商务智能
Business Intelligence

数据采集与集成

朱卫平 博士
计算机学院
武汉大学

目录

数据采集

- 数据源
  - 内部数据：企业各种应用系统、办公自动化系统等产生的业务数据、文档等。
  - 外部数据：市场、竞争对手的数据以及各类外部统计数据等。
  - 业务数据需要经过数据评价、数据筛选以及ETL（数据抽取、转换、装载）后才可存储在数据仓库中。
- 数据仓库与数据集市
  - 数据仓库是面向主题的，其数据包括元数据和经过ETL的业务数据。数据集市是数据仓库的一个子集。

数据采集

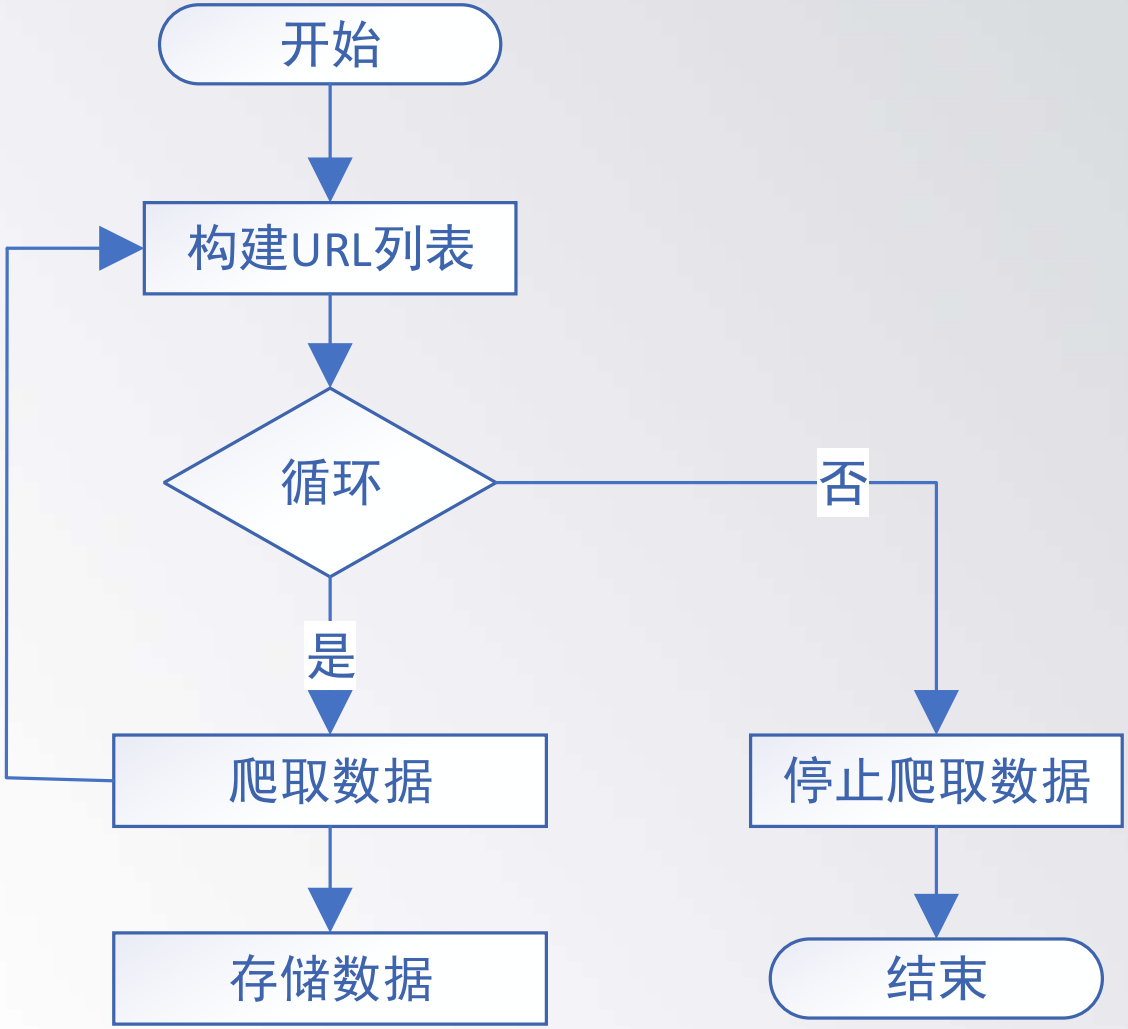
- 我们已经进入大数据的时代。
- 每天，有大量的（TB、PB数量级）的数据从商业、社会、科学、医药以及生活中的方方面面涌入我们的信息系统、万维网、以及各种数据存储设备。
- 这些爆炸性增长的、广泛可获取的、大量的数据使我们真正的处于数据时代。

你知道产生大数据那些具体例子？

网络爬虫

- 网络爬虫是什么?
  - 网络爬虫是一种按照一定的规则，自动地抓取因特网信息的程序。
  - 最初，它为搜索引擎从因特网上下载并保存网页，是搜索引擎的重要组成。
  - 传统爬虫从一个或若干初始网页的URL开始信息抓取，在抓取网页的过程中，不断从当前页面上（如新闻列表页）抽取新的URL放入队列并进一步抓取, 直到满足系统的停止条件。

网络爬虫工作原理

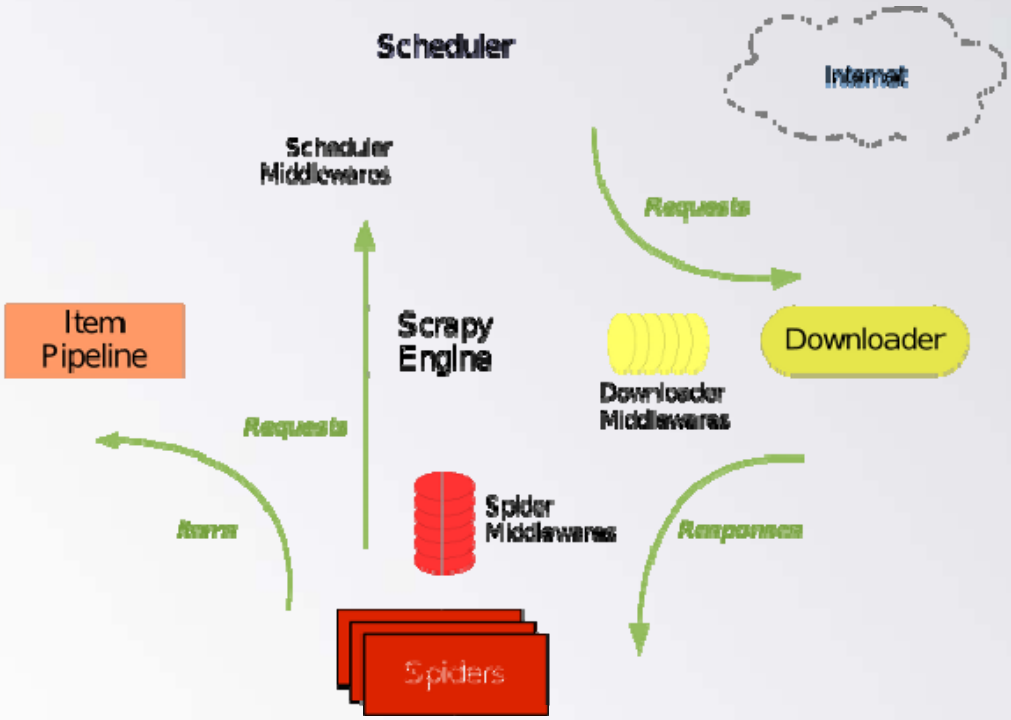


多页面爬虫流程

爬虫框架Scrapy

- Scrapy介绍
  - Scrapy一个开源的网络资源获取框架。其最初是为了页面抓取所设计的，使用它可以以快速、简单、可扩展的方式从网站中提取所需的数据。
  - 目前Scrapy的用途十分广泛，可用于如数据挖掘、监测和自动化测试等领域，也可以用于获取API所返回的数据或者通用网络爬虫。
  - Scrapy用纯Python实现，作为一个优秀的框架，用户只需要定制开发几个模块就可以轻松的实现一个爬虫，用来抓取网页内容以及各种结构化和非结构化信息。

爬虫框架Scrapy



绿线是数据流向，首先从初始 URL 开始，Scheduler 会将其交给 Downloader 进行下载，下载之后会交给 Spider 进行分析，Spider 分析出来的结果有两种：一种是需要进一步抓取的链接，例如 “下一页” 的链接，这些东西会被传回 Scheduler；另一种是需要保存的数据，它们则被送到 Item Pipeline 那里，那是对数据进行后期处理（详细分析、过滤、存储等）的地方。另外，在数据流动的通道里还可以安装各种中间件，进行必要的处理。