DASHUJU
GUANJIAN JISHU YU ZHANWANG

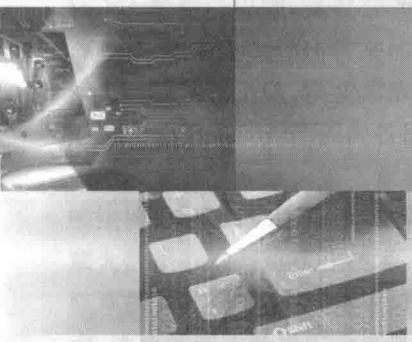
大数据关键技术与展望

孔令云 著

四川大学出版社

DASHUJU
GUANJIAN JISHU YU ZHANWANG

大数据
关键技术与展望



孔令云　著

四川大学出版社

责任编辑：唐　飞
责任校对：王　锋
封面设计：陈　勇
责任印制：王　炜

图书在版编目（CIP）数据

大数据关键技术与展望 / 孔令云著. —成都：四川大学出版社，2018.4
ISBN 978-7-5690-1707-6

Ⅰ.①大… Ⅱ.①孔… Ⅲ.①数据处理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字（2018）第 073094 号

书名　**大数据关键技术与展望**

著　　者　孔令云
出　　版　四川大学出版社
地　　址　成都市一环路南一段 24 号（610065）
发　　行　四川大学出版社
书　　号　ISBN 978-7-5690-1707-6
印　　刷　四川盛图彩色印刷有限公司
成品尺寸　170 mm×240 mm
印　　张　10.75
字　　数　203 千字
版　　次　2019 年 2 月第 1 版
印　　次　2019 年 2 月第 1 次印刷
定　　价　45.00 元

◆读者邮购本书，请与本社发行科联系。
电话：(028)85408408/(028)85401670/
(028)85408023　邮政编码：610065
◆本社图书如有印装质量问题，请
寄回出版社调换。
◆网址：http://press.scu.edu.cn

前　言

大数据是继互联网、云计算技术后世界又一热议的信息技术。随着网络信息化时代的日益普遍，移动互联、社交网络、电子商务大大扩展了互联网的疆界和应用领域，我们正处在一个数据爆炸性增长的“大数据”时代，大数据在社会经济、政治、文化、生活等方面必将产生深远的影响。

大数据的出现，不仅带来了机遇，也带来了困难与挑战。无论是科学界，还是企业界，都对大数据所带来的巨大冲击寄予厚望。对于一个国家而言，能否紧紧抓住大数据发展机遇，快速形成核心技术和应用参与新一轮的全球化竞争，将直接决定未来若干年世界范围内各国科技力量博弈的格局。

本书共分 6 章。第 1 章概论，主要阐述大数据基础、大数据度量以及大数据系统结构；第 2 章和第 3 章，主要阐释大数据存储技术以及大数据分析与挖掘；第 4 章大数据查询与可视化技术，主要对数据的查询、网络数据索引与查询技术、基于概率的大数据查询系统——Probery、数据可视化进行研究；第 5 章大数据安全技术，主要阐明大数据安全威胁、大数据安全与隐私保护技术、大数据安全趋势及应对策略以及云环境下教育大数据安全策略；第 6 章大数据技术的展望，主要探讨大数据技术的应用趋势以及数据聚合商。

本书在撰写过程中参考了大量的文献与资料，并汲取了多方人士的宝贵经验，在此向这些文献的作者表示感谢。由于大数据技术的发展日新月异，加之作者水平和学识有限，书中难免存有不妥之处，敬请广大读者批评指正。

作　者

2018 年 8 月

目　录

第1章 概 论

在这日新月异发展的社会中，人们发现未知领域的规律主要依赖抽样数据、局部数据和片面数据，甚至在无法获得真实数据时只能纯粹依赖经验、理论、假设和价值观去认识世界。因此，人们对世界的认识往往是表面的、肤浅的、简单的。然而大数据时代的来临使人类拥有更多的机会和条件在各个领域更深入地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。大数据的出现帮助商家了解用户、锁定资源、规划生产、做好运营及开展服务。本章主要阐述大数据基础、大数据度量以及大数据系统结构。

1.1 大数据基础

1.1.1 大数据的概念与特点

1.1.1.1 大数据的概念

大数据本身是一个比较抽象的概念，单从字面来看，它表示数据规模的庞大。但是仅仅数量上的庞大显然无法看出大数据这一概念和以往的“海量数据”(Massive Data)、“超大规模数据”(Very Large Data)等概念之间有何区别。针对大数据，目前存在多种不同的理解和定义。

麦肯锡在其报告《大数据：下一个创新、竞争和生产率的前沿》(*Big Data*：*The Next Frontier for Innovation*，*Competition and Productivity*)中给出的大数据定义是：大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集。但它同时强调，并不是说一定要超过特定 TB 值的数据集才能算是大数据。

维基百科对“大数据”的解读是：“大数据”(Big Data)，或称巨量数据、海量数据、大资料，指的是所涉及的数据量规模巨大到无法通过人工在合理时间内达到截取、管理、处理并整理成为人类所能解读的信息。

百度百科对“大数据”的定义为：“大数据”(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透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的目的的资讯。

Gartner Group 公司认为，“大数据”是需要新处理模式才能具有更强的决策力、洞察力和流程优化能力的海量、高增长率和多样化的信息资产。从数据的类别上看，“大数据”指的是无法使用传统流程或工具处理或分析的信息。它定义了那些超出正常处理范围和大小，迫使用户采用非传统处理方法的数据集。

按照美国国家标准与技术研究院(National Institute of Standards and Technology，NIST)发布的研究报告的定义，大数据是用来描述在我们网络的、数字的、遍布传感器的、信息驱动的世界中呈现出的数据泛滥的常用词语。大量数据资源为解决以前不可能解决的问题带来了可能性。

大数据是一个宽泛的概念，每个人的见解都不一样。笔者在综合各种观点的基础上，给出了自己的定义：“大数据”是在体量和类别特别大的杂乱数据集中，深度挖掘分析取得有价值信息的能力。大数据不仅仅在于数量的大，“大”只是信息技术不断发展所产生的海量数据的表象而已。我们更加关注“数据”的深度分析和应用，对于数据有价值的深度挖掘分析和在新形势下的数据应用是需要探讨的重点。

1.1.1.2 大数据的特点

大数据的特点主要体现为数据量大、数据类型多、处理速度快和价值密度低。

1)数据量大

人类进入信息社会以后，数据以自然方式增长，其产生不以人的意志为转移。从 1986 年到 2010 年的 20 多年时间里，全球数据的数量增长了 100 倍，今后的数据量增长速度将更快，我们正生活在一个“数据爆炸”时代。今天，世界上只有 25%的设备是联网的，大约 80%的上网设备是计算机和手机，而在不远的将来，将有更多的用户成为网民，汽车、电视、家用电器、生产机器等各种设备也将接入互联网。随着 Web 2.0 和移动互联网的快速发展，人们已经可以随时随地、随心所欲发布包括博客、微博、微信等在内的各种信息。以后，随着物联网的推广和普及，各种传感器和摄像头将遍布我们工作和生活的各个角落，这些设备每时每刻都在自

动产生大量数据。

综上所述，人类社会正经历第二次“数据爆炸”(如果把印刷在纸上的文字和图形也看作数据的话，那么，人类历史上第一次数据爆炸发生在造纸术和印刷术发明的时期)。各种数据产生速度之快，产生数量之大，已经远远超出人类可以控制的范围，“数据爆炸”成为大数据时代的鲜明特征。根据著名咨询机构 IDC(Internet Data Center)做出的估测，人类社会产生的数据一直都在以每年 50%的速度增长，也就是说，每两年就增加一倍，这被称为“大数据摩尔定律”。这意味着，人类在最近两年产生的数据量相当于之前产生的全部数据量之和。预计到 2020 年，全球将总共拥有 35ZB(见表 1-1)的数据量，与 2010 年相比，数据量将增长近 30 倍。

表 1-1　数据存储单位之间的换算关系

单位	换算关系
Byte(字节)	1Byte=8bit
KB(Kilobyte，千字节)	1KB=1024Byte
MB(Megabyte，兆字节)	1MB=1024KB
GB(Gigabyte，吉字节)	1GB=1024MB
TB(Trillionbyte，太字节)	1TB=1024GB
PB(Petabyte，拍字节)	1PB=1024TB
EB(Exabyte，艾字节)	1EB=1024PB
ZB(Zettabyte，泽字节)	1ZB=1024EB

2)数据类型多

广泛的数据来源，决定了大数据形式的多样性。以往的数据尽管数量庞大，但通常是事先定义好的结构化数据。结构化数据是将事物向便于计算机存储、处理的方向抽象后的结果，结构化数据在抽象的过程中，忽略了一些在特定的应用下可以不考虑的细节。相对于以往的结构化数据，非结构化数据越来越多，包括网络日志、音频、视频、图片、地理位置信息等，这一类数据的大小、内容、格式、用途可能完全不一样，对数据的处理能力提出了更高的要求。无论是企业还是人们日常生活中接触到的数据，绝大部分都是非结构化的。而半结构化数据介于完全结构化数据和完全非结构化数据之间，HTML 文档就属于半结构化数据，它一般是自描述的，数据的结构和内容混在一起，没有明显的区分。

3)处理速度快

大数据时代的数据产生速度非常迅速。在 Web 2.0 应用领域,在 1 分钟内,新浪可以产生 2 万条微博,Twitter 可以产生 10 万条推文,苹果可以下载 4.7 万次应用,淘宝可以卖出 6 万件商品,人人网可以发生 30 万次访问,百度可以产生 90 万次搜索查询,Facebook 可以产生 600 万次浏览量。大名鼎鼎的大型强子对撞机(LHC),大约每秒产生 6 亿次的碰撞,每秒生成约 700MB 的数据,有成千上万台计算机分析这些碰撞。

大数据时代的很多应用,都需要基于快速生成的数据给出实时分析结果,用于指导生产和生活实践,因此,数据处理和分析的速度通常要达到秒级响应,这一点和传统的数据挖掘技术有着本质的不同,后者通常不要求给出实时分析结果。

为了实现快速分析海量数据的目的,新兴的大数据分析技术通常采用集群处理和独特的内部设计以谷歌公司的 Dremel 为例,它是一种可扩展的、交互式的实时查询系统,用于只读嵌套数据分析,通过结合多级树状执行过程和列式数据结构,它能做到几秒内完成对万亿张表的聚合查询。系统可以扩展到成千上万的 CPU 上,满足谷歌上万用户操作 PB 级数据的需求,并且可以在 2～3 秒内完成 PB 级别数据的查询。

4)价值密度低

大数据虽然看起来很美,但是价值密度却远远低于传统关系数据库中已经有的那些数据。在大数据时代,很多有价值的信息都是分散在海量数据中的。以小区监控视频为例,如果没有意外事件发生,连续不断产生的数据都是没有任何价值的,当发生偷盗等意外情况时,也只有记录了事件过程的那一小段视频是有价值的。但是,为了能够获得发生偷盗等意外情况时的那一段宝贵的视频,我们不得不投入大量资金购买监控设备、网络设备、存储设备,耗费大量的电能和存储空间,来保存摄像头连续不断传来的监控数据。

如果这个实例还不够典型的话,那么可以想象另一个更大的场景。假设一个电子商务网站希望通过微博数据进行有针对性营销,为了实现这个目的,就必须构建一个能存储和分析新浪微博数据的大数据平台,使之能够根据用户微博内容进行有针对性的商品需求趋势预测。愿景很美好,但是,现实代价很大,可能需要耗费几百万元构建整个大数据团队和平台,而最终带来的企业销售利润增加额可能会比投入低许多,从这点来看,大数据的价值密度是较低的。

1.1.2 大数据的类型、潜在价值及挑战

1.1.2.1 大数据的类型

1)按照数据结构分类

按照数据结构,数据分为结构化、半结构化、非结构化数据。结构化数据是存储在数据库里,可以用二维表结构来逻辑表达实现的数据。而相对于结构化数据而言,不方便用数据库二维表结构来表现的数据即称为非结构化数据和半结构化数据。除了以上3种基本的数据类型以外,还有一种重要的数据类型为元数据。

(1)结构化数据。结构化数据指的是关系模型数据,即以关系型数据库表形式管理的数据。绝大多数的企业业务数据都以此格式进行存放。

(2)非结构化数据。相对于结构化数据(即行数据,存储在数据库里,可以用二维表结构来逻辑表达实现的数据)而言,不方便用数据库二维逻辑表来表现的数据即称为非结构化数据,包括所有格式的办公文档、文本、图片、标准通用标记语言下的子集XML、HTML、各类报表、图像和音频/视频信息等。

非结构化数据库是指其字段长度可变,并且每个字段的记录又可以由可重复或不可重复的子字段构成的数据库,用它不仅可以处理结构化数据(如数字、符号等信息),而且更适合处理非结构化数据(全文文本、图像、声音、影视、超媒体等信息)。

非结构化Web数据库主要是针对非结构化数据而产生的,与以往流行的关系数据库相比,其最大区别在于它突破了关系数据库结构定义不易改变和数据固定长度的限制,支持重复字段、子字段以及变长字段并实现了对变长数据和重复字段进行处理和数据项的变长存储管理,在处理连续信息(包括全文信息)和非结构化信息(包括各种多媒体信息)中有着传统关系型数据库所无法比拟的优势。

(3)半结构化数据。所谓半结构化数据,就是介于完全结构化数据(如关系型数据库、面向对象数据库中的数据)和完全无结构的数据(如声音、图像文件等)之间的数据,HTML文档就属于半结构化数据。它一般是自描述的,数据的结构和内容混在一起,没有明显的区分。

这样的数据和上面两种类别都不一样,它是结构化的数据,但是结构变化很大,也不能够简单地建立一个表相对应。因为要了解数据的细节,所以不能将数据简单地组织成一个文件按照非结构化数据处理。

事实上,结构化、半结构化与非结构化数据的区分,只是按数据格式进行分类,

并且由来已久。严格来讲，结构化与半结构化数据都是有基本固定结构模式的数据(即专业意义上的结构化数据)。

但将其中的关系模型数据单独定义为结构化数据，这对企业数据管理现状是可取的，并具有一定的现实意义。

另外，半结构与非结构化数据与目前流行的大数据之间只有领域重叠的关系。本质来讲，两者并无必然联系。现在有人将大数据认同为半结构化与非结构化数据的说法，是因为大数据技术最先是在半结构化数据领域发挥作用。

(4)元数据。元数据提供了一个数据集的特征和结构信息。这种数据主要由机器生成，并且能够添加到数据集中。搜寻元数据对于大数据存储、处理和分析是至关重要的一步，因为元数据提供了数据系谱信息，以及数据处理的起源。元数据的例子包括：XML 文件中提供作者和创建日期信息的标签，数码照片中提供文件大小和分辨率的属性文件。

2)按照数据作用方式分类

按照数据作用的方式，分为交易数据和交互数据。

交易数据是指来自电子商务和企业应用的数据，包括 ERP、企业对企业(B2B)、企业对个人(B2C)、个人对个人(C2C)、团购等系统，这些数据存储在关系型数据库和数据仓库中，可以执行联机事务处理(OLTP)和联机分析处理(OLAP)。这些数据的规模和复杂性一直在提高。

交互数据指来自相互作用的社交网络的数据，包括社交媒体交互(人为生成交互)和机器交互(设备生成交互)的新型数据。

两类数据的有效融合将是大势所趋，大数据应用要有效集成这两类数据，并在此基础上，实现这些数据的处理和分析。

3)按照产生主体分类

按照产生主体，数据分为企业数据、机器数据、社会化数据。

(1)企业数据(Enterprise Data)。2010 年，全球企业新存储的数据超过了7000PB，全球消费者新存储的数据约为 6000PB，每一天都有无数的数据被收集、交换、分析和整合。数据已经如一股“洪流”注入了世界经济，成为全球各个经济领域的重要组成部分，数据将和企业的固定资产、人力资源一样，成为生产过程中的基本要素。

2011 年，麦肯锡在其研究报告《大数据：下一个创新、竞争和生产率的前沿》中指出，在美国，仅仅制造行业就拥有比美国政府多一倍的数据，此外，新闻业、银行业、医疗业、投资业、零售业都拥有可以和美国政府相提并论的海量数据。

据互联网数据中心IDC发布的《中国大数据技术与服务市场2012—2016年预测与分析》报告显示该市场规模将会从2011年的7760万美元增长到2016年的6.17亿美元，未来5年的复合增长率达51.4%，市场规模增长近7倍。庞大的数据来源所带来的量化转变在企业界已经迅速蔓延。

(2)机器数据(IT Data)。大数据中，机器数据是份额最大且增长最快的一部分。每个现代企业机构，无论规模大小，都会产生海量的机器数据，如何管理和利用机器数据，进行业务创新并获取竞争优势，已经成为目前企业或机构所面临的关键任务。

机器数据，顾名思义，是由机器(软硬件系统)产生的数据，也是大数据最原始的数据类型，它通常包括所有软硬件设备生产的信息，这些数据包括了日志文件、交易记录、网络消息、传感器采集的数据等，这些信息几乎包含了所有客户、交易、设备等元素的动作行为。

在大数据时代，结合IT运维、系统安全、搜索引擎、电子商务等特定应用的需求实现大数据环境下机器数据的存储、管理、检索和分析将是目前企业或机构管理和利用机器数据的重点所在。

(3)社会化数据(Social Data)。随着社交网络的流行，国内外社会化媒体得到了迅猛发展。截至2012年10月，Facebook的用户数超过10亿，Twitter的用户数超过5亿。据中国互联网络信息中心(CNNIC)最新发布的报告显示，中国的网民已达5.55亿，其中超过4亿的用户分布在微博、SNS、个人空间等社会化媒体上。

集中在社会化媒体上庞大的用户群及发生的用户行为将会产生巨量的数据回馈，这些包括评论、视频、照片、地理位置、个人资料、社交关系等由用户在社会化媒体中产生或分享的各类信息即为社会化数据。

社会化数据与以前采集的静态的、事务性数据完全不一样，它具有实时性和流动性。人们在社会化媒体上通过交流、购买、出售和其他日常活动以免费的方式提供着大量信息。这些数据由每个网民的微行为汇集而成，蕴含着巨大的价值，将带来政府在公共管理方面、企业在市场调研和营销方面的变革。

1.1.2.2 大数据的潜在价值

大数据的潜在价值可以通过数据结构的复杂性和关联性体现出来。当提到大数据时，我们最先想到的一定是体量大，但体量大的数据如果仅是简单的数据堆砌，或者仅是对单一类型数据的记录，那么这种重复性高、结构简单的数据还不能

被称为大数据。例如，在一个购物商场内，商品种类有上千种，每种商品又有来自不同公司的产品，再加上购物、休闲、娱乐、餐饮等信息，则它拥有的数据就能从各个维度反映出顾客的行为特征，从而蕴含更大的数据价值。

大数据潜在价值的另一个体现是其关联性。大数据的重要来源之一是互联网行业。随着移动互联网的发展及互联网普及率的提升，网民上网行为呈现出跨网站、跨终端、跨平台等特点，用户数据不仅包括人与人交流产生的数据，还包括人机交互及机器与机器间通信产生的数据。这些数据之间如果没有较明显的逻辑关系和确定的关联关系，则数据价值的挖掘就会变得相当困难，同时数据价值也相应低很多。所以数据之间的逻辑性和关联性也是数据潜在价值的蕴藏点。

大数据潜在价值的实现包括 3 个层次，即社会领域、行业领域及企业发展领域。大数据最终需要解决的问题主要集中在这样 3 个层面上：一是宏观层面，主要是应用于社会领域，如智慧交通、智慧城市和灾难预警等；二是中观层面，主要表现在提升行业生产率水平、促进行业的融合发展以及促进行业内商业模式的变革等；三是微观层面，主要表现在促进客户服务水平的提升、企业流程的创新、内部运营成本的降低及供应链的协调和改善等。

1.1.2.3 大数据的挑战

1)外部业务需求的数据转换

由移动智能终端、物联网、云计算引发的大数据趋势，不仅改变了人们的生活方式，也要求企业重新设计考虑原来的运作模式，以数据驱动满足新的外部业务需求。但是，通常业务管理人员和后台技术人员使用的语言是不同的。业务管理人员会加入自己领域的术语和解释，技术人员会从系统实现的角度解释需求，两者的转换变得较为困难。因此，需要了解面向业务级的数据应用，针对不同业务部门的具体需求，统一业务语义模型和数据逻辑建模，根据需求合并、汇总业务数据，满足业务分析、挖掘和查询需求的变化。

2)大数据技术运用仍存在困难

在实际生产中，有些行业的数据涉及上百个参数，其复杂性不仅体现在数据样本本身，更体现在多源异构、多实体和多空间之间的交互动态性，难以用传统的方法描述与度量。而现有的数据处理方法仅适用于结构化数据，无法将大量的非结构化数据与结构化数据进行统一、整合。如何对跨业务平台的数据进行关联，并全面实时地给出分析结果，也是大数据技术需要面临的一个挑战。

3)用户隐私与便利性的冲突

通过对大量用户数据的分析,可以有效提升用户服务。但是,搜集的用户数据成为一个具有价值的整体,无论是对用户隐私还是数据本身,都成为具有争议的灰色地带。例如,华尔街一位股票炒家利用电脑程序分析全球3.4亿微博账户的留言,以此判断民众情绪。这时提供数据的众多微博用户便成为被利用的对象。因此,如何在挖掘数据价值和个人隐私保护之间寻求平衡,防止数据窃取、非法添加或篡改等情况的出现,是大数据需要解决的另一个难题。

4)数据安全风险更加凸显

大数据的发展需要加大信息的开放程度,设计出新的信息收集设备,并为海量数据的存续和分析提供支持。由于数据存储和应用方式出现新的变化,可能带来的副作用是:IT基础架构将变得越来越一体化和外向型,对数据安全和知识产权构成更大风险。若企业不了解大数据内涵,则更增加了其风险成本。因此,企业需要关注完整的数据生命周期,包括数据质量、数据保留度、数据整合、数据安全性和信息隐私等内容。

5)数据分析与管理人才紧缺

大数据时代,企业、组织、政府需要大量既精通业务又能进行大数据分析的人才。研究表明,在美国对拥有深厚海量数据分析(包括机器学习和高级统计分析)技能人才的需求,可能超出预测供应量的50%~60%。因此,如何培养大量大数据分析人才是当务之急,这对现有人才培养机制提出了新的挑战。

1.2 大数据度量

1.2.1 大数据度量概述

大数据已经在社会政治、经济、文化、教育及科技等各个方面产生了巨大的推动作用,各行各业都在研究、分析大数据。如何看待大数据,大数据到底有什么能力,对如何度量其本身能力和价值的研究分析少之又少,本节就大数据的度量做简单阐述。

大数据能够带来巨大的效益,但是到目前为止,对于大数据的度量,业界并没有形成任何有效的度量体系和架构。下面从5个方面对大数据的度量建立一个初

步的度量体系指标，如图 1-1 所示。

图 1-1　大数据度量指标

大数据的度量指标包含 5 个部分：大数据能耗度量、大数据计算能力度量、大数据的数据中心服务能力度量、大数据商业与社会价值度量、大数据冷热度度量。

1.2.2　大数据度量分析

1.2.2.1　大数据能耗度量

大数据的飞速发展带来的一个直接影响就是需要建立能耗量极大的数据中心，数据中心的服务器基本处于连续运行状态，对能源的消耗巨大。

众多的数据密集型应用公司需要建立自己的发电站才能满足本公司的数据中心正常运转。以 Google 公司为例，其分布在全球各地的上百万台服务器，如果一直不停机地运转，消耗的能量与一座中等规模城市的用电量相当。因此，美国很多的大型互联网公司将数据中心建在沙漠、河边或者电站旁边，以利用良好的冷却系统和电站对数据中心进行供电。

大数据与生俱来的最直接的一个需求就是用来存储和计算大数据的数据中心建设。数据中心一个很大的度量指标是能耗问题。大数据的存储利用率、计算利用率、存储副本数量及死数据率均是影响能耗的关键因素。以下是一种能源消耗的度量指标。

$$Energy_{\text{consumptionRate}} = \frac{1}{4}\left(\frac{Storage_{\text{used}}}{Storage_{\text{all}}} \cdot Storage_{\text{security}} + \frac{Computing_{\text{used}}}{Computing_{\text{all}}} \cdot Computing_{\text{security}} + \frac{Replias_{\text{rational}}}{Replias_{\text{reality}}}\right.$$

$$+ Data_{wholeVolume} - \frac{Data_{dead}}{Datawhole_{Volume}})$$

式中：$Energy_{consumptionRate}$——数据中心的能耗比率；

$\frac{Storage_{used}}{Storage_{all}}$——实际使用的存储容量和存储总容量的比率；

$Storage_{security}$——存储预留的安全系数，例如，安全系数为1.2，则表明至少要多预留20%的安全存储空间；

$\frac{Computing_{used}}{Computing_{all}}$——实际计算所需资源（CPU、内存等）和总共计算资源（CPU、内存等）的比率；

$Computing_{security}$——计算预留的安全系数，例如，安全系数为1.2，则表明至少要多预留20%的安全计算资源（CPU、内存等）；

$\frac{Replias_{rational}}{Replias_{reality}}$——合理的副本数量和实际采用的副本数量的比率；

$\frac{Data_{wholeVolume} - Data_{dead}}{Data_{wholeVolume}}$——实际应该带电存储的大数据（需要除去死数据）和实际带电存储的大数据的比率。

1.2.2.2 大数据计算能力度量

很难有一个计算标准来度量大数据的计算能力，因此大数据计算能力的度量指标非常复杂。

关于大数据的计算能力度量，有如下两个度量指标：大数据的可计算性度量、可降维性度量。

1）大数据的可计算性度量

可计算性是大数据计算能力度量中一个最为重要的技术指标，主要包括大数据的时间计算复杂度与大数据的空间计算复杂度。由于大数据的类型众多、结构各异及数据容量极大，它是否具有可计算性、可计算性有多大是决定该大数据是否具有价值的重要衡量指标。

2）大数据的可降维性度量

如何实现降维是大数据计算的一个重要方面。大数据结构复杂并且维度众多，如果不进行一定程度的降维，将很难甚至无法实现计算。如何实现大数据计算的降维、对其可降维性如何进行度量也是一个重要指标。现在众多的聚类算法等在实施计算之前，均需降低维度。