数据流

spider =》 Kafka =》 ES =》 ES =》 Mysql =》ES =》API =》App

约定

1. ES索引约定

- gnews_raw_data（alias）: 存储爬虫抓取的引力资讯原始新闻数据的内容（真实索引使用gnews_raw_data+第一次写入数据时的日期）；
- gnews_valid_data（alias）: 将gnews_raw_data中的数据经过处理逻辑后，计算出的有效数据（真实索引使用gnews_valid_data+第一次写入数据时的日期）；
- publish_niuer（alias）: 从gnews的mysql中导入的数据（真实索引使用publish_niuer+第一次写入数据时的日期）；
- 每个新闻类型使用一个type；

2. 逻辑

- 每个需求使用一个独立的模块；

3. 字段

- 原始数据中的字段名要与最终使用的字段名一致；
- 需要计算后产生的字段要有明确标注和计算方法说明；

4. Kafka Topic

- gnew_raw_data: 用于存放爬虫抓取的原始数据；
- 分区数不要小于40个；
- 保留天数为15天；

ES替代MongoDB需求分析

准备工作

- spider 需要同时向kafka 和mongoDB写所有数据
- 规范数据格式，确定数据中包含的字段及含义

第一阶段

A. Streaming

- 保存评论数据，要考虑去重机制
  - 使用独立的评论索引保存

- 重复数据去重：存入ES时通过程序生成的id（对url取md5）去重
- "toutiao*category*class","data*source*class"

B. Batch

- 补齐publishtime：使用当前时间补
- 计算data valid既计算数据的有效性（沈，已确定，此部分为导入mysql逻辑）
  - 不同类型的数据计算逻辑不同

- 提取文章的标签；
  - 调用其他部门写的算法；

- 新闻去重（此部分为导入mysql逻辑）
  - 增加独立字段用作重复标记
  - 重复的数据保留publishtime小的
  - 通过其他部门写的算法进行计算，回溯过去15天数据

- 图片后缀是webp将data value设置成7（目标webp已经不特殊处理，不标注为7）；
- 如果有多个视频，把第一个视频的属性设置为此数据的视频属性；
- 小说数据author+" "+title作为新的标题（此部分为导入mysql逻辑）；
- 小说、漫画、电商处理逻辑一致，正文、带html标签正文都默认设置为已经存在（此部分为导入mysql逻辑）；
- 给status字段设置值，默认为1（此部分为导入mysql逻辑）；
  - 对于图集、视频数据如果data valid为0则status设置为0（此部分为导入mysql逻辑）；

- tags标签要转为用,（逗号）连接的字符串（此部分为导入mysql逻辑）；
- 将所有空值转为空字符串（此部分为导入mysql逻辑）；
- 将视频时长的时间转为秒（此部分为导入mysql逻辑）；
- 将列表页url长度大于500只去问号前的部分（引力资讯本身没有使用）（此部分为导入mysql逻辑）；
- 另外运行一个batch来对历史数据进行去重处理，回溯过去180天数据；
- 将含有乱码的数据做特殊标记（此部分为导入mysql逻辑）；

C. ES To Mysql

- 将时间字符串转为时间戳
- 视频、图片中video*path/img*path 为空的不导入；
- 如果发布时间存在且大于抓取时间，将数据忽略；
- 视频长度存在且小于5s的忽略；
- 如果字段为空值(Null,none,"",NULL)，写入时不写此值；

- 文章类型的数据，正文字数小于150不导入（因为视频数据导致，有些数据无法区分类型）；
- 缩略图、文章图集、视频集，将python中的list json dump成字符串；
- 图集如果图片数量大于100，则只保留100个图，img*location*list_count设置为100；
- 将news mode 的字符串替换为数字；

D. 数据统计

- spider写入kafka的量，从spider log中获取；
- streaming从kafka消费的量，streaming写入mysql；
- ES中streaming写入的原始数据的量；
- ES中batch加工后的有效数据量；
- ES导入到mysql 中的数据量；
- ES中streaming写入的原始数据与mongoDB中数据量对比；
- mongo导入mysql的量与es导入msyql量的对比；

E. Mysql To ES

- 将已经对外的数据从mysql导回ES

待确定的内容列表

1. 需要确定导入mysql时的过滤规则；

其他

- 广告表（？）

第二阶段

- 图片处理（B）
- 开发ES的crud接口（F）
- 视频、图片资源大小统计（D）
- 提供sql查询工具

注意事项

- 数据分类：AD、新闻、图集、视频、小说、电商、漫画；
- 通过news_mode来区分；
- 每条数据中包含多个url：

  万世涛

  - source_url: 列表页url；

- url: 详情页url；
- response_url：详情页跳转后的url，不跳转与url一致；

季家震

- url： 抓取目标（抓取接口）=》；
- response url： 抓取目标返回的url（可以用浏览器打开查看的url，不一定都是手机端的）；
- 原 url： 新闻原始出处的url（列表页url）；
- 原 response url： 新闻原始出处返回的url；

技术结构划分

给爬虫提出运行基础数据，从抓回来的数据中提取 中间咱们自己的逻辑 给引力资讯 mysql的

搜索输入时的关联词推荐

搜索优化