北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

大数据时代的管理

Management in Big Data Era

马宝君 博士 讲师

经济管理学院
电子商务中心

2014.09.29

上次课程小结

- 数字化生活：大数据时代的体现
- “工业革命”与技术时代沿革：大数据时代的背景
- “大数据”的相关概念

大数据概念和特征

2.“大数据”的特征

回顾：大数据理念（知著、见微、晓意）

“大数据”的数据类型

- 大数据不仅仅体现在数量大，也体现在数据类型多
- 仅有不到5%左右的数据属于结构化数据，超过95%的数据属于非结构化数据
- 下表是按照数据结构分类

类别	概念/特点	示例
结构化数据	一般存储在数据库中、可以用二维表结构来逻辑表达的数据	客户号 | 客户姓名 | 交易额 | 所购产品 200048901 | 张伟 | 1000 | 微波炉 200057903 | 李东 | 456 | 烤炉
半结构化数据	介于完全结构化数据和完全无结构的数据之间，格式较为规范，一般为纯文本数据，可以通过某种方式解析得到每项的数据	• XML(Extensible Markup Language)文档 • JSON(JavaScript Object Notation) • 日志文件 • 点击流
无结构的非结构化数据	非纯文本类数据，没有标准格式，无法直接解析出相应的值	• Web网页 • 电子邮件 • 富文本文档（RTF） • 富媒体文件 • 即时消息或事件数据(如微博、微信)

半结构化数据类型的一些示例

- XML文档

```
<?xml version="1.0"?>
<book category="COOKING">
  <title lang="en">Everyday Italian</title>
  <author>Giada De Laurentiis</author>
  <year>2005</year>
  <price>30.00</price>
</book>
```

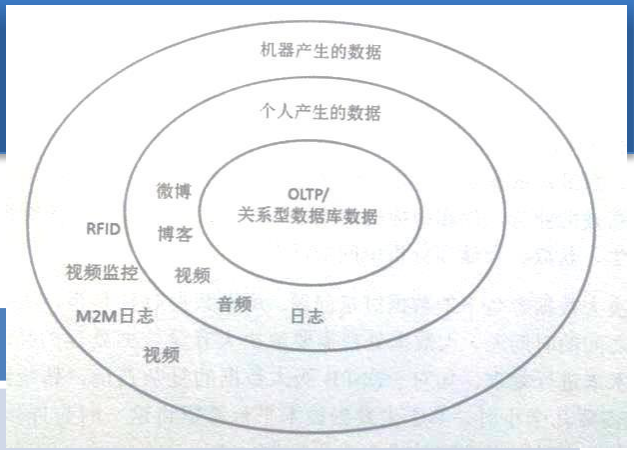
- JSON：基于JavaScript的轻量级的数据交换格式

```
{
"employees": [
{ "firstName":"Bill" , "lastName":"Gates" },
{ "firstName":"George" , "lastName":"Bush" },
{ "firstName":"Thomas" , "lastName":"Carter" }
]
}
```

- 日志文件：用于记录业务或信息系统内执行的自动功能的详细信息，最常见的是Web日志

58.61.164.141 – - [22/Feb/2010:09:51:46 +0800] "GET /reference-and-source/weblog-format/ HTTP/1.1" 206 6326 " http://www.google.cn/search?q=webdataanalysis" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"

“大数据”的数据类型（续）



- 下表是按照产生主体分类

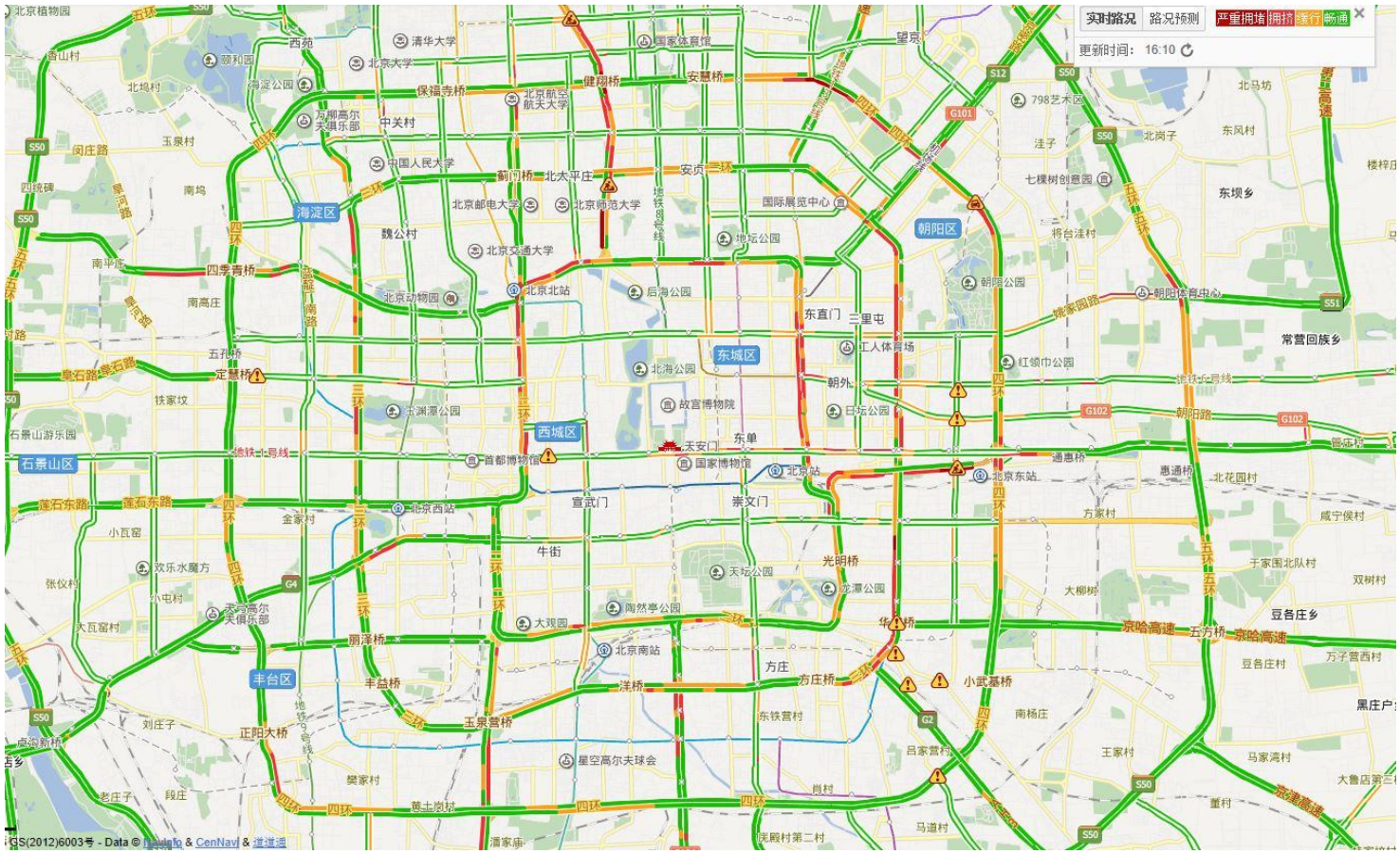
类别	示例
最里层： 少量企业应用产生的数据	• 关系型数据库中的数据 • 数据仓库中的数据
次外层： 大量人产生的数据	• Twitter，微博，微信（文字、图片、音频、视频等） • 博客、评论、图片和视频分享 • 企业博客、企业微博、企业微信 • 电子商务在线交易的日志数据、供应商交易的日志数据 • 呼叫中心的评论、留言或电话投诉等 • 企业应用相关的评论数据
最外层： 巨量机器产生的数据	• 应用服务器日志（Web站点、游戏） • 传感器数据（天气、水、智能电网）、 • 图像和视频（摄像头、监控头的视频、音频数据） • RFID(射频识别技术)、二维码或者条形码扫描的数据

大数据的潜在价值

- 总体来看，大数据由于数据容量大、数据种类多、信息量大，可以从中获取的知识多，能够发挥的潜在价值也很大
- 然而，大数据的价值密度相对较低
  - 存储和计算PB级的数据是需要非常高的成本的，大数据虽然看起来很美，但是价值密度却远远低于传统关系型数据库中已经有的那些数据
  - “如果用石油行业来类比大数据分析，那么在互联网金融领域甚至整个互联网行业中，最重要的并不是如何炼油（分析数据），而是如何获得优质原油（优质元数据）”。
  - 以股市为例，真正有价值的数据都只会在很小范围内（例如庄家之间）传播，极少可能会流落到互联网上来，所以你如果想去只靠分析微博上网民对股票涨跌的评论来做行情预测的话，真的是要小心了。
  - Big-Data-As-a-Service
- 大数据中往往包含着大量的重复信息或者对所关注问题没有意义的信息
- 大数据的价值挖掘是“沙里淘金”和“海里捞针”

大数据的速度

- 数据创建、处理和分析的速度，由数据从客户端采集、装载并流动到处理器和存储设备、在处理器和存储设备中进行计算的速度决定
- 流信息、实时数据、连续商务
- 批处理、离线处理 => 实时流处理



大数据的“4V”特征

体量Volume	**海量数据：**比传统数据仓库增长速度快10倍到50倍
多样性Variety	**多源异构性：**不同形式（文本、图像、视频数据）、无模式或者模式不明显、不连贯语法或句义
价值密度Value	**低价值密度：**大量的不相关信息、需深度分析
速度Velocity	**实时分析：**流信息、即时需求、连续商务

大数据时代的变革

1. 大数据时代的思维变革

大数据时代的思维变革

- 更多
  - 不是随机样本，而是**全体数据**
- 更杂
  - 不是精确性，而是**混杂性**
- 更好
  - 不是因果关系，而是**相关关系**

（1）不是随机样本，而是全体数据

- **利用所有的数据，而不再仅仅依靠一小部分数据**
- **统计学**
  - 用尽可能少的数据来证实尽可能重大的发现
- **小数据时代的随机采样，最少的数据获得最多的信息**
  - 人口普查(Census / Censere 推测，估算)的可行性、时效性
  - 选取最具有代表性的样本：随机性
  - 采样分析的精确性随着采样随机性的增加而大幅提高，但与样本数量的增加关系不大
  - 随机采样是在不可收集和分析全部数据的情况下的选择，存在许多固有缺陷
  - 社会科学过去的很多研究依赖于样本分析、调查问卷