elastic | 中国开发者大会

华为云-云搜索服务Elasticsearch实践

胡斐然
华为，云服务技术总监

HUAWEI

华为云-云搜索服务Elasticsearch实践

2018.11

云搜索服务
Cloud Search Service

站内搜索	多媒体	日志&指标	运维
智能分词	协同搜图	数据协同	服务化
越搜越准	音乐检索	内核增强	专业看护

智能分词

商品检索

坐席知识库

自动对话服务

- 新词或特有名词无法搜索
- 短语问题
- 多租户的支持

泊松分词器

- 支持未登录词识别
- 对英文更友好
- 支持多租户，对每个Index分别设定词库
- 支持多粒度分词，对每个词进粗细粒度的控制
- 支持简繁体、拼音分词
- 支持同义词

问题一： 新词或者特有名词无法搜索

举例：李世石是人名

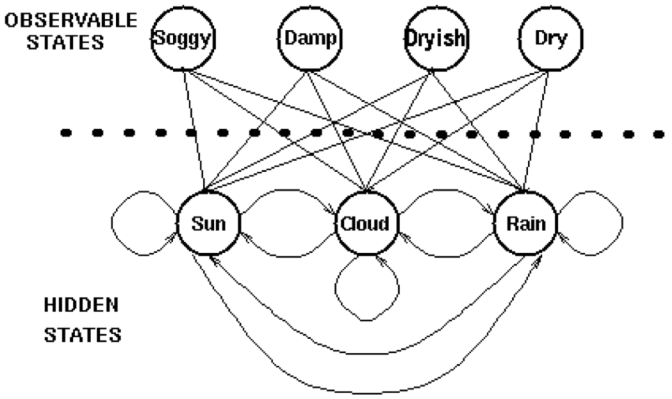
Ik_max_word:
李/世/石/是/人/人名/名

Ik_smart:
李/世/石/是/人名

poission_hmm:
李世石/是/人名

说明：
默认词库中没有的人名、地名、特有名词不能准确分词

解决思路：



使用隐马尔可夫模型
观察态：字符
隐藏态：BMES

使用华为积累的知识库、论坛语料训练出来基础模型

问题二： 英文短语

举例：windows system

```
GET _analyze
{
  "analyzer": "ik_max_word",
  "text": ["windows system"]
}

GET _analyze
{
  "analyzer": "ik_smart",
  "text": ["windows system"]
}
```

```
{
  "tokens": [
    {
      "token": "windows",
      "start_offset": 0,
      "end_offset": 7,
      "type": "ENGLISH",
      "position": 0
    },
    {
      "token": "system",
      "start_offset": 8,
      "end_offset": 14,
      "type": "ENGLISH",
      "position": 1
    }
  ]
}
```

说明：

IK分词器中对英文是按照空格切分的，即便是在词库中配置windows system，也无法当做短语来识别

解决思路：

自定义词库文件中，一个词一行，空格识别为普通字符

```
GET token_test1/_analyze
{
  "analyzer": "my_analyzer",
  "text": ["windows system"]
}
```

```
{
  "tokens": [
    {
      "token": "windows system",
      "start_offset": 0,
      "end_offset": 14,
      "type": "word",
      "position": 0
    },
    {
      "token": "windows",
      "start_offset": 0,
      "end_offset": 7,
      "type": "word",
      "position": 0
    },
    {
      "token": "system",
      "start_offset": 8,
      "end_offset": 14,
      "type": "word",
      "position": 1
    }
  ]
}
```

问题三： 词库无法在索引上单独配置，全局词库有干扰

举例：
有两个词库文件A.dict/B.dict，想在索引A中使用A.dict，索引B中使用B.dict。

IK分词器配置 IKAnalyzer.cfg.xml：

```
<properties>
   <comment>IK Analyzer 扩展配置</comment>
   <entry key=“ext_dict”>/path/A.dict; /path/B.dict</entry>
</properties>
```

说明：
IK分词器的词库配置在配置文件中，属于全局配置，和索引没有产生任何联系。因此，任何词库都是全局生效的，无法达到想要的结果。

解决思路：
建立索引的时候在setting中配置

```
PUT A
{
  "settings": {
    "analysis": {
      "tokenizer": {
        "my_tokenizer_index": {
          "type": "poisson_index",
          “poisson_dict”: [“资源管理器","P20 pro","喜大普奔"],
          “poisson_dict_path”: “A.dict，A1.dict",
          "poisson_stopword_dict_path":"stopwordA.dict“
        },
        "analyzer": {
          "my_analyzer_index": {
            "tokenizer": "my_tokenizer_index"
          }
        }
      }
    }
}
```

智能搜索（越搜越准）

传统搜索引擎都是基于文本匹配的算法来打分排序，没有考虑其它因素，比如搜索内容的热度等。华为云云搜索服务提供基于用户历史数据提升搜索排序的功能，可以提高此类场景的用户体验。

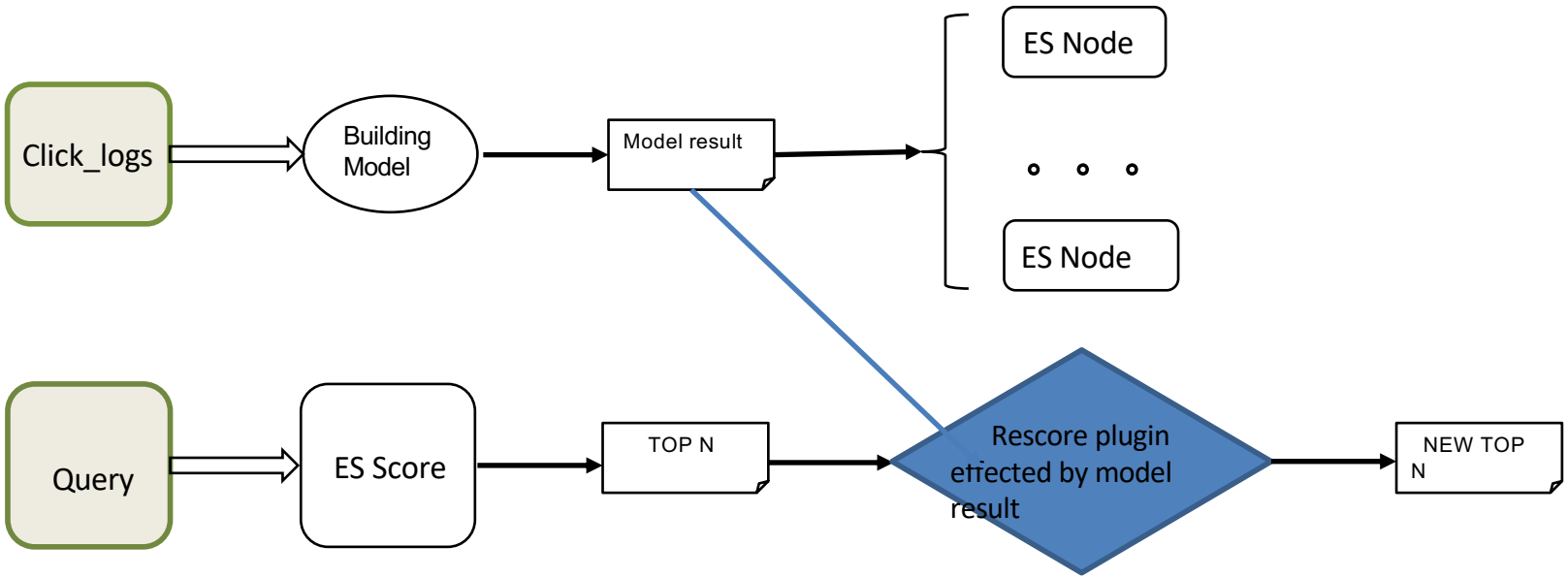
适用场景：

- 帖子搜索
- 新闻搜索
- 视频摘要搜索
- 内容推荐

HUAWEI

根据历史数据训练模型进行重排序——方案

搜索内容	点击文档	Index	点击时间	原始排序
华为	华为Mate 20发布会	product	2018-10-11 12:24:00	6



越搜越准训练模型后重排序——效果

原始搜索

```
{
  "query": {
    "match": {
      “desc”: “华为"
    }
  }
}
```

文档	score
探访华为总部：华为大学开2500门课	0.7911257
华为Mate 20发布会	0.5753642
华为技术逆天！每年跟苹果收几十亿专利费	0.5753642

越搜越准模型搜索：

```
{
  "query": {
    "match": {
      “value”: “华为"
    }
  },
  "rescore": {
    "window_size": 10,
    "poisson_rank": {
      “keyword”: “华为",
      "boost": 4,
      “total_match": 0
    }
  }
}
```

效果：重打分后历史被点击的越多的记录排序越靠前

文档	score
华为Mate 20发布会	1.5241129
华为技术逆天！ 每年跟苹果收几十亿专利费	1.0712540
探访华为总部：华为大学开2500门课	0.7925697

协同搜图

以文搜图

鲜花预定_厦门鲜花店送花 ...
xm.hua.com

以图搜图

协同搜图

“2018 感冒药”

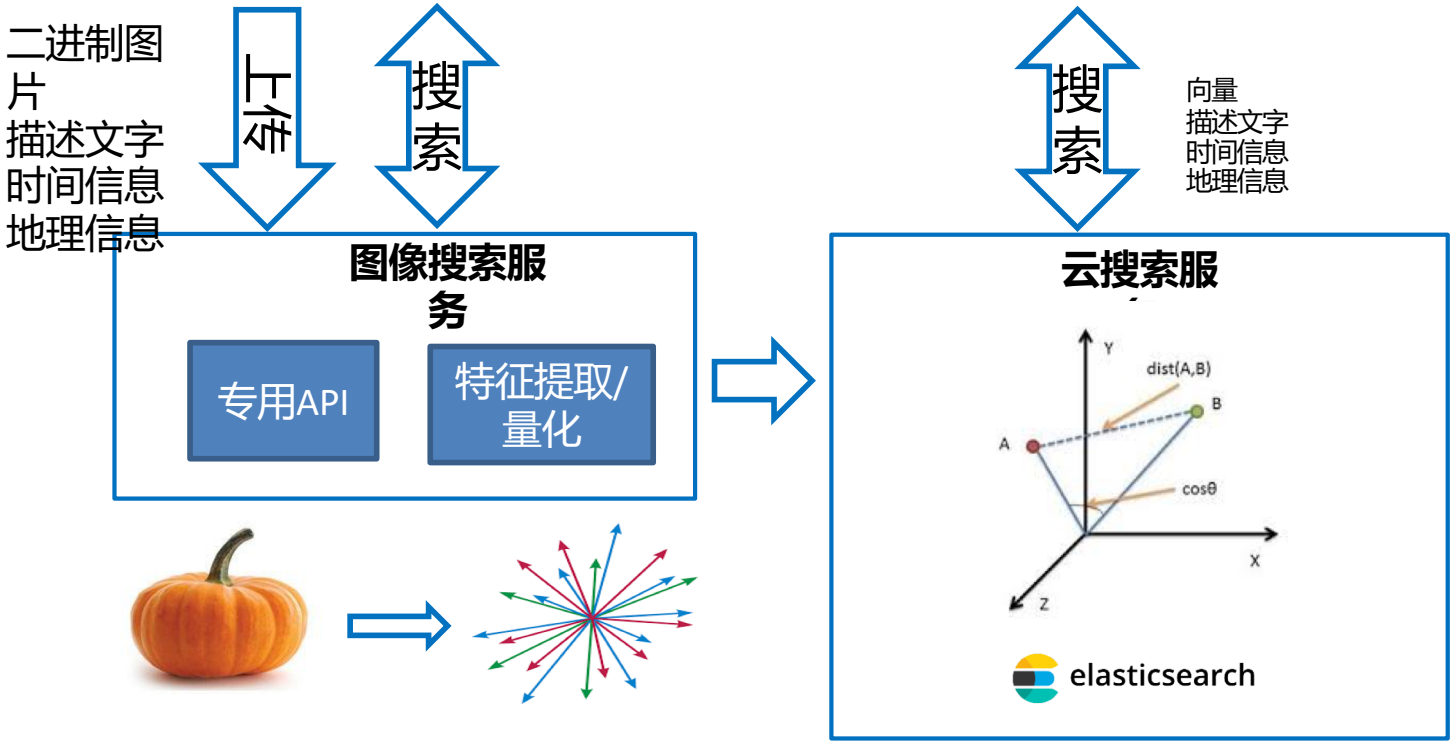
适用场景：

- 图像版权
- AR呈现
- 商品检索
- 素材检索

HUAWEI

协同搜图——方案

云搜索服务在elasticsearch基础上增加向量检索的能力，和elasticsearch已有的检索能力相融合



支持的向量表达方式

- 原始浮点向量
- 乘积向量
- 局部性Hash
- 混合

支持的距离函数

- 欧式距离
- 点积
- 夹角
- 混合

协同搜图——效果

南瓜的做法_健康饮食_疾病网

南瓜- Danbaoli blog Danbaoli blog

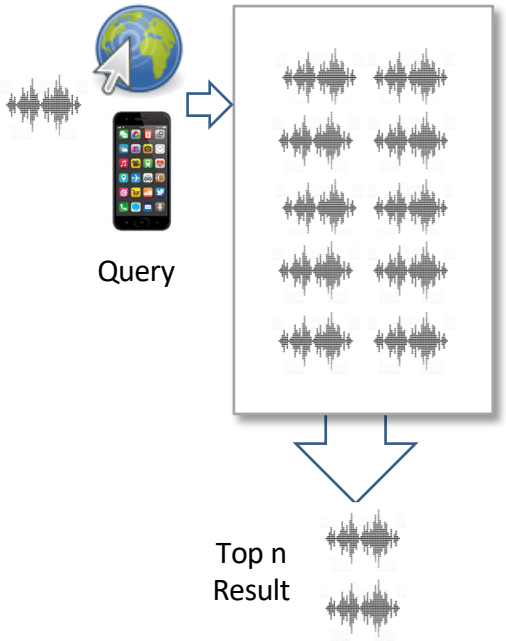
南瓜（南瓜葫芦科植物）_百度百科

+万圣节+绿叶

协同搜图场景：亿级图片，毫秒级完成查询，准确率99%

音频检索

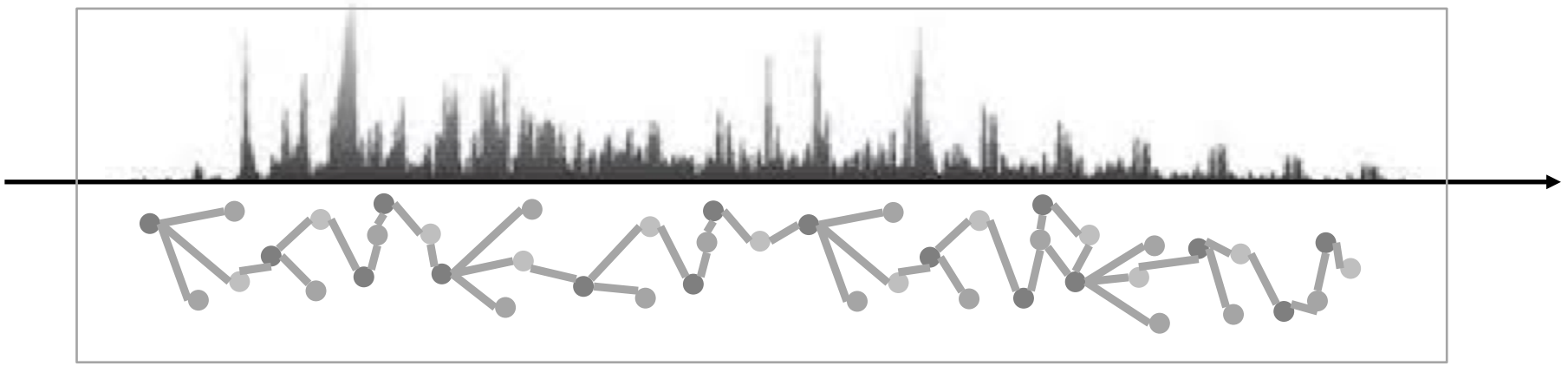
通过一段音频片段，检索出音频库中相似的音频内容



适应场景：

- 音频、音乐版权
- 第二屏关联
- 旋律找歌
- 媒资管理
- 媒体去重

音频检索——问题



[SHA1(P,P,dt),SHA1(P,P,dt),……]　[874dae83e3805c3bdd7d,e99eef00f40c704f620c,……]

业界普遍的做法，是将一段波形通过频谱变换找到一些高能量的点，然后将高能量点序列组成音频中的指纹。所以，音频检索的问题在于：

如何在大量音频内容高效的找到尽可能多的匹配指纹？

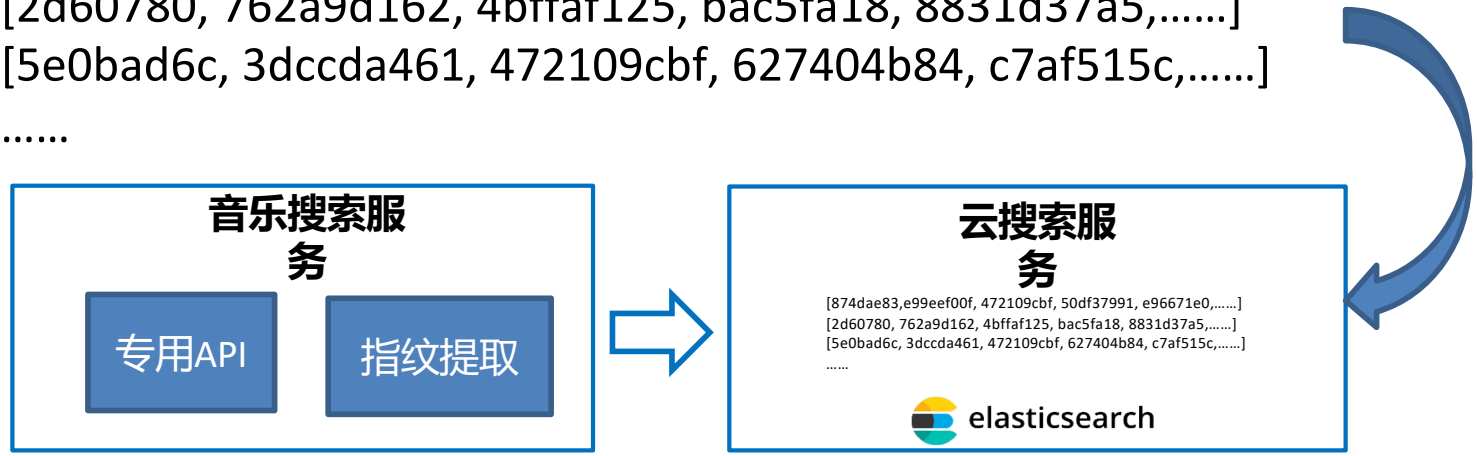
音频检索——方案

Query：

[874dae83,e99eef00f,……]

Documents

[874dae83,e99eef00f, 472109cbf, 50df37991, e96671e0,……]
[2d60780, 762a9d162, 4bffaf125, bac5fa18, 8831d37a5,……]
[5e0bad6c, 3dccda461, 472109cbf, 627404b84, c7af515c,……]
……

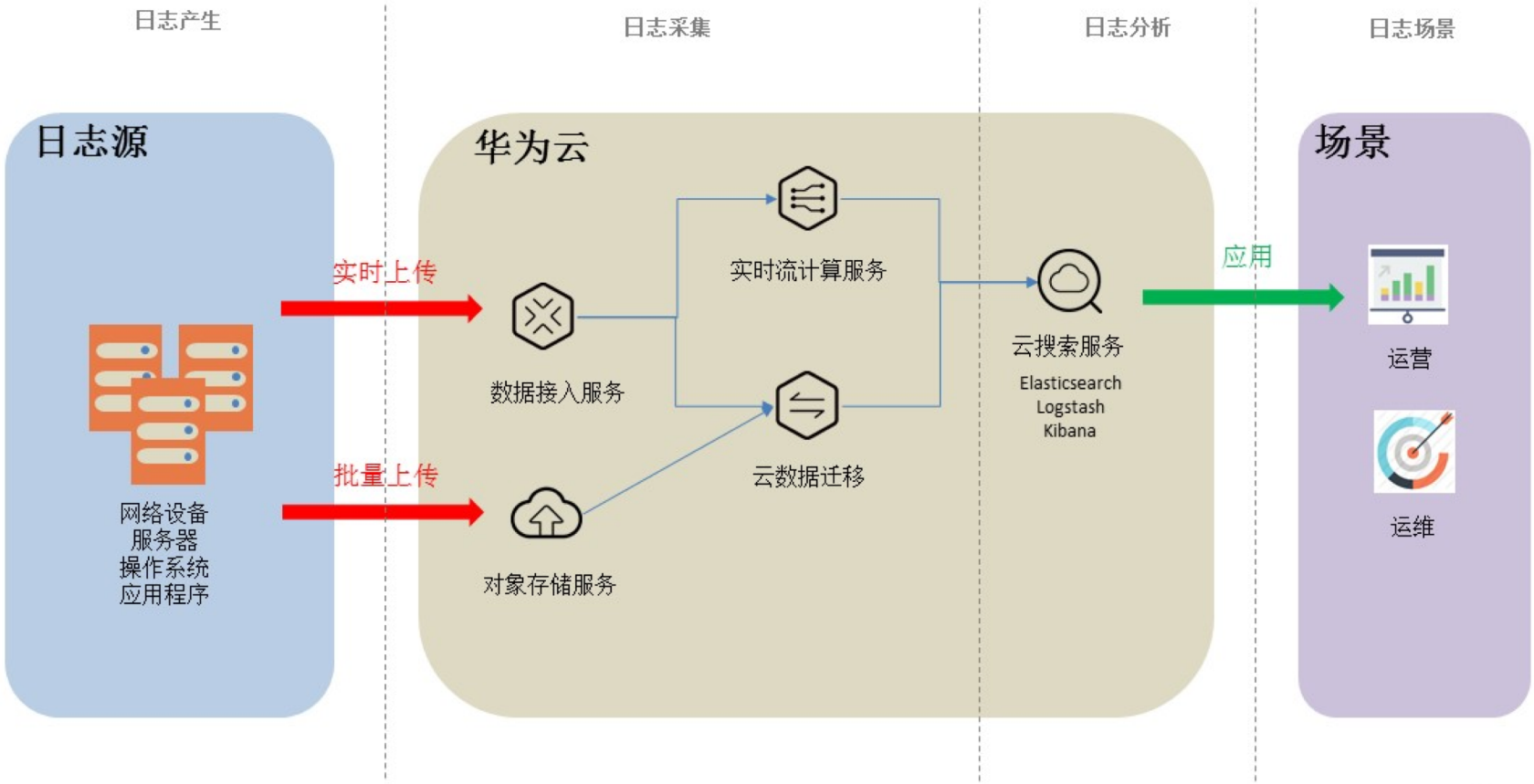


音频检索——效果

音乐库大小	ES检索
5万首音乐	100ms
20万首音乐	200ms
100万首音乐	230ms

经过验证，Elasticsearch的索引机制天然适合音频的检索模型，通过将音频转换为指纹组成的文档。能够非常好的处理音频检索问题

数据协同——云搜索服务日志解决案例



HUAWEI

内核优化——基于运维应用的优化案例

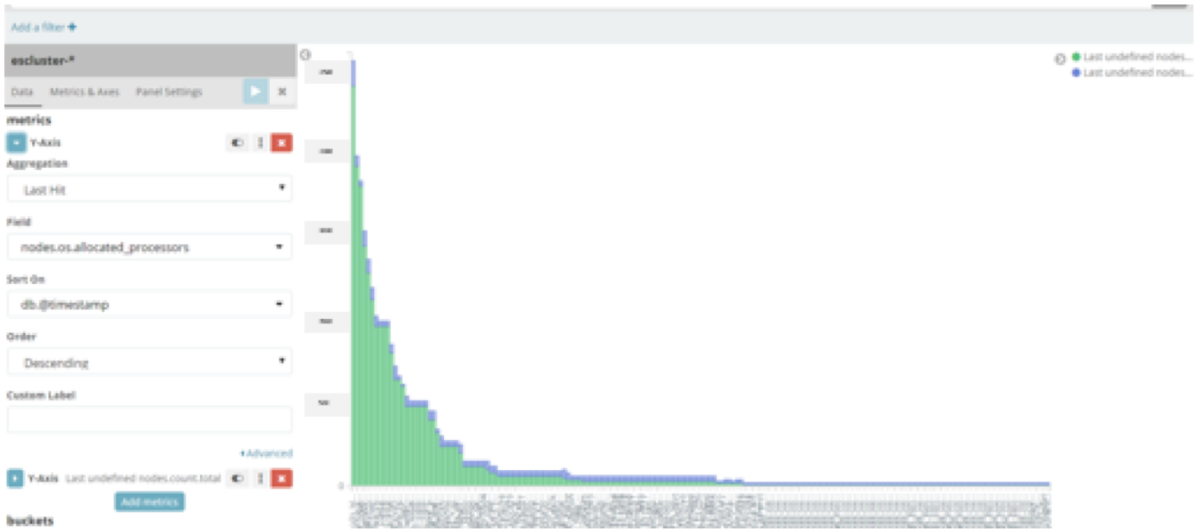
需求：时间戳、主机名、当前内存占用量，收集周期为每分钟收集一次，统计当前集群的占用内存总量

根据时间倒排，获取每个节点最新上报的数据 ← Top Hit

获取每个节点最新上报的数据 ← Sum Bucket

然而：Sum Bucket无法处理Top Hit返回的结果

Top Hit → 优化 → Last Hit



HUAWEI

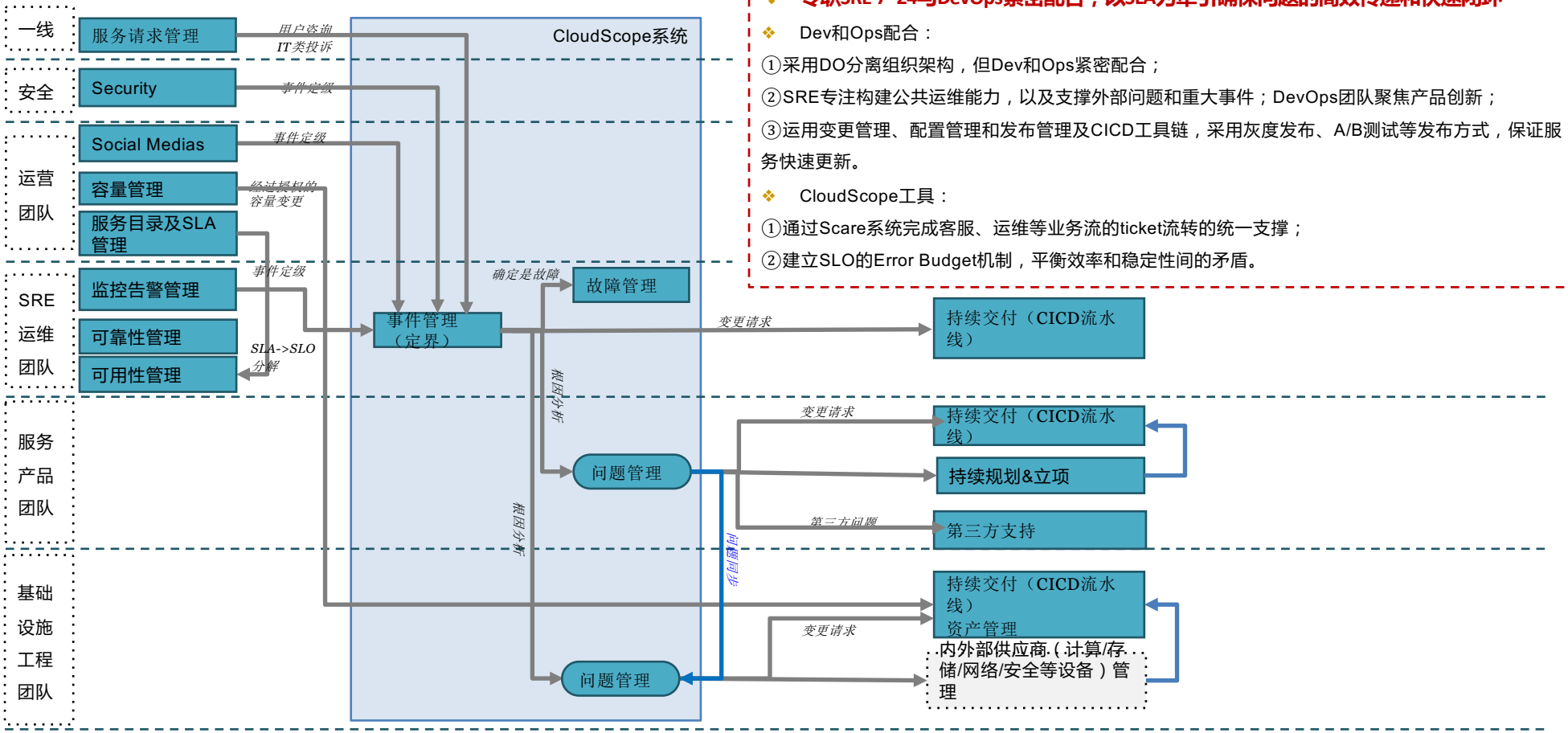
服务化的相关改进

痛点：
- 集群部署
- 多集群管理
- 节点扩容
- 磁盘扩容
- 监控告警
- 数据备份与恢复
- 集群状态诊断分析

云搜索服务

解决方案：
- 一键部署，15分钟内完成100节点的部署
- 提供openAPI管理多集群
- 一键扩容节点，不中断现网业务
- 一键扩容磁盘，不中断现网业务
- 基于华为云监控告警平台及云监控服务
- 基于OBS可配置的手动和定时备份任务
- 集群监控指标的统计分析

专业看护——华为云的运维看护流程



- **专职SRE 7*24与DevOps紧密配合，以SLA为牵引确保问题的高效传递和快速闭环**
- Dev和Ops配合：

①采用DO分离组织架构，但Dev和Ops紧密配合；

②SRE专注构建公共运维能力，以及支撑外部问题和重大事件；DevOps团队聚焦产品创新；

③运用变更管理、配置管理和发布管理及CICD工具链，采用灰度发布、A/B测试等发布方式，保证服务快速更新。

- CloudScope工具：

①通过Scare系统完成客服、运维等业务流的ticket流转的统一支撑；

②建立SLO的Error Budget机制，平衡效率和稳定性间的矛盾。

吃自己的狗粮

自建ES集群运维ES集群

- 收集服务管理面和业务面日志，日志可视化、统计分析、搜索定位
- 收集ES集群状态、监控统计指标，定制界面呈现并可转告警通知

日志监控可视化

业务指标监控可视化

什么是云搜索服务

云搜索服务

Cloud Search Service

=

Elasticsearch

+

Kibana

+

云化增强

云搜索服务是一个**基于Elasticsearch且完全托管**的在线**分布式搜索服务**，为用户提供结构化、非结构化文本的多条件检索、统计、报表。完全兼容开源Elasticsearch软件原生接口。

它可以帮助网站和APP**搭建搜索框**，提升用户寻找资料和视频的体验；还可以**搭建日志分析平台**，在运维上进行业务日志分析和监控，在运营上进行流量分析等等。

Thank You.