**可配置化网络爬虫项目需求分析说明书**

**小组名称：**可配置化网络平台小组

**小组口号：**我们爱学习

**指导教师：**陈博老师

**文档撰写人：**吴祥龙 关玉婷

**文档撰写时间：**2017年6月20日

团队分工记录表

|  |  |  |  |
| --- | --- | --- | --- |
| **项目名称** | **学号** | **姓名** | **分工** |
| 可配置化网络爬虫 | SA17225405 | 吴祥龙 | 爬虫框架需求分析与报告撰写 |
| SA17225094 | 关玉婷 | 运行需求分析与报告撰写 |
| SA17225362 | 王润 | 功能需求分析与报告撰写 |
| SA17225317 | 唐冬鑫 | 功能需求分析与报告撰写 |

目录

[1.系统概述 1](#_Toc1170)

[2.分布式爬虫框架的需求 1](#_Toc31734)

[2.1 需求概述 1](#_Toc2448)

[2.2 运行环境的需求 2](#_Toc14236)

[2.2.1 硬件环境的要求 2](#_Toc23708)

[2.2.2 软件环境的支持 2](#_Toc1311)

[2.3 功能需求 2](#_Toc5696)

[2.3.1. 对爬虫集群资源的管理与监控 2](#_Toc3924)

[2.3.2 动态在集群中增加新的爬虫任务 2](#_Toc31075)

[2.3.3 对运行在集群中爬虫的监控 2](#_Toc27460)

[2.3.4 动态的加减一个任务的节点 2](#_Toc13735)

[2.3.5 提供访问丰富的集群访问方式 2](#_Toc9694)

[3 本章小结 2](#_Toc6342)

# 

# 1.系统概述

# 本系统将开发爬虫的通用的操作封装了起来，例如html页面的下载，dom树的解析数据的存储等，同时还为用户屏蔽了开发分布式爬虫程序的很多细节，例如任务的同步，任务的调度，资源的管理等。

# 2.分布式爬虫框架的需求

## 2.1 需求概述

本文的主要目标是基于java设计并实现一个分布式的爬虫系统，这个爬虫框架能够自动管理爬虫集群的资源，能够在该系统内运行多种不同类型的爬虫任务，任务之间互不影响。该系统还能监控整个集群的所有爬虫任务，方便爬虫的开发与维护。

## 2.2 运行环境的需求

### 2.2.1 硬件环境的要求

本文实现的分布式爬虫框架需要的硬件如下

安装了ssh工具的客户端，主要用来远程连接服务器，进行提交任务、监控任务等操作一个具有存储能力的计算机集群，用来运行爬虫任务。

### 2.2.2 软件环境的支持

本文实现的分布式的爬虫框架需要一些基础的软件支持，客户端主要用来连接服务端执行一系列的控制命令，服务端是本系统实现的基本支撑，以下是软甲的需求

客户端：Xshell，securecrt，等ssh远程连接的工具

服务端:Mysql+JDK+ 13assandra+redis+zookeeper

## 2.3 功能需求

### 2.3.1. 对爬虫集群资源的管理与监控

爬虫系统要能对整个集群的资源进行管理。对集群的资源进行管理，是分布式爬虫系统实现的基础。避免了多个爬虫任务在同一个机器上运行造成的资源浪费的现象。方便管理员对整个爬虫集群的管理。

### 2.3.2 动态在集群中增加新的爬虫任务

爬虫系统要实现动态的在爬虫集群中增加新的爬虫任务的功能。管理员可以动态的在集群中启动新的爬虫任务，这样极大的节省了爬虫开发和维护的成本，节省时间。

### 2.3.3 对运行在集群中爬虫的监控

爬虫系统要能监控整个集群中运行的爬虫。对集群中运行的爬虫监控的功能能帮助管理员迅速定位有问题的爬虫，并快速发现问题出现的原因，极大的方便了爬虫的维护

### 2.3.4 动态的加减一个任务的节点

爬虫系统要能够对一个爬虫任务进行动态的调控，能够动态的增加或减小这个爬虫任务正在运行的节点，只要集群的机器资源足够，能够为一个爬虫任务增加大量的节点，来提高爬虫的运行效率。

### 2.3.5 提供访问丰富的集群访问方式

爬虫系统能够很方便的进行访问。本系统是一个分布式的系统，满足多语言对爬虫集群系统的访问非常重要，支持多语言的访问可以很方便的与其他系统进行集成，提高了系统的扩展

## 3 本章小结

本章对本片文章要实现的分布式爬虫的需求做了一个详细的分析。本章首先对整体的需求做了一个概述，然后分析了软件的运行环境的需求，最后分析了系统的功能需求。