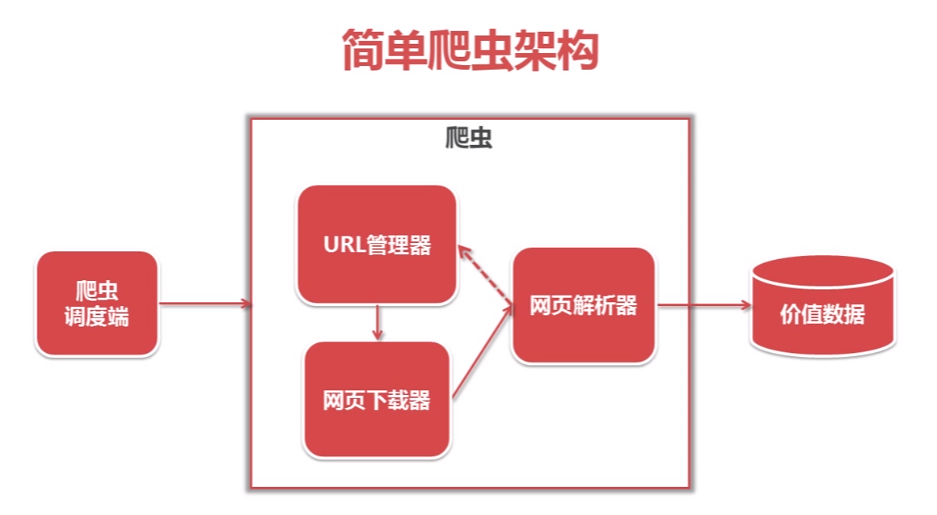
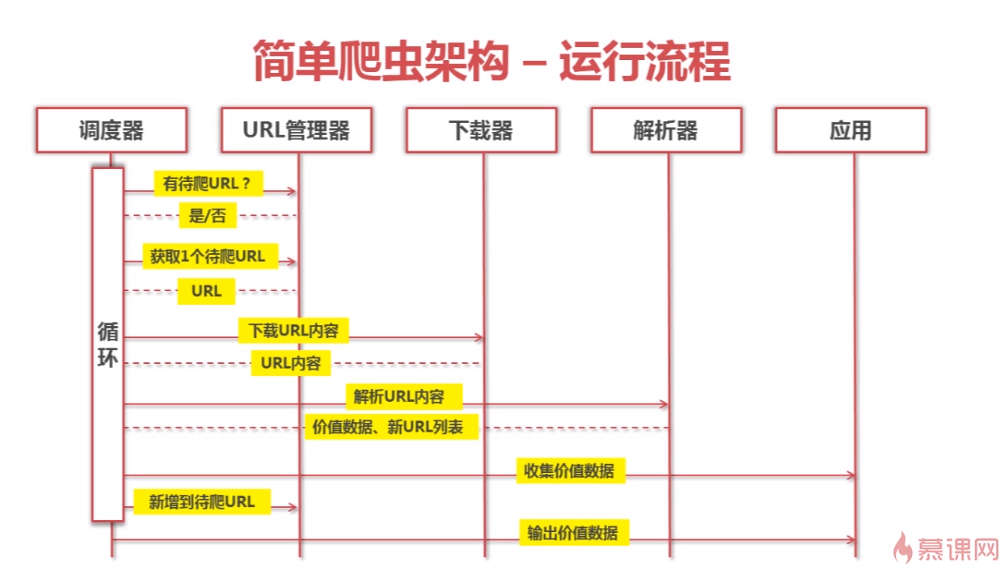
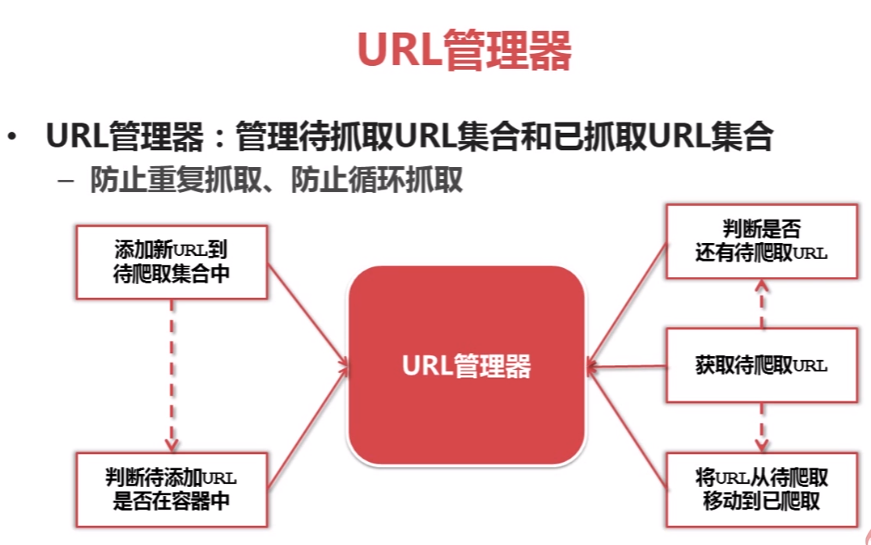
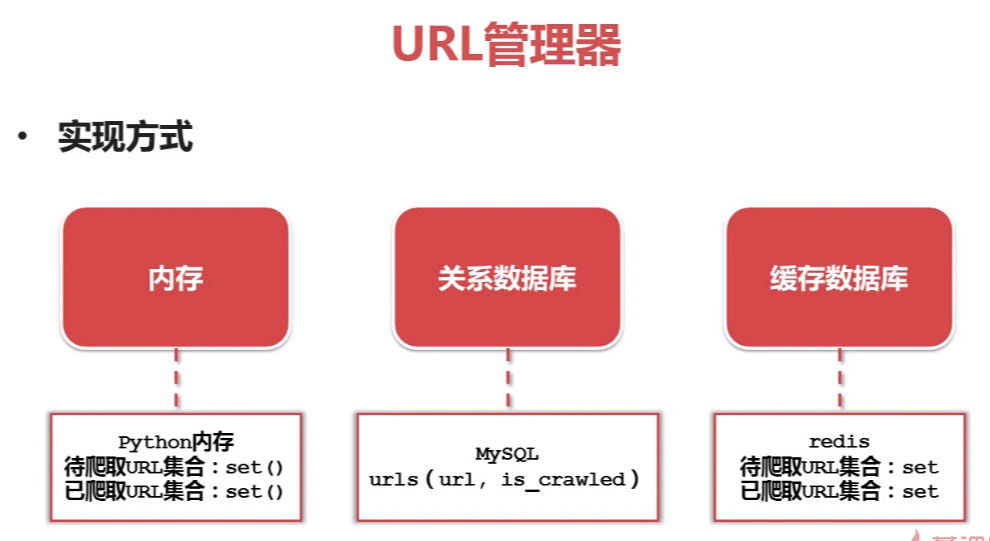
**Structure**





**URL Monitor**





内存：用于小型存储需求（个人）

缓存：用于企业存储需求

数据库： 用于永久性或者大数据量存储

**Webpage Downloader**



类似于网页浏览器



1. Basic

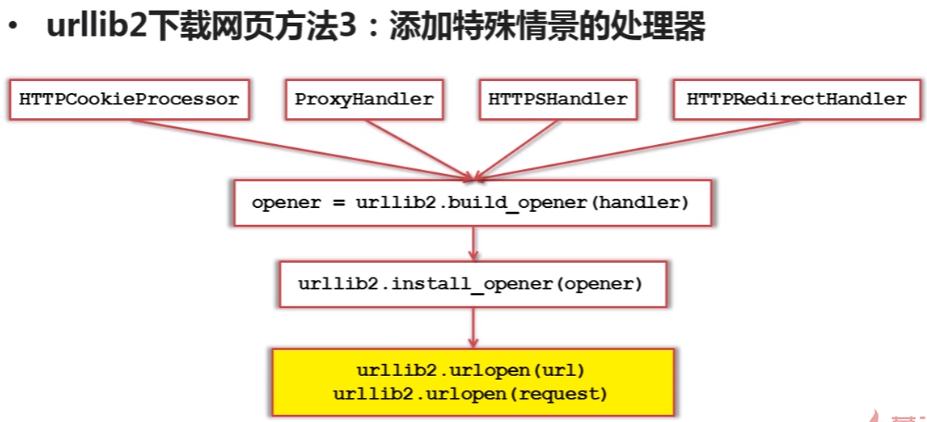


1. Camouflage



Self-define the url data and header. Disguise as Mozilla Browser here.

1. Special scenarios



有些网页需要用户登录才能处理，这里添加Cookie处理

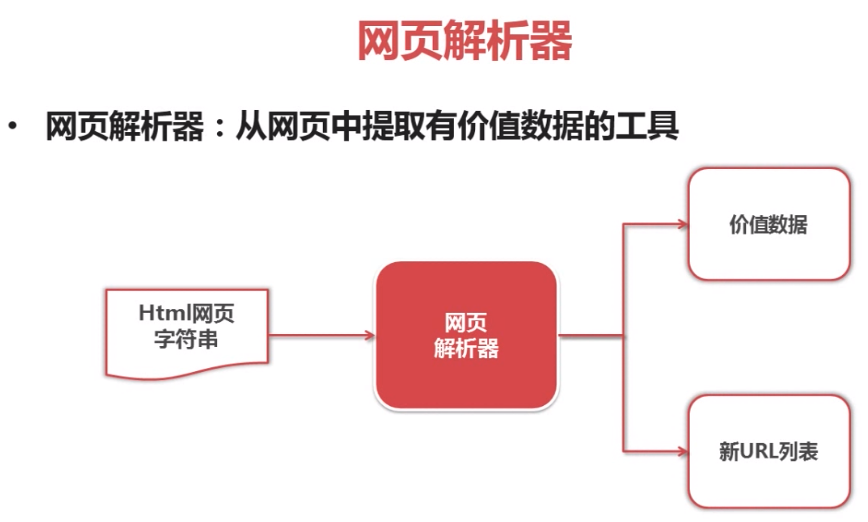
有些网页需要代理才能访问

有些网页使用https加密协议

有些网页url之间有相互自动的跳转关系

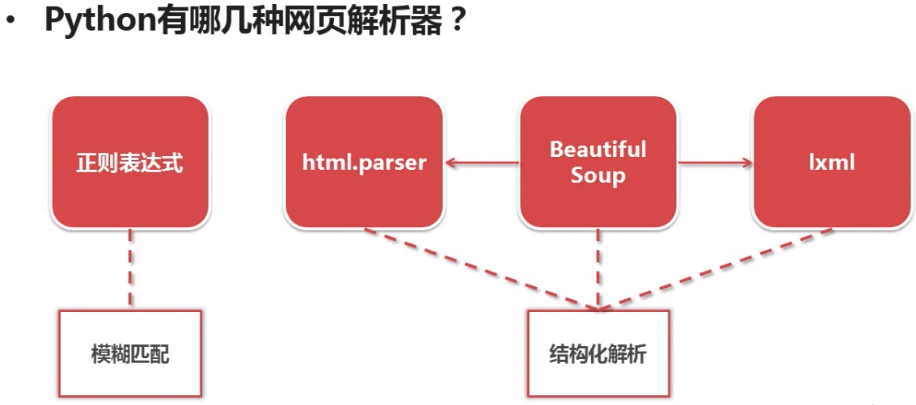


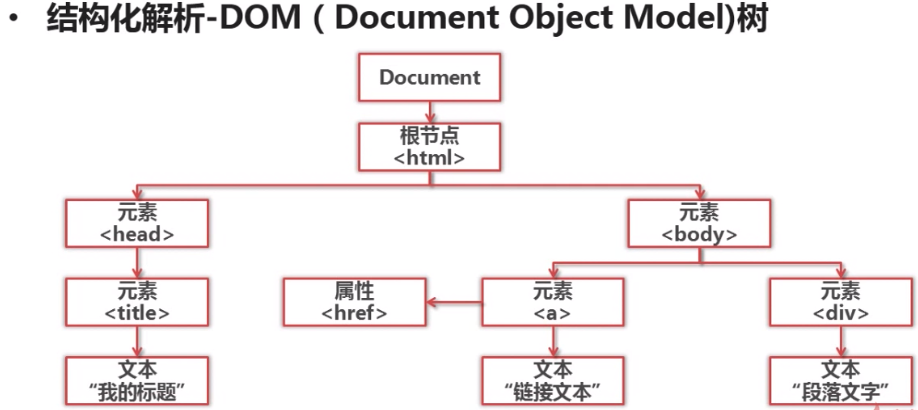
**Webpage Analyser**



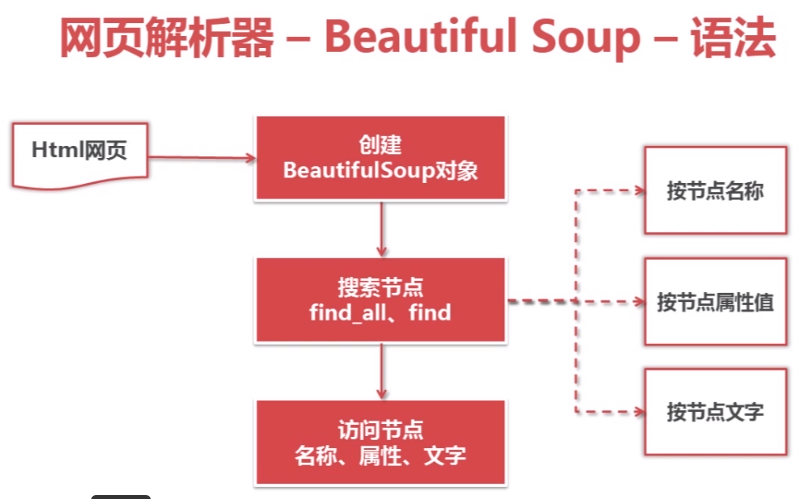
正则表达式：将整个网页当作一个字符串提取数据

不利于处理复杂文档

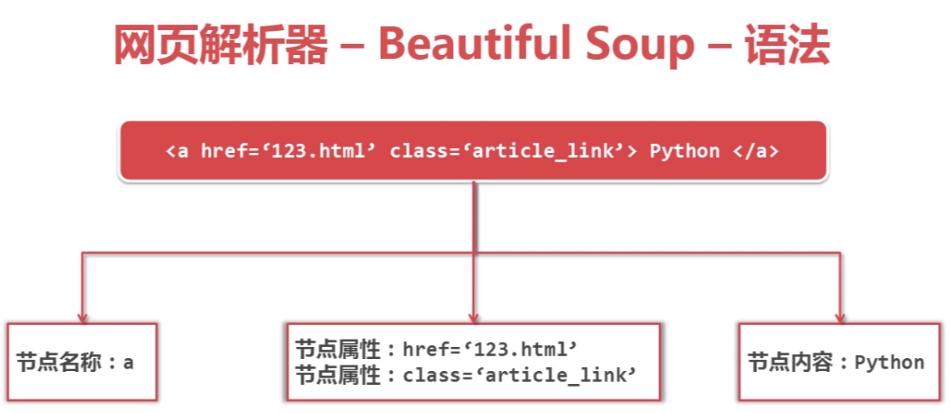




**Beautiful Soup**



搜索方法：







Attrs 中搜索时可以使用正则表达式

**Crawl**

