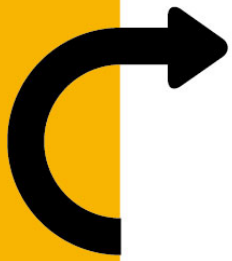
苏宁ES平台化实践之路

——苏宁大数据中心 大数据平台 搜索平台组



1.ES平台总体介绍

2.ES平台化之路

3.实战经验

关于我们

苏宁云商IT总部
大数据平台研发中心

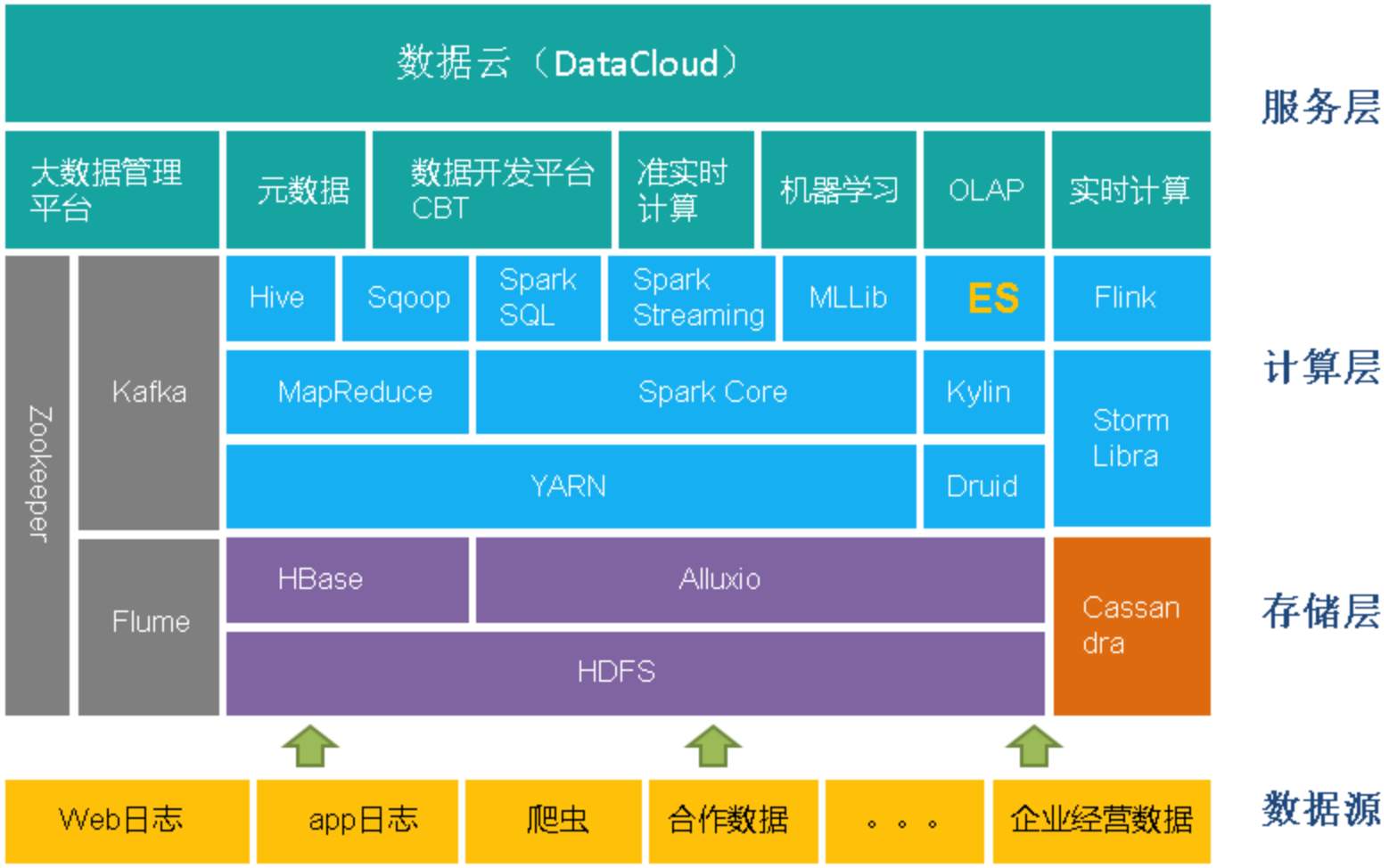
大数据平台职责：

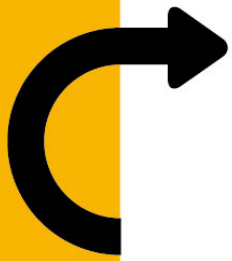
提供苏宁集团大数据存储和计算能力。
保证平台的稳定、高效运行。
提高平台易用性。

个人职责：

Elasticsearch组件负责人

苏宁大数据平台总体架构



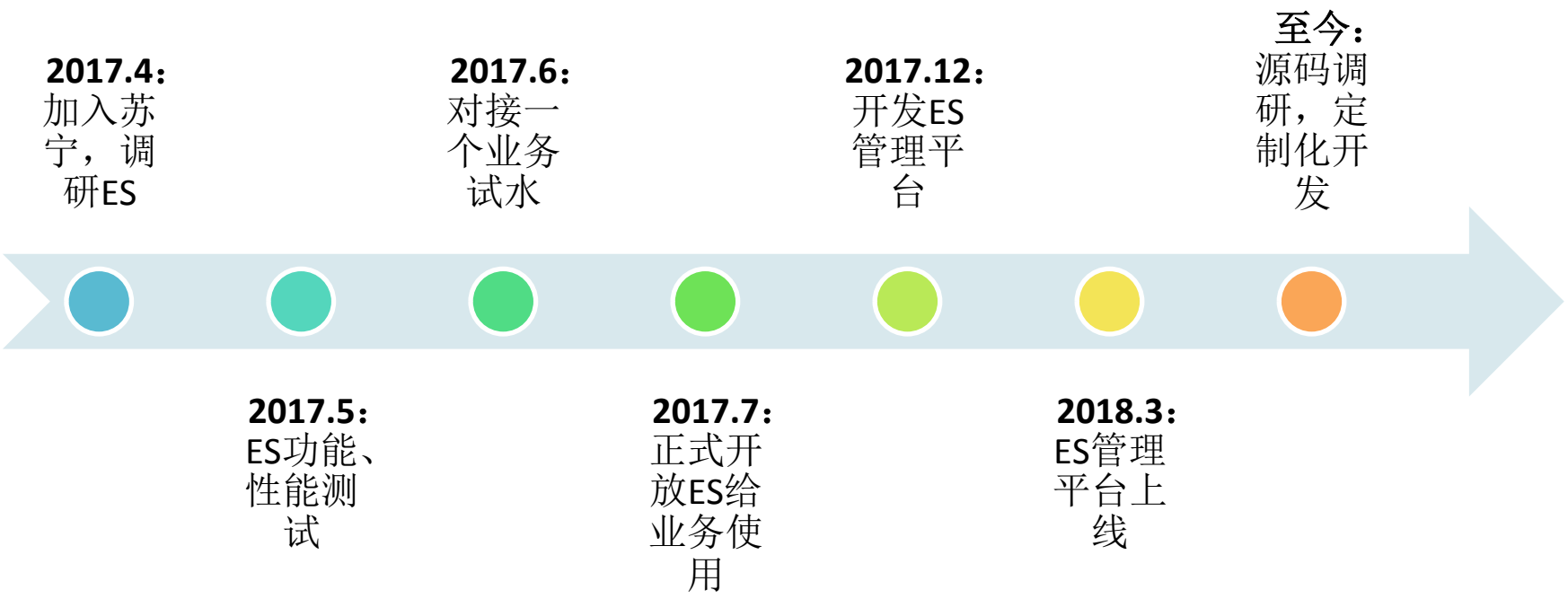


1. ES平台总体介绍
2. ES平台化之路
3. 实战经验

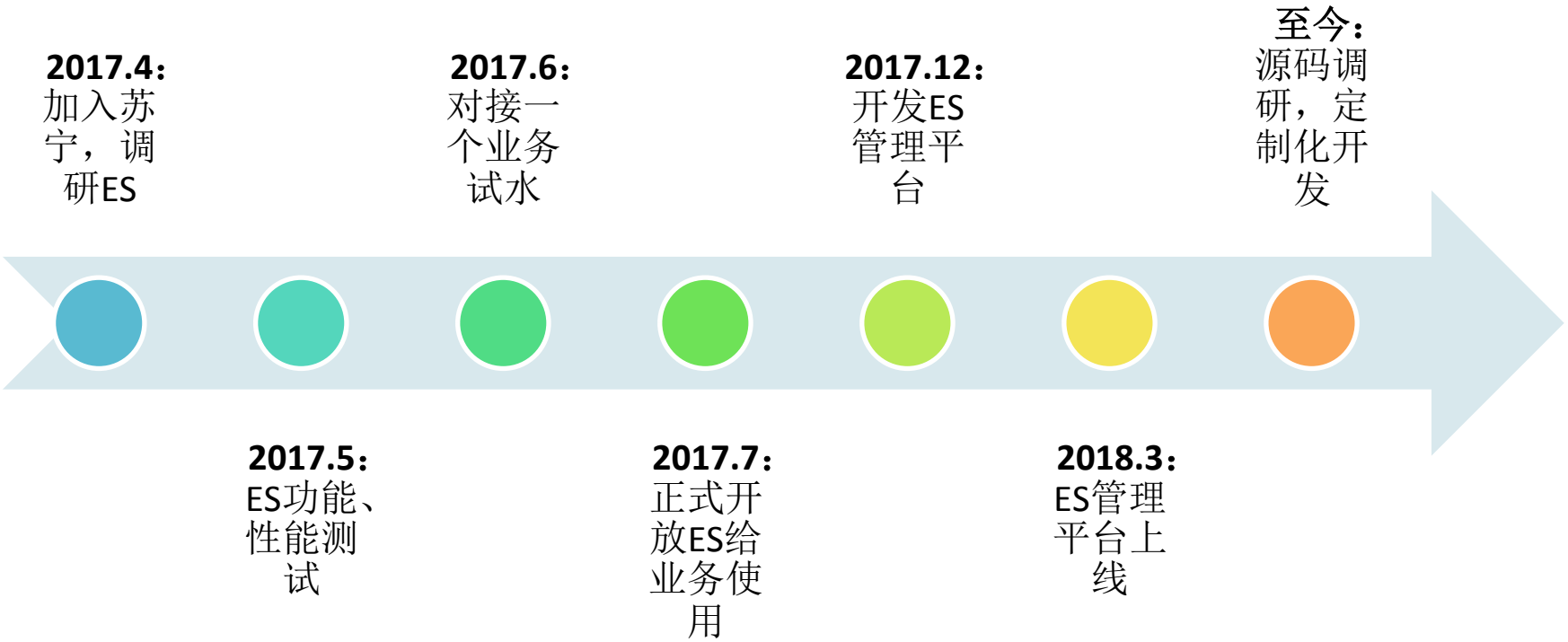
为什么做ES平台？

在还没有做ES平台之前，公司有些业务部门已经使用了ES，他们自己各自搭建自己的集群，而且版本也不统一。

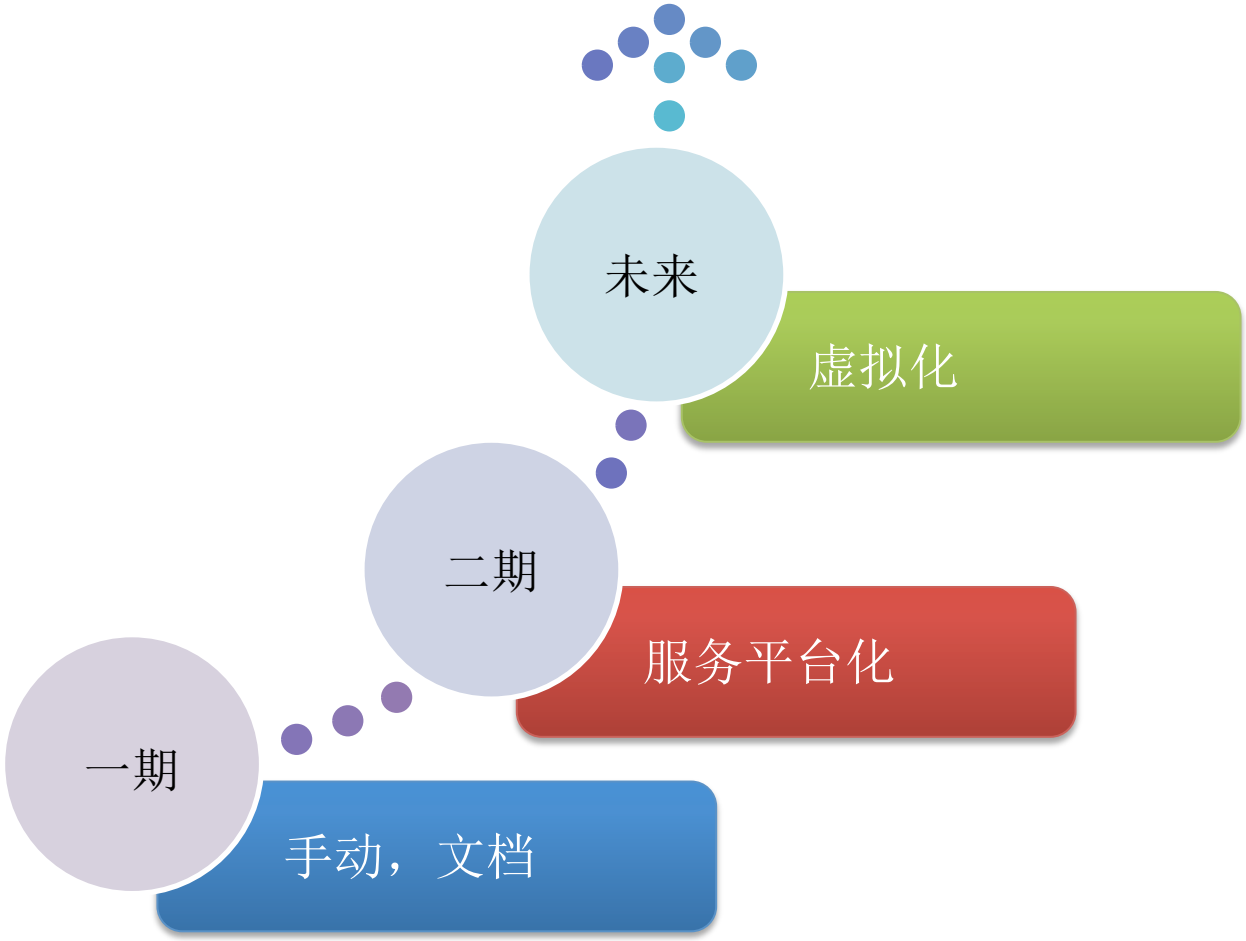
每逢大促，有些业务的ES集群经常出现这样那样的问题，而且业务部门由于大部分时间都在写业务代码，对ES技术的积累也不够，这时有些业务就希望我们平台统一管理ES（像Spark、Hive、Hbase，Hadoop，Storm等一样）。



ES平台时间轴


2017.4：
加入苏宁，调研ES
2017.5：
ES功能、性能测试
2017.6：
对接一个业务试水
2017.7：
正式开放ES给业务使用
2017.12：
开发ES管理平台
2018.3：
ES管理平台上线
至今：
源码调研，定制化开发

我们做了什么？



一期：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 □高级数查询 □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

一期：业务字段

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

二期：ES管理平台

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

二期：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

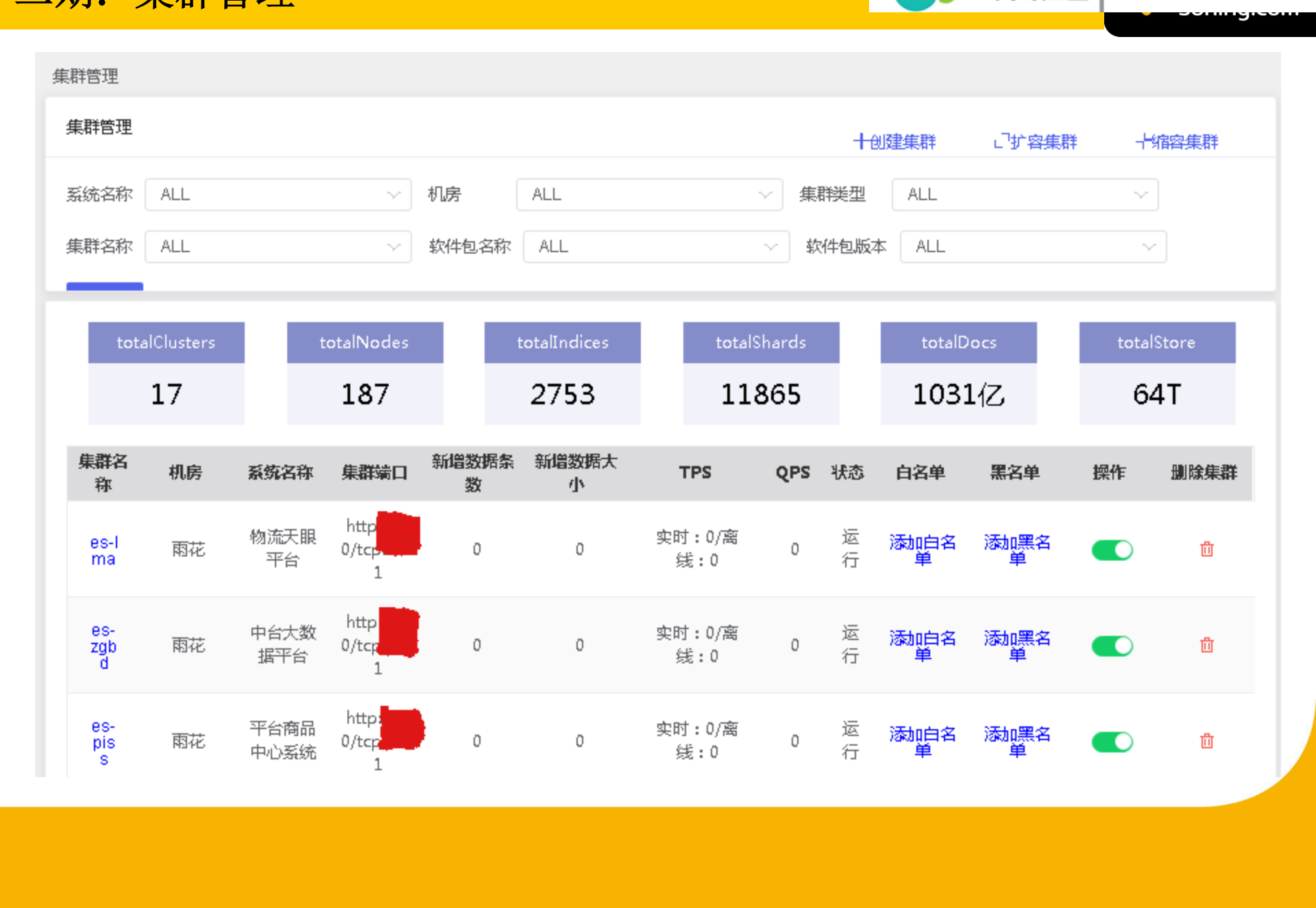
查询

Clusters	Nodes	Indices	Memory
17	187	2753	46GB/128GB

Total Shards	Data	Unassigned Shards	Documents
11865	64T	0	1031亿

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	2018-06-17T16:47:02.000Z	数据云公司	大数据平台研发中心	王志强	雨花	es-olap	red	独占	拒绝任务. null	拒绝任务: 10.247.24.151 bul...
2018-05-24T06:39:26.000Z	2018-06-17T16:47:03.000Z	消费者平台研发中心	搜索研发中心	贾洪园	雨花	es-ecias	red	独占	拒绝任务. 拒绝任务. 拒绝任务...	拒绝任务: 10.247.24.48 searc...

二期：集群管理


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
totalClusters 17
totalNodes 187
totalIndices 2753
totalShards 11865
totalDocs 1031亿
totalStore 64T
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
es-lma 雨花 物流天眼平台 http 0/tcp 1 0 0 实时：0/离线：0 0 运行 添加白名单 添加黑名单
es-zgbd 雨花 中台大数据平台 http 0/tcp 1 0 0 实时：0/离线：0 0 运行 添加白名单 添加黑名单
es-piss 雨花 平台商品中心系统 http 0/tcp 1 0 0 实时：0/离线：0 0 运行 添加白名单 添加黑名单

二期：项目管理


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
15157
大数据ES服务化平台
16050633@cnsuning.com
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
共 1 条
20条/页
1
前往 1 页

二期：资源管理

资源管理 > 机器列表

机器列表

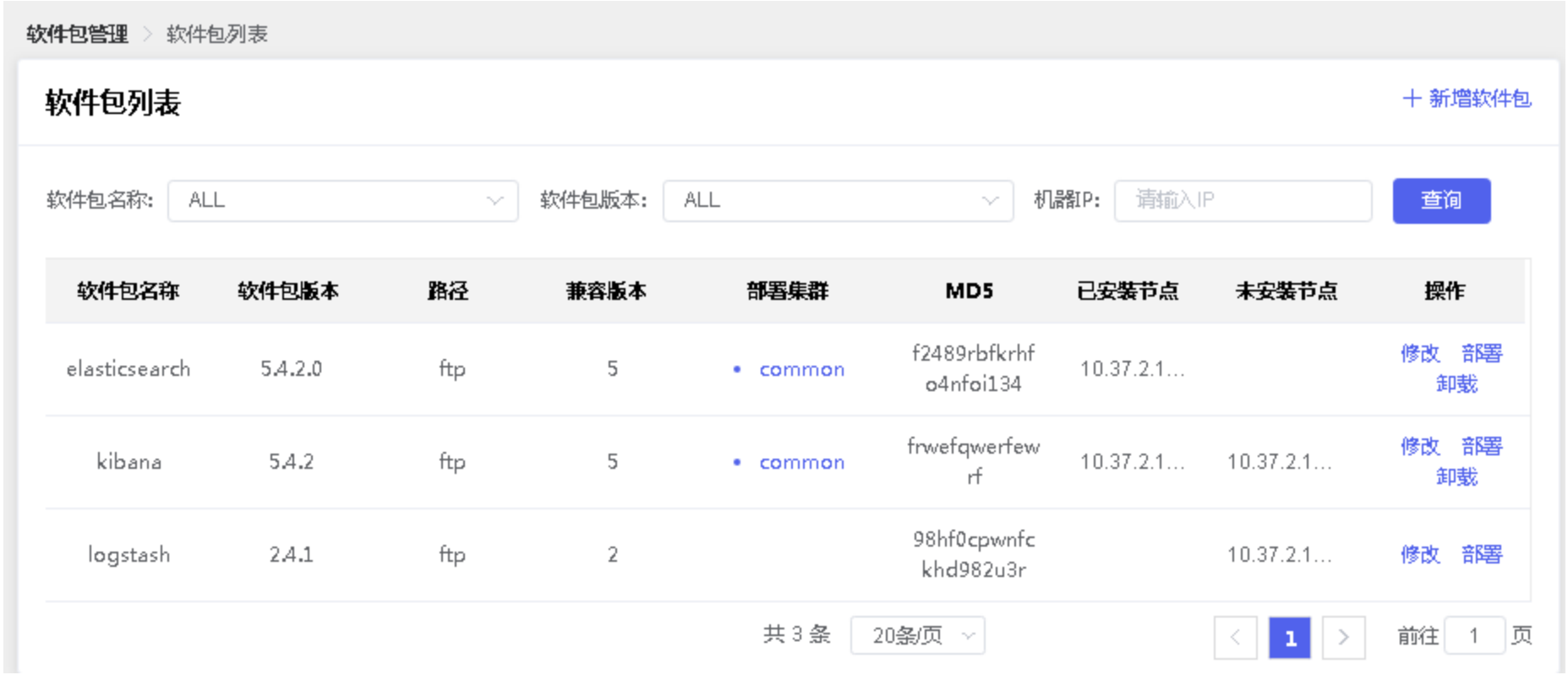
+ 新增资源

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
es01-prd1.cnsuning.com	[redacted]	32	131072 M	172032/data00 281600/data01 281600/data02	9100-9300	es-zgbd	elasticsearch:5.4.2.0	修改配置 删除
es02-prd1.cnsuning.com	[redacted]	32	131072 M	172032/data00 281600/data01 281600/data02	9100-9300	es-zgbd	elasticsearch:5.4.2.0	修改配置 删除
es03-prd1.cnsuning.com	[redacted]	32	131072 M	172032/data00 281600/data01 281600/data02	9100-9300	es-zgbd	elasticsearch:5.4.2.0	修改配置 删除

二期：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 common f2489rbfkrhfo4nfoi134 10.37.2.1... 修改 部署 卸载
kibana 5.4.2 ftp 5 common frwefqwerfewrf 10.37.2.1... 10.37.2.1... 修改 部署 卸载
logstash 2.4.1 ftp 2 98hf0cpwnfckhd982u3r 10.37.2.1... 修改 部署
共 3 条
20条/页
1
前往 1 页

二期：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
clm
bs_tracking_log
中台大数据平台
雨花
es-zqbd
2018-06-10T16:00:00.000Z
1000
callTime
查看 删除
ddposorderinfo
orders
体育内容标签系统
雨花
es-sport
2018-05-31T16:00:00.000Z
99999
isDelete
查看 删除
ecias
products
电商情报分析系统
雨花
es-ecias
2018-06-04T16:00:00.000Z
1
offSellDate
查看 删除
eciaspricedetail
priceDetail
电商情报分析系统
雨花
es-ecias
2018-06-14T16:00:00.000Z
84
statDate
查看 删除
eppaccountctrl
eppaccountctrl
金融账务核心数据处理系统
雨花
es-faepp
2018-06-10T16:00:00.000Z
360
TERMDATE
查看 删除
eppaccountctrldetail
eppaccountctrldetail
金融账务核心数据处理系统
雨花
es-faepp
2018-06-10T16:00:00.000Z
360
TERMDATE
查看 删除

二期：工单管理

序号	工单名称	工单类型	申请系统	申请人	审批人	申请时间	上线时间	审批状态	操作
1	中台大数据平台-es-zgbd-创建索引-z_inv_qty_serial-z_inv_qty_serial	创建索引	中台大数据平台	杭军(16050323)	韩宝君(17033293)	2018-06-15 09:48:24	2018-06-19 00:00:00	不通过	查看
2	中台大数据平台-es-zgbd-创建索引-z_gaia_all_log-z_gaia_all_log	创建索引	中台大数据平台	杭军(16050323)	韩宝君(17033293)	2018-06-15 09:46:08	2018-06-19 00:00:00	通过	查看
3	中台大数据平台-es-zgbd-创建索引-z_inv_status_log-inv_status_log	创建索引	中台大数据平台	杭军(16050323)	韩宝君(17033293)	2018-06-14 17:23:42	2018-06-18 00:00:00	通过	查看
4	中台大数据平台-es-zgbd-创建索引-z_inv_status_log-inv_status_log	创建索引	中台大数据平台	杭军(16050323)	韩宝君(17033293)	2018-06-14 08:52:31	2018-06-18 00:00:00	不通过	查看

二期：计量管理—计量概览

计量管理

计量概览

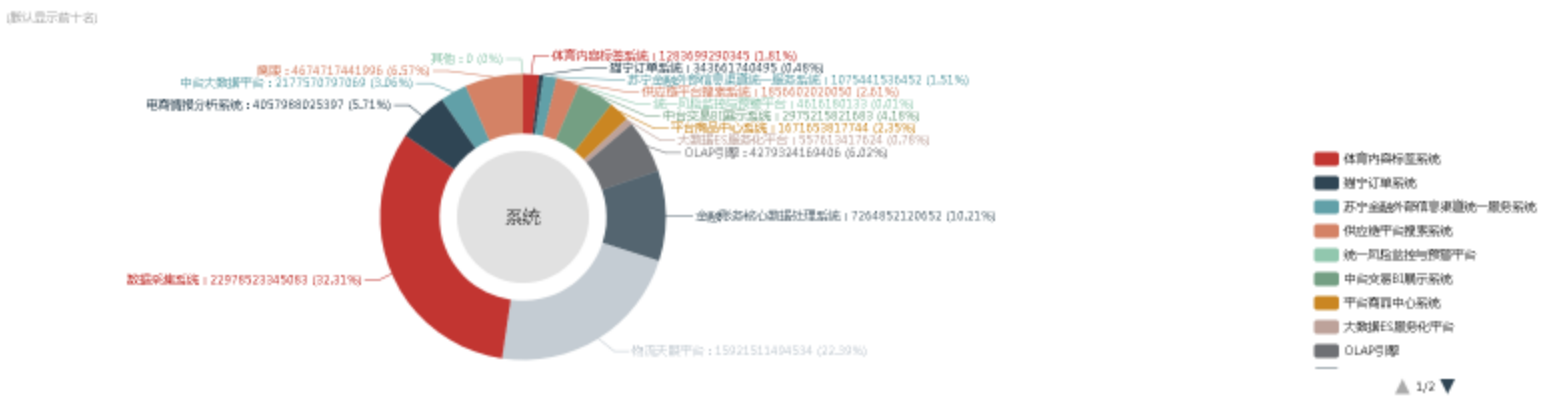
所属系统: ALL 所属集群: ALL 查询

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：2703107435	6月： 40756311	6月： 2662351124
5月：0	5月： 0	5月： 0
4月：0	4月： 0	4月： 0

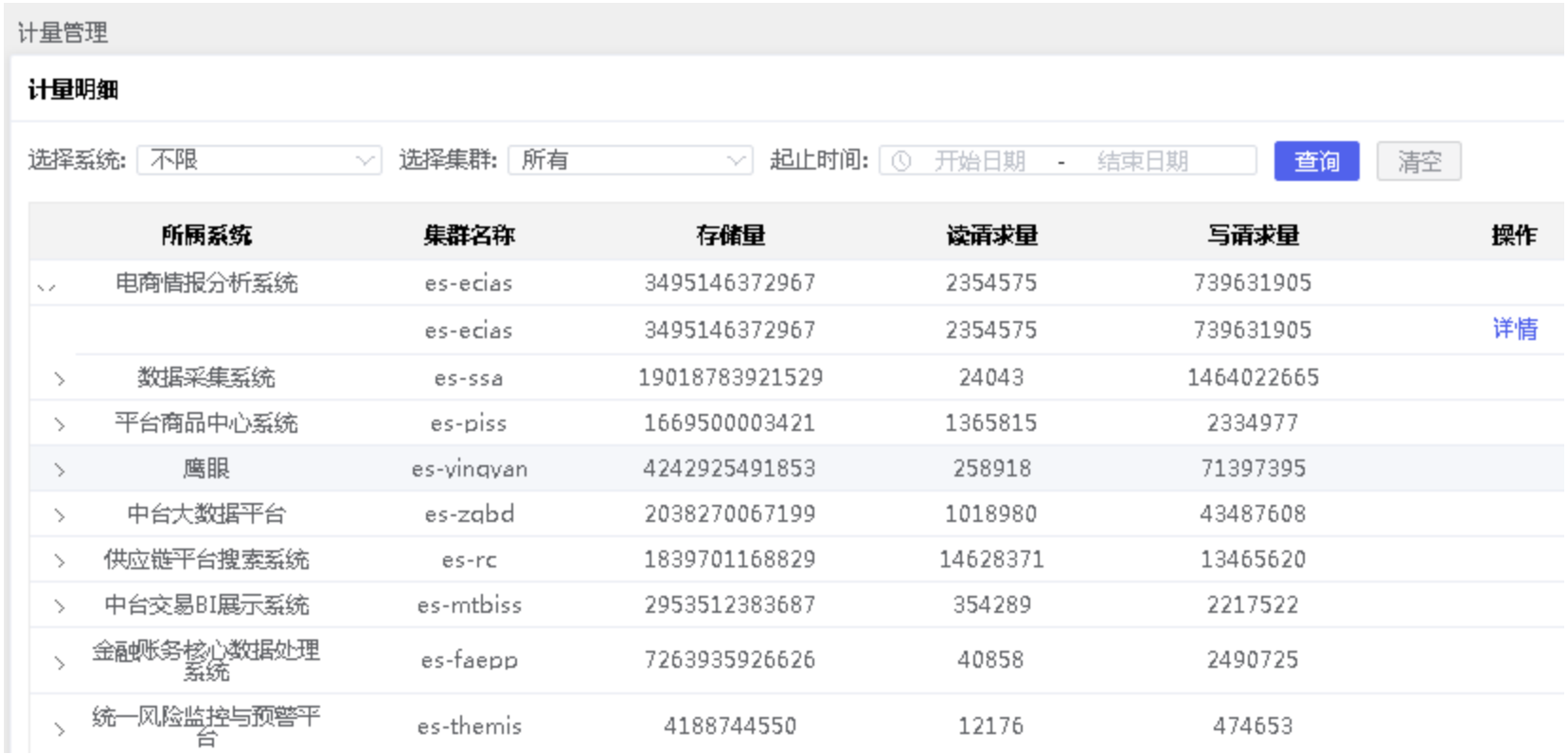
占比分布

占比对象 系统 集群

统计维度 存储量 读请求量 写请求量



二期：计量管理—计量明细



计量管理

计量明细

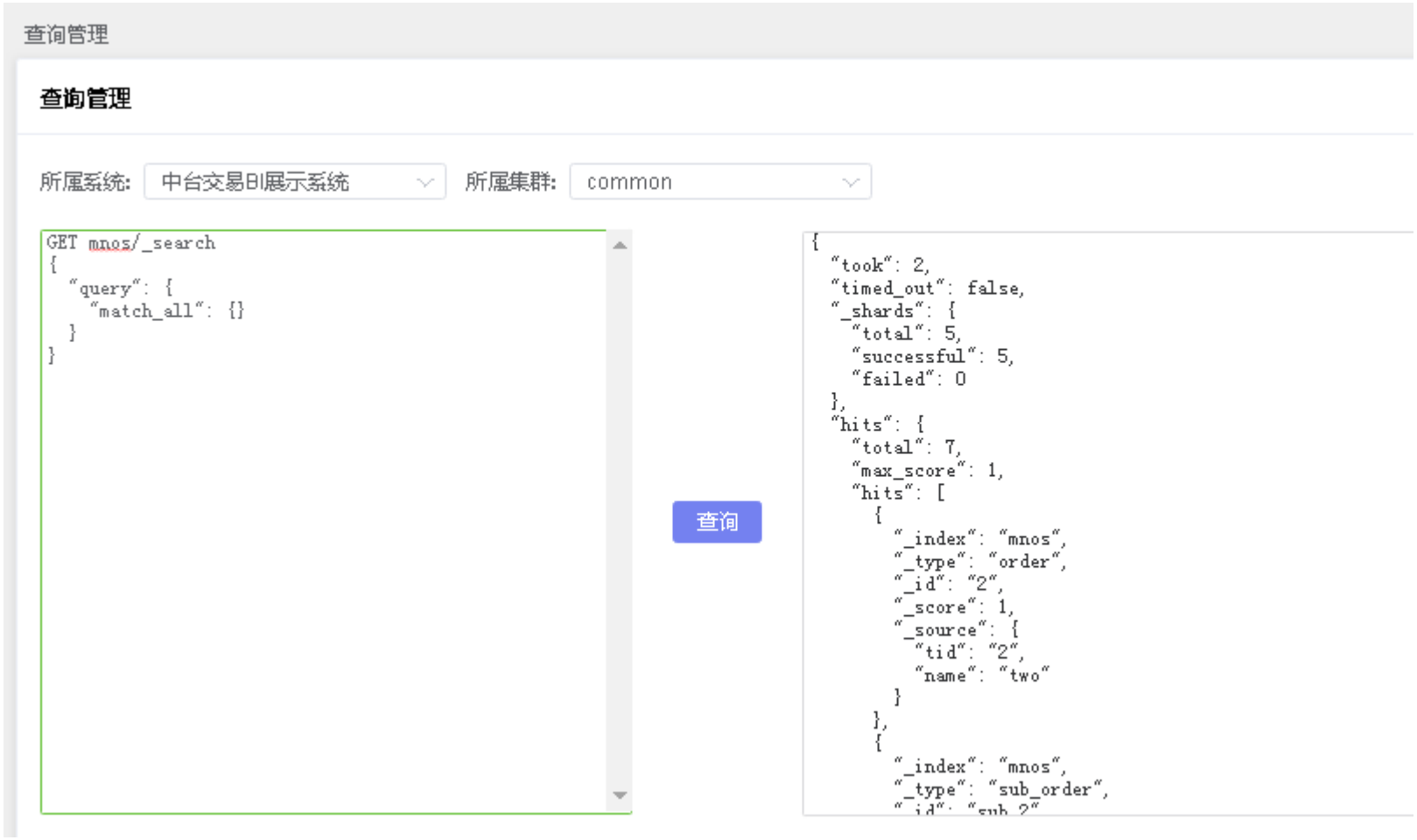
选择系统: 不限　选择集群: 所有　起止时间: 开始日期 - 结束日期　查询　清空

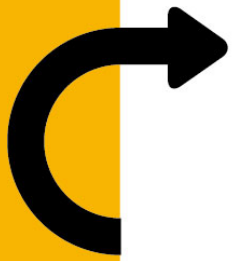
所属系统	集群名称	存储量	读请求量	写请求量	操作
电商情报分析系统	es-ecias	3495146372967	2354575	739631905	
	es-ecias	3495146372967	2354575	739631905	详情
数据采集系统	es-ssa	19018783921529	24043	1464022665	
平台商品中心系统	es-piss	1669500003421	1365815	2334977	
鹰眼	es-yingyan	4242925491853	258918	71397395	
中台大数据平台	es-zqbd	2038270067199	1018980	43487608	
供应链平台搜索系统	es-rc	1839701168829	14628371	13465620	
中台交易BI展示系统	es-mtbiss	2953512383687	354289	2217522	
金融账务核心数据处理系统	es-faepp	7263935926626	40858	2490725	
统一风险监控与预警平台	es-themis	4188744550	12176	474653	

二期：计量管理—计量详情


计量详情
计量详情
返回
选择索引: ecias
查询
清空
索引 存储量 读请求量 写请求量
ecias 2222924960066 2102120 738274904
eciascatalogstatdetail 5465524467 0 0
eciasmissproductdetail 947809189618 40 0
eciaspricedetail 206830197856 1240 0
eciaspricestatdetail 18168529399 0 0
eciasstoredetail 84349749934 0 0
共 6 条
1
前往 1 页

二期：查询管理


查询管理
查询管理
所属系统:
中台交易BI展示系统
所属集群:
common
GET mnos/_search
{
"query": {
"match_all": {}
}
}
查询
{
"took": 2,
"timed_out": false,
"_shards": {
"total": 5,
"successful": 5,
"failed": 0
},
"hits": {
"total": 7,
"max_score": 1,
"hits": [
{
"_index": "mnos",
"_type": "order",
"_id": "2",
"_score": 1,
"_source": {
"tid": "2",
"name": "two"
}
},
{
"_index": "mnos",
"_type": "sub_order",



1. ES平台总体介绍
2. ES平台化之路
3. 实战经验

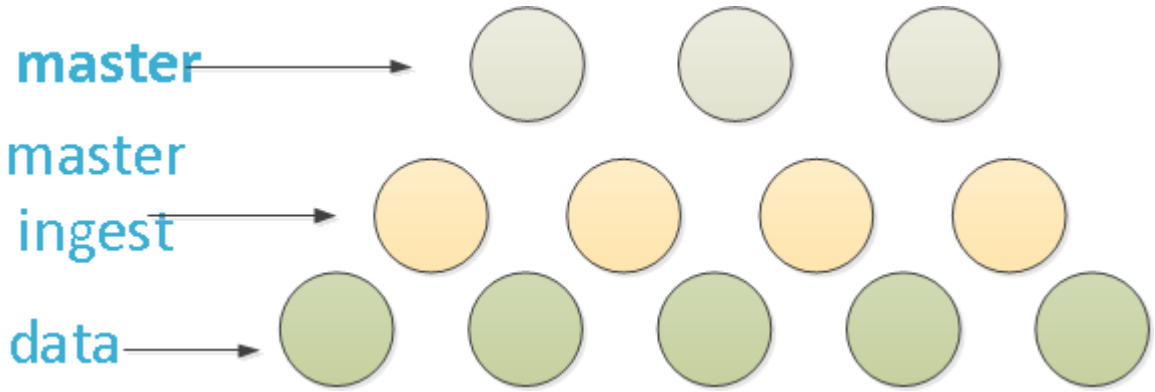
Master选举修改

背景：

默认情况下，elasticsearch 集群中每个节点都有成为主节点的资格，也都存储数据，还可以提供读写服务，由于我们其中一个olap集群索引数接近1万，分片数将近到达20万，在高并发、高基数查询和高写入量并存的场景下，节点负载高有时导致该节点脱离集群或者触发重新选举master。

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master、ingest节点）
    1. https://github.com/hanbj/elasticsearch.git (hanbj_v5.4.2)
    2. **Commits：elect master**



集群名称校验

背景：

ES平台目前有17个ES集群（每个大中心一个集群），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，就可以访问到别的中心的集群。曾经出现了其他业务访问了别的中心集群的现象。

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：

1. 修改elasticsearch源码（支持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

HTTP请求过长

背景：

业务在使用elasticsearch-spark查询es时，一个简单的macth_all查询报如下错误：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。查询es的索引字段较多，700个field左右。经过查看源码发现elasticsearch-hadoop底层查询es时将索引的字段拼接在url中导致url过长。

方案：

1. 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

SparkSession conf

背景：

olap业务的特殊需求：

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，每一个查询从缓存里命中SparkSession实例，那么该查询就是查询该SparkSession实例里配置的索引。

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**

```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", "false")
    val session1 = SparkSession.builder().config(sparkConf).getOrCreate()
    session1.conf.set("es.index.filter", "index_name1,index_*")
    session1.conf.set("es.nodes", "localhost")
    session1.conf.set("es.port", "9200")
    session1.sql( sqlText = "create table if not exists es_test using es OPTIONS (es.nodes 'localhost', es.port '9200',path 'index_name')")
    val dataFrame = session1.sql( sqlText = "select * from es_test")
    dataFrame.show()

    val session2 = SparkSession.builder().config(sparkConf).getOrCreate()
    session2.conf.set("es.nodes", "localhost")
    session2.conf.set("es.port", "9100")
    val df = EsSparkSQL.esDF(session2,  resource = "ddposorderinfo")
    df.show()
}
```

过滤出要查询的索引

不同的sparkSession可以访问不同的集群或者索引

导出的痛

背景：

ES平台目前有17个ES集群（每个大中心一个集群），一个中心下的所有业务共用该中心的集群，有的中心多个业务有导出需求，甚至他们可能在同一时刻导出，更有甚者一个业务可能上百个人并发导出。这种情况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、GC频繁等。

方案：

1. 修改elasticsearch源码
    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
    4. **功能：**
        1. 默认情况下同一个集群同一时刻只允许一个索引可以导出
        2. 默认2小时之内一个索引累计只能导出20w条数据
        3. 默认2小时之内只允许有一个索引导出，防止业务频繁导出

API格式统一

背景：

delete_by_query和update_by_query API 和其他的API 格式不统一，业务在使用时需要查资料或者咨询ES平台，为了减轻工作量和实现苏宁ES transport client API格式统一，故在相关类中增加几个方法。

方案：

1. 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：

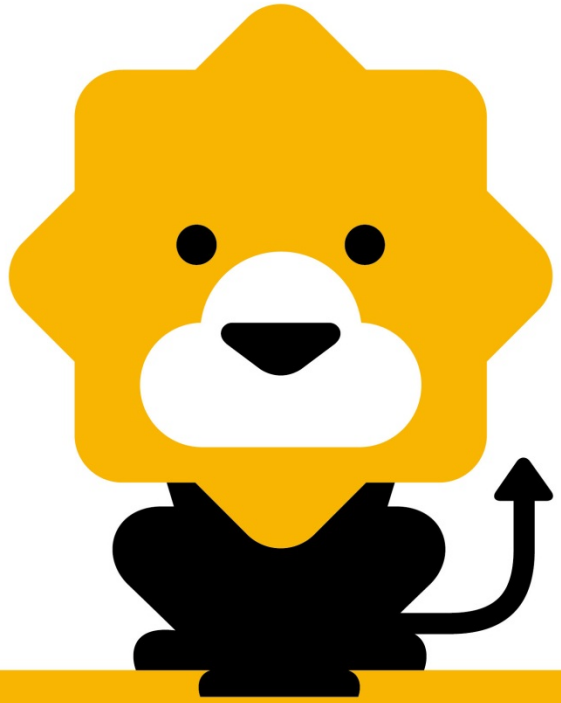
Transport模块有一个专用的跟踪记录器，当被激活时，记录传入和进出请求。您还可以使用一组包含和排除通配符模式来控制哪些操作将被跟踪。默认情况下，每个请求将被跟踪，除了故障检测和ping。

但是日志中并没有打印请求源，我在ES的基础上增加了请求源IP、请求发送、响应接收的详细日志。

方案：

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

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/