

这个是Scrapy的整体框架结构。其中包含了因特网 Internet, 下载器Downloader, 爬虫Spiders, 数据处理Item Pipeline, 调度器Scheduler， Scrapy引擎。

Scrapy引擎是整个架构的核心，用于处理整个架构的数据流。

Downloader下载器，用于从WEB中下载网络页面资源，并且作为结果返回给网络爬虫，作为网络爬虫的数据源。Scrapy已经为下载网络资源做了许多后台的准备，我们一般不需要手动干预，只需要制定URL地址。

网络爬虫Spiders: 定制Web资源的解析规则，过滤Web资源,对Web资源进行数据切片。使用XPath技术。再交给项目管道。

项目管道Item Pipeline: 对数据进行保存到文件，进一步清洗数据。

调度器Scheduler: 调度触发Downloader。

三个中间件：Spider中间件，Downloader中间件，调度器中间件。