苏宁ES平台化实践之路

苏宁大数据平台.搜索平台组 韩宝君

2018年6月

苏宁易购 suning.com | 进极2018 ULTIMATE CREATION

关于我们

苏宁云商IT总部
大数据平台研发中心

大数据平台职责：

提供苏宁集团各个业务所需要的大数据存储和计算能力。
保证平台的稳定、高效运行。
提高平台易用性。

个人职责：

Elasticsearch组件负责人

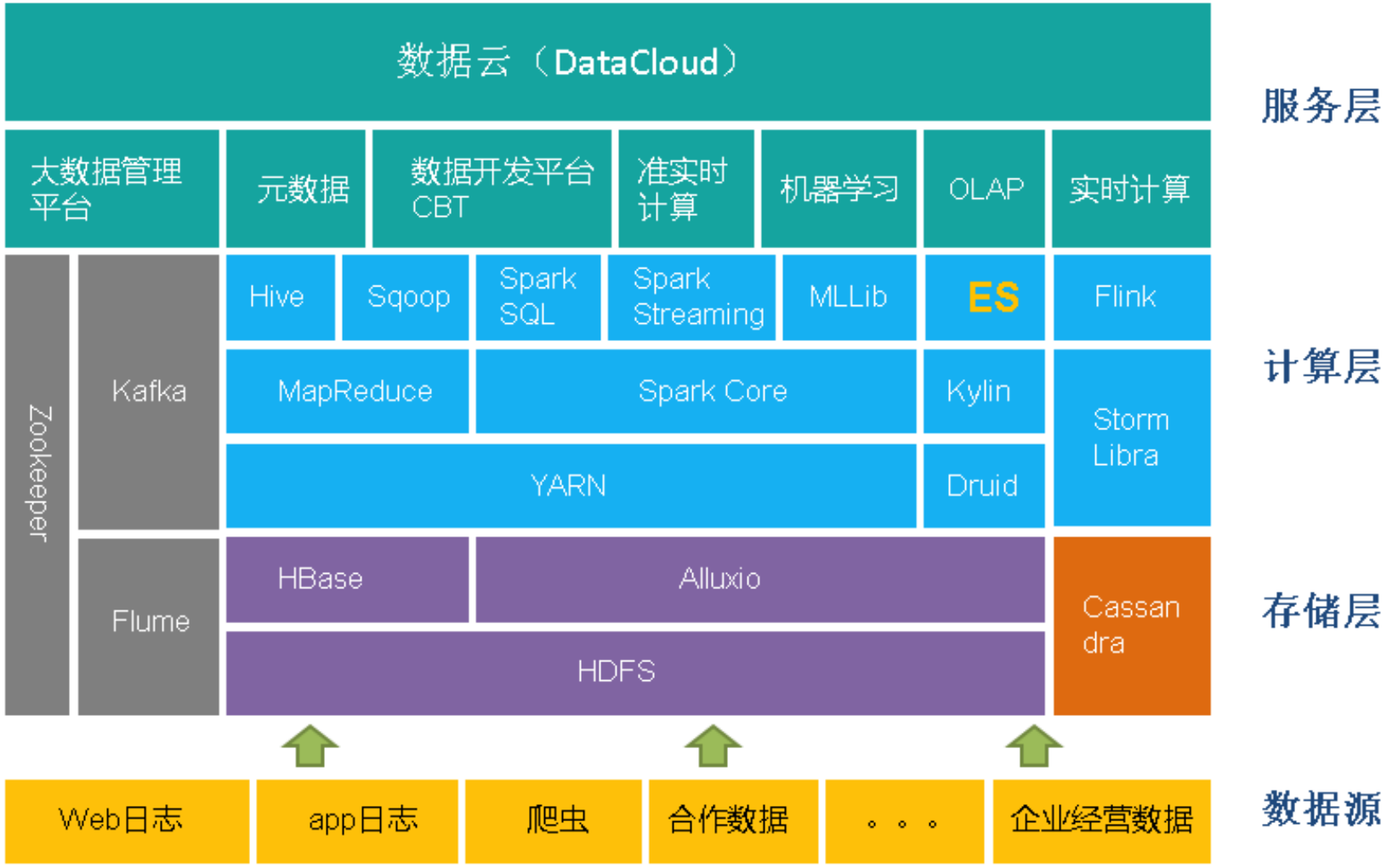
苏宁易购 suning.com

1.ES平台总体介绍

2.ES平台化之路

3.实战经验

苏宁大数据平台总体架构



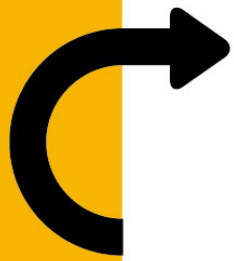
Elasticsearch平台概况

集群规模

18个集群，195个节点
接入100+个业务，4500+个索引
数据量64+TB

平台功能

独占服务/共享服务：资源利用率和业务隔离的权衡
计量计费：强调成本概念
页面化服务：使用便利性，减少误操作，节约时间



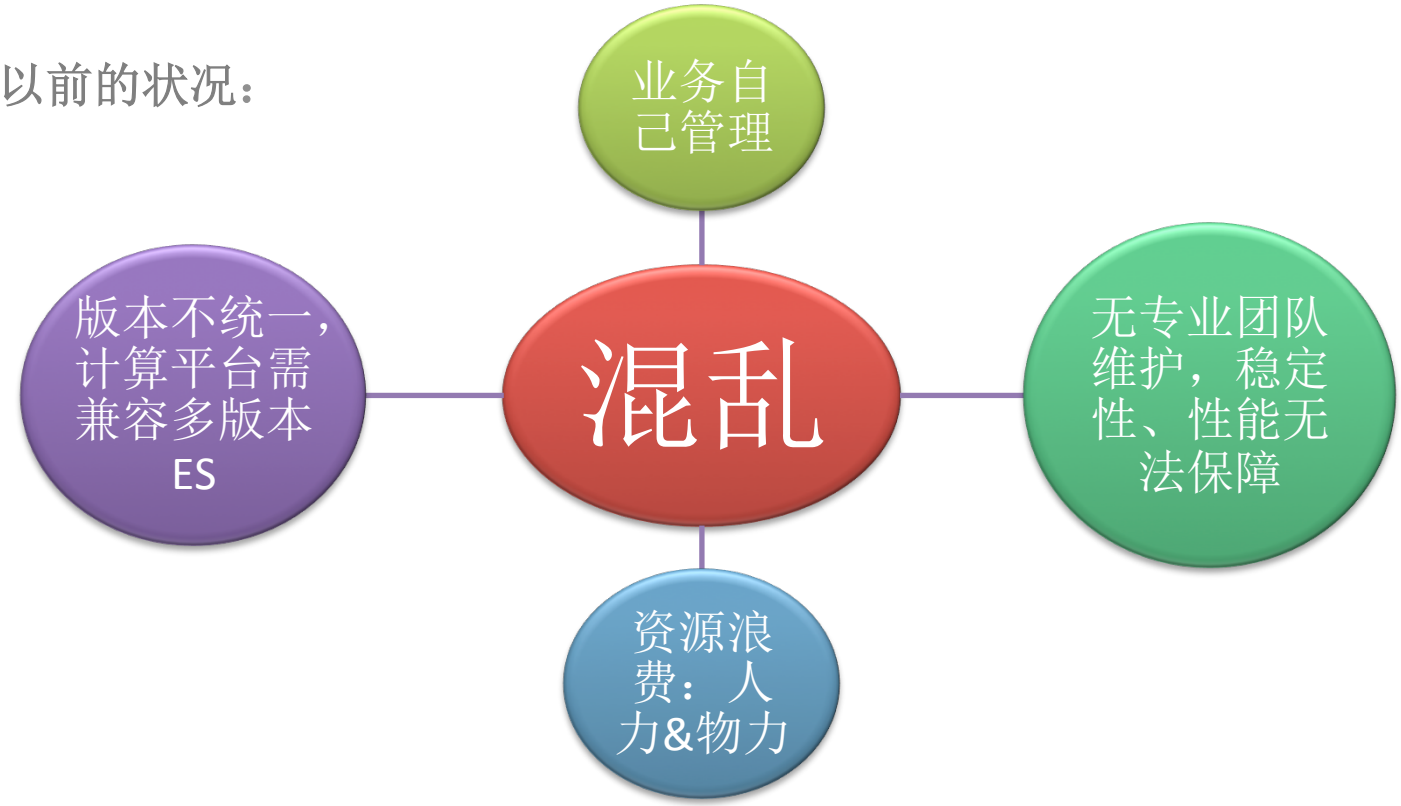
1.ES平台总体介绍

2.ES平台化之路

3.实战经验

为什么做ES平台？

一年以前的状况：



目标：做一个专业的ES平台

两个阶段

人肉部署集群

人肉对接业务

人肉运维集群

自动化部署集群

申请服务页面化

计量计费功能

可用性、性能监控

安全增强

数据探查

纯人肉时代

真正的平台化服务

人肉阶段的对接文档

ES 业务对接

·1、业务信息

一级中心	部门名称	业务对接人	工号

·2、项目信息

项目名称	
项目英文简称	
项目介绍	

·3、应用场景

使用方式	□实时入库 □离线入库 □结构化查询 □排序 □全文检索 □聚合计算 □高级查询 □导出
场景描述	

·4、数据量级

数据量级	每天新增数据数量	每天新增数据大小	大促每天新增数据数量	大促每天新增数据大小
	初始化数据数量	初始化数据大小	保存时间	检索 QPS
	实时入库 TPS	批量入库间隔	承受检索延时范围	承受导出延时范围

人肉阶段的对接文档

索引(INDEX)	类型(TYPE)	字段名	数据类型	字段值	主键	备注
kbmp_knowledge	answer（子文档）	id	integer	47165		精确匹配
		sq_id	integer	13972		精确匹配
		answer	text	信用卡转入涉及资金套现问题呢，所以零钱包是不支持信用卡转入的哦，不过您可以使用借记卡转入。		全文搜索，模糊查询(IK分词)
		scene_code	keyword	10002		精确匹配，分组字段
		scene	keyword	非辅助应答		精确匹配，排序字段
		terminal_code	keyword	20001		精确匹配
		terminal	keyword	pc		不查询
		create_dttm	date	2017-10-16 17:18:11		时间区间，范围搜索，排序字段
		validity_date	date	2030-01-01		时间区间，范围搜索，
		validity_dttm	date	2030-01-01 00:00:00		时间区间，范围搜索，
		is_permanent	integer	1		精确匹配，聚合字段

服务化平台

监控管理： ES集群的各项监控指标展示

集群管理： 支持多集群，可以查看具体集群信息和节点列表

项目管理： 使用ES服务的项目列表，分配给各自系统的秘钥

资源管理： 机器列表，该机器归属于哪些集群及部署哪些软件包

软件包管理： 支持选择的ES版本相关软件包

索引管理： 该系统下的索引列表，索引的详细信息

工单管理： 用户可以提交相应的工单，管理员审批

计量管理： 系统、集群、索引的请求量和存储量及详情

查询管理： 可以像kibana一样查询es

监控管理

监控管理

集群巡检 集群监控 节点监控 索引监控 线程池监控

刷新周期 15s

系统 所有系统 机房 所有机房 告警级别 所有告警 集群 所有集群

查询

Clusters	Nodes	Indices	Memory
18	195	4554	2117GB/5138GB

Total Shards	Data	Unassigned Shards	Documents
19501	65T	394	1218亿

触发时间	解决时间	所属中心	所属部门	管理员	机房	集群名称	集群状态	集群类型	风险项	风险说明
2018-05-24T13:29:19.000Z	2018-06-29T06:16:02.000Z						red	独占	拒绝任务. 分片未分配. 部分...	拒绝任务:
2018-05-24T06:39:26.000Z	2018-06-29T06:16:02.000Z						red	独占	拒绝任务. 拒绝任务. 拒绝任务...	拒绝任务:

资源管理

资源管理 > 机器列表

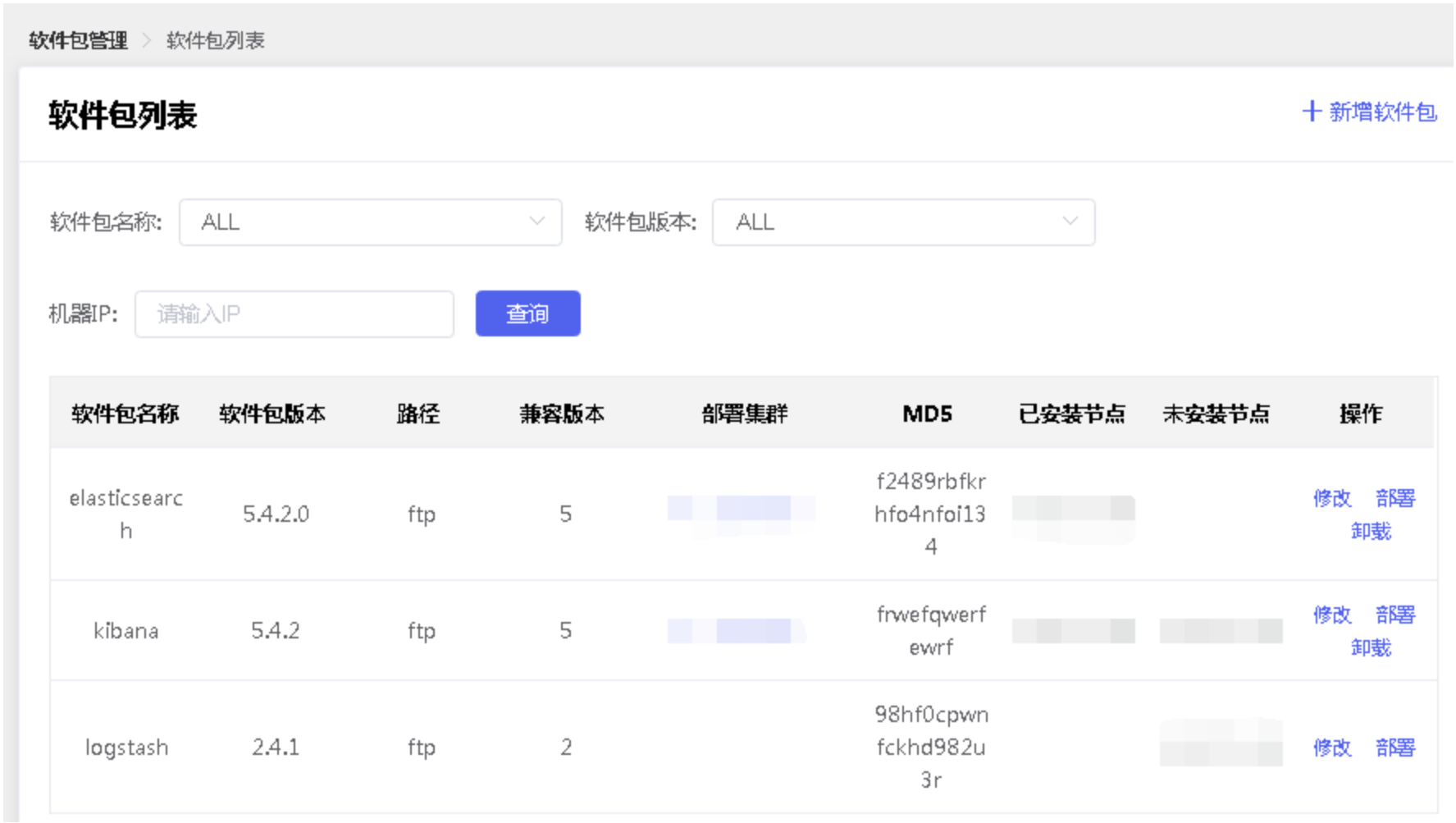
机器列表 + 新增资源

机房: ALL 机器类型: ALL 机器标签: ALL

机器状态: ALL 部署软件包: ALL 部署版本: ALL 查询

主机名	节点IP	节点CPU	节点内存	节点磁盘	节点端口	归属集群	部署版本	操作
		32	131072 M	172032/data00 281600/data01 281600/data02	9100-9300		elasticsearch: 5.4.2.0	修改配置 删除
		32	131072 M	172032/data00 281600/data01 281600/data02	9100-9300		elasticsearch: 5.4.2.0	修改配置 删除
		32	131072 M	172032/data00 281600/data01 281600/data02	9100-9300		elasticsearch: 5.4.2.0	修改配置 删除
		32	131072 M	172032/data00 281600/data01 281600/data02	9100-9300		elasticsearch: 5.4.2.0	修改配置 删除

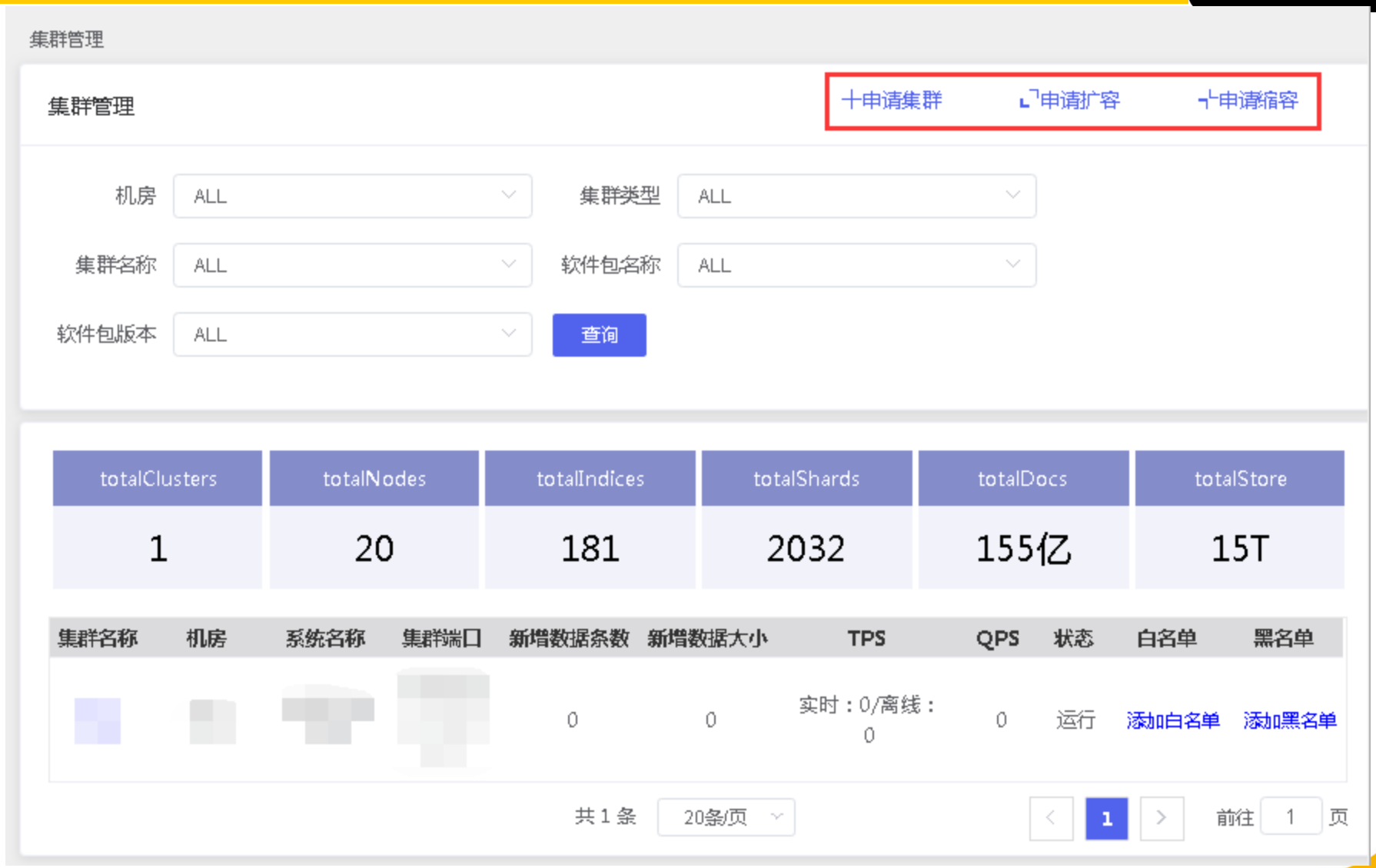
软件包管理


软件包管理 > 软件包列表
软件包列表
+ 新增软件包
软件包名称: ALL
软件包版本: ALL
机器IP: 请输入IP
查询
软件包名称 软件包版本 路径 兼容版本 部署集群 MD5 已安装节点 未安装节点 操作
elasticsearch 5.4.2.0 ftp 5 f2489rbfkrhfo4nfoi134 修改 部署 卸载
kibana 5.4.2 ftp 5 frwefqwerfewrf 修改 部署 卸载
logstash 2.4.1 ftp 2 98hf0cpwnfckhd982u3r 修改 部署

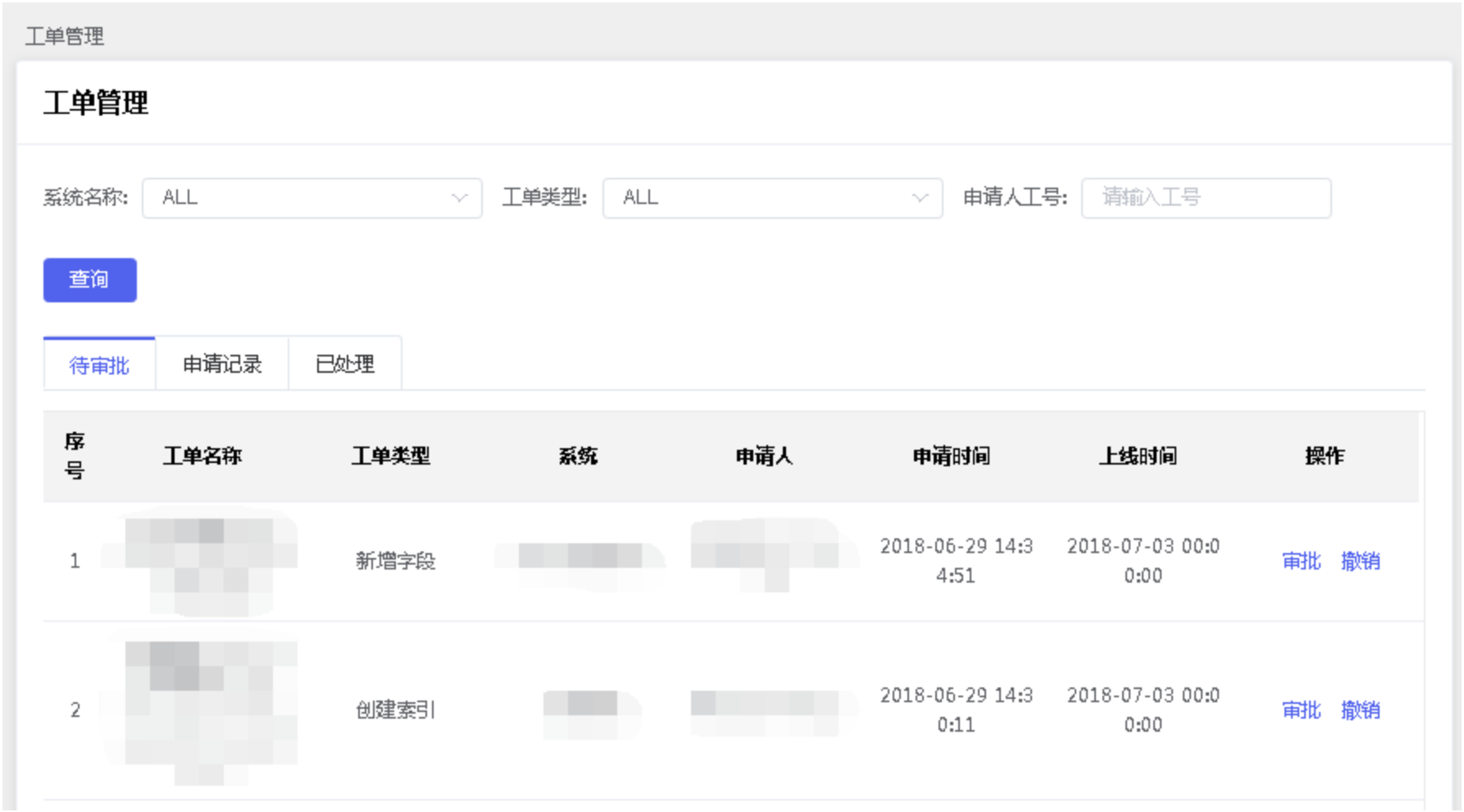
集群管理-管理员视图


集群管理
集群管理
+创建集群
扩容集群
缩容集群
系统名称
ALL
机房
ALL
集群类型
ALL
集群名称
ALL
软件包名称
ALL
软件包版本
ALL
查询
totalClusters
18
totalNodes
195
totalIndices
4554
totalShards
19501
totalDocs
1219亿
totalStore
65T
集群名称
机房
系统名称
集群端口
新增数据条数
新增数据大小
TPS
QPS
状态
白名单
黑名单
操作
删除集群
0
0
实时：0/离线：0
0
运行
添加白名单
添加黑名单
0
0
实时：0/离线：0
0
运行
添加白名单
添加黑名单

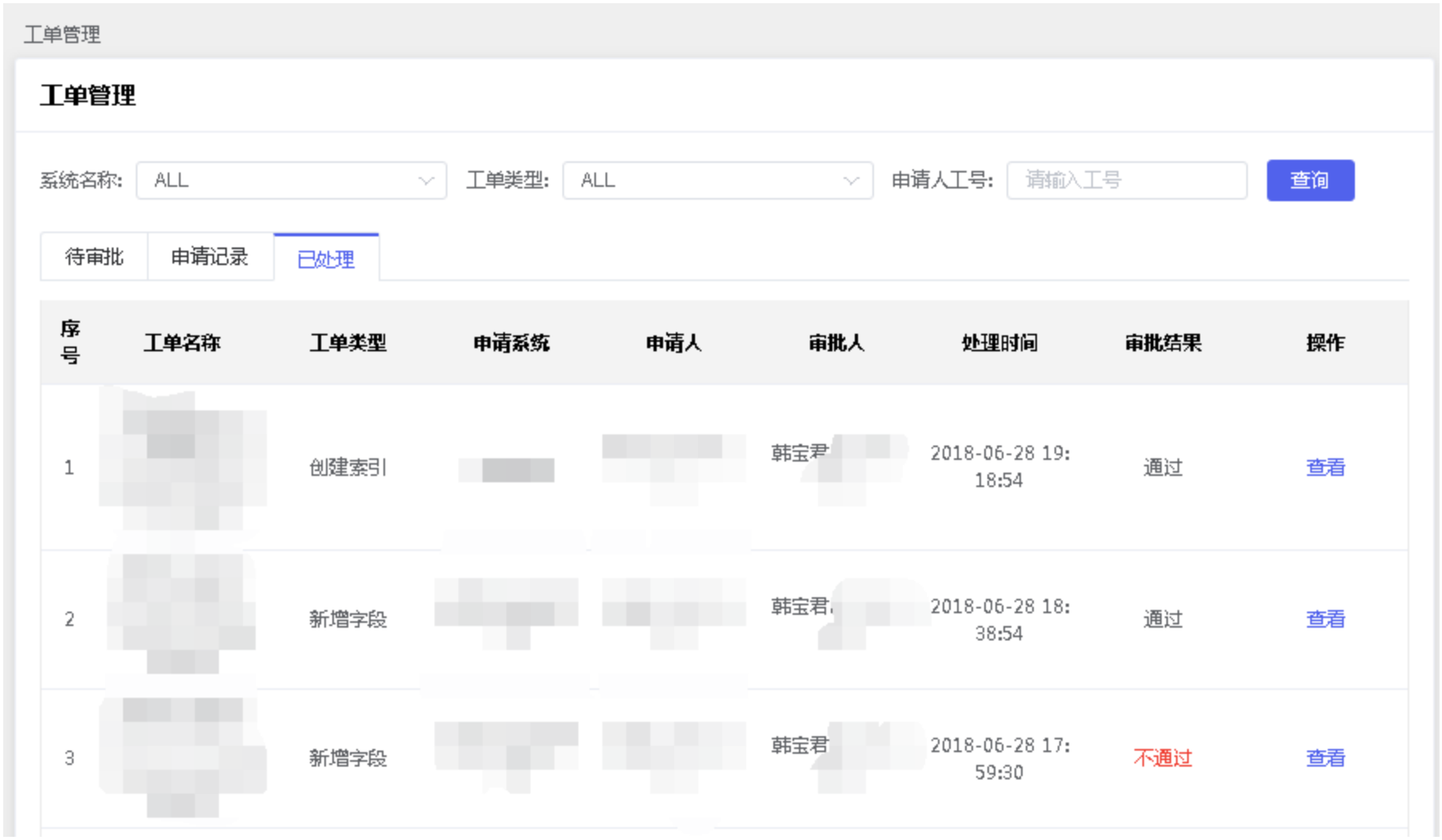
集群管理-用户视图


集群管理
集群管理
+申请集群
申请扩容
申请缩容
机房
ALL
集群类型
ALL
集群名称
ALL
软件包名称
ALL
软件包版本
ALL
查询
totalClusters
totalNodes
totalIndices
totalShards
totalDocs
totalStore
1
20
181
2032
155亿
15T
集群名称
机房
系统名称
集群端口
新增数据条数
新增数据大小
TPS
QPS
状态
白名单
黑名单
0
0
实时：0/离线：0
0
运行
添加白名单
添加黑名单
共 1 条
20条/页
1
前往
1
页

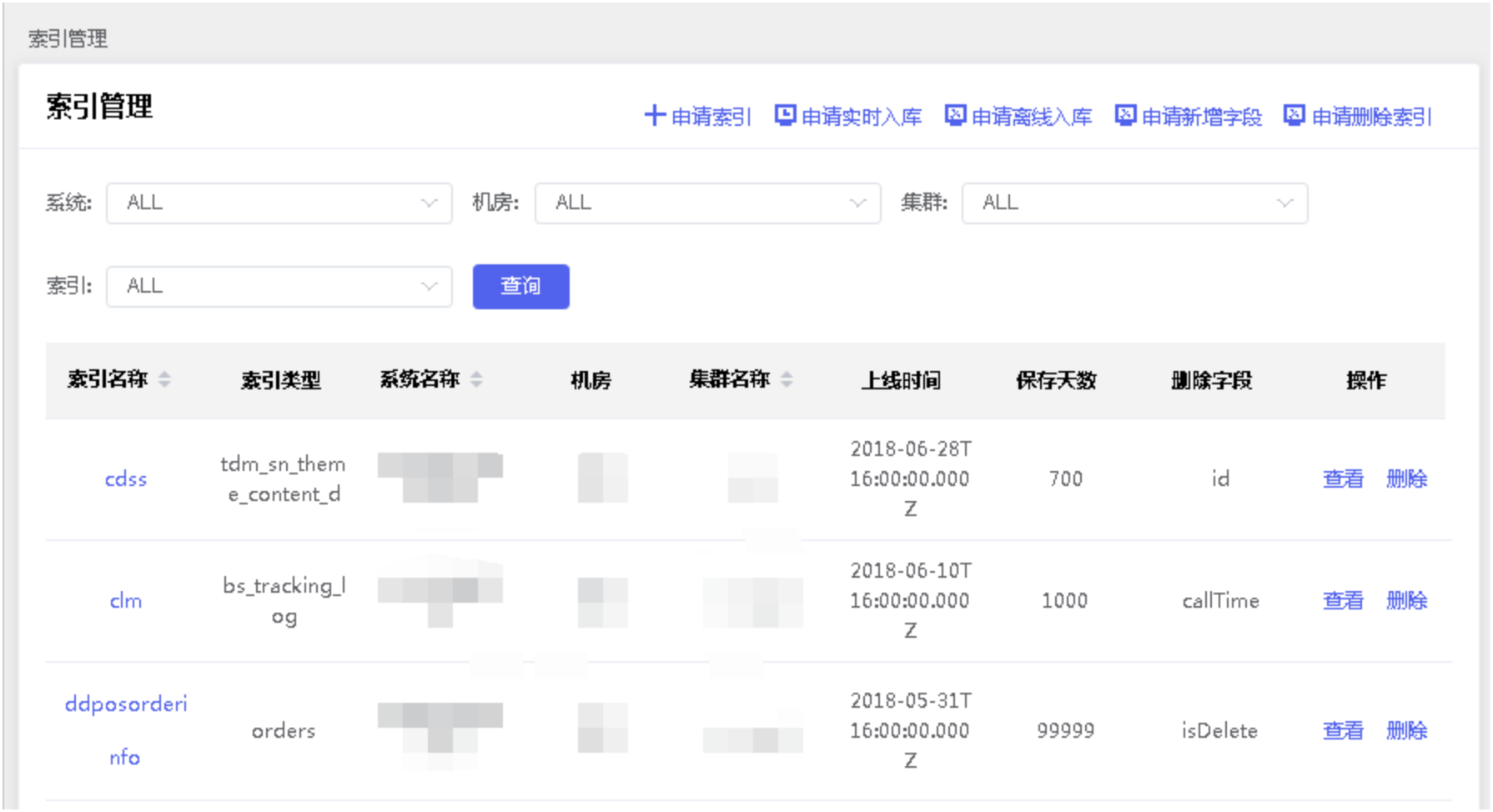
工单管理-待审批


工单管理
工单管理
系统名称:
ALL
工单类型:
ALL
申请人工号:
请输入工号
查询
待审批
申请记录
已处理
序号
工单名称
工单类型
系统
申请人
申请时间
上线时间
操作
1
新增字段
2018-06-29 14:34:51
2018-07-03 00:00:00
审批
撤销
2
创建索引
2018-06-29 14:30:11
2018-07-03 00:00:00
审批
撤销

工单管理-已处理


工单管理
工单管理
系统名称: ALL
工单类型: ALL
申请人工号: 请输入工号
查询
待审批
申请记录
已处理
序号
工单名称
工单类型
申请系统
申请人
审批人
处理时间
审批结果
操作
1
创建索引
韩宝君
2018-06-28 19:18:54
通过
查看
2
新增字段
韩宝君
2018-06-28 18:38:54
通过
查看
3
新增字段
韩宝君
2018-06-28 17:59:30
不通过
查看

索引管理


索引管理
索引管理
申请索引
申请实时入库
申请离线入库
申请新增字段
申请删除索引
系统: ALL
机房: ALL
集群: ALL
索引: ALL
查询
索引名称
索引类型
系统名称
机房
集群名称
上线时间
保存天数
删除字段
操作
cdss
tdm_sn_theme_content_d
2018-06-28T16:00:00.000Z
700
id
查看 删除
clm
bs_tracking_log
2018-06-10T16:00:00.000Z
1000
callTime
查看 删除
ddposorderinfo
orders
2018-05-31T16:00:00.000Z
99999
isDelete
查看 删除

索引管理-申请索引

索引管理 〉 申请新建索引

申请新建索引

*集群名: 请选择集群

*索引名: 请输入索引名

*类型名: 请输入类型名

*上线时间: 请选择日期

*保存天数: 请输入保存天数

*删除字段: 按此字段删除过期数据

索引主键: 请输入索引主键；支持大小写字母，数字，下划线 ，字母开头；多个主键英文分号分割

*申请原因: 请输入申请原因

*索引mapping: 请输入索引mapping

申请配额:

*每天新增数据数量: 请输入每天新增数据数量

*初始化数据数量: 请输入初始化数据数量

*实时TPS: 请输入实时TPS

*每天新增数据大小(M): 请输入每天新增数据大小

*初始化数据大小(M): 请输入初始化数据大小

*离线TPS: 请输入离线TPS

*集群QPS: 请输入集群QPS

确定 取消

计量管理—计量概览

计量概览

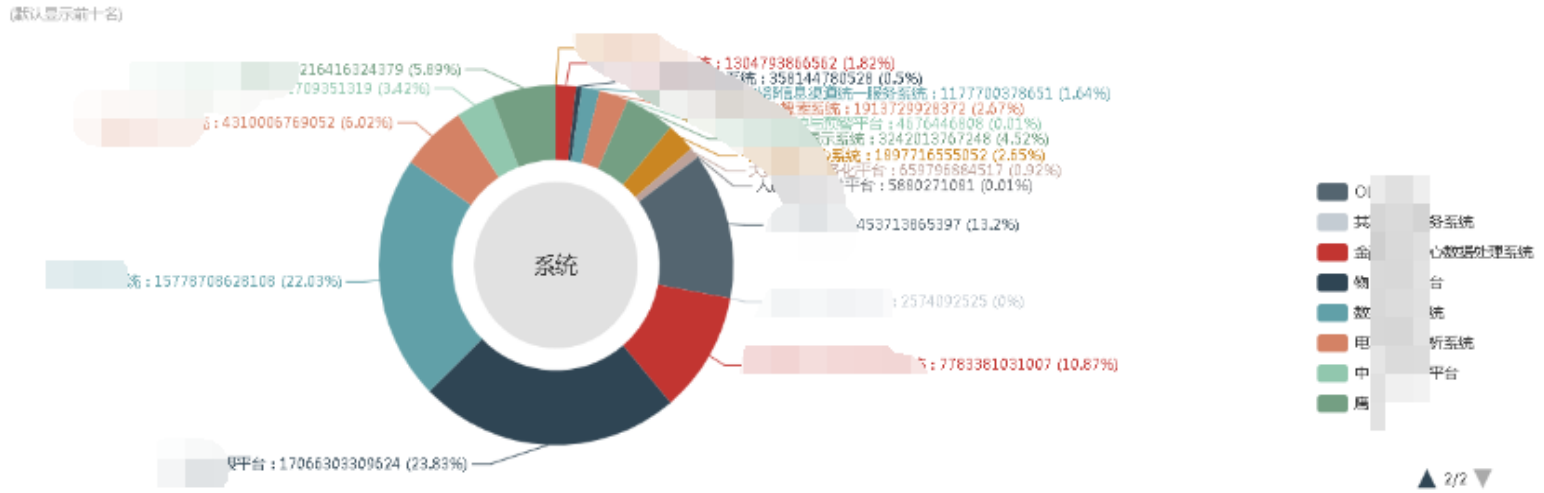
所属系统: ALL　　所属集群: ALL　　查询

最近三个月总请求量	最近三个月读请求量	最近三个月写请求量
6月：126[redacted]70502	6月：[redacted]3965	6月：12[redacted]6537
5月：0	5月：0	5月：0
4月：0	4月：0	4月：0

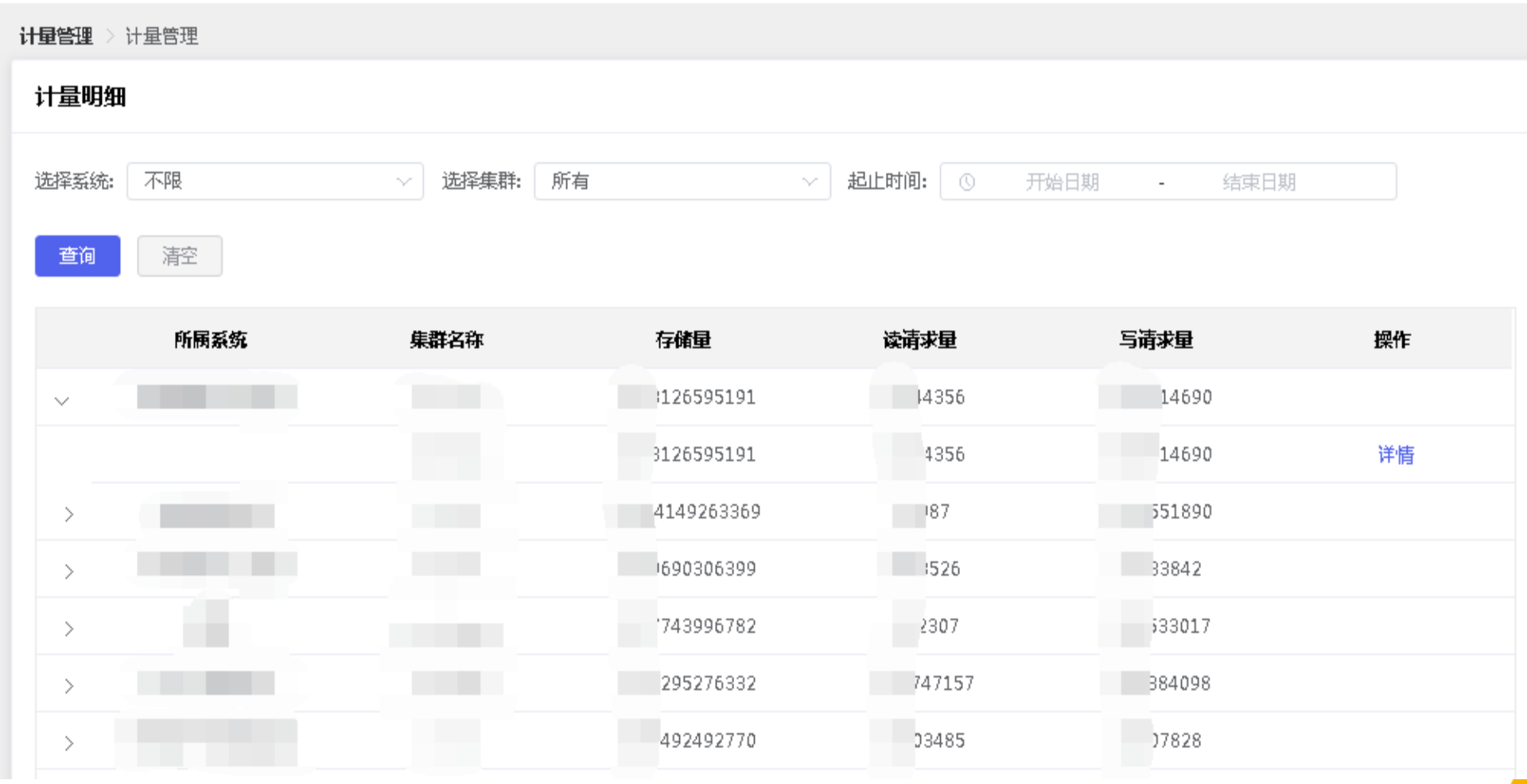
占比分布

占比对象　系统　集群

统计维度　存储量　读请求量　写请求量



计量管理—计量明细


计量管理 > 计量管理
计量明细
选择系统: 不限
选择集群: 所有
起止时间: 开始日期 - 结束日期
查询
清空
所属系统
集群名称
存储量
读请求量
写请求量
操作
详情

计量管理—计量详情

计量详情

返回

选择索引: 请输入索引 查询 清空

索引	存储量	读请求量	写请求量
ssa2_abnormally_log_20180617	4397218	8	680425
ssa2_abnormally_log_20180618	6039480	6	174596
ssa2_abnormally_log_20180619	1146687	3	88346
ssa2_abnormally_log_20180620	71263400	3	22806
ssa2_abnormally_log_20180621	2948190	0	37600
ssa2_abnormally_log_20180622	76822381		0173
ssa2_abnormally_log_20180623	46063794		35144

项目管理-项目列表

项目管理 > 项目概览

项目概览

系统: 所有系统 查询

系统全称	系统帐号	技术总监	系统管理员	所属中心	创建时间	集群名称
					018-05-24T01:47:09.000Z	
					2018-05-24T01:47:09.000Z	
					2018-05-24T01:47:09.000Z	
					2018-05-24T01:47:11.000Z	
					2018-05-24T01:47:09.000Z	

项目管理-秘钥列表

项目管理 > 密钥管理

密钥管理

+ 新增密钥

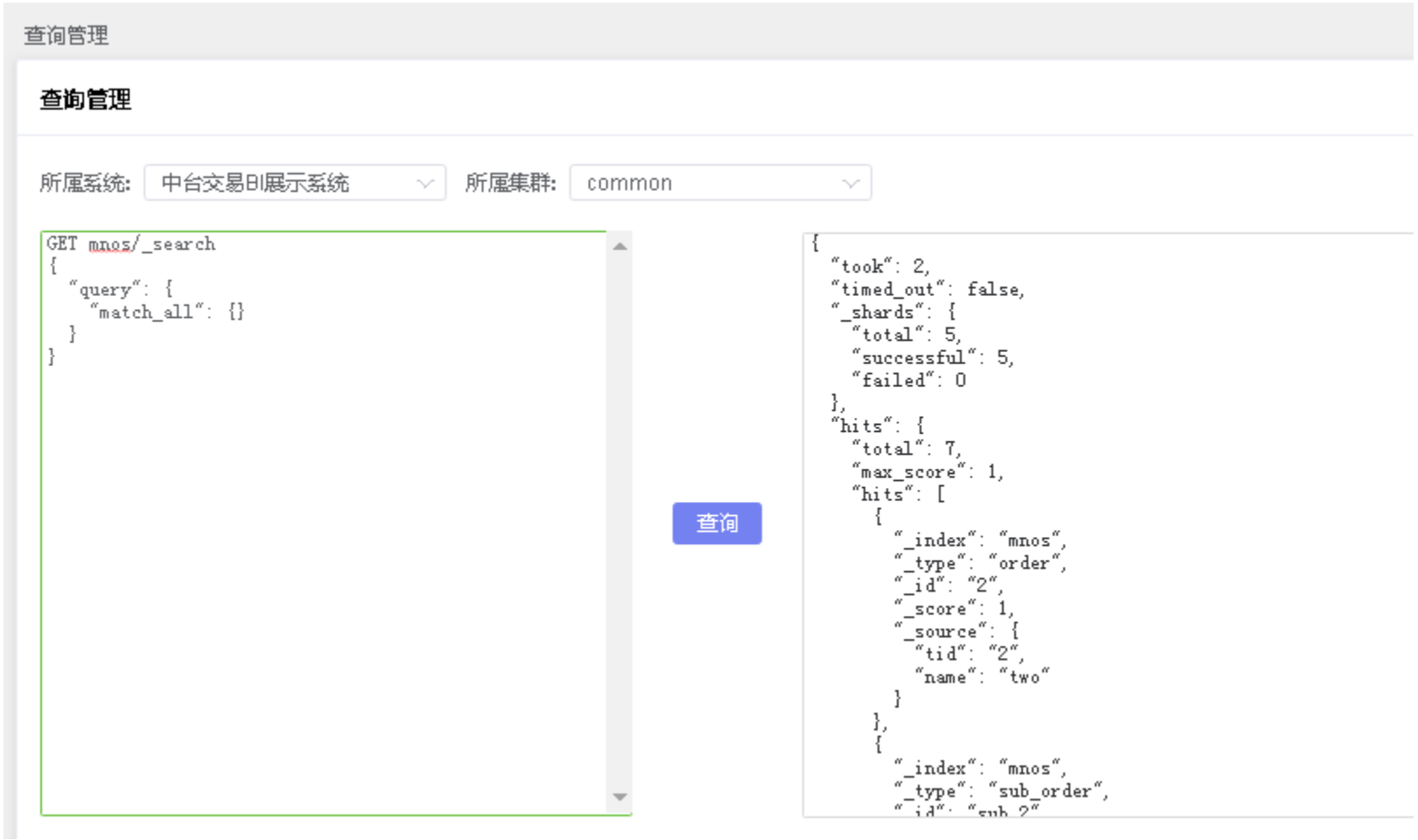
系统: 所有系统 查询

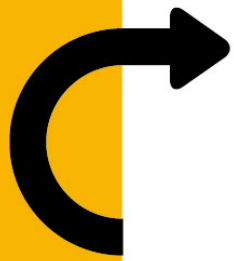
+ 全部展开

系统ID	系统名称	密钥ID	发送邮箱	操作
[redacted]	[redacted]		[redacted]	系统删除
		AK:oRuH7PRLQZSBTZcQUiGYUg SK:xyArxC4SRSCYMSCumhiLDQ		密钥发送 密钥删除
		AK:xAlVQWuJQzKLMpx_20Wi4Q SK:MQ6XpoyiTZ6inYTZRJGlYQ		密钥发送 密钥删除

查询管理

权限：用户只有查看权限、不能修改、删除





1.ES平台总体介绍

2.ES平台化之路

3.实战经验

HTTP请求过长

背景：

Caused by: org.elasticsearch.hadoop.rest.EsHadoopInvalidRequest: An HTTP line is larger than 4096 bytes. （该错误表示http请求超过了es的http请求长度限制）

一个简单的macth_all查询报如上错

方案：

修改elasticsearch-hadoop源码

1. https://github.com/hanbj/elasticsearch-hadoop.git (hanbj_v5.4.2)
2. **Commits**：**An HTTP line is larger than 4096 bytes**

详情见**PR**： https://github.com/elastic/elasticsearch-hadoop/pull/1154

ThreadContextStruct:putHeaders

背景：

线程池模块在处理header时有一点小问题

方案：

修改elasticsearch源码

详情见**PR：** https://github.com/elastic/elasticsearch/pull/26068

elasticsearch-hadoop build sql to dsl

背景：

创建一个Hive外部表指向ES中的某个索引，通过Hive HQL直接操作ES。当运行下面的HQL时，发现找不到数据。

select * from table_name where city_code in (‘010’, ‘791’);

但是下面两条Hql都可以找到数据：

select * from table_name where city_code in (‘010’);
select * from table_name where city_code in ('791');

方案：

修改elasticsearch-hadoop源码

1. https://github.com/hanbj/elasticsearch-hadoop.git (hanbj_v5.4.2)
2. **Commits：** **change match to terms**

详情见**PR：** https://github.com/elastic/elasticsearch-hadoop/pull/1168

SparkSession conf

背景：

Spark2.x之后，SparkConf是全局的配置，一个系统中可以有多个SparkSession实例，而且每个SparkSession实例可以有自己单独的配置，OLAP系统中有很多个业务，每个业务有多个模型，每个模型对应多个索引（每天一个索引），每个模型对应一个SparkSession 实例缓存在内存里，当要查询一个模型时，从内存里找出对应的SparkSession实例去查找对应的索引。

方案：

1. 修改elasticsearch-hadoop源码
    1. https://github.com/hanbj/elasticsearch-hadoop.git (hanbj_v5.4.2)
    2. **Commits：check cluster name and add SparkSession conf**

详情见PR： https://github.com/elastic/elasticsearch-hadoop/pull/1135

```scala
def main(args: Array[String]): Unit = {
    val sparkConf = new SparkConf().setAppName("HiveToES").setMaster("local[*]")
    sparkConf.set("spark.sql.hive.metastorePartitionPruning", "false")
    sparkConf.set("es.index.auto.create", "true")
    sparkConf.set("es.mapping.date.rich", "false")

    val session = SparkSession.builder().config(sparkConf).getOrCreate()
    session.conf.set("es.cluster.name", "common1")
    session.conf.set("es.index.filter", "hanbj-20180502,hanbj-201803*")
    session.sql( sqlText = "create table if not exists es_test using es OPTIONS (es.nodes 'localhost', es.port '9200',path 'hanbj')")
    val dataFrame = session.sql( sqlText = "select * from es_test")
    dataFrame.show()

    val session1 = SparkSession.builder().config(sparkConf).getOrCreate()
    session1.conf.set("es.cluster.name", "common2")
    session1.conf.set("es.index.filter", "t_wh_full_link_d-2018")
    session1.conf.set("es.nodes", "localhost")
    session1.conf.set("es.port", "9100")
    val df = EsSparkSQL.esDF(session1, resource = "t_wh_full_link_d")
    df.show()
}
```

查找大别名下的个别索引

大别名（包含很多索引）

不同的SparkSession可以访问不同的集群或者索引

集群名

Master选举修改

背景：

当集群规模越来越大，在高并发、高基数查询和高写入量并存的场景下，节点负载高有时可能导致该节点脱离集群或者触发重新选举master，这两种情况下可能都会导致分片漂移，造成资源的浪费。

方案：

1. 部分节点单机多实例部署
2. 适量增大ping的超时时间
3. master、data、ingest节点分离（至少设置3个实例为master节点实现多可靠）
4. 修改master选举算法（master节点成为master的优先级最高，其次是ingest节点做担保）
    1. https://github.com/hanbj/elasticsearch.git (hanbj_v5.4.2)
    2. **Commits：elect master，include ingest node**

集群名称校验

背景：

目前有近20个ES集群，部分集群http端口一样。REST方式访问ES集群时，只要IP和端口配置正确，就可以进行访问

方案：

1. 修改elasticsearch-hadoop源码
    1. https://github.com/hanbj/elasticsearch-hadoop.git (hanbj_v5.4.2)
    2. **Commits：check cluster name and add SparkSession conf，params append cluster.name**
2. 修改elasticsearch源码
    1. https://github.com/hanbj/elasticsearch.git (hanbj_v5.4.2)
    2. **Commits：RestClient add check cluster name，RestClient add check cluster name （netty3）**

访问集群方式改变成：见下页

```
RestClient client = RestClient.builder(
        new HttpHost( hostname: "localhost",  port: 9200,  scheme: "http"))
        .setClusterName("elasticsearch")
        .build();
```

```
val sparkConf = new SparkConf().setAppName("HiveToES").setMaster("local[*]")
sparkConf.set("es.index.auto.create", "true")
sparkConf.set("es.cluster.name", "elasticsearch")
sparkConf.set("es.nodes", "localhost")
sparkConf.set("es.port", "9200")
val sc = new SparkContext(sparkConf)
```

黑白名单控制

背景：

集群安全稳定至关重要，为了控制集群以外的机器对集群进行无效查询和攻击，所以应该支持动态允许/防止一些机器访问集群。

方案：

修改elasticsearch源码（支持IPv4、IPv6和通配符）

1. https://github.com/hanbj/elasticsearch.git (hanbj_v5.4.2)
2. **Commits：black and white list**
3. **参数：**都可动态修改
    1. http.filter.enabled
    2. transport.filter.enabled
    3. http.filter.allow
    4. http.filter.deny
    5. transport.filter.allow
    6. transport.filter.deny

导出的痛

背景：

在共享集群的模式下，频繁或大并发的导出操作，会给集群造成比较大的压力，导致该集群下其他业务的正常查询延迟。

方案：

修改elasticsearch源码，添加相应的控制参数，可以控制导出的并发量和数据量。

1. https://github.com/hanbj/elasticsearch.git (hanbj_v5.4.2)
2. **Commits：limit scroll**
3. **参数：**都可动态修改
    1. scroll.enabled
    2. scroll.interval
    3. scroll.concurrent.indices
    4. scroll.limit

elasticsearch-hadoop支持删除

背景：

目前的elasticsearch-hadoop版本不支持删除索引和数据，但业务有这方面的需求。

方案：

修改elasticsearch-hadoop源码，增加删除索引和根据query删除数据的逻辑。

1. https://github.com/hanbj/elasticsearch-hadoop.git (hanbj_v5.4.2)
2. **Commits：** **delete index and delete by query**

API格式统一化，标准化

背景：

delete_by_query和update_by_query API 和其他API 格式不统一，容易对业务造成困扰。

方案：

修改elasticsearch源码

1. https://github.com/hanbj/elasticsearch.git (hanbj_v5.4.2)
2. **Commits：delete by query，update by query**

```
public void testDeleteByQuerySync() {
    client.prepareDeleteByQuery().source("hanbj")
            .filter(QueryBuilders.matchQuery("name", "hanbj")).get();
}
public void testDeleteSync() {
    DeleteByQueryAction.INSTANCE.newRequestBuilder(client)
            .filter(QueryBuilders.matchQuery("name", "hanbj")).source("hanbj").get();
}
```

修改后

修改前

请求跟踪

背景：

Transport模块有一个专用的跟踪记录器，当被激活时，记录传入和进出请求。可以使用一组通配符模式来控制哪些操作将被跟踪。默认情况下，每个请求将被跟踪，除了故障检测和ping。但是日志中并没有打印请求源，不方便进行追踪。我在ES的基础上增加了请求源IP、内部请求转发、请求发送、响应接收的详细日志。

方案：

修改elasticsearch源码，增加了请求源IP、内部请求转发、请求发送、响应接收的详细日志。

1. https://github.com/hanbj/elasticsearch.git (hanbj_v5.4.2)
2. **Commits**：请求追踪

元数据查看

背景：

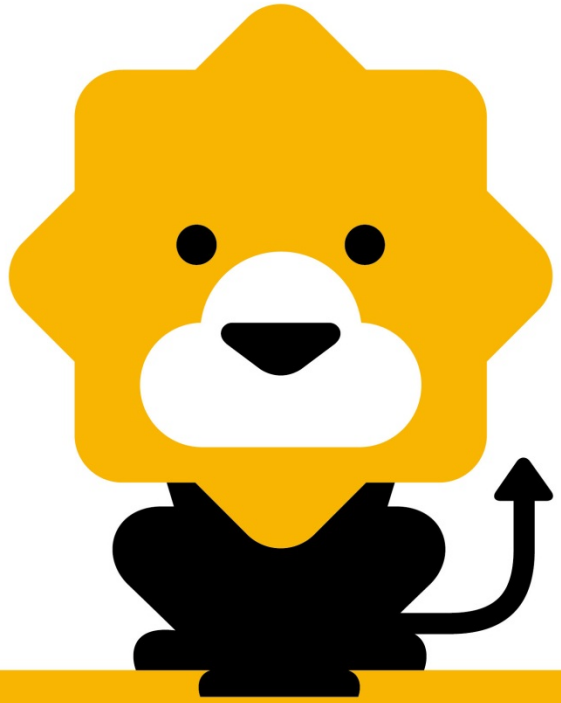
Gateway模块用于存储es集群的MetaData。MetaData每一次改变（比如增加、删除索引等），都要通过Gateway模块进行持久化。当集群第一次启动的时候，这些信息就会从Gateway模块中读出并应用。状态文件存的都是二进制，不具备可读性。

方案：

修改elasticsearch源码，实现查看所有、全局、单个索引的状态文件内容。

1. https://github.com/hanbj/elasticsearch.git (hanbj_v5.4.2)
2. **Commits：cat metadata**

Thanks!



elastic
中文社区

专业、垂直、纯粹的 Elastic 开源技术交流社区
https://elasticsearch.cn/