# 采集kafka消息定义

# 算法部分

|  |
| --- |
| **中科软科技股份有限公司** |
| **2018年1月** |

## 关于本文档

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 主题 | 关于采集后发送kafka消息定义，提供算法分析部分定义 | | | | |
| 说明 | 阐述了采集信息发送kafka消息的详细定义说明，指导开发工作 | | | | |
| 适用对象 | 信息采集工程师，算法分析工程师 | | | | |
| 修订历史 | | | | | |
| 版本 | 章节 | 类型 | 日期 | 作者 | 说明 |
| v1.0 | A | C | 2018年1月27日 | 苏展，杨经天，魏伟保，陈宏江 |  |
|  |  |  | 2018-10-22 | 陈宏江 | 增加评论消息和抖音相关描述 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**说明：类型－创建（C）、修改（U）、删除（D）、增加（A）；**

## 1 消息内容及形式

### 消息定义

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | topic名称 | 对应媒体 | 备注 |
| 1 | sc\_kafka\_content | 文本消息 | ~~新闻 partition=3~~  ~~微信 partition=2（不工作了）~~  ~~微博 partition=1（已经不工作了）~~  ~~TB\_FLOOR\_TABLE partition=0 （和微博）~~ |
| 2 | sc\_kafka\_picture | 图片消息 |  |
| 3 | sc\_kafka\_audio | 音频消息 |  |
| 4 | sc\_kafka\_video | 视频消息 |  |
| 5 | sc\_kafka\_comment | 评论消息 | 新增确认 |

### 消息内容

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 字段名称 | 内容 | 备注 |
| 1 | blockRowkey | 表BLOCK\_TABLE的rowkey, 与媒体类型无关 | 同时也是媒体的原始url |
| 2 | dfsName | 媒体fastdfs地址，形如：  ”group13/M00/14/37/CgoeDVn0JtyAPZI9AAESetN4KWY895.png” | 文本消息此字段为空，  媒体的fastdfs地址 |
| 3 | infoRowkey | info表rowkey，图片或音视频发送所属的文章的info表rowkey。 形如微博：  ”12345678-3f3vn49df3” | 同时也是媒体所在网页的url |
| 4 | infoTableName | info表名称，图片或音视频发送所属文章的info表名。含：  INFO\_TABLE，  WECHAT\_INFO\_TABLE，  WEIBO\_INFO\_TABLE，  DOUYIN\_INFO\_TABLE，  DOUYIN\_COMMENT\_TABLE |  |
| 5 | spiderTime | 采集时间戳，（碎片采集时间，可能对应block表或info表的时间，不一定），  形如：”1507613334” |  |

### 消息形式

消息形式为json格式，如下：

{

“blockRowkey”: blockRowkey,

“dfsName”: dfsName,

“infoRowkey”: infoRowkey,

“infoTableName”: infoTableName，

“spiderTime”: spiderTime

}

### 消息中的key

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | info表名称 | 消息中key值 | 备注 |
| 1 | INFO\_TABLE | news或inner或outer | 新闻表发出的消息 |
| 2 | WECHAT\_INFO\_TABLE | wechat | 微信表发出的消息 |
| 3 | WEIBO\_INFO\_TABLE | weibo | 微博表发出的消息 |
| 4 | DOUYIN\_INFO\_TABLE | douyin | 抖音表发出的消息 |
| 5 | DOUYIN\_COMMENT\_TABLE | douyin\_c | 抖音评论表发出的消息 |

## 2 算法分析工程师备注

媒体原始网页获取方法（图片，音频，视频消息）：

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 发送消息相关info表名称 | 媒体原始网页获取方法 | 备注 |
| 1 | INFO\_TABLE | 所发消息的infoRowkey字段内容即为媒体原始网页地址 | block表rowkey同info表rowkey |
| 2 | WECHAT\_INFO\_TABLE | block表rowkey同info表rowkey |
| 3 | WEIBO\_INFO\_TABLE | BLOCK\_TABLE表中，blockRowkey指向的记录中origin字段指向的内容为媒体原始地址 | block表的rowkey是info表的wb\_weibo\_url字段 |
| 4 | DOUYIN\_INFO\_TABLE | 所发消息的infoRowkey字段内容即为媒体原始网页地址ID(视频还原地址需请求数据采集组接口) |  |
| 5 | DOUYIN\_COMMENT\_TABLE | -- |  |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_