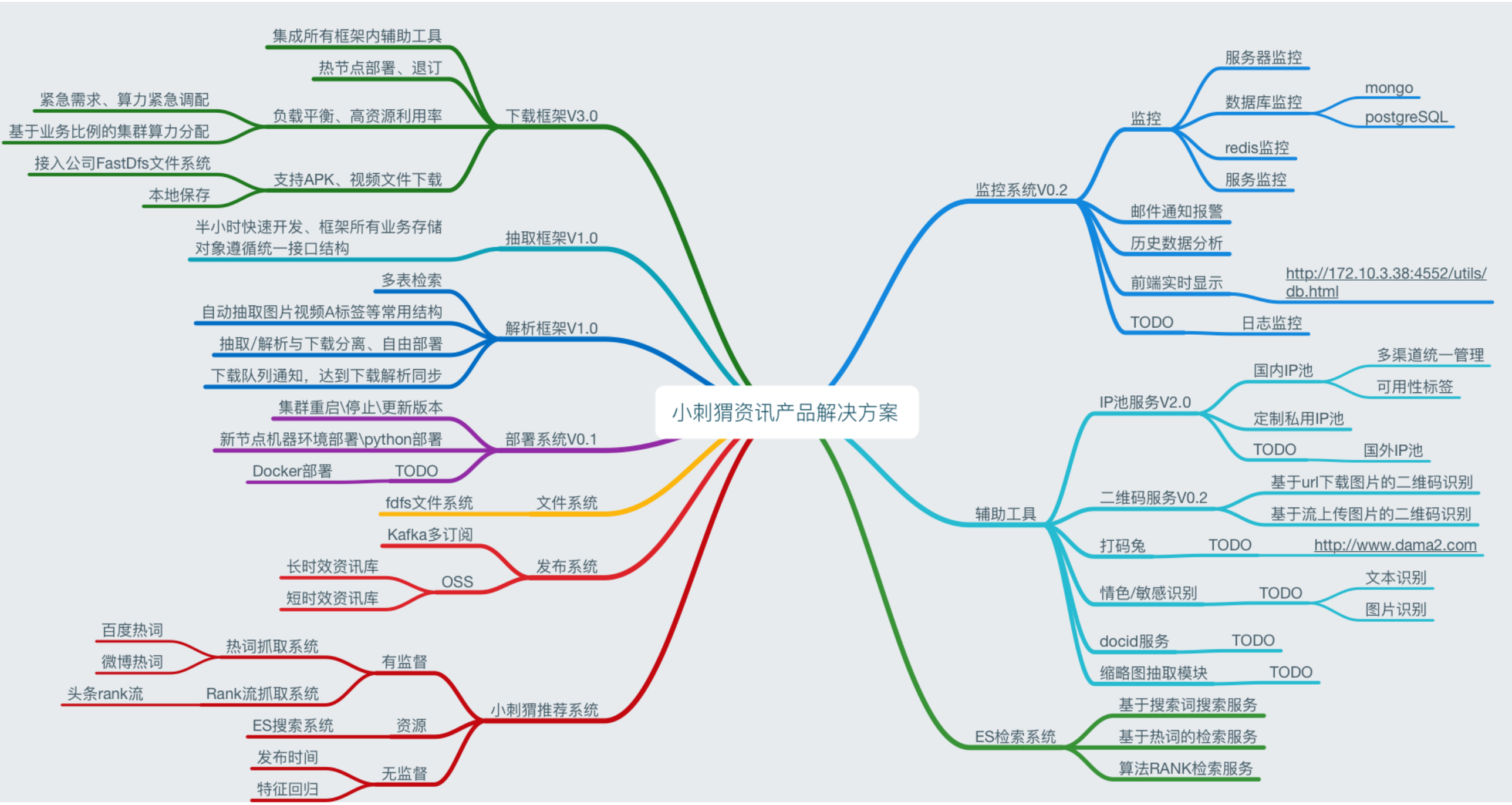
小刺猬爬虫框架

–FOR BEYEBE

历史过往



千万级别的抓取落地业务

阶段1：验证性阶段（搜狐）

即使大厂，对于业务之间的抓取复用依然很差。

- 浮躁
- 项目迭代快
- 爬虫复杂度高难以抽象共性
- 爬虫人员普遍技术栈浅,黑科技多

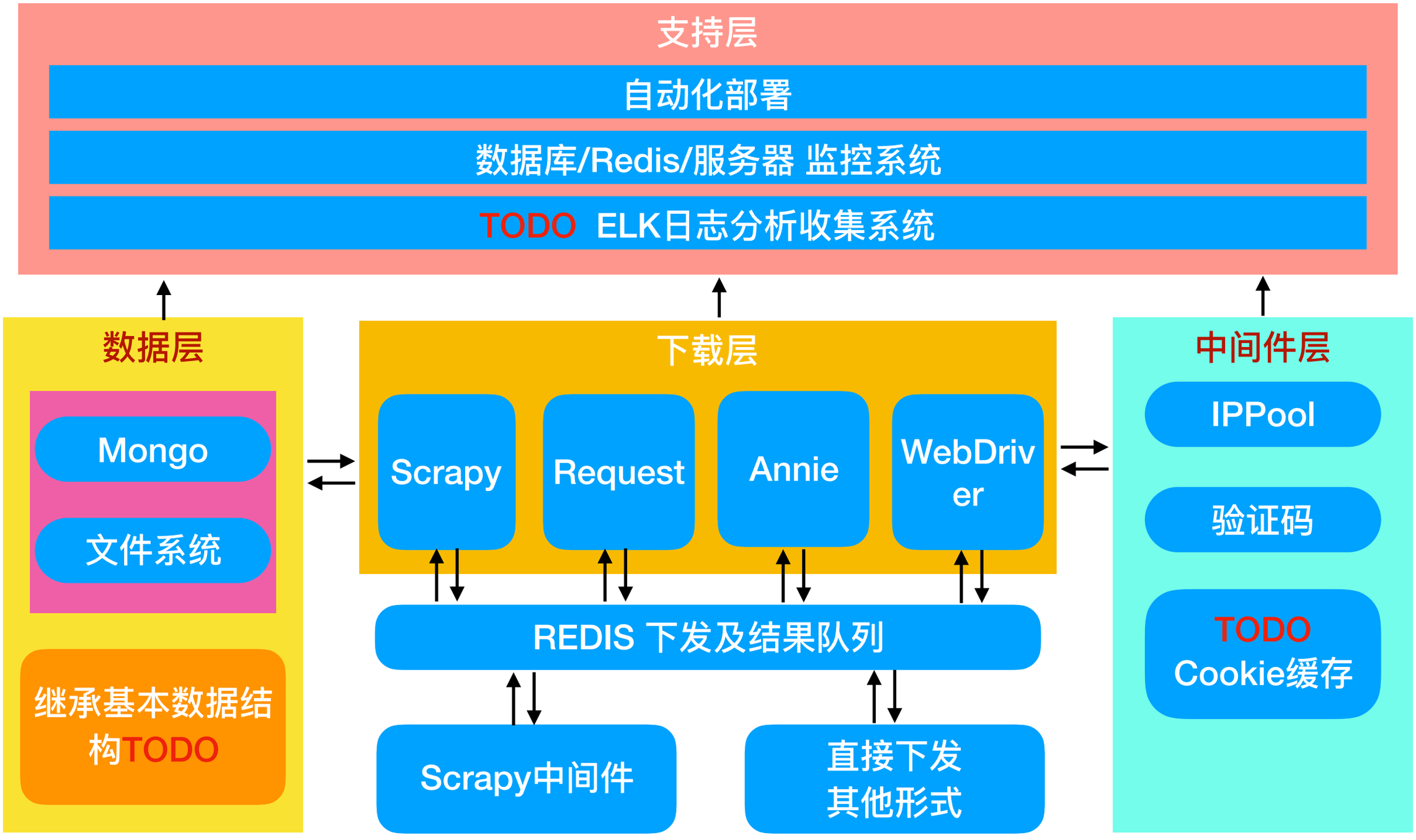
阶段2：开发阶段（喜文及全民点游）

完成落地，新渠道平均2小时开发完毕
支持千万级别数据抓取\大文件抓取

阶段3：开源阶段（比一比）

完善框架支持层
优化底层效率
规范化

抓取架构


支持层
自动化部署
数据库/Redis/服务器 监控系统
TODO ELK日志分析收集系统
数据层
Mongo
文件系统
继承基本数据结构TODO
下载层
Scrapy
Request
Annie
WebDriver
REDIS 下发及结果队列
Scrapy中间件
直接下发
其他形式
中间件层
IPPool
验证码
TODO
Cookie缓存

使命

将小刺猬爬虫框架做成开源框架级别的项目

资源

1、一些重要但不紧急的业务（尤其是既有业务优化）

2、能够跟业务匹配的硬件资源,系统集群冗余30%

3、前期需要一个可以熟练使用Python的同事一起兼做

节奏

1、首先完成抓取架构中的TODO，再优化NLP相关。

2、抓取架构稳定、产生效益后个人即申请转正。