架构大数据

大数据技术与算法的深入解析研究

黄思行 段 昂 韦鹏程 著

中国原子能出版社
China Atomic Energy Press

架构大数据

大数据技术与算法的深入解析研究

黄思行　段　昂　韦鹏程　著

中国原子能出版社
China Atomic Energy Press

图书在版编目（CIP）数据

架构大数据 ：大数据技术与算法的深入解析研究 /
黄思行，段昂，韦鹏程著．-- 北京 ：中国原子能出版社，
2019.10
ISBN 978-7-5221-0114-9

Ⅰ．①架… Ⅱ．①黄… ②段… ③韦… Ⅲ．①数据处
理－研究 Ⅳ．①TP274

中国版本图书馆CIP数据核字（2019）第248417号

内容简介

本书从大数据架构角度全面解析大数据技术与算法，梳理了大数据技术算法，分析了大数据技术分类。如基础架构支持、大数据采集、大数据存储、大数据处理、大数据展示及交互，使人们更加深层次地、更全面地了解大数据技术以及算法。希望借助此次研究，向人们提供一个全景的大数据技术，推动我国大数据技术的发展。

架构大数据 ：大数据技术与算法的深入解析研究

出版发行	中国原子能出版社（北京市海淀区阜成路43号　100048）
策划编辑	高树超
责任编辑	王　丹　高树超
装帧设计	河北优盛文化传播有限公司
责任校对	冯莲凤
责任印制	潘玉玲
印　　刷	定州启航印刷有限公司
开　　本	710 mm×1000 mm　1/16
印　　张	20.5
字　　数	385千字
版　　次	2019年10月第1版　　2019年10月第1次印刷
书　　号	ISBN 978-7-5221-0114-9
定　　价	88.00元

发行电话：010-68452845

黄思行，男，汉族，1992 年生，湖北黄冈人，硕士研究生，研究方向：数据挖掘与自然语言处理。

段昂，男，汉族，1991 年生，重庆人，硕士研究生，重庆第二师范学院讲师，研究方向：认知无线电、大数据分析与计算。

韦鹏程，男，1975 年生，博士后，教授，重庆第二师范学院数学与信息工程学院院长，重庆邮电大学硕士生导师，先后破格晋升为讲师、副教授和教授，重庆市高校中青年骨干教师，重庆市高校优秀人才支持计划，留港、留美访问学者，重庆市儿童大数据工程实验室负责人，重庆市工程技术研究中心负责人。研究方向：保密通信、计算智能和大数据分析。主持科研项目 20 余项，发表论文 50 余篇，出版专著 5 部，获重庆市自然科学奖、科技进步三等奖各 1 项，获重庆市高等教育教学成果奖一等奖 2 项。

前　言

近年来，以物联网、移动互联网、云计算和大数据为代表的新一代信息技术发展迅猛，其中大数据风头最劲。无所不在的移动终端、智能设备、无线传感器等每分每秒都在产生数据，拥有数以亿计用户的互联网服务时时刻刻都在产生巨量的交互，如百度每天大约要处理几十拍字节的数据，Twitter 每天会产生 7 TB 的数据，Facebook 每天生成 300 TB 以上的日志数据，等等。数据产生的速度太快，要处理的数据量十分庞大。据互联网数据中心预测，到 2020 年全球将拥有 35 ZB 的数据。与此同时，数据的价值不断凸显，数据被类比为新时代的黄金和石油，现代企业快节奏的业务需求和竞争压力对数据处理的实时性和有效性提出了更高的要求，传统的数据处理技术已经完全不能满足大量数据的实时处理需求，大数据全面爆发了。大数据涉及国家战略、区域及企业发展、社会民生的方方面面，把握大数据的核心理念、模式和技术就是把握了新时代的脉搏。

预计到 2020 年，全球以电子形式存储的数据量会比 2016 年全球存储量增长 30 倍。正是在这种背景下，“大数据”的概念应运而生。大数据具有数据体量大、数据类型繁多、要求处理速度快的特征。大数据技术涵盖了从数据的海量存储、处理到应用多方面的技术，包括海量分布式文件系统、并行计算框架、NoSQL 数据库、实时流数据处理以及智能分析技术（模式识别、自然语言理解等）。

本书主要介绍大数据的含义与特征、大数据技术、大数据安全技术、大数据算法、大数据架构与分析的实现工具、大数据技术的应用领域、大数据与云计算的结合、大数据技术发展趋势、知名企业大数据架构分析等，旨在帮助大数据从业人士了解、掌握和架构大数据。

该专著是由重庆第二师范学院数学与信息工程学院黄思行、段昂、韦鹏程三位教师共同完成，并得到重庆市儿童大数据工程实验室、重庆市交互式教育电子工程技术研究中心、重庆市计算机科学与技术重点学科、重庆市计算科学与技术特色专业支持、重庆市教育委员会科学技术研究计划重点项目资助（NO. KJZD-K201801601）的支持。

目 录

第一章　大数据与大数据算法

变化是永恒的主题。由云计算、社交计算和移动计算三大趋势推动的大数据正在重塑业务流程、IT 基础设施以及我们对企业、客户和互联网信息的捕获与使用方式。近年来，“大数据”概念的提出为中国数据分析行业的发展提供了无限的空间，越来越多的人认识到了数据的价值。

第一节　大数据概述

一、大数据的含义

简单地讲，大数据就是那些超过传统数据库系统处理能力的数据，是难以用常用的软件工具在可容忍时间内抓取、管理以及处理的数据集，具有数据体量巨大、数据类型繁多、要求的处理速度快等显著特征。

大数据技术涵盖了从数据的海量存储、处理到应用多方面的技术，包括海量分布式文件系统、并行计算框架、非关系型数据库（NoSQL 数据库）、实时流数据处理以及智能分析技术（模式识别、自然语言理解、应用知识库）等。

大数据有 4 个“V”字开头的特征：volume（容量）、variety（种类）、velocity（速度）和 value（价值）。

大数据最主要的作用是服务，即面向人、机、物的服务。对于机器而言，需要数据有一些关联，如非结构化、半结构化、结构化等，使人能够从中分析出有用的信息。人、机、物对数据的参与度非常高：从数据规模上看，人到物理世界是从小到大的；从数据质量来讲，人提供的数据质量是最高的。

二、大数据技术的发展趋势

企业越来越希望能将自己的各类应用程序及基础设施转移到云平台上。就像其他 IT 系统那样，大数据的分析工具和数据库也将走向云计算。

云计算能为大数据带来哪些变化呢？首先，云计算为大数据提供了可以弹性扩展、相对便宜的存储空间和计算资源，使中小企业可以像亚马逊一样通过云计算完成大数据分析。其次，云计算 IT 资源庞大、分布较为广泛，是异构系统较多的企业及时准确处理数据的有力方式，甚至是唯一的方式。当然，大数据要走向云计算，还有赖数据通信带宽的提高和云资源池的建设，需要确保原始数据能迁移到云环境以及资源池可以弹性扩展。

三、大数据技术的研究现状与展望

大数据分析相较于传统的数据仓库应用，具有数据量大、查询分析复杂等特点。为了设计适合大数据分析的数据仓库架构，本节列举了大数据分析平台需要具备的几个重要特性，对当前的主流实现平台——并行数据库、MapReduce 及基于两者的混合式架构进行了分析归纳，指出了各自的优势及不足，同时对各个方向的研究现状及大数据分析方面进行了介绍，并展望未来。

（一）研究现状

并行数据库的最大问题在于有限的扩展能力和待改进的软件容错能力；MapReduce 的最大问题在于性能，尤其是连接操作的性能；混合式架构的关键是怎样能尽可能多地把工作推向合适的执行引擎（并行数据库或 MapReduce）。下面对近年来在这些问题上的研究进行分析归纳。

1. 并行数据库扩展性和容错性研究

华盛顿大学的文献中提出可以生成具备容错能力的并行执行计划优化器。该优化器可以依靠输入的并行执行计划、各个操作符的容错策略及查询失败的期望值等，输出一个具备容错能力的并行执行计划。在该计划中，每个操作符都可以采取不同的容错策略，在失败时重新执行其子操作符（在某节点上运行的操作符）的任务即可，避免了整个查询的重新执行。

麻省理工学院（MIT）于 2010 年设计的 Osprey 系统基于维表在各个节点全复制、事实表横向切分冗余备份的数据分布策略，将一星形查询划分为众多独立子查询。每个子查询在执行失败时都可以在其备份节点上重新执行，而不用重做整

个查询，使数据仓库查询获得了类似 MapReduce 的容错能力。

2. MapReduce 性能优化研究

MapReduce 的性能优化研究集中在对关系数据库的先进技术和特性的移植上。Facebook 和美国俄亥俄州立大学合作，将关系数据库的混合式存储模型应用于 Hadoop 平台，提出了 RCFile 存储格式。Hadoop 系统运用传统数据库的索引技术，并通过分区数据并置的方式来提升性能。基于 MapReduce 实现了以流水线方式在各个操作符间传递数据，有效缩短了任务执行时间；在线聚集的操作模式使用户可以在查询执行过程中看到部分较早返回的结果。

3. HadoopDB 的改进

HadoopDB 于 2011 年针对其架构提出了两种连接优化技术和两种聚集优化技术。

两种连接优化的核心思想都是尽可能地将数据的处理推入数据库层执行。第一种优化方式是根据表与表之间的连接关系，通过数据预分解，使参与连接的数据尽可能分布在同一数据库内，从而实现将连接操作下压进数据库内执行。该方式的缺点是应用场景有限，只适用于链式连接。第二种优化方式是针对广播式连接而设计的，在执行连接前，先在数据库内为每张参与连接的维表建立一张临时表，使连接操作尽可能在数据库内执行。该方式的缺点是较多的网络传输和磁盘 I/O 操作。

两种聚集优化技术分别是连接后聚集和连接前聚集。前者是执行完 Reduce 端连接后，直接对符合条件的记录执行聚集操作；后者是将所有数据先在数据库层执行聚集操作，然后基于聚集数据执行连接操作，并将不符合条件的聚集数据做减法操作。该方式适用的条件有限，主要用于参与连接和聚集的列的基数相乘后小于表记录数的情况。

总的来说，HadoopDB 的优化技术大都局限性较强，对复杂的连接操作仍不能下推到数据库层执行，并未从根本上解决其性能问题。

（二）研究展望

当前三个方向的研究都不能完美地解决大数据分析问题，这意味着每个方向都有极具挑战性的工作等待着我们。

并行数据库的扩展性虽有较大改善（如 Greenplum 和 Aster Data 都是面向 PB 级数据规模设计开发的），但距离大数据的分析需求仍有较大差距。因此，怎样改善并行数据库的扩展能力是一项非常有挑战性的工作。该项研究将同时涉及数

据一致性协议、容错性、性能等数据库领域的诸多方面。

混合式架构方案可以复用已有成果，开发量较小。但只是简单的功能集成，似乎并不能有效解决大数据的分析问题，因此该方向还需要更加深入的研究工作，如从数据模型及查询处理模式上进行研究，使两者能较自然地结合起来，这将是一项非常有意义的工作。中国人民大学的 Dumbo 系统即是在深层结合方向上努力的一个例子。

相较于前两者，MapReduce 的性能优化进展迅速，其性能正逐步逼近关系数据库。该方向的研究又分为两个方向：理论界侧重利用关系数据库技术及理论改善 MapReduce 的性能；工业界侧重基于 MapReduce 平台开发高效的应用软件。针对数据仓库领域，如下几个研究方向比较重要，且目前的研究较少涉及。

1. 多维数据的预计算

MapReduce 更多针对的是一次性分析操作。大数据上的分析操作难以预测，目前仍以基于报表和多维数据的分析居多。因此，MapReduce 平台可以利用预计算等手段加快数据分析的速度。基于存储空间的考虑，MOLAP 是不可取的，混合式 OLAP（HOLAP）应该是 MapReduce 平台的优选 OLAP 实现方案。具体研究如下：

（1）基于 MapReduce 框架的高效 Cube 计算算法。

（2）物化视图的选择问题，即选择物体的哪些数据问题。

（3）不同分析的物化手段及怎样基于物化的数据进行复杂分析操作。

2. 各种分析操作的并行化实现

大数据分析需要高效的复杂统计分析功能的支持。国际商业机器公司（IBM）将开源统计分析软件 R 集成进 Hadoop 平台，增强了 Hadoop 平台的统计分析功能。但更具挑战性的问题是，怎样基于 MapReduce 框架设计可并行化的、高效的分析算法。需要强调的是，鉴于移动数据的巨大代价，这些算法应基于移动计算的方式来实现。

3. 查询共享

MapReduce 采用步步物化的处理方式，导致其 I/O 代价及网络传输代价较高。一种有效的方式是在多个查询间共享物化的中间结果，甚至原始数据，以分摊代价并避免重复计算。因此，怎样在多查询间共享中间结果将是一项非常有实际应用价值的研究。

4. 用户接口

用户接口的研究方向是怎样较好地实现数据分析的展示和操作，尤其是复杂分析操作的直观展示。

5. Hadoop 可靠性

当前，Hadoop 采用的是主从结构，这就意味着主节点一旦失效，将会出现整个系统失效的局面。因此，怎样在不影响 Hadoop 现有实现的前提下，提高主节点的可靠性，将是一项切实的研究。

6. 数据压缩

MapReduce 的执行模型决定了其性能取决于 I/O 和网络传输代价。实验发现，压缩技术并没有改善 Hadoop 的性能。但实际情况是，压缩不仅可以节省空间、节省 I/O 及网络带宽，还可以利用当前中央处理器（CPU）的多核并行计算能力平衡 I/O 和 CPU 的处理能力，从而提高性能。例如，并行数据库利用数据压缩后，性能往往可以大幅提升。

7. 多维索引

多维索引的研究方向是怎样基于 MapReduce 框架实现多维索引，加快多维数据的检索速度。

此外，仍有许多其他研究工作，如基于 Hadoop 的实时数据分析、弹性研究、数据一致性研究等，这些都非常有挑战性和意义。

第二节　大数据的实用价值

一、模拟实境

运用大数据模拟实境可发掘新的需求和提高投入的回报率。现在越来越多的产品中都装有传感器，汽车和智能手机的普及更使可收集数据呈现爆炸性增长。云计算和大数据分析技术使商家可以在成本效率较高的情况下，实时地把这些数据连同交易行为的数据进行储存和分析。交易过程、产品使用和人类行为都可以数据化。大数据技术可以把这些数据整合起来进行数据挖掘，从而在某些情况下通过模型模

拟判断不同变量（如不同地区、不同促销方案）下何种方案投入回报率最高。

二、个性化精准推荐

在企业运营商内部，根据用户喜好推荐各类业务及应用是常见的，如应用商店软件推荐、交互式网络电视（IPTV）视频节目推荐等，而通过关联算法、文本摘要抽取、情感分析等智能分析算法可以将之延伸到商用化服务，利用数据挖掘技术帮助客户进行精准营销。互联网的出现更是放大了广告发送消费的特点，如人们发现自己搜索过的或者买过的商品都能被针对性地推荐，出现在浏览的网页广告中。随着信息数量的持续增加、大数据的到来，这些数据中隐藏了消费者的消费习惯、市场的变化、产品的趋势以及大量的历史记录，对企业和组织的后续运营与发展意义重大。更准确的营销手段已经成为一种广告工具，这种个性化的广告推广主要是为了缩小范围，专门针对某一类人群。

三、数据存储空间出租

企业和个人有着海量信息存储的需求，只有将数据妥善存储，才有可能进一步挖掘其潜在价值。具体而言，这块业务模式可以细分为针对个人文件存储和针对企业用户存储两大类。用户可以通过易于使用的应用程序编程接口（API）方便地将各种数据对象放在云端，然后再像使用水、电一样按用量交费。

四、数据精准搜索

数据搜索是一个并不新鲜的应用，尤其随着“大数据”时代的到来，实时性、全范围搜索的需求也变得越来越强烈。我们需要能搜索各种社交网络、用户行为等数据。运营商掌握的用户网上行为信息使所获取的数据更全面，也更具商业价值。隐私安全大数据已经与我们的生活息息相关。微博的社交关系、淘宝的购物记录、全球定位系统（GPS）的移动数据、快递的物流信息等，这些形形色色的数据包括了人们的各种行为细节，同时记录了人们大量的个人隐私。不难看出，大数据时代的到来，给传统的网络与信息安全带来了新的问题，传统防御威胁的手段已逐渐失效。大数据将安全带入了一个全新、复杂和综合的时代，不安全的那些蛛丝马迹在浩瀚数据的掩护下，正在精准地发起一次又一次的攻击。

近年来，有关网络威胁导致服务器宕机、个人和企业信息泄露事件频繁发生，网络信息安全问题已成为全球关注的焦点。然而，任何事物都具有两面性，人们常常担心大数据带来的不安全性，但大数据技术也是一种保护信息安全的工具。对于互联网而言，利用传统安全设备从终端数据或本地网络中发现未知的威胁，

就如在森林中找到指定的叶子，效率极低。

第三节　大数据算法

一、大数据算法的基本概念

我们看一看大数据问题求解的过程。我们面对的是一个计算问题，也就是说，要用计算机处理一个问题。

拿到一个计算问题之后，先要判定这个问题是否可以用计算机进行计算，如果学习过可计算性理论，就可以了解有许多问题计算机是无法计算的，如判断一个程序是否有死循环，或者是否存在能够杀所有病毒的软件。从“可计算”的角度来看，大数据上的判定问题和普通的判定问题是一样的，也就是说，如果还是用电子计算机模型（图灵机模型），那么在小数据上不可计算的问题，在大数据上也不可计算。这是因为计算模型的计算能力是一样的，只是算得快慢不同而已。

那么，大数据计算问题与传统的计算问题有什么本质区别呢？

第一个不同之处是数据量，就是说大数据处理的数据量要比传统的数据量大。第二个不同之处是有资源约束，就是说数据量可能很大，但是能真正用来处理数据的资源是有限的，这个资源包括 CPU、内存、磁盘、计算所消耗的能量。第三个不同之处是对计算时间存在约束。最简单的一个例子是基于无线传感器网络的森林防火，如果能在几秒之内自动发现火情，这个信息就是非常有价值的，若三天之后才发现火情，那么这个信息就没有价值。所以，大数据计算问题需要有一个时间约束，即到底需要多长时间得到计算结果才是有价值的。判定能否在给定数据量的数据上，在计算资源存在约束的条件下，在时间约束内完成计算任务，是大数据计算的可行性问题，需要计算复杂性理论来解决。然而，当前面向大数据的计算复杂性理论研究刚刚开始，有大量的问题需要解决。

值得说明的一点是，大数据算法分析尤为重要。这是为什么呢？对于小数据上的算法，通过实验的方法测试性能便可以很快得到结果，但是在大数据上，实验就不是那么简单了，经常需要成千上万的机器才能够得出结果。为了避免耗费如此高的计算成本，大数据算法分析就十分重要了。

经过算法设计与分析，得到了算法，接着用计算机语言来实现算法，得到一些程序模块，再用这些程序模块构建软件系统。这些软件系统需要相应的平台来实现，如 Hadoop、Spark。

大数据算法是在给定的资源约束下，以大数据为输入，在给定时间约束内可以计算出给定问题结果的算法。这个定义和传统的算法有相同的地方，即大数据算法也是一个算法，有输入有输出，且算法必须是可行的，是机械执行的计算步骤。

大数据的特点决定了大数据算法的设计方法。正如前面介绍的，大数据的特点通常用四个“V”来描述。这四个“V”中和大数据算法密切相关的有两个。一个是数据量（volume）大，也就是大数据算法必须处理足够大的数据量。另一个是速度（velocity）。速度有两方面：第一，大数据的更新速度很快，相应的大数据算法必须考虑更新算法的速度；第二，要求算法具有实时性，因此大数据算法要考虑到运算时间。对于另外两个“V”，我们可假设大数据算法处理的数据是经过预处理的，其多样性（variety）已经被屏蔽掉了。

二、大数据算法的难度

大数据具有规模大、速度快的特点，因此要设计一个大数据算法并不容易。大数据算法设计的难点主要体现在四个方面。

（一）访问全部数据时间过长

有的时候算法访问全部数据时间过长，应用无法接受。特别是数据量达到PB级甚至更大的时候，即使有多台机器一起访问数据，也是很困难的。这种情况下只能放弃使用全部数据的想法，选择通过部分数据得到一个还算满意的结果，这个结果不一定是精确的但基本满意。这就涉及一个“时间亚线性算法”的概念，即算法的时间复杂度低于数据量，算法运行过程中需要读取的数据量小于全部数据。

（二）数据难以放入内存计算

数据量非常大时，可能无法放进内存。一个有效的策略是把数据放到磁盘上，基于磁盘上的数据来设计算法，这就是所谓的外存算法。外存算法的特点是以磁盘块为处理单位，其衡量标准不再是简单的CPU时间，而是磁盘的I/O。另外一个处理方法是不对全部的数据进行计算，只向内存中放入小部分数据，仅使用内存中的小部分数据，就可以得到一个有质量保证的结果，这样的算法通常称为“空间亚线性算法”，就是说执行这一类算法所需要的空间是小于数据本身的。

（三）单个计算机难以保存全部数据，计算需要整体数据

在某些情况下，单个计算机难以保存全部数据或者在时间约束内处理全部数

据，而计算又需要整体数据，这时可以采取并行处理技术，即使用多台计算机协同工作。并行处理对应的算法是并行算法，大数据处理中常见的 MapReduce 就是一种大数据的编程模型，Hadoop 是基于 MapReduce 编程模型的计算平台。

（四）计算机计算能力不足或计算所需要的知识不足

还有一种情况是计算机的计算能力不足或者计算所需要的知识不足。例如，判断一幅图片里是不是包含猫或者狗。这时候计算机并不知道什么是猫、什么是狗，如果仅利用计算机而没有人的知识参与计算，那么这个问题会变得非常困难，计算机可能要从大量的标注图像里进行学习。但如果让人来参与，这个问题就变得简单了。更难一点儿的问题，如两个相机哪个更好，这是一个比较主观的问题，计算机是无法判断的，怎么办呢？正确的做法是采用“众包算法”，即把计算机难以计算但人计算相对容易的任务交给人来做。有时，众包算法的成本更低，算得更快。

三、大数据算法的应用

大数据算法在大数据中将扮演什么样的角色呢？我们通过下面一些例子分析大数据算法的应用。

（一）预测中的大数据算法

如何利用大数据进行预测？一种可能的方法是从多个数据源（如社交网络、互联网等）提取和预测与主题相关的数据，然后根据预测主题建立统计模型，通过训练集学习得到模型中的参数，最后基于模型和参数进行预测，其中每一个步骤都涉及大数据算法问题。在数据获取阶段，因为从社交网络或者互联网上获取的数据量很大，所以从非结构化数据（如文本）提取出关键词或者结构化数据（如元组、键值对）需要适用大数据的信息提取算法。在特征选择过程中，发现预测结果和哪些因素相关需要关联规则挖掘或者主成分分析算法。在参数学习阶段，需要机器学习算法，如梯度下降等。虽然传统的机器学习有相应的算法，但是这些算法复杂度通常较高，不适合处理大数据，因此需要面向大数据的新的机器学习算法来完成任务。

（二）推荐中的大数据算法

当前，推荐已经成为一个热门的研究分支，有大量的推荐算法提出。例如，为了减少处理数据量的奇异值分解（SVD），基于以前有哪些用户购买这个商品和

这些用户购买哪些商品的信息构成一个矩阵，这个矩阵规模非常大，以至于在进行推荐时无法使用，这就需要SVD技术对这个矩阵进行分解，将矩阵变小。同时，基于这样大规模的稀疏矩阵上的推荐需要相应的大规模矩阵操作算法。

（三）商业情报分析中的大数据算法

商业情报分析先要从互联网或者企业自身的数据仓库（如沃尔玛PB级的数据仓库）中发现与需要分析的内容密切相关的内容，继而根据这些内容分析出有价值的商业情报，这一系列操作如果利用计算机自动完成，需要算法来解决。其中，涉及的问题包括文本挖掘、机器学习，涉及的大数据算法包括分类算法、聚类分析、实体识别、时间序列分析、回归分析等。这些问题在统计学和计算机科学方面都有相关的方法提出，但面向大数据，这些方法的性能和可扩展性难以满足要求。

（四）科学研究中的大数据算法

科学研究中涉及大量的统计计算，如利用回复分析发现统计量之间的相关性，利用序列分析发现演化规律。美国能源部支持的项目中专门有一部分给了大数据算法，在其公布的指南里包括相应的研究内容，即如何从庞大的科学数据集合中提取有用的信息，如何发现相关数据间的关系（相关规则发现），以及大数据上的机器学习、数据流上的实时分析。这些都在科学研究中扮演着重要的角色。

第四节　大数据算法设计与分析

一、大数据算法设计技术

（一）精确算法设计方法

精确算法设计方法就是传统算法设计与分析课里讲授的算法，如贪心法、分治法、动态规划、搜索、剪枝。这些算法设计方法也是大数据算法设计中不可缺少的。

（二）并行算法

并行算法是一类很重要的大数据算法设计技术。在很多人的理解中，大数据算法等同于并行算法，但是大数据算法不完全是并行算法。