



# Elastic 中文社区

## 深圳 Meetup

2019/11/16

Saturday 13:00

合作伙伴



合作社区



# 阿里云Elasticsearch

## 内核优化与应用实践

阿里巴巴 欧阳楚才

# 阿里云Elasticsearch

全托管式的Elasticsearch服务，100%兼容开源，针对性优化内核性能，提供安全功能，支持多租户，高可用服务，弹性伸缩，支持公有云和专有云部署

## 目标场景



# 公有云业务规模

**17**

区域数

**4000+**

集群数量

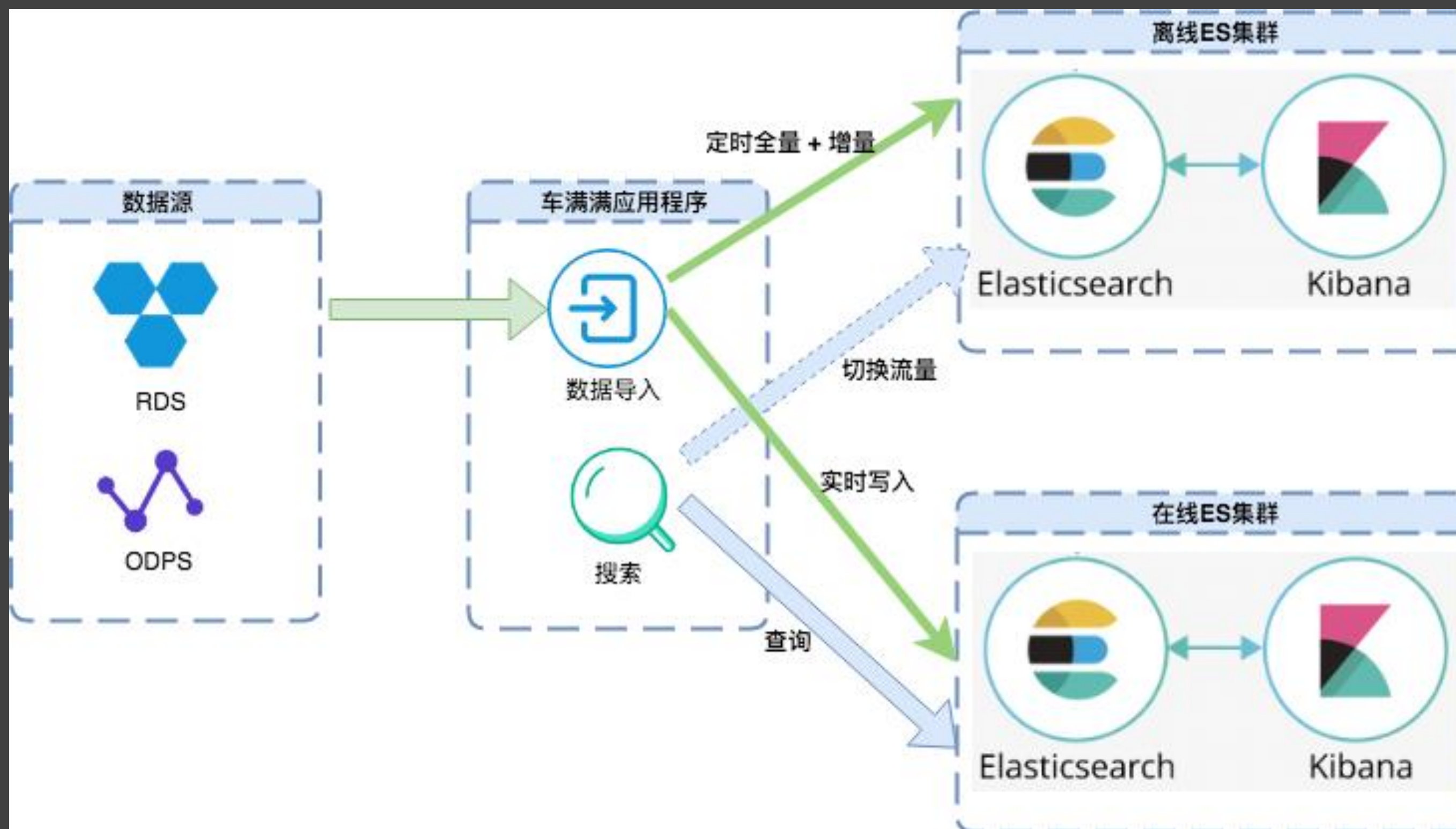
**5PB+**

数据量

# 索引写入痛点

**痛点:** 因全量写入影响在线查询，始终运行着一套备集群，成本提高了一倍

**痛点:** 全量写入速度慢，大数据量场景下耗时过长



用户双集群架构

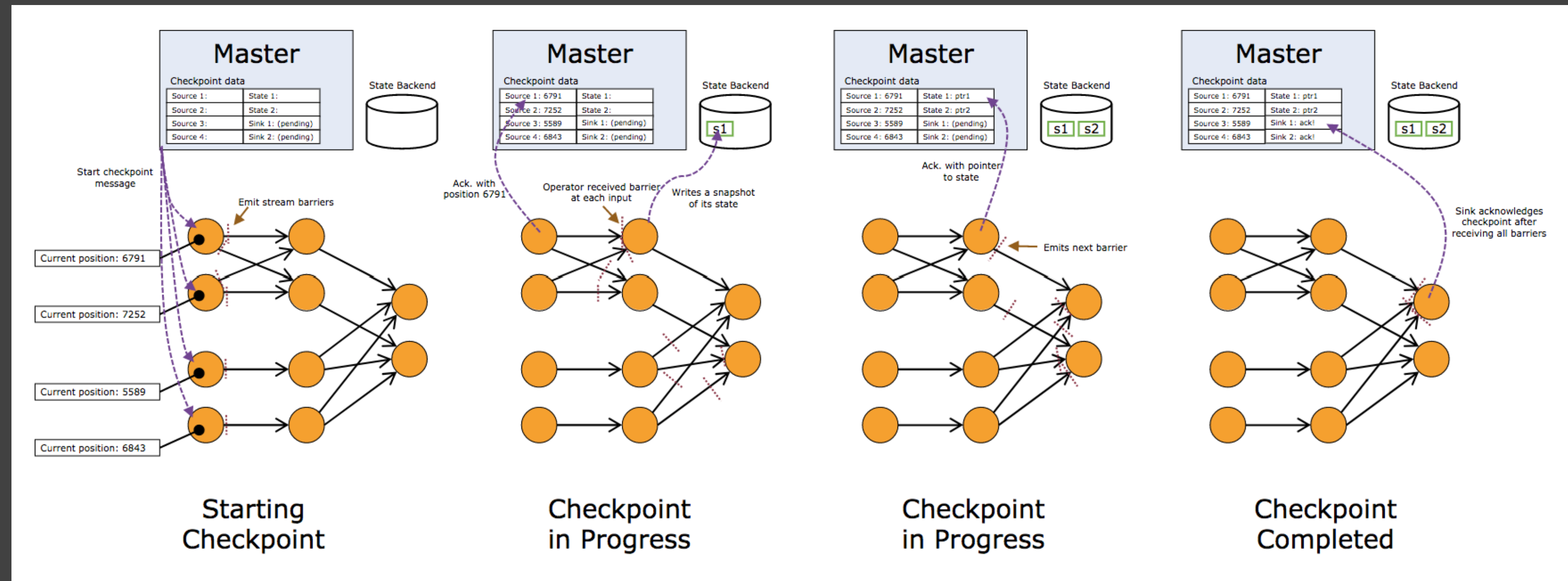
# ElasticBuild - 全量

Blink Checkpoints取代Translog

**优势:** 移除Translog后  
降低了近一倍的IO开销

**优势:** At Least Once保  
证数据准确

**优势:** 秒级Failover恢复



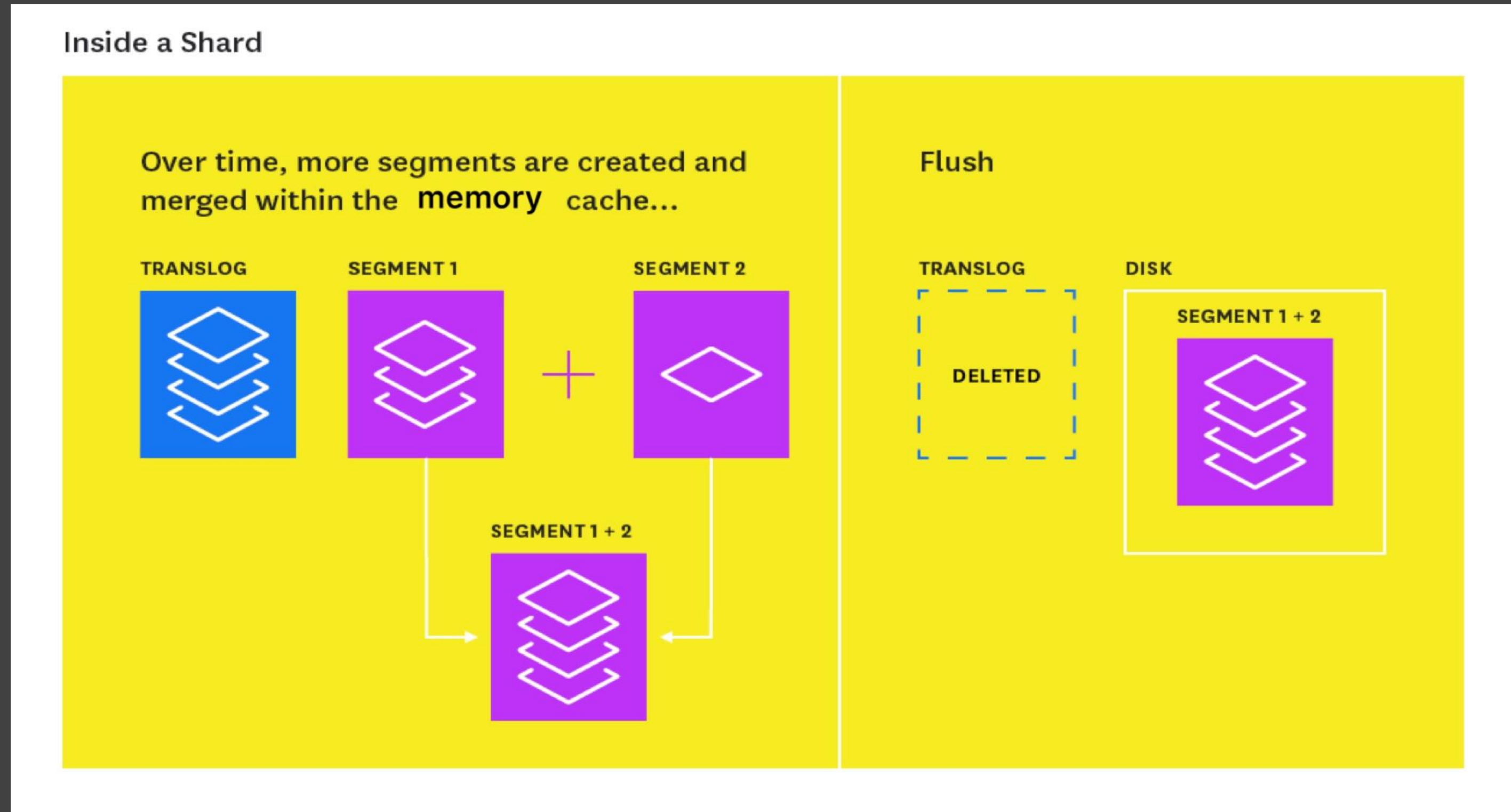
完整的Blink Checkpoints过程如上图所示

# ElasticBuild - 全量

## 内存索引合并

**瓶颈:** Segment写入磁盘后, 又会被反复读到内存中合并, 接着又写入磁盘。这个过程会有大量重复IO开销

**方案:** 将索引合并过程放在内存中 如图, 多个segment会在内存中达到一定大小后, 才Flush到磁盘, 从而避免了IO开销



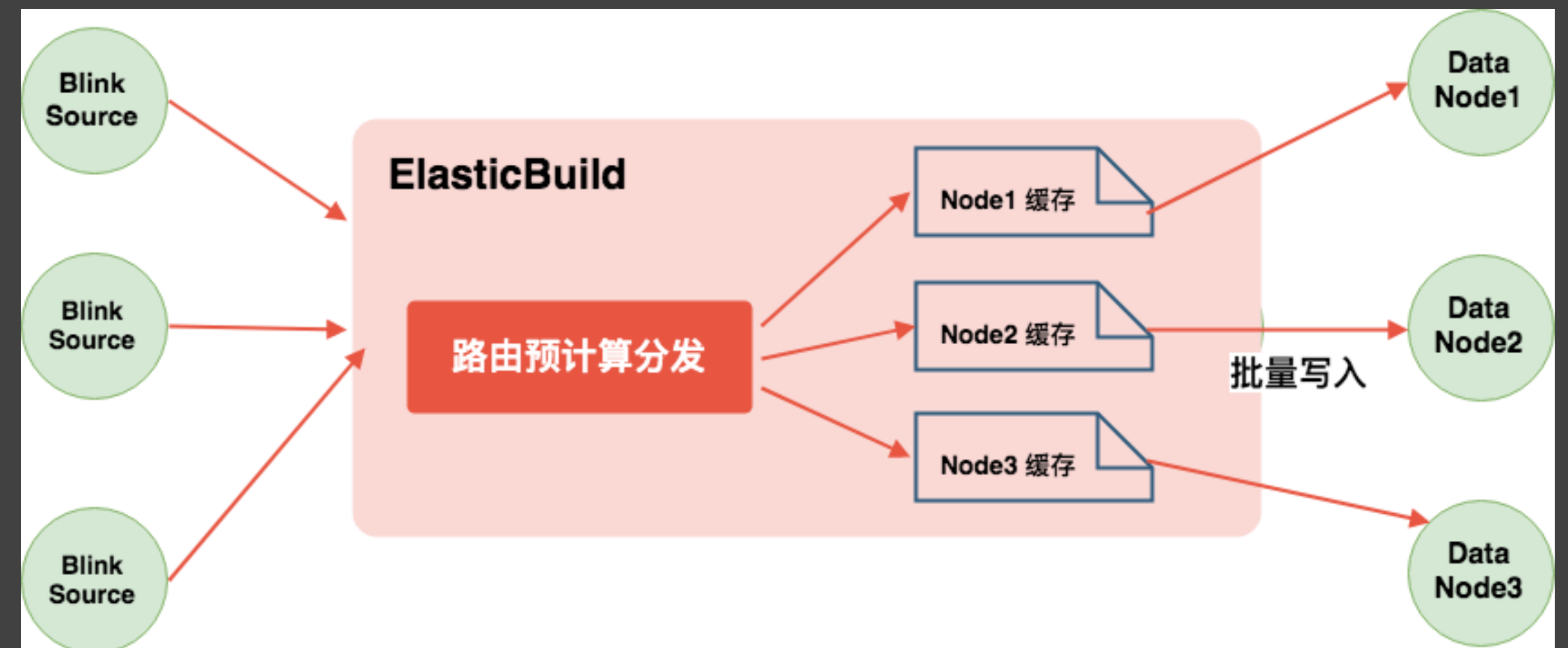
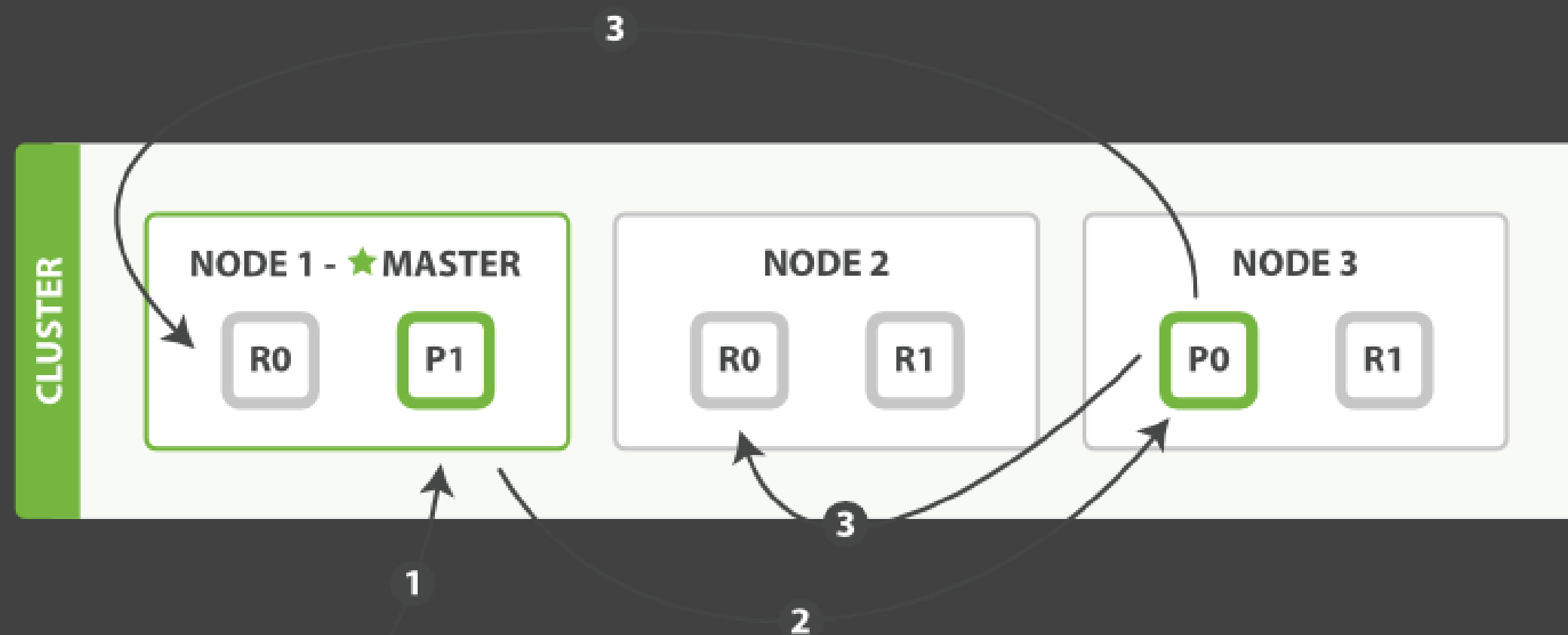
内存索引合并

# ElasticBuild - 增量

**用户痛点:** 整体实时写入TPS为200W/s, 单机写入TPS 3W+/s, 实时写入需优化

**瓶颈分析:** CPU瓶颈, 将路由计算转移至blink中, 减少client node的计算和网络开销

**结果:** ElasticBuild实时部分加入路由预计算批量发送优化, 实时写入性能提升30%





# ElasticFlow产品化输出

一系列优化后相比在线索引写入，性能提升**3**倍

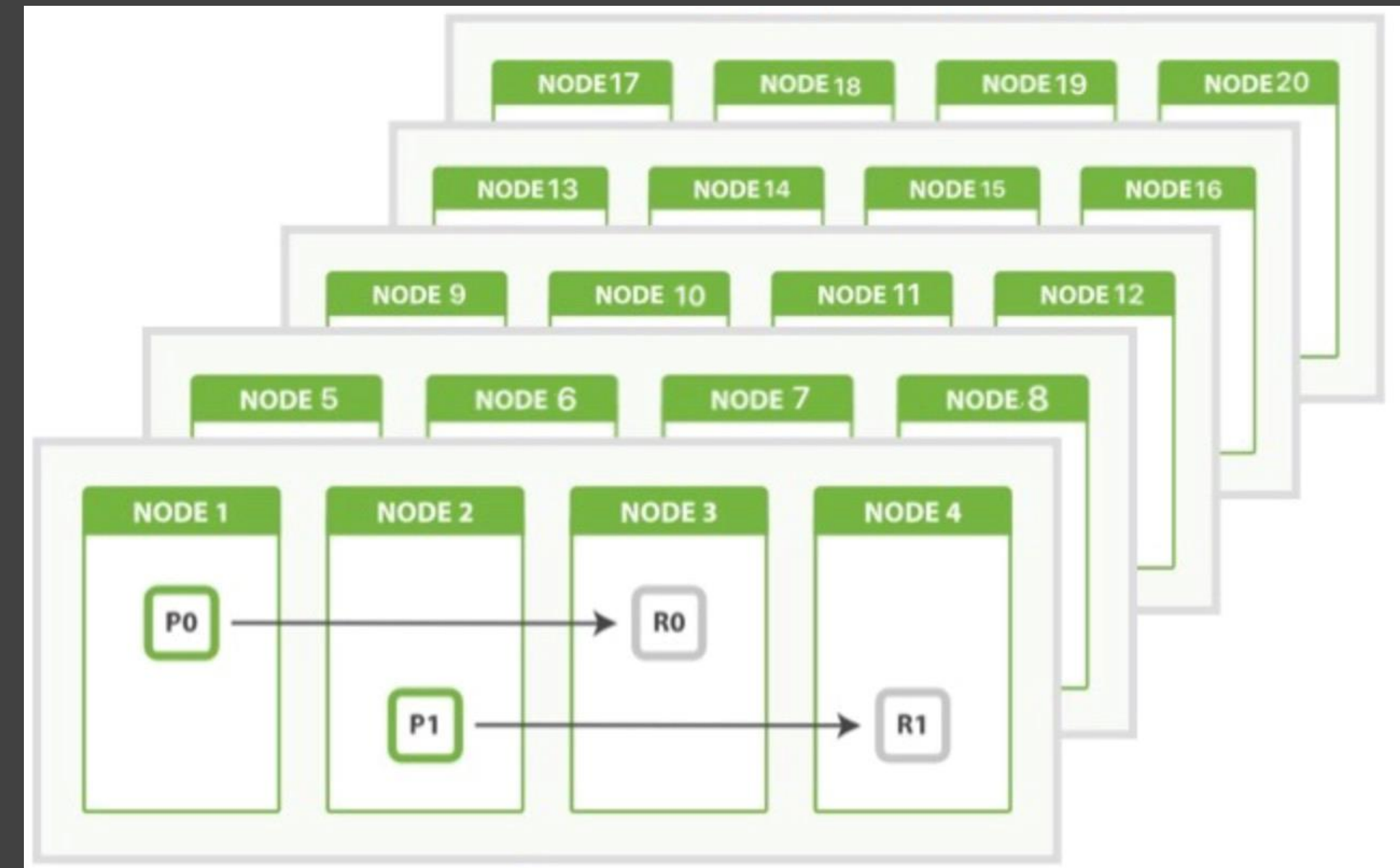
# 计算存储分离 - 用户痛点

早8点至10点流量高峰

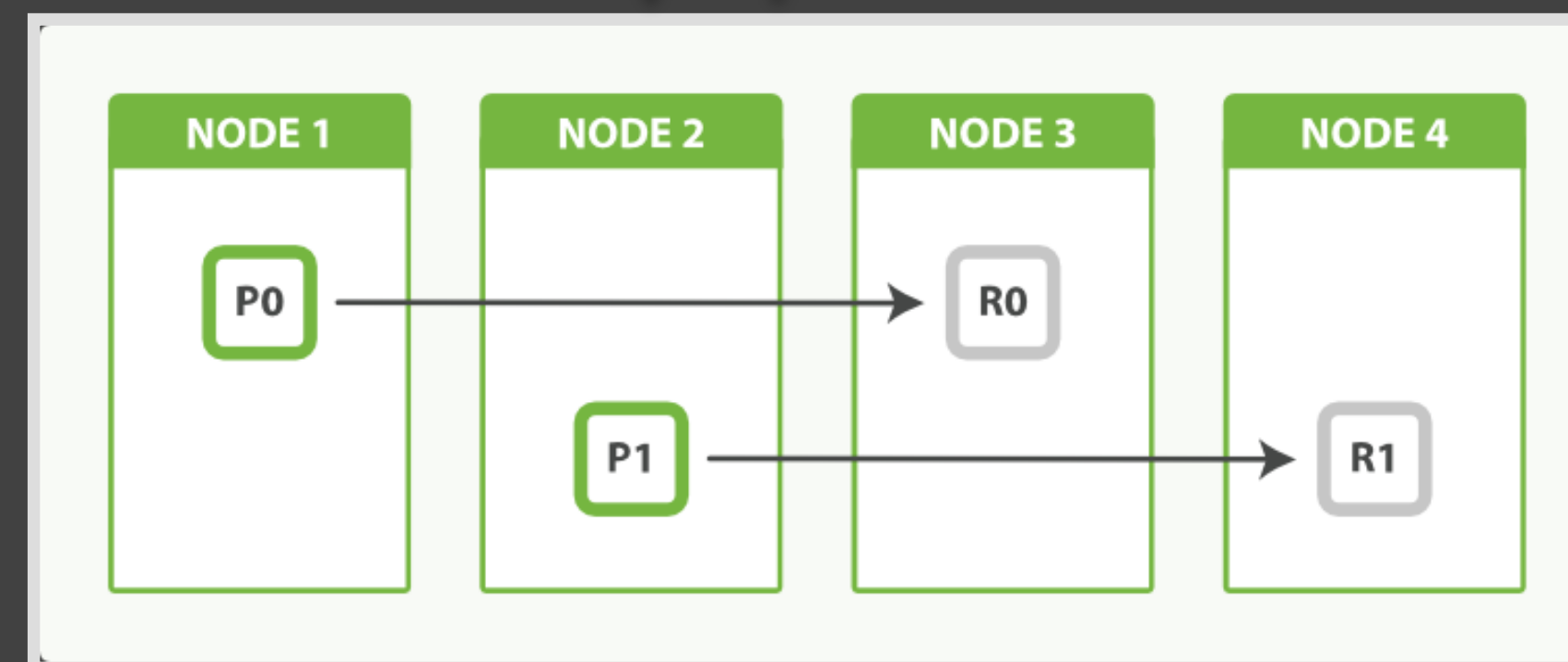
**用户痛点:** 早高峰两小时流量是其他时段的五倍，造成大部分时间资源浪费

**用户痛点:** 弹性扩缩容存在数据迁移慢问题，无法有效实施

**用户痛点:** 副本过多，存储成本高，写入慢

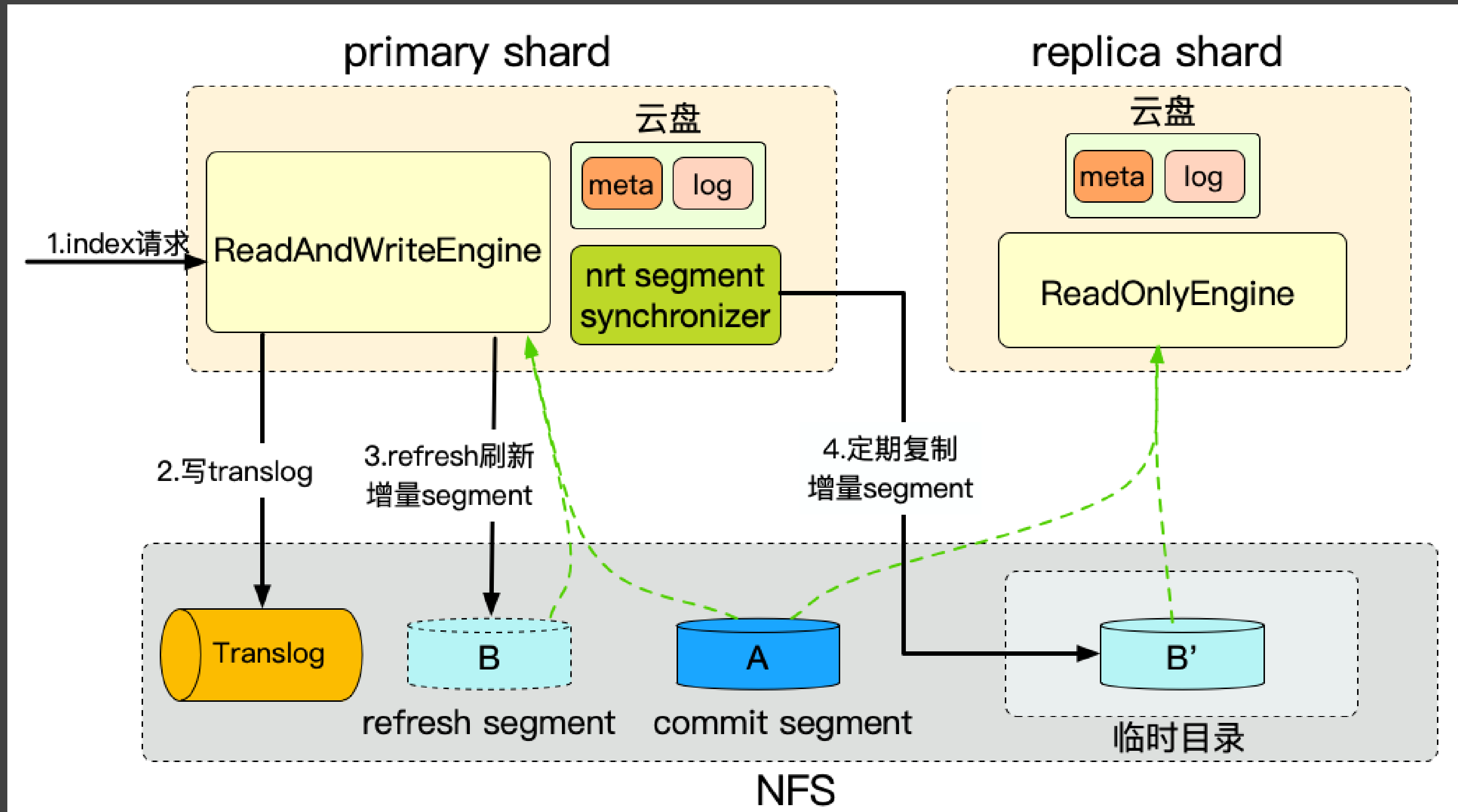


↓ ↑ 弹性扩缩容?



其他时段流量低谷

# 计算存储分离- 架构概述



阿里云高速网络环境

索引分片一写多读

依赖云存储保证数据可靠性

状态与索引分离

IO fence机制保证数据一致性

内存物理复制降低主备延迟

# 计算存储分离优势



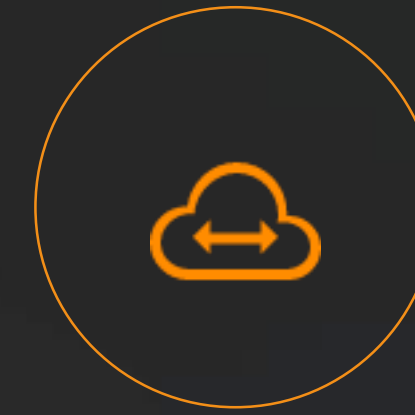
**写入性能提升100%**

计算上避免了副本写入的cpu开销



**存储成本倍数级降低**

业务数据只存一份



**秒级弹性扩缩容**

副本秒级快速扩缩容和迁移，轻松应对高峰流量

# 中文分词插件 - aliws plugin



## 基于阿里巴巴alinlp分词技术

支持多种模型和分词算法包括CRF、结合词典的CRF、MMSEG等，应用于多种业务场景包括淘宝搜索、优酷、口碑等，提供近1G的海量词库



## 支持热更新alinlp词典

通过控制台上传新词典干预分词效果

# 阿里中文分词 VS 开源中文分词

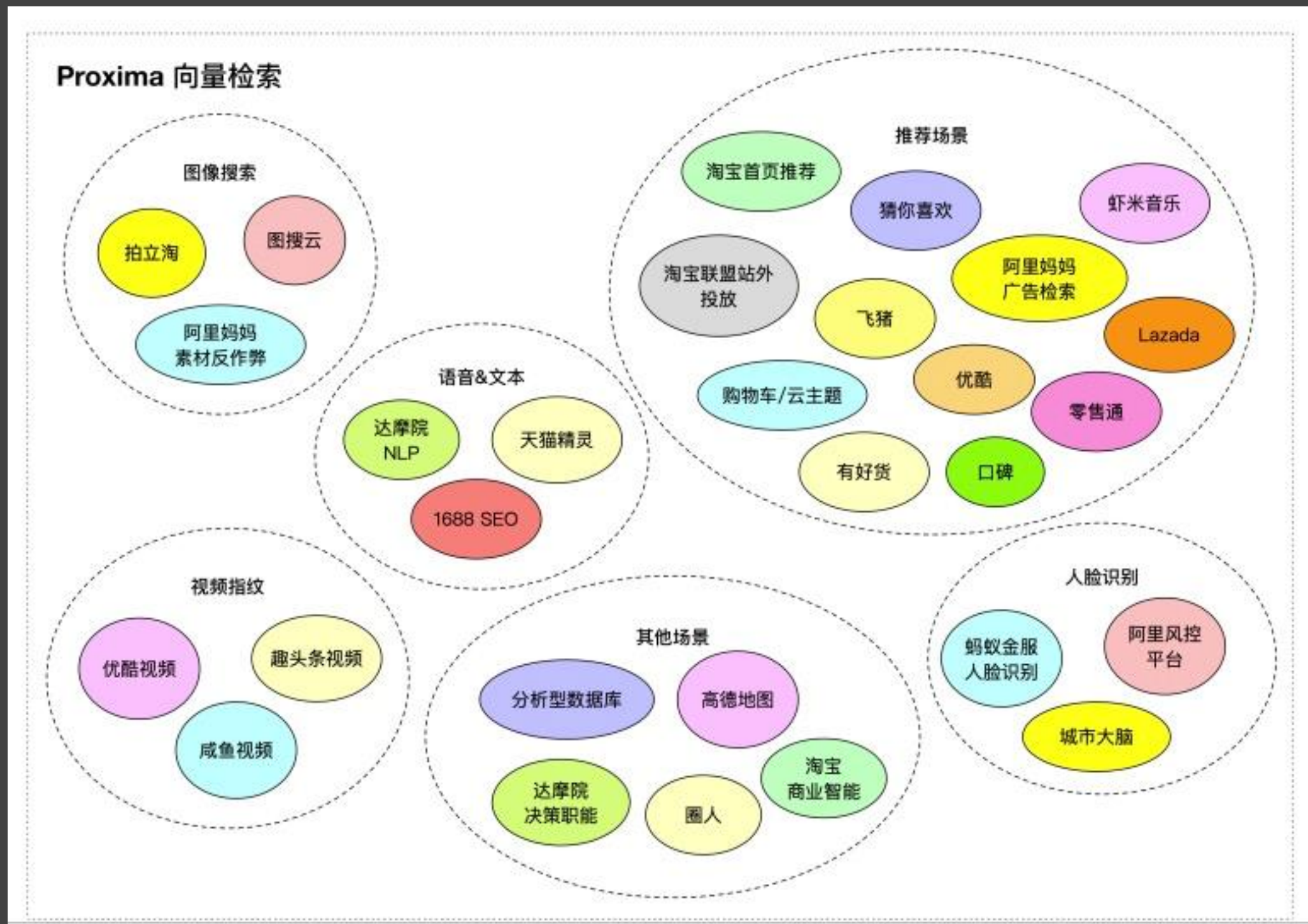
分词效果测试：  
POST \_analyze

```
{  
  "analyzer": "ik_max_word",  
  "text": "南京市长江大桥"  
}
```

“长春市长春街长春药店”会被怎样分词？

分词器	分词结果	查全率	查准率	查询性能
标准分词器standard	南/京/市/长/江/大/桥	高	低	低
中日韩文分词器cjk	南京/京市/ <b>市长</b> /长江/江大/大桥	高	低	高
IK中文分词器ik_max_word	南京市/南京/ <b>市长</b> /长江大桥/长江/大桥	高	低	高
IK中文分词器ik_smart	南京市/长江大桥	低	高	高
阿里中文分词器aliws	南京/市/长江/大桥	高	高	高

# 向量检索插件- aliyun-knn plugin



Proxima向量检索应用场景

## 基于阿里巴巴Proxima向量检索引擎

支持多种检索方法, 如BF (Brute Force)、PQ (Product Quantization)、HC (Hierarchical Clustering)、QGraph (Quantization Graph)、HNSW (Hierarchical Navigable Small World) 等

## 基于Lucene扩展codec功能

与ES原生索引无缝结合

## 支持水平弹性扩展

降低用户使用成本, 支持大规模数据实时读写

# 向量检索插件性能与效果测评

测试环境：2节点阿里云Elasticsearch6.7集群，单节点16核CPU、64GB内存、100G SSD云盘。  
测试数据：sift128维float向量 (<http://corpus-texmex.irisa.fr/>)，数据总量为2千万。

指标\算法	hnsw算法	linear算法
Top10召回率	98.6%	100%
Top100召回率	97.4%	100%
响应时间(P99)	0.093s	0.934s
响应时间(P90)	0.018s	0.305s



# Elasticsearch master调度性能优化

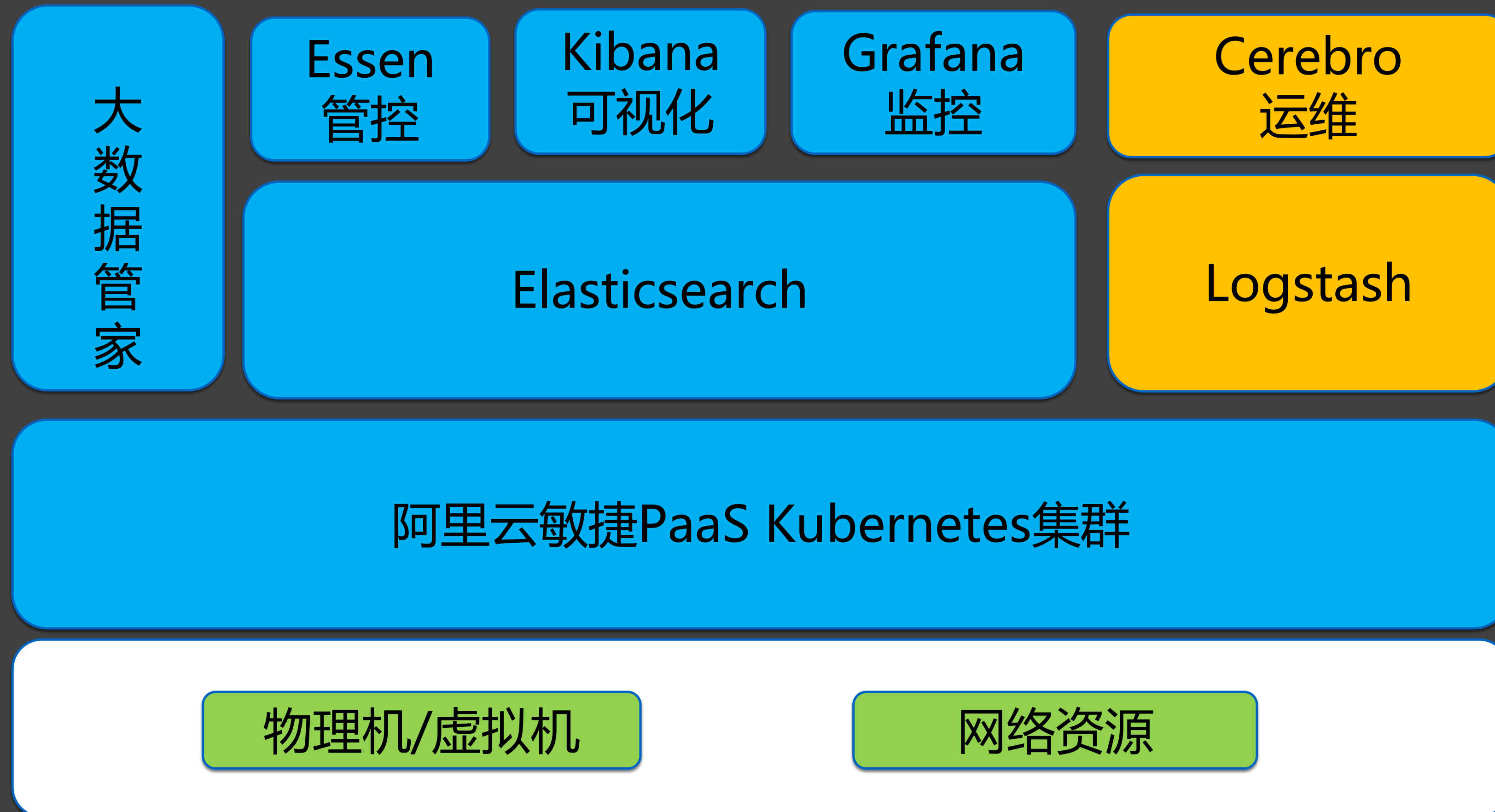
**用户痛点:** 集群有3个专有主节点、10个热节点、2个冷节点，超过5万个shard，创建索引和删除耗时超过1分钟。

**解决方案:** reroute的调度算法复杂度 $O(n^2)$ 降为 $O(n)$ ，大规模集群索引创建和删除耗时降到1s。

```
@@ -144,51 +156,34 @@ public int numberOfShardsWithState(ShardRoutingState... states) {
144     * @return List of shards
145     */
146     public List<ShardRouting> shardsWithState(ShardRoutingState... states) {
147 -         List<ShardRouting> shards = new ArrayList<>();
148 -         for (ShardRouting shardEntry : this) {
149 -             for (ShardRoutingState state : states) {
150 -                 if (shardEntry.state() == state) {
151 -                     shards.add(shardEntry);
152 -                 }
153 -             }
154 -         }
155 -         return shards;
156     }

156     * @return List of shards
157     */
158     public List<ShardRouting> shardsWithState(ShardRoutingState... states) {
159 +         return Arrays.stream(states)
160 +             .map(state -> statesToShards.get(state))
161 +             .flatMap(Collection::stream)
162 +             .collect(Collectors.toList());
163     }
```

# ES on K8S – 专有云部署



# 阿里云敏捷PAAS版优势

产品特性	阿里云ES	客户自建开源ES
弹性扩展	扩容简单，支持一键扩节点、规格	需要手工部署软硬件扩容新节点
集群管理	白屏化管理界面，多租户，支持自定义插件、词典管理	不支持多租户，无附加增值功能
安全性	支持安全认证	默认无安全认证，容易出现数据泄露
可靠性	自动数据备份	无灾备能力和容错机制
存储类型	物理机本地盘,支持选择SSD盘/机械盘	物理机本地盘
资源隔离	通过容器技术隔离资源，支持混部	软件混部无法做到资源隔离，业务互相影响
中文分词	支持阿里自研中文分词，搜索结果准确	ES自带中文分词效果差，搜索结果难以优化
检索能力	支持文本、图像、视频检索，阿里自研向量检索	只支持文本检索
技术支持	阿里云官方技术支持	技术瓶颈解决不了线上问题
运维成本	自动化部署，运维成本低	部署运维成本高
生态	ELK产品组合、阿里云大数据产品组合	ELK产品组合

# 全文检索应用



姓名	笑笑	性别	女	年龄	26岁
祖籍	浙江	学历	本科	职业	KTV从业人员
婚姻	已婚	常居住地	浙江	车牌号	浙A*****

个人标签: 涉毒前科, 昼伏夜出, 昼伏夜出

### 综合人物画像



风险系数	[中]	消费能力	[高]
身份特征	[危险]	行为模式	[正常]
内容特征	[危险]	关系网络	[正常]

### 行为模式: 正常

时间: 2018年04月18日




规律: 09:12:12-18:34:11

异常:

01:12:12-04:12:15	*****酒店
06:20:11-07:00:10	*****KTV

### 推荐

#### 亲属关系



父: 李\*\*, 王\*\*  
母: 吴\*\*, 陈\*\*  
目标: 李\*\*  
夫: 吴\*\*  
女: 李\*\*  
儿: 李\*\*

#### 相似&相关人员

李**	铁路同乘车2次
张**	民航同订票1次
李**	铁路同乘车2次
张**	民航同订票1次
李**	铁路同乘车2次
张**	民航同订票1次

#### 疑似相关案件

某某某某某某案件  
某某某某某某案件

## 公安智能搜索

专题检索、批量检索、时空检索、标签搜索、以图搜图以及轨迹假设、轨迹对比、轨迹展现

## 数据量

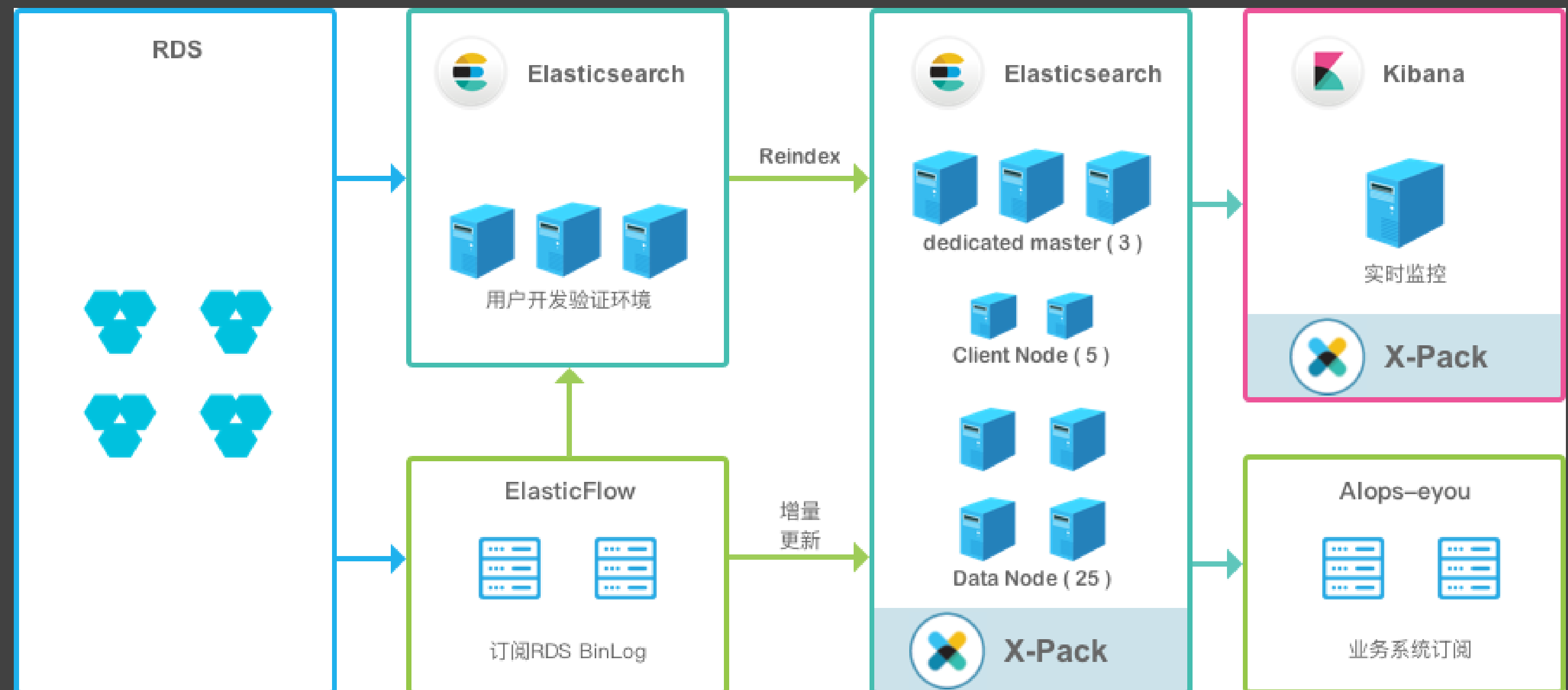
1000亿+条人员、物品、案件、地址、组织等要素信息

## 集群规模

单个业务150个节点Elasticsearch, 节点配置16核CPU、64GB内存、2TB SSD盘, 支持跨集群搜索

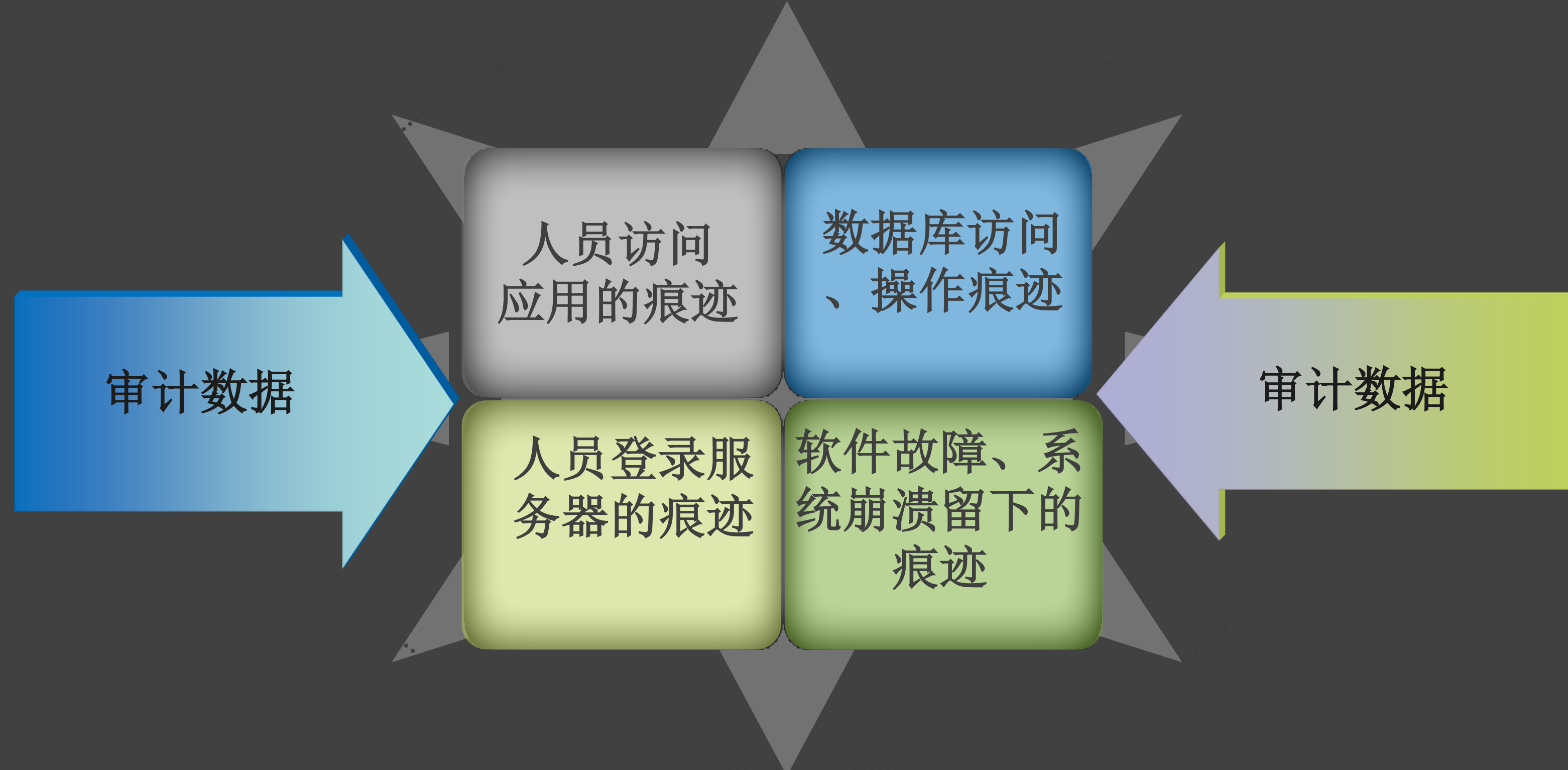
# 数据库加速应用

- 1 保单数据更新**秒级**实效性
- 2 **1万**并发请求下服务稳定性
- 3 **同城多活**的服务容灾能力
- 4 构建索引之前的**流式数据预处理**



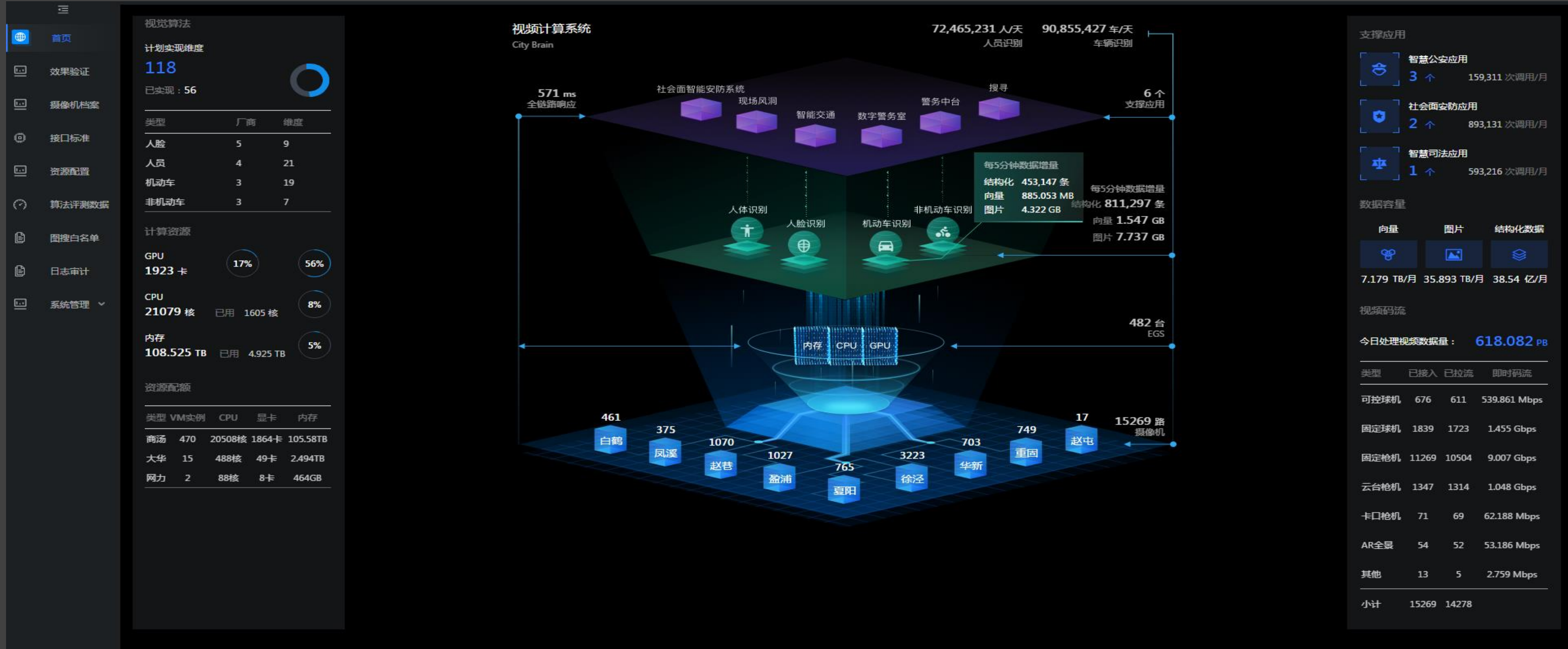
# 日志审计应用

采集应用系统、数据库日志，提供日志搜索、风险告警、分析报告等服务，解决数据盗取、越权访问、信息泄露等问题。



# 交通视频应用

结构化数据检索: 车牌号、时间、经纬度、卡口  
向量数据检索: 人体、人脸、机动车、非机动车



# THANKS !

Elasticsearch中文技术社区   
1817人



 扫一扫群二维码，立刻加入该群。

欢迎加入Elasticsearch技术交流钉钉群！





## 关于

您所阅读的资料出自 Elastic 中文社区 深圳 Meetup 活动 @2019-11-16

<https://meetup.elasticsearch.cn/event/shenzhen/1002.html>

Elastic 中文社区 <http://elasticsearch.cn>

Elastic Meetup 是由 Elastic 中文社区定期举办的线下交流活动，主要围绕 Elastic 的开源产品（Elasticsearch、Logstash、Kibana 和 Beats）及 Elastic Stack 周边技术，探讨在搜索、数据实时分析、日志分析、安全等领域的实践与应用。

欢迎加入 Elastic 中文社区，**参与分享交流** 或 **赞助社区活动**！

深圳联络人：杨振涛

微信：nodexy

邮箱：nodexy@qq.com

本次活动回顾及现场照片在“vivo互联网技术”公众号发布，欢迎关注浏览。



微信扫码关注