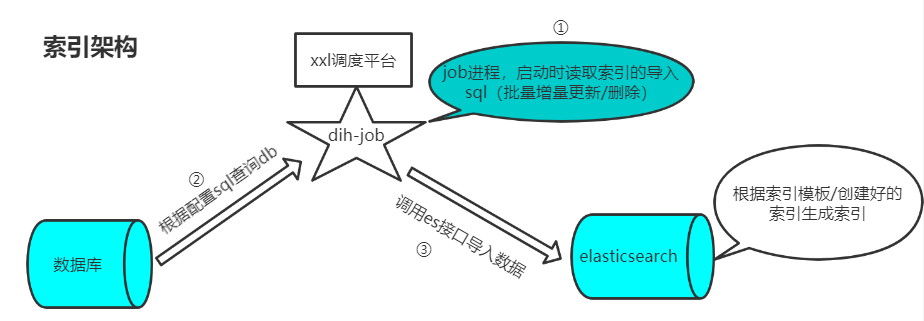
一、搜索引擎安装软硬件要求



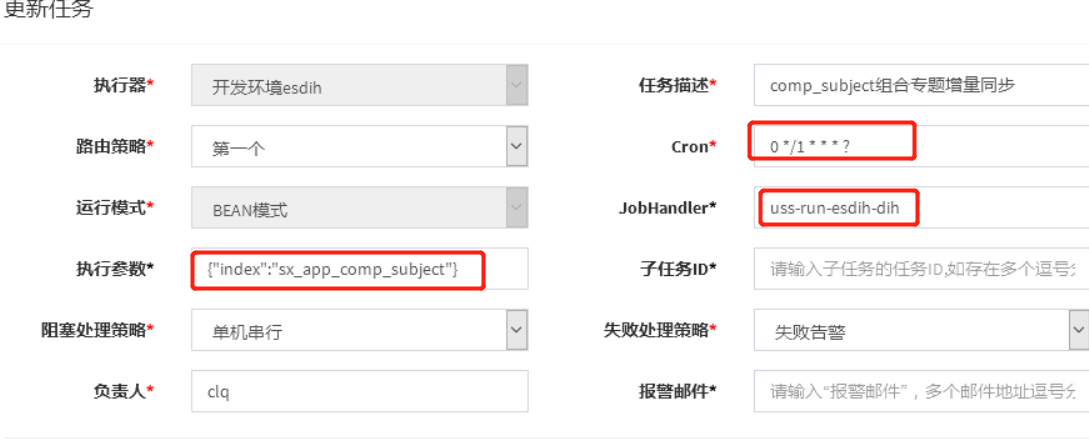
二、搜索引擎对比es vs solr

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | Elasticsearch | solr | 优势方 |
| Dih | 自研 | 自带 | 自研快 |
| 索引配置 | 索引模板 | Scema文件 | 各有利弊 |
| 集群 | 自动组成集群 | 通过zk组成集群 | Es不需要维护额外的组件 |
| 安装 | 7.X之后版本自带jdk，安装简单，修改elasticsearch.yml即可 | 需要借助容器 | Es稍好 |

三、elasticsearch索引

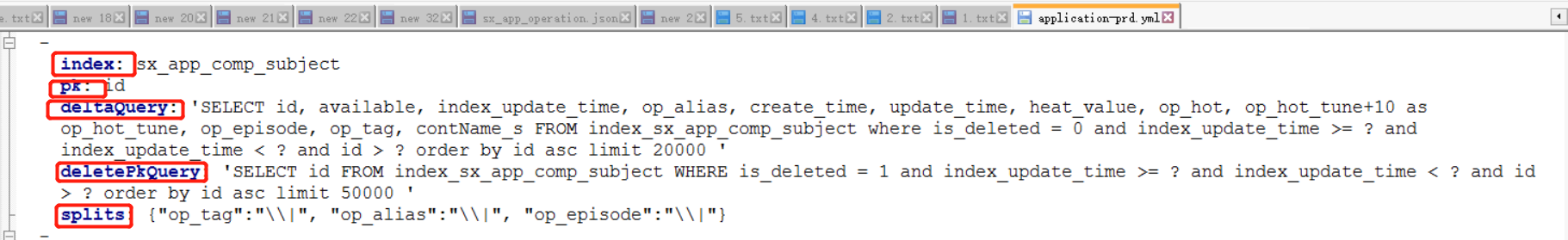


3.1 dih配置



|  |  |
| --- | --- |
| 属性名 | 描述 |
| Cron | Job执行周期 |
| JobHandler | 代码写死的job处理bean实例名称，固定uss-esdih |
| 执行参数 | Dih任务的导入索引名称，json格式{“index”:”索引名”} |

3.1.1 dih导入sql配置



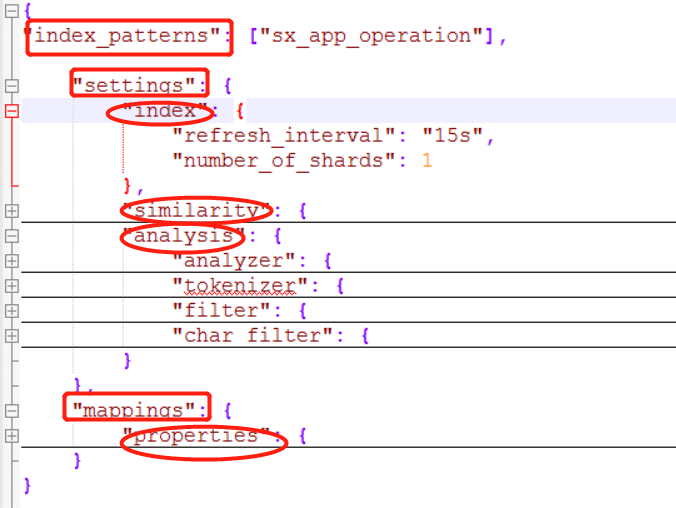
|  |  |
| --- | --- |
| 属性名 | 描述 |
| Index | 导入到es中的索引名字 |
| Pk | 表的主键，dih利用该字段为主键索引，提升sql查询性能 |
| deltaQuery | 增量导入sql语句，查询时代码内部已经将查询开始时间往前回拨30s |
| DeletePkQuery | 增量删除的sql语句，查询时代码内部已经将查询开始时间往前回拨30s |

3.2 索引模板

索引模板和索引的关系

索引模板可以指导索引的创建，**索引模板**只**是**一个**配置**，而**索引**是**真实存在**的存储。**搜索**时**可以没有索引模板**，但**必须有索引**。

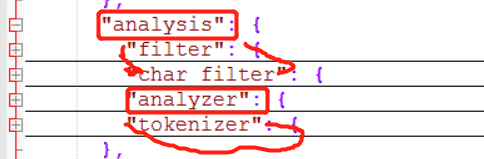
3.2.1 模板常用属性介绍



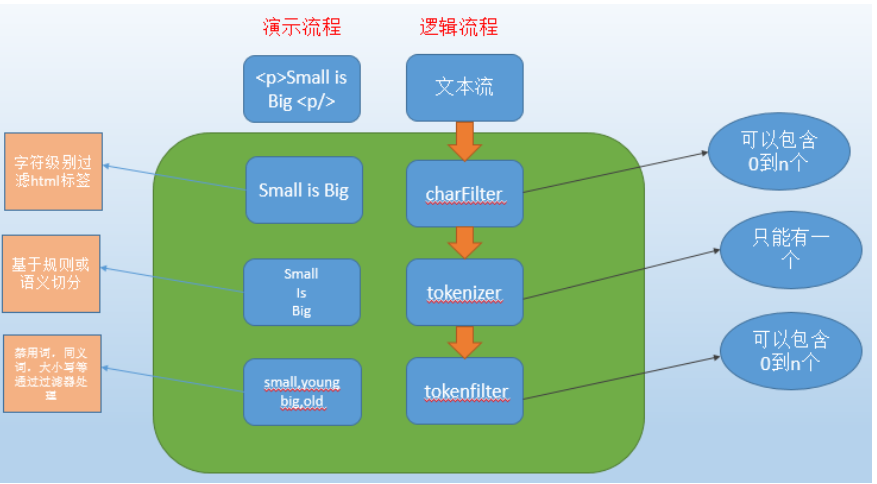
|  |  |  |  |
| --- | --- | --- | --- |
| 名称 | 描述 | 子项 | 说明 |
| **Index\_patterns** | 根据当钱索引模板创建索引的 索引名称正则表达式 |  |  |
| **Settings** | 配置分片和副本数、索引分词定义 |  | refresh\_interval统一填15s,意思是入索引15s后可见 |
| **Mapping** | 定义字段的使用哪个分词/是否存储/索引类型等（字段和分词类型等） |  |  |

3.2.2 执行顺序

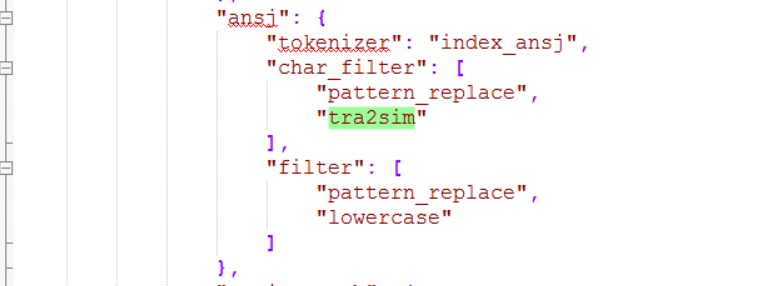
3.2.2.1 setting



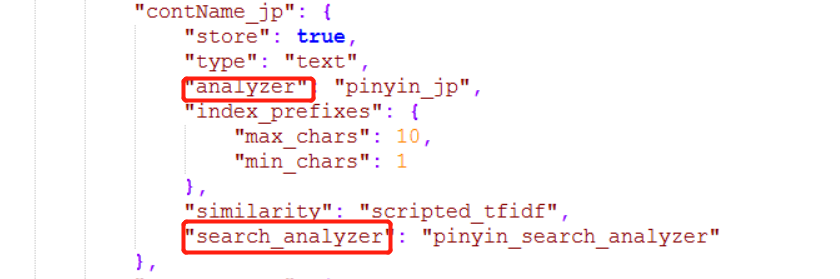
一个analysis集合 可以包含多个analyzer，而一个analyzer则由一个单个的tokenizer，零个或多个的tokenfilter组成，而一个tokenizer又可以包含零个或多个的charFilter。总体的执行流程如下图



例如ansj分词的setting--analyzer



3.3 mapping配置



|  |  |  |
| --- | --- | --- |
| 名称 | 描述 | 注意 |
| store | 是否存储原词 | Copyfiled不管是否为true都不会存储 |
| type | 字段类型，integer/keyword/text等 |  |
| analysis | Index阶段分词选项 |  |
| search\_analysis | search阶段分词选项 |  |
| Index\_prefixes | 前缀索引（性能优化） |  |
| similarity | 定制的相关计算算法 |  |

3.4 踩过的坑

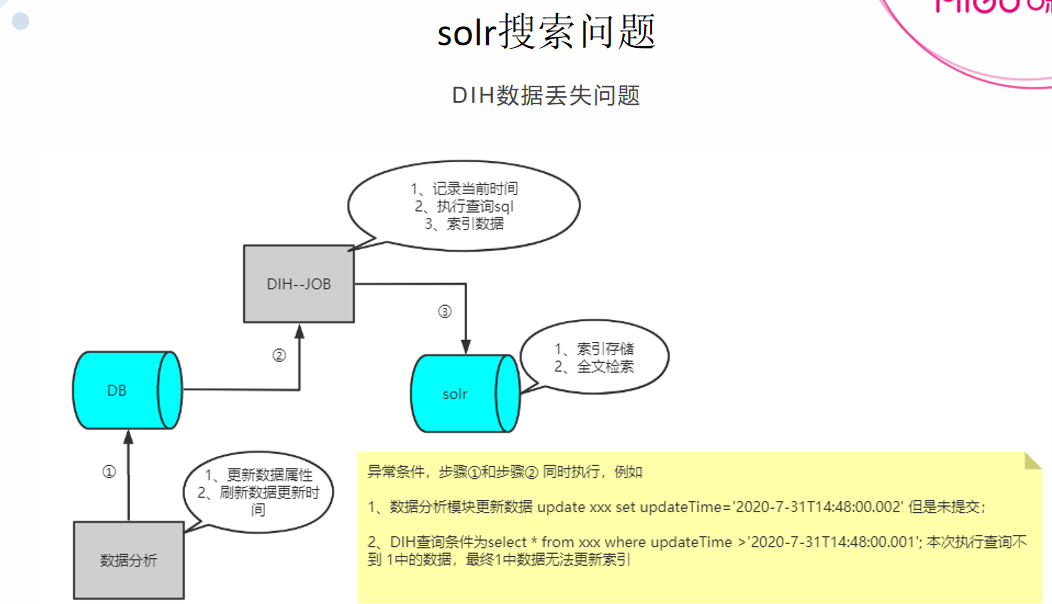
3.4.1 dih时区配置

dih导入数据的定时任务，每次会根据时间范围查询mysql数据，查询条件例如：index\_update\_time >= {lastExcTime} AND index\_update\_time < {currExcTime},

lastExcTime和currExcTime是和数据集连接中配置的时区相关的，首次没有配置该时区，默认按照操作系统默认时区计算了时间，导致lastExcTime和currExcTime和数据库时间不一致，最终查询到的数据比实际数据晚 （数据库时区 - linux时区）小时

3.4.2 dih数据丢失

老生常谈，请阅



3.4.3 分词插件安装遗漏

现网安装es的时候，遗漏知会运维需要安装，hanlp以及繁简转换的分词插件。建议对于新安装的搜索引擎可以直接copy已经调试好的安装包直接解压，修改部分配置即可。

3.4.4 dih任务配置错误

Gk数据下线之后，es数据未发生变化，查看dih - gk配置，配置的索引名称错误。导致增量未入索引库。

3.4.5 分词类型配置错误

analyzer分词的tokenizer配置的非分词字段，导致单字以及分词类型字段搜索不到数据。

dz分词第一步是进行单个字的term匹配，然后根据匹配结果的slop进行筛选，相邻字段slop为1的符合要求。

tokenizer配置的非分词字段导致所有的index分词term的slop都变成了0，匹配不到

3.4.6 原生的bm25打分算法丢失权重计算结果

视频原有代码使用固定得分（对应solr语法 A:b^=1）,为兼容es使用索引时取消tf-idf算法所需参数，导致分词带权重的匹配无法计算bm25相关度。

3.4.7 分词特殊符号问题

特殊符号需要在tokenizer之前的char\_filter先执行，否则，进行分词tokenizer之后，在进行去除，改position对应的term的词会变成null，导致匹配也会出现问题

3.4.8反序列化

es反序列化一开始使用fastjson，压测时发现总在时间反序列化函数上等待，导致进程卡死。后来换成jackson，没有出现。

定义es索引字段最好和java bean字段一致，否则要添加jackjson注解，容易和fastjson冲突。

3.4.9重启es时使用安全重启

重启es时务必先停止分片分配。否则造成各节点副本数据不同步，集群找不到主节点，只能重建集群。

参考链接： <https://www.cnblogs.com/bonelee/p/7285878.html>

3.4.10 性能

Integer、前缀索引、停用词、脚本评分、常量得分（0得分）/缓存

|  |  |  |
| --- | --- | --- |
| 名称 | 检索 | 优化 |
| Integer | 转rangeQuery | 索引字段，在不进行范围查询时，尽量定义为keyword使用term查询 |
| A\* | 先循环匹配项，在做term匹配，量大性能低，内存占用多 | 根据需要建索引时使用index\_prefixes选项 |
| 停用词 | 在大多数内容几乎都出现过，召回数据多，score计算慢 | 导入停用词，的、地等 |
| 脚本评分 | 实现while(hasnext){  Doc doc=doc(i);  Script(xxx);  } | 使用functions替换 |
| 常量得分（0得分） | 权重设置为0的会缓存 | 设置很小的权重，如0.0000001 |

三、查询语法

|  |  |
| --- | --- |
| 名称 | 描述 |
| query | 查询dsl：  GET sx\_app\_content/\_search{  "query": {  "match": {  "contName": "elasticsearch"  }  }} |
| filter | 过滤：  GET sx\_app\_content/\_search{  "query": {  "bool": {  "should": [  ...  ],  "filter": [  { "term": { "available": 1 }},  { "range": { "publish": { "gte": "2015-01-01" }}}  ]  }  }} |
| function | 结果提升权重:  GET sx\_app\_content/\_search{  "query": {  "function\_score": {  "field\_value\_factor": {  "field": "isLong",  "factor": 0.5,  "modifier": "none",  "missing": 0  }  }  }} |
| agg | 结果进行统计:  GET sx\_app\_content/\_search{  "aggs": {  "xxx": {  "terms": { "field": "genre" }  }  }} |

四、工具

4.1 kibana

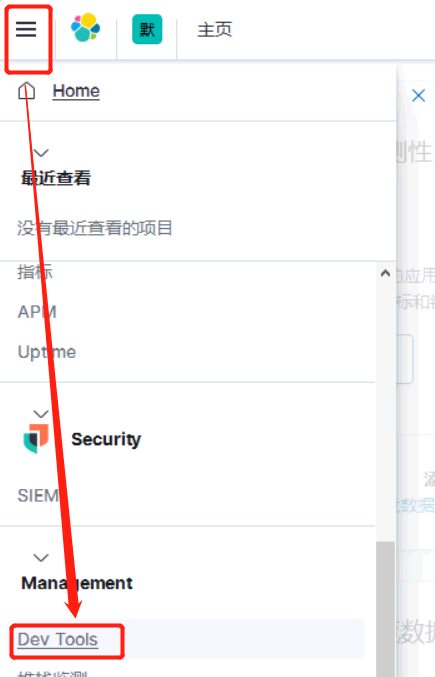
4.1.1 地址



4.1.2 查看索引

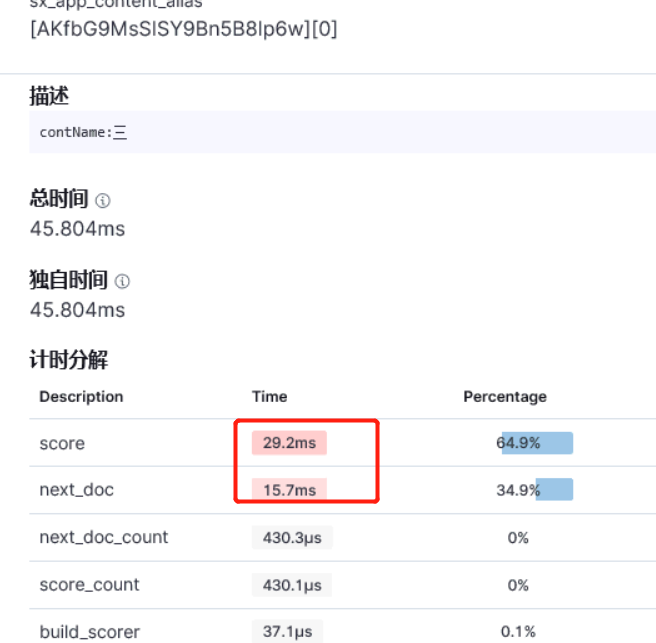


4.1.3 开发工具



4.2 查询分析profile



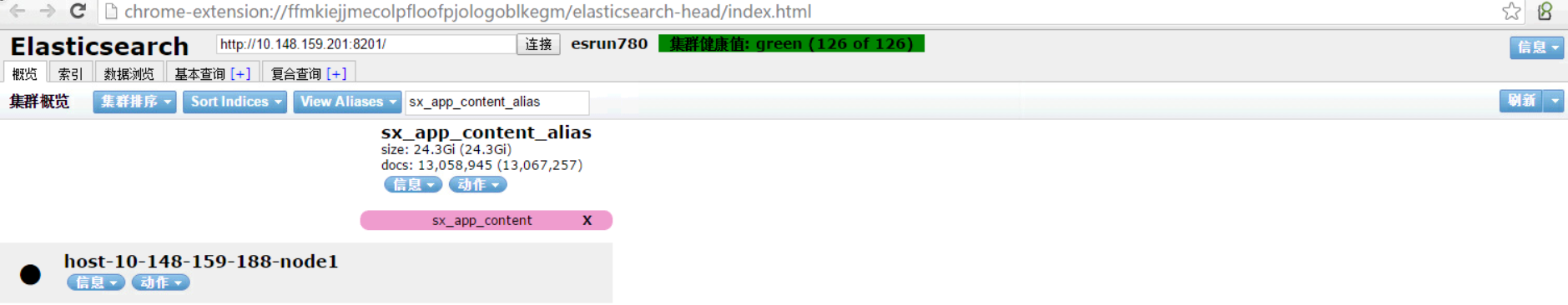


4.3 head插件

安装包如下，直接拖入谷歌浏览器即可，部分不能安装的，修改包后缀名，本地解压在导入到浏览器。



其使用界面如下



4.4 浏览器选择，带快捷自动联想的

个人推荐IE