

基于ES的音乐搜索引擎

Jakes

2017年11月25日

目录

酷狗音乐搜索引擎架构变迁

构建音乐搜索引擎经验之谈

架构变迁

伪集群



真集群

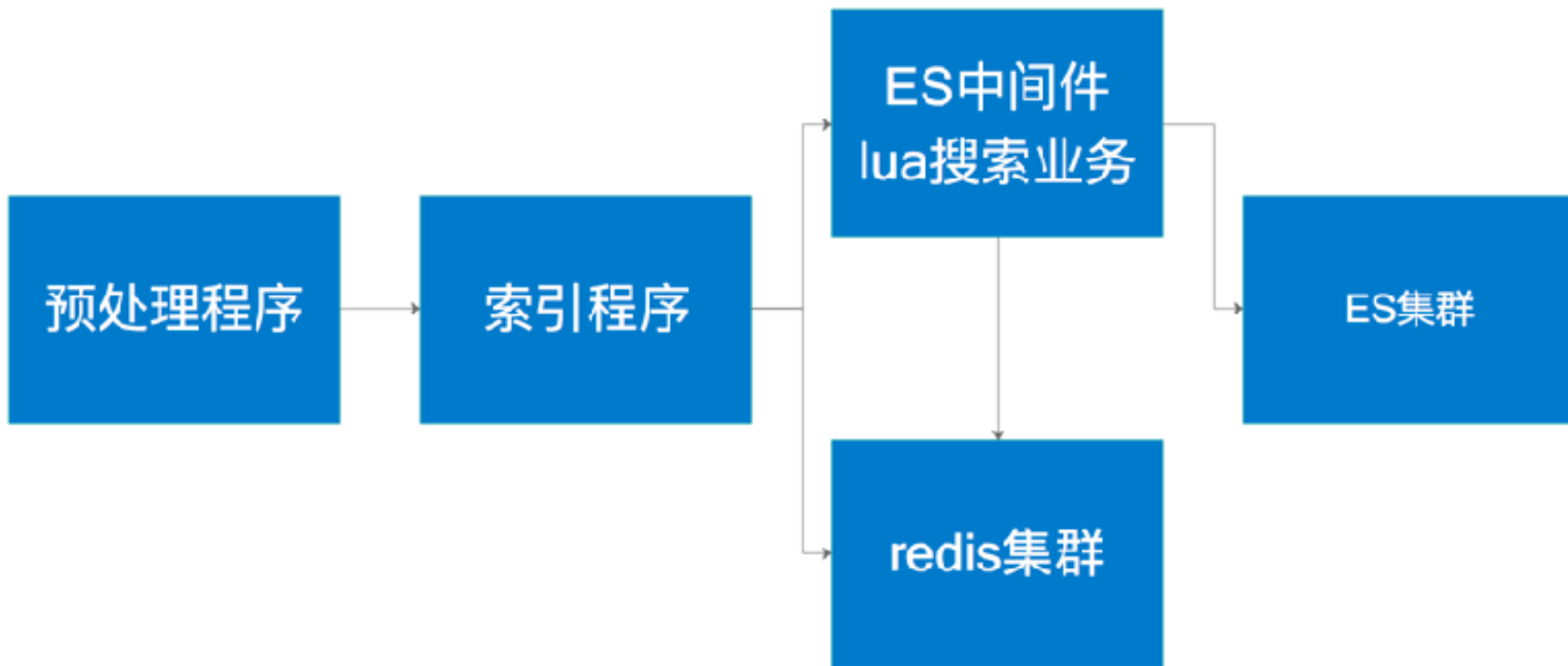
预处理程序

索引程序

ES中间件
lua搜索业务

ES集群

ES + redis集群



中文分词器

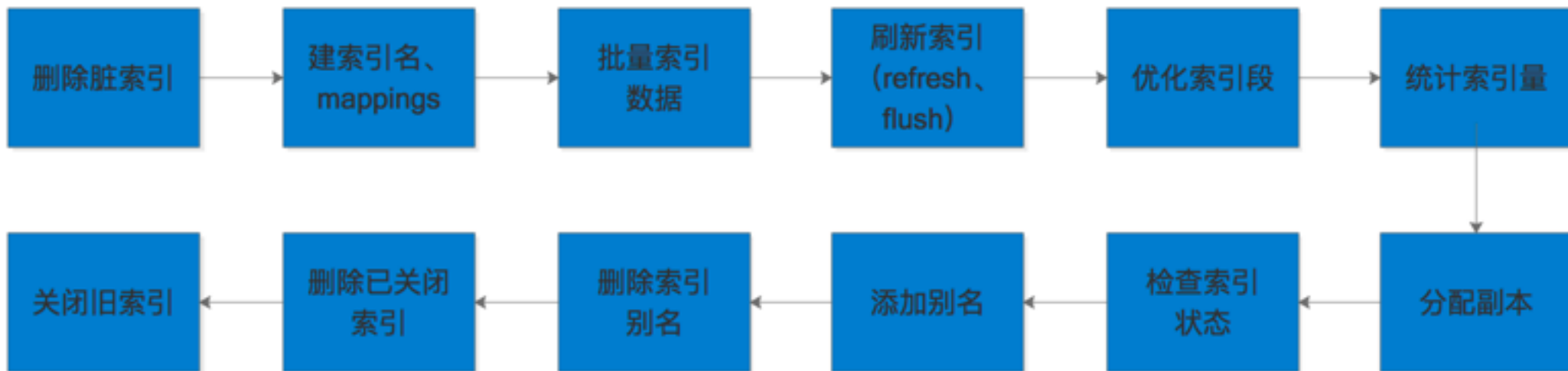
中文分词器

- 开源中文分词器（IK、ansj、jcseg、HanLP、word、jieba等）
- 如何选择？
- 为什么要自己写？为了KPI？
- “不污”中文分词器
 - 具有核心、繁简体、标点符号、排除、替换词典
 - 支持混合字符分词（英文词组、中英文、中文数字、数字英文、标点非标点等）
 - 基于MMSEG算法，含有多种分词策略
 - 支持词典热更新

索引篇

索引篇

- 索引流程



索引篇

- 优化及注意点

- 索引名加上时间戳
- mappings写在文件，方便修改
- number_of_shards: 1、number_of_replicas: 0、refresh_interval: “0s”
- 排序、聚合字段采用doc_values
- 根据需求应用不同分词策略
- 采用生产者消费者方式批量索引数据
- 优化索引段为1
- 统计索引量，防止建空索引
- 恢复副本数 ($R = N / P - 1$) ，修改refresh_interval
- 关闭索引以作备份
- 涉及到索引状态变更的操作一定要有重试



elastic
中文社区

IT大咖说
知识共享平台

搜索篇

搜索篇

- 常用query
 - multi_match query
 - bool query
 - function_score query
 - nested query
- 常用filter
 - bool filter
 - term filter
 - range filter
- 插件

搜索篇

- 优化及注意点

- query vs filter

- query计算得分，filter不计算得分
 - filter可以缓存结果，query“不行”（当search_type=count&query_cache=true时，可以缓存hits.total, aggregations, suggestions，当分片刷新时，缓存失效）

- filtered query vs post filter

- filtered query影响aggregations 和 hits
 - post filter只影响hits

- query写成文件，方便修改

- 限制最大from、size、搜索关键词长度

- 排序调优，要善用_explain，各种query搭配不同分词方式



elastic
中文社区

IT大咖说
知识共享平台

其他

其他小技巧

- 增量和全量分开方法索引，以免增量堆积
- `index.mapper.dynamic: false` 禁止自动生成mappings
- 优化小索引段为1个



elastic
中文社区

IT大咖说
知识共享平台

Q&A

Thanks

就是歌多