scrapy_redis概念作用和流程

学习目标

1. 了解 分布式的概念及特点
2. 了解 scarpy_redis的概念
3. 了解 scrapy_redis的作用
4. 了解 scrapy_redis的工作流程

> 在前面scrapy框架中我们已经能够使用框架实现爬虫爬取网站数据,如果当前网站的数据比较庞大, 我们就需要使用分布式来更快的爬取数据

1. 分布式是什么

> 简单的说 分布式就是不同的节点（服务器，ip不同）共同完成一个任务

2. scrapy_redis的概念

scrapy_redis是scrapy框架的基于redis的分布式组件

3. scrapy_redis的作用

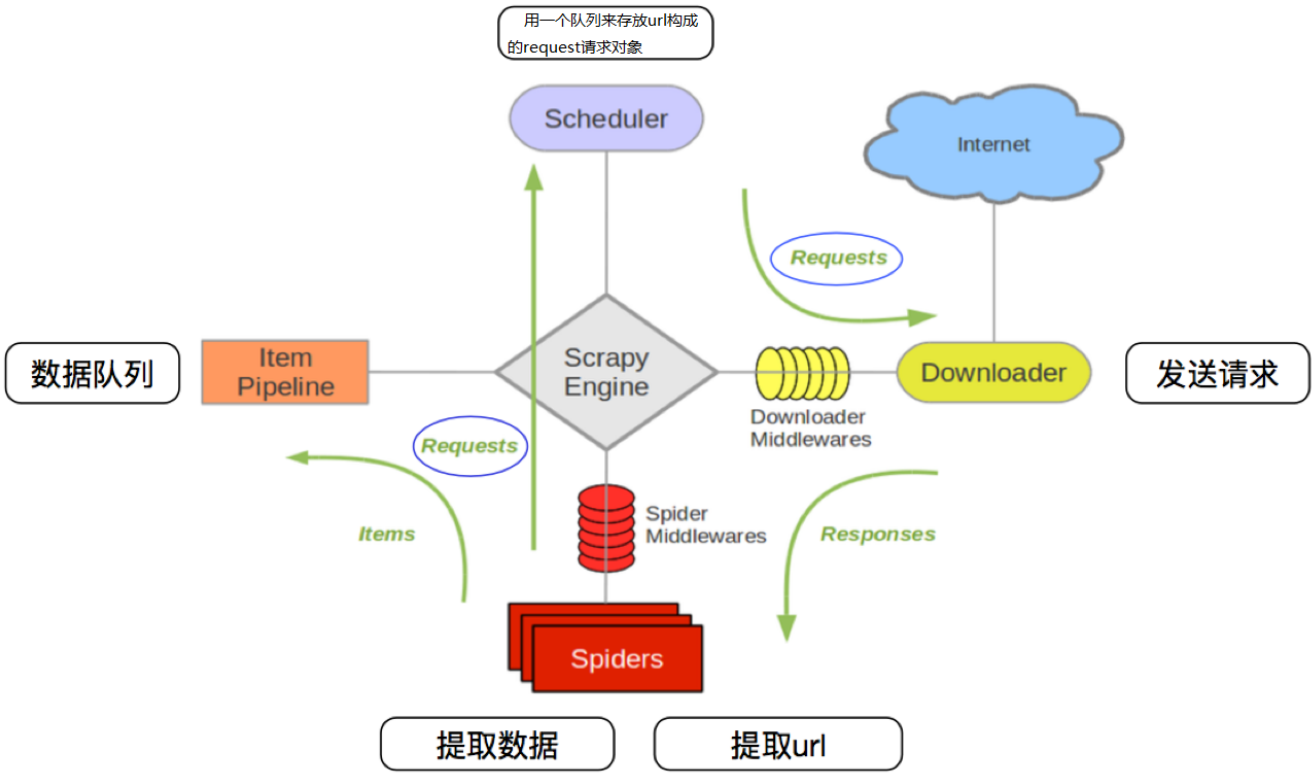
> Scrapy_redis在scrapy的基础上实现了更多，更强大的功能，具体体现在：

通过持久化请求队列和请求的指纹集合来实现：

- 断点续爬
- 分布式快速抓取

4. scrapy_redis的工作流程

4.1 回顾scrapy的流程

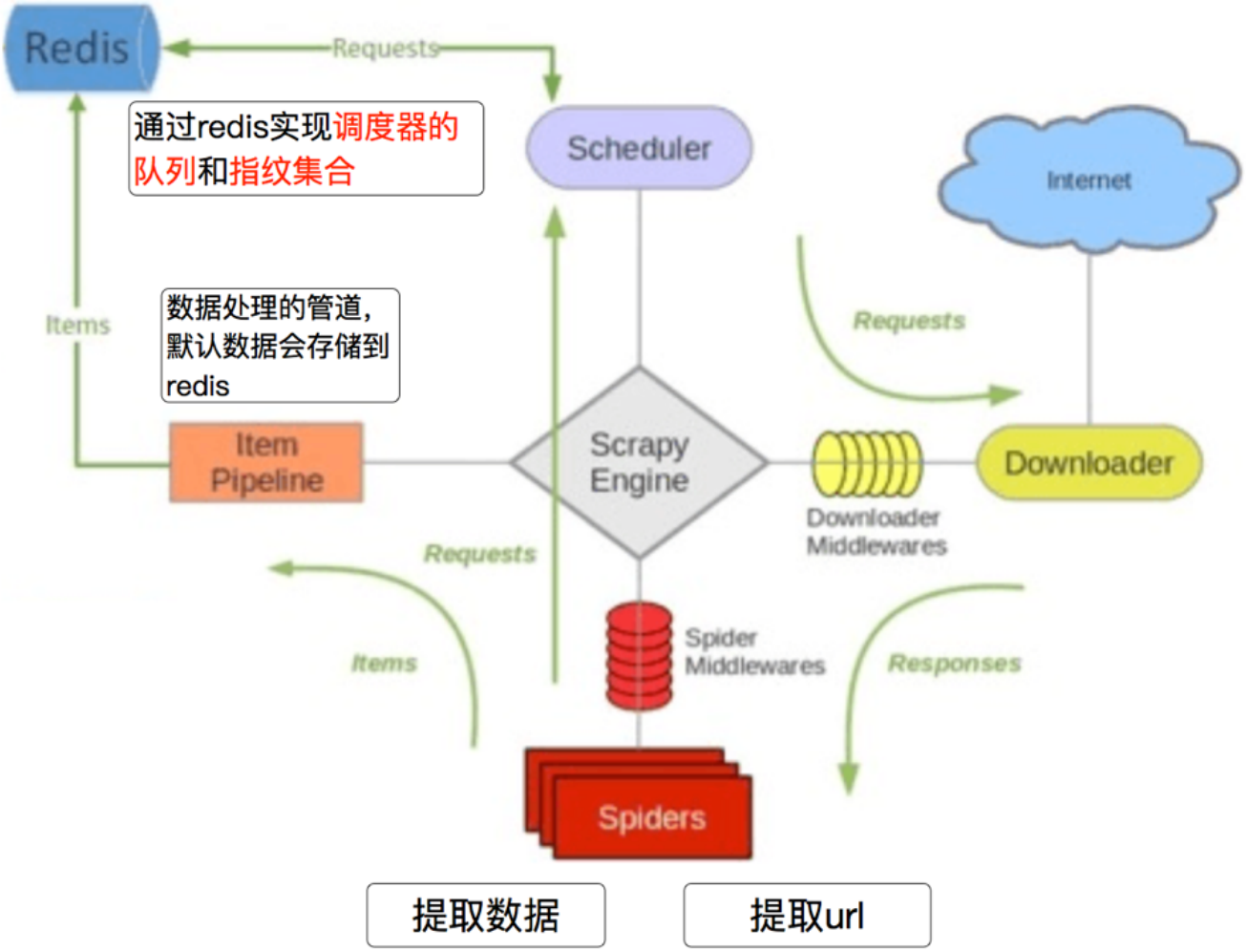


思考：那么，在这个基础上，如果需要实现分布式，即多台服务器同时完成一个爬虫，需要怎么做呢?

4.2 scrapy_redis的流程

- 在scrapy_redis中，所有的待抓取的request对象和去重的request对象指纹都存在所有的服务器公用的redis中
- 所有的服务器中的scrapy进程公用同一个redis中的request对象的队列
- 所有的request对象存入redis前，都会通过该redis中的request指纹集合进行判断，之前是否已经存入过
- 在默认情况下所有的数据会保存在redis中

具体流程如下：



小结

scarpy_redis的分布式工作原理

- 在scrapy_redis中，所有的待抓取的对象和去重的指纹都存在公用的redis中
- 所有的服务器公用同一redis中的请求对象的队列
- 所有的request对象存入redis前，都会通过请求对象的指纹进行判断，之前是否已经存入过