Broadview
www.broadview.com.cn

十二五国家重点图书出版规划项目 云计算实践指南丛书

纵横大数据

云计算数据基础设施

从策略、技术、应用、数据架构等多个维度
指点企业大数据规划

BIG DATA

何小朝 著

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

十二五国家重点图书出版规划项目 云计算实践指南丛书

纵横大数据

云计算数据基础设施

何小朝 著

電子工業出版社
Publishing House of Electronics Industry

内 容 简 介

大数据的概念很火，但人们对它的认识却是混乱的：有人说大数据就是指所有的数据，有人说大数据是指线上行为、日志等半结构/非结构化的数据形态，有人说大数据就是以Hadoop为代表的新技术……到底什么是大数据？同样风风火火了很久的云计算与大数据有什么关系？令人眼花缭乱的众多大数据技术的本质是什么？各有什么优缺点？争论不休的“小变大”与“大变小”技术策略到底孰正孰邪？企业究竟应该如何定位与使用大数据，难道是为了Hadoop而Hadoop？未来的技术方向究竟如何？

本书结合现代企业数据管理实践，从策略、技术、应用、企业数据架构等多个维度，体系化地对大数据及相关技术进行了全面深入的论述：首先对大数据相关概念予以澄清；接着深入剖析各种大数据技术的内在本质，指出其各自的优缺点、适用场景与相互关系；同时对大数据技术“分”与“合”这两种广受争议的技术策略的内在联系进行了分析与讨论，明确指出现代数据管理技术的发展趋势；最后结合大数据时代企业新一代数据架构规划的实际，对大数据及相关技术在企业数据体系中的具体定位给出了切实可行的建议，并且面向云数据中心建设，提出了大数据云——云计算数据基础设施的概念与方法。

图书在版编目（CIP）数据
纵横大数据：云计算数据基础设施 / 何小朝著. —北京：电子工业出版社，2014.5
（云计算实践指南丛书）
十二五国家重点图书出版规划项目
ISBN 978-7-121-23213-8

Ⅰ. ①纵… Ⅱ. ①何… Ⅲ. ①云计算 Ⅳ. ①TP393.027

中国版本图书馆 CIP 数据核字（2014）第 099379 号

策划编辑：张月萍
责任编辑：周宏敏
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×1092　1/16　　印张：16.5　　字数：395 千字
版　　次：2014 年 5 月第 1 版
印　　次：2014 年 5 月第 1 次印刷
定　　价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。
质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei. com.cn。
服务热线：（010）88258888。

推荐序一

花五个小时，你获得什么？

这两年市面上大数据的书也很多了，大体分为两类，一类是讲趋势，重点是大数据对商业、社会、政府的影响；另一类是讲工具，比如，专门讲 Hadoop 指南，但对于企业 CIO，CTO 和企业架构师们来说，缺少一本承上启下的“中坚论著”，尤其是对于数据管理类各种技术流派做出一致性评价的书籍，可能是因为从业人员里面能够拥有如此宽的架构视野，具备具体技术实践，同时具备思辨能力的人本来就很少，何博士的经历刚好满足这几个要求。

本书有这么几个特质，导致了它的可读性和实用价值。

解惑：具备清晰的分析逻辑和对多种技术的内在理解

从关系型数据库到 MapReduce，从 NoSQL 到分布式文件系统，笔者讨论了今天 IT 环境里的多种技术，并围绕着 CAP，BASE，Codd 原则等基本原则对多个技术流派的特点进行了精辟的分析，使得多种技术体系的利弊在一个整体分析框架里来讨论，并结合具体应用案例，笔者对相关技术（包括硬件技术潮流），对主要技术演进的“所以然”给出了清晰的分析与解释。能够把纷繁复杂的多种技术讲得环环相扣，体现了笔者的技术功底和思辨精神，尤其可贵的是，笔者对主要技术优缺点的分析是相对中立和客观的。何种情况采用何种技术策略，读者如果“吃透”了本书的见解，会有一个基本的判断方向。

书中对有些问题的讲解，甚至有侦探小说的味道，这个是非常难得的，带着问题来读会使读者收获更大。同时本书敢于表达自己的“定义”能力，对很多技术的本质阐述用词精准。

授业：具备实践性的架构经验，有实战指导价值

笔者本人是直接动手做过一个已在应用中的大数据系统的，在本书的后半段也直接分享了

这个实践，这使得整个书中的见解和建议都具备实战价值，这点对于今天“雷声大，雨点小”的中国大数据市场是非常可贵了。尽管做过 Hadoop 工程师、关系型数据库架构师的人不少，但能够根据一个特定的企业架构需求，给出恰当的技术建议，包括一套整体思考方法的人，却非常稀有，何博士此书有些实践性的“干货”也是本书的特色之一。

传道：帮助从业人员提升对数据管理的整体视野

对于更多的 IT 业者，希望对云计算与大数据的体系有一个整体的理解，尤其能够历史地理解数据管理的前世今生，包括互联网大潮下的架构变迁，这本书尤为可贵。此书对于技术的理解，既包括了技术的深入理解力，又超过了技术本身，能够站在取舍之道的角度来看技术趋势，所以这本书，等于对云计算的数据基础设施与架构选择等方面，给出了一个方法论与原则方面的参考。对于从事数据中心规划、大数据架构设计、云计算数据基础设施的从业人员都有很大帮助。

一本非常有可读性的书！

甲骨文大中华区技术战略部总经理

刘松

推荐序二

大数据，从麦肯锡完整给出定义到现在还不到三年的时间。也就是说，这个概念本身依然很新，时间不长。但是，在当今互联网时代，三年已经是一个不短的时间，尤其对于知识传播来讲。这是一个数据爆炸的时代，是一个信息飞速传递的时代，是一个随时创造奇迹的时代。只有到了今天我才真正有了科技水平日新月异的感觉，新科技、新应用层出不穷。一个“余额宝”几个月就可以到几千万，超越基金公司老大十多年积累的规模；微信一个发红包的简单应用，短短几天时间就有几千万人乐在其中；一个求关爱买保险方式，几天就有几十万人参与。

仅仅几年的时间，大数据的理念已经深入人心。这个概念已经突破了行业界限，突破了地域界限，突破了国家界限，成为全世界都在关心的理念。对于大数据技术和应用的研究已经成为很多国家的科技战略，甚至已经成为国家战略。从各国的投入和重视程度，我们不难看到它的重要性和未来前景的广阔。不论是美国总统奥巴马计划投入几亿美元、英国准备用近 2 亿英镑投入大数据的研究，还是新加坡的大数据人才发展战略都充分凸显了各国把大数据作为研究的重点方向和战略高地。在我们中国，这个概念的普及和重视程度丝毫不亚于任何一个国家。从国务院、工信部到各个地方政府，我们都能够听到和看到领导们对大数据的重视程度。短短几个月的时间，各地政府的想法和做法相继出台。 山西太原首先宣布建立中国金融业大数据中心，贵州公布了雄心勃勃的大数据发展战略与规划（800 亿的投入和 2020 年 4500 亿的产值），广东省准备设立专门的大数据部门，中关村建立大数据交易中心，可以说是政策和策略层出不穷。各个企业更是不甘落后，不论是互联网企业，还是技术应用前沿的金融和电信，以及众多的 IT 服务公司，大数据无疑都已经成为大家关注的焦点。

对于任何一个事物，不论热到什么程度，在我们参与和推广之前，我们一定要理解其本质是什么，它能够干什么，我们可以如何用，应用这些技术和理念能够给我们的生活和工作带来什么好处。要回答这些问题，我们就需要了解什么是大数据，大数据如何定义才比较全面，大数据有哪些特点和特征，大数据在各个行业的潜在应用有哪些，大数据的关键技术有哪些，推广大数据需要什么样的人才，凡此等等。要用好大数据，有太多的问题需要我们回答，有太多的知识需要我们去学习。 何博士的这本书可以很好地帮助我们去理解和回答这些问题，尤其是他多年的实战经验，可以弥补目前这一领域实战应用的相对匮乏。

我非常同意本书对大数据的定义：大数据就是指**企业以“数据驱动业务与运营”的相关战略与战术**。也就是说，在大数据时代，企业在进行一门决策、开展一项活动、设计一个产品时，需要养成一种习惯（甚至可以说是一种制度或规范）：让数据说话！以数据分析的结果来指导

这些决策与设计活动。这样看来，很显然，大数据就不是单指某一种数据类型，也不是单指某一种技术了。

那么，要实现“数据驱动业务与运营”，过去企业所拥有的以交易行为为核心的数据显然是远远不够的。例如，根据一个银行以前所拥有的交易数据，无法知道这个人的爱好、行为习惯、社会关系等信息，也就无法全面地了解该客户，那么对该客户做出的营销，就无法达到满足客户要求的个性化，自然也就很难说是精准营销了。我们必须将新的数据源补充到现有的数据体系中来，而这些数据正好就是那些被大多数人称作是大数据的社交媒体、线上行为数据等。也就是说，在这样的定义下，新数据源成为了企业实现大数据策略所需要补充的数据，是大数据体系中必不可少的重要成员。传统的结构化数据、内部数据和众多的非结构化数据、外包数据的结合，才可以让我们更加准确地理解我们的客户和服务对象。

再看看在该定义之下的 Hadoop 技术。同样，企业要实现“数据驱动业务与运营”，过去所掌握的关系数据库技术也已经不够了，同样必须引入新的技术手段，而以 Hadoop 为代表的新技术手段也成为了企业大数据体系中需要补充的成员。

在我看来，至今为止，本书这样的定义，既有相当的高度与长期的可适用性，并且还合理地囊括与兼容了我们过去对大数据的普遍理解，是目前见到的最为恰当的解释了。

再有就是纯技术层面的问题了：Hadoop、NoSQL 等技术在企业内到底应该如何使用？这确实是令很多企业头痛的实际问题。我很赞同本书作者的看法：**没有一种观点是完全可以拿来照办的！要做出正确的决策，就必须先对各种技术的本质特点有一个全面正确的了解，然后结合企业自身的实际，做出自己的判断**。于是，该书将很大一部分内容都放在了对各种技术手段的深入分析上了，并且还给出了各种技术在企业数据管理实际中具体的定位参考的相关实例，同时还对实现大数据的技术策略，以及未来数据管理技术的发展趋势等进行了分析与判断，内容十分翔实丰富。而到目前为止，在涉及大数据话题的资料与书籍中，能像本书这样全面分析与介绍大数据技术的还非常少！我认为，对企业与技术人员来讲，本书的这些内容远远比介绍 Hadoop 到底如何使用要重要、要有意义得多！

另外，从本书的内容中，我们还了解到一个非常重要的前沿趋势：即使是纯从技术上看大数据，目前以互联网数据源及 Hadoop 技术为主导的固有思路已经需要调整与提高了。在不久的将来，除了数据库技术之外，大数据技术的重点可能会**逐渐以更加实时高效的、面向海量数据对象或海量计算任务的大规模并行处理技术为主，而 Hadoop 应该只是其中的一员**而已。

基于上述原因，我认为该书是目前大数据领域内不可多得的一本好书，无论是对企业来讲，还是对技术人员来讲，都有相当的参考意义，我乐于将该书推荐给各位读者。

刘世平
中科院大学教授，博导
金融科技研究中心主任
吉贝克信息技术（北京）有限公司董事长

前言

“云计算”与“大数据”应该说是目前 IT 界最为热门的两个概念了。云计算以各种软硬件资源新的消费与交付模式为核心理念，被普遍认为将会成为未来社会最为深远的革新。而现实却是：在多“云”的天空，成功的实践却少得可怜，致使其很多情况下只是充当了一个时髦的噱头。

令人遗憾的是，如今风头已远远盖过“云计算”的“大数据”，其现实情况与此类似。大数据概念最初是伴随着 Hadoop 等开源技术的推广而出现的，在国内外众多互联网公司依靠它们取得巨大成功的强力推动下，传统数据管理技术的地位受到了严重的挑战，似乎不知 Hadoop、不用 Hadoop 就会落后！但如何才能在本企业或者某个具体需求中正确有效地使用这些新技术呢？这至今依然是众多企业技术决策者的困惑。

大多数企业目前对大数据潮流的热烈响应其实是“雷声大，雨点小”，其中相当一部分是不分青红皂白，纯粹为了 Hadoop 而 Hadoop，很少有产生实际成效的案例。本书认为，要正确回答这些问题，给出合适的决策，必须对这些技术本身进行较为深入的了解与分析，然后结合自己企业的实际，做出自己的判断。**任何其他企业的经验都不可以照搬照抄；任何资料中关于各种技术的适用场景描述，即使是正确的，也都有其特殊的上下文环境，不可以当成普遍真理去盲目遵从**。这里所说的对技术的了解，并不是指具体如何去使用它，而是指其内在本质、特点与相互联系，这些往往比使用方法更重要，也是本书区别于其他大数据资料的主要特点之一。

首先，让我们看看云计算与大数据的关系，目前人们对此的理解更是混乱不堪，有人认为两者完全不同，有人则认为大数据技术其实就是云计算。对“云”，最开始，人们普遍认为那是一种采用一堆闲散资源完成一件重大任务的技术。后来，人们又意识到现代社会对“云”的诠释，其实更多的是指一种以服务为主的商业模式，而不是一种技术。现在，绝大多数人对“云”的理解停留于此，认为“云计算”与技术无关的人大都是这种思路。但在对“云”业务模式的实践中却发现，要搞“云”服务，必须从技术手段与商业模式两个维度同时入手才有意义，只拥有其中任何一个方面都是不行的，甚至可以说前者要比后者重要得多。大多数情况下，在“云”能适用的领域内，如果没有前者，后者所能提供的服务水平自然也就很有限，从而也就自然失

去了“云”的含义。所以说，云计算的本质是商业模式，但其核心却仍然是技术问题。

而云在技术层面的核心问题又是什么呢？有人认为是“小变大”的分布式计算，有人认为是“大变小”的虚拟化，而本书认为，云计算最核心的问题是数据，具体地讲，是现代业务环境下的数据管理问题，也就是能实现海量、多类型、高负载、高性能、低成本需求的数据管理技术，这实际上就是传统数据管理技术在现代的最大挑战。这其中最耀眼的，就是各种新兴的大数据家族成员的出现，包括开源体系的 Hadoop、各种 NoSQL 数据库、NewSQL 数据库（关系数据库联邦）、分布式文件系统等，甚至还包括非开源体系的新一代关系数据库。这样看来，“大数据”应该是“云计算”业务模式得以实现在数据管理层面的核心技术支撑，两者密不可分。

而从纯技术的角度看，“云计算”概念最初出现时就是指采用网络互联起来的设备共同完成一项庞大任务的技术策略，而 Hadoop 等流行大数据技术的核心思路大多如此。因此，我们又可以说：“云计算”是大数据的技术实现方法。这便是云计算与大数据的联系，两者无论是在业务上，还是技术上，都是相互依存的。一句话，无论叫什么名称，其实都是代表现代 IT 发展的最新进展而已。

再来看看各种流行的大数据技术本身，包括 Hadoop，NoSQL，NewSQL，甚至一些新一代的关系数据库等。对它们，在现代数据管理领域内，目前的状态却是：人们普遍困惑的并不是能不能掌握这些技术的具体用法，而是到底什么时候，在什么场景下，如何定位与使用这些技术？这主要表现在以下几个方面。

一是如何定位新旧技术。即指新兴的以 Hadoop 为代表的开源技术，与传统的关系数据库技术，到底是新技术彻底颠覆传统技术，还是两者共存？如果是共存，如何共存？这是目前各个企业普遍感到困惑的最重要的问题。

二是部分技术人员对新事物只是盲从。大家在应用实践中或多或少地会遇到一些困难，于是很多技术人员就会把希望寄托在新出现的技术上，认为只要一用上如 Hadoop 或 NoSQL 这些新东西，目前的问题就会迎刃而解。接下来就立即紧张地投入到新技术的学习与使用上去，而不做是否适合自己需求的合理判断。很显然，这种对新技术的崇拜是盲目的。

三是各种技术之间出现了互相攻击、互相否定的态势。一度以来，传统的主流关系数据库（如 Oracle，DB2 等）在实践中出现了一些问题，主要是对高负荷环境下的海量数据应用出现了力不从心的现象，同时，其水平扩展性的限制与高昂的成本问题使客户越来越难以忍受。于是，一些非关系型的 NoSQL 数据库，或者一些低端数据库集群方案（如 MySQL 集群）就在一些场合替代了主流的商业数据库，并且表现出很优秀的性价比；另外，有些企业在分析领域也出现了以 Hadoop MapReduce 等开源产品全面替代关系型数据仓库的现象。于是，便出现了一种思潮，认为关系数据库最终将退出历史舞台。而另有一部分人则认为，所谓极其成功的新技术，只是昙花一现的暂时现象而已，传统的关系数据库经过改良以后，依然会是数据管理领域的王者，

其他的技术会像 30 多年前关系数据库与其他数据管理技术之争的结果一样，逐渐消失。这些观点中，大多都是凭直觉、凭感觉、凭个人经验的判断得出，虽然不能说是武断，但如果没有令人信服的技术分析做支撑，就很难说谁对谁错。

四是新技术本身在实践中也出现了很多的问题。例如 Hadoop MapReduce，虽然已经出现了 Hadoop 2.0 中的各项重要改进，但相信只要是真正用过它的人都知道，其在方便性、可靠性、可用性、效率等方面都还很不尽如人意。笔者记得一位很熟悉 Hadoop 的朋友说："如果企业能用关系数据库解决问题，就尽量不要用它！"再如 Twitter 放弃了用 Cassandra 替代 MySQL 的决策，Digg 使用 Cassandra 后出现的一系列严重问题等，都使很多人开始重新审视这些新技术。

其实，究其根本，以上现象出现的主要原因是：人们只是去学习如何使用这些新技术，却很少独立思考，对它们进行较为深入的学习与剖析；很少在设计思想、技术架构、内在本质等方面将它们与其他技术进行对比，以能在真正掌握后，做出属于自己、适合自己的判断。而这些又正是本书的主体内容。

如果在数据库技术领域继续探究，会发现 NoSQL 技术虽然适合海量数据的快速存取，却无法满足较复杂的关系模型数据管理及人们对习惯使用 SQL 语言的要求，而标准的关系数据库在水平扩展性上又严重受限。那么，是否存在一种技术，既可以使用关系模型存储数据，使用 SQL 操作数据，又可以像 NoSQL 一样方便扩展？于是，本书还与读者分享了笔者自主研发的一个关系型云数据库的设计与实践，它既不同于目前流行的 Hadoop/NoSQL 等开源技术，也不同于传统的关系数据库，是一种介于两者之间的技术模式，目前的状态正好满足 Hadoop 与传统关系数据库都不太适用的企业级海量历史数据管理的需求，并已经在实践中取得一定的成果。接着，由该自主产品的设计实践活动出发，我们产生了对 Hadoop 本身许多固有技术问题更大胆的、更进一步的深入思考：PB 级海量数据的批量分析能不能比 Hadoop 再提高一个数量级，例如，达到秒级？在保守的认识中，这样的要求似乎是不合理的，也是不可能实现的。然而大数据领域最新的技术进展——Hadoop 的缔造者 Google 近年来一系列更前沿的、被称为"Google 新三驾马车"的研究成果，通过模式（Schema）的回归与精巧的设计，已经向这样似乎是"不可能的任务"的宏伟目标迈出了一大步。这使我们意识到：技术的发展瞬息万变，Hadoop 本身已不见得有多么先进了，想要在实践中做出正确的决策，就必须不断学习，勇于创新，不断经历破与立的过程，而不能故步自封，原地不动。

除了需要对各种大数据技术手段进行深入剖析以外，当今 IT 界还在云计算技术两个不同的技术策略上有着广泛的争议，即"分"为云与"合"为云，前者是指数据切分后以小变大，后者是指以大变小，将分散的小资源集中整合起来管理后，再将资源进行统一的按需调度与分配。两者都称自己是云计算技术（或者说是大数据技术）的正宗，相互攻击与否定的现象极为激烈，并且各自都有坚实的成功实践为基础。表现最明显的就是以淘宝为代表的新兴互联网技术力量与 IBM、Oracle 等老牌的数据库厂商之间关于以"分"为主的开源技术及以"合"为主的一体

机技术之间的争论与竞争，可以说已经到了白热化的阶段。他们各说各话，各有千秋，已经成为企业技术决策者的主要困惑之一。而实际上，经过研究与分析，很容易就可以发现，他们所争论的“分”与“合”，看起来是完全相反的，实际上并不矛盾，其实是你中有我，我中有你，两者是有机结合的统一体，在现代数据管理的需求中都有各自的定位。企业所要做的并不是对技术策略进行非你即他的选择，而是根据自己的实际情况与需求，对各种技术与产品进行合理的定位；同时，更加重要的工作并不是某一项技术的正确定位与使用，而是能站在云数据中心建设的高度，将传统关系数据库资源与 Hadoop 集群资源集中起来形成 PaaS 平台，再对外提供分散的、数据相关的云服务，包括数据库云与 Hadoop 平台云，可以将之统称为大数据云。将大数据的话题提高到这样的层面，虽然相关的资源池调度与分配技术也非常重要，但更重要的却已经是面向云计算的大数据服务模式了。

另一方面，虽然关系数据库将与 Hadoop 等技术共存的思想被大多数人接受，也是本书所认同的观点，但广大读者可能还注意到一个现象：新兴的 Hadoop/NoSQL 等非 SQL 技术在不断发展的过程中，已经在逐步引进一些原本属于 SQL 技术体系的功能，如索引与事务；而关系数据库领域，也在逐步将这些新兴的技术引入其技术体系，如 AsterData 与最新 Oracle 12C 所具备的 InDB MapReduce 功能，都是除原有的 SQL 引擎以外，在其数据库内引入 MapReduce 处理引擎。那么，未来数据管理技术的发展趋势究竟如何呢？我们说，在物理基础设施上，分布式集群架构应该是未来发展的大趋势，而在软基础设施层面，虽然 SQL 与非 SQL 技术体系在相当长的时间内会共存，但未来的趋势是相互融合的。现在看来，起码**对数据管理技术来讲，开源是大趋势，摒弃产品销售为主导的商业模式，以技术服务为主体应该是各大厂商应该尽早考虑的策略**。

在本书最后，笔者结合企业数据架构规划的实际，针对当今各个企业在响应大数据潮流时最为关心、最为困惑的问题：“到底如何在本企业实施与推广大数据”给出了切实可行的建议。可以看到，企业引入大数据的本质就是：以适合更多更广的数据源，以及提供更强大的数据管理处理能力为目标，面向新时代的业务规划（如互联网金融），对现有数据体系的各个层面（包括采集、传输、加工、集成、分析、展现等）进行全面改造，推出大数据时代的新一代企业级数据架构，并将其作为现代企业 IT 架构的重要组成部分之一。笔者认为，**企业引入云计算与大数据的战略思想应该是：“业务上是改造，技术上是改进；业务上是创新，技术上是补充”，**仅供企业参考。

最后借此机会向王建波、李鹏、葛荪葳等朋友表示感谢，与他们的讨论使我受益匪浅，也一并感谢所有对我的写作有过帮助的人。希望本书是一个成功的尝试，同时也希望能为广大读者与企业的相关设计、规划与实践活动提供有用的借鉴与帮助。

何小朝

2014 年 2 月

目录

第1部分　大数据概论

第1章　大数据与云计算 2

1.1　云计算概论 3
1.2　大数据概论 4
1.2.1　现代数据管理需求分析 4
1.2.2　大数据的引入 9
1.2.3　大数据的定义与特征 10
1.2.4　大数据与互联网 12
1.2.5　大数据战略、大数据与大数据技术 14
1.3　大数据的技术实现——云计算 15
1.4　本章小结 16

第2章　关系数据库的挑战与应对 17

2.1　关系数据库技术的核心特征 18
2.2　主流关系数据库的挑战 22
2.2.1　经典 DBMS 的挑战 22
2.2.2　Shared Disk 23
2.2.3　Shared Nothing 24

2.3 改进型关系数据库 26
2.3.1 技术改进 26
2.3.2 主要产品代表 30
2.4 本章小结 40

第 3 章 非 SQL 技术简介 41

3.1 大数据技术家族 42
3.1.1 NoSQL 42
3.1.2 关系数据库联邦 NewSQL 42
3.1.3 分布式海量文件管理 43
3.1.4 Map Reduce 43
3.2 分与合——云计算的两种技术路线 44
3.3 本章小结 44

第 2 部分 “分”为云——数据切分

第 4 章 NoSQL 46

4.1 NoSQL 的引入 47
4.1.1 概念诠释与特征分析 47
4.1.2 NoSQL 的本质 50
4.2 NoSQL 家族 52
4.2.1 NoSQL 产品目录与分类 52
4.2.2 Hadoop 之 HBase 54
4.2.3 Facebook 之 Cassandra 58
4.2.4 MongoDB 与 CouchDB 61
4.2.5 Oracle NoSQL DB 63
4.2.6 Memcached 与 Redis 65

4.2.7 图数据库 Neo4J 65
4.2.8 其他 NoSQL 数据库 67
4.2.9 问题与疑惑 67
4.3 NoSQL 技术探研 68
4.3.1 NoSQL 理论基础 68
4.3.2 NoSQL 技术手段 75
4.3.3 NoSQL 技术解析 83
4.4 NoSQL 与关系数据库 88
4.5 本章小结 89

第 5 章 NewSQL——关系数据库联邦 90

5.1 数据库联邦的引入 91
5.1.1 企业业务数据管理面临的问题 91
5.1.2 垂直分库 92
5.1.3 水平分表 93
5.1.4 读写分离 95
5.1.5 联邦的引入 97
5.2 “联邦”的设计与实践 99
5.2.1 企业级“联邦”架构设计 99
5.2.2 公共基础服务设计 103
5.2.3 联邦的元数据库 106
5.2.4 联邦的应用实践 107
5.3 “联邦”技术分析 108
5.3.1 关于“垂直分库” 108
5.3.2 如何“水平分表” 110
5.3.3 关于“读写分离” 112
5.3.4 基本方法——分布与聚合 114
5.3.5 关于分布式事务 116

5.3.6 关联操作 117
5.3.7 冗余策略 119
5.3.8 异步解耦策略 120
5.3.9 使用缓存 122
5.3.10 其他问题 123
5.4 数据库联邦、NoSQL 与主流关系数据库 124
5.4.1 技术与应用——八仙过海，各显神通 124
5.4.2 互联网的神话 126
5.5 本章小结 128

第 6 章 文件系统联邦 129

6.1 问题的引入 130
6.1.1 关于几个数据概念的澄清 130
6.1.2 文件数据管理的困难 131
6.1.3 文件系统联邦的引入 133
6.2 典型开源技术介绍 135
6.2.1 MogileFS 135
6.2.2 FastDFS 136
6.2.3 MogileFS 与 FastDFS 的对比 138
6.3 技术分析 139
6.4 本章小结 140

第 7 章 平民化的分布计算——MapReduce 141

7.1 分布式计算概述 142
7.1.1 几个概念的澄清 142
7.1.2 分布式计算技术综述 143
7.1.3 MapReduce 的引入 147

7.2 MapReduce 技术介绍 148
7.2.1 设计思想 148
7.2.2 MapReduce 框架介绍 152
7.3 MapReduce 技术分析 160
7.3.1 关于效率 160
7.3.2 关于扩展性 162
7.3.3 关于可靠性与可用性 163
7.3.4 关于 MapReduce 与关系数据库 164
7.3.5 关于适用的数据类型 167
7.3.6 关于数据存储与管理 168
7.4 MapReduce 的应用实践 169
7.5 本章小结 170

第 8 章 后 Hadoop 时代 171

8.1 Hadoop 体系及其困惑 172
8.2 Google 的新三驾马车 173
8.2.1 新一代搜索引擎 Caffeine 173
8.2.2 大规模图处理系统 Pregel 174
8.2.3 Dremel——秒级实现 PB 级数据分析 175
8.3 Symphony MapReduce 181
8.4 后 Hadoop 时代即将来临 181
8.5 本章小结 183

第 9 章 InfiniData——一种关系型云数据库的设计与实践 184

9.1 现代企业数据管理需求再分析 185
9.1.1 新的企业数据需求——海量关系数据管理 185
9.1.2 技术分析 187

9.2 关系型云数据库架构设计 188
9.2.1 关系型云数据库的引入 188
9.2.2 技术架构设计 189
9.3 云存储层 192
9.3.1 逻辑架构 193
9.3.2 物理架构 194
9.3.3 关系模型云存储元 196
9.4 云计算层 198
9.4.1 MapReduce 云计算引擎 198
9.4.2 集群式云计算引擎 200
9.4.3 两种引擎的比较 201
9.5 云存储索引层 202
9.5.1 存储索引的管理 202
9.5.2 索引云运行时动态创建 203
9.6 技术分析 203
9.7 本章小结 205

第 3 部分 云计算的分与合

第 10 章 合为“云”——数据整合 208

10.1 数据整合的需求分析 209
10.2 存储整合云 210
10.3 数据库整合云 211
10.4 本章小结 213

第 11 章 关于分与合的讨论 214

11.1 困惑——分与合，孰是孰非？ 215