



2013中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2013

大数据 数据库架构与优化 数据治理与分析

SequeMedia
盛拓传媒

IT168.com

ITPUB

ChinaUnix

Database
BDaas
flowingdata
DB2
NoSQL MySQL
Oracle Big Data

Hadoop 2012 & 2013 @ baidu

提纲

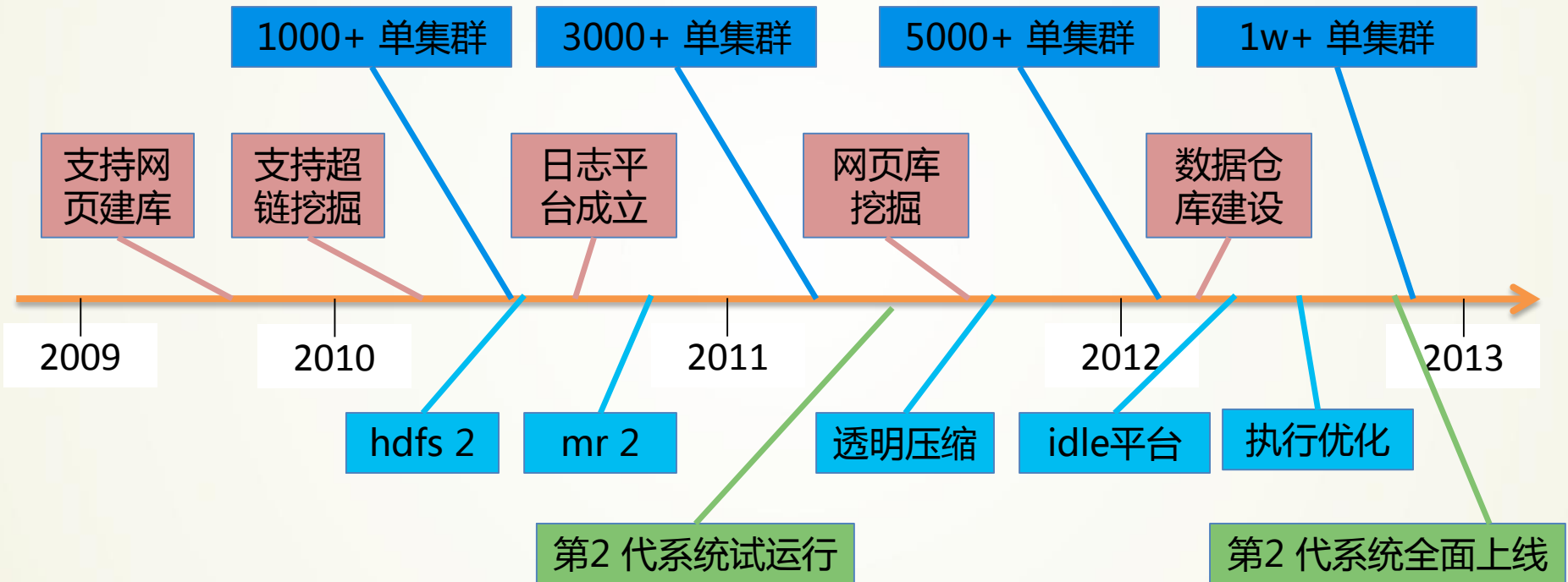
1. baidu hadoop现状
2. 存储系统进展
3. 计算系统进展

baidu hadoop 现状 – 规模

- 2008年
 - 始于 社区 0.18~0.19 之间的trunk版本
 - 300台机器，2个集群

- 2013年
 - 总机器 4.8w+
 - 单集群最大规模 1w+
 - CPU利用率70%+
 - 日均作业数？日均输入数据量？
 - 总inode数？使用磁盘空间？

baidu hadoop 现状 - 发展



存储 – hdfs2全面上线 – 背景

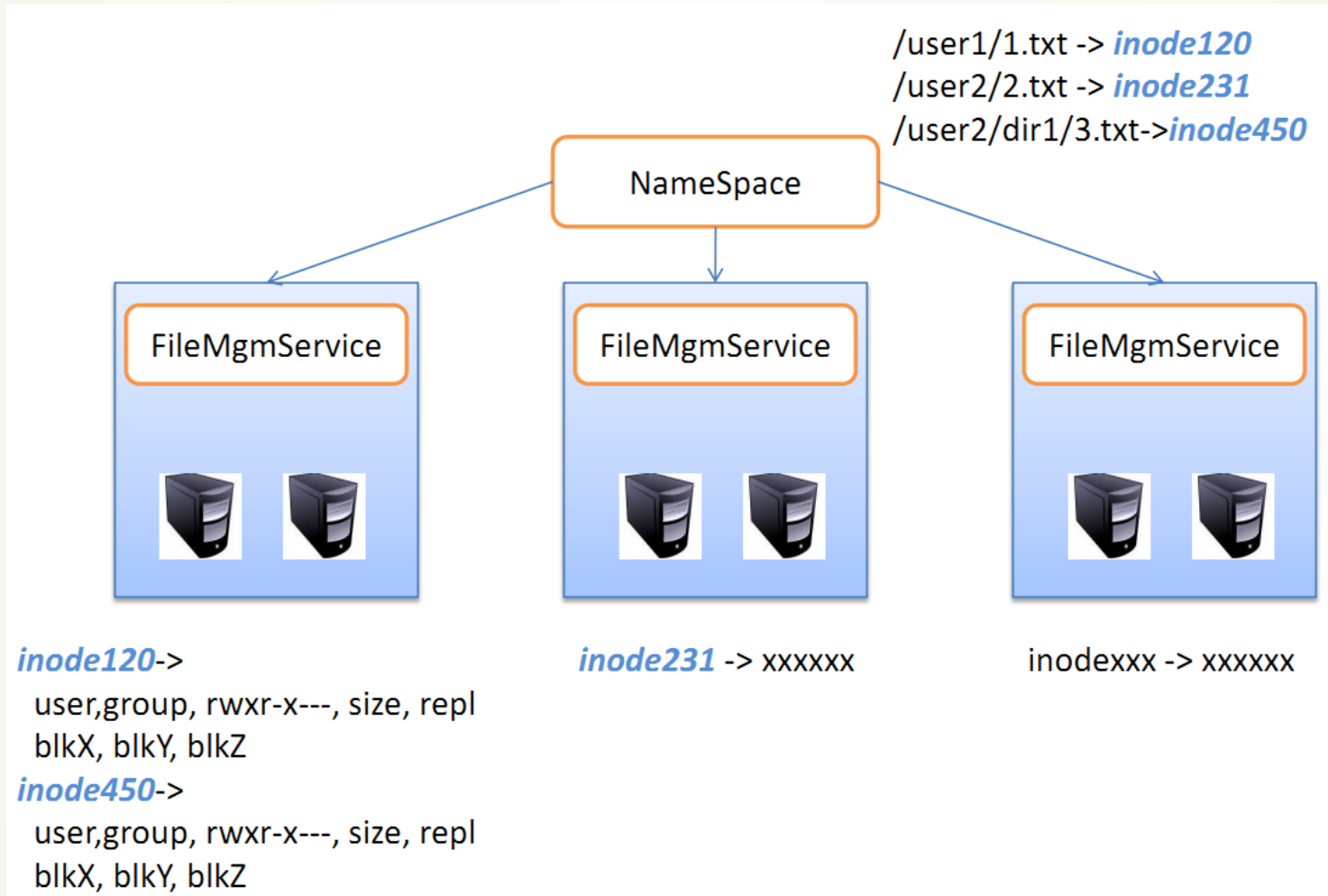
➤ 需求

- $10000 * 12 * 2T * 3\text{压缩比} / 256M / 3\text{副本} = 9.8\text{亿}$

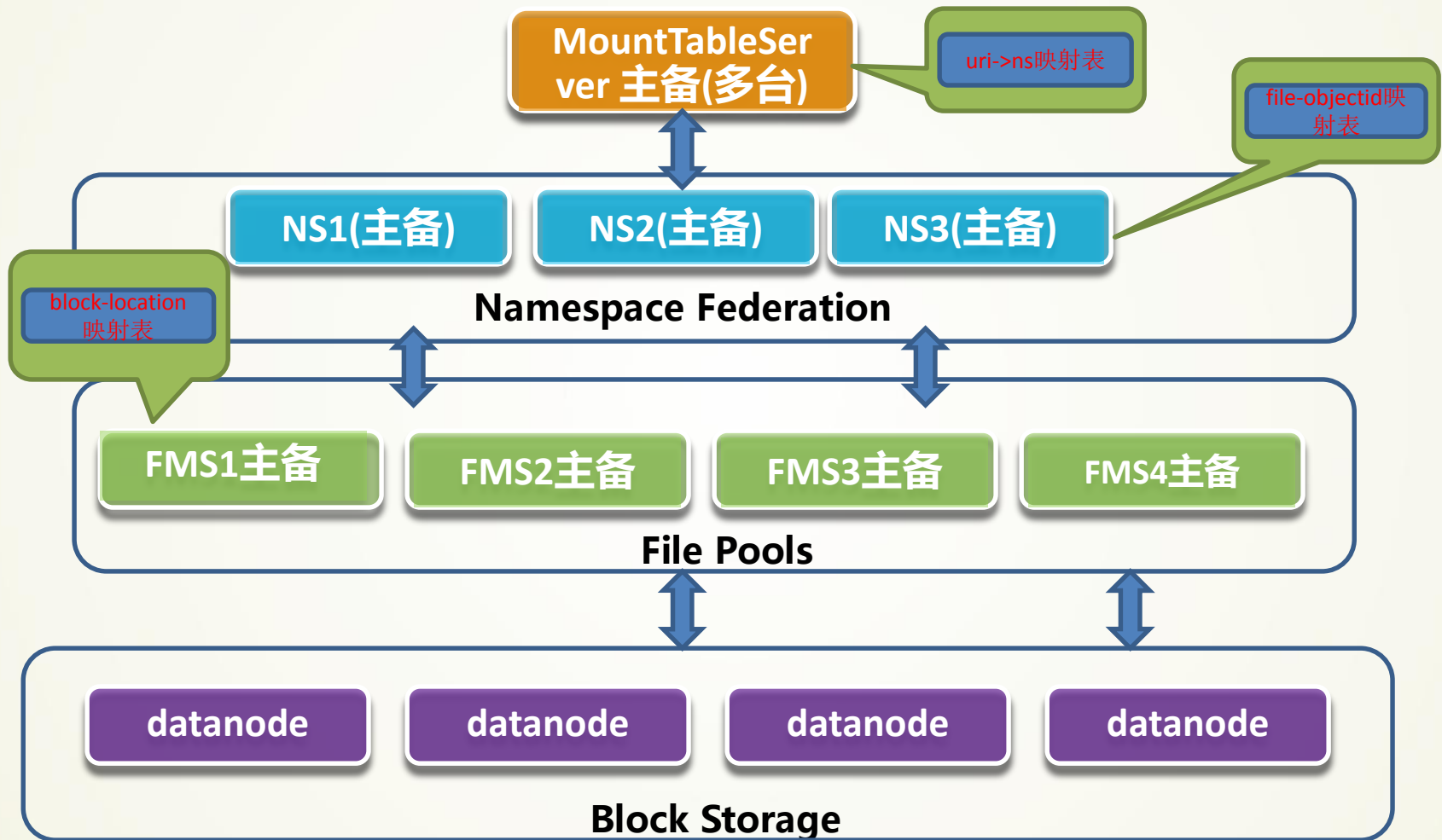
➤ 问题

- 内存 : 9.8亿文件 (file:block = 1:1) 占用内存 380G
- 负载 : 吞吐有限, latency 增加
- 稳定性 : GC影响
- 可用性 : 2 亿 inode重启一次花费 1小时左右

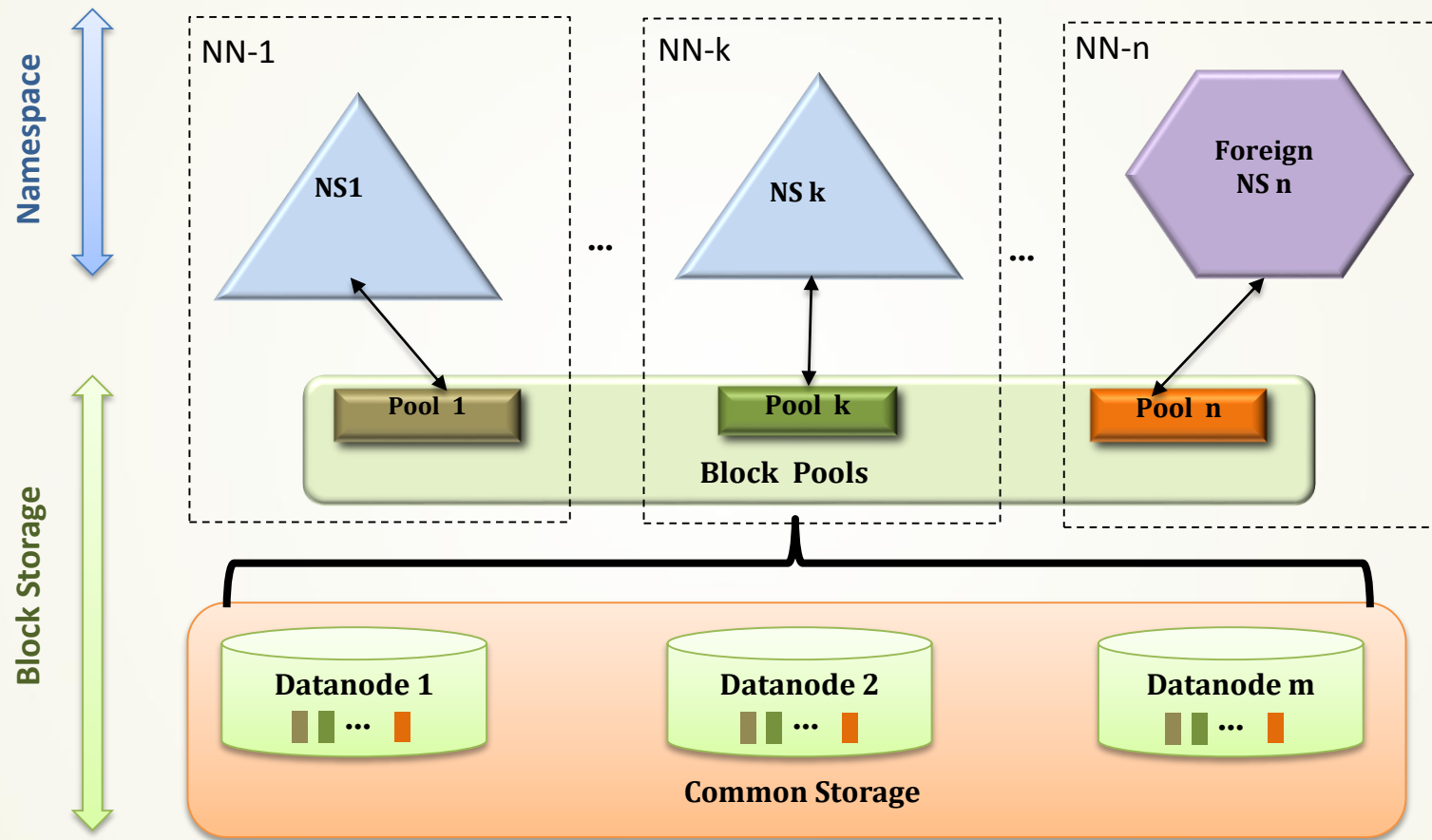
存储 – hdfs2全面上线 – 架构



存储 - hdfs3 - 架构



存储 – hdfs3 – 社区方案未来



存储 – 存储成本 – 思路

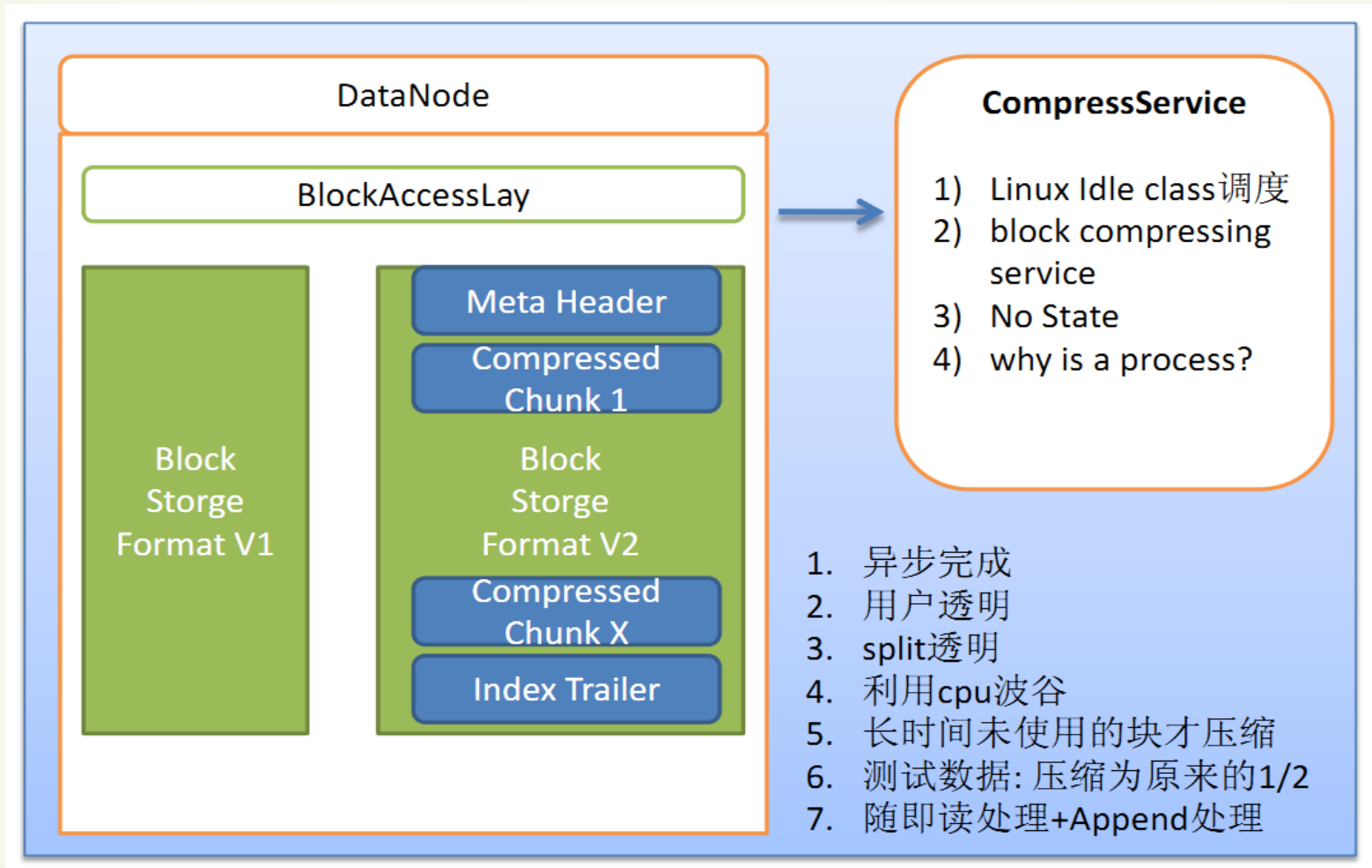
➤ 统计数据

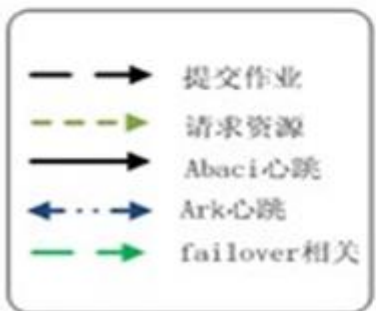
- 62% 的数据在一个月內没有访问
- 79% 的数据两周内没有访问
- 49% 的数据没有被压缩

➤ 优化思路

- 压缩
 - 透明压缩
 - 列存储，提高压缩比和访问效率
- 降副本
 - RAID

存储 – 存储成本 – 透明压缩





计算 – mr2全面上线 – 收益

可扩展性

- 计算模型和资源管理分开
- 单集群1w+，并发运行task 16w

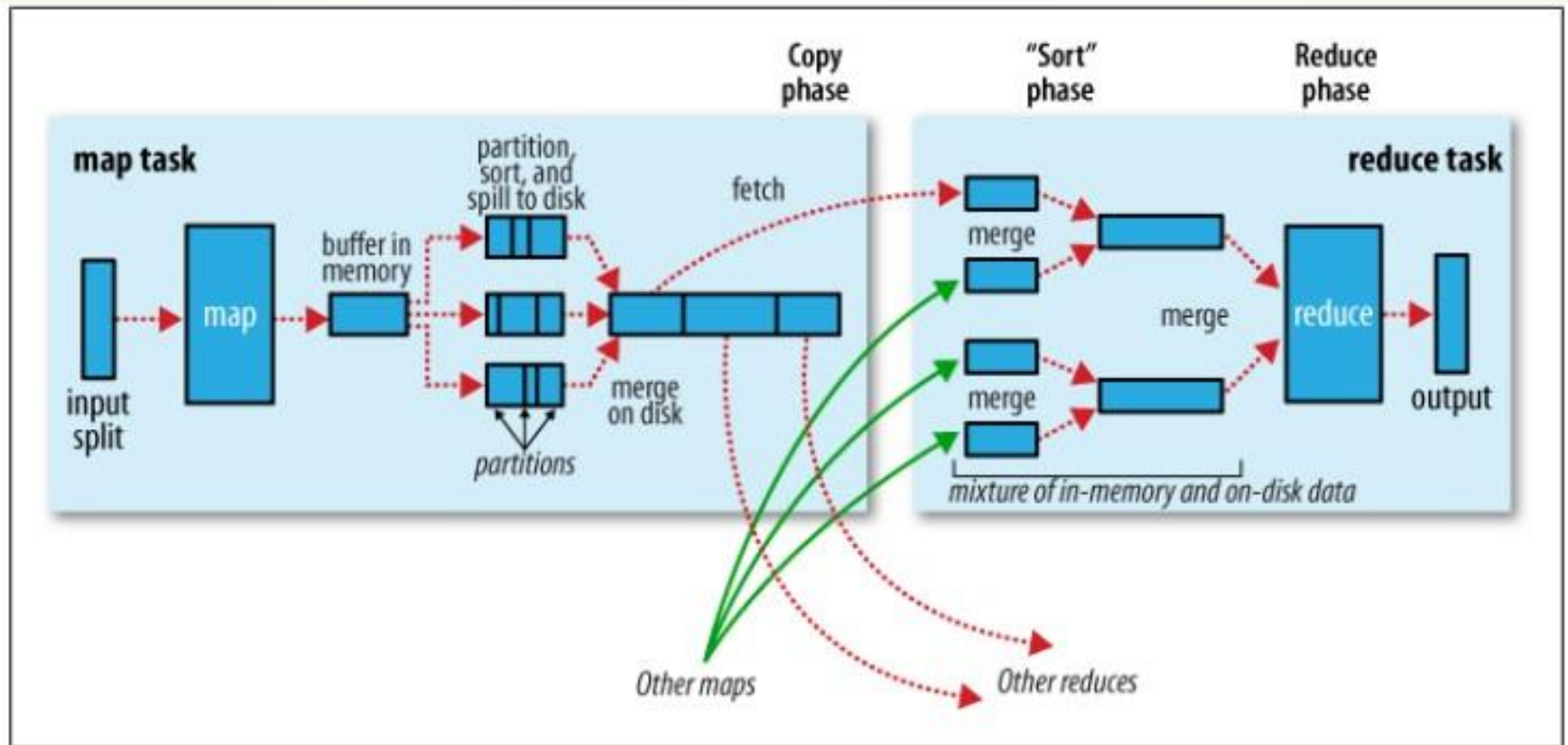
热升级

- MR计算模型升级，更新系统hdfs上abaci包
- 资源管理升级，可以正常查看提交作业

资源利用率提升

- (cpu, mem, disk, net) 多维资源描述
- Over-commit调度

计算 - mr执行优化 - 原理



计算 – mr执行优化 – 方案

Shuffle独立

- 尽量减少map/reduce之间的barrier
- 同时充分利用资源，减少资源浪费
- IO密集作业加速20%作业，资源利用提升6%

Map sort优化

- 优化map sort/spill过程，结合MAPREDUCE-64
- 并行sort，加速sort，减少block time
- 简单统计类应用map加速 30%

欢迎莅临

2013中国数据库技术大会

Database
BDaaS
flowingdata
DB2
NoSQL MySQL
Oracle Big Data