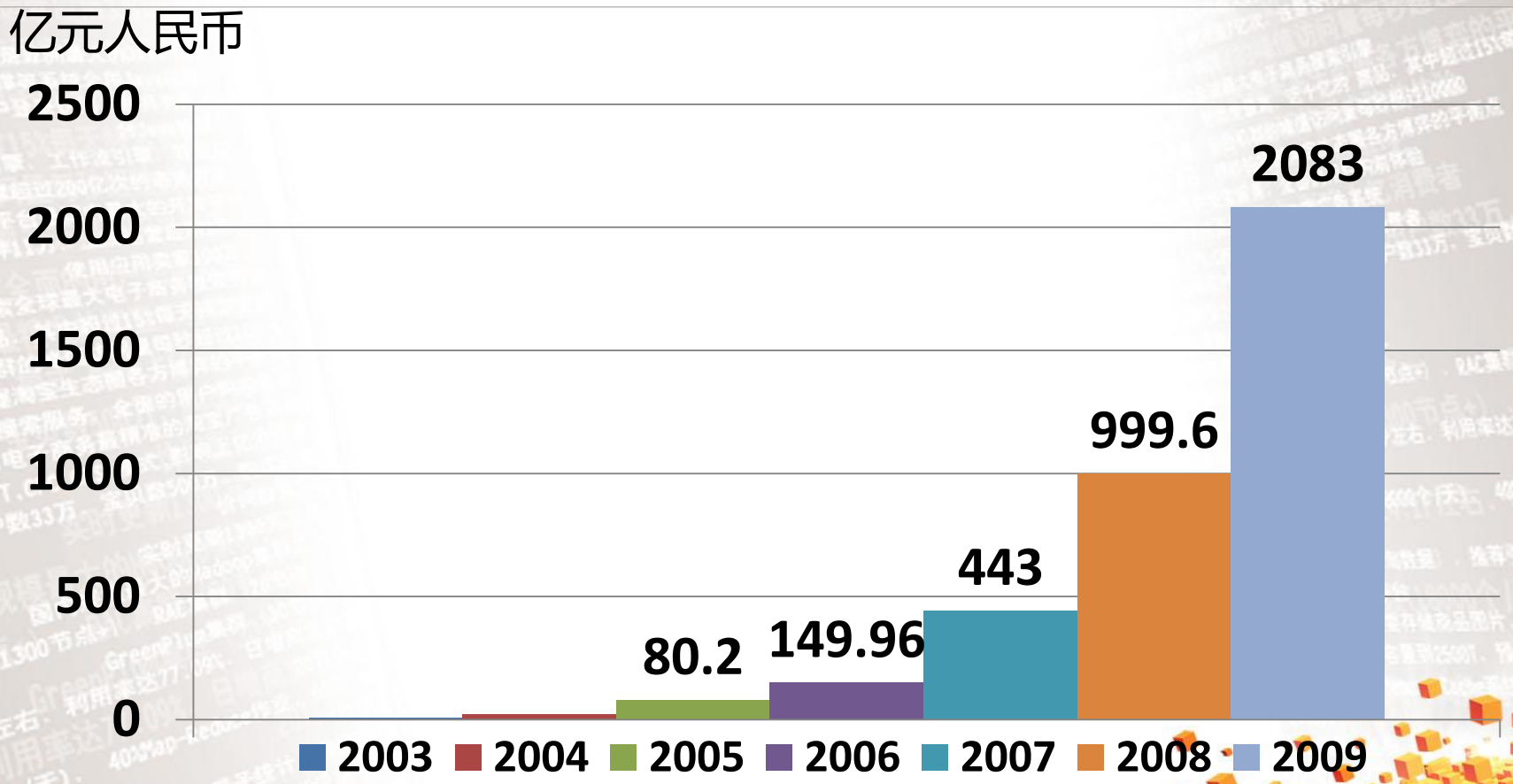
**淘宝数据库** 有何特别之处？OceanBase是什么？

# 关于淘宝

- 2003年5月10日成立，办公地点杭州/北京，员工6000+人，工程师40%...
- 3.7亿注册用户
- 300万人就业
- 平均每天20亿PV，6000万登录，800万笔交易
- 平均每分钟卖出：4.8万件商品，包括864件衣服，36部手机、880件化妆品、85本书...

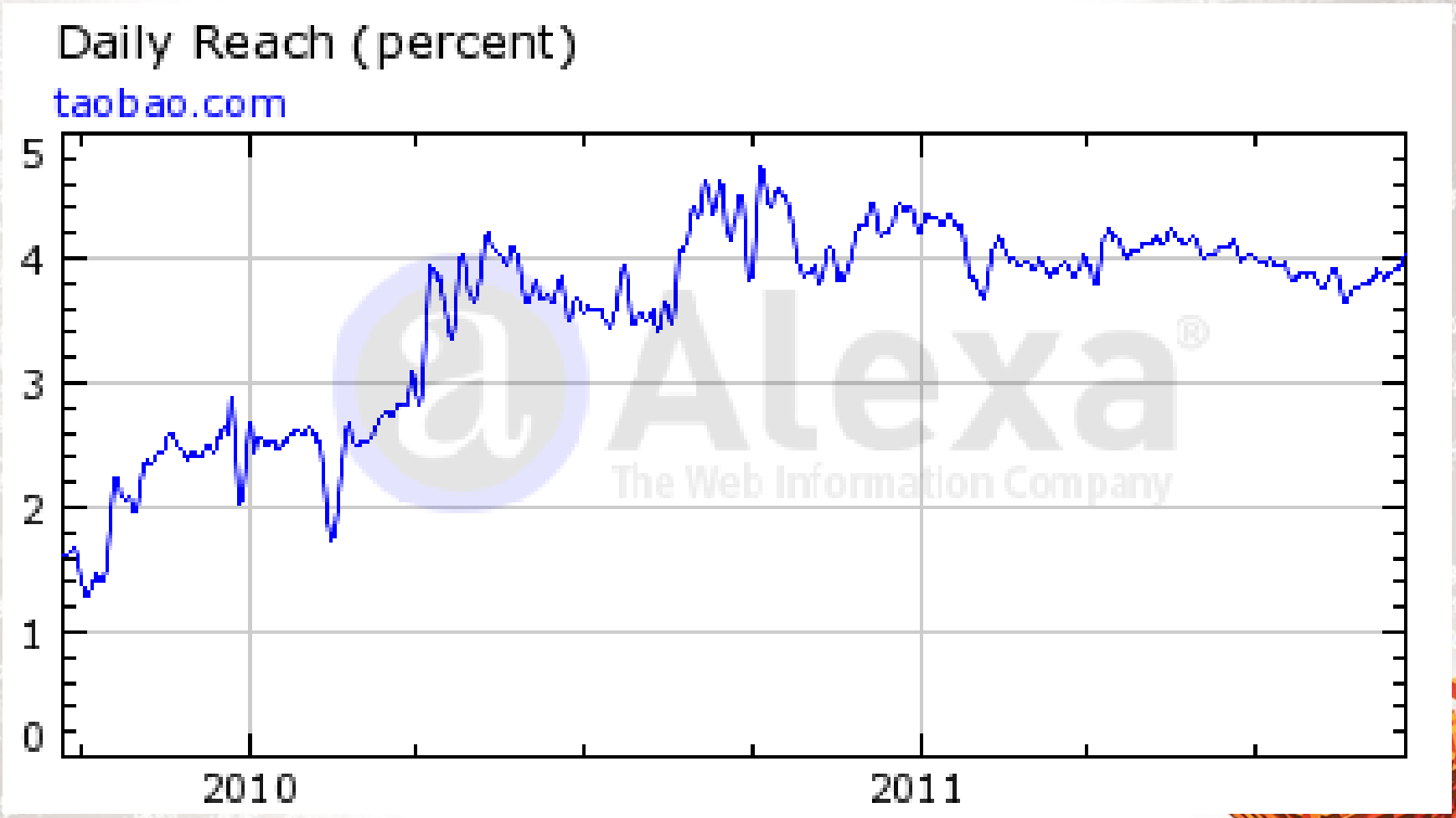


# 淘宝年度交易额




数据来自公开媒体

# 淘宝网站流量



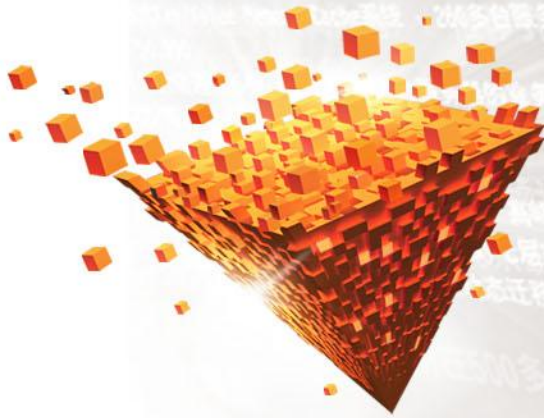

数据来源：www.alexa.com

# Agenda

| | |
|---|---|
| 关于淘宝 | 你可能不知道的…… |
| 存储与处理 | 哪些数据？如何存储？<br>怎么处理？ |
| 淘宝数据库 | 有何特别之处？<br>OceanBase是什么？ |

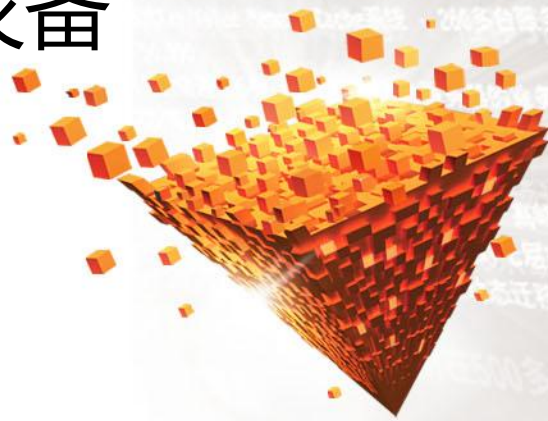
# 淘宝数据(1)：离线数据

- 39PB+，每天新增50TB
- 2000+台服务器Hadoop机群，每天40000+个MapReduce作业
- 2010年淘宝上最畅销手机价格？
- 2010年什么年货最畅销？
- 什么地方人最钟爱大闸蟹？



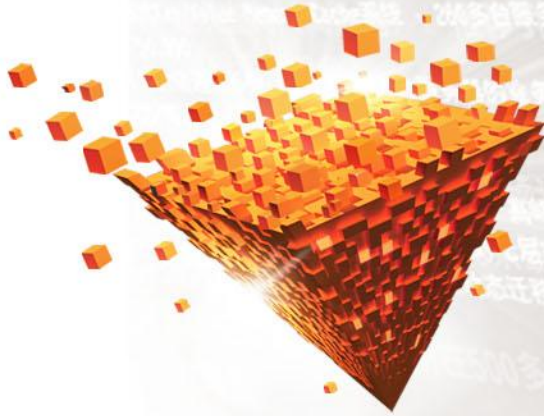
# 淘宝数据(2)：在线图片

- 商品图片(效果图，描述图)，交易快照，2700TB+...
- 淘宝为什么要存储这些图片？
- 淘宝文件系统TFS
- 实时响应，同城热备+异地灾备



# 淘宝数据(3)：结构化数据

- 商品、用户、订单、评价……
- 商品等的存储、修改及展示依赖于对应的结构化数据的存储、修改和查询
- **数据库**是可靠、高效地实现以上操作的最佳手段

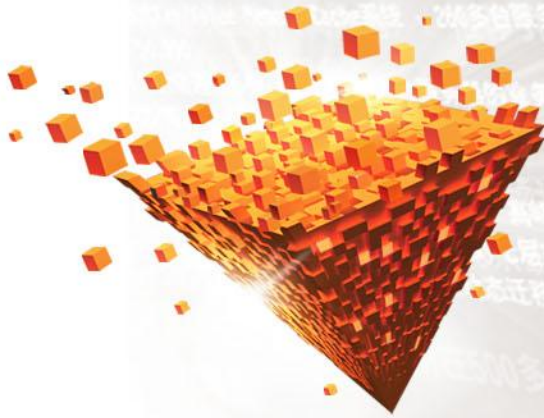


# Agenda

| | |
|---|---|
| 关于淘宝 | 你可能不知道的…… |
| 存储与处理 | 哪些数据？如何存储？<br>怎么处理？ |
| **淘宝数据库** | **有何特别之处？<br>OceanBase是什么？** |

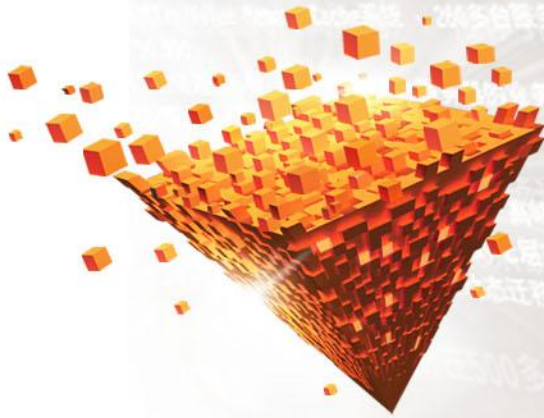
# 淘宝数据库特点

- 非常关键：几乎所有淘宝业务都依赖
- 数量多：以千计的数据库服务器
- 数据量大：单表几亿~几百亿条记录
- 访问量大：每天几亿~几百亿次访问
- **“Oracle数据库 + 小型机 + 高端存储”已成为历史**



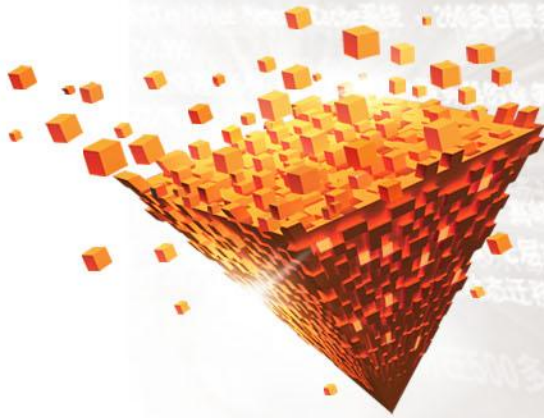
# DBMS：分库与分表

- 缓解了数据量大与访问量大的挑战
- 业务逻辑支持
- 扩展性、容错及故障恢复
- 分表后事务
- IOPS vs. 固态盘(SSD)



# OceanBase数据库

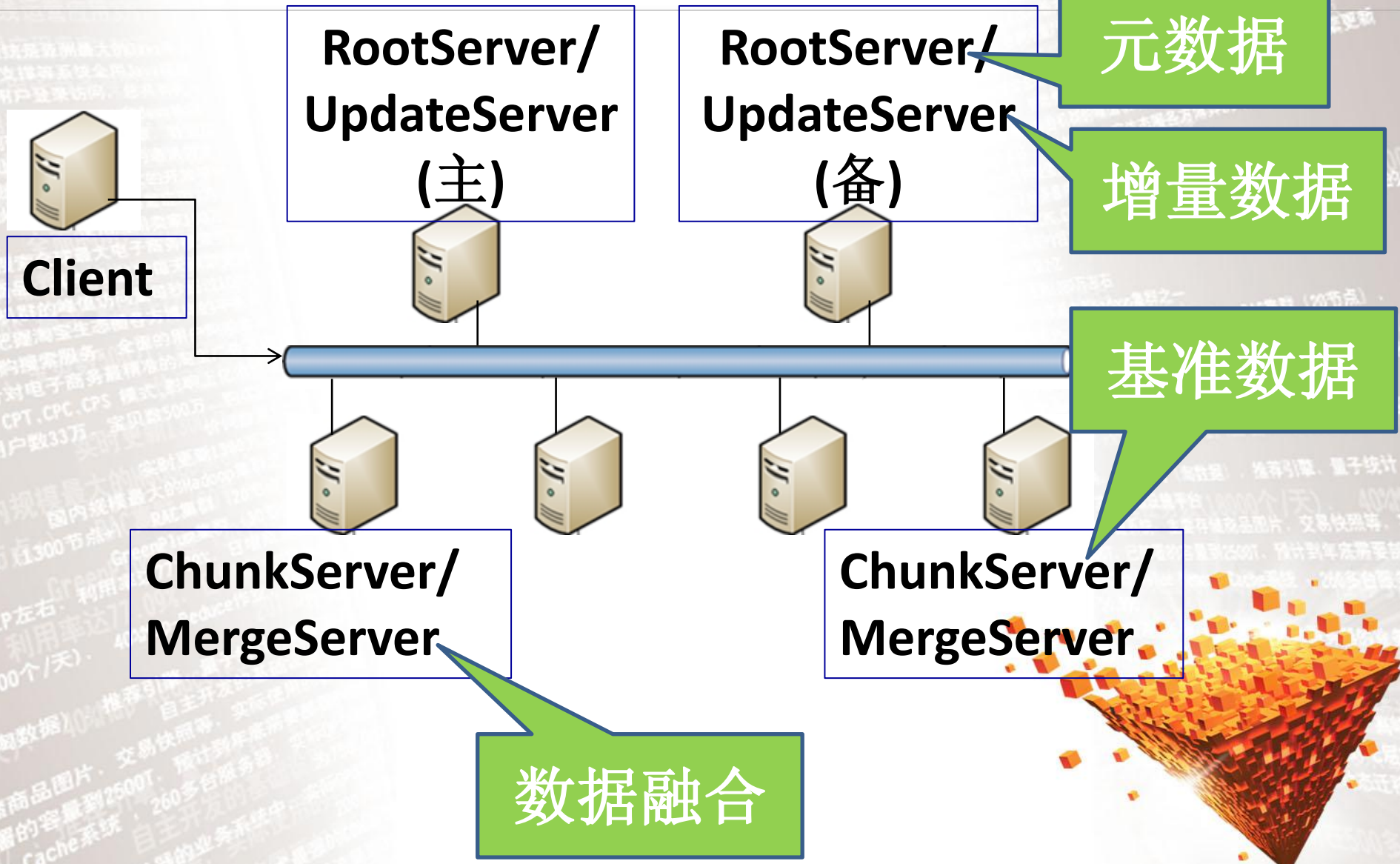
- 淘宝研发，基于普通PC服务器
- 千亿条记录，无需分库分表
- 跨行跨表事务(ACID)
- 自动容错和故障恢复
- 容量与性能可随服务器数量而增加
- 适合SSD(无随机磁盘写)



# 设计思想

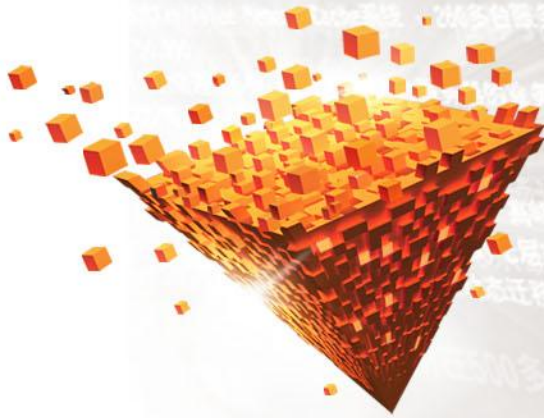
- 数据总量大，但更新比例小
- **数据 = 基准数据 + 增量数据**
- 增量数据：动态B树，内存 + SSD
- 基准数据：静态B+树、分布式存储(磁盘或SSD)

# 系统架构


RootServer/
UpdateServer
(主)
RootServer/
UpdateServer
(备)
元数据
增量数据
Client
基准数据
ChunkServer/
MergeServer
ChunkServer/
MergeServer
数据融合

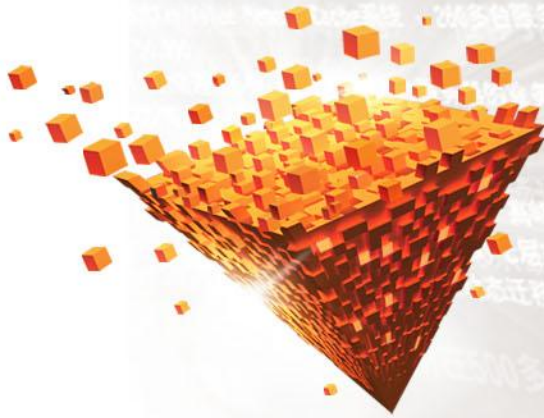
# 基准数据更新

- 新的基准数据 = 旧的基准数据 + 增量数据快照，ChunkServer(s)
- 低优先级 + 低负载时段
- 基准数据多副本**不必要**同步
- 基准数据多副本**必须**一致



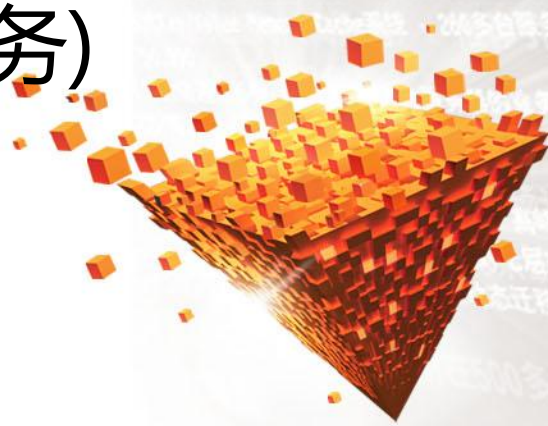
# 写性能&扩展性

- UpdateServer：B树 + Copy-on-write，10万TPS + 100万QPS(内存)
- Group commit + 带电池/电容RAID卡
- 万兆网卡
- 主备机 + 主备机群
- 一主多备，主写备读
- 内存 + SSD



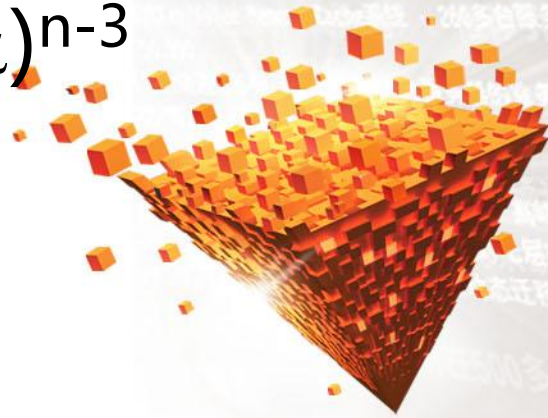
# 容错&故障恢复

- 单机群：3数据副本
- 同城热备(实时日志同步)：2+2数据副本
- 远程灾备(准实时日志同步)：2+2+2数据副本
- 数据记录自校验(磁盘&网络)
- 在线切换、在线升级(不停服务)



# 数据丢失几率分析

- n台设备，年度故障率λ，则单机t小时内故障概率为$\alpha = \lambda \times t/(360*24)$
- 恰好0台故障：$(1-\alpha)^n$
- 恰好1台故障：$n \times \alpha \times (1-\alpha)^{n-1}$
- 恰好2台故障：$n \times (n-1)/2! \times \alpha^2 \times (1-\alpha)^{n-2}$
- 恰好3台故障：$n \times (n-1) \times (n-2)/3! \times \alpha^3 \times (1-\alpha)^{n-3}$
- ……

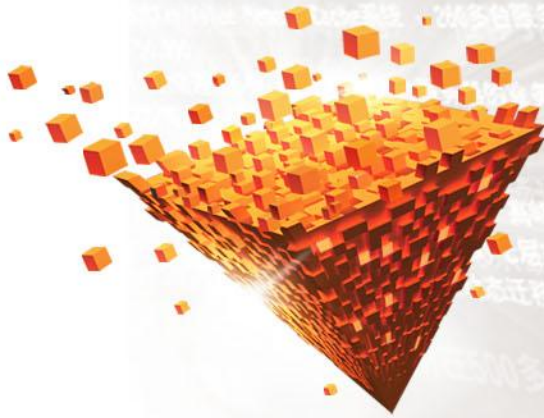


# 数据丢失概率

| 年故障率 | 设备数量 | 处理时长 | 单台故障率 | ≥2台故障率 | ≥3台故障率 | ≥4台故障率 |
|---|---|---|---|---|---|---|
| 5% | 10 | 0.5 | 2.9E-05 | 3.8E-08 | 2.9E-12 | 4.0E-16 |
| 5% | 10 | 1 | 5.8E-05 | 1.5E-07 | 2.3E-11 | 2.8E-15 |
| 5% | 10 | 2 | 1.2E-04 | 6.0E-07 | 1.9E-10 | 3.8E-14 |
| 5% | 50 | 0.5 | 1.4E-04 | 2.6E-05 | 5.9E-08 | 1.0E-10 |
| 5% | 50 | 1 | 2.9E-04 | 1.0E-04 | 4.7E-07 | 1.6E-09 |
| 5% | 50 | 2 | 5.8E-04 | 4.0E-04 | 3.7E-06 | 2.5E-08 |
| 5% | 100 | 0.5 | 2.9E-04 | 4.1E-04 | 3.8E-06 | 2.7E-08 |
| 5% | 100 | 1 | 5.8E-04 | 1.6E-03 | 3.0E-05 | 4.2E-07 |
| 5% | 100 | 2 | 1.2E-03 | 6.2E-03 | 2.3E-04 | 6.4E-06 |

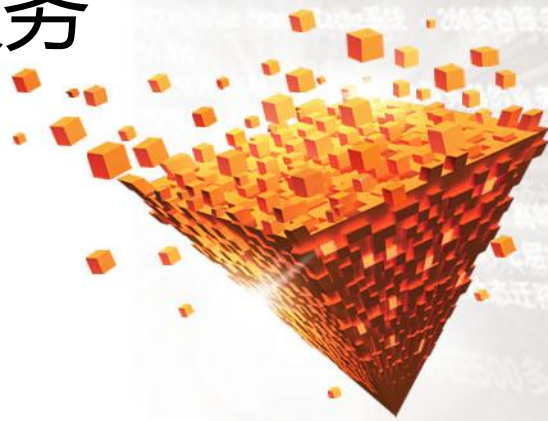
# 收藏夹线上运行状况

- 半年前上线
- 单表超过60亿条记录
- 1.2亿次/天访问
- 服务器数量：16*2 → 14*2 → 6*2
- 平均查询响应时间：~80ms



# 源码开放

- 淘蝌蚪：数十个开源项目/工具
  http://code.taobao.org/
- **OceanBase：淘宝海量数据库**
  **http://oceanbase.taobao.org/**
- TFS：淘宝分布式文件系统
- TAIR：淘宝分布式CACHE服务
- ……



# Agenda

| | |
|---|---|
| 关于淘宝 | 你可能不知道的…… |
| 存储与处理 | 哪些数据？如何存储？<br>怎么处理？ |
| 淘宝数据库 | 有何特别之处？<br>OceanBase是什么？ |

# Q&A

# Thanks

邮件: yangzhenkun@gmail.com
微博(**淘正祥**): http://weibo.com/1070095910
博客: http://blog.sina.com.cn/kern0612