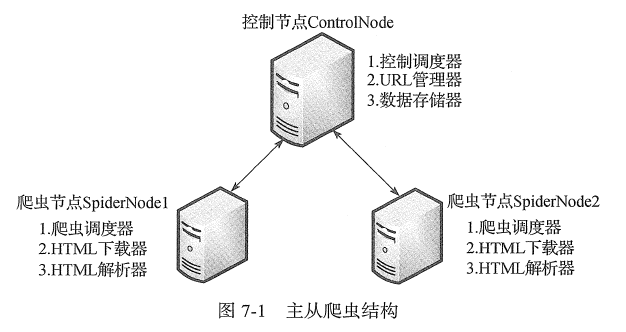
# 分布式爬虫

三台主机进行分布式爬取，一台主机作为控制节点，另外两台作为爬虫节点

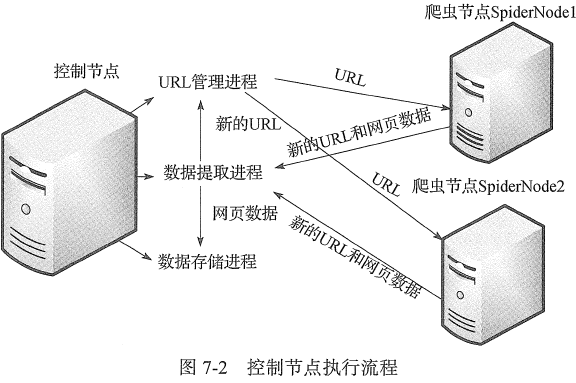
爬虫结构如图所示



## 控制节点

ControlNode主要分为**URL管理器，数据存储器和控制调度器**。控制调度器通过三个进程来协调URL管理器和数据存储器的工作：一个是**URL管理进程**，负责URL的管理和将URL传递给爬虫节点；一个是**数据提取进程**，负责读取爬虫节点返回的数据，将返回数据中的URL交给URL管理进程，将标题和摘要等数据交给数据存储进程；最后一个是**数据存储进程**，负责将数据提取进程中提交的数据进行本地存储。

执行流程如下图



### URL管理器

采用set内存去重的方式，字符串经过MD5算法处理（Python中算法生成的是256位，取中间的128位，将未爬取和已爬取的URL集合序列化到本地，保存当前的进度）

### 数据存储器

生成的文件按照当前时间进行命名，以避免重复，同时对文件进行缓存写入

### 控制调度器

主要是产生并启动URL管理进程、数据提取进程和数据存储进程，同时维护4个队列保持进程间的通信，分别为url\_queue、result\_queue、conn\_q、store\_q。

url\_q队列是URL管理进程将URL传递给爬虫节点的通道

result\_q队列是爬虫节点将数据返回给数据提取进程的通道

conn\_q队列是数据提取进程将新的URL数据提交给URL管理进程的通道

store\_q队列是数据提取进程将获取到的数据交给数据存储进程的通道