大数据科学与应用丛书

云端数据治理

刘小茵 李尧 程广明 等编著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

大数据科学与应用丛书

云端数据治理

刘小茵　李　尧　程广明　高智伟
谢灵群　钟世敏　张寒坤　朱楠楠　编　著

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

本书的编写团队结合最新的科研成果以及多年来在成熟度评估方面的实战经验，在借鉴国际先进数据治理理论和方法的基础上，针对云端数据特点，开发了云端数据治理模型，构成了本书的主要内容。本书旨在帮助读者和有数据治理需求的组织了解云计算环境下的数据治理方法，为云端数据治理体系建立和云端数据治理实施提供指导。

图书在版编目（CIP）数据

云端数据治理/刘小茵等编著. —北京：电子工业出版社，2017.6
（大数据科学与应用丛书）
ISBN 978-7-121-30374-6

Ⅰ. ①云… Ⅱ. ①刘… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2016）第 276348 号

策划编辑：牛平月
责任编辑：王敬栋
文字编辑：牛平月
印　　刷：北京季蜂印刷有限公司
装　　订：北京季蜂印刷有限公司
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：720×1 000　1/16　印张：21.75　字数：520 千字
版　　次：2017 年 6 月第 1 版
印　　次：2017 年 6 月第 1 次印刷
定　　价：68.00 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）88254888，88258888。

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。

本书咨询联系方式：（010）88254454，niupy@phei.com.cn。

信息技术与经济社会的交汇融合引发了数据的爆发式增长，数据蕴含着重要的价值，已成为国家基础性战略资源。数据正日益对全球生产、流通、分配、消费活动以及经济运行机制、社会生活方式和国家治理能力产生重要影响。

数据治理对于确保数据的准确、适度分享和保护是至关重要的，也是发挥数据在各种决策中支持作用的必由途径。近年来，数据治理理论得到广泛关注，在一些行业也得到了广泛应用。然而随着云计算、大数据等技术的发展与应用，数据本身的特点也发生了诸多变化，如数据来源的多样性、数据的远程存储与传输、数据的合规要求更具复杂性、数据面临的安全威胁更多等，这些变化让传统数据治理难以全面深入地应对。因此针对云计算环境下数据的特点和治理需求，研究开发一套针对云端数据治理的理论体系和实践指导是很有必要的。

编写团队结合最新的科研成果以及所在机构多年来在成熟度评估方面的实战经验，在借鉴国际上先进数据治理理论和方法的基础之上，针对云端数据特点，开发了云端数据治理模型，构成了本书的主要内容，旨在帮助读者和有数据治理需求的组织了解云计算环境下的数据治理方法，为建立云端数据治理体系和云端数据治理实施提供指导。

本书共 11 章，分为 3 篇：第 1 篇云端数据治理成熟度概述；第 2 篇云端数据治理成熟度模型解读；第 3 篇云端数据治理体系与实施。围绕“概念—内容—实施”循序渐进的总体思路，本书有望帮助读者了解数据治理的概念，建立对云端数据治理的初步认识；帮助读者深入学习云端数据治理模型及其内涵，掌握开展云端数据治理活动所需的理论知识；指导读者构建云端数据治理体系，开展具体的云端数据治理活动。

第 1 篇 云端数据治理成熟度概述（第 1～3 章），通过对比云计算环境和传统 IT 环境下数据的特点，结合数据治理发展趋势，有针对性地提出了云端数据治理成熟度模型，帮助读者建立云端数据治理的初步认识。

第 1 章 数据治理概述，介绍了数据治理的相关概念，讨论了数据治理的定义，并简要回顾了数据治理理论的研究历程以及数据治理面临的挑战。

第 2 章 大数据与云端数据治理，介绍了云计算和大数据的特点以及他们对数据本身带来的影响，针对云计算环境下数据的特点，提出了云端数据治理的定义并解释了其内涵，对比分析了云端数据治理与数据治理的关系，论述了开展云端数据治理的作用和价值。

第 3 章 云端数据治理模型与治理成熟度评估模型，从治理内容、治理原则、治理实施 3 方面解读了云端数据治理模型，帮助读者建立对云端数据治理的总体认识；定义了云端数据治理能力成熟度等级并描述了相应等级的特征，将云端数据治理能力成熟度分为 5 个等级，用于衡量组织的云端数据治理水平；介绍了云端数据治理成熟度评估模型与评估方法。

第 2 篇 云端数据治理成熟度模型解读（第 4～9 章），从云端数据战略、云端数据管理、云端数据质量、云端数据操作、云端数据架构、安全与隐私 6 个职能域分别介绍云端数据治理模型的具体内容，帮助读者掌握云端数据实施的基本知识和开展成熟度评估的依据。

第 4 章 云端数据治理战略，从战略、组织与角色、业务案例、资源保障和沟通 5 个过程域提出了开展云端数据治理战略管理的要求，帮助组织制订一个能够全局指导云端数据治理活动开展的顶层规划。

第 5 章 云端数据管理，从业务词汇表、元数据管理和主数据管理 3 个过程域提出了对云端数据进行管理的要求，为云端数据管理应该达到的水平指明了方向。

第 6 章 云端数据质量，从数据质量战略、数据概要分析、数据质量评估和数据清洗 4 个过程域提出了对云端数据质量进行管理的要求，能够指导组织如何提高数据质量。

第 7 章 云端数据操作，从数据提供者管理、数据集成与互操作和数据生命周期管理 3 个过程域提出了对云端数据操作进行管理的要求，帮助组织建立一个完整的数据生命周期管理体系，提高数据全流程的管理水平。

第 8 章 云端数据架构，从架构方法、架构标准、数据管理平台和历史数据管理 4 个过程域对云端数据架构建设提出了要求。

第 9 章 安全与隐私，从风险管理、数据安全、隐私保护和合规管理 4 个过程域对数据安全与隐私保护提出了要求，帮助组织提高数据安全保护能力，降低数据安全风险，指导组织如何在符合法律法规等要求的前提下开展数据治理。

第 3 篇 云端数据治理体系与实施（第 10～11 章），围绕云端数据治理体系构建和云端数据治理实施论述，指导组织如何具体开展云端数据治理活动。

第 10 章 云端数据治理体系，介绍了体系的必要性和体系的基本框架，并从宏观上介绍了云端数据治理体系的实施过程。

第 11 章 云端数据治理体系实施，系统全面地介绍了云端数据治理的实施过程，能为读者提供一个清晰的云端数据治理实施指南。

本书各章执笔分别是：第 1 章由刘小茵、程广明、朱楠楠编写，第 2、3 章由刘小茵、程广明编写，第 4 章由程广明编写，第 5 章由李尧编写，第 6 章由高智伟、李尧编写，第 7 章由张寒坤、李尧编写，第 8、9 章由钟世敏、程广明编写，第 10、11 章由刘小茵、谢灵群、程广明编写。刘小茵、程广明负责全书的组织、策划、汇总和校审工作，其他执笔人分别负责了相关章节的审阅工作。

本书得到了国家科技支撑计划项目、广东省科技计划项目和广州市科技计划项目产学研协同创新重大专项资金等的支持，在此表示感谢。本书在撰写过程中，得到了来自学术界和产业界众多专家的帮助，感谢他们给出了很多非常有价值的意见和建议。

本书的宗旨是为读者提供最新的云端数据治理的参考，但由于云端数据治理研究刚刚起步，本书提出的云端数据治理模型难免会有一些偏颇和不当之处。由于编写时间仓促，加上编写水平有限，书中难免会有错误和不足之处，请读者不吝赐教，提出意见和建议，以便我们不断进步。

编著者

第 1 篇　云端数据治理成熟度概述

第 1 篇

云端数据治理成熟度概述

第 1 章　数据治理概述

从 IT（信息技术）时代到 DT（数据技术）时代，都离不开一个关键词——数据。数据已成为 21 世纪最重要的战略资源之一。数据成为可以变现交易的资产，但又不同于传统的财务资产，数据的可拷贝、可重用以及数据的收集、存储、使用都有其特殊性，数据还涉及个人隐私、运行的安全。数据治理是以数据为研究对象的一门学科，它涉及对组织内的人员、流程、技术和策略的分配，让组织能够从数据中获取更优的价值。

本章首先介绍数据治理相关的概念，以及这些概念之间的关系；其次，介绍数据治理理论的研究历程；再次，分析了因数据引发的典型事件以及数据面临的主要问题；最后，阐述数据治理的意义和面临的挑战。

1.1　数据治理相关概念

1.1.1　数据、信息、知识

数据（Data）是客观事实经过获取、存储和表达后得到的结果，用于表示客观事物未经加工的原始素材。通常以符号、文本、数字、图形、图像、声音和视频等表现形式存在[1]。

信息（Information）不是具体的事物，也不是某种物质，而是客观事物的一种属性。它是包含上下文语境的数据（Data with Context），必须依附于某个客观事物（媒介）而存在。同一个信息可以借助不同的信息媒介表现出来，如：文字、手势、表情、图形、图像、声音、影视和动画等。

知识（Acknowledge）是对情境的理解、意识、认识和识别，以及对复杂性的把握。知识只有在经过广泛深入的实践检验，被人类消化吸收，并成为个人的信念和判断取向之后才能成为知识[2]。

表 1-1 为数据、信息、知识之间的关系实例。

表 1-1 数据、信息、知识之间的关系实例

<table>
<tr><td colspan="3">姓名：李昊天　　性别：男
年龄：11 岁　　年级：六年级
数学考试成绩：40 分
自述：在 2016 年 9 月末考试中李昊天数学成绩 40 分，满分为 100 分。</td></tr>
<tr><td>数据</td><td>40</td><td>单拿数据 40 来说，在没有上下文语境的情况下，它没有任何意义。</td></tr>
<tr><td>信息</td><td>数学考试成绩：40 分</td><td>这时 40 就有其代表的意义，代表六年级小朋友在一次月末考试中，数学成绩为 40 分。在这样的背景下，40 成为信息中的一个关键指标。</td></tr>
<tr><td>知识</td><td>孩子是不是厌学了，家长该如何解决问题</td><td>总分 100 分，考试成绩 40 分为不及格，可认为孩子这次没有考好。家长可以从以下 3 个方面判断：（1）孩子最近的表现是不是厌学，情绪低落，如果是近期厌学或心情不好引起的成绩下滑，要及时调整；（2）是不是考试当天身体不舒服以至于影响考试成绩，如果是这种情况就没必要太担心，毕竟只是一次考试；（3）孩子以往学习成绩都是在 40 分左右，如果是这种情况，需要家长和学校都给予高度重视，及时帮孩子补习，找准学习方法，培养学习兴趣。
根据多方面观察和分析，找出孩子数学成绩差的主要原因，并及时解决问题。</td></tr>
</table>

通过以上分析可以看出，数据、信息与知识之间既有联系，又有区别。数据是记录下来可以被鉴别的符号。它是最原始的素材，未被加工解释，没有回答特定的问题，没有任何意义；信息是已经被处理、具有逻辑关系的数据，是对数据的解释，这种信息对其接收者具有意义。知识是在对信息进行了筛选、综合、分析等过程之后产生的，它不是信息的简单累加，往往还需要加入基于以往的经验所作的判断。因此，知识可以解决较为复杂的问题，可以回答“如何”的问题，能够积极地指导任务的执行和管理，进行决策和解决问题。

数据是信息的表现形式和载体，数据和信息是不可分离的，数据是信息的表达，信息是数据的内涵。如果没有数据和信息，知识就难以发挥作用。数据和信息的获取则相对简单，而只有知识能够帮助解决问题。

数据信息知识智慧模型（DIKW）将数据、信息、知识纳入到一种金字塔形的层次体系，每一层比下一层都赋予了一些新的特质。原始观察及量度获得了数据，分析数据间的关系获得了信息。在行动上应用信息产生了知识[3]。图 1-1 所示的是数据、信息和知识之间的关系。

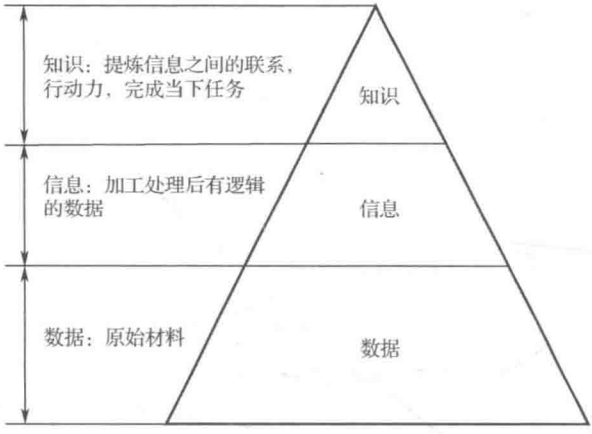

图 1-1　数据、信息和知识之间的关系

1.1.2　数据治理的定义

随着以移动互联网、物联网、社交网络为代表的新型信息发布方式的不断涌现，数据规模不断增长，非结构化的数据所占比重将越来越高，有价值的数据密度会越来越低。应用信息系统建设发展到一定阶段，数据资源将成为战略资产，而有效的数据治理才是数据资产形成的必要条件。由于侧重点和切入视角的不同，目前业界对数据治理的定义尚未形成统一的标准。下面主要介绍 6 个典型数据治理的定义。

1. ITSS 给出的数据治理的定义

ITSS（中国电子工业标准化技术协会信息技术服务分会）指出，数据是资产，通过服务产生价值，数据治理是在数据产生价值的过程中治理团队对其的评价、指导、控制，是数据治理的最基本概念。ITSS 认为数据治理应包含以下 3 个方面内容[4]。

（1）确保信息利益相关者的需要评估，以达成一致的企业目标，这些企业目标需要通过对信息资源的获取和管理实现；

（2）确保有效助力业务的决策机制和方向；

（3）确保绩效和合规进行监督。

ITSS 认为数据治理是一种体系，是一个关注于信息系统执行层面的体系。这一体系的目的是整合 IT 部门与业务部门的知识和意见，通过一个类似于监督委员会或项目小组的虚拟组织对企业的信息化建设进行全方位的监管，这一组织的基础是企业高层的授权和业务部门与 IT 部门的建设性合作。从范围来讲，数据治理涵盖了从前端事务处理系统、后端业务数据库到终端的数据分析，从源头到终端再回到源头形成一个闭环负反馈系统（控制理论中趋稳的系统）。从目的来讲，数据治理就是要对数据的获取、处理、使用进行监管（监管就是我们在执行层面对信息系统的负反

馈），而监管的职能主要通过五个方面的执行力来保证——发现、监督、控制、沟通、整合。

2. DAMA 给出的数据治理的定义

DAMA（国际数据管理协会）在其出版的 DMBOK（The DAMA Guide to the Data Management Body of Knowledge）[5] 中认为数据管理（Data Management，DM）是规划、控制和提供数据及信息资产的一组业务职能，包括开发、执行和监督有关数据的计划、政策、项目、流程、方法和程序，从而控制、保护、交付和提高数据资产的价值。该定义突出了数据管理的职能、过程和规范三个关键词。在职能上认为数据管理是业务数据管理专员和技术数据管理专员共同承担的责任，业务数据管理专员是企业数据资产的托管人，技术数据管理专员是这些资产的专业管理人和监护人。对数据管理职能治理可以协调 IT 和企业之间的协同合作。在过程上，数据管理是数据资产管理的权威性和控制性活动，是在数据管理和使用层面上进行规划、监管和控制。在规范上，数据管理必须遵守相关的规则和规范，才能确保数据管理过程能够顺利进行。

3. ISACA 给出的数据治理的定义

ISACA（国际信息系统审计协会）在其出版物 COBIT（Control Objectives for Information and related Technology）[6] 中认为数据治理是一个由关系和过程所构成的体制，用于指导和控制企业，通过平衡信息技术与过程的风险、增加价值来确保实现企业的目标。数据治理通过评估利益相关者的需求、条件和选择以达成平衡一致的企业目标，通过优先排序和决策机制来设定方向，再根据方向和目标来监测绩效和合规性问题。

COBIT 认为数据治理的关键在于授权和控制并举，应该做出什么决策、谁来做决策、如何做出决策和监督，促成 IT 创造有利于战略的价值。例如，企业要不要上 ERP 一事是要按流程体系办事，经过评估分析其战略价值，按照企业实际状况来决定的，而不是由领导一人决定的。

4. DGI 给出的数据治理的定义

DGI（数据治理研究所）[7]认为数据治理是针对数据信息相关过程的决策权和职责体系，这些过程遵循“在什么时间和什么情况下、用什么方式、由谁、对哪些数据、采取哪些行动”的方法来执行。

5. IBM 给出的数据治理的定义

IBM 认为数据治理是一门将数据视为一项企业资产的学科。它涉及到以企业资产的形式对数据进行优化、保护和利用的决策权利。它涉及到对组织内的人员、流

程、技术和策略的编排，以从企业数据获取最优的价值[8]。从一开始，数据治理就在协调不同的、孤立的且常常冲突的策略（可能导致数据异常）的过程中扮演着重要角色。类似于客户关系管理（CRM）诞生之初，组织开始任命全职或兼职数据治理负责人。与任何新兴学科一样，数据治理有许多定义，但市场已经开始围绕将数据视为资产的定义进行具体化行动。

6．Oracle 给出的数据治理的定义

Oracle 认为数据治理是制订决策权和问责的框架，以规范企业在估值、创建、存储、使用、归档及删除数据和信息的行为。它包括流程、角色、标准和指标，以确保组织能有效和高效地利用数据和信息以实现其目标[9]。

1.2 数据治理理论研究历程

1.2.1 数据治理理论的研究进展

在数据治理理论研究领域，以 ISACA、DAMA、DGI、IBM 等组织为代表，提出了自成体系的数据治理框架或与数据治理相关的治理理论模型。

1．ISACA 提出的 COBIT 5

COBIT 是 ISACA 制订的面向过程的信息系统审计和评价的标准，以支持企业实现其企业 IT 治理和管理的目标。COBIT5 是以五项关键原则为基础进行企业 IT 治理和管理[6]，如图 1-2 所示。这五项原则结合在一起能够使企业构建一种能优化信息和技术投资的、用于利益相关者收益的、有效的治理和管理框架。其中有一条原则，COBIT5 将治理和管理明确区分开来。这两种科目包含不同类型的活动，需要不同的组织结构，并服务于不同的用途。从 COBIT5 的角度来看，治理和管理之间的关键区别在于：治理确保利益相关者的需要、条件和选项得到评估，以决定平衡、协商一致、需要实现的企业目标；通过优先等级和决策来设定导向；并监控商定的导向和目标的绩效以及合规性。在大多数企业中，整体治理是董事长领导下的董事会的责任。具体的治理责任可能授予适当级别的特别组织结构，尤其是在较大型综合性企业中更是如此。管理层计划、构建、运行和监控与治理机构设立导向一致的活动，以实现企业目标。大多数企业中，管理是首席执行官领导下的行政管理层的责任。

COBIT5 以持续改进的生命周期为基础提供了实用和广泛的实施指南。COBIT5 实施生命周期被划分为七个阶段，形成持续改进的生命周期循环，如图 1-3 所示。

图 1-2 COBIT 5 数据治理基本原则

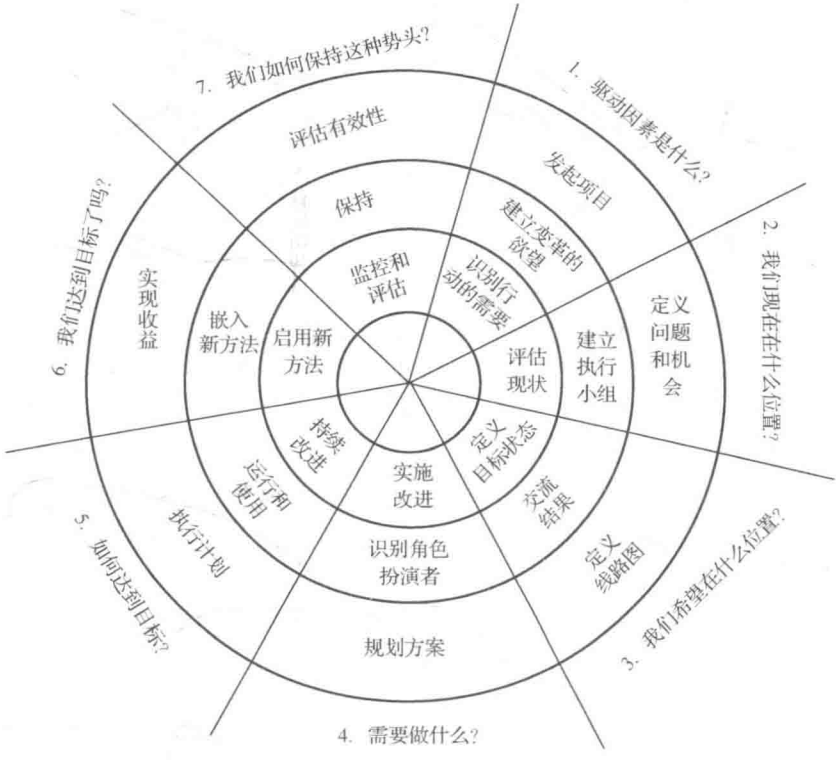

图 1-3 COBIT 实施生命周期的七个阶段

生命周期的三个互为联系的组成部分分别是：（1）生命周期不是一次性项目，而是持续改进的过程（内环）；（2）通过启动变更来解决行为和文化方面的问题（中环）；（3）项目集管理（外环）。