云时代的大数据技术与应用实践

朱利华 著

YUNSHIDAI DE

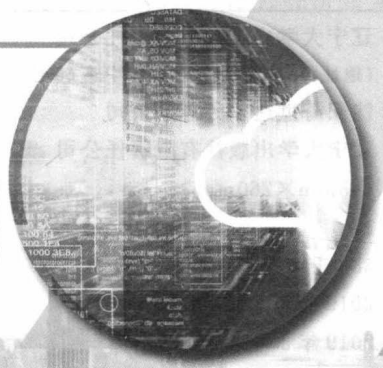
DASHUJU JISHU YU YINGYONG SHIJIAN

辽宁大学出版社
Liaoning University Press

本文受江苏高校品牌专业建设工程资助项目（PPZY2015A090）资助
英文标志：Top-notch Academic Programs Project of Jiangsu Higher Education Institutions

云时代的大数据技术与应用实践

朱利华　著



辽宁大学出版社
Liaoning University Press

图书在版编目（CIP）数据

云时代的大数据技术与应用实践/朱利华著. —沈阳：辽宁大学出版社，2018.12
ISBN 978-7-5610-9340-5

Ⅰ.①云… Ⅱ.①朱… 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 139722 号

云时代的大数据技术与应用实践
YUNSHIDAI DE DASHUJU JISHU YU YINGYONG SHIJIAN

出 版 者：辽宁大学出版社有限责任公司
（地址：沈阳市皇姑区崇山中路 66 号　邮政编码：110036）
印 刷 者：沈阳海世达印务有限公司
发 行 者：辽宁大学出版社有限责任公司
幅面尺寸：185mm×260mm
印　　张：17
字　　数：380 千字
出版时间：2019 年 3 月第 1 版
印刷时间：2019 年 3 月第 1 次印刷
责任编辑：窦重山
封面设计：徐澄玥
责任校对：齐　悦

书　　号：ISBN 978-7-5610-9340-5
定　　价：62.00 元

联系电话：024-86864613
邮购热线：024-86830665
网　　址：http://press.lnu.edu.cn
电子邮件：lnupress@vip.163.com

内容简介

云时代是指计算时代，而云计算是分布式处理、并行处理和网格计算的发展。当前，谷歌、亚马逊、阿里巴巴正在采用云计算搭建数据，大数据已发展成为一种新兴技术。尤其互联网、物联网、云计算的快速兴起，数据的爆炸式增长更加超乎人们的想象。据统计，预计到2020年，全球以电子形式存储的数据量将达到35ZB，比2016年全球存储量增长了30倍。总体来看，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，涵盖了从数据的海量存储、处理到应用等方面的技术，如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）。本书以云时代为背景，研究大数据分析、大数据挖掘、大数据算法、大数据链接分析技术、大规模文件系统 MapReduce、HDFS 海量存储数据、HBase 存储百科数据以及数据安全问题，希望为研究大数据的相关人员提供帮助，促进“大数据技术”未来的发展，为社会经济带来更大的利益。

前言

进入21世纪，云时代已经成为热点技术。亚马逊弹性计算云的商业化应用，美国电话电报公司推出的Synaptic Hosting（动态托管）服务，使云计算从节约成本的工具到盈利的推动器，从ISP（网络服务提供商）到电信企业，已经成功地从内置IT系统演变成公共的服务。

随着云时代的来临，大数据吸引了更多的人关注。大数据通常用来形容一个公司创造的大量结构化和半结构化数据，这些数据在下载到关系数据库中用于分析时，会花费过多的时间和金钱。大数据分析常和云计算联系到一起，因为实时的大型数据集分析需要像MapReduce一样的框架来向数百甚至数千台计算机分配工作。

关于“大数据”有许多种定义。多数定义都反映了那种不断增长的捕捉、聚合与处理数据的技术能力。换言之，数据可以更快获取，有着更大的广度和深度，并且包含了以前做不到的新的观测和度量类型。大数据多用来描述为更新网络搜索索引需要同时进行指量处理或分析的大量数据集。随着谷歌MapReduce和Google File System（GFS）的发布，大数据不仅用来描述大量的数据，还涵盖了处理数据的速度。

IT行业中都需要对数据进行分析，而数据分析都需要数据源。互联网公司通过搜索引擎、访问记录、App追踪等技术手段可以获得大量的用户浏览信息，但这些信息的收集、存储、提取、访问等环节不对外公开。而大数据可应用于多个领域，如医疗各种疾病数据、农业上的作物等数据；工业制造的原料、加工流程、设备信息、产品规格等数据；金融行业的客户资料、金融产品等数据；教育领域的学生、学校、教师、教材等数据；国防领域卫星、海域等数据，环境保护中的空气污染物、源质量分析等实时数据，为相关领域的数据分析提供了重要依据，有着极大的现实意义。

目录

第一章　云时代下的大数据技术研究绪论

云时代是指云计算时代，云计算（Cloud Computing）是分布式处理（Distributed Computing）、并行处理（Parallel Computing）和网格计算（Grid Computing）的发展，或者说是这些计算机科学概念的商业实现，这将是一个时代的来临。

第一节　研究背景——云时代

大数据的兴起，既是信息化发展的必然，也是云计算面临的挑战。云计算与大数据的关系是“动”与“静”的关系。一方面，大数据需要处理大数据的能力（数据获取、清洁、转换、统计等能力），其实就是强大的计算能力；另一方面，云计算的“动”也是相对而言，比如基础设施即服务中的存储设备提供的主要是数据存储能力，所以可谓是“动中有静”。

一、什么是云计算

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。

狭义的云计算是指IT基础设施的交付和使用模式，指通过网络以按需、易扩展的方式获得所需资源。广义的云计算是指服务的交付和使用模式，指通过网络以按需、易扩展的方式获得所需服务。这种服务可以是IT和软件、互联网相关，也可以是其他服务。它意味着计算能力也可作为一种商品通过互联网进行流通。

目前，“云计算”概念被大量运用到生产环境中，国内的阿里云、云谷公司的XenSystem、在国外已经非常成熟的Intel和IBM，各种“云计算”的应用服务范围正日渐扩大，影响力也无可估量。总体来说，云计算具有以下四个特征：以网络为中心、以服务为提供方式、资源的池化与透明化、高扩展与高可靠性。

二、云计算的体系结构

云计算平台可以看成一个强大的“云”网络，连接大量并发计算的网络计算和服务，可利用虚拟化技术扩展每个服务器的能力，将各自的资源通过计算平台结合起来，最终提供超级计算和存储能力。云计算体系结构如图 1–1 所示。

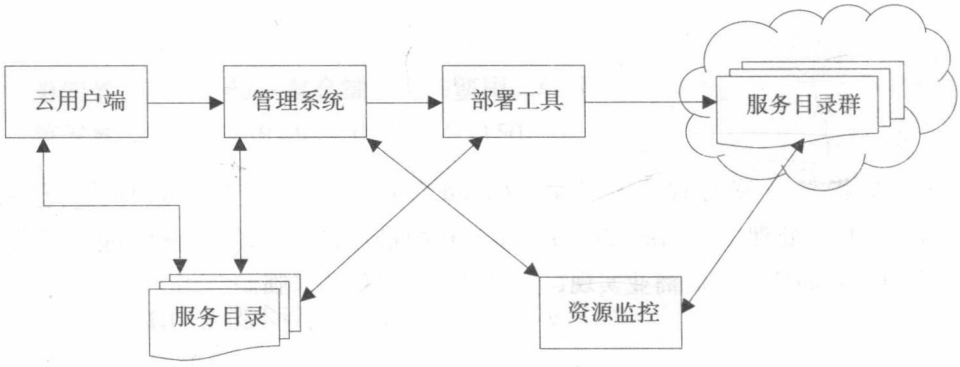


图 1–1　云计算体系结构示意图

云计算体系结构功能如下。

（1）云用户端：为用户提供云请求服务的交互界面，是用户使用云的入口。用户通过 Web 注册账户、登录、制定服务、配置管理用户，打开应用实例就像本地操作桌面系统一样。

（2）管理系统：主要为用户提供管理和服务，能对用户授权、认证、登录进行管理，还能够管理用户的可用资源和服务。

（3）部署工具：接收用户发送的请求，根据用户请求转发相应的程序，部署资源应用，配置、回收资源。

（4）服务目录群：管理系统管理的是虚拟的物理服务器，负责高并发量的用户请求处理、大运算时计算处理、用户 Web 应用服务，云数据存储时采用相应数据切割法，并行方式上传 / 下载大容量数据。

（5）服务目录：云用户通过付费取得相应的权限后可以选择定制服务列表，或者对服务的退订操作。在管理系统中，云用户还可以在界面生成图标或列表等服务。

（6）资源监控：监控云资源的使用情况，以便做出反应，完成节点同步配置工作，确保资源能够分配到每个用户。

三、云计算的关键技术

云计算将动态易扩展的被虚化的计算资源通过网络提供服务，其中的关键技术有以下几种。

（一）虚拟化技术

虚拟技术是在虚拟的环境中运行，它可以扩展硬件的容量，简化软件的重新配置过程，减少软件虚拟机相关开销，支持更广泛的操作系统。在云计算中，计算系统虚拟化是一切建立在“云”上的服务与应用基础。目前，虚拟技术主要应用于服务器、操作系统、中央处理器（CPU）等方面，提高了工作效率。通过虚拟技术可以实现软件应用与底层硬件的隔离，它可以将单个资源划分成多个虚拟资源的分裂模式，也可以将多个资源整合成一个虚拟资源的整合模式。虚拟技术根据应用对象可分为三类：存储虚拟化、计算虚拟化、网络虚拟化。

（二）弹性规模扩展技术

云计算像是一个巨大的资源池，为存储使用提供了空间。但云计算的应用使用有着不同的负载周期，并根据负载对应的资源进行动态伸缩（高负载时动态扩展资源，低负载时释放多余资源），如此一来可以充分调用或提高资源的利用率，不会出现冗余拥挤的情况。弹性规模扩展技术为不同的应用架构设定不同的集群类型，每一种集群类型都有特定的扩展方式，然后通过监控负载的动态变化，自动为应用集群增加或减少资源。

（三）分布式海量数据存储技术

云计算系统由大量服务器组成，为用户提供服务。云计算系统采用了分布式存储方式存储数据，用冗余存储方式（集群计算、数据冗余和分布式存储）保证数据的可靠性。冗余的方式通过任务分解和集群，用低配机器替代超级计算的性能来保证低成本，这种方式保证了分布式数据的高可用、高可靠和经济性。分布式存储目标是利用云环境中多台服务器的存储资源来满足单台服务器所不能满足的存储需求，使存储资源能够被抽象表示和统一管理。

（四）分布式计算技术

MapReduce 编程模型是云平台最经典的分布式计算模式。MapReduce 将大型任务分离成许多细粒度的子任务，将这些子任务分布在多个计算节点上进行调度和计算，从而在云平台上获得对海量数据的处理能力。

（五）多租户技术

多租户技术的目的在于大量用户能够共享同一堆栈的软件硬件资源，并且针对每个用户的需求分配适当的资源，实现软件服务客户化配置。这种技术的核心包括数据隔离、客户化配置、架构扩展和性能定制。

（六）海量数据管理技术

数据管理技术必须能够高效地管理大量的数据，云计算才能对海量式或分布式的数据进行处理、分析。谷歌的 BT（BigTable）数据管理技术和 Hadoop 团队所开发的开源数据管理模块 HBase 是计算系统中的数据管理技术。由于云数据存储管理不同于传统的 RDBMS 数据管理方式，云计算数据管理技术必须解决如何在规模巨大的分布式数据中找到特定的数据。

四、大数据中的云时代

（一）政府的服务云

我国早在2004年就提出了“服务型政府”的概念，要努力建设服务型政府，要把公共服务和社会管理放在更加重要的位置上，努力为人民群众提供方便、快捷、优质和高效的公共服务。某日报也曾指出政府的服务云4条路径。

（1）引入政府公共关系，运用传播的手段与社会公众建立互相了解、互相适应的持久联系。

（2）推进公共服务社会化，即把不一定要政府承担或政府无法承担的公共事务交由非政府组织来承担和处理，通过市场机制提高效率。

（3）完善电子政务建设。

（4）推进回应型政府建设，提升政府对社会呼声和突发事件的反应、驾驭和处理的能力，提升各级政府信息部门的反应能力。

（二）政府的服务云架构

政府的服务云的架构如图1-2所示。

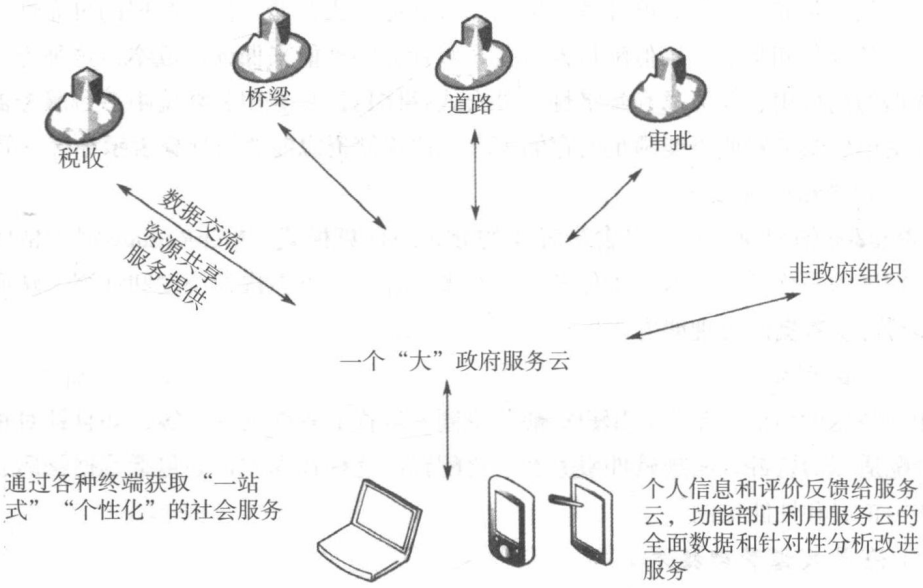


图1-2 “大”政府服务云架构

税收、桥梁、道路、审批等职能通过数据交换连接到服务云中，并且通过服务云实现各自的资源共享。非政府组织所提供的公共资源，诸如世界卫生组织、世界野生动物保护组织等也通过数据交换加入云中，与政府各职能部门实现基于角色的信息共享。云的另一端则连接着无数个体民众和法人（营利和非营利组织）。他们上网不需要单独与税务局打交道，对他们来说，对象即为一个政府，而这个政府即存在于无所不在的云中。

（三）数据平台和创新中心

（1）职能部门间的数据共享

在 Web2.0 时代，正如谷歌哺育了一代由广告养活的中小网站，苹果通过联合庞大的第三方开发者开发各种应用软件而颠覆手机行业游戏规则一样，政府的数据平台应该成为新的创新中心。数据平台一方面整合各个职能部门的资源，另一方面如果让每个人都能够接入政府数据，这将可能给企业带来新的商业机会。

（2）无缝用户体验与公共服务消费数据的共享

政府云可通过民众的访问和操作获得大量终端用户的行为信息。一个整合的政府提供给终端民众的不仅仅是一站式的服务，也是无缝的用户体验。未来政府的服务云将实现不同终端的一致性接入，民众不仅可通过计算机，还可通过手机、iPad 等获取端到端的服务。Web 已推动民众间建立起可信的人际关系网络，并日益向真实化的社交社区演进。政府也在利用这个工具加强与民众的沟通，并实时监控，进行前馈性分析，从而提高对突发事件和热点话题的反应速度。白宫将政府信息实时发送到 MySpace、Facebook、Twitter 上。英国政府甚至向机关发文，要求公务员学习使用 Twitter，各政府部门都要开 Twitter 每天发布 2 ~ 10 条信息，且间隔不得小于半小时。民众与政府间的围墙正在消失。

同时，英国政府正在推动为每个公民设立一个网页的计划，这样学生可与老师间做关于课程的讨论，医生和患者间可以成为保持沟通的朋友。如此多的网页，不可能由市民自己想办法建立，而是统一运行在一个云平台上面。相比之前围墙高耸的情况，无论政府部门之间还是政府与民众间，我们都有理由相信，当所有的政府数据、民众对公共服务的消费数据、互动数据在同一平台上汇聚时，将引发新一轮商业创新的爆发。

第二节　研究内容——大数据技术

变化是永恒的主题。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于企业、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，使越来越多的人认识到了数据的价值。

一、什么是大数据

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。大数据是指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。大数据具有数据体量巨大、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、NoSQL 数据库、实时流数据处理以及智能分析技术，如模式识别、

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）和最重要的 Value（价值）。

大数据最主要的作用是服务，即面向人、机、物的服务。例如，机器需要数据有一些关联，能够从中分析出有用的信息。人、机、物对数据的贡献和参与度非常高。从数据规模上，可看到人到物理世界是从小到大；从数据质量讲，人提供数据质量是最高的。

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，使中小企业也可以像亚马逊一样通过云计算来完成大数据分析。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，还有赖于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

大数据分析相比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介绍，并展望未来。

（一）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。

1. 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。该优化器可以通过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条件，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，使数据仓库查询获得类似 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。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。

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 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，对于复杂的连接操作（如环形连接等）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）展望研究

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

对并行数据库来说，其扩展性近年虽有较大改善（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。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。

相比前两者，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：理论界侧重于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，可认为如下几个研究方向比较重要，且目前研究还较少涉及。

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虽然难以预测，但基于报表和多维数据的分析仍占多数。因此，MapReduce 平台也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（如预测分析操作的物化）及怎样基于物化的数据进行复杂分析操作（如数据访问路径的选择问题）。

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。IBM 将开源统计分析软件 R 集成进 Hadoop 平台，增强了 Hadoop 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。尤其需要强调的是，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 及网络传输代价较高。一种有效地降低该代价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

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 可靠性研究

当前 Hadoop 采用主从结构，由此决定了主节点一旦失效，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、I/O 及网络带宽，还可以利用当前 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. 多维索引研究

基于 MapReduce 框架实现多维索引，加快多维数据的检索速度。当然，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，都是非常有挑战和意义的研究。

第二章　大数据存储技术的研究

目前，淘宝每天的活跃数据量已经超过 50TB，共有 4 亿注册用户，数亿条产品信息显现，平台每天超过 4 000 万人次访问。如此巨大的数据访问量，使淘宝数据仓库成为国内最忙碌的数据仓库之一，每天大约要处理几亿次的用户行为。从淘宝数据仓库的访问量看，大数据要实现高效智能的存储是至关重要的。本章就来研究大数据存储技术问题。

第一节　大数据存储技术的要求

存储本身就是大数据中一个很重要的组成部分，或者说存储在每一个数据中心中都是一个重要的组成部分。随着大数据的到来，对于结构化、非结构化、半结构化的数据存储也呈现出新的要求，特别对统一存储也有了新变化。对于企业来说，数据对于战略和业务连续性都非常重要。然而，大数据集容易消耗巨大的时间和成本，从而造成非结构化数据的雪崩。因此，合适的存储解决方案的重要性不能被低估。如果没有合适的存储，就不能轻松访问或部署大量数据。

如何平衡各种技术以支持战略性存储并保护企业的数据？组成高效的存储系统的因素是什么？通过将数据与合适的存储系统相匹配以及考虑何时、如何使用数据，企业机构可确保存储解决方案支持，而不是阻碍关键业务驱动因素（效率和连续性）。通过这种方式，企业可自信地引领这个包含大量、广泛信息的新时代。

一、数据存储面临的问题

数据存储主要面临三类典型的大数据问题：

OLTP（联机事务处理）系统里的数据表格子集太大，计算需要的时间长，处理能力低。

OLAP（联机分析处理）系统在处理分析数据的过程中，在子集之上用列的形式去抽取数据，时间太长，分析不出来，不能做比对分析。

典型的非结构化数据，每一个数据块都比较大，带来了存储容量、存储带宽、I/O 瓶颈