# 普林云爬虫架构设计

目录

[普林云爬虫架构设计 1](#_Toc446400299)

[1 平台特点 2](#_Toc446400300)

[1.1 分布式 2](#_Toc446400301)

[1.2 反爬技术 2](#_Toc446400302)

[1.3 可视化管理监控 2](#_Toc446400303)

[1.4 自然语言处理 2](#_Toc446400304)

[2 系统架构 2](#_Toc446400305)

[2.1 设计目标 2](#_Toc446400306)

[2.2 设计图 2](#_Toc446400307)

[2.3 对象名词 2](#_Toc446400308)

[3 采集层 2](#_Toc446400309)

[3.1 设计目标 2](#_Toc446400310)

[3.2 主要对象 3](#_Toc446400311)

[3.3 设计图 3](#_Toc446400312)

[3.4 流程图 3](#_Toc446400313)

[4 结构化层 3](#_Toc446400314)

[4.1 设计目标 3](#_Toc446400315)

[4.2 主要对象 3](#_Toc446400316)

[4.3 设计图 3](#_Toc446400317)

[4.4 流程图 3](#_Toc446400318)

[5 存储层 3](#_Toc446400319)

[5.1 设计目标 3](#_Toc446400320)

[5.2 主要对象 3](#_Toc446400321)

[5.3 设计图 3](#_Toc446400322)

[5.4 流程图 3](#_Toc446400323)

[6 系统层 4](#_Toc446400324)

[6.1 设计目标 4](#_Toc446400325)

[6.2 主要对象 4](#_Toc446400326)

[6.3 设计图 4](#_Toc446400327)

[6.4 流程图 4](#_Toc446400328)

# 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

## 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：

* URI生成器
* 下载器
* 结构化配置规则信息
* 存储信息
* 父Job

# 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

## 主要对象

## 设计图

## 流程图

# 系统层

## 设计目标

## 主要对象

## 设计图

## 流程图