基于ES的企业搜索中台

付国庆
2019-12-7, 总经理, 北京顶尖时代科技发展有限公司

目录

elastic

北京顶尖时代科技发展有限公司 TOPTIME

背景

- 2005年成立
- 技术团队来自百度
- 十几年专注于搜索技术领域

产品服务

- 产品 ： 企业搜索中台
- 产品 ： 企业情报系统
- 服务 ： 互联网云采集
- 服务 ： 互联网云搜索

客户500+

- 企业
- 政府
- 金融机构
- 研究机构

目录

elastic

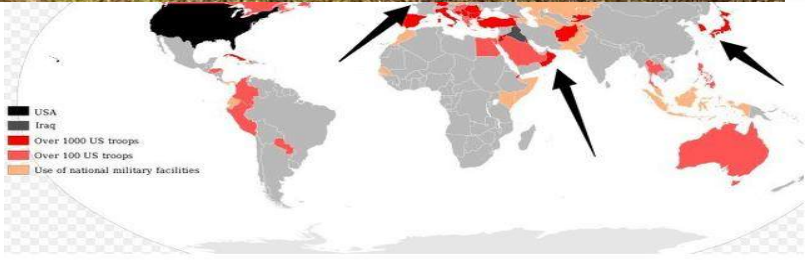
中台概念

战斗单元

前台

中台

军事基地



后台

elastic

互联网公司的中台战略

提出中台战略 ： 大中台 小前台 ；

中台运营数据能力、产品技术能力，帮助前台适应市场、业务支撑、业务创新。

腾讯副总裁：腾讯将进一步开放数据中台和技术中台

京东集团黎科峰：作为一家技术企业 京东将打造产业互联网中台

企业中台

应用形式多

需求变化快

灵活、机动

前台

（业务）

网页　APP　微信　微博　小程序　小视频　大屏　API

营销　渠道　售后　物业　变化？

供应商　监管　生产　决策　创新？

营销部门
服务部门
客户
合作伙伴
供应商

资源、能力、数据

复用、赋能平台

中台

（能力）

业务中台

用户中心　统一认证

统一登录　消息中心

数据中台

数据仓库　建模工具

文件系统　可视化

中台建设团队

稳定、安全、
审计、合规

后台

（管理）

财务管理　人力管理　合同管理　协同办公

库存系统　风控系统　成本系统　营销系统

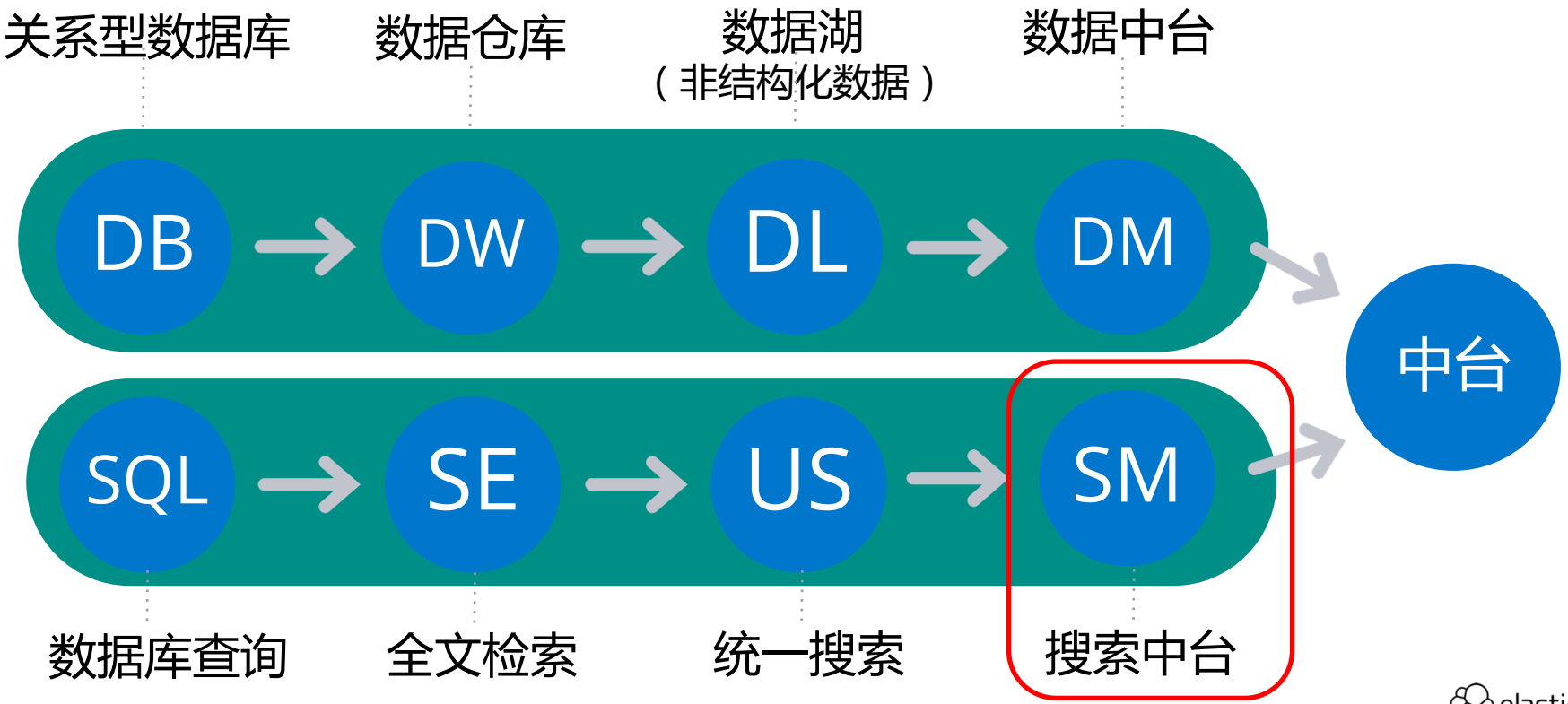
集团内部
运营部门

IAAS 企 业 云

目录

elastic

数据与搜索的交汇


关系型数据库
数据仓库
数据湖
（非结构化数据）
数据中台
DB
DW
DL
DM
中台
SQL
SE
US
SM
数据库查询
全文检索
统一搜索
搜索中台

企业搜索中台

应用形式多

需求变化快

灵活、机动

前台（业务）

网页 APP 微信 微博 小程序 小视频 大屏 API

营销 渠道 售后 物业 变化？

供应商 监管 生产 决策 创新？

营销部门
服务部门
客户
合作伙伴
供应商

资源、能力、数据

复用、赋能平台

中台（能力）

搜索中台（SM）

非结构化数据汇集、管理、利用、服务

业务中台

数据中台

中台研发团队

稳定、安全、
审计、合规

后台（管理）

财务管理 人力管理 合同管理 协同办公

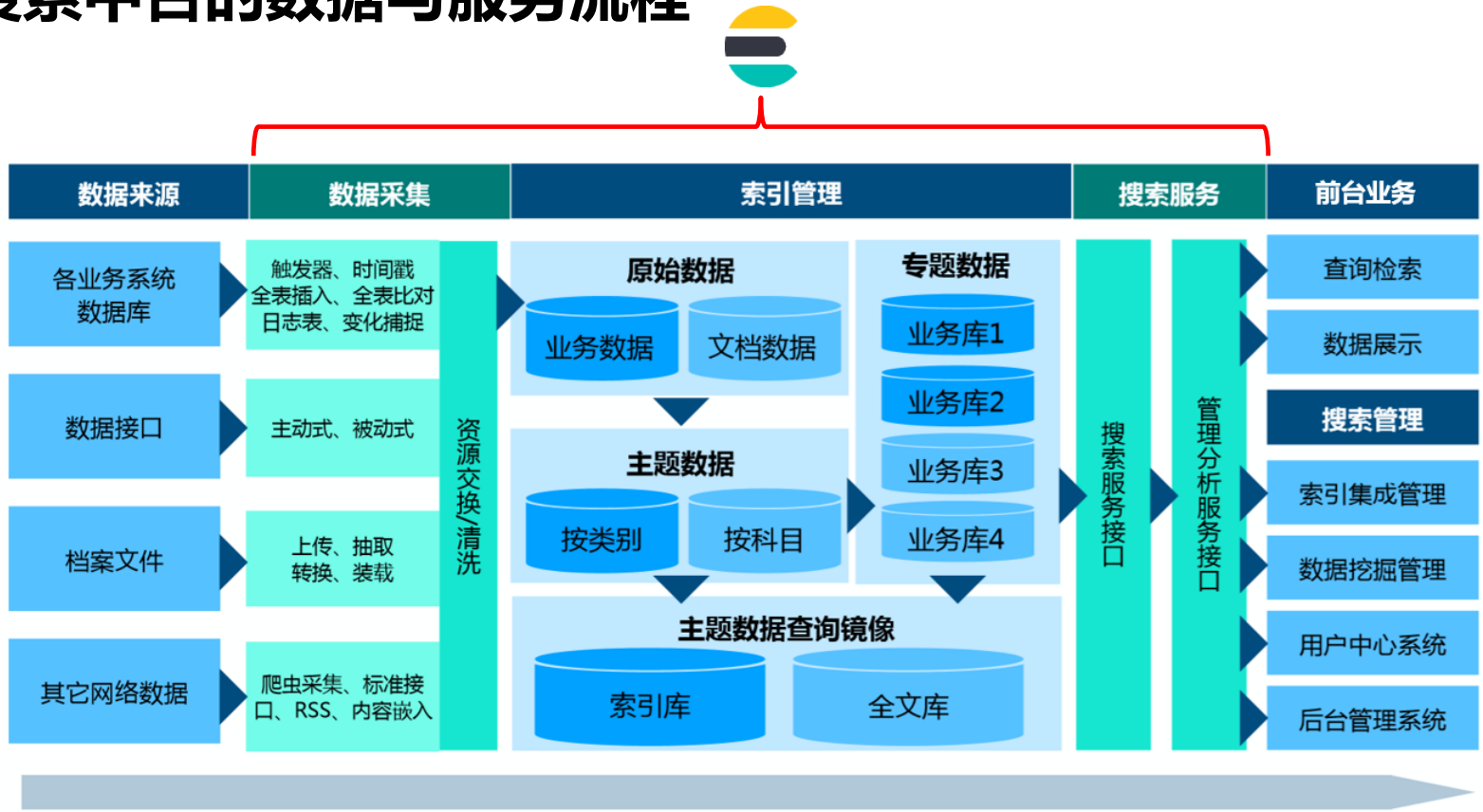
库存系统 风控系统 成本系统 营销系统

集团内部
运营部门

IAAS 企 业 云

elastic

搜索中台的数据与服务流程



elastic

目录

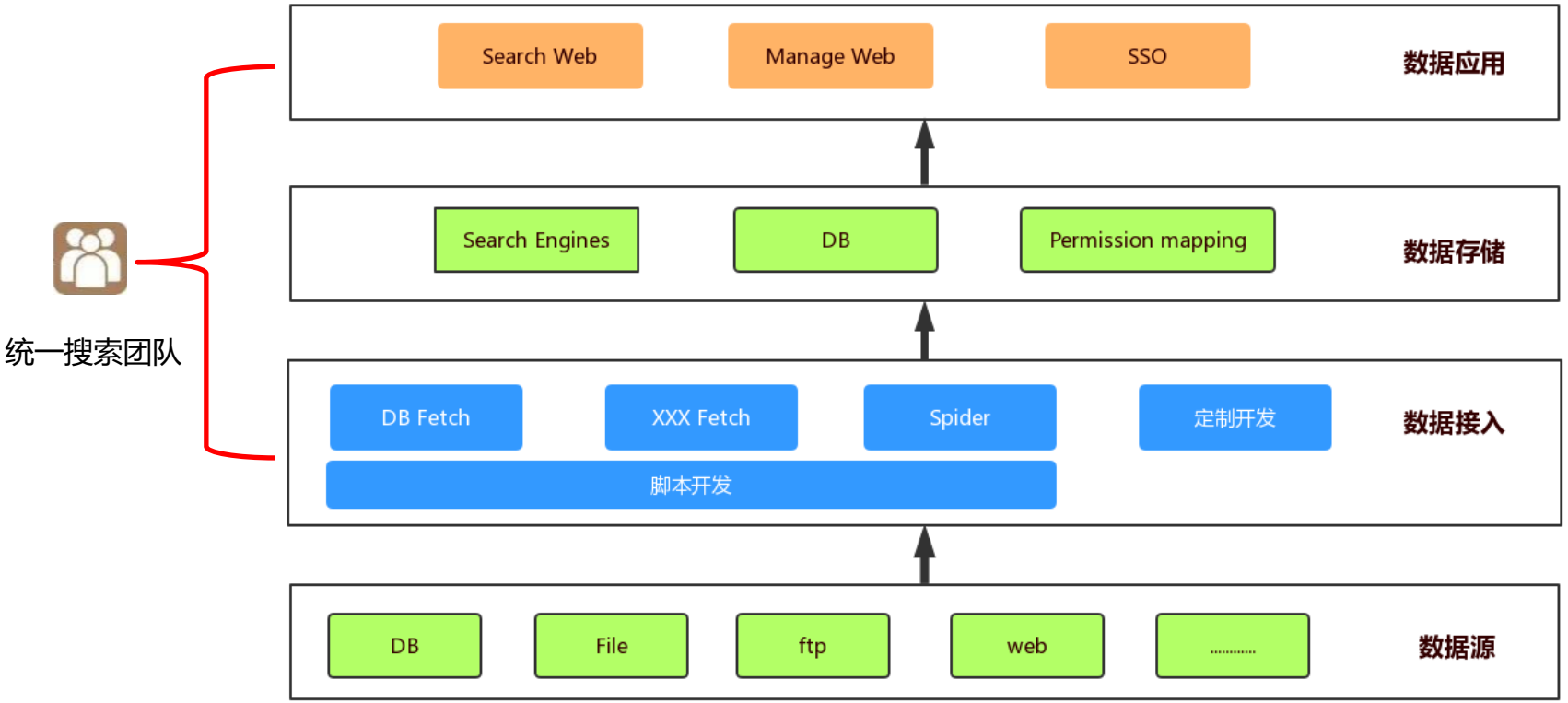
elastic

统一搜索

搜索中台

发展演变

统一搜索流程



elastic

统一搜索方案

优点

实现了企业内部资源的一键搜索，解决信息孤岛

统一登录，继承业务系统权限进行搜索

缺点

开发、实施周期长，工作量大，

搜索技术团队，需适应各业务系统进行开发

缺少复用性

后期各系统升级、调整后，搜索功能需要调整

企业搜索中台优势 专业的事情交给专业团队完成

快速开发

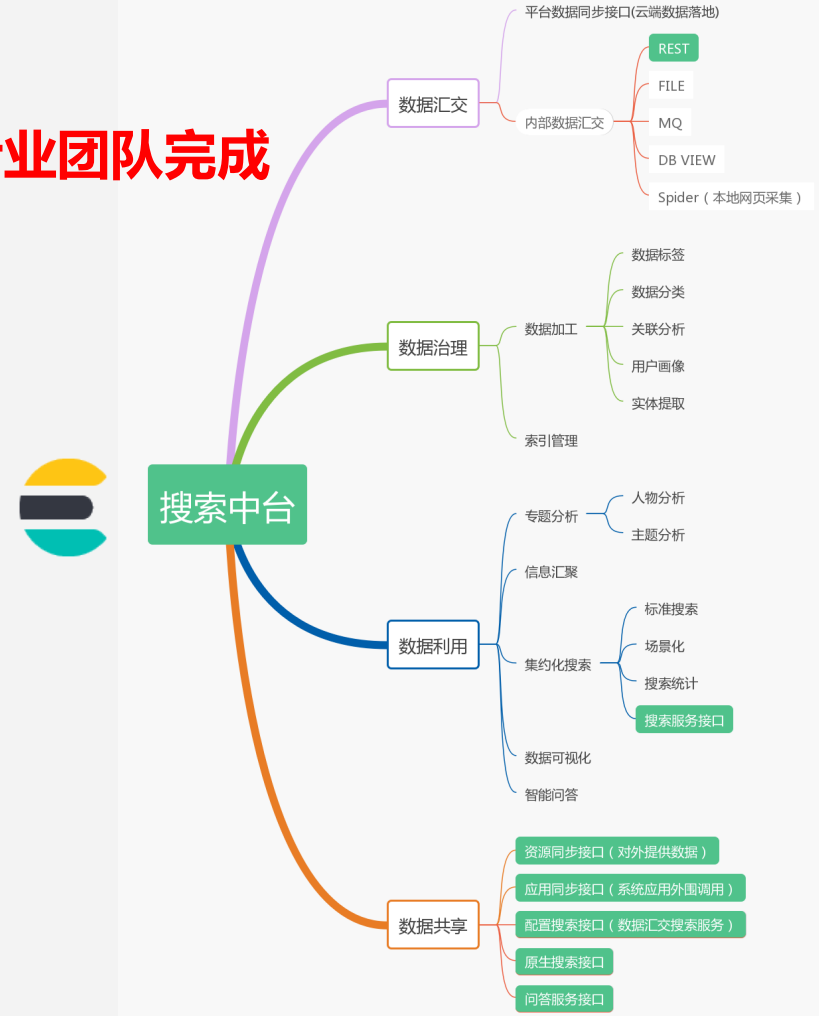
接入、搜索实现了可视化配置

统一解决方案

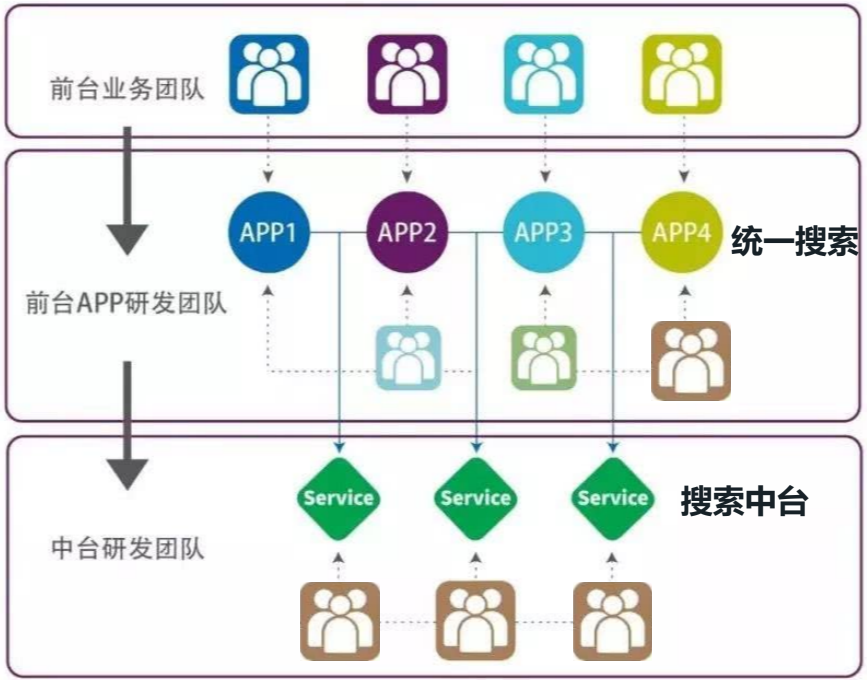
不直接面向应用,提供统一解决方案，支持二次开发，
各系统变化时工作量大幅减少

灵活复用

功能、数据多次利用，解决方案不局限在某一领域

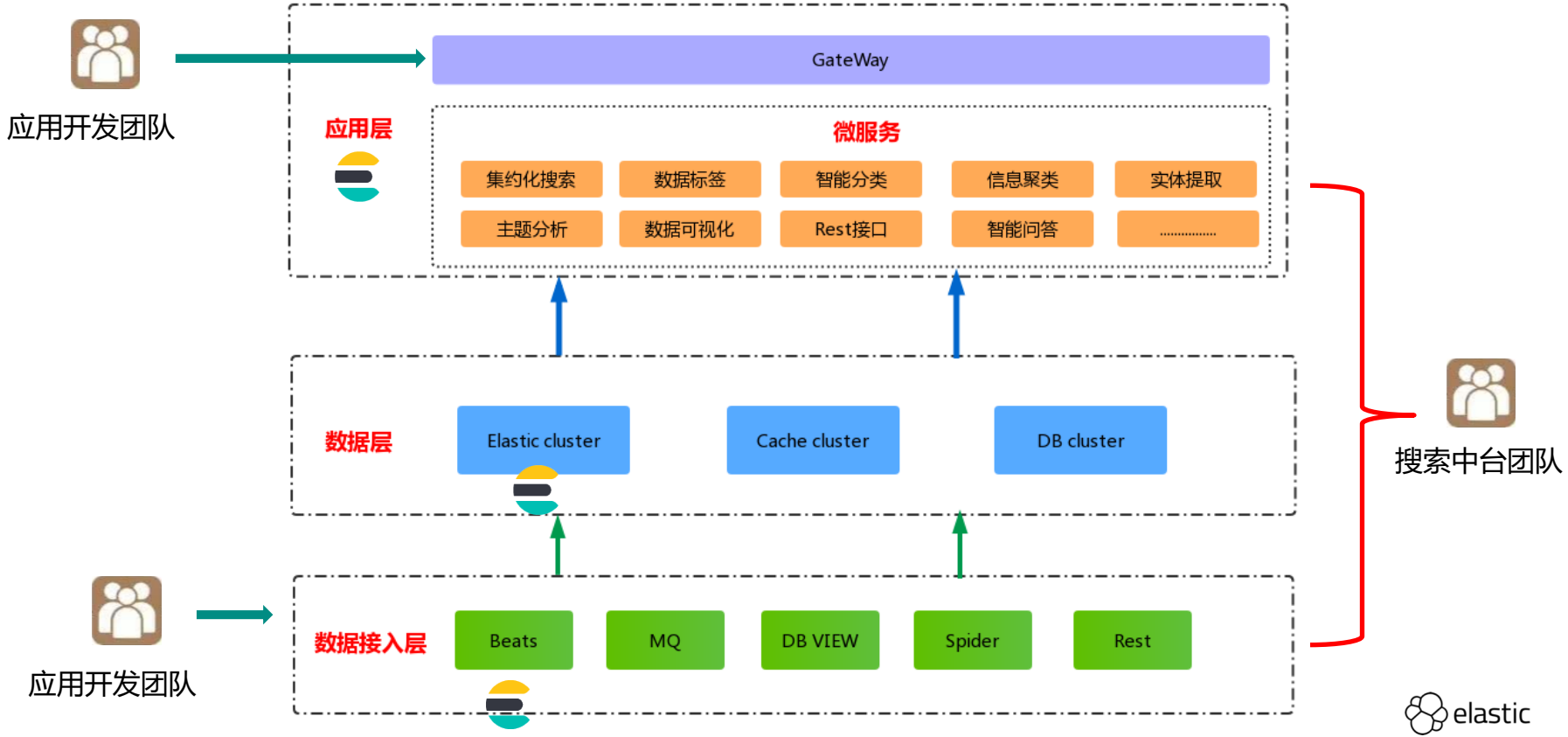


采用搜索中台架构后，研发团队定位变化



elastic

基于ES搜索中台架构


应用开发团队
GateWay
应用层
微服务
集约化搜索
数据标签
智能分类
信息聚类
实体提取
主题分析
数据可视化
Rest接口
智能问答
.....
数据层
Elastic cluster
Cache cluster
DB cluster
搜索中台团队
应用开发团队
数据接入层
Beats
MQ
DB VIEW
Spider
Rest

数据汇交

REST接口

支持EleasticSearch原始脚本进行数据提交

MQ方式

支持Kafka、RabbitMQ、RocketMQ方式进行增量数据提交

DB视图方式

支持ORACLE、MYSQL、SQLSERVER等数据库接口视图进行可视化配置

文件方式

支持csv文件方式进行初始化、增量数据提交

数据汇交

REST接入样例

接口参数说明

调用地址：http:/[redacted]est/bulk
请求方式：POST
返回类型：JSON

请求头：

名称	类型	是否必须	描述
X-Gaia-Api-Key	STRING	必填	网关校验信息，请在网关管理处申请

请求参数：

名称	类型	是否必须	描述
indexDb	STRING	必填	索引库别名
requestBody	STRING	必填	批量数据

测试样例：

```
indexDb=resttestdb0
requestBody=
{ "index" : { "_id" : "1" } }
{ "field1" : "value1" }
```

接口请求测试

indexDb *: oa

requestBody *:

发送请求

测试返回结果

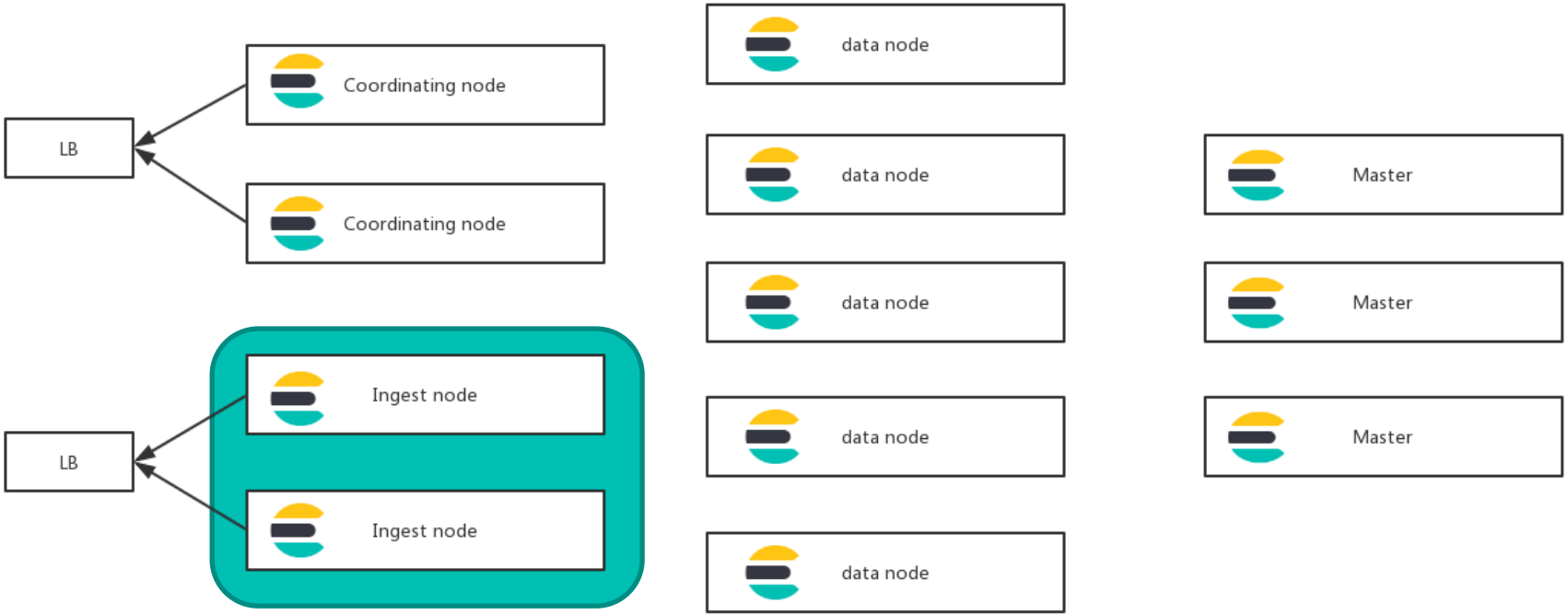
elastic

数据汇交

MQ接入样例

```
{
  “customerId”: “56431”,//租户ID
  “indexName”: “oadb”,//索引
  “mappingName”: “cps”,//索引结构
  “type”: 0,//操作类别（增删改）
  “dataList”: [//数据详情
      {
      "DREREFERENCE": "",
      "DRETITLE": "",
      "URL": ""
      }
  ]
}
```

数据治理



elastic

数据治理

Pipeline

```
curl -XPUT "http://192.168.0.7:9200/_ingest/pipeline/toptime-default-pipeline" -u
'elastic:toptime' -H 'Content-Type: application/json' -d'
{
  "description": "toptime default pipeline",
  "processors": [
    {
      "set": {
        "field": "DREDBNAME",
        "value": "{{_index}}"
      }
    },
    {
      "script": {
        "source": "if(ctx._index==~ /(.*?)[\\d-_]*?$/){ctx.DREDBNAME=/(.*?)[\\d-
_]*?$/.matcher(ctx._index).replaceAll(\"$1\")}"
      }
    }
  ]
}'
```

数据利用

基于ES搜索中台微服务

搜索功能	智能化服务	场景化服务	索引管理 搜索统计服务	接口服务
关键词搜索	下拉提示	主题场景	搜索次数统计	互联网云采集接口 Cloud Spider
分类导航搜索	智能纠错	应用场景	搜索热词统计	
搜索结果分类统计	拼音搜索	机构名片	搜索点击统计	数据库同步接口 DB View
复合条件搜索	关键词推荐	人物名片	搜索转化率统计	
标签过滤搜索	通俗语言	知识图谱	搜索改进率统计	文档提交与解析接口 FILE
地域过滤搜索	敏感词屏蔽	图形化	索引库管理	
图片搜索	框计算	智能问答	索引数据管理	API提交接口（REST）
附件文档搜索	智能排序	地图搜索	词汇管理	API搜索接口（REST）
微信、微博搜索	搜索结果置顶	用户画像		API删除接口（REST）
视频搜索	智能推荐			MQ
	智能语义识别			

数据利用

- Query DSL
  - Query and filter context
  - + Compound queries
  - + Full text queries
  - + Geo queries
  - + Joining queries
  - Match all
  - + Span queries
  - + Specialized queries
  - + Term-level queries
  - `minimum_should_match` parameter
  - `rewrite` parameter
  - Regular expression syntax

elastic

数据利用

搜索代码解耦-Search Template

```
POST _scripts/template1
{
  "script": {
    "lang": "mustache",
    "source": "{"from": {{from}}{{^from}}0{{/from}},"size":
{{size}}{{^size}}20{{/size}},"query": {"bool": {"should":
[{"bool": {"must": [{"bool": {"should": [{"query_string":
{"query":
"{{queryText}}","fields":["{{queryField1}}"{{#queryField2}},"{
{queryField2}}"{{/queryField2}}],"default_operator":
"and","auto_generate_phrase_queries": true,"boost":
{{completeBoost}}{{^completeBoost}}2{{/completeBoost}}}}}"}"
  }
}
```

数据共享

搜索接口配置

elastic

系统安全

ES访问安全

文件和原生 Realm，可用于创建和管理用户TLS功能，可对通信进行加密

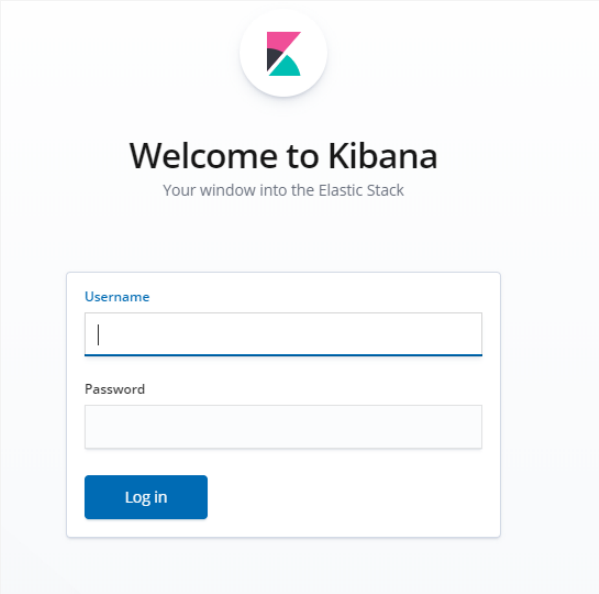
每个Elasticsearch节点生成私钥和X.509证书

租户权限分离

租户单独进行授权，禁止跨库操作

公私钥分发

租户单独分发秘钥，通过RSA方式进行接口通信



Elasticsearch 安全功能现免费提供
（从 6.8.0 和 7.1.0 版本开始）

系统运维&统计

DevOps &Count

elastic

APM-运维监控分析

日志分析

- 中间件日志
- 应用日志

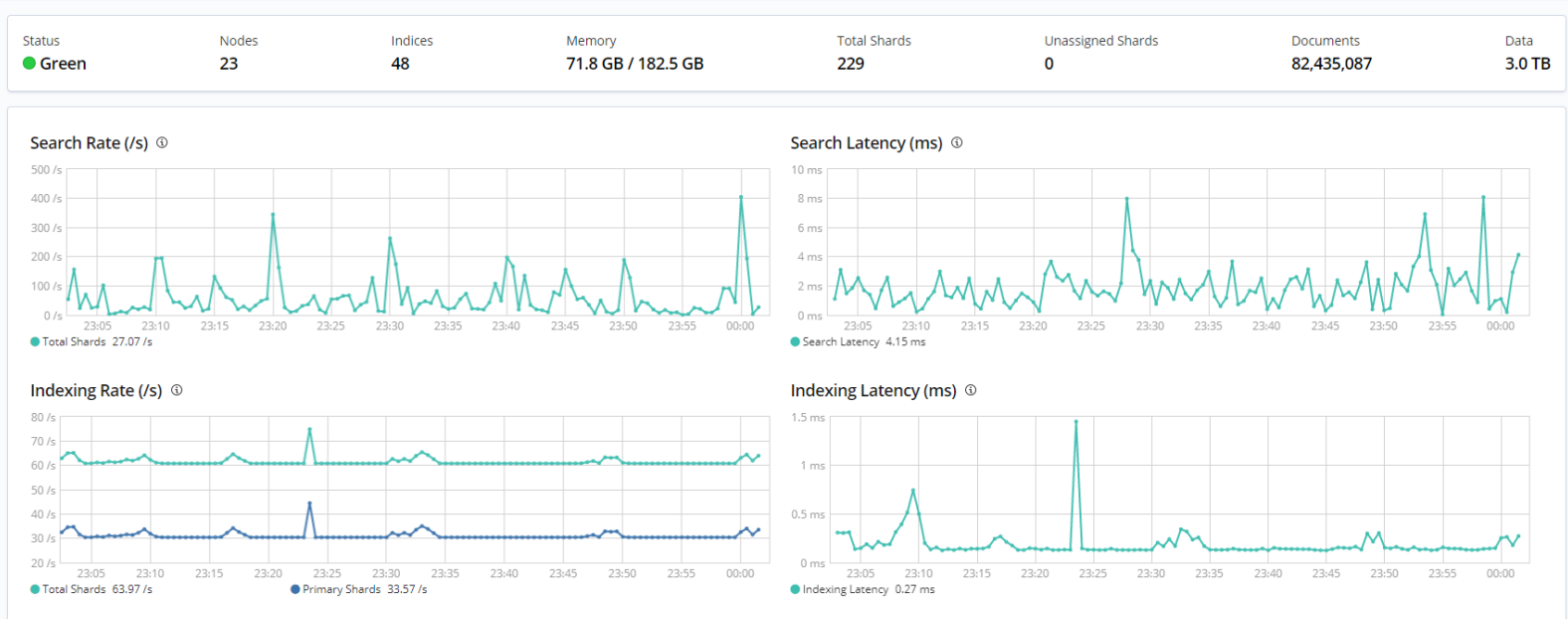
系统指标

- CPU
- RAM
- HDD

安全报警

- 机器人攻击
- 恶意刷榜

系统监控



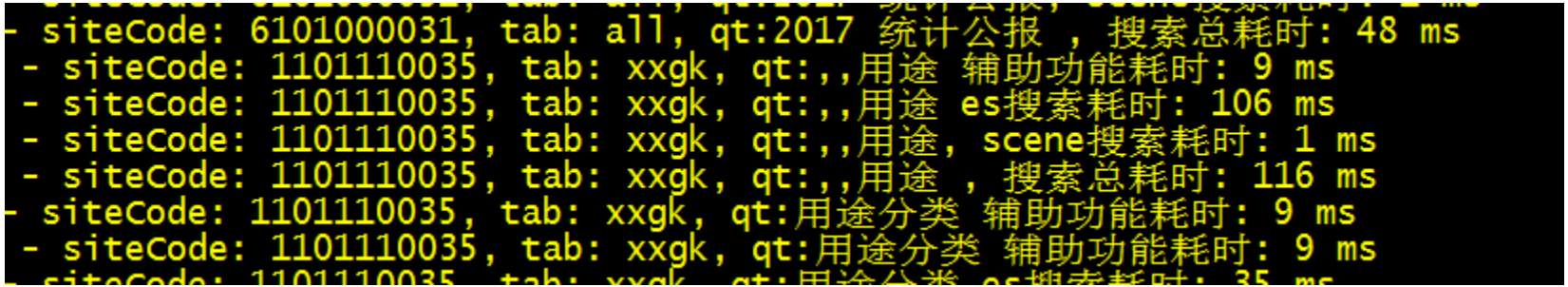
elastic

系统监控

Name	Status	CPU Usage	Load Average	JVM Memory	Disk Free Space ↑	Shards
★ master-01 192.168.0.7:9300	Online	0% ↑ 1% max 0% min	0.16 ↓ 0.23 max 0 min	44% ↑ 45% max 40% min	52.0 GB ↓ 52.0 GB max 52.0 GB min	0
master-03 192.168.1.99:9300	Online	0% ↑ 1% max 0% min	0 ↓ 0.27 max 0 min	22% ↑ 26% max 20% min	52.3 GB ↓ 52.3 GB max 52.3 GB min	0
master-02 192.168.1.98:9300	Online	0% ↑ 1% max 0% min	0.01 ↑ 0.17 max 0 min	21% ↑ 26% max 20% min	52.3 GB ↓ 52.3 GB max 52.3 GB min	0
data-04 192.168.0.12:9300	Online	0% ↓ 5% max 0% min	0.03 ↓ 0.14 max 0 min	61% ↑ 63% max 58% min	101.5 GB ↑ 101.5 GB max 101.3 GB min	12
data-05 192.168.0.18:9300	Online	0% ↑ 1% max 0% min	0.08 ↑ 0.11 max 0 min	44% ↑ 46% max 41% min	102.8 GB ↓ 102.8 GB max 102.7 GB min	11
data-11 192.168.1.106:9300	Online	0% ↑ 1% max 0% min	0.03 ↓ 0.15 max 0 min	43% ↑ 47% max 41% min	105.3 GB ↓ 105.3 GB max 105.3 GB min	12
data-01 192.168.0.8:9300	Online	0% ↑ 2% max 0% min	0.15 ↑ 0.34 max 0 min	24% ↑ 24% max 19% min	108.2 GB ↓ 108.2 GB max 108.1 GB min	11
data-08 192.168.0.15:9300	Online	0% ↑ 1% max 0% min	0.05 ↑ 0.43 max 0 min	38% ↑ 39% max 34% min	111.3 GB ↓ 111.3 GB max 111.2 GB min	11
data-06 192.168.0.14:9300	Online	0% ↑ 1% max 0% min	0.01 ↑ 0.21 max 0 min	59% ↑ 60% max 55% min	111.3 GB ↓ 111.3 GB max 111.3 GB min	11
data-19 192.168.1.102:9300	Online	0% ↑ 3% max 0% min	0.08 ↑ 0.72 max 0 min	57% ↑ 60% max 54% min	113.7 GB ↓ 113.7 GB max 113.7 GB min	12

elastic

日志分析预警



IP访问信息

IP	Num
121.9.216.126	474
218.19.215.148	195
222.95.171.222	180
103.85.144.198	151

elastic

有效控制企业构建的成本

ES服务器硬件要求

(PC server|虚拟化主机)

- CPU　8核
- RAM　16G/32G
- HDD　SSD
- OS　Centos 7.X
- 参数配置
  - Ulimit
  - Swappiness
  - JVM等

未来扩展
2020

elastic

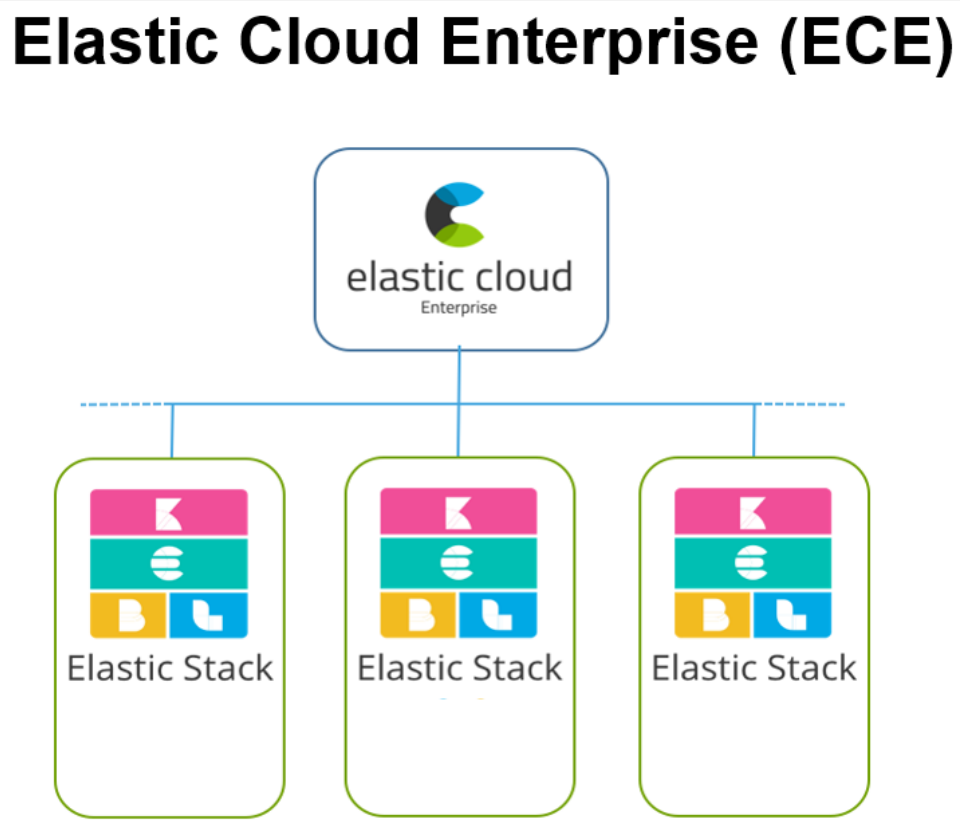
未来扩展

集成ECE

扩展多集群应用场景，简化Elastic Stack集群维护成本

租户安全分离

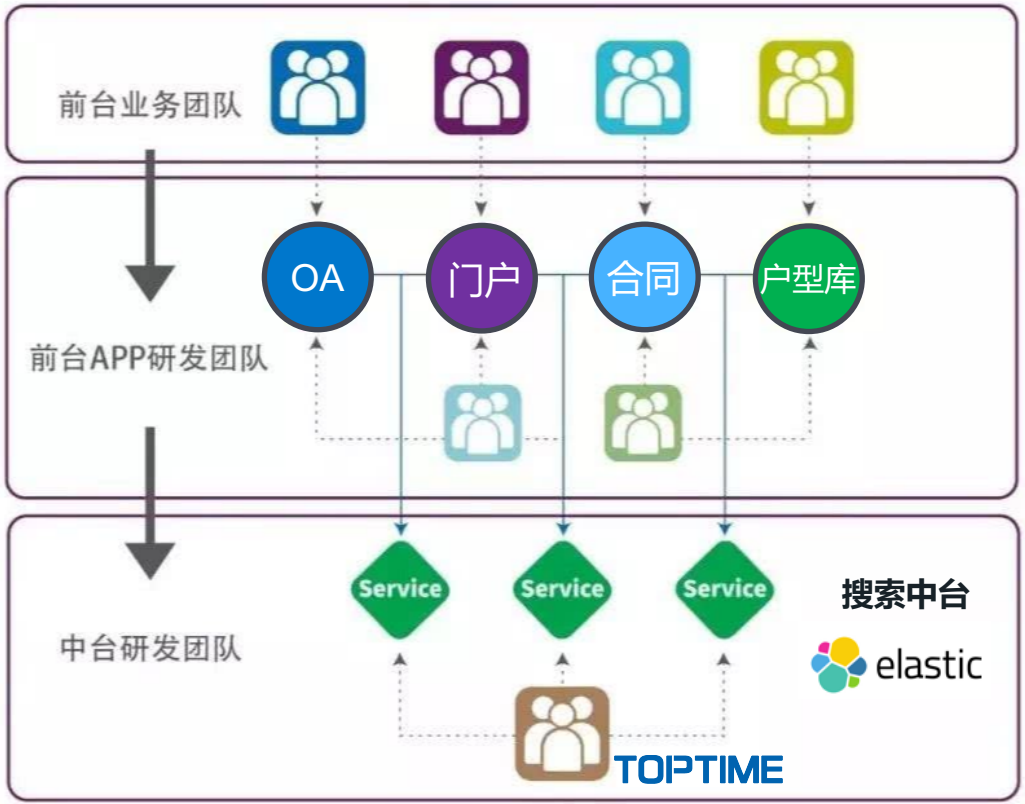
租户在应用端、后台管理端均进行隔离，有效保障数据安全性



目录

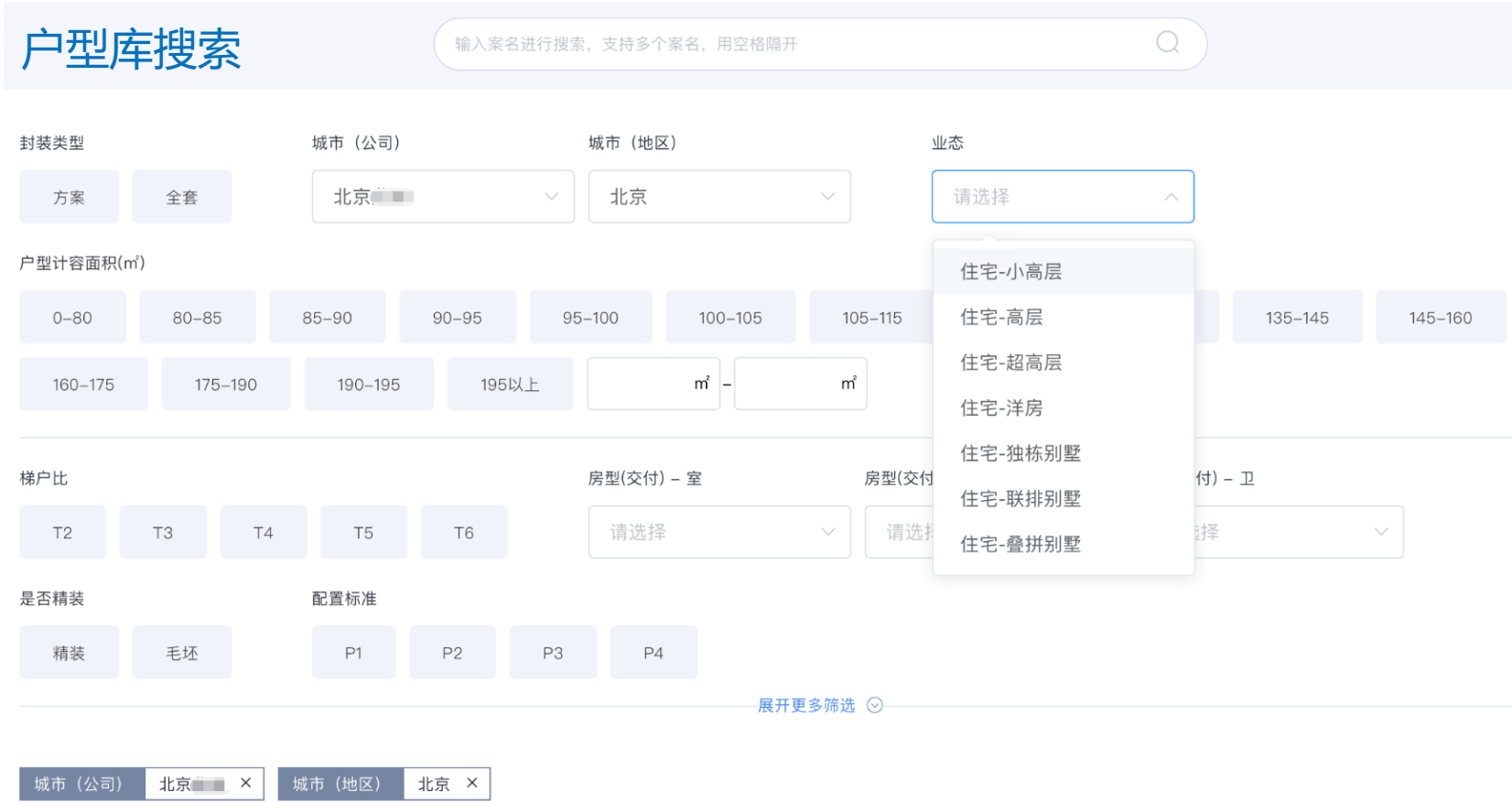
elastic

搜索中台支撑了企业若干系统的搜索功能建设

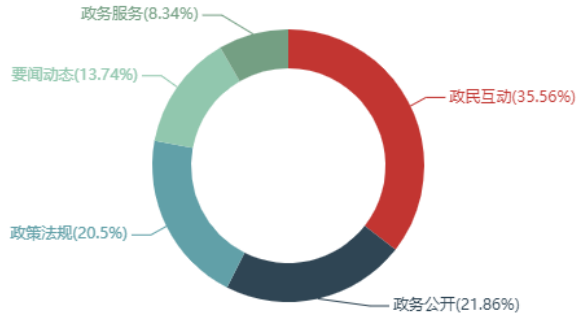


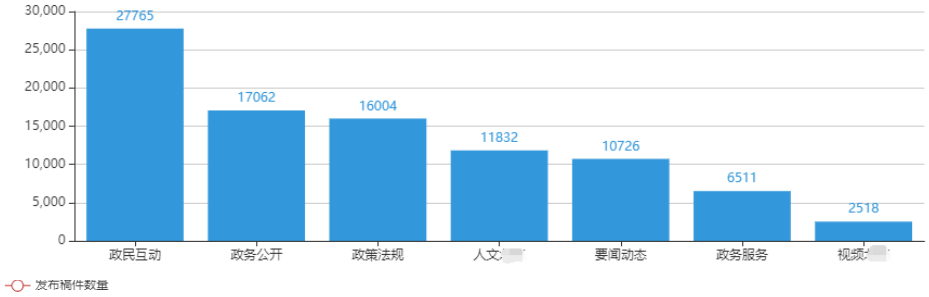
elastic

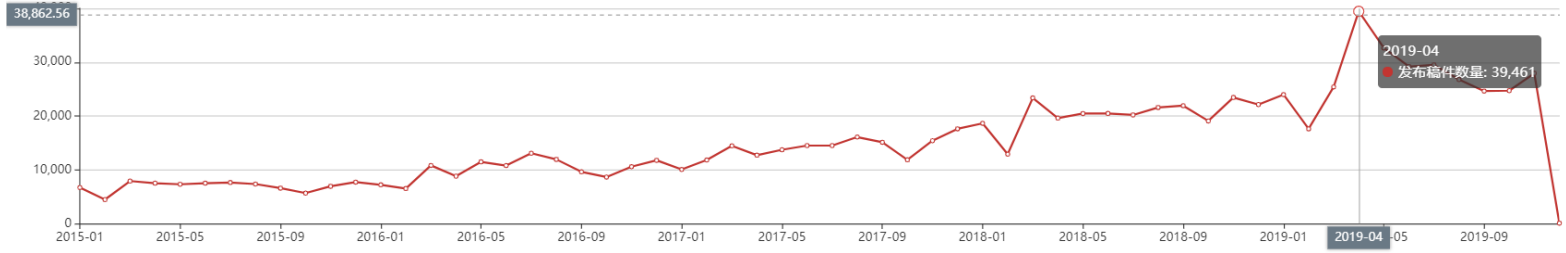
搜索中台-灵活构建复合条件搜索应用

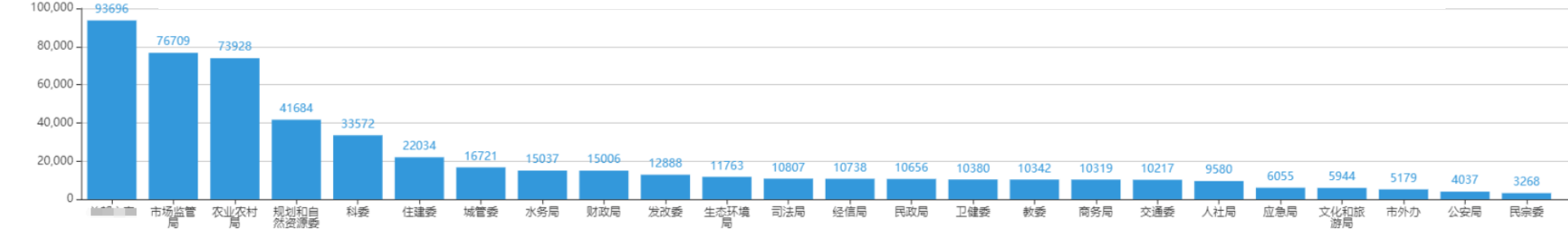


搜索中台-快速构建数据可视化分析专题 ：大数据统计



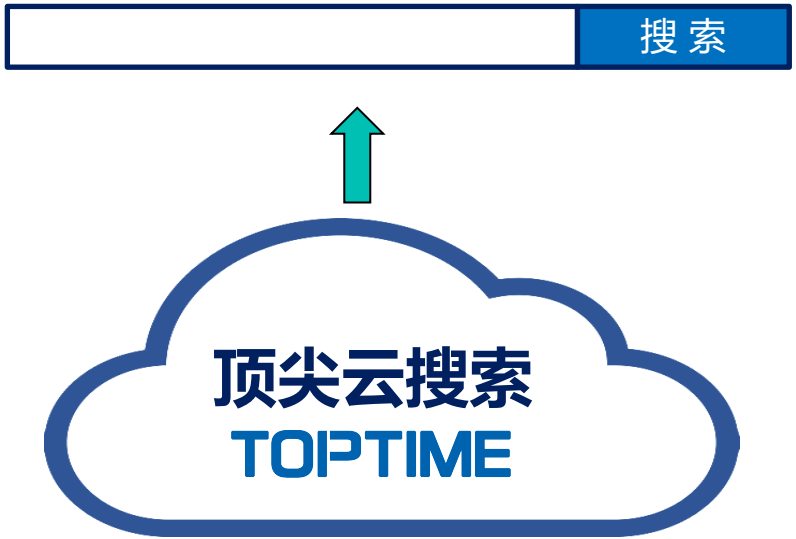






总结

支撑1000+网站搜索服务



elastic 搜索中台

谢 谢

elastic