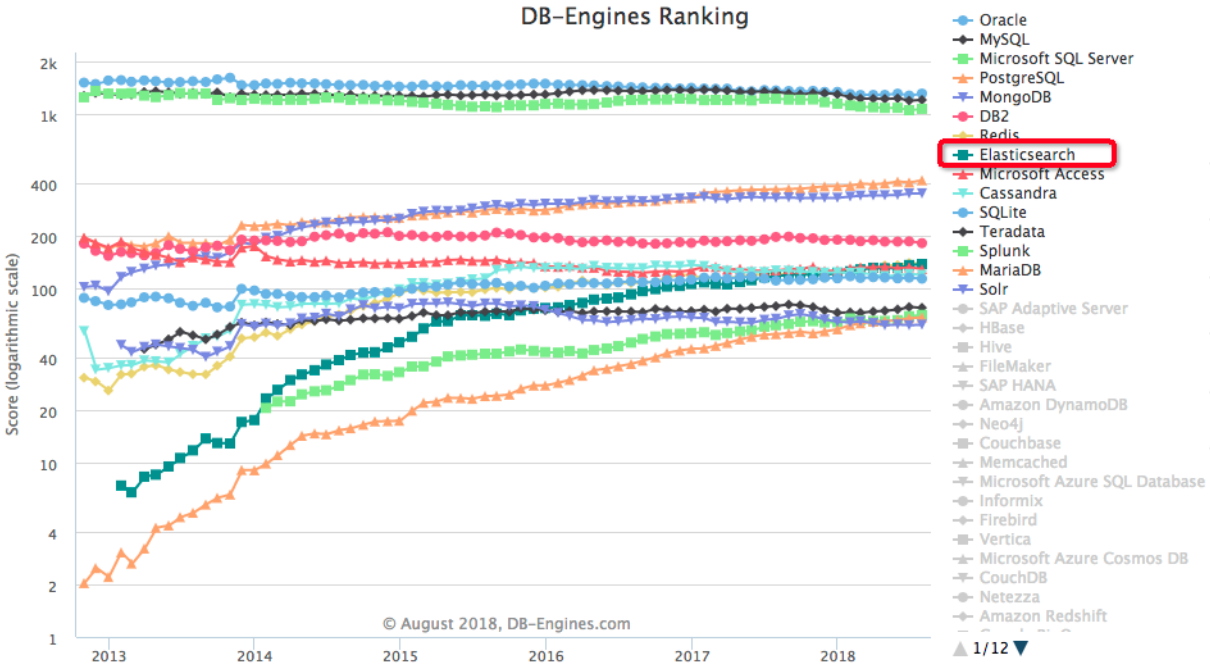
ali-es双十一最佳实践

邓建军（路放@alibaba 数据库事业部）

Elasticsearch 发展最快的数据库



. 1.0 2014.2

. 2.0 2015.10

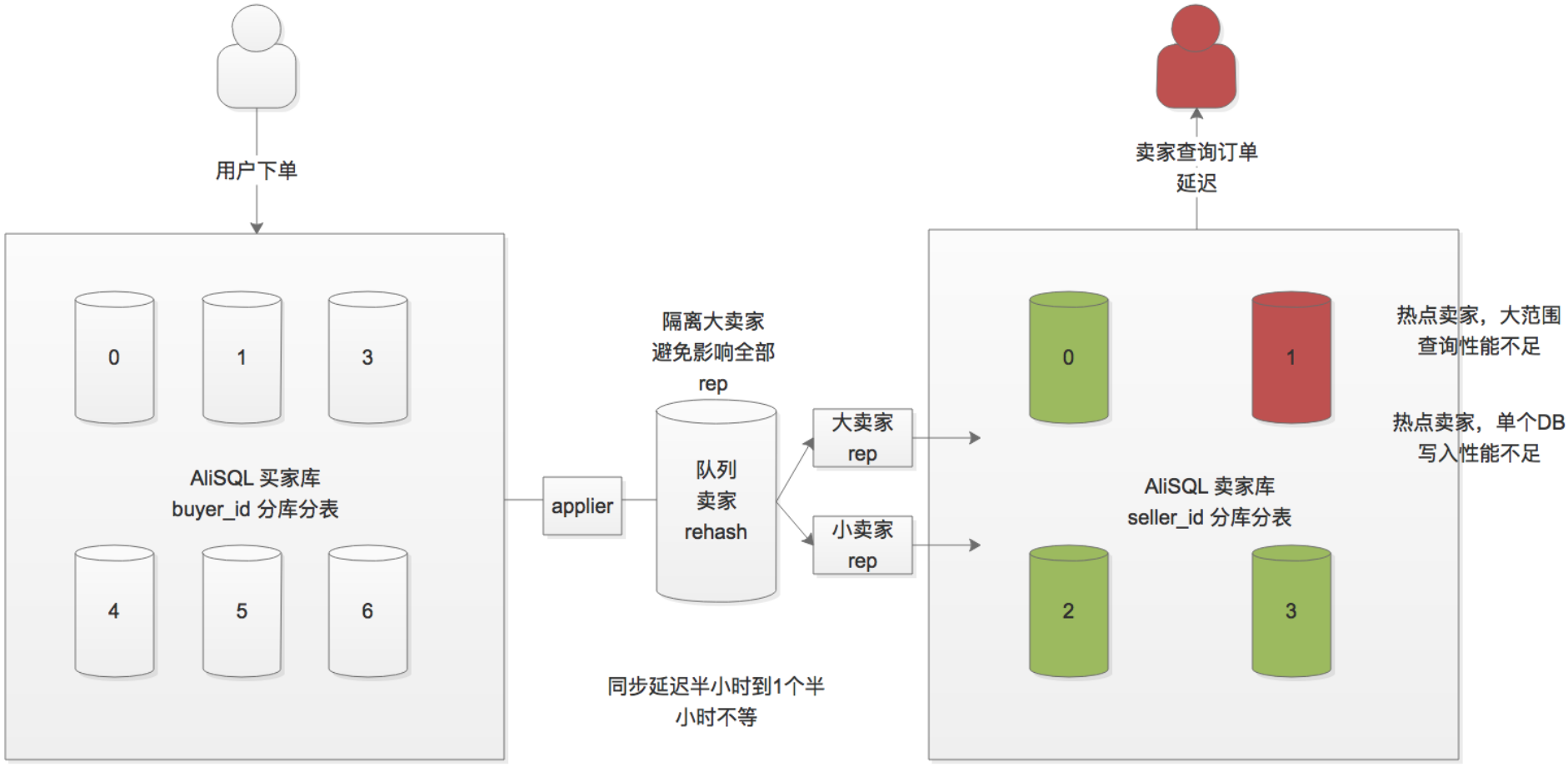
. 5.0 2016.10

. 5.3 2017.5 ali-es 1.x

. 6.0 2017.9

. 6.2 2018.3 ali-es 2.x

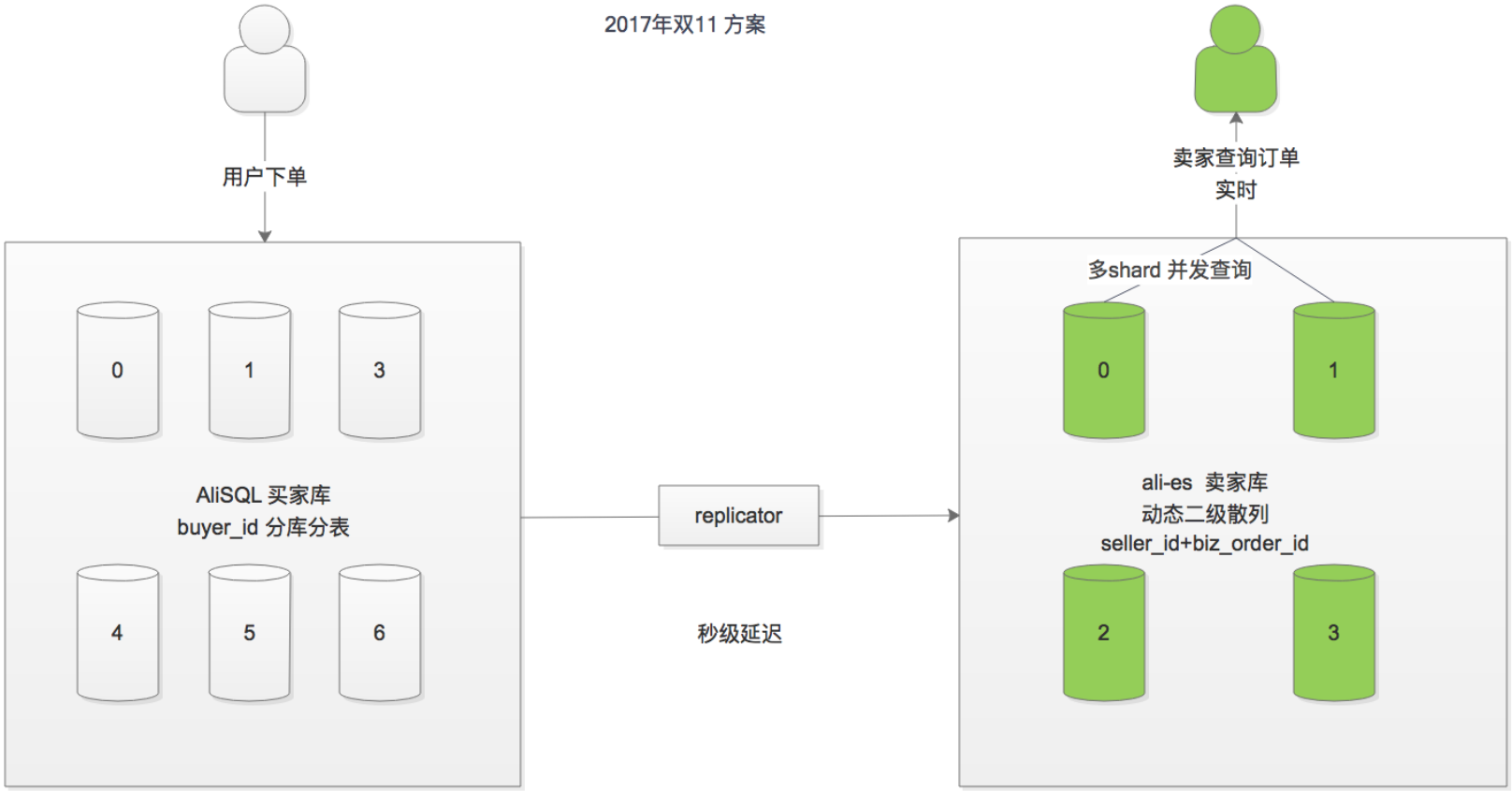
双11 核心场景


用户下单
0
1
3
AliSQL 买家库
buyer_id 分库分表
4
5
6
applier
隔离大卖家
避免影响全部
rep
队列
卖家
rehash
大卖家
rep
小卖家
rep
同步延迟半小时到1个半
小时不等
卖家查询订单
延迟
0
1
AliSQL 卖家库
seller_id 分库分表
2
3
热点卖家，大范围
查询性能不足
热点卖家，单个DB
写入性能不足

核心场景面临的问题

- 热点
  - buyer维度写入，seller维度查询（卖家负载不均衡造成数据同步延迟）
- 查询能力
  - 30ms->5ms
- 跨地域
  - 机房/城市级别容灾
  - 一地写入，多地查询

2017年双11最终方案



内核改造点

- 写入性能
  - 基于5.3 的动态二级散列
- 查询性能
  - 联合索引
  - 深翻
- 用户迁移零成本（jdbc）
  - 支持sql查询的xdriver
- 稳定性
  - 大卖家读和深翻隔离*
  - 大规模断网
  - GC问题-池化
  - 告警&诊断*

热点数据同步

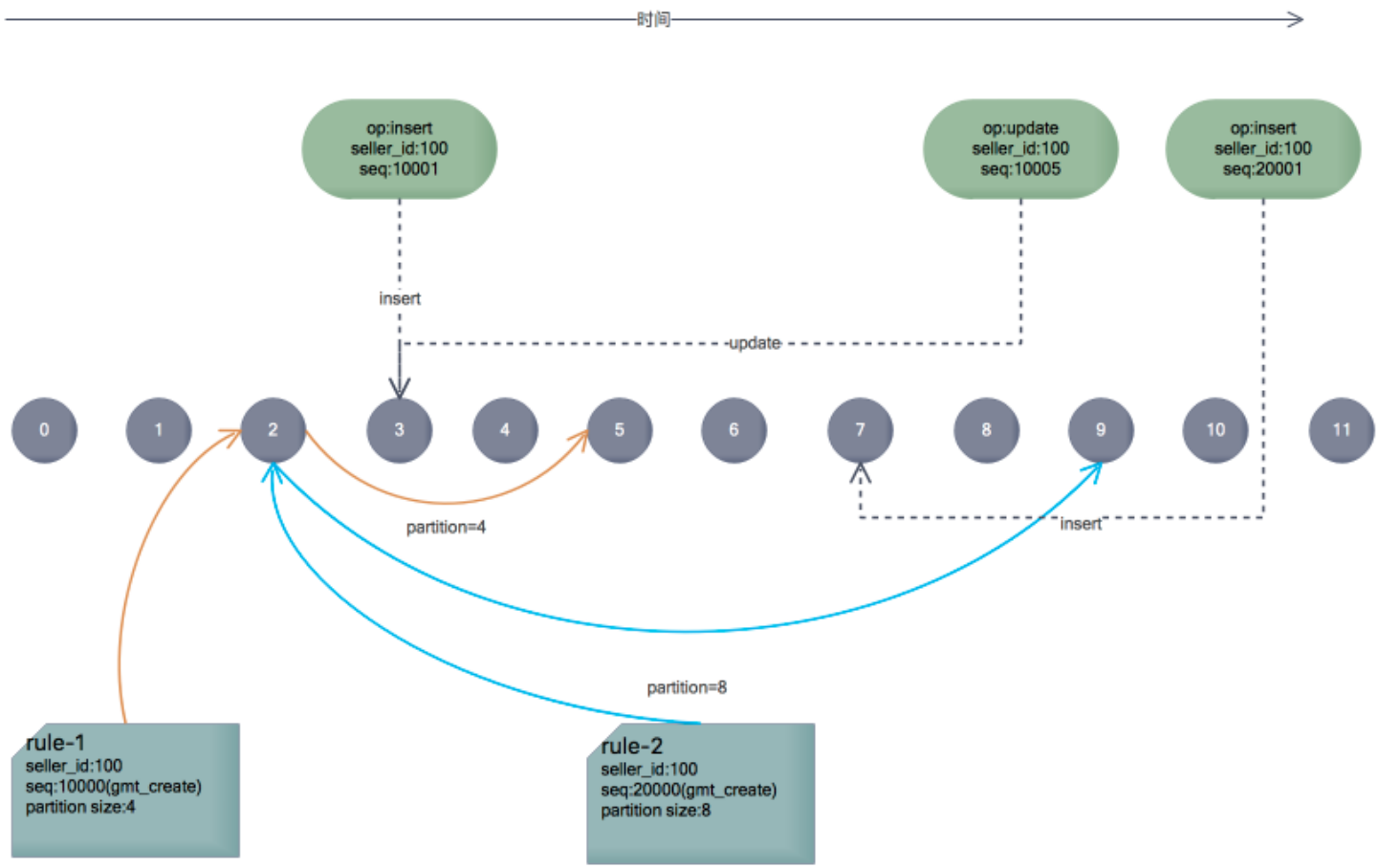
- 散列方案对比

	方案	特点	性能
传统散列方案	静态一级散列	热点大卖家无法散列 存在单点写入瓶颈	1x
ES 5.3	静态二级散列	固定全局二级散列 小卖家查询劣化	2x
ali-es 1.x	动态二级散列	基于规则二级散列， 可实时推送散列规则 兼顾大卖家写入和小卖家查询	6x

Alibaba Group 阿里巴巴集团
阿里巴巴 Alibaba.com 淘宝网 Taobao.com Tmall.com 天猫 一淘 阿里云 aliyun.com 聚划算

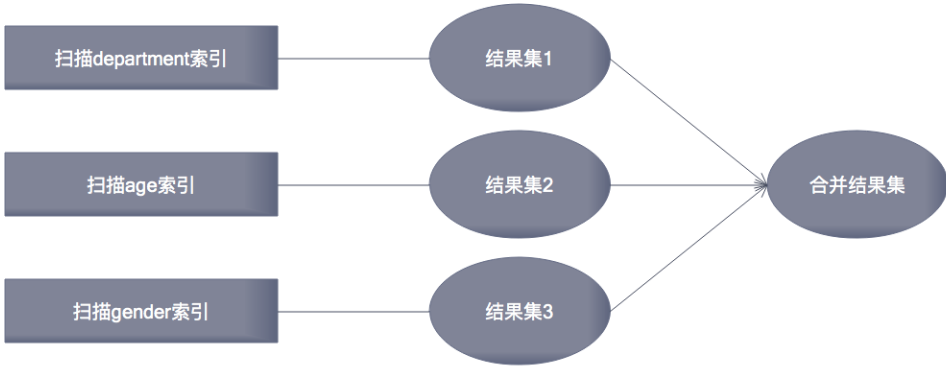
动态二级散列

有选择性的推送

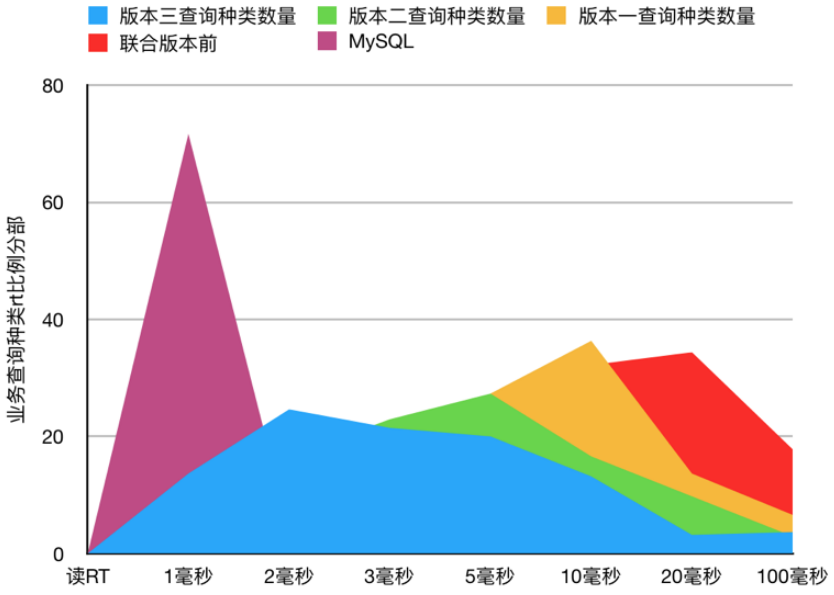


联合索引

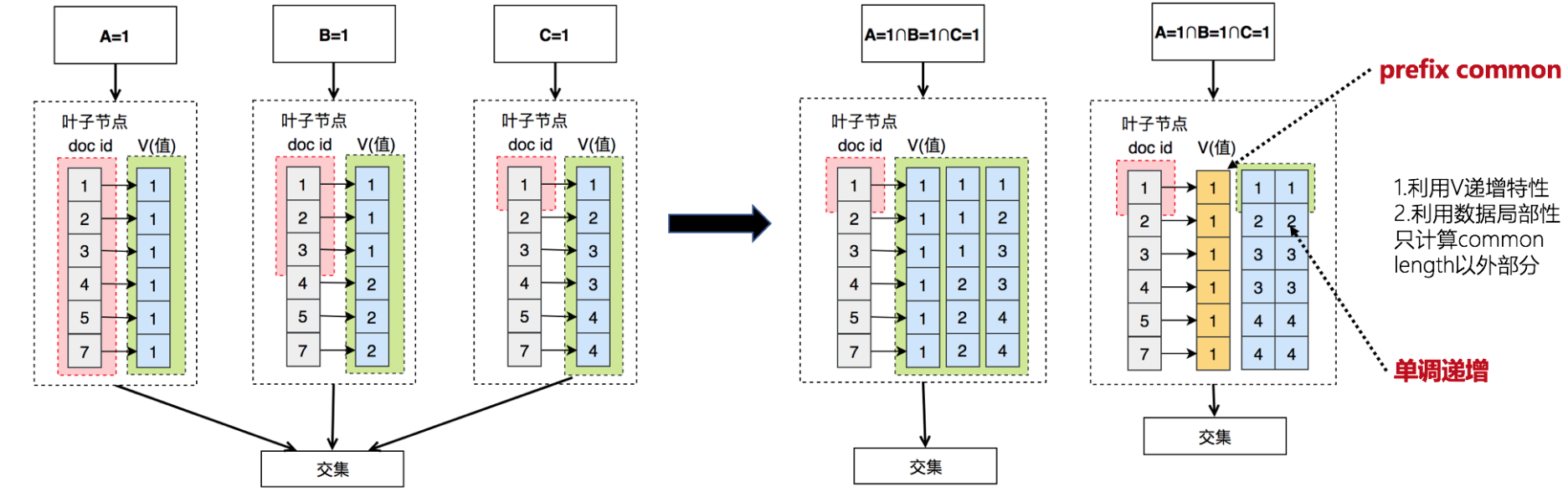
where (department = 10 OR department = 11) and age >= 25 and age <=30 and gender = 'male'



elasticsearch查询计划



联合索引



联合索引基础之上性能提升30%

从SQL到DSL（xdriver sql转换优化）

- SQL语法

```
SELECT * FROM myindex WHERE (department = 10 OR department = 11) AND (age >=25 AND age < 30) AND gender =
'male'
```

- DSL语法（druid -- elasticsearch-sql）

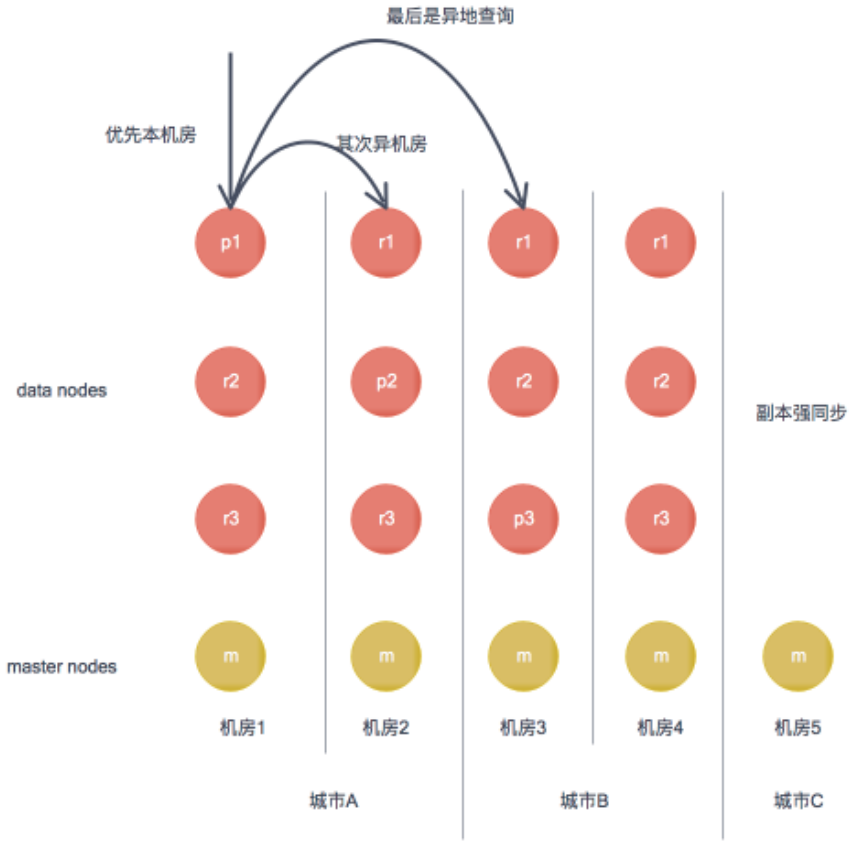
```
{"query": {"bool": {"must": {"bool": {"must": [
     {"bool": {"should": [
        {"match": {"department": {"query": 10}}},
        {"match": {"department": {"query": 11}}}]}},
    {"range": {"age": {"from": 25,"to": null,"include_lower": true,"include_upper": true}}},
    {"range": {"age": {"from": null,"to": 30,"include_lower": true,"include_upper": false}}},
    {"match": {"gender": {"query": "male"}}}]}}}}}
```

- DSL语法（ESDB）

```
{"query": {"bool: {"must":[
{"terms": {"department":[10,11]},
{"range":{"age":{"gte":25,"lt":30}}},
{"term":{"gender":"male"}}]}"}}
```

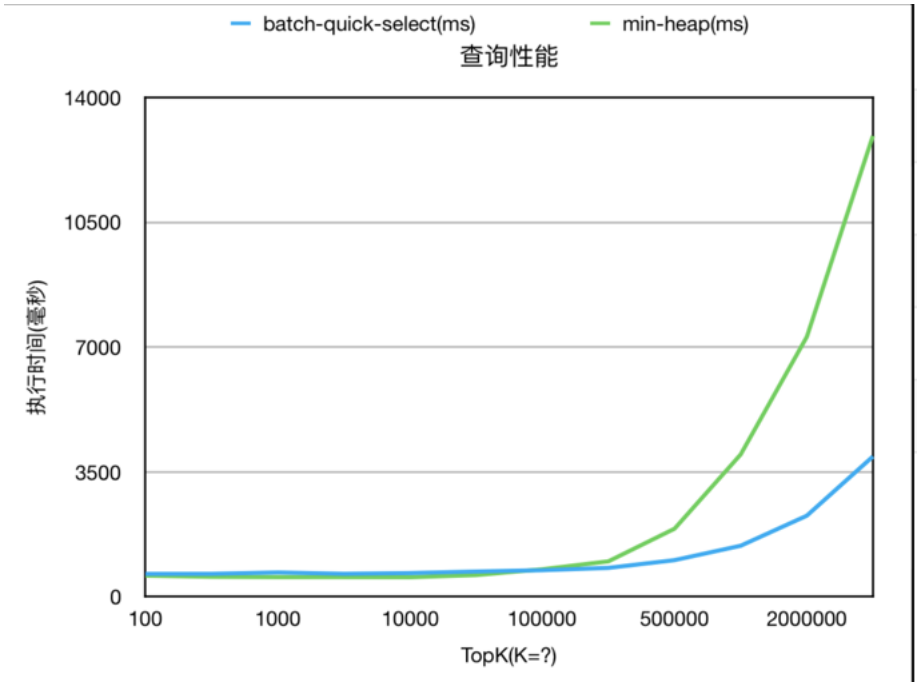
异地部署（youtu核心场景）

- 3地5机房部署
  - 机房/城市级故障，自动容灾
  - 查询本地化
    - 优先选本机房副本
    - 次选同城机房副本
    - 最后选异地副本



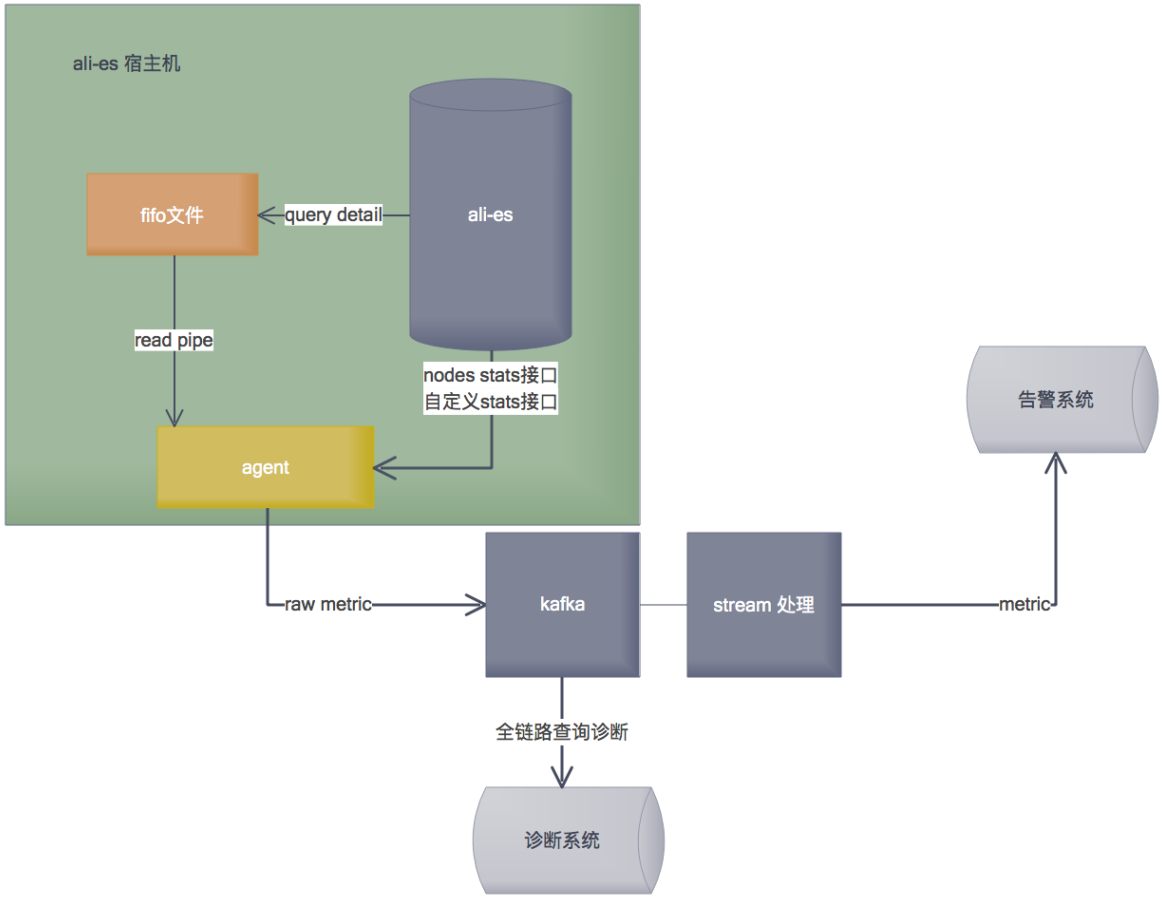
Ali-es稳定性

- 大卖家的查询隔离（低优先级线程池）
- 深翻的查询隔离（深翻线程池）
- 深翻优化
  - quick select
  - 池化对象，避免晋升



Ali-es监控诊断体系

- 采集方式
  - Nodes stats接口
  - 自定义stats接口
  - 查询pipe文件
    - 父查询时间
    - Routing
    - 执行节点信息
    - Sql/dsl
    - 任务task



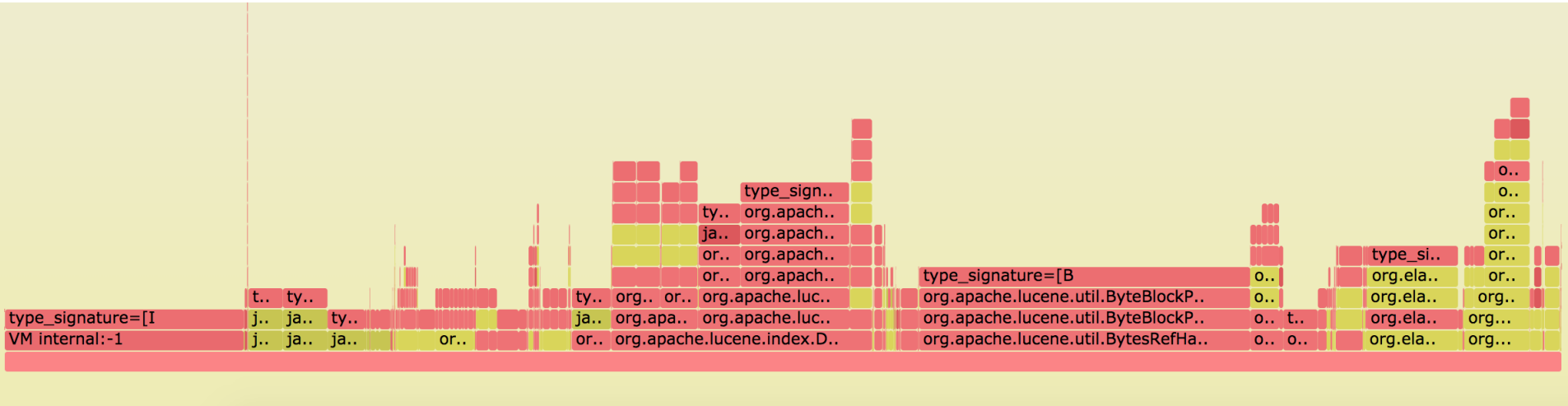
Ali-es 2017双十一电商表现

- 买卖家同步，相比于2016年，在零成本增加的前提下，首次实现零点高峰秒级延迟
  - 数据库峰值32.5万笔每秒
- 双十一当天平均rt 5ms，日常3ms
  - 与MySQL相当

阿里巴巴 Alibaba.com 淘宝网 Taobao.com Tmall.com 天猫 一淘 阿里云 aliyun.com 聚划算

Jvm相关优化

- SIMD
  - 增强byte数组compare速度
- 使用jetspaner（晋升/survivor内存诊断工具）
  - 降低应用young gc时间
    - 展示内存move的每一个事件
    - 火焰图展示内存产生调用盏



Ali-es总结

- 能力定位
  - 多维度查询索引
  - 日志分析引擎
- 和社区一起发展
  - 贡献优化patch和新功能组件
  - 兼容社区版本

Q&A

Thanks

邮箱地址: jianjun.djj@alibaba-inc.com