xiaomi | elastic

小米ElasticSearch服务平台化演进

贺祥/张鑫刚

2019-04-21

目录
CONTENTS

发展过程

2017/03
ES 5.2

- 测试+立项（为公司提供es服务）
- 集群个数:5
- 实例数:20
- 业务数:50
- 集群总数据量:百亿条/20T
- 数据写入量(天):1亿条/1T

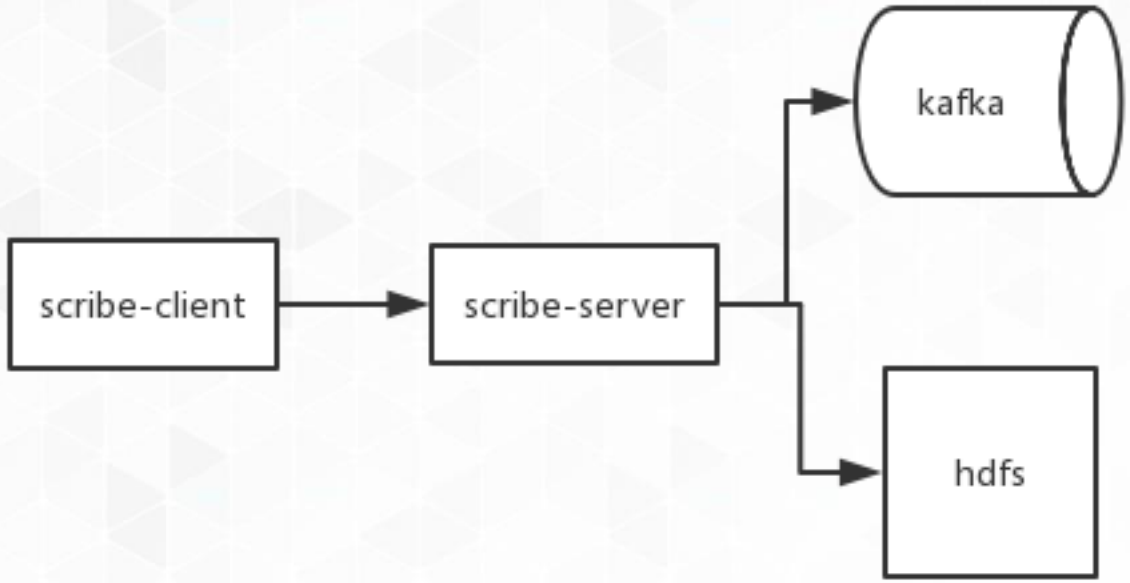
2017/11
ES 5.6

- es平台化
- 集群个数:15
- 实例数:200
- 业务数:500
- 集群总数据量:千亿条/200T
- 数据写入量(天):百亿/10T

2018/11-至今
ES 6.4/6.7

- es服务化
- 集群个数:物理集群30/容器集群10
- 实例数:500
- 业务数:1000
- 集群总数据量:万亿/近1PB
- 数据写入量(天):千亿条/80T

一期选型



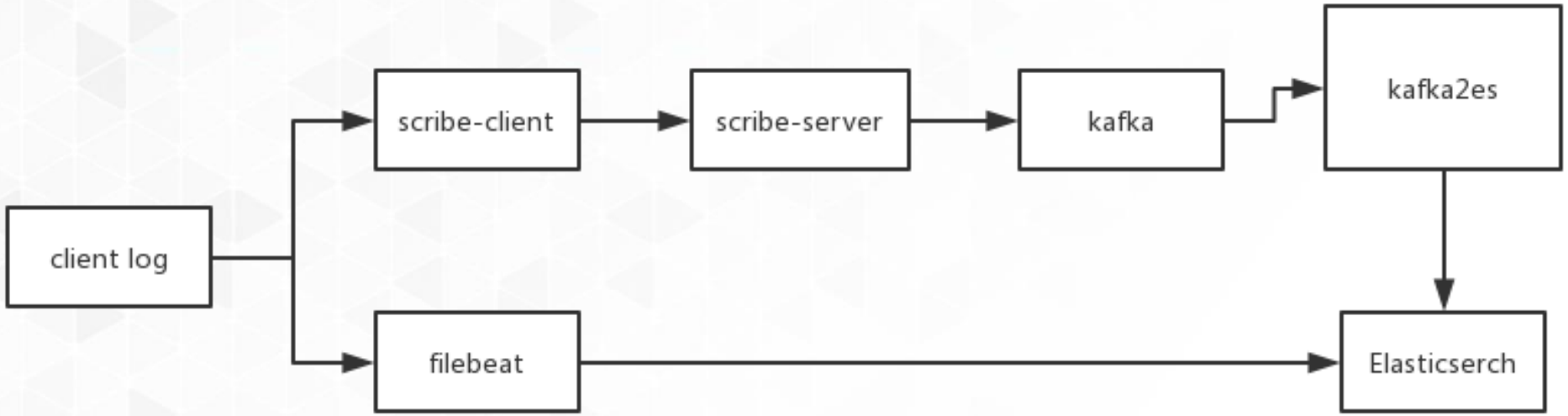
简图

数据如何接入到ES?

- logstash?
- filebeat?

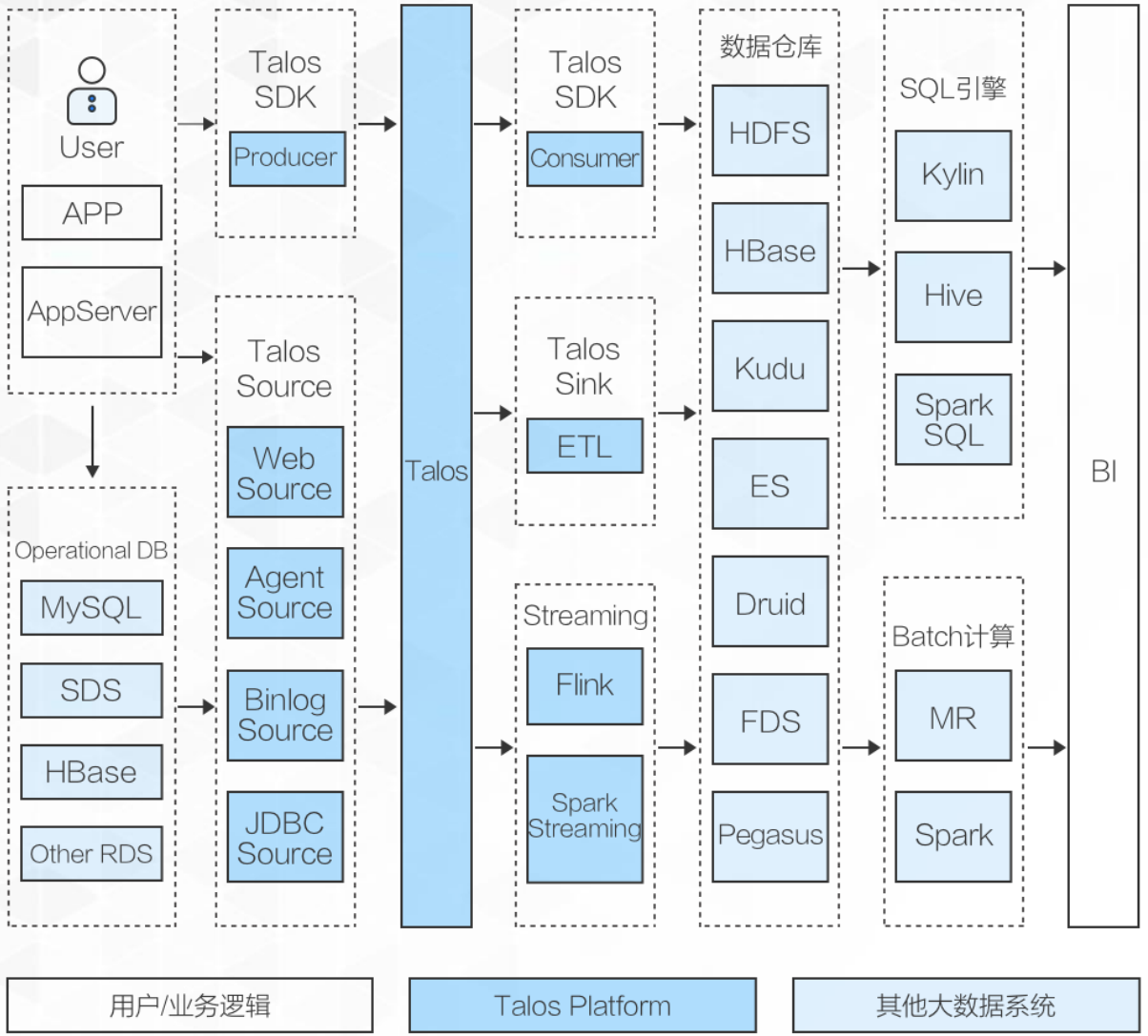
我米大部分数据为thrift格式，如何解析?

写入ES

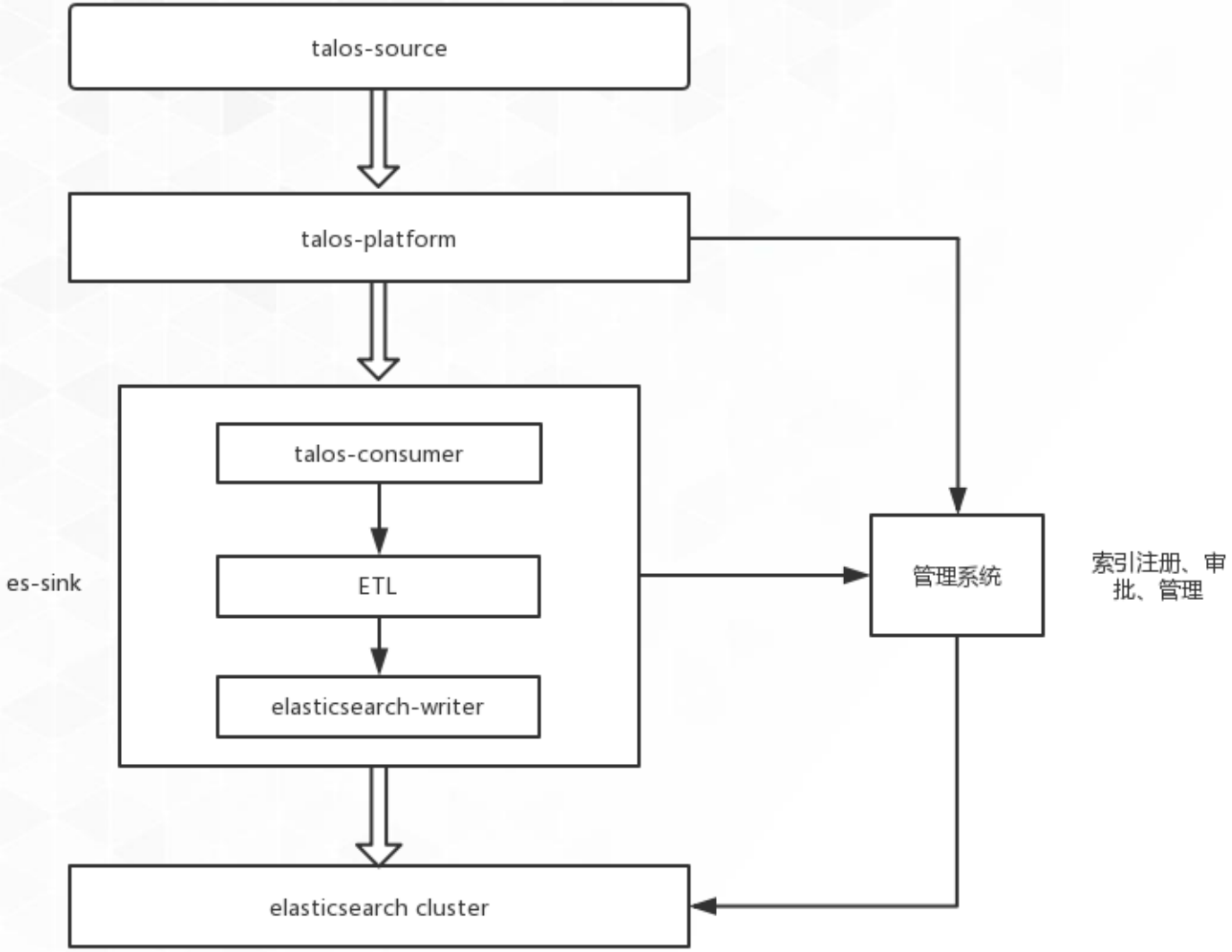


最初版接入Elasticsearch 数据链路图

数据流架构升级


User
APP
AppServer
Operational DB
MySQL
SDS
HBase
Other RDS
Talos SDK
Producer
Talos Source
Web Source
Agent Source
Binlog Source
JDBC Source
Talos
Talos SDK
Consumer
Talos Sink
ETL
Streaming
Flink
Spark Streaming
数据仓库
HDFS
HBase
Kudu
ES
Druid
FDS
Pegasus
SQL引擎
Kylin
Hive
Spark SQL
Batch计算
MR
Spark
BI
用户/业务逻辑
Talos Platform
其他大数据系统

es-sink


talos-source
talos-platform
talos-consumer
ETL
elasticsearch-writer
es-sink
管理系统
索引注册、审批、管理
elasticsearch cluster

接入申请

Elasticsearch集群	类型 / 集群	*
索引名称	部门_组_业务_描述	*
数据描述	Enter ...	*
数据工场Hive表编号		
数据源	talos	
ETL Class		
ES Pipeline		
数据保留时间/天	7	

以下为高级设置:

索引切分规则	按天	
自动管理索引	是	
分片大小/GB	30	*
副本数	1	
热数据转移时间/天	3	
HDFS快照	否	

帮助文档

提交

kafka2es VS es-sink

kafka2es

功能与特点
- 对接消息队列，数据实时处理并写入es集群
- 支持用户自定义ETL
- 支持用户自定义ingest pipeline
- 无状态高并发，单实例支持10W/s的数据处理

存在的问题
- 应用需手动部署，且每个业务需部署多个实例做一定的灾备冗余
- 随着业务增多，资源消耗较大，无法做到资源动态扩缩，维护困难
- 所有配置和代码都是通过git维护，管理繁琐

es-sink

功能与特点
- kafka2es功能的继承
- spark streaming实现
- 自动化部署+监控
- 处理延时动态感知，支持动态扩缩容

存在的问题
- 实时性相比kafka2es略差，实为批处理模式
- 动态扩缩容敏感度暂时无法做到自定义配置，对于延时非常敏感的业务，堆积扩容也需要分钟级

ES集群管理者的建议

- **数据准入流程严格规范**

数据格式、查询需求、保留时间、用户隐私、确定sla

- **数据写入的收敛与控制**

尽可能禁止业务直连集群写数据，实时数据通过消息队列缓冲

- **权限控制**

- **监控、告警全方位覆盖**

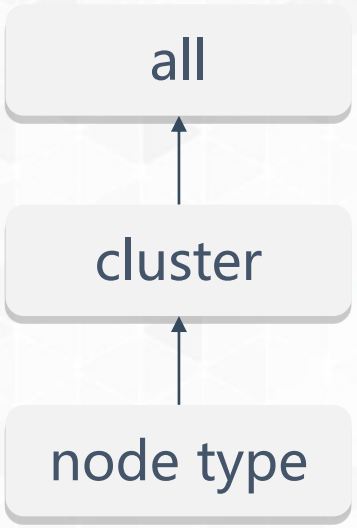
查询监控：保留所有查询记录、监控慢查询
集群使用率监控：线程池、队列、索引内存、cpu、内存、磁盘、jvm gc等等

目录
CONTENTS

自动化运维：部署

Ansible

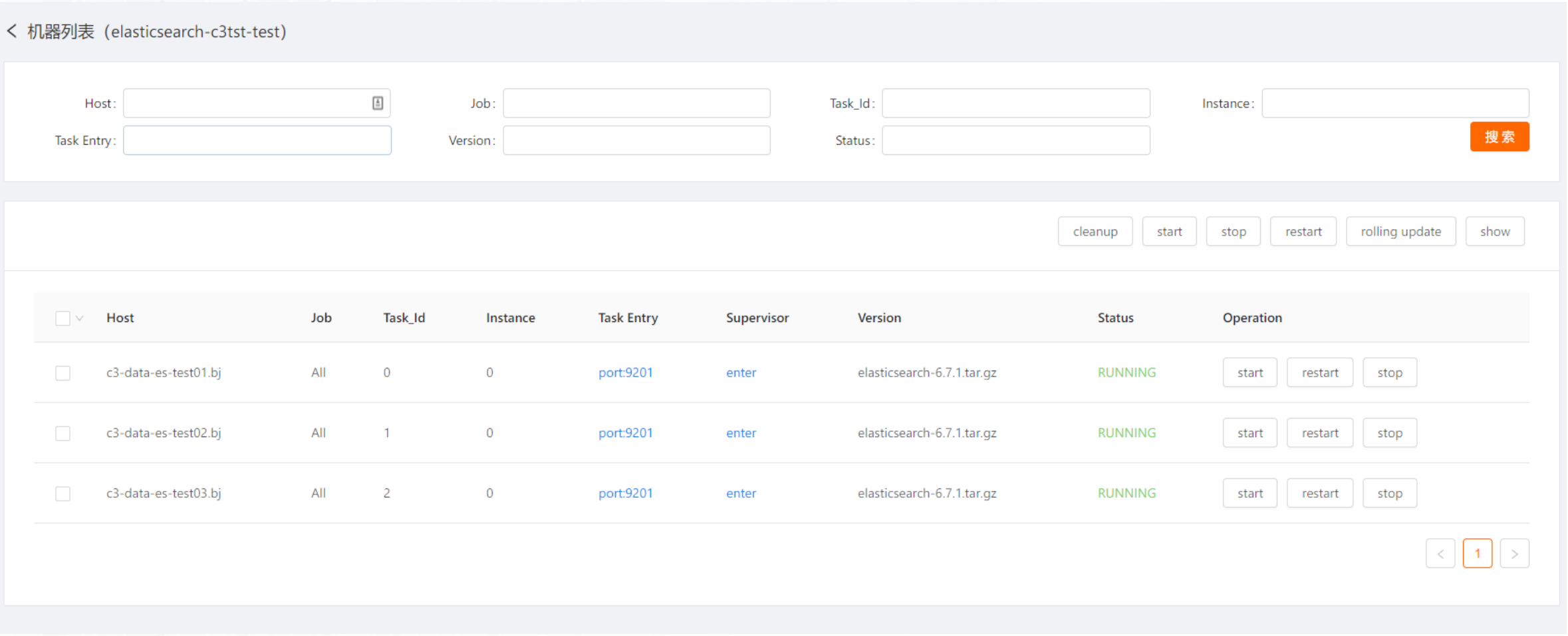
使用group_vars管理集群配置



Minos 2.0

Web交互式部署
控制实例启停
部署一机多实例
集群配置的版本管理

自动化运维

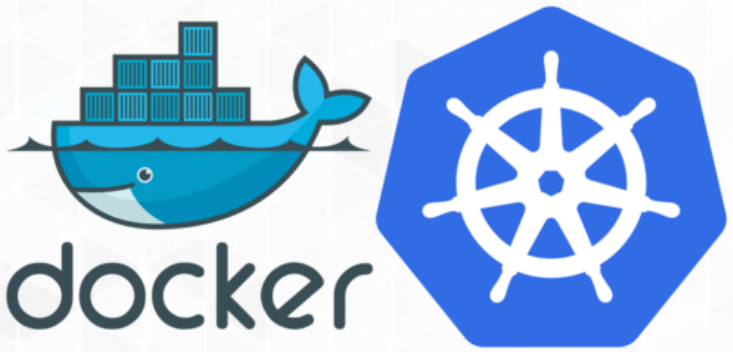

机器列表（elasticsearch-c3tst-test）
Host:
Job:
Task_Id:
Instance:
Task Entry:
Version:
Status:
搜索
cleanup
start
stop
restart
rolling update
show
Host
Job
Task_Id
Instance
Task Entry
Supervisor
Version
Status
Operation
c3-data-es-test01.bj
All
0
0
port:9201
enter
elasticsearch-6.7.1.tar.gz
RUNNING
start
restart
stop
c3-data-es-test02.bj
All
1
0
port:9201
enter
elasticsearch-6.7.1.tar.gz
RUNNING
start
restart
stop
c3-data-es-test03.bj
All
2
0
port:9201
enter
elasticsearch-6.7.1.tar.gz
RUNNING
start
restart
stop
1

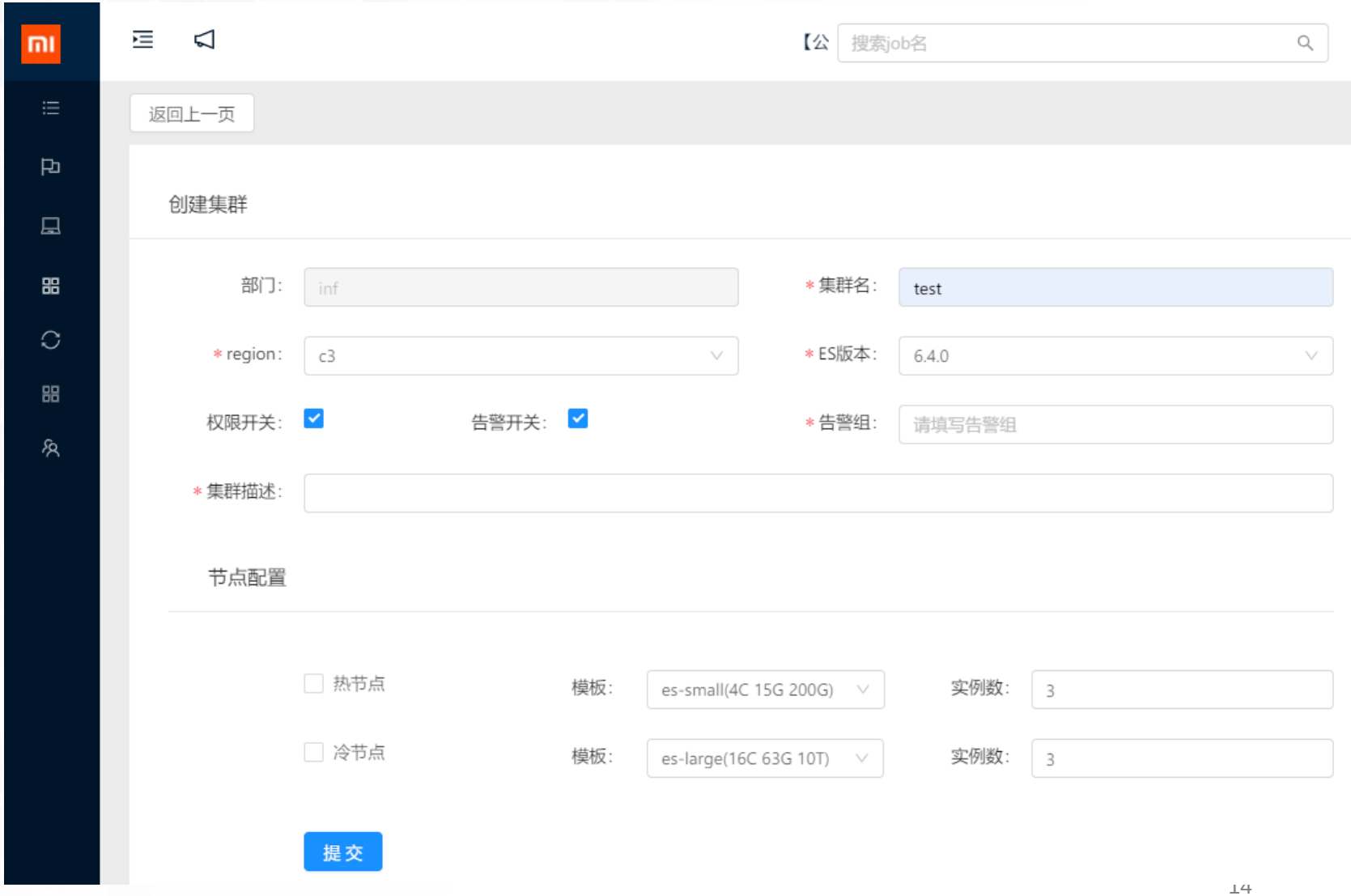
自动化运维：ES on Ocean

背景

- 用户对独立集群的需求大
- 不同配置及规模的集群
- 部署升级效率需提高

ES容器化





运维管理:成本优化

Master节点混布
- Heap：4G

数据冷热分离
- 热节点：
  4T SSD RAID50/RAID0
- 温节点：
  44T SATA 单盘/RAID10
- 冷节点：
  88T SATA 单盘/RAID10

多盘推荐使用LVM/RAID组成单盘

1TB存储：2GB内存
多实例部署，充分利用物理内存
使用cgroup/numactl进行资源隔离

ES 6.6: Frozen Index
无常驻内存，每次搜索动态加载

ES 监控

监控架构

收集模块
- ES_Monitor
- Filebeat

存储模块
- Elasticsearch
- Falcon

可视化
- Grafana
- Kibana

关键告警

- 健康状态
- 写入可用性
- 查询可用性
- 内存使用率
- 线程池拒绝数
- 最大分片大小

其他功能

- Kill异常查询和邮件
- 状态异常原因邮件
- 自动扩分片数

……

ES 数据管理

集群信息

- 集群名字
- 集群属性
- API地址
- 机房信息

……

+

索引信息

- 索引名字
- 所属集群
- 保留时间
- 主副本数
- 分片大小

……

ES_Manager

- 预创建索引
- 删除过期数据
- 转移冷数据
- 优化大索引
- 用户报表
- 管理员报表
- 备份Template

……

索引管理: 别名

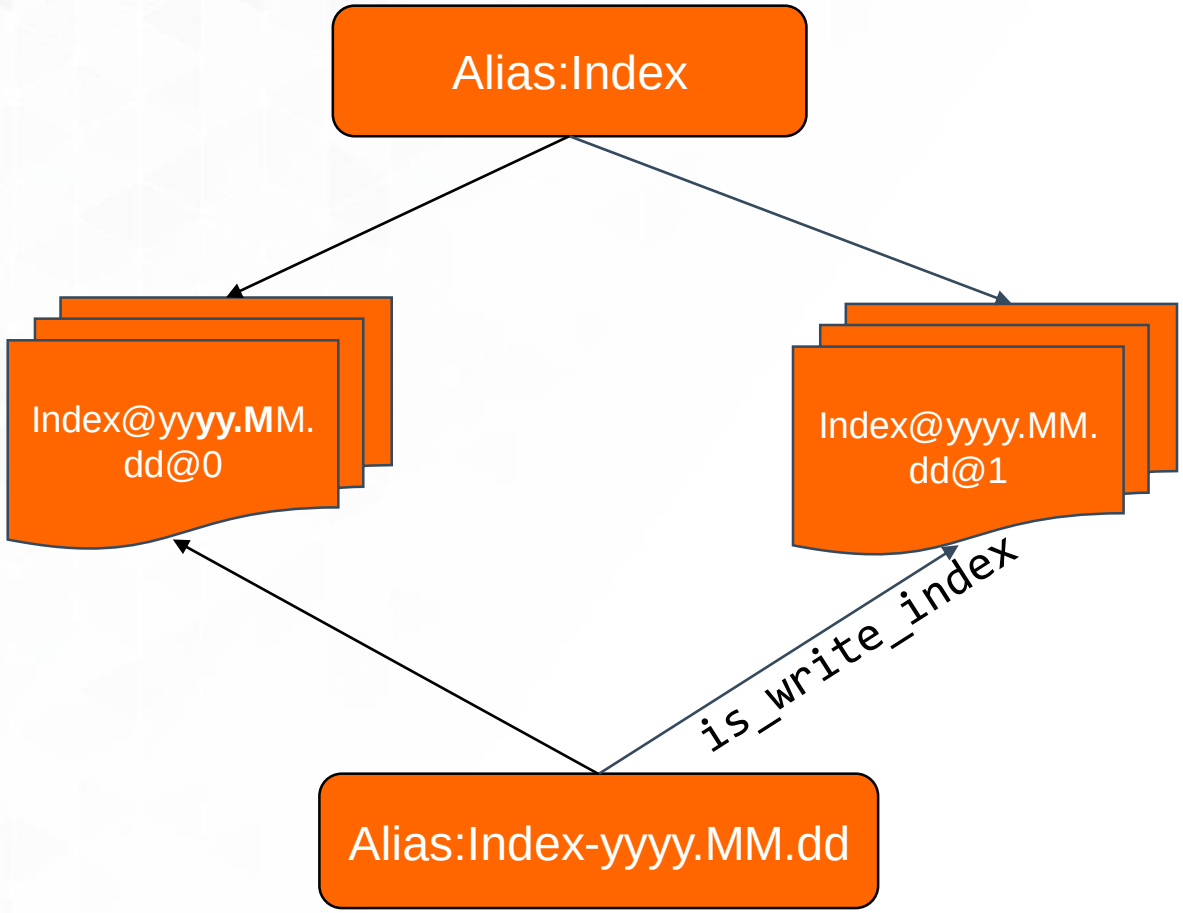
问题
Mapping确定后无法更改
分片数量增减不方便

缺乏灵活性

索引名:
index@yyyy.MM.dd@N

别名
1. index
2. index-yyyy.MM.dd

配合`is_write_index`参数



优化：JVM

JVM
- GC
- OOM

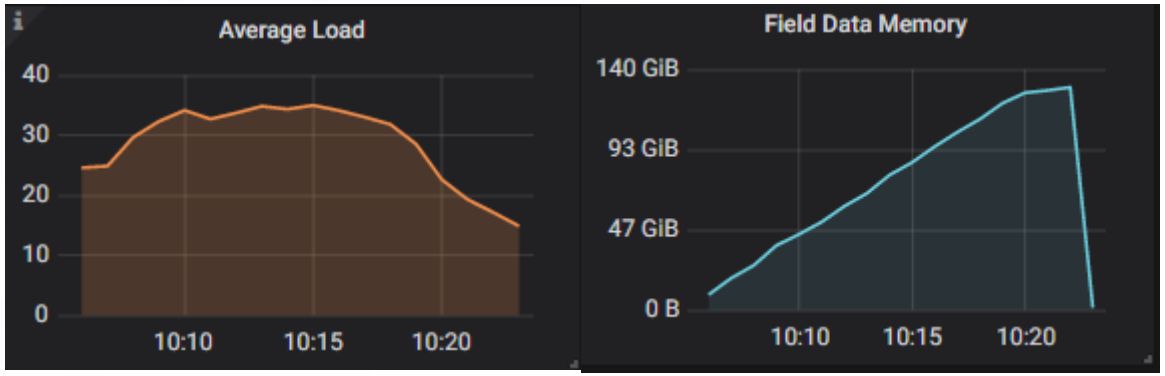
解决方案
- G1 GC
- 压缩指针 -> 32736M≈31.9G
  零基压缩 -> 30720M
- 参数限制:
  - search.max_buckets
  - Kibana: Disable suggest
  - indices.fielddata.cache.size
- 升级到ES7.x

```
> ES_HEAP_SIZE=32736m ./bin/elasticsearch

heap address: 0x00007f8c10000000, size: 31744 MB, Compressed Oops
mode: Non-zero based:0x00007f8c0ffff000, Oop shift amount: 3
```

```
> ES_HEAP_SIZE=30720m ./bin/elasticsearch

heap address: 0x0000000080000000, size: 30720 MB, Compressed Oops
mode: Zero based, Oop shift amount: 3
```



优化：索引

问题

- Shard数量
- 分配不匀
- 大索引写入消耗太多资源

解决方案

- 根据前几天索引的加权均大小计算后一天的Shard数量
- 设置total_shards_per_node
- 大于1T的数据针对写入优化

大索引参数

```
{
  "index.merge.scheduler.max_merge_count": "32",
  "index.merge.scheduler.max_thread_count": "1",
  "index.merge.policy.max_merged_segment": "2gb",
  "index.merge.policy.floor_segment": "20mb",
  "index.routing.allocation.total_shards_per_node": "8",
  "index.translog.durability": "async",
  "index.translog.flush_threshold_size": "2g",
  "index.translog.sync_interval": "30s",
  "index.store.type": "niofs"
}
```

Thanks&QA

专业、垂直、纯粹的 Elastic 开源技术交流社区

https://elasticsearch.cn/