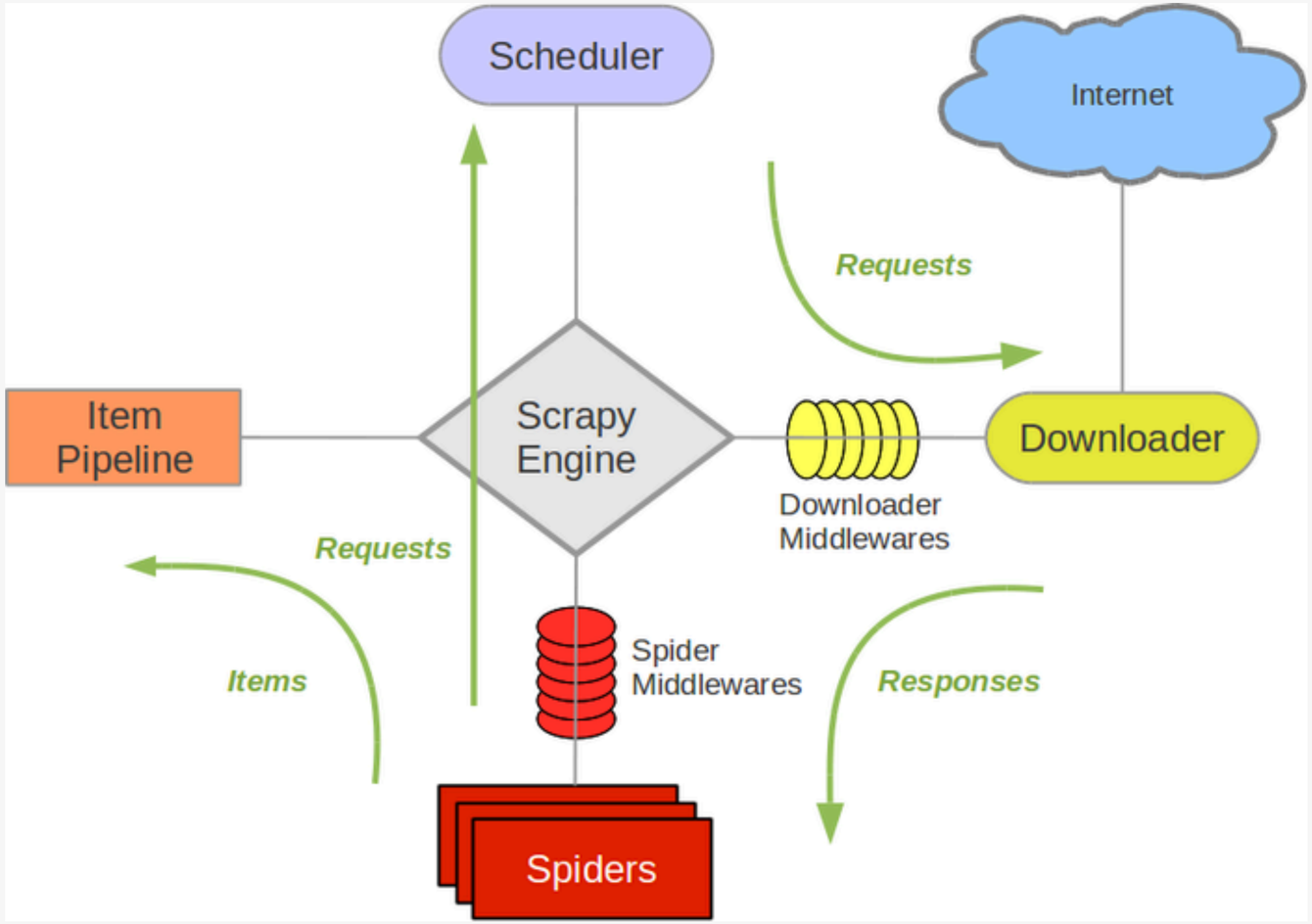
Scrapy分布式原理

Scrapy单机架构

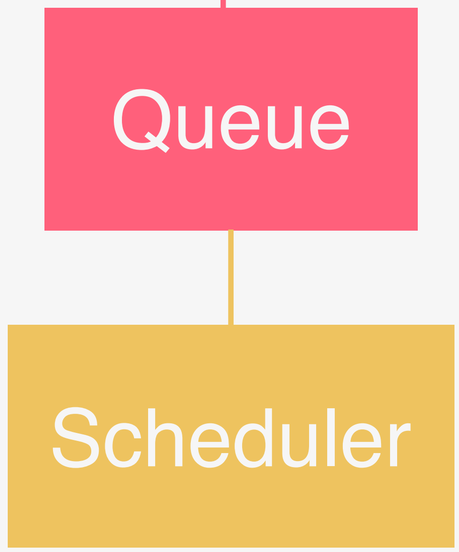


在本机维护一个爬取队列，Scheduler进行调度

多台主机协作的关键是什么？

共享爬取队列

单主机爬虫架构



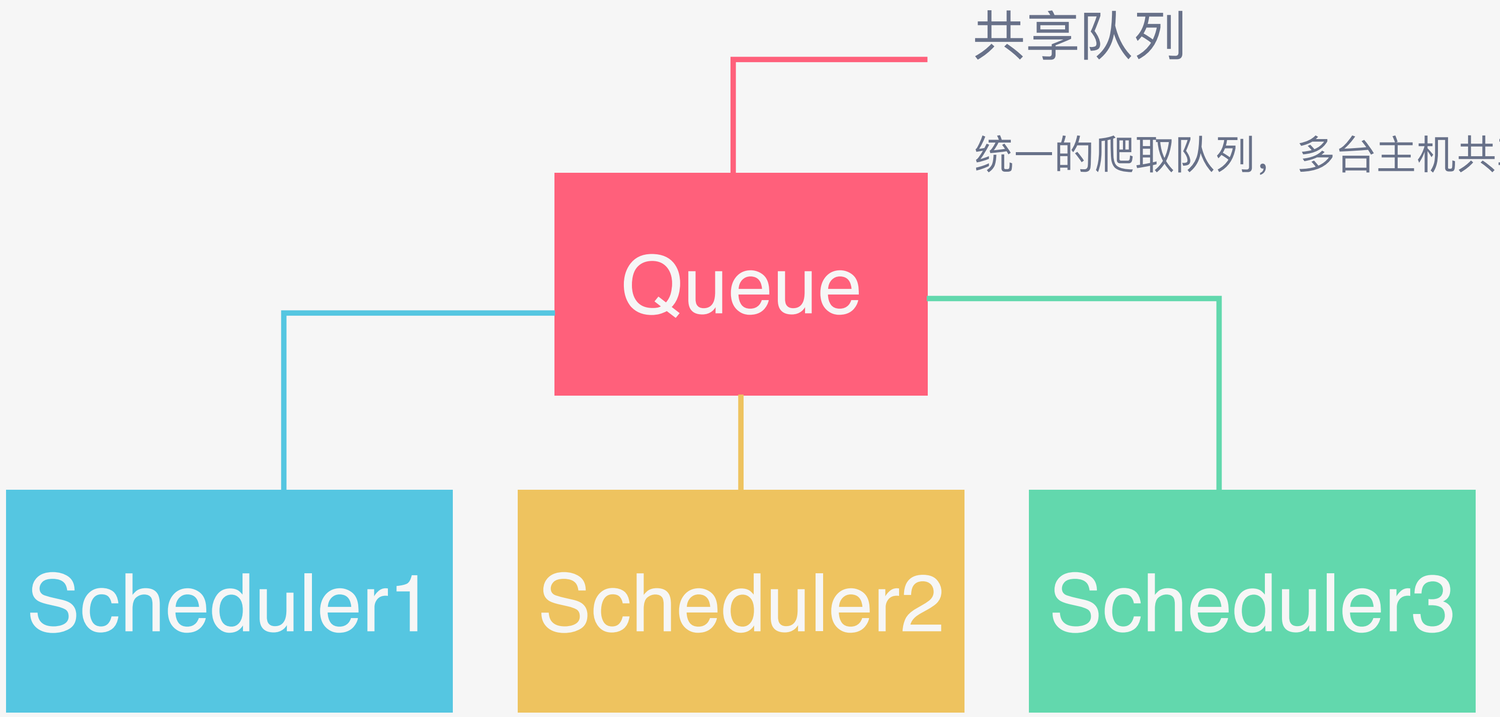
爬取队列

Requests队列，每台主机分别维护

调度器

负责从队列中调度Reqeusts进行爬取

分布式爬虫架构



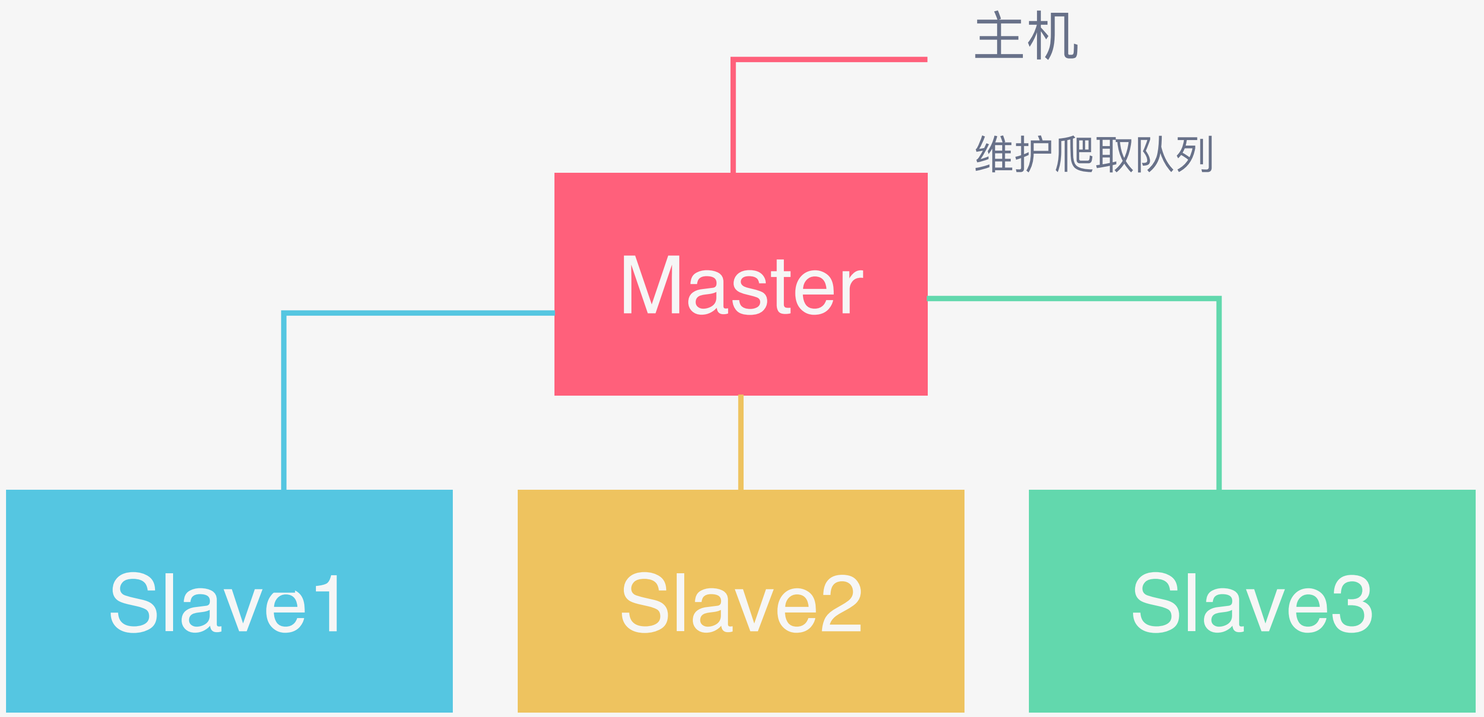
共享队列

统一的爬取队列，多台主机共享

调度器

各台主机的调度器统一从Queue调度

分布式爬虫架构



从机

负责数据抓取、数据处理、数据存储

队列用什么维护?

Redis队列

Redis，非关系型数据库，Key-Value形式存储，结构灵活。

是内存中的数据结构存储系统，处理速度快，性能好。

提供队列、集合等多种存储结构，方便队列维护。

怎样来去重?

Redis集合

Redis提供集合数据结构，在Redis集合中存储每个Request的指纹。

在向Request队列中加入Request前首先验证这个Request的指纹是否已经加入集合中。

如果已存在，则不添加Request到队列，如果不存在，则将Request添加入队列并将指纹加入集合。

怎样防止中断?

启动判断

在每台从机Scrapy启动时都会首先判断当前Redis Request队列是否为空。

如果不为空，则从队列中取得下一个Request执行爬取。

如果为空，则重新开始爬取，第一台从机执行爬取向队列中添加Request。

怎样实现该架构?

Scrapy-Redis

Scrapy-Redis库实现了如上架构，改写了Scrapy的调度器，队列等组件。利用它可以方便地实现Scrapy分布式架构。

https://github.com/rolando/scrapy-redis

源码分析

谢谢