## 附件一 分布式网络数据采集器需求规格说明

(Version 1.0)

1. 数据采集终端的相关功能

数据采集终端的主要功能包含以下几部分。

1. 网页操作功能
2. 加载指定URL
3. 在输入框中填写指定值
4. 在下拉菜单中选择某一项
5. 勾选复选框或者单选框
6. 提交指定表单
7. 休眠等待指定时间
8. 页面滚动条滚动
9. 点击页面上的任意元素
10. 鼠标悬浮到指定元素之上
11. 循环操作
12. 判断操作
13. 数据采集功能
14. 获取指定元素的innerHTML
15. 获取指定元素的outerHTML
16. 获取元素的内容文本（ContentText）
17. 获取当前网址的URL
18. 获取指定图片的URL
19. 获取指定超链接的URL
20. 数据清洗功能
21. 字符串替换
22. 正则匹配
23. 去除数据中的空格
24. 添加前缀
25. 添加后缀
26. 日期格式化
27. 转码
28. 其他功能
29. 日志功能：可以将系统运行时数据采集终端发生的以上1-3部分的任意事件全部进行记录到日志，并且进行本地保存，当主控机发起日志获取命令的时候，自动将日志信息发送给主控机。
30. 数据传输功能：在执行数据采集后，数据都会在数据采集终端进行本地存储，当主控机需要这些数据的时候，会向数据抓取终端发出获取数据命令，终端收到命令后将本地抓取到的数据传输到主控机端。
31. 数据清单功能：在执行数据采集后，数据都会在数据采集终端进行本地存储，当主控机需要这些数据的时候，需要首先知道要获取那些数据，主控机向数据抓取终端发出“获取数据清单命令”后，数据采集终端会把本地的所有数据整理成列表的形式发送给主控机。
32. 任务队列：本地任务会以一个队列的形式进行整理，所有主控机派发的任务都会加入任务队列，进入队列的任务会依次被执行。任务执行允许设置超时时间，也允许设置同时并发数，设置并发数大于1之后，即可在数据抓取终端中同时执行多个任务。同时当主控机发出管理任务操作（增删改查）任务命令时，本地会对应执行对应的响应。
33. 异常处理：异常分为“普通异常”和“致命异常”。普通异常指不会影响后续操作的异常，如：普通的获取数据失败。这类异常发生后对后续操作无影响，因此可以根据预先命令中的描述来选择继续执行还是其他操作。致命异常指当前异常会导致后续操作无法执行的异常，如：点击某个按钮获取下一页内容时按钮点击错误。因为无法点击下一页按钮，可能造成获取不到下一页的数据。因此，致命异常发生后会直接终止当前任务的执行。所有异常的对应解决方案都必须在任务描述中正确的描述完整。并且所有的异常都会在日志中进行详细记载。
34. 心跳机制：为证明当前终端是在线的，所有数据抓取终端都会与主控机保持长连接，并且每隔一定时间发送心跳包来描述自己在线状态。
35. 终端用户系统：终端若想连接到主控机，必须通过用户系统登录到主控机，登录成功后，开始准备就绪并等待主控机分配任务。
36. 主控机的相关功能
37. 终端管理相关功能
38. 用户系统：主控机和管理员终端都需要通过这套完整的用户系统来登入到主控机，并且主控机可以根据登入的用户来区别各终端的身份，以及根据这个用户对终端进行管理。同一用户有多终端登录的能力，在多终端同时登录时，服务器可以根据机器码来区别各个终端。
39. 任务管理：主控机可以对各个已登入的终端进行任务管理（增删改查），以及对当前各个终端进行任务监视。
40. 数据管理功能
41. 数据管理：主控机可以对任意终端进行数据回收、远程管理数据。对回收进来的数据，主控机可以对其解析并以指定要求将其存储到数据库中的指定库、指定表中。并且主控机可以执行SQL语句对数据库中以存储的数据进行管理。
42. 日志管理功能
43. 日志管理：主控机可以对任意终端进行日志回收、远程管理日志。对回收进来的日志，主控机可以本地存储等管理操作（增删改查）。也可以进行导出操作。
44. 管理终端的相关功能
45. 用户系统：管理终端若想对各个数据采集终端进行管理，则必须通过管理终端登入管理员用户账号到主控机中。
46. 终端管理：管理终端登入管理员账号之后，可以间接的通过主控机的中转实现对数据采集终端的管理，管理的内容包括任务的管理（增删改查），数据的管理（增删改查），日志的管理（增删改查）。
47. 任务添加：管理终端可以编写任务并分发到任意数据采集终端。
48. 数据下载：管理终端可以随时从主控机中导出任意数据。
49. 任务相关说明

任务可以设置执行周期、执行时间、过期时间以及每次任务的超时时间。并且每个任务在任何时候都有自己的状态信息，主控机以及各个终端机随时可以根据任务的状态信息来明确当前任务执行的情况。

1. 程序说明

数据采集终端允许安装到任意操作系统。并且支持自动更新功能。当有新的版本时，终端能自动下载新的版本并更新。