JIYU YUNJISUAN DE DASHUJU
CHULI JISHU FAZHAN YU YINGYONG

基于云计算的大数据处理技术发展与应用

申时凯　佘玉梅　著

电子科技大学出版社
University of Electronic Science and Technology of China Press

基于云计算的大数据处理技术发展与应用

申时凯　佘玉梅　著

电子科技大学出版社
University of Electronic Science and Technology of China Press

图书在版编目(CIP)数据

基于云计算的大数据处理技术发展与应用 / 申时凯，佘玉梅著. -- 成都：电子科技大学出版社，2018.6
ISBN 978-7-5647-6486-9

Ⅰ. ①基… Ⅱ. ①申… ②佘… Ⅲ. ①云计算－数据处理 Ⅳ. ①TP393.027②TP274

中国版本图书馆CIP数据核字(2018)第147594号

基于云计算的大数据处理技术发展与应用

申时凯　佘玉梅　著

策划编辑　李述娜

责任编辑　李述娜

出版发行　电子科技大学出版社

成都市一环路东一段159号电子信息产业大厦九楼　邮编　610051

主　　页　www.uestcp.com.cn

服务电话　028-83203399

邮购电话　028-83201495

印　　刷　定州启航印刷有限公司

成品尺寸　185mm × 260mm

印　　张　15

字　　数　335千字

版　　次　2019年3月第一版

印　　次　2019年3月第一次印刷

书　　号　ISBN 978-7-5647-6486-9

定　　价　53.00元

版权所有，侵权必究

前言

现在的社会是一个高速发展的社会，科技发达，信息流通，人们之间的交流越来越密切，生活也越来越方便，大数据就是这个高科技时代的产物。

随着网络带宽的不断增长，通过网络访问非本地的计算服务（包括数据处理、存储和信息服务等）的条件越来越成熟，于是就有了今天我们称作“云计算”的技术。之所以称作“云”，是因为计算设施不在本地而在网络中，用户不需要关心它们所处的具体位置，于是我们就像以前画网络图那样，用“一朵云”来代替了。其实，云计算模式的形成由来已久（Google 公司从诞生之初就采用了这种模式），但只有当宽带网普及到一定程度，且网格计算、虚拟化、SOA 和容错技术等成熟到一定程度并融为一体，又有业界主要大公司的全力推动和吸引人的成功应用案例时，它才如同一颗新星闪亮登场。

云计算（cloud computing）是基于互联网的相关服务的增加、使用和交付模式，通常涉及通过互联网来提供动态易扩展且经常是虚拟化的资源。云是网络、互联网的一种比喻说法。过去在图中往往用云来表示电信网，后来也用来表示互联网和底层基础设施的抽象。因此，云计算甚至可以让你体验每秒 10 万亿次的运算能力，拥有这么强大的计算能力可以模拟核爆炸、预测气候变化和市场发展趋势。用户通过电脑、笔记本、手机等方式接入数据中心，按自己的需求进行运算。

大数据时代已经来临，它将在众多领域掀起变革的巨浪。但我们要冷静地看到，大数据的核心在于为客户挖掘数据中蕴藏的价值，而不是软硬件的堆砌。因此，针对不同领域的大数据应用模式、商业模式研究将是大数据产业健康发展的关键。我们相信，在国家的统筹规划与支持下，通过各地方政府因地制宜制定大数据产业发展策略，通过国内外 IT 龙头企业以及众多创新企业的积极参与，大数据产业未来发展前景十分广阔。

本书共八章，从概念到应用，由浅入深，全面深入的探析了基于云计算的大数据处理技术。结合云计算技术对大数据架构进行分析，并对大数据的发展和应用进行探索，根据大数据的巨量分析，进一步对机器学习和数据挖掘等领域进行研究，结合实际应用，从而让读者更好的了解云计算与大数据，对未来有更高的期待。

全书由申时凯、佘玉梅共同策划，其中，第 1 章～第 5 章由申时凯撰写，第 6 章～第 8 章由佘玉梅撰写，云南师范大学研究生钱晓如对第 1 章～第 5 章进行了仔细检查和修改。

本书的研究工作得到了中央财政支持地方高校发展专项资金项目“应用型高校教育大数据平台建设”和“物联网工程专业教学实验平台”、教育部－华为技术有限公司2017年第二批产学合作协同育人项目“昆明学院路由交换与信息安全创新创业教育改革项目”、教育部2017年第二批产学合作协同育人项目“北京赛佰特产学合作协同育人实践条件项目”、昆明市物联网应用技术科技创新团队（合同编号：昆科计字2016-2-R-07793号）、昆明学院物联网应用技术科研创新团队（NO.2015CXTD04）、昆明学院应用型人才培养改革创新项目－应用型本科计算机类专业实践教学基地的资助。

由于作者水平所限，书中错误和不足之处在所难免，恳请专家、读者批评指正。

申时凯　佘玉梅

2018年2月于昆明

此为试读，需要完整PDF请访问：www.ertongbook.com

目录

第一章　云计算与大数据的概述

第一节　云计算的概述

一、定义

Gartner 公司在其报告中将云计算放在战略技术领域的前沿，进一步重申了云计算是整个行业的发展趋势。在这份报告中，Gartner 公司将云计算正式定义为："……一种计算方式，能通过 Internet 技术将可扩展的和弹性的 IT 能力作为服务交付给外部用户。"

这个定义对 Gartner 公司 2008 年的原始定义做了一点修订，将原来的"大规模可扩展性"修改为"可扩展的和弹性的"。这表明了可扩展性与垂直扩展能力相关的重要性，而不仅仅与规模庞大相关。

Forrester Research 公司将云计算定义为："……一种标准化的 IT 性能（服务、软件或者基础设施），以按使用付费和自助服务方式，通过 Internet 技术进行交付。"

该定义被业界广泛接受，它是由美国国家标准与技术研究院（NIST）制定的。早在 2009 年，NIST 就公布了其对云计算的原始定义，随后在 2011 年 9 月，根据进一步评审和企业意见，发布了修订版定义："云计算是一种模型，可以实现随时随地、便捷地、按需地从可配置计算资源共享池中获取所需的资源（例如，网络、服务器、存储、应用程序及服务），资源可以快速供给和释放，使管理的工作量和服务提供者的介入降低至最少。这种云模型由五个基本特征、三种服务模型和四种部署模型构成。"

二、云资源

云（cloud）是指一个独特的 IT 环境，其设计目的是为了远程供给可扩展和可测量的 IT 资源。这个术语原来用于比喻 Internet，意为 Internet 在本质上是由网络构成的网络，用于对一组分散的 IT 资源进行远程访问。在云计算正式成为 IT 产业的一部分之前，云符号作为 Internet 的代表，出现在各种基于 Web 架构的规范和主流文献中。现在，同样的符号则专门用于表示云环境的边界。

区分术语"云"、云符号与 Internet 是非常重要的。作为远程供给 IT 资源的特殊环境，

云具有有限的边界。通过 Int 进制 ernet 可以访问到许多单个的云。

Internet 提供了对多种 Web 资源的开放接入，与之相比，云通常是私有的，而且对提供的 IT 资源的访问也是需要计量的。

Internet 主要提供了对基于内容的 IT 资源的访问，这些资源是通过万维网发布的。而对于由云环境提供的 IT 资源来说，主要提供的是后端处理能力和对这些能力进行基于用户的访问。另一个关键区别在于，虽然云通常是基于 Internet 协议和技术的，但是它并非必须基于 Web。这里的协议是指一些标准和方法，它们使得计算机能以预先定义好的结构化方式相互通信。而云可以基于任何允许远程访问其 IT 资源的协议。

IT 资源（ITresource）是指一个与 IT 相关的物理的或虚拟的事物，它既可以是基于软件的，比如虚拟服务器或定制软件程序，也可以是基于硬件的，比如物理服务器或网络设备。

一个给定云符号边界中画出的 IT 资源并不代表这个云中包含的所有可用 IT 资源。为了说明一个特定的话题，通常只突出显示一部分 IT 资源。当重点集中在一个问题的某些方面时，就需要特意用抽象图示来表示底层技术架构。这就意味着，在图示中只会显示实际技术的部分细节。

作为一个独特且可以远程访问的环境，云代表了 IT 资源的一种部署方法。处于一个组织边界（并不特指云）中的传统 IT 企业内部承载的 IT 资源被认为是位于 IT 企业内部的，简称为内部的。换句话说，术语“内部的”是指“在一个不基于云的可控的 IT 环境内部的”，它和“基于云的”是对等的，用来对 IT 资源进行限制。一个内部的 IT 资源不可能是基于云的，反之亦然。

有三点需要注意：

- 一个内部的 IT 资源可以访问一个基于云的 IT 资源，并与之交互 C。
- 一个内部的 IT 资源可以被迁移到云中，从而成为一个基于云的 IT 资源。
- IT 资源既可以冗余部署在内部的环境中，也可以在云环境中。

如果在私有云中，难以区分是企业内部的 IT 资源还是基于云的 IT 资源，那么就需要使用明确的限定词。

提供基于云的 IT 资源的一方称为云提供者（cloud provider），使用基于云的 IT 资源的一方称为云用户（cloud consumer）。这两个术语通常代表的是与云及相应云供应合同相关的组织所承担的角色。

第二节　大数据的发展历程

一、国际发展历程

大数据的历史最早可以追溯到十八世纪八十年代，1885–1890 美国统计学家赫尔曼·霍

尔瑞斯为了统计1890年的人口普查数据，发明了一台电动器来读取卡片上的洞数，该设备让美国用一年时间就完成了原本耗时8年的人口普查活动，由此在全球范围内引发了数据处理的新纪元。

1944年，卫斯理大学图书馆员弗莱蒙特·雷德对大数据时代的到来进行了预见。他出版了《学者与研究型图书馆的未来》一书，在书中他估计美国高校图书馆的规模每16年就翻一番。

1961年德里克·普赖斯出版了《巴比伦以来的科学》，在这本书中，普赖斯通过观察科学期刊和论文的增长规律来研究科学知识的增长。他得出以下结论：新期刊的数量以指数方式增长而不是以线性方式增长，每15年翻一番，每50年以10为指数倍进行增长。普赖斯将其称之为“指数增长规律”。

1980年4月I·A·特詹姆斯兰德在第四届美国电气和电子工程师协会（IEEE）“大规模存储系统专题研讨会”上做了一个报告，题为《我们该何去何从？》。在报告中，他指出所有数据正在被无选择地保存以避免错失有价值的信息。

1981年匈牙利中央统计办公室开始实施了一项调查国家信息产业的研究项目，包括以比特为单位计量信息量。这项研究一直持续至今。

1986年7月哈尔·B·贝克尔在《数据通信》上发表了《用户真的能够以今天或者明天的速度吸收数据吗？》一文，预计数据记录密度将大幅增长。

1993年，匈牙利中央统计办公室首席科学家伊斯特万·迪恩斯编制了一本国家信息账户的标准体系手册。

1997年10月，迈克尔·考克斯和大卫·埃尔斯沃思在第八届美国电气和电子工程师协会（IEEE）关于可视化的会议论文集中发表了《为外存模型可视化而应用控制程序请求页面调度》的文章。这是在美国计算机学会的数字图书馆中大数据发展历程综述第一篇使用“大数据”这一术语的文章。

1999年8月，史蒂夫·布赖森、大卫·肯怀特、迈克尔·考克斯、大卫·埃尔斯沃思以及罗伯特·海门斯在《美国计算机协会通讯》上发表了《千兆字节数据集的实时性可视化探索》一文。这是《美国计算机协会通讯》上第一篇使用“大数据”这一术语的文章。

2001年，美国一家在信息技术研究领域具有权威地位的咨询公司Gartner首次开发了大数据模型。

2001年2月，梅塔集团分析师道格·莱尼发布了一份研究报告，题为《3D数据管理：控制数据容量、处理速度及数据种类》。十年后，3V作为定义大数据的三个维度而被广泛接受。

2005年Hadoop项目诞生。Hadoop是由多个软件产品组成的一个生态系统，这些软件产品共同实现全面功能和灵活的大数据分析。

2007年，著名图灵奖获得者Jim Gray在的一次演讲中提出，“数据密集型科学发现”（Data-Intensive Sci-entific Discovery）将成为科学研究的第四范式。

2008 年末，“大数据”得到部分美国知名计算机科学研究人员的认可，业界组织计算社区联盟（Computing Community Consor-tium），发表了一份有影响力的白皮书《大数据计算：在商务、科学和社会领域创建革命性突破》。它使人们的思维不仅局限于数据处理的机器，此组织可以说是最早提出大数据概念的机构。

2008 年，在 Google 成立 10 周年之际，著名的《自然》杂志出版了一期专刊，专门讨论未来的大数据处理相关的一系列技术问题和挑战，其中就提出了“Big Data”的概念。

大约从 2009 年开始，“大数据”逐渐成为互联网信息技术行业的流行词汇。

2009 年印度政府建立了用于身份识别管理的生物识别数据库，联合国全球脉冲项目已研究了对如何利用手机和社交网站的数据源来分析预测从螺旋价格到疾病爆发之类的问题。

2009 年中，美国政府通过启动 Data.gov 网站的方式进一步开放了数据的大门，这个网站向公众提供各种各样的政府数据，这一行动激发了从肯尼亚到英国范围内的政府们相继推出类似举措。

2010 年 2 月，肯尼斯库克尔在《经济学人》上发表了长达 14 页的大数据专题报告《数据，无所不在的数据》。库克尔在报告中提到：“世界上有着无法想象的巨量数字信息，并以极快的速度增长。科学家和计算机工程师已经为这个现象创造了一个新词汇：‘大数据’。”库克尔也因此成为最早洞见大数据时代趋势的数据科学家之一。

2010 年 12 月，美国总统办公室下属的科学技术顾问委员会 (PCAST) 和信息技术顾问委员会 (PITAC) 向奥巴马和国会提交了一份《规划数字化未来》的战略报告，把大数据收集和使用的工作提升到体现国家意志的战略高度。

2011 年 2 月，IBM 的沃森超级计算机每秒可扫描并分析 4TB（约 2 亿页文字量）的数据量，并在美国著名智力竞赛电视节目“Jeopardy”（危险边缘）上击败两名人类选手而夺冠。

后来纽约时报认为这一刻为一个“大数据计算的胜利”。

2011 年 5 月，全球知名咨询公司麦肯锡的全球研究院（MGI）发布了一份报告——《大数据：创新、竞争和生产力的下一个新领域》，这项研究估计 2010 年所有的公司存储了 7.4EB 新产生的数据，消费者存储了 6.8EB 新数据。大数据开始备受关注，这也是专业机构第一次全方面的介绍和展望大数据。

2012 年 1 月份，瑞士达沃斯召开的世界经济论坛上，大数据是主题之一，会上发布的报告《大数据，大影响 (Big Data，Big Impact)》宣称，数据已经成为一种新的经济资产类别。

2012 年美国总统选举中，那些精于数字计算的数据挖掘团队把传统的投票放在一边不用，而是利用“大数据”来规划这次选举将如何进行。如利用房产记录、选举记录甚至是期刊的订阅注册等来预测人们对候选人的看法、这些看法是否能被改变，以及为此要采取怎样的措施等。

2012 年 3 月，美国奥巴马政府在白宫网站发布了《大数据研究和发展倡议》，这一倡议

此为试读，需要完整PDF请访问：www.ertongbook.com

标志着大数据已经成为重要的时代特征。

2012 年 3 月 22 日，奥巴马政府宣布 2 亿美元投资大数据领域，是大数据技术从商业行为上升到国家科技战略的分水岭，在次日的电话会议中，政府对数据的定义“未来的新石油”，大数据技术领域的竞争，事关国家安全和未来。

2012 年 4 月，美国软件公司 Splunk 于 19 日在纳斯达克成功上市，成为第一家上市的大数据处理公司。

Splunk 成功上市促进了资本市场对大数据的关注，同时也促使 IT 厂商加快大数据布局。

2012 年 7 月，联合国在纽约发布了一本关于大数据政务的白皮书《大数据促发展：挑战与机遇》，全球大数据的研究和发展进入了前所未有的高潮。这本白皮书总结了各国政府如何利用大数据响应社会需求，指导经济运行，更好地为人民服务，并建议成员国建立“脉搏实验室”(Pulse Labs)，挖掘大数据的潜在价值。

2014 年 4 月，世界经济论坛以“大数据的回报与风险”主题发布了《全球信息技术报告（第 13 版）》。报告认为，在未来几年中针对各种信息通信技术的政策甚至会显得更加重要，接下来将对数据保密和网络管制等议题展开积极讨论。

2014 年 5 月，美国白宫发布了 2014 年全球“大数据”白皮书的研究报告《大数据：抓住机遇、守护价值》。报告鼓励使用数据以推动社会进步，同时，也需要相应的框架、结构与研究，来帮助保护美国人对于保护个人隐私、确保公平或是防止歧视的坚定信仰。

由于大数据技术的特点和重要性，目前国内外已经出现了“数据科学”的概念，即数据处理技术将成为一个与计算科学并列的新的科学领域。

二、国内发展状况

为了紧跟全球大数据技术发展的浪潮，我国政府、学术界和工业界对大数据也予以了高度的关注。

2011 年 12 月，工信部发布的物联网十二五规划上，把信息处理技术作为 4 项关键技术创新工程之一被提出来，其中包括了海量数据存储、数据挖掘、图像视频智能分析，这都是大数据的重要组成部分。

2012 年 7 月，为挖掘大数据的价值，阿里巴巴集团在管理层设立“首席数据官”一职，负责全面推进“数据分享平台”战略，并推出大型的数据分享平台“聚石塔”，为天猫、淘宝平台上的电商及电商服务商等提供数据云服务。

随后，阿里巴巴董事局主席马云在 2012 年网商大会上发表演讲，称从 2013 年 1 月 1 日起将转型重塑平台、金融和数据三大业务。阿里巴巴也是最早提出通过数据进行企业数据化运营的企业。

为了推动我国大数据技术的研究发展，2012 年中国计算机学会 (CCF) 发起组织了 CCF 大数据专家委员会，CCF 专家委员会还特别成立了一个“大数据技术发展战略报告”撰写组，并已撰写发布了《2013 年中国大数据技术与产业发展白皮书》。

2013 年 4 月 14 日和 21 日，央视著名“对话”节目邀请了《大数据时代——生活、工作与思维的大变革》作者维克托·迈尔·舍恩伯格，以及美国大数据存储技术公司 LSI 总裁阿比分别做客“对话”节目，做了两期大数据专题谈话节目“谁在引爆大数据”、“谁在掘金大数据”，国家央视媒体对大数据的关注和宣传体现了大数据技术已经成为国家和社会普遍关注的焦点。

国内的学术界和工业界也都迅速行动，广泛开展大数据技术的研究和开发。

2013 年以来，国家自然科学基金、973 计划、核高基、863 等重大研究计划都已经把大数据研究列为重大的研究课题。

清华信息学院、国家实验室也成立了数据科学院，并于 2014 年 12 月 22 日举办了“大数据论坛——数据科学与技术”，对大数据发展战略和各大数据专项进行了探讨。

第三节　大数据的现状与挑战

大数据分析相比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，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、Map Reduce 及基于两者的混合架构进行了分析归纳，指出了各自的优势及不足，同时也对各个方向的研究现状及大数据分析方面进行介绍，并展望未来。

一、大数据的挑战

最近几年，数据仓库又成为数据管理研究的热点领域，主要原因是当前数据仓库系统面临的需求在数据源、需提供的数据服务和所处的硬件环境等方面发生了根本性的变化，这些变化是我们必须面对的。

（一）三个变化

1. 数据量

由 TB 级升到 PB 级，并仍在持续爆炸式增长。2011 年经调查显示，最大的数据仓库中的数据量，每两年增加 3 倍（年均增长率为 173%），其增长速度远超摩尔定律增长速度。照此增长速度计算，最近几年最大数据仓库中的数据量将逼近 100PB。

2. 分析需求

由常规分析转向深度分析（Deep Analytic s）。数据分析日益成为企业利润必不可少的支撑点。根据 TDWI（中国商业智能网）对大数据分析的报告，如图 1–1 所示，企业已经不满足于对现有数据的分析和监测，而是期望能对未来趋势有更多的分析和预测，以增强企业竞争力。这些分析操作包括诸如移动平均线分析、数据关联关系分析、回归分析、市场分析等复杂统计分析，我们称之为深度分析。

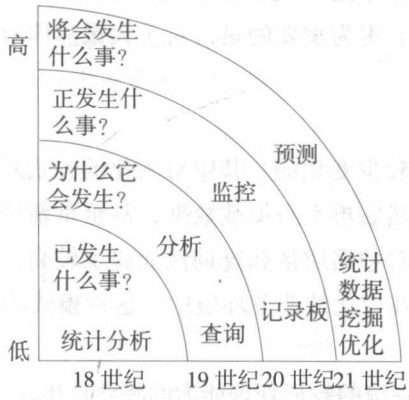

图 1-1　分析的趋势图

3. 硬件平台

由高端服务器转向由中低端硬件构成的大规模机群平台。由于数据量的迅速增加，并行数据库的规模不得不随之增大，从而导致其成本的急剧上升。出于成本的考虑，越来越多的企业将应用由高端服务器转向了由中低端硬件构成的大规模机群平台。

（二）两个问题

图 1-2 所示为一个典型的数据仓库架构。

由图 1-2 可以看出，传统的数据仓库将整个实现划分为 4 个层次，数据源中的数据首先通过 ETL 工具被抽取到数据仓库中进行集中存储和管理，再按照星形模型或雪花模型组织数据，然后由 OLAP 工具从数据仓库中读取数据，生成数据立方体（MOLAP）或者直接访问数据仓库进行数据分析（ROLAP）。在大数据时代，此种计算模式存在以下两个问题。

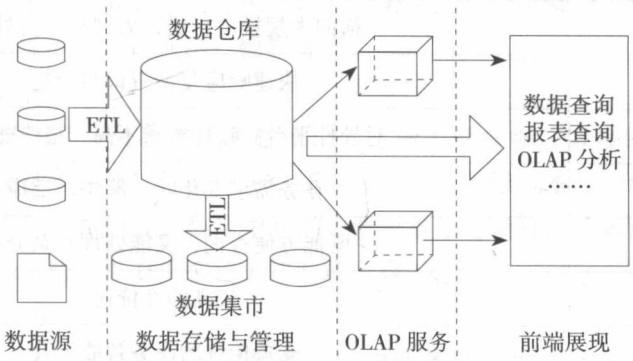

图 1-2　典型的数据仓库架构

1. 数据移动代价过高

在数据源层和分析层间引入一个存储管理层，可以提升数据质量并针对查询进行优化，但也付出了较大的数据迁移代价和执行时的连接代价。数据首先通过复杂且耗时的 ETL 过程存储到数据仓库中，在 OLAP 服务器中转化为星形模型或者雪花模型；执行分析时，又通

过连接方式将数据从数据库中取出，这些代价在 TB 级时也许可以接受，但面对大数据，其执行时间至少会增长几个数量级。更为重要的是，对于大量的即时分析，这种数据移动的计算模式是不可取的。

2. 不能快速适应变化

传统的数据仓库假设主题是较少变化的，其应对变化的方式是对数据源到前端展现的整个流程中的每个部分进行修改，然后再重新加载数据，甚至重新计算数据，导致其适应变化的周期较长。这种模式比较适合对数据质量和查询性能要求较高，而不太计较预处理代价的场合。但在大数据时代，分析处在变化的业务环境中，这种模式将难以适应新的需求。

（三）一个鸿沟

在大数据时代，巨量数据与系统的数据处理能力间将会产生一个鸿沟：一边是至少 PB 级的数据量，另一边是面向传统数据分析能力设计的数据仓库和各种 BI 工具。如果这些系统工具发展缓慢，该鸿沟将会随着数据量的持续爆炸式增长而逐步拉大。

虽然，传统数据仓库可以采用舍弃不重要数据或者建立数据集市的方式来缓解此问题，但毕竟只是权益之策，并非系统级解决方案。而且，舍弃的数据在未来可能会重新使用，以发掘出更大的价值。

二、大数据的期望特性

数据仓库系统需具备几个重要特性，如表 1-1 所示。

表 1-1　数据仓库系统需要具备的特性

特性	说明
高度可扩展性	横向大规模可扩展，大规模并行处理
高性能	快速响应复杂查询与分析
高度容错性	对硬件平台一致性要求不高，适应能力强
支持异构环境	业务需求变化时，能快速反应
较低的分析延迟	既能方便查询，又能处理复杂分析
较低成本	较高的性价比
向下兼容性	支持传统的商务智能工具

（一）高度可扩展性

一个明显的事实是，数据库不能依靠一台或少数几台机器的升级（scale-up 纵向扩展）满足数据量的爆炸式增长，而是希望能方便地做到横向可扩展（scale-out）来实现此目标。

普遍认为 shared-nothing 无共享结构（每个节点拥有私有内存和磁盘，并且通过高速网

络与其他节点互连）具备较好的扩展性。分析型操作往往涉及大规模的并行扫描、多维聚集及星形连接操作，这些操作也比较适合在无共享结构的网络环境下运行。Teradata 即采用此结构，Oracle 在其新产品 Exadata 中也采用了此结构。

（二）高性能

数据量的增长并没有降低对数据库性能的要求，反而有所提高。软件系统性能的提升可以降低企业对硬件的投入成本、节省计算资源，提高系统吞吐量。巨量数据的效率优化，并行是必由之路。1PB 数据在 50MB/S 速度下串行扫描一次，需要 230 天；而在 6000 块磁盘上，并行扫描 1PB 数据只需要一小时。

（三）高度容错性

大数据的容错性要求在查询执行过程中，一个参与节点失效时，不需要重做整个查询，而机群节点数的增加会带来节点失效概率的增加。在大规模机群环境下，节点的失效将不再是稀有事件（根据谷歌报告，平均每个 Map Reduce 数据处理任务即有 1.2 个工作节点失效）。因此在大规模机群环境下，系统不能依赖于硬件来保证容错性，要更多地考虑软件容错。

（四）支持异构环境

建设同构系统的大规模机群难度较大，原因在于计算机硬件更新较快，一次性购置大量同构的计算机是不可取的，而且也会在未来添置异构计算资源。此外，不少企业已经积累了一些闲置的计算机资源，此种情况下，对异构环境不同节点的性能是不一样的，可能出现“木桶效应”，即最慢节点的性能决定整体处理性能。因此，异构的机群需要特别关注负载均衡、任务调度等方面的设计。

（五）较低的分析延迟

分析延迟是分析前的数据准备时间。在大数据时代，分析所处的业务环境是变化的，因此也要求系统能动态地适应业务分析需求。在分析需求发生变化时，减少数据准备时间，系统能尽可能快地做出反应，快速地进行数据分析。

（六）较低的成本

在满足需求的前提下，使技术成本越低，其生命力就越强。值得指出的是，成本是一个综合指标，不仅仅是硬件或软件的代价，还应包括日常运维成本（网络费用、电费、建筑等）和管理人员成本等。据报告，数据中心的主要成本不是硬件的购置成本，而是日常运维成本，因此，在设计系统时需要更多地关注此项内容。

（七）向下兼容性

数据仓库发展的 30 年，产生了大量面向客户业务的数据处理工具（如 Informactica、Data Stage 等）、分析软件（如 SPSS、R、MATLAB 等）和前端展现工具（如水晶报表）等。这些软件是一笔宝贵的财富，已被分析人员所熟悉，是大数据时代中小规模数据分析的必要补充。因此，新的数据仓库需考虑同传统商务智能工具的兼容性。由于这些系统往往提供标准驱动程度，如 ODBC、JDBC 等，这项需求的实际要求是对 SQL 的支持。

总而言之，以较低的成本投入、高效地进行数据分析是大数据分析的基本目标。

三、研究现状

对并行数据库来讲，其最大问题在于有限的扩展能力和待改进的软件级容错能力；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做分析归纳。

（一）并行数据库扩展性和容错性研究

华盛顿大学在文献中提出了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，在失败时仅重新执行其子操作符（在某节点上运行的操作符）的任务来避免整个查询的重新执行。

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，使得数据仓库查询获得类似 MapReduce 的容错能力。

（二）MapReduce 性能优化研究

MapReduce 的性能优化研究集中于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 系统运用了传统数据库的索引技术，并通过分区数据并置（Co-Partition）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，从而缩短了任务执行时间：在线聚集（online aggregation）的操作模式使得用户可以在查询执行过程中看到部分较早返回的结果。两者的不同之处在于前者仍基于 sort-merge 方式来实现流水线，只是将排序等操作推向了 Reduce，部分情况下仍会出现流水线停顿的情况，而后者利用 Hash 方式来分布数据，能更好地实现并行流水线操作。

（三）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 1 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。该算法的缺点是应用场景有限，只适用于链式连接。第 2 种连接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，使得连接操作尽可能在数据库内执行。该算法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

此为试读，需要完整PDF请访问：www.ertongbook.com