Computer Science and Application 计算机科学与应用, 2018, 8(10), 1503-1509
Published Online October 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.810163

Hans汉斯

Overview on Big Data

Kaiyue Liu

China University of Mining & Technology (Beijing), Beijing
Email: liukaiyue@ict.ac.cn

Received: Oct. 1st, 2018; accepted: Oct. 11th, 2018; published: Oct. 19th, 2018

Abstract

As a current popular technical, big data has received wide attention from every industry. In order to further understand big data, this paper comprehensively describes big data from the six aspects: The basics of big data, the origin and development status of big data, big data processing, big data application, big data challenges and the future of big data. The basics of big data include the concepts and differences between big data and traditional databases, and the characteristics of big data. The big data processing includes generating and getting data, preprocessing data, data storage, analyzing and mining data. This article is a systematic review of big data, and can establish a good knowledge system for scholars who are new to big data.

Keywords

Big Data, Data Storage, Data Mining, Data Visualization, Big Data Application

大数据综述

刘凯悦

中国矿业大学(北京)，北京
Email: liukaiyue@ict.ac.cn

收稿日期：2018年10月1日；录用日期：2018年10月11日；发布日期：2018年10月19日

摘　要

大数据作为当今的热点技术，受到了各行各业的广泛关注。为了进一步认识大数据，本文从大数据的基础、大数据的起源和发展现状、大数据的处理流程、大数据的应用、大数据面临的挑战、大数据未来展望六个方面对大数据进行了综合性描述。其中大数据基础包括大数据和传统数据库的概念和区别、大数据的特性，处理流程包括数据生成和获取、数据预处理、数据存储、数据分析挖掘。本文是大数据的系统性综述，可以对初次接触大数据的学者建立了良好的知识体系。

文章引用: 刘凯悦. 大数据综述[J]. 计算机科学与应用, 2018, 8(10): 1503-1509.
DOI: 10.12677/csa.2018.810163

关键词

大数据，数据存储，数据挖掘，数据可视化，大数据应用

1. 引言

近几年，由于移动互联网、物联网、云计算等技术的迅速发展，产生了海量的大规模数据。数据爆炸将人们带入了一个新的时代——大数据时代，如何存储、处理批量数据已经成为一个值得研究和讨论的问题。大数据技术的迅速发展推动了社会进步，国内外的许多甚至高校开展了新的一门学科即数据科学。大数据可以应用到诸多领域，人们通过大数据技术获取海量数据，并对数据进行处理和分析，得到许多对未来具有研究和改进意义的结果。大数据对现代社会的发展做出的贡献千千万万，本文主要针对大数据的概念、技术、应用等各方面进行综合性描述。

2. 大数据基础

2.1. 大数据和传统数据库

数据库是按照数据结构来组织、存储和管理数据的仓库，在大数据这个概念出现以前，人们一直是应用数据库来存储和管理一些相对简单小型的数据。随着信息技术的发展和数据量的迅速增长，传统数据库在有些方面已经不能满足人们的需求，由此衍生出大数据这一概念。

大数据又称为巨量数据、海量数据、大资料等，是指无法在一定时间范围内通过人工或计算机进行捕捉、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[1]。

大数据和传统数据库有许多区别：首先，从数据规模和类型来看，传统数据库通常是以 MB 为单位且数据种类单一。但从大数据的数据单位很大，通常以 GB、TB、PB 甚至 EB、ZB 为单位，且数据种类繁多；其次，从模式和数据关系来看，传统数据库是先有模式再产生数据的。而大数据很难预先确定模式，甚至有些时候模式是会随着数据量的增加而改变的；最后，从处理对象上来看，传统数据库中的数据仅仅作为处理对象。但大数据中是要将数据作为一种资源来帮助分析其他领域的诸多问题的。

2.2. 大数据的特性

大数据具有 5V 特性，即大量(Volume)、高速(Velocity)、多样(Variety)、价值密度低(Value)、真实性(Veracity)。

Volume，主要体现在数据存储量大和数据增量大。数据规模庞大是大数据最主要的特性，而随着云计算等技术的发展，数据量也不断在增长，数据量已从 GB、TB 再到 PB 字节，甚至已经开始以 EB 和 ZB 字节来计量[2]。

Velocity，高速性指的是数据的产生和处理速度快。数据可以通过社交媒体、定位系统等应用快速大量地产生。同时数据的处理速度也应加快，只有快速适时处理才可以更加有效的利用得到的数据。

Variety，多样化主要体现在格式多和来源多两个方面。大数据产生的数据类型繁多，其中包括结构

化、半结构化和非结构化数据，甚至包括非完整和错误数据[2]。这是因为数据的来源多种多样，例如网页日志、电子邮件、传感器等。

Value，价值密度低是指，虽然数据量庞大但其中具有利用价值的信息并不多。需要通过特定的技术进行处理和进一步挖掘，提取最有用的信息来加以利用[3]。

Veracity，数据的真实性和质量决定数据带给我们的价值[3]。高质量的数据一定是具有真实性的，但有时真实的数据并不一定代表着高质量。我们可以通过一些大数据技术，在保证数据真实性的同时提高数据的质量，使数据能够更好的为我们所用。

3. 大数据起源和发展现状

3.1. 大数据的起源

目前，IT 界普遍认为大数据起源于谷歌的“三驾马车”：谷歌文件系统、MapReduce 和 BigTable。谷歌工程师在 2003 年至 2006 年先后公开发表了这几项核心技术的学术论文，引起了巨大反响，吸引了众多互联网公司的注意。在各大互联网公司的技术推动下，最终诞生了 Hadoop 系统，并在 2008 年 6 月处于相对稳定的状态。Hadoop 发展过程中一个标志性的公司是 Facebook，其在 Hive 上投入大量资源。Hadoop 高扩展、高容错的优点很受业内欢迎，被广泛应用于离线分析场景。2011 年 5 月，在“云计算相遇大数据”为主题的 EMC World 2011 会议中，EMC 抛出了 Big Data 这一概念。Facebook 公司在 2012 年将重点转移至 Presto，其查询速度很快，没有用到 MapReduce，很快便超过了 Hive。另外，伯克利大学 AMP 实验室开发了高速、灵活的 Spark 系统，Spark 的兴起是 Hadoop 生态圈一个比较关键的转折点，在迭代计算和实时分析领域占领了绝对优势。

3.2. 大数据的发展现状

随着大数据技术的不断发展，许多国家都认识到大数据对国家发展的重要性。以美国为首的多个国家先后发布了大数据的国家发展战略，联合国也发布了“全球脉搏”项目的重要成果——名为《大数据促发展: 挑战与机遇》的大数据政务白皮书。美国政府投入了巨资到大数据的研究领域，将其作为重要的战略发展方向，并将大数据技术发展提升到国家安全和未来的发展战略的高度[4]。

我国科技界与信息技术密切相关的产业领域对大数据技术与应用的关注程度正在逐渐增强，并引起了政府相关部门的重视。2013 年 3 月在上海召开了题为“大数据技术与应用中的挑战性科学问题”双清论坛，并将“大数据技术与应用中的挑战性科学问题”列入 2014 年的项目指南中，拟以重点项目群的方式支持和推动相关领域的基础研究[2]。自 2016 年开始，国家信息中心已经连续 3 年利用大数据技术反映“一带一路”的建设进展和成效。除此之外，大数据技术目前已经在很多领域有了具体应用案例。2018 年 9 月 19 日，国家信息发布中心在天津举办的 2018 年夏季达沃斯论坛上发布了《“一带一路”大数据报告 2018》。该报告的发布，能够为国内外各界了解、参与“一带一路”建设提供更为丰富的信息。2018 年 9 月 20 日，国家发展改革委国际合作中心(以下简称“国际合作中心”)举办第三期“国合党建讲堂”，邀请国家信息中心大数据发展部主任于施洋作题为“以大数据思维助力创新发展改革工作”的专题讲座。

目前，大数据行业主要分为三类产业：数据服务产业、基础支撑产业、融合应用产业。数据服务产业是以大数据为核心资源，以大数据应用为主业开展商业经营的产业，包括数据交易、数据采集、数据应用服务、基于大数据的信息服务、数据增值服务等。基础支撑产业是指提供直接应用于大数据处理相关的软硬件、解决方案及其他工具的产业，例如提供大数据存储管理、大数据预处理软硬件、大数据计算、大数据可视化产品等。融合应用产业是指在业务应用中产生大数据，并与行业资源相结合开展商业经营的产业，例如政务大数据、金融大数据、交通大数据、工业大数据等。

4. 大数据技术流程

大数据技术的主要流程可以分为：数据生成和获取、数据预处理、数据存储、数据计算分析挖掘、数据结果应用。

4.1. 数据生成和获取

数据的来源多种多样，可以来自物联网、互联网、各类传感器等。同时数据的方式也是多种多样的，可以是数字、文字、声音、图片、视频等。中国工程院李德毅院士认为：大数据的主要来源有三方面: 自然界的大数据、生命和生物的大数据和社交大数据。自然界的大数据主要是机器与机器交互产生的数据，主要通过各类传感器来采集[4]。生命和生物的大数据主要研究基因组学、蛋白质组学、代谢组学等生物学数据。社交大数据主要来源于人类社会活动，而互联网通常为其载体。目前大数据的主要研究对象集中在社交数据和自然数据，同时生命和生物的大数据对医学方面的贡献也不容小视。

4.2. 数据预处理

现实中收集到的真实数据通常都是不完整的脏数据，没有办法直接进行数据挖掘和处理。所以为了提高数据的质量，通常需要对获取到的数据进行预处理，也就是在主要的数据处理之前对数据做出的一些基本处理。

数据预处理的内容主要有：数据审核、数据筛选、数据排序。数据审核主要审核数据的准确性、适用性、及时性、一致性。数据筛选是对审核过程中发现的错误进行纠正的过程，通常包括两方面内容：剔除不符合要求的数据、筛选出符合条件的数据。数据排序是按照一定的顺序把数据进行排列，以便于研究者进一步观察和分析。

数据预处理的主要方法有数据清理、数据集成、数据变换、数据规约。数据清理的主要目的为格式标准化、清除异常数据、纠正错误。数据集成是将多个数据源中的数据结合起来统一存储。数据变换是利用规范化、平滑聚集、数据概化等方式将数据转变成有利于数据挖掘的形式。数据规约可以得到规约表，节省挖掘分析时间且仍然能保持数据的完整性。

4.3. 数据存储

传统的数据存储方式可以分为块存储、文件存储、对象存储，大数据的存储方式可以分为分布式系统、NoSQL 数据库、云数据库。分布式系统主要包含分布式文件系统 HDFS、分布式键值系统。其中分布式文件系统是一个高度容错性系统，适用于批量处理并且能够提供高吞吐量的数据访问。分布式键值系统可以用于存储关系比较简单的半结构化数据，其存储和管理的是对象而不是数据块。NoSQL 数据库的发展是由于关系型数据库已经无法满足大量数据的管理需求，而 NoSQL 数据库可以存储超大规模的数据，具有较好的横向扩展能力。云数据库是基于云计算技术发展的一种共享基础构架的方法，是部署和虚拟化在云计算环境中的数据库。

大数据存储技术路线包括分布式架构、大数据一体机、MPP 混合构架。其中分布式构架包括 Hadoop、MapReduce 等，随着相关技术的不断进步，其应用场景也将逐步扩大。大数据一体机是专门为大数据分析处理而设计的软硬件结合的产品，具有良好的稳定性和纵向扩展性。MPP 混合构架重点面向行业大数据，通过列存储、粗粒度索引等多项大数据处理技术，再结合 MPP 架构高效的分布式计算模式，完成对分析类应用的支撑，具有高性能和高扩展性的特点。

4.4. 数据分析挖掘

数据分析和挖掘就是从大量的数据中提取出隐含在其中的、具有潜在价值的信息，是统计学、人工

智能、数据库技术的综合运用。

大数据的计算分析框架主要包括批处理框架、流处理框架、交互式计算框架、混合处理框架、图数据处理框架。

批处理框架是以 MapReduce 为代表的，MapReduce 是一个基于集群的高性能并行计算平台。Map 表示映射，Reduce 表示化简，所以 MapReduce 可以理解为把一堆杂乱无章的数据按照某种特征归纳起来，然后处理并得到最后的结果。MapReduce 具有易于编程、良好的扩展性、高容错性、适合 PB 级以上海量数据的离线处理等特点。

流处理框架 Storm 是一个分布式的、容错的实时计算系统。Storm 集群包含一个主控节点和若干个工作节点，主控节点接受任务并分配给工作节点执行。Strom 具有可持续流处理、可扩展、高容错、高可靠、结构丰富、支持多语言开发等优势。

交互式计算框架 Spark 是基于内存计算的大数据并行计算框架，提高了在大数据环境下数据处理的实时性，同时还保证了高容错性和高可伸缩性。Spark 的核心概念是 RDD(resilient distributed dataset)，指的是一个只读的、可分区的分布式数据集。RDD 除了提供内存存储和各种类型并行计算外，还可以自动从故障中恢复，实现了基于 Lineage 的容错机制[4]。

混合处理框架 Lambda 是在结合了批处理和流处理两种架构的混合架构。在处理数据时，分别将数据送入批处理层和实时处理层，这样可以使得得到的结果更加快速且精确，该混合框架对不同业务需求进行了良好的平衡[4]。

图处理框架 Pregel 是一个用于分布式图计算、基于整体同步并行计算模型的计算框架。Pregel 是以节点为中心进行计算的，每个节点在初始时处于活跃状态，完成计算后进入不活跃状态。Pregel 具有良好的容错机制、可以持久性存储、采用主/从结构实现整体功能等特点。

在数据分析挖掘的过程中，可视分析是十分重要的方法，其可以充分挖掘人对于可视化信息的认知能力优势，帮助人们更直观高效地了解大数据隐含的信息。数据可视化分析是指在数据分析的同时利用支持信息可视化的用户界面或人机交互方式，融合人的认知能力和计算机的计算能力，便于获得大规模复杂数据洞察力的一门技术[5]。其通常可以分为 2D 法和多维法，例如时间可视化、层次法可视化。

5. 大数据应用

大数据已被广泛的认为是创造新价值的利器，具有无可估量的资源价值。大数据对人类的贡献已扩散到各个领域，推动各个领域的快速发展，例如医疗、教育、商业、工业、农业等。

在医疗方面，通过采集和分析医疗机构产生的实验室记录、用药记录、手术记录、住院记录、急诊记录等各类大数据，总结出其隐含价值并应用在未来的医学研究中[6]。

大数据还可以应用在智能电网、工业互联网、排水系统、智能交通等工业领域，通过传感器等采集数据装置，收集大量数据进行分析和处理，最终得到优化方案或改进措施。

在商业领域，股票、保险、银行等行业仍离不开大数据的应用。通过大数据技术可以利用股票走势数据、保险报销人群数据、银行汇率数据等得到一些预测数据，帮助人们更好的选择有利的商业方向。

总之，大数据对社会做出的贡献体现在方方面面，使得各行各业的发展更具有规划性和方向性。大数据时代为人们提供了便利、高效、高品质的生活环境，给人们的生活带来了翻天覆地的变化。

6. 大数据面临的挑战

6.1. 技术方面的挑战

首先从网络方面考虑，大数据的传输需要一个超高速的网络来进行支撑，这对目前的网络技术是一

个十分重要的挑战。其次从机器学习方面来看，机器学习等分析算法需要更加智能化、高效化发展，才能更好的适应大数据时代。最后从数据存储方面，获取到的数据经过预处理后需要更加高效的存储方式。

6.2. 政策方面的挑战

大数据的兴起也为人们带来一些负面的影响，例如隐私泄漏这一关键问题。尤其是在网络方面，人们的个人信息基本上一览无余。在平时生活中甚至会在不知情的情况下被泄漏了个人信息，总是会接到各类推销电话。所以政府机关应尽快建立大数据背景下完善的信息安全法律法规体系，建立大数据技术的行业通用标准，才能有效减少大数据带来的消极影响[7]。

6.3. 国际关系方面的挑战

大数据蕴含着丰富的政治经济文化社会信息，一个国家的科技发展、社会动向、经济浮动、军事行动、国家安全等信息均可以利用大数据技术分析并传递出来[8]。所以各个国家应时刻注意本国重要信息安全问题，警惕非法泄漏信息，力争掌握数据信息的主动权。这样以来，可以有效防止因数据信息权力争夺导致的科技战争。

7. 大数据未来展望

大数据时代的有些未来是可以预见的。首先是数据库能力的提升，谷歌的 Spanner 和亚马逊的 Redshift 都体现了这种变化：数据库的能力越来越强，它可以解决很多大数据的问题。而同时数据也将逐渐趋于资源化，资源化是指大数据成为企业和社会关注的重要战略资源，并已成为大家争相抢夺的新焦点。因此，企业必须要提前制定大数据营销战略计划，抢占市场先机。

其次，大数据未来会与云计算更加紧密深入地结合。大数据离不开云处理，云处理为大数据提供了弹性可拓展的基础设备，是产生大数据的平台之一。物联网、工业互联网等新兴计算形态，将让大数据营销发挥出更大的影响力[9]。目前，工业互联网平台的应用还处于发展的初级阶段，而工业互联网平台的未来，则需要设备物联和系统互联全面打通。所以应当在数据管理和分析应用方面为工业互联网平台赋能[10]。而大数据技术未来在物联网方面的应用可以在统计技术标准、优化数据安全管理、控制成本投入等方面进行着重发展和改进[11]。

最后，未来的大数据会和人工智能这一当今热门核心技术进行完美地结合。我们可以通过人工智能技术给大数据建立更好的索引，人工智能促进大数据发展和大数据融合会是一个很重要的发展方向。虽然人工智能技术是大数据分析的利器，但面临大数据问题时，现有的机器学习、深度学习、计算智能等人工智能分析方法、大数据平台都存在许多不足，难以有效解决大数据的诸多问题[12]。目前进一步研究的主要方向有：分布式深度学习算法、设计机器学习模型并行策略、分布式优化算法、优化分布式集群环境、分配深度神经网络的并行训练、优化深度学习参数、建立先进的大数据平台等。

8. 小结

本文针对大数据的概念、特性、起源和现状进行了详细讲解。并对大数据的主要技术：数据生成和获取、数据预处理、数据存储、数据分析挖掘进行了综合性描述。最后阐述了大数据对各行各业的积极贡献和影响，并提出了大数据技术现在仍面临的技术、政策、国际关系方面的挑战以及大数据未来发展趋势的展望。

在大数据技术发展的关键阶段，我国应该积极倡导大量科学研究投入到大数据的应用研究中。将大数据技术应用到各个领域，并与人工智能、深度机器学习、云计算等关键技术相结合，建立自己的数据科学体系、政策体系、人才体系等。

致　谢

首先感谢我的导师兰德品老师，在兰老师的耐心指导和帮助下，我完成了学校的大学生创新项目——大数据技术的应用研究。在这次项目中我从一个对大数据知识一无所知的学生逐渐学习和成长，最终建立了良好的大数据知识体系，并对这个行业有了自己的见解和感受。他严谨求学的治学作风和一丝不苟的敬业精神对我影响深刻，使我不仅在学习和研究中受益颇深，在为学之道上也有巨大的收获，是我学习和生活上的榜样。

其次要感谢同我一起学习和成长的小组同学们，在研究和学习过程中我们积极配合，不懂的问题大家共同探讨共同学习，使我分析解决问题的能力得到了很好的锻炼。

最后，感谢本文最后所列的参考文献作者们，在认识大数据的道路上是他们的研究成果及文献给了我很大的帮助。

基金项目

中国矿业大学(北京)科学研究基金项目“大数据技术的应用研究”(编号 C201707543)。

参考文献

[1] 陈军成, 丁治明, 高需. 大数据热点技术综述[J]. 北京工业大学学报, 2017, 43(3): 358-367.

[2] 王成红, 陈伟能, 张军, 宋苏, 鲁仁全. 大数据技术与应用中的挑战性科学问题[J]. 中国科学基金, 2014(2): 92-98.

[3] 王晨晨, 孙睿. 浅析大数据的发展[J]. 中国市场, 2018(27): 194-196.

[4] 张锋军. 大数据技术研究综述[J]. 通信技术, 2014, 47(11): 1240-1248.

[5] 任磊, 杜一, 马帅, 张小龙, 戴国忠. 大数据可视分析综述[J]. 软件学报, 2014, 25(9): 1909-1936.

[6] 邢丹, 姚俊明. 医疗健康大数据: 概念、特点、平台及数据集成问题研究[J]. 物联网技术, 2018(8): 104-106.

[7] 张茂月. 大数据时代个人信息数据安全的新威胁及其保护[J]. 中国科技论坛, 2015(7): 117-122.

[8] 孙睿, 王晨晨. 大数据时代面临的挑战[J]. 中国市场, 2018(26): 187-196.

[9] 王建民. 工业大数据是工业互联网的核心[J]. 中国信息化周报, 2018(14): 1-2.

[10] 李琼. 大数据应用于工业互联网平台的融合[J]. 软件和集成电路, 2018(8): 52-53.

[11] 李鹏飞. 大数据时代物联网技术的应用与发展[J]. 西部皮革, 2018(15): 69.

[12] 王万良, 张兆娟, 高楠, 赵燕伟. 基于人工智能技术的大数据分析方法研究进展[J]. 计算机集成制造系统. http://kns.cnki.net/kcms/detail/11.5946.tp.20180817.1005.011.html