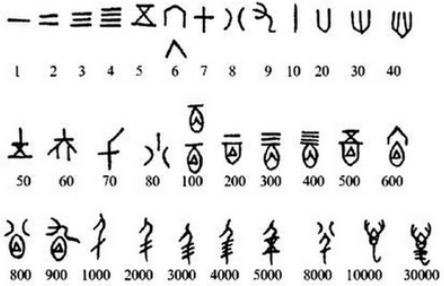
第二章 大数据感知与获取

肖江
Mail：jiangxiao@hust.edu.cn
Office: 东五楼 222 室

提纲

2.1 引言

古代埃及文字



古代中国数字

古印度文字

对数据的记录能力是原始社会与现今社会的重要分界标志

很早以前，人类就已开始收集“数据”

中国传统药学著作

神农本草经

上古，先秦，秦汉时期多位医家集结整理

上中下三卷，载药365种

明朝李时珍，历时27年 编纂，1590年出版

共52卷，载药1892种， 方剂11096个

美国海军上尉与他的大数据实践

- 马修·方丹·莫里（Matthew Fontaine Maury）
- 1806年出生于美国弗吉尼亚
- 1824年刚刚达到入伍年龄便进入了美国海军学校

- 1839年，已经晋升为海军上尉的莫里在一次事故中不幸腿部致残
- 不适合于服役远航的莫里在1842年被任命为主管海图和仪器库的负责人

莫里杂乱的库房

六分仪

经典的书籍、教材

指南针

莫里的目标

陈旧的图表

变废为宝

大量快发霉的航海日志

前人视为垃圾

莫里想要更多的数据

- 与船商交换信息，在自愿基础上互利互惠的合作方式锻造了国际气象界公开交换环境资料的传统

莫里

更新后的图表

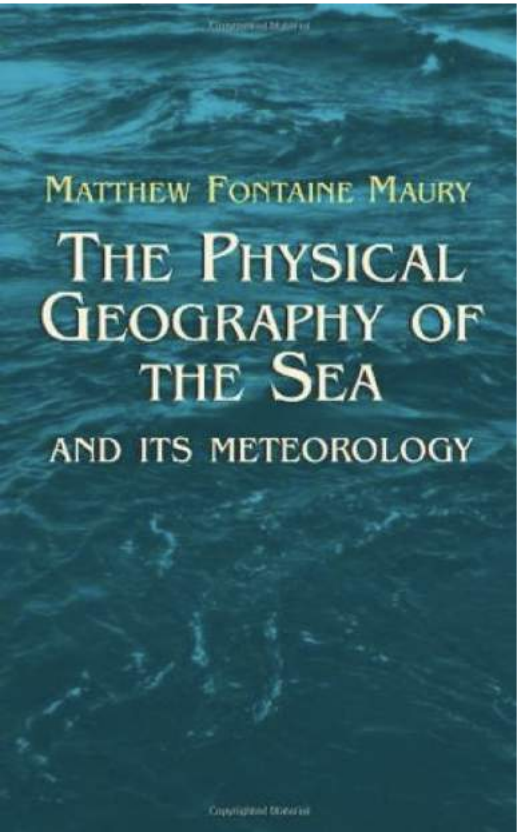
航海日志

船商

更多的数据！

- 让船商定期向海里扔掷标有日期、位置、风向以及当时洋流情况的瓶子
- 寻回瓶子，记录信息，更新数据

名垂千古：海洋学的奠基人



- 1855年,莫里出版权威著作《关于海洋的物理地理学》，被誉为海洋学的奠基人
- 当时，他已经绘制了120万数据点
- 四个国家授予了他爵士爵位,包括梵蒂冈在内的其他八个国家还颁给了他金牌奖章
- 即使到今天,美国海军颁布的导航图上仍然有他的名字

信息科技令人类收集数据的能力和渠道大幅提高延展，
创造了全新的生产力

智能设备——数据收集每时每刻

智能手机

地理位置数据
运动数据
环境亮度数据
图像数据
语音数据
…

…

数据

…

空气质量数据
温湿度数据
气压数据
…

智能家居设备

身体状况数据
运动习惯数据
实时图像数据
…

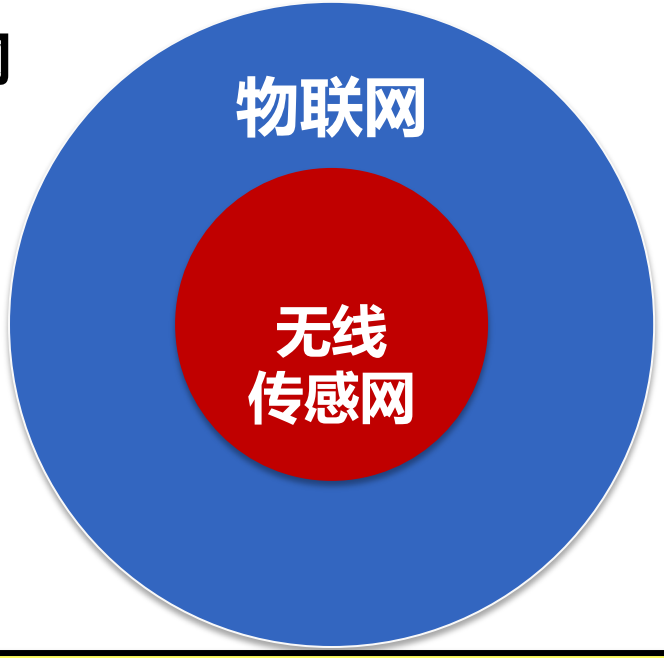
可穿戴设备

无线传感网 VS. 物联网

无线传感网≠物联网

无线传感器作为数据采集的入口，物联网的“心脏

无线传感器网络为物联网奠定传感和监控的技术基础



物联网>无线传感网

物联网不仅仅是传感器
(不仅需要感知，还要控制处理

物联网带来
更广泛的互联
更透彻的感知

物联网 —— 对物理世界实时控制、精确管理和科学决策

2.1 引言

- **海量数据的产生**

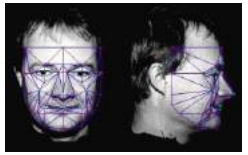
智能终端拍照拍视频

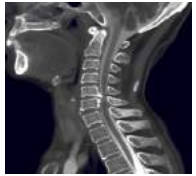
发微博、发微信

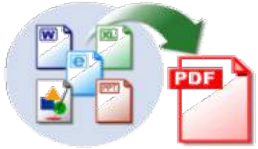
其他互联网数据

来自"大人群"泛互联网数据

来自大量传感器的物联网数据

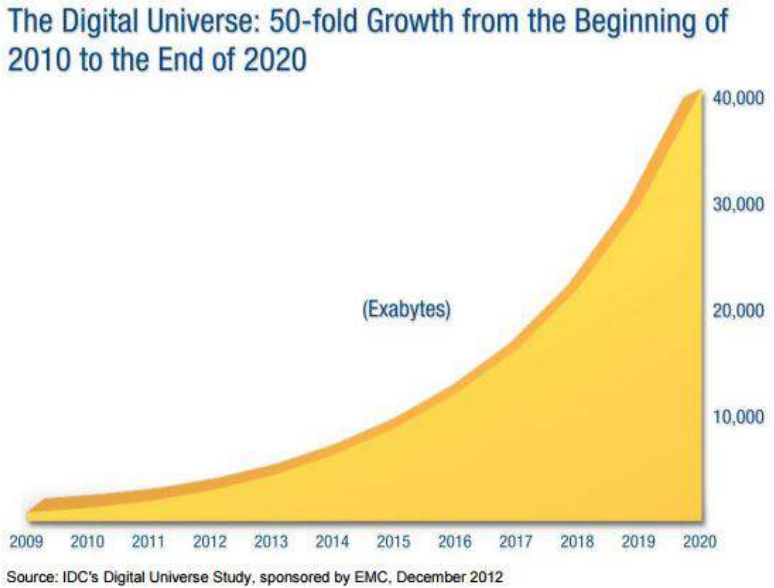






科学研究及行业异构专业数据

2.1 引言



全世界数据量急剧增长

移动互联网进一步激励数据增长

大数据的产生是计算机和网络通信技术发展的必然结果。

2.1 引言

- 多种多样的数据格式


交易记录


系统日志


支付信息


搜索历史

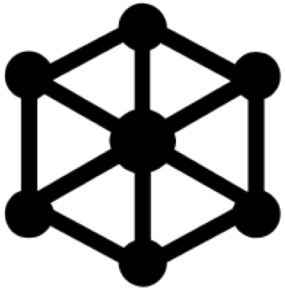

社交关系


浏览记录

数据是信息时代的基础生活资料与市场要素

2.1 引言

数据
大小



数据
深度

数据
质量

数据量的大小、数据涉及业务领域的深度
以及数据的质量将对大数据分析结果产生直接影响

2.2 数据渠道

- 数据获取 (大数据应用的第一个环节)

数据源分布

本单位自营

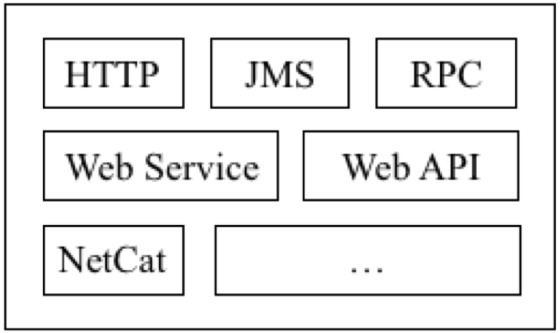
- 自营系统
- 历史遗留数据

外单位他营

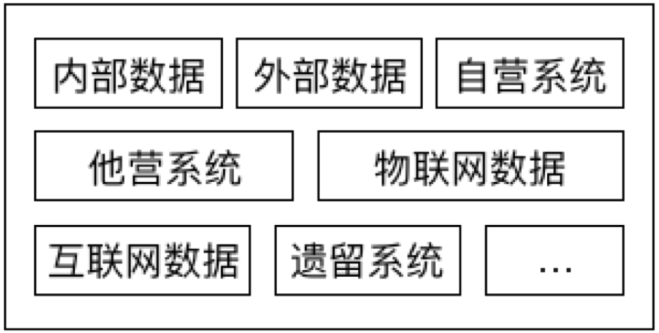
- 其他利益主体运营平台
- 物联网数据
- 政府数据
- 互联网/移动互联网数据

2.2 数据渠道

- **数据获取 (大数据应用的第一个环节)**



不同的协议



不同的来源

不同的技术/策略

- 不同数据源需不同的获取协议
- 多源性是大数据应用典型特征
- 不同数据源需不同的获取技术/策略

2.2 数据渠道

- **按照技术流分类**

内部数据

散布于各个利益主体，包括政府各级部门及企事业单位的服务器中，数据的富集与整合是在数据库层面或者软件系统层面进行数据导入导出，包括本单位自营系统、外单位利益主体营运的系统、政府数据，物联网数据等。

互联网数据

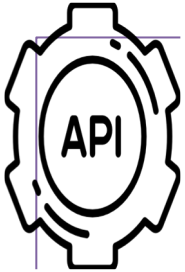
散布于互联网中,也称为网络大数据,数据的富集与整合是通过网络爬虫自动从URL中获得数据。

2.2 数据渠道

- **几种典型的数据获取途径**

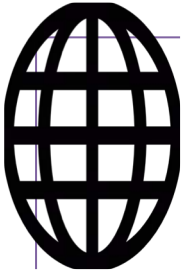
数据库层次 (内部数据从数据库中直接获取)

- 存储过程
- ETL



API层次 (以接口的形式提供/获取数据, 安全、灵活)

- 接口

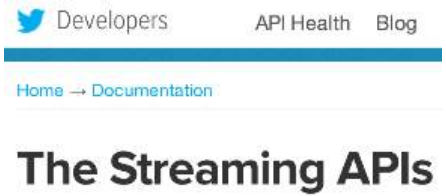


互联网 (网络爬虫)

- 表面网
- 深网

2.2 数据渠道

- **API**



海量推特
文字信息

通过推特开放的API，能够让第三方近乎实时地获取公开数据的各个子集，使得许多基于推特数据的研究与分析应用成为可能

开源工具
OpinionFinder

正面情绪

负面情绪

情绪测试工具
GPOMS

幸福

冷静

警惕

确信

活力

友善

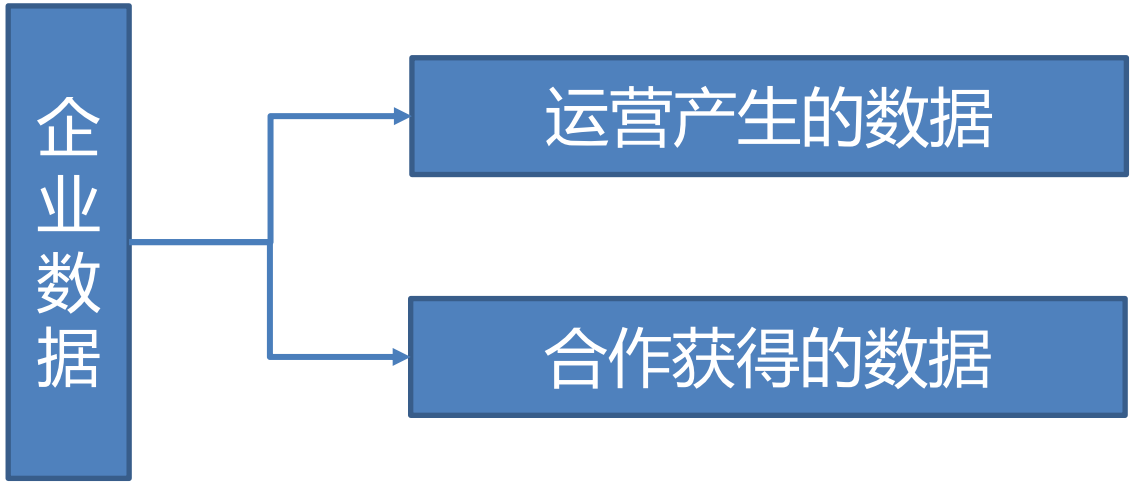
情绪分值走势图

如果将推特情绪走势图向后挪3天，跟道琼斯工业指数的走势图吻合度达87.6%

公开数据——不可忽视的一方力量

2.3 内部数据及获取方法

2.3.1 目标任务-获取企业数据



2.3 内部数据及获取方法

2.3.1 目标任务

- 内部数据资源整合的重要性

构建数据驱动应用
推进扩展价值实现

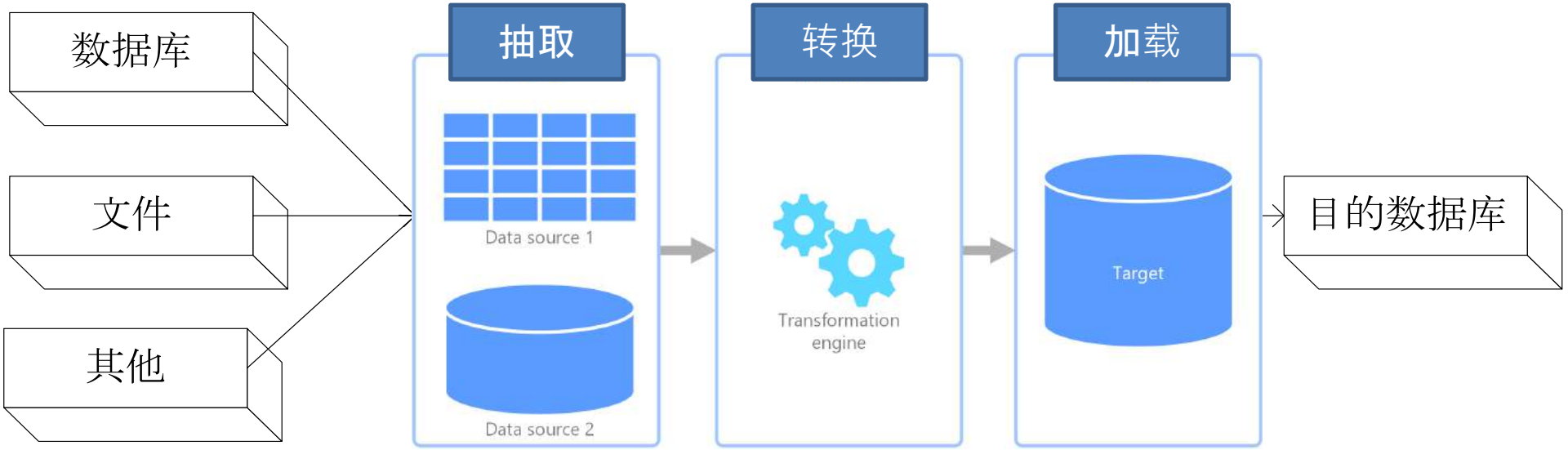
统一数据规范标准
推动数据共享开放

重视数据安全管理
完善数据安全保障

推进数据融合管理
增加数据语义厚度

2.3 内部数据及获取方法

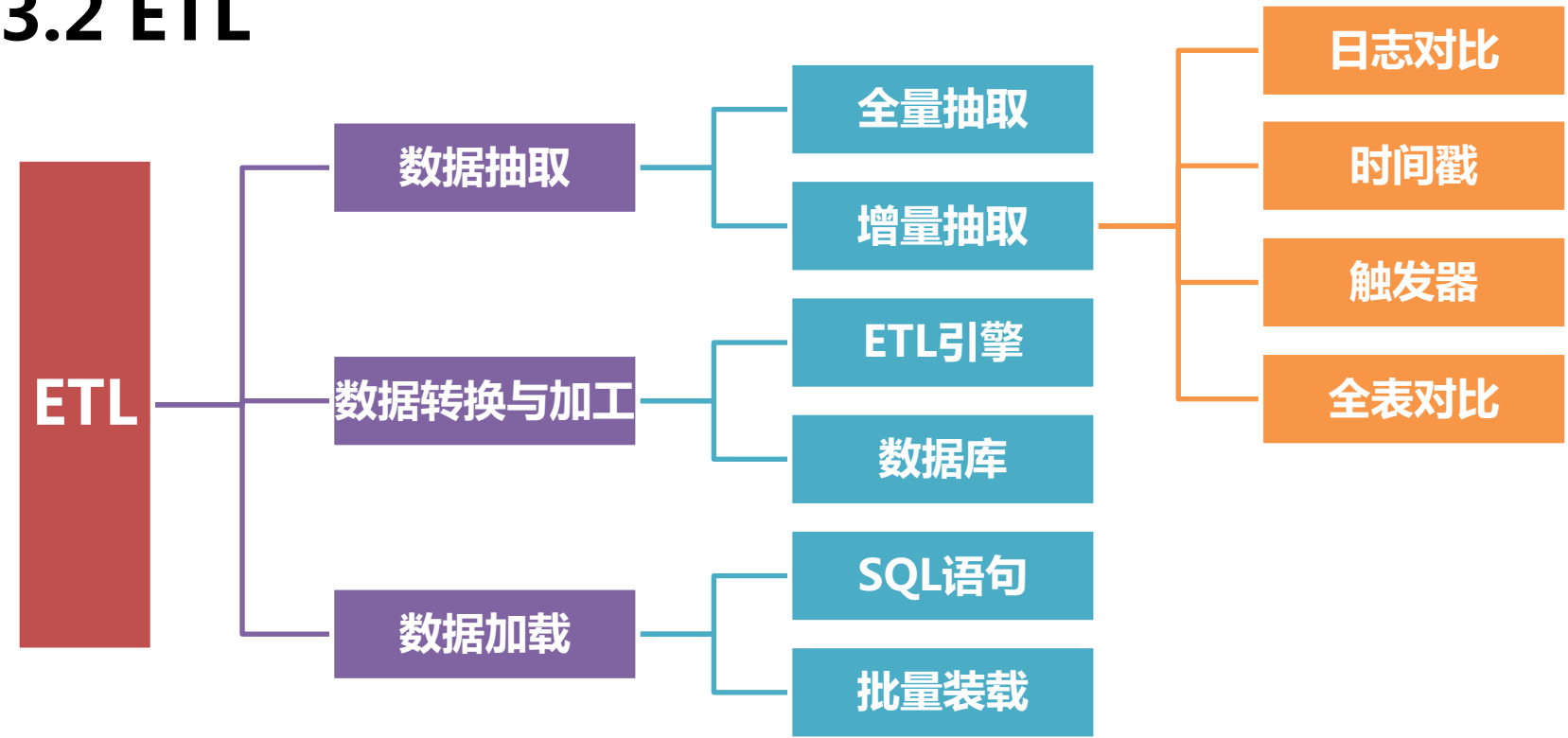
2.3.2 ETL (*Extract-Transform-Load*)



ETL 是数据获取的重要手段

2.3 内部数据及获取方法

2.3.2 ETL



2.3 内部数据及获取方法

2.3.3 三种主流 ETL 工具

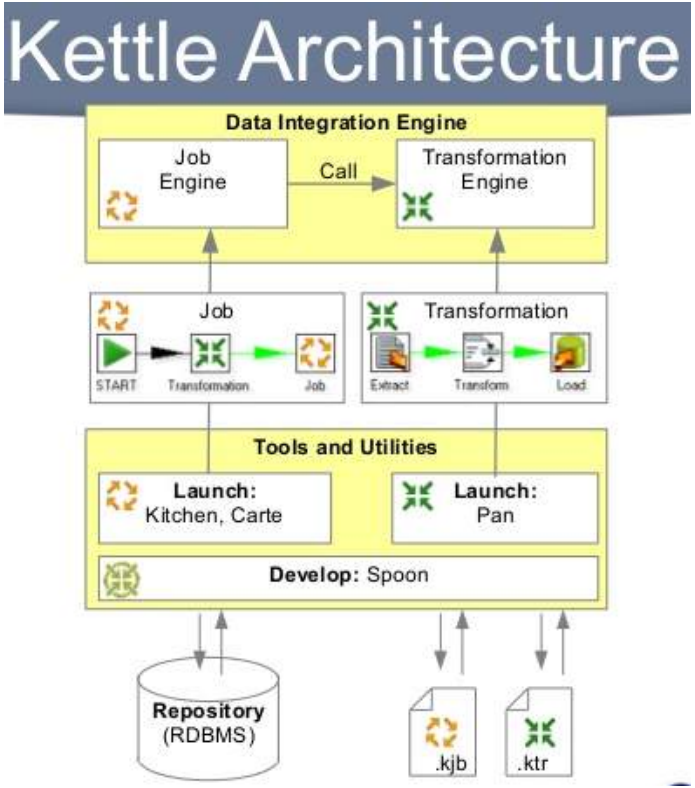
DataStage

Informatica PowerCenter

Kettle

2.3 内部数据及获取方法

- **Kettle**



2.3 内部数据及获取方法

2.3.3 三种主流 ETL 工具

比较维度	DataStage	Informatica PowerCenter	Kettle
数据源	目前市场上的大部分主流数据库，并且具有优秀的文本文件和XML文件读取和处理能力	大部分主流数据库，用于访问和集成几乎任何业务系统、任何格式的数据	大部分主流数据库
免费与否	需购买		免费开源
运行平台	windows/unix/linux		
软件安装和升级	图形安装，安装步骤较为复杂	完全图形化安装，无需额外安装平台软件且不需修改系统内核参数	绿色安装，直接使用
处理性能	支持并行处理，此外DataStage企业版可以在多台装有DataStage Server的机器上并行执行。并行执行能力使得DataStage所能处理数据的速度可以得到趋近于线性的扩展，轻松处理大量数据	可并行运行多个Session提高性能，可使用分区写目标数据以提高速度可建立多个PowerCenter Server，并发运行多个Session和workflow，结合Streaming和文件交换区的技术，优化硬盘和内存的资源利用。	使用JDBC，性能与Datastage，Informatica相比要差很多 适合于数据量较小的ETL加工使用

2.3 内部数据及获取方法

2.3.3 三种主流 ETL 工具

比较维度	DataStage	Informatica PowerCenter	Kettle
元数据管理	元数据信息不公开	元数据资料库可基于所有主流系统平台的关系型数据库(Oracle、DB2、teradata、Informix、Sql server等)	无元数据管理
抽取容错性	没有真正的恢复机制	抽取出错可恢复，可实现断点续传的功能	无恢复功能
操作便捷性	全图化开发，无编码	全图化开发，无编码操作性简便	全图化开发，无编码，操作简单
编码支持	几乎支持目前所有编码格式	支持编码格式十分丰富	支持常见的编码格式
系统安全性	只提供Developer和Operator两个角色，系统较安全	多范围的用户角色和操作权限（只读、操作和设计等），权限可以分到用户或组	简单的用户管理功能

2.4 外部数据及获取方法

2.4.1 目标任务

- **网络大数据(Network Big Data)**通常是指在互联网上可获得的大数据，价值巨大。
- 分为表面网和深网，前者一般获取手段是通过爬虫，后者是通过深网爬虫。

2.4 外部数据及获取方法

2.4.1 目标任务

- **网络大数据特性：**

2.4 外部数据及获取方法

2.4.2 网络爬虫

- 是一种按照一定规则，自动抓取万维网信息的程序或者脚本
- **三类典型爬虫**

批量型爬虫
- 根据用户配置进行网络数据的爬取
- 另两类爬虫的基础

增量型爬虫
- 根据用户配置持续进行网络数据的爬取
- 准实时任何应用场景，如Google、Baidu等

垂直型爬虫
- 根据用户配置持续进行指定网络数据的爬取
- 面向指定网站或者主题准实时任何应用场景

2.4 外部数据及获取方法

2.4.2 网络爬虫

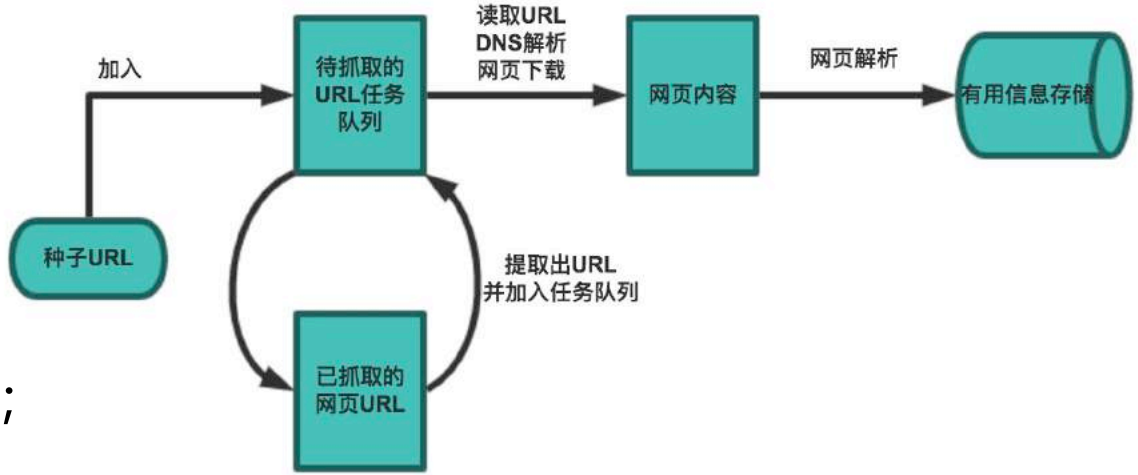
- 是一种按照一定**规则**，**自动**抓取万维网信息的**程序**或者**脚本**
- 网络爬虫开始于一张被称作**种子**的统一资源地址列表（也称**URL池**或**URL队列**），将其作为抓取的链接入口。
- 当网络爬虫访问这些网页时，识别出页面上所有的所需**网页链接**并将它们加入到**待爬队列**中。
- 伺候从待爬队列中取出网页链接按照一套**策略**循环访问，这样一直循环，直到待爬队列为空时爬虫程序停止运行。

2.4 外部数据及获取方法

2.4.2 网络爬虫

通用爬虫框架流程

1. 指定入口 URL，将其放入种子URL队列；
2. 将种子 URL 加入待抓取 URL 队列；
3. 读取URL，下载对应页面；
4. 解析页面，嗅探新的URL去重加入队列；
5. 持续上述1-4步直到待抓取 URL 队列为空

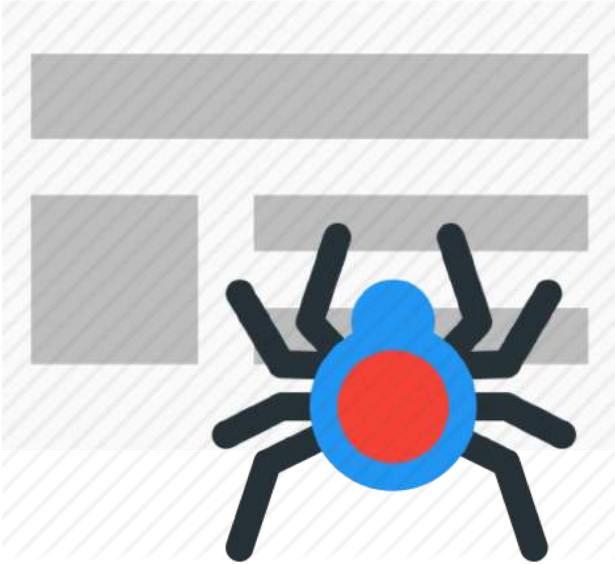


注：每一个URL都是互联网中的一个网页,而互联网中的每一个网页都是通过网页中的URL链接扇出到另外的URL中

2.4 外部数据及获取方法

2.4.2 网络爬虫

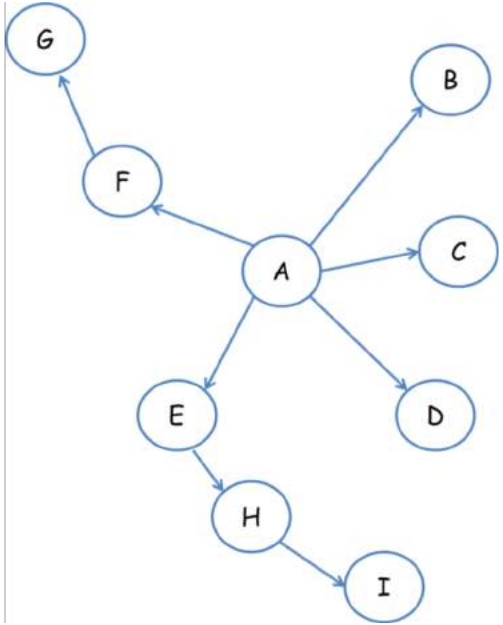
- **针对问题：在爬取一个具体URL中数据时，如何处理这个URL中扇出的URL链接?**
- **网络爬取策略**
  - 指在网络爬虫系统中决定 URL 在待抓取队列中排列顺序的方法。



2.4 外部数据及获取方法

2.4.2 网络爬虫

- **网络爬取策略**



深度优先遍历策略

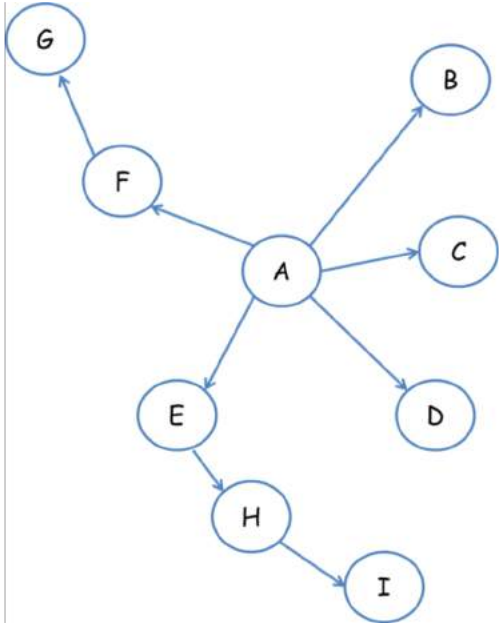
指网络爬虫会从起始页开始，一个链接一个链接跟踪下去，处理完这条线路之后再转入下一个起始页，继续跟踪链接;

遍历的路径：A-F-G E-H-I B C D

2.4 外部数据及获取方法

2.4.2 网络爬虫

- **网络爬取策略**



宽度优先遍历策略

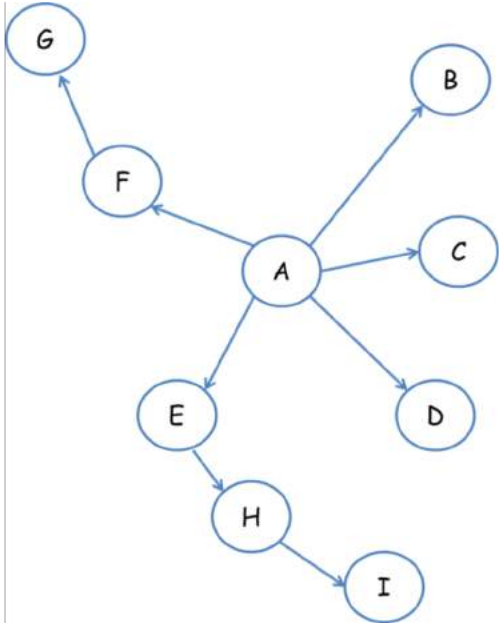
基本思路是将新下载网页中发现的链接直接插入待抓取URL队列的末尾。也就是指网络爬虫会先抓取起始网页中链接的所有网页，然后再选择其中的一个链接网页，继续抓取在此网页中链接的所有网页。

遍历的路径：A-B-C-D-E-F G H I

2.4 外部数据及获取方法

2.4.2 网络爬虫

- **网络爬取策略**



深度优先遍历策略

指网络爬虫会从起始页开始，一个链接一个链接跟踪下去，处理完这条线路之后再转入下一个起始页，继续跟踪链接;

遍历的路径：A-F-G E-H-I B C D

宽度优先遍历策略

基本思路是将新下载网页中发现的链接直接插入待抓取URL队列的末尾。也就是指网络爬虫会先抓取起始网页中链接的所有网页，然后再选择其中的一个链接网页，继续抓取在此网页中链接的所有网页。

遍历的路径：A-B-C-D-E-F G H I

若干实际问题：URL去重、反爬虫技术、高性能爬虫（分布式）

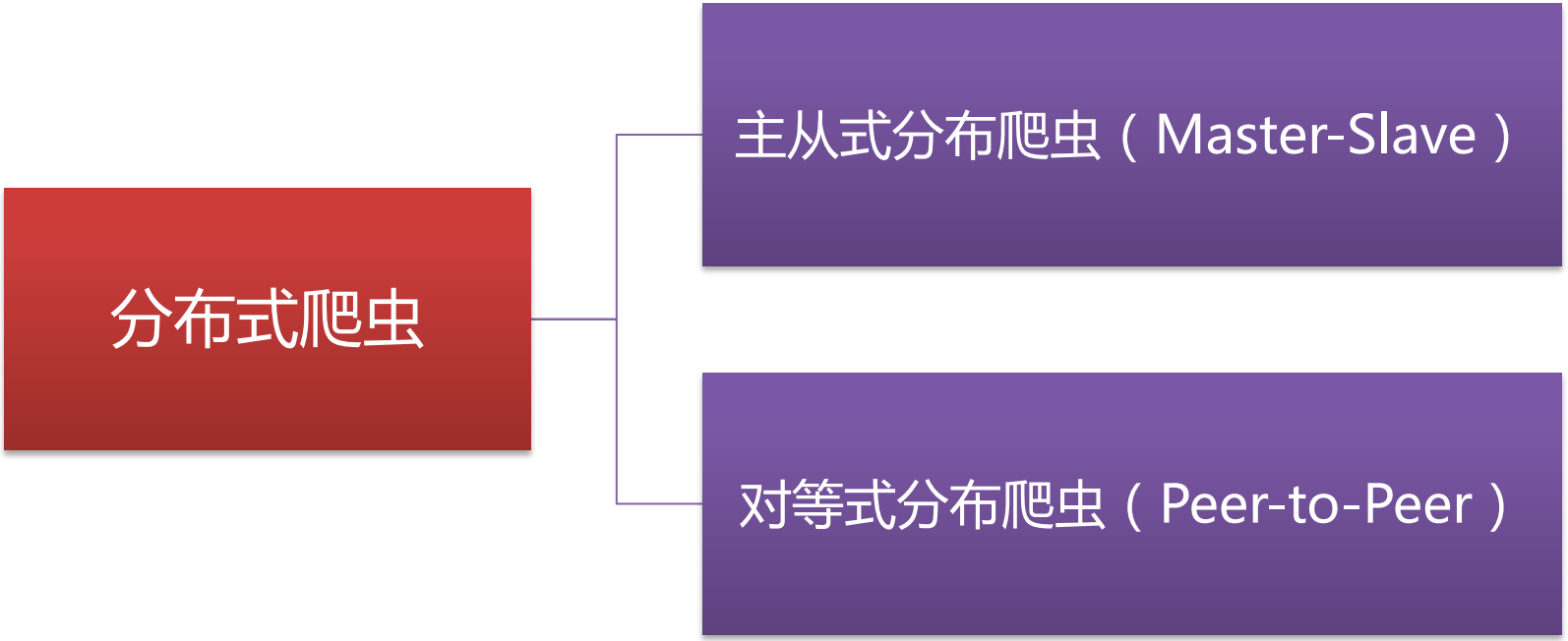
2.4 外部数据及获取方法

2.4.2 网络爬虫

抓取策略	描述	特点
深度优先	从URL池中选择某URL，然后按深度优先遍历以该URL为根节点的所有URL网页内容，然后取出URL池中下一个URL，继续上述策略循环至URL池遍历完	抓取深度大； 但易导致无限抓取，使得爬取过程无法收敛
广度优先	逐层抓取URL池中每一个URL内容并将每一层的扇出URL纳入URL池中，按照宽度优先策略继续遍历	抓取宽度广，易控制，有效减轻服务器的负载； 但易造成URL大量聚集而导致URL池溢出
局部PageRank	在URL池和已抓取网页组成的网页集合中计算URL池中PageRank值并以此进行排序，然后按照此序顺序遍历各个URL	网络环境中，由于广告链接、作弊链接的存在，易导致PageRank值不能完全刻画其重要程度，从而导致实际抓取数据无效
OPIC策略	将每个网页赋予相同的“金币”，每当下载某个页面P，则将P拥有的“金币”平均分配给网页中包含的链接页面。待爬队列中链接依“金币”排序	OPIC计算速度快于局域PageRank策略，是一种较好的重要性衡量策略，适合实时计算场合

2.4 外部数据及获取方法

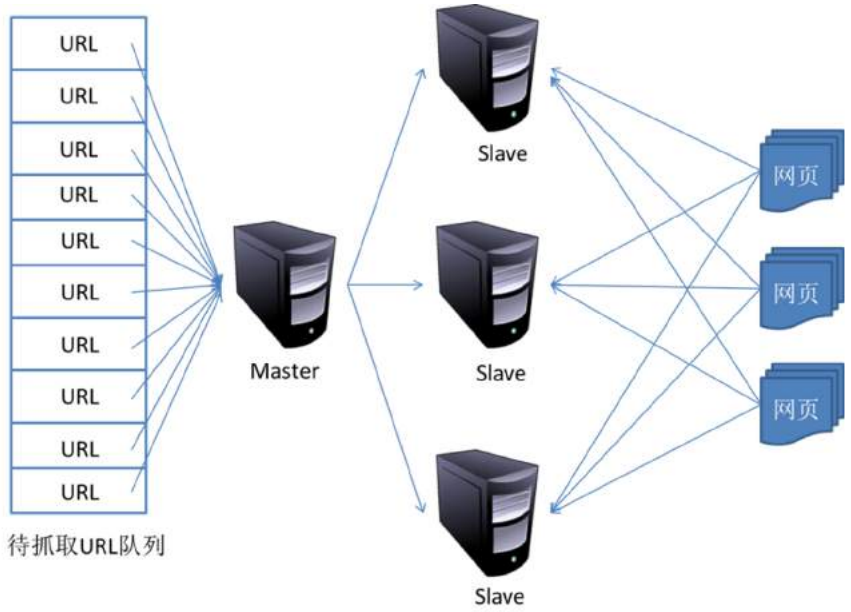
2.4.2 网络爬虫



2.4 外部数据及获取方法

2.4.2 网络爬虫

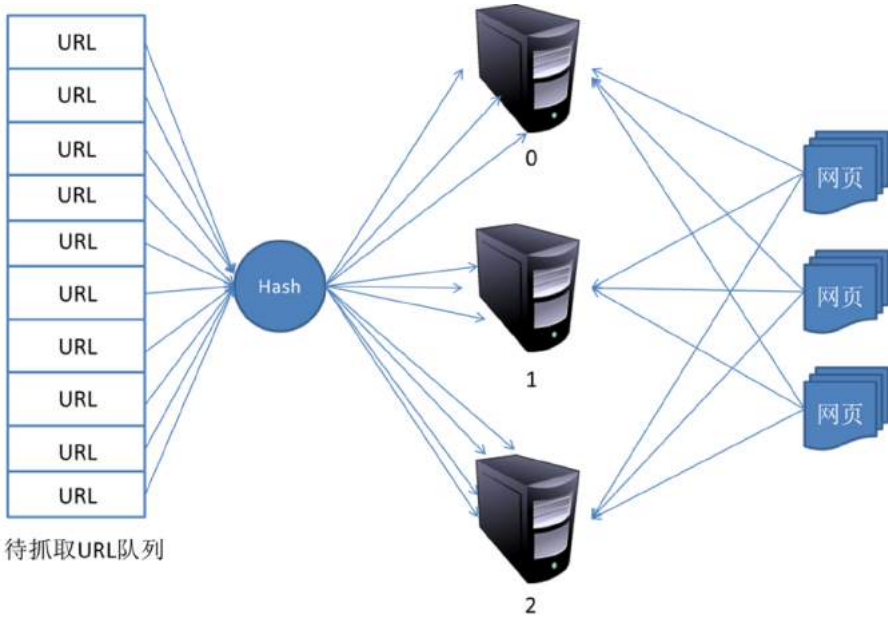
- **主从式分布式爬虫**
  - Master负责URL分发、负载均衡、心跳检测，全局去重等服务；
  - Slave负责实际任务的抓取
- 缺点：主服务器容易成为系统瓶颈



2.4 外部数据及获取方法

2.4.2 网络爬虫

- **对等式分布式爬虫**
  - 每台服务器功能相同，无主从之分；
  - 将主域名哈希取模决定所属服务器；
- 缺点：某台服务器宕机会造成所有任务重新分配



2.4 外部数据及获取方法

2.4.2 网络爬虫

❖ 爬虫评价视角

视角	评价维度	评价细节
程序开发	高效性	每秒钟抓取的网页数量，每秒钟抓取的网页数据越大，爬虫程序越高效
	可扩展性	不同的网页具有不同的（模板）结构，针对不同的应用场景，针对网络爬虫数据抓取的需要也不一样，均需网络爬虫具有良好的扩展性
	健壮性	必须具有良好的容错性，能够正确处理相关异常情况，保障抓取过程正常进行
	友好性	1)网络爬虫程序应该易于管理URL池；2) 减少被抓取网站的网络负载
抓取内容	抓取网页覆盖率	网络爬虫应具有较大的抓取网页覆盖率（指抓取的网页占整个互联网的比例），抓取网页覆盖率越大，表明抓取的网络大数据越可能全面
	抓取网页及时性	应及时获取最新网络数据，以保持抓取的网络大数据的“活性”
	抓取网页重要性	应该抓取具有重要价值的网页，使抓取过程具有较高的性价比

2.4 外部数据及获取方法

爬虫	优点	缺点	适用场景	语言
Nutch	1. 集爬取、索引于一体。基于Hadoop的分布式系统；存储层剥离，支持存储HBase, Cassandra, MySql等数据库； 2. 基于插件式设计，扩展和定制比较方便 3. 支持网页解析和索引，可以对接至Solr，搭建通用的搜索引擎	Nutch更侧重于索引，和Hadoop结合之后，会消耗更多资源在非爬虫部分，爬取效率较低	不仅需要爬取数据，同时对索引有一定需求。大数量、考虑分布式的场景下，可以直接使用Nutch的分布式解决方案	Java
Scrapy	1. 插件设计，扩展性比较好； 2. 爬虫规则定制简单； 3. 支持抓取和抽取，数据抽取结构化； 4. 抽取支持xpath和css提取网页数据；	1. 单机多线程实现，默认不支持分布式 2. 数据存储方案支持 Local filesystem、FTP、S3、Standard output 3. 默认中间过程网页不会保存，只保存抽取结果；	1. 没有分布式需求，或者有其他分布式解决方案； 2. 只需要抽取结果，对原始网页不感兴趣；	Python
Larbin	单纯的爬取功能，简单，单机效率高；	1. 不支持分布式系统抓取存储； 2. 功能相对简单，提供的配置项也不够多； 3. 不支持网页自动重访，更新功能	1. 只需要爬虫工具，其他功能通过其他方案解决； 2. 硬件有限，但对效率要求较高； 3. 适合作定制化爬虫系统爬虫器	C++

2.5 深网数据及获取方法

2.5.1 深网的基本概念

2000年由Bright Planet公司首创，表述将信息内容存储在检索数据库中而仅响应直接查询提问的网站。

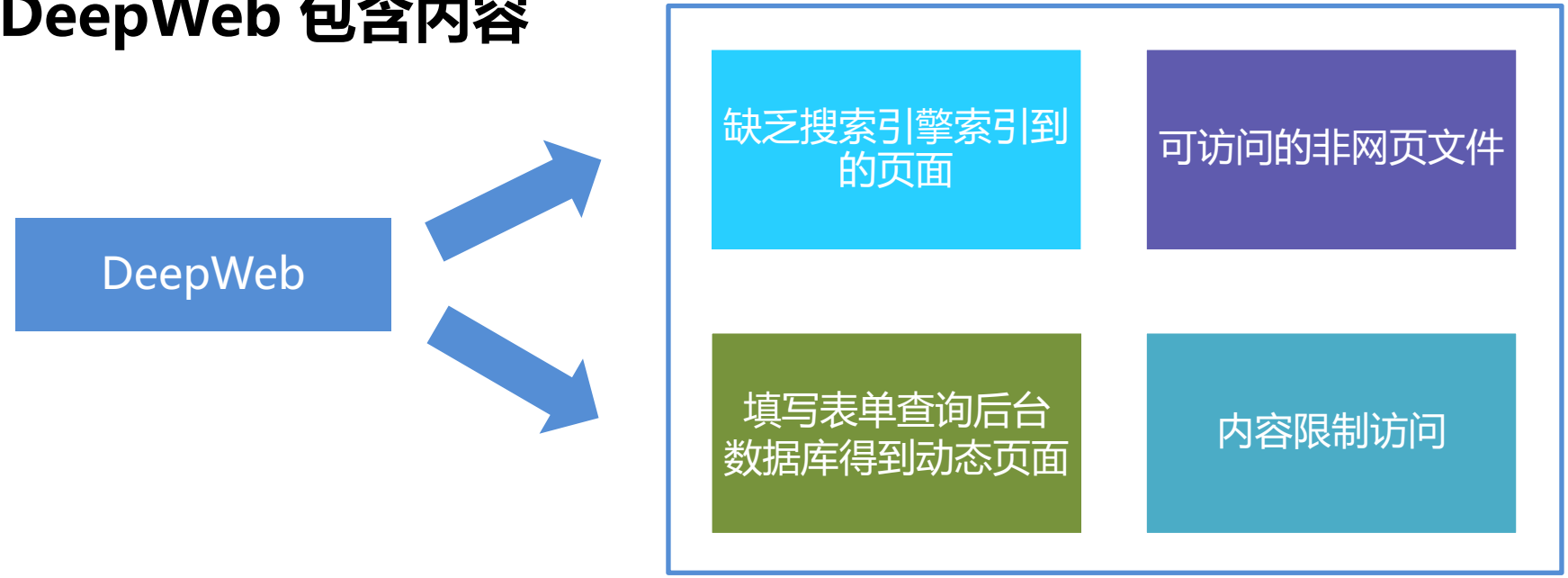
美国互联网专家、图书馆员Chris Sher-man和Gary Price将其定义为:“在互联网上可获得的、但传统搜索引擎由于技术限制不能或者经过慎重考虑后不愿意作索引的那些本网页、文件或其他高质量、权威的信息”。

https://brightplanet.com

2.5 深网数据及获取方法

2.5.2 深网信息的特点

- DeepWeb 包含内容



2.5 深网数据及获取方法

2.5.2 深网信息的特点

深网内容的全部价值是表面网的1000-2000倍。

95%的深网信息可以公共获取而无需付费或订阅。

深网的信息内容与所有的信息需求、市场和领域高度相关。

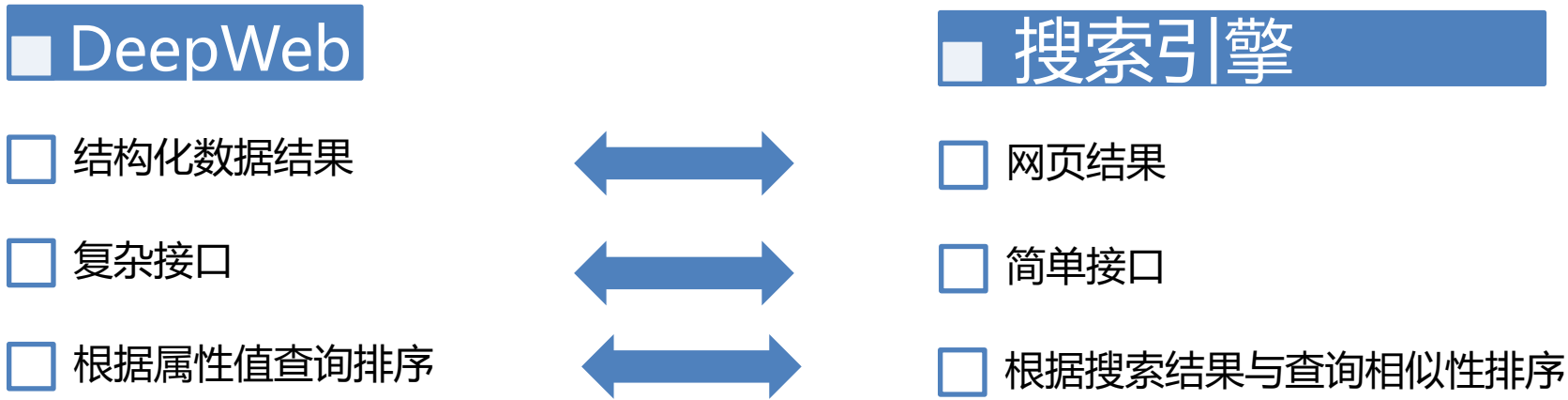
一半以上的深网内容存储在专题数据库中。

深网的规模远大于表面网且持续性地高速增长。

2.5 深网数据及获取方法

2.5.2 深网信息的特点

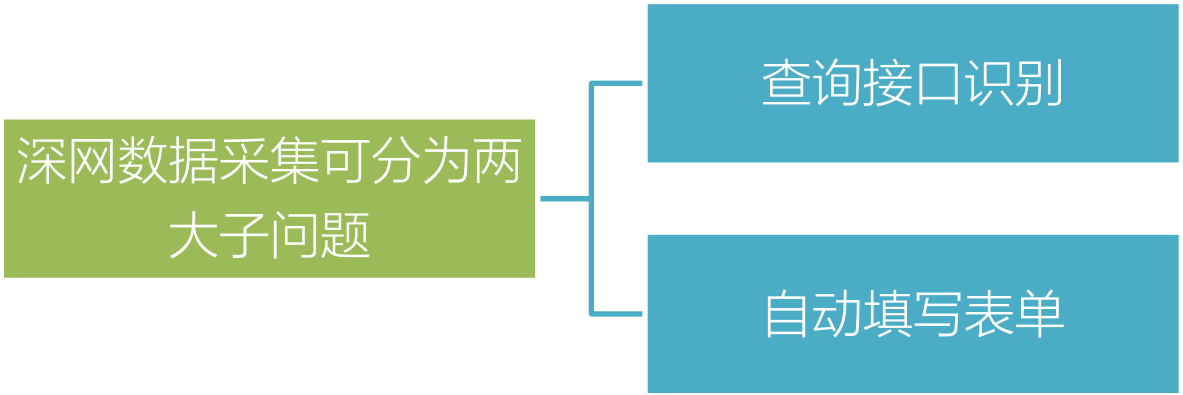
- DeepWeb 与搜索引擎区别



2.5 深网数据及获取方法

2.5.3 深网数据获取方法

- 数据存于后台数据库中，很少有显式链接指向这些数据
- 需输入相关查询条件，传统爬虫无法获取这些数据



2.5 深网数据及获取方法

2.5.3 深网数据获取方法

查询接口识别

- 采用包含视觉布局在内的多种方法来解析HTML表单或通过对HTML表单进行语法分析来自动发现深网数据资源
- 在视觉布局的基础上，增加了文本相似性启发式规则，从而能够将HTML表单与特定领域关联起来以实现表单自动填写功能
- 假定HTML表单遵循一个隐藏的语法规则，构造了一个存在二义性的语法并编写出一个解析器对其进行处理
- 通过构造页面分类器和表单分类器从而自动寻找到与集成任务相关的深网数据库

2.5 深网数据及获取方法

2.5.3 深网数据获取方法

- 自动填写表单
  - 基于领域知识：使用启发式规则将表单的域与领域概念关联起来并由此输入与领域概念相关参数
  - 领域无关探测：基于采样迭代式地从查询结果中获取查询关键词，以较少查询次数获取尽可能多查询结果

2.5 深网数据及获取方法

- **延展阅读**
  - [VLDB'01]Crawling the Hidden Web
  - [PVLDB'08]Google's Deep-Web Craw

2.6 本章小结

- 数据的获取直接决定了大数据分析的结果。
- 数据源梳理是基础中的基础。鉴于数据源分布的潜在多源、异构（获取协议异构、平台异构等）特性，数据获取必须有效响应各层次的技术难题。
- 本章简单介绍了数据源的一般分布以及通常意义下的不同数据源的获取方法和策略，重点介绍了ETL、表面网数据获取、深网数据获取等关键技术和方法。