

大数据编程

3-1

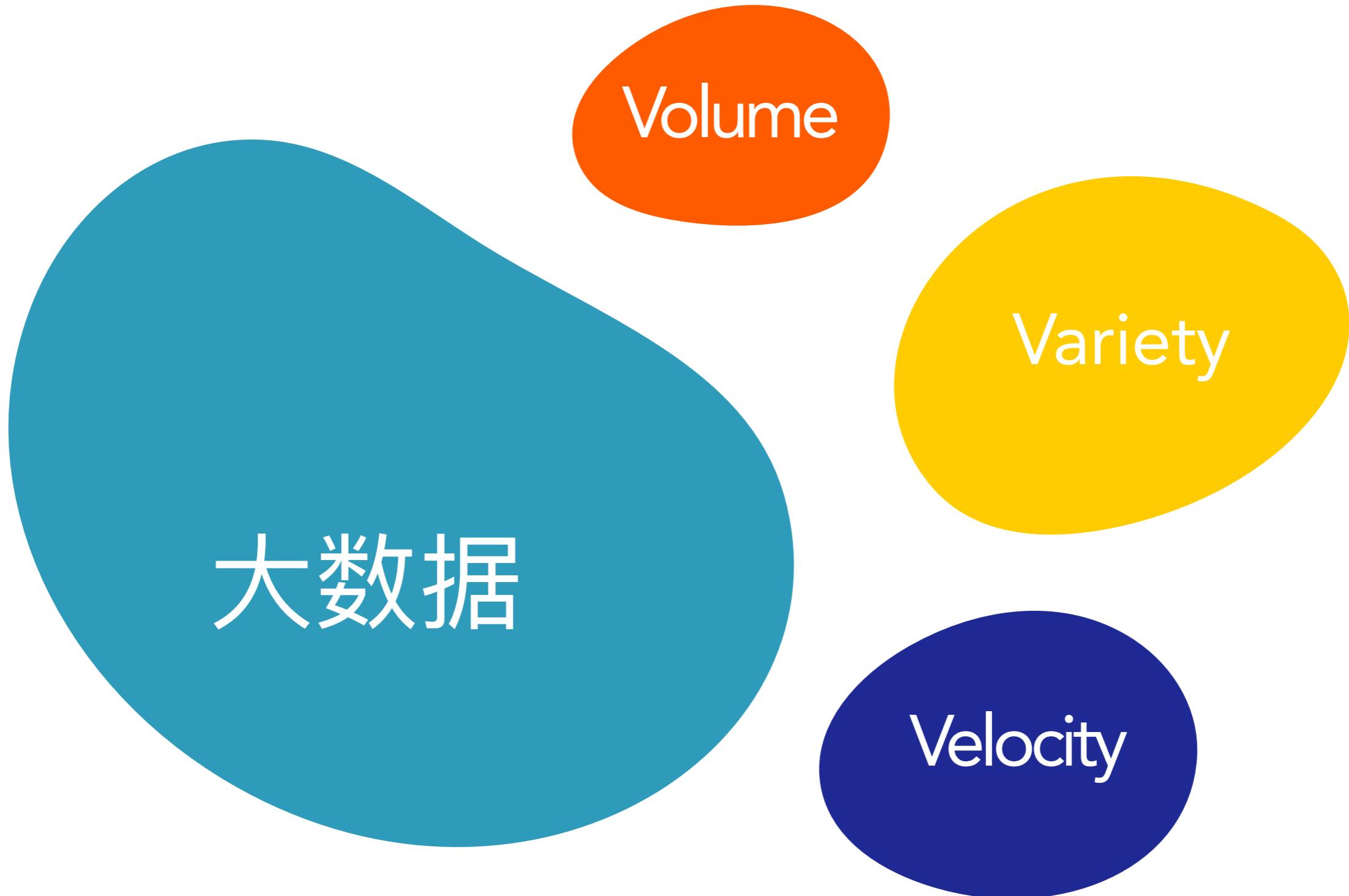
数据获取方法

中央财经大学 商学院
姚凯
2016

主要内容

- ❖ 知识回顾
- ❖ 数据来源
- ❖ 数据爬取方法
- ❖ HTML基础

大数据难题



大数据难题

13000+个
iPhone
应用下载

Skype上
37万+分钟的
语音通话

Twitter上发
布98000+新
微博

上传6600张
新照片到
flickr

发出1.68亿+
条Email

Facebook上
更新69.5万+
条新状态

YouTube上
上传600+新
视频

淘宝光棍节
10680+个新
订单

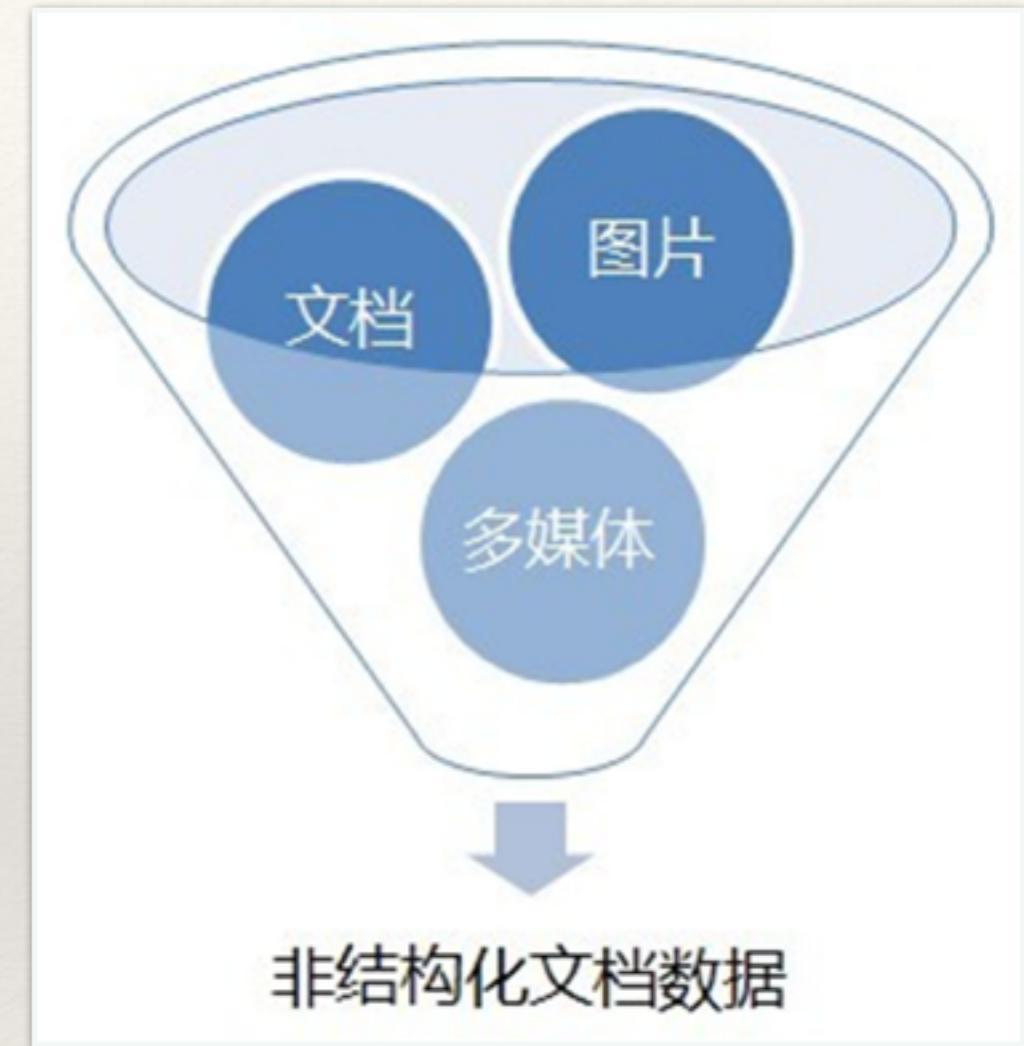
12306出票
1840+张



传统存储方式无法满足海量数据的要求

大数据面临的挑战

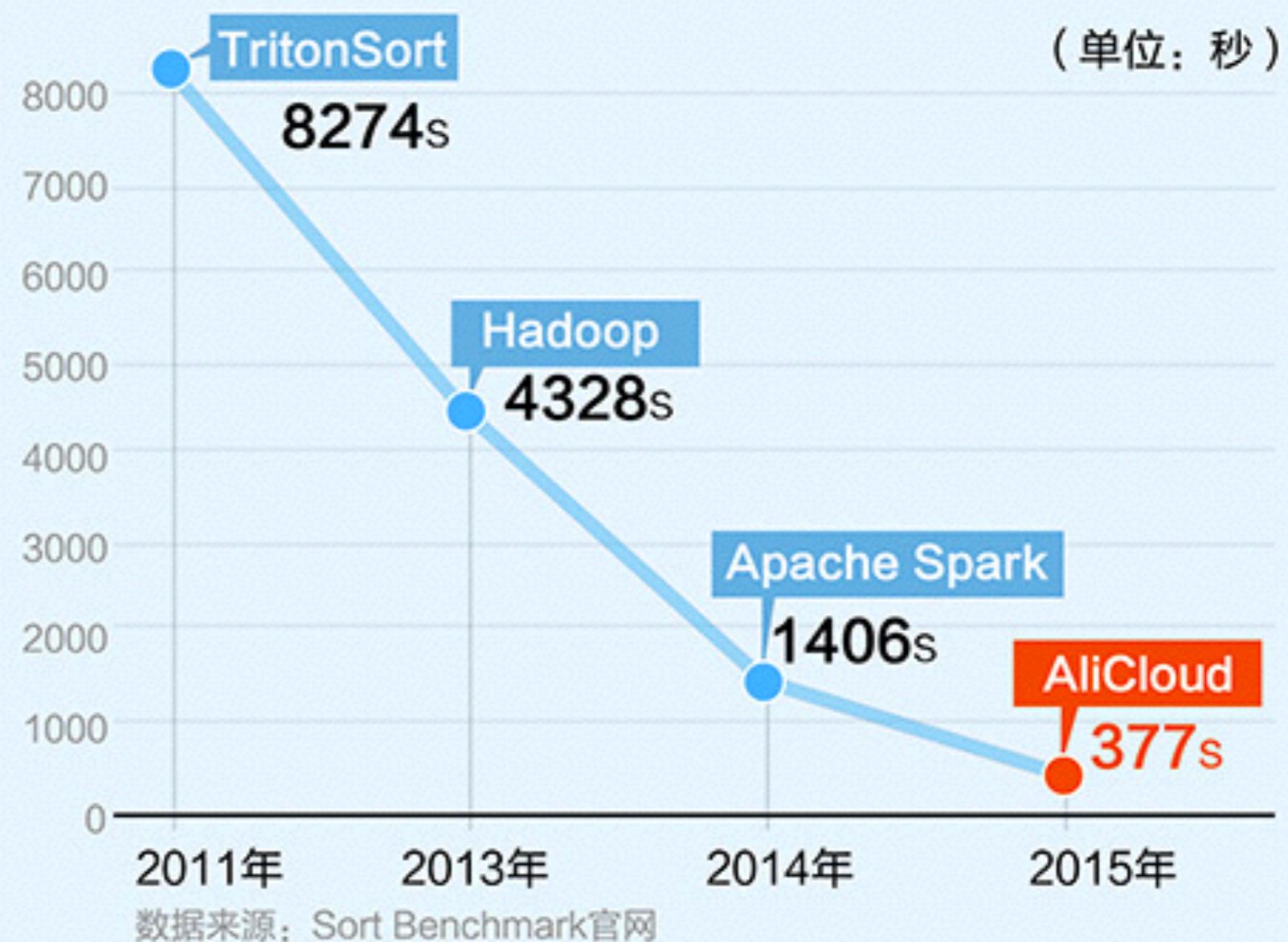
Variety



大数据面临的挑战



全球大数据计算性能变迁史 (100TB数据排序耗时)



大数据面临的挑战



1. 海量存储
2. 数据备份与恢复



1. 新型数据库
2. 新分析方法



1. 实时处理能力
2. 新处理框架

大数据核心技术

- ❖ 分布式文件系统
- ❖ 非关系型数据库
- ❖ 并行处理和分布式处理
- ❖ Hadoop生态圈
- ❖ 实时计算框架

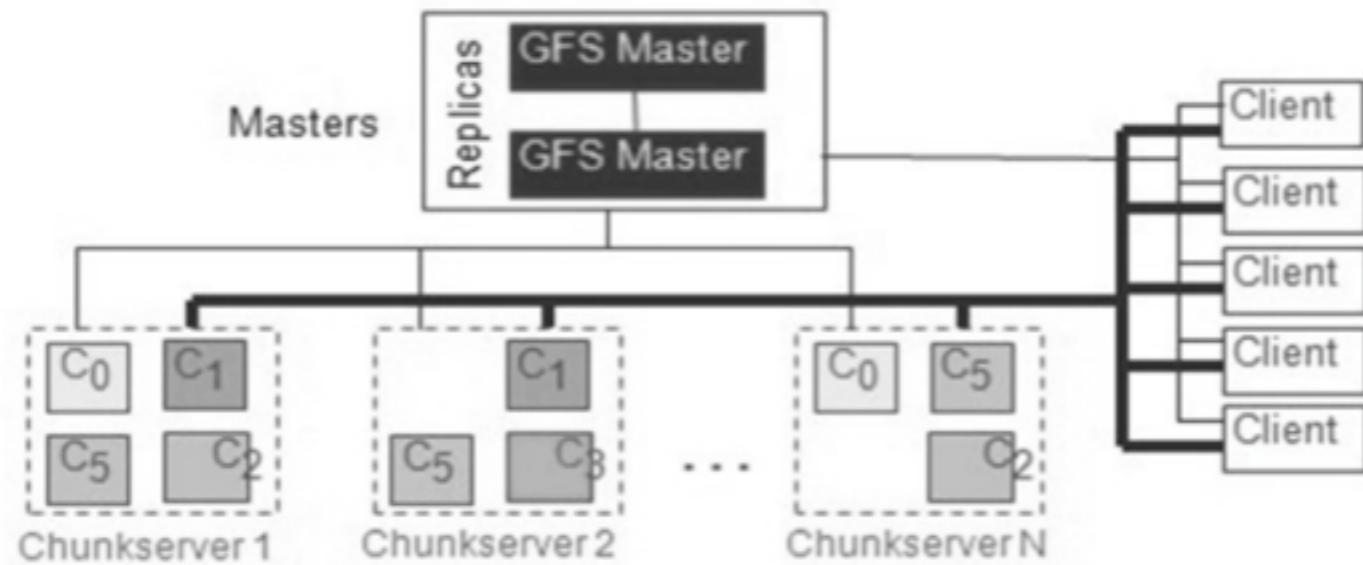
分布式文件系统



GFS将整个系统分为三类角色：Client（客户端）、Master（主服务器）、Chunk Server（数据块服务器）。

GFS Architecture

Heartbeat



NoSQL数据库

NoSQL = Not Only SQL



- 关系型数据库的局限性
 - 难以满足高并发读写的需求
 - 难以满足对海量数据高效率存储和访问的需求
 - 难以满足对数据库高可扩展性和高可用性的需求

NoSQL数据库

- ❖ 键值(Key-Value)存储数据库
- ❖ 列存储数据库
- ❖ 文档型数据库
- ❖ 图形(Graph)数据库

NoSQL 四大家族

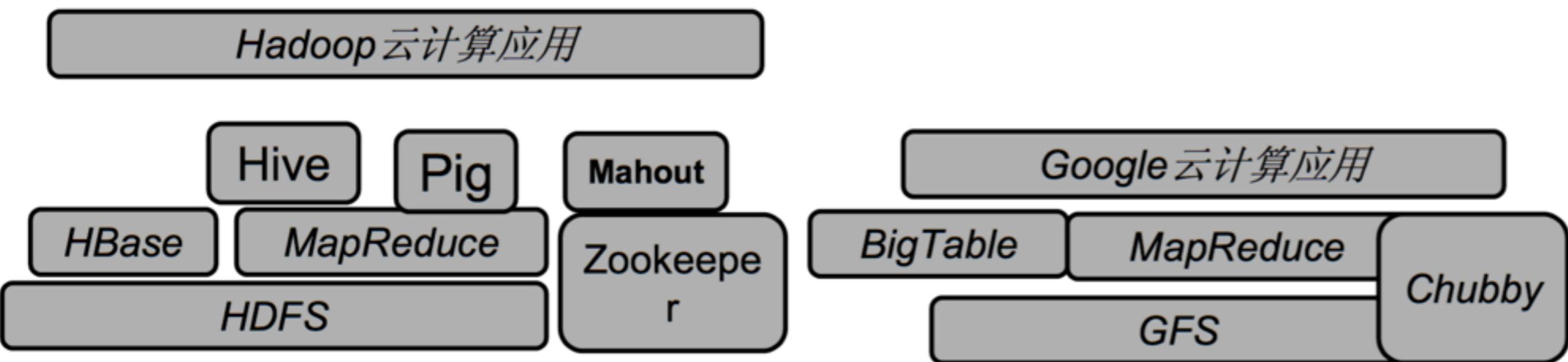
NoSQL 数据库类型	代表性产品	性能	扩展性	灵活性	复杂性	优点	缺点
键/值数据库	Redis Riak	高	高	高	无	查询效率高	不能存储结构化信息
列式数据库	HBase Cassandra	高	高	一般	低	查询效率高	功能较少
文档数据库	CouchDB MongoDB	高	可变的	高	低	数据结构灵活	查询效率较低
图形数据库	Neo4J OrientDB	可变的	可变的	高	高	支持复杂的图算法	只支持一定的数据规模

大数据核心技术

- ❖ 分布式文件系统
- ❖ 非关系型数据库
- ❖ 并行处理和分布式处理
- ❖ Hadoop生态圈
- ❖ 实时计算框架

Hadoop和Google架构比较

- ❖ 并行计算模型: MapReduce->MapReduce
- ❖ 分布式文件系统: HDFS->GFS
- ❖ 数据结构化管理组件: Hbase->BigTable
- ❖ 分布式锁服务: Zookeeper->Chubby

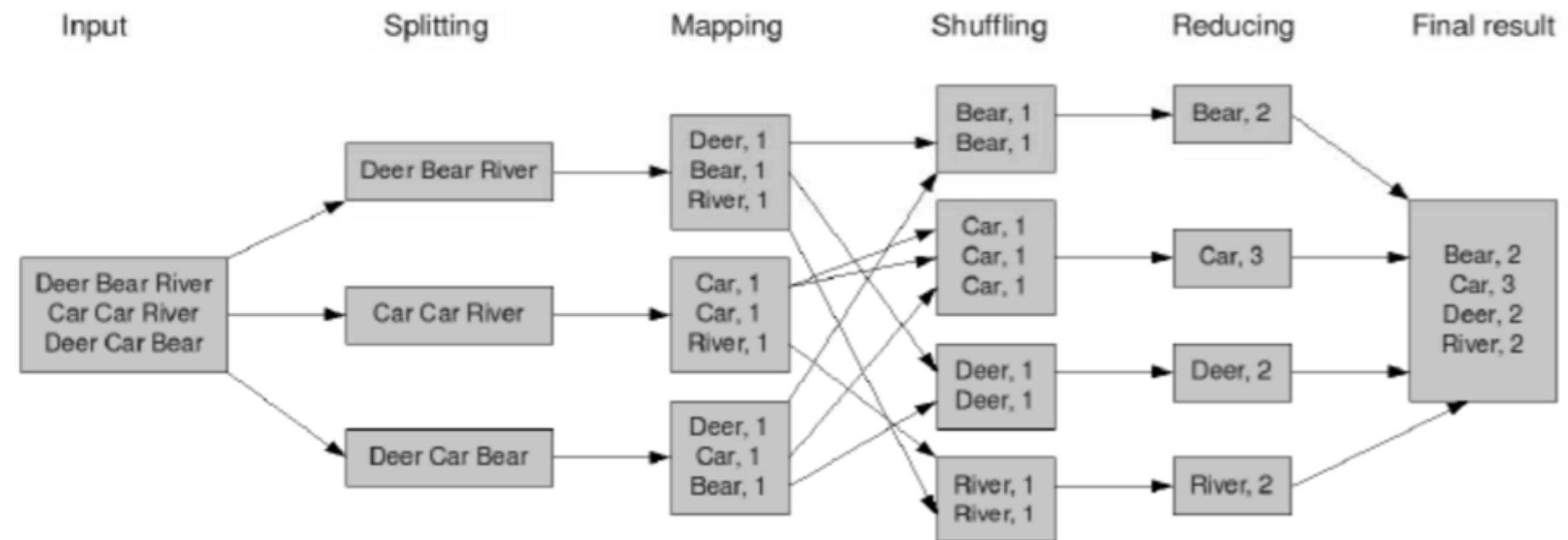


Hadoop核心组件

- ❖ HDFS: Hadoop Distributed File System
- ❖ MapReduce
- ❖ ZooKeeper
- ❖ HBase

MapReduce 示例

The overall MapReduce word count process



作业1

- ❖ 比较Google和Hadoop大数据框架三个核心组件
- ❖ 解释Spark和Storm的特点与区别
- ❖ 作业： 姓名+作业1.pdf
- ❖ 邮箱： yaokai@cufe.edu.cn

主要内容

- ❖ 知识回顾
- ❖ 数据来源
- ❖ 数据爬取方法
- ❖ HTML基础

数据来源

- ❖ 公司内部数据
- ❖ 公开数据库
- ❖ 企业年报
- ❖ 焦点访谈
- ❖ 实验法
- ❖ 网页爬取数据

公司内部数据

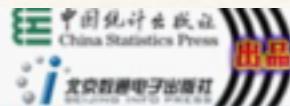
Objects	clickdata @test (local)
	email click_time click_useragent link
linchao 10/1/15 8:16	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
26871 10/1/15 8:37	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38051/GLxxuokz8OXbDCR2IsovuLOMA-Q_/p063/0/http:/
45401 10/1/15 8:50	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
16687 10/1/15 9:01	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
64546 10/1/15 9:04	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
69533 10/1/15 9:07	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
42327 10/1/15 9:13	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
83962 10/1/15 9:17	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
56441 10/1/15 9:20	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38049/9ysX575g8ZVUSIUa.6-9v.ORFXY_/p06c/0/http://
30541 10/1/15 9:20	Mozilla/5.0 (iPhone; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
56441 10/1/15 9:21	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38049/9ysX575g8ZVUSIUa.6-9v.ORFXY_/p06c/0/http://
56441 10/1/15 9:21	Mozilla/5.0 (iPhone; http://a1722.oadz.com/link/C/1722/38049/9ysX575g8ZVUSIUa.6-9v.ORFXY_/p06c/0/http://
52809 10/1/15 9:25	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38066/SfwdiRcJ4mohdNdOs8kw2AMCRnc_/p079/0/http:
38293 10/1/15 9:46	Mozilla/5.0 (iPhone; http://edm.baifendian.com/redirect?bfd_nid=lingdong_test&cid=Q2ppdXhpYW4=&em=\${em}&
38293 10/1/15 9:46	Mozilla/5.0 (iPhone; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
38293 10/1/15 9:47	Mozilla/5.0 (iPhone; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
38293 10/1/15 10:01	Mozilla/5.0 (iPhone; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
38293 10/1/15 10:03	Mozilla/5.0 (iPhone; http://a1722.oadz.com/link/C/1722/38066/SfwdiRcJ4mohdNdOs8kw2AMCRnc_/p079/0/http
93454 10/1/15 11:41	Mozilla/5.0 (iPhone; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
45925 10/1/15 12:27	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
27679 10/1/15 12:27	Mozilla/5.0 (iPhone; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
60115 10/1/15 14:31	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
60115 10/1/15 14:31	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
19619 10/1/15 14:36	Mozilla/5.0 (Linux; http://edm.baifendian.com/redirect?bfd_nid=lingdong_test&cid=Q2ppdXhpYW4=&em=\${em}&
60115 10/1/15 14:37	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
50177 10/1/15 16:21	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/
14986 10/1/15 18:56	Mozilla/5.0 (Linux; http://edm.baifendian.com/redirect?bfd_nid=lingdong_test&cid=Q2ppdXhpYW4=&em=\${em}&
42119 10/1/15 23:43	Mozilla/5.0 (Linux; http://a1722.oadz.com/link/C/1722/38065/uPeHpshDEDTLUEi8AJo6CE2m4fs_/p07e/0/http:/

公开数据库

国务院人口普查办公室
国家统计局人口和就业统计司

中国2010年人口普查资料

TABULATION ON THE 2010 POPULATION CENSUS OF
THE PEOPLE'S REPUBLIC OF CHINA



EXCEL
版权声明
资料使用说明
编委会
<ul style="list-style-type: none">第一部分全部数据资料<ul style="list-style-type: none">第一卷概要<ul style="list-style-type: none">1-1 各地区户数、人口数和性别比1-1a 各地区户数、人口数和性别比(城市)1-1b 各地区户数、人口数和性别比(镇)1-1c 各地区户数、人口数和性别比(乡村)1-2 各地区分性别、户口登记状况的人口<ul style="list-style-type: none">1-2a 各地区分性别、户口登记状况

1-1 各地区户数、人口数和性别比

地区	户 数				合 计			人 口 数		
	合计	家庭户	集体户	合计	男	女	性别比 (女=100)	小计	家	庭 户
全国	417722698	401934196	15788502	1332810869	682329104	650481765	104.90	1239981250	627410399	6125708
北京	7355291	6680552	674739	19612368	10126430	9485938	106.75	16389723	8173161	82165
天津	3963604	3661992	301612	12938693	6907091	6031602	114.52	10262186	5129604	51325
河北	20813492	20395116	418376	71854210	36430286	35423924	102.84	68538709	34552649	339860
山西	10654162	10330207	323955	35712101	18338760	17373341	105.56	33484131	16988087	164960
内蒙古	8470472	8205498	264974	24706291	12838243	11868048	108.17	23071690	11725291	113463
辽宁	15334912	14994046	340866	43746323	22147745	21598578	102.54	41755874	20956756	207991
吉林	9162183	8998492	163691	27452815	13907218	13545597	102.67	26457769	13358390	130993
黑龙江	13192935	13000088	192847	38313991	19426106	18887885	102.85	36884039	18603181	182808
上海	8893483	8253257	640226	23019196	11854916	11164280	106.19	20593430	10318168	102752
江苏	25635291	24381782	1253509	78660941	39626707	39034234	101.52	71685839	35542124	361437
浙江	20060115	18854021	1206094	54426891	27965641	26461250	105.69	49425543	25037320	243882
安徽	19322432	18861956	460476	59500468	30245513	29254955	103.39	56493891	28462853	280310
福建	11971873	11206317	765556	36894217	18981054	17913163	105.96	33397663	16901083	164965
江西	11847841	11542527	305314	44567797	23003521	21564276	106.67	42181417	21600070	205813
山东	30794664	30105454	689210	95792719	48446944	47345775	102.33	89855501	45023357	448321
河南	26404973	25928729	476244	94029939	47493063	46536876	102.05	90028072	45262137	447659
湖北	17253385	16695121	558264	57237727	29391247	27846480	105.55	52745625	26826301	259193
湖南	19029894	18625710	404184	65700762	33776459	31924303	105.80	61911446	31611459	302999
广东	32222752	28630609	3592143	104320459	54400538	49919921	108.98	88979305	45465958	435133
广西	13467663	13151404	316259	46023761	23924704	22099057	108.26	43970320	22733969	212363

企业年报



[企业信用信息](#) [经营异常名录](#) [严重违法企业名单](#) [抽查检查公示](#) [信息公告](#)

请输入企业名称或注册号

搜索

[企业公示信息填报](#) [相关部门信息交换](#) [小微企业名录](#) [电商信用](#)

企业年报

cninfo 巨潮资讯

海量权威数据 首次公开下载
6.28 数据下载器 强势上线

代码/简称/拼音缩写 搜索

信息披露 市场资讯 产品服务

最新公告

代码	简称	公告标题	本网站公告采用了PDF文件格式，阅读需下载并安装Adobe Acrobat Reader软件
300441	鲍斯股份	关于召开2016年第六次临时股东大会的提示性公告	
300173	智慧松德	关于股东部分股票质押的公告	
300526	中潜股份	关于重大资产重组进展的公告	
300142	沃森生物	关于筹划重大资产重组事项停牌的进展公告	
300166	东方国信	关于延期回复深圳证券交易所重组问询函暨继续停牌的进展公告	
300044	赛为智能	关于对外投资设立控股子公司的公告	
300284	苏交科	关于部分限售股份上市流通的提示性公告	
300017	网宿科技	关于控股股东、实际控制人股份减持计划的提示性公告	

工具箱

个股财务数据 历史收盘行情

代码： 类别： 利润表 利润表

起： 2016 止： 2016 下载 免责声明

交易特别提示 2016-09-15 更多 >

增发新股招股书刊登日

300317 珈伟股份

更多增发新股招股书刊登日

基金发行结束日 >

焦点访谈



Focus Group Studio

Observation Room

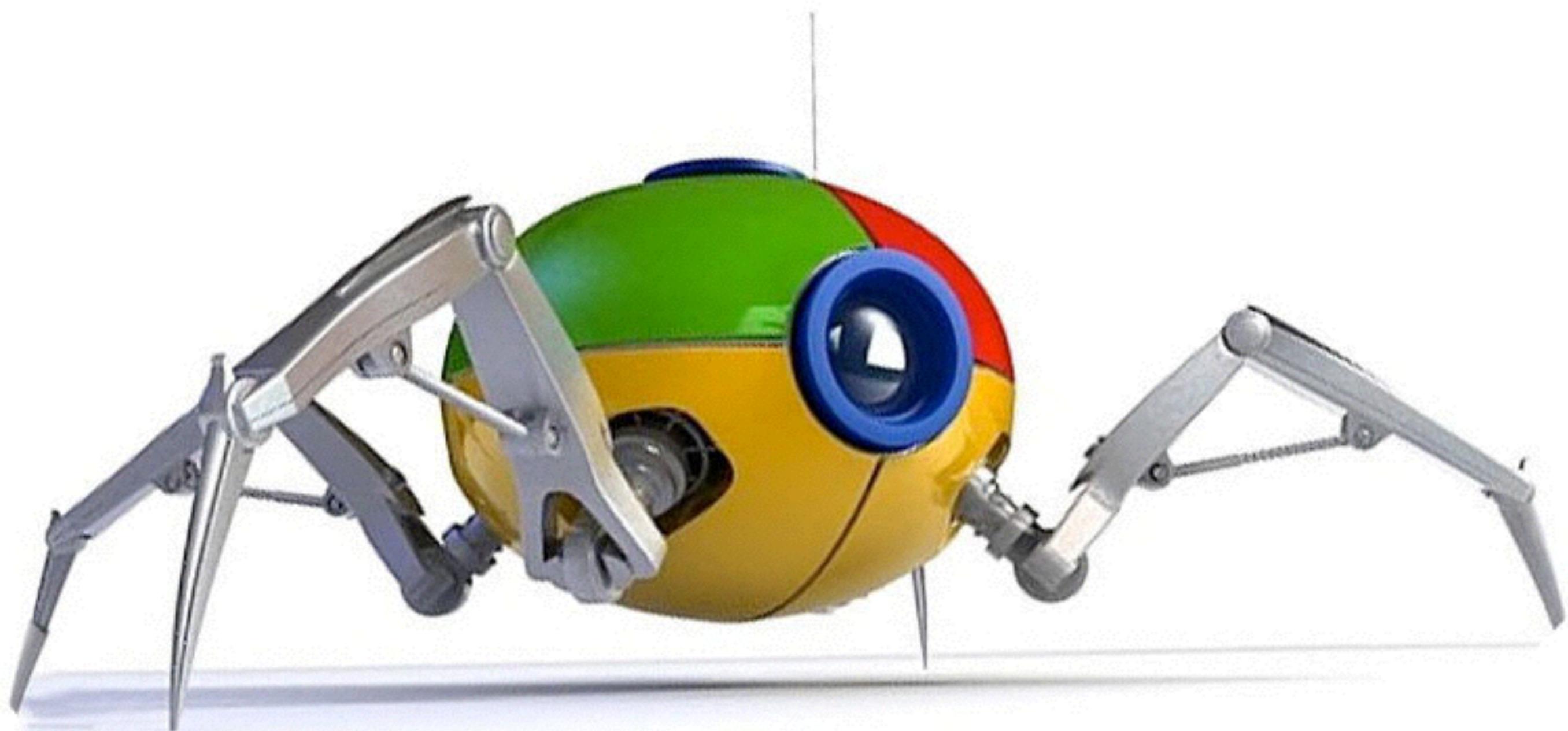
实验法

◆ 实验法

研究人员在一定的控制条件下，改变一个或多个因素，来观察这一变化对其他因素的影响，以找出各因素之间的因果关系。

为了避免实验结果受其他因素的干扰，一方面要尽量控制实验条件，另一方面在实验法中要设置实验组和控制组两组样本来进行观察比较。

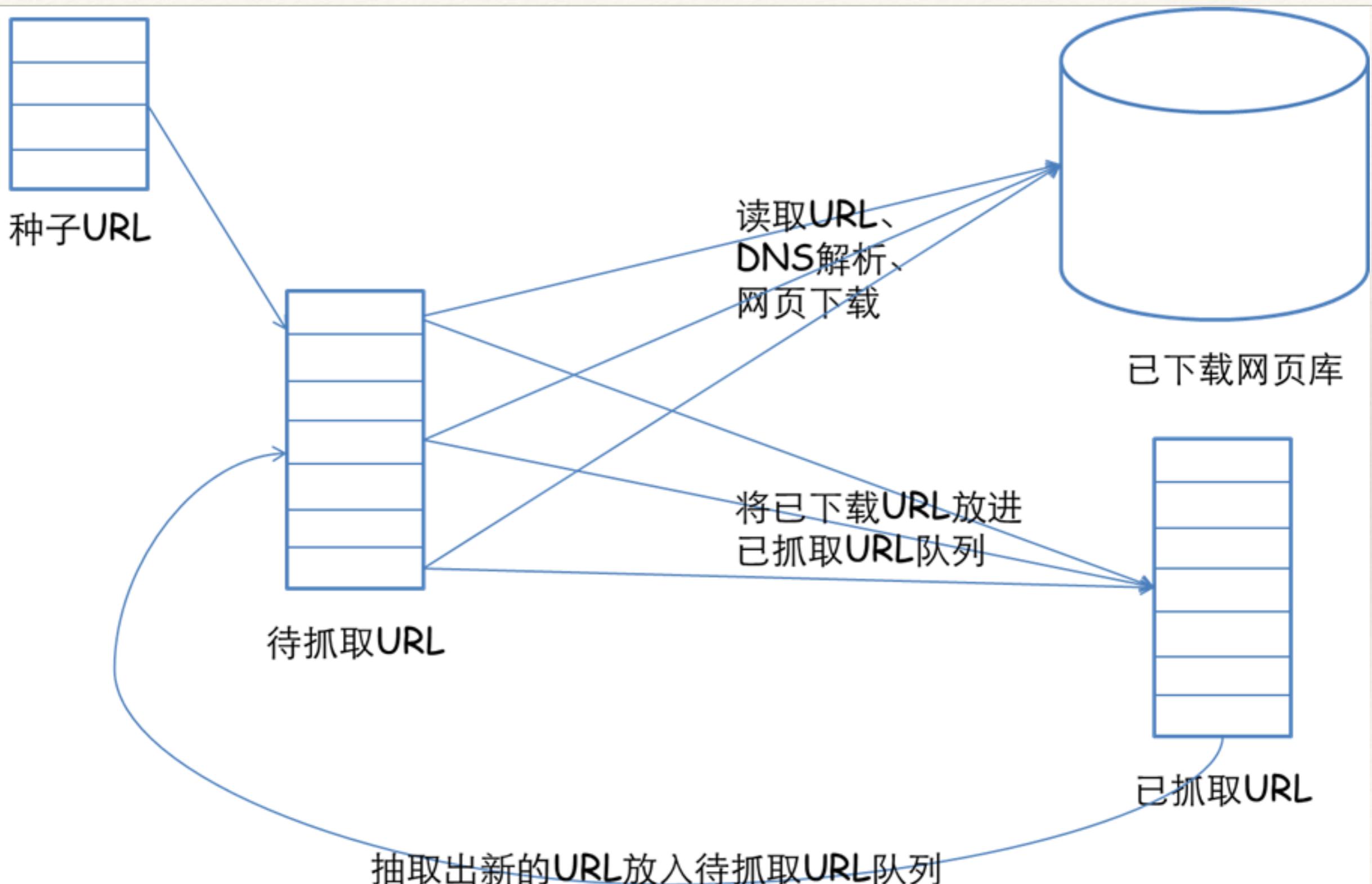
网页爬取数据



主要内容

- ❖ 知识回顾
- ❖ 数据来源
- ❖ 数据爬取方法
- ❖ HTML基础

网页爬虫的基本原理



数据爬取流程

1. 获取URL网页内容
2. 对网页内容进行解析
3. 将解析后的数据存储到本地文件或数据库

中

例子

知乎

搜索你感兴趣的内容...

Q

首页

话题

发现

消息

提问

kkk

个人中心

提问

回答

写文章

草稿

最新动态

设置

热门回答, 来自 心理学 关注话题

什么是假性亲密关系 (Irrelationship) ?

33K

KY教主, 微信公号knowyourself2015。审美、态度、...

原文发表于: 在假性亲密关系中逃避真的亲密感? | 你以为很近, 其实你们很远 ↳ 真正的亲密关系是你与另一个人之间深刻的、自由的、互相回应的联结。然而, 并不是所有的情侣之间都真的有这样的感受和联结。如果我们仔细观察自己身边的人, 也许就是你的父母, 甚... 显示全部

+ 关注问题 2451 条评论 感谢 分享 收藏 没有帮助 举报 禁止转载

热门回答, 来自 牙齿正畸 关注话题

戴牙套正畸为什么那么昂贵? 有没有什么后遗症?

38K

武霖, 口腔科医生

谢邀。口腔正畸学——也就是民众所说的箍牙, 牙套——也就是我所学的专业, 远比大家想象中复杂的多, 也困难的多, 当然, 也神奇的多。如知友“曹杨”所言, 医生卖的是经验, 技术, 手艺。不是那些材料。正畸基本材料——托槽从国产到进口, 方丝弓到自锁, ... 显示全部

+ 关注问题 4014 条评论 作者保留权利

热门回答, 来自 医疗 关注话题

麻醉过后做手术, 在手术中醒来, 会发生什么状况?

10K

郭相治, 麻醉专业毕业生

麻醉医生的职责就是保证病人无痛无知觉的情况下完成手术, 在全身麻醉下, 病人的各种反射不同程度受到抑制, 无法保护自己, 麻醉医师就充当起病人的保护者的角色, 需要时刻关注病人的各项指标, 随时关注出血量, 在紧急情况下要承担起挽救病人生命的任务, 在病... 显示全部

+ 关注问题 2024 条评论 作者保留权利

我的收藏

我关注的问题

邀请我回答的问题

话题广场

公共编辑动态

社区服务中心

版权服务中心

知乎专栏

专栏 · 发现

知乎圆桌

查看全部

空间技术揭秘

还有 3 天结束

主要内容

- ❖ 知识回顾
- ❖ 数据来源
- ❖ 数据爬取方法
- ❖ HTML基础

网页概述

❖ 什么是网页

Web直译过来就是“网”，可以理解为通过超级连接将各种文档连接起来的一个大规模的信息集合。

网页（Web页）实际上就是HTML文件，是一种可以在WWW网上传输，并能被浏览器认识和翻译成页面的文件。WWW是“World wide web”的缩写；HTML则是“HyperText Markup Language”的缩写，意为“超文本标记语言”。超文本就是指页面可以包含图片、链接、音乐等非文字的元素。

网页概述

HTML 是用来描述网页的一种语言。

- HTML 指的是超文本标记语言 (Hyper Text Markup Language)
- HTML 不是一种编程语言，而是一种标记语言 (markup language)
- 标记语言是一套标记标签 (markup tag)
- HTML 使用标记标签来描述网页

网页中的基本元素

1. 文本

文本是人类最重要的信息载体和交流的工具，网页的主体一般以文本为主。在制作网页时，可以根据需要设置文本的字体、字号、颜色以及所需要的其他格式。

2. 图像

图像在网页中可以起到提供信息、展示作品、美化网页以及体现风格等效果。图像可以用作标题、网站标志、网页背景、链接按纽、导航条、网页主图等，网页中使用图像的格式主要有：GIF、JPEG、PNG等格式。

网页中的基本元素

主要图像格式

(1)GIF图像。GIF(graphics interchange format)由Compu-Serve公司1987年6月制订。GIF通常对于卡通、图像、Logo、以及带有透明区域的图像、动化很有作用。

GIF文件格式的扩展名是“.gif”.

(2)JPEG图像。JPEG (joint photographic experts group 联合照片专家组) 是一种特别为照片图像设计的图片压缩处理格式。JPEG文件采用先进的压缩算法，可以保持较好的图像保真度和较高的压缩比。JPEG文件格式的扩展名为“.jpg”.

(3) PNG图像。PNG (portable network graphic) 即可移植网络图形。PNG图像是专门针对Web开发的一种无损压缩图像，它的压缩比要大大超过许多传统的图像无损压缩算法，同时还支持透明背景和动态效果。PNG文件格式的扩展名是“.png”

网页中的基本元素

3. 动画

动画实质上就是动态的图像。在网页中使用动画可以有效地吸引浏览者的注意。由于活动的对象比静止的对象更具有吸引力，因而，网页上通常有大量的动画。网页中使用较多的动画是GIF动画和Flash动画。

4. 声音

声音是多媒体网页的一个重要组成部分。用于网络声音文件格式非常多，常用的是MIDI、MAV、MP3和AIF等。

一般来讲，不要使用声音文件作为网页的背景音乐，那样会影响网页的下载速度。可以在网页中添加一个链接来打开声音文件作为背景音乐，让播放音乐变得可以控制。

浏览器的不同，处理声音文件的方式也会有很大的差异和不一致的地方，最好将声音文件添加到Flash影片中，然后嵌入SWF文件以改善一致性。

网页中的基本元素

5. 视频

在网页中视频文件格式也非常多，常见的有RealPlayer、MPEG、AVI和DivX、flash等。

6. 表格

表格是一种用来控制网页中页面布局的有效方式。为了达到理想的视觉效果，通常都不显示边框，我们所看到的网页如果具有横竖分明的风格，一般都是用表格来辅助布局的。

许多设计人员使用表格对网页进行布局。Dreamweaver提供两种方式来查看和操作表格：标准视图和布局视图。在标准视图中，表格显示为行和列的网格，而布局视图允许创建者在将表格用做基础结构的同时在网页上绘制、移动方框以及调整方框的大小。

网页中的基本元素

7. 表单

表单是一种特殊的网页元素。表单的主要用途是：收集联系信息、接受用户要求、获得反馈意见、设置访问者签名、让浏览者输入关键字去搜索相关网页、让浏览者注册会员或以会员身份登录。登录的用户名、密码、搜索引擎等都是表单。

表单由不同功能的表单元素组成，最简单的表单也要包含一个输入区域和一个提交按钮。站点浏览者填写表单的方式通常是输入文本、选中单选按钮和复选框，以及从下拉列表框中选择选项。根据表单功能和处理方式的不同，通常可以将表单分为用户反馈表单、流言簿表单、搜索表单和用户注册表单等类型。

HTML基本标志

1. 文档标志

<HTML></HTML>。<HTML>标志用于HTML文档的最前面，用来标识HTML文档的开始。而</HTML>标志恰恰相反，它放在HTML文档的最后边，又来标识HTML文档的结束，两个标志必须成对使用。

2. 文件头标志

<head></head>。<head>和</head>构成HTML文档的开头部分，在此标志之间可以使用<title></title>、<script></script>等标志对。<head></head>标志对之间的内容是不会在浏览器的框内显示出来。两个标志必须成对使用。

HTML基本标志

3. 文件主体标志

<body></body>. <body></body>是HTML文档的主体部分，在此标志对之间可以包含<p></p>、<h1></h1>、
、<hr>等众多标志。它们所定义的文本、图像等将会在浏览器的框内显示出来。两个标志必须成对使用。<body>标志中还可以有如表1-1所示的属性。

属性	用途	示例
<body bgcolor="#rrggbb">	设置背景颜色	<body bgcolor="red">红色背景
<body text="# rr ggb">	设置文本颜色	<body text="#0000ff">蓝色文本
<body link="# rr ggb">	设置链接颜色	<body link="blue">链接为蓝色
<body vlink="# rr ggb">	设置已经使用的链接的颜色	<body vlink="#ff0000">

HTML基本标志

4. 文件标题标志

<title></title>。使用过浏览器的人可能都会注意到浏览器窗口最上边的蓝色部分显示的文本信息，那些信息一般是网页的“主题”。要将网页的主题显示到浏览器的顶部其实很简单，只要在<title></title>标志对之间加上显示的文本即可。

注意：<title></title>标志对只能放在<head></head>标志对之间。

下面是一个综合的例子，说明了HTML文档中最基本标志的使用。

```
<HTML>
<head>
<title>显示在浏览器最上边蓝色条中的文本</title>
</head>
<body bgcolor="red" text="blue">
<p>红色背景、蓝色的文本</p>
</body>
</HTML>
```

页面格式标志

1. 段落标志

(1).<p></p>

< p > < / p > 标志对是用来创建一个段落，在此标志对之间加入的文本将按照段落的格式显示在浏览器上。另外，< p > 标志还可以使用 align 属性，它用来说明对齐方式，语法是 < p align = " " > < / p > 。 align 可以是 Left （左对齐）、Center （居中）和 Right （右对齐）三个值中间的一个。

如：< p align = " Center " > < / p > 表示标志对中的文本使用居中对齐方式。

(2).<per></per>

< per > < / per > 标志队有来对文本进行预处理操作。

页面格式标志

2.换行标志

是一个很简单的标志，它没有结束标志，因为它是用来创建一个回车换行的。在
的使用上面还有一定的技巧，如果把
加在<p></p>标志对的外边,将创建一个很大的回车换行,即
前面和后面的文本的行与行之间的距离很大,若放在<p></p>的里面,则
前面和后面的文本行与行之间的距离比较小.

页面格式标志

3. 列表标志

(1) <dl></dl>、<dt></dt>、<dd></dd>

<dl></dl>用来创建一个普通的列表,<dt></dt>用来创建列表中的上层项目,<dd></dd>用来创建列表中最下层项目,<dt></dt>和<dd></dd>都必须放在<dl></dl>标志对之间。

下面是一个创建普通列表的例子

```
<html>
  <head>
    <title>一个普通的列表</title>
  </head>
  <body style="text-align: center;">
    <dl>
      <dt>中国城市</dt>
      <dd>北京</dd>
      <dd>上海</dd>
      <dd>广州</dd>
    </dl>
  </body>
</html>
```

<dt>美国城市</dt>
<dd>华盛顿</dd>
<dd>芝加哥</dd>
<dd>纽约</dd>
</dl>
</body>
</html>

页面格式标志

标志	含义
<table>	最外层，创建一个表格
<tr>	创建一行
<td>要输出的文本只能放在此处</td> <td>要输出的文本只能放在此处</td> <td>要输出的文本只能放在此处</td>	创建一个单元格（这里总共创建了三个单元格）
</tr>	行末尾
</table>	最外层

表格标志

<th></th>

<th></th>标志对用来设置表格头，文字通常是黑体、居中。

表格标志

```
<html>

<head>

<title>表格标志的综合示例</title>

</head>

<body>

<table border="1" width="80%" bgcolor="#e8e8e8" cellpadding="2" bordercolor="30000ff">

<tr>

<th width="33%" colspan="2" valign="bottom">意大利</th>

<th width="36%" colspan="2" valign="bottom">英格兰</th>

<th width="36%" colspan="2" valign="bottom">西班牙</th>

<tr>

<td width="16%" align="center">AC米兰</td>

<td width="16%" align="center">佛罗伦莎</td>

<td width="17%" align="center">曼联</td>

<td width="17%" align="center">纽卡斯尔</td>

<td width="17%" align="center">巴塞罗那</td>

<td width="17%" align="center">皇家社会</td>
```

表格标志

```
<tr>
<td width="16%" align="center">尤文图斯</td>
<td width="16%" align="center">桑普多利亚</td>
<td width="17%" align="center">利物浦</td>
<td width="17%" align="center">阿申纳</td>
<td width="17%" align="center">皇家马德里</td>
<td width="17%" align="center">.....</td>
<tr>
<td width="16%" align="center">拉奇奥</td>
<td width="16%" align="center">国际米兰</td>
<td width="17%" align="center">切尔西</td>
<td width="17%" align="center">米德尔斯堡</td>
<td width="17%" align="center">马德里竞技</td>
<td width="17%" align="center">.....</td>
</table>
```

小结

- 数据来源方式
- 数据爬取原理
- 网页基础知识

下次内容

- ❖ 利用R语言编写爬虫程序

