大数据与云计算

■ 何 清* 中国科学院计算技术研究所 北京 100190

摘 要：大数据（Big Data）这个概念近年来在越来越多的场合、被越来越多的人提及，并且经常和云计算联系在一起，云计算与大数据之间到底是什么关系成为热点话题。本专题报告包含以下四个方面内容：1. 大数据的价值；2. 大数据带来的挑战；3. 大数据研究成果；4. 云计算是大数据挖掘的主流方式。通过本报告阐述我们对大数据的理解，以及对大数据的价值的认识，探讨大数据处理与挖掘技术，论述以下观点：没有互联网就没有云计算模式，没有云计算模式就没有大数据处理技术，也就没有大数据挖掘技术。

关键词：大数据 云计算 数据挖掘

DOI:10.11842/chips.2014.01.006

一、大数据的价值

根据维基百科的定义，大数据（Big Data）是用于数据集的一个术语，是指大小超出了常用的软件工具在运行时间内可以承受的收集，管理和处理数据能力的数据集。换句话说，在单一数据集里，数据规模超出目前常用软件工具在合理的可容忍时间里可以访问、管理、处理能力的数据集就是大数据。由于软件的能力是与时俱进的，因而大数据规模的定量界限就是随着技术进步而不断增大。大数据的规模大小是一个不断演化的指标，目前范围是指在一个单一的数据集从数十 TB 到十几 PB 级的数据规模。大数据逐渐有替代此前类似的海量数据（Massive Data）、大规模数据（Large Scale Data）、庞大数据（Enormous Data）、巨量数据（Huge data）等概念的趋势。实际上，不能简单地以数据规模来界定大数据，而要考虑满足用户需求的数据处理与分析的复杂程度。针对简单的用户需求（如关键字搜索），数据量为 TB 至 PB 级时可称为大数据；而针对复杂的用户需求（如数据挖掘），数据量为 GB 至 TB 级时即可称为大数据。存在数据量很大，计算任务简单的“小数据”；也存在数据量不大，但数据和计算复杂性高的“大数据”。

大数据的特征可以用所谓的 3 个“V”表示：体量（Volume）、多样性（Variety）与速度（Velocity）。体量（Volume）是指聚合在一起供分析的数据量必须是非常庞大的。无所不在的移动设备、RFID、无线传感器每分每秒都在产生数据，数以亿计用户的互联网服务时时刻刻在产生巨量的交互。Web 日志、RFID、传感网、社会网、社会数据、互联网文本文档、互联网搜索索引、呼叫记录、天文记录、大气科学、基因学、生物化学、

* 何清，中国科学院计算技术研究所研究员，博士生导师，2008 年底开发完成了我国最早的基于云计算的并行数据挖掘平台，用于中国移动 TB 级实际数据的挖掘，实现了高性能、低成本的数据挖掘，先后主持完成多个有关数据挖掘的国家自然科学基金项目和 863 项目，提出了一系列有效的数据挖掘算法，组织开发的多个数据挖掘软件获得了软件著作权，并实际应用到电信、国家电网、信息安全、环保等多个行业，为企业带来了可观的经济效益和社会效益。

生物学、其他复杂的交叉学科的科学研究、军事监控、医学记录、照片摄像档案、视频档案、大规模的电子商务都是大数据的来源。在美国拥有 1000 名员工的公司有至少 200TB 的存储数据。例如沃尔玛每小时处理超过一百万客户交易，这些交易数据放到数据库估计超过 2.5PB，这等价于美国国会图书馆所有书包含信息的 167 倍。多样性（Variety）是指数据类型的复杂性。如企业内部的信息主要包括联机交易数据和联机分析数据，这些数据一般都是结构化的静态、历史数据，可以通过关系型数据进行管理和访问，数据仓库是处理这些数据的常用方法。而来自于互联网上的数据，如用户创造的数据、社交网络中人与人交互的数据、物联网中的物理感知数据等，都是非结构化且动态变化，这些非结构化的数据占到整个数据的 80% 以上。在金融服务、政府管理、零售业会产生文本和数字数据，而制造业、医疗保健、新闻传媒等多产生多媒体数据。而速度（Velocity）则是指数据处理的速度必须满足实时性要求。像离线数据挖掘对处理时间的要求并不高，因此这类应用往往运行 1、2 天获得结果依然是可行的。但对于大数据的某些应用而言，必须要在 1 秒钟内形成答案，否则这些结果可能就因过时无效而失去其商业价值，例如实时路况导航、全球股价波动。

这些特点也反映了大数据所潜藏的价值（Value），或许可以认为，这四个 V 就是大数据的基本特征。

大数据无疑将给人类社会带来巨大的价值。科研机构可以通过大数据业务协助进行研究探索，如环境、资源、能源、气象、航天、生命等领域的探索。产业方面，大数据是现有产业升级与新产业诞生的重要推动力量。数据为王的大数据时代的到来，产业界需求与关注点发生了重大转变：企业关注的重点转向数据，计算机行业正在转变为真正的信息行业，从追求计算速度转变为关注大数据处理能力，软件也将从编程为主转变为以数据为中心。

大数据正在影响企业商业模式的转变，对数据进行分析、优化正成为提升核心竞争力的有效方式。制药企业可借助大数据进行更多药品实验和分析。对于销售和服务可以提供消费者偏好与需求模式等方面的信息，帮助企业提高计划、决策和预测的准确性。

当然大数据相关的产业链也必然带来巨大影响。首先，信息数据产生将会是第一个环节。其次，信息数据的大量产生需要存储。再次，信息数据需要采集整理。最后，信息数据的分析产出。这个环节是整个“大数据”产业链的最末端，也可能是最具技术含量和产业附加值的子行业。任何数据不经过分析这一环节，都无法落实到实际应用。在同样的数据面前，谁分析出的结果最快最有效，将决定谁才是真正的“大数据”产业领跑者。

二、大数据带来的挑战

1. 描述与存储的挑战

云计算环境下对大数据管理技术提出了新的挑战，主要反映在传统的关系数据库不能满足大数据处理的需求，如海量用户的高并发读写、海量数据的高效存储与访问、系统的高可用性与高扩展性等。随着数据规模的增大，原来高效的算法会变得低效，关系数据库事务处理要求的 ACID 特性，即原子性 (Atomicity)、一致性（Consistency）、隔离性（Isolation）、持久性（Durability）的开销巨大。目前的 NoSQL 运动正在通过放弃关系型数据库强大的 SQL 查询语言、事务的一致性以及范式的约束，或者采用键—值数据格式存储，以获得高效灵活的大数据处理能力。在业界，全球著名的 Google、EMC、惠普、IBM、微软等互联网公司都已经意识到大数据存储的重要意义，研发了一批包含分布式数据缓存、分布式文件系统（GFS、HDFS）、非关系型 NoSQL 数据库（Amazon 的 Dynamo、Apache Cassandra、HBase）和新关系型 NewSQL 数据库等新技术。Gupta 等人提出分析大数据过程中面临的挑战，包括静态数据与动态数据。对于静态的大数据，Gupta 等人描述了面向交互数据服务环境的 NoSQL 系统以及基于 MapReduce 编程模式的面向大规模数据分析的系统。

2. 分析与理解的挑战

大数据具有复杂性是不言而喻的，这种复杂性不仅体现在数据类型的多样性以及数据来源的广泛性上，更重要的是体现在分布的不确定性上。大数据集往往来源于对多源异构数据的融合和集成，具有超高维、稀疏、多模态等内在分布特征。这些内部特征导致现有机器

学习算法的性能和效率降低，导致对大数据的理解如同盲人摸象。

3. 挖掘与预测的挑战

大数据中所蕴含的价值需要挖掘。大数据挖掘增加样本容易，降低算法复杂度难。很多传统的数据挖掘算法不一定能够适用于大数据环境，目前常用的数据挖掘的算法并不都能够被并行化，也就是说并非所有的算法都具有高度的并行性，并行不能降低算法复杂度，因此需要研究和开发新的适应大数据环境的算法。

三、大数据研究成果

1. 大数据处理技术

由于海量数据的大数据量和分布性的特点，使得传统的数据处理技术不适合于处理海量数据。这对海量数据的分布式并行处理技术提出了新的挑战，开始出现以 MapReduce 为代表的一系列工作。

(1) 数据并行处理

MapReduce 是 2004 年谷歌提出的一个用来并行处理大数据集的并行处理模型。而 Hadoop 是 MapReduce 的开源实现，是企业界及学术界共同关注的大数据处理技术。MapReduce 并行编程模型具有强大的处理大规模数据的能力，因而是大数据处理的理想编程平台。Map-Reduce 通过动态负载均衡及资源调配机制，可以根据需求的变化，对计算资源自动进行分配和管理，实现“弹性”的缩放和优化使用，对复杂问题采用分而治之的策略，把问题拆分后进行并行的运算，再将结果进行整合，从而得到最终的结果，表现出良好的扩展性、容错性和大规模并行处理的优势，在大数据管理和分析等方面得到广泛应用。

针对并行编程模型易用性，出现了多种大数据处理高级查询语言，如 FaceBook 的 Hive、Yahoo 的 Pig、Google 的 Sawzall 等。这些高层查询语言通过解析器将查询语句解析为一系列的 MapReduce 作业在分布式文件系统上执行。与基本的 MapReduce 系统相比，高层查询语言更适合用户方便地进行大规模数据的并行处理。MapReduce 及高级查询语言在应用中也暴露了在实时性和效率方面的不足，因此有很多研究针对它们进行优化提高效率。

MapReduce 作为典型的离线计算框架，无法适应于很多在线实时计算需求。目前在线计算主要基于两种模式研究大数据处理问题，一种基于关系型数据库研究提高其扩展性，增加查询通量来满足大规模数据处理需求；另一种基于新兴的 NoSQL 数据库，通过提高其查询能力丰富查询功能来满足现有大数据处理需求的应用。使用关系型数据库为底层存储引擎，上层对主键空间进行切片划分，数据库全局采用统一的哈希方式将请求分发到不同的存储节点以达到可以水平扩展要求，这种方案一般不能对上层提供原存储引擎的全部查询能力。Oracle NoSQL DB、MySQL Cluster、MyFOX 即是典型系统，通过扩展 NoSQL 数据库的查询能力的方法来满足大规模数据处理需求的最典型的例子就是 Google 的 BigTable 及其一系列扩展系统。

如何处理海量分布式的复杂数据也是目前的研究热点。Google MapReduce 的设计初衷是分析 Web Graph，但处理图数据常常需要大量的迭代运算，而 MapReduce 不是很适合处理这类复杂数据，已有的并行图算法库 Parallel BGL 或者 CGMgraph 又没有提供容错功能。于是 Google 开发了 Pregel，一个可以在分布式通用服务器上处理 PB 级别图数据的大型同步处理应用，与之对应的开源项目 Giraph 也得到学术界的关注。

(2) 增量处理技术

如何采用增量处理技术来设计高效的增量算法来解决分布式大数据的动态更新问题也是目前的研究热点。Google 已经采用增量索引过滤器（Percolator for incremental indexing），而不是 MapReduce 来对频繁变化的数据集进行分析，使得的搜索结果返回速度越来越接近实时。通过只处理新增的、改动过的或删除的文档和使用二级指数来高效率建目录，返回查询结果。Percolator 将文档处理延迟缩短了 100 倍，其索引 Web 新内容的速度比用 MapReduce 快很多。

(3) 流式计算技术

目前流式计算是一个业界研究的热点，最近 Twitter、LinkedIn 等公司相继开源了流式计算系统 Storm、Kafka 等，加上 Yahoo! 之前开源的 S4，流式计算研究在互联网领域持续升温。百度已经引入了流

计算系统 DStream，能提供灵活的、可伸缩的效率解决方案，又能在数据完整性、高可用、可扩展性及收缩性方面支撑上层业务。

2. 大数据挖掘

数据的价值只有通过数据挖掘才能从低价值密度的数据中发现其潜在价值，而大数据挖掘技术的实现离不开云计算技术。在业界，全球著名的 Google、EMC、惠普、IBM、微软等互联网公司都已经意识到大数据挖掘的重要意义。上述 IT 巨头们纷纷通过收购大数据分析公司，进行技术整合，希望从大数据中挖掘更多的商业价值。

数据挖掘通常需要遍历训练数据获得相关的统计信息，用于求解或优化模型参数，在大规模数据上进行频繁的数据访问需要耗费大量运算时间。数据挖掘领域长期受益于并行算法和架构的使用，使得性能逐渐提升。过去 15 年来，效果尤其显著。试图将这些进步结合起来，并且提炼。GPU 平台从并行上得到的性能提升十分显著。这些 GPU 平台由于采用并行架构，使用并行编程方法，使得计算能力呈几何级数增长。即便是图形处理、游戏编程是公认的复杂，它们也从并行化受益颇多。研究显示数据挖掘、图遍历、有限状态机是并行化未来的热门方向。

MapReduce 框架已经被证明是提升 GPU 运行数据挖掘算法性能的重要工具。D.Luo 等提出一种非平凡的策略用来并行一系列数据挖掘与数据挖掘问题，包括一类分类 SVM 和两类分类 SVM，非负最小二乘问题，及 L1 正则化回归（lasso）问题。由此得到的乘法算法，可以被直截了当地在如 MapReduce 和 CUDA 的并行计算环境中实现 [1]。K. Shim 在 MapReduce 框架下，讨论如何设计高效的 MapReduce 算法，对当前一些基于 MapReduce 的数据挖掘和数据挖掘算法进行归纳总结，以便进行大数据的分析 [2]。Junbo Zhang 等提出一种新的大数据挖掘技术，即利用 MapRedue 实现并行的基于粗糙集的知识获取算法，还提出了下一步的研究方向，即集中于用基于并行技术的粗糙集算法处理非结构化数据 [3]。F.Gao 提出了一种新的近似算法使基于核的数据挖掘算法可以有效的处理大规模数据集。当前的基于核的数据挖掘算法由于需要计算核矩阵面临着可伸缩性问题，计算核矩阵需要 $O(N^2)$ 的时间和空间复杂度来计算和存储。该算法计算核矩阵时大幅度降低计算和内存开销，而且并没有明显影响结果的精确度。此外，通过折中结果的一些精度可以控制近似水平。它独立于随后使用的数据挖掘算法并且可以被它们使用。为了阐明近似算法的效果，在其上开发了一个变种的谱聚类算法，此外设计了一个所提出算法的基于 MapReduce 的实现。在合成和真实数据集上的实验结果显示，所提出的算法可以获得显著的时间和空间节省 [4]。

Christian Kaiser 等还利用 MapReduce 框架分布式实现了训练一系列核函数学习机，该方法适用于基于核的分类和回归。Christian Kaiser 还介绍了一种扩展版的区域到点建模方法，来适应来自空间区域的大量数据 [5]。

Yael Ben-Haim 研究了三种 MapReduce 实现架构下并行决策树分类算法的设计，并在 Phoenix 共享内存架构上对 SPRINT 算法进行了具体的并行实现 [6]。

F. Yan[7] 考虑了潜在狄利克雷分配（LDA）的两种推理方法——塌缩吉布斯采样（collapsed Gibbs sampling，CGS）和塌缩变分贝叶斯推理（collapsed variational Bayesian，CVB）在 GPU 上的并行化问题。为解决 GPU 上的有限内存限制问题，F. Yan 提出一种能有效降低内存开销的新颖数据划分方案。这种划分方案也能平衡多重处理器的计算开销，并能容易地避免内存访问冲突。他们使用数据流来处理超大的数据集。大量实验表明 F. Yan 的并行推理方法得到的 LDA 模型一贯地具有与串行推理方法相同的预测能力，但在一个有 30 个多核处理器的 GPU 上，CGS 方法得到了 26 倍的加速，CVB 方法得到了 196 倍的加速。他们提出的划分方案和数据流方式使他们的方法在有更多多重处理器时可伸缩，而且可被作为通用技术来并行其它数据挖掘模型。Bao-Liang Lu 提出了一种并行的支持向量机，称为最小最大模块化网络 (M3)，它是基“分而治之”的思想解决大规模问题的有效的学习算法 [8]。针对异构云中进行大数据分析服务的并行化问题，G. Jung 提出了最大覆盖装箱算法来决定系统中多少节点、哪些节点应该应用于大数据分析的并行执行。这种方

法可以使大数据进行分配使得各个计算节点可以同步的结束计算，并且使数据块的传输可以和上一个块的计算进行重叠来节省时间。实验表明，这种方法比其他的方法可以提高大约60%的性能[9]。在分布式系统方面，Cheng等人[10]提出一个面向大规模可伸缩数据分析的可伸缩的分布式系统——GLADE。GLADE通过用户自定义聚合（UDA）接口并且在输入数据上有效地运行来进行数据分析。文章从两个方面来论证了系统的有效性。第一，文章展示了如何使用一系列分析功能来完成数据处理。第二，文章将GLADE与两种不同类型的系统进行比较：一个用UDA进行改良的关系型数据库（PostgreSQL）和MapReduce(Hadoop)。然后从运行结果、伸缩性以及运行时间上对不同类型的系统进行了比较。

3. 大数据实践

随着云计算概念的不断普及与推广，云计算核心技术的不断突破，云计算应用的不断深入，云计算得到了国内外工业界、学术界乃至政府部门的热烈响应。国内高校与科研院所针对云计算的不同领域开展了深入的研究。例如，清华大学的云存储平台着力于构建存储云，中科院计算所利用云计算开展数据挖掘工作，上海交通大学注重于数据的安全和隐私关键性技术研究。

清华大学在云存储研究方面，以分布式文件系统为基础的云存储平台，为校园网用户设计开发了用于数据存储与共享的云存储服务，利用底层云存储平台所提供的基础存储服务，提供用户管理与目录管理功能，增加了文件检索功能，并对数据传输进行了优化，为用户提供简单实用的云存储访问接口。

中国科学院计算技术研究所在Hadoop基础上开发实现了并行数据挖掘工具平台。其数据处理规模远远超出商用软件，在商用软件能承受的相同数据规模下，采用相同方法和相同参数设置，获得了一致的挖掘结果，实现了高性能、低成本的海量数据挖掘。

上海交通大学针对云计算中存在的数据安全问题，利用密码理论与技术，网络与信息安全技术，编码理论等方向所取得的成果，解决数据安全存在的一些基础问题，提高云计算的安全性。

另外，在云计算大潮中，许多本土IT厂商或是自主创新，或是强强联合，在不同的行业和领域开展了丰富多样的创新商业实践。

2008年底，中国移动建设了256台服务器，1000个CPU，256TB存储组成的“大云”试验平台，在该平台支持下，中国科学院计算技术研究所开发了基于Hadoop的并行分布式数据挖掘平台PDMiner，这是一个集成各种并行算法的数据挖掘工具平台，包括数据预处理（ETL）、数据挖掘算法、结果展示等功能。开发的并行ETL算法达到了线性加速比；可实现TB级海量数据的预处理及之后的并行挖掘分析处理，且挖掘算法随节点数线性增加，加速比随之增加。其中的并行计算模式不仅包括算法之间的并行，而且包括算法内部的并行。该系统具有运行稳定，容错能力强，扩展性好等特点。目前已用于中国移动通信企业TB级实际数据的挖掘。图1展示了PDMiner的系统架构图。

安徽科大讯飞公司针对当前移动互联网时代智能语音技术的人机交互需求，立足智能语音交互和云计算的结合，实现了面向移动互联网最终用户及开发者的科大讯飞智能语音云平台，使得手机等各种移动终端均可以通过自然的语音交互方式获取移动互联网上的各种信息和服务，提升用户获取信息的效率，以获得更好的用户体验。

四、总　结

大数据的超大容量自然需要容量大，速度快，安全的存储，满足这种要求的存储离不开云计算。高速产生的大数据只有通过云计算的方式才能在可等待的时间内对其进行处理。同时，云计算是提高对大数据的分析与理解能力的一个可行方案。大数据的价值也只有通过数据挖掘才能从低价值密度的数据中发现其潜在价

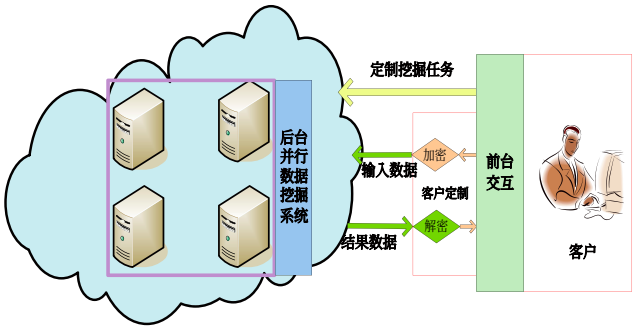

图1　PDMiner系统架构图

值，而大数据挖掘技术的实现离不开云计算技术。总之，云计算是大数据处理的核心支撑技术，是大数据挖掘的主流方式。没有互联网，就没有虚拟化技术为核心的云计算技术，没有云计算就没有大数据处理的支撑技术。

参考文献：

[1] D.Luo, C. Ding and H. Huang. Parallelization with Multiplicative Algorithms for Big Data Mining.IEEE 12th International Conference on Data Mining, 2012.

[2] K. Shim. MapReduce algorithms for big data analysis, and storage of big data, In Proceedings of the VLDB Endowment, Istanbul, Turkey, pages 2016-2017. 2012.

[3] Junbo Zhang, Tianrui Li andYi Pan, Parallel Rough Set Based Knowledge AcquisitionUsingMapReduce from Big Data, BigMine'12, pages:20-27,2012.

[4] F. Gao,W. Abd-Almageed and M. Hefeeda. Distributed Approximate Spectral Clustering forLarge-Scale Datasets. In proceedings of the 21st International ACM Symposium on High-Performance Parallel and Distributed Computing, pages 223-234, 2012.

[5] Christian Kaiser and Alexei Pozdnoukhov. Enabling real-time city sensing with kernel stream oracles and MapReduce. Pervasive and MobileComputing(2012).doi:10.1016/j.pmcj.2012.11.003

[6] Yael Ben-Haim and Elad Tom-Tov., A streaming parallel decision tree algorithm, Journal of Machine Learning Research, 11, 849-872 , 2010.

[7] F. Yan, N. Xu and Y Qi.Parallel Inference for Latent Dirichlet Allocation onGraphics ProcessingUnits.In NIPS, 2009.

[8] Bao-Liang Lu, et al. A part-versus-part method for massively parallel training of support vector machines, IEEE International Joint Conference on Neural Networks, 2004.

[9] G. Jung, N. Gnanasambandam, and T. Mukherjee. Synchronous parallel processing of big-data analytics services to optimize performance in federated clouds. In proceedings of 5th IEEEInternational Conference on Cloud Computing, CLOUD 2012, pages 811-818,2012.

[10] Yu Cheng, Chengjie Qin, Florin Rusu. GLADE: Big Data Analytics Made Easy. SIGMOD'12 , Scottsdale, Arizona, USA, 2012.

Big Data Using Cloud Computing

He Qing

Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190

Abstract: The concept of big data has been mentioned by more and more people in more and more occasions, and it is often related to cloud computing. The relationship between the cloud computing and big data becomes the hot topic. This report contains the following four thematic areas: First, the value of big data. Second, the challenges brought by big data. Third, the big data research; Fourth, Cloud Computing is a mainstream way of big data mining. In this report, we describe the understanding of big data, as well as awareness of the big data value, explore large data processing and mining technology and discuss the following points: with the absence of internet cloud computing will not exist, without cloud computing, there will be no big data processing and mining technology.

Keywords: Big Data, Could Computing, Data Mining

（责任编辑：何岸波，张志华；责任译审：龚　宇）