## 爬虫需要抓取的字段信息

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **编号** | **APP名称** | **需求来源** | **数据采集表** | **优先级** | **数据标签** |
| A1 | 秒拍 | 版权云 | 用户数据 |  | 用户ID |
| A2 | 秒拍 | 版权云 | 用户数据 |  | 用户头像 |
| A3 | 秒拍 | 版权云 | 用户数据 |  | 用户昵称 |
| A4 | 秒拍 | 版权云 | 用户数据 |  | 用户关注数 |
| A5 | 秒拍 | 版权云 | 用户数据 |  | 用户粉丝数 |
| A6 | 秒拍 | 版权云 | 用户数据 |  | 用户作品数 |
| A7 | 秒拍 | 版权云 | 用户数据 |  | app包名 |
| B1 | 秒拍 | 版权云 | 作品数据 |  | 作品ID |
| B2 | 秒拍 | 版权云 | 作品数据 |  | 作品标题 |
| B3 | 秒拍 | 版权云 | 作品数据 |  | 作品链接 |
| B4 | 秒拍 | 版权云 | 作品数据 |  | 内容类型 |
| B5 | 秒拍 | 版权云 | 作品数据 |  | 作品发布者ID |
| B6 | 秒拍 | 版权云 | 作品数据 |  | 作品发布者昵称 |
| B7 | 秒拍 | 版权云 | 作品数据 |  | app包名 |
| B8 | 秒拍 | 版权云 | 作品数据 |  | 作品发布时间 |
| C1 | 秒拍 | 版权云 | 评论数据 |  | 评论作品ID |
| C2 | 秒拍 | 版权云 | 评论数据 |  | 评论用户ID |
| C3 | 秒拍 | 版权云 | 评论数据 |  | 评论用户昵称 |
| C4 | 秒拍 | 版权云 | 评论数据 |  | 评论内容 |
| C5 | 秒拍 | 版权云 | 评论数据 |  | 评论发布时间 |
| C6 | 秒拍 | 版权云 | 评论数据 |  | app包名 |

其中：

* 作品发布时间需要由爬虫确认，能否获取到。如果不能获取，则由大数据平台入库时，由插入时间替代。
* 需要爬虫明确字段是否都支持

## 爬虫抓取数据策略

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **需求** | **集合分类** | **集合名称** | **定义** | **item定义** | **item update 时效性** | **item in** |
| 传统 | 全量用户 | 全量用户集合 | 所有有作品的用户 | 秒拍APPA1-C6 | 初始3天为一个周期 |  |
| 重点关注 | 新增用户 | 新增用户集合 | 当天新增入库的用户 | 秒拍APPA1-C7 |  | 实时关注2天 |
| 粉丝暴增用户 | 粉丝暴增用户集合 | 在全量集合中通过两个周期数据对比，计算出用户粉丝上升UP。设置粉丝数上升超过设定量值时，则放入粉丝集合UP | 秒拍APPA1-C8 | 1.粉丝二次基数在1000以上 2.粉丝增绝对值：10000以上 3.粉丝增值百分比：30% 其中2、3是或的关系 | 实时关注2天 |
| 违规用户 | 违规用户集合 | 由版权云确认是违规的用户 | 秒拍APPA1-C9 | 没日更新，通过大数据的短视库中获取 |  |

其中：

* 符合以上条件的用户，即为优先用户。
* 关注数据的时效性。“初始3天为一个周期”和“实时关注2天”是预估的值，需要能够根据实际情况手动调整
* 数据生效时间T+1

## 数据存储

数据存储。主要分为：点播视频信息存储、评论信息存储、用户状态信息存储

注意，具体的存储结构需要大数据根据实际情况调整。

全局说明：

时间字段用： unix时间戳即：当前时间与 1970年1月1日（UTC/GMT的午夜）之间的时间差（以毫秒为单位测量）。Java：System.currentTimeMillis()

**AI审核状态和人工审核状态的更新操作，目前支持分钟级更新同步完成。**

### 3.1、点播（短视频）信息

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名 | 英文名称 | 解释 | 是否必填 | 爬虫获取 | 大数据补填 | 版权是否会更新 |
| 数据主键 | id | 由大数据平台维护，生成唯一标识 | 必填 |  | 是 |  |
| APP包名 | app\_package\_name |  | 必填 | 是 |  |  |
| app唯一标识 | app\_id | Appid。默认为空 | 非必填 |  |  | 是 |
| 用户昵称 | user\_name |  | 必填 | 是 |  |  |
| 用户ID | user\_id | 用户ID：标识视频属于哪个用户 | 必填 | 是 |  |  |
| 用户头像 | user\_dp | dp是采用的display photo缩写 | 必填 | 是 |  |  |
| 用户关注数 | user\_follow\_num |  | 必填 | 是 |  |  |
| 用户粉丝数 | user\_fans\_num |  | 必填 | 是 |  |  |
| 用户作品数 | user\_video\_num |  | 必填 | 是 |  |  |
| 作品发布者ID | release\_user\_id |  | 必填 | 是 |  |  |
| 作品发布者昵称 | release\_user\_name |  | 必填 | 是 |  |  |
| 视频标题 | video\_title |  | 必填 | 是 |  |  |
| 视频描述 | video\_describe |  | 默认为空 |  |  |  |
| 视频ID/房间ID | fid | 点播视频为视频id，直播视频为房间（频道）id | 必填 | 是 |  |  |
| 发布时间 | release\_time | 标识视频发布时间，要求倒序排列 | 必填 | 是 |  |  |
| 视频/文件链接 | video\_url |  | 必填 | 是 |  |  |
| 视频/文件链接状态 | video\_url\_status | 0:初始状态1：可用2：不可用默认0 | 必填 |  |  | 是 |
| 最近一次失效时间 | url\_failure time |  | 非必填 |  |  | 是 |
| 视频类型：点播/直播 | video\_type | 1:点播，默认值2:直播 | 必填 | 是 |  |  |
| 用户全局唯一标识 | uid | 跨APP的全局唯一标识。全局指的是互联网不良内容巡检平台。可采用app\_id+用户id | 非必填 |  |  | 是 |
| 租户id | tenant\_id |  | 非必填 |  |  | 是 |
| 入库时间 | insert\_time | 大数据入库时间，要求倒序排列 | 必填 |  | 是 |  |
| AI审核时间 | ai\_audit\_time | AI最近审核时间 | 非必填 |  |  | 是 |
| AI审核状态 | ai\_status | 0：未审核，1：AI已审核。默认值：0 | 必填 |  |  | 是 |
| 人工审核状态 | manual\_status | 0：未审核，1：人工审核未违规 2：人工审核违规。默认值：0 | 必填 |  |  | 是 |
| 用户是否优先 | user\_priority | 用户是否为优先用户：0：非优先用户1：优先用户默认：0 | 必填 |  | 是 |  |
| 用户是否白名单 | user\_pure | 用户是否是白名单用户， 0：非白名单用户1：白名单用户默认：0 | 必填 |  | 是 |  |
| AI审核版本 | ai\_version | AI审核版本 默认v0 | 必填 |  |  | 是 |
| 数据来源 | data\_sources | 数据来源：0：爬虫1：租户默认：0 | 必填 |  | 是 |  |

其中：

* 用户是否优先：符合新增用户、暴增用户、违规用户其中条件之一，即为优先状态
* 数据来源为0，即爬虫获取的数据，需要大数据统计用户信息，非爬虫来源的数据，目前不需要参与用户信息的统计。
* 人工审核状态和AI审核状态需要支持实时更新和查询
* 支持数据的insert和update操作API

**数据新增去重**

点播数据插入时，需要大数据根据app\_package\_name+fid去重。去重的原则是要更新最新获取的播放串更新。即更新：video\_url

**数据修改诉求**

* 当视频为我方会根据视频唯一标识，来修改数据的最app\_id、uid、tenant\_id、ai\_audit\_time、ai\_status、manual\_status、ai\_version、url\_failure time

更新数据时，条件可以使用id或app\_package\_name+fid

* 非爬虫数据来源的数据，允许批量和单条的数据插入insert

**数据获取**

**公共参数**

* size：可以根据size设置返回数据的条数
* taotal: 可以根据查询条件返回目前的总条数
* 支持分页查询

**查询条件**

支持字段的简单查询。支持按照时间字段的排序（倒序和正序）

### 3.2、点播（短视频）评论信息

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 英文名称 | 解释 | 是否必填 | 爬虫是否支持 | 大数据平台是否补充 |
| 短视频数据主键作为外键 | fk\_id | 对应短视频数据主键 | 必填 |  | 是 |
| APP包名 | app\_package\_name |  | 必填 | 是 |  |
| 评论作品ID | fid |  | 必填 | 是 |  |
| 评论用户ID | user\_id |  | 必填 | 是 |  |
| 评论用户昵称 | user\_name |  |  | 是 |  |
| 评论内容 | content |  |  | 是 |  |
| 评论发布时间 | release\_time |  | 必填 | 是 |  |
| 数据入库时间 | insert\_time |  | 必填 |  | 是 |

**数据修改诉求**

* 非爬虫数据来源的数据，允许批量和单条的数据插入insert

**数据获取**

**公共参数**

* size：可以根据size设置返回数据的条数
* taotal: 可以根据查询条件返回目前的总条数
* 支持分页查询

**查询条件**

支持字段的简单查询。支持按照时间字段的排序（倒序和正序）

### 3.3、用户信息

1、大数据平台根据点播（短视频）信息。统计用户信息并维护用户状态

2、版权需要同步白名单用户到用户信息。（设置白名单和取消白名单）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 字段名 | 英文名称 | 解释 | 是否必填 | 大数据计算补充 | 版权更新 |
| APP包名 | app\_package\_name |  | 必填 | 是 |  |
| app唯一标识 | app\_id | Appid。默认为空 | 非必填 |  | 是 |
| 用户全局唯一标识 | uid | 跨APP的全局唯一标识。全局指的是互联网不良内容巡检平台。可采用app\_id+用户id | 非必填 |  | 是 |
| 用户昵称 | user\_name |  | 必填 | 是 |  |
| 用户ID | user\_id | 用户ID：标识视频属于哪个用户 | 必填 | 是 |  |
| 用户头像 | user\_dp | dp是采用的display photo缩写 | 必填 | 是 |  |
| 用户关注数 | user\_follow\_num |  | 必填 | 是 |  |
| 用户粉丝数 | user\_fans\_num |  | 必填 | 是 |  |
| 用户作品数 | user\_video\_num |  | 必填 | 是 |  |
| 新增用户 | new\_user | 1:新增用户0：非新增用户默认0 | 必填 | 是 |  |
| 粉丝暴增用户 | increase\_user | 1:暴增用户0：非暴增用户默认0 | 必填 | 是 |  |
| 业务制定用户 | user\_status | 1:违规用户0：非违规用户默认0 | 必填 | 是 |  |
| 白名单用户 | user\_pure | 用户是否是白名单用户， 0：非白名单用户1：白名单用户默认：0 | 必填 |  | 是 |
| 用户入库时间 | insert\_time |  | 必填 | 是 |  |
| 用户状态更改时间 | update\_time |  | 必填 | 是 | 是 |

**其中：**

业务制定用户中的违规用户算法：即在统计周期内用户的点播（短视频）中存在“人工审核后状态为违规”的用户

白名单用户：支持版权系统的用户白名单同步。要支持白名单的设定和取消功能。

**数据修改诉求**

* 当视频为我方会根据视频唯一标识，来修改数据的最app\_id、uid、user\_pure、update\_time

更新数据时，条件可以使用uid或app\_package\_name+ user\_id

* 非爬虫数据来源的数据，允许批量和单条的数据插入insert

**数据获取**

**公共参数**

* size：可以根据size设置返回数据的条数
* taotal: 可以根据查询条件返回目前的总条数
* 支持分页查询

**查询条件**

支持字段的简单查询。支持按照时间字段的排序（倒序和正序）