斗鱼ElasticSearch集群运维实践

文远

斗鱼 DOUYU.COM

目的

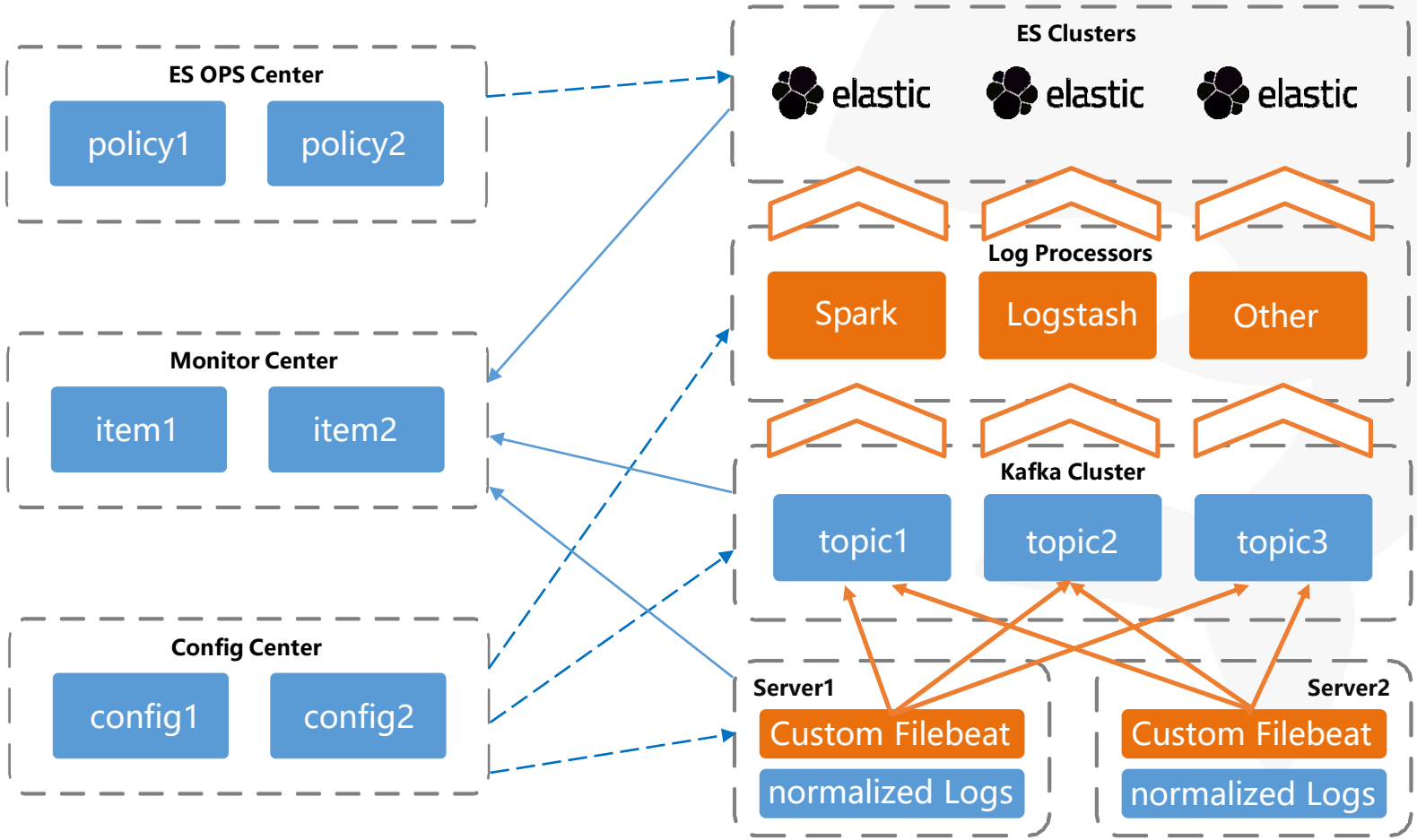
- 集群规模/版本/组成
- 日志系统架构
- 自动维护
- 性能优化

斗鱼 DOUYU.COM

集群组成

- 总共12个独立的ES集群，60+台服务器
- 最大集群 8台服务器 每天: 52+亿条日志 / 10+TB数据
- 总量：430+ TB数据
- 业务：日志(持续写入)、静态(一次写入)、半静态(一次全量，持续更新)
- 日志集群按照部门/业务+语言栈划分
- 多个日志集群之上开启跨集群搜索

日志收集系统


ES OPS Center
policy1
policy2
Monitor Center
item1
item2
Config Center
config1
config2
ES Clusters
elastic
elastic
elastic
Log Processors
Spark
Logstash
Other
Kafka Cluster
topic1
topic2
topic3
Server1
Custom Filebeat
normalized Logs
Server2
Custom Filebeat
normalized Logs

斗鱼 DOUYU.COM

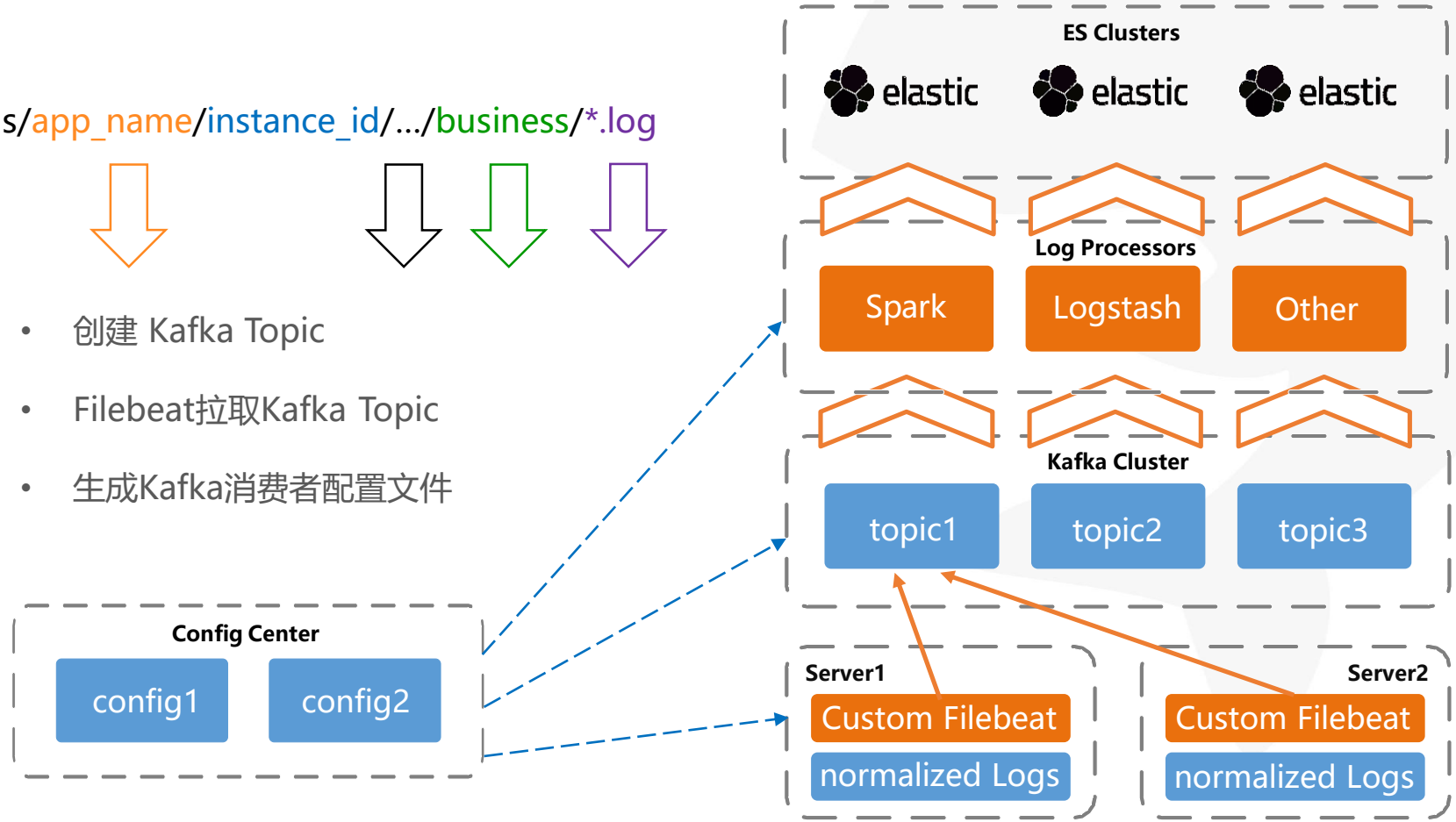
日志系统规范

- 日志规范
  - 固定字段定义
  - 日志格式
- 日志采集
  - 落盘规则
  - 滚动策略
  - 采集方法
- 日志传输
  - 消息队列
  - 消费方式
  - Topic规范
  - 保存时间
- 日志切分
  - 采样
  - 过滤
  - 自定格式
- 日志检索
  - 索引分割
  - 分片设置
  - 检索优化
  - 权限设置
  - 保存时间
- 日志流监控
  - 采集异常
  - 传输异常
  - 检索异常
  - 不合规范
  - 监控报警

日志接入

/logs/app_name/instance_id/.../business/*.log

- 创建 Kafka Topic
- Filebeat拉取Kafka Topic
- 生成Kafka消费者配置文件

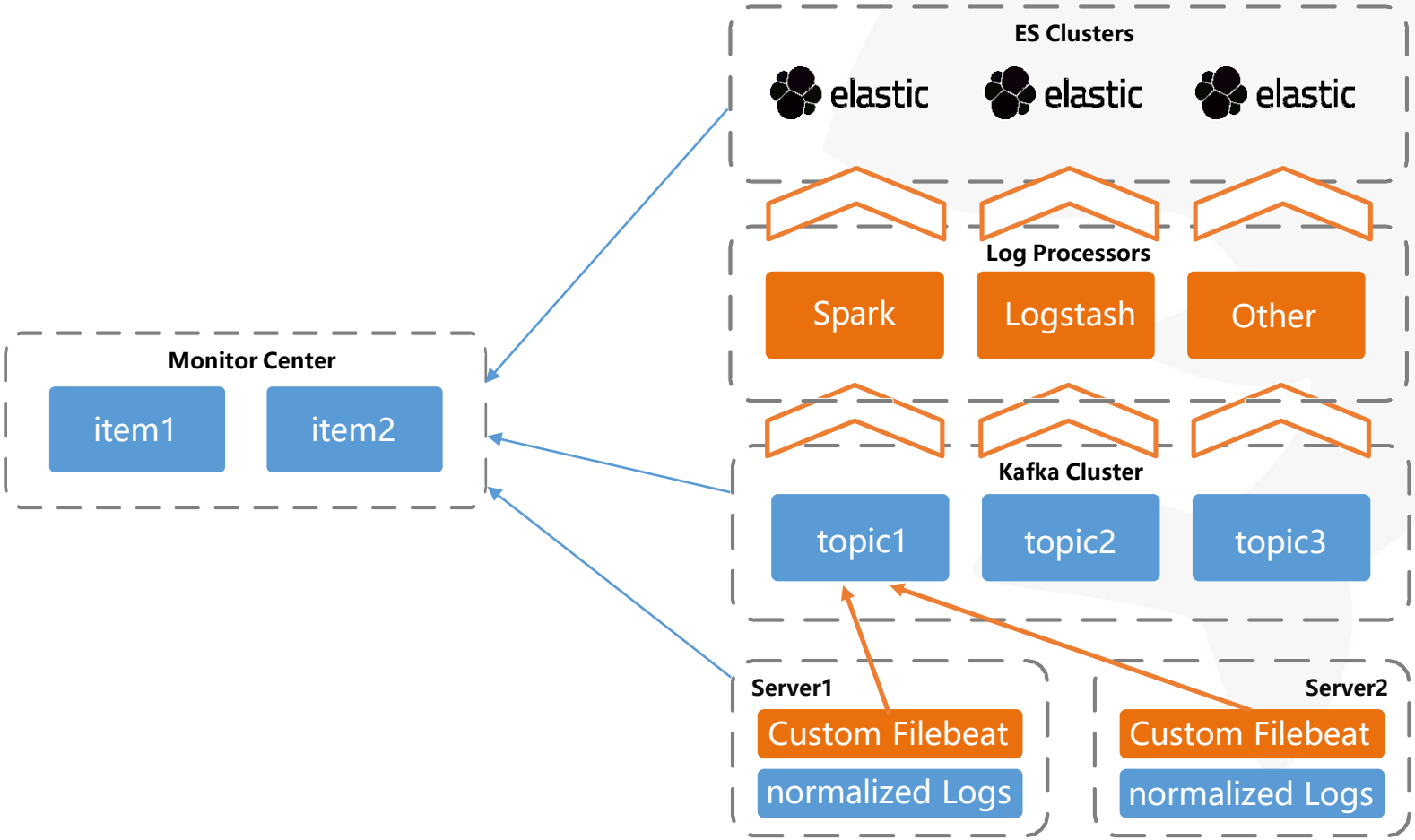


斗鱼 DOUYU.COM

日志接入

- 固定字段定义 – 统一字段名，且部分可以从日志路径得来
  - timestamp / level / app_id / app_name / instance_id
- 日志格式 – 同一语言栈采用统一格式
- 滚动策略 – 同一语言栈采用统一策略
- 索引分割
  - 定时任务自动分析 – 由日志量的大小，提前建立下一天索引为按天或按小时分割
- 分片设置
  - 手动配置初始固定分片数 / 关闭时间 / 删除时间 等
  - 定时任务自动分析 – 由日志量的大小，提前建立下一天索引分片数
- Kibana设置
  - 自建网关，反向代理多个Kibana，接入公司OA系统

采集监控


ES Clusters
elastic
elastic
elastic
Log Processors
Spark
Logstash
Other
Monitor Center
item1
item2
Kafka Cluster
topic1
topic2
topic3
Server1
Custom Filebeat
normalized Logs
Server2
Custom Filebeat
normalized Logs

采集监控

- 采集异常
  - 收集端使用定制版Filebeat，自身日志同样收集到ES监控集群
  - Registry文件内容定时全量采集到ES监控集群
  - Logstash，Java和Spark消费Kafka的日志收集到ES监控集群
  - ES集群自身日志收集到ES监控集群
  - 自建网关日志收集到ES监控集群
- 限流 / 采样
  - 默认全量导入所有日志
  - 可以手动配置采样比例，重新启动Kafka消费者进行采样导入
- 监控报警
  - 使用自研监控系统接入以上索引，针对异常情况报警
  - Kibana是否可用
  - ES集群是否可用及索引性能

斗鱼 DOUYU.COM

日志问题

- 字段过多
  - 业务代码写日志奔放，直接把对象序列化写到日志
- 字段重名
  - 除了固定字段，业务日志字段存在重名而不同类型的问题，导致丢日志
  - 两种解决方案：key1/value1 和 提练固定字段
- 日志格式不合规
  - 格式不符合标准JSON
  - 字段中有JSON不能解析的特殊字符
  - 字段超长
  - Ngnix日志中有中文Unicode

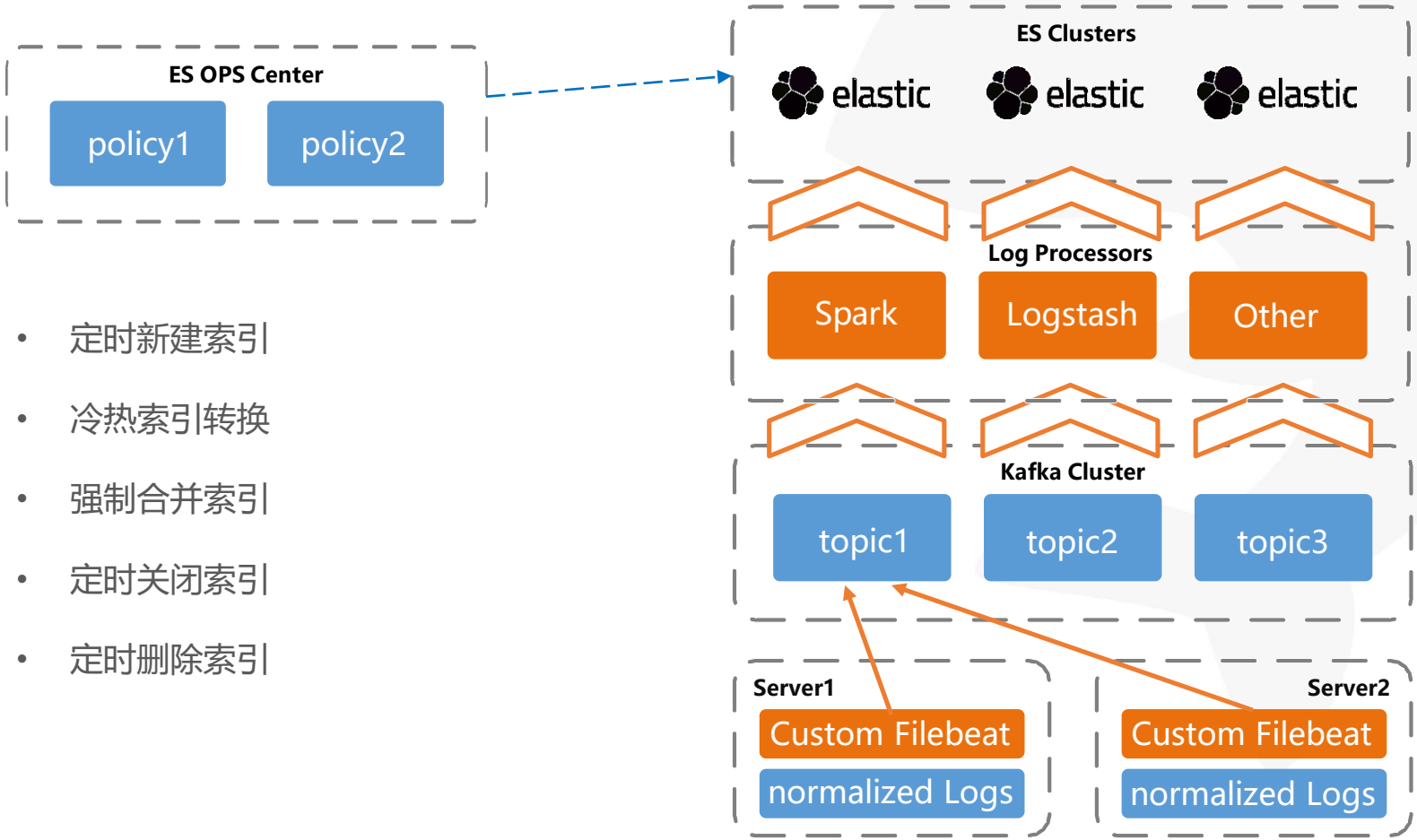
集群设置

- 硬件设置
  - CPU 32核，内存 128GB
  - 对传统硬盘采用6~8块盘组成RAID 0提高IO写入速度
  - 对延迟要求低的APM日志、网关日志等，使用SSD存储热索引
  - 服务器预留大约一半内存作为系统缓存提升查询性能
- 集群配置
  - 版本：大部分 Java 1.8.0_162，Elasticsearch 6.2.4
  - Heap：Master节点 4~6GB，Client节点 10~15GB，Data节点 30GB
  - 一台物理机上启动两个Data节点
  - 开启cluster.routing.allocation.same_shard.host选项
  - 写入的Client节点 和 读取的Client节点 分开
  - 对外提供查询服务的cross-search小集群分开，专门提供跨集群搜索服务

索引设置

- 索引刷新间隔 index.refresh_interval 为30s或更大
- 事务日志持久化模式 index.translog.durability 为 async
- 推迟分片分配 index.unassigned.node_left.delayed_timeout 为 10m
- 合并线程数 index.merge.scheduler.max_thread_count 对机械硬盘为1
- 限制节点分片数 index.routing.allocation.total_shards_per_node 一般为1或2
- Mapping
  - norms 不需要计算字段的评分，将该参数设为false
  - ignore_malformed 忽略这个字段错误并保留该文档中的其他字段，设为ture
  - _source 文档原始内容，可以考虑删除
  - dynamic_templates 动态模板将默认的字符串类型映射为keyword
  - 如果不需要按数字大小排序或过来，不使用int类型，使用keyword

日志收集系统

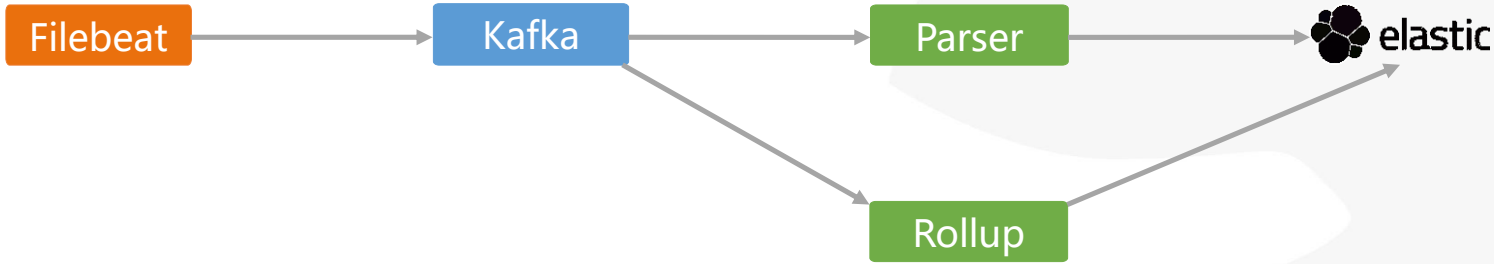


- 定时新建索引
- 冷热索引转换
- 强制合并索引
- 定时关闭索引
- 定时删除索引

索引管理

- 索引生命周期
  - 定时提前新建下一天用的空索引
  - 冷热数据分离，48小时以前的索引数据搬移到冷节点
  - 索引变冷后（无写入后），进行强制合并（ForceMerge）
  - 定时关闭和删除索引
- 合理切分索引和制定分片大小，避免节点故障后索引恢复速度慢的问题
- 日志收集场景下写请求远远大于读请求，尽量均匀分配分片在每个节点上，分摊写压力
- 节点磁盘独立，避免节点间产生IO竞争

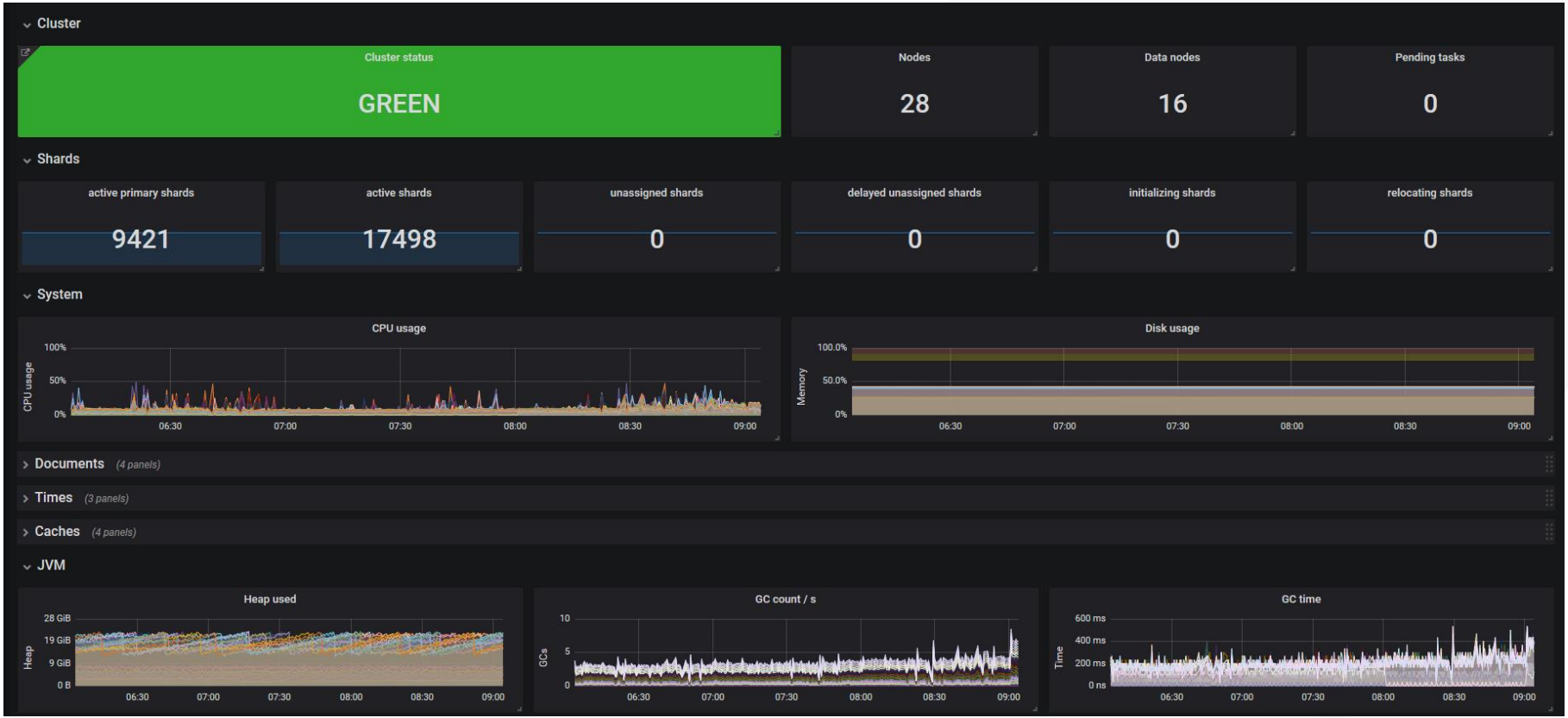
预聚合



- 使用日志进行报警的场合
  - 使用Spark Streaming进行指标预聚合(Rollup)
  - 每一批(10000条)聚合一次，按每30秒级别的聚合结果为一条记录，写入预聚合索引
  - 正在预研Flink/Blink进行聚合

斗鱼 DOUYU.COM

监控可视化



THANKS!

斗鱼全民直播平台
DOUYU.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/