elastic 中国开发者大会 2023

腾讯云大数据 Elasticsearch 在日志领域的系统性优化

黄　华
腾讯云 Elasticsearch 内核负责人
2023/04/08

分享嘉宾

黄　华

Elasticsearch Top 100 Contributor

Elastic 中文社区主席团成员

- **日志领域的挑战**
- **高吞吐写入**
- **低成本存储**
- **高性能查询**

日志领域的挑战

半/非结构化

JSON

Dynamic mapping

Runtime fields

高吞吐写入

定向路由

物理复制

海量存储

压缩编码

混合存储

高性能查询

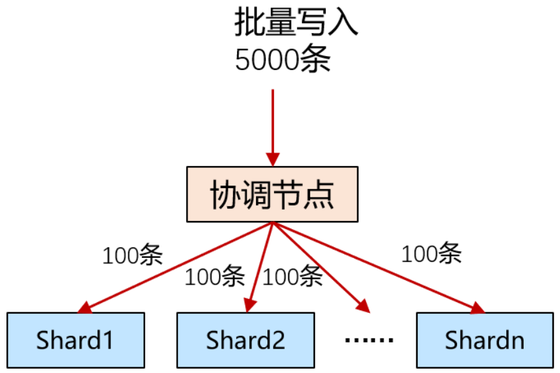
倒排索引

时序裁剪

并行化

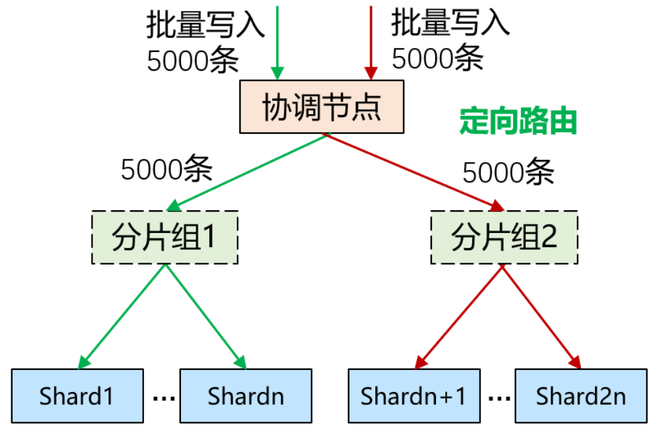
高吞吐写入 – 定向路由

问题



写入拒绝率高，资源利用率低

优化



分组定向路由，低扇出，容忍慢节点
不可靠的环境中，提供可靠的服务

原因

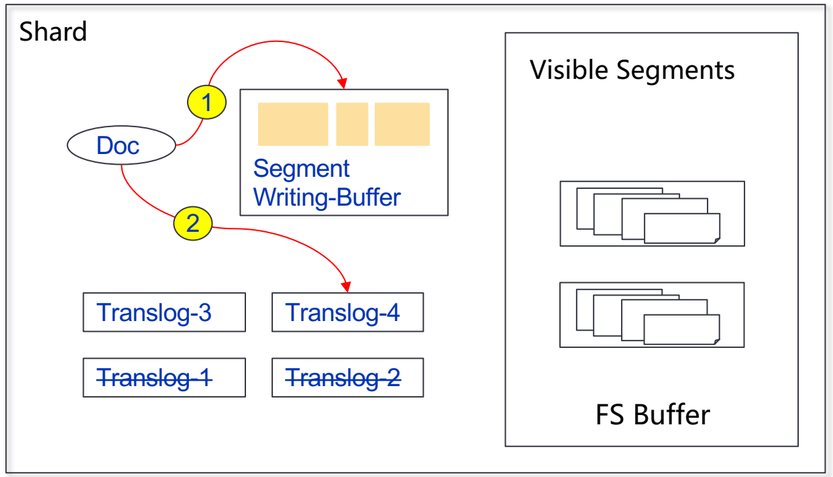
- 长尾节点：GC、慢盘、网络抖动
- 分布式写入：高扇出放大影响，写入堆积

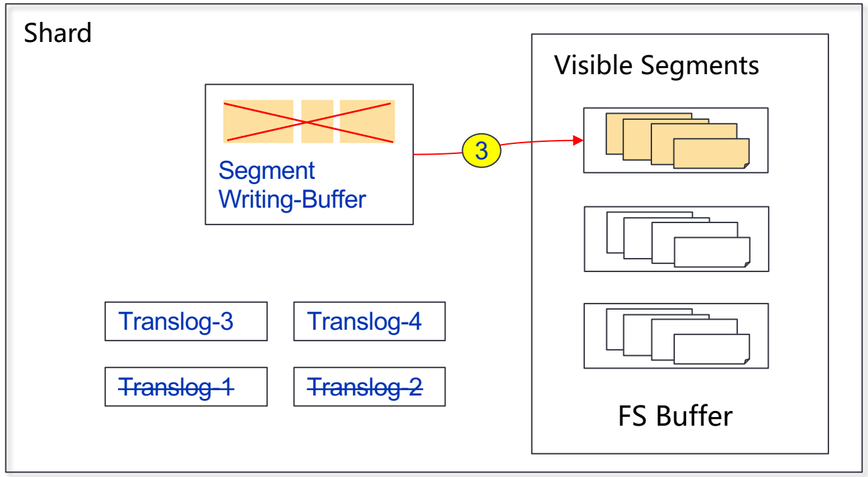
效果

	写入吞吐(TPS)	CPU	拒绝率
开源版本	51w	31%	3‰
定向路由	113w (+121%)	49% (+58%)	0‰

高吞吐写入 – 物理复制

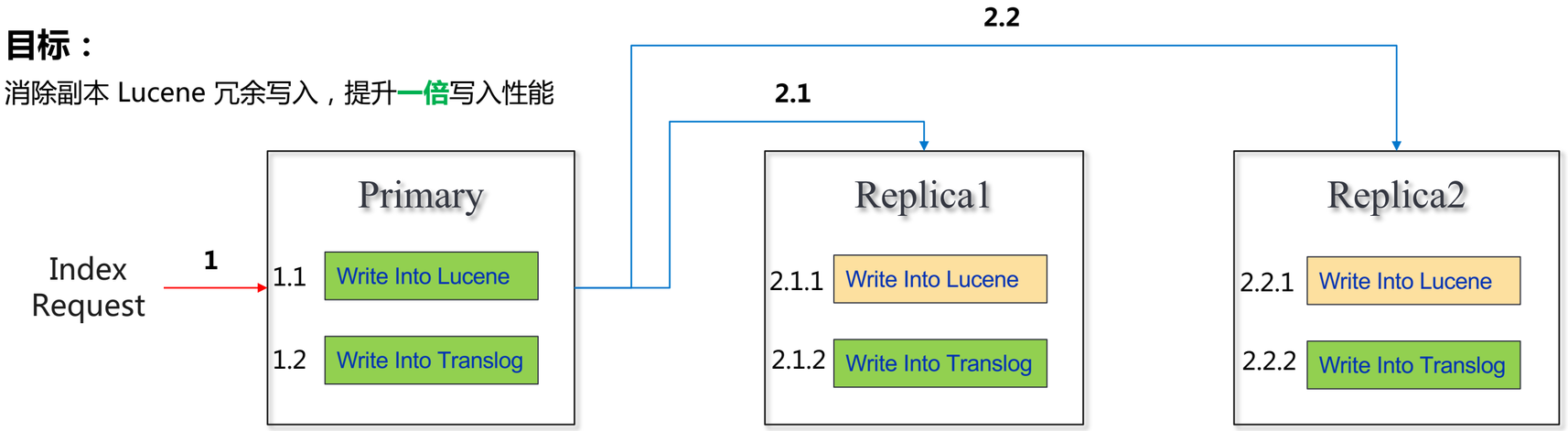
背景





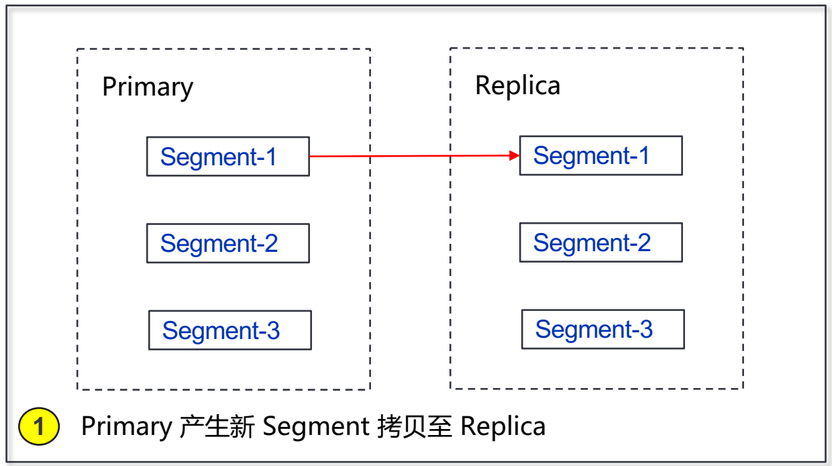
目标：

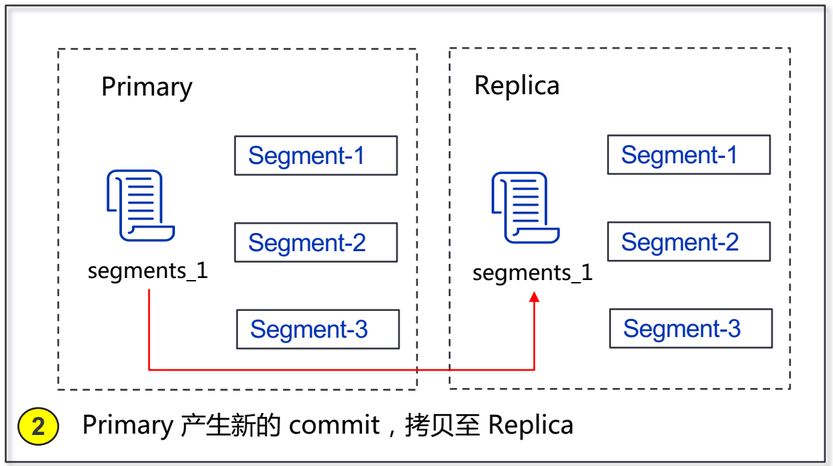
消除副本 Lucene 冗余写入，提升一倍写入性能

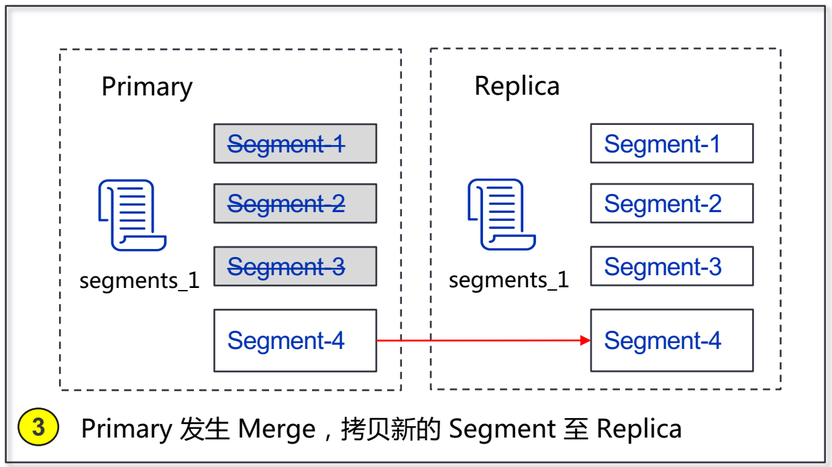


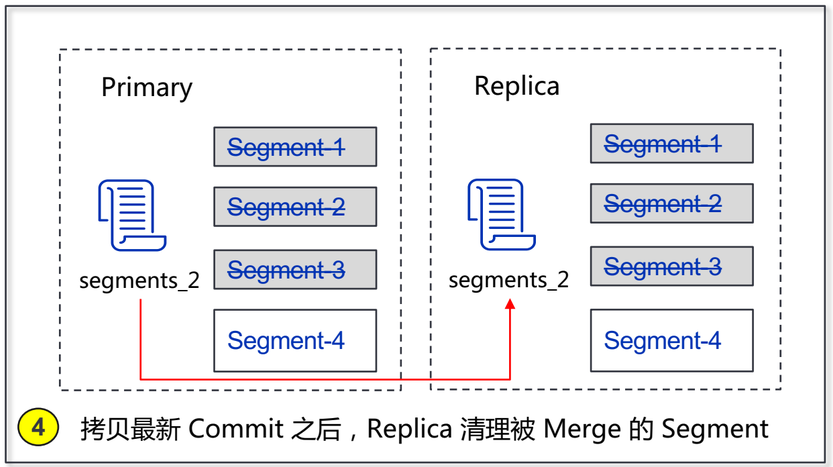
高吞吐写入 – 物理复制

方案









低成本存储

背景

分布式架构

云原生冗余副本

2 x 3

存储引擎

行、列、索引混存

单位 doc 存储放大

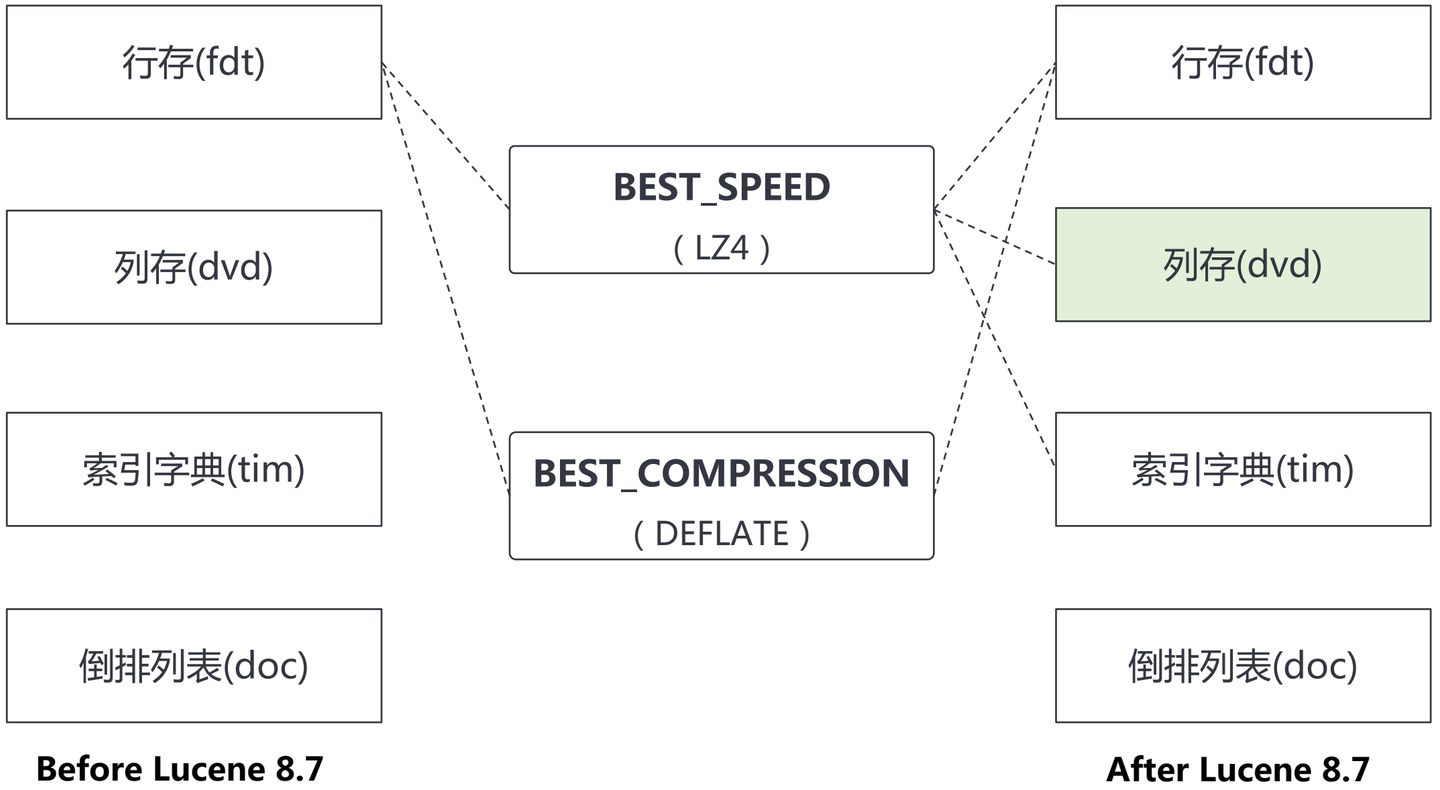
基础设施

SSD 成本、寿命

磁盘、机器运维成本

低成本存储 – 压缩编码优化

压缩文件类型扩展



Before Lucene 8.7

压缩算法仅针对行存数据

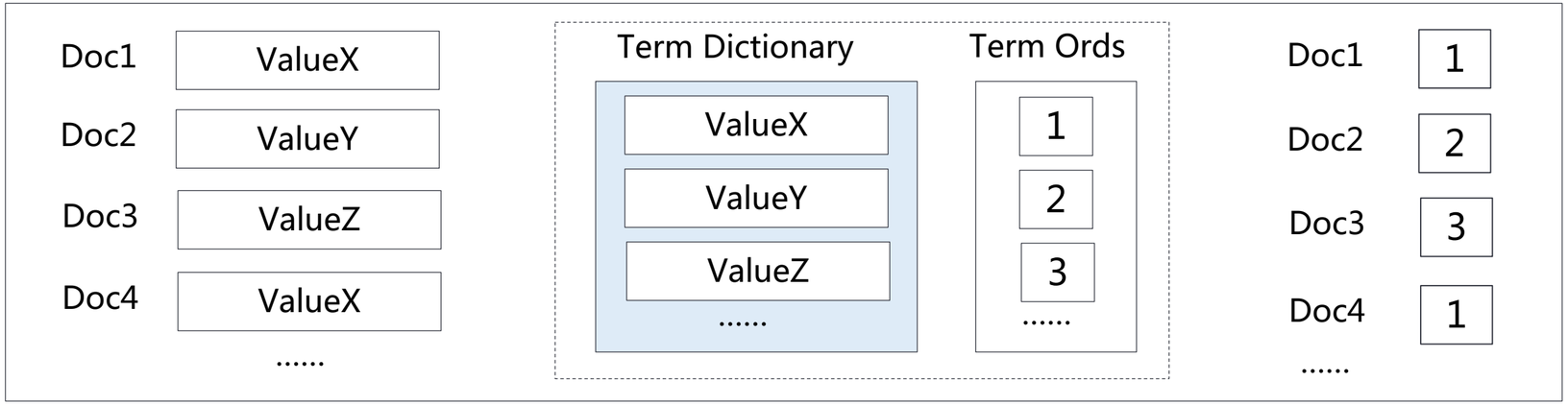
After Lucene 8.7

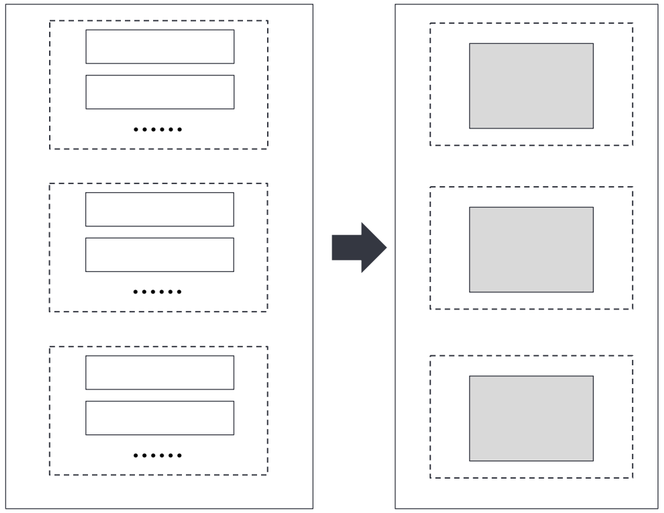
腾讯云贡献社区

压缩算法扩展至更多文件类型

低成本存储 – 压缩编码优化

SortedSet DocValue 列存压缩





Term Dictionary Compression

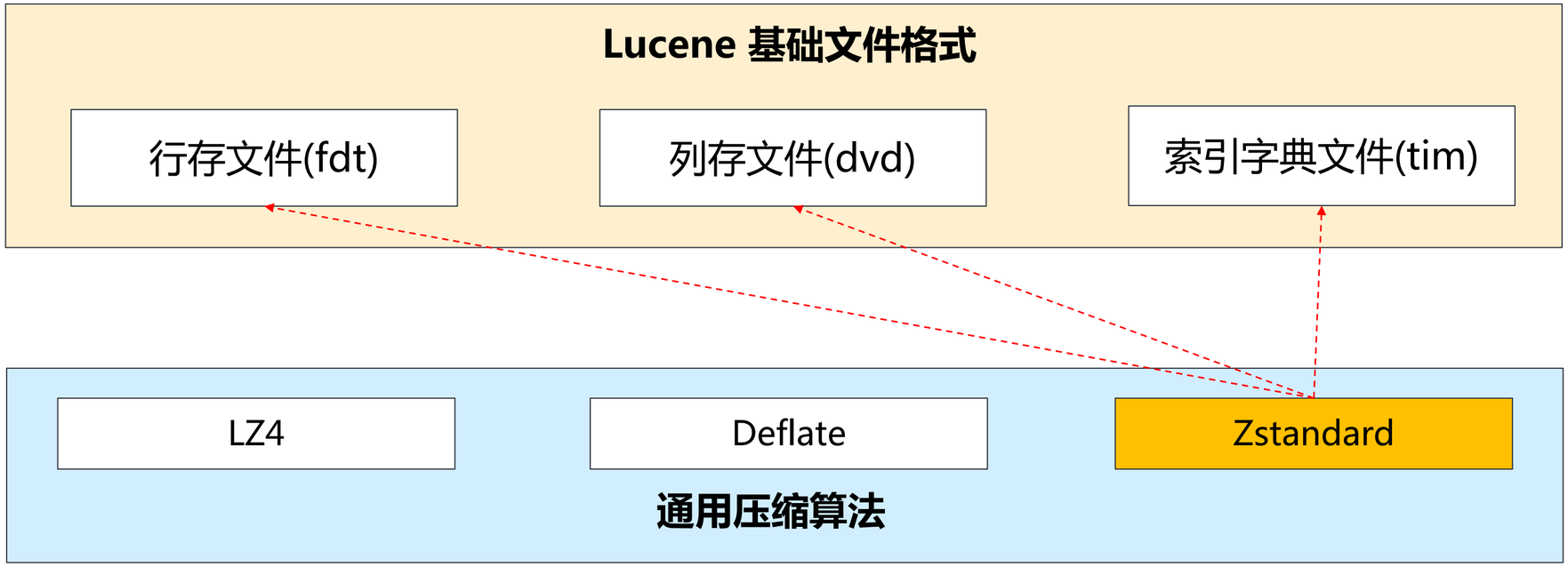
针对 High Cardinality 类型的 Field，启用压缩。

优化效果 已贡献社区 LUCENE-9663

压缩算法	优化前	优化后
写入耗时	591972 ms	618200 ms
Merge耗时	270661 ms	294663 ms
dvd文件大小	1.95 GB	**1.15** GB （↓40%+ ）

低成本存储 – 压缩编码优化

扩展通用压缩算法

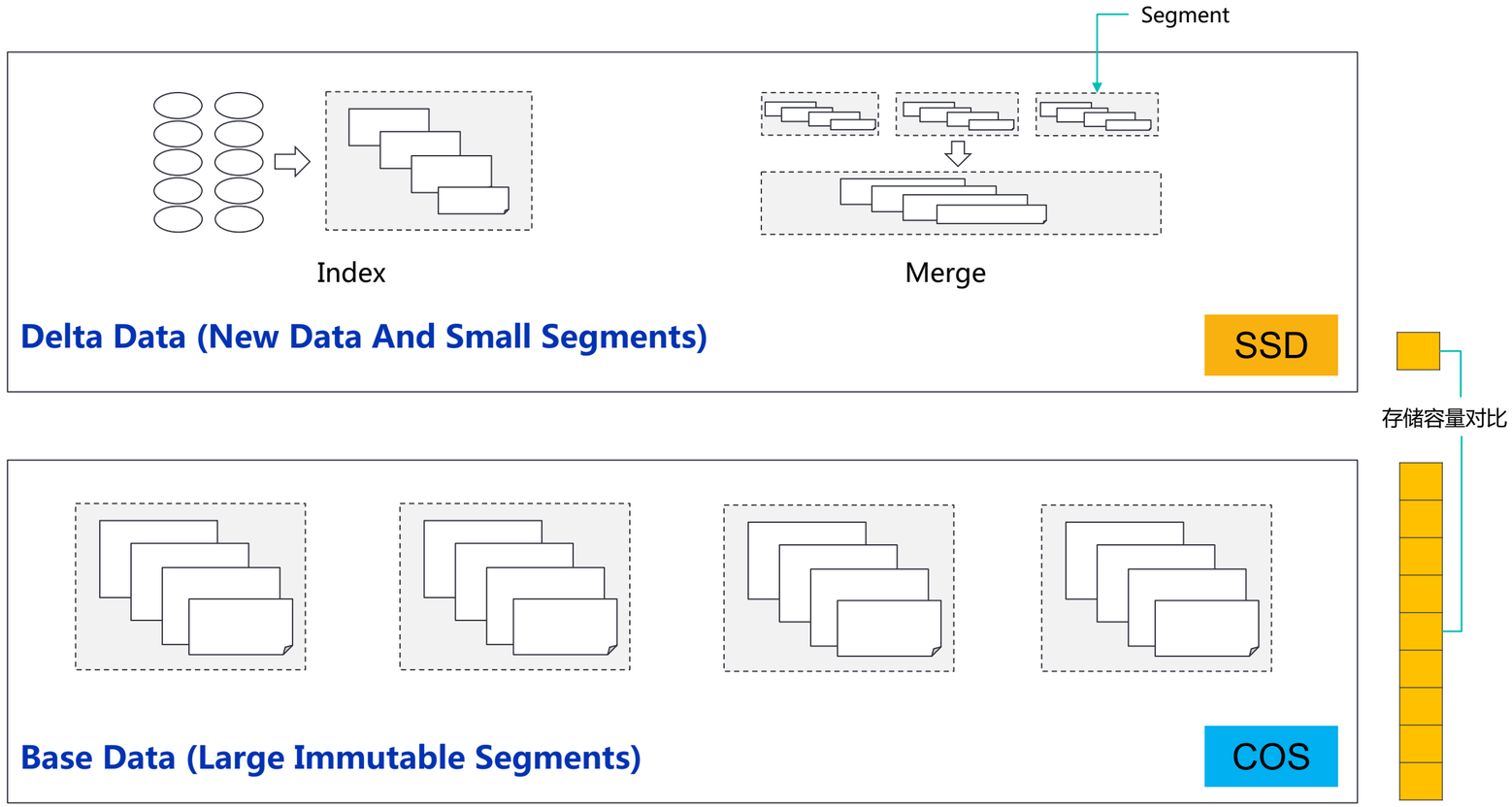


引入 Zstandard 高压缩比、高性能压缩算法，**整体压缩效果提升 30%+**

低成本存储 – 混合存储引擎

设计思路

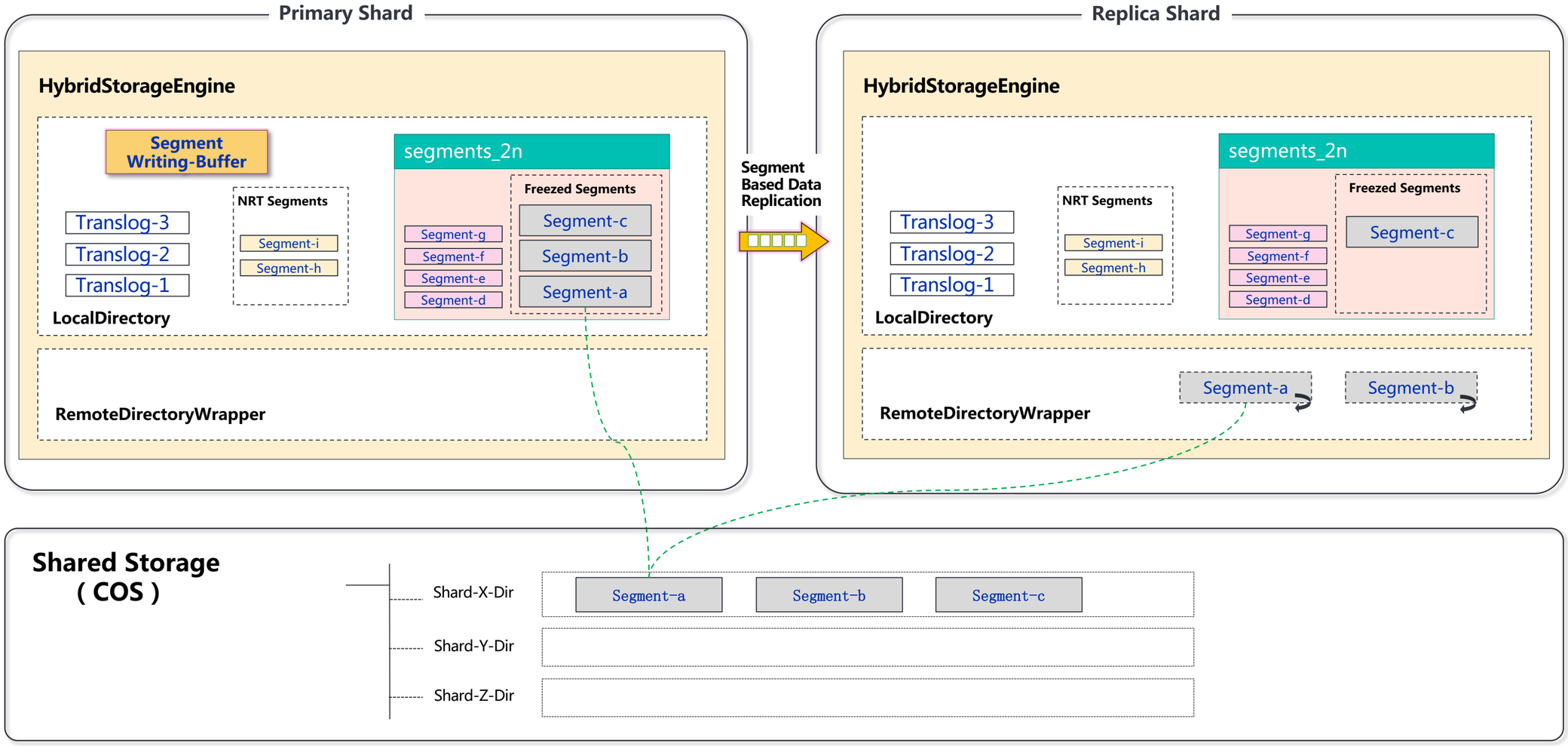
Delta(SSD) + Base(COS) 混合存储架构



COS：腾讯云对象存储，可用性 99.995%，可靠性12个9，支持标准、低频、归档等智能分层

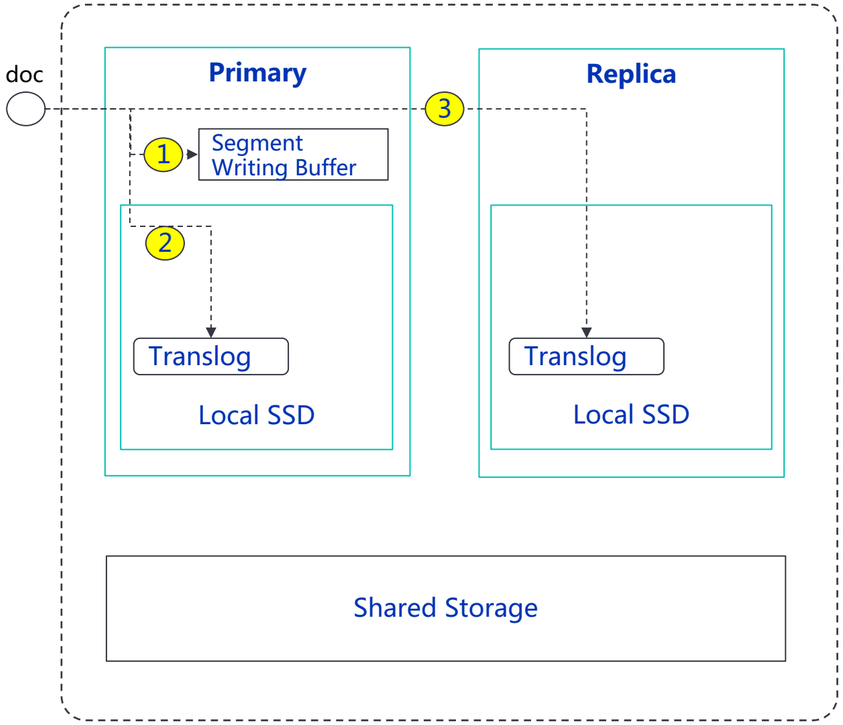
低成本存储 – 混合存储引擎

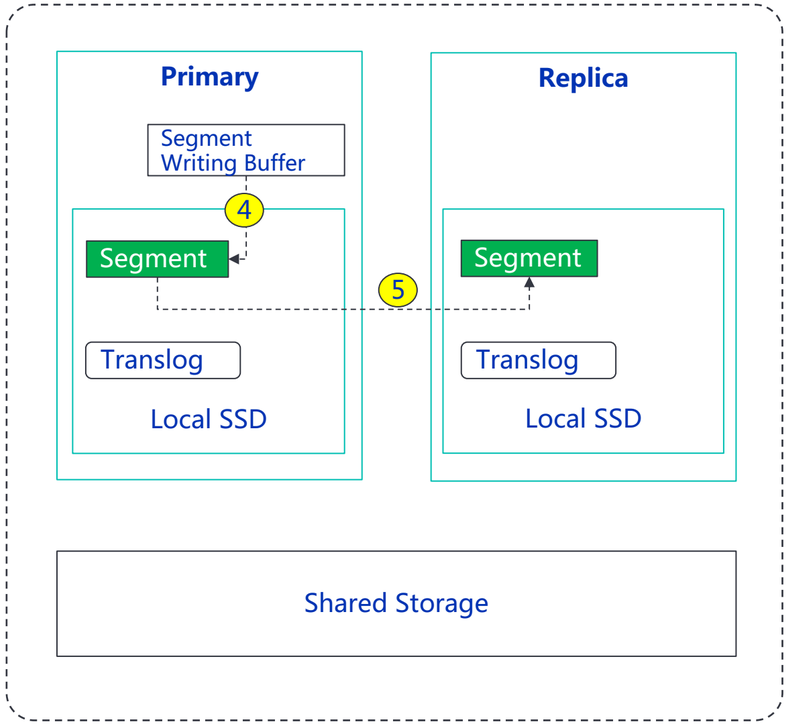
整体方案



低成本存储 – 混合存储引擎

设计流程1



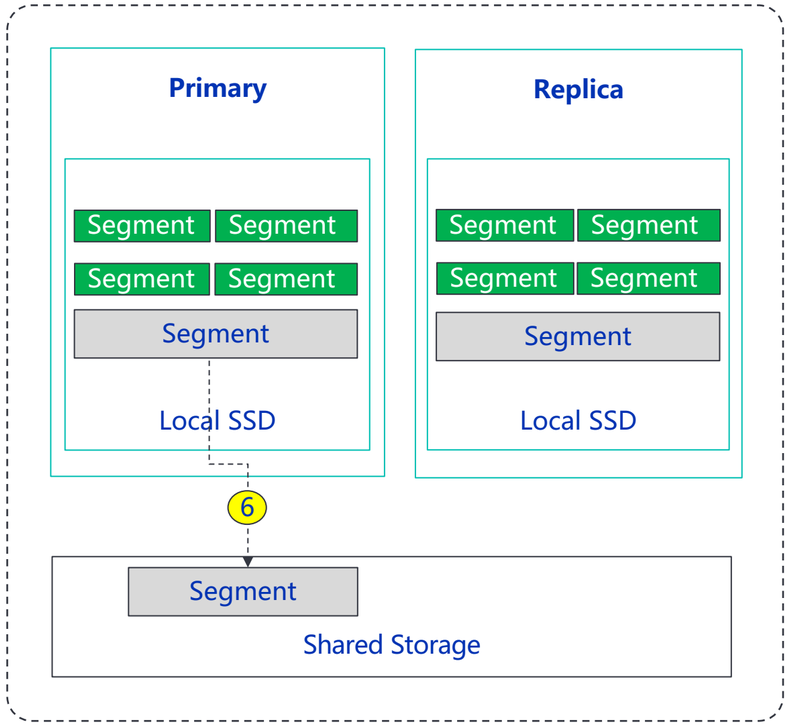


① Primary Segment 内存构建
② Primary Translog 写入
③ Replica Translog 写入 (**跳过 Segment 构建**)

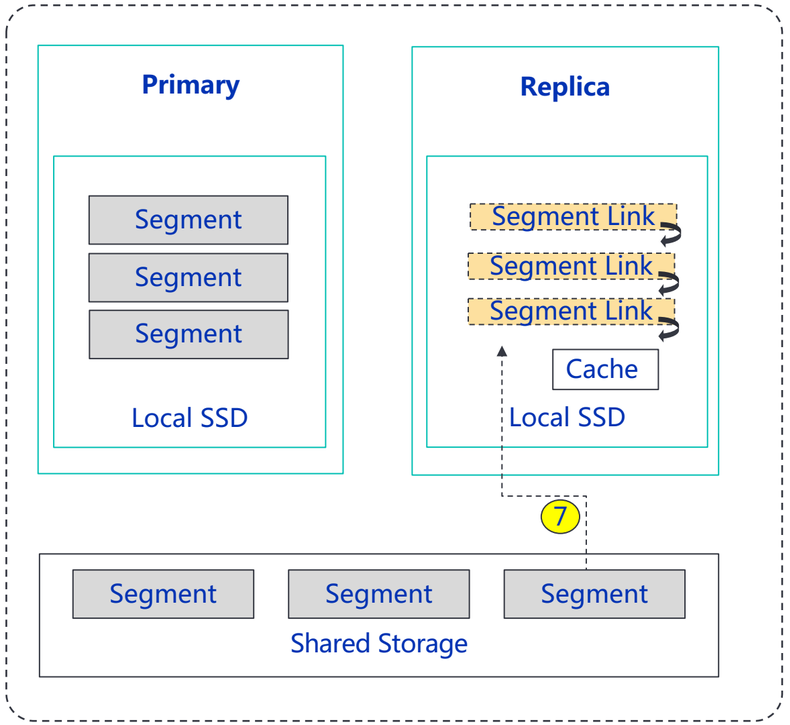
④ Primary Segment Flush
⑤ Primary **复制 Segment** 给 Replica

低成本存储 – 混合存储引擎

设计流程2



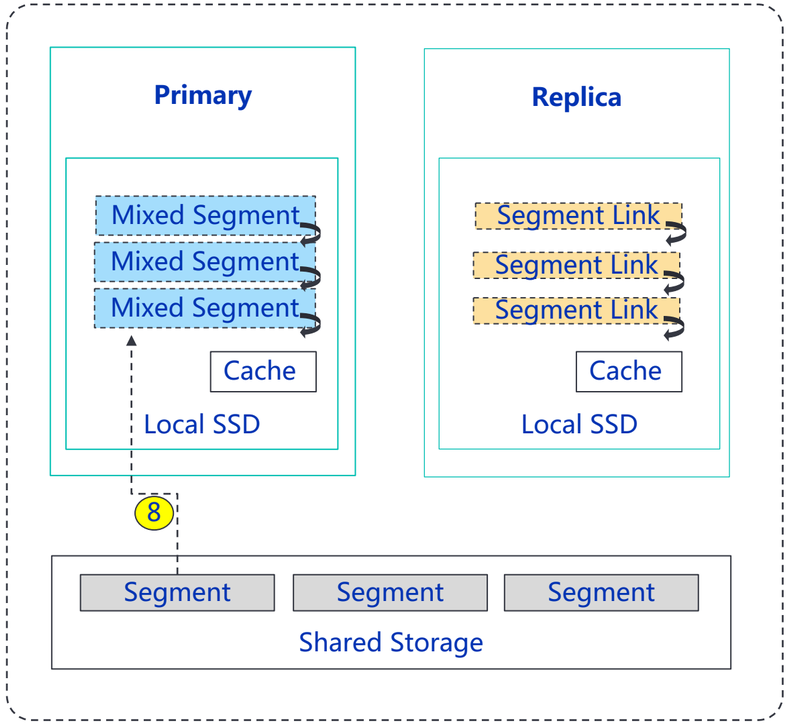
⑥ 冻结**大 Segment** 并下沉至共享存储



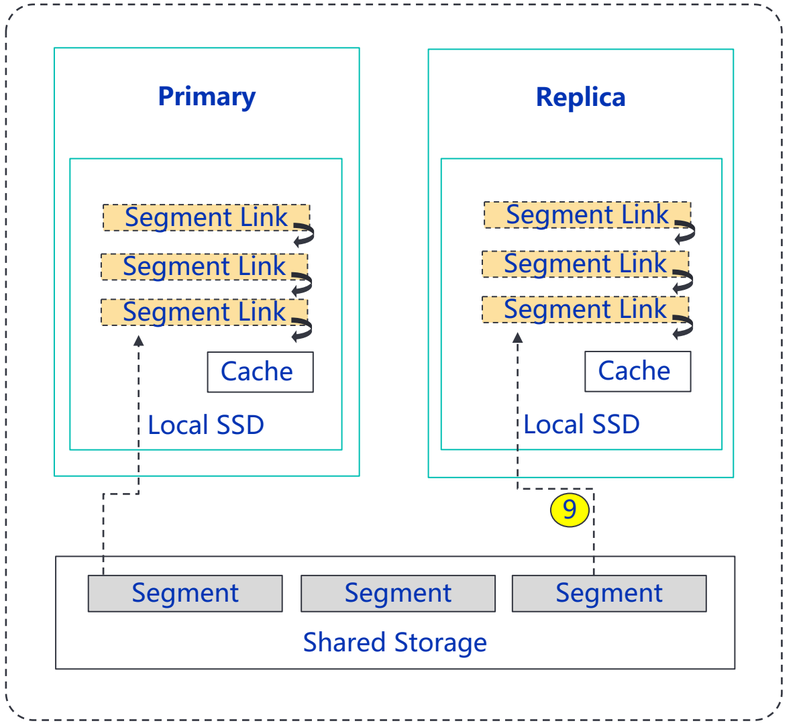
⑦ Replica 卸载，**缩减 ~50% SSD** 容量

低成本存储 – 混合存储引擎

设计流程3



⑧ Primary Segment 部分卸载，**缩减 ~70% SSD**



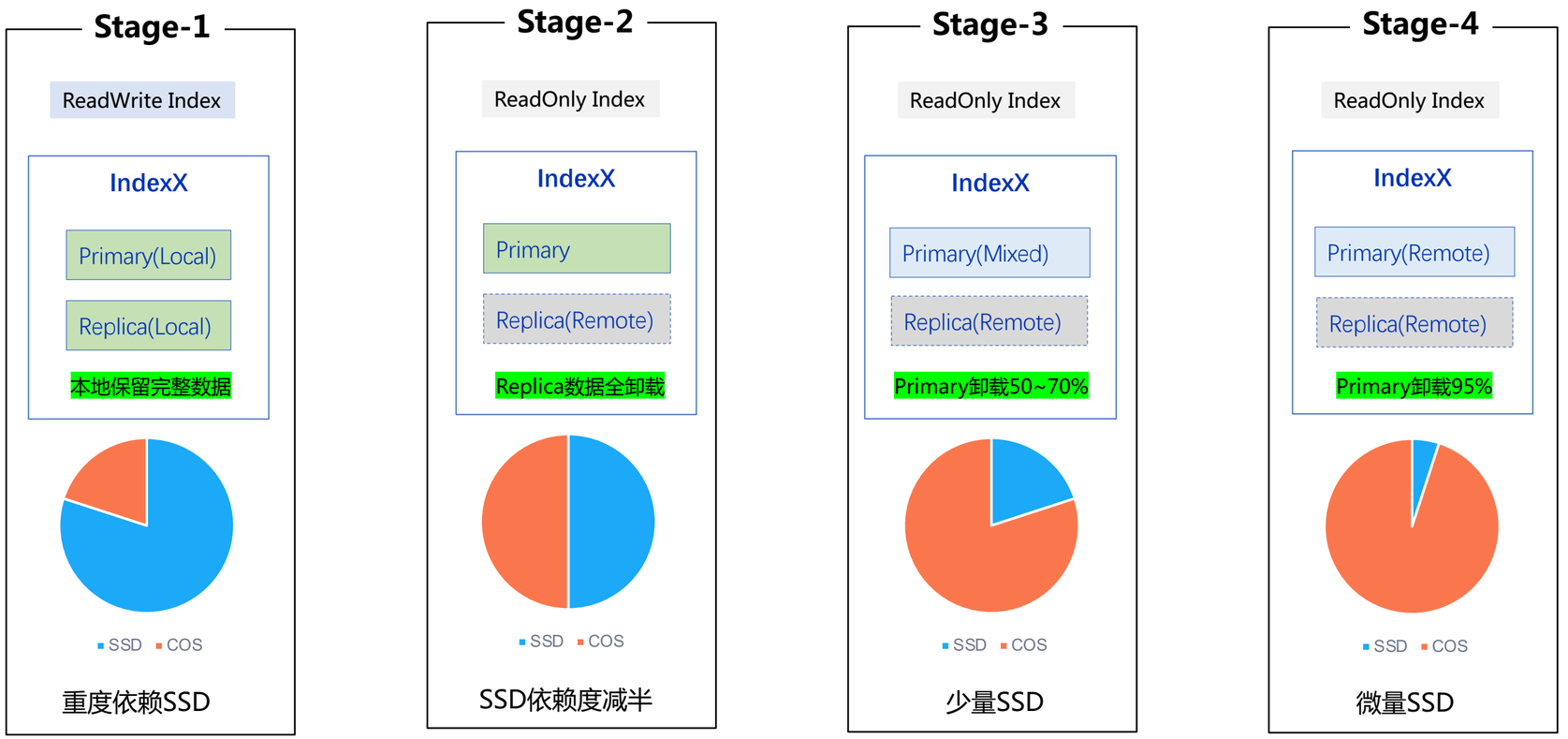
⑨ Primary Segment 完全卸载，**缩减 ~90% SSD**

低成本存储 – 混合存储引擎

Segment 三种形态

低成本存储 – 混合存储引擎

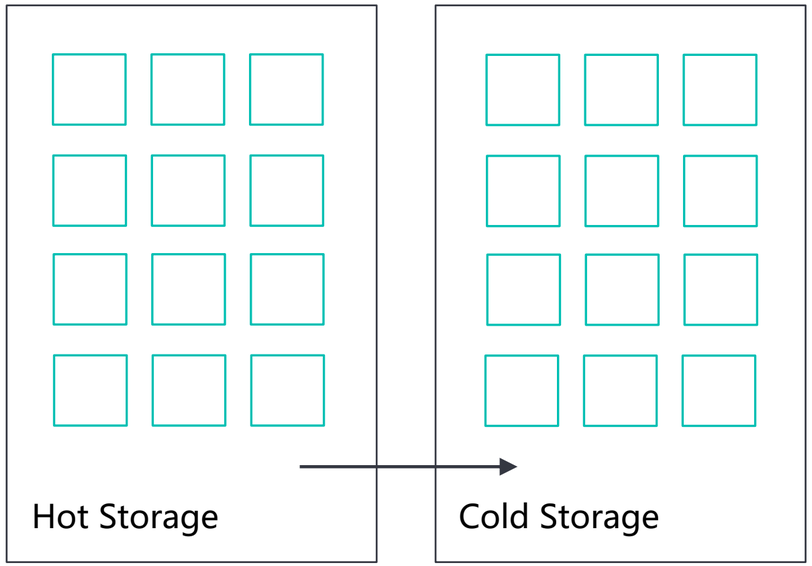
收益：存储重心逐步迁移，整体存储成本下降 50% - 80%



低成本存储 – 混合存储引擎

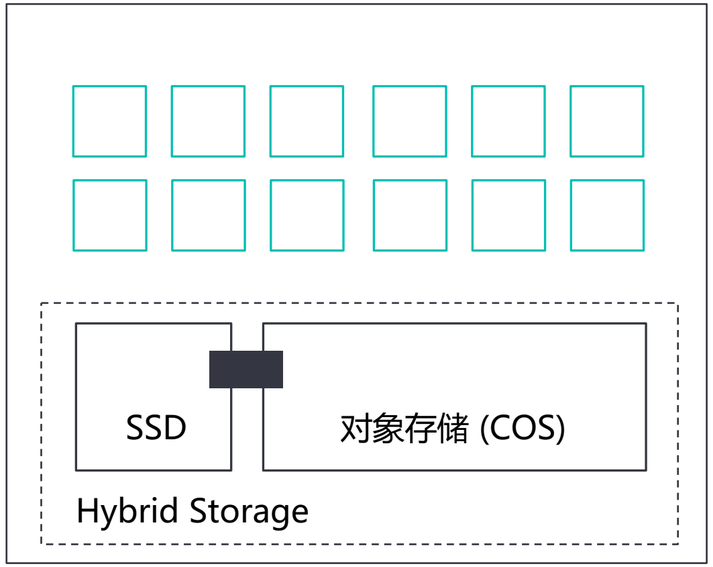
与现有方案区别

原生冷热分层方案



- 索引级别存储介质固定
- 冗余副本
- 数据降温后对整个索引进行搬迁
- 依赖用户的静态配置策略

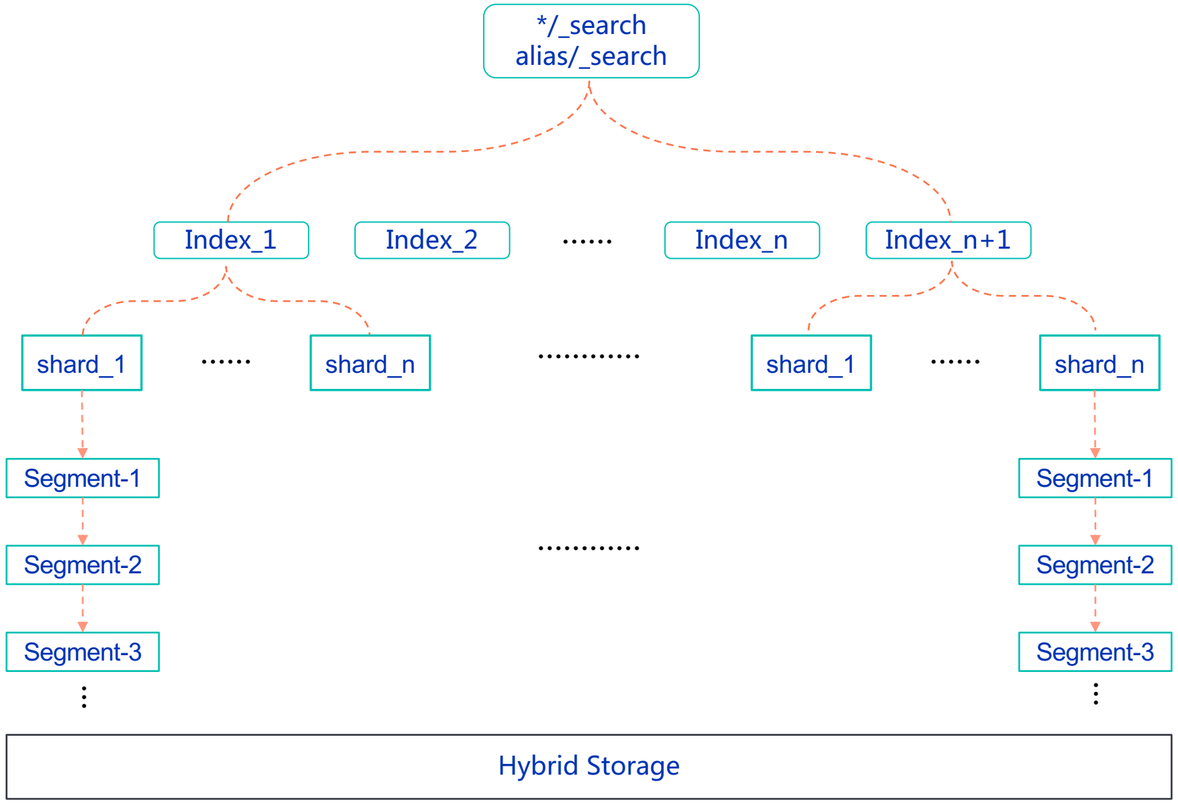
混合存储引擎



- 定制化存储引擎，天然支持了多种存储介质
- 逻辑分片，共享存储
- Segment 级别的热数据下沉
- 可根据用户访问统计信息自动决策数据卸载

高性能查询

背景



问题

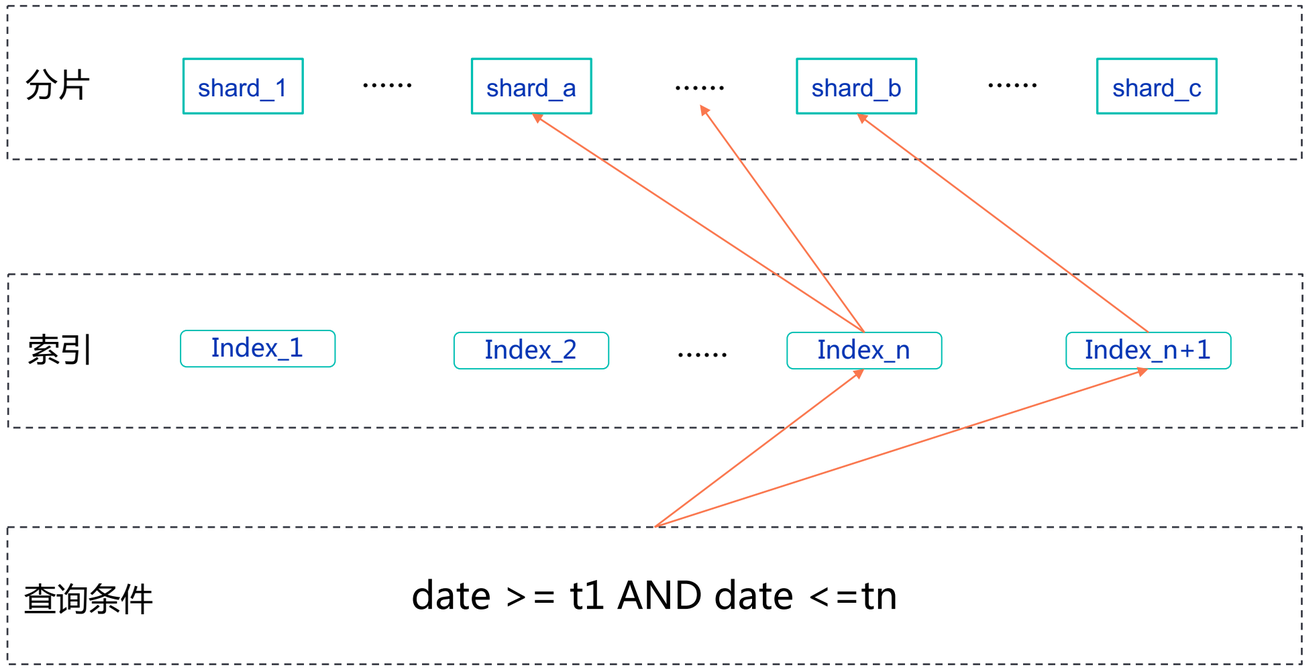
- 多数情况大量无用索引、分片、Segment 查询扫描
- Segment 串行化查询

优化思路

- 裁剪
- 并行化

高性能查询 – 索引分片级时序裁剪

优化方案



- 腾讯自研自治索引，托管索引分片管理
- 索引分片维度时序 Min/Max 索引维护
- 查询请求入口高效裁剪，避免无用扫描

优化效果：内部某业务查询性能提升8倍

高性能查询 – Segment 级查询裁剪

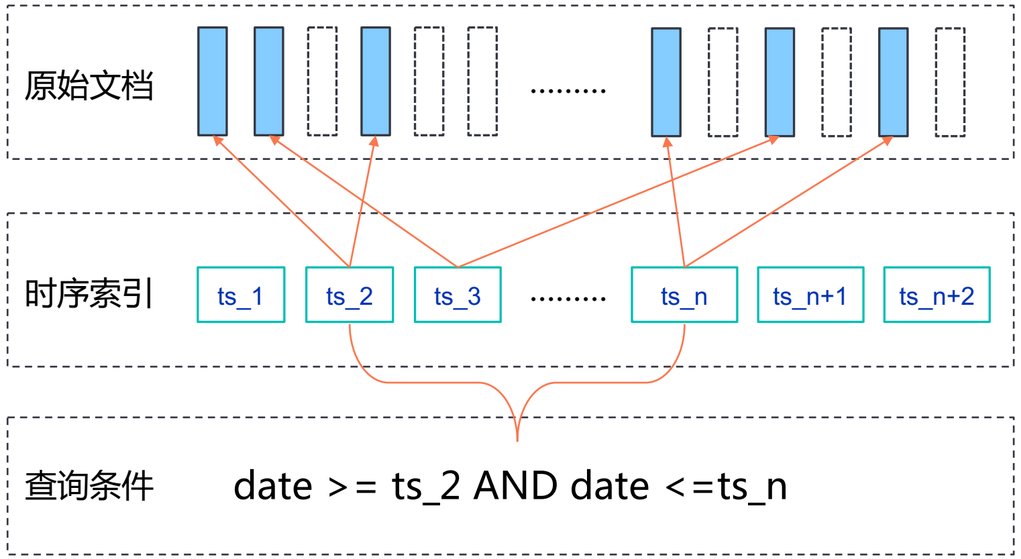
Segment 整体裁剪

Terms min/max 裁剪 Segment：查询提升 25.7%，已贡献社区 LUCENE-8980

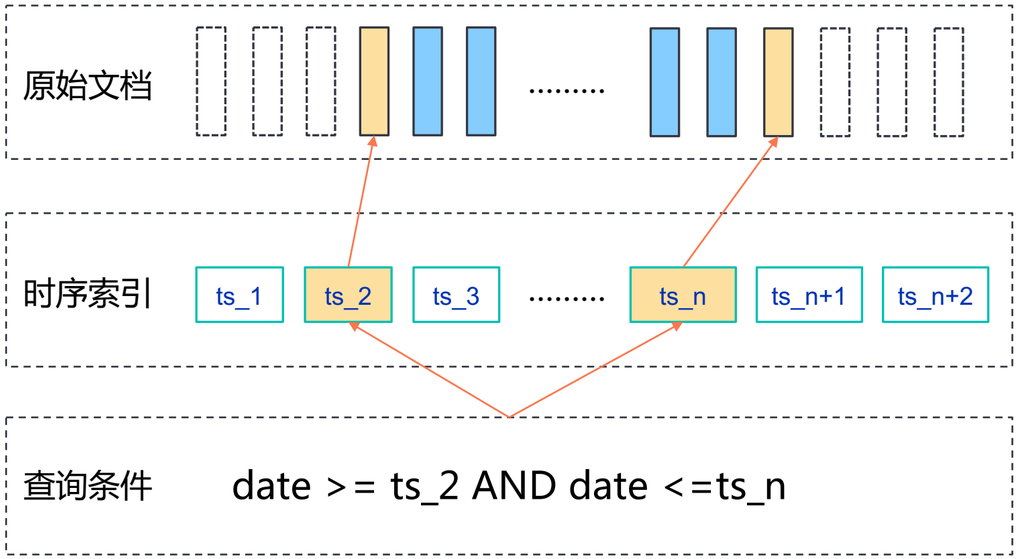
Segment 内数据裁剪

流式聚合裁剪：start doc 跳转 + early termination，聚合性能提升7倍，已贡献社区 ES-48399

时序裁剪：index sorting + 二级索引 + 逆序二分查找，时序搜索性能提升40倍，已发表顶会论文 VLDB 2022



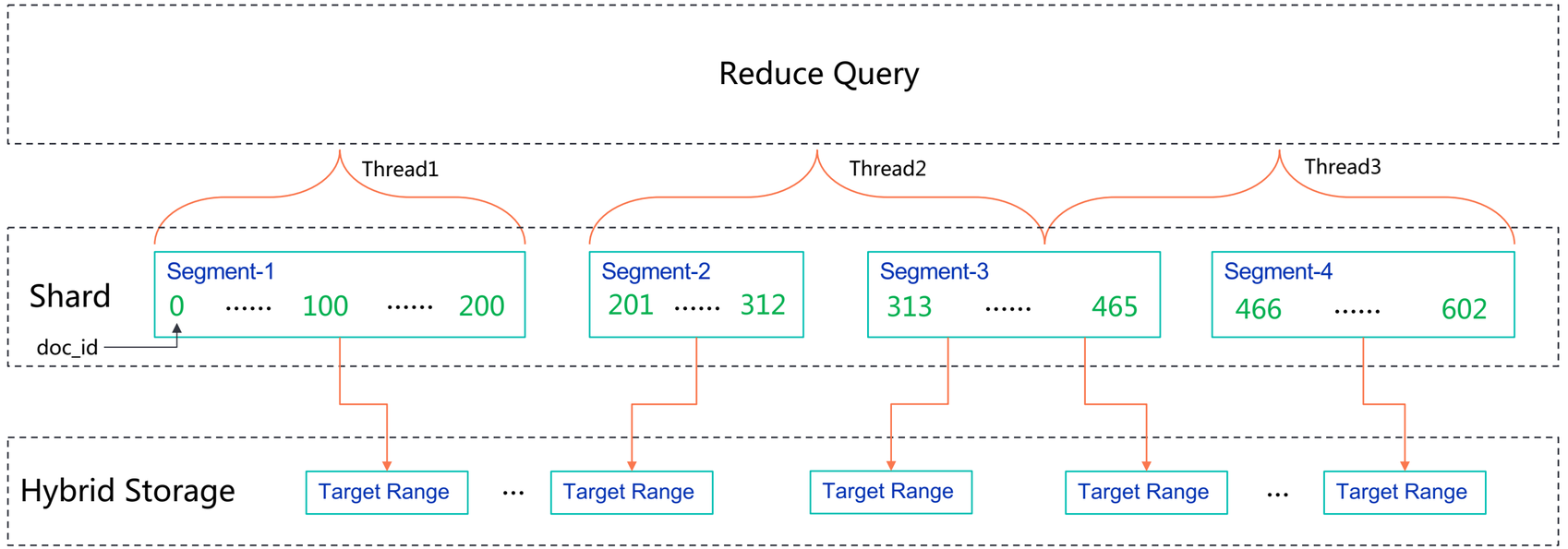
优化前大量随机 IO 遍历



优化后边界点裁剪

高性能查询 – Segment 并行化

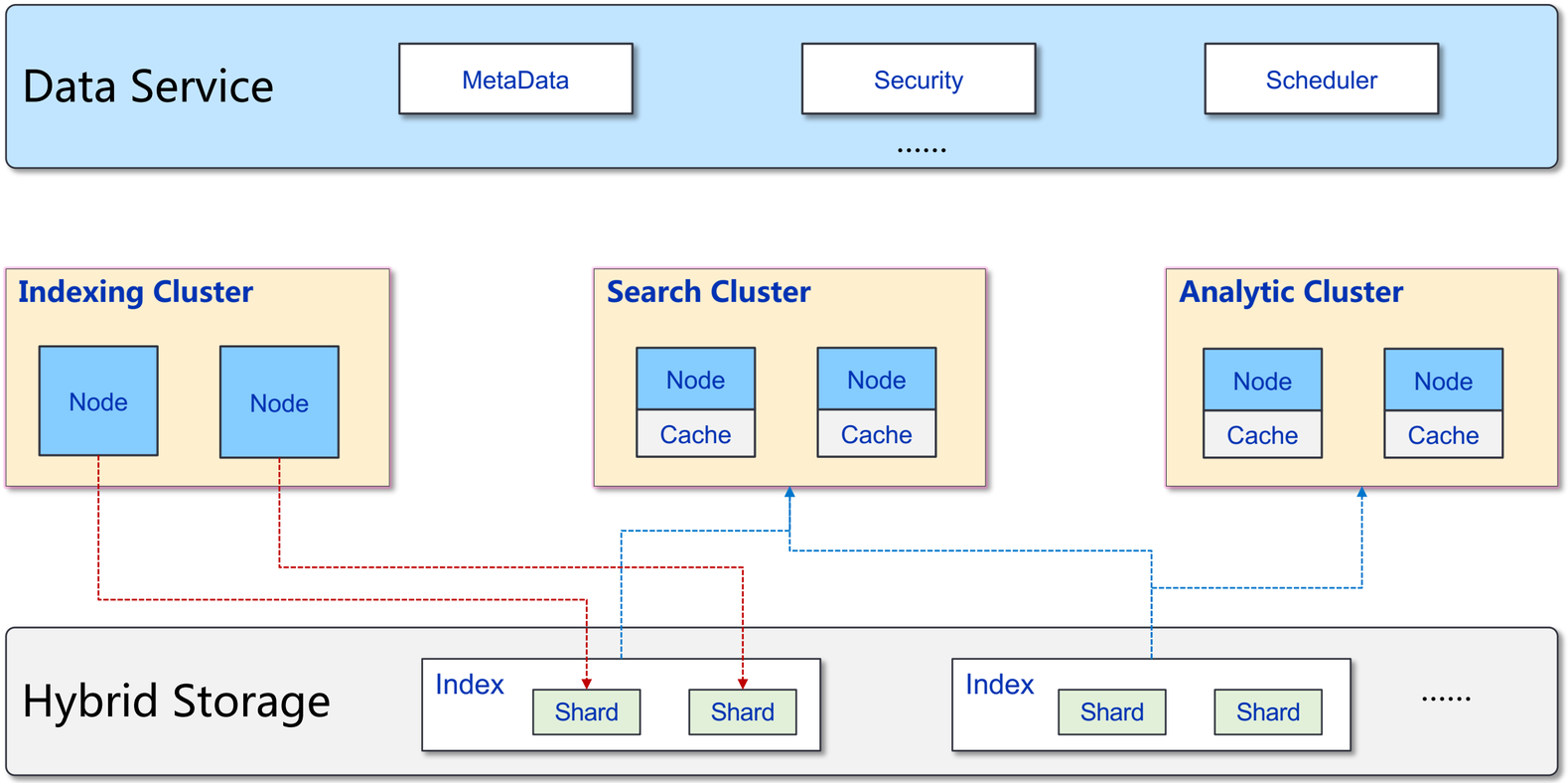
优化方案



- 多线程文档切分
- 多 Segment 并行化处理
- 精准拉取文件数据段

优化效果：查询性能提升**5**倍

云原生数据平台



腾讯云大数据 ES 服务：闭环 PB 级检索、分析

- **存算分离：**共享存储，消除副本
- **读写分离：**资源隔离，业务隔离
- **资源调度：**数据跨节点、集群挂载

elastic 中国开发者大会 2023

Thanks

elastic
中文社区

专业、垂直、纯粹的 Elastic 开源技术交流社区

https://elasticsearch.cn/