数据采集工作汇报

###### 汇报范围：*2020.11-2021.10* 汇报时间：*2021.10.15* 汇报对象：*研发部 / 产品部*

|  |
| --- |
| **部门：**研发部 **中心：**MCN  **姓名：**李浩 **职位：**数据采集工程师 |
| **近一年工作总结 (临时工作和版本迭代不算在列)** |
| **关键项目：**   1. 采集管理系统。 - 目前支持 (抖音/快手/哔站/红书/微博) 5个平台基础数据和作品数据采集。 2. 数据分析面板。 - 目前支持 (账号/自媒体主/广告主/任务/IM) 5个指标的数据分析和展示。 3. 逆向破解手机。 - 手机Android系统的Root破解并进行App逆向，目前只支持抖音。 4. 手机模拟运行。 - 进行Auto.js的开发，目前只支持抖音的账号刷机。 5. 破解红书滑块。 - 红书大幅改版只能采集小程序数据，进行了滑块突破，目前通过率50%。 6. 采集星图数据。 - 完成抖音星图平台的数据采集，目前筛选匹配准确率85%以上。 7. 采集红人数据。 - 完成Weiq平台全量数据采集，当前采集工作已停，   粉丝画像数据仍作为平台补充数据使用。   1. 资源管理工具。 - 对接资源管理后台进行工具支持的开发，   目前支持 (审核/作品/快速/账号) 的数据异步采集。 |
| **重点工作：**   1. 绕过微信验证码。 - 目前微信基础数据采集已停。 2. 绕过快手滑块。 - 目前快手需要登录采集，一个账号刷新采集，效率慢。 3. 破解哔站播放数。 - 播放数需要登录，已解决。 4. 解决图片跨域。 - 开发实时下载保存图片功能，之后已由前端解决，已弃用。 5. 解决微播易打码。 - 采集过微播易平台数据，目前已弃用。 6. 反推抖音账号。 - 抖音链接不全，利用App反推出真实账号，找回账号2100个。 7. 采集知乎MCN。 - 采集过知乎MCN平台数据，目前已弃用。 8. 采集统计图表。 - 开发采集后台实时统计图表，由于版本升级已弃用。 9. 增加作品数据。 - 平台开始不支持作品数据采集，目前已支持。 10. 增加平台指数。 - 平台互动指数和粉丝指数的算法开发。 |
| **核心数据：**   1. 采集平台对应 (正式/预发布/测试/开发) 4个环境共享数据16W，还在持续增加。 2. 星图匹配平台数据1W8， 占比85%以上，还在持续增加。 3. 原Weiq粉丝画像数据10W，已停止增加。 4. 数据库服务器1台4U8G，辅助服务器1台2U4G，采集服务器7台1U1G，域名1个。 5. 每月服务器支出约1000元左右，上下浮动不等，每年域名支出约130元。 6. 测试手机1部，采集星图数据笔记本一台。 |
| **问题难点：**   1. 目前没有使用代理资源，海外代理较贵，都是按流量收取费用。 2. 采集效率受资源制约，随着账号数量增加，服务器成本需增加。 3. 购买海外服务器制度和流程欠佳，受制约。 4. 手机逆向刷机采集数据受资源严重制约，成本过高。 5. 手机逆向工程利用程序去采集数据难度过高，受技术和人员制约。 6. 手机App逆向除抖音和微博以外，其他平台粉丝数量展示有限，逆向意义不大。 7. 快手采集受账号封禁制约，采集通过率和效率极低。 8. 星图数据采集受技术制约，只能本地Windows系统采集。 9. 红书采集目前受官方改版影响，技术条件制约，采集策略提高，账号数量增大，采集效率低下。 |
| **明年预计计划** |
| **重点工作：**   1. 各平台手机App逆向工作继续推进。 2. 各平台粉丝画像数据继续推进。 3. 大数据相关语音和语义分析的相关开发。 4. 数据面板的不断开发和完善。 5. 其他由产品分配的相关工作开发。 |