



# PB级ElasticSearch集群架构设计与平台化演进

---

周祝群

2019.12.07, 同程艺龙-研发中心



# 生产集群

50+

---

集群

---

500+

---

节点

---

4万亿

---

文档数

---

# 日志集群

300+

---

节点

---

3万亿

---

文档

---

1500亿

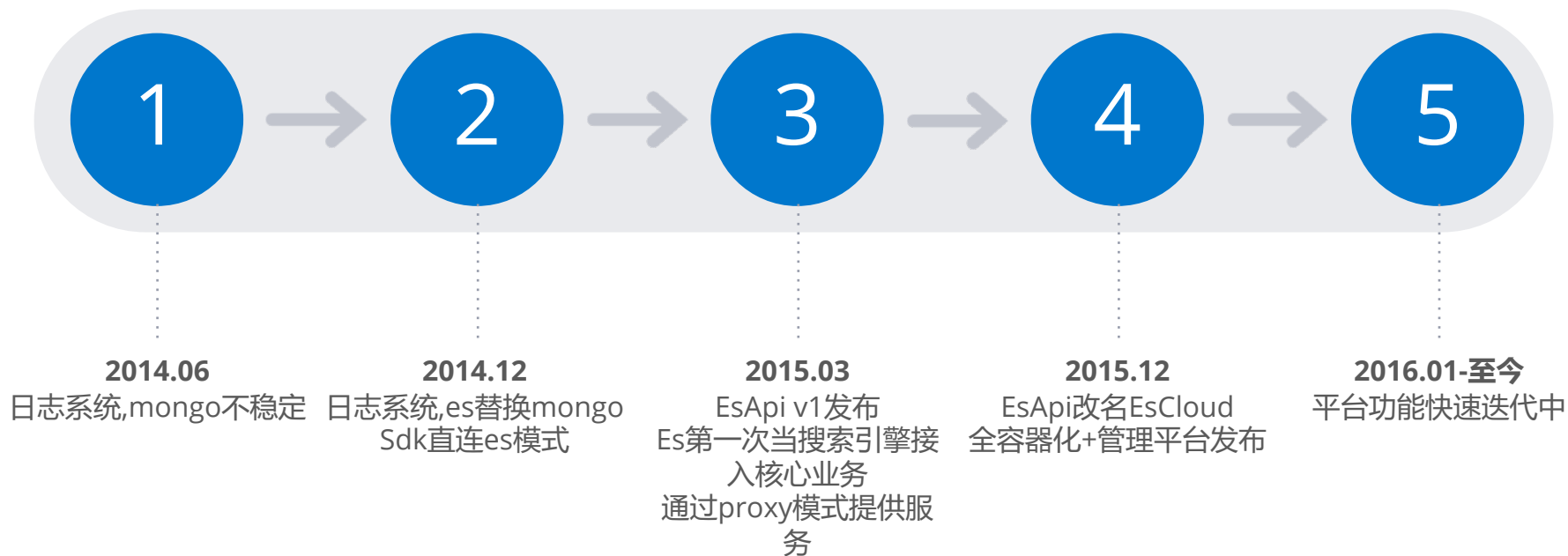
---

日增条数

---

# Part1:背景介绍

# 背景介绍



# Part2:平台与架构演进

# 架构演进

v1

1个人

2天



# 架构演进

## V1-问题

集群**难**稳定,OOM...

版本**难**升级

性能**难**保证

**复杂**



# 架构演进

## 小目标

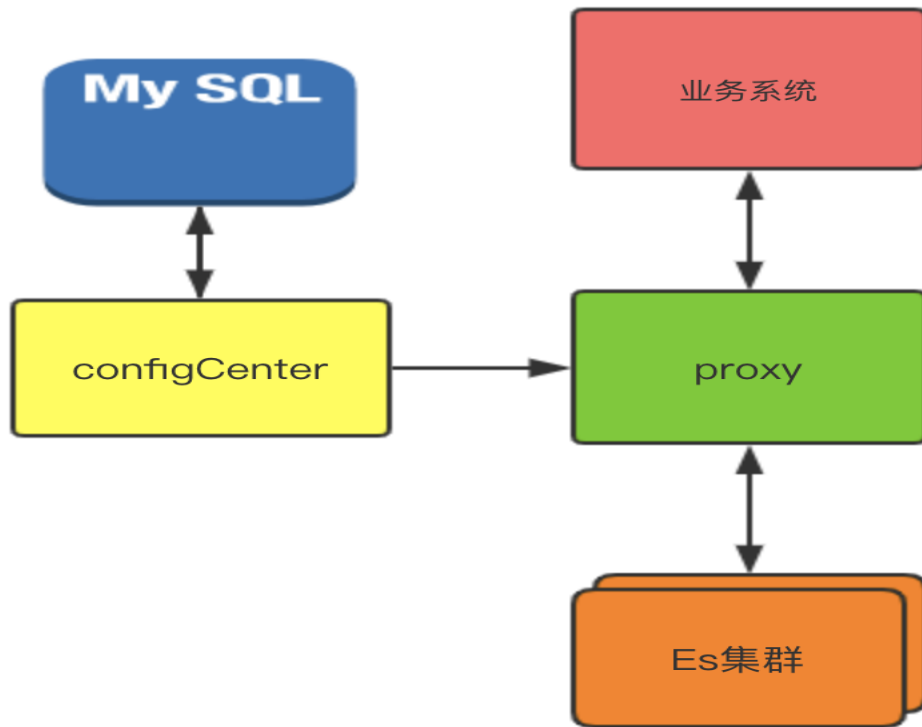
# 让不熟悉es的用户更优地使用es

简单

高性能,稳定

# 架构演进

简单-集群难稳定



# 架构演进

简单-性能难保证

写入

```
[{
  "field1": "1",
  "field2": "value2"
},
{
  "field1": "1",
  "field2": "value2"
}]
```

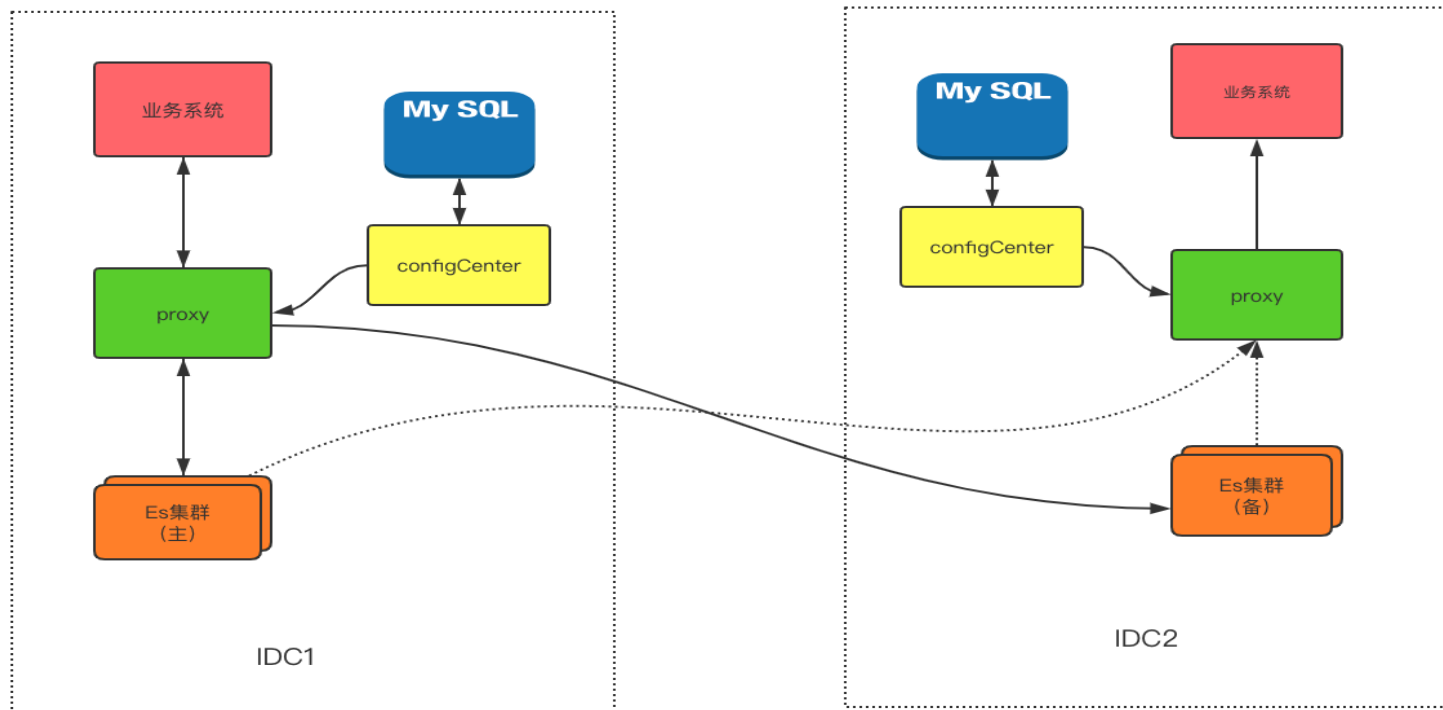
查询

```
{
  "field1": "gte:1",
  "field2": "eq:value2"
}
```

会发HTTP即可

# 架构演进

## 稳定-高可用



# 架构演进

版本难升级



升级，那是不可能的！

# 架构演进

一大波需求。。。。

# 架构演进

## V2-问题-查询

语法难兼容

对业务黑盒

问题查询不能限制

# 架构演进

## V3-查询

All in 查询模板

预定义

# 架构演进

## V3-查询

首页 / 索引 / 查询模板

帮助文档



新建模板 辅助工具

| 名称                                | 版本数 | 最新版本                   | 最后修改人 | 最后修改时间              | 描述                                                                          | 操作                                                                            |
|-----------------------------------|-----|------------------------|-------|---------------------|-----------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| > test                            | 2   | <a href="#">废弃</a>     | 3     | 2018-12-28 15:12:34 | test                                                                        | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > cmdb                            | 11  | <a href="#">1.0.10</a> |       | 2018-12-28 15:07:37 | cmdb查询                                                                      | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > ops-tblog-query-timestamp       | 1   | <a href="#">1.0.0</a>  |       | 2018-12-10 15:57:38 | ops-tblog-query-timestamp                                                   | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > duty                            | 4   | <a href="#">1.0.3</a>  |       | 2019-01-07 15:49:20 | 值班运维查询                                                                      | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > duty-uri                        | 6   | <a href="#">1.0.5</a>  |       | 2019-04-03 18:49:37 | 二级目录                                                                        | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > ops-tblog-check-by-domain       | 1   | <a href="#">1.0.0</a>  |       | 2019-02-13 11:27:34 | 根据域名查询单位时间内的总量，按照状态码聚合倒序                                                    | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > geturibylighthost               | 5   | <a href="#">1.0.4</a>  |       | 2019-02-26 14:35:47 | 根据机票host查询uri                                                               | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > ops-tblog-pv-by-idc             | 1   | <a href="#">1.0.0</a>  |       | 2019-02-25 14:51:56 | 获取各机房PV流量+接入点                                                               | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > testgetrequesturl               | 2   | <a href="#">1.0.1</a>  |       | 2019-02-27 16:18:15 | 聚合查询                                                                        | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > interfacetrafficinfo            | 2   | <a href="#">1.0.1</a>  |       | 2019-03-07 16:14:53 | 应用接口访问基本信息查询                                                                | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > duty-uri-ip                     | 1   | <a href="#">1.0.0</a>  |       | 2019-05-10 14:20:13 |                                                                             | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > ops-sz-host-uri-Nxx             | 3   | <a href="#">1.0.2</a>  |       | 2019-08-14 16:22:10 | 根据 host、uri、时间区间、Http Status 区间 统计文档数                                       | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > ops-sz-host-uri-Nxx-seg         | 4   | <a href="#">1.0.3</a>  |       | 2019-08-14 16:24:07 | 根据 host、uri、时间区间，获取不同 http status 区间的统计文档数                                  | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > ops-sz-host-uri-Nxx-seg-exclude | 4   | <a href="#">1.0.3</a>  |       | 2019-08-14 16:24:22 | 根据 host、uri、时间区间，获取不同 http status 区间的统计文档数<br>可以使用 excludeUri 排除统计某些 uri 数据 | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > ops-sz-host-Nxx                 | 3   | <a href="#">1.0.7</a>  |       | 2019-08-14 16:23:50 | 根据 host、时间区间、Http Status 区间 统计文档数                                           | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |
| > ops-sz-host-seg                 | 1   | <a href="#">1.0.0</a>  |       | 2019-08-02 15:14:39 |                                                                             | <input checked="" type="checkbox"/> <a href="#">更新模板</a> <a href="#">语法分析</a> |

# 架构演进

## V3-查询



duty

4

1.0.3



2019-01-07 15:49:20

值班运维查询



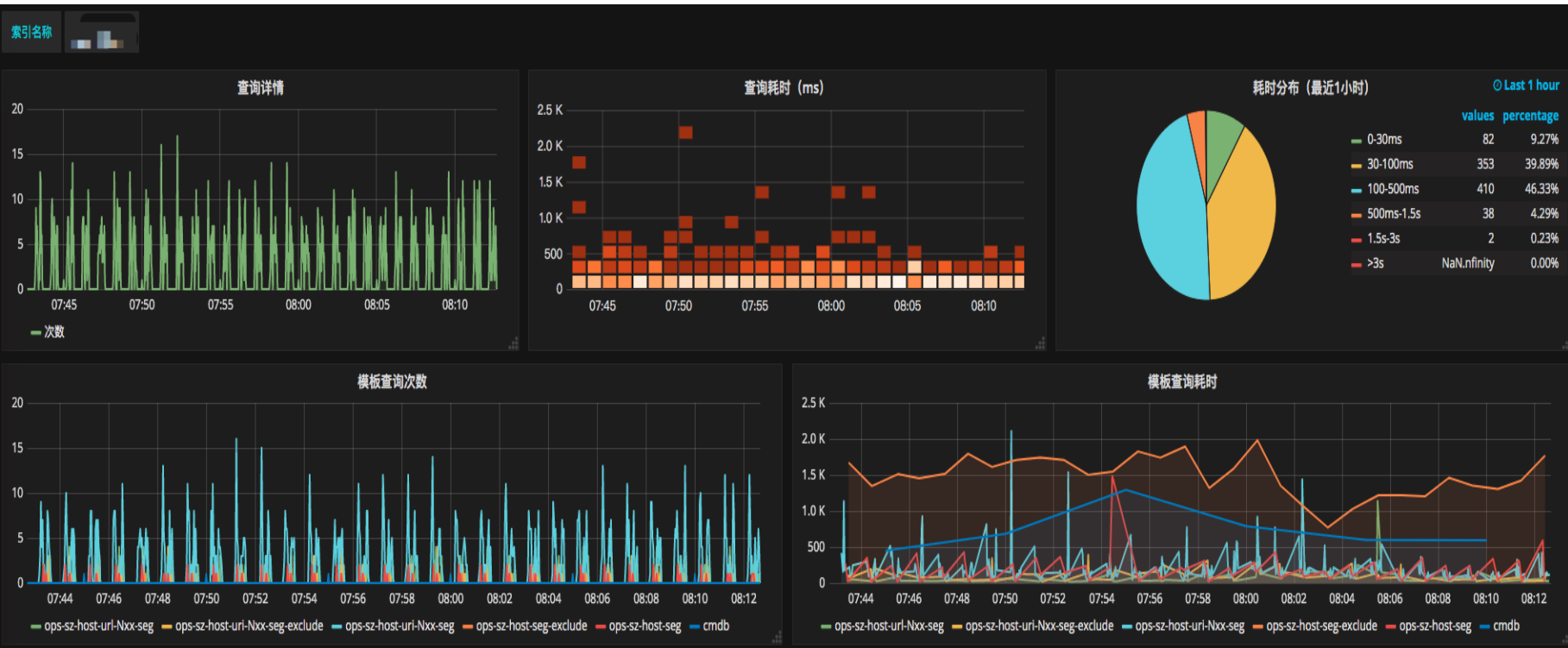
更新模板

语法分析

| 版本    | 复制模板 | 模板内容                                                                                                                                                                                                                                                                                   | 创建人 | 创建时间                |
|-------|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|---------------------|
| 1.0.3 |      | <pre>{ "from": 0, "size": 0, "query": { "bool": { "filter": [{ "range": { "@timestamp": { "gte": "{{start}}" } } }, { "range": { "@timestamp": { "lt": "{{end}}" } } } ], "aggregations": { "group_by_host.keyword": { "terms": { "field": "host.keyword", "size": 100 } } } }</pre>   |     | 2019-01-07 15:49:20 |
| 1.0.2 |      | <pre>{ "from": 0, "size": 100, "query": { "bool": { "filter": [{ "range": { "@timestamp": { "gte": "{{start}}" } } }, { "range": { "@timestamp": { "lt": "{{end}}" } } } ], "aggregations": { "group_by_host.keyword": { "terms": { "field": "host.keyword", "size": 100 } } } }</pre> |     | 2019-01-04 09:42:40 |
| 1.0.1 |      | <pre>{ "from": 0, "size": 100, "query": { "bool": { "filter": [{ "range": { "@timestamp": { "gte": "{{start}}" } } }, { "range": { "@timestamp": { "lt": "{{end}}" } } } ], "aggregations": { "group_by_host.keyword": { "terms": { "field": "host.keyword", "size": 10 } } } }</pre>  |     | 2019-01-02 15:44:32 |
| 1.0.0 |      | <pre>{</pre>                                                                                                                                                                                                                                                                           |     | 2019-01-02 15:26:35 |

# 架构演进

## V3-查询

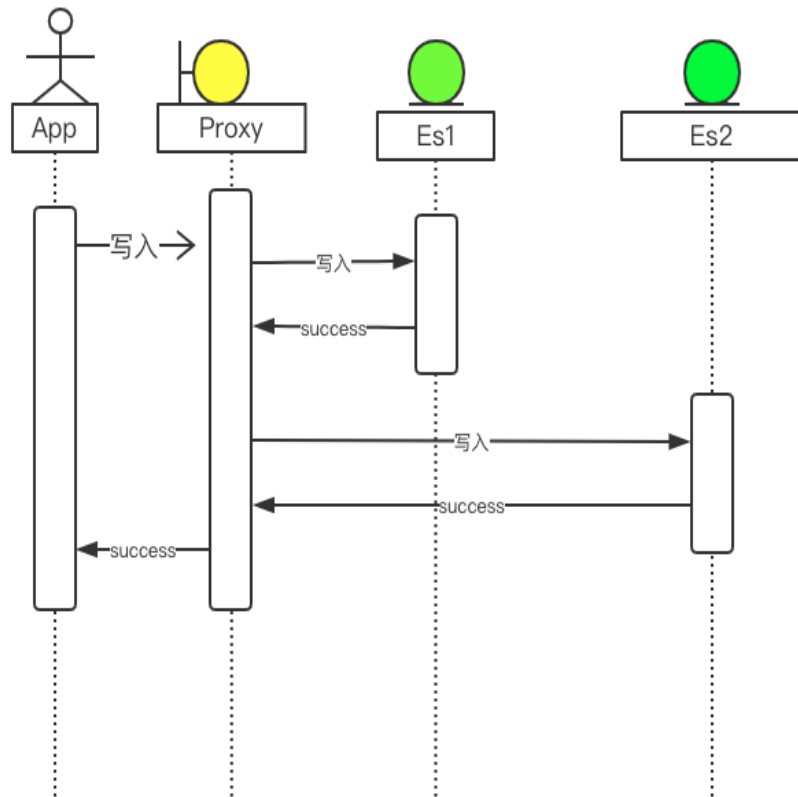


# 架构演进

稳定-高可用

写入性能下降

数据不一致



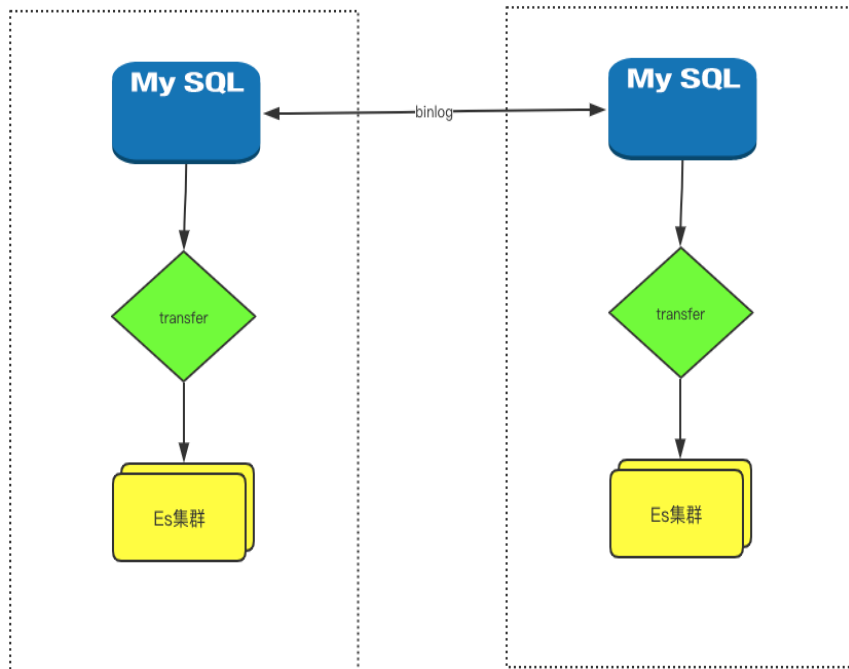
# 架构演进

## v3-可靠性

Mysql -> 消息队列 ( kafka or mq )

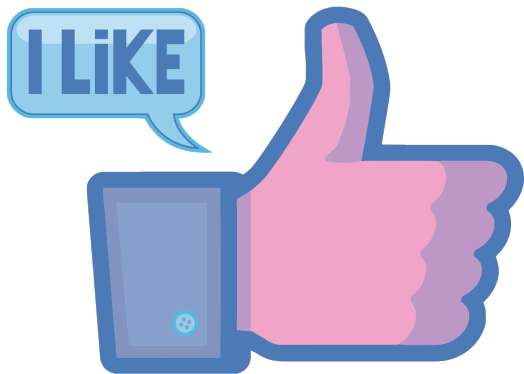
接入成本太高

推进不了



# 架构演进

v3-可靠性



v6.5 Released



**CCR**

# 架构演进

## v3-可靠性



# 架构演进

v3-可靠性

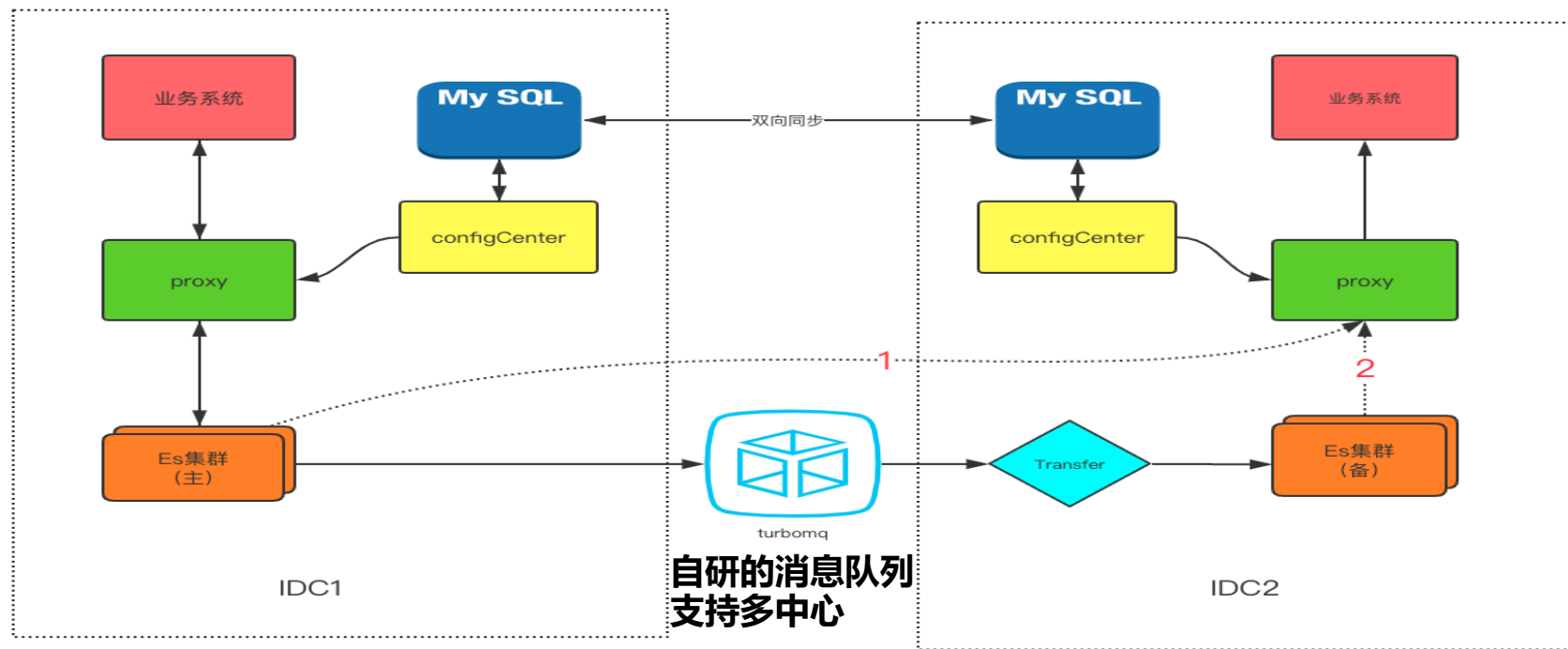
自



研

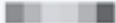
# 架构演进

## v3-可靠性



# 架构演进

## v3-可靠性

| 条数                                                                                |                                                                                   |
|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
|  |  |
| 1,242,960,676                                                                     | 1,242,960,676                                                                     |
| 150,667,084                                                                       | 150,667,084                                                                       |
| 459,640,273                                                                       | 459,640,273                                                                       |
| 11,172,981                                                                        | 11,172,981                                                                        |



## 数据幂等处理

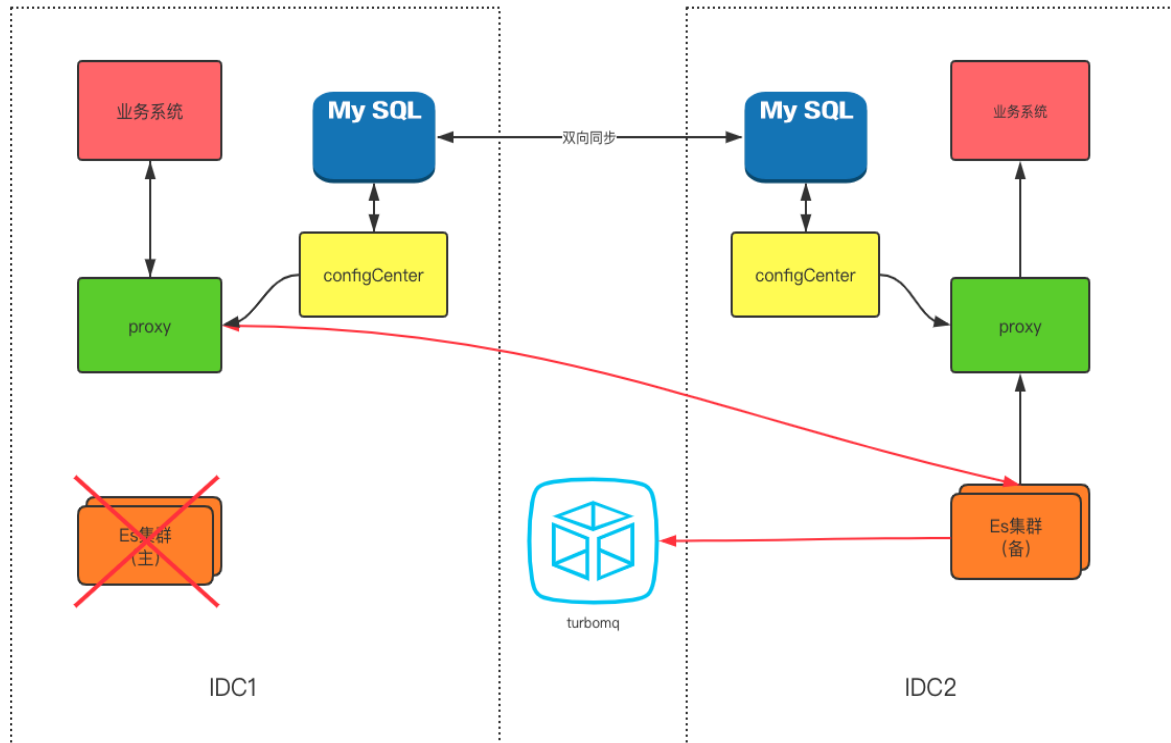
Version = 全局增长id

# 架构演进

v3-可靠性

## 单集群故障

### Step1 切换

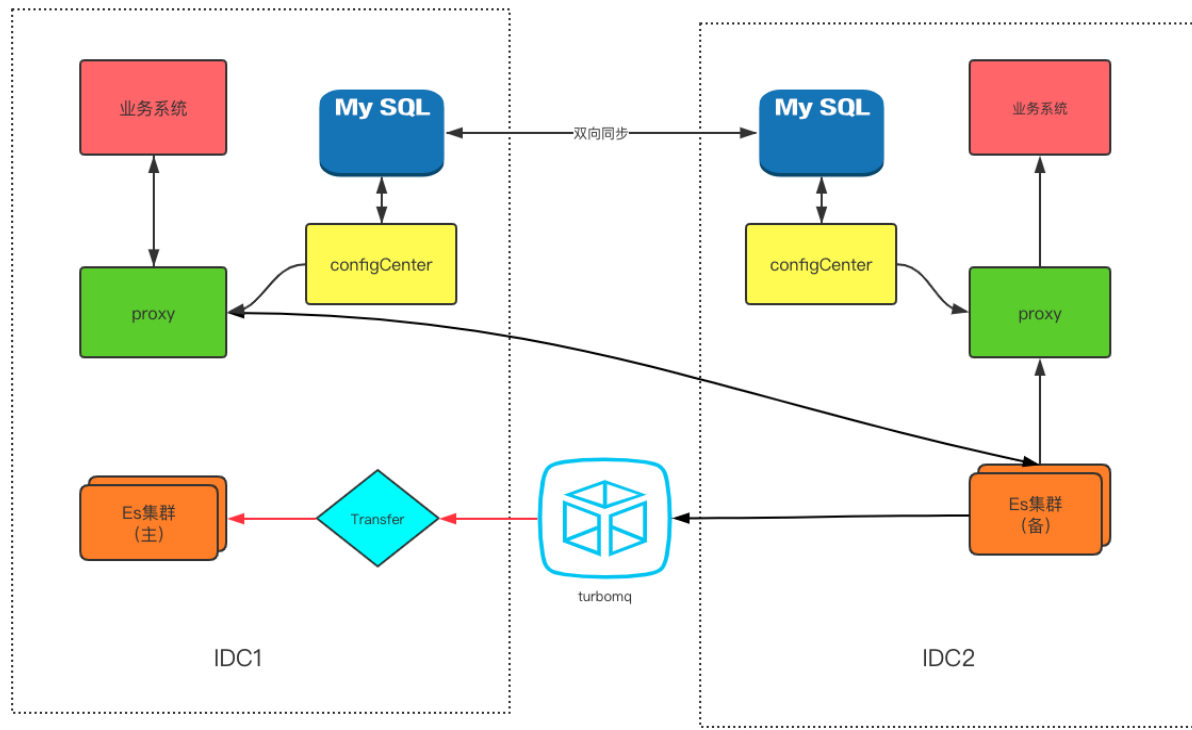


# 架构演进

v3-可靠性

单集群故障

Step2  
打开数据同步

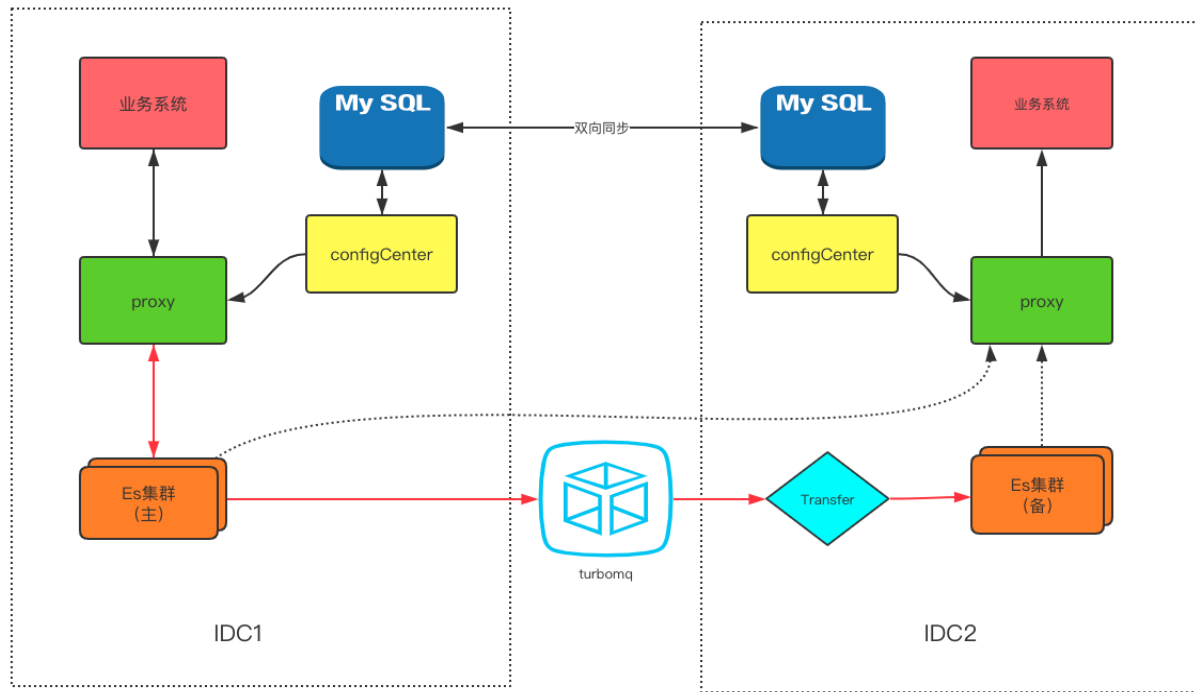


# 架构演进

v3-可靠性

## 单集群故障

### Step3 恢复最初配置



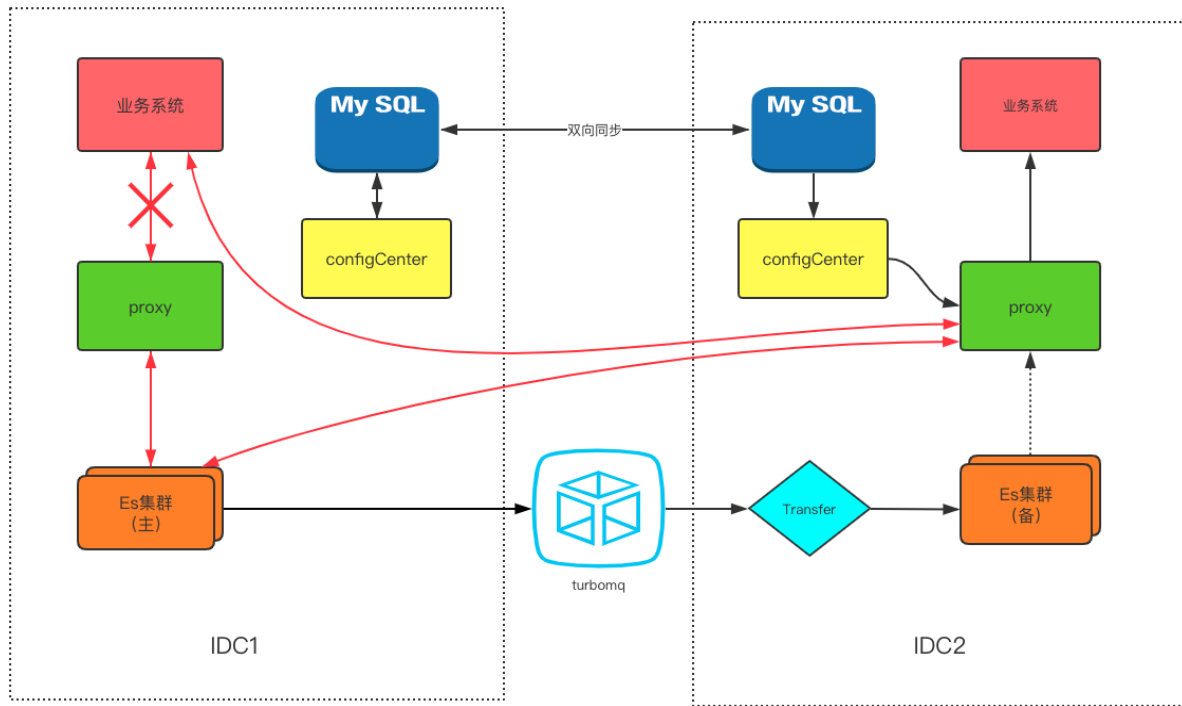
# 架构演进

v3-可靠性

Proxy故障

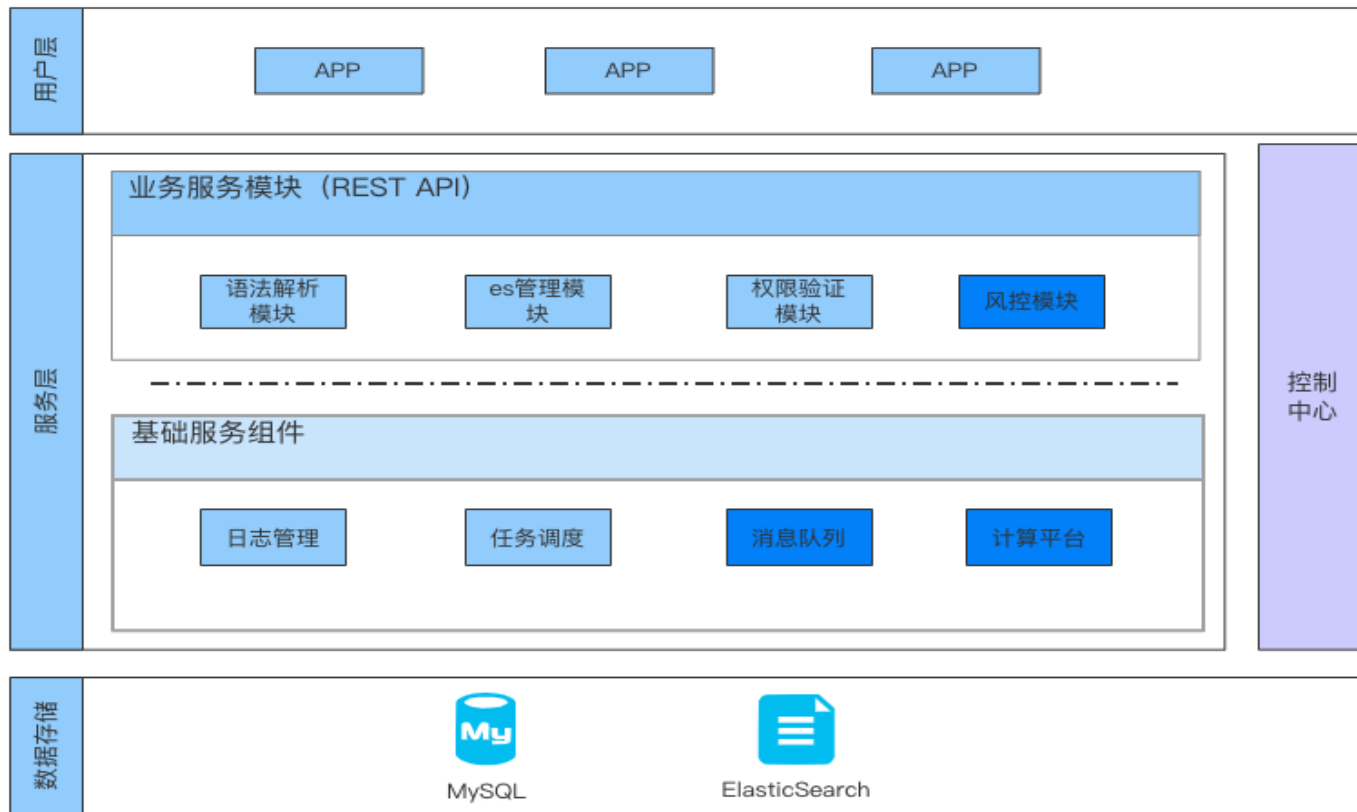
Step1  
切换

Step2  
恢复



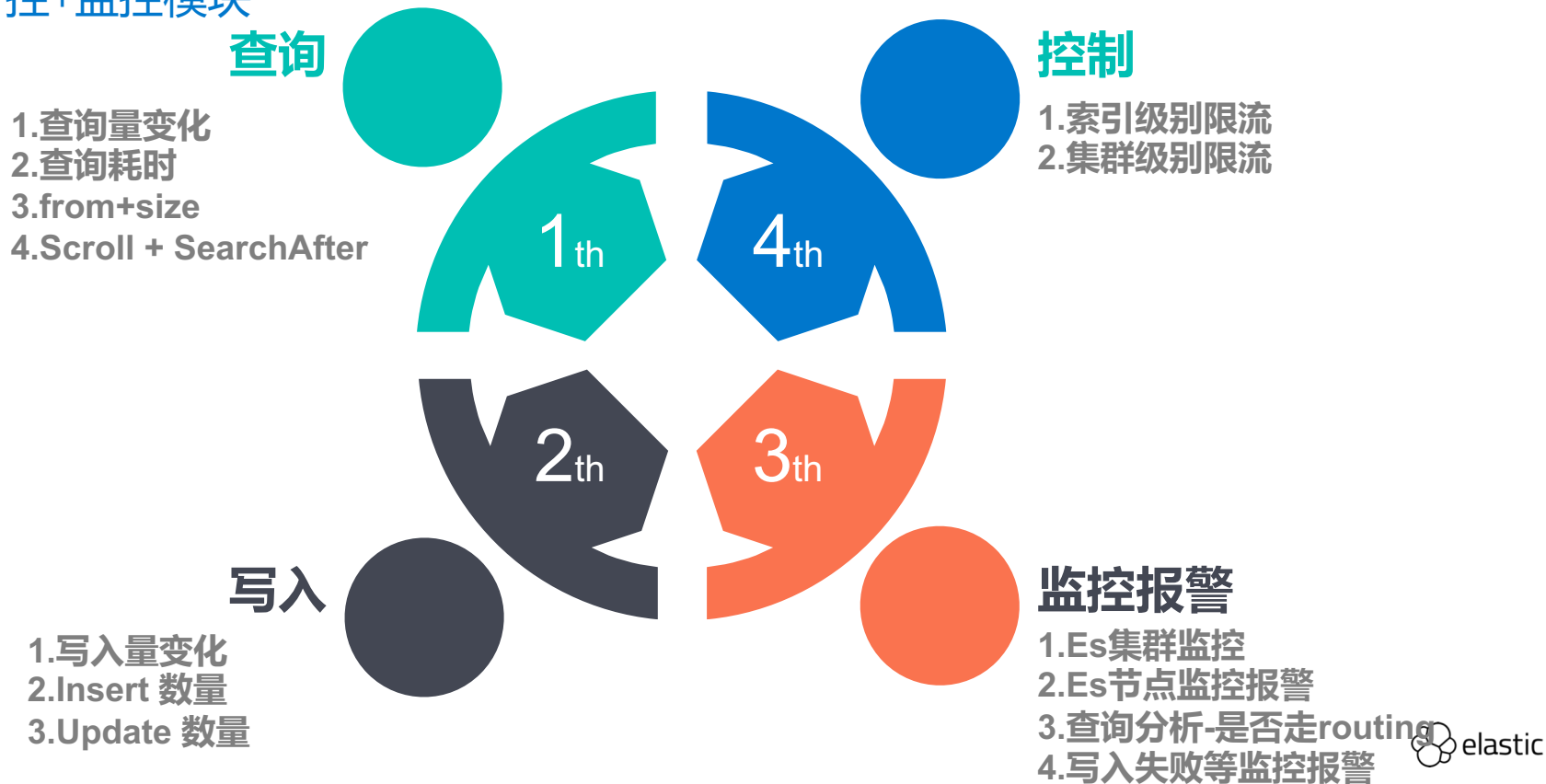
# 架构演进

## v3-可靠性



# 架构演进

## v3-分控+监控模块



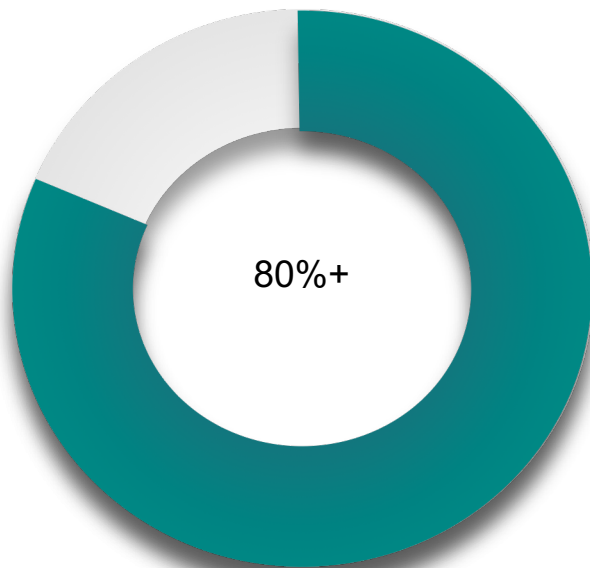
# 总结

- **简单**-所有用户都可以快速入门使用
- **稳定**-自研双中心方案下的高可用+实时计算的风控
- **高性能**-离线分析等方式提供更好使用建议

# Part3:性能优化

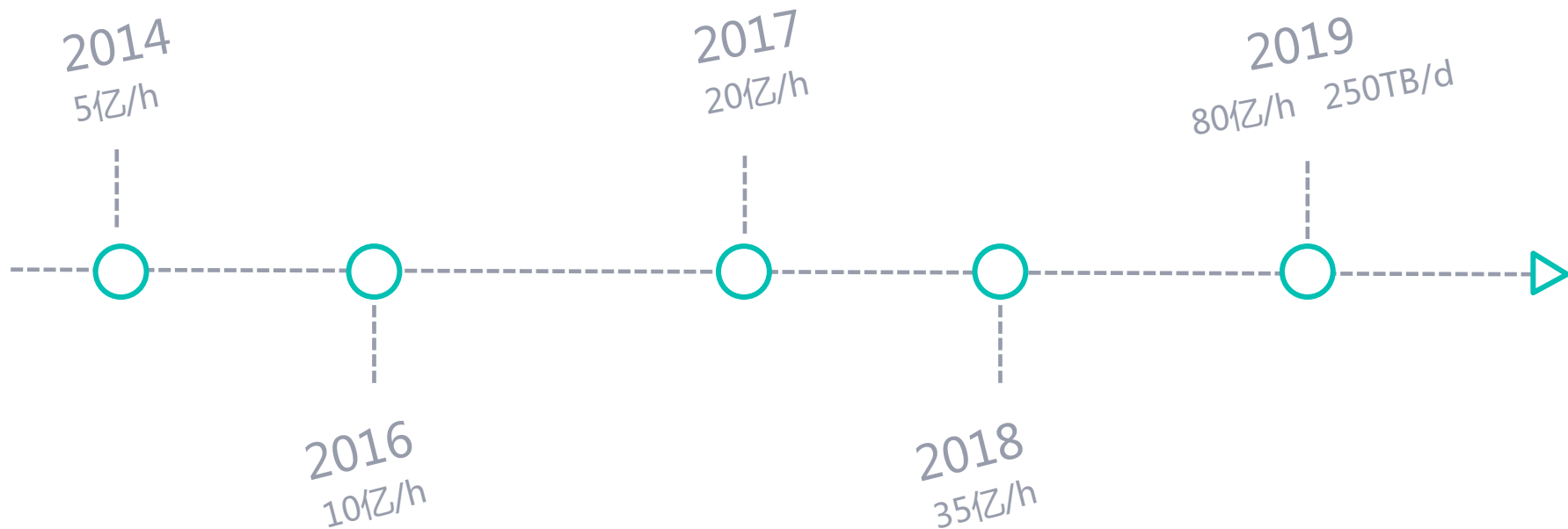
# 性能优化

## 日志集群



# 性能优化

## 日志集群



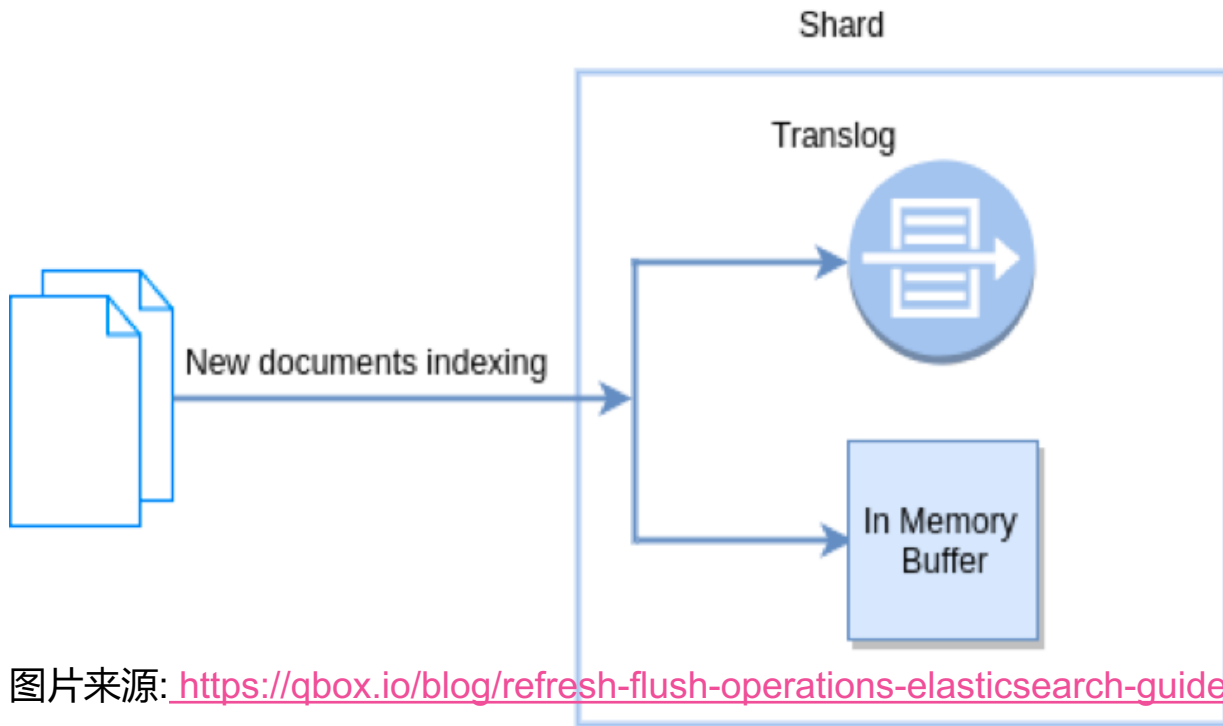
# 性能优化

## 日志集群

### Tranlog的作用

- 1.恢复时数据回放
- 2.Realtime get

~~99%~~



图片来源: <https://qbox.io/blog/refresh-flush-operations-elasticsearch-guide>

# 性能优化

## 日志集群

### Dynamic Translog

1. Replicas为0
2. Shard状态非started时

**自动打开**



# 性能优化

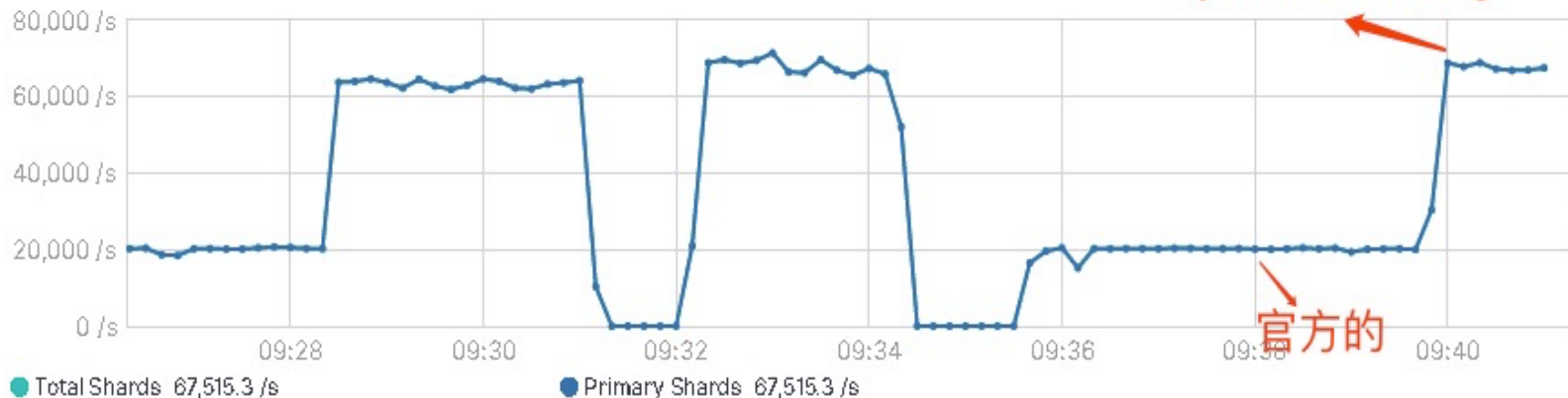
## 日志集群



# 性能优化

## 日志集群

Indexing Rate (/s) ①



真实 70%+ 提升

# 性能优化

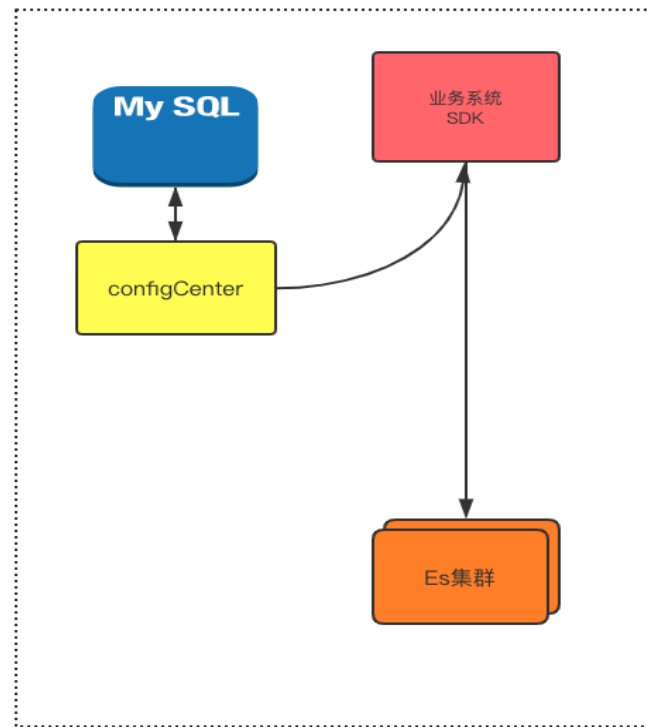
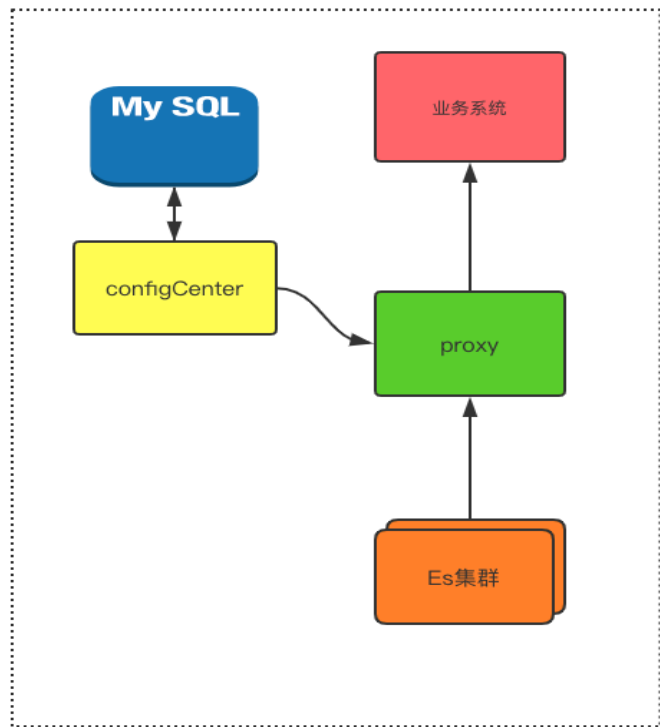
## 优化总结

1. Translog
2. Shards number
3. Mapping
4. Server/Node config
5. .....

# Part4:未来展望

# 未来展望

## 去proxy



# QA