# 普林云爬虫架构设计

目录

[普林云爬虫架构设计 1](#_Toc446404338)

[1 平台特点 3](#_Toc446404339)

[1.1 分布式 3](#_Toc446404340)

[1.2 反爬技术 3](#_Toc446404341)

[1.3 可视化管理监控 3](#_Toc446404342)

[1.4 自然语言处理 3](#_Toc446404343)

[2 系统架构 3](#_Toc446404344)

[2.1 设计目标 3](#_Toc446404345)

[2.2 设计图 3](#_Toc446404346)

[2.3 对象名词 3](#_Toc446404347)

[2.3.1 Job 3](#_Toc446404348)

[2.3.2 Job Group 4](#_Toc446404349)

[2.3.3 生成器 4](#_Toc446404350)

[2.3.4 下载器 4](#_Toc446404351)

[2.3.5 解析器 4](#_Toc446404352)

[3 采集层 4](#_Toc446404353)

[3.1 设计目标 4](#_Toc446404354)

[3.2 主要对象 5](#_Toc446404355)

[3.3 设计图 5](#_Toc446404356)

[3.4 流程图 5](#_Toc446404357)

[4 结构化层 5](#_Toc446404358)

[4.1 设计目标 5](#_Toc446404359)

[4.2 主要对象 5](#_Toc446404360)

[4.3 设计图 5](#_Toc446404361)

[4.4 流程图 5](#_Toc446404362)

[5 存储层 5](#_Toc446404363)

[5.1 设计目标 5](#_Toc446404364)

[5.2 主要对象 5](#_Toc446404365)

[5.3 设计图 5](#_Toc446404366)

[5.4 流程图 5](#_Toc446404367)

[6 系统层 6](#_Toc446404368)

[6.1 设计目标 6](#_Toc446404369)

[6.2 主要对象 6](#_Toc446404370)

[6.3 设计图 6](#_Toc446404371)

[6.4 流程图 6](#_Toc446404372)

# 平台特点

## 分布式

## 反爬技术

## 可视化管理监控

## 自然语言处理

# 系统架构

## 设计目标

* 分层化，减少耦合
* 错误跟踪，能够定位到具体层面
* 报警提示，可定制
* 完善的运行日志

## 设计图



## 对象名词

### Job

Job是核心对象。Job定义为一个数据源，比如法院破产网。包含如下内容：

1. 生成器
2. 生成器调度规则
3. 下载器
4. 下载器调度规则
5. 解析器
6. 解析器调度规则
7. 源数据存储
8. 解析后数据存储
9. 生成器日志
10. 下载器日志
11. 解析器日志
12. 错误日志
13. 报警提示
14. 运行日志
15. 父Job
16. 所属Job Group

### Job Group

定义为：Job的分组

### 生成器

定义为：能够生成URI，并输出到stdout的工具

### 下载器

定义为：根据URI，获得对方数据的工具。

### 解析器

定义为：专门解析下载器下载的数据，并转化为JSON格式输出的工具

### 调度规则

定义：类似Crontab的语法

### 防重器

定义：一种可以根据指定规则，来防止重复的工具。规则有：

1. 内容MD5
2. ID主键

# 采集层

## 设计目标

* 获得源数据，并提供给结构化层进行下一步的处理
* 能够支持拓展到1000台+机器
* 支持主流的URI，比如HTTP、HTTPS。
* 可以扩展新的URI协议
* 具备如下的反爬虫模块：动态页面、验证码识别、代理自动切换
* 错误追踪
* 报警提示，阈值设置

## 主要对象

## 设计图

## 流程图

# 结构化层

## 设计目标

## 主要对象

## 设计图

## 流程图

# 存储层

## 设计目标

* 分业务、分片存储
* 容量可以扩展到P级
* 可水平扩展
* 便于检索

## 设计图



## 主要对象

### MongoDB

定义为：MongoDB的集群服务

### MySQL

定义为：MySQL的集群服务，主要用于存储配置类数据。

### NFS

定义为：类似本地磁盘的网络文件系统集群

### 小文件

定义：大小在(0, 100M)的二进制的文件

### 大文件

定义：大小在(100M, +)的二进制的文件

### 文件索引

定义：记录文件位置的对象。包含如下属性：

1. ID。全局的唯一主键
2. 存储介质。MongoDB或者是NFS
3. URI。访问路径
4. 添加时间
5. 文件字节数
6. 来源主机IP
7. 数据内容MD5值

## 流程图



# 系统层

## 设计目标

## 主要对象

## 设计图

## 流程图