**HITSpider分布式云爬虫平台开发**

**任务计划表**

**开发模式：**总体规划、分布实施、敏捷开发、快速迭代

**开发语言：**Java

|  |  |  |
| --- | --- | --- |
| **序号** | **起止时间** | **任务内容** |
| **1** | 2017.2.6-  2017.3.10 | （1）云爬虫平台的总体结构设计  （2）爬虫分布式调度系统的设计与框架开发 |
| **2** | 2017.3.11-  2017.5.31 | 完成分布式爬虫平台v0.5版  （1）master端系统：包括worker集群管理模块、爬行project执行管理模块  （2）worker端系统：包括基础爬行组件、基础爬行限定模块、基础爬行策略组件、基础信息抽取组件、页面存储模块  （3）前端管理系统：XML配置，API调用、无前端界面 |
| **3** | 2017.6.1-  2017.7.15 | 完成分布式爬虫平台v1.0内测版  （1）master端系统：增加project的自动负载均衡（重点是反反爬算法的设计与实现）、project断点管理和崩溃恢复功能、project状态的实时监控功能等  （2）worker端系统：增加动态页面的爬行组件、SDF爬行策略组件、BBS爬行策略、智能信息抽取组件  （3）前端管理系统：完成完整的Web前端管理界面 |
| **4** | 2017.7.15-  2017.7.31 | 完成分布式爬虫平台v1.0外测版发布、优化 |
| **5** | 2017.8.16-  2017.8.31 | 完成v2.0版本的详细功能设计：重点支持微博、贴吧、微信公众号等爬取功能，增加众包计费功能 |
| **6** | 2017.9.1-  2017.12.31 | 完成分布式爬虫平台v2.0版 |