

南京云利来基于ELK的 大数据分析平台实践分享

主题介绍



主题

分享

启发

主题：Elasticsearch的搭建过程和运维经验。

分享：工作期间主要遇到的问题和解决思路。

启发：快速发现并解决问题，提升运维效率。



CONCENTS



公司简介



数据分类



运维之路



告警分析

The background is a light blue gradient. There are three dark blue hexagons: one at the top center, one on the left side, and one at the bottom center. The bottom hexagon contains the white number '01'.

公司简介

Company profile

01

公司介绍



专注实时网络使用分析，世界领先
20Gbps分析能力

为数据中心搭建大数据分析平台

提供智能运维，网络安全和预警分析
能力

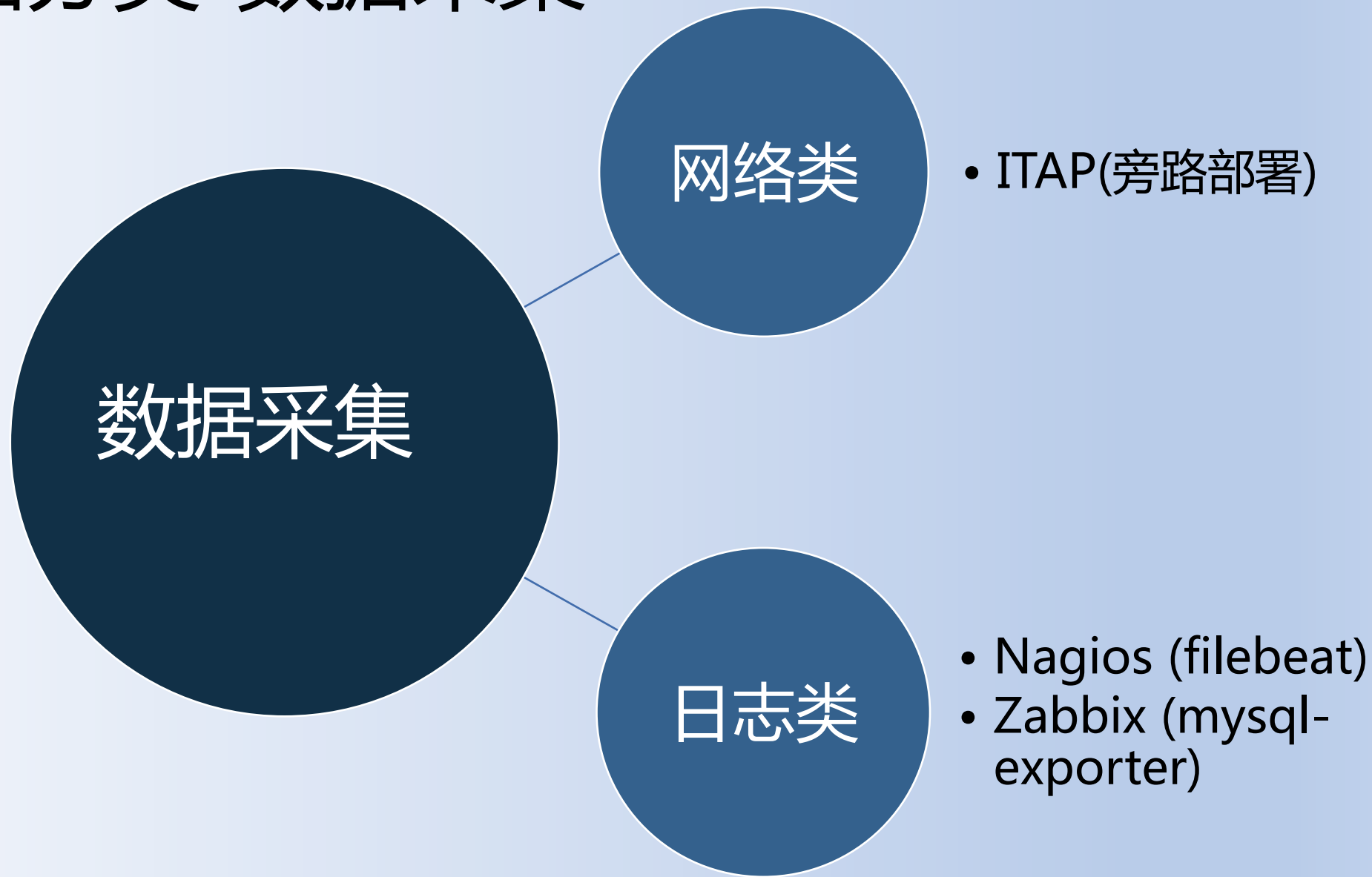
02

数据分类

Data classification

1. 数据采集
2. 数据类型
3. 数据量
4. 使用场景

数据分类-数据采集



数据分类-数据类型



数据分类-数据量



Everyday

每天数据量至少2TB，
记录数22亿，不含副
本



Peak

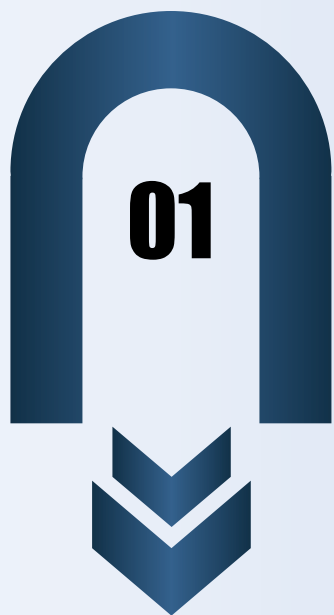
高峰数据量每秒6万
条记录



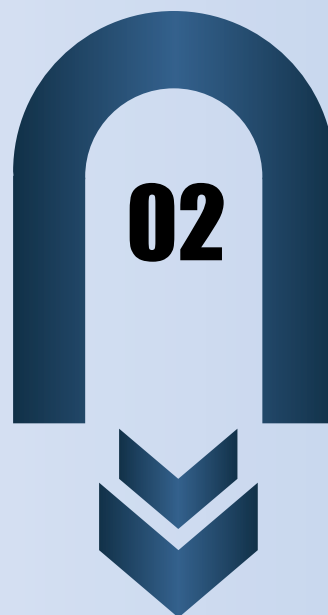
Fast

单个索引最快处理12
万条记录每秒

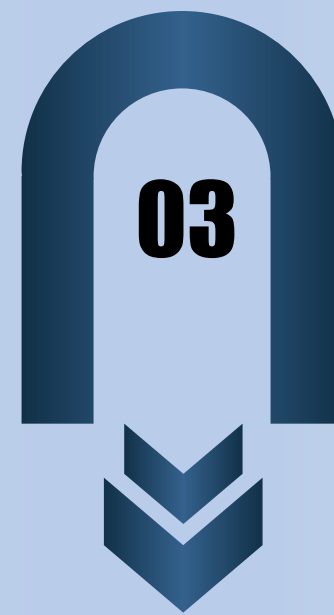
数据分类-使用场景



查询
聚合



大屏分析
预测告警



网络指标
业务指标
安全指标

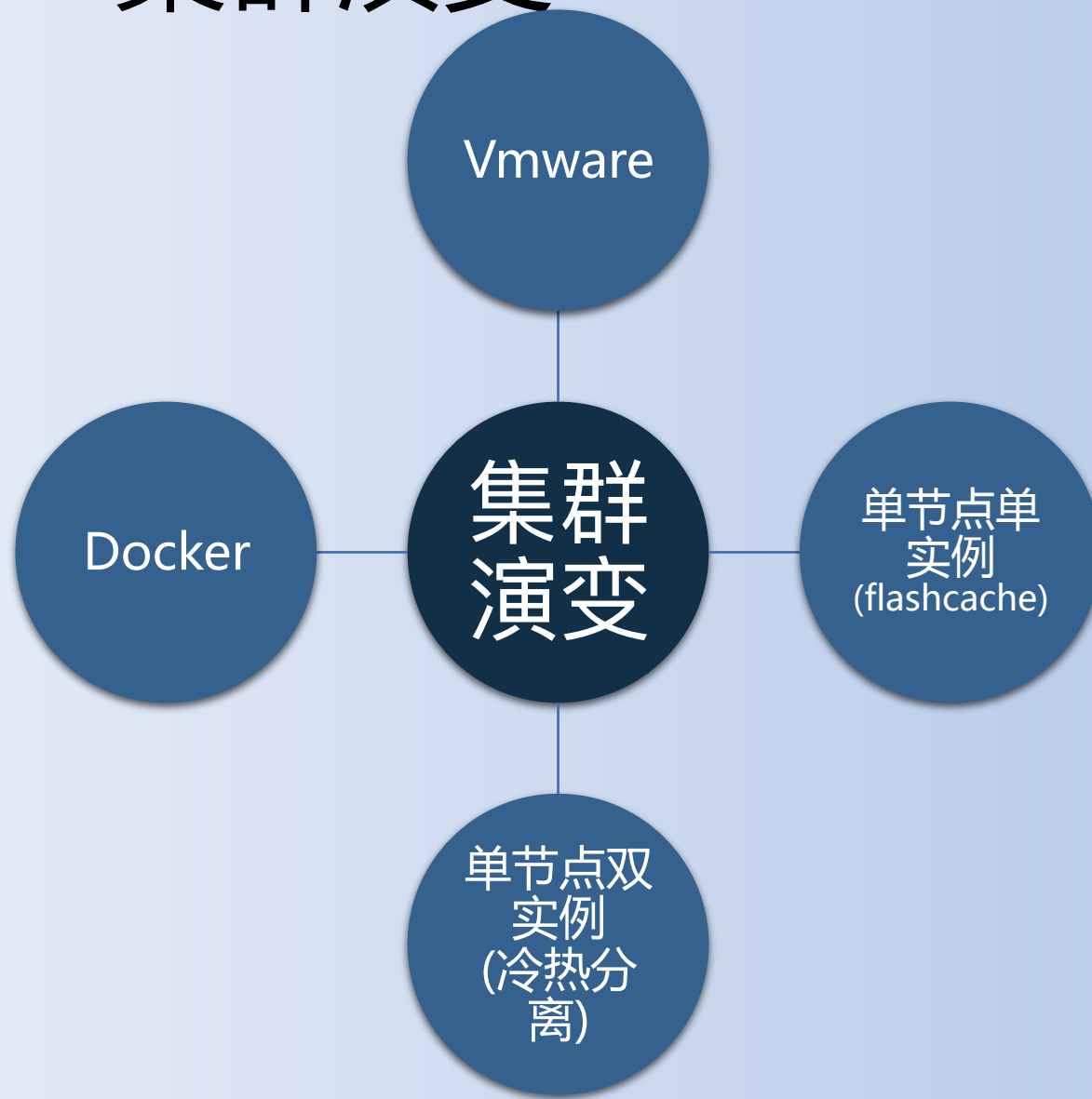
03

运维之路

The road of operations

1. 集群演变
2. 冷热分离
3. 重要选型
4. 性能分析
5. 存储规划
6. 性能提升
7. 集群监控

运维之路-1.集群演变



运维之路-2.冷热分离

采用flashcache
模式

- 磁盘IO连续小块读
- 负载高，IOwait高
- 分析发现存在抖动


采用单机双实例
冷热分离模式

- 充分利用1.6TB的SSD
- 只保存每天的热数据
- 隔夜迁移到HDD Raid0

升级主要目的：

- 内存隔离-当天和历史JAVA对象分离在不同的JVM里。
- IO隔离-当天和历史数据的磁盘IO分离在不同的磁盘上。

运维之路-前后效果对比

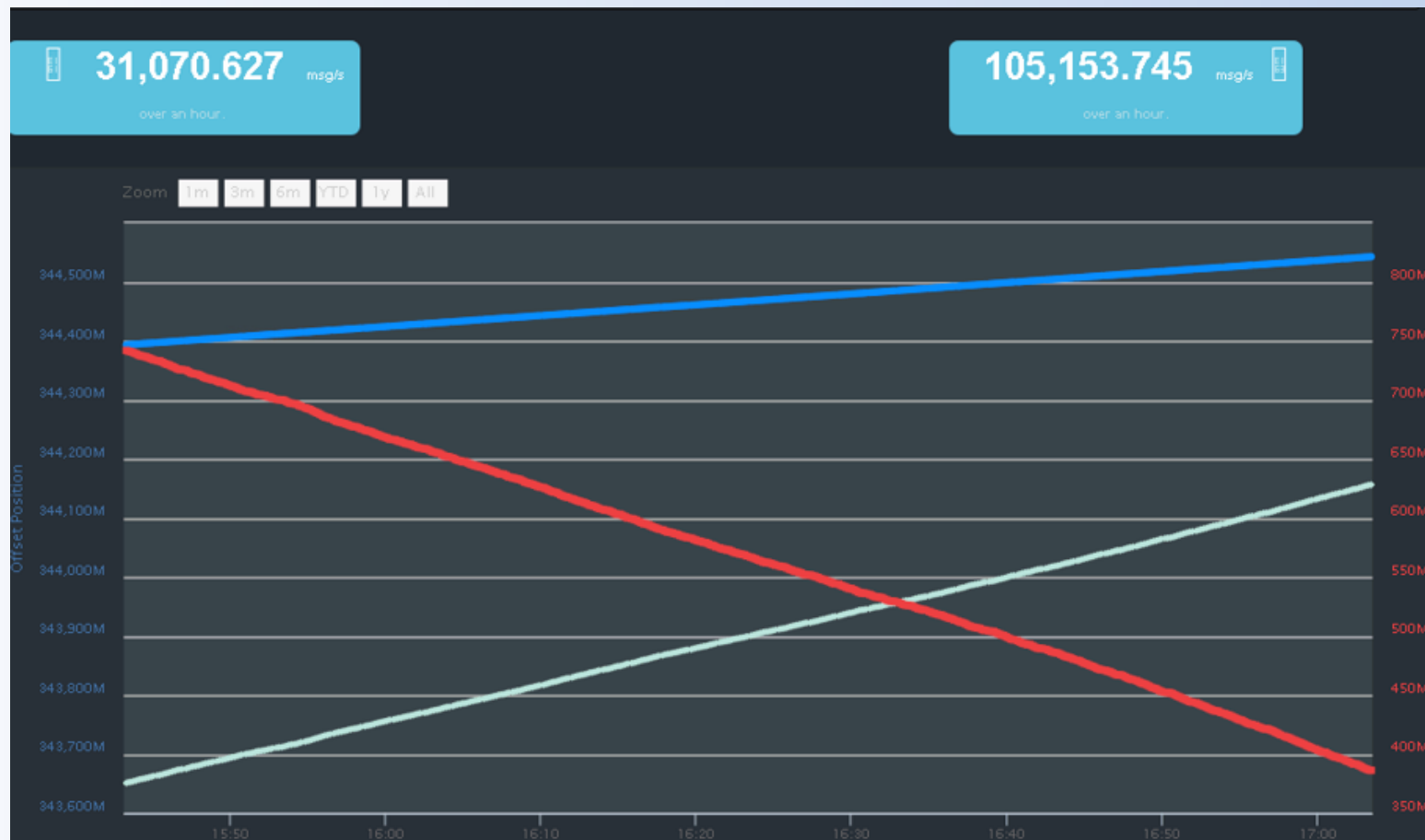


----total-cpu-usage----						-dsk/total-	
usr	sys	idl	wai	hiq	siq	read	writ
14	2	77	7	0	0	11M	126M
13	3	77	7	0	0	16M	126M
12	2	78	8	0	0	20M	130M
26	2	65	6	0	0	11M	140M
11	2	80	6	0	0	17M	131M
7	2	85	5	0	0	24M	127M

----total-cpu-usage----						-dsk/total-	
usr	sys	idl	wai	hiq	siq	read	writ
11	1	89	0	0	0	0	0
9	1	90	0	0	0	0	66M
13	2	84	1	0	0	229M	0
7	0	93	0	0	0	0	25M
6	0	91	2	0	0	0	252M
5	0	94	0	0	0	0	0

升级后，有效减少了cpu wait和磁盘读，降低了系统负载，有效提升查询和写入性能。

运维之路-前后效果对比



- 单个索引最高速度从之前的60,000条每秒提升到120,000条记录每秒，平均10万条每秒
- 聚合查询性能提升1倍

运维之路-3.重要选型

CPU Xeon E5-2600 V4系列

Mem 128TB,SSD 1.6TB,HDD 40TB

OS file system

Shard,Replica

Client,Master,Data

运维之路-3.重要选型

Xeon E5-2600 V4系列

01

比V3系列提升JAVA性能60%

02

指令预取, cache line预取, Numa Set

Refer to
Server-side
Java*
Benchmarks
@[https://
www.intel.cn/
content/www/
cn/zh/
benchmarks/
server/xeon-
e5-v4/xeon-
e5-v4-server-
side-java.html](https://www.intel.cn/content/www/cn/zh/benchmarks/server/xeon-e5-v4/xeon-e5-v4-server-side-java.html)

运维之路-3.重要选型

Mem 128GB,SSD 1.6TB,HDD
40TB

01

大内存, Cache加速

02

写负载高上SSD, 定期Trim优化

03

利用ssd,sas和sata盘分级存储

IO
scheduler
CFQ/NOOP

运维之路-3.重要选型

OS file system(ext4,xfs)

01

针对HDD,SSD 4k对齐优化

02

每个分区的Start Address能被8整除

03

解决跨扇区访问，减少读写次数和延迟

运维之路-3.重要选型

Shard, Replica

01

Shard count 根据节点数*(1~3)

02

Shard size 控制在30GB以内

03

Shard docs 控制在5百万记录以内

04

Replica 至少为1

运维之路-3.重要选型

可靠性

01

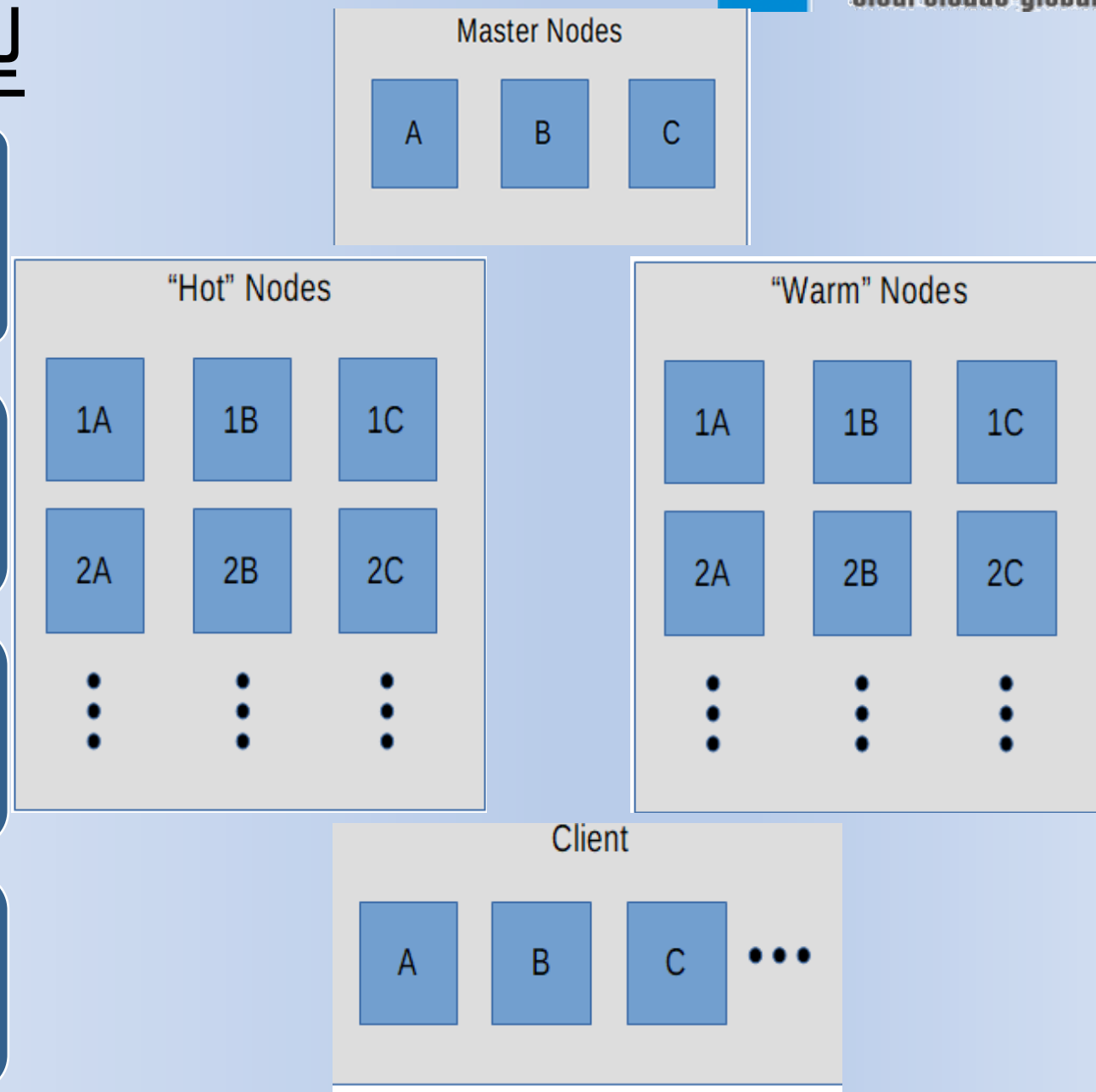
• 单点失败

02

• 冷热分离

03

• 负载均衡



运维之路-4.性能分析



运维之路-4.性能分析



• 高负载

- Cpu or IO 负载型
- Sar , Vmstat , IOstat , Dstat
- Systemtap , Perf

运维之路-4.性能分析

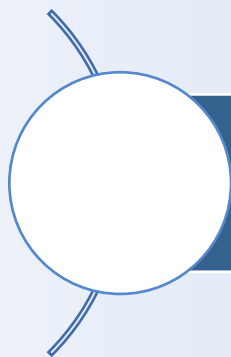


•线程池

Index , Query , Merge , Bulk:

- Thread , Queue Size
- Active , Queue , Rejected

运维之路-4.性能分析



•内存占用

- Node, Indice, Shard level stat api
- Querycache, Fielddata

运维之路-4.性能分析



• 查询

- 慢查询
- 请求，响应，延时，峰值统计
- Query profile
- Cache filling, hit, miss, eviction
- 日志埋点采集，query replay

运维之路-4.性能分析



• 集群健康

- `_cluster/health`
active, reallocating, initializing, unassigned
- Ping timeout
- Shard allocation, recover latency

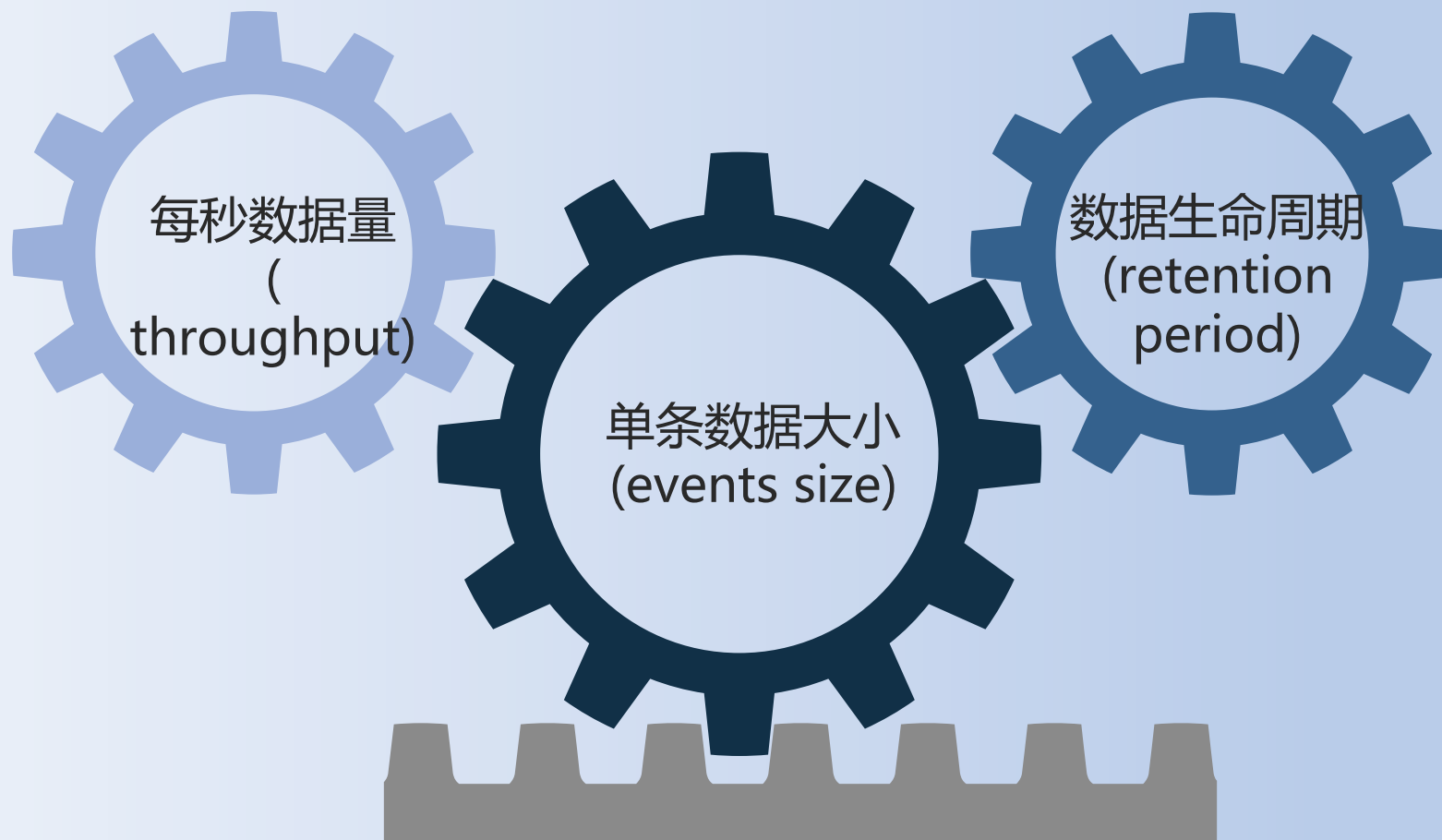
运维之路-4.性能分析



•GC效率

- GC时长占比，GC回收量占比
- 内存增长速率，内存回收速率
- 各代回收耗时，频率
- Dump profile
- Jstack，Jmap，Jstat

运维之路-5.存储规划



$$\text{Storage} = \text{throughput} * \text{events size} * \text{retention period}$$

运维之路-6.性能提升

合理设计

优化查询

内存保护

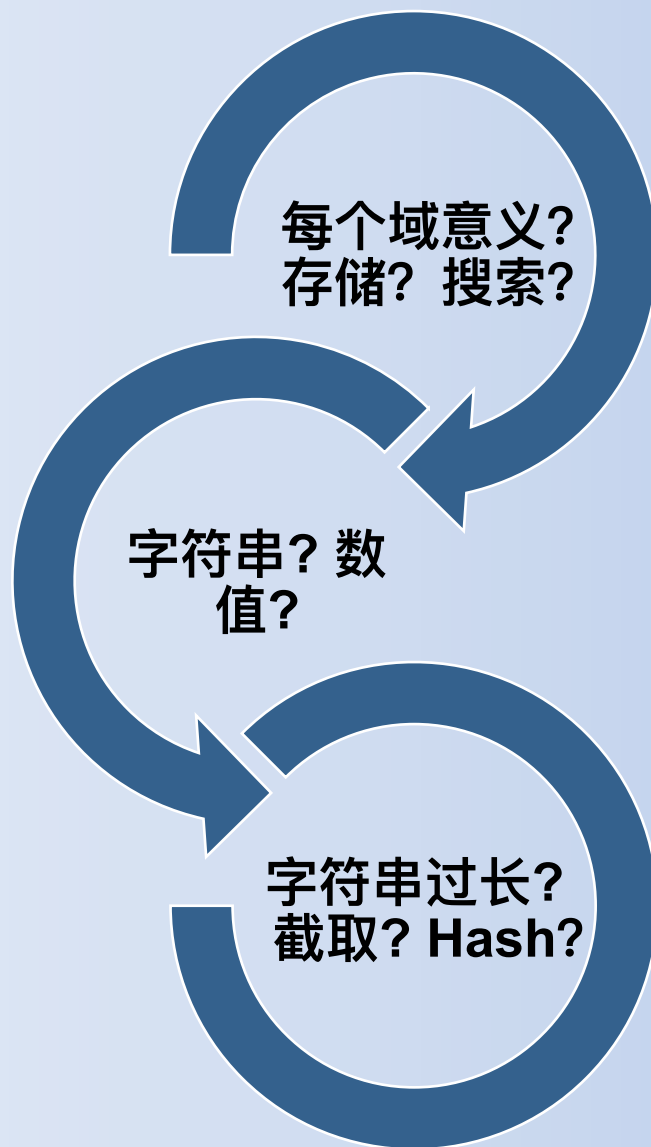
批处理

按需隔离

规划计算

运维之路-6.性能提升

合理设计



运维之路-6.性能提升

优化查询

01

- From&Size VS Scroll

02

- 聚合? 高基数? 深度?

03

- Term Execution_hint

04

- Analyzed&Wildcards

05

- Dashboard , 5*5

06

- Warm-up

运维之路-6.性能提升

内存保护

Fielddata

Circuit breaker

indices.
fielddata.
cache.
size

indices.
breaker.
total.
limit

indices.
breaker.
fielddata.
limit

indices.
breaker.
request.
limit

运维之路-6.性能提升

批处理

Index Refresh

Translog

Bulk batch size

Merge

运维之路-6.性能提升

按需隔离-分表分级分组

logstash-\$(date +%Y-%m-%d)

Index&Search _routing

Dynamic_templates

- tag schema : include, exclude, require
- order level: 0, 1, 2

运维之路-6.性能提升

规划计算

01

提前聚合后插入

02

超过生命周期后只保留基线

03

近似值

04

Pipeline

运维之路-7.集群监控

Netdata

_cat api

Sense

Cerebro

Prometheus



告警分析

The alarm analysis

04

告警引擎

Sql语法

工作日

同比环比

平均值&标准差

基线学习



不断迭代，严苛细节，
最终性能是否满足？
可接受？

Q&A

谢 谢

欢迎您加入我们!

<mailto:info@clearclouds-global.com>



elastic
中文社区

专业、垂直、纯粹的 Elastic 开源技术交流社区

<https://elasticsearch.cn/>