京东云存储服务和应用探索

京东商城云平台部架构师　柳刘

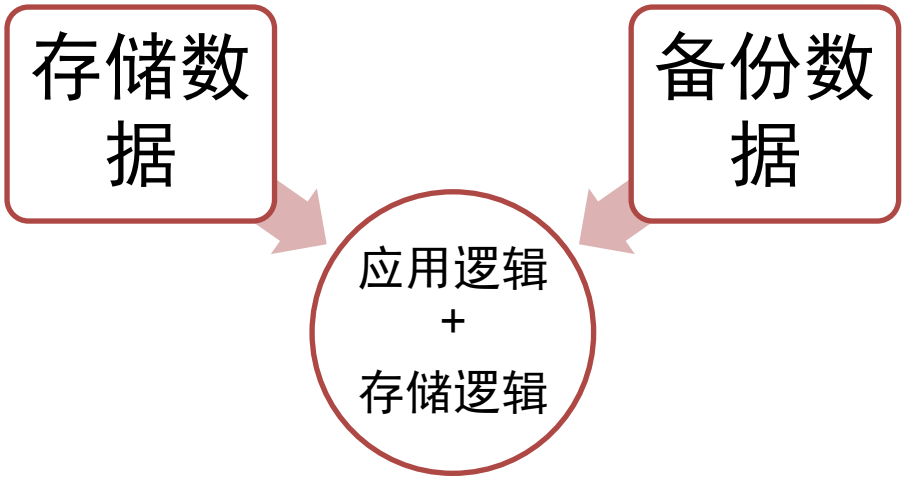
cdliuliu@jd.com

2013.04

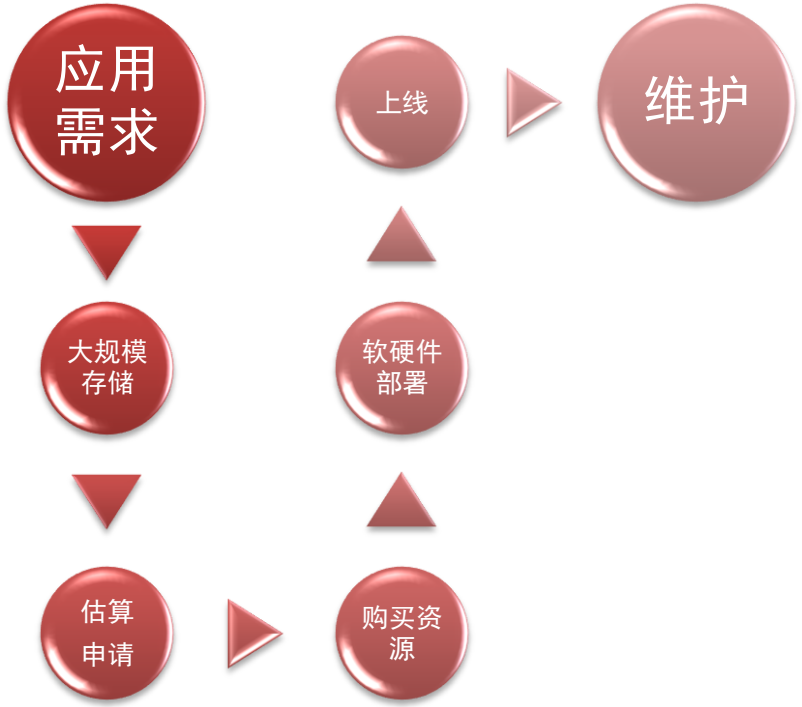
大纲

- 我们为什么要做云存储？
  - 应用场景
- 总体架构
  - 挑战
  - 开放API
  - 安全
  - 小插曲：分库分表和全局ID实现
- 架构演进
  - 三个阶段，不同规模
  - 技术堆叠
- 典型应用——网盘

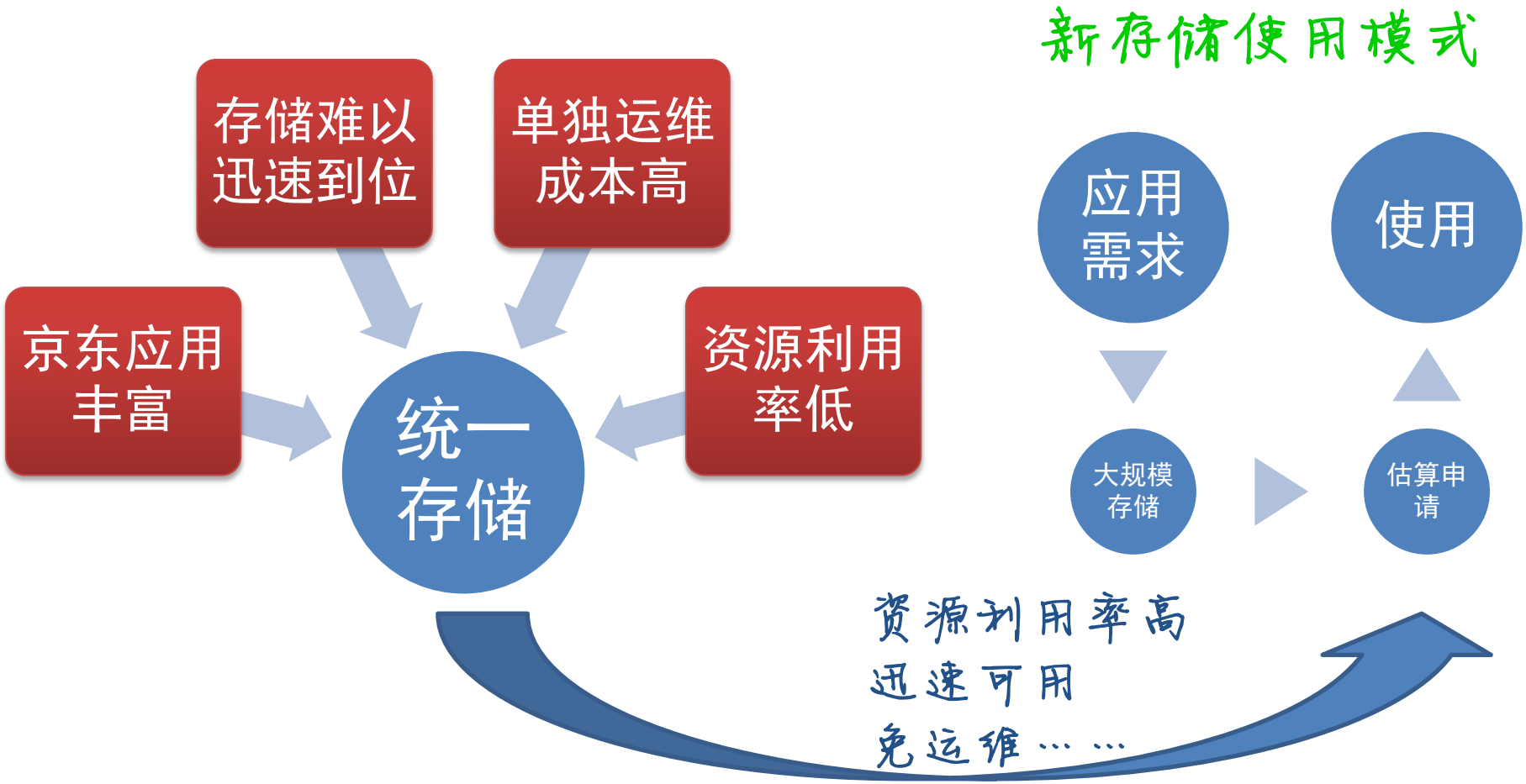
原存储使用模式



资源分散
重复建设
单独运维



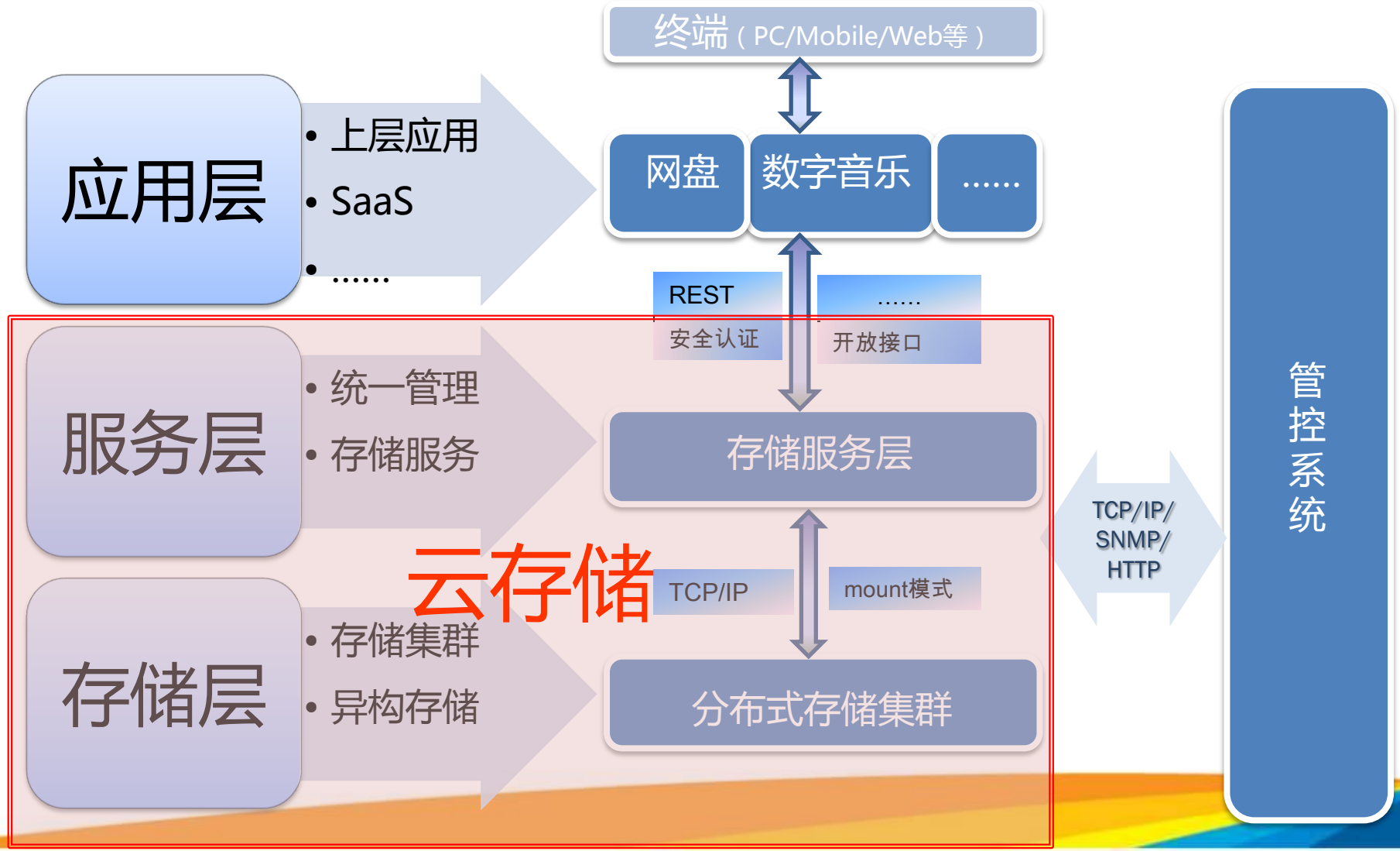
为什么需要云存储


新存储使用模式
存储难以迅速到位
单独运维成本高
京东应用丰富
统一存储
资源利用率低
应用需求
使用
大规模存储
估算申请
资源利用率高
迅速可用
免运维……

统一存储挑战

可伸缩性	• 灵活的水平扩展以应对数据不断扩张的压力
高响应	• 如何保证响应时间比单独存储无明显增加
安全性	• 应用不备份数据、云存储必须保证数据安全
吞吐量	• 满足业务吞吐量要求、业务之间互不影响
运维	• 如何及时发现并解决问题

云存储生态系统



存储服务层

安全

- 数字签名&安全认证
- 数据加密存储

高效

- 大文件分块传输
- 断点上传下载

运维

- 提供关键业务数据监控，即时了解存储情况，为业务分析提供数据支持。

底层存储层

可靠

- 存储数据多份冗余，防止数据丢失。
- 可异地容灾备份

高可用

- 多存储节点热备。
- 存储节点负载自动均衡。

可扩展

- 存储空间水平扩展

最初架构

存储服务层

- 签名认证
- 加解密
- 访问控制列表
- SDK

元数据存储

- MySQL Sharding

存储层

- 分布式
- 多副本
- 去中心化

开放API

RESTful风格API，易用兼容

基于SDK快速开发

- 上传
- 下载
- 删除
- 外链

外链分享

- 分享为url外链
- 统计和防盗链

云存储的安全

认证：客户端生成签名Token

Token的双Key设计+时间戳+.....

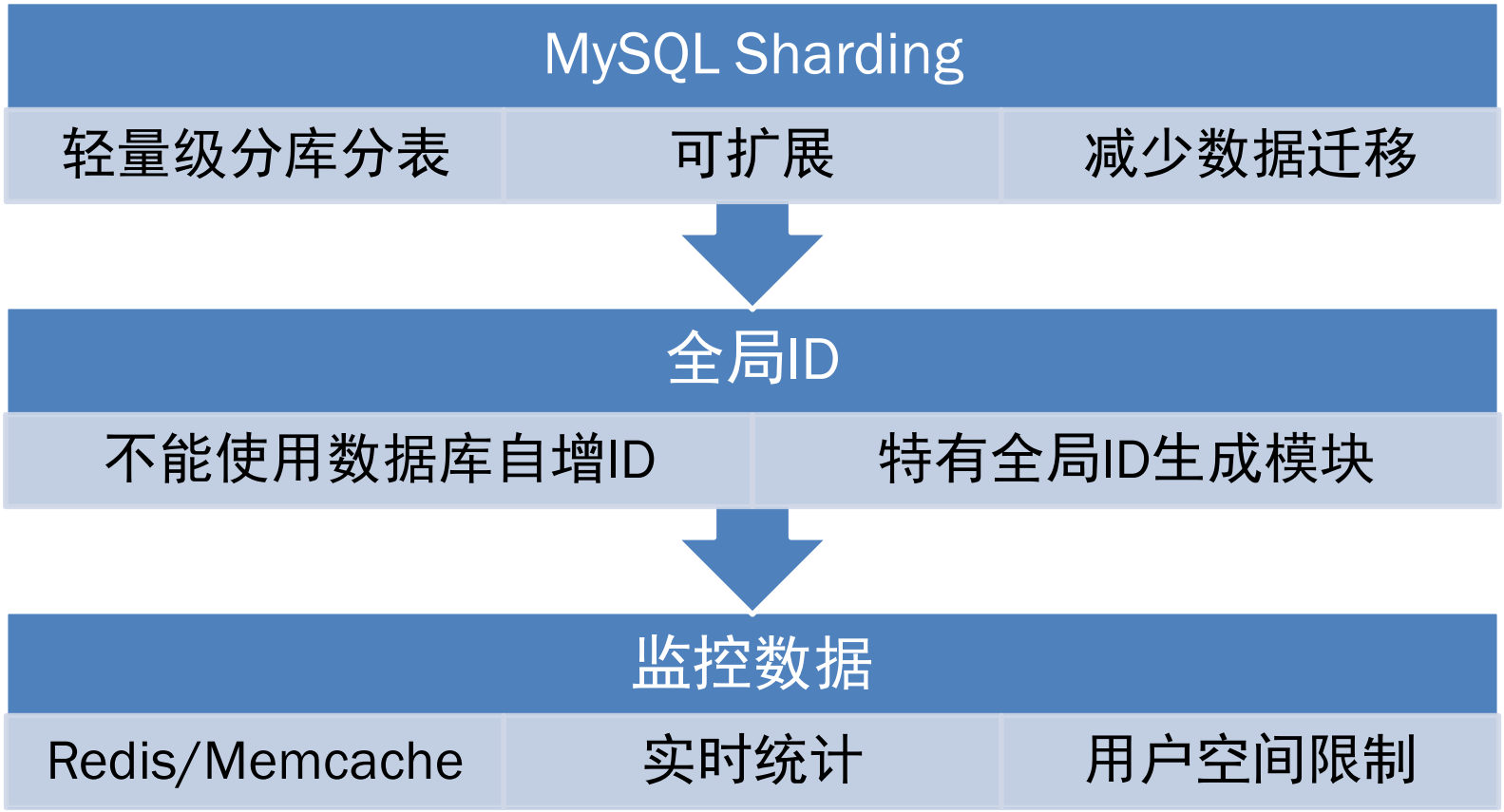
防盗链

用户访问控制设计

加解密：AES256资源消耗

自主研发快速加密算法

文件元数据存储


MySQL Sharding
轻量级分库分表
可扩展
减少数据迁移
全局ID
不能使用数据库自增ID
特有全局ID生成模块
监控数据
Redis/Memcache
实时统计
用户空间限制

轻量级分库分表

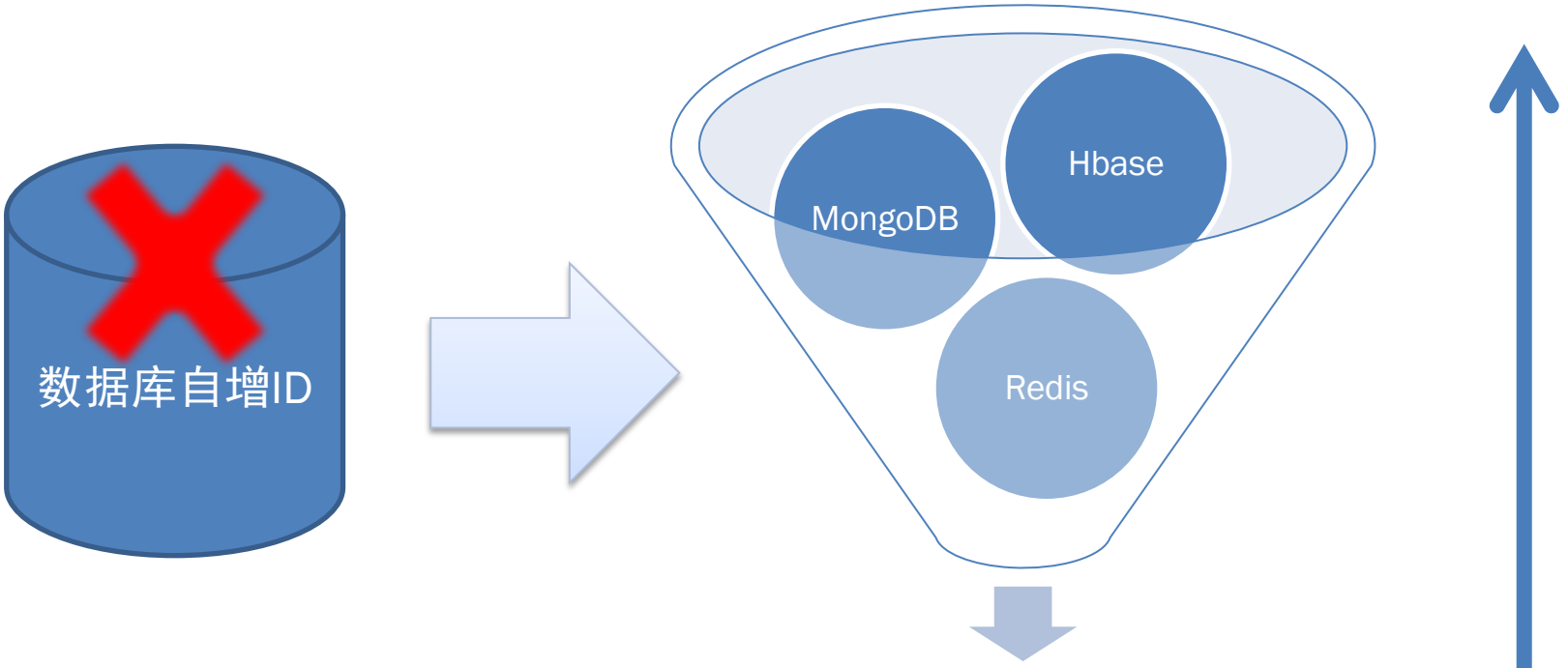
业务特性

- 用户数据空间逻辑独立
- 空间使用情况统计监控
- 半预测式

设计

- 使用轻量级Sharding层，半预测可扩展
- Sharding层基本无附加延时（No Payload）
- 通过规则引擎灵活配置Shard规则，例：
  - if［老用户的元数据 between 阀值1 and 阀值2］then 新用户元数据分配新的数据库Shard
  - elseif［老用户元数据 > 阀值2］then 使用新的计算方式为老用户新增元数据分配新增数据库Shard

全局ID



MySQL

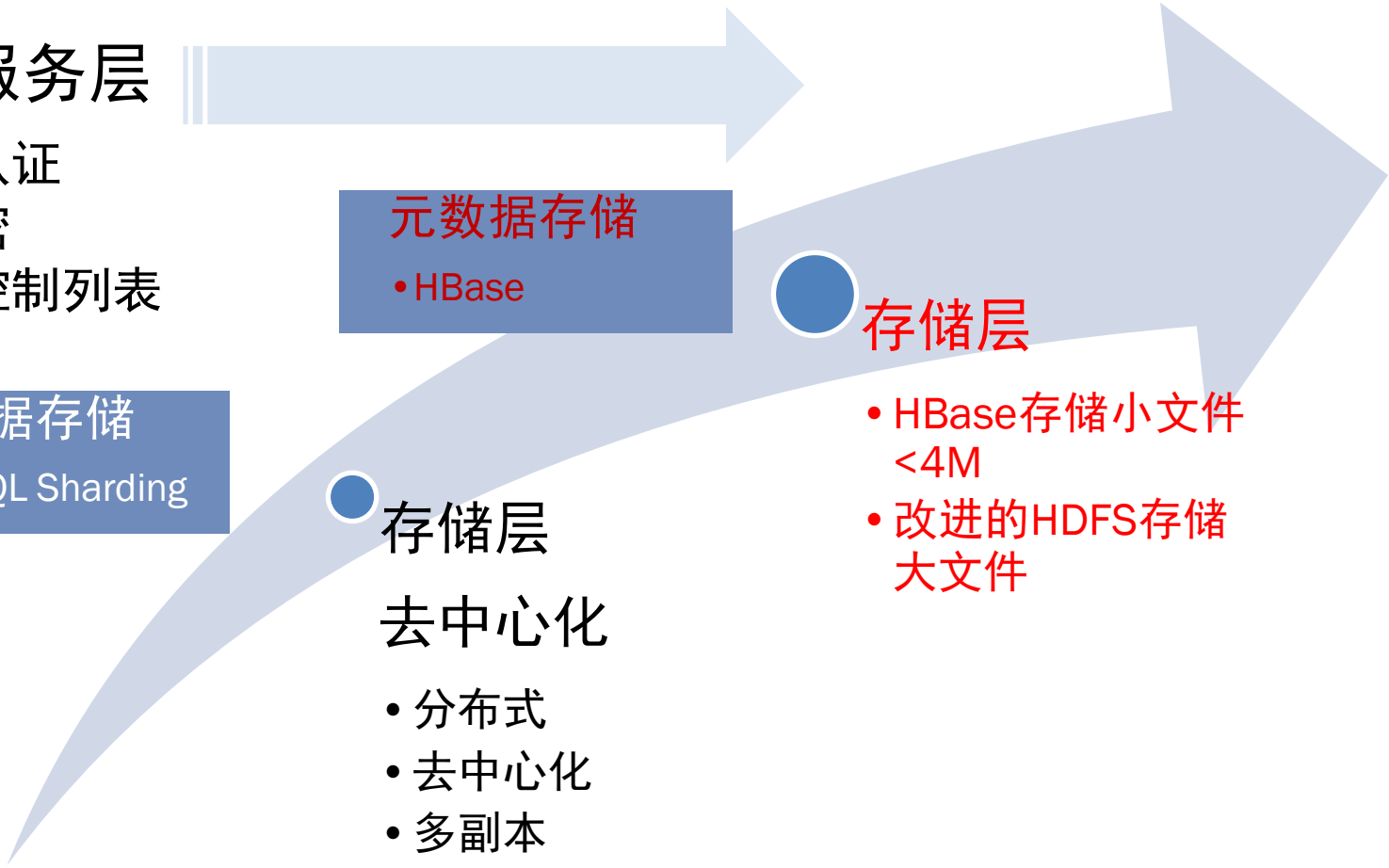
- 只用一张单表
- 客户端每次获取一段ID空间，缓存起来供使用 高TPS

遇到的瓶颈

- 文件存储
  - 分布式环境运维困难
  - 存储规模——中小规模

- 元数据存储
  - 扩展需要修改规则
  - 半预测式，对运维和监控要求高

成熟架构


存储服务层
• 签名认证
• 加解密
• 访问控制列表
• SDK
元数据存储
• MySQL Sharding
元数据存储
• HBase
存储层
去中心化
• 分布式
• 去中心化
• 多副本
存储层
• HBase存储小文件 <4M
• 改进的HDFS存储大文件

改进的HDFS

- 多NameNode
- NameNode自动切换
- 锁机制优化
- 日志同步优化、大大提高写并发
- 优化元数据——压缩和信息紧缩
- 增加缓存设计
- 更加可靠，适合在线应用

改进的HDFS数据

- 并发
  - 测试环境30节点
  - 90,000并发
- 吞吐量
  - 测试环境30节点
  - 读 >2.5GB/s
  - 写 >1.4GB/s

技术堆叠

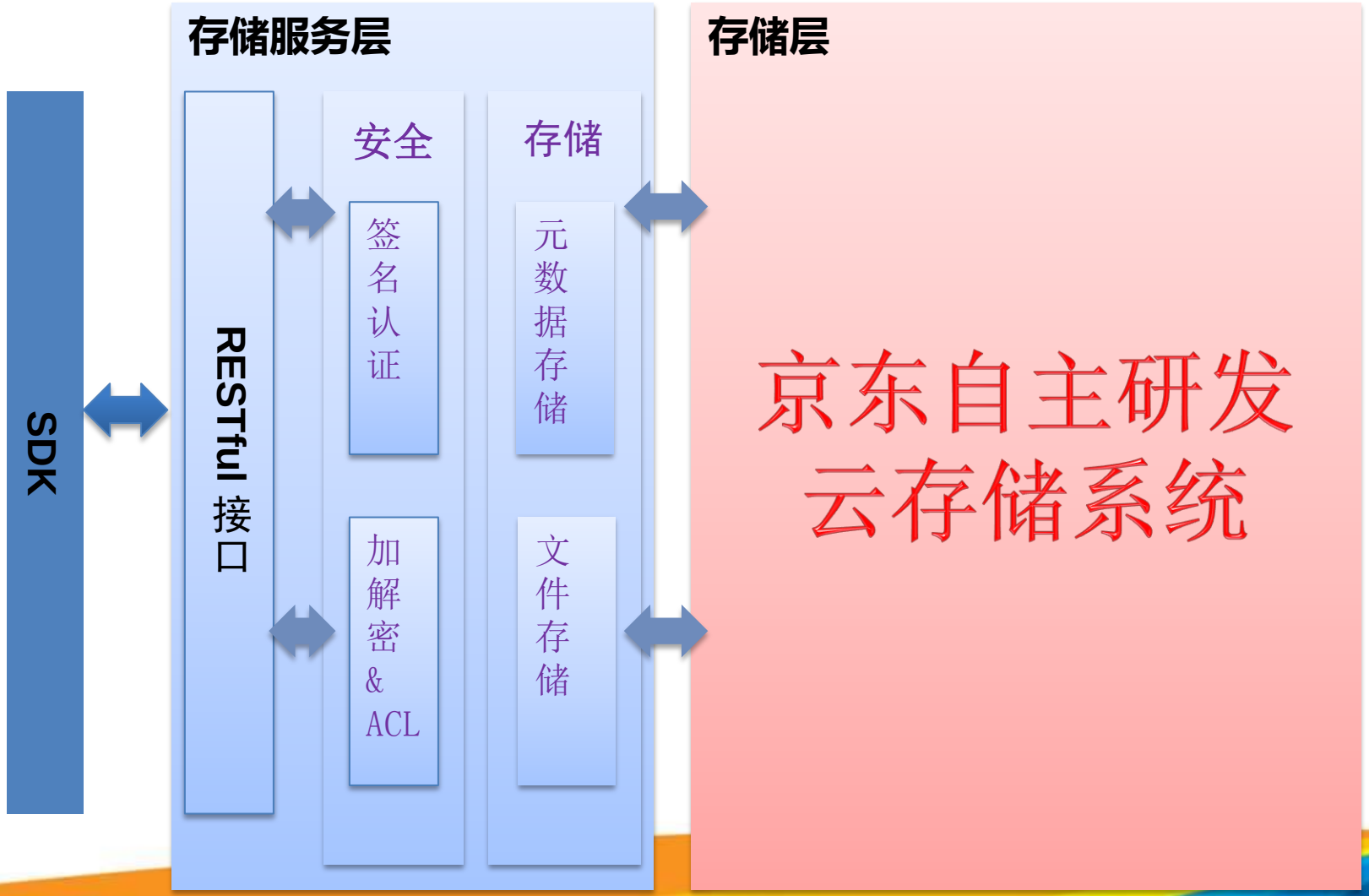
Restful → 认证 → 元数据 → 安全控制 → 底层存储

RestEasy	SHA-HMAC	HBase	自主加密算法	JHDFS
		索引 分析 事务 Redis/Memcache		
• 高性能流服务器		• ~~MySQL Shard~~	• ~~AES256~~	• ~~原JDFS~~ • 多副本

优化空间

- 目前分布式实现在软件层面
- 普通服务器没有针对存储优化
- 硬件底层有很大优化空间
- 硬件架构可以更好设计以支持集群存储
- 从硬件底层到驱动、到软件、到服务层整体优化设计

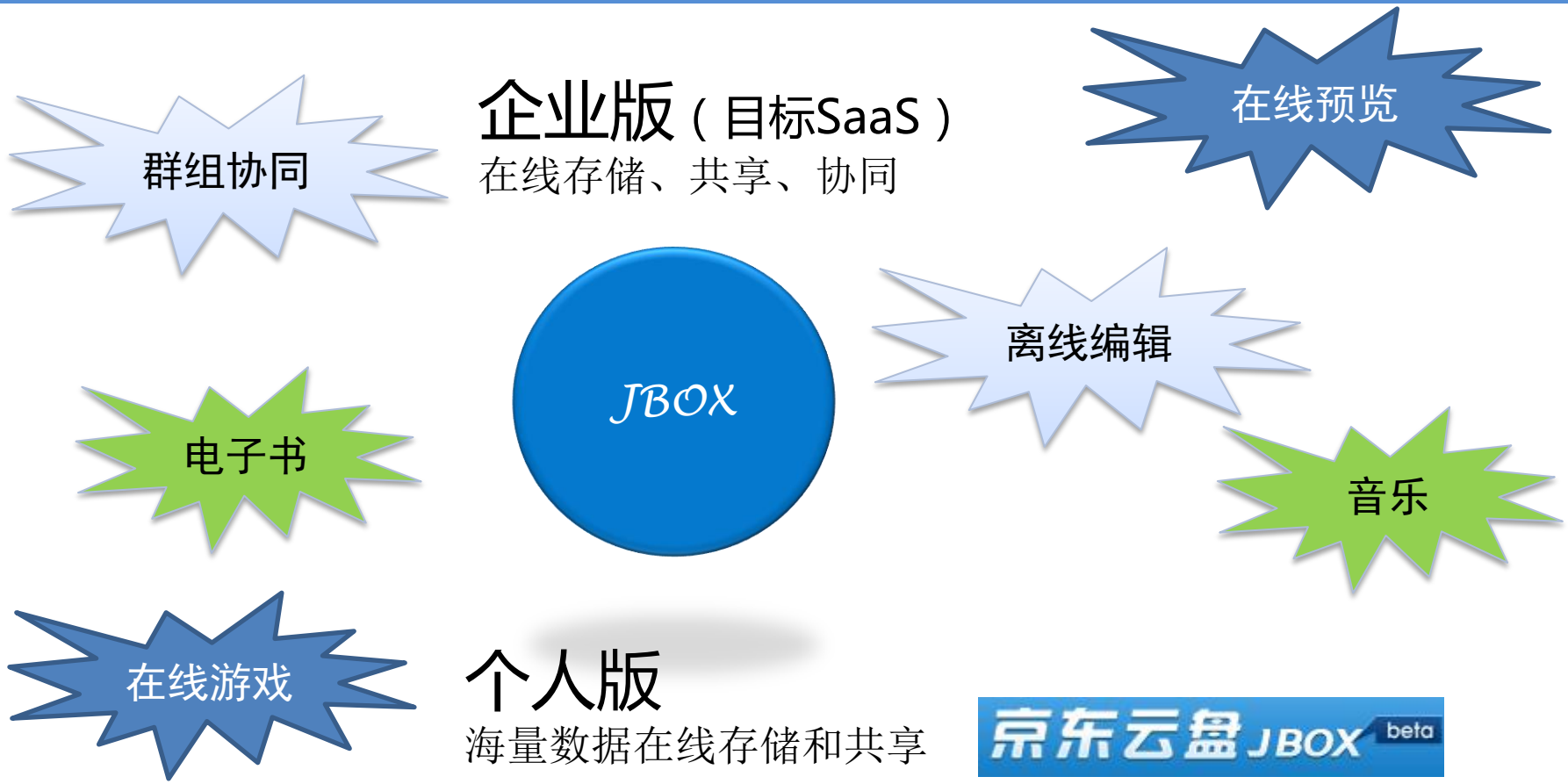
架构展望


存储服务层
存储层
SDK
RESTful 接口
安全
存储
签名认证
元数据存储
加解密&ACL
文件存储
京东自主研发
云存储系统

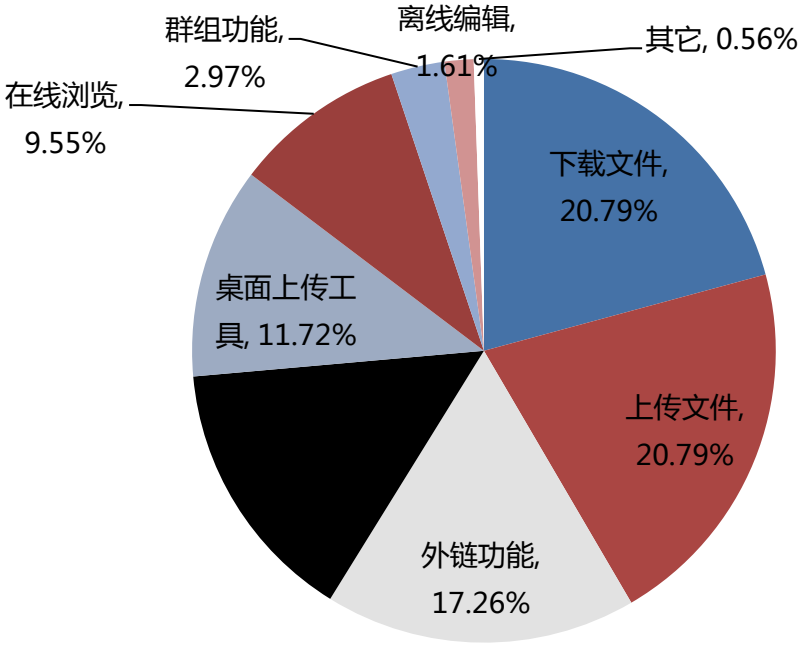
经验

- 敏捷快速开发，不断实践，积累经验
- 管理系统对云存储服务至关重要
- 系统运维和快速反应
- 节点之间延时过大：部署中减少物理环节
- 自动扩展（AS）、虚拟化、智能升级

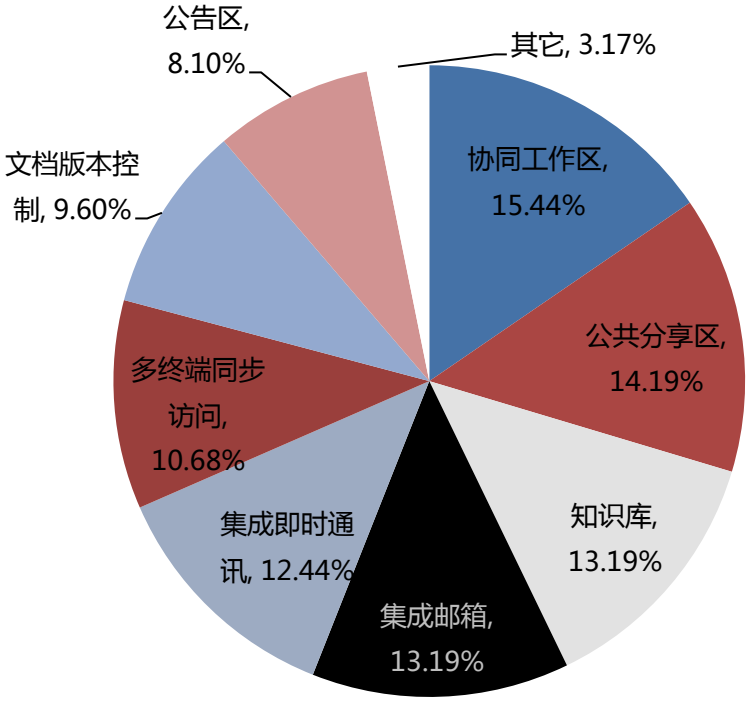
基于云存储典型应用介绍——JBox



JBox统计数据



使用功能占比



用户最喜爱功能

谢 谢！