



# 基于Kibana和ES的苏宁实时日志分析平台

苏宁云商IT总部技术总监-彭燕卿

2016.11.21

Elastic{ON} DevChina – Dec 10, 2016

# Agenda

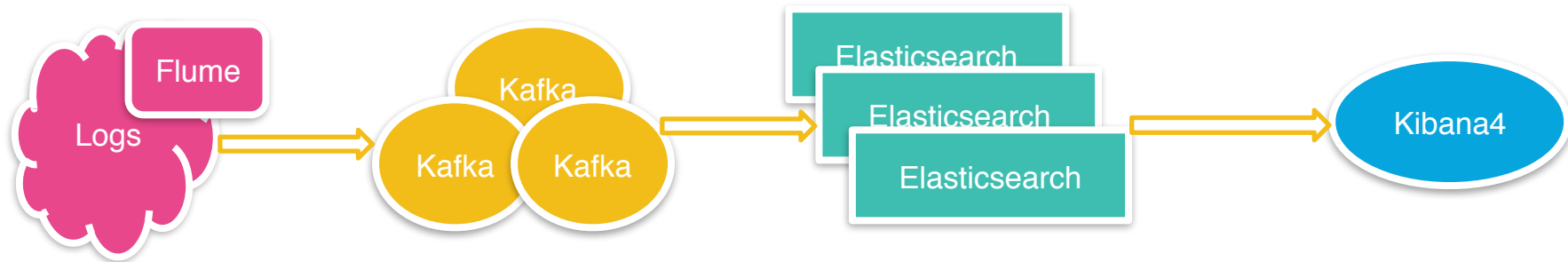
- 集群现状
- 日志平台架构演进
- 日常优化总结
- 运维小技巧
- Kibana4二次开发

## 集群现状

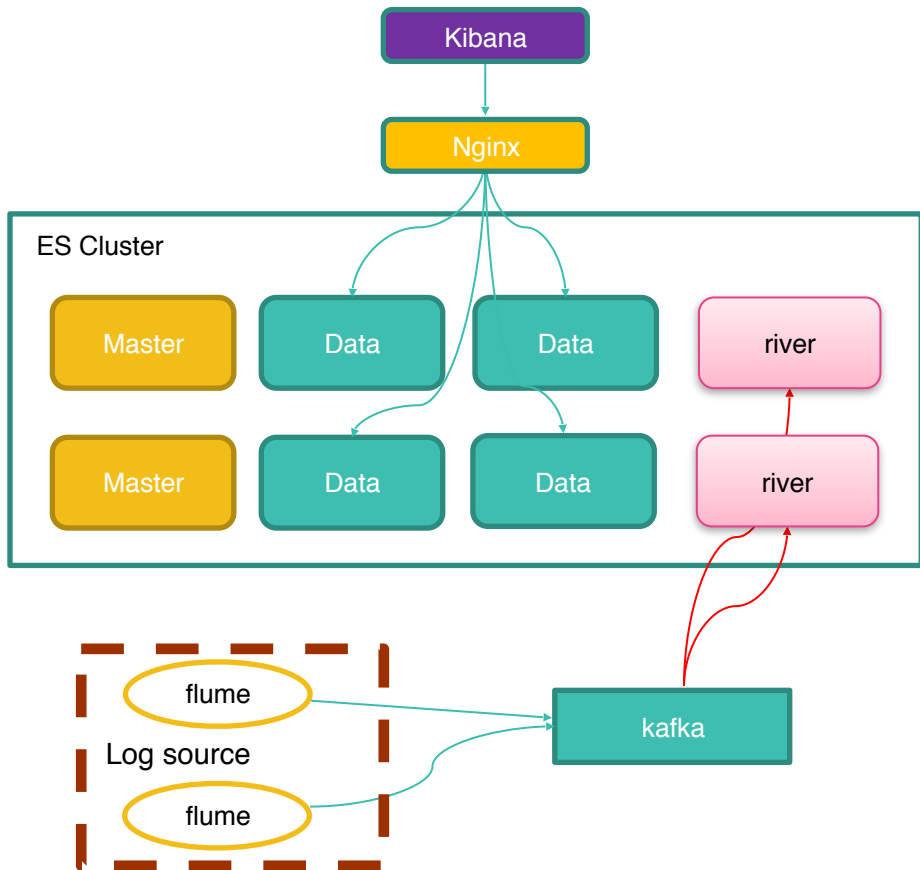
- 126个数据节点，7个cluster，12C/128G/2T SATA、16C/128G/3T SSD、12C/128G/16T
- 接入苏宁近2000个系统的应用日志、web访问、缓存、应用防火墙等日志
- 大促每天新索引25T数据，doc数超过450亿条
- open 1100索引、130T、2500亿数据、20000 shard、7天存储
- 峰值90W/s数据写入
- 平均每个doc 0.6kb

# 整体架构

- 实时日志平台采用的是flume+Kafka+Elasticsearch+Kibana的部署架构。
- 与ELK架构有点不同，我司使用flume实时采集日志，Kafka作为数据通道，ES river插件消费Kafka里面的数据，将Kafka中的数据清洗过滤后，index到ES集群中。



# 日志平台架构演进-①



## 配置:

- 虚拟机节点
- 按天生成索引

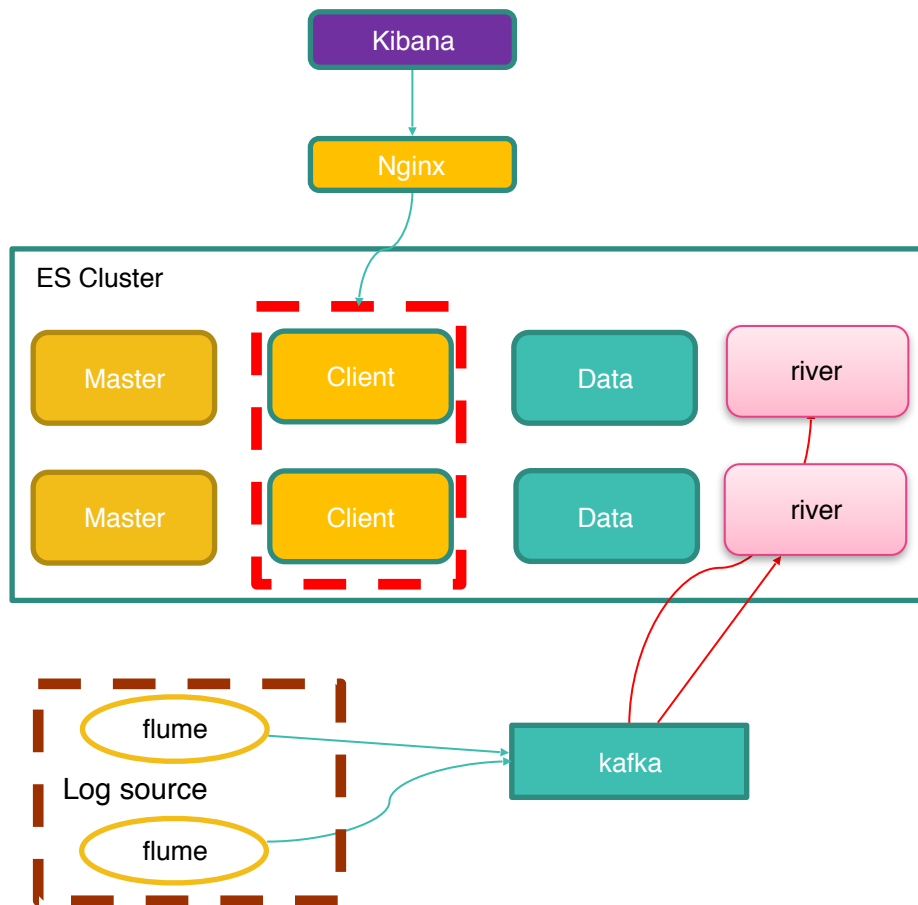
## 问题:

- 部分数据节点负载非常高
- 数据消费延时
- 查询响应慢
- QPS低

## 原因:

- 同一个主机上存在多个ES 虚拟机Node
- 数据节点同时承担索引和检索， 负荷重
- enabled\_all
- 按天索引体量大
- 集群节点少

## 日志平台架构演进-②



### 主要优化:

- 增加client节点
- 解决同一物理机上多虚拟机data节点
- 增加部分物理机
- 关闭\_all字段,
- 小时生成索引
- 根据日志类型划分不同的索引

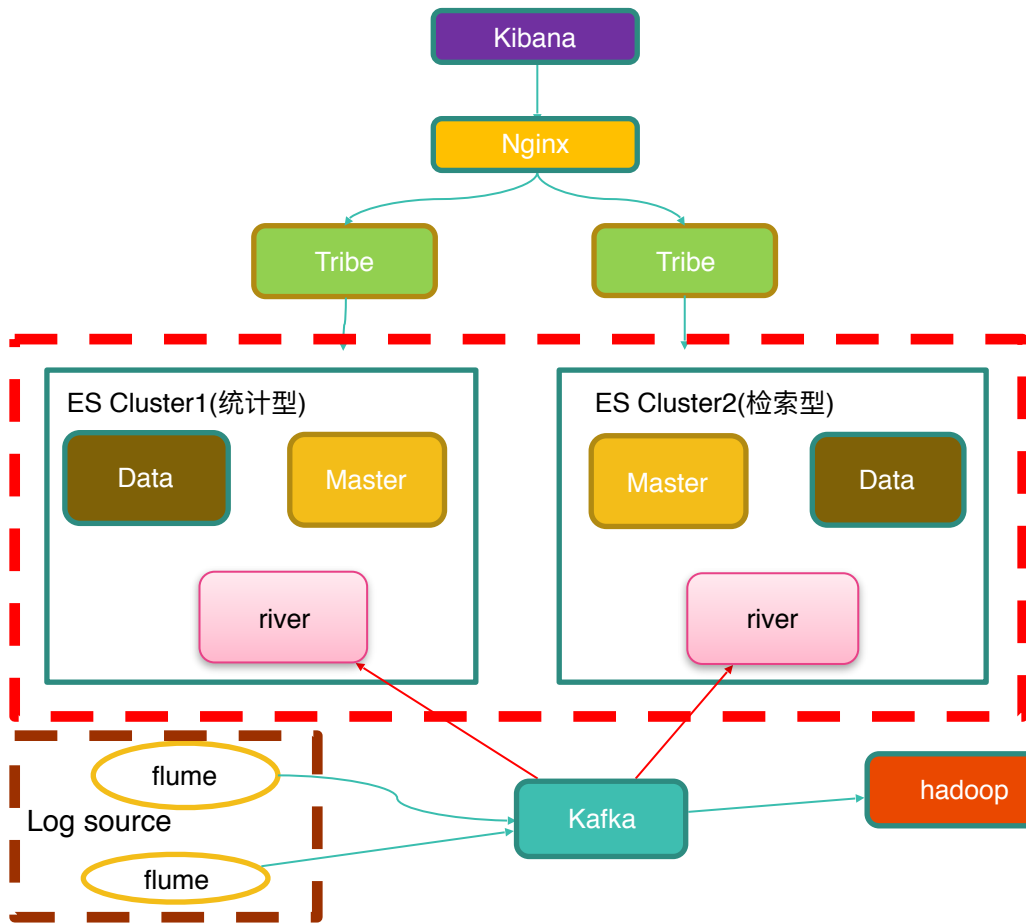
### 运行状况:

- 非大促期间, 索引和检索速度基本能达到秒级
- 大促期间, 日志量膨胀以及访问人数过多时, 索引和检索速度任然很慢

### 原因分析:

- 单集群能力受限
- 不同类型的数据混在一个大集群中, 互相影响。
- client 节点性能提升不明显
- 虚拟机和物理机混合, 制约物理机的能力

# 日志平台架构演进-③



## 主要优化:

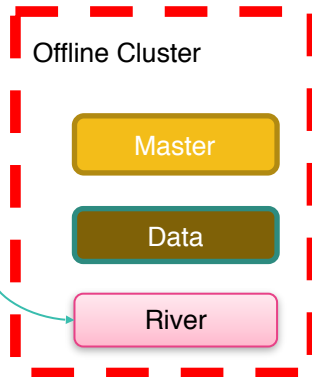
- 使用tribe做多集群路由
- 根据不同的分析类型进行集群拆分
- 将data节点全部替换成物理机
- 提供按照系统、文件路径等应对大促期间的日志洪峰系统进行降级的功能
- 统计型集群使用SSD

## 运行状况:

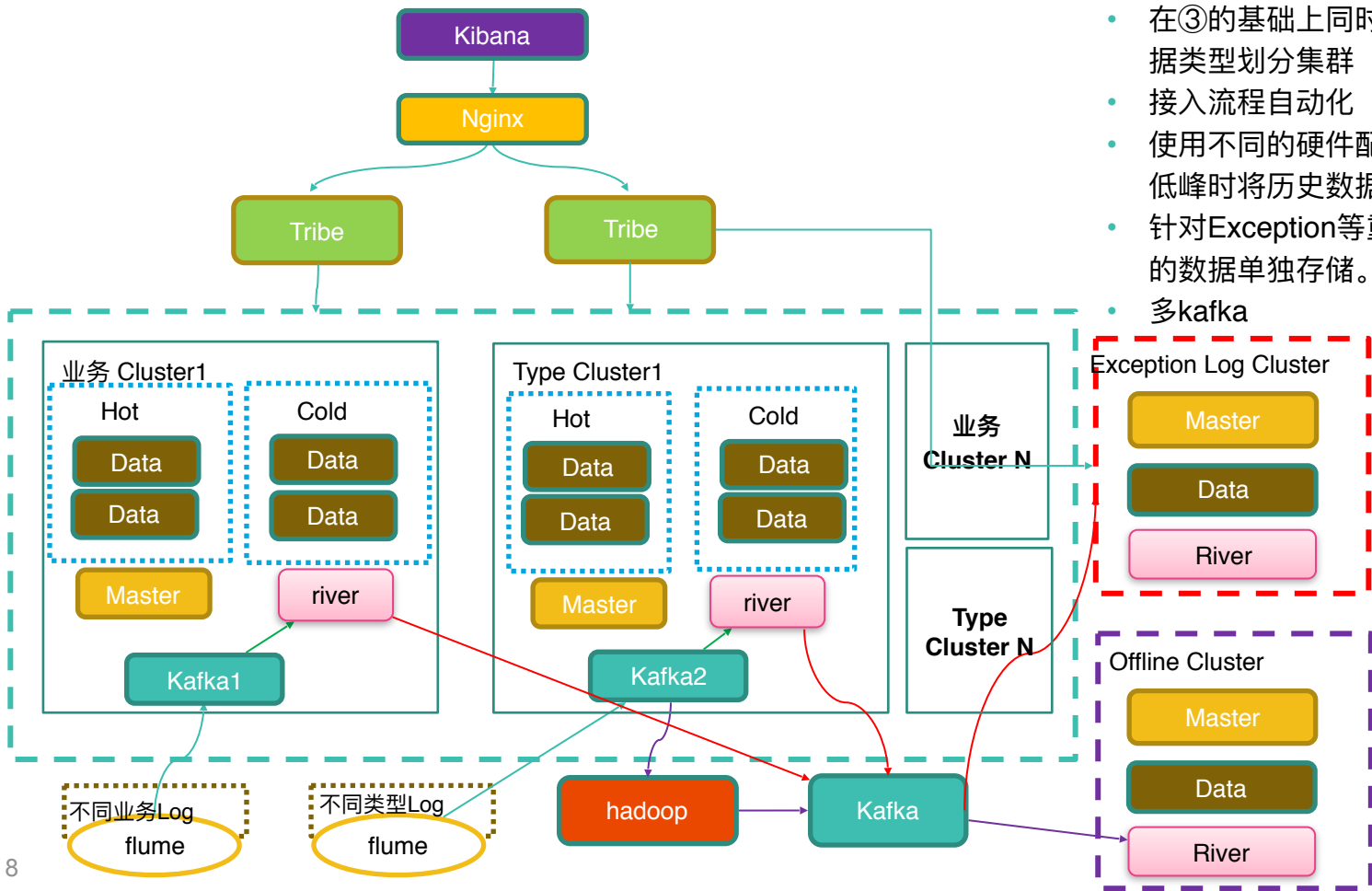
- 流量的剧增通过降级, 能够保证核心系统数据的实时性

## 存在的问题:

- 随着业务量的快速增长,简单的按照分析类型进行集群拆分已不能满足要求,并且增加机器资源容量并不是线性增加。
- 用户希望保留7天以前以及大促期间数据,集群容量以及性能也存在严重问题



# 日志平台架构演进-现状



## 主要优化:

- 在③的基础上同时支持按照业务划分以及数据类型划分集群
- 接入流程自动化
- 使用不同的硬件配置划分Hot/Cold节点,业务低峰时将历史数据迁移到Cold节点
- 针对Exception等重点关注且长期保留分析的数据单独存储。

## 问题:

- 多kafka
- 跨多索引进行统计分析性能差以及对集群压力任然很大



# 日常优化总结-硬件

- 优先独立物理机
- 对于实时性要求非常高的需求，优先SSD
- 适当调整OS的max\_file\_descriptors,解决Too many open files 异常
- 单服务器运行多个node时,调整max user processes, 否则容易native thread OOM.
- 关闭swap交换或锁内存 `ulimit -l unlimited/bootstrap.mlockall: true`

## 日常优化总结-ES

- 根据数据量合理的规划索引pattern和shard数
- disabled `_all` 节省存储空间、提升索引速度
- 不需要分词的字段设成 `not_analyzed`
- 对于不要求100%高可用的内部系统，可不设置副本，提升index速度和减少存储

# 日常优化总结-ES

- 设置合理的refresh时间

`index.refresh_interval: 300S`

- 设置合理的flush间隔

`index.translog.flush_threshold_size: 4g`

`index.translog.flush_threshold_ops: 50000`

- 合理配置throttling

`indices.store.throttle.max_bytes_per_sec: 200mb`

- 适当调整bulk队列

11 `threadpool.bulk.queue_size: 1000`

# 日常优化总结-ES

- 有时可能因为gc时间过长，导致该数据节点被主节点踢出集群的情况，导致集群出现不健康的状态，为了解决这样的问题，我们适当的调整ping参数。(master)

discovery.zen.fd.ping\_timeout: 40s

discovery.zen.fd.ping\_interval: 5s

discovery.zen.fd.ping\_retries: 5

- 调整数据节点的JVM新生代大小

数据节点young gc频繁,适当调转新生代大小 (-Xmn3g)，降低young gc的频率。

- 在进行检索和聚合操作时，ES会读取反向索引，并进行反向解析，然后进行排序，将结果保存在内存中。这个处理会消耗很多Heap，有必要进行限制，不然会很容易出现OOM。

Disabled analyzed field fielddata

限制Field Data的Heap Size的使用

indices.fielddata.cache.size: 40%

indices.breaker.fielddata.limit: 50%

```
    "properties": { -  
      "message": { -  
        "store": true,  
        "fielddata": { -  
          "format": "disabled"  
        }  
      },  
      "analyser": "ik",  
      "type": "string"  
    },  
    "createTime": 0 -
```

# ES运维小技巧

## 增加节点

- 调整shard数

```
index.routing.allocation.total_shards_per_node: 2
```

index在每个node的shard数据

(如果后期需要移除节点,保证每个node有可分配的shard)

## 移除节点

- 移除node前可以先exclude要移除的node

```
cluster: cluster.routing.allocation.exclude._name : node1
```

```
index: index.routing.allocation.exclude._name: node1
```

```
index.routing.allocation.require.node_type: hot
```

以上参数可根据\_ip、\_host等来进行配置

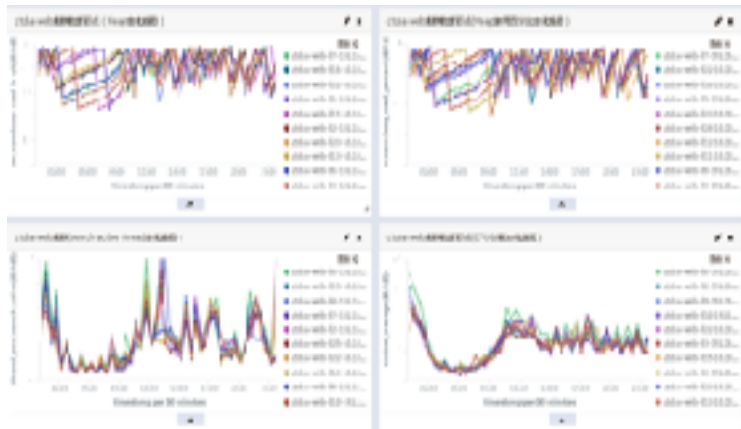
以上参数可实现hot-cold cold数据的自动迁移

# ES运维小技巧-工具

- 监控插件满天飞,各有千秋

Head、Kopf、bigdesk、elasticsearch-sql(NLPchina)

- 定时关闭和删除index: curator
- 基于python脚本实现采集ES集群指标数据,并使用kibana展示

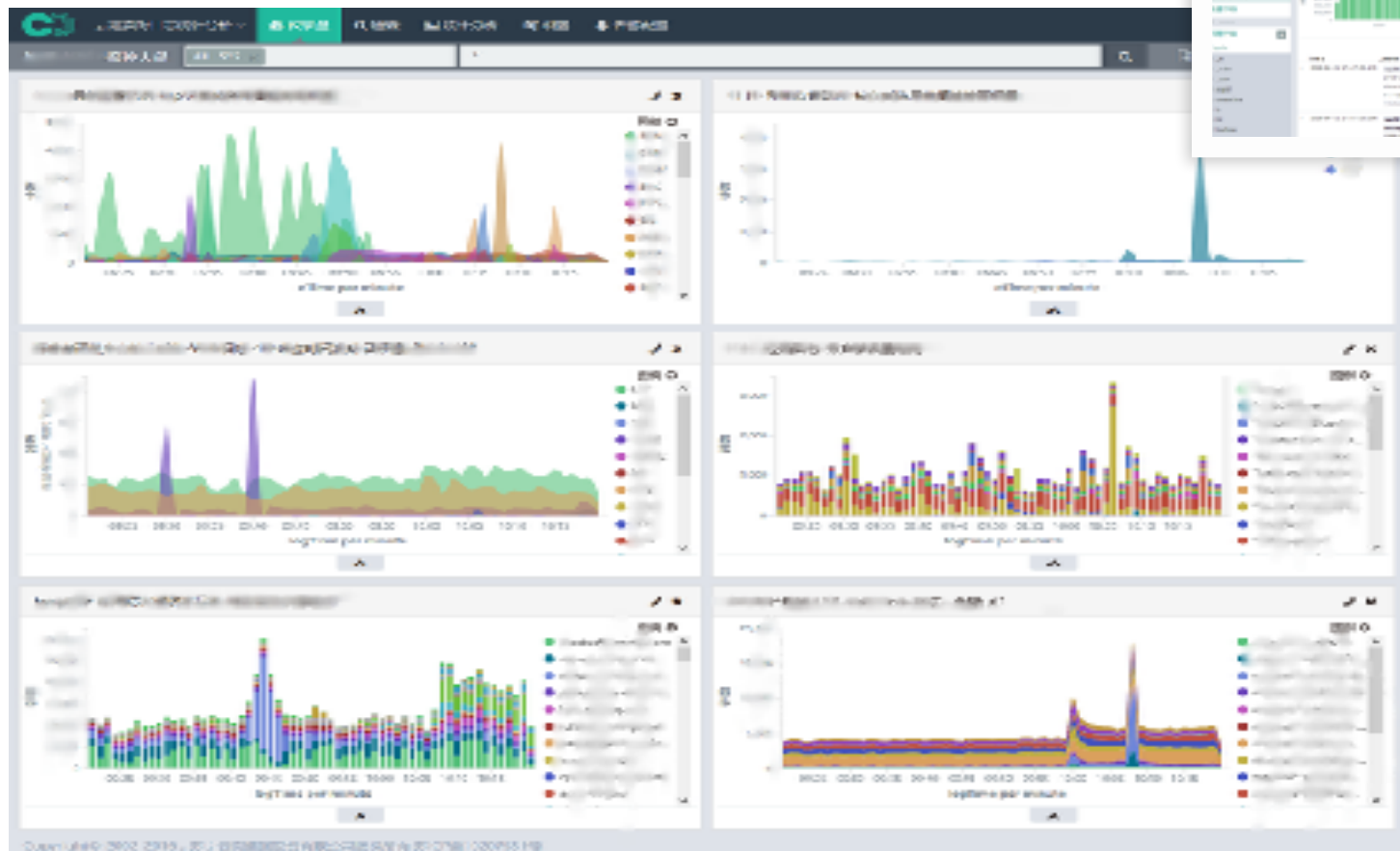


- 日志平台自监控ES日志,重点对以下关键词进行监控告警

- OutOfMemoryError
- removed AND cluster.service                    node 退出cluster
- unable to create new native thread
- master\_left    master退出cluster

- Slow log监控,重量级query可能把整个集群拖慢,对slow log重点监控分析(待实现)

# Kibana二次开发-从汉化开始



# Kibana二次开发-权限

- nodejs实现cas单点登录

```
app.use('/', cas.balancer, ctlsa, systest, checkkys1st);
```

- 授权数据权限



- 不同业务可查询的数据范围

- 禁用通过kibana访问\_plugin、\_shutdown等

- 用户关联 仪表盘、检索、统计分析



# Kibana二次开发-discover可查询更多

- Discover可以前后查询更多数据，解决查询discover:sampleSize行数限制

外数据查看的问题

The screenshot shows the Kibana Discover interface. On the left, there are filters for '选择系统名' (Select System Name), '选择日志类型' (Select Log Type), and '选择字段' (Select Fields). The main area displays a bar chart titled 'logTime per 5 seconds' with a y-axis labeled '计数' (Count) ranging from 0 to 300. The x-axis shows time intervals from 19:00:00 to 19:13:00. A prominent bar is visible at 19:12:30. Below the chart, a message states '以下是符合条件的 500 条数据。加载更多 回到顶部' (Below are 500 records that meet the conditions. Load More Return to Top). A red box highlights the '加载更多' button. Below this, a table of log entries is shown with columns for Time, sortNum, and message.

Time	sortNum	message
2016-06-22 19:12:00.826	4,665,340,000,300,001	{"name":"ctdsa-front-web","hostname":"ctdsaprdapp01","pid":8004,"level":30,"req":{"method":"POST","u":383,200 - 108ms","time":"2016-06-22T11:13:20.409Z","v":10}}
2016-06-22 19:13:20.456	4,665,940,004,560,001	{"name":"ctdsa-front-web","hostname":"ctdsaprdapp208","pid":15272,"level":30,"req":{"method":"POST","u":585,200 - 155ms","time":"2016-06-22T11:13:20.214Z","v":0}}
2016-06-22 19:13:20.369	4,665,940,003,690,001	{"name":"ctdsa-front-web","hostname":"ctdsaprdapp206","pid":15436,"level":30,"req":{"method":"POST","u":383,200 - 72ms","time":"2016-06-22T11:13:19.992Z","v":0}}



# Kibana二次开发-禁用check es version

- Kibana

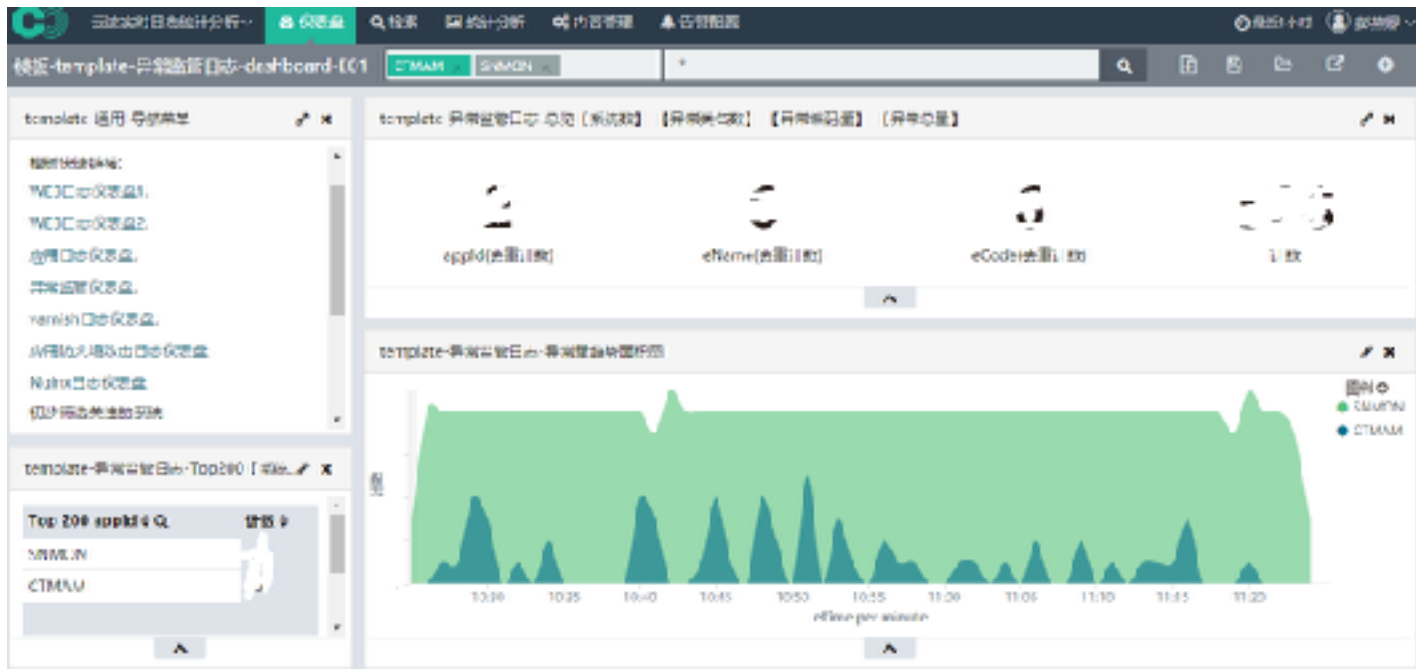
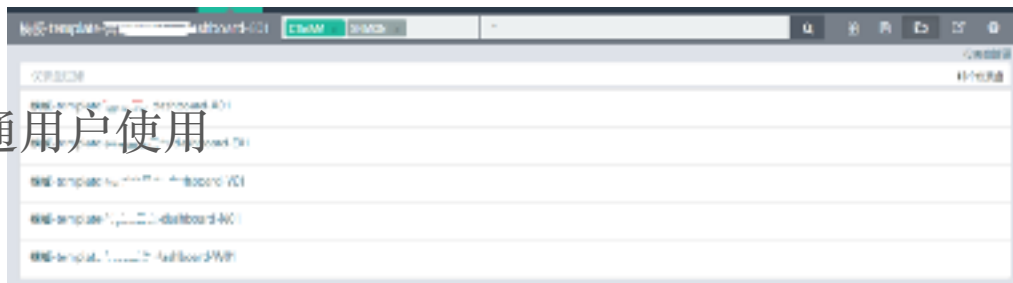
- 当使用  
得不到

- 可去掉

```
7   var mnhify = new Mnhify({
8     location: 'setup: elasticsearch version check'
9   });
10
11  return actify.times(function checkEsVersion() {
12    /*
13     * var SetupError = Private(require('components/setup/_setup_error'));
14     *
15     * return es.nodes.info()
16     *   .then(function (info) {
17     *     var badNodes = _.filter(info.nodes, function (node) {
18     *       // remove client nodes (Legstash)
19     *       var isClient = _.get(node, 'attributes.client');
20     *       if (isClient != null && isClient === true) {
21     *         return false;
22     *       }
23     *
24     *       // remove nodes that are greater than the min version
25     *       var v = node.version.split('-')[0];
26     *       return !versionGreaterThan(minimumElasticsearchVersion, v);
27     *     });
28     *
29     *     if (!badNodes.length) return true;
30     *
31     *     var badNodeNames = badNodes.map(function (node) {
32     *       return 'Elasticsearch ' + node.version + ' @ ' + node.kibana_address + ' (' + node.ip + ')';
33     *     });
34     *
35     *     throw SetupError(
36     *       'This version of Kibana requires Elasticsearch ' +
37     *       minimumElasticsearchVersion + ' or higher on all nodes. ' +
38     *       'I found the following incompatible nodes in your cluster:\n\n' +
39     *       badNodeNames.join('\n')
40     *     );
41     *   });
42     *
43     *   return true;
44     * }
45   });
```

# Kibana二次开发-dash-board

- 管理员可分享dashboard模板给普通用户使用
- dashboard可选择系统查看数据



# Kibana二次开发-other

- indexPattern映射为中文
- 禁用分词字段的统计分析(误操作容易导致长GC甚至OOM,亲身经历),如果template中指定fileddata为disabled, 进行统计分析或排序会报异常。
- 根据不同的数据类型默认展示不同的默认field
- 修改默认排序字段,根据自定义seqNum排序
- visualize 支持模板共享
- Discover 左侧field快速计数展示更多



```
});  
//每2小时刷新一次页面  
setInterval(function(){  
  window.location.reload();  
},1800*60*120); //指定120分钟刷新一次*/  
/*print*/
```

sh问题

# NEXT

- 升级ES5
- 多数据中心
- 服务化

**Thanks!**  
**Questions?**