| 学　　号： | 0121310880330 |
| --- | --- |

# 武漢理工大學

# 《专业教育》

# 课 程 报 告

| | |
| --- | --- |
| 题　　目 | 浅析大数据时代下的软件工程中的“取舍”之道 |
| 学　　院 | 计算机科学与技术学院 |
| 专　　业 | 软件工程 |
| 班　　级 | 软件 sy1301 |
| 姓　　名 | 郑文博 |
| 指导教师 | 饶文碧 |

2015 年 1 月 22 日

# 浅析大数据时代下的软件工程中的“取舍”之道

郑文博
计算机科学与技术学院 软件工程 0121310880330 13554068450

**摘要：**文章在阐述大数据时代的现状的基础上，叙述了软件工程专业的专业内涵以及专业内容，提出了在大数据时代的背景下，软件工程专业的未来发展方向以及所面临的挑战。

**关键词：**大数据时代；软件工程；现状；挑战；前景

在人类历史长河中，即使是在现代社会日新月异的发展中，人们还主要是依赖抽样数据、局部数据和片面数据，甚至在无法获得实证数据的时候纯粹依赖经验、理论、假设和价值观去发现未知领域的规律。因此，人们对世界的认识往往是表面的、肤浅的、简单的、扭曲的或者是无知的。大数据时代的来临使人类第一次有机会和条件，在非常多的领域和非常深入的层次获得和使用全面数据、完整数据和系统数据，深入探索现实世界的规律，获取过去不可能获取的知识，得到过去无法企及的商机。

## 1 我们所处的时代——大数据时代

### 1.1 大数据时代的简要介绍

大数据技术的出现是随着计算机技术、互联网技术和图像(视频)采集技术的快速发展而发展起来的。近几年全球每天产生的图像和视频数据都以 PB 为单位增长，而我们对数据处理的实时性、准确性的要求却在不断提高，因此，主要包括分布式缓存、分布式运算、分布式文件系统、分布式数据库的大数据解决方案应运而生。

大数据技术(big data)，或称巨量资料，指的是所涉及的资料量规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理、并整理成为帮助企业经营决策更积极目的的资讯。在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》中大数据指不用随机分析法（抽样调查）这样的捷径，而采用所有数据进行分析处理。大数据的 4V 特点：Volume（大量）、Velocity（高速）、Variety（多样）、Value（价值）。

随着云时代的来临，大数据（Big data）也吸引了越来越多的关注。《著云台》的分析师团队认为，大数据（Big data）通常用来形容一个公司创造的大量非结构化数据和半结构化数据，这些数据在下载到关系型数据库用于分析时会花费过多时间和金钱。大数据分析常和云计算联系到一起，因为实时的大型数据集分析需要像 MapReduce 一样的框架来向数十、数百或甚至数千的电脑分配工作。

大数据需要特殊的技术，以有效地处理大量的容忍经过时间内的数据。适用于大数据的技术，包括大规模并行处理（MPP）数据库、数据挖掘电网、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统。

## 1.2 研究大数据的意义

大数据，尤其是网络大数据，存在数据规模巨大、数据关联复杂、数据状态演变等显著特征.其规模和复杂度的增长远远超出了符合摩尔定律增长的软硬件计算能力.这种矛盾为我们带来了问题，同时也存在着巨大机遇和挑战。

## 2 大数据时代的问题范式

大数据具有 4V 的特点，其特点如图 1 所示。其具体问题范式分为以下四个部分。

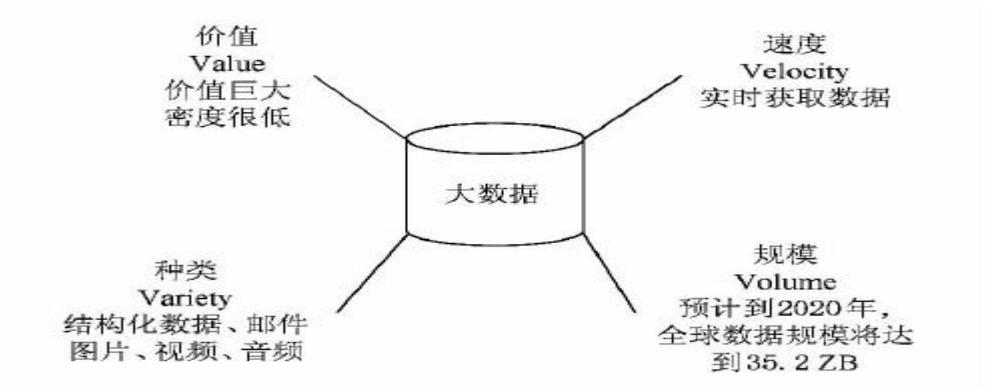


图 1 “4V”特点[1]

## 2.1 数据复杂性

大数据处理面临三个主要因素:大容量数据、多格式数据和速度。

大容量数据:计算机技术、互联网技术和图像(视频)采集技术的快速发展导致海量的数据被不断产生，这些数据对 IT 系统带来了极大的挑战，其存储和安全以及在未来访问和使用这些数据已成为难点。

多格式数据:由于计算机的应用越来越广泛，其处理的信息也越来越丰富，因此，也产生了越来越多的数据格式，不同的数据格式因其包含的内容不同，其结构和存储方式也多种多样。从简单到复杂，光文本格式就有几十种，更别说图像数据、视频数据、专用数据采集仪器采集的数据、各种传感器采集的数据、互联网上各种应用产生的数据，这些数据都对应几种甚至几十种不同的格式。大数据处理技术要对这些不同格式的数据采取不同的处理方法，合理高效地处理好不同的数据格式是大数据处理技术面临的另一个问题。

速度:高速的处理器和优化的算法可以提高数据处理速度，但在对大数据的处理问题上仅通过这两种途径是难以胜任的。

## 2.2 数据计算的复杂性

对于数据计算的处理，通常有两种——批处理和流处理。

流处理的基本理念是数据的价值会随着时间的流逝而不断减少，因此尽可能快地对最新的数据作出分析并给出结果是所有流数据处理模式的共同目标。流处理的处理模式将数据视为流，源源不断的数据组成了数据流.当新的数据到来时就立刻处理并返回所需的结果.图 1 是流处理中基本的数据流模型:

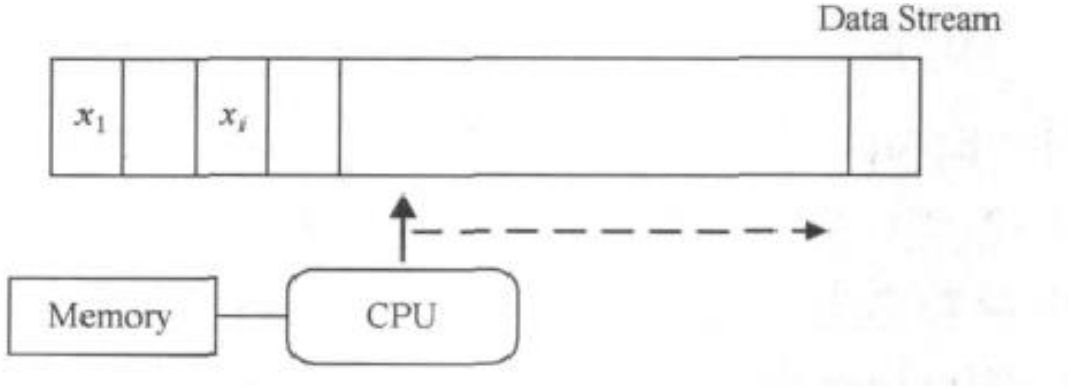


**图 2 基本的数据流模型[4]**

批处理是一种简化的脚本语言，也称作宏。它应用于 DOS 和 Windows 系统中，它是由 DOS 或者 Windows 系统内嵌的命令解释器（通常是 COMMAND. COM 或者 CMD.EXE)解释运行。类似于 Unix 中的 Shell 脚本。批处理文件具有.bat 或者.cmd 的扩展名，其最简单的例子，是逐行书写在命令行中会用到的各种命令。更复杂的情况，需要使用 if，for，goto 等命令控制程序的运行过程，如同 C，Basic 等中高级语言一样。如果需要实现更复杂的应用，利用外部程序是必要的，这包括系统本身提供的外部命令和第三方提供的工具或者软件。批处理文件，或称为批处理程序，是由一条条的 DOS 命令组成的普通文本文件，可以用记事本直接编辑或用 DOS 命令创建，也可以用 DOS 下的文本编辑器 Edit.exe 来编辑。在“命令提示”下键入批处理文件的名称，或者双击该批处理文件，系统就会调用 Cmd.exe 运行该批处理程序。一般情况下，每条命令占据一行；当然也可以将多条命令用特定符号（如：&;、&&;、|、||等）分隔后写入同一行中；还有的情况就是像 if、for 等较高级的命令则要占据几行、几十甚至几百行的空间。系统在解释运行批处理程序时，首先扫描整个批处理程序，然后从第一行代码开始向下逐句执行所有的命令，直至程序结尾或遇见 exit 命令或出错意外退出。

## 2.3 数据处理系统的复杂性

现实网络信息的爆炸对数据存储与处理的效能提出了挑战:Facebook 用户近 10 亿，每天产生近 300TB 日志数据，淘宝超过 4 亿会员，每天交易量数千万，每天产生 30TB 的数据分析产品，因此存在数据快速增长与数据中心扩容周期相对缓慢的矛盾，数据业务深度的不断加强与数据处理性能相对不足的矛盾。Uartner 等信息咨询机构预计:2020 年前后，60%以上的数据将因无法有效存储而失去价值或丢失。大数据而临的首要问题就是海量数据的存储。

传统的数据分析方法已经满足不了大数据时代的要求，传统数据分析是指用适当的统计方法对收集来的大量第 1 手资料和第 2 手资料进行分析，把隐没在一大批看来杂乱无章的数据中的信息集中、萃取和提炼出来，找出所研究对象的内在规律，以求最大化地开发数据资料的功能，发挥数据的作用.数据分析对国家制定发展计划，对企业了解客户需求、把握市场动向都有巨大的指导作用。大数

据分析可以视为对一种特殊数据的分析，因此很多传统的数据分析方法也可用于大数据分析.以下是可用于大数据分析的传统数据分析方法，这些方法源于统计学和计算机科学等多个学科。

### 2.4 基于数据的学习复杂性

随着大数据时代的到来，大数据逐渐成为学术界和产业界的热点，已在很多技术和行业广泛应用，从大规模数据库到商业智能和数据挖掘应用;从搜索引擎到推荐系统;推荐最新的语音识别、翻译等。大数据算法的设计、分析和工程涉及很多方面，包括大规模并行计算、流算法、云技术等.由于大数据存在复杂、高维、多变等特性，如何从真实、凌乱、无模式和复杂的大数据中挖掘出人类感兴趣的知识，迫切需要更深刻的机器学习理论进行指导。

传统机器学习的问题主要包括如下 4 个方面：

1)理解并且模拟人类的学习过程；

2)针对计算机系统和人类用户之间的自然语言接口的研究；

3)针对不完全的信息进行推理的能力，即自动规划问题；

4)构造可发现新事物的程序。

传统机器学习面临的一个新挑战是如何处理大数据.目前，包含大规模数据的机器学习问题是普遍存在的，但是，由于现有的许多机器学习算法是基于内存的，大数据却无法装载进计算机内存，故现有的诸多算法不能处理大数据.如何提出新的机器学习算法以适应大数据处理的需求，是大数据时代的亟待解决的问题。

## 3 软件工程中的“取舍”之道

### 3.1 数据挖掘新模式

主要考虑如何提出面向网络数据界的结构规则度量与网络模式表达，摆脱现有方法在大数据处理而临的时空挑战.当前主要的途径有:搜索方法(计算每条传播路径的可能性，即时间换空间)和判定方法(记录每条信息的传播路径，即空间换时间)，但这两种方法存在精确计算不可行、近似计算难以保证效率与精度的问题;基于结构规则性的识别方法:依靠结构化计算，寻找数据空间的新度量，在时间和空间上一致性约简。

在数据的复杂模式处理和数据的网络化效应方面，已经有了一些成果，如图 3 所示。

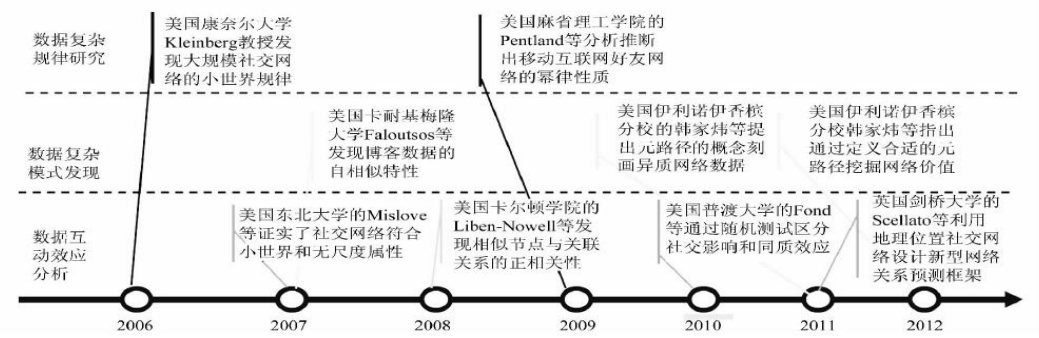


**图 3 数据挖掘新模式成果图[2]**

## 3.2 Hadoop 平台的出现和 Map-reduce 的应用

Hadoop 是一个能够对大数据进行分布式处理的基础架构平台.其架构底层是 Hadoop 分布式文件系统(HDFS),主要负责存储 Hadoop 集群中所有节点上的文件.HDFS 由一组特定节点构成，包括:1 个 NameNode 结点负责管理文件系统名称空间并控制外部客户机的访问;多个 DataNode 用于存储文件被切割成的多个 Block，同时负责响应来自 HDFS 客户机的读写请求。

HDFS 的上层是 Map-Reduce 执行引擎，该引擎由 1 个单独运行在主节点上的 JobTracker 和多个运行在集群节点上的 TaskTracker 组成。JobTracker 负责协调调度在 TaskTracker 上运行
的任务.

Map-Reduce 是一个软件架构，主要用于大规模数据集的分布式计算及任务处理一个 Map-Reduce 任务过程主要包括 2 个阶段:映射(map)阶段和化简(reduce)阶段.每个阶段都以键一值对<key,value>作为输入和输出。map 函数接受一组数据并将其转换为<key, value>列表，传递给 reduce 函数，reduce 函数接受列表后根据键缩小<key,value>列表。

## 3.3 大数据时代的数据分析方法

①Bloom Filter:布隆过滤器，其实质是一个位数组和一系列 Hash 函数。布隆过滤器的原理是利用位数组存储数据的 Hash 值而不是数据本身，其本质是利用 Hash 函数对数据进行有损压缩存储的位图索引。其优点是具有较高的空间效率和查询速率，缺点是有一定的误识别率和删除困难.布隆过滤器适用于允许低误识别率的大数据场合。

②Hashing:散列法，也叫做 Hash 法，其本质是将数据转化为长度更短的定长的数值或索引值的方法.这种方法的优点是具有快速的读写和查询速度，缺点是难以找到一个良好的 Hash 函数。

③索引:无论是在管理结构化数据的传统关系数据库，还是管理半结构化和非结构化数据的技术中，索引都是一个减少磁盘读写开销、提高增删改查速率的有效方法.索引的缺陷在于需要额外的开销存储索引文件，且需要根据数据的更新而动态维护。

④Trie 树:又称为字典树，是 Hash 树的变种形式，多被用于快速检索，和

词频统计.Trie 树的思想是利用字符串的公共前缀，最大限度地减少字符串的比较，提高查询效率。

⑤并行计算:相对于传统的串行计算，并行计算是指同时使用多个计算资源完成运算.其基本思想是将问题进行分解，由若干个独立的处理器完成各自的任务，以达到协同处理的目的.目前，比较典型的并行计算模型有 MPI，MapReduce，Dryad 等。

### 3.4 深度学习

机器学习算法进入大数据时代后，面临诸多考验。大数据时代下的复杂性，建模的复杂性，以及确定性的概率性分析等等棘手关键问题。

深度学习旨在解决千层模型特征设计和统计分类两步走的学习范式中存在的问题:

1)特征设计费时费力，普适性没有保障；

2)生物视觉系统等参照系支持两者（特征设计和统计分类）合二为一考虑；

3)解决过拟合的现有方法存在缺陷.

深度学习模拟生物视觉系统的层级抽象结构，并充分利用大量无监督数据中隐含的丰富信息.逐层的无监督学习方式能够实现层叠的学习结构，实现从原始特征开始层级抽象，逐级提高更高级语义信息的特征学习过程.目前的深度模型在应用领域产生了一定成功的应用，例如百度和 Google 的类似图像搜索和推荐等。具体的成果图如图 4 所示。

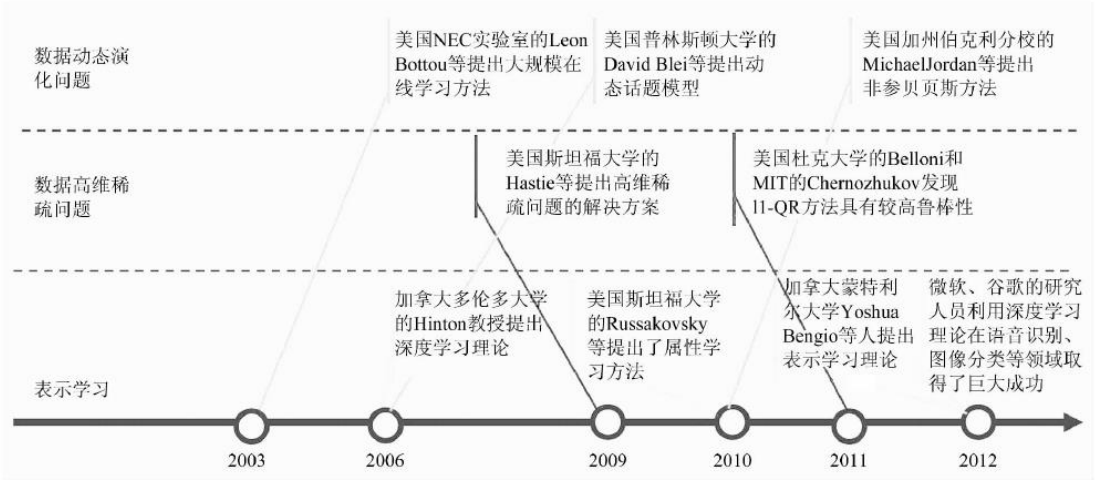


图 4 深度学习成果图[3]

## 4 在大数据时代下的我们——未来的软件工程从业者

### 4.1 思维转变

①分析数据时，要尽可能地利用所有数据，而不只是分析少量的样本数据。

②相比于精确的数据，我们更乐于接受纷繁复杂的数据.

③我们应该更为关注事物之间的相关关系，而不是探索因果关系.

④大数据的简单算法比小数据的复杂算法更为有效.

⑤大数据的分析结果将减少决策中的草率和主观因素，数据科学家将取代“专家”.

### 4.2 安全机制

大数据时代，数据的隐私问题包括两个方面:一方面是个人隐私的保护，随着数据采集技术的发展，在用户无法察觉，个人的兴趣、习惯、身体特征等隐私信息可以被更容易地获取;另一方面，即使得到用户的许可，个人隐私数据在存放、传输和使用的过程中，也有被泄露的风险.大数据的分析能力导致看似简单的信息可能会被挖掘出其中的隐私,因此而对大数据时代的隐私保护将成为新的命题。

### 4.3 交叉学科

在大数据时代，学科之间的联系性更加广泛。大数据不仅促进了云计算、物联网、计算中心、移动网络等技术的充分融合，还催生了许多学科的交叉融合.大数据的发展，既需要立足于信息科学，探索大数据的获取、存储、处理、挖掘和信息安全等创新技术与方法,也需要从管理的角度探讨大数据对于现代企业生产管理和商务运营决策等方而带来的变革与冲击.而在特定领域的大数据应用，更需要跨学科人才的参与[6]。

### 4.4 面向数据

程序是数据结构和算法，而数据结构就是存储数据的。在程序设计的发展历程中,也可以看出数据的地位越来越重要。在逻辑比数据复杂的小规模数据时代，程序设计以而向过程为主;随着业务数据的复杂化,催生了面向对象的设计方法。如今，业务数据的复杂度已经远远超过业务逻辑，程序也逐渐从算法密集型转向数据密集型。可以预见，一定会出现面向数据的程序设计方法，如同面向对象一样，在软件工程、体系结构、模式设计等方而对 IT 技术的发展产生深远的影响[5]。

## 参考文献：

[1]《网络大数据:现状与展望》，王元卓、靳小龙、 程学旗，计算机学报
[2]《大数据的一个重要方面:数据可用性》，李建中、刘显敏，计算机研究与发展
[3]《大数据分析——RDBMS 与 MapReduce 的竞争与共生》，覃雄派、王会举、杜小勇，王珊，软件学报
[4]《大数据时代下数据分析理念的辨析》，朱建平、章贵军、刘晓葳，统计研究
[5]《大数据应用的现状与展望》，张引、陈敏、廖小飞，计算机研究与发展
[6]《大数据时代的机遇与挑战》，邬贺铨，求是

# 《专业教育》成绩评定表

班级：软件 sy1301　　姓名：郑文博　　学号：　0121310880330

| 序号 | 评分项目 | 满分 | 实得分 |
|---|---|---|---|
| 1 | 学习态度认真、遵守纪律 | 10 | |
| 2 | 选题与课程内容符合程度 | 10 | |
| 3 | 报告格式的规范性 | 10 | |
| 4 | 报告结构层次、论述的条理性 | 20 | |
| 5 | 正文内容的准确性<br>（概念清晰，分析与归纳准确） | 20 | |
| 6 | 课程报告的使用价值 | 30 | |
| | | 总得分<br>/等级 | |
| 评语： | | | |

注：最终成绩以五级分制记。优（90-100 分）、良（80-89 分）、中（70-79 分）、及格（60-69 分）、60 分以下为不及格

指导教师签名：

2015 年　月　日