### 爬虫技术概要

**采集源**: 快看 腾讯漫画

**模拟登录**：普通用户 VIP

**技术概要**: 分三大线程流水作业

线程1 爬取总分类大纲漫画列表 分类列表逐一抓取并记录

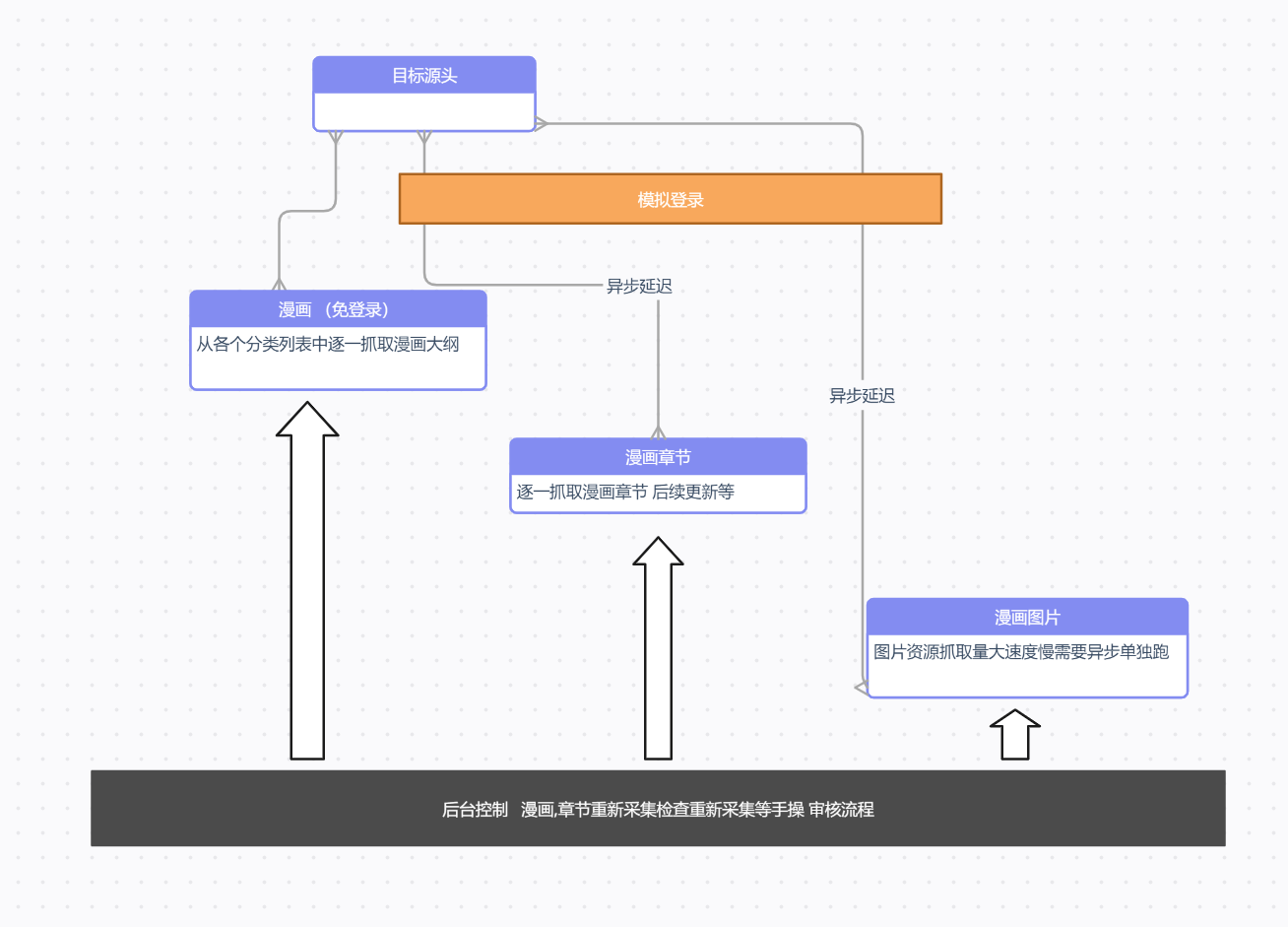
线程2 爬取漫画详情页内容 包括 id url 名称 详细简介 章节列表...

线程3 爬取章节内容 主要为图片资源

**站点分析：**

1. 网站分类 以题材 区域 属性 进度 4个点做拆分
2. 漫画包含 标题 封面 详情 作者 人气 点赞 标签(腾讯)
3. 章节包含 标题 封面(快看) 点赞 发布时间

**流程图**

****

**技术细节**：

1. 爬虫可能会有反扒图片拦截或者js加载等可根据具体情况使用selenium模拟用户实操抓取资源
2. 爬虫采集点记录和防重复采集
3. 漫画后期更新可根据最新发布 连载中关键字 或者后台设置等 定点每日监测
4. 后期还需要一个可控的后台可观测资源数据问题等

**服务器方面**：需要ip可不定时更换以防止反扒监测 爬取图像资源会占用大量存储可提前准备oss云存储相关

网络问题最好可以选择国内云做爬取服务器或者架设专门vpn 国外云做资源存储服务器

本地开发也需要vpn

**付费问题：**漫画存在vip或付费 需要开通

**技术评估：**

1. 第一周 采集快看漫画样例
2. 第二周 抽样更新 控制后台
3. 第三周 腾讯源采集 后台数据审核流程,数据整理流程等
4. 第四周 多源多任务等性能优化 掉线监测等

**漫画站**

漫画站点图片资源量消耗最高

单张图800k-1.5M不等

章节平均 20-40页 16M-40M

一部漫画 平均 短篇40< 中长篇40-200 长篇500-800+

快看漫画总量 6000+ 腾讯5600+

漫画估测统计存量占用 40T