

University of South China

**课程作业**

|  |  |
| --- | --- |
| **题 目** | **大数据技术研究综述** |
| **学 院** | **计算机学院** |
| **课 程** | **计算机科研写作** |
| **学 号** | **201620810172** |
| **学生姓名** | **孙 溢** |

2017年6月20日

摘要

大数据的产生给海量信息处理技术带来新的挑战。为了更全面深入地了解大数据的内涵，首先从大数据若干个版本的定义出发，对大数据的基本概念做一个简单了解，然后详细阐述大数据处理流程，以及包括云计算、分布式文件系统、分布式并行数据库和大数据可视化在内的大数据技术，并在其中穿插介绍一些工具。最后总结大数据的为例发展方向。目录

[0 引言 1](#_Toc20336)

[1 大数据的基本概念 1](#_Toc25219)

[2大数据处理流程 2](#_Toc23667)

[2.1数据采集 2](#_Toc8601)

[2.2数据处理与集成 2](#_Toc5108)

[2.3数据分析 2](#_Toc10095)

[2.4数据解释 3](#_Toc23940)

[3 大数据关键技术 3](#_Toc6438)

[3.1云计算和MapReduce 3](#_Toc32333)

[3.1.1云计算 3](#_Toc12702)

[3.1.2 MapReduce 4](#_Toc11913)

[3.2分布式文件系统 4](#_Toc30111)

[3.3分布式并行数据库 4](#_Toc8074)

[3.4开源实现平台Hadoop 5](#_Toc14900)

[3.5大数据可视化 5](#_Toc7976)

[4 大数据的研究与发展方向 6](#_Toc5859)

[4.1关系数据库和非关系数据库的融合 6](#_Toc11163)

[4.2数据的不确定性与数据质量 6](#_Toc27308)

[4.3跨领域的数据处理方法的可移植性 7](#_Toc8955)

[4.4大数据的预测性作用日益凸显 7](#_Toc4596)

[5结语 7](#_Toc21512)

[参考文献 8](#_Toc7362)

# 0 引言

大数据这个概念并不是近几年才有的，早在1980年，著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，将大数据热情地赞颂为“第三次浪潮的华彩乐章”。在20世纪80年代我国已经有一些专家学者谈到了海量数据的加工和管理，但是由于计算机技术和网络技术的限制大数据未能引起足够的重视，它蕴藏的巨大信息资源也暂时隐藏了起来。

随着移动互联网、物联网和云计算技术的迅速发展，开启了移动云时代的序幕，大数据（big data）也越来越吸引人们的视线。正如1982年世界预测大师、未来学家约翰· 奈斯比特（John Naisbitt）在他的著作《Mega trends: Ten new directions transforming our live》一书中所提到的：“我们现在大量生产信息，正如过去我们大量生产汽车一样”、“人类正被信息淹没，却饥渴知识”，等等诸如此类的预言均在当下得到了充分的证实，这也恰恰说明，世界正处于一个信息爆炸的时代。

借助Internet的高速发展、数据库技术的成熟和普及、高内存高性能的存储设备和存储介质的出现，人类在日常学习、生活、工作中产生的数据量正以指数形式增长，呈现“爆炸”状态。“大数据问题”（Big Data Problem）就是在这样的背景下产生的，成为科研学术界和相关产业界的热门话题，并作为信息技术领域的重要前沿课题之一，吸引着越来越多的科学家研究大数据带来的相关问题。

# 1 大数据的基本概念

大数据本身就是一个很抽象的概念，提及大数据很多人也只能从数据量上去感知大数据的规模，如：百度每天大约要处理几十PB的数据；Facebook每天生成300TB以上的日志数据；据著名咨询公司IDC的统计，2011年全球被创建和复制的数据总量为1.8ZB，但仅仅是数据量并不能区分大数据与传统的海量数据的区别。

到目前为止，对大数据的概念并没有一个统一的定论：

在2008年《Science》杂志出版的专刊中，大数据被定义为“代表着人类认知过程的进步，数据集的规模是无法在可容忍的时间内用目前的技术、方法和理论去获取、管理、处理的数据”。

比较有影响力的Gartner公司也给出了大数据的定义：大数据是高容量、高生成速率、种类繁多的信息价值，同时需要新的处理形式去确保判断的作出、洞察力的发现和处理的优化。这种定义不仅是数据规模大，更重要的是如何从这些动态快速生成的数据流或数据块中获取有用的具有时效性价值的信息，但是这些数据类型众多，结构化、半结构化、非结构化的数据对已有的数据处理模式带来了巨大的挑战，其中也体现了大数据在3V基础上发展的4V定义。

4V定义即volume，variety，velocity，value，关于第4个V的说法并不统一，国际数据公司（International Data Corporation，IDC）认为大数据还应当具有价值性(value)，大数据的价值往往呈现出稀疏性的特点；而IBM认为大数据必然具有真实性(veracity)，这样有利于建立一种信任机制，有利于领导者的决策。

百度百科对大数据的定义是：大数据（big data），指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

大数据的科学家Rauser提到一个简单的定义：大数据就是超过了任何一个计算机处理能力的庞大数据量。

# 2大数据处理流程

从大数据的特征和产生领域来看，大数据的来源相当广泛，由此产生的数据类型和应用处理方法千差万别。但是总的来说，大数据的基本处理流程大都是一致的。目前，中国人民大学网络与移动数据管理实验室（WAMDM）开发了一个学术空间“Scholar Space”，从计算机领域收集的相关文献可以总结出大数据处理的一般流程。

在此基础上，笔者认为大数据的处理流程基本可划分为数据采集、数据处理与集成、数据分析和数据解释4个阶段。整个处理流程即经数据源获取的数据，因为其数据结构不同（包括结构、半结构和非结构数据），用特殊方法进行数据处理和集成，将其转变为统一标准的数据格式方便以后对其进行处理，然后用合适的数据分析方法将这些数据进行处理分析，并将分析的结果利用可视化等技术展现给用户，这就是整个大数据处理的流程。

## 2.1数据采集

大数据的“大”，原本就意味着数量多、种类复杂，这从大数据的特征volume和variety就可以看出来。因此，通过各种方法获取数据信息便显得格外重要。数据采集是大数据处理流程中最基础的一步，目前常用的数据采集手段有传感器收取、射频识别（RFID）、数据检索分类工具如百度和谷歌等搜索引擎，以及条形码技术等。并且由于移动设备的出现，如智能手机和平板电脑的迅速普及，使得大量移动软件被开发应用，社交网络逐渐庞大，这也加速了信息的流通速度和采集精度。

## 2.2数据处理与集成

数据的处理与集成主要是完成对于已经采集到的数据进行适当的处理、清洗去噪以及进一步的集成存储。

根据前文所述，大数据特点之一是“variety”，也就是大数据的多样性。这就决定了经过各种渠道获取的数据种类和结构都非常复杂，给之后的数据分析处理带了极大的困难。

通过数据处理与集成这一步骤，首先将这些结构复杂的数据转换为单一的或是便于处理的结构，为以后的数据分析打下良好的基础，因为这些数据里并不是所有的信息都是必需的，而是会掺杂很多噪音和干扰项，因此，还需对这些数据进行“去噪”和清洗，以保证数据的质量以及可靠性。

常用的方法是在数据处理的过程中设计一些数据过滤器，通过聚类或关联分析的规则方法将无用或错误的离群数据挑出来过滤掉，防止其对最终数据结果产生不利影响；然后将这些整理好的数据进行集成和存储，这是很重要的一步，若是单纯随意的放置，则会对以后的数据取用造成影响，很容易导致数据访问性的问题，现在一般的解决方法是针对特定种类的数据建立专门的数据库，将这些不同种类的数据信息分门别类的放置，可以有效地减少数据查询和访问的时间，提高数据提取速度。

## 2.3数据分析

数据分析是整个大数据处理流程里最核心的部分，因为在数据分析的过程中，会发现数据的价值所在。

经过上一步骤数据的处理与集成后，所得的数据便成为数据分析的原始数据，根据所需数据的应用需求对数据进行进一步的处理和分析。传统的数据处理分析方法有数据挖掘、机器学习、智能算法、统计分析等，而这些方法已经不能满足大数据时代数据分析的需求。

在数据分析技术方面，Google公司无疑是做得最先进的一个。Google作为互联网大数据应用最为广泛的公司，于2006年率先提出了“云计算”的概念，其内部各种数据的应用都是依托Google自己内部研发的一系列云计算技术，例如分布式文件系统GFS、分布式数据库BigTable、批处理技术 MapReduce，以及开源实现平台Hadoop等。这些技术平台的产生，提供了对大数据进行处理、分析很好的手段。

## 2.4数据解释

对于广大的数据信息用户来讲，最关心的并非是数据的分析处理过程，而是对大数据分析结果的解释与展示，因此，在一个完善的数据分析流程中，数据结果的解释步骤至关重要。若数据分析的结果不能得到恰当的显示，则会对数据用户产生困扰，甚至会误导用户。

传统的数据显示方式是用文本形式下载输出或用户个人电脑显示处理结果。但随着数据量的加大，数据分析结果往往也越复杂，用传统的数据显示方法已经不足以满足数据分析结果输出的需求，因此，为了提升数据解释、展示能力，现在大部分企业都引入了“数据可视化技术”作为解释大数据最有力的方式。

通过可视化结果分析，可以形象地向用户展示数据分析结果，更方便用户对结果的理解和接受。常见的可视化技术有基于集合的可视化技术、基于图标的技术、基于图像的技术、面向像素的技术和分布式技术，等等。

# 3 大数据关键技术

在大数据处理流程中，最核心的部分就是对于数据信息的分析处理，所以其中所运用到的处理技术也就至关重要。提起大数据的处理技术，就不得不提起“云计算”，这是大数据处理的基础，也是大数据分析的支撑技术。分布式文件系统为整个大数据提供了底层的数据贮存支撑架构；为了方便数据管理，在分布式文件系统的基础上建立分布式数据库，提高数据访问速度；在一个开源的数据实现平台上利用各种大数据分析技术可以对不同种类、不同需求的数据进行分析整理得出有益信息，最终利用各种可视化技术形象地显示给数据用户，满足用户的各种需求。

## 3.1云计算和MapReduce

### 3.1.1云计算

Google作为大数据应用最为广泛的互联网公司之一，2006年率先提出“云计算”的概念。所谓“云计算”，就是一种大规模的分布式模型，通过网络将抽象的、可伸缩的、便于管理的数据能源、服务、存储方式等传递给终端用户。根据维基百科的说法，狭义云计算是指IT基础设施的交付和使用模式，指通过网络以按照需求量的方式和易扩展的方式获得所需资源；广义云计算指服务的交付和使用模式，指通过网络以按照需求量和易扩展的方式获得所需服务。

目前，云计算可以认为包含3个层次的内容：服务（IaaS）、平台即服务（PaaS）和软件即服务（SaaS）。国内的“阿里云”与云谷公司的Xen System，以及在国外已经非常成熟的Intel和IBM都是“云计算”的忠实开发者和使用者。

云计算是大数据分析处理技术的核心原理，也是大数据分析应用的基础平台。Google内部的各种大数据处理技术和应用平台都是基于云计算，最典型的就是以分布式文件系统GFS、批处理技术MapReduce、分布式数据库BigTable为代表的大数据处理技术以及在此基础上产生的开源数据处理平台Hadoop。

### 3.1.2 MapReduce

MapReduce技术是Google公司于2004年提出，作为一种典型的数据批处理技术被广泛的应用于数据挖掘、数据分析、机器学习等领域，并且，MapReduce因为它并行式数据处理的方式已经成为大数据处理的关键技术。

MapReduce系统主要由两个部分组成：Map和Reduce。MapReduce的核心思想在于“分而治之”，也就是说，首先将数据源分为若干部分，每个部分对应一个初始的键值对（Key/Value），并分别给不同的Map任务区处理，这时的Map对初始的键-值（Key/Value）对进行处理，产生一系列中间结果Key/Value对，MapReduce的中间过程Shuffle将所有具有相同Key值的Value值组成一个集合传递给Reduce环节；Reduce接收这些中间结果，并将相同的Value值合并，形成最终的较小Value值的集合。

MapReduce系统的提出简化了数据的计算过程，避免了数据传输过程中大量的通信开销，使得MapReduce可以运用到多种实际问题的解决方案里，公布之后获得了极大的关注，在各个领域均有广泛的应用。

## 3.2分布式文件系统

在Google之前，没有哪一个公司曾需要处理数量如此多、种类如此繁杂的数据，因此，Google公司结合自己的实际应用情况，自行开发了一种分布式文件系统GFS（Google File System）。

这个分布式文件系统是个基于分布式集群的大型分布式处理系统，作为上层应用的支撑，为MapReduce计算框架提供低层数据存储和数据可靠性的保障。GFS同传统的分布式文件系统有共同之处，比如性能、可伸缩性、可用性等。然而，根据应用负载和技术环境的影响，GFS和传统的分布式文件系统的不同之处使其在大数据时代得到了更加广泛的应用。

GFS采用廉价的组成硬件并将系统某部分出错作为常见情况加以处理，因此具有良好的容错功能。从传统的数据标准来看，GFS能够处理的文件很大，尺寸通常都是100MB以上，数GB也很常见，而且大文件在GFS中可以被有效地管理。另外，GFS主要采取主从结构（Master-slave），通过数据分块、追加更新等方式实现海量数据的高速存储。

随着数据量的逐渐加大、数据结构的愈加复杂，最初的GFS架构已经无法满足对数据分析处理的需求，Google公司在原先的基础上对GFS进行了重新设计，升级为Colosuss，单点故障和海量小文件存储的问题在这个新的系统里得到了很好的解决。

## 3.3分布式并行数据库

由上述数据处理过程可看出，从数据源处获得的原始数据存储在分布式文件系统中，但是用户的习惯是从数据库中存取文件。传统的关系型分布式数据库已经不能适应大数据时代的数据存储要求，主要原因如下：

1、数据规模变大。大数据的特征之一“volume”，就是指巨大的数据量，因此必须采用分布式存储方式。传统的数据库一般采用的是纵向扩展（scale-up）的方法，这种方法对性能的增加速度远远低于所需处理数据的增长速度，因此不具有良好的扩展性。大数据时代需要的是具备良好横向拓展（scale-out）性能的分布式并行数据库。

2、数据种类增多。大数据的特征之二“variety”，就是指数据种类的多样化。也就是说，大数据时代的数据类型已经不再局限于结构化的数据，各种半结构化、非结构化的数据纷纷涌现。如何高效地处理这些具有复杂数据类型、价值密度低的海量数据，是现在必须面对的重大挑战之一。

3、设计理念的差异。传统的关系型数据库讲求的是“one size for all”，即用一种数据库适用所有类型的数据。但在大数据时代，由于数据类型的增多、数据应用领域的扩大，对数据处理技术的要求以及处理时间方面均存在较大差异，用一种数据存储方式适用所有的数据处理场合明显是不可能的。因此，很多公司已经开始尝试“one size for one”的设计理念，并产生了一系列技术成果，取得了显著成效。

非关系型数据库方案现在被统称为NoSQL（not only SQL）。就目前来说，对于NoSQL没有一个确切的定义，一般普遍认为NoSQL数据库应该具有以下特征：模式自由（schema-free）、支持简易备份（easy replication support）、简单的应用程序接口（simple API）、一致性、支持海量数据（huge amount of data）。

## 3.4开源实现平台Hadoop

大数据时代对于数据分析、管理都提出了不同程度的新要求，许多传统的数据分析技术和数据库技术已经不足以满足现代数据应用的需求。

为了给大数据处理分析提供一个性能更高、可靠性更好的平台，Doug Cutting模仿GFS，为MapReduce开发了一个云计算开源平台Hadoop，用Java编写，可移植性强。现在Hadoop已经发展为一个包括分布式文件系统（Hadoop Distributed File System，HDFS）、分布式数据库（HBase、Cassandra）以及数据分析处理MapReduce等功能模块在内的完整生态系统，现已经发展成为目前最流行的大数据处理平台。

在这个系统中，以MapReduce算法为计算框架，HDFS是一种类似于GFS的分布式文件系统，可以为大规模的服务器集群提供高速度的文件读写访问。HBase是一种与BigTable类似的分布式并行数据库系统，可以提供海量数据的存储和读写，而且兼容各种结构化或非结构化的数据。

## 3.5大数据可视化

可视化技术作为解释大数据最有效的手段之一最初是被科学与计算领域运用，它对分析结果的形象化处理和显示，在很多领域得到了迅速而广泛应用。

数据可视化（data visualization）技术是指运用计算机图形学和图像处理技术，将数据转换为图形或图像在屏幕上显示出来，并进行交互处理的理论、方法和技术。由于图形化的方式比文字更容易被用户理解和接受，数据可视化就是借助人脑的视觉思维能力，将抽象的数据表现成为可见的图形或图像，帮助人们发现数据中隐藏的内在规律。

可视分析起源于2005年，它是一门通过交互可视界面来分析、推理和决策的科学，通过将可视化和数据处理分析方法相结合，提高可视化质量的同时也为用户提供更完整的大规模数据解决方案。如今，针对可视分析的研究和应用逐步发展，已经覆盖科学数据、社交网络数据、电力等多个行业。面对海量数据的涌现，如何将其恰当、清楚地展现给用户是大数据时代的一个重要挑战。学术科研界以及工业界都在不停致力于大数据可视化的研究，已经有了很多经典成功的应用案例。

1. 互联网宇宙（The Internet Map）。为了探究互联网这个庞大的宇宙，俄罗斯工程师Ruslan Enikeev根据2011年底的数据，将196个国家的35万个网站数据整合起来，并根据这些网站相互之间的链接关系将这些“星球”联系起来，命名为“The Internet Map”。一个“星球”代表一个网站，每一个“星球”的大小根据其网站流量来决定，而“星球之间”的距离远近则根据链接出现的频率、强度和用户跳转时创建的链接等因素决定。
2. 标签云（Tag Cloud）。标签云的本质就是一种“标签”，用不同的标签标示不同的对象。标签的排序一般按照字典的顺序排列，并根据其热门程度确定字体的颜色和大小，出现频率越高的词语字体就越大，反之越小，这就方便用户按照字典或是该标签的热门程度来寻找信息。

# 4 大数据的研究与发展方向

尽管大数据的时代已经到来，各界也发现了大数据的巨大价值，并开发出了各种工具用于大数据的采集、处理集成、分析以及解释，但是大数据的研究还处在初始阶段。随着研究的不断深入，大数据所面临的问题也越来越多，如何让大数据朝着有利于全社会的方向发展就需要全面地研究大数据，以下是几种可能的大数据未来的研究与发展方向。

## 4.1关系数据库和非关系数据库的融合

众所周知，关系数据库系统在数据分析中占据着主要地位，但是随着后来半结构化和非结构化数据的大量涌现，关系数据库系统就无所适从了。而类似于MapRduce的大数据处理工具在容错性、可扩展性、数据的移动性上明显优于关系数据库系统，但在处理数据的实时性能上，MapReduce与RDBMS相比还有一定的差距。关系数据库和非关系数据库各有所长，如果在以后的大数据的研究处理过程中，能将关系数据库系统和分布式并行处理系统进行有效的结合，而不是将二者明显地区分开来，那么大数据的分析效率将在很大程度上得到提高。

## 4.2数据的不确定性与数据质量

在网络环境下，不确定性的数据广泛存在，并且表现形式多样，这样大数据在演化的过程中也伴随着不确定性。大数据的不确定性要求人们在处理数据时也要应对这种不确定性，包括数据的采集、处理、分析都需要新的方法来应对，这样也给学习者和研究者带来了很大的挑战，数据质量就很难得到保证，况且大数据的研究领域尚浅，本身就有很多亟待解决的问题。

面对不断快速产生的数据，在数据分析的过程中很难保证有效的数据不丢失，而这种有效的数据才是大数据的价值所在，也是数据质量的体现。所以需要研究出一种新的计算模式，一种高效的计算模型和方法，这样数据的质量和数据的时效性才能有所保证。

## 4.3跨领域的数据处理方法的可移植性

大数据自身的特点决定了大数据处理方法的多样性、灵活性和广泛性。而今几乎每个领域都有涉及到大数据，在分析处理大数据的建模过程中除了要考虑大数据的特点外还可以结合其他领域的一些原理模型。广泛吸纳其他研究领域的原理模型，然后进行有效的结合，从而提高大数据处理的效率，这可能会成为以后大数据分析处理的重要方法。

## 4.4大数据的预测性作用日益凸显

提及大数据，它的作用自然是不言而喻，也有不少专家进行了总结，大数据有变革价值的力量、大数据有变革经济的潜力、大数据有变革组织的潜能。但是从很多大数据的应用案例分析不难发现，无论是大数据的研究者还是普通人，大数据给人们带来的最直接的利益就是对未来的预见。气象部门可以根据气象数据预测未来的天气变化；经销商可根据商品的销量分析客户的喜好从而制定未来的采购计划及时调整经营模式，增加利润；通信部门通过对大数据的分析实时了解市场行情，从而作出合理决策。由已知推测未知，通过大数据可以提高对未知预测的可靠性和精准性，这对整个人类来说都是一种进步。

# 5结语

这是一个信息爆炸的时代，不管是研究领域、商业领域还是工业领域，都要同数据打交道。随着科技的迅猛发展，更加先进的存储技术的出现，使得人们必须面对规模更加巨大、结构更加复杂的数据，并亟待从中挖掘出有用的信息。

本文介绍了大数据的基本概念，概括出了大数据的一般处理流程，详细介绍了大数据的几种关键技术：云计算、分布式文件系统、分布式并行数据库和大数据可视化。最后，对大数据的研究与发展方向进行了相应的总结：关系数据库和非关系数据库的融合、数据的不确定性和数据质量、跨领域的数据处理方法的可移植性、大数据的预测性作用日益凸显。

总体来说，目前对于大数据的研究尚属起步阶段，还有很多问题亟待解决。大数据时代已经来临，如何从海量数据中发现知识、获取信息，寻找隐藏在大数据中的模式、趋势和相关性，揭示社会运行和发展规律，以及可能的科研、商业、工业等应用前景，都需要我们更加深入的了解大数据，并具有更加深刻的数据洞察力。

# 参考文献

[1] Naisbitt J.MeGatrends: TenNewDirectionsTransformingOurLive. [M]. NewYork:Warner Books，1982：40-42.

[2] 阿尔文·托勒夫.第三次浪潮.[M].黄明坚译.北京：中信出版社，2006：19-25.

[3] 邬雪艳，孙永杰.云计算和大数据助力医疗协同[N].通信世界，2013.04-17.

[4] 林子雨.大数据技术原理与应用.[M].北京：人民邮电出版社，2015-08

[5]IBM.Whatisbigdata?[EB/OL].[2012-10-02].http//www-01.ibm.eom/soflware/data/bigdata/what-is-big-data.htm1.

[6] GRAHAM-ROWED，GOLDSTOND，DOCTOROWC，eta1.Big Data:Science in The petabyteera.[J].Nature，2008，455(7209)：8-9.

[7] Lab of Web and Mobile Data Management.WAMDM HomePage[EB/OL].[2013-07-24].http：//idke.ruc.edu.cn/index.htm

[8] 李乔，郑啸.云计算研究现状综述[J].计算机科学，2011，38（4）：32-37.Li Qiao，Zheng Xiao.Research Survey of Cloud Computing[J].Computer Science，2011，38（4）：32-37.

[9] GheMawat s，GoBioff h，leunG s t.the Google File System[J].ACM Sigops Operating Systems Review，2003，37（5）：29-43.

[10] Chang F，Dean J，Ghemawats，et al.BigTable：A Distributed Storage System for Structured Data[J].ACM Transactions on Computer Systems，2008，26（2）：4.

[11] Dean J，Ghemawats.MapReduce：Simplified Data Processing on Large Clusters[J].Communications of The ACM 51，2008（1）：107-113.

[12] 杨宸铸.基于Hadoop的数据挖掘研究[d].重庆：重庆大学，2010.Yang Chen-zhu.The Research of Data Mining Based on Hadoop[d].Chongqing：Chongqing University，2010.

[13]贺全兵.可视化技术的发展及应用[J].中国西部科技，2008，7（4）：4-7.He Quan-Bing.The Development and Application of Visualization Technique[J].Science and Technology of West China，2008，7（4）：4-7.

[14]维基百科.云计算[EB/OL].[2013-07-24].https://zh.wikipedia.org/wiki/%E9%9B%B2%E7%AB%AF%E9%81%8B%E7%AE%97