elastic 中国开发者大会 2023

十亿级人脸搜索的实践和优化

刘刚，CTO
谱时智能云，2023/04/08

分享嘉宾

刘刚 ， 谱时智能云 CTO

2014年Elasticsearch中国技术分享会分享嘉宾

- 我们的业务
- 我们的问题
- 我们的方案
- 我们的规划

我们的业务

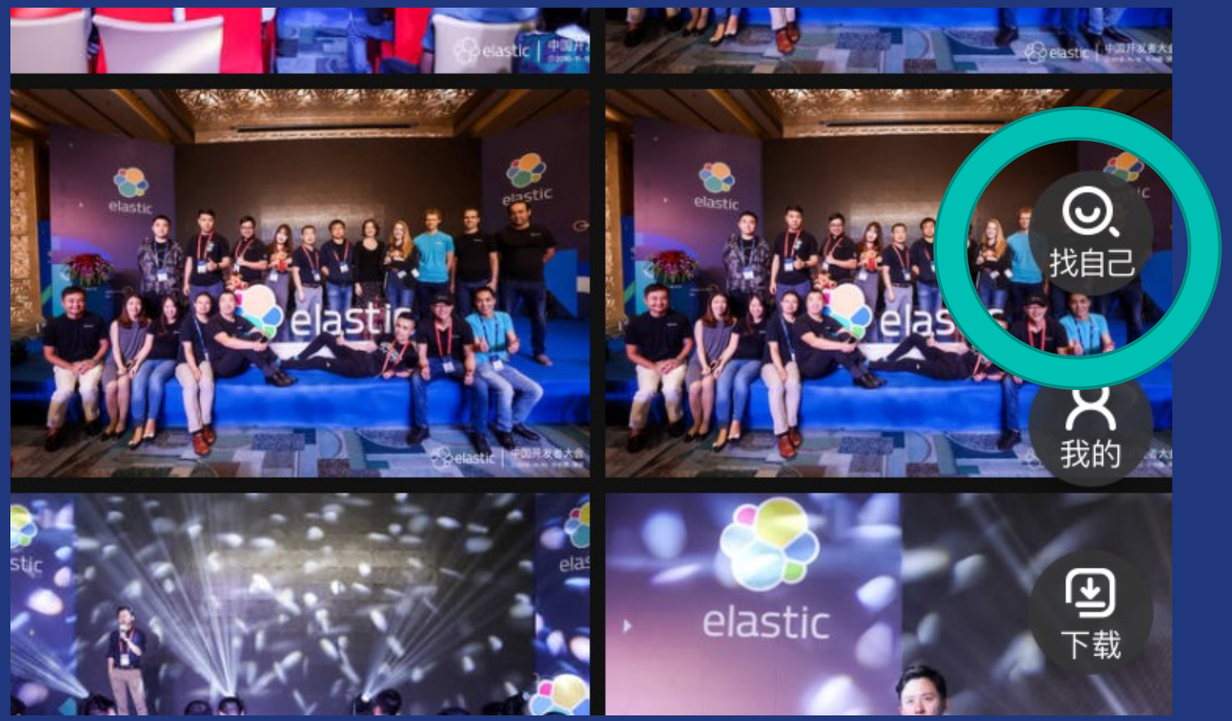
图片直播

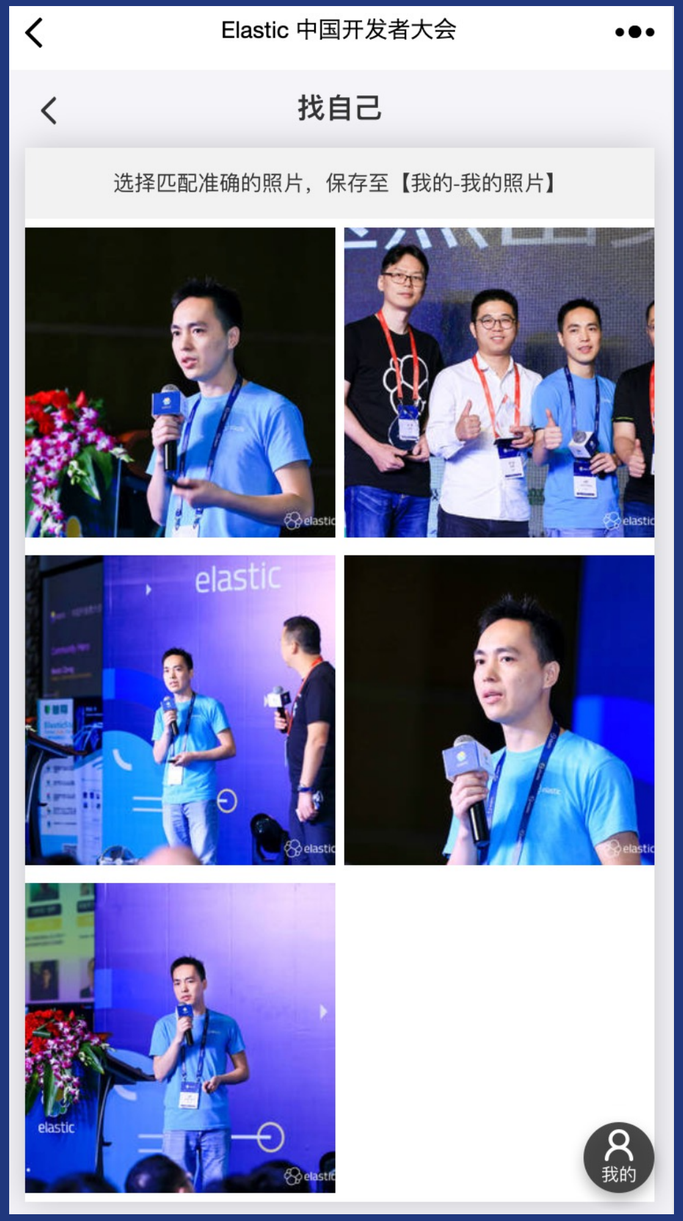
Elastic 中国开发者大会 2021
图片直播二维码


Elastic 中国开发者大会
图片
全部
热门
展示排序
16 / 1088
TOP
找自己
我的
下载


Elastic 中国开发者大会
图片
全部
热门
展示排序
16 / 1088
TOP

我们的业务






2021上海静安女子半程马拉松...
海静安女子半程马拉松暨女王
跑上海站
11 / 107047
TOP
请输入
搜索
人脸识别
图片
全部
展示排序
扫码关...
翻页
我的


19:48
玄奘之路
理想 · 行动 · 坚持 · 超越
商学院戈壁挑战赛
时光博物馆
走进玄奘之路，请发心
缘起 2005
更多 >
戈一 2006
更多 >
参赛队伍：6支 / 参赛人数：57人
2007 戈二
更多 >
参赛队伍：8支 / 参赛人数：106人

我们的问题

依赖第三方人脸搜索

遇到问题

性能限制

2QPS

5个搜索结果

功能限制

多活动聚合

活动人物相册

费用暴增

业务指数增长

不能额外收费

我们的问题

2QPS------->20QPS

依赖第三方人脸搜索

性能限制

5个搜索结果-------> 无解（没有预期解决时间）

我们的问题

依赖第三方人脸搜索

功能限制

多活动聚合

一场马拉松几十万张图

多场马拉松聚合

活动人物相册

我们的问题

业务指数增长

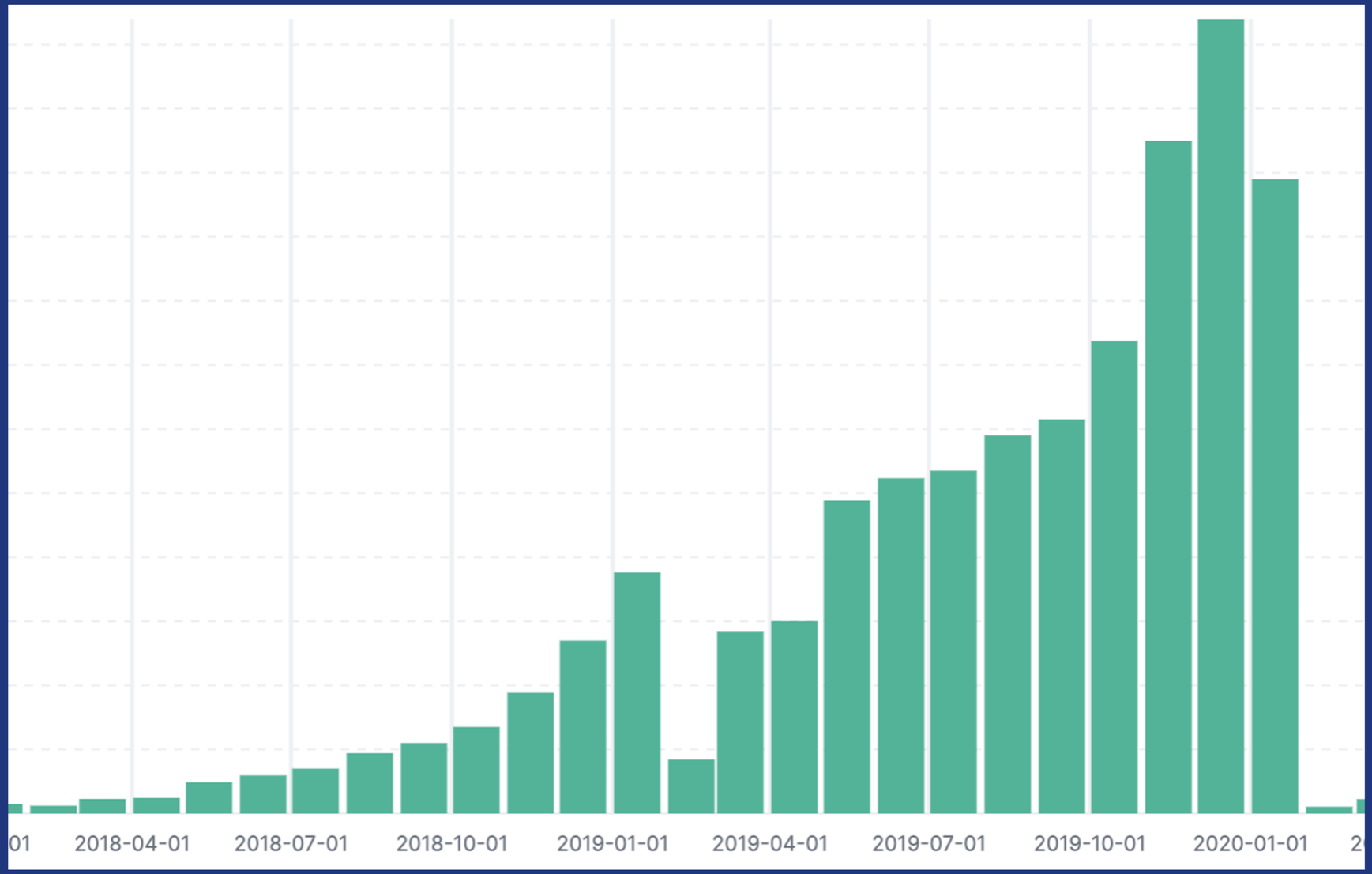
依赖第三方人脸搜索

费用暴增

不能额外收费

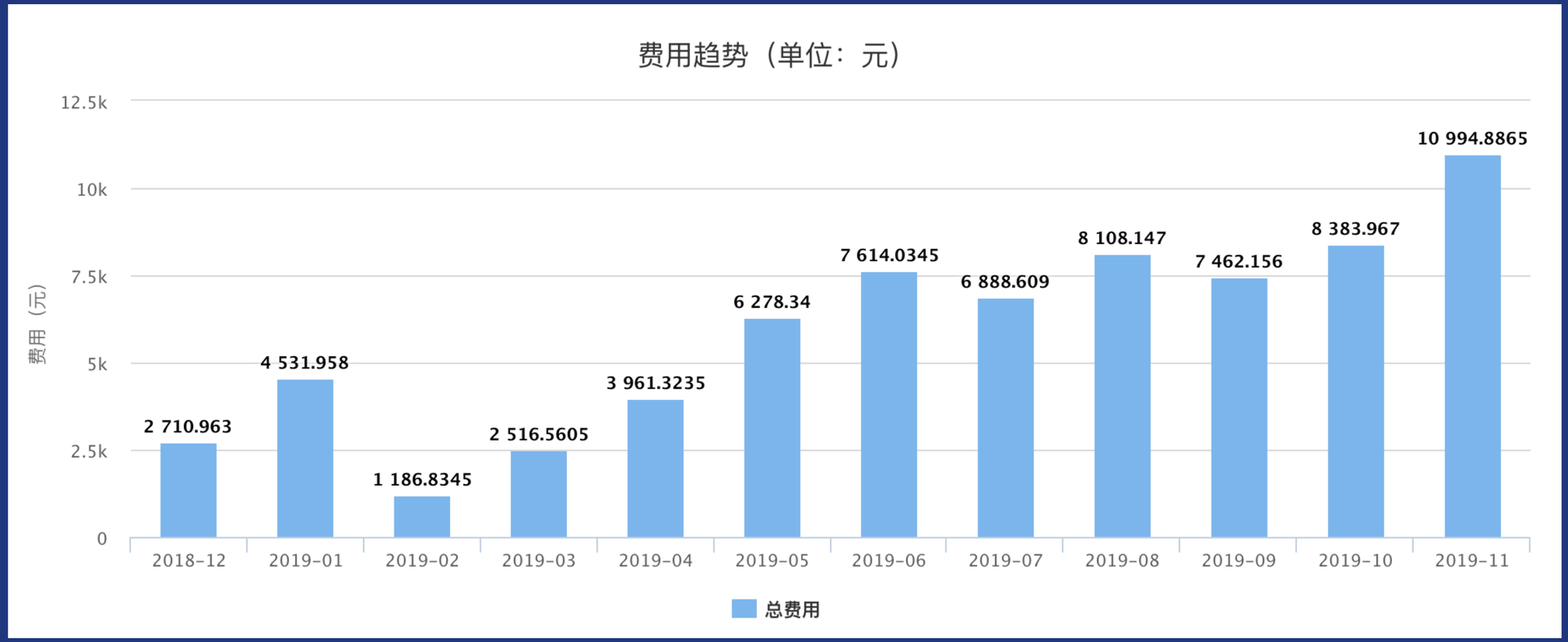
我们的问题

业务指数增长



我们的问题

费用暴增


费用趋势（单位：元）
费用（元）
12.5k
10k
7.5k
5k
2.5k
0
2 710.963
4 531.958
1 186.8345
2 516.5605
3 961.3235
6 278.34
7 614.0345
6 888.609
8 108.147
7 462.156
8 383.967
10 994.8865
2018-12
2019-01
2019-02
2019-03
2019-04
2019-05
2019-06
2019-07
2019-08
2019-09
2019-10
2019-11
总费用

我们的问题

怎么解决？？？

小公司

没钱

没人

没时间

“万能的”

elastic

Search. Observe. Protect.

我们的方案

解决问题

第一阶段：

elastic方案探索

理论基础

搜索是计算多维词向量距离

人脸搜索是计算512位向量欧几里得距离

数据类型支持

Elasticsearch7.0之后新数据类型 Dense_vector

支持最多500维数据

算法支持？

已经支持余弦相似度计算

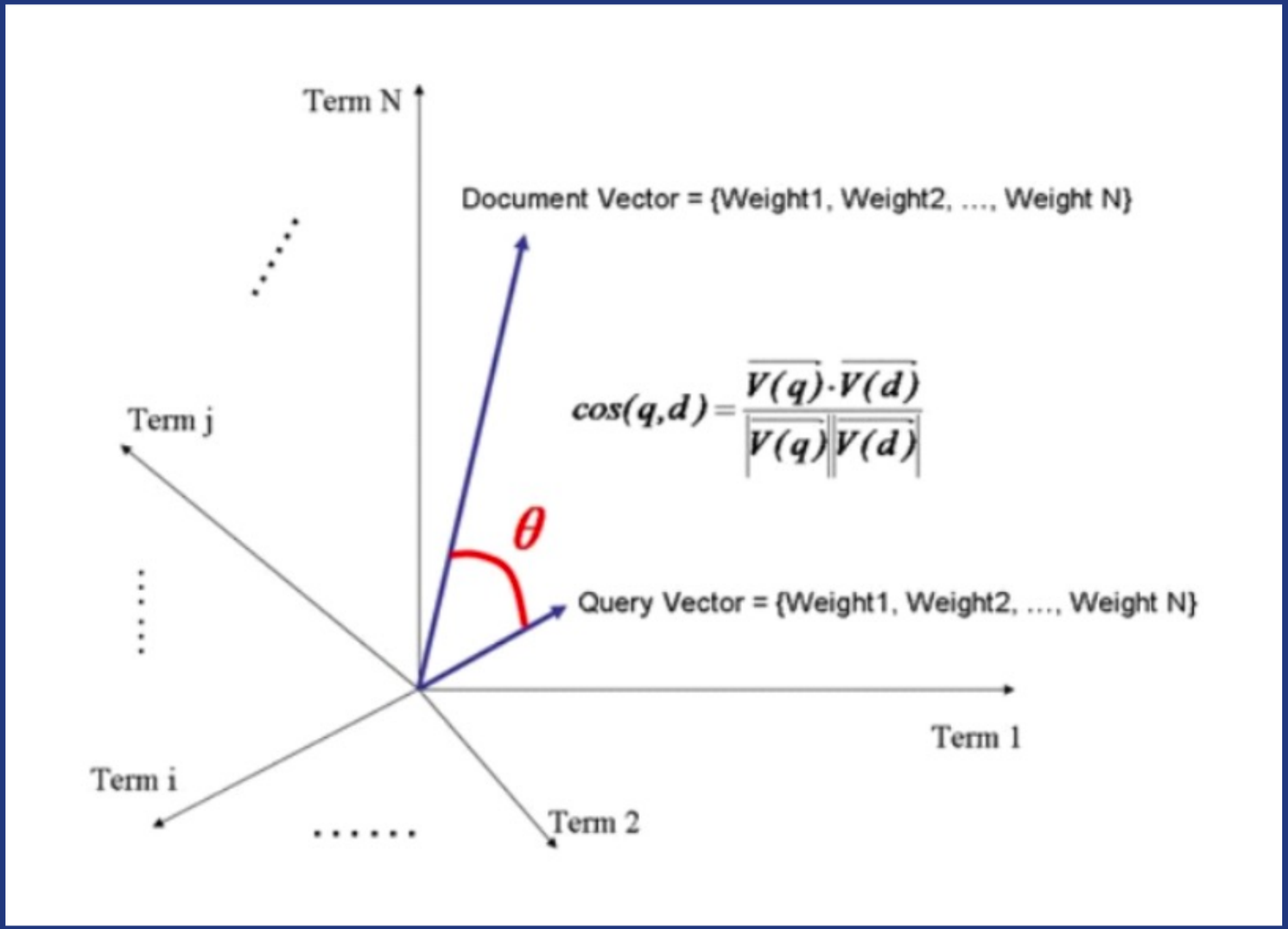
大佬承诺后续版本支持欧几里得距离

我们的方案

理论基础

TF/IDF算法

空间向量算法/
余弦相似度

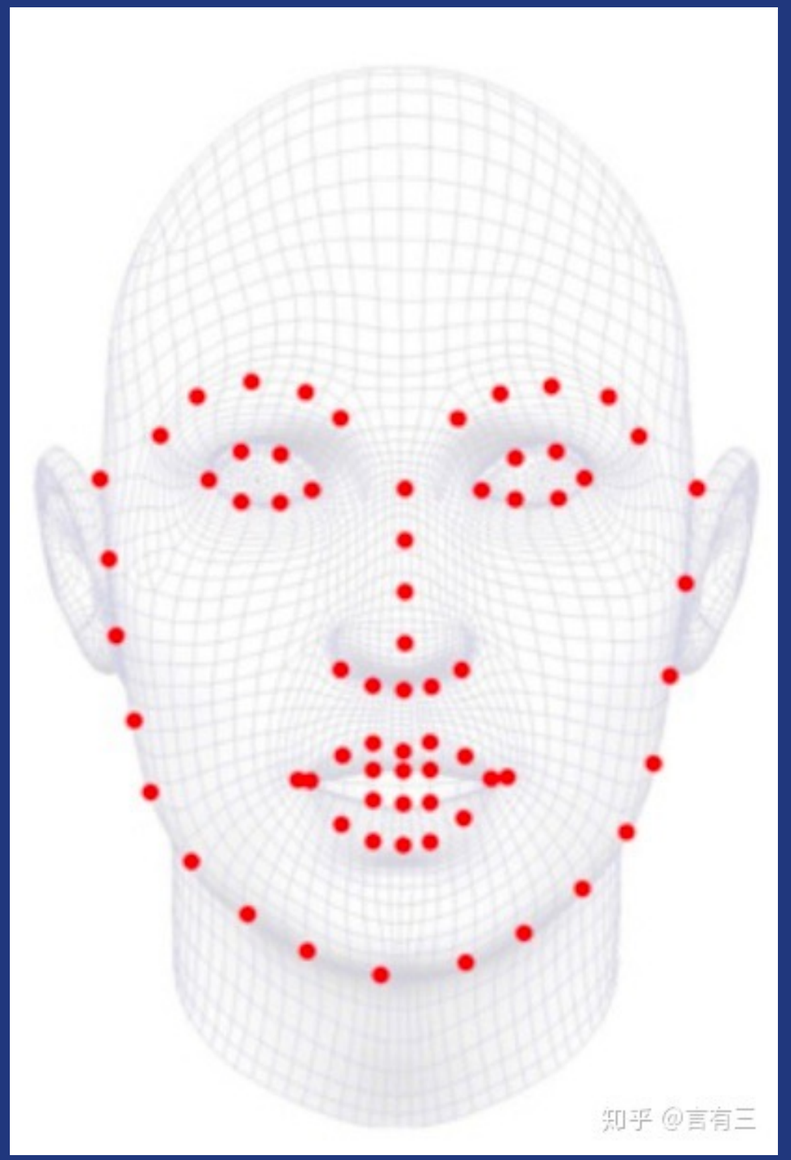


我们的方案

理论基础

欧式空间特征512维

欧几里得距离

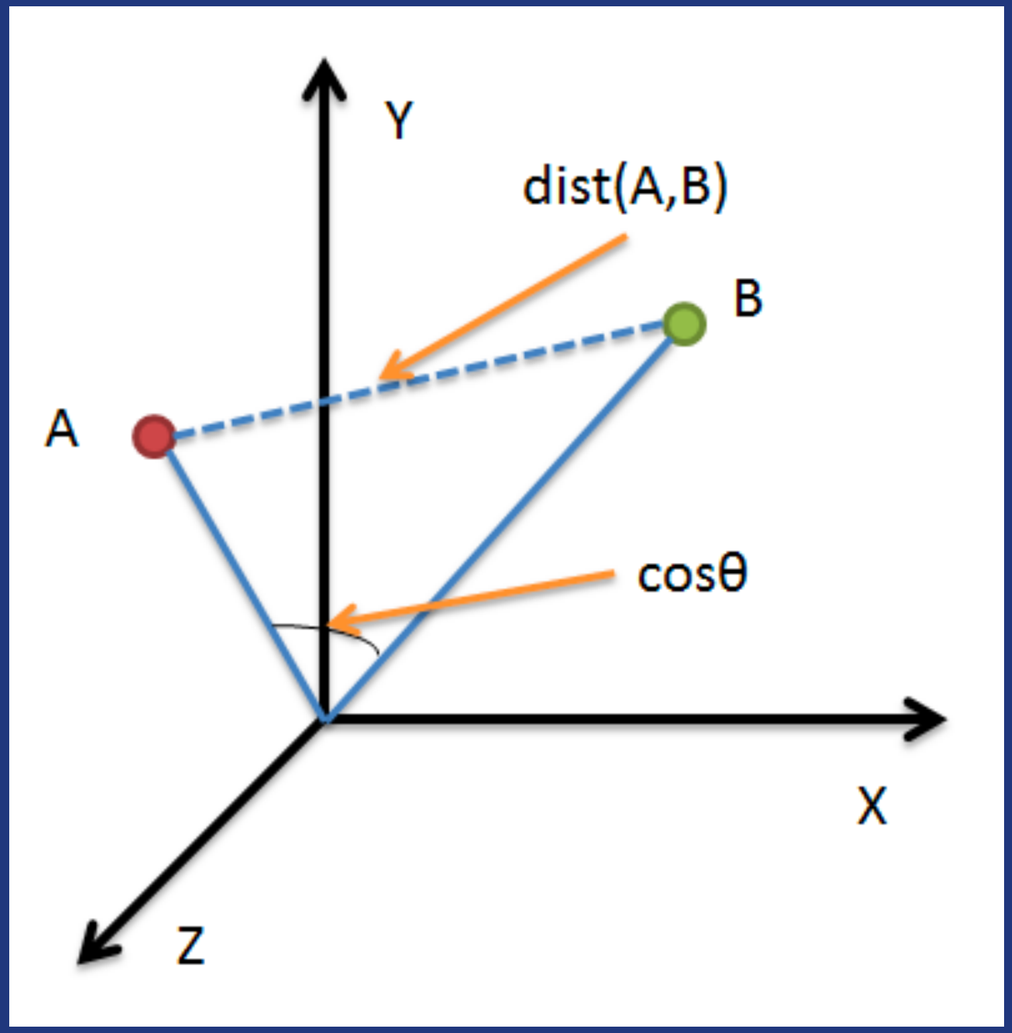


我们的方案

理论基础

欧式空间特征512维

欧几里得距离



$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + \cdots + (y_{n_1} - y_{n_2})^2}$$

大佬救命

How to search dense_vector

Elastic Stack · Elasticsearch

LiuGangR 3 Jan 2019

which query is used to search dense_vector. There is no query type to work.

dense_vector is the new feature of es 7.0
https://www.elastic.co/guide/en/elasticsearch/reference/master/dense-vector.html 181

1

created	last reply	replies	views	user	like	links
Jan 2019	Feb 2019	2	1.8k	1	1	4

LiuGangR Jan 2019

there is a way.

github.com/elastic/elasticsearch

Issue: Introduce vector field, vector query and rescoring based on them 71

opened by mayya-sharipova on 2018-06-27

```
Introduce a new field of type vector on which vector calculations can
be done during rescoring phase
PUT my_index
{
 "mappings": {
 "_doc":...
```

:Search/Ranking

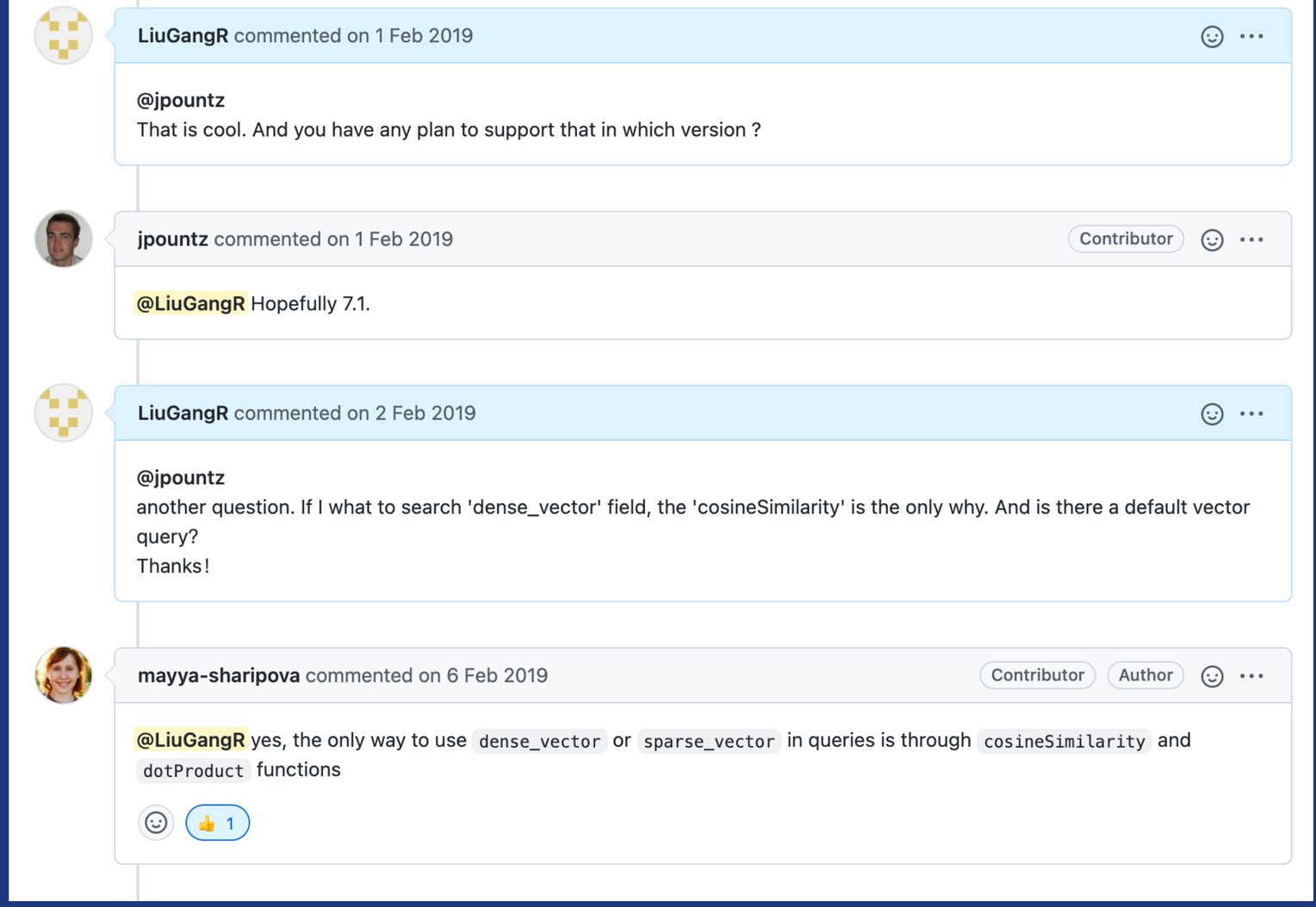
Jan 2019

1 / 3
Jan 2019

Feb 2019

大佬救命


Mayya Sharipova
mayya-sharipova
Follow
61 followers · 0 following
@elastic
Canada
mayya@apache.org


LiuGangR commented on 1 Feb 2019
@jpountz
That is cool. And you have any plan to support that in which version ?
jpountz commented on 1 Feb 2019
Contributor
@LiuGangR Hopefully 7.1.
LiuGangR commented on 2 Feb 2019
@jpountz
another question. If I what to search 'dense_vector' field, the 'cosineSimilarity' is the only why. And is there a default vector query?
Thanks!
mayya-sharipova commented on 6 Feb 2019
Contributor
Author
@LiuGangR yes, the only way to use dense_vector or sparse_vector in queries is through cosineSimilarity and dotProduct functions
1

我们的方案

第一阶段结论：

es作为人脸搜索引擎

第一阶段验证：

es7.x（未发布）+ script

```
"script": {
 "lang": "painless",
 "source": """
  double total = 0;
  for (int i = 0; i < doc['face_detail'].length; ++i) {
   total +=(params['_source']['face_detail'][i] - params.queryVector[l])
            *(params['_source']['face_detail'][i] - params.queryVector[i]);
   if (total>params.confidence){
    return total;
   }
  }
  return total;
 """,
 "params": {
  "confidence”:0.16,
  "queryVector": [
   0.070275992155075 07,
   0.09286174178123474,
   0.036543644964694 98,
```

我们的方案

第二阶段：

elastic方案上线

7.2版本之前

没有正式版，是等还是用开发板?

人脸向量512位，dense_vector最多支持500位

不支持欧几里得距离计算

7.2版本

人脸向量512位，dense_vector支持1024位

不支持欧几里得距离计算

集群管理引入x-pack

7.4版本之后

支持欧几里得距离计算

搜索优化

异步搜索

我们的方案

自研人脸向量512维

es7.2版本之前

20kb/人脸

dense_vector只支持500位

非es方案一

数据存数据库

计算用java

非es方案二

数据存redis

计算用python

我们的方案

es7.2版本

测试+集群整体升级

7.2版本dense_vector支持1024位

功能测试

性能测试

建表

7.1版本x-pack免费

更安全的集群管理

身份管理

我们的方案

es7.2版本

测试

功能测试

1.数据类型，欧几里得距离计算脚本

2.几十万样本计算

性能测试

1.预估正式数据量和压力

2.不同硬件，索引结构

3.确认集群硬件和索引结构

信心

1.能不能替代行业独角兽大厂的服务？

2.能不能一直稳定?

3.能不能省钱 ？

场景

1场活动　　500-50万张图　　500-100万张人脸

硬件

cpu　　内存　　硬盘

索引结构和分布

几台物理机？几个索引？几个分片？几个备份？

我们的方案

场景

1场活动　　500-50万张图　　500-100万张人脸

3*8c32G3T高性能云盘　　性价比

10个index(face_01)　　分表--扩展性

7个shards　　3,5,7,11,13测试后确认

2个replicas　　写多读少

我们的方案

es7.2版本

信心哪里来？

是骡子是马，遛出来

生产数据在双方同时生成

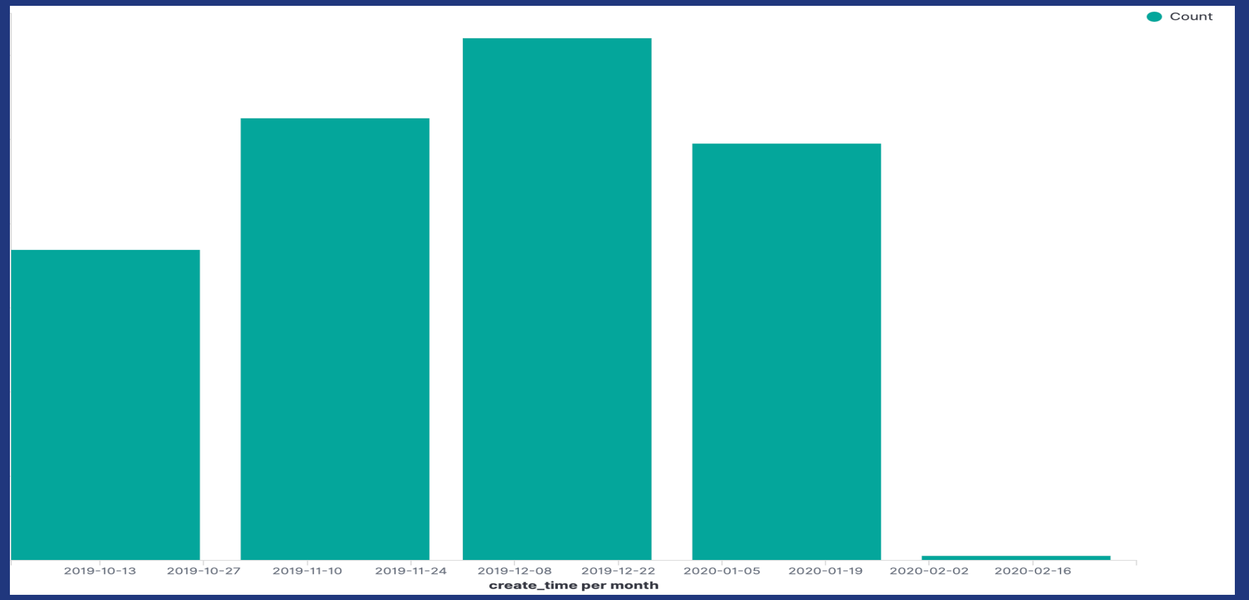
同时搜索互为补充

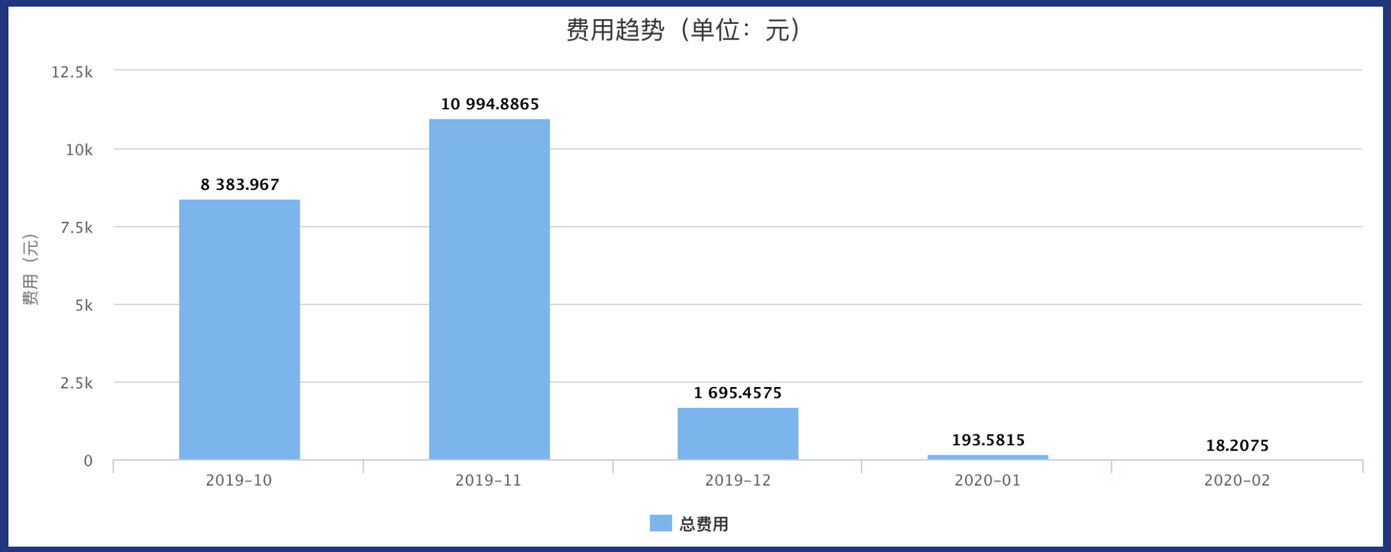
对比结果的数量，准确率，失败率，召回率

我们的方案

es7.2版本

正式上线





我们的方案

es7.2版本

测试+集群整体升级

7.2版本dense_vector支持1024位

功能测试

性能测试

建表

7.1版本x-pack免费

更安全的集群管理

身份管理

我们的方案

es7.4版本之后

优化

支持欧几里得距离计算

搜索优化

异步搜索

我们的方案

es7.4版本之后

优化

```
"sort": [
 {
  "_score": {
   "order": "desc"
  }
 }
],
"size": 100,
"from": 0,
"query": {
 "script_score": {
    "min_score":10-confidence,
  "script": {
   "source": "10-l2norm(params.face, 'face_detail')",
   "params": {
    "face": [-0.0016885467,-0.08441351,-0.020592572,
```

我们的方案

第二阶段

回顾

性能限制

2QPS

5个搜索结果

功能限制

多活动聚合

活动人物相册

费用暴增

业务指数增长

不能额外收费

我们的方案

第三阶段

集群优化+省钱

疫情影响

行业是重灾区

降本增效

业务特点

相册有时效性，可以分层

统计，监控，业务都在一个集群

硬件降价

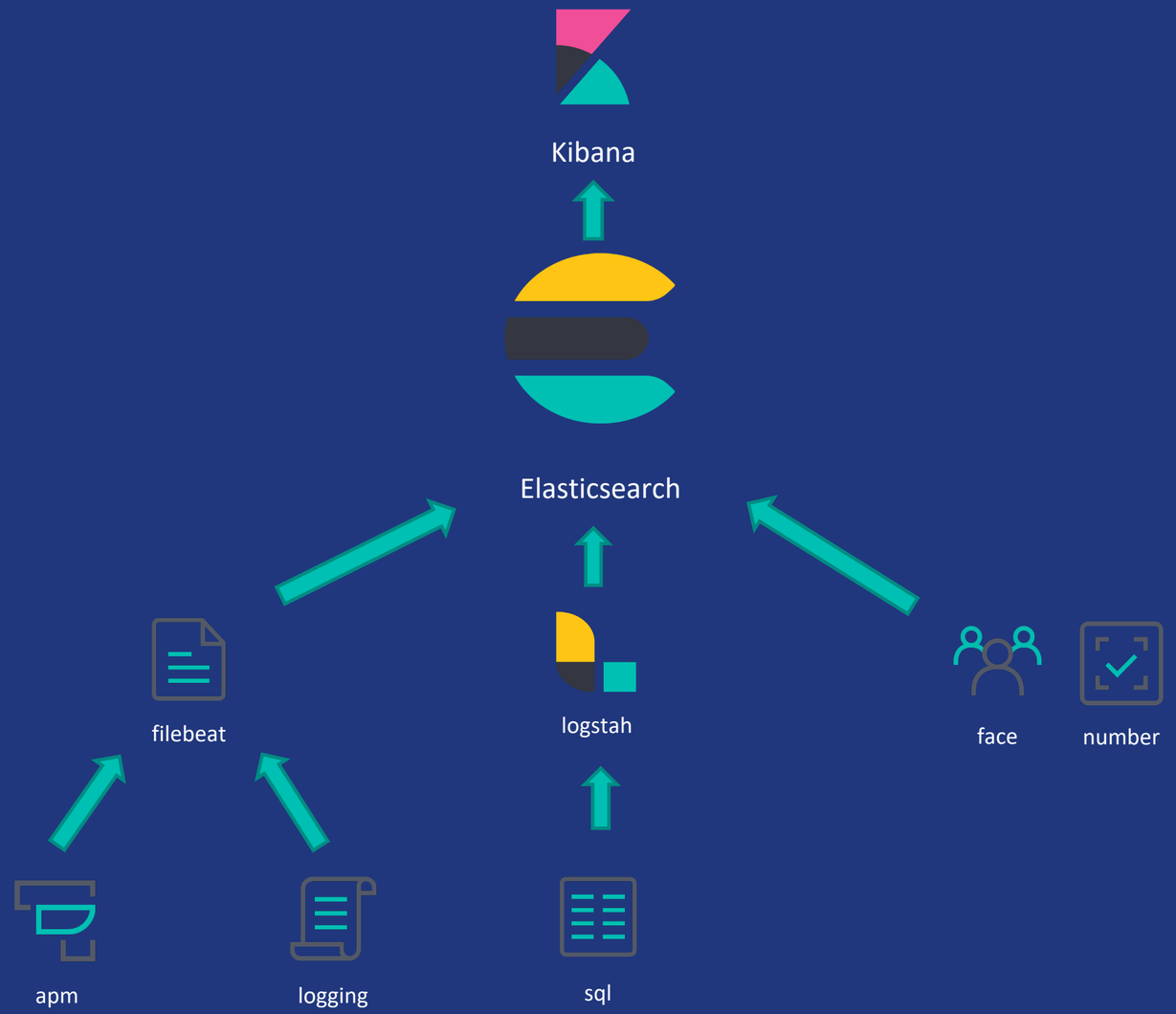
大数据类型服务器出现

HHD和SSD都可选

我们的方案

第三阶段

业务特点



我们的方案

第三阶段

方案

生命周期方案

硬件分层方案

索引汇总作业（rollup jobs）

我们的方案

第三阶段

方案

硬件分层+索引汇总

硬件分层

服务器分为两部分-------HHD和SSD

热数据放SSD，冷数据放HHD

定期手动配置

索引汇总（rollup jobs）

捅进数据按照业务需求定期汇总到新索引

调整业务模式

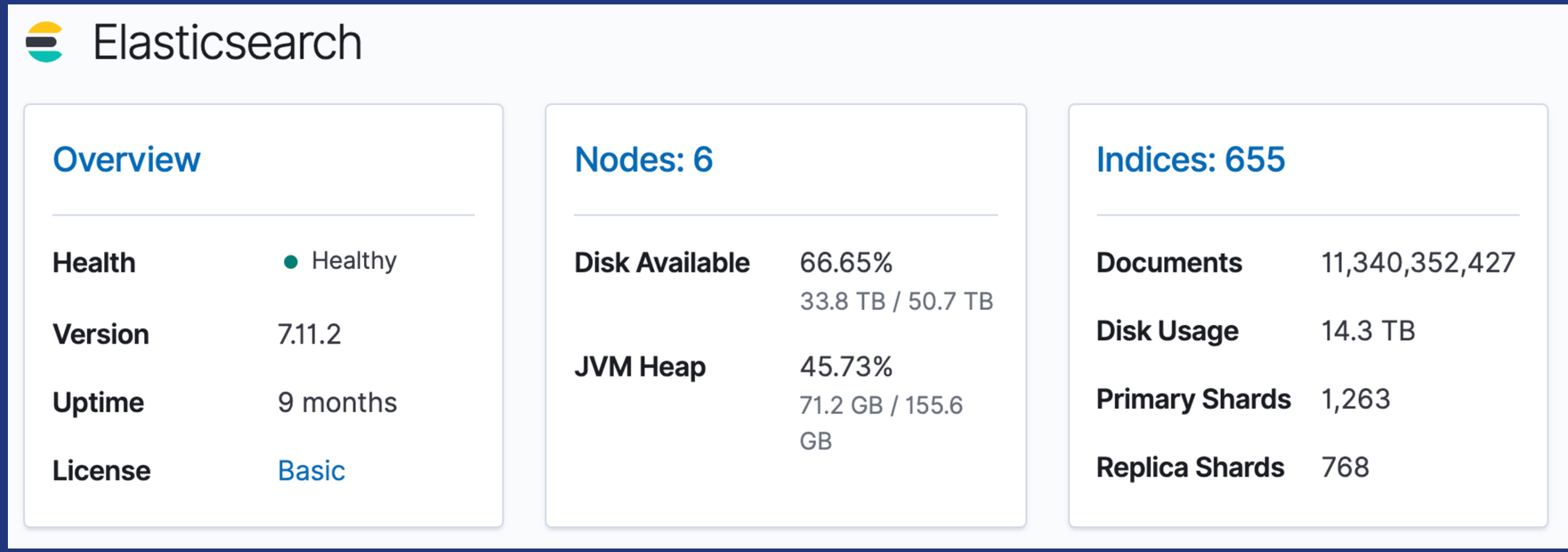
我们的方案

第三阶段

2*8c32G22T(HHD)

4*8c64G2T(SSD)

回顾



后续计划

我们的规划

后续计划

百亿级人脸搜索规划，引入es8.x特性

数据迁移到其他云服务商----省钱

视频上人脸搜搜的方案探索

特别感谢

Medcl 黎霖 林瑞凯 徐奥平 王金雷

十亿级人脸搜索的实践和优化

刘刚 CTO @谱时智能云

2023年4月8日

elastic 中国开发者大会 2023

感谢观看

elastic
中文社区

专业、垂直、纯粹的 Elastic 开源技术交流社区

https://elasticsearch.cn/